SHOCK CAPTURING, FITTING AND RECOVERY

K. W. MORTON

NUMERICAL ANALYSIS REPORT 5/82

Invited talk presented at the 8th International Conference on Numerical Methods in Fluid Dynamics, Aachen, 28th June - 2nd July, 1982.

SHOCK CAPTURING, FITTING AND RECOVERY

K. W. Morton

University of Reading, Reading, England

1. INTRODUCTION

During the 1950's two very different approaches to numerical shock modelling were developed - shock fitting, in which the position and evolution of shocks is approximated explicitly, and shock capturing by finite difference methods operating on a fixed mesh. There was a prolonged lull in developments from the late 1950's to the early 1970's, as the two methods seemed to have reached a stalemate, but the last ten years have seen rapid progress on many fronts - theory of conservation laws, numerical methods and analysis of convergence. An indication of the situation is provided by the excellent survey of Gary Sod (1978). This stimulated widespread interest and even the best results given there now look very poor by comparison with those from current methods.

Yet still these mainly shock capturing methods lack the precision achievable with shock fitting techniques on the coarse meshes that must often be used in two and three dimensions. One way of achieving this and bridging the gap between the two viewpoints is to use shock recovery: the results obtained on a fixed grid by a shock capturing method are scanned for the presence of shocks whose positions and strengths are found and this information used in the subsequent evolution. The mesh may or may not be adjusted.

Such a viewpoint is natural to those working with finite elements and we shall consider here some recent developments of these methods for evolutionary problems which are pertinent to shock recovery.

In contrast to finite difference methods, the objective with finite elements is to create at each time level the best least squares fit to the solution from an appropriate space of trial functions. Typical choices are piecewise constants or piecewise linears. Obtaining intermediate values in order to model advection accurately or to calculate flux functions requires formulae comparable with those for polynomial interpolation. Recovery of sub-gridscale information, such as shocks or boundary layers, requires more specialised techniques which take maximum advantage of any available knowledge about the solution.

Let us consider first the problem of obtaining intermediate results from those given on a uniform mesh. Typically a finite difference method will provide grid-point values u_j at $x_j = jh$, j an integer; then, for example, mid-point values are given by truncating the interpolation formula

$$u_{j-\frac{1}{2}} \approx (1 - \tfrac{1}{8}\delta^2 + \tfrac{3}{128}\delta^4 - \ldots)\left(\frac{u_{j-1} + u_j}{2}\right), \qquad (1.1)$$

where $\delta^2 u_j := u_{j-1} - 2u_j + u_{j+1}$, in the standard notation. On the other hand suppose U_j are the nodal values defining a continuous piecewise linear approximation which is the best L_2 fit to $u(x)$: that is, the U_j are given by the Galerkin equations

$$\int [\sum_{(j)} U_j \phi_j(x) - u(x)] \phi_i(x) dx = 0, \qquad \forall i, \tag{1.2}$$

where $\phi_j(x)$ are the hat-shaped linear basis functions. Then intuitively we can see that the U_j will overshoot the true nodal values u_j of $u(x)$ and indeed we have the well-known recovery formula

$$u_j \sim (1 + \frac{1}{12}\delta^2 - \frac{1}{360}\delta^4 + \ldots)U_j . \tag{1.3}$$

For reasonably smooth $u(x)$, the three-term formula $12u_j \sim U_{j-1} + 10U_j + U_{j+1}$ is remarkably accurate. To compare with (1.1), suppose now we need the nodal values for a similar approximation on a mesh shifted by $\frac{1}{2}h$: then the natural formula is

$$(1 + \frac{1}{6}\delta^2)U_{j-\frac{1}{2}} \sim (1 + \frac{1}{24}\delta^2 + \frac{1}{384}\delta^4 \ldots)(\frac{U_{j-1} + U_j}{2}). \tag{1.4}$$

The operator on the left is natural since it corresponds to the mass matrix in the tridiagonal system of equations given by (1.2). We see that the implicit four point formula given by truncating (1.4) is nine times more accurate than the four point explicit formula given by (1.1). This is one of the bases of the greater accuracy that is often attainable with finite element methods - at the cost of more work because of the implicitness.

