MATHEMATICS OF COMPUTATION
VOLUME 37, NUMBER 156
OCTOBER, 1981

On a Dimensional Reduction Method. III. A Posteriori Error Estimation and an Adaptive Approach*

By M. Vogelius and I. Babuška

Abstract. This paper is the last in a series of three which analyze an adaptive approximate approach for solving $(n + 1)$-dimensional boundary value problems by replacing them with systems of equations in n-dimensional space.

In this paper we show how to find reliable a posteriori estimates for the error and how these can also be used in the design of an adaptive strategy. Various numerical examples are contained in the paper.

1. Introduction. In a recent paper, [6], we introduced the concept of dimensionally reduced solutions to an elliptic boundary value problem. These are obtained by projecting (in the energy) the true solution of the boundary value problem in the $(n + 1)$-dimensional domain $\omega \times [-h, h]$ onto spaces of the form

$$V_N^h = \left\{ \sum_{j=0}^{N} w_j(\mathbf{x})\psi_j(y/h) \middle| w_j \text{ arbitrary} \right\},$$

where $\{\psi_j\}_{j=0}^{\infty}$ is a given set of functions on $[-1, 1]$, ($\mathbf{x}$ are coordinates on ω and y ranges over $[-h, h]$). For some basic ideas behind this concept, see [6] and the introduction to [5]. In [6] the focus was on the right selection of the ψ_j's. It was shown there that for a very wide class of problems the ψ_j's should be selected such that

$$\operatorname{span}\{\psi_j\}_{j=0}^{2k-1} = \mathcal{N}(P^k),$$

where P is a second order ordinary differential operator intrinsic to the elliptic boundary value problem.

In [7] we analyzed the convergence properties of such methods as the order, N, increases.

The present paper, which is a direct continuation of the previous work, deals with the problem of reliable a posteriori error estimation. It also designs an adaptive algorithm for the selection of the right dimensionally reduced solution. As it follows from [6] and [7], a high number of basis functions ψ_j may be needed (depending on the desired accuracy) either if the thickness of the domain, h, is not

---

Received November 25, 1980.

1980 *Mathematics Subject Classification.* Primary 35B25, 41A25, 65N15, 65N30, 73K15.

*This work was in part supported by the Office of Naval Research under contract N00014-77-C-0623. In addition the first author was partially supported by Army Research Office grant #DAA G29-78-G-0177. The computations were supported by the Computer Science Center of the University of Maryland.

0025-5718/81/0000-0158/$07.00

sufficiently small or there are singularities in the true solution to the boundary value problem. (Because of the corner in the domain such singularities are often present in the neighborhood of $\partial\omega \times \{-h\}$ and $\partial\omega \times \{h\}$.)

Since singularities are local phenomena, it is of the utmost practical importance to introduce dimensionally reduced solutions that permit N, the order, to vary throughout the domain ω. This aspect, specifically the adaptive choice of the distribution for N, is also addressed here.

We now give a short review of the contents of this paper.

In Section 2 we give a precise formulation of the model problem (which is identical to that of [6]) and prove some auxiliary results.

Section 3 is devoted to the construction of an estimator for the error. The main theoretical results in this section are Theorem 3.1 and Theorem 3.2, which show that the introduced estimator is an upper bound for the error but on the other hand is not too conservative (away from singularities and for reasonably small h). Numerical experiments verify this and furthermore indicate that even for relatively large h, or strong singularities, the estimator is of the same magnitude as the error. The problems of how to detect if the estimator is unacceptably conservative and how to improve it are addressed in Section 4.

In Section 5 we extend the concept of dimensional reduction to include a possibly different number of basis functions, ψ_j, in different parts of the domain. We also design an adaptive strategy to select the right distribution for the number of basis functions. This strategy is based on our ability to give reliable estimates for the error much in the same way as the strategy used by the finite element solver F.E.A.R.S. to generate an 'optimal' grid; cf. [1].

Finally Section 6 (and also 4) contains a numerical example that illustrates how well the error estimation and the adaptive strategy perform in practice.

2. Notation and the Model Problem. Let $\mathcal{H}$ be a separable Hilbert space with inner product $\langle u, v\rangle$ and norm $\|u\| = \langle u, u\rangle^{1/2}$.

A denotes a (possibly unbounded) selfadjoint linear operator in $\mathcal{H}$ with domain of definition $\mathcal{D}(A)$.

Furthermore, we assume that A is a strictly positive-definite operator, i.e., there exists $C > 0$ such that

$$\forall\, u \in \mathcal{D}(A)\colon C\|u\|^2 \leqslant \langle Au, u\rangle.$$

Let M be a selfadjoint bounded linear operator in $\mathcal{H}$. M is also assumed to be a strictly positive-definite operator.

$\mathcal{D}(A^{1/2})$ is itself a Hilbert space with inner product $\langle u, v\rangle + \langle A^{1/2}u, A^{1/2}v\rangle$. The same is true about $\mathcal{D}((M^{-1}A)^k)$ for any integer $k \geqslant 0$.

I denotes an interval on the real line. $L^2(I; \mathcal{H})$ is defined as the set of strongly measurable functions $u\colon I \to \mathcal{H}$ such that $\|u(\cdot)\|$ is an element of $L^2(I)$; cf. [4]. The same goes for $L^2(I; \mathcal{D}(A^{1/2}))$ and $L^2(I; \mathcal{D}((M^{-1}A)^k))$.

We also need Sobolev spaces of functions with values in $\mathcal{H}$, $\mathcal{D}(A^{1/2})$ and $\mathcal{D}((M^{-1}A)^k)$. $H^1(I; \mathcal{H})$ denotes the space of functions $u\colon I \to \mathcal{H}$ such that $u(\cdot) \in L^2(I; \mathcal{H})$ and $(d/dy)u(\cdot) \in L^2(I; \mathcal{H})$; cf. [2]. The spaces for $\mathcal{D}(A^{1/2})$ and $\mathcal{D}((M^{-1}A)^k)$ are defined similarly. The derivative is taken in the distributional sense.

$H^1(I)$ denotes the standard Sobolev space on I.

Assume a and b are real valued functions in $L^\infty([-1, 1])$ such that

$$a_0 \leqslant a(y), \quad b_0 \leqslant b(y)$$

for some constants $a_0 > 0$, $b_0 > 0$. a_h and $b_h \in L^\infty([-h, h])$ are then defined as

$$a_h(y) = a(y/h), \quad b_h(y) = b(y/h).$$

By $P_h(d/dy)$ we denote the differential operator $-(d/dy)(a_h d/dy)$. Let f and g be two arbitrary vectors from $\mathcal{H}$. We consider the following model problem

$$(1) \qquad \begin{cases} P_h\left(\dfrac{d}{dy}\right)Mu^h + b_h A u^h = 0 & \text{in }]-h, h[, \\ a_h \dfrac{d}{dy} Mu^h = g & \text{for } y = h, \\ a_h \dfrac{d}{dy} Mu^h = f & \text{for } y = -h. \end{cases}$$

The precise formulation of (1) is

$$(2) \qquad \begin{cases} u^h \in H^1([-h, h]; \mathcal{H}) \cap L^2([-h, h]; \mathcal{D}(A^{1/2})), \\ \mathcal{B}_h(u^h, v) = \langle g, v(h)\rangle - \langle f, v(-h)\rangle, \\ \forall\, v \in H^1([-h, h]; \mathcal{H}) \cap L^2([-h, h]; \mathcal{D}(A^{1/2})), \end{cases}$$

where $\mathcal{B}_h$ denotes the bilinear form

$$\mathcal{B}_h(u, v) = \int_{-h}^{h} a_h \left\langle M^{1/2}\frac{d}{dy}u, M^{1/2}\frac{d}{dy}v \right\rangle dy + \int_{-h}^{h} b_n \langle A^{1/2}u, A^{1/2}v\rangle\, dy.$$

For more details, see [6]. In that paper we introduced the notion of dimensionally reduced solutions to (1). Let $\{\psi_j\}_{j=0}^{\infty} \subseteq H^1([-1, 1])$ be a given sequence of linearly independent functions (referred to as basis functions).

Definition. The dimensionally reduced solution u_N^h of order N is the projection of u^h onto the space

$$V_N^h = \left\{ \sum_{j=0}^{N} \psi_j(y/h)x_j \,|\, x_j \in \mathcal{D}(A^{1/2}), \quad j = 0, \ldots, N \right\}.$$

The projection is with respect to the inner product $\mathcal{B}_h(u, v)$.

We proved that in order to obtain optimal rate error estimates for $h \to 0$ there is essentially only one choice for the sequence $\{\psi_j\}_{j=0}^{\infty}$. This is related to the operator $P = b^{-1}(d/dy)(ad/dy)$.

