INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN ENGINEERING, VOL. 12, 1597–1615 (1978)

A-POSTERIORI ERROR ESTIMATES FOR THE FINITE ELEMENT METHOD†

I. BABUŠKA‡ AND W. C. RHEINBOLDT§
University of Maryland, Maryland, U.S.A.

SUMMARY

Computable *a-posteriori* error estimates for finite element solutions are derived in an asymptotic form for $h \to 0$ where h measures the size of the elements. The approach has similarity to the residual method but differs from it in the use of norms of negative Sobolev spaces corresponding to the given bilinear (energy) form. For clarity the presentation is restricted to one-dimensional model problems. More specifically, the source, eigenvalue, and parabolic problems are considered involving a linear, self-adjoint operator of the second order. Generalizations to more general one-dimensional problems are straightforward, and the results also extend to higher space dimensions; but this involves some additional considerations. The estimates can be used for a practical *a-posteriori* assessment of the accuracy of a computed finite element solution, and they provide a basis for the design of adaptive finite element solvers.

INTRODUCTION

In recent years very effective procedures have been designed and analysed for the solution of initial value problems for ordinary differential equations. On many test problems (see, e.g. References 1–3) the practical reliability and efficiency of these processes has been ascertained. All of them use adaptive techniques for the selection of the step-size and the order of the numerical method involved. The principal tool for these adaptive approaches is the availability of an error analysis with a local, *a-posteriori* character. It relates the accuracy of the computed solution to a perturbation of the original problem. The adaptive control is then designed to keep the size of this perturbation below a given tolerance.

The indicated error analysis is asymptotic in nature. More specifically, higher-order terms in the step-size h are neglected; that is, asymptotic expressions of the form $1 + o(1)$ as $h \to 0$ are always approximately considered to be equal to 1. Despite this asymptotic character of the principal estimates, practical experience has shown that for reasonable tolerances of the order of 10^{-2} the results obtained with them are excellent and very reliable.

The success of this approach suggests that similar asymptotic analyses may provide also for reliable and effective error controls of the finite element method. This is the topic of the paper. More specifically, we provide here some further details of the general ideas about error estimates for the finite element method developed in References 4–7.

Basically, the present approach is similar to that of the well-known residual method. But the principal difference is our use of the norm of negative Sobolev spaces. In the case of the finite

† The work of the first author was partially supported by the U.S. Energy Research and Development Administration under Contract E(40-1)3443, and that of the second author by the National Science Foundation under Grant MCS 72-03721A06.

‡ Institute for Physical Science and Technology.
§ Computer Science Center.

0029-5981/78/1012-1597$01.00

Received 20 September 1977
Revised 19 December 1977

element method, this leads to error estimates which involve only local rather than global computations. As in the case of the solution procedures for ordinary differential equations, the error estimates are given in an asymptotic form for $h \to 0$ where h is now the size of the elements. However, the asymptotic terms $1+o(1)$ can also be specified precisely and non-asymptotically. It is only for ease of computation that the asymptotic estimates are preferable.

For clarity of presentation, we restrict the discussion to some simple, one-dimensional model problems involving a linear, self-adjoint operator of the second order. The approaches allow for rather straightforward extensions to a variety of more general problems in one dimension. For higher space dimensions some additional considerations have to be taken into account, and we refer to References 5–7. But the general concepts remain the same.

The estimates given here can, of course, be used for an *a-posteriori* assessment of the accuracy of a computed finite element solution, and, as in the case of initial value problems, they may provide a basis for the design of adaptive finite element solvers.

The paper consists of the following parts. Section 'Notation' gives some notational conventions and theoretical preliminaries. For more details on the spaces used and on the formulation of the finite element method in terms of bilinear forms, see, e.g., References 4, 8. Then Section 'A boundary value problem' presents the basic error analysis for a linear, self-adjoint (elliptic), two-point boundary value problem. These ideas are extended to an eigenvalue problem in Section 'An eigenvalue problem' and to a parabolic problem in Section 'A parabolic problem'.

NOTATION

Throughout this paper $I \subset R^1$ is the open unit interval $\{x \in R^1;\ 0<x<1\}$ and $\bar{I}$ its closure. As usual $L_p(I) = H_p^0(I)$, $1<p<\infty$, are the Banach spaces of functions on I with integrable pth power. The norm on these spaces is written as $\|\cdot\|_p$ and the inner product on the Hilbert space $H_2^0(I)$ as $(.,.)_0$.

Let $\mathscr{E}(I)$ be the space of real, infinitely differentiable functions on I for which all derivatives have continuous extensions on $\bar{I}$. Moreover, define $\mathscr{D}(I) \subset \mathscr{E}(I)$ as the subspace of all functions with compact support in I. Then, for any integer $k \geq 1$ and $1<p<\infty$, the Sobolev spaces $H_p^k = H_p^k(I)$ and $\mathring{H}_p^k = \mathring{H}_p^k(I)$ are the completions of $\mathscr{E}(I)$ and $\mathscr{D}(I)$, respectively, under the norm

$$\|u\|_{k,p} = \left[\sum_{0\leq\alpha\leq k} \left\|\frac{d^\alpha u}{dx^\alpha}\right\|_p^p\right]^{1/p} \tag{1}$$

For $p=2$ these are Hilbert spaces, and we denote their inner product by $(.,.)_k$ (see, e.g., Reference 9).

On $\bar{I}$ we consider partitions

$$\Delta : 0 = x_0 < x_1 < \ldots < x_{m-1} < x_m = 1 \tag{2}$$

and introduce the notations

$$\left.\begin{aligned} I_j &= I_j(\Delta) = \{x \in R^1;\ x_j < x < x_{j+1}\} \\ h_j &= h_j(\Delta) = x_{j+1} - x_j \end{aligned}\right\} j = 0, 1, \ldots, m-1$$

$$h = h(\Delta) = \max_{j=0,\ldots,m-1} h_j(\Delta) \tag{3}$$

Let $r \geq 1$ be a given integer and $\mathbf{k} = (k_0, \ldots, k_{m-1})$ a vector of integers corresponding to the m subintervals I_j such that $k_j \geq 2r$, $j = 0, \ldots, m-1$. Then $S^{r,\mathbf{k}} = S^{r,\mathbf{k}}(\Delta)$ is defined as the space of $(r-1)$-times continuously differentiable functions on $\bar{I}$ for which the restriction to I_j, $j = 0, \ldots, m-1$, is a polynomial of degree at most $k_j - 1$.

The structure of these spaces $S^{r,\mathbf{k}}$ is rather simple. In fact, for any $u \in S^{r,\mathbf{k}}$ let u_j be the restriction to I_j, $j = 0, \ldots, m-1$. Then for $k_j = 2r$, the polynomial u_j is defined by its value and that of its $r-1$ first derivatives at the end points of I_j. For $k_j \geq 2r$, u_j is a linear combination of a polynomial for which the value and that of the $r-1$ first derivatives is prescribed at the end points of I_j and $k_j - 2r$ polynomials which together with their $(r-1)$ first derivatives vanish at both these end points.

Clearly we have $S^{r,\mathbf{k}} \subset H_p^r(I)$ for any $1 < p < \infty$. Moreover, it is well-known that the set

$$\bigcup_{\hat{k} \geq 2r} S^{r,\mathbf{k}}(\Delta), \qquad \hat{k} = \min_{j=0,\ldots,m-1} k_j$$

is dense in H_p^r. In other words, for any $u \in H_p^r$ and $\varepsilon > 0$ there exists a $k_0(\varepsilon) \geq 2r$ such that for any $\mathbf{k}$ with $\hat{k} \geq k_0(\varepsilon)$ we have $\|w - g\|_{r,p} \leq \varepsilon$ for some $w \in S^{r,\mathbf{k}}$.