One of the overlap areas between finite differences and finite elements is provided by the use of piecewise constant approximations, as for instance in some finite volume methods. We shall use a notation in which the value U_j is understood to extend from $(j-\frac{1}{2})h$ to $(j+\frac{1}{2})h$ on a uniform mesh, or from $x_{j-\frac{1}{2}}$ to $x_{j+\frac{1}{2}}$ on a non-uniform mesh: the basis function, which we shall also denote by ϕ_j when there is no confusion, has unit value over this interval and is zero elsewhere. The Galerkin equations (1.2) for an L_2 best fit have a diagonal matrix in this case and U_j is the average of $u(x)$ over the interval, an interpretation which is shared by most finite difference schemes. In contrast to the piecewise linears, U_j undershoots the point values and the recovery formula corresponding to (1.3) becomes

$$u_j \sim (1 - \frac{1}{24}\delta^2 + \frac{3}{640}\delta^4 - \ldots)U_j . \tag{1.5}$$

For the shifted projection, we now have exactly the same formula for the $U_{j-\frac{1}{2}}$ as (1.1) for the point values. Thus the only important point about the interpretation of U_j as element averages is to use (1.5) to obtain u_j before, for instance, calculating flux functions.

All the above formulae apply to smooth underlying functions $u(x)$. In the presence of boundary layers or shocks, the recovery process is the same in principle but must use different recovery functions: while formulae like (1.3) may be derived by replacing u in (1.2) by an interpolating spline, Barrett & Morton (1980) in their work on diffusion-convection problems used recovery formulae of the general form

$$\langle U - \tilde{u}, \phi_i \rangle = 0, \tag{1.6}$$

in which the recovery function $\tilde{u}$ was exponential in form and the recovery was performed very locally. (Here and below we use the notation $\langle \cdot, \cdot \rangle$ to denote the L_2 inner product in the space variables). The main point is to use whatever information is known about the approximated function $u(x)$ and to exploit the fact that $U(x)$ is its L_2 best fit - or nearly so in evolutionary problems.

In the next and main section of the paper, we shall describe Characteristic Galerkin methods for approximating unsteady conservation laws and their application to shock capturing and recovery. Finite element methods are strongly based on the Galerkin formulation but as their development for problems other than those which are steady, linear, elliptic and self-adjoint progresses one finds that the formulation has to be generalised and the Characteristic Galerkin methods result from such a generalisation. This part of the paper is not meant as a review nor is it primarily intended to promulgate new methods which have been widely tested on practical problems: it is rather my aim to present a new viewpoint and to show how this naturally links together important ideas and algorithms developed by Godunov (1959), Boris & Book (1973), van Leer (1979), Engquist & Osher (1980), Roe (1981) and several others. In the final section of the paper the moving finite element method due to Gelinas, Doss & Miller (1981) will be described from the same viewpoint and developments of it to capture shocks presented briefly.

2. CHARACTERISTIC GALERKIN SCHEMES

2.1 Basic ECG scheme

Consider the scalar conservation law in one dimension

$$\partial_t u + \partial_x f(u) = 0 \tag{2.1a}$$

or

$$\partial_t u + a(u)\partial_x u = 0, \tag{2.1b}$$

where $a(u) = \partial f/\partial u$: and suppose $u(x,t)$ is approximated at time level $n\Delta t$, in terms of basis functions $\phi_j(x)$, by

$$U^n(x) = \sum_{(j)} U_j^n \phi_j(x). \tag{2.2}$$

Then the characteristic Galerkin method based on Euler time-stepping, and hence called the Euler Characteristic Galerkin or ECG method by Morton & Stokes (1981)

and Morton (1982), takes the form

$$\langle U^{n+1} - U^n, \phi_j \rangle + \Delta t \langle \partial_x f(U^n), \phi_j^n \rangle = 0. \tag{2.3}$$

Here ϕ_j^n is the upwind-averaged test function,

$$\phi_j^n(x) := \frac{1}{a^n(x)\Delta t} \int_x^{x+a^n(x)\Delta t} \phi_j(z)dz, \tag{2.4}$$

where $a^n(x) := a(U^n(x))$. The method was introduced in order to model accurately the advection of continuous profiles. It is equivalent to exactly tracing the evolution of $U^n(x)$ along characteristics by the equation

$$u(y,t + \Delta t) = u(x,t) \quad \text{where} \quad y = x + a(u(x,t))\Delta t \tag{2.5}$$

and following this by L_2 projection onto $\text{span}\{\phi_j\}$.