THEOREM. *There exists a sequence of linearly independent functions* $\{\psi_j\}_{j=0}^{\infty}$, *with*

(i) $\mathcal{N}(P^i) = \operatorname{span}\{\psi_j\}_{j=0}^{2i-1}$, $i \geqslant 1$,

that has the following property:

(ii) *For any integer* $N \geqslant 0$ *and for any given set of vectors* $f, g \in \mathcal{D}((AM^{-1})^N)$ *there exists a constant* C_N (*independent of* h) *such that*

$$|||u^h - u_{2N}^h|||_E \leqslant C_N h^{2N+1/2}.$$

$\mathcal{N}(P^i)$ here denotes the nullspace of P^i, and $|||\cdot|||_E$ is the energy-norm associated with the bilinear form $\mathcal{B}_h$. This is slightly different from the formulation in [6],

where we used the norm

$$\left(\int_{-h}^{h}\left\|\frac{d}{dy}u(y)\right\|^2 dy + \int_{-h}^{h}\|A^{1/2}u(y)\|^2 dy\right)^{1/2}.$$

It is obvious though, that these two norms are equivalent with constants independent of h. For more details concerning this theorem and its converse we refer to [6]. It is now convenient to introduce

Definition. Any sequence $\{\psi_j\}_{j=0}^{\infty}$ that has the two properties listed in the previous theorem is said to be an optimal sequence of basis functions.

It follows immediately from Theorem 4.1 of [6] that any two optimal sequences of basis functions $\{\phi_j\}_{j=0}^{\infty}$ and $\{\psi_j\}_{j=0}^{\infty}$ satisfy

$$\operatorname{span}\{\phi_j\}_{j=0}^{N} = \operatorname{span}\{\psi_j\}_{j=0}^{N} \quad \forall N \geqslant 0.$$

We shall often use this fact without explicitly mentioning so.

In the present paper we need a slightly different but weaker version of the result contained in Theorem 4.1 of [6].

LEMMA 2.1. *Let $\{\psi_j\}_{j=0}^{\infty}$ be an optimal sequence of basis functions, and let N be an integer $\geqslant 0$.*

For any nontrivial set of vectors $f, g \in \mathcal{H}$, there exists a constant C_N (independent of h) such that

$$C_N h^{2N+1/2} \leqslant |||u^h - u_{2N}^h|||_E$$

for h sufficiently small.

Proof. The proof is by contradiction, i.e., we assume

$$|||u^{h_i} - u_{2N}^{h_i}|||_E = o\left(h_i^{2N+1/2}\right)$$

for some sequence h_i, with $h_i \to 0$ as $i \to \infty$.

If f and g are linearly independent, Theorem 4.1 of [6] then gives that

$$\operatorname{span}\{\psi_j\}_{j=0}^{2N+2} \subseteq \operatorname{span}\{\psi_j\}_{j=0}^{2N},$$

and this is obviously a contradiction.

We therefore only have to consider the case when f and g are linearly dependent, say $f = \alpha \cdot g$, $g \neq 0$. As in the proof of Theorem 4.1 of [6] it now follows that

$$\frac{d}{dy}\psi_{N+1}^0 - \alpha\frac{d}{dy}\psi_{N+1}^1 \in \operatorname{span}\left\{\frac{d}{dy}\psi_j\right\}_{j=0}^{2N}, \tag{3}$$

where ψ_j^0 and ψ_j^1 are as introduced in Lemma 3.1 of [6]. Since

$$\operatorname{span}\left\{\frac{d}{dy}\psi_j\right\}_{j=0}^{2N} = \operatorname{span}\left\{\frac{d}{dy}\psi_j^0, \frac{d}{dy}\psi_j^1\right\}_{j=0}^{N}$$

and

$$b^{-1}\frac{d}{dy}a\frac{d}{dy}\psi_j^l = \psi_{j-1}^l, \qquad l = 0, 1,$$

(3) immediately leads to the conclusion

$$\psi_1^0 - \alpha\psi_1^1 \in \operatorname{span}\{\psi_0^0, \psi_0^1\}.$$

Because of the fact that ψ_0^0 and ψ_0^1 are both constant, we get

$$\frac{d}{dy}\psi_1^0 - \alpha\frac{d}{dy}\psi_1^1 = 0. \tag{4}$$

But, according to [6], ψ_1^0 and ψ_1^1 satisfy

$$a\frac{d}{dy}\psi_1^0(1) = 1, \quad a\frac{d}{dy}\psi_1^1(1) = 0,$$

so (4) is obviously a contradiction. □

For the analysis in this paper we also need two simple regularity results, one concerning the true solution u^h and one concerning the dimensionally reduced solutions u_N^h.

LEMMA 2.2. *Let u^h denote the solution to* (2). *If for some integer $k \geqslant 0$, $f, g \in \mathcal{D}((AM^{-1})^k)$, then*

$$u^h \in H^1\big([-h, h]; \mathcal{D}\big((M^{-1}A)^k\big)\big) \cap L^2\big([-h, h]; \mathcal{D}\big(A^{1/2}(M^{-1}A)^k\big)\big).$$

Proof. Let $\tilde{u}^h$ denote the solution to (2) in the case where f and g have been replaced by $(AM^{-1})^k f$ and $(AM^{-1})^k g$, respectively, i.e.,

$$\mathcal{B}_h(\tilde{u}^h, v) = \big\langle (AM^{-1})^k g, v(h)\big\rangle - \big\langle (AM^{-1})^k f, v(-h)\big\rangle$$
$$\forall\, v \in H^1\big([-h, h]; \mathcal{H}\big) \cap L^2\big([-h, h]; \mathcal{D}(A^{1/2})\big).$$

We then get that

$$\begin{aligned}\mathcal{B}_h\big((A^{-1}M)^k\tilde{u}^h, v\big) = \mathcal{B}_h\big(\tilde{u}^h, (A^{-1}M)^k v\big) &= \big\langle (AM^{-1})^k g, (A^{-1}M)^k v(h)\big\rangle \\ &\quad -\big\langle (AM^{-1})^k f, (A^{-1}M)^k v(-h)\big\rangle \\ &= \langle g, v(h)\rangle - \langle f, v(-h)\rangle \\ &\qquad \forall\, v \in H^1\big([-h, h]; \mathcal{H}\big) \cap L^2\big([-h, h]; \mathcal{D}(A^{1/2})\big),\end{aligned}$$

and hence $(A^{-1}M)^k\tilde{u}^h = u^h$. Since $\tilde{u}^h \in H^1([-h, h]; \mathcal{H}) \cap L^2([-h, h]; \mathcal{D}(A^{1/2}))$, the desired result now immediately follows. □

The regularity result we need for the dimensionally reduced solutions u_N^h is the following.

LEMMA 2.3. *Let $u_N^h = \sum_{j=0}^N \psi_j(y/h)x_j$ denote a dimensionally reduced solution of order N. If, for some integer $k \geqslant 0, f, g \in \mathcal{D}((AM^{-1})^k)$, then*

$$x_j \in \mathcal{D}\big((M^{-1}A)^{k+1}\big) \quad \forall\, j\colon 0 \leqslant j \leqslant N.$$

Proof. Let $\mathbf{x} \in \mathcal{H}^{N+1}$ denote the vector $(x_0, \ldots, x_N)$. It is clear that $\mathbf{x}$ is the solution to

$$h\langle \mathbf{C}A^{1/2}\mathbf{x}, A^{1/2}\mathbf{y}\rangle + h^{-1}\langle \mathbf{D}M^{1/2}\mathbf{x}, M^{1/2}\mathbf{y}\rangle = \langle \mathbf{r}, \mathbf{y}\rangle \quad \forall\, \mathbf{y} \in \big[\mathcal{D}(A^{1/2})\big]^{N+1},$$

where $\mathbf{C} = \{c_{ij}\}_{i,j=0}^N$, $\mathbf{D} = \{d_{ij}\}_{i,j=0}^N$, and $\mathbf{r} = \{r_i\}_{i=0}^N$ are given by

$$c_{ij} = \int_{-1}^1 \psi_i(y)\psi_j(y)\, dy, \quad d_{ij} = \int_{-1}^1 \left(\frac{d}{dy}\psi_i\right)\left(\frac{d}{dy}\psi_j\right) dy,$$

and $r_i = \psi_i(1)g - \psi_i(-1)f$, respectively. (We have here used $\langle \cdot, \cdot \rangle$ also to denote the inner product $\sum_{i=0}^N \langle x_i, y_i\rangle$ in $\mathcal{H}^{N+1}$.)

Due to the fact that A is selfadjoint and M bounded, we get that $\mathbf{x} \in [\mathcal{D}(A)]^{N+1}$ and

$$h\mathbf{C}A\mathbf{x} + h^{-1}\mathbf{D}M\mathbf{x} = \mathbf{r}, \quad \text{i.e.,} \quad A\mathbf{x} = h^{-1}\mathbf{C}^{-1}\mathbf{r} - h^{-2}\mathbf{C}^{-1}\mathbf{D}M\mathbf{x}.$$

By successive application of this equality it follows that $\mathbf{r} \in [\mathcal{D}((AM^{-1})^k)]^{N+1}$ implies $\mathbf{x} \in [\mathcal{D}((M^{-1}A)^{k+1})]^{N+1}$. This finishes the proof of the lemma. □

3. A Posteriori Error Estimation in the General Case. As already mentioned in the introduction, one purpose of this paper is to derive a reliable technique for a posteriori estimation of the error introduced by dimensional reduction. The error here is measured in the energy-norm. The key ingredient of this technique is a so-called estimator Est, which we now proceed to define.