In the following we consider a given family $\mathcal{S}$ of partitions Δ of I. Then for fixed r and $\mathbf{k} = k^* = (k, k, \ldots, k)$, $k \geq 2r$, and any given $u \in H_p^r$ there exists an $h_0(\varepsilon) > 0$ such that for any $\Delta \in \mathcal{S}$ with $h(\Delta) \leq h_0(\varepsilon)$ we have $\|w - u\|_{r,p} \leq \varepsilon$ for some $w \in S^{r,k^*}(\Delta)$.

A BOUNDARY VALUE PROBLEM

Basic formulation

As mentioned in the introduction, we restrict the discussion to some simple model problems. In this chapter we consider the equation

$$L(u) = -\frac{\mathrm{d}}{\mathrm{d}x} a(x) \frac{\mathrm{d}u}{\mathrm{d}x} + b(x)u = f, \qquad \forall x \in I \tag{4}$$

together with the Dirichlet boundary conditions

$$u(0) = u(1) = 0 \tag{5}$$

Here the functions a, $\mathrm{d}a/\mathrm{d}x$, b, and f are assumed to be continuous on $\bar{I}$ and $0 < \alpha_0 \leq a(x) \leq \alpha_1 < \infty$, $0 \leq b(x) < \infty$, $\forall x \in \bar{I}$.

As usual, we associate with (4) the bilinear form

$$B(u, v) = \int_0^1 \left(a \frac{\mathrm{d}u}{\mathrm{d}x} \frac{\mathrm{d}v}{\mathrm{d}x} + buv \right) \mathrm{d}x \tag{6}$$

which is well-defined and continuous on $\mathring{H}_p^1 \times \mathring{H}_q^1$ with $(1/p) + (1/q) = 1$. The weak solution of our problem is then the unique $u_0 \in \mathring{H}_p^1$ with

$$B(u_0, v) = F(v), \qquad \forall v \in \mathring{H}_q^1 \tag{7}$$

where

$$F(v) = \int_0^1 fv \, \mathrm{d}x \tag{8}$$

For $p = q = 2$ the 'energy-norm'

$$\|u\|_{\mathrm{E}} = B(u, u)^{1/2} \tag{9}$$

is equivalent to the norm $\|\cdot\|_{1,2}$ on $\mathring{H}_2^1$; that is, there are constants $c_1, c_2 > 0$ such that

$$c_1\|u\|_{1,2} \leq \|u\|_E \leq c_2\|u\|_{1,2}, \qquad \forall u \in \mathring{H}_2^1 \tag{10}$$

This shows that for $u, v \in \mathring{H}_2^1$

$$\sup_{\|v\|_{1,2}\leq 1} |B(u,v)| \geq c_1\|u\|_1, \qquad \sup_{\|u\|_{1,2}\leq 1} |B(u,v)| \geq c_1\|v\|_{1,2} \tag{11}$$

It is not difficult to show that, more generally, we have for any $1 < p, q < \infty$, $1/p + 1/q = 1$,

$$\begin{aligned} &\sup_{\|v\|_{1,q}\leq 1} |B(u,v)| \geq c(p)\|u\|_{1,p}, \qquad \forall u \in H_p^1 \\ &\sup_{\|u\|_{1,p}\leq 1} |B(u,v)| \geq \bar{c}(p)\|v\|_{1,q}, \qquad \forall v \in H_q^1 \end{aligned} \tag{12}$$

For a given partition Δ and space $S^{r,\mathbf{k}}(\Delta)$, $r \geq 1$, we set

$$\mathring{S}^{r,\mathbf{k}} = \mathring{S}^{r,\mathbf{k}}(\Delta) = \{u \in S^{r,\mathbf{k}}(\Delta);\ u(0) = u(1) = 0\} \tag{13}$$

Of course, this implies that $\mathring{S}^{r,\mathbf{k}} \subset \mathring{H}_p^1$ for any $1 < p < \infty$. Now the finite element solution $\bar{u}_0 \in \mathring{S}^{r,\mathbf{k}}(\Delta)$ is uniquely defined by the condition

$$B(\bar{u}_0, v) = F(v), \qquad \forall v \in \mathring{S}^{r,\mathbf{k}}(\Delta) \tag{14}$$

Clearly then, the error $e = \bar{u}_0 - u_0$ satisfies

$$B(e, v) = 0, \qquad \forall v \in \mathring{S}^{r,\mathbf{k}}(\Delta) \tag{15}$$

Two auxiliary lemmas

In this section we prove some lemmas which play an essential role in the subsequent considerations.

Lemma 1. Let $\mathcal{S}$ be a given family of partitions (2) and $r \geq 1$. Then for any $\Delta \in \mathcal{S}$ and $v \in \mathring{H}_2^1$ we have

$$\inf\{\|v - w\|_E;\ w \in \mathring{S}^{r,\mathbf{k}}(\Delta),\ w(x_j) = v(x_j),\ j = 0, \ldots, m\} \leq C_2^r(\Delta)\|v\|_E \tag{16}$$

The constant $C_2^r(\Delta)$ is independent of v and $\mathbf{k}$ and satisfies

$$C_2^r(\Delta) \leq \bar{C}_2^r(1 + o(h)) \quad \text{as} \quad h(\Delta) \to 0 \tag{17}$$

where

$$\bar{C}_2^r = (1 + Q^2(r))^{1/2} \tag{18}$$

is independent of Δ and more specifically

$$\begin{aligned} 1 + Q^2(r) &= \frac{1}{A(r)^2}\int_{-1/2}^{+1/2}\left(x^2 - \frac{1}{4}\right)^{2(r-1)} dx \\ A(r) &= \int_{-1/2}^{+1/2}(x^2 - \tfrac{1}{4})^{r-1}\, dx \end{aligned} \tag{19}$$

Proof. Since $v \in \mathring{H}_2^1$ is continuous, the conditions $w(x_j) = v(x_j)$ in (16) are well-defined. For given $v \in \mathring{H}_2^1$, let w be the function that satisfies these conditions and is otherwise linear on

each I_j, $j=0,\dots,m-1$. Then $z=v-w\in \mathring{H}_2^1$, and

$$\|v\|_E^2=\|w+z\|_E^2=B(w,w)+2B(w,z)+B(z,z)=\|w\|_E^2+\|z\|_E^2+2\sum_{j=0}^{m-1}\int_{x_j}^{x_{j+1}}\left(a\frac{dw}{dx}\frac{dz}{dx}+bwz\right)dx \tag{20}$$

With

$$a_{j+1/2}=a\left(\tfrac{1}{2}(x_j+x_{j+1})\right)$$

we obtain

$$\int_{x_j}^{x_{j+1}} a\frac{dz}{dx}\frac{dw}{dx}\,dx=a_{j+1/2}\int_{x_j}^{x_{j+1}}\frac{dz}{dx}\frac{dw}{dx}\,dx+\int_{x_j}^{x_{j+1}}(a-a_{j+1/2})\frac{dz}{dx}\frac{dw}{dx}\,dx$$