A continuous piecewise linear approximation is appropriate for accurate advection and the corresponding test function, when the characteristic speed a^n is constant and for various positive values of the CFL number $\mu = a\Delta t/h$, is shown in Fig. 1. The effectiveness of such a scheme in advecting a steep ramp function over a coarse grid is shown by Fig. 2: the small scale oscillation shown there is typical of least squares best fits by piecewise linears and it is seen that there is little change in it from that produced by the projection of the initial data. In this constant coefficient case on a uniform mesh, the scheme (2.3) reduces, in terms of nodal values, to

$$(1 + \tfrac{1}{6}\delta^2)(U_j^{n+1} - U_j^n) + \mu(\Delta_0 - \tfrac{\mu}{2}\delta^2 + \tfrac{\mu^2}{6}\delta^2\Delta_-)U_j^n = 0 \tag{2.6}$$

and corresponds to shifting the projection by the distance $a\Delta t$ that the characteristic has travelled in one time step. Clearly for $\mu = \frac{1}{2}$ the formula (2.6) reduces to that obtained from the first two terms of (1.4) and the ninefold improvement of this over (1.1) is consistent with the advantage of (2.6) over any of the usual explicit four point difference schemes for advection. As it stands (2.3) may be quite expensive to evaluate exactly but several efficient approximations of the ideal test function ϕ_j^n which reproduce (2.6) are given by Morton (1982).

2.2 Shock modelling with piecewise constants

In shock modelling, however, discontinuous approximations are often favoured for either finite difference or finite element schemes. Let us consider piecewise constants first. Then (2.3) needs careful interpretation even for rarefaction waves because not only is $f(U^n)$ discontinuous but so is ϕ_j^n, through the dependence on $a(U^n)$, and these discontinuities coincide. Several finite difference methods have used the idea of resolving the jumps in $U^n(x)$ by the correct physical rarefaction fans. This could be done here but is unnecessary if we remember that, even if the

objectives of our calculation are perfectly achieved, $U^n(x)$ is only the L_2 projection of the exact solution at time $n\Delta t$: thus the exact evolution of this approximate solution is hardly justified. Suppose then that $a(U_k^n) =: a_k^n > a_{k-1}^n$ so that (2.5) leaves a gap in the definition of the mapping $y \to x$ and hence (2.3) is ill-defined. We need use our knowledge of the exact solutions of the differential equation only to the extent of recognising that the jump in $U^n(x)$ would have been resolved. So let us suppose, for instance, that $U^n(x)$ is the projection of a function $\tilde{u}^n(x)$ which is linear between $(k-\frac{1}{2}-\frac{1}{2}\theta)h$ and $(k-\frac{1}{2}+\frac{1}{2}\theta)h$, with $0 < \theta < 1$, and takes the constant value $\tilde{u}_k^n$ in $[(k-\frac{1}{2}+\frac{1}{2}\theta)h, kh]$ and the value $\tilde{u}_{k-1}^n$ in $[(k-1)h,(k-\frac{1}{2}-\frac{1}{2}\theta)h]$. Then $\partial_x f(\tilde{u}^n)$ is well defined and a little algebra shows that, if $a(\tilde{u}^n) > 0$, $\frac{1}{2}\theta < \min a(\tilde{u}^n)\Delta t/h$ and $\max a(\tilde{u}^n)\Delta t/h \le 1$ for $x \in [(k-1)h, kh]$,

$$\int_{(k-1)h}^{kh} \partial_x f(\tilde{u}^n)\tilde{\Phi}_j^n \, dx = \begin{cases} \Delta_- \tilde{f}_k^n - \frac{\theta h}{8\Delta t}\Delta_-\tilde{u}_k^n & \text{for } j = k \\ \frac{\theta h}{8\Delta t}\Delta_-\tilde{u}_k^n & \text{for } j = k-1, \end{cases} \tag{2.7}$$

where $\tilde{f}$ and $\tilde{\Phi}$ are defined using $\tilde{u}$ rather than U. Suppose also that a similar jump exists at $(k+\frac{1}{2})h$ and is similarly resolved. Then combining the two results we obtain from (2.3)

$$h(U_k^{n+1} - U_k^n) + \Delta t(\Delta_-\tilde{f}_k^n + \frac{\theta h}{8\Delta t}\delta^2\tilde{u}_k^n) = 0 \, . \tag{2.8}$$