Let $\varepsilon \in H^1([-h, h]; \mathcal{H})$ be the solution to

$$\text{(5)} \qquad \begin{cases} P_h\left(\dfrac{d}{dy}\right)M\varepsilon = -P_h\left(\dfrac{d}{dy}\right)Mu_N^h - b_hAu_N^h & \text{in }]-h, h[, \\ a_h\dfrac{d}{dy}M\varepsilon = g - a_h\dfrac{d}{dy}Mu_N^h & \text{for } y = h, \\ a_h\dfrac{d}{dy}M\varepsilon = f - a_h\dfrac{d}{dy}Mu_N^h & \text{for } y = -h. \end{cases}$$

The exact meaning of (5) is

$$\text{(6)} \qquad \begin{cases} \varepsilon \in H^1([-h, h]; \mathcal{H}) \quad \text{and} \\ \displaystyle\int_{-h}^{h} a_h\left\langle M^{1/2}\frac{d}{dy}\varepsilon, M^{1/2}\frac{d}{dy}v\right\rangle dy = \langle g, v(h)\rangle - \langle f, v(-h)\rangle \\ \displaystyle -\int_{-h}^{h} a_h\left\langle M^{1/2}\frac{d}{dy}u_N^h, M^{1/2}\frac{d}{dy}v\right\rangle dy - \int_{-h}^{h} b_h\langle Au_N^h, v\rangle\, dy \\ \forall\, v \in H^1([-h, h]; \mathcal{H}). \end{cases}$$

In Eqs. (5) and (6) we have used the fact that $u_N^h(y) \in \mathcal{D}(A)$, which immediately follows from Lemma 2.3 with $k = 0$.

Since (6) is a Neuman problem for ε, it only has a solution provided

$$\langle g, x\rangle - \langle f, x\rangle - \int_{-h}^{h} b_h\langle Au_N^h, x\rangle\, dy = 0 \quad \forall\, x \in \mathcal{H}.$$

Because of the equations defining u_N^h, it follows that this is true if

$$\text{(7)} \qquad 1 \in \operatorname{span}\{\psi_j\}_{j=0}^{N}.$$

Note. According to Theorems 3.1 and 4.1 of [6], condition (7) is in general necessary and always sufficient to ensure that

$$|||u^h - u_N^h|||_E \to 0 \quad \text{for } h \to 0.$$

Certainly (7) is satisfied for any optimal sequence of basis functions.

We now define

$$\text{(8)} \qquad \text{Est} = \left(\int_{-h}^{h} a_h\left\|M^{1/2}\frac{d}{dy}\varepsilon\right\|^2 dy\right)^{1/2}.$$

The function ε is clearly not uniquely determined, but, since only $d\varepsilon/dy$ is involved, Est is well-defined. We shall now show that the estimator Est exhibits some very attractive properties.

THEOREM 3.1. *Let u^h be the solution to* (2) *and u_N^h a dimensionally reduced solution of order $N \geqslant 0$ corresponding to a sequence of basis functions that satisfies* (7). *If* Est *is as defined in* (8), *then*

$$|||u^h - u_N^h|||_E \leqslant \text{Est}.$$

Note. In the terminology of [1] this theorem says that Est is a 'guaranteed' upper estimator.

Proof. Clearly

$$|||u^h - u_N^h|||_E = \sup|\mathcal{B}_h(u^h - u_N^h, v)|/|||v|||_E,$$

where the sup is taken over $v \in H^1([-h, h]; \mathcal{H}) \cap L^2([-h, h]; \mathcal{D}(A^{1/2}))$. According to the definition of ε, this is nothing but

$$\sup\left|\int_{-h}^{h} a_h\left\langle M^{1/2}\frac{d}{dy}\varepsilon, M^{1/2}\frac{d}{dy}v\right\rangle dy\right|/|||v|||_E,$$

and, using Schwarz's inequality, we now get

$$|||u^h - u_N^h|||_E \leqslant \left(\int_{-h}^{h} a_h\left\|M^{1/2}\frac{d}{dy}\varepsilon\right\|^2 dy\right)^{1/2} = \text{Est}. \qquad \square$$

For use of the estimator Est in actual computations, it is important that it is very close to the real error in a wide class of situations. In the terminology of [1] this is expressed by the requirement that Est be asymptotically exact. The following theorem contains a precise formulation of the asymptotical exactness for the estimator Est. It is essential here that the dimensional reduction is based on an optimal sequence of basis functions.

THEOREM 3.2. *Let $\{\psi_j\}_{j=0}^{\infty}$ be an optimal sequence of basis functions. Let u^h be the solution to* (2) *and u_N^h be the dimensionally reduced solution of order $N \geqslant 0$. Assume furthermore that f and g are elements of $\mathcal{D}((AM^{-1})^{[(N+1)/2]+1})$.*

If Est *is as defined in* (8), *then*

$$\text{Est} = |||u^h - u_N^h|||_E(1 + O(h^2)).$$

Note. Here $[\cdot]$ denotes the integer part.

Proof. Since $f, g \in \mathcal{D}(AM^{-1})$, it follows from Lemma 2.3 and (5) that ε in this case can be selected so that $\varepsilon \in H^1([-h, h]; \mathcal{D}(A))$. We now have

$$\begin{aligned}\text{Est} &= \left|\int_{-h}^{h} a_h\left\langle M^{1/2}\frac{d}{dy}\varepsilon, M^{1/2}\frac{d}{dy}\varepsilon\right\rangle dy\right|/\left(\int_{-h}^{h} a_h\left\|M^{1/2}\frac{d}{dy}\varepsilon\right\|^2 dy\right)^{1/2} \\ &\leqslant \left(\sup\left|\int_{-h}^{h} a_h\left\langle M^{1/2}\frac{d}{dy}\varepsilon, M^{1/2}\frac{d}{dy}v\right\rangle dy\right|/|||v|||_E\right) \\ &\quad \cdot \left(|||\varepsilon|||_E/\left(\int_{-h}^{h} a_h\left\|M^{1/2}\frac{d}{dy}\varepsilon\right\|^2 dy\right)^{1/2}\right),\end{aligned}$$

where the sup is taken over

$$v \in H^1([-h, h]; \mathcal{H}) \cap L^2([-h, h]; \mathcal{D}(A^{1/2})).$$

As shown in the proof of Theorem 3.1, this last expression is equal to

$$|||u^h - u_N^h|||_E \cdot |||\varepsilon|||_E / \left(\int_{-h}^{h} a_h \left\| M^{1/2} \frac{d}{dy} \varepsilon \right\|^2 dy \right)^{1/2}.$$

The theorem will therefore be proven if we show that ε can be chosen such that

$$\int_{-h}^{h} b_h \|A^{1/2}\varepsilon\|^2 \, dy \leqslant C_N h^2 \int_{-h}^{h} a_h \left\| M^{1/2} \frac{d}{dy} \varepsilon \right\|^2 dy. \tag{9}$$

It is clear that, by appropriately selecting the undetermined constant of ε, we can obtain

$$\begin{aligned} \int_{-h}^{h} b_h \|A^{1/2}\varepsilon\|^2 \, dy &\leqslant Ch^2 \int_{-h}^{h} a_h \left\| \frac{d}{dy} A^{1/2}\varepsilon \right\|^2 dy \\ &= Ch^2 \int_{-h}^{h} a_h \left\langle \frac{d}{dy}\varepsilon, \frac{d}{dy} A\varepsilon \right\rangle dy. \end{aligned} \tag{10}$$

Now from (6), the definition of ε, we have

$$\begin{aligned} \int_{-h}^{h} a_h \left\langle \frac{d}{dy}\varepsilon, \frac{d}{dy} A\varepsilon \right\rangle dy &= \int_{-h}^{h} a_h \left\langle M^{1/2} \frac{d}{dy}\varepsilon, M^{1/2} \frac{d}{dy} M^{-1} A\varepsilon \right\rangle dy \\ &= \langle g, M^{-1}A\varepsilon(h) \rangle - \langle f, M^{-1}A\varepsilon(-h) \rangle \\ &\quad - \int_{-h}^{h} a_h \left\langle M^{1/2} \frac{d}{dy} u_N^h, M^{1/2} \frac{d}{dy} M^{-1} A\varepsilon \right\rangle dy \\ &\quad - \int_{-h}^{h} b_h \langle Au_N^h, M^{-1}A\varepsilon \rangle \, dy, \end{aligned}$$

and by introduction of u^h this is, because of Lemma 2.3, equal to

$$\mathcal{B}_h(M^{-1}A(u^h - u_N^h), \varepsilon).$$

By an application of Schwarz's inequality, this expression can be bounded by

$$|||M^{-1}A(u^h - u_N^h)|||_E \cdot |||\varepsilon|||_E.$$

From the fact that A is strictly positive-definite together with (10), it follows that

$$|||\varepsilon|||_E \leqslant C(1 + h^2) \left(\int_{-h}^{h} a_h \left\| \frac{d}{dy} A^{1/2}\varepsilon \right\|^2 dy \right)^{1/2},$$

and, since we only need to consider small h, this now gives

$$\begin{aligned} \int_{-h}^{h} a_h \left\| \frac{d}{dy} A^{1/2}\varepsilon \right\|^2 dy &= \mathcal{B}_h(M^{-1}A(u^h - u_N^h), \varepsilon) \\ &\leqslant C |||M^{-1}A(u^h - u_N^h)|||_E \left(\int_{-h}^{h} a_h \left\| \frac{d}{dy} A^{1/2}\varepsilon \right\|^2 dy \right)^{1/2}. \end{aligned}$$

By insertion in (10) we conclude that

$$\int_{-h}^{h} b_h \|A^{1/2}\varepsilon\|^2 \, dy \leqslant Ch^2 |||M^{-1}A(u^h - u_N^h)|||_E^2. \tag{11}$$

For the rest of this proof let us assume that N is even (the procedure for N odd is quite similar, only there are slight variations of Lemma 2.1 and Theorem 3.1 of [6] needed in this case).