Since $z(x_j)=z(x_{j+1})=0$ and w is linear on I_j, the first term on the right is zero. In order to estimate the second term note that

$$|a(x)-a_{j+1/2}|\leq \operatorname{con} h_j,\qquad \forall x\in I_j \tag{21}$$

Here, and in subsequent inequalities 'con' denotes a generic constant with generally different values in each instance. From (21) we find that

$$\left|\int_{x_j}^{x_{j+1}}(a-a_{j+1/2})\frac{dz}{dx}\frac{dw}{dx}\,dx\right|\leq \operatorname{con} h_j\left(\int_{x_j}^{x_{j+1}}\left(\frac{dz}{dx}\right)^2dx\right)^{1/2}\left(\int_{x_j}^{x_{j+1}}\left(\frac{dw}{dx}\right)^2dx\right)^{1/2}$$

$$\leq \operatorname{con} h(\Delta)\int_{x_j}^{x_{j+1}}\left[\left(\frac{dz}{dx}\right)^2+\left(\frac{dw}{dx}\right)^2\right]dx$$

whence, because of the equivalence of the norms $\|\cdot\|_{1,2}$ and $\|\cdot\|_E$ on $\mathring{H}_2^1$,

$$\left|\sum_{j=0}^{m-1}\int_{x_j}^{x_{j+1}} a\frac{dz}{dx}\frac{dw}{dx}\,dx\right|\leq \operatorname{con} h(\Delta)(\|w\|_E^2+\|z\|_E^2) \tag{22}$$

It is also easily verified that

$$\int_{x_j}^{x_{j+1}} z^2\,dx\leq \operatorname{con} h^2\int_{x_j}^{x_{j+1}}\left(\frac{dz}{dx}\right)^2dz$$

which leads in a corresponding way to

$$\left|\sum_{j=0}^{m-1}\int_{x_j}^{x_{j+1}} bwz\,dx\right|\leq \operatorname{con} h(\|w\|_E^2+\|z\|_E^2) \tag{23}$$

Together (20), (22), (23) show that

$$\|v\|_E^2=(\|w\|_E^2+\|z\|_E^2)(1+O(h))\quad\text{as}\quad h(\Delta)\to 0 \tag{24}$$

In the case $r=1$, we have $w\in S^{r,\mathbf{k}}(\Delta)$ and thus

$$\|v-w\|_E^2=\|z\|_E^2\leq(\|w\|_E^2+\|z\|_E^2)=\|v\|_E^2(1+O(h))$$

which directly implies (16).

For $r>1$ we introduce the function W which reduces on each I_j to a polynomial of degree $2r-1$ with

$$W(x_j)=c(x_j)=w(x_j),\ \frac{d^\alpha W}{dx^\alpha}(x_j)=0,\ j=0,\dots,m,\qquad \alpha=1,\dots,r-1$$

$$\begin{aligned}\|w-W\|_E^2 &= \sum_{j=0}^{m-1} a_{j+1/2}\int_{x_j}^{x_{j+1}}\left(\frac{d}{dx}(w-W)\right)^2 dx\\ &\quad+\sum_{j=0}^{m-1}\int_{x_j}^{x_{j+1}}\left[(a-a_{j+1/2})\left(\frac{d}{dx}(w-W)\right)^2+b(w-W)^2\,dx\right]\\ &\leqslant \sum_{j=0}^{m-1}[a_{j+1/2}+O(h)]\int_{x_j}^{x_{j+1}}\left(\frac{d}{dx}(w-W)\right)^2 dx \quad \text{as} \quad h_j(\Delta)\to 0\end{aligned}\tag{25}$$

The conditions on $w-W$ imply that

$$\frac{d(w-W)}{dx}(x)=B(q(x)^{r-1}-A),\qquad x_j\leqslant x\leqslant x_{j+1}$$

where

$$q(x)=(x-x_j)(x-x_{j+1})$$

and

$$A=\frac{1}{h_j}\int_{x_j}^{x_{j+1}} q(x)^{r-1}\,dx,\qquad B^2=\frac{1}{A^2h_j}\int_{x_j}^{x_{j+1}}\left(\frac{dw}{dx}\right)^2 dx$$

A simple substitution gives

$$\begin{aligned}A=h^{2(r-1)}\int_{-1/2}^{1/2}(x^2-\tfrac14)^{r-1}\,dx &= A(r)h^{2(r-1)}\frac{1}{h_jA^2}\int_{x_j}^{x_{j+1}}[q(x)^{r-1}-A]^2\,dx\\ &=\frac{1}{A(r)^2}\int_{-1/2}^{1/2}\left[\left(x^2-\frac14\right)^{r-1}-A(r)\right]^2 dx=Q(r)^2\end{aligned}$$

whence

$$\int_{x_j}^{x_{j+1}}\left[\frac{d}{dx}(w-W)\right]^2 dx=Q^2(r)\int_{x_j}^{x_{j+1}}\left(\frac{dw}{dx}\right)^2 dx$$

and thus by (25)

$$\|w-W\|_E\leqslant(1+O(h))Q(r)\|w\|_E$$

But then by (24)

$$\begin{aligned}\|v-W\|_E &\leqslant \|z\|_E+\|w-W\|_E\\ &\leqslant(\|z\|_E+Q(r)\|w\|_E)(1+O(h))\\ &\leqslant(\|z\|_E^2+\|w\|_E^2)^{1/2}(1+Q(r)^2)^{1/2}(1+O(h))\\ &\leqslant\|v\|_E^2(1+Q(r)^2)^{1/2}(1+O(h))\quad\text{as}\quad h(\Delta)\to 0\end{aligned}$$

completes the proof of the lemma.

The following Table I shows some values of the constant $\bar C_2^r=(1+Q(r)^2)^{1/2}$:

Table I

r	$\bar{C}_2^r$
1	1
2	1·0954
3	1·1952
4	1·2774

Note that these constants are not large. But since they increase with r, it will be advantageous to keep r small.

The results extends also to the case $p \neq 2$; but then the norm of $\mathring{H}_p^1$ has to be used in (16), instead of the energy norm.

Lemma 2. Let $\mathscr{S}$ be a given family of partitions and $r \geq 1$, $1 < p < \infty$. Then for any $\Delta \in \mathscr{S}$ and $v \in \mathring{H}_p^1$ we have

$$\{\inf \|v - w\|_{1,p};\ w \in \mathring{S}^{r,\mathbf{k}}(\Delta),\ w(x_j) = v(x_j),\ j = 0, \ldots, m\} \leq C_p^r(\Delta)\|v\|_{1,p} \tag{26}$$

with

$$C_p^r(\Delta) = \bar{C}_p^r(1 + o(1)) \quad \text{as} \quad h(\Delta) \to 0$$

The proof proceeds analogously as that of Lemma 1 and we shall not go into details here. The derivation of the constants is now more complicated than before.

A-posterior error estimates in the energy norm

We apply now the results of the previous section to an *a-posteriori* error analysis based on the energy norm for finite element solutions of the problem (4)–(5).

