FINITE ELEMENT METHODS FOR NON-SELF-ADJOINT ELLIPTIC AND FOR HYPERBOLIC PROBLEMS: OPTIMAL APPROXIMATIONS AND RECOVERY TECHNIQUES

K. W. MORTON

NUMERICAL ANALYSIS REPORT 7/83

Based on lectures given at the IMA Conference on 'The Mathematical Basis of Finite Element Methods with Applications to Partial Differential Equations', Imperial College, London, 5th-7th January, 1983.

FINITE ELEMENT METHODS FOR NON-SELF-ADJOINT ELLIPTIC AND FOR HYPERBOLIC PROBLEMS: OPTIMAL APPROXIMATIONS AND RECOVERY TECHNIQUES

K. W. Morton
Dept. of Mathematics, University of Reading

1. INTRODUCTION

In this article we shall consider the progress that has been made in extending and developing the finite element method so that it may be applied to much wider classes of problems than that for which it was originally developed. Within the context of stress problems, where engineers originated many of the early ideas, the method could be based on an extremal principle - for the strain energy. Mathematically, such principles always lead to self-adjoint elliptic problems: alternative principles are therefore needed for more general problems. Some form of variational principle or weak formulation is usually available but the way forward is less clear-cut and the results can often be disappointing.

Just as with finite difference methods, it is not too difficult to devise approximation schemes for quite general problems which will converge as some mesh parameter tends to zero: one may even achieve the optimal order of convergence. But the marvellous economy and robustness on coarse meshes, which were key features of finite element methods in their original context, will be lost unless rather special steps are taken to preserve the optimal approximation properties and superconvergence properties which lay behind them.

How this may be done is the prime theme linking the three lectures on which this article is based. We shall concentrate on essential ideas, giving references where more details and more specific applications may be found.

Much of the author's thinking is based on work with students and colleagues which is gratefully acknowledged and which can be only partially referenced.

To fix our notation, let us begin by recalling the main properties of finite element approximations to the following self-adjoint problem for the real scalar function u: in classical form,

$$Lu = f \quad \text{in} \quad \Omega \tag{1.1a}$$

$$u = g \quad \text{on} \quad \Gamma_D, \quad \partial u/\partial n = 0 \quad \text{on} \quad \Gamma_N , \tag{1.1b}$$

where we will assume Ω is a bounded, open region of $\mathbb{R}$, with boundary composed of the two non-intersecting portions Γ_D and Γ_N, and L is a second order linear elliptic operator. In weak form, we write this:-
find $u \in H^1_E$ such that

$$A(u,w) = (f,w) \qquad \forall\, w \in H^1_{E_0} , \tag{1.2}$$

where $(\cdot,\cdot)$ denotes the usual L_2 inner product, H^1_E denotes elements of the usual Sobolev space $H^1(\Omega)$ which satisfy the essential boundary condition $u = g$ on Γ_D and $H^1_{E_0}$ denotes the associated subspace of elements satisfying the corresponding homogeneous condition; here the symmetric form $A(u,w)$ can be obtained formally from (Lu,w) by integrating by parts and applying the boundary conditions $\partial u/\partial n = 0$ on Γ_N and $w = 0$ on Γ_D. We define the

$$\text{energy norm,} \quad \|v\|_A := \{A(v,v,)\}^{\frac{1}{2}} , \tag{1.3}$$

its positive definiteness following from the coercivity of $A(\cdot,\cdot)$, i.e. the ellipticity of L (see below). In many fields of application, the basic physical principles are best expressed in the extremal principle form:-
find $u \in H^1_E$ to

$$\text{minimise} \quad A(v,v) - 2(f,v) \qquad \text{over} \quad H^1_E. \tag{1.4}$$

Variation of this expression clearly leads directly to (1.2).

Suppose now Ω can be precisely divided into elements, on which a finite element basis $\{\phi_i(\underline{x})\}$ can be constructed, and g is such that a conforming

trial space S_E^h can be defined as:-

$$H_E^1 \supset S_E^h := \{U(\underline{x}) = \sum U_j \phi_j(\underline{x}) \mid U = g \text{ on } \Gamma_D\} , \tag{1.5a}$$

together with a corresponding test space S_0^h given by

$$H_{E_0}^1 \supset S_0^h := \{V(\underline{x}) = \sum V_j \phi_j(\underline{x}) \mid V = 0 \text{ on } \Gamma_D\} . \tag{1.5b}$$

Then carrying out the minimisation (1.4) over S_E^h leads to the Galerkin approximation U to u given by:-

$U \in S_E^h$ such that

$$A(U,W) = (f,W) \qquad \forall\, W \in S_0^h. \tag{1.6}$$

Because of the conforming property, (1.2) is true with W substituted for w so that subtracting (1.6) from the result gives

$$A(u-U,W) = 0 \qquad \forall\, W \in S_0^h. \tag{1.7}$$

That is, the error $u-U$ is orthogonal to the test space S_0^h in the inner product defined by $A(\cdot,\cdot)$. Because of the symmetry of $A(\cdot,\cdot)$, the optimal approximation property of U follows immediately:-

$$\| u-U \|_A = \inf_{V \in S_E^h} \| u-V \|_A \quad . \tag{1.8}$$

This is the key property of finite element methods which we wish to carry over into more general problems: first to non-self-adjoint elliptic operators L, which we shall do in section 2 mainly by reference to diffusion-convection problems; and then to hyperbolic problems which we shall treat in section 4.

To appreciate the significance of (1.8), let us suppose L is the Laplacian operator so that $A(u,w) = (\underline{\nabla} u, \underline{\nabla} w)$ and (1.8) becomes

$$\int_\Omega | \underline{\nabla}(u-U)|^2 d\Omega = \inf_{V \in S_E^h} \int | \underline{\nabla}(u-V)|^2 d\Omega \ . \tag{1.9}$$

Suppose also that the simplest conforming elements are used, that is piecewise linear elements over triangles, with the U_j in (1.5a) corresponding to the nodal values at the triangle vertices. Then $\underline{\nabla} U$ is piecewise constant and (1.9)

shows that it is the best such piecewise constant approximation to $\underline{\nabla} u$ in the least squares sense. For practical calculations it is very often the field $\underline{\nabla} u$ rather than the potential u which is of most interest and the piecewise constant approximation can only be at best first order accurate at most points. However, we shall see in section 3 that a second order approximation can be constructed from U because of the superconvergence properties implied by (1.9): unlike the case of bilinear elements on rectangles, there are no points (such as the centroid) of each triangle where $\underline{\nabla} U$ is second order accurate and a simple recovery procedure is needed; moreover, this construction also depends on the arrangement of the triangles such that exactly six have each vertex in common.

2. NON-SELF-ADJOINT ELLIPTIC PROBLEMS

In weak form and for second order operators, these can be written as in (1.2): find $u \in H^1_E$ such that

$$B(u,w) = (f,w) \qquad \forall\, w \in H^1_{E_0}\,, \tag{2.1}$$

where now $B(\cdot,\cdot)$ is an unsymmetric bilinear form. The theory is simplest if we have homogeneous boundary conditions, while of course we wish to allow for inhomogeneous Dirichlet data in practice, as in (1.1b). Thus suppose the boundary and the Dirichlet data are smooth enough that g is the restriction to Γ_D of a function $G \in H^1(\Omega)$: it is sufficient, for example, that the boundary is locally Lipschitz continuous and $g \in L_2(\Gamma_D)$ - see Ciarlet (1978) or any similar text. Then (2.1) can be rewritten as: find $u^0 \in H^1_{E_0}$ such that

$$B(u^0,w) = (f^0,w) := (f,w) - B(G,w) \qquad \forall\, w \in H^1_{E_0}\,, \tag{2.2}$$

where $u = u^0 + G$ and we have introduced f^0 which lies in the dual space of $H^1_{E_0}$, to be denoted by $H^{-1}_{E_0}$. Existence and uniqueness of the solution to (2.2) results from the following lemma.

Lax-Milgram Lemma. Suppose $B(\cdot,\cdot)$ is a bilinear form on $H^1(\Omega) \times H^1(\Omega)$, where $H^1(\Omega)$ is equipped with the norm $\|\cdot\|_B$, and it is

(i) continuous, i.e. $\exists$ a constant K such that

$$|B(v,w)| \le K \|v\|_B \|w\|_B \qquad \forall\, u,w \le H^1_{E_0} \;; \tag{2.3a}$$

and (ii) coercive, i.e. $\exists$ a constant $\alpha > 0$ such that

$$B(u,u) \ge \alpha \|u\|_B^2 \qquad \forall\, u \in H^1_{E_0} \;. \tag{2.3b}$$

Then there exists a unique solution $u^0 \in H^1_{E_0}$ to (2.2) forevery $f^0 \in H^{-1}_{E_0}$.
(Note that if $\|\cdot\|_B$ is taken from the symmetric part of $B(\cdot,\cdot)$, i.e. $2(u,w)_B = B(u,w) + B(w,u)$ and $\|u\|_B^2 = (u,u)_B$, then $\alpha = 1$ and K measures the asymmetry of $B(\cdot,\cdot)$.)