Furthermore, from the fact that U^n is the projection of $\tilde{u}^n$, we easily deduce that

$$U_k^n = [1 + (\theta/8)\delta^2]\tilde{u}_k^n, \tag{2.9}$$

so that (2.8) becomes

$$U_k^{n+1} = \tilde{u}_k^n - (\Delta t/h)\Delta_-\tilde{f}_k^n. \tag{2.10}$$

In other words, U^{n+1} depends only on the constant sections of $\tilde{u}^n$ and the dependence on the parameter θ appears only in the recovery formula (2.9). We will below generally use the limiting case $\theta \to 0$ so that (2.10) becomes just the familiar first order upwind scheme: however, it should be noted that increasing θ will reduce the false diffusion that ruins this scheme for smooth flows and, indeed, taking $\theta = 1$ to make $\tilde{u}^n$ piecewise linear with knots at jh gives a scheme similar to that of Fromm (1968).

For a general rarefaction case, suppose $f(u)$ has just a single sonic point $\bar{u}$; that is $a(\bar{u}) = 0$ and otherwise $a(u) \neq 0$. Then from (2.4), we see that

$$\tilde{u}^n(x) = \bar{u}, \ \phi_j(x) = 1 \Rightarrow \tilde{\phi}_j^n(x) = 1. \tag{2.11}$$

It is also easy to see that if the CFL condition

$$a(u)\Delta t/h < 1 \quad \text{for } u \in [\tilde{u}_{k-1}^n, \tilde{u}_k^n] \tag{2.12}$$

is satisfied, then

$$\tilde{\phi}^n_{k-1}(x) + \tilde{\phi}^n_k(x) = 1 \quad \text{for} \quad x \in [(k-1)h, kh]. \tag{2.13}$$

Thus we can deduce that, in the limit $0 \to 0$, (2.3) yields the following algorithm:

$$a^n_{k-1},\ a^n_k \geq 0 \ : \ \text{use} \ \Delta_- f^n_k \ \text{to update} \ U^n_k \to U^{n+1}_k \tag{2.14a}$$

$$a^n_{k-1},\ a^n_k \leq 0 \ : \ \text{use} \ \Delta_- f^n_k \ \text{" "} \ U^n_{k-1} \to U^{n+1}_{k-1} \tag{2.14b}$$

$$\begin{aligned} a^n_{k-1}\, a^n_k < 0 \ : \ &\text{use} \ f^n_k - f(\bar{u}) \ \text{" "} \ U^n_k \to U^{n+1}_k \\ &\text{use} \ f(\bar{u}) - f^n_{k-1} \ \text{to update} \ U^n_{k-1} \to U^{n+1}_{k-1}. \end{aligned} \tag{2.14c}$$

It is clear that this algorithm is exactly equivalent in this case to the key flux-splitting idea of Engquist & Osher (1980), in which U_j is updated using $\Delta_- f^+_j + \Delta_+ f^-_j$. On the other hand, the derived fluxes f^-_j, f^+ are not introduced and the algorithm has the form of those due to Roe (1981), in which the total flux difference $\Delta_- f_j$ between each pair of intervals is in turn broken up into contributions to update U in these (and possibly neighbouring) intervals. We shall see that this feature of sharing the properties of these two finite difference schemes occurs naturally in all our ECG schemes. Note that if $f(u)$ has several sonic points then the contribution from each which is correctly given by the Engquist-Osher scheme is picked up by the ECG scheme (2.3) only if the correct physical waves are used in the recovery process rather than piecewise linears for $\tilde{u}$.