Let P be the orthogonal projection of $\mathring{H}_2^1$ onto the subspace

$$\{u \in \mathring{H}_2^1;\ u(x_j) = 0, j = 0, 1, \ldots, m\} \tag{27}$$

with respect to the scalar product defined by $B(u, v)$. As in subsection 'Basic formulation', let $e = u_0 - \bar{u}_0$ be the finite element error; then by (15) we have

$$\|e\|_E^2 = |B(e, e)| = |B(e, e - v)|$$
$$= |B(Pe, e - v)| \leq \|Pe\|_E \|e - v\|_E, \qquad \forall v \in \mathring{S}^{r,\mathbf{k}}(\Delta), \qquad v(x_j) = e(x_j), \qquad j = 0, \ldots, m$$

Now Lemma 1 implies that

$$\|e\|_E^2 \leq \|Pe\|_E C_2^r(\Delta)\|e\|_E;$$

that is,

$$\|e\|_E \leq C_2^r(\Delta)\|Pe\|_E \tag{28}$$

In addition, the lower bound

$$\|e\|_E \geq \|Pe\|_E \tag{29}$$

follows directly from the definition of P.

A comparison with Table I shows that the error interval given by (28)–(29) is relatively small. The question is only how to compute $\|Pe\|_E$.

It is readily seen that on any subinterval I_j, $j = 0, \ldots, m-1$, the function $\kappa(\cdot) = Pe$ satisfies the differential equation

$$L(\kappa) = L(\bar{u}_0) - f \tag{30}$$

and the boundary conditions

$$\kappa(x_j) = \kappa(x_{j+1}) = 0 \tag{31}$$

Let u_j denote the exact solution of (4) on I_j with the boundary conditions

$$u_j(x_j) = \bar{u}_0(x_j),\ u_j(x_{j+1}) = \bar{u}_0(x_{j+1}) \tag{32}$$

then clearly we have $\kappa(\cdot) = \bar{u}_0 - u_j$ on I_j. In other words, κ can be computed locally on I_j by solving (4) on that interval with the boundary conditions (32).

Obviously, it would be undesirably expensive to compute κ rather accurately on each I_j, and we are therefore interested in some simple estimate of this function. For this, recall that in (4) the functions a, da/dx, and b were assumed to be continuous on $\bar{I}$. Moreover, $\bar{u}_0$ is on I_j a polynomial of degree at most $k_j - 1$. Thus, the residual

$$r_j(x) = L(\kappa) = L(\bar{u}_0) - f, \qquad x \in I_j, \qquad j = 0, 1, \ldots, m-1 \tag{33}$$

on I_j is for small $h(\Delta)$ close to a polynomial of degree at most $k_j - 1$. The quantities

$$\mu_j^2 = \int_{x_j}^{x_{j+1}} r_j(x)^2 \, dx, \qquad j = 0, 1, \ldots, m-1 \tag{34}$$

are readily computable using, for instance, Gaussian quadrature—as in the finite element method itself.

The smallest eigenvalue of the differential operator L on $\bar{I}_j$ with zero boundary conditions at x_j and x_{j+1} is bounded below by the smallest eigenvalue $a_{\min}\pi^2/h_j^2$ of the operator $a_{\min}\, d^2/dx^2$ on the same interval, where

$$a_{\min} = \min_{x_j \le x \le x_{j+1}} |a(x)| = a_{j+1/2}(1 + O(h)) \quad \text{as} \quad h_j \to 0$$

and $a_{j+1/2}$ is defined as before. Thus from (33) it follows that

$$\left(\int_{x_j}^{x_{j+1}} \kappa(x)^2 \, dx \right)^{1/2} \le \frac{h_j^2}{\pi^2 a_{\min}} \left(\int_{x_j}^{x_{j+1}} r_j(x)^2 \, dx \right)^{1/2} \tag{35}$$

Therefore, using the Schwarz inequality together with (34) we obtain

$$\int_{x_j}^{x_{j+1}} \left(a\left(\frac{d\kappa}{dx}\right)^2 + b\kappa^2 \right) dx = \int_{x_j}^{x_{j+1}} r_j \kappa \, dx \le \frac{h_j^2}{\pi^2 a_{j+1/2}} \mu_j^2 (1 + O(h)) \quad \text{as} \quad h_j(\Delta) \to 0 \tag{36}$$

The computable quantities

$$\varepsilon_j^2 = \frac{h_j^2 \mu_j^2}{\pi^2 a_{j+1/2}}, \qquad j = 1, \ldots, m-1 \tag{37}$$

are called the error indicators of the approximate solution $\bar{u}_0$. From (16) and (28) it follows that

$$\|e\|_E \le \bar{C}_2^r \left(\sum_{j=0}^{m-1} \varepsilon_j^2 \right)^{1/2} (1 + O(h)) \quad \text{as} \quad h(\Delta) \to 0 \tag{38}$$

where $\bar{C}_2^r$ is given by (18/19). This estimate may be improved somewhat. For example, in Reference 7 we show that for linear elements the constant $1/\pi^2$ in (37) may be replaced by $1/12$.

The estimate (38) may be used in various ways. It provides us with an easily computable asymptotic bound for the error under the energy norm. At the same time the error indicators ε_j of (37) provide a measure of the error contribution of each interval. Note that the calculation of ε_j consists principally of the evaluation of the L_2-norm μ_j of the residual r_j on $\bar{I}_j$. For this a suitable quadrature formula is all that is needed. The error indicators can also be used to characterize optimal meshes. More specifically, we show in Reference 7 that a mesh is asymptotically optimal if all error indicators are equal, that is, for any other partition with sufficiently small $h(\Delta)$ the error is larger. These results are proved for linear elements but the approaches allow for rather straightforward extensions to more general one-dimensional problems as well as to higher-order elements. In Reference 7 it is also shown that the value of the optimal error is rather stable under perturbations of the optimal mesh. Hence it is unnecessary to compute this mesh with excessive accuracy. This suggests the use of an adaptive mesh-refinement algorithm of the type presented in Reference 5. We refer to that paper for further details.

A-posteriori error estimates in the $\|\cdot\|_{1,p}$ norm

Sometimes in practice the energy norm is note entirely appropriate. In this section we show that corresponding error estimates can also be obtained in the $\|\cdot\|_{1,p}$ norm. Especially for large p these norms provide some desirable information. Nevertheless, we shall sketch here only the derivation of the estimates and not enter into a discussion of the specific constants occurring in them.

By (12) we have

$$\|e\|_{1,p} \leq \frac{1}{c(p)} \sup \{|B(e, v)|; \|v\|_{1,q} \leq 1, v \in \mathring{H}_q^1\}$$

and, as before, with (15) and Lemma 2, it follows that

$$\begin{aligned} |B(e, v)| &= |B(e, v-w)| = |B(Pe, v-w)| \\ &\leq \bar{C}(p)\|Pe\|_{1,p}\|v-w\|_{1,q} \\ &\leq \bar{C}(p)C_q^r(\Delta)\|Pe\|_{1,p}\|v\|_{1,q}, \qquad \forall v \in \mathring{H}_q^1,\ w \in S^{r,\mathbf{k}} \qquad w(x_j) = v(x_j), \qquad j = 0, \ldots, m \end{aligned}$$

where $\bar{C}(p)$ is the constant of continuity of B on $\mathring{H}_p^1 \times \mathring{H}_q^1$. Thus together we obtain

$$\|e\|_{1,p} \leq \frac{\bar{C}(p)}{c(p)} \bar{C}_q^r \|Pe\|_{1,p}(1 + o(1)) \quad \text{as} \quad h(\Delta) \to 0 \tag{39}$$

As in the previous section we define the residuals $r_j(x)$ on I_j by (33) and, in this case, the quantities μ_j by

$$\mu_j = \left(\int_{x_j}^{x_{j+1}} |r_j(x)|^p \, dx \right)^{1/p}, \qquad j = 0, 1, \ldots, m-1 \tag{40}$$

then one can show that

$$\|e\|_{1,p} \leq \text{con} \left(\sum_{j=0}^{m-1} |\mu_j|^p h_j(\Delta)^p \right)^{1/p} (1 + o(1)) \quad \text{as} \quad h(\Delta) \to 0 \tag{41}$$

Hence, essentially the same procedure provides us with asymptotic error estimates under different norms.