Although we no longer have an extremal principle from which the Galerkin approximation can be derived, such an approximation can be defined directly from (2.1) or (2.2). The latter actually allows a larger class of g to be treated than that based on (1.5a). Thus we replace this definition by

$$H^1_E \supset S^h_E := \{U = G + V \,|\, V \in S^h_0\} \tag{2.4}$$

and then have the discrete problem:- find $U \in S^h_E$ such that

$$B(U,W) = (f,W) \qquad \forall\, W \in S^h_0 \tag{2.5}$$

and, just as with (1.7), we have

$$B(u-U,W) = 0 \qquad \forall\, W \in S^h_0. \tag{2.6}$$

Unfortunately, instead of (1.8) all we can now prove from (2.3) is the following: for any $V \in S^h_E$, setting $W = U-V$ in (2.6),

$$\alpha \|u-U\|_B^2 \le B(u-U,u-U) = B(u-U,u-V)$$

$$\le K \|u-U\|_B \, \|u-V\|_B \;.$$

Hence we have

$$\|u-U\|_B \le (K/\alpha) \inf_{V \in S^h_E} \|u-V\|_B \;. \tag{2.7}$$

This means that, while the same order of convergence in the norm $\|\cdot\|_B$ is obtained as in the self-adjoint case, the constant in the asymptotic rate may be greatly increased: moreover, the crucial superconvergence properties are lost.

Diffusion-convection problems

These form an important class of practical problems which illustrate the difficulties. They are of the following form:

$$-\underline{\nabla}\cdot(a\underline{\nabla}u - \underline{b}u) = f \quad \text{in} \quad \Omega \tag{2.8a}$$

$$u = g \quad \text{on} \quad \Gamma_D\,, \qquad \partial u/\partial n = 0 \quad \text{on} \quad \Gamma_N. \tag{2.8b}$$

Here a is a (positive) diffusion coefficient and $\underline{b}$ a convective velocity field which we shall assume is incompressible (i.e. $\underline{\nabla}\cdot\underline{b} = 0$). The corresponding bilinear form is

$$B(u,w) := (a\underline{\nabla}u,\ \underline{\nabla}w) + (\underline{\nabla}\cdot(\underline{b}u),w) \tag{2.9}$$

and to ensure its coercivity we assume that Γ_D includes all points of the boundary on which $\underline{b}\cdot\underline{n} < 0$, so that u is prescribed on the inflow boundary. Indeed, it is easy to see that

$$B(u,v) = (a\underline{\nabla}u,\underline{\nabla}u) + \tfrac{1}{2}\int_{\Gamma_N}(\underline{b}\cdot\underline{n})u^2 ds \qquad \forall\ u \in H^1_{E_0}\,. \tag{2.10}$$

Thus, if $a\in C^0(\Omega)$ and $\underline{b}\in[H^1(\Omega)]^2$, $B(\cdot,\cdot)$ clearly satisfies all the hypotheses of the Lax-Milgram Lemma.

Suppose we introduce the symmetric form

$$B_1(u,w) := (a\underline{\nabla}u,\underline{\nabla}w)\,, \tag{2.11}$$

with associated norm $\|\cdot\|_{B_1}$, and let U_1^* be the best fit to u in this norm from the trial space S_E^h : that is,

$$B_1(u-U_1^*,W) = 0 \qquad \forall\ W \in S_0^h. \tag{2.12}$$

Then for the Galerkin approximation U given by (2.5) we have, comparing (2.10) with (2.11),

$$\| u-U \|^2_{B_1} \le B(u-U,u-U) = B(u-U,u-U_1^*)$$

$$= B_1(u-U,u-U_1^*) + (\underline{b}\cdot\nabla(u-U),u-U_1^*)$$

$$\le \| u-U \|_{B_1} \{ \| u-U_1^* \|_{B_1} + \max_{\Omega} (|\underline{b}|/a) \| a^{\frac{1}{2}}(u-U_1^*) \| \} . \qquad (2.13)$$

From the Aubin-Nitsche duality argument (see earlier chapter on self-adjoint problems) it is easy to deduce that there exists a constant K, independent of $\underline{b}$, such that

$$\| a^{\frac{1}{2}}(u-U_1^*) \| \le Kh \| u-U_1^* \|_{B_1} , \qquad (2.14)$$

where h is the largest diameter of any element. It therefore follows that

$$\| u-U \|_{B_1} \le [1 + Kh \max_{\Omega} (|b|/a)] \| u-U_1^* \|_{B_1} . \qquad (2.15)$$

This is sharper than (2.7), showing the dependence on element size through the important dimensionless parameter, the mesh Péclet number bh/a.

A useful simple test problem in one dimension is:

$$-au'' + bu' = f \quad \text{on } (0,1) \qquad (2.16a)$$

$$u(0) = 0, \quad u(1) = 1, \qquad (2.16b)$$

where a and b are positive constants. For $f = 0$, the solution is easily seen to be

$$u(x) = (e^{bx/a} - 1)/(e^{b/a} - 1): \qquad (2.17)$$

piecewise linear elements on a uniform mesh of size h give the Galerkin equations for $j = 1,2,\ldots,J-1$ with $Jh = 1$

$$-\delta^2 U_j + (bh/a)\Delta_0 U_j = 0, \qquad (2.18)$$

where we have used the usual finite difference notation $\delta^2 U_j := U_{j+1} - 2U_j + U_{j-1}$, $\Delta_0 U_j := \frac{1}{2}(U_{j+1} - U_{j-1})$; these have the solution

$$U_j = (\mu_0^j - 1)/(\mu_0^J - 1), \qquad \mu_0 = (2+bh/a)/(2-bh/a). \qquad (2.19)$$

Clearly, the approximation for $bh/a > 2$ exhibits oscillations which bear no relation to the exponential solution (2.17) and the error bound (2.15) is seen to be quite realistic, the K in this case being calculated as $1/\pi$.

These spurious oscillations are a well-known result of the central differences yielded by the Galerkin method in (2.18). In difference methods they are overcome by using upwind differencing, replacing $\Delta_0 U_j$ by $\Delta_- U_j := U_j - U_{j-1}$ or by a weighted average of the two. The best-known scheme is that due to Allen & Southwell (1955) which with the average $(1-\xi)\Delta_0 + \xi\Delta_-$ can be written as

$$-[1+\tfrac{1}{2}\xi(bh/a)]\delta^2 U_j + (bh/a)\Delta_0 U_j = 0; \qquad (2.20)$$

for the choice

$$\xi = \coth(\tfrac{1}{2}bh/a) - (\tfrac{1}{2}bh/a)^{-1} \qquad (2.21)$$

we obtain the so-called exponentially-fitted scheme which gives exact nodal values for this model problems.

Petrov-Galerkin methods

The first finite element methods to overcome the deficiencies of the Galerkin method followed similar lines and used different weight functions from the trial functions ϕ_j with a view to generating these upwind difference schemes. In general, for a Petrov-Galerkin method we introduce a test space T_0^h different from but with the same dimension as the S_0^h of (1.56): with basis functions $\psi_j(\underline{x})$, usually over the same elements, we have

$$H_{E_0}^1 \supset T_0^h := \{V(\underline{x}) = \sum V_j\psi_j(\underline{x}) \mid V = 0 \text{ on } \Gamma_D\} . \qquad (2.22)$$

Then the Galerkin method of (2.5) is generalised to find $U \in S_E^h$ such that

$$B(U,W) = (f,W) \qquad \forall W \in T_0^h. \qquad (2.23)$$

The important question is "how should T_0^h be chosen for a given trial space S_E^h?". In particular, can it be done satisfactorily without reference to the upwind difference schemes that it might be induced to give on a regular mesh? There is a large literature concerned with the development of Petrov-Galerkin

methods for diffusion-convection problems and most approaches make some use of this idea: see, for example, the conference proceedings Hughes (1979). We shall however follow Barrett & Morton (1980, 1981) and Morton (1982) in basing our approach on symmetrizing the bilinear form $B(\cdot,\cdot)$.

Suppose $B_S(\cdot,\cdot)$ is a symmetric bilinear form giving an inner product and norm $\|\cdot\|_{B_S}$ on $H^1_{E_0}$ with respect to which $B(\cdot,\cdot)$ satisfies the hypotheses (2.3a) and (2.3b) of the Lax-Milgram Lemma. Then, for any fixed w, $B(u,w)$ is a bounded linear functional of u and by the Riesz Representation Theorem can be written $B_S(u,Rw)$ where Rw is an element of $H^1_{E_0}$: indeed, by the linearity of $B(\cdot,\cdot)$ and (2.3), R is a linear operator on $H^1_{E_0}$ and we can write

$$B(u,w) = B_S(u,Rw) \qquad \forall u,\ w \in H^1_{E_0} . \tag{2.24}$$

Effectively, R is a symmetrizer for $B(\cdot,\cdot)$. Note too that the coercivity condition (2.3b) ensures that R is invertible on $H^1_{E_0}$. Now for the Petrov-Galerkin approximation given by (2.23) we have, corresponding to (2.6),

$$B(u-U,W) = 0 \qquad \forall W \in T^h_0 \tag{2.25}$$

which by (2.24) we can now write as

$$B_S(u-U,RW) = 0 \qquad \forall W \in T^h_0 . \tag{2.26}$$

The extent to which this leads to the optimal approximation property of (1.7) and (1.8) is then given by the following theorem.

