# RECENT DEVELOPMENTS IN FINITE ELEMENT METHODS FOR FLUID DYNAMICS PROBLEMS

K.W. MORTON

---

NUMERICAL ANALYSIS REPORT 9/83

# RECENT DEVELOPMENTS IN FINITE ELEMENT METHODS FOR FLUID DYNAMICS PROBLEMS

K. W. Morton
Department of Mathematics, University of Reading, U.K.

Summary: A single theme is concentrated on. Namely, the development of finite element methods so as to retain one of their principal attractions, that of yielding optimal approximations, as they are extended from self-adjoint elliptic problems to more general problems. Two types of problem and corresponding methods are discussed: non-self-adjoint elliptic problems, typified by diffusion-convection problems, are dealt with by Petrov-Galerkin methods; hyperbolic problems are treated by both these methods and Characteristic Galerkin methods.

## 1. INTRODUCTION

In the fields where they were originally developed finite element methods had two inherent advantages. One was the flexibility that they brought to the modelling of awkwardly shaped regions. The other was the fact that they yield approximations to the unknown functions which are optimal in a natural energy norm. The first is presumably of little interest in meteorology but the second, with its promise of accurate representation on coarse grids, should be a great attraction. However, while the first depends little on the type of problem being solved the latter is very problem dependent. Thus it is taking some time for the methods to be developed to a point where they can make a practical impact on typical problems in fluid dynamics and this is a very active research area at the present time.

We begin by recalling how the optimal approximation property arises. Consider the following extremal problem for functions $v$ defined in a region $\Omega$ and satisfying certain essential boundary conditions:-

$$\text{minimise } \{\tfrac{1}{2}\|Tv\|^2 - \langle f,v\rangle\}\ , \tag{1.1}$$

where $T$ is a linear differential operator of order $m$, $f$ is a given function and $\langle\cdot,\cdot\rangle$, $\|\cdot\|$ denote respectively the $L_2$ inner product and norm over $\Omega$. In the minimisation, $v$ is to lie in $H^m(\Omega)$, the Sobolev space of functions with square integrable $m^{th}$ derivatives. The solution $u$ of (1.1) satisfies the differential equation of order $2m$

$$T^*Tu = f\ , \tag{1.2}$$

together with the essential and possibly some natural boundary conditions, where $T^*$ is the formal adjoint of $T$. Now suppose $u$ is approximated from a conforming finite element space $S^h \subset H^m(\Omega)$ spanned by basis functions

$\phi_j(\underline{x})$, that is

$$S^h := \{V \in H^m(\Omega) \mid V(\underline{x}) = \sum_{(j)} V_j \phi_j(\underline{x})\} \quad . \tag{1.3}$$

Then carrying out the minimisation in (1.1) over $S^h$ gives the approximation $U$ satisfying the Galerkin equations

$$\langle TU, T\phi_i \rangle = \langle f, \phi_i \rangle \qquad \forall \phi_i \in S_0^h \quad . \tag{1.4}$$

Here, $U$ like $u$ is to satisfy the essential boundary conditions so we write $U \in S_E^h$, while the variations lie in the subspace of $S^h$ satisfying homogeneous essential conditions, which we have denoted by $S_0^h$.

Now $u$ also satisfies these equations. So on subtraction we get

$$\langle T(u-U), T\phi_i \rangle = 0 \qquad \forall \phi_i \in S_0^h. \tag{1.5}$$

From this orthogonality relationship for the error $u-U$ it follows immediately that

$$\|T(u-U)\| = \inf_{V \in S_E^h} \|T(u-V)\| \quad . \tag{1.6}$$

This is the optimal approxmation property of $U$. From it follow various superconvergence properties. For example, suppose $T$ is the gradient operator $\underline{\nabla}$ so that $m = 1$ and (1.4) corresponds to Poisson's equation: and suppose $S^h$ consists of piecewise linear functions over a triangulation of $\Omega$. Then $\underline{\nabla}U$ will be piecewise constant and generally can only hope to have first order accuracy. But for triangulations which are reasonably regular and everywhere have six triangles meeting at each node, the derivative of $U$ along each edge will be second order accurate at the mid-point: at the same point the average of the normal derivative either side of the edge will also be second order accurate. So the whole vector $\underline{\nabla}U$ can be "recovered" to second order at these points (Levine, 1983). On practical meshes, the increase in accuracy is very substantial. Such phenomena are very widespread and have long been exploited by engineers in stress calculations and similar applications.

The Galerkin equations (1.4) can be written down and solved for a wide variety of problem types. But unfortunately for the Galerkin approximation $U$ to have the crucial property (1.6) the bilinear form $\langle T\cdot, T\cdot \rangle$ on the left of (1.4) has to be symmetric. In the next section we consider how Petrov-Galerkin methods can yield this property for steady flow problems having unsymmetric forms. For unsteady flows, particularly hyperbolic problems, there is a different but related difficulty: Galerkin methods have very desirable properties for very small time steps but lose them long before CFL numbers near unity are reached. We shall see in section 3 how Characteristic Galerkin methods, and even some Petrov-Galerkin

methods, can maintain these properties to unit CFL numbers.

## 2. PETROV-GALERKIN METHODS FOR STEADY DIFFUSION-CONVECTION

Diffusion-convection problems not only form an important class of practical problems in their own right: their successful approximation is a necessary preliminary to tackling the Navier-Stokes equations at moderate Reynolds numbers, that is considerably larger than the $Re = O(10^2)$ cases which are presently solved very successfully by mixed Galerkin methods. A typical problem takes the form:-

$$-\underline{\nabla}\cdot(a\underline{\nabla}u - \underline{b}u) = f \quad \text{in} \quad \Omega \tag{2.1a}$$

$$u = g \quad \text{on} \quad \Gamma_D, \quad \partial u/\partial n = 0 \quad \text{on} \quad \Gamma_N, \tag{2.1b}$$

where $a$ is a diffusion coefficient and $\underline{b}$ is a convective velocity field which we shall assume is incompressible (i.e. $\underline{\nabla}\cdot\underline{b} = 0$). We shall assume $\Omega$ is a bounded region of the plane with boundary $\Gamma_D \cup \Gamma_N$ and $\partial/\partial_n$ is in the outward normal direction. The inhomogeneous Dirichlet boundary condition is an essential condition which we shall impose on the finite element approximation by assuming that $g$ is the restriction of a function $G \in H^1(\Omega)$ to the boundary $\Gamma_D$ : thus we define the <u>trial space</u> as $S_E^h$, where

$$H^1(\Omega) \supset S_E^h := \{U = G + V \mid V \in S_0^h\} \tag{2.2a}$$

and, as in (1.3) and the remarks following,

$$H^1(\Omega) \supset S_0^h := \{V(\underline{x}) = \textstyle\sum_{(j)} V_j \phi_j(\underline{x}) \mid V = 0 \quad \text{on} \quad \Gamma_D\}. \tag{2.2b}$$

The bilinear form corresponding to (2.1) is

$$B(v,w) := \langle a\underline{\nabla}v, \underline{\nabla}w\rangle + \langle \underline{\nabla}\cdot(\underline{b}v), w\rangle . \tag{2.3}$$

It is easy to see that for $w = v$ and for $v = 0$ on $\Gamma_D$ we have, because of the incompressibility,

$$B(u,u) = \langle a\underline{\nabla}u, \underline{\nabla}u\rangle + \tfrac{1}{2}\int_{\Gamma_N} (\underline{b}\cdot\underline{n})v^2 ds. \tag{2.4}$$

By assuming further that $\Gamma_D$ includes all points of the boundary on which $\underline{b}\cdot\underline{n} < 0$, so that $u$ is prescribed on the inflow boundary, we ensure that $B(u,v)$ is positive definite. This in turn ensures that a unique solution exists to (2.1) of the form $u = u^0 + G$, where if we define

$$H^1_{E_0} := \{v \in H^1(\Omega) \mid v = 0 \quad \text{on} \quad \Gamma_D\} , \tag{2.5}$$

$u^0$ is given by the so-called <u>weak form</u> of the problem : find $u^0 \in H^1_{E_0}$ such that

$$B(u^0,w) = \langle f,w\rangle - B(G,w) \qquad \forall w \in H_{E_0} . \tag{2.6}$$

By the same arguments, the Galerkin approximation is uniquely defined by $U = U^0 + G$, where $U^0 \in S_0^h$ is given by

$$B(U^0,W) = \langle f,W\rangle - B(G,W) \qquad \forall W \in S_0^h . \tag{2.7}$$

Since $S_0^h \subset H_{E_0}^1$ we can substitute $W$ for $w$ in (2.6) and then subtracting (2.7) from the result we get, corresponding to (1.5),

$$B(u^0 - U^0,W) = B(u - U,W) = 0 \qquad \forall W \in S_0^h . \tag{2.8}$$

However, we cannot form a norm from $B(\cdot,\cdot)$ to get (1.6) because it is unsymmetric. Let us therefore take the main symmetric part and define

$$B_1(v,w) := \langle a\underline{\nabla}u, \underline{\nabla}w\rangle \tag{2.9}$$

with corresponding norm given by $B_1(u,v) =: \| v \|^2_{B_1}$ : and define $U_1^* \in S_E^h$ as the best fit in this norm to $u$. Then we have, comparing (2.9) with (2.4) and using (2.8),

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

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

$$\le \| u-U \|_{B_1} \{ \| u-U_1^* \|_{B_1} + \max_{\Omega} (|b|/a) \| a^{\frac{1}{2}}(u-U_1^*) \| \} . \tag{2.10}$$

Now it can be shown by a standard argument that, if $h$ is the maximum diameter of the elements in the discretisation of $\Omega$, there is a constant $K$ independent of $h$ such that

$$\| a^{\frac{1}{2}}(u-U_1^*) \| \le Kh \| u-U_1^* \|_{B_1} . \tag{2.11}$$

It therefore follows that

$$\| u-U \|_{B_1} \le [1 + Kh \max_{\Omega} (|b|/a)] \| u-U_1^* \|_{B_1} . \tag{2.12}$$

The dimensionless parameter $bh/a$ is called the mesh Péclet number and is the key factor in the loss of optimality in the Galerkin approximation $U$.