AN EIGENVALUE PROBLEM

Problem formulation

Analogous to section 'A boundary value problem' we restrict ourselves to the model problem

$$\begin{aligned} Lu &\equiv -\frac{\mathrm{d}}{\mathrm{d}x}a(x)\frac{\mathrm{d}u}{\mathrm{d}x}+b(x)u=\lambda u \quad \text{on} \quad I \\ u(0)&=u(1)=0 \end{aligned} \tag{42}$$

where a and b satisfy the same conditions as before. This is a self-adjoint eigenvalue problem with eigenvalues $0<\lambda_1\leqslant\lambda_2\leqslant\lambda_3\ldots$. Let $\varphi_1, \varphi_2, \varphi_3, \ldots$ be the corresponding (normalized) eigenfunctions. Then $\varphi_i \in \mathring{H}_2^1$ and the φ_i form a complete, orthonormal sequence in $\mathring{H}_2^1$. Hence for any $u \in \mathring{H}_2^1$ we have

$$\begin{aligned} u &= \sum_{j=1}^{\infty} a_j\varphi_j \\ \|u\|_0^2 &= \sum_{j=1}^{\infty} a_j^2,\ \|u\|_{\mathrm{E}}^2 = \sum_{j=1}^{\infty} a_j^2\lambda_j \end{aligned} \tag{43}$$

Moreover, with B and F defined by (6) it follows that

$$B(\varphi_j, v)=\lambda_j F_j(v),\ \forall v \in \mathring{H}_2^1 \tag{44}$$

where

$$F_j(v)=\int_0^1 \varphi_j v\ \mathrm{d}x \tag{45}$$

For a given partition Δ and space $S^{r,\mathbf{k}}(\Delta)$ let $\mathring{S}^{r,\mathbf{k}}(\Delta)$ again be given by (13). Then the finite element solution $\bar{\lambda}_j, \bar{u}_j \in \mathring{S}^{r,\mathbf{k}}(\Delta)$, $\|\bar{u}_j\|_{0,2}=1$ of the eigenvalue problem is uniquely defined by

$$B(\bar{u}_j, v)=\bar{\lambda}_j\bar{F}_j(v), \qquad \forall v \in \mathring{S}^{r,\mathbf{k}}, \qquad \bar{F}_j(v)=\int_0^1 \bar{u}_j v\ \mathrm{d}x \tag{46}$$

This represents a finite-dimensional, generalized eigenproblem. The convergence of the process is well understood. In particular, in our case it follows that $\bar{\lambda}_j \to \lambda_j$ and $\|\bar{u}_j-\varphi_j\|_{\mathrm{E}}\to 0$ as $h(\Delta)\to 0$ (see, e.g. Reference 8).

A-posteriori error analysis–I

Once a finite element solution $\bar{\lambda}_j, \bar{u}_j$ of (46) has been computed, we may interpret $\bar{u}_j$ as a finite element solution of the boundary value problem (4) with $f=\bar{\lambda}_j\bar{u}_j$. Let $w_j \in \mathring{H}_2^1$ be the corresponding exact solution; that is,

$$B(w_j, v)=\bar{\lambda}_j\bar{F}_j(v), \qquad \forall v \in \mathring{H}_2^1 \tag{47}$$

Then the results of section 'A boundary value problem' show that $\eta_j=\bar{u}_j-w_j$ satisfies

$$\|\eta_j\|_{\mathrm{E}} \leqslant \theta_j(1+\mathrm{O}(h)) \quad \text{as} \quad h(\Delta)\to 0 \tag{48}$$

where the constant θ_j is computed as in section 'A boundary value problem'.

We have now the relations

$$B(\bar{u}_j, v) = \bar{\lambda}_j \bar{F}_j(v) + B(\eta_j, v), \qquad \forall v \in \mathring{H}_2^1 \tag{49}$$

and, of course,

$$B(\varphi_j, v) = \lambda_j F_j(v), \qquad v \in \mathring{H}_2^1,$$

which, together, imply for the error $e_j = \bar{u}_j - \varphi_j$ that

$$B(e_j, v) - \lambda_j \int_0^1 e_j v \, dx = (\bar{\lambda}_j - \lambda_j) \int_0^1 \bar{u}_j v \, dx + B(\eta_j, v), \qquad \forall v \in \mathring{H}_2^1$$

Because $B(u, v) = B(v, u)$, the left side is zero for $v = \varphi_j$ and therefore

$$(\bar{\lambda}_j - \lambda_j) \int_0^1 \bar{u}_j \varphi_j \, dx + B(\eta_j, \varphi_j) = 0 \tag{50}$$

Since $\bar{u}_j \to \varphi_j$ as $h(\Delta) \to 0$, it follows that

$$|\bar{\lambda}_j - \lambda_j| \leq \theta_j \|\varphi_j\|_E (1 + O(h)) \quad \text{as} \quad h(\Delta) \to 0 \tag{51}$$

Because of (43) we have

$$\|\varphi_j\|_E = \lambda_j^{1/2}$$

and thus (50) can be written as

$$|\bar{\lambda}_j - \lambda_j| \leq \theta_j \lambda_j^{1/2} (1 + O(h)), \quad \text{as} \quad h(\Delta) \to 0 \tag{52}$$

It should be noted that this estimate was obtained without any restrictions upon the distribution of the spectrum.

A-posteriori error analysis-II

In this section we consider the case of a simple, well-separated eigenvalue. More specifically, an eigenvalue λ_k of (42) is α_k-separated if

$$\inf_{\forall j, j \neq k} \left| 1 - \frac{\lambda_k}{\lambda_j} \right| = \alpha_k > 0 \tag{53}$$

and we call λ_k well-separated if α_k is not too close to zero.

In the notation of the previous section, let

$$\bar{u}_k = \sum_{j=1}^{\infty} a_j^{(k)} \varphi_j, \; \eta_k = \sum_{j=1}^{\infty} c_j^{(k)} \varphi_j, \qquad k = 1, 2, \ldots, \hat{k}(\Delta) \tag{54}$$

Then we have for any fixed $k = 1, \ldots, \hat{k}(\Delta)$,

$$\sum_{j=1}^{\infty} (a_j^{(k)})^2 = 1 \tag{55}$$

and

$$\|\bar{u}_k\|_E^2 = \sum_{j=1}^{\infty} (a_j^{(k)})^2 \lambda_j = \bar{\lambda}_k \tag{56}$$

Moreover, (48) leads to

$$\|\eta_k\|_E^2 = \sum_{j=1}^{\infty} (c_j^{(k)})^2 \lambda_j \leq \theta_k^2 (1 + O(h)) \quad \text{as} \quad h(\Delta) \to 0 \tag{57}$$

and (49) has the form

$$\sum_{j=1}^{\infty} a_j^{(k)} b_j \lambda_j = \bar{\lambda}_k \sum_{j=1}^{\infty} a_j^{(k)} b_j + \sum_{j=1}^{\infty} \lambda_j c_j^{(k)} b_j$$

where $\{b_i\}$ is any sequence such that $\sum_{j=1}^{\infty} b_j^2 \lambda_j < \infty$. This implies that

$$c_j^{(k)} = a_j^{(k)} \alpha_{jk}, \ \alpha_{jk} = \alpha_{jk}(\Delta) = 1 - \frac{\bar{\lambda}_k}{\lambda_j}, \qquad \forall j, k \tag{58}$$