Theorem (Morton, 1982). Suppose the test space T^h_0 has the same dimension as S^h_0 and that there exists a constant $\Delta\in[0,1)$ such that

$$\inf_{W\in T^h_0} \|V-RW\|_{B_S} \le \Delta \|V\|_{B_S} \qquad \forall V \in S^h_0. \tag{2.27}$$

Then there exists a unique solution $U^0 \in S^h_0$ to (2.23) for every $f^0 \in H^{-1}_{E_0}$ and the error between U^0 and the solution u^0 of (2.2) satisfies

$$\|u^0 - U^0\|_{B_S} \le (1-\Delta^2)^{-\frac{1}{2}} \inf_{V \in S^h_0} \|u^0-V\|_{B_S} . \tag{2.28}$$

The effect of inhomogeneous Dirichlet data is dealt with as described at the beginning of this section. The result (2.28) is also shown in the above reference to include and be somewhat sharper than that of the Generalised Lax-Milgram theorem of Babuška & Aziz (1972).

This theorem in principle enables an error bound to be calculated for any Petrov-Galerkin method, so long as sufficient knowledge of R is available for the approximation result (2.27) to be obtained. Note that this result holds for all data f and if (2.27) is sharp then so is (2.28): however for specific data (2.28) may not be particularly sharp.

This framework also allows two alternative approaches to the task of constructing effective Petrov-Galerkin methods. The first, conventional, approach is to construct basis functions ψ_i of T_0^h in such a way that the constant Δ in (2.27) is small but also so as to have local support so that the stiffness matrix $B(\phi_j,\psi_i)$ resulting from their substitution in (2.23) is easily evaluated and has small bandwidth. Note however that this matrix will be unsymmetric and the solution of (2.23) correspondingly more difficult than that of the Galerkin equations for a self-adjoint problem. The alternative approach is based on the ideal test functions ψ_i^* which are the solution of the equations

$$R\psi_i^* = \phi_i \qquad \forall \phi_i \in S_0^h \ . \tag{2.29}$$

Use of this test space would give $\Delta = 0$ in (2.27) and yield the optimal approximation to u^0 in the $\|\cdot\|_{B_S}$ norm, i.e. that which achieves the infimum on the right-hand side of (2.28): so we would have completely achieved our original objective. Moreover, the Petrov-Galerkin equations (2.23) could then be written for this $U^* \in S_0^h$ as

$$B_S(U^*,\phi_i) = (f,\psi_i^*) - B(G,\psi_i^*) \qquad \forall \phi_i \in S_0^h \ , \tag{2.30}$$

where as before G is the extension of the boundary data. These equations have the practical advantage of being symmetric: indeed they are the same as the Galerkin equations for a self-adjoint problem corresponding to $B_S(\cdot,\cdot)$. What

remains is to approximate ψ_i^* sufficiently well for the right-hand side of (2.30) to be calculated to adequate accuracy. This is a linear functional of ϕ_i which we will write as

$$F^*(\phi_i) := (f, R^{-1}\phi_i) - B(G, R^{-1}\phi_i) \qquad \forall \phi_i \in S_0^h . \tag{2.31}$$

Suppose now that this is approximated by $F(\phi_i)$, an approximation for which we can establish the error bound

$$|F(V) - F^*(V)| \leq \delta_F \|V\|_{B_S} \qquad \forall V \in S_0^h . \tag{2.32}$$

Then the corresponding approximation U^0 to u^0 is given by

$$B_S(U^0, \phi_i) = F(\phi_i) \qquad \forall V \in S_0^h \tag{2.33}$$

and clearly satisfies

$$\|U^*-U^0\|_{B_S}^2 = |F^*(U^*-U^0) - F(U^*-U^0)| \leq \delta_F \|U^*-U^0\|_{B_S} . \tag{2.34}$$

There results, using the optimality property of U^*, the error bound

$$\|u^0-U^0\|_{B_S}^2 \leq \|u^0-U^*\|_{B_S}^2 + \delta_F^2 . \tag{2.35}$$

That is, we have a term added to the optimal error estimate as compared with the multiplicative factor of (2.28); and, moreover, this term can be determined for specific data. Note that a data independent error bound may also be obtained if required but that it will still be additive: for example, suppose $F(\phi_i)$ is computed through a linear operator $T: S_0^h \to H_{E_0}^1$, approximating R^{-1},

$$\begin{aligned} F(\phi_i) &:= (f, T\phi_i) - B(G, T\phi_i) \\ &= B(u^0, T\phi_i) = B_S(u^0, RT\phi_i) ; \end{aligned} \tag{2.36}$$

then we have

$$\begin{aligned} |F(V) - F(V^*)| &= |B_S(u^0, RTV - V)| \\ &\leq \|[I-(RT)^*]u^0\|_{B_S} \|V\|_{B_S} \qquad \forall V \in S_0^h . \end{aligned} \tag{2.37}$$

where $(RT)^*$ is the adjoint of RT in the inner product $B_S(\cdot,\cdot)$.

Application to diffusion-convection.

We conclude this section by outlining the application of these ideas to the diffusion-convection problems given by (2.8) and (2.16): more details can be found in Barrett & Morton (1983) and the references therein. The operator in (2.8a) can be factored to give

$$L_1^* L_2 u = f, \tag{2.38}$$

where

$$L_1 v := a^{\frac{1}{2}} \underline{\nabla} v \ , \quad L_2 v := a^{\frac{1}{2}} \underline{\nabla} v - (\underline{b}/a^{\frac{1}{2}}) v$$

and L_1^* is the formal adjoint of L_1. This suggests two distinct symmetric bilinear forms, one based on L_1 and one on L_2. We have already introduced the former and called it $B_1(\cdot,\cdot)$ in (2.11): we now define the second choice by

$$B_2(v,w) := (a\underline{\nabla} v, \underline{\nabla} w) + ((b^2/a)v, w) \tag{2.39a}$$

$$= (L_2 v, L_2 w) + \int_\Gamma (\underline{b} \cdot \underline{n}) v w ds. \tag{2.39b}$$

Let us denote by R_1 and R_2 the Riesz representer in (2.24) and by Δ_1 and Δ_2 the smallest constant in the approximation estimate (2.27) in these two cases. Then a little manipulation shows that these constants are given by the following discrete minimisation problems:

$$1 - \Delta_m^2 = \min_{\underline{V}} \left\{ \frac{\underline{V}^T B^T A^{-1} B \underline{V}}{\underline{V}^T C \underline{V}} \right\} \ , \quad m = 1,2 \tag{2.40}$$

where the matrices A,B,C have components given by

$$A_{ij} = B_m(R_m \psi_j, R_m \psi_i)$$

$$B_{ij} = B_m(\phi_j, R_m \psi_i) = B(\phi_j, \psi_i)$$

$$C_{ij} = B_m(\phi_i, \phi_j)$$

and $$\underline{V}^T = \{V_1, V_2, \ldots, V_N\}, \quad V = \sum_1^N V_j \phi_j \in S_0^h.$$

For the model problem (2.16), these have been calculated explicitly for several test spaces and piecewise linear trial spaces by Scotney (1982).

Symmetrization with $B_1(\cdot,\cdot)$

In one dimension with constant a, the best piecewise linear fit in the norm $\|\cdot\|_{B_1}$ is exact at the nodes. Thus it is natural to analyse methods based on finite difference arguments under the $\|\cdot\|_{B_1}$ norm. The Riesz representer R_1 for the model problem (2.16) can be written explicitly as

$$(R_1 w)(x) = w(x) + (b/a)\int_0^x (w(t) - \bar{w})dt, \tag{2.41}$$

where $\bar{w} = \int_0^1 w(t)dt$, which shows its non-local character. For the same problem the earliest upwind test functions were those due to Christie et al. (1976) and Heinrich et al. (1977): if $\phi_i(x)$ are the piecewise linear basis functions, typical of such test functions are

$$\psi_i(x) := \phi_i(x) + \alpha\sigma_i(x) \tag{2.42a}$$

where

$$\sigma_i(x) := \begin{cases} 3(x-x_{i-1})(x_i-x)/(x_i-x_{i-1})^2 & x_{i-1} \leq x \leq x_i \\ -3(x_{i+1}-x)(x-x_i)/(x_{i+1}-x_i)^2 & x_i \leq x \leq x_{i+1}. \end{cases} \tag{2.42b}$$

On a uniform mesh, setting the parameter α equal to ξ defined in (2.21) leads to the Allen & Southwell difference operator.

With variable coefficients local values of α are used and, in two dimensions if bilinear elements on rectangles are used, the trial basis functions are the product functions $\phi_i(x)\phi_j(y)$ which are matched with product test functions $\psi_i(x)\psi_j(y)$ with the two parameters α based on the x and y components of $\underline{b}$.