It follows immediately from the proofs of Lemmas 2.2 and 2.3 that $M^{-1}Au^h$ and $M^{-1}Au_N^h$ are solutions to the same problems as u^h and u_N^h just with f and g replaced by $AM^{-1}f$ and $AM^{-1}g$. Because of Theorem 3.1 in [6] and the assumption that $f, g \in \mathcal{D}((AM^{-1})^{N/2+1})$, we get

$$|||M^{-1}A(u^h - u_N^h)|||_E \leqslant C_N h^{N+1/2},$$

and this combined with (11) immediately leads to

$$\int_{-h}^{h} b_h \|A^{1/2}\varepsilon\|^2 \, dy \leqslant C_N h^{2N+3}.$$

On the other side from Lemma 2.1 it follows that

$$C_N h^{2N+1} \leqslant |||u^h - u_N^h|||_E^2.$$

From the definition of Est and Theorem 3.1 of this paper, we therefore get

$$C_N h^{2N+1} \leqslant \int_{-h}^{h} a_h \left\| M^{1/2} \frac{d}{dy} \varepsilon \right\|^2 dy,$$

that is, we have finally proven

$$\int_{-h}^{h} b_h \|A^{1/2}\varepsilon\|^2 \, dy \leqslant C_N h^2 \int_{-h}^{h} a_h \left\| M^{1/2} \frac{d}{dy} \varepsilon \right\|^2 dy. \qquad \square$$

Since the estimator Est is to be used in actual computation, it is of the utmost importance that it can be calculated very simply. The equations (6) and (8) that include the solution of an O.D.E. are therefore not well suited as a formula for the calculation of Est. In the following we shall show how easily Est can be calculated by means of different formulae. For these formulae to be valid it is essential not only that the dimensional reduction is based on an optimal sequence of basis functions but also that this sequence satisfies a special orthogonality condition.

Let $\{\psi_j\}_{j=0}^{\infty}$ denote an optimal sequence of basis functions which is orthogonal in the semi-inner-product $\int_{-1}^{1} ad/dy \cdot d/dy \cdot dy$. ($\psi_j$ is clearly uniquely determined modulo a constant and a scalar.)

We define a sequence $\{\phi_j\}_{j=0}^{\infty}$ by

$$\begin{aligned} \phi_0 &= 1, \\ \phi_1(y) &= \psi_1(y) - \left(\int_{-1}^{1} b(t)\, dt \right)^{-1} \cdot \int_{-1}^{1} b(t)\psi_1(t)\, dt \\ \phi_j(y) &= \psi_j(y) - \psi_j(-1) \quad \text{for } j \geqslant 2. \end{aligned}$$

($\phi_j, j \geqslant 1$, are therefore uniquely determined modulo a scalar.)

LEMMA 3.1. *Let $\{\phi_j\}_{j=0}^{\infty}$ be a sequence as defined above. Let $u_N^h = \Sigma_{j=0}^{N} \phi_j(y/h)x_j$ denote the corresponding dimensionally reduced solution of order N. If ε is as defined in* (6), *then there exist constants $\{c_k\}_{k=0}^{\infty}$ and $\{\tilde{c}_k\}_{k=0}^{\infty}$ such that*

$$\begin{aligned} \frac{d}{dy}\varepsilon = &\left(\frac{d}{dy}\phi_1\right)(y/h)M^{-1}(c_0(f+g) + \tilde{c}_0(f-g)) \\ &+ \left(\frac{d}{dy}\phi_2\right)(y/h)hM^{-1}Ac_1x_0 \quad \textit{for } N = 0 \end{aligned}$$

and

$$\frac{d}{dy}\varepsilon = \left(\frac{d}{dy}\phi_{N+1}\right)(y/h)hM^{-1}A(c_N x_{N-1} + \tilde{c}_N x_N) + \left(\frac{d}{dy}\phi_{N+2}\right)(y/h)hM^{-1}Ac_{N+1}x_N \quad \text{for } N \geqslant 1.$$

Proof. Since $\{\phi_j\}_{j=0}^{\infty}$ is an optimal sequence of basis functions, it follows immediately that

$$\varepsilon = \sum_{j=0}^{N+2} \phi_j(y/h)\varepsilon_j$$

for some $\varepsilon_j \in \mathcal{H}$, $0 \leqslant j \leqslant N+2$, i.e.,

$$\frac{d}{dy}\varepsilon = \sum_{j=1}^{N+2} h^{-1}\left(\frac{d}{dy}\phi_j\right)(y/h)\varepsilon_j.$$

For any $1 \leqslant j \leqslant N$ and $x \in \mathcal{D}(A^{1/2})$,

$$\langle M\varepsilon_j, x\rangle \int_{-1}^{1} a\left(\frac{d}{dy}\phi_j\right)^2 dy = \langle M^{1/2}\varepsilon_j, M^{1/2}x\rangle h\int_{-h}^{h} a_h\left(\frac{d}{dy}[\phi_j(y/h)]\right)^2 dy = h\int_{-h}^{h} a_h\left\langle M^{1/2}\frac{d}{dy}\varepsilon, M^{1/2}\frac{d}{dy}[\phi_j(y/h)x]\right\rangle dy,$$

where the last equality is due to the fact that $\{\phi_j\}_{j=0}^{\infty}$ is orthogonal in the semi-inner-product $\int_{-1}^{1} ad/dy \cdot d/dy \cdot dy$. Now from (6) we get that the last expression is equal to $h\mathcal{B}_h(u^h - u_N^h, \phi_j(y/h)x)$, and this vanishes because u_N^h is defined as a projection. We have therefore proven that

$$\langle M\varepsilon_j, x\rangle = 0 \quad \forall\, 1 \leqslant j \leqslant N, x \in \mathcal{D}(A^{1/2}),$$

or

$$\varepsilon_j = 0 \quad \forall\, 1 \leqslant j \leqslant N. \tag{12}$$

For $j = N+1$ we get as before

$$\langle M\varepsilon_{N+1}, x\rangle \int_{-1}^{1} a\left(\frac{d}{dy}\phi_{N+1}\right)^2 dy = h\int_{-h}^{h} a_h\left\langle M^{1/2}\frac{d}{dy}\varepsilon, M^{1/2}\frac{d}{dy}[\phi_{N+1}(y/h)x]\right\rangle dy,$$

and this is, according to (6), equal to

$$h\langle g, \phi_{N+1}(1)x\rangle - h\langle f, \phi_{N+1}(-1)x\rangle - h\int_{-h}^{h} b_h\langle Au_N^h, \phi_{N+1}(y/h)x\rangle\, dy.$$

In this identity we also used the orthogonality of the ϕ_j's. Concerning the last term

$$h\int_{-h}^{h} b_h\langle Au_N^h, \phi_{N+1}(y/h)x\rangle\, dy = h^2\sum_{j=0}^{N}\langle Ax_j, x\rangle\int_{-1}^{1} b(y)\phi_j(y)\phi_{N+1}(y)\, dy.$$

This vanishes for $N = 0$, because $\int_{-1}^{1} b(y)\phi_1(y)dy = 0$, and therefore establishes the formula for ε_{N+1} in the particular case $N = 0$.

In the following we are left to consider $N \geqslant 1$. Let j be any integer $0 \leqslant j < N - 1$. Then

$$\int_{-1}^{1} b(y)\phi_j(y)\phi_{N+1}(y)\,dy = \int_{-1}^{1} \frac{d}{dy} a \frac{d}{dy}\left(\sum_{i=0}^{j+2} \alpha_i \phi_i(y)\right)\phi_{N+1}(y)\,dy, \tag{13}$$

where the α_j's are selected such that $P(\sum_{i=0}^{j+2} \alpha_i \phi_i) = \phi_j$.

This is possible because $\{\phi_j\}_{j=0}^{\infty}$ is an optimal sequence of basis functions.

Since $N \geqslant 1$ and $\int_{-1}^{y} ds/a(s) \in \mathfrak{N}(P)$, we also have that

$$\phi_{N+1}(1) - \phi_{N+1}(-1) = \int_{-1}^{1} a \frac{d}{dy}\left(\int_{-1}^{y} \frac{1}{a(s)}\,ds\right)\frac{d}{dy}\phi_{N+1}\,dy = 0.$$

We already know that $\phi_{N+1}(-1) = 0$, i.e., we conclude

$$\phi_{N+1}(1) = \phi_{N+1}(-1) = 0. \tag{14}$$

It follows immediately from an integration by parts in (13) and application of (14) that for any integer $0 \leqslant j < N - 1$

$$\int_{-1}^{1} b(y)\phi_j(y)\phi_{N+1}(y)\,dy = \sum_{i=0}^{j+2} \alpha_i \int_{-1}^{1} a(y) \frac{d}{dy}\phi_i \frac{d}{dy}\phi_{N+1}\,dy = 0.$$

The last identity is due to the orthogonality of $\{\phi_j\}_{j=0}^{\infty}$ and the fact that $i < N + 1$. In summary we have therefore proven

$$\langle M\varepsilon_{N+1}, x\rangle \int_{-1}^{1} a\left(\frac{d}{dy}\phi_{N+1}\right)^2 dy = -h\int_{-h}^{h} b_h \langle Au_N^h, \phi_{N+1}(y/h)x\rangle\,dy$$

(here again we used (14))

$$= -h^2\left(\langle Ax_{N-1}, x\rangle \int_{-1}^{1} b\phi_{N-1}\phi_{N+1}\,dy + \langle Ax_N, x\rangle \int_{-1}^{1} b\phi_N\phi_{N+1}\,dy\right),$$

that is,

$$\varepsilon_{N+1} = h^2 M^{-1} A(c_N x_{N-1} + \tilde{c}_N x_N), \tag{15}$$

with the explicit expressions for c_N and $\tilde{c}_N$ as given above.