This result is quite sharp, for consider the simple one-dimensional test problem:-

$$-au'' + bu' = f \quad \text{on } (0,1) \tag{2.13a}$$

$$u(0) = 0, \quad u(1) = 1, \tag{2.13b}$$

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

$$u(x) = (e^{bx/a} - 1)/(e^{b/a} - 1), \tag{2.14}$$

giving a sharp boundary layer on the right when $b/a \gg 1$. Piecewise linear elements on a uniform mesh of size $h$ give the Galerkin equations for $j = 1,2,\ldots,J-1$ with $Jh = 1$

$$-\delta^2 U_j + (bh/a)\Delta_0 U_j = 0 \tag{2.15}$$

in the usual difference notation $\delta^2 U_j := U_{j+1} - 2U_j + U_{j-1}$, $\Delta_0 U_j := \tfrac{1}{2}(U_{j+1} - U_{j-1})$. These have the solution

$$U_j = (\mu_0^j - 1)/(\mu_0^J - 1), \quad \mu_0 = (2 + bh/a)/(2 - bh/a). \tag{2.16}$$

When $bh/a > 2$, $U$ exhibits spurious oscillations which bear no relation to the true solution and in fact the bound (2.12) can be attained with $K = 1/\pi$. This is a very familiar consequence of using central differences for the first order term $bu'$. With difference methods it is overcome by some form of upwinding, 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 of Allen & Southwell (1955) which with the average $(1-\xi)\Delta_0 + \xi\Delta_-$ can be written as

$$-[1 + \tfrac{1}{2}\xi(bh/a)]\delta^2 U_j + (bh/a)\Delta_0 U_j = 0; \tag{2.17}$$

with the choice

$$\xi = \coth(\tfrac{1}{2}bh/a) - (\tfrac{1}{2}bh/a)^{-1} \tag{2.18}$$

this is often called an exponentially-fitted scheme since for this model problem it gives exact nodal values, matching the exponential of (2.14).

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 $\phi_j$ with a view to generating these upwind difference schemes. Generally we define what are now called Petrov-Galerkin methods as follows: we introduce a test space $T_0^h$ different from but with the same dimension as the $S_0^h$ of (2.2b) and suppose it has basis functions $\psi_j(\underline{x})$ over the same elements,

$$H^1_{E_0} \supset T_0^h := \{V(x) = \textstyle\sum_{(j)} V_j \psi_j(\underline{x}) \mid V = 0 \text{ on } \Gamma_D\}; \tag{2.19}$$

then the Petrov-Galerkin approximation $U \in S_E^h$ is given by

$$B(U,W) = \langle F, W \rangle \qquad \forall W \in T_0^h, \tag{2.20a}$$

and the error satisfies

$$B(u-U,W) = 0 \qquad \forall W \in T_0^h \,. \tag{2.20b}$$

The problem then is to choose test spaces $T_0^h$ which are practically convenient and give good approximations - in some sense.

The earliest upwind test functions were those due to Christie et al. (1976) and Heinrich et al. (1977): a useful review is that of Heinrich & Zienkiewicz (1979) and other articles in the conference proceedings edited by Hughes (1979) give valuable background. For piecewise linear trial functions $\phi_i(x)$, typical test functions of this type have the form

$$\psi_i(x) := \phi_i(x) + \alpha\sigma_i(x) \tag{2.21a}$$

where

$$\sigma_i(x) := \begin{cases} 3(x-x_{i-1})(x_i-x)/(x_i-x_{i-1})^2 & x_{i-1} \le x \le x_i \\ -3(x_{i+1}-x)(x-x_i)/(x_{i+1}-x_i)^2 & x_i \le x \le x_{i+1} \,. \end{cases} \tag{2.21b}$$

On a uniform mesh setting the parameter $\alpha$ equal to $\xi$ defined in (2.18) leads to the Allen & Southwell difference operator and exact nodal values for the problem (2.13) when $f$ is a constant. With variable coefficients local values of $\alpha$ are used: and in two dimensions if bilinear elements are used on rectangles the trial basis functions have the form $\phi_i(x)\phi_j(y)$ and it is natural to use corresponding test functions $\psi_i(x)\psi_j(y)$ with the two parameters $\alpha$ based on the two components of $\underline{b}$.

An alternative but related approach is that due to Hughes & Brooks (1979, 1982): their streamline diffusion method starts from regarding the Allen & Southwell scheme written in the form (2.20) as enhancing the diffusion in the direction of the flow vector $\underline{b}$. Then the scalar diffusion coefficient of (2.1a) is replaced by a tensor diffusivity with components

$$A_{\ell m} = a\delta_{\ell m} + \tilde{a}\, b_\ell b_m \tag{2.22a}$$

where

$$\tilde{a} = \tfrac{1}{2}(\xi_1 b_1 h_1 + \xi_2 b_2 h_2) \tag{2.22b}$$

and $b_1$, $b_2$ are the components of $\underline{b}$ along the sides of a rectangular element of sides $h_1$, $h_2$ : $\xi_1$ and $\xi_2$ are corresponding values of the parameter given by (2.18). When bilinear elements $\phi$ are used and this modified operator is used with the Galerkin method, it can be shown that one obtains a difference operator equivalent to that obtained using a Petrov-Galerkin method with test functions

$$\psi = \phi + (\tilde{a}/|b|^2)\underline{b}\cdot\nabla\phi \,. \tag{2.23}$$

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.

Not only these test functions but clearly also any others which have the right amount of asymmetry will reproduce the Allen & Southwell difference operator for (2.13): but they will generally differ in two dimensions and even for (2.13) they will give different results for general source functions $f$. However, Morton (1982a) has given a general framework in which one can identify the ideal test functions for all $f$ and also can estimate the performance of any given test space. We apply it first to the symmetric form $B_1(\cdot,\cdot)$ of (2.9). Since for any fixed $w$, $B(v,w)$ is a bounded linear functional of $v$ in the norm $\|\cdot\|_{B_1}$, by the Riesz Representation Theorem it can be written as $B_1(v,R_1w)$ for some function $R_1w$ : and since this is true for any $w$ in $H^1_{E_0}$ and $R_1w$ depends linearly on $w$ we can define a linear operator $R_1 : H^1_{E_0} \to H^1_{E_0}$ such that

$$B(v,w) = B_1(v,R_1w) \qquad \forall v,\ w \in H^1_{E_0} \ . \tag{2.24}$$

In effect $R_1$ is a symmetriser for $B(\cdot,\cdot)$. Applying (2.24) to (2.20a) since $u-U \in H^1_{E_0}$, we get

$$B_1(u-U,R_1W) = 0 \qquad \forall W \in T^h_0 \ . \tag{2.25}$$

Clearly the ideal test functions $\psi^*_i$ would be such that

$$R_1\{\psi^*_i\} = \{\phi_i\} \qquad \phi_i \in S^h_0 \tag{2.26}$$

for then we could substitute $\phi_i$ for $R_1W$ in (2.25) and we would have recovered the orthogonality condition of (1.5). Moreover for general test spaces we have the following theorem:-

Theorem (Morton, 1982a) - Suppose the test space $T^h_0$ has the same dimension as $S^h_0$ and that there exists a constant $\Delta_1 \in [0,1)$ such that

$$\inf_{W\in T^h_0} \|V-R_1W\|_{B_1} \le \Delta_1 \|V\|_{B_1} \qquad \forall V \in S^h_0 \ . \tag{2.27}$$

Then the error in the corresponding Petrov-Galerkin approximation satisfies

$$\|u-U\|_{B_1} \le (1-\Delta_1^2)^{-\frac{1}{2}} \inf_{V\in S^h_E} \|u-V\|_{B_1} \ . \tag{2.28}$$

In particular if the ideal test space $\{\psi^*_i\}$ of (2.26) is used then $\Delta_1 = 0$ and $U$ is the optimal approximation to $u$, in the norm $\|\cdot\|_{B_1}$. Of course it will seldom be possible to find $R_1$ explicitly so that $\Delta_1$ will usually be difficult

to estimate. However, for the model problem (2.13) we have

$$(R_1 w)(x) = w(x) + (b/a)\int_0^x [w(t)-\bar{w}]dt, \tag{2.29}$$

where $\bar{w} = \int_0^1 w(t)dt$ : the ideal test functions are exponential in form and correspond to those used by Hemker (1977), namely

$$\psi_i(x) := \begin{cases} \left[1-e^{-b(x-x_{i-1})/a}\right] \Big/ \left[1-e^{-b(x_i-x_{i-1})/a}\right] & , x_{i-1} \le x \le x_i \\ \left[e^{-b(x-x_i)/a} - e^{-b(x_{i+1}-x_i)/a}\right] \Big/ \left[1-e^{-b(x_{i+1}-x_i)/a}\right] & , x_i \le x \le x_{i+1}. \end{cases} \tag{2.30}$$

Scotney (1982) has calculated the constant in (2.28) for both the test space of (2.21) and that of (2.23) and we reproduce his results in the table below. (Note however that since in the last case the method is non-conforming the theorem above does not strictly apply). The improvement over the Galerkin method is obvious: note particularly how the bound is virtually independent of bh/a.

| 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 : Ratios of Petrov-Galerkin error to optimal error given by $(1-\Delta_1^2)^{-\frac{1}{2}}$ - cf. (2.27) and (2.28).