An alternative approach is that due to Hughes & Brooks (1979, 1982): their streamline diffusion method starts from regarding the Allen & Southwell scheme as written in (2.20) as enhancing the diffusion in the direction of the flow vector $\underline{b}$. Thus the scalar diffusion coefficient a of (2.8a) is replaced by the tensor diffusivity with components

$$A_{\ell m} = a\,\delta_{\ell m} + \tilde{a}\,b_\ell\,b_m \tag{2.43a}$$

where $$\tilde{a} = \tfrac{1}{2}(\xi_1 b_1 h_1 + \xi_2 b_2 h_2) \qquad (2.43b)$$

and b_1, b_2 are the components of $\underline{b}$ along the sides of a rectangular element of sides h_1, h_2 : ξ_1, ξ_2 are corresponding values of the parameter (2.21). If this modified operator is used with the Galerkin method and bilinear elements, it can be shown that it is equivalent to using a Petrov-Galerkin method with the test functions

$$\psi = \phi + (\tilde{a}/|b|^2)\underline{b}\cdot\underline{\nabla}\phi \quad . \qquad (2.44)$$

These are discontinuous and therefore non-conforming elements. So the terms in the bilinear form corresponding to $(a\underline{\nabla}\phi, \underline{\nabla}\psi)$ have to be evaluated element by element and also the error analysis leading to (2.28) does not strictly apply. Nevertheless evaluation of (2.40) for these two test functions (2.42) and (2.44) does show how effectively these Petrov-Galerkin methods overcome the deficiencies of the pure Galerkin method. The results obtained by Scotney (1982) are given in Table 1.

bh/a	Galerkin	Heinrich et al	Hughes & Brooks
2	1.1547	1.0060	1.0924
5	1.7559	1.0468	1.1509
50	14.468	1.2022	1.1547
500	144.34	1.2344	1.1547
10^5	28868	1.2383	1.1547

TABLE 1 : Ratios of Petrov-Galerkin error to optimal error given by $(1-\Delta_1^2)^{-\frac{1}{2}}$ - see (2.28) and (2.40).

The optimal test space for the model problem under $B_1(\cdot,\cdot)$ was used by Hemker (1977), though derived in a different way and with a local basis. The inversion of R_1 to calculate the ψ_i^* of (2.29) gives rise to rather awkward exponentials which are difficult to handle in the formulation (2.30): but Hemker's test functions are quite simple in form (though still difficult to evaluate), namely

$$\psi_i(x) := \begin{cases} [1-e^{-b(x-x_{i-1})/a}]/[1-e^{-b(x_i-x_{i-1})/a}] & , \quad x_{i-1} \le x \le x_i \\ [e^{-b(x-x_i)/a} - e^{-b(x_{i+1}-x_i)/a}]/[1-e^{-b(x_{i+1}-x_i)/a}], & x_i \le x \le x_{i+1}. \end{cases} \tag{2.45}$$

These again give rise to the Allen & Southwell difference approximation to $-au'' + bu'$ but now sample the right-hand side f of (2.16a) so as to always give exact nodal values. Unfortunately it is much more difficult to extend these test functions into two dimensions and this has not yet been satisfactorily achieved.

Symmetrization with $B_2(\cdot,\cdot)$

With its lack of dependence on b/a it is not clear that $\|\cdot\|_{B_1}$ is an appropriate norm, especially for singular perturbation problems. The alternative bilinear form $B_2(\cdot,\cdot)$ defined in (2.39) has therefore been used by Barrett & Morton (1980, 1981, 1983) in their work. For the model problem, R_2^{-1} now has a simpler form than R_2. Thus the solution of (2.29) can be written explicitly as

$$\psi_i^*(x) = \phi_i(x) + (b/a)\int_x^1 [\phi_i(t) - ce^{-bt/a}]dt, \tag{2.46}$$

where the constant c is such as to ensure that $\psi_i^*(0) = 0$. Moreover it is easy to use the symmetric form of the equations given by (2.30) which becomes

$$B_2(U_2^*,\phi_i) = (f_2,\phi_i) - cb[u(0) - e^{-b/a}u(1)] \quad , \tag{2.47}$$

where

$$f_2(x) = f(x) + (b/a)[F(x) - \bar{F}] \quad , \tag{2.48}$$

$F(x) = \int_0^x f(t)dt$ and $\bar{F}$ is the average of F with weighting function $e^{-bx/a}$. These formulae are generalised to variable coefficients and Neumann boundary conditions in the above references where examples of their use are given as well as sharp error bounds of the form (2.35): see also Rheinhardt (1982). In two dimensions it is unlikely that R_2^{-1} can be calculated explicitly and various approximate approaches have been tried: several work well for limited classes of problem but at the moment a direct approach to (2.30) appears to be the most generally successful.

A distinctive feature of the optimal piecewise linear approximations in the $\|\cdot\|_{B_2}$ norm obtained by these methods is that steep boundary layers appear as damped oscillations at the mesh frequency. This is because for large Péclet numbers $\|\cdot\|_{B_2}$ tends to the L_2 norm. How sub-gridscale information can be recovered from these results will be discussed in the next section.

3. SUPERCONVERGENCE AND OPTIMAL RECOVERY

As we have seen, one of the main features of a finite element approximation is its optimal, or almost optimal, approximation property in an energy norm, as in (1.8) or (2.28). We then have the problem of recovering from this pointwise estimates of the solution u or its derivatives. Clearly one could use corresponding point values of the approximation U or its derivatives. This is seldom very efficient, for one usually has more a priori knowledge of u than was used in constructing U - such as extra smoothness - and by using this one can achieve much more.

As a simple starting point consider the trivial problem :-

$$-u'' = f \quad \text{on} \quad (0,1) \quad \text{with} \quad u(0) = u(1) = 0. \tag{3.1}$$

Let $U(x) = \sum U_j \phi_j(x)$ be the piecewise linear Galerkin approximation on a non-uniform mesh with points x_j. Then the Galerkin equations (1.7) reduce to

$$\int_0^1 (u'-U')\phi_j' dx = 0, \quad \text{i.e.} \quad \frac{\Delta_- \epsilon_j}{\Delta_- x_j} - \frac{\Delta_- \epsilon_{j+1}}{\Delta_- x_{j+1}} = 0 \quad , \tag{3.2}$$

because ϕ_j' is piecewise constant, where $\epsilon_j = u(x_j) - U_j$. It follows from the boundary conditions that $\epsilon_j = 0$ for all j so that U has exact nodal values: note that this requires that the integrals in (f,ϕ_j) are evaluated exactly. To obtain values of u at intermediate points interpolation may be used: linear interpolation just reproduces the corresponding values of U; but if the smoothness of f implies continuity of higher derivatives of u, higher order interpolation using more nodal values will give greater accuracy - or at

least a higher order of accuracy. Interpolation theory also indicates how the derivative of u may be recovered to any order of accuracy allowed by its smoothness. Clearly U' itself is only first order accurate at most points: but it will be second order accurate at the mid-point of each interval - the simplest example of the phenomenon of superconvergence.

Another way of looking at the pointwise superconvergence of U, and indeed that which led Hemker to the test functions (2.45), is the following. For the problem (2.1) define the adjoint Green's function $G_{\underline{\xi}}$ by

$$B(v,G_{\underline{\xi}}) = (\delta_{\underline{\xi}},v) \qquad \forall\, v \in H^1_{E_0}\,, \tag{3.3}$$

where $\delta_{\underline{\xi}}$ is the delta function centred at $\underline{\xi}$. Then if U is the Petrov-Galerkin approximation given by (2.23) and (2.25) we have

$$\begin{aligned} u(\underline{\xi}) - U(\underline{\xi}) &= B(u-U,G_{\underline{\xi}}) \\ &= B(u-U,G_{\underline{\xi}}-W) \qquad \forall W \in T^h_0\,. \end{aligned} \tag{3.4}$$

Thus from (2.3a) we get,

$$|u(\underline{\xi}) - U(\underline{\xi})| \leq K \|u-U\|_B \, \|G_{\underline{\xi}}-W\|_B \qquad \forall W \in T^h_0. \tag{3.5}$$

In the case of (3.1), $B(\cdot,\cdot)$ is symmetric and we use the Galerkin method with piecewise linears: the crucial fact is that G_{ξ} is also piecewise linear, with a change of gradient at ξ. Thus if ξ is a node, G_{ξ} can be exactly matched from $S^h_0 = T^h_0$ and $U(G)$ is exact. For the model problem (2.16) the Green's function has exponential form and this led to the choice of test functions (2.45) to obtain exact nodal values.

In two dimensions (or for more complicated problems in one dimension) both of these arguments break down and one cannot achieve exact nodal values: the best piecewise linear fit in the Dirichlet norm (1.9) no longer interpolates at the nodes; and the Green's function is no longer piecewise linear. But much of value can be achieved, especially for the gradients or fluxes, by pointwise sampling.

Superconvergence of gradients for Poisson's equation.

For a dozen years or more use has been made of the experimentally observed fact that for bilinear elements over rectangles the gradient has exceptional accuracy at the centroid of each element. Subsequently in a series of papers Zlamal (1977, 1978) and LeSaint & Zlamal (1979) have shown rigorously that the bilinear element is superconvergent at the centroids and that similar higher order elements are superconvergent at corresponding Gauss points: moreover, this is true for more general self-adjoint equations and for mildly non-rectangular quadrilaterals. Meanwhile various corresponding results have been hypothesised for linear elements over triangles but nothing had been established clearly until quite recently. Now in a report which has yet to be published Levine (1983) has clearly set out the true situation. His results provide a good illustration of the recovery problem in a relatively simple situation: the methods of proof that he used are based on those of Zlamal so we begin by outlining these.