Along exactly the same lines we also get

$$\varepsilon_{N+2} = h^2 M^{-1} A c_{N+1} x_N,$$

and therefore the desired result for the function $d\varepsilon/dy$. □

Lemma 3.1 immediately leads to

THEOREM 3.3. *Let $\{\phi_j\}_{j=0}^{\infty}$ be a sequence with properties as in the previous lemma. Let $u_N^h = \Sigma_{j=0}^{N} \phi_j(y/h)x_j$ denote the corresponding dimensionally reduced solution of order N. If* Est *is as defined in* (8), *then there exist constants $\{C_k\}_{k=0}^{\infty}$ and $\{\tilde{C}_k\}_{k=0}^{\infty}$ such that*

$$\text{Est} = h^{3/2}\left(\|M^{-1/2}(C_0(f+g) + C_0(f-g))\|^2 h^{-2} + \|M^{-1/2}AC_1x_0\|^2\right)^{1/2} \quad \textit{for } N = 0$$

and

$$\text{Est} = h^{3/2}\left(\|M^{-1/2}A(C_N x_{N-1} + \tilde{C}_N x_N)\|^2 + \|M^{-1/2}AC_{N+1}x_N\|^2\right)^{1/2} \quad \textit{for } N \geqslant 1.$$

From the formulae contained in the proof of Lemma 3.1 it follows that the constants $\{c_k, \tilde{c}_k\}_{k=0}^{\infty}$ and $\{C_k, \tilde{C}_k\}_{k=0}^{\infty}$, respectively, are readily computable based on the sequence $\{\phi_j\}_{j=0}^{\infty}$. The following example gives the exact values in the case of constant coefficients a and b.

Example 3.1. If the functions a and b are both constant, it immediately follows that every optimal sequence of basis functions $\{\psi_j\}_{j=0}^{\infty}$ satisfies

$$\operatorname{span}\{\psi_j\}_{j=0}^{N} = \text{all polynomials of degree} \leqslant N$$

for any $N \geqslant 0$, and vice versa.

The specific optimal sequence of basis functions $\{\phi_j\}_{j=0}^{\infty}$ used in Lemma 3.1 and Theorem 3.3 is now (modulo a scalar)

$$\phi_0 = 1, \quad \phi_1(y) = y,$$
$$\phi_j(y) = \int_{-1}^{y} l_{j-1}(t)\, dt, \qquad j \geqslant 2,$$

where l_k denotes the Legendre polynomial of degree k (normalized so that $\int_{-1}^{1} l_k^2(t)\, dt = 2/(2k+1)$).

Because of the alternating even and odd polynomials, it immediately follows that

$$\tilde{c}_k = \tilde{C}_k = 0 \quad \forall\, k \geqslant 0.$$

Simple algebraic manipulations with the Legendre polynomials now give

$$c_0 = 1/2a, \quad c_1 = b/a$$

and

$$c_k = b/\left((2k-1)(2k-3)a\right) \quad \text{for } k \geqslant 2.$$

From the definition of C_k it immediately follows that

$$C_k = \left(\int_{-1}^{1} a\left(\frac{d}{dy}\phi_{k+1}\right)^2 dy\right)^{1/2} \cdot c_k,$$

i.e.,

$$C_0 = \left[1/2a\right]^{1/2}, \quad C_1 = \left[2b^2/3a\right]^{1/2},$$

and

$$C_k = \left[2b^2/\left((2k+1)(2k-1)^2(2k-3)^2 a\right)\right]^{1/2} \quad \text{for } k \geqslant 2.$$

We shall now show how one can derive another set of formulae for the a posteriori error estimator. As it will turn out these formulae are much better suited for practical applications. For the case $N = 0$ it is clear that $Ax_0 = Ch^{-1}(f - g)$. In the following we therefore only consider $N \geqslant 1$.

LEMMA 3.2. *Let* $\{C_k\}_{k=0}^{\infty}$, $\{\tilde{C}_k\}_{k=0}^{\infty}$ *and* $u_N^h = \sum_{j=0}^{N} \phi_j(y/h)x_j$ *be as in the previous theorem. Then there exist constants* D_N^{ij}, $1 \leqslant i, j \leqslant 2$, *such that*

$$A(C_N x_{N-1} + \tilde{C}_N x_N) = h^{-1} D_N^{11} M\left(g - a_h \frac{d}{dy} M u_N^h(h)\right)$$
$$+ h^{-1} D_N^{12} M\left(f - a_h \frac{d}{dy} M u_N^h(-h)\right)$$

and

$$AC_{N+1}x_N = h^{-1}D_N^{21}\left(g - a_h\frac{d}{dy}Mu_N^h(h)\right) + h^{-1}D_N^{22}\left(f - a_h\frac{d}{dy}Mu_N^h(-h)\right) \quad \text{for } N \geqslant 1.$$

Proof. From (5) and Lemma 3.1 it follows that

$$-P_h\left(\frac{d}{dy}\right)Mu_N^h - b_hAu_N^h = -\left(\frac{d}{dy}a\frac{d}{dy}\phi_{N+1}\right)(y/h)A(c_Nx_{N-1} + \tilde{c}_Nx_N) - \left(\frac{d}{dy}a\frac{d}{dy}\phi_{N+2}\right)(y/h)Ac_{N+1}x_N. \tag{16}$$

If we integrate the right-hand side of (6) by parts, set $v = \phi_0(y/h)\cdot x$, and apply the identity (16), the result is

$$hd_N^{11}A(c_Nx_{N-1} + \tilde{c}_Nx_N) + hd_N^{12}Ac_{N+1}x_N = g - f - a_h\frac{d}{dy}Mu_N^h|_{-h}^h$$

with

$$d_N^{11} = \int_{-1}^{1}\frac{d}{dy}a\frac{d}{dy}\phi_{N+1}\,dy \quad \text{and} \quad d_N^{12} = \int_{-1}^{1}\frac{d}{dy}a\frac{d}{dy}\phi_{N+2}\,dy.$$

Performing the similar procedure with $v = \phi_1(y/h)\cdot x$, instead we get

$$hd_N^{21}A(c_Nx_{N-1} + \tilde{c}_Nx_N) + hd_N^{22}Ac_{N+1}x_N = \phi_1(1)\left(g - a_h\frac{d}{dy}Mu_N^h(h)\right) - \phi_1(-1)\left(f - a_h\frac{d}{dy}Mu_N^h(-h)\right)$$

with

$$d_N^{21} = \int_{-1}^{1}\frac{d}{dy}a\frac{d}{dy}\phi_{N+1}\phi_1\,dy \quad \text{and} \quad d_N^{22} = \int\frac{d}{dy}a\frac{d}{dy}\phi_{N+2}\phi_1\,dy.$$

The lemma therefore immediately follows if the matrix

$$\mathbf{d}_N = \{d_N^{ij}\}_{i,j=1}^2$$

is always invertible. The proof of the invertibility follows by contradiction. Assume $\mathbf{d}_N$ is not invertible for some N. This implies that some nontrivial linear combination of the columns of $\mathbf{d}_N$ vanishes. In terms of the functions ϕ_{N+1} and ϕ_{N+2} this says that some nontrivial linear combination $s_1\phi_{N+1} + s_2\phi_{N+2}$ exists such that

$$\int_{-1}^{1}\frac{d}{dy}a\frac{d}{dy}(s_1\phi_{N+1} + s_2\phi_{N+2})v\,dy = 0 \tag{17}$$

for $v = \phi_0$ and $v = \phi_1$. By performing an integration by parts we easily see that this identity must also hold for $v = \phi_j$, $2 \leqslant j \leqslant N$. Since the sequence $\{\phi_j\}_{j=0}^{\infty}$ is an optimal sequence of basis functions, we know that

$$\frac{d}{dy}a\frac{d}{dy}(s_1\phi_{N+1} + s_2\phi_{N+2}) = b\sum_{j=0}^{N}\alpha_j\phi_j$$

for some set of constants $\{\alpha_j\}_{j=0}^N$. Combining this with (17) we conclude that

$$\frac{d}{dy}a\frac{d}{dy}(s_1\phi_{N+1} + s_2\phi_{N+2}) = 0.$$

Due to the fact that $s_1\phi_{N+1}(y) + s_2\phi_{N+2}(y) = 0$ for $y = \pm 1$, this implies

$$s_1\phi_{N+1} + s_2\phi_{N+2} = 0,$$

which obviously contradicts the fact that this is a nontrivial linear combination. □

The difference between the formulae given in Theorem 3.3 and those that are based on Lemma 3.2 is that while the first include elements of the form Ax_j the latter are expressed solely in terms of f, g and the x_j's. In practical applications we seldom know the exact values of the x_j's. Instead we compute some approximate values $\tilde{x}_j$, e.g. using a finite element method. The error introduced by using approximate values, derived from finite elements, in the expressions of Lemma 3.2 can be neglected. The reason is that the difference between x_j and $\tilde{x}_j$ in the $\mathcal{H}$ norm (viz. L^2) is normally very small. The problem with the expressions of Theorem 3.3 in this context is that in general $A\tilde{x}_j$ is not at all defined.