There are still two weaknesses in this development, however. The first is that $\|\cdot\|_{B_1}$ is not obviously the most appropriate norm: because of the high gradients there, it concentrates attention on any thin boundary layer, where a good approximation is not possible without local mesh refinement; and it is quite independent of $\underline{b}$. There is a natural alternative which has been used by Barrett & Morton (1980, 1981, 1982, 1983) and with which they introduced the idea of symmetrisation. The diffusion-convection operator is of the form $T_1^* T_2$, where $T_1 := a^{\frac{1}{2}}\underline{\nabla}$ and $T_2 := a^{\frac{1}{2}}\underline{\nabla} - (\underline{b}/a^{\frac{1}{2}})$: the form $B_1(\cdot,\cdot)$ was based on $T_1$ and the alternative based on $T_2$ can be defined as

$$B_2(u,w) := \langle a\underline{\nabla}v, \underline{\nabla}w\rangle + \langle (b^2/a)v, w\rangle \tag{2.31a}$$

$$= \langle T_2 v, T_2 w\rangle + \int_\Gamma (\underline{b}\cdot\underline{n})vw ds. \tag{2.31b}$$

For increasing Peclet number this becomes closer to the $L_2$ norm with less emphasis on fitting gradients. Corresponding to (2.24) we can define an operator $R_2$ and an optimal test space as in (2.26): and the theorem of (2.27), (2.28) holds with a constant $\Delta_2$. For the model problem (2.13), $R_2^{-1}$ now has a simpler form than $R_2$ and we have

$$(R_2^{-1}w)(x) = w(x) + (b/a)\int_x^1 [w/t) - ce^{-bt/a}]dt. \qquad (2.32)$$

where the constant $c$ is such as to ensure that $(R_2^{-1}w)(0) = 0$ : thus it is easy to write down the ideal test functions as $\{R_2^{-1}\phi_i\}$ in this case, as it is also for variable coefficients $a$ and $b$.

The second disadvantage of using such test functions as (2.21) and (2.23) in a conventional Petrov-Galerkin formulation is that the system of equations to be solved is unsymmetric. This is true even for test functions (2.30) used by Hemker as they are linear combinations of the set $\{R_1^{-1}\phi_i\}$ which would give the symmetric matrix $\{B_1(\phi_j,\phi_i)\}$. The alternative is to write the problem (2.6) in the form, with $m = 1$ or $2$,

$$B_m(u,R_m w) = \langle f,w\rangle - B(G,w) \qquad \forall w \in H^1_{E_0} ; \qquad (2.33)$$

and then to approximate it, using the ideal test functions $\psi_i^* = R_m^{-1}\phi_i$ to give $U_m^0 \in S_0^h$, as

$$B_m(U_m^0,\phi_i) = \langle f,\psi_i^*\rangle - B(G,\psi_i^*) \qquad \forall\phi_i \in S_0^h . \qquad (2.34)$$

This is now a symmetric system of equations. What it requires is sufficient knowledge of $\psi_i^*$ to be able to calculate the terms of the right-hand side, which express the effect of the inhomogeneous data - both the source function $f$ and the Dirichlet boundary data. In practice it will need to be approximated. In one dimension and with $B_2(\cdot,\cdot)$ this can be done extremely accurately because of the explicit form available for $R_2^{-1}$ : and this is the form used by Barrett & Morton (1980, 1981, 1983) and Rheinhardt (1982) to obtain their very good results. Note that for the model problem (2.13) we now obtain the self-adjoint difference operator

$$-\delta^2 U_j + (bh/a)^2(1 + \tfrac{1}{6}\delta^2)U_j \qquad (2.35)$$

instead of the Galerkin operator in (2.15) or the ubiquitous Allen & Southwell operator (2.17) which occurred with the use of $B_1(\cdot,\cdot)$ and the standard Petrov-Galerkin formulation. The development of this approach for practical two-dimensional problems is still continuing - see Scotney (1982) for early results.

There is no reliable error analysis for either approach in two dimensions so we must consider model problems such as that due to Hutton (1981) and modified by Morton (1982b). In this the flow field on $-1 \leq x \leq 1$, $0 \leq y \leq 1$ is derived from a stream function $(1-x^2)(1-y^2)$. In Hutton's problem a tanh input profile is specified for $u$ on $y = 0$, $-1 \leq x \leq 0$ with Dirichlet conditions on the tangential boundary consistent with pure advection: the main test was the outlet profile on $y = 0$, $0 \leq x \leq 1$ for various values of the Peclet number. In the modified problem a zero input profile is specified but on the right-hand boundary ($x = 1$) we set $u = 100$ so that a tangential boundary layer forms there corresponding to a cold fluid flowing over a hot surface. Results for the Heinrich et al. upwind scheme of (2.21), the Hughes & Brooks stream-line diffusion scheme of (2.23) and a scheme based on the methods of Barrett & Morton (2.34) with $m = 2$, all with bilinear elements on rectangles, have been given by Scotney (1982). A selection are reproduced in Figs. 1-4, except that Fig. 3 represents more recent results.

For the original Hutton problem, the two $B_1(\cdot,\cdot)$ based schemes work well for mesh Peclet numbers $bh/a \leq 100$: at higher values they give some overshoot but generally fail to indicate the presence of a sharp input profile. The $B_2(\ ,\ )$ based method on the other hand works better for higher Peclet numbers. The difference between the two approaches is shown more clearly with the modified problem. Here the Heinrich scheme seems to be more reliable than the streamline diffusion method in showing the thickening boundary layer for decreasing $y$: but if we regard them as aiming at the best fit in the $\|\cdot\|_{B_1}$ norm neither should show any overshoot. The advantage of the $\|\cdot\|_{B_2}$ norm is shown in Fig. 4: not only is the thickening of the boundary layer as $y$ decreases shown well when this is larger than the mesh length; but even when $bh/a = 100$ the degree of overshoot in this norm gives a measure of the boundary layer thickness when it is substantially less than $h$ - i.e. it gives sub-grid scale information. The recovery of this information can be made quantitative by performing the <u>local recovery operation</u> of setting

$$B_2(U - \tilde{u}, \phi_i) = 0 \qquad (2.36)$$

for a sufficient number of neighbouring basis functions $\phi_i \in S_0^h$ to determine the free parameters in a hypothesised exponential $\tilde{u}$ for $u$ as described in Barrett & Morton (1980) and applied by Scotney (1982). This is highly effective for this problem even when the boundary layer is a fifth of the mesh length.

These last results serve to show why the clear objective of an optimal approximation in an integral norm is worth pursuing, as well as the importance

of choosing an appropriate norm. These points and theprocess of recovering sub-grid scale information will be further illustrated in the next section.