Consider the approximation of Poisson's equation using bilinear elements on rectangles of diameter h. Let u be the exact solution and u^I its interpolant by bilinears. Then writing $a(u,w)$ for $(\underline{\nabla}u,\underline{\nabla}w)$ the first result to be established is that, for some constant C,

$$a(u-u^I,W) \leq Ch^2|u|_{3,\Omega}\,|W|_{1,\Omega} \qquad \forall W \in S_0^h\,, \tag{3.6}$$

where $|\cdot|_{p,\Omega}$ denotes the p^{th} Sobolev semi-norm over Ω (L_2 norm of all p^{th} order derivatives). This is established through use of the Bramble-Hilbert lemma as follows: define the following linear functional for functions $\tilde{u}$ of the local co-ordinates (ξ,η) over the unit square S on which W is bilinear,

$$F(\tilde{u}) := \int_S \partial_\xi(\tilde{u}-\tilde{u}^I)\partial_\xi\tilde{W}d\xi d\eta\,; \tag{3.7a}$$

then it is easy to see that

$$|F(\tilde{u})| \leq C\,\|\tilde{u}\|_{3,S}\,\|\partial_\xi\tilde{W}\|_{0,S} \tag{3.7b}$$

where $\|\cdot\|_{p,S}$ denotes the p^{th} Sobolev norm over S; moreover, a little computation enables one to show that for any quadratic polynomial over S we have

$$F(q) = 0; \tag{3.7c}$$

it is then a direct result of the lemma that

$$|F(\tilde{u})| \leq C \ |\tilde{u}|_{3,S} \ \|\partial_\xi \tilde{W}\|_{0,S} \ . \tag{3.7d}$$

The required result (3.6) can then be obtained by scaling and summing over all rectangles. We next use this result with $W = U - u^I$, where U is the Galerkin approximation satisfying (1.7), to obtain

$$|U-u^I|^2_{1,\Omega} = a(U-u^I, U-u^I) = a(u-u^I, U-u^I)$$

$$\leq Ch^2|u|_{3,\Omega} \ |U-u^I|_{1,\Omega}$$

$$\text{i.e. } |U-u^I|_{1,\Omega} \leq Ch^2|u|_{3,\Omega} \ . \tag{3.8}$$

Thus U is an order of magnitude closer to u^I than it is to u, in the energy norm.

Suppose now that D_P is a sampling operator at a point P, for instance for the derivative ∂_x at the centroid of a rectangular element R. Then by a similar argument to the above, using the Bramble-Hilbert lemma and a computation for quadratic polynomials, one can show that

$$|D_P(u-u^I)| \leq Ch|u|_{3,R} \ . \tag{3.9}$$

Finally, by writing $D_P(u-U) = D_P(u-u^I) + D_P(u^I-U)$, using both (3.8) and (3.9) and summing over all rectangles R_j in Ω we obtain

$$\sum_{(j)} h^2 \ |D_{P_j}(u-U)|^2]^{\frac{1}{2}} \leq Ch^2|u|_{3,\Omega} \ . \tag{3.10}$$

This holds for ∂_x at points on the vertical bisector of each rectangle and for ∂_y at points on the horizontal bisectors: hence it holds for the gradient at the centroids and so confirms the superconvergence phenomenon in this ℓ_2 sense.

For linear elements over triangles, Levine (1983) has shown both theoretically and numerically that there are no points where the gradient is

superconvergent. However, suppose the triangulation is such that there are six triangles per vertex and the triangles can be grouped in pairs to form parallelograms with vertical diagonals and also in pairs to form parallelograms with horizontal diagonals. Then he has proved the conjecture of Strang & Fix (1973, p. 169) that the derivatives along the edges of the triangles are superconvergent at the mid-points. Moreover, he has shown that the average of the normal derivative in the two triangles either side of an edge is also superconvergent at the mid-point. Thus the gradient can be recovered to second order accuracy at the edge mid-points by this very simple device: averaging between the three mid-points of a triangle will also give the gradient to second order accuracy at the centroids. The proofs of these results follow similar lines to those of Zlamal, outlined above, but more constructive methods than the use of the Bramble-Hilbert lemma give sharper bounds for several of the results. Numerical experiments confirm the practical value of the results and the importance of the triangulation giving six triangles per node. The regularity of the mesh can otherwise be considerably relaxed and there is some hope that similar results can be proved in the supremum norm.

It is interesting to note that the recovery procedures described above, followed by interpolation, will often coincide with the use of divided differences of the Galerkin approximation as advocated, for instance, by Long & Morton (1977) and Thomée (1977). The analysis of the first reference, however, though also covering quadratic elements was essentially limited to regular meshes: that in the latter did not cover linear elements. Finally, before leaving this topic we should note complementary results of Douglas, Dupont & Wheeler (1974) and Wheeler (1974) for recovering the normal gradient at a Dirichlet boundary, which is often of very great practical importance: indeed, the procedure whose superconvergence is established in this reference was proposed for calculating boundary fluxes in heat-transfer problems by Wheeler (1973).

Optimal recovery

In their seminal paper, Golumb & Weinberger (1959) explored many of the basic ideas of optimal recovery which are pertinent to finite element methods: see also Micchelli & Rivlin (1976) for a more recent survey. The general situation is as follows: we are given the values of n linear data functionals $F_1(u), F_2(u), \ldots, F_n(u)$ of an unknown function u togehter with some (non-linear) constraint on u, such as $|u|_p \leq K$; then the problem is to define an optimal estimator for another linear functional $F(u)$, that is one with a minimal a priori error bound. As applied to finite element methods, the ideas are related to that of the hypercircle (Synge, 1955). For consider the problem (1.2), but with homogeneous boundary data, for $u \in H^1_{E_0}$: the Galerkin approximation U of (1.5a), (1.6) is determined from the data functionals

$$F_i(u) := A(u,\phi_i) = \langle f,\phi_i \rangle \qquad \forall \phi_i \in S^h_0 ; \tag{3.11}$$

and it is easy to check that it coincides with the centre $\bar{u}$ of Synge's hypercircle defined by

$$\|\bar{u}\|_A = \inf \{ \|v\|_A \mid v \in H^1_{E_0} \text{ s.t. } A(v,\phi_i) = (f,\phi_i) \ \forall \phi_i \in S^h_0 \} . \tag{3.12}$$

Now suppose $F(u)$ is to be estimated and define $\bar{y}$ by

$$F(\bar{y}) = \sup \{ |F(v)| \mid v \in H^1_{E_0} \text{ s.t. } \|v\|_A = 1, \ A(v,\phi_i) = 0 \ \forall \phi_i \in S^h_0 \} . \tag{3.13}$$

Then $F(U)$ is an optimal estimator of $F(u)$ with

$$|F(u) - F(U)|^2 \leq |F(\bar{y})|^2 \, [\|u\|^2_A - \|U\|^2_A], \tag{3.14}$$

given that the constraint on u is of the form $\|u\|_A \leq r$. This bound is sharp with $F(U)$ lying at the centre of the range of possible values for $F(u)$ obtained by taking $u = U + \alpha(r^2 - \|U\|^2_A)\bar{y}$ with $|\alpha| \leq 1$. It is important to note that although $\bar{y}$ depends on F, U is quite independent of the linear functional to be estimated.

For example, consider the one-dimensional self-adjoint problem

$$\begin{aligned} -(pu')' + qu &= f \qquad \text{on } (a,b) \\ \text{with} \qquad u(a) &= u(b) = 0 \end{aligned} \tag{3.15}$$

and $p > 0$, $q \geq 0$. Suppose we wish to estimate $u(\xi) =: F(u)$ for $\xi \in (a,b)$. Then it is easy to check that $\bar{y}$ is the difference $G(x,\xi) - G^h(x,\xi)$ of the Green's function from its best fit from S_0^h. We then obtain

$$|u(\xi) - U(\xi)|^2 \leq [G(\xi,\xi) - G^h(\xi,\xi)][\,\|u\|_A^2 - \|U\|_A^2], \tag{3.16}$$

which is actually the same as (3.5) in this case. When the trial space is piecewise linear, one can deduce that the nodal values are optimal sampling points (in the sense of giving a minimal error bound) although in general one will still obtain only first order accuracy unless stronger smoothness hypotheses are made on u than merely the boundedness of $\|u\|_A$: this indicates how singular was the situation of $p = 1$, $q = 0$ covered by (3.2).

This last point also shows the limitations of this framework: for it is not clear how one can exploit any greater smoothness, that u is known to possess, in the estimation of $F(u)$; we shall take this up in the next sub-section on defect correction. In the meantime let us consider the other extreme, appropriate to the diffusion-convection problems of section 2 and the hyperbolic problems of the next section, where the solution u may be far from smooth and typified by $|q| >> |p|$ in (3.15). The Galerkin approximation to (3.15) gives the best fit from the trial space $S_0^h = \text{span}\,\{\phi_j\}$ in the energy norm

$$\{\int_a^b (pu'^2 + qu^2)dx\}^{\frac{1}{2}}; \tag{3.17}$$

in the limit $p \to 0$ this becomes the weighted L_2 norm. For the recovery problem in this limit there are two natural sets of data functionals, the moment functionals

$$F_i^M(u) := \int_a^b qu\phi_i dx \tag{3.18a}$$

and the point functionals given by the nodal parameters of the best fit

$$F_i^P(u) := U_i. \tag{3.18b}$$

Barrett, Moore & Morton (1983) consider the local recovery problem for both types: that is, the problem corresponding to classical interpolation of approximating u from n consecutive values of either (3.18a) or (3.18b).