Now consider the quantity

$$\xi_k^2 = \sum_{\substack{j=1 \\ j \neq k}}^{\infty} (a_j^{(k)})^2 \lambda_j$$

By (58), (53), and (56) we have

$$\xi_k^2 = \sum_{\substack{j=1 \\ j \neq k}}^{\infty} \left(\frac{c_j^{(k)}}{\alpha_{jk}} \right)^2 \lambda_j \leq \frac{1}{\alpha_k^2} \sum_{j=1}^{\infty} (c_j^{(k)})^2 \lambda_j (1 + o(1)) \leq \frac{1}{\alpha_k^2} \theta_k^2 (1 + o(1)) \quad \text{as} \quad h(\Delta) \to 0 \tag{59}$$

On the other hand, from (55), (56) and the known fact that $\bar{\lambda}_k \geq \lambda_k$ it follows that

$$1 \geq (a_k^{(k)})^2 = \frac{1}{\lambda_k} (\bar{\lambda}_k - \xi_k^2) \geq 1 - \frac{\xi_k^2}{\lambda_k} + \frac{\bar{\lambda}_k - \lambda_k}{\lambda_k} \geq 1 - \frac{\xi_k^2}{\lambda_k}$$

and therefore, because $\xi_k \to 0$ as $h(\Delta) \to 0$,

$$\begin{aligned} \|\varphi_k - \bar{u}_k\|_E^2 &= \xi_k^2 + \lambda_k (1 - a_k^{(k)})^2 \\ &\leq \xi_k^2 + \lambda_k \left[1 - \left(1 - \frac{\xi_k^2}{\lambda_k} \right)^{1/2} \right]^2 \\ &\leq \xi_k^2 \left[1 + \frac{1}{4} \lambda_k \left(\frac{\xi_k}{\lambda_k} \right)^2 + o(1) \right] = \xi_k^2 (1 + o(1)) \quad \text{as} \quad h(\Delta) \to 0 \end{aligned} \tag{60}$$

Together with (59) this gives the estimate

$$\|\varphi_k - \bar{u}_k\|_E \leq \frac{1}{\alpha_k} \theta_k (1 + o(1)) \quad \text{as} \quad h(\Delta) \to 0 \tag{61}$$

By (46) and (47) we have

$$B(\eta_k, v) = 0, \qquad \forall v \in S^{r,k}(\Delta) \tag{62}$$

and hence it follows from (50) that

$$0 = (\bar{\lambda}_k - \lambda_k) \int_0^1 \bar{u}_k \varphi_k \, dx + B(\eta_k, \varphi_k - \bar{u}_k)$$

which by (57) and (61) leads to

$$|\bar{\lambda}_k - \lambda_k| \leq \frac{1}{\alpha_k} \theta_k^2 (1 + o(1)) \quad \text{as} \quad h(\Delta) \to 0 \tag{63}$$

Evidently, α_k cannot be computed directly. But in practice, we may obtain the quantity

$$\inf_{\forall j, j \neq k} \left| 1 - \frac{\bar{\lambda}_k}{\bar{\lambda}_j} \right| = \bar{\alpha}_k$$

and, because $\alpha_k = \bar{\alpha}_k(1+o(1))$, the estimates (61) and (63) also hold with $\bar{\alpha}_k$ in place of α_k.

A-posteriori error estimates–III

The results of the previous section may be extended by introducing some measure of the distribution of the eigenvalues around λ_k. Suppose again that all eigenvalues are simple. For given integer k and tolerance $\tau > 0$ we introduce the index set

$$J = J(k, \tau) = \left\{ i \,\middle|\, \left| 1 - \frac{\lambda_k}{\lambda_i} \right| \geq (1+\tau)^{-1} \right\} \tag{64}$$

clearly then the complement of J is a finite set. Analogously to (59) we obtain

$$\xi_J^2 = \sum_{j \in J} (a_j^{(k)})^2 \lambda_j \leq (1+\tau)^2 \theta_k^2 (1+o(1)) \quad \text{as} \quad h(\Delta) \to 0 \tag{65}$$

By (62) we have

$$\sum_{j=1}^{\infty} \lambda_j c_j^{(k)} a_j^{(i)} = 0, \qquad \forall i, k$$

and thus, because of $a_k^{(k)} = 1 + o(1)$ as $h \to 0$, for all $i \neq k$

$$|\lambda_i c_i^{(k)}| = \left| \sum_{\substack{j=1 \\ j \neq i}}^{\infty} \lambda_i c_j^{(k)} a_j^{(i)} \right| (1+o(1)) \quad \text{as} \quad h(\Delta) \to 0 \tag{66}$$

Hence, using (66) and (58) we find that

$$\chi_J^2 = \sum_{\substack{i \notin J \\ i \neq k}} (a_i^{(k)})^2 \lambda_i = \sum_{\substack{i \notin J \\ i \neq k}} (c_i^{(k)} \lambda_i)^2 \frac{\lambda_i}{(\lambda_i - \bar{\lambda}_k)^2}$$

$$\leq \left[\sum_{\substack{i \notin J \\ i \neq k}} \left(\sum_{\substack{j=1 \\ j \neq i}}^{\infty} \lambda_j c_j^{(k)} a_j^{(i)} \right)^2 \frac{1}{\lambda_i \alpha_k^2} \right] (1+o(1)) \quad \text{as} \quad h(\Delta) \to 0 \tag{67}$$

Now by the Schwarz inequality and (57) and (59) it follows that

$$\left(\sum_{\substack{j=1 \\ j \neq i}}^{\infty} \lambda_j c_j^{(k)} a_j^{(i)} \right)^2 \leq \left(\sum_{\substack{j=1 \\ j \neq i}}^{\infty} (c_j^{(k)})^2 \lambda_j \right) \left(\sum_{\substack{j=1 \\ j \neq i}}^{\infty} (a_j^{(i)})^2 \lambda_j \right)$$

$$\leq \theta_k^2 \frac{\theta_i^2}{\alpha_i^2} (1+o(1)) \quad \text{as} \quad h(\Delta) \to 0 \tag{68}$$

Because the complement of J is a finite set (independent of the partition) and $\theta_i^2 \to 0$ as $h(\Delta) \to 0$, (67) and (68) together imply that

$$\chi_J^2 \leq \theta_k^2 \frac{1}{\lambda_1 \alpha_k^2} \sum_{\substack{i \notin J \\ i \neq k}} \frac{\theta_i^2}{\alpha_i^2} (1+o(1)) = \theta_k^2 o(1) \quad \text{as} \quad h(\Delta) \to 0 \tag{69}$$

Thus by (65) and (69) the quantity ξ_k^2 of (59) satisfies

$$\xi_k^2 = \xi_J^2 + \chi_J^2 \leq (1+\tau)^2 \theta_k^2 (1+O(1)) \quad \text{as} \quad h(\Delta) \to 0 \tag{70}$$

and as before we obtain the estimate

$$|\lambda_k - \bar{\lambda}_k| \leq (1+\tau) \theta_k^2 (1+O(1)) \quad \text{as} \quad h(\Delta) \to 0 \tag{71}$$

which is the same as (63) except that $1/\alpha_k$ is replaced by $(1+\tau)$. Of course, the asymptotic behaviour of the right side of (71) depends on the set J, that is, on k and the choice of τ. It is possible to compute ξ_J^2 and χ_J^2 numerically and hence to obtain a more specific estimate (71). We mention also that the assumption of the simplicity of the eigenvalues is not essential and may be removed at the expense of slightly more difficult arguments.