Suppose now that we have a shock at $(k-\frac{1}{2})h$, that is that $a^n_{k-1} > a^n_k$. Then the characteristics drawn from points just to the left and just to the right of $(k-\frac{1}{2})h$ overlap, and the mapping (2.5) gives a multivalued $y(x)$ and hence a multivalued $u(x,t+\Delta t)$. However, (2.3) can still be properly defined and we have purposely delayed the derivation of this formula until we came to this case. We suppose that we have recovered from $U^n(x)$ a continuous function $\tilde{u}^n(x)$, such as that used in obtaining (2.7), which is determined by the projection relation

$$\langle U^n - \tilde{u}^n, \phi_j \rangle = 0 \qquad \forall j. \tag{2.15}$$

Then $y(x) = x + a(\tilde{u}^n(x))\Delta t$ defines a continuous (x,y) path which gives a possibly multivalued mapping $y \to x$. We define $U^{n+1} \in \text{span}\{\phi_j\}$ by

$$\langle U^{n+1}, \phi_j \rangle = \int_{-\infty}^{\infty} \tilde{u}^n(x(y))\phi_j(y)dy, \tag{2.16}$$

the integral being defined along the (x,y) path. Then, introducing $\tilde{\phi}^n_j(x)$ by (2.4), we have

$$\begin{aligned} \langle U^{n+1} - U^n, \phi_j \rangle &= \int_{-\infty}^{\infty} \tilde{u}^n(x)\, d\Big[\int_x^y \phi_j(z)dz\Big] \\ &= -\int_{-\infty}^{\infty} d\,[\tilde{u}^n(x)] \int_x^y \phi_j(z)dz = -\Delta t \int_{-\infty}^{\infty} f_x(\tilde{u}^n(x))\tilde{\phi}^n_j(x)dx, \end{aligned}$$

i.e. $$\langle U^{n+1} - U^n, \phi_j \rangle + \Delta t \langle \partial_x f(\tilde{u}^n), \tilde{\phi}_j^n \rangle = 0. \qquad (2.17)$$

This is now the general formula with which we shall work. For the case of a shock at $(k-\frac{1}{2})h$ with piecewise constant elements, it is clear that where $f_x \neq 0$ we have $dy < 0$ and the calculation is the same as for the rarefaction case but with $\tilde{u}^n_{k-1}$ and $\tilde{u}^n_k$ interchanged. The net effect is to yield the same algorithm as given in (2.14) which is now shown to cover all possible cases.

2.3 Piecewise linear basis functions

For continuous piecewise linear functions no recovery is necessary and (2.3) may be used directly even in the presence of shocks, though there may be considerable loss of accuracy then, which we shall discuss below in 2.4. It is convenient again to arrange the algorithm to deal in turn with the contributions from $\partial_x f(U^n)$ arising from each interval. For $x \in [x_{k-1}, x_k]$, we have $\phi_{k-1}(x) + \phi_k(x) = 1$ and if we assume the CFL condition holds locally then

$$\Phi^n_{k-2}(x) + \Phi^n_{k-1}(x) + \Phi^n_k(x) + \Phi^n_{k+1}(x) = 1 \quad \text{for} \quad x \in [x_{k-1}, x_k] \qquad (2.18)$$

with only the first three non-zero if $a(x) > 0$ and the last three if $a(x) < 0$. Thus the total contribution to updating up to four nodal values of $U^n(x)$ is again $\Delta_- f^n_k$. In general the integrals will need to be approximated by quadrature rules but always this property should be retained. Also, of course, as in (2.6), a tri-diagonal system has to be solved for the nodal values once all contributions to the updating have been accumulated.

For shock modelling, van Leer (1979) has suggested that discontinuous piecewise linear approximations should be used and has developed in his MUSCL code an approximate Riemann solver for them. Usually, basis functions consisting of piecewise constants and of linear functions varying from -1 to +1 over an interval have been used for such schemes. However, for an ECG scheme it is more convenient to use the two parts of the continuous linear basis functions, namely $\phi_{j+}(x)$, which varies from 1 at x_j to 0 at x_{j+1}, and $\phi_{j-}(x)$, which varies from 1 at x_j to 0 at x_{j-1}. Then we can write

$$U^n(x) = \sum_{(j)} [U^n_{j+} \phi_{j+}(x) + U^n_{j-} \phi_{j-}(x)] \qquad (2.19)$$

with the jump at x_j equal to $U^n_{j+} - U^n_{j-}$. The sum of test functions in (2.18) splits into eight parts, but each is non-negative: the two outermost terms Φ^n_{k-2-} and Φ^n_{k+1+} can be discarded and for each particular x only four terms are non-zero. Thus contributions from $\partial_x f(U^n)$ for $x \in (x_{k-1}, x_k)$ are distributed at worst to six nodal values of U^n, and usually only to four. As in the continuous case above, these contributions need to be approximated by quadrature rules.