They show that for piecewise linear ϕ_i there exists a unique $(n-1)^{th}$ degree polynomial P_{n-1} with the same set of data functional values and that if $u \in H^n(I)$ then

$$|u(x) - P_{n-1}(x)| \leq Ch^n |u|_{n,I} , \qquad x \in I, \tag{3.19}$$

where h is the maximal node spacing and I is the union of the support of the basis functions ϕ_i which are involved. Moreover, if $u \in H_I^{n+1}$ there are n points of superconvergence in I where an extra order of accuracy is achieved. These results generalise and extend those given by Bramble & Schatz (1977). As a particular case they include the following well-known result: for $n = 3$, $q = 1$ and on a uniform mesh, the parabola which has the point functionals U_{j-1}, U_j and U_{j+1} yields the superconvergent recovery result

$$|u(x_j) - \frac{1}{12}[U_{j-1} + 10U_j + U_{j+1}]| \leq \frac{1}{360} h^4 |u^{(iv)}|. \tag{3.20}$$

Such local recovery formulae are of great practical value in extracting the best results from Galerkin approximations and have been used by a number of authors. They hold also for $p \neq 0$ so long as $p/q = O(h^2)$: thus they apply to the $\|\cdot\|_{B_2}$ norm of (2.39a) used by Barrett & Morton (1980) in the diffusion-convection problem, for any moderately large mesh Peclet number; these authors also used special formulae based on exponentials rather than polynomials for recovery of boundary layers much narrower than the mesh spacing.

Defect correction

For general p,q in (3.15), however, local recovery is not possible: and direct global recovery would often be prohibitively expensive. So consider the following approach based on assuming that $u \in H^2(a,b)$. Introduce the bilinear form

$$\tilde{A}(v,w) = \int_a^b (pv''w'' + qv'w')dx \tag{3.21}$$

and the representers χ_i of the data functionals F_i^M in this form:-

$$F_i^M(u) := \int_a^b (pu'\phi_i' + qu\phi_i)dx \equiv A(u,\phi_i) = \tilde{A}(u,\chi_i), \tag{3.22}$$

where we will take the ϕ_i as the piecewise linears. Then applying the hypercircle result to this form, the centre of the hypercircle for the optimal estimation problem in which the $F_i^M(u)$ are given is an element $\tilde{U}$ in the span of these representers given by

$$A(u - \tilde{U}, \phi_i) = 0 \qquad \forall i, \tag{3.23a}$$

$$\text{i.e.} \quad \tilde{A}(u - \tilde{U}, \chi_i) = 0 \qquad \forall i. \tag{3.23b}$$

That is, we have a Petrov-Galerkin approximation in the original bilinear form $A(\cdot,\cdot)$ which is also a Galerkin approximation in the new form $\tilde{A}(\cdot,\cdot)$. When p and q are constants it is clear that $\chi_i'' = -\phi_i$ and hence that the χ_i are <u>natural cubic splines</u>: thus from an error analysis of either (3.23b) or of (3.23a) (using the theorem of (2.27) and (2.28) we find that $\tilde{U}$ achieves fourth order accuracy when $u \in H^4(a,b)$.

Now solving either of the forms (3.23) directly is a completely separate computation from calculating the piecewise linear Galerkin approximation U. So instead we consider the calculation of $\tilde{U}$ as a recovery operation: it is more convenient to use cubic splines even for variable p,q so we define a natural cubic spline U^C by

$$A(U - U^C, \phi_i) = A(u - U^C, \phi_i) = 0 \qquad \forall i \tag{3.24}$$

$$U^C(a) = U^C(b) = 0. \tag{3.24}$$

Let P be the nodal interpolatory mapping from piecewise linears to natural cubic splines. Then we define the iteration for piecewise linears $U^{(\ell)}$ by

$$A(U^{(\ell+1)}, \phi_i) = A(U^{(\ell)}, \phi_i) - A(PU^{(\ell)} - U, \phi_i) \qquad \forall i$$

$$U^{(\ell+1)}(a) = U^{(\ell+1)}(b) = 0, \tag{3.25}$$

with $U^{(0)} = U$. The last term of (3.25) can be rewritten as

$$A(PU^{(\ell)}, \phi_i) - (f, \phi_i)$$

which reveals its rôle as a 'defect' in the calculation of the Petrov-Galerkin approximation U^C to u and (3.25) as a defect correction technique.
In Barrett Moore & Morton (1983) it is shown that the convergence factor is

$O(h)$ if $p' \in L^\infty(a,b)$ and $O(h^2)$ if p is constant. Thus one or two iterations, which we note involve only the original stiffness matrix $A(\phi_j,\phi_i)$ for the piecewise linears, are sufficient to obtain the full fourth order accuracy of U^c.

4. HYPERBOLIC PROBLEMS

At first sight hyperbolic problems offer an unpromising field for the use of finite element methods and, indeed, finite difference methods continue their domination in practical problems with perhaps the strongest challenge at the moment coming from spectral methods. This inauspicious prospect is because of a lack of useful variational principles and the fact that many of the phenomena are local in character and less suitable for global approximation. To accentuate the difference from elliptic problems we shall exclude from our consideration the sort of steady hyperbolic problems that occur in supersonic gas flow. Thus we shall assume that the time t is one independent variable and consider first order systems of the form

$$\underset{\sim}{u}_t + L(\underset{\sim}{u}) = 0, \tag{4.1}$$

where $\underset{\sim}{u} = \underset{\sim}{u}(\underline{x},t)$ is a vector of unknowns, the subscript t or the operator ∂_t denotes partial differentiation with respect to t and L is a (generally non-linear) operator involving the first order spatial derivatives.

Then the first choice is whether to use finite element approximation in time as well as space. We shall not do so but use finite differences in the time variable. This is partly for simplicity and flexibility: but it is mainly because with finite elements we seek approximations which are optimal in an integral norm and this would seem to be more appropriate in just the space variables rather than in space-time. We shall moreover concentrate on one-step methods in time and indeed, mainly on the explicit Euler method, with which much can be achieved: the methods derived can then be extended to implicit methods or, by predictor-corrector or Runge-Kutta schemes, to higher order methods. We

shall say little about boundary conditions in this section.

Petrov Galerkin methods

These have been widely used to overcome the disadvantages of pure Galerkin methods, just as with non-self-adjoint elliptic problems. Thus at time-step n we write the approximation as

$$\underset{\sim}{U}^n(\underline{x}) = \sum_{(j)} \underset{\sim}{U}_j^n \phi_j(\underline{x}) \tag{4.2}$$

in terms of trial space basis functions $\phi_j(\underline{x})$. Then the Petrov-Galerkin method for (4.1) based on Euler time-stepping and test space basis functions $\psi_i(\underline{x})$ has the form

$$\left(\frac{\underset{\sim}{U}^{n+1} - \underset{\sim}{U}^n}{\Delta t} + L(\underset{\sim}{U}^n), \psi_i \underset{\sim}{e}_{(k)} \right) = 0 \qquad \forall i,\ k \tag{4.3}$$

where the vector $\underset{\sim}{e}_{(k)}$ has a single unit component in k^{th} position. The Galerkin method, with ψ_i taken as ϕ_i, has advantages for small Δt: if $L(\cdot)$ is a conservative operator, that is $(L(\underset{\sim}{u}), \underset{\sim}{u}) = 0$ so that the L_2 energy $\|\underset{\sim}{u}\|^2$ is conserved this same property is retained for $\|\underset{\sim}{U}\|^2$; also as $\Delta t \to 0$, the Galerkin equations can be regarded as giving the best L_2 fit to $\partial_t \underset{\sim}{U}$ when $L(\underset{\sim}{U}^n)$ is known. However Galerkin methods generally have very poor stability properties as well as poor accuracy for moderate values of Δt. For example, for the linear advection equation with $L(\underset{\sim}{u})$ replaced by $a\partial u/\partial x$, just as for (2.16), (2.18) we obtain a central difference approximation scheme and with explicit time-stepping this is stable only for $\Delta t = O(h^2)$.

The linear advection problem is a natural model problem for the development of more effective test functions and most of those described in the literature owe something to the idea of upwinding. On a uniform mesh the key parameter is $a\Delta t/h$, the CFL number (after Courant, Friedrichs and Lewy): if this has an integral value, clearly the solution can be exactly advected on the mesh. This is a particularly pertinent property within the framework that we have taken, and is usually satisfied by difference methods but not by Galerkin methods. Thus the unit CFL property was taken as a specific objective by

Morton & Parrott (1980) in devising a variety of Petrov-Galerkin methods. It would be inappropriate to describe their results in detail here and we give merely a flavour. For linear ϕ_j in one dimension and various time-stepping schemes they found special test functions χ_i with the unit CFL property and then for scalar problems set

$$\psi_i(x) = (1 - \nu)\phi_i(x) + \nu\chi_i(x) \tag{4.4}$$

where ν is determined from the CFL number. With the Euler scheme this gives a method close to but more accurate than the well-known Lax-Wendroff difference scheme: with Crank-Nicolson it gives a third order accurate scheme and with leapfrog one of fourth order. The methods also extend to systems of equations.

However, Morton & Stokes (1982) found that some properties were difficult to extend into two space dimensions. While the unit CFL property could be retained with bilinear elements on rectangles it could not be made to hold along all the edge directions of a uniform triangular mesh when piecewise linear elements were used. Also the test functions χ_i are discontinuous and in the case of the Euler scheme do not span the unit function: thus even the conservation of the first moment $\int u dx$ is lost. The attention of this author has therefore shifted to the Characteristic Galerkin methods described below.