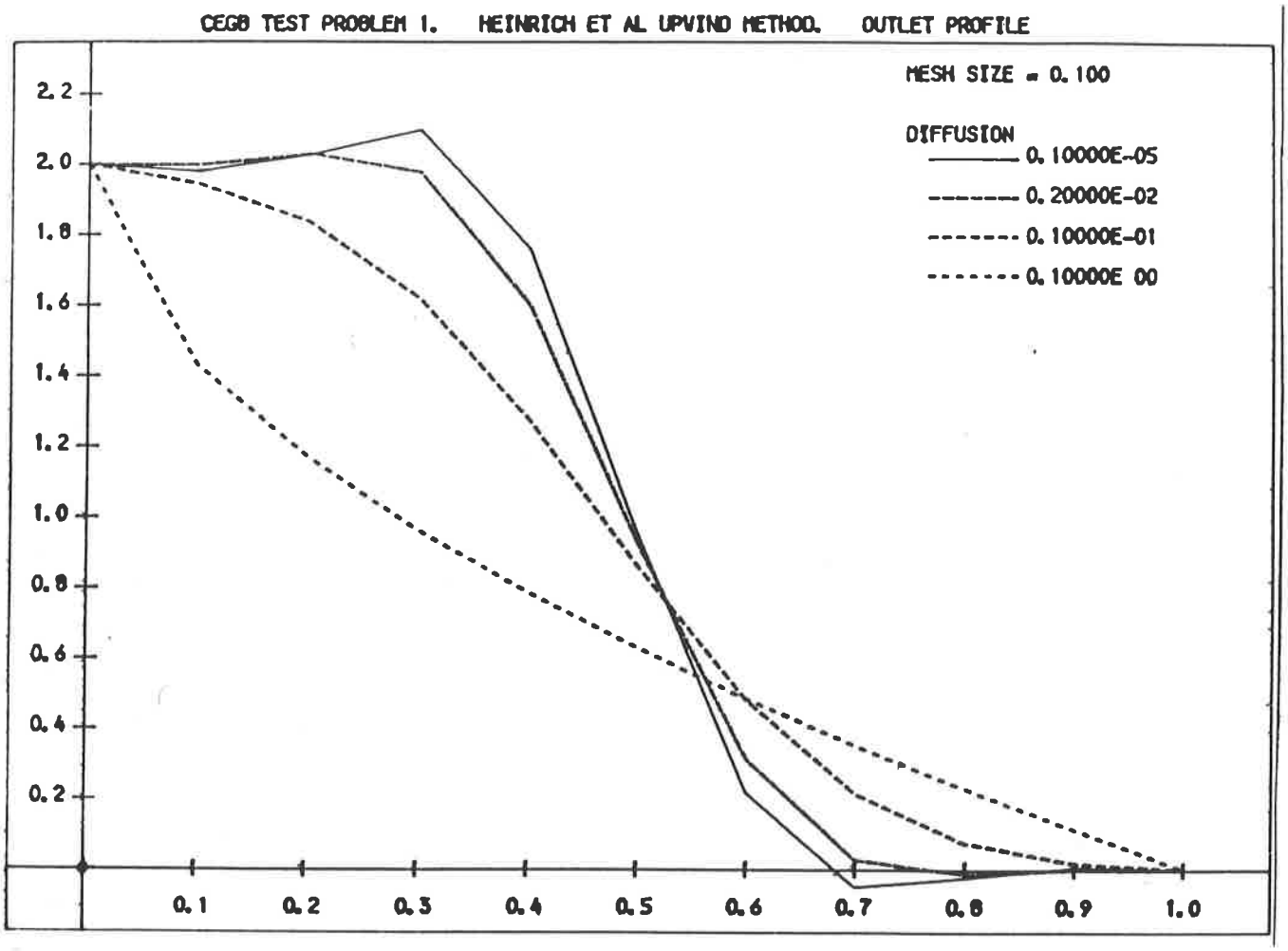



Fig. 1 - Outlet profile for test problem corresponding to an inlet profile of $u(x) = 1 + \tanh 10(2x + 1)$ on $(-1,0)$, using the Heinrich et al. (1977) scheme.

Figs. 2 & 3 (on next page)

Corresponding results using the Hughes & Brooks (1982) scheme (Fig. 2) and a method based on Barrett & Morton (1982)(Fig. 3).

Fig. 2

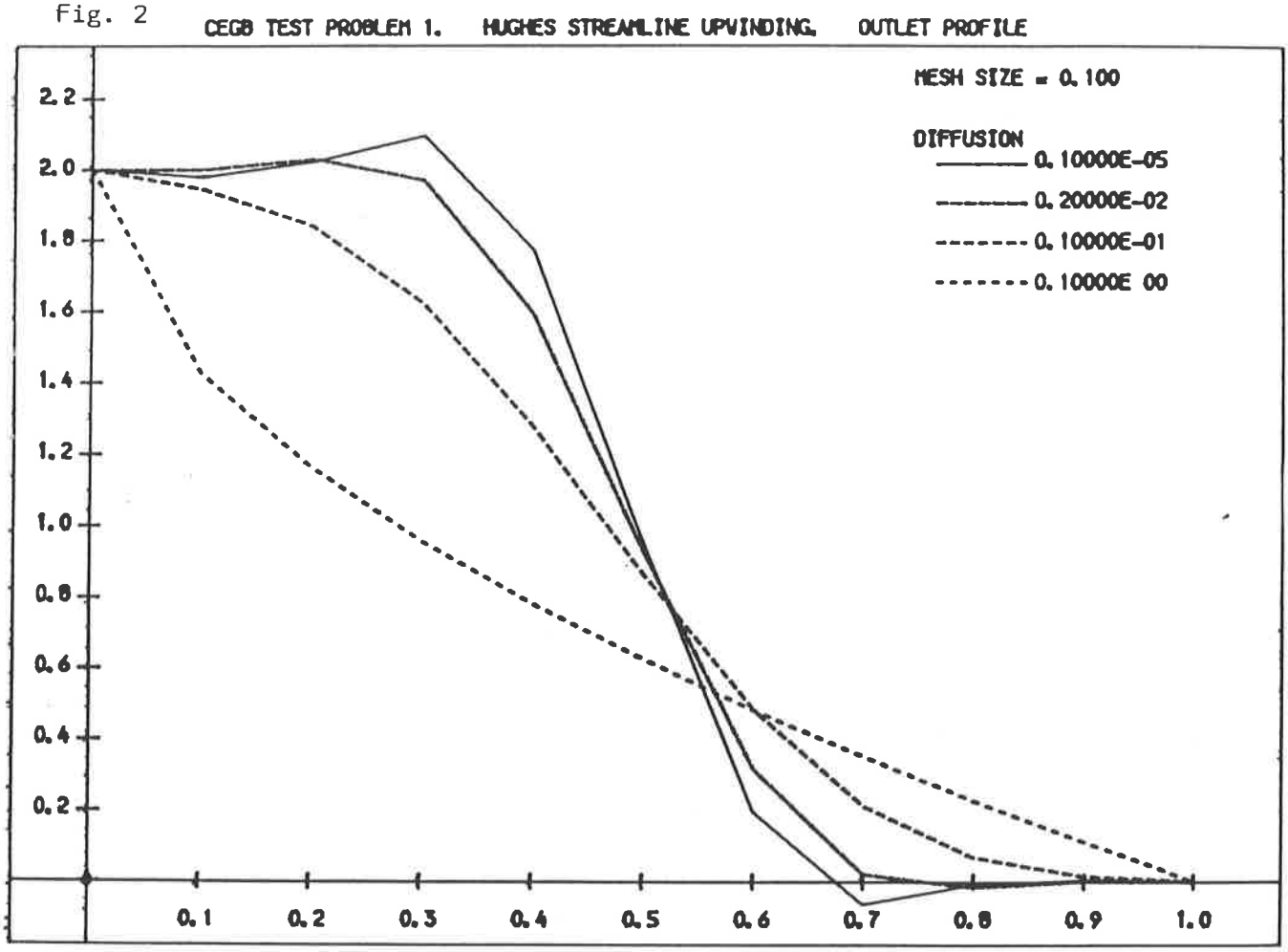
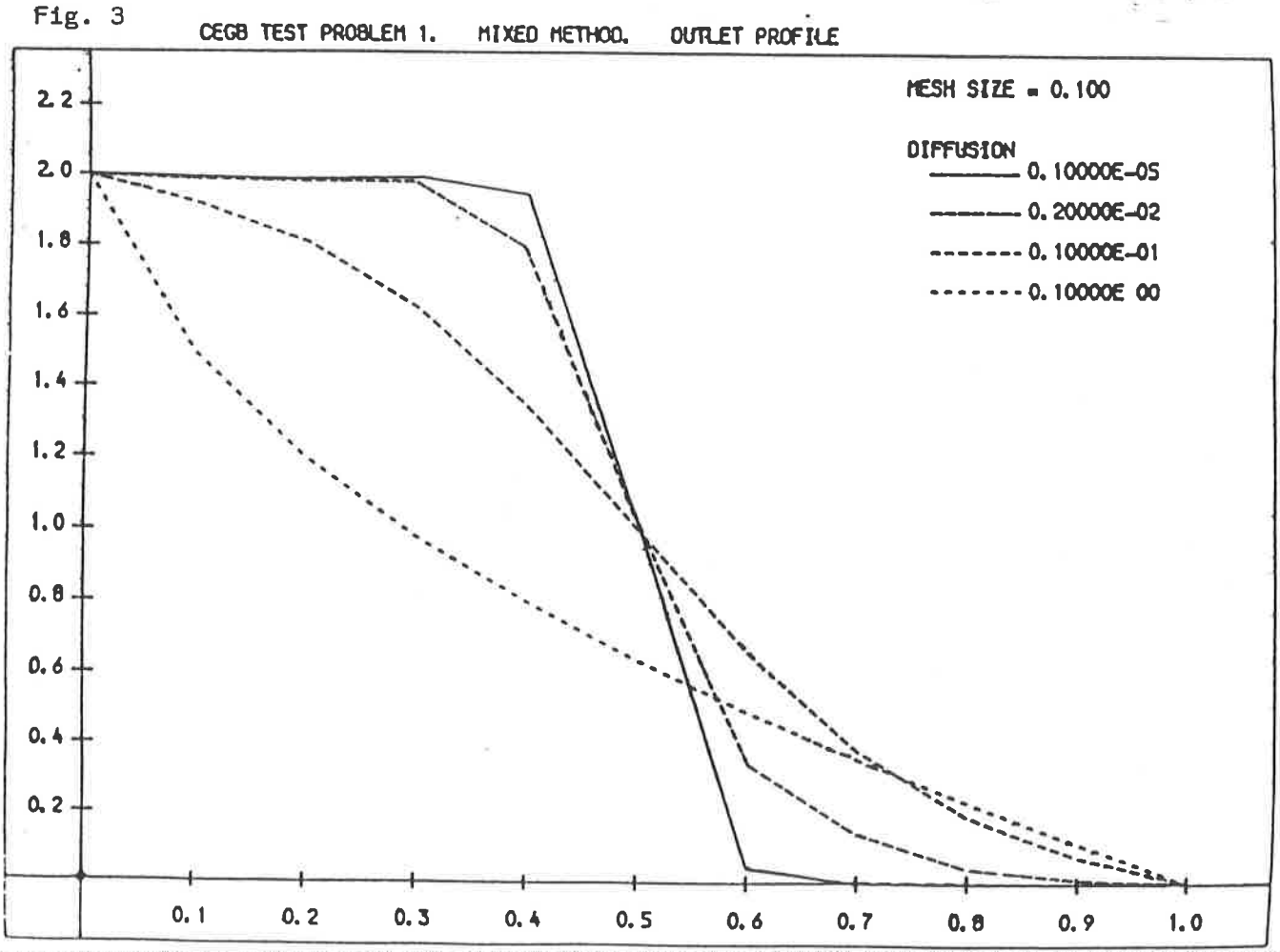