As before we now give the values of the constants D_N^{ij} in the case where a and b are both constants.

Example 3.2. Assume a and b are constants. In this case we already know that $\tilde{C}_N = 0$. Lemma 3.2 therefore reduces to formulae for AC_Nx_{N-1} and $AC_{N+1}x_N$, $N \geqslant 1$. It is easily seen that these are

$$AC_Nx_{N-1} = h^{-1}(2a(2N+1))^{-1/2}\left[\left(g - a_h\frac{d}{dy}Mu_N^h(h)\right) + (-1)^N\left(f - a_h\frac{d}{dy}Mu_N^h(-h)\right)\right]$$

$$AC_{N+1}x_N = h^{-1}(2a(2N+3))^{-1/2}\left[\left(g - a_h\frac{d}{dy}Mu_N^h(h)\right) + (-1)^{N+1}\left(f - a_h\frac{d}{dy}Mu_N^h(-h)\right)\right]$$

for $N \geqslant 1$.

4. Improved Error Estimation–A Specific Example. As mentioned before it is very important that we are able to estimate the error accurately. Theorem 3.2 shows that our estimator Est does exactly that provided the data is sufficiently regular and h is not too large. In this section we shall address the problem of how to detect if the estimator Est is too conservative, due to singularities in the data or large h, and what can be done to correct it. For simplicity we consider the model problem

$$\left[\left(\frac{\partial}{\partial x}\right)^2 + \left(\frac{\partial}{\partial y}\right)^2\right]u^h = 0 \quad \text{in } \Omega^h =]0, 1[\times]-h, h[,$$

$$\frac{\partial u^h}{\partial y} = g \quad \text{for } y = h,$$

$$\frac{\partial u^h}{\partial y} = -g \quad \text{for } y = -h,$$

$$u^h = 0 \quad \text{for } x = 0, 1,$$

where g is an element of $L^2([0, 1])$.

Let $\{\phi_j\}_{j=0}^{\infty}$ be the sequence of polynomials introduced in Example 3.1. The dimensionally reduced solution of order N, u_N^h, has the form

$$u_N^h(x, y) = \sum_{j=0}^{N} \phi_j(y/h)v_j(x),$$

where $v_j \in \mathring{H}^1([0, 1])$, $0 \leqslant j \leqslant N$. Let e_N denote the exact error, i.e., $e_N = u^h - u_N^h$. Since everything is even in y, all the terms in the dimensionally reduced solutions corresponding to odd indices vanish. From here on we only consider dimensionally reduced solutions of even order $2N$. Let us start by giving a table that shows the relative error ($= |||e_{2N}|||_E/|||u^h|||_E$) and the efficiency index of the estimator Est (Eff $= |||e_{2N}|||_E/\text{Est}$) in the case $g(x) = \pi/4$, for $N = 0$ and $N = 1$, respectively, and for different values of h.

TABLE 4.1

	$N = 0$:			$N = 1$:	
h	Rel.Error	Eff.	h	Rel.Error	Eff.
1/2	0.67	0.88	1/2	0.121	0.69
1/4	0.43	0.94	1/4	0.051	0.68
1/8	0.24	0.97	1/8	0.019	0.68
1/16	0.12	0.99	1/16	0.007	0.68
1/32	0.06	0.99	1/32	0.002	0.68

From the table it is evident that the efficiency index approaches 1, as $h \to 0$, for $N = 0$, but that this is not so for $N = 1$. The numbers therefore clearly show that some smoothness condition, as in Theorem 3.2, is essential in order to ensure that this index converges to 1 for $h \to 0$. For most practical applications though, an efficiency index of 0.7 is completely satisfactory and no corrections to Est are needed. It is also important to note that $h = 1/2$ corresponds to a square and that Est still gives a very reliable estimate for the error. The next table lists the efficiency index, also in the case where $g(x) = \pi/4$, but for $N = 2$ and 3, and three different values of h.

By a comparison of Tables 4.1 and 4.2, it is seen that the efficiency significantly decreases as we include more and more polynomials.

TABLE 4.2

	$N = 2$:			$N = 3$:	
h	Rel.Error	Eff.	h	Rel.Error	Eff.
1/2	0.045	0.53	1/2	0.022	0.43
1/4	0.019	0.52	1/4	0.008	0.38
1/8	0.008	0.55	1/8	0.003	0.41

In the following we shall take a closer look at the derivation of the estimator Est for the purpose of suggesting corrections that can increase the efficiency to any desired level. We shall only work out the details of a first correction.

The exact error e_{2N} is the solution to the boundary value problem

$$\left[\left(\frac{\partial}{\partial x}\right)^2 + \left(\frac{\partial}{\partial y}\right)^2\right] e_{2N} = r_{2N} \qquad \text{in } \Omega^h,$$

$$\frac{\partial e_{2N}}{\partial y} = \rho_{2N} \qquad \text{for } y = h,$$

$$\frac{\partial e_{2N}}{\partial y} = -\rho_{2N} \qquad \text{for } y = -h,$$

$$e_{2N} = 0 \qquad \text{for } x = 0, 1,$$

with

$$r_{2N} = -\left[\left(\frac{\partial}{\partial x}\right)^2 + \left(\frac{\partial}{\partial y}\right)^2\right] u_{2N}^h(x, y)$$

and

$$\rho_{2N} = g(x) - \frac{\partial}{\partial y} u_{2N}^h(x, h).$$

In terms of complementary energy the norm of e_{2N}, $|||e_{2N}|||_E$, can now be characterized by

$$|||e_{2N}|||_E^2 = \min_{(s,t)\in\mathfrak{M}} \int_0^1 \int_{-h}^h (s^2(x, y) + t^2(x, y))\, dy\, dx,$$

where

$$(s, t) \in \mathfrak{M} \quad \text{if and only if } (s, t) \in (L^2(\Omega^h))^2,$$

$$\frac{\partial}{\partial x} s + \frac{\partial}{\partial y} t = r_{2N},$$

$$t = \rho_{2N} \quad \text{for } y = h \quad \text{and} \quad t = -\rho_{2N} \quad \text{for } y = -h.$$

If we define t_0 by

$$\frac{\partial}{\partial y} t_0 = r_{2N}, \quad t_0 = \rho_{2N} \quad \text{for } y = h$$

then

$$t_0 = -\rho_{2N} \quad \text{for } y = -h,$$

and it is clear that $(0, t_0) \in \mathfrak{M}$. On the other hand, it is also clear that Est $= (\int_0^1 \int_{-h}^h t_0^2(x, y)\, dy\, dx)^{1/2}$, i.e., Est is simply a particular value of a functional, the minimal value of which is the exact norm of the error. One way to improve the estimator Est is therefore to take the minimum over more than just the single function t_0. This should not be exaggerated since we also have to keep the formulae simple. Define

$$s(x, y) = \int_0^x \left(r_{2N}(z, y) - \bar{r}_{2N}(z)\left(\frac{d}{dy}\zeta\right)(y/h)\right) dz$$

and

$$t(x, y) = h\bar{r}_{2N}(x)\zeta(y/h),$$

where $\bar{r}_{2N}(x)$ denotes the function

$$\bar{r}_{2N}(x) = \frac{1}{2h}\int_{-h}^h r_{2N}(x, y)\, dy,$$

and ζ is an arbitrary element of $H^1([-1, 1])$. It is obvious that

$$\frac{\partial}{\partial x} s + \frac{\partial}{\partial y} t = r_{2N}.$$

If furthermore ζ satisfies $\zeta(1) = 1$, $\zeta(-1) = -1$, it also follows that

$$t = \rho_{2N} \quad \text{for } y = h \quad \text{and} \quad t = -\rho_{2N} \quad \text{for } y = -h.$$

From Lemma 3.1 we get the identity

$$r_{2N}(x, y) = \left(\frac{d}{dy} l_{2N+1}\right)(y/h) \cdot \bar{r}_{2N}(x),$$

where l_{2N+1} as before denotes the Legendre polynomial of degree $2N + 1$, so it immediately follows that

$$\left(\int_0^1 \int_{-h}^{h} (s^2(x, y) + t^2(x, y))\, dy\, dx\right)$$

$$= \left(h^3 \int_{-1}^{1} \zeta^2(y)\, dy \int_0^1 (\bar{r}_{2N}(x))^2\, dx \right.$$

$$\left. + h \int_{-1}^{1} \left(\frac{d}{dy}(l_{2N+1}(y) - \zeta(y))\right)^2 dy \int_0^1 \left(\int_0^x \bar{r}_{2N}\right)^2 dx\right)^{1/2}.$$

It should be noticed that the estimator Est is obtained from this with the choice $\zeta(y) = l_{2N+1}(y)$. If we define

$$A(h) = h^3 \int_0^1 (\bar{r}_{2N}(x))^2\, dx$$

and

$$B(h) = h \int_0^1 \left(\int_0^x \bar{r}_{2N}\right)^2 dx,$$

then the previous expression can be written as

$$\left(\int_{-1}^{1} \zeta^2(y)\, dy A(h) + \int_{-1}^{1} \left(\frac{d}{dy}(l_{2N+1}(y) - \zeta(y))\right)^2 dy B(h)\right)^{1/2}. \tag{18}$$