An example

Consider the problem (42) with $a \equiv 1$, $b = 0$, for which it is well known that $\lambda_k = k^2\pi^2$, $k = 1, 2, \ldots$. We use a uniform partition with $h_j \equiv h(\Delta) = 1/n$ and choose $r = 1$, $k = 2$, $\mathbf{k} = k^*$. Then the corresponding finite element formulation is the finite-dimensional generalized eigenvalue problem

$$Ax = h^2\bar{\lambda}Bx, \qquad x \in R^{n-1}, \qquad x^{\mathrm{T}}Bx = 1 \tag{72a}$$

where

$$A = \begin{bmatrix} 2 & -1 & & & 0 \\ -1 & 2 & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & -1 \\ 0 & & & -1 & 2 \end{bmatrix}, \qquad B = \tfrac{1}{6}\begin{bmatrix} 4 & 1 & & & 0 \\ 1 & 4 & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & 1 \\ 0 & & & 1 & 4 \end{bmatrix} \tag{72b}$$

It is easily verified that the eigenvectors and eigenvalues of (72) are

$$\begin{aligned} &x^k = (x_{k,1}, \ldots, x_{k,n-1})^{\mathrm{T}}, \qquad x_{k,j} = c_k \sin\frac{k\pi}{n}j, \qquad j, k = 1, \ldots, n-1 \\ &\bar{\lambda}_k = c_k^2 n^2\left(1 - \cos\frac{\pi k}{n}\right), \qquad k = 1, \ldots, n-1 \end{aligned} \tag{73}$$

where

$$c_k = \left(\frac{3}{1 + \frac{1}{2}\cos k\pi/n}\right)^{1/2}, \tag{74}$$

From this it follows that asymptotically

$$\bar{\lambda}_k = \lambda_k + \frac{1}{12}\frac{[\pi^2k^2]^2}{n^2} + O(n^{-4}) \quad \text{as} \quad n \to \infty \tag{75}$$

On the other hand, for the quantities (34) we obtain after a simple computation that

$$\sum_{j=0}^{n-1} (\mu_j^{(k)})^2 = \lambda_k^2(1 + O(n^{-2})) \quad \text{as} \quad n \to \infty$$

which by (38) and (48) leads to the constants

$$\theta_k^2 = \frac{1}{\pi^2}(\bar{C}_2^1)^2 \sum_{j=0}^{n-1} (\mu_j^{(k)})^2 h_j^2 = \frac{1}{\pi^2}\frac{(\pi^2k^2)^2}{n^2} + O(n^{-4}) \quad \text{as} \quad n \to \infty \tag{76}$$

in the estimate (71). Thus comparing (71) and (75) we see that the estimated bound is by a factor $(1+\tau)12/\pi^2 \leq 1{\cdot}22(1+\tau)$ larger than the exact value. Note that this factor is independent of k.

Set $\varepsilon_k = \bar{\lambda}_k - \lambda_k$ and $\omega_k = \varepsilon_k/\lambda_k$ and define

$$\kappa_k = \frac{\theta_k^2}{\varepsilon_k}$$

as efficiency coefficient for the error estimate. Table II below shows κ_k and ω_k for different values of n.

Table II

n	κ_1	ω_1		
2	1·71	2·15 (−1)		
3	1·411	9·42 (−2)		
5	1·282	3·33 (−2)		
10	1·231	8·25 (−3)		
n	κ_4	ω_4	κ_9	ω_9
5	3·008	4·42 (−1)		
10	1·509	1·31 (−1)	3·986	3·96 (−1)
30	1·244	1·47 (−2)	1·371	7·60 (−2)
50	1·226	5·27 (−3)	1·269	2·69 (−2)

Clearly the efficiency of the estimates does not change too much with the accuracy. Moreover, for a reasonable relative error of, say, 10 per cent the efficiency coefficient is practically constant and close to the limiting factor. Note that the efficiency coefficients tend to $12/\pi^2$. As noted earlier in connection with (37), for linear elements the factor $1/\pi^2$ in the error indicators may be replaced by $1/12$ (see Reference 7), in which case the efficiency indeed tends to one as should be expected.

The above example is, of course, a special case. In general, the error indicators

$$\varepsilon_j^{(k)} = \frac{\mu_j^{(k)} h_j}{\pi \sqrt{a_{j+1/2}}}, \qquad j = 0, 1, \ldots, n-1$$

have to be computed and then θ_k^2 is asymptotically equal to

$$(\bar{C}_2^r)^2 \sum_{j=0}^{n-1} (\varepsilon_j^{(k)})^2$$

A PARABOLIC PROBLEM

Problem formulation

Once again we consider a simple model problem, namely,

$$\frac{\partial u}{\partial t} - \frac{\partial}{\partial x} a(x) \frac{\partial u}{\partial x} + b(x) u = f(x, t), \qquad \forall x \in I, \qquad t > 0 \tag{77}$$

with the boundary conditions

$$u(0, x) = 0, \qquad \forall x \in \bar{I}, \qquad u(t, 0) = u(t, 1) = 0, \qquad \forall t \geq 0 \tag{78}$$

As before, a, da/dx, b, and f are assumed to be continuous. Then the solution of (77)–(78) exists (see, e.g. Reference 8).

Let $\varphi_j \in H_2^1$, $j = 1, 2, \ldots$, be the eigenfunctions of the problem (42) discussed in section 'An eigenvalue problem'. Then the expansion

$$f(t, x) = \sum_{j=1}^{\infty} b_j(t)\varphi_j(x), \qquad \forall x \in \bar{I}, \qquad t \geqslant 0 \tag{79}$$

converges in H_2^0 for every fixed t. With it the solution of (77)–(78) can be written in the form

$$u(t, x) = \sum_{j=1}^{\infty} a_j(t)\varphi_j(x) \tag{80}$$

where the a_j satisfy the initial value problem

$$\begin{gathered} \frac{da_j}{dt} + \lambda_j a_j = b_j, \qquad t > 0 \\ a_j(0) = 0 \end{gathered} \tag{81}$$

Hence we have

$$a_j(t) = \int_0^t e^{-\lambda_j(t-\tau)} b_j(\tau)\, d\tau \tag{82}$$

and

$$|a_j(t)|^2 \leqslant \frac{1}{2\lambda_j} \int_0^t |b_j(\tau)|^2\, d\tau \tag{83}$$

For any fixed t this implies that

$$\|u(x, t)\|_E^2 = \sum_{j=1}^{\infty} |a_j(t)|^2 \lambda_j \leqslant \frac{1}{2} \sum_{j=1}^{\infty} \int_0^t |b_j(\tau)|^2\, d\tau = \frac{1}{2} \int_0^t \|f(x, \tau)\|_0^2\, d\tau \tag{84}$$

and therefore the series (80) converges for fixed t in the energy norm. Moreover, we see that

$$\|u(x, t)\|_0^2 \leqslant \frac{1}{2} \sum_{j=1}^{\infty} \int_0^t \frac{1}{\lambda_j} |b_j(\tau)|^2\, d\tau \tag{85}$$

For fixed t let now $\bar{u}(t) \in \mathring{S}^{r,\mathbf{k}}(\Delta)$ be the finite element solution defined by

$$\left(\frac{\partial \bar{u}}{\partial t}, v\right)_0 + B(\bar{u}, v) = (f, v)_0, \qquad \forall v \in \mathring{S}^{r,\mathbf{k}}(\Delta) \tag{86}$$

This leads to an initial value problem for a system of ordinary differential equations and is well known that for $h(\Delta) \to 0$ we have convergence to the exact solution of (77)–(78) for the given t (see, e.g., Reference 8).