CEGB TEST PROBLEM 1. HUGHES STREAMLINE UPWINDING. OUTLET PROFILE
MESH SIZE = 0.100
DIFFUSION
0.10000E-05
0.20000E-02
0.10000E-01
0.10000E 00
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Fig. 3


CEGB TEST PROBLEM 1. MIXED METHOD. OUTLET PROFILE
MESH SIZE = 0.100
DIFFUSION
0.10000E-05
0.20000E-02
0.10000E-01
0.10000E 00
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

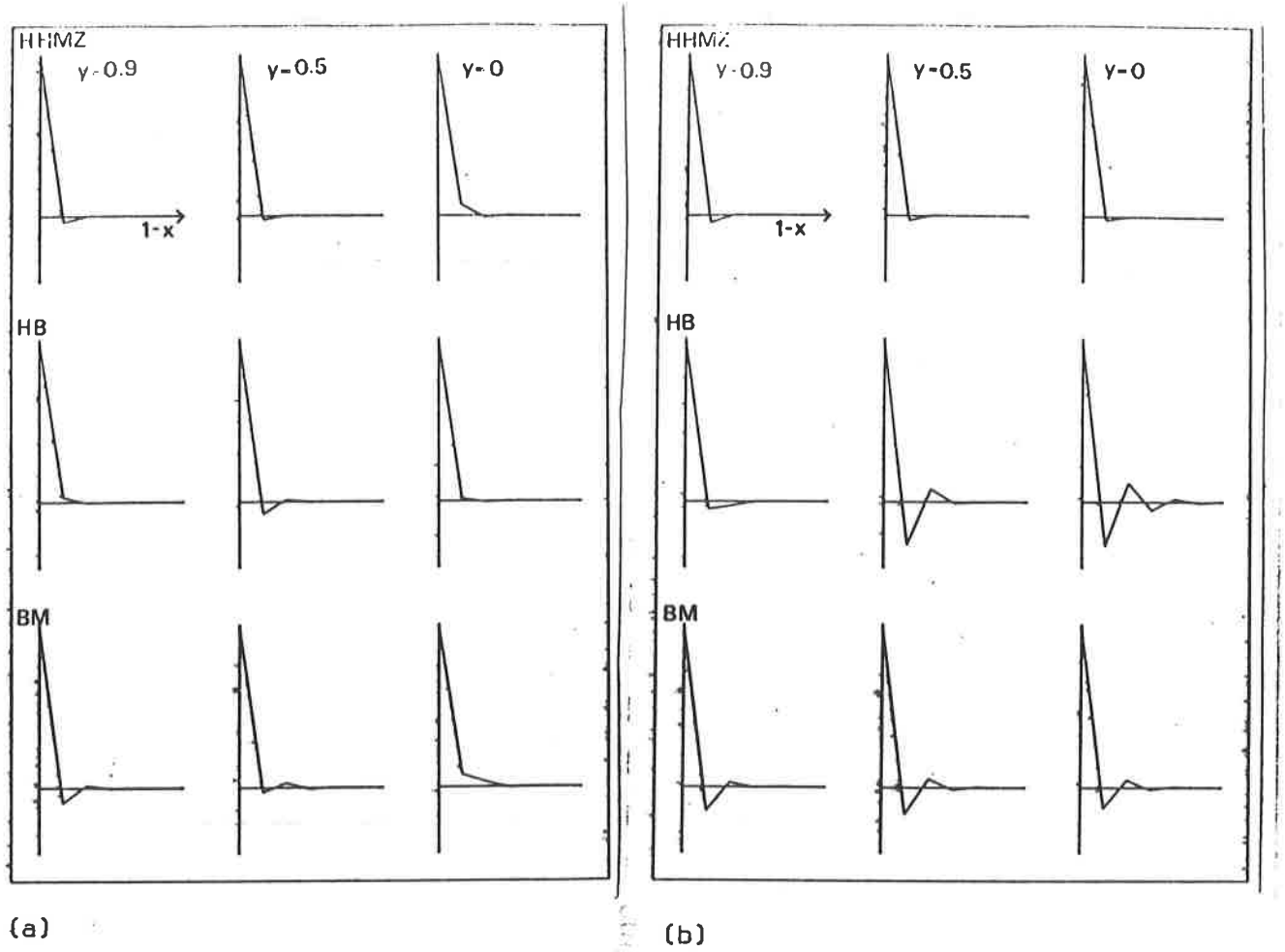



Fig. 4 - Boundary layers for modified test problem and each method (a) for mesh Péclet number $\beta = 20$ and (b) for $\beta = 100$.

## 3. GENERALISED GALERKIN METHODS FOR HYPERBOLIC PROBLEMS

For over a decade it has been realised that Galerkin approximations to hyperbolic equations have some very attractive properties - see, for example, Swartz & Wendroff (1969, 1974), Thomée & Wendroff (1974), Wahlbin (1974), Dendy (1974), Jespersen (1974). Consider the first order system of equations for the vector of unknowns $\underset{\sim}{u}(\underline{x},t)$

$$\underset{\sim}{u}_t + L(\underset{\sim}{u}) = 0, \tag{3.1}$$

where we use the subscript $t$ and the operator $\partial_t$ interchangeably to denote the partial derivatives with respect to time and $L$ is a (generally non-linear) operator involving the spatial derivatives. By the semi-discrete finite element approximation we mean a system of ordinary differential equations in time for the coefficients or nodal parameters in the expansion

$$\underset{\sim}{U}(\underline{x},t) = \sum_{(j)} \underset{\sim}{U}_j(t)\phi_j(\underline{x}), \tag{3.2}$$

where we have assumed for simplicity that the same basis functions are used for

all the components of $\underset{\sim}{U}$. The semi-discrete Galerkin equations are then

$$\langle \partial_t \underset{\sim}{U} + L(\underset{\sim}{U}), \phi_i \underset{\sim}{e}_{(k)} \rangle = 0 \qquad \forall i,k \tag{3.3a}$$

where $\underset{\sim}{e}_{(k)}$ is the unit vector consisting of just a unit component in the $k^{th}$ position: this gives the system of ordinary differential equations for the nodal parameter vectors $\underset{=}{U}_{(k)}$ of the components $U_{(k)}(\underset{\sim}{x},t)$ of $\underset{\sim}{U}(\underset{\sim}{x},t)$

$$\underset{=}{M}\dot{\underset{=}{U}}_{(k)} + \underset{=}{K}_{(k)}(\underset{\sim}{U}) = 0 \quad , \tag{3.3b}$$

where $M$ is the mass matrix $\{M_{ij}\} := \{\langle \phi_i, \phi_j \rangle\}$ and $\underset{=}{K}_{(k)}$ the stiffness vector $\{\langle L(\underset{\sim}{U}), \phi_i \underset{\sim}{e}_{(k)} \rangle\}$.

Now suppose $L(.)$ is a conservative operator in the sense that

$$\langle L(\underset{\sim}{u}), \underset{\sim}{u} \rangle = 0 \qquad \forall \underset{\sim}{u} \; . \tag{3.4}$$

Then, since equations (3.3a) can be multiplied by coefficients $U_{(k)i}$ and summed over $i$ and $k$, we have

$$\langle \partial_t \underset{\sim}{U} + L(\underset{\sim}{U}), \underset{\sim}{U} \rangle = 0 \tag{3.5}$$

so that

$$\frac{d}{dt} \| \underset{\sim}{U} \|^2 = 0; \tag{3.6}$$

that is the "energy" $\| \underset{\sim}{U} \|^2$ is conserved for the approximation as for the exact solution. Clearly by the same argument when any quadratic functional of $\underset{\sim}{u}$ is conserved then the same functional of the semi-discrete Galerkin approximation is also conserved. Note too that, if the so-called "lumped mass" equations obtained by replacing $M$ in (3.3b) by the diagonal matrix with the same row sums are used, then a discrete sum over the nodes $|\underset{\sim}{U}_i|^2$ is conserved. In this way, as Jespersen (1974) has shown, one may generate much more easily all the energy-conserving difference schemes derived originally by Arakawa (1966) and by a more direct technique by Morton (1970). Moreover one can see much more readily the advantages of these schemes as regards the suppression of non-linear instabilities because of the simpler Galerkin technique - see Morton (1977) for details.