It is easy to see that the minimum of this expression over ζ approaches

$$\left(\int_{-1}^{1} l_{2N+1}^2(y)\, dy A(h)\right)^{1/2} = \text{Est}$$

as $A(h)/B(h) \to 0$. In the case that $A(h)/B(h) \to \infty$, the minimum of the expression (18) approaches

$$(4A(h)B(h))^{1/4}.$$

Based on these asymptotics we introduce a new estimator Est_1 by

$$\text{Est}_1 = \begin{cases} \text{Est} & \text{if } A(h)/B(h) < \tau, \\ (4A(h)B(h))^{1/4} & \text{if } A(h)/B(h) > \tau, \end{cases}$$

where τ is some specified constant. The quantity $A(h)/B(h)$ will therefore tell us whether we shall use the value of the old estimator Est or not. Since we know that $h\bar{r}_{2N}(x) = \rho_{2N}(x)$, $A(h)$ and $B(h)$ can very easily be expressed in terms of ρ_{2N}. It follows from this that $A(h)/B(h)$ is small exactly if h is small and ρ_{2N} is sufficiently

smooth. Since the smoothness of ρ_{2N} may sharply vary on the interval [0, 1], we get that the estimator Est may well give a good estimation of the error in parts of [0, 1] but not in others. We therefore divide the interval [0, 1] into K disjoint subintervals I_j, $1 \leqslant j \leqslant K$, and define

$$\mathrm{Est}_2 = \left(\sum_{j=1}^{K} \mathrm{Est}_1(I_j)^2 \right)^{1/2},$$

where $\mathrm{Est}_1(I_j)$ refers to the estimator Est_1 computed on the interval I_j. The following table shows the efficiency index of the estimator Est_2 ($\mathrm{Eff}_2 = |||e_{2N}|||_E/\mathrm{Est}_2$) in the case $g(x) = \pi/4$ for $N = 2, 3$ and three different values of h. The interval [0, 1] was divided into eight subintervals of equal length, i.e. $K = 8$, and the constant τ in the definition of Est_1 was chosen to be 10^2.

Although the efficiency index is not quite 1, Table 4.3 shows a definite improvement over Table 4.2. Again note that the efficiency is almost independent of h. If additional accuracy of the estimator is deemed necessary, this can obviously be obtained by an extension of the technique used to derive Est_2. Whether such additional corrections are worthwhile in the end of course depends on the balance between the cost of computing the estimator and what is computationally to be gained from a more accurate estimator.

TABLE 4.3

	$N = 2$:			$N = 3$:	
h	Rel.Error	Eff_2	h	Rel.Error	Eff_2
1/2	0.045	0.62	1/2	0.022	0.60
1/4	0.019	0.59	1/4	0.008	0.53
1/8	0.008	0.62	1/8	0.003	0.56

5. Some Remarks on an Adaptive Strategy. As already mentioned in the introduction, the goal of this paper is not only to derive reliable estimates for the error, but also to use these estimates as tools in an automatic selection of the right order dimensionally reduced solution for a given problem.

First let us introduce a slight generalization of the concept of dimensionally reduced solution. Instead of projecting onto the space $\{\Sigma_{j=0}^{N} \psi_j(y/h)x_j | x_j \in \mathcal{D}(A^{1/2}), j = 0, \ldots, N\}$, we project onto $\{\Sigma_{j=0}^{N} \psi_j(y/h)x_j | x_j \in \mathcal{K}_j, j = 0, \ldots, N\}$, where $\{\mathcal{K}_j\}_{j=0}^{N}$ is a family of closed subspaces of $\mathcal{D}(A^{1/2})$.

To see the importance of this generalization and describe the ideas behind the self-adaptive strategy, we shall consider the case that A is a differential operator on some domain Ω. Let Ω be divided into k disjoint subdomains Ω_i, $1 \leqslant i \leqslant k$, and let N_i, $1 \leqslant i \leqslant k$, be k nonnegative integers. Set $\mathcal{K}_j = \{u \in \mathcal{D}(A^{1/2}) | u(x) = 0 \text{ for } x \in \cup_{N_i < j} \Omega_i\}$, the extended concept of dimensionally reduced solutions with this family $\{\mathcal{K}_j\}_{j=0}^{N}$, $N = \max_i\{N_i\}$, is one that permits different order of the dimensionally reduced solution in different parts of the domain Ω. This is extremely important for practical applications, where a low order dimensionally reduced solution may very well be satisfactory in the interior of the domain and away from singularities in the loads and at the same time a high order solution is required near

the boundary or near singularities. As a total estimator for the error, let us use an expression of the form

$$\left(\sum_{i=1}^{k} [\eta_i(N_i)]^2 \right)^{1/2}, \tag{19}$$

where $\eta_i(N_i)$ refers to some estimator on the domain Ω_i with respect to dimensional reduction of order N_i. (η_i could for example be the estimator Est or the corrected estimator Est_2 of the previous section.)

To set a goal for the 'best' distribution of the orders $\{N_i\}_{i=1}^k$ for a dimensionally reduced solution we need the concept of cost. Let us assume that the cost of (solving) the dimensionally reduced problem with orders $\{N_i\}_{i=1}^k$ is given by $\Sigma_{i=1}^k(\beta N_i + 1)^\alpha m(\Omega_i)$, where α and β are two positive constants and $m(\Omega_i)$ is some measure of Ω_i.

As a 'best' distribution of the orders for a dimensionally reduced solution we define one which for a given cost minimizes the energy norm of the error. (We could also have defined a 'best' distribution as one that for a given value of the energy norm of the error minimizes the cost. Which of these two definitions we take makes no difference in the strategy we propose.)

The following is very heuristic in nature and by no means an exact verification that the strategy works. Let us use the expression (19) as if it were the exact norm of the error. Secondly, let us assume this expression to be defined for all positive values of the N_i's, and not only integers. By introduction of Lagrangean multipliers it is easily seen that a 'best' distribution of the N_i's has to satisfy

$$\exists C \text{ (independent of } i\text{) such that } \frac{(\partial/\partial N_i)\left([\eta_i(N_i)]^2\right)}{(\beta N_i + 1)^{\alpha-1} m(\Omega_i)} \sim C, \quad \forall\, i. \tag{20}$$

In practice we only have the values of the η_i's at integer points and a discrete equivalent of (20) is then

$$\frac{[\eta_i(N_i + 1)]^2 - [\eta_i(N_i)]^2}{(\beta N_i + 1)^{\alpha-1} m(\Omega_i)} \sim C, \quad \forall\, i. \tag{21}$$

We shall also assume that $\eta_i(N_i + 1)$ is significantly smaller than $\eta_i(N_i)$, so that instead of (21) we get

$$\frac{[\eta_i(N_i)]^2}{(\beta N_i + 1)^{\alpha-1} m(\Omega_i)} \sim C, \quad \forall\, i. \tag{22}$$

The strategy we propose is one that aims at equilibrating the left-hand sides of (22). We do this in a way similar to the adaptive finite element solver F.E.A.R.S.; cf. [1]. Let us assume that we have arrived at a distribution $\{N_i^0\}_{i=1}^k$ and that the estimate for the error is unacceptably large. Our strategy is simply to find j such that

$$\delta_j = \frac{[\eta_j(N_j^0)]^2}{(\beta N_j^0 + 1)^{\alpha-1} m(\Omega_j)}$$

is maximal, and then increase N_j^0 by 1. In the next section we shall see how well this performs in a practical example.

6. A Numerical Example. Consider the same problem as in Section 4, namely

$$\left[\left(\frac{\partial}{\partial x}\right)^2 + \left(\frac{\partial}{\partial y}\right)^2\right] u^h = 0 \quad \text{in } \Omega^h =]0, 1[\times]-h, h[,$$

$$\frac{\partial u^h}{\partial y} = g \quad \text{for } y = h,$$

$$\frac{\partial u^h}{\partial y} = -g \quad \text{for } y = -h,$$

$$u^h = 0 \quad \text{for } x = 0, 1.$$

Let [0, 1] be divided into the four subintervals $I_i = [(i-1)/4, i/4]$, $1 \leqslant i \leqslant 4$. A dimensionally reduced solution can now have different order in the different intervals I_i, $1 \leqslant i \leqslant 4$. As basis functions we choose the polynomials introduced in Example 3.1.