The contributions from the jump in $f(U^n)$ at x_k can be calculated however as for the piecewise constant case. There is a little more dependence on the form

used to resolve the discontinuity but we note that

$$\Phi_{k+}(x) + \Phi_{k-}(x) + \Phi_{k+1-}(x) + \Phi_{k-1+}(x) = 1 \quad \text{for} \quad |x-x_k| \le c \qquad (2.20)$$

for sufficiently small c so that only these four corresponding nodal values can be affected. The only uncertainty in the calculation results from the fact that each term in (2.20) may be discontinuous at x_k because of the discontinuity in $a(U^n)$: the simplest resolution of the uncertainty is to define $\Phi_j(x_k)$ as the mean of the limits from above and below. Then, dropping unnecessary superscripts, the allocation of updates corresponding to (2.14) can be set out as follows:

$$\begin{aligned} a_{k-},\ a_{k+} \ge 0 : \quad & [f_{k+}-f_{k-}]\Phi_{k+}(x_k) \ \text{to} \ \tfrac{1}{3}U_{k+} + \tfrac{1}{6}U_{k+1-} \\ & [f_{k+}-f_{k-}]\Phi_{k+1-}(x_k) \ \text{to} \ \tfrac{1}{6}U_{k+} + \tfrac{1}{3}U_{k+1-} \end{aligned} \qquad (2.21a)$$

$$\begin{aligned} a_{k-},\ a_{k+} \le 0 : \quad & [f_{k+}-f_{k-}]\Phi_{k-}(x_k) \ \text{to} \ \tfrac{1}{6}U_{k-1+} + \tfrac{1}{3}U_{k-} \\ & [f_{k+}-f_{k-}]\Phi_{k-1+}(x_k) \ \text{to} \ \tfrac{1}{3}U_{k-1+} + \tfrac{1}{6}U_{k-} \end{aligned} \qquad (2.21b)$$

$$\begin{aligned} a_{k+}a_{k-} < 0 : \quad & [f_{k+}-f(\bar{u})]\Phi_{k+}(x_k) \ \text{to} \ \tfrac{1}{3}U_{k+} + \tfrac{1}{6}U_{k+1-} \\ & [f_{k+}-f(\bar{u})]\Phi_{k+1-}(x_k) \ \text{to} \ \tfrac{1}{6}U_{k+} + \tfrac{1}{3}U_{k+1-} \\ & [f(\bar{u})-f_{k-}]\Phi_{k-}(x_k) \ \text{to} \ \tfrac{1}{6}U_{k-1+} + \tfrac{1}{3}U_{k-} \\ & [f(\bar{u})-f_{k-}]\Phi_{k-1+}(x_k) \ \text{to} \ \tfrac{1}{3}U_{k-1+} + \tfrac{1}{6}U_{k-} \end{aligned} \qquad (2.21c)$$

Choosing the basis functions $\Phi_{j\pm}$ to be non-negative, and therefore ensuring that the corresponding $\Phi_{j\pm}$ have the same property, makes it very much easier to introduce approximations to the latter to be used in these formulae and in the quadrature over each open interval (x_{k-1},x_k). The penalty for not having an orthogonal basis is very minor: for each interval is independent of the others and the two nodal parameters at its ends, U_{j+} and U_{j+1-} say, are given by a simple pair of equations.

2.4 Shock recovery

It was shown by Cullen & Morton (1980) with the shallow water equations how for smooth flows the accuracy of a purely Galerkin procedure could be greatly improved on by a two stage procedure for approximating the non-linear terms $\underline{u}\cdot\nabla\underline{u}$: from a piecewise linear approximation to $\underline{u}$ they formed a best least squares fit to $\nabla\underline{u}$ by piecewise linears before doing the same for the product $\underline{u}\cdot\nabla\underline{u}$. This can be regarded as a simplified recovery procedure: for in one dimension the first step corresponds to forming a quadratic spline approximation to $\underline{u}$. More general use of recovery procedures was envisaged by Barrett & Morton (1980) for diffusion-convection problems and a number of results in one dimension are collected together by Barrett, Moore & Morton (1982).