We have not explicitly mentioned the imposition of boundary conditions in the above outline. However, certain homogeneous boundary conditions will be required on $\underset{\sim}{u}$ for (3.4) to hold. We suppose these same conditions are imposed on $\underset{\sim}{U}$ and the remaining non-physical conditions necessary for (3.3b) to be solved are implied naturally. Then the set of $\phi_I$ which are used in (3.3a) span the expansion (3.2) for $\underset{\sim}{U}$ so that (3.5) still holds and hence the energy is conserved in the presence of such boundary conditions. In this way the Galerkin

method gives a much simpler means of deriving conservative boundary conditions than is possible with difference methods, as for instance in Morton (1970).

Such global properties as the above are natural to the Galerkin formulation: what are not so obvious are the superconvergence properties obtained with one-dimensional spline basis functions. In line with our earlier viewpoint of seeking optimal approximations and following the analysis of Cullen & Morton (1980) we write the error $\underset{\sim}{u} - \underset{\sim}{U}$ in the form

$$\underset{\sim}{u} - \underset{\sim}{U} = (\underset{\sim}{u} - \underset{\sim}{U}^*) + (\underset{\sim}{U}^* - \underset{\sim}{U}) \qquad (3.7)$$

where $\underset{\sim}{U}^*$ is the $L_2$ projection of $\underset{\sim}{u}$ onto span $\{\phi_i\}$ so that $\underset{\sim}{u} - \underset{\sim}{U}^*$ is the projection error and $\underset{\sim}{U}^* - \underset{\sim}{U}$ is the evolution error, that is the difference of $\underset{\sim}{u}$ from the optimal approximation: the finite element space $\{\phi_i\}$ is chosen to minimise the first and the evolutionary procedure designed to minimise the second. Splines of order $\mu$ consist of piecewise polynomials of order $\mu - 1$ which have $\mu - 2$ continuous derivatives, the most practical for present purposes being the piecewise linears corresponding to $\mu = 2$. Clearly then the optimal approximation $\underset{\sim}{U}^*$ will generally have an accuracy of order $h^\mu$ while, as we have seen, specific features of $\underset{\sim}{u}$ can be recovered from this to a higher, superconvergent accuracy. But for these to be recovered from the approximation $\underset{\sim}{U}$ we need the evolutionary error $\underset{\sim}{U}^* - \underset{\sim}{U}$ to be of this higher order. It was this quantity which Thomée & Wendroff (1974) showed was of order $h^{2\mu}$ for linear problems with either constant or variable coefficients. Subsequently, Cullen & Morton (1980) showed this was also true for non-linear problems on a uniform mesh. Thus one has the rather remarkable property of piecewise linear elements yielding fourth order accuracy in this sense. Moreover, the implied constant can also be made quite small if the two-stage Galerkin process of the latter authors is used.

However, one still has to discretise in time. To start from (3.3b) and use a standard ODE package to solve the system seems rather unnatural and inefficient for hyperbolic equations: for space and time are linked through the characteristics and therefore the discretisation of one should have some influence over that of the other. Also a multi-level or multi-stage scheme can involve heavy storage penalties. Thus most authors favour a fairly simple one- or two-step method related to the spatial discretisation. For finite elements then the first choice is whether to use this type of approximation in time as well as space. We shall not do so but use finite differences in the time variable - though one could usually, rather artificially, produce the same schemes by using tensor product finite elements. This choice is partly for simplicity and flexibility: but it is mainly because neither of the key features of finite element methods pointed out in the Introduction are relevant to the time variable; geometric flexibility would be helpful only for moving boundary problems; and best

approximation in a time-integral norm is seldom of interest.

Because of the characteristics it is also very often advantageous to use explicit time-stepping. However, one then soon comes across disadvantages of the pure Galerkin approach. The price of the enhanced order of accuracy is generally a reduced range of stability and in some cases this can be severe. For example with leap frog time differencing for the advection equation the presence of the mass matrix with linear elements reduces the stability range by a factor $\sqrt{3}$ : but with Euler time-stepping, the central differences that the Galerkin method produces, as in (2.15), makes the scheme completely impractical with a limit $\Delta t = O(h^2)$. Linked to this phenomenon one also finds a very rapid loss of accuracy with increasing $\Delta t$. In particular, the common Galerkin schemes do not possess the unit CFL property: that is, when the characteristics of the linear advection equation pass through successive nodes or mesh points, they do not give exact advection.

Many authors have sought to overcome these disadvantages by moving to the more general Petrov-Galerkin methods, already described for the steady diffusion-convection problems of section 2. That is a test space is used, with basis functions $\psi_i$ replacing $\phi_i$ in (3.3a). In nearly all cases the idea of upwinding is involved, either by conscious choice at the outset or as a natural consequence. Thus the linear advection problem is a natural starting point for developing the choice of the $\{\psi_i\}$ and clearly this choice will depend on the time-stepping to be used. Thus let us start with Euler time-stepping for linear advection of a scalar and consider the scheme

$$< \frac{U^{n+1}-U^n}{\Delta t} + a\partial_x U^n, \psi_i > = 0 \qquad \forall i. \tag{3.7}$$

Morton & Parrott (1980) sought special test functions $\chi_i$ which on a uniform mesh with $a\Delta t/h = 1$ lead to exact advection of $U^n$: for piecewise linear $\phi_i$ they found such test functions of the form

$$\chi_i(x) := \begin{cases} 4 - 6(x_i - x)/h & , x_{i-1} \le x \le x_i \\ 0 & \text{otherwise .} \end{cases} \tag{3.8}$$

Then for more general meshes and $a > 0$ they set

$$\psi_i(x) = (1 - \nu_i)\phi_i(x) + \nu_i\chi_i(x) \tag{3.9}$$

where $\nu_i \in [0,1]$ is determined from the local CFL number. Clearly if $\nu_i = 1$ when $a\Delta t/h = 1$ this scheme has the unit CFL property. For $\nu_i = a\Delta t/h$ it gives the same spatial operator as the Lax-Wendroff method but due to its mass matrix it has improved accuracy while retaining the same stability range. This is clearly a vast improvement over the corresponding Galerkin method in

most respects. Unfortunately however since the $\{\chi_i^-\}$ do not span the unit constant the scheme does not conserve the first moment of U, let alone the second: thus it would be unsatisfactory for non-linear problems without further development. For later reference it is worth noting that for $a < 0$ one could replace $\chi_i^-$ by its mirror image $\chi_i^+$. A much improved and much more convenient scheme however is what Morton & Parrott called EPG II for which the test functions are

$$\psi_i = (1-\mu_i^2)\phi_i + \tfrac{1}{2}\mu_i^2(\chi_i^+ + \chi_i^-) + \tfrac{1}{2}\mu_i(\chi_i^+ - \chi_i^-) \qquad (3.10)$$

where $\mu_i$ is the local CFL number: this is third order accurate.

On the other hand, with leap frog time-differencing the corresponding $\chi_i$ are similar to (3.8) but symmetric about $x_i$. The resulting scheme as a consequence conserves first moments. Moreover with $\nu_i = (a\Delta t/h)^2$ it is fourth order accurate in both $h$ and $\Delta t$ with no dissipation and remarkably good phase accuracy. Also with Crank-Nicolson time-stepping the $\chi_i$ are just the characteristic functions for the intervals $(x_{i-1}, x_i)$ and so lead to conservation of first moments: the scheme with $\nu_i = (a\Delta t/h)^2$ is third order accurate and slightly dissipative but again has very good phase accuracy.

The above schemes extend without too much difficulty to systems of equations, that for the leap-frog time-stepping giving a particularly simple modification to the Galerkin equations. However, Morton & Stokes (1982) found that some of the properties were difficult to extend into two dimensions: while the 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. Thus their interest switched to Characteristic Galerkin methods.

An alternative approach to dealing with finite time-steps is that of Donea (1982) based on the same approach that led to the Lax-Wendroff methods. For Euler time-stepping and the linear advection equation we write the Taylor expansion

$$u(t+\Delta t) = u(t) + \Delta t u_t + \tfrac{1}{2}(\Delta t)^2 u_{tt} + \frac{1}{6}(\Delta t)^3 u_{ttt} + \ldots. \qquad (3.11a)$$

and then replace the time-derivatives by space derivatives through the differential equation to obtain the approximations:-

$$\begin{aligned} u_t &= -au_x \to -aU_x^n \\ u_{tt} &= -au_{xt} = a(au_x)_x \to a(aU_x^n)_x \\ u_{ttt} &\to (a/\Delta t)[a(U_x^{n+1} - U_x^n)]_x \end{aligned} \qquad (3.11b)$$

This can then be incorporated in a Galerkin formulation to give the

Taylor-Galerkin method:-

$$\left\langle \frac{U^{n+1}-U^n}{\Delta t} , \phi_i \right\rangle + \frac{1}{6}(\Delta t)^2 \left\langle a\partial_x \frac{U^{n+1}-U^n}{\Delta t} , a\partial_x \phi_i \right\rangle$$

$$+\langle a\partial_x U^n, \phi_i\rangle + \tfrac{1}{2}\Delta t \langle a\partial_x U^n, a\partial_x \phi_i\rangle = 0. \qquad (3.12)$$

For constant $a$ this is exactly the same as the EPG II scheme of (3.10). Similarly for leap-frog time-stepping the same scheme as that based on (3.9) is produced: but for Crank-Nicolson the schemes are different. We shall return to these schemes in a moment.