In practice, the initial value problem for each t-value has to be solved by means of a suitable solution routine for such systems. In general, 'stiff-solvers' have to be used; and there is a need to adapt them to the sparsity structure of the equations under consideration. For a discussion of such solvers see, e.g., Reference 10.

A-posteriori error analysis—I

In (86) not only $\bar{u}$ but also $\partial\bar{u}/\partial t$ belongs to $\mathring{S}^{r,\mathbf{k}}(\Delta)$. As before, we may introduce for each t a function $z(t)\in\mathring{H}_2^1$ so that

$$\left(\frac{\partial\bar{u}}{\partial t}, v\right)_0 + B(\bar{u}, v) - (f, v)_0 = B(z(t), v), \qquad \forall v \in \mathring{H}_2^1, \qquad t \text{ fixed.} \tag{87}$$

As in section 'A boundary value problem' this leads to an estimate of the form

$$\|z(t)\|_E \leq K(t) \qquad (t \text{ fixed}) \tag{88}$$

where the constant $K(t)$ is computed as in section 'A boundary value problem'. With

$$z(t) = \sum_{j=1}^{\infty} c_j(t)\varphi_j$$

(88) becomes

$$\sum_{j=1}^{\infty} |c_j(t)|^2 \lambda_j \leq K(t)^2, \qquad (t \text{ fixed}) \tag{89}$$

The error $e(t) = \bar{u}(t) - u(\cdot, t)$ satisfies

$$\left(\frac{\partial e}{\partial t}, v\right)_0 + B(e, v) = B(z, v), \qquad \forall v \in \mathring{H}_2^1$$

and because of $B(z(t), \varphi_j) = c_j(t)\lambda_j$ it follows as in the case of (85) that

$$\|e(t)\|_0^2 \leq \frac{1}{2}\sum_{j=1}^{\infty} \int_0^t \frac{|c_j(\tau)|^2 \lambda_j^2}{\lambda_j} \, d\tau \leq \frac{1}{2}\int_0^t K(\tau)^2 \, d\tau, \qquad (t \text{ fixed}) \tag{90}$$

This represents an error estimate for each fixed t. It is also possible to obtain an estimate in the energy norm provided that the data are sufficiently smooth.

A-posteriori error analysis—II

As mentioned in the introduction, modern solution routines for an initial value problem

$$\frac{dy}{dx} = F(y, x), \qquad y(0) = y_0$$

are adaptive in nature. Their steps are chosen such that the computed solution is an exact solution of the perturbed system

$$\frac{dy}{dx} = F(y, x) + \varepsilon(x)$$

where, under some chosen norm,

$$\|\varepsilon(x)\| \leq \tau$$

with a given tolerance τ.

Accordingly, suppose that for given t our computed solution $\hat{u}$ of (86) represents the exact solution of the perturbed problem

$$\left(\frac{\partial\hat{u}}{\partial t}, v\right)_0 + B(\hat{u}, v) = (f, v)_0 + (g, v)_0, \qquad \forall v \in \mathring{S}^{r,\mathbf{k}} \tag{91}$$

where

$$\|g\|_0 \leq \kappa(t) \tag{92}$$

If the indicated procedures are used, then $g \in \mathring{S}^{r,\mathbf{k}}(\Delta)$. Now it is natural to require that

$$\kappa(t) = K(t) \tag{93}$$

In other words, we obtain K directly from the error estimates of the solution routine. But then (90) provides us also with a computable *a-posteriori* estimate, namely,

$$\|e(t)\|_0^2 \leq \frac{1}{2}\left(1 + \frac{1}{\lambda_1}\right) \int_0^t \kappa(\tau)^2 \, d\tau, \qquad (t \text{ fixed}) \tag{94}$$

The evaluation of the bound is here more delicate than before. The *a-posteriori* estimate can be computed exactly as before if all ordinary differential equations are solved exactly. If adaptive solvers are used for them, the error bound is computed by combining our bounds with the tolerance of the initial value problem solver. It may be noted that this also solves the problem how to adjust the tolerance to the t-discretization. We shall not enter into details here.

CONCLUSIONS AND REMARKS

As noted already in the Introduction, the approaches allow for rather straightforward extensions to a variety of more general problems in one dimension. This includes, in particular, the following types of (one-dimensional) problems:

1. Equations of higher order.
2. Non-selfadjoint problems.
3. Problems on an infinite domain (in R^1). Here we obtain also a bound for the error introduced by the restriction to a bounded domain.
4. The 'lumped mass' approach to eigenvalue problems. This approach may be treated as a finite element method involving the use of numerical quadratures. Here we have to estimate the discretization error as well as the quadrature error, and the total error is obtained by summation. Of course, this leads to a less effective estimate than in the non-lumped approach, because there may be cancellation between the two errors. In fact, this will happen in the case of the simple example of subsection 'An example'. The error of the lumped mass approach may also be estimated in the parabolic case.

For details about these and other problems we refer to Reference 6. Our experience with the approach has shown it to be practically reliable much in the same way as the estimation approaches used in the numerical solution of initial value problems for ordinary differential equations. Moreover, these estimates can be used for the automatic generation of (near) optimal meshes for boundary value problems. These and related questions will be discussed elsewhere.

REFERENCES

1. T. E. Hull, W. H. Enright, B. M. Fellen and A. E. Sedgwick, 'Comparing numerical methods for ordinary differential equations', *SIAM J. Num. Anal.* **9**, 603–637 (1972).
2. L. F. Shampine, H. A. Watts and S. M. Davenport, 'Solving non-stiff-ordinary differential equations—the state of the art', *SIAM Rev.*, **18**, 376–411 (1976).
3. F. T. Krogh, 'On testing a subroutine for the numerical integration of ordinary differential equations', *J. ACM*, **4**, 545–562 (1973).
4. I. Babuška and W. Rheinboldt, 'Computational aspects of finite element analysis', in *Mathematical Software—III* (Ed. J. R. Rice), Academic Press, New York, 1973, pp. 223–253.

5. I. Babuška and W. Rheinboldt, 'Error estimates for adaptive finite element computations', University of Maryland, Institute for Physical Science and Technology, *Technical Note BN–854* (1977); *SIAM J. Num. Anal.*, **15** (1978), in press.
6. I. Babuška and W. Rheinboldt, *Theoretical and Computational Analysis of the Finite Element Method*, in preparation.
7. I. Babuška and W. Rheinboldt, 'Analysis of optimal finite element meshes in R^1, University of Maryland, Institute for Physical Science and Technology, *Technical Note BN–869* (1977).
8. I. Babuška and A. K. Aziz, 'Survey lectures on the mathematical foundations of the finite element method', in *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations* (Ed. A. K. Aziz), Academic Press, New York, 1972.
9. R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
10. C. A. Hall and J. M. Watt (Ed.), *Modern Numerical Methods for Ordinary Differential Equations*, Clarendon Press, Oxford, 1976.