In the context of modelling conservation laws recovery of u from a best fit in any integral norm is clearly very important before evaluating the flux $f(u)$. When a piecewise constant approximation is used, peaks are cut off and need to be restored where possible: and for a continuous piecewise linear approximation, overshoots occur which need to be smoothed out. In the former case we have already seen, in the course of interpreting the basic ECG scheme (2.3), that recovery by piecewise linears to give (2.17) can restore some of the information lost by false diffusion. There is a similarity here with the philosophy of SHASTA codes (Boris & Book, 1973) and an appropriate algorithm for the choice of the parameter θ in (2.7) would seem to put this scheme into the more general class of flux-corrected transport algorithms described by Zalesak (1979). For the continuous piecewise linears, we have also already given the spline-derived recovery formula (1.5) which can be used very simply to improve accuracy in smooth parts of the flow.

But it is in the neighbourhood of shocks that enhancement of accuracy is most important. Loss of information clearly results primarily from the use of a fixed set of mesh points if discontinuous approximations are used: even for continuous linear elements, fixed nodes together with the resulting fixed spacing is the prime cause of inaccuracy. Thus for each approximation U, we define a recovery function $\tilde{u}$ which has discontinuities just where we deduce that the exact solution u has shocks. There are two steps involved:

(i) recognition of the presence and position of the shock;

(ii) estimation of the shock parameters.

As Rusanov has pointed out (see Rusanov (1981) and references therein) every difference scheme has its own limiting shock profile, and these can be used to deal with (i). Generally speaking this will mean scanning for a local maximum in $-\Delta_- U_j$, that is checking a criterion such as

$$\delta^2 \Delta_- U_j > \text{tol}. \tag{2.22}$$

and this was what was used in the work reported in (Morton, 1980). However, it is more natural with ECG schemes to monitor the crossing of characteristics in each interval and to use a criterion (with ν a free parameter) like

$$a(U_{j-1}) - a(U_j) > (x_j - x_{j-1})/\nu\Delta t \tag{2.23}$$

to detect a shock. This can in fact be used for either piecewise constants or piecewise linear elements: note that the criterion does not have to be completely foolproof since our schemes are valid even in the presence of a shock and we are merely seeking to enhance their accuracy. The estimation of the shock parameters will however be dependent on the approximation used.

For the piecewise constant approximation, the simplest configuration of a steady shock between $x_{k-\frac{1}{2}}$ and $x_{k+\frac{1}{2}}$ linking two constant states clearly leads to a

projection with one intermediate value U_k. If the shock is strong enough this situation will be recognised by (2.23) being satisfied for the two isolated values $j = k, k+1$: alternatively it may be desirable to replace the left-hand side of (2.23) by $a(U_{j-1}) - a(U_{j+1})$ which would be recognised for $j = k$. In either case from the three nodal values involved we can recover

$$\tilde{u}_L = U_{k-1},\ \tilde{u}_R = U_{k+1},\ \eta = \Delta_+ U_k/(\Delta_+ U_k + \Delta_- U_k) \tag{2.24}$$

with $x_S = (1-\eta)x_{k-\frac{1}{2}} + \eta x_{k+\frac{1}{2}}$ giving the shock position. In contrast to shock fitting, the idea is now not to attempt to follow the shock movement but to incorporate this recovered information into the general ECG scheme (2.17): there is an independent choice as to whether (2.7) with $\theta > 0$ should be used for recovery between the shocks that have been identified. Taking $\theta = 0$ for simplicity, the effect of introducing the recovered shock defined by (2.24) into the ECG formula (2.17) is just to change the allocation of contributions to the updating process arising from $\Delta_+ f_k^n$ and $\Delta_- f_k^n$. The shock jumps from U_{k-1}^n to U_{k+1}^n so compared with (2.14) the allocation process is applied to $\Delta_- f_k^n + \Delta_+ f_k^n$ and is dependent on a_{k-1}^n and a_{k+1}^n. The other difference from (2.14) is the dependence on the shock position x_S: if for instance $a_{k-1}^n,\ a_{k+1}^n \geq 0$ and η is sufficiently small then $\tilde{\phi}_{k+1}^n(x_S^n) = 0$ and an allocation as in (2.14a) takes place; but for larger η, $\tilde{\phi}_{k+1}^n(x_S^n) \neq 0$