However we now turn to methods which make more explicit use of the characteristics, in particular the characteristic-Galerkin methods. We will consider from the outset the scalar conservation law in one space dimension

$$\partial_t u + \partial_x f(u) = 0 \qquad (3.13a)$$

or

$$\partial_t u + a(u)\partial_x u = 0 \qquad (3.13b)$$

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

$$u^{n+1}(y) = u^n(x) \qquad \text{where} \quad y = x + a^n(x)\Delta t = x + a^{n+1}(y)\Delta t. \qquad (3.14)$$

Thus for the $L_2$ projection onto the trial space $S^h = \text{span}\{\phi_i\}$ we have

$$\langle u^{n+1}, \phi_i\rangle = \int u^{n+1}(y)\phi_i(y)dy = \int u^n(x)\phi_i(y)dy. \qquad (3.15)$$

This has been directly incorporated into schemes for an approximation $U^n$ by several authors [see, for example, Douglas & Russell (1980), Bercovier et al. (1982)]. We can set

$$\langle U^{n+1}, \phi_i\rangle = \int U^n(x)\phi_i(y)dy \qquad \forall \phi_i \in S^h, \qquad (3.16a)$$

in which, if this is regarded as an explicit method, the right-hand side is evaluated by solving for each $y$ an implicit equation to give $x$, the foot of the backward-drawn characteristic from $y$: or if this is regarded as an implicit method we can use $x = y - a^{n+1}(y)\Delta t$. Alternatively we can rewrite (3.16a) as

$$\langle U^{n+1}, \phi_i\rangle = \langle U^n, \phi_i^s\rangle \qquad \forall \phi_i \in S^h \qquad (3.16b)$$

where $\phi_i^s(x) = \phi_i(y)(dy/dx)$.

We prefer to follow the latter route and, taking it somewhat further, note that

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

That is, we have the exact relationship for the true solution:-

$$\langle u^{n+1} - u^n, \phi_i\rangle + \Delta t\langle \partial_x f^n, \Phi_i^n\rangle = 0 \qquad \forall \phi_i \in S^h, \tag{3.17}$$

where

$$\Phi_i^n(x) := \frac{1}{a^n(x)\Delta t}\int_x^{x+a^n(x)\Delta t} \phi_i(z)dz. \tag{3.18}$$

This is now much more clearly related to a generalised Galerkin method and its form suggests immediately the basic ECG method.:-

$$\langle U^{n+1}-U^n, \phi_i\rangle + \Delta t\langle \partial_x f(U^n), \overline{\Phi}_i^n\rangle = 0 \qquad \forall \phi_i \in S^h \tag{3.19}$$

where $\overline{\Phi}_i^n$ has the same form as $\Phi_i^n$ but with $a^n$ replaced by $a(U^n)$. The resulting $U^{n+1}$ is exactly the same as that given by (3.16a): that is, it is the result of tracing the evolution of $U^n(x)$ through one time-step by means of (3.14) and then projecting the result onto $S^h$. However, (3.19) leads to several further improvements and approximations as well as pointing up the relationship with the Petrov-Galerkin formulation (3.7) and linking to some of the schemes derived from it.

Suppose that the $\phi_i$ are piecewise linear. Then the fact that (3.18) is a simple averaging operation means that $\overline{\Phi}_i^n$ is very easily approximated. Morton (1982b) gives several such approximations when the CFL number $\mu = a\Delta t/h$ lies in (0,1), all of which reproduce the results of (3.19) when $a$ is constant. One family of these takes the form

$$\begin{aligned}\Phi_i \simeq (1 - \tfrac{1}{2}\mu)\,\phi_i + \tfrac{1}{2}\mu\phi_{i-1} + \frac{11}{12}\mu(3 - 2\mu)(\phi_i' - \phi_{i-1}') \\ + M[(\phi_i - \phi_{i-1}) + \tfrac{1}{2}(\phi_i' - \phi_{i-1}')]\end{aligned} \tag{3.20}$$

where the choice $M = \frac{1}{2}\mu(1-\mu)^2$ gives the best $L_2$ fit to $\Phi_i$ by a linear fit in each interval: there are clearly several relationships here with Petrov-Galerkin methods proposed by Wahlbin (1974), Dendy (1974), and Hughes et al. (1982) based on test functions of the form $\phi_i + \alpha\phi_i'$, (c.f. the streamline-diffusion scheme (2.23)). Another approximates the inner product

$$\langle\phi_j',\phi_i\rangle \text{ by } (1-\mu)^2\langle\phi_j',\phi_i\rangle - \mu(1-\mu)\langle\phi_j',\phi_{i-1}\rangle$$

$$+ \mu(3-2\mu)\langle\phi_j,\phi_i-\phi_{i-1}\rangle \quad . \tag{3.21}$$

a form which requires no more inner products than the Galerkin method and is particularly suitable for use with the product approximation (see Christie et al. 1981)

$$\partial_x f(U^n) \simeq \sum_{(j)} f(U_j^n)\phi_j' \quad . \tag{3.22}$$

Instead of deriving (3.17) with the $L_2$ norm we could have used a mixed norm, based on the inner product

$$\langle u,v\rangle + \langle\gamma\partial_x u,\gamma\partial_x v\rangle \tag{3.23}$$

for some weight function $\gamma$ . Then as for (3.18), with piecewise linear $\phi_i$ on a uniform mesh and $\mu\epsilon(0,1)$, the corresponding special test functions have support over the three intervals $(x_{i-2},x_{i+1})$. However if $\gamma^2 = \frac{1}{6}(a\Delta t)^2$ the average value over $(x_{i-2},x_{i-1})$ is zero and a good approximation is given by $\phi_i + \frac{1}{2}a\Delta t\phi_i'$ which yields exactly the same scheme when $a$ is constant. But then the resulting scheme is precisely the Taylor-Galerkin scheme (3.12): indeed all the Taylor-Galerkin schemes can be generated in this manner. Similar schemes based on mixed norms are used by Baker & Soliman (1982).

It should be noted that in principle there is no stability limit on (3.19). Indeed since if the terms in (3.19) 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 (3.14) will not hold and shocks will often intervene to destroy the basic assumption above that the solution is smooth. Similarly there would be increasing complication for large time-steps with the natural generalisation of (3.19) to multidimensional problems: in these one has a flux vector $\underline{f}(u)$, a velocity vector $\underline{a}(u) = \partial\underline{f}/\partial u$ and (3.18) is replaced by

$$\phi_i^n(\underline{x}) = \frac{1}{|\underline{a}^n(x)|\Delta t}\int_{\underline{x}}^{\underline{x}+\underline{a}^n(\underline{x})\Delta t} \phi_i(\underline{z})d\ell \ ; \tag{3.24}$$

otherwise the form of the scheme is unchanged. However, for conventional time-steps with CFL numbers of the order of unity, the basic ECG method is extremely accurate. With piecewise linears it is third order accurate and closely related to well-known difference schemes studied by Warming et al. (1973): for example,

under the mixed norm (3.23) with $\gamma^2 = \frac{1}{6}h^2$ on a uniform mesh, the mass matrix becomes the identity so that the scheme is fully explicit, identical to one of the schemes given there.

The potential of the identity (3.17) goes further, however: to exploit it we need to recall our objective of maintaining as near a best fit as possible to the true solution. So suppose $U^n$ is the best fit to $u^n$ from $S^h$ in either the $L_2$ norm or the mixed norm. Then any further information that we have about $u^n$ or the underlying problem can be exploited by the recovery techniques briefly described in the last section, in order to obtain a better approximation than that given by (3.19). Thus suppose this further information - for example, smoothness, monotonicity, positivity - is embodied in a recovery function $\tilde{u}^n$ which in the $L_2$ case satisfies

$$\langle \tilde{u}^n - U^n, \phi_i \rangle = 0 \qquad \forall \phi_i \in S^h \ . \tag{3.25}$$

Then we can replace (3.19) by

$$\langle U^{n+1} - U^n, \phi_i \rangle + \Delta t \langle \partial_x f(\tilde{u}^n), \tilde{\phi}_i^n \rangle = 0 \qquad \forall \phi_i \in S^h, \tag{3.26}$$

where $\tilde{\phi}_i^n$ has the same form as $\phi_i^n$ in (3.18) but with $a^n$ replaced by $a(\tilde{u}^n)$. For example, if $u^n$ is smooth enough we can recover from piecewise linears with cubic splines. Of even greater interest however is the possibility of using non-conforming elements, in particular the very simple piecewise constants. Indeed one can show that for the linear advection equation with constant $a$, quadratic spline recovery from piecewise constants yields through (3.26) precisely the same formula as (3.19) with piecewise linears. There is in fact a whole hierachy of similarly related Characteristic Galerkin methods based on splines [c.f. results of Swartz & Wendroff (1974)].