Euler Characteristic Galerkin (ECG) methods.

These make more explicit use of characteristics. Consider the scalar conservation law in one space dimension (on the whole real line)

$$\partial_t u + \partial_x f(u) = 0 \tag{4.5a}$$

or

$$\partial_t u + a(u)\partial_x u = 0 \tag{4.5b}$$

where $a(u) = \partial f/\partial u$. Then u is constant along the characteristics $dx/dt = a$ so that if we write $u^n(x)$ for $u(x,n\Delta t)$ and use a similar notation for f and a, we have for smooth flows

$$u^{n+1}(y) = u^n(x) \quad \text{where} \quad y = x + a^n(x)\Delta t. \tag{4.6}$$

Thus the L_2 projection of u^{n+1} onto the trial space S^h spanned by $\{\phi_j\}$ is related to that of u^n by

$$\begin{aligned}(u^{n+1}-u^n,\phi_i) &= \int_{-\infty}^{\infty} u^{n+1}(y)\phi_i(y)dy - \int_{-\infty}^{\infty} u^n(x)\phi_i(x)dx \\ &= \int_{-\infty}^{\infty} u^n(x)[\phi_i(y)\,\frac{dy}{dx} - \phi_i(x)]dx \\ &= \int_{-\infty}^{\infty} u^n(x)\,[\,\frac{d}{dx}\int_x^y \phi_i(z)dz]dx \\ &= -\int_{-\infty}^{\infty} \partial_x u^n(x)\,[\int_x^y \phi_i(z)dz]dx:\end{aligned}$$

that is, we have the exact relationship for the true solution

$$(u^{n+1} - u^n,\phi_i) + \Delta t(\partial_x f^n,\Phi_i^n) = 0 \qquad \forall\phi_i \in S^h \qquad (4.7)$$

where

$$\Phi_i^n(x) := \frac{1}{a^n(x)\Delta t}\int_x^{x+a^n(x)\Delta t} \phi_i(z)dz \quad . \qquad (4.8)$$

This form strongly suggests the following basic ECG method :-

$$(U^{n+1} - U^n,\phi_i) + \Delta t(\partial_x f(U^n),\overline{\Phi}_i^n) = 0 \qquad \forall\phi_i \in S^h, \qquad (4.9)$$

where $\overline{\Phi}_i^n$ has the same form as Φ_i^n in (4.8) but with a^n replaced by $a(U^n)$. This method in effect exactly traces the evolution of $U^n(x)$ through one time-step, as given by the relationship (4.6), and then projects this onto S^h. However, it is implemented like a Petrov-Galerkin method, from which it is distinguished by the fact that the time difference involves a Galerkin inner product (and hence only symmetric equations have to be solved) while the spatial operator is combined with a test function directly derived from the trial function. The fact that the derivation of $\overline{\Phi}_i^n$ involves only an averaging process means that it may be very efficiently approximated. Several approximations when the ϕ_j are piecewise linear are given in Morton (1982b) for the CFL number range $|a\Delta t/h| \leq 1$: these reproduce the results of (4.9) when a is constant and one of them involves evaluation of only the same inner products

used in the Galerkin method if the product approximation scheme is used for $\partial_x f$, namely

$$\partial_x f(U^n) \cong \sum_{(j)} f(U_j^n)\phi_j' \quad . \tag{4.10}$$

In the same reference, it is also shown how (4.9) extends naturally into more space dimensions: one then has a flux vector $\underline{f}(u)$ and a characteristic velocity vector $\underline{a}(u) = \partial \underline{f}/\partial u$ so that (4.8) is replaced by

$$\Phi_i^n(\underline{x}) = \frac{1}{|\underline{a}^n(\underline{x})|\Delta t} \int_{\underline{x}}^{\underline{x}+\underline{a}^n(\underline{x})\Delta t} \phi_i(\underline{z})d\ell, \tag{4.11}$$

integration being along the straight line between $\underline{x}$ and $\underline{x} + \underline{a}^n(\underline{x})\Delta t$.

It should be noted that in principle there is no stability limit for (4.9). Indeed, since if the terms in (4.9) are evaluated exactly the only error is at the projection stage, the least error is committed in going from $t = 0$ to $t= T$ if one large step $\Delta t = T$ is used! This is not very practical of course because for a system of equations the characteristics will be curved, the simple relation (4.6) will not hold and shocks will often intervene to destroy the basic assumption above that the solution is smooth. However, for conventional time-steps, with CFL numbers of the order of unity, these methods are extremely accurate,piecewise linear elements giving third order accuracy for instance. And when they are written out in terms of nodal values they show many relationships with high accuracy difference methods. These relationships and a detailed analysis of the basic ECG scheme will be given in Morton (1983b). Here we point out just one such set of links. Suppose that in deriving (4.7) and (4.9) we used a mixed norm instead of the L_2 norm, i.e. the inner product

$$(u,v) + \gamma^2(\partial_x u, \partial_x v) \tag{4.12}$$

for some constant γ. Then, as for (4.8) with piecewise linear ϕ_j on a uniform mesh and $a\Delta t/h \in (0,1)$, the corresponding special test function has

support over three intervals (x_{i-2},x_{i+1}): however, if $\gamma^2 = \frac{1}{6}(a\Delta t)^2$ its average value over (x_{i-2},x_{i-1}) is zero and it can therefore be approximated by $\phi_i + \frac{1}{2}a\Delta t\phi_i'$, which gives the same scheme if a is constant. The resulting scheme is then one of the Taylor-Galerkin schemes given by Donea (1982), which were derived in an entirely different manner, and in turn is equivalent to the EPG II scheme given by Morton & Parrott (1980) which was derived as indicated above in (4.3) and (4.4). Indeed, all of Donea's schemes can be derived in a like manner from the Characteristic-Galerkin methods.

Much more yet can be derived from the basic formulae (4.7) and (4.9). If U^n is the best fit to u^n from S^h in either the L_2 norm or the mixed norm derived from (4.12) any further knowledge of u^n (derived for example from studies of the original differential equation (4.5)) can be exploited through the recovery techniques discussed in section 3. Thus suppose this further information is incorporated in a recovery function $\tilde{u}^n$ which in the L_2 case satisfies

$$(\tilde{u}^n - U^n,\phi_i) = 0 \qquad \forall\phi_i \in S^h \ . \tag{4.13}$$

Then (4.9) can be replaced by

$$(U^{n+1} - U^n,\phi_i) + \Delta t\,(\partial_x f(\tilde{u}^n),\tilde{\Phi}_i^n) = 0 \qquad \forall\phi_i \in S^h, \tag{4.14}$$

where $\tilde{\Phi}_i^n$ has the same form as Φ_i^n in (4.8) but with a^n replaced by $a(\tilde{u}^n)$. For example, if u^n is smooth enough one can as in section 3, cf. (3.24), recover from piecewise linears to cubic splines. More interestingly, this enables one to use the non-conforming piecewise constant basis functions: as is shown in Morton (1983b) for the linear advection problem with constant a, quadratic spline recovery from piecewise constants yields through (4.14) a formula identical to (4.9) with piecewise linears. There is in fact a whole hierarchy of similarly related Characteristic Galerkin methods based on splines.

Piecewise constant ϕ_i are a natural choice for shock-modelling and their use in connection with (4.13) and (4.14) has been explored in (Morton, 1982c). Clearly the basic ECG scheme (4.9) is not defined when U^n, $f(U^n)$ and $a(U^n)$ all have discontinuities at the cell boundaries which we can take to be at $x_{i+\frac{1}{2}}$. This is true even for smooth flows: but we can then spread the discontinuities by a linear variation over $\frac{1}{2}\theta h$ either side of $x_{i+\frac{1}{2}}$ to join constant values $\tilde{u}_i$ and $\tilde{u}_{i+1}$ either side; then it is easy to see that on a uniform mesh (4.13) gives

$$\tilde{u}_i + \frac{\theta}{8}\delta^2\tilde{u}_i = U_i \qquad \forall i. \tag{4.15}$$

For sufficiently small θ and if $a(\tilde{u}_{i-1})$, $a(\tilde{u}_i)$ and $a(\tilde{u}_{i+1})$ are all non-negative, we also find that (4.14) reduces to

$$h(U^{n+1}_i - U^n_i) + \Delta t\,[\Delta_- f(\tilde{u}_i) + \frac{\theta}{8}\,\frac{h}{\Delta t}\,\delta^2 u_i] = 0 \qquad \forall i \tag{4.16a}$$

i.e.
$$U^{n+1}_i = \tilde{u}^n_i - (\Delta t/h)\Delta_- f(\tilde{u}_i) \qquad \forall i. \tag{4.16b}$$

Clearly as $\theta \to 0$ it reduces to the familiar first order upwind scheme: for $\theta > 0$ it has a similar form in (4.16b) but from (4.16a) it is seen to incorporate an anti-diffusive flux, as in many modern difference schemes; in fact it can be seen from (4.15) that it is the recovery process that is sharpening up the profiles broadened by the averaging process which is presumed to have led to U_i.