TABLE 6.1

$2N_1$ \ $2N_2$	0	2	4	6
0	0.3974 — 1	0.3257 — 4.5	0.3168 — 14	0.3160 — 32.5
2	0.3662 — 4.5	0.0730 — 8	0.0721 — 17.5	0.0720 — 36
4	0.3517 — 14	0.0383 — 17.5	0.0270 — 27	0.0270 — 45.5
6	0.3499 — 32.5	0.0292 — 36	0.0134 — 45.5	0.0134 — 64

error
—
work

$h = \frac{1}{2}$

$2N_1$ \ $2N_2$	0	2	4	6
0	0.3014 — 1	0.2296 — 4.5	0.2270 — 14	0.2267 — 32.5
2	0.2540 — 4.5	0.0359 — 8	0.0359 — 17.5	0.0359 — 36
4	0.2484 — 14	0.0134 — 17.5	0.0130 — 27	0.0130 — 45.5
6	0.2479 — 32.5	0.0063 — 36	0.0055 — 45.5	0.0055 — 64

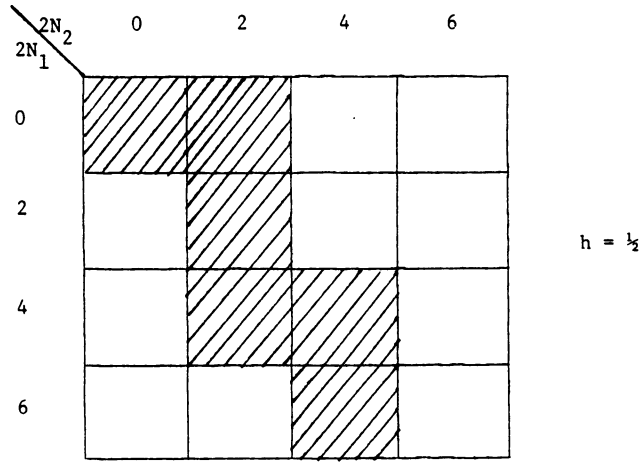
error
—
work

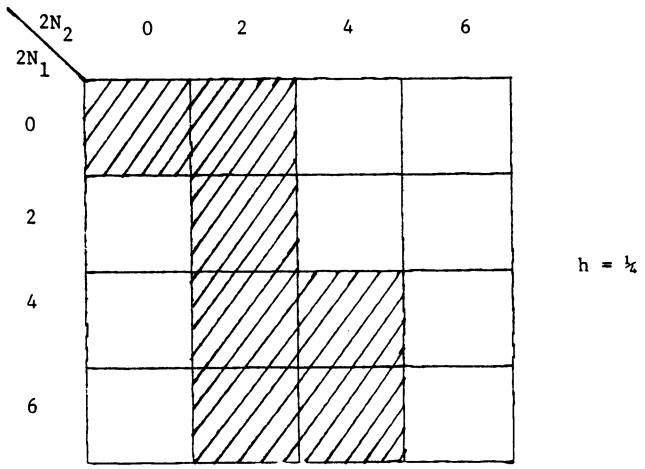
$h = \frac{1}{4}$

The equations that define the dimensionally reduced solutions are solved by introducing a finite element discretization in the x-direction. Piecewise linear functions on a regular mesh are used as test and trial functions for the finite element method. Since we want to illustrate the behavior of the dimensional reduction, and are here not interested in any contribution from the x-discretization, we choose a very fine grid of meshsize $= 2^{-9}$. The involved linear equations are solved by a Cholesky decomposition combined with iterative refinement. In the computations that we present here $g(x)$ is chosen $= \pi/4$. Since this choice of boundary data makes the problem symmetric in the line $x = \frac{1}{2}$ we only need consider x in the interval $[0, \frac{1}{2}]$. Let $2N_i$, $1 \leq i \leq 2$, denote the order of the dimensionally reduced solution in I_i. The above table shows the error on the whole interval $[0, 1]$ $(= |||e_{2\bar{N}}|||_E)$ and the work (here defined by $\frac{1}{2}(N_1 + 1)^3 + \frac{1}{2}(N_2 + 1)^3$) as a function of the pair $\bar{N} = 2\bar{N} = (2N_1, 2N_2)$ for two different values of h.

Based on the numbers in this table we can now find the entries with the property that the error is smaller than any error obtained with the same or less work. These entries are marked in the following table.

TABLE 6.2.





Tables 6.1 and 6.2 clearly illustrate the advantage of a nonuniform distribution of the polynomials. It is easy to see that the true solution u^h in the limit as $h \to 0$ has a parabolic behavior in the y direction, also for x in the middle of the interval [0, 1]. This is reflected in the fact that the pair (0, 2) is slightly better than (2, 0), it also explains the significant decrease in the error obtained by choosing the pair (2, 2). For the higher order polynomials there is a clear tendency towards concentration near the boundary $x = 0$ (and $x = 1$) in the entries marked in Table 6.2. This concentration is more visible the smaller h is; for $h = \frac{1}{2}$ the pair (6, 2) is not as good as (4, 4) but for $h = \frac{1}{4}$ the error obtained by (6, 2) is less than half the error by (4, 4) with only a slight increase in the work.

We now want to test the adaptive strategy outlined in Section 5 on this example. We consider the case $h = \frac{1}{4}$, where nonuniformity in the distribution of the polynomials is most advantageous. As an estimator we use Est_2 of Section 4, with the constant τ set to 10^2 and each interval I divided into 2 subintervals of length $\frac{1}{8}$. The following table shows the efficiency of Est_2 ($\mathrm{Eff}_2 = |||e_{2\bar{N}}|||_E / \mathrm{Est}_2$) as a function of the pair $2\bar{N} = (2N_1, 2N_2)$. It is evident from Table 6.3 that Est_2 provides a reliable estimate for the error even in the case of variable order. We also note that Est_2 is not necessarily an upper bound for the error, when the orders of the polynomials are allowed to vary. Steps could be taken to correct this, but on the other hand computational experience shows that this effect is insignificant, and that Est_2 is very close to an upper bound in most cases.

TABLE 6.3

$2N_1$ \ $2N_2$	0	2	4	6
0	0.94	0.99	1.00	1.00
2	1.06	0.68	0.68	0.68
4	1.09	0.60	0.59	0.59
6	1.09	0.57	0.53	0.53

Eff_2

$h = \frac{1}{4}$

Let us start with an initial distribution for the orders of the polynomials given by

$$(2N_1, 2N_2) = (0, 0).$$

Based on the present formula for the work and the error estimate, we now compute $\delta_j, j = 1, 2$, as in Section 5. The result is

$$\delta_1 = \delta_2 = 0.10.$$

We can therefore proceed to both (0, 2) and (2, 0). According to Table 6.1, (0, 2) is only slightly better than (2, 0), so this apparent "failure" of our strategy is of very little significance.

For the pair (0, 2) we compute

$$\delta_1 = 0.10, \quad \delta_2 = 0.14 \times 10^{-2}$$

and for the pair (2, 0)

$$\delta_1 = 0.31 \times 10^{-2}, \quad \delta_2 = 0.10.$$

In both of these two cases we are told to proceed to the distribution given by

$$(2, 2).$$

For this pair we get

$$\delta_1 = 0.14 \times 10^{-2} \quad \text{and} \quad \delta_2 = 0.59 \times 10^{-6},$$

i.e., if we want higher accuracy with dimensional reduction, our strategy selects the pair

$$(4, 2).$$

In this case

$$\delta_1 = 0.14 \times 10^{-3} \quad \text{and} \quad \delta_2 = 0.16 \times 10^{-5},$$

so that additional requirements to the accuracy will lead us to the distribution

$$(6, 2).$$

The path that our strategy goes through can schematically be represented as

$$\begin{array}{ccccccc} & & (0,2) & & & & \\ & \nearrow & & \searrow & & & \\ (0,0) & & & & (2,2) & \to (4,2) & \to (6,2) \\ & \searrow & & \nearrow & & & \\ & & (2,0) & & & & \end{array}$$

and based on Tables 6.1 and 6.2 this is clearly seen to be a very good choice. The strategy has been tried in a variety of other situations and has consistently been very effective. It has also been tried with different measures for the work. Here it should be noted that by changing the measure of the work we may entirely change the "best" distributions for the polynomials, but the strategy detects that easily.

7. Conclusions. In the following we list some conclusions concerning the approach of dimensional reduction developed in a series of three papers ([6], [7], and the present).

(a) It is common in engineering to distinguish between structures with large and small thickness (see, e.g., [3]). The approach presented here entirely avoids this somewhat artificial categorization.

(b) This approach gives, in an optimal and adaptive way, the advantages of asymptotic expansion (when the thickness is small) and the effectivity of the spectral or p-version methods (when the thickness is not small, or strong singularities are present). It has been shown that these two requirements uniquely characterize the approach.

(c) Reliable a posteriori error estimates can be obtained for this approach, and they lead immediately to an effective adaptive strategy.

(d) The approach is numerically very robust and works well independent of the thickness and the regularity of input data.

(e) The underlying mathematical theory and numerical experiments show the direction for various generalizations. These shall be dealt with elsewhere.

Courant Institute of Mathematical Sciences
New York University
New York, New York 10012

Institute for Physical Science and Technology
University of Maryland
College Park, Maryland 20742

1. I. BABUŠKA & W. C. RHEINBOLDT, "Reliable error estimation and mesh adaptation for the finite element method," *Computational Methods in Nonlinear Mechanics* (J. T. Oden, Ed.), North-Holland, Amsterdam, 1980, pp. 67–108.

2. V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Groningen, 1976.

3. V. DUNDER & S. RIDLON, "Practical applications of the finite element method," *ASCE J. Structures Division ST*1, January 1978, pp. 9–21.

4. N. DUNFORD & J. T. SCHWARTZ, *Linear Operators*, Part I, Interscience, New York, 1958.

5. M. VOGELIUS, Ph.D. Thesis, University of Maryland, December 1979.

6. M. VOGELIUS & I. BABUŠKA, "On a dimensional reduction method. I. The optimal selection of basis functions," *Math. Comp.*, v. 37, 1981, pp. 31–46.

7. M. VOGELIUS & I. BABUŠKA, "On a dimensional reduction methood. II. Some approximation-theoretic results," *Math. Comp.*, v. 37, 1981, pp. 47–68.