Piecewise constant elements are a natural choice for shock-modelling and we end this section by illustrating the potential of the recovery process allied to (3.26) with some of the results that have been obtained with these simple elements - see (Morton, 1982c). Clearly the basic ECG scheme (3.19) is not defined when $U^n$, $f(U^n)$ and $a(U^n)$ all have discontinuities at the cell boundaries, which we take to be at $x_{i+\frac{1}{2}}$. This is true even for smooth flows. However we can then justifiably smear the discontinuities in the recovery process: suppose we regard $U^n$ as the projection on to piecewise constants of a function which is piecewise linear with flat sections in the centre of each cell; specifically, on a uniform mesh we spread the discontinuity at $x_{i+\frac{1}{2}}$ 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 $\tilde{u}_{i+1}$ on either side. Then the recovery formula (3.25), with piecewise constant $\phi_i$, gives

$$\tilde{u}_i + \frac{\theta}{8}\delta^2\tilde{u}_i = U_i \qquad \forall i \ . \tag{3.27}$$

For sufficiently small $\theta$ and if $a(\tilde{u}_{i-1})$, $a(\tilde{u}_i)$ and $a(\tilde{u}_{i+1})$ are all non-negative, we then find that (3.26) reduces to

$$h(U_i^{n+1} - U_i^n) + \Delta t[\Delta_- f(\tilde{u}_i) + \frac{\theta}{\theta}\frac{h}{\Delta t}\delta^2\tilde{u}_i] = 0 \qquad \forall i, \qquad (3.28a)$$

$$\text{i.e.} \quad U_i^{n+1} = \tilde{u}_i^n - (\Delta t/h)\Delta_- f(\tilde{u}_i) \qquad \forall i. \qquad (3.28b)$$

Clearly as $\theta \to 0$ this reduces to the familiar first order upwind scheme: for $\theta > 0$ it has a similar form but as (3.28a) shows it incorporates an anti-diffusive flux, as in many modern difference schemes. In fact the recovery process in (3.27) sharpens up the profiles broadened by the averaging process which is presumed to have led to $U_i$.

Regions of smooth flow where it is legitimate to use such recovery techniques are recognised by characteristics not crossing, typically that is $a(\tilde{u}_{i-1}) \le a(\tilde{u}_i)$. In the contrary case, the crossing of characteristics leads to the presence of shocks and the breakdown of the basic formula (3.14) because the mapping from $y$ to $x$ is not unique. Even if recovery by a smooth function were appropriate here the exact evaluation of $\tilde{u}^n$ followed by its projection would not be described by (3.26): instead this gives the projection of a multi-valued solution produced by the crossing characteristics. This can however be used for small enough $\Delta t$ as a good approximation to the true evolution and accurate results are given by Morton (1982c) for breaking waves using the inviscid Burger's equation. It is appropriate in the neighbourhood of shocks to take the limit $\theta \to 0$ in (3.27) and (3.28) and it turns out then that precisely the same upwind formula is obtained whether $a(\tilde{u}_{i-1}) \le a(\tilde{u}_i)$ 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 to that of Engquist & Osher (1980, 1981), which has the desirable property of avoiding non-physical shocks.

The use of the recovery process can be taken further in the modelling of shocks. For instance, suppose that through some such criterion as

$$[a(U_{i-1}^n) - a(U_{i+1}^n)]\Delta t > h \qquad (3.29)$$

a shock is recognised as present in cell $i$. Then we can suppose that $U_i^n$ is the average between two values either side of the shock and the position of the shock can be deduced as $x_s = (1-\eta)x_{i-\frac{1}{2}} + \eta x_{i+\frac{1}{2}}$ where

$$\eta = \Delta_+ U_i^n/(\Delta_+ U_i^n + \Delta_- U_i^n) \; . \qquad (3.30)$$

The way in which this simple procedure satisfies the Rankine-Hugoniot shock conditions and greatly improves the accuracy of the results can be found in (Morton, 1982c). Since then the method has been developed for the Euler equations of gas dynamics, using the approximate Riemann-solver methods of Roe (1981) to generalise the scalar methods to systems of equations. The form of the ECG algorithm for piecewise constants is particularly appropriate for this purpose, as it is for extending these techniques to two space dimensions.

## 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. and Appl. Math. VIII, 129-145.

Arakawa, A 1966 A computational design for the long term integration of the equations of atmospheric motion. J. Comp. Phys. 1, 119-143.

Baker, A.J. & Soliman, M.O. 1982 An accurate and efficient finite element Euler equation algorithm. Proc. 8th Intl. Conf. on Numerical Methods in Fluid Dynamics (ed. E. Krause), Lecture Notes in Physics 170, Springer-Verlag, Berlin, 115-123.

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. 1982 Optimal finite element approximation for diffusion-convection problems. Proc. MAFELAP 1981 Conf. (ed. J.R. Whiteman) Academic Press, London (1982), 403-411.

Barrett, J.W. & Morton, K.W. 1983 Approximate symmetrization and Petrov-Galerkin methods for diffusion-convection problems. To appear.

Bercovier, M., Pironneau, O., Hasbani, Y. & Livne, E. 1982 Characteristics and the finite element methods applied to the equation of fluids. Proc. MAFELAP 1981 Conf. (ed. J.R. Whiteman), Academic Press, London (1982), 471-478.

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.

Christie, I., Griffiths, D.F., Mitchell, A.R. & Sanz-Serna, J.M. 1981 Product approximation for non-linear problems in the finite element problem. IMA J. Num. Anal. 1, 253-266.

Cullen, M.J.P., & Morton, K.W. 1980 Analysis of evolutionary error in finite element and other methods. J. Comp. Phys. 34, 245-268.

Dendy, J.E. 1974 Two methods of Galerkin type achieving optimal $L^2$ rates of convergence for first order hyperbolics. SIAM J. Numer. Anal. 11, 637-653.

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

Douglas, Jr.J. & Russell, T.F. 1982 Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures. SIAM J. Numer. Anal. 19, 871-885.

Engquist, B. & Osher, S. 1980 Stable and entropy satisfying approximations for transonic flow calculations. Math. Comp. 34, 45-75.

Engquist, B. & Osher, S. 1981 One sided difference equations for non-linear conservation laws. Math. Comp. 36, 321-352.

Heinrich, J.C., Huyakorn, P.S., Mitchell, A.R. & Zienkiewicz, O.C. 1977 An upwind finite element scheme for two-dimensional convective transport equation. Int. J. Num. Meth. Engng. 11, 131-143.

Heinrich, J.C., & Zienkiewicz, O.C. 1979 The finite element method and 'upwinding' techniques in the numerical solution of convection dominated flow problems. Finite Element Methods for Convection Dominated Flows (ed. T.J.R. Hughes), AMD Vol. 34, Am. Soc. Mech. Eng. (New york), 105-136.

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

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

Hughes, T.J.R. & Brooks, A. 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., Tezduyar, T.E. & Brooks, A.N. 1982 A Petrov-Galerkin finite element fomrulation for systems of conservation laws with special reference to the compressible Euler equations. Proc. IMA Conf., Numerical Methods for Fluid Dynamics (eds. K.W. Morton & M.J. Baines), Academic Press, London, 95-125.

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.

Hutton, A.G. 1981 The numerical representation of convection. IAHR Working Group meeting, May 1981.

Jespersen, D.C. 1974 Arakawa's Method is a finite element method. J. Comp. Phys. 16, 383-390.

Levine, N. 1983 Superconvergent recovery of the gradient from finite element approximation on linear triangles. University of Reading, Numer. Anal. Report 6/83.

Morton, K.W. 1970 The design of difference schemes for evolutionary problems. SIAM-AMS Proc. 2, Amer. Math. Soc., 1-10.

Morton, K.W. 1977 Initial-value problems by finite difference and other methods. in The State of the Art in Numerical Analysis, Proc. IMA Conf. (ed. D.A.H. Jacobs), Academic Press, 699-756.

Morton, K.W. 1982(a) 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. 1982(b) 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. 1982(c) 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. & 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. 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.

Swartz, B. & Wendroff, B. 1969 Generalised finite difference schemes. Math. Comp. 23, 37-50.

Swartz, B. & Wendroff, B. 1974 The relation between the Galerkin and collocation methods using smooth splines. SIAM J. Numer. Anal. 11, 994-996.

Thomee, V. & Wendroff, B. 1974 Convergence estimates for Galerkin methods for variable coefficient initial-value problems. SIAM J. Numer. Anal. 11, 1059-1068.

Wahlbin, L. 1974 A dissipative Galerkin method for the numerical solution of first order hyperbolic equations. in Mathematical Aspects of Finite Elements in Partial Differential Equations (ed. C. de Boor), Academic Press, New York, 147-169.

Warming, R.F., Kutler, P. & Lomax, H. 1973 Second and third-order non-centred difference schemes for non-linear hyperbolic equations. AIAA Jnl. 11, 189.