Regions of smooth flow are recognised by characteristics not crossing, typically that is $a(\tilde{u}_{i-1}) \leq a(\tilde{u}_i)$. On the other hand $a(\tilde{u}_{i-1}) > a(\tilde{u}_i)$ will lead to crossing of characteristics and the breakdown of (4.6) because the mapping from y to x is not unique. Even if recovery by a smooth function were appropriate in this case, (4.14) would not describe the exact evolution of $\tilde{u}^n$ followed by projection: instead it gives the projection of the multi-valued solution produced by the crossing characteristics from (4.6). The resulting approximation was used successfully by Morton (1982c) to model breaking waves but of course now Δt must be limited if good accuracy is to be

achieved. In fact it turns out that the same upwind formula is obtained in the limit $\theta \to 0$ where $a(\tilde{u}_{i-1}) \leq a(\tilde{u}_i)$ or $a(\tilde{u}_{i-1}) > a(\tilde{u}_i)$. Moreover, if $f(\cdot)$ is convex with a single sonic point $\bar{u}$ at which $a(\bar{u}) = 0$, the intermediate case in which $a(\tilde{u}_{i-1})a(\tilde{u}_i) < 0$ is dealt with very naturally through the recovery process: $f(\tilde{u}_i) - f(\tilde{u}_{i-1})$ is split into $f(\tilde{u}_i) - f(\bar{u})$ and $f(\bar{u}) - f(\tilde{u}_{i-1})$ with the first contributing to the updating of U_i^n and the second to that of U_{i-1}^n; the scheme is then identical with that of Engquist & Osher (1980).

It would not be appropriate to go into greater details on shock modelling with the ECG scheme here: the scalar problem is dealt with fully in Morton (1982c) and the use of the approximate Riemann solvers of Roe (1981) to deal with the Euler equations of gas dynamics is described in Morton (1983a). The important point here is to recognise the crucial role played by the recovery process and the exploitation of the best L_2 fit property of the approximation in the understanding and development of the methods.

Thus these shock problems represent the furthest point in a consistent line of development that we have presented, starting from the self-adjoint elliptic problems in section 1. There the solutions were smooth and the appropriate norms dominated by derivative terms: here the solutions are discontinuous (and the analysis should ideally be in L_1) and we have used mainly the L_2 norm. But the objective of optimal approximation is consistent throughout.

REFERENCES

ALLEN, D. & SOUTHWELL, R. (1955) Relaxation methods applied to determine the motion, in two dimensions, of a viscous fluid past a fixed cylinder. Quart. J. Mech. & Appl. Math. VIII, 129-145.

BABUSKA, I. & AZIZ, A.K. (1972) Survey lectures on the mathematical foundations of the finite element method. The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations (ed. A.K. Aziz), Academic Press, New York, 3-363.

BARRETT, J.W., MOORE, G. & MORTON, K.W. (1983) Optimal recovery and defect correction in the finite element method. Univ. of Reading, Num. Anal. Report 11/83.

BARRETT, J.W. & MORTON, K.W. (1980) Optimal finite element solutions to diffusion-convection problems in one dimension. Int. J. Num. Meth. Engng. 15, 1457-1474.

BARRETT, J.W. & MORTON, K.W. (1981) Optimal Petrov-Galerkin methods through approximate symmetrization. IMA J. Numer. Anal. 1, 439-468.

BARRETT, J.W. & MORTON, K.W. (1983) Approximate symmetrization and Petrov-Galerkin methods for diffusion-convection problems. Comp. Meths. in Appl. Mech. Engng (to appear)

BRAMBLE, J.H. & SCHATZ, A.H. (1977) Higher order local accuracy by averaging in the finite element method. Math. Comp. Vol. 31, 94-111.

CHRISTIE, I., GRIFFITHS, D.F., MITCHELL, A.R. & ZIENKIEWICZ, O.C. (1976) Finite element methods for second order differential equations with significant first derivatives. Int. J. Num. Meth. Engng. 10, 1389-1396.

CIARLET, P.G. (1978) The Finite Element Method for Elliptic Problems. North-Holland, Amsterdam.

DONEA, J. (1982) A Taylor-Galerkin method for convective transport problems. Int. J. Num. Meth. in Engng. (To appear).

DOUGLAS, J.Jr., DUPONT, T. & WHEELER, M.F. (1974) A Galerkin procedure for approximating the flux on the boundary for elliptic and parabolic boundary value problems. R.A.I.R.O. R-2, 47-59.

ENGQUIST, B. & OSHER, S. (1980) Stable and entropy satisfying approximations for transonic flow calculations. Math. Comp. 34, 45-75.

GOLUMB, M. & WEINBERGER, H.F. (1959) Optimal approximation and error bounds. Symp. on Numerical Approximation (ed. R.E. Langer), Madison, 117-190.

HEINRICH, J.C., HUYAKORN, P.S., MITCHELL, A.R. & ZIENKIEWICZ, O.C. (1977) An upwind finite element scheme for two-dimensional convective transport equations. Int. J. Num. Meth. Engng. 11, 131-143.

HEMKER, P.W. (1977) A numerical study of stiff two-point boundary problems. Thesis, Mathematisch Centrum, Amsterdam.

HUGHES, T.J.R. (1979) Finite Element Methods for Convection Dominated Flows, AMD Vol. 34, Am. Soc. of Mech. Eng. (New York).

HUGHES, T.J.R. & BROOKS, A.N. (1979) A multi dimensional upwind scheme with no crosswind diffusion. Finite Element Methods for Convection Dominated Flows (ed. T.J.R. Hughes), AMD Vol. 34, Am. Soc. Mech. Eng., (New York), 19-35.

HUGHES, T.J.R. & BROOKS, A.N. (1982) A theoretical framework for Petrov-Galerkin methods with discontinuous weighting functions: application to the streamline-upwind procedure. Finite Elements in Fluids Vol. 4 (eds. R.H. Gallagher, D.H. Norrie, J.T. Oden & O.C. Zienkiewicz), J. Wiley & Sons, New York, 47-65.

LESAINT, P. & ZLAMAL, M. (1979) Superconvergence of the gradient of finite element solutions. R.A.I.R.O. Numer. Anal. 13, 139-166.

LEVINE, N. (1983) Superconvergent recovery of the gradient from finite element approximation on linear triangles. Univ. of Reading, Num. Anal. Report 6/83.

LONG, M.J. & MORTON, K.W. (1976) The use of divided differences in finite element calculations. J. Inst. Maths. Applics. 19, 307-323.

MICCHELLI, C.A. & RIVLIN, T.J. (1976) A survey of optimal recovery. Optimal Estimation in Approximation Theory (eds. C.A. Micchelli & T.J. Rivlin), Plenum Press, New York, 1-54.

MORTON, K.W. (1982a) Finite element methods for non-self-adjoint problems. Proc. SERC Summer School, 1981 (ed. P.R. Turner), Lect. Notes in Maths 965, Springer-Verlag, Berlin, 113-148.

MORTON, K.W. (1982b) Generalised Galerkin methods for steady and unsteady problems. Proc. IMA Conf. on Num. Meth. for Fluid Dynamics (eds. K.W. Morton & M.J. Baines), Academic Press, 1-32.

MORTON, K.W. (1982c) Shock capturing, fitting and recovery. Proc. 8th Int. Conf. on Numerical Methods in Fluid Dynamics, Aachen. (ed. E. Krause), Lect. Notes in Physics 170, Springer-Verlag, Berlin, 77-93.

MORTON, K.W. (1983a) Characteristic Galerkin methods for hyperbolic problems. Proc. 5th GAMM Conf. on Numerical Methods in Fluid Mechanics. (To appear).

MORTON, K.W. (1983b) Analysis of Characteristic Galerkin methods for scalar problems in one dimension. (in preparation).

MORTON, K.W. & PARROTT, A.K. (1980) Generalised Galerkin methods for first order hyperbolic equations. J. Comp. Phys. 36, 249-270.

MORTON, K.W. & STOKES, A. (1982) Generalised Galerkin methods for hyperbolic equations. Proc. MAFELAP 1981 Conf. (ed. J.R. Whiteman) Academic Press, London, 421-431.

RHEINHARDT, H.J. (1982) A-posteriori error analysis and adaptive finite element methods for singularly perturbed convection-diffusion equations. Math. Methods Appl. Sci. (To appear).

ROE, P.L. (1981) Approximate Riemann solvers, parameter vectors and difference schemes. J. Comp. Phys., 43, 357-372.

SCOTNEY, B.W. (1982) Error analysis and numerical experiments for Petrov-Galerkin methods. Univ. of Reading Num. Anal. Report 11/82.

STRANG, G. & FIX, G.J. (1973) An Analysis of the Finite Element Method. Prentice-Hall, New York, 169.

SYNGE, J.L. (1957) The hypercircle in mathematical physics. Cambridge University Press, London.

THOMÉE, V. (1977) High order local approximations to derivatives in the finite element method. Math. Comp. 31, 652-660.

WHEELER, J.A. (1973) Simulation of heat transfer from a warm pipeline buried in permafrost. Presented to the 74th National Meeting of Am. Inst. of Chem. Eng., New Orleans.

WHEELER, M.F. (1974) A Galerkin procedure for estimating the flux for two-point boundary value problems. SIAM J. Numer. Anal. 11, 764-768.

ZLAMAL, M. (1977) Some superconvergence results in the finite element method. Mathematical Aspects of Finite Element Methods, Springer-Verlag, 353-362.

ZLAMAL, M. (1978) Superconvergence and reduced integration in the finite element method. Math. Comp. 32, 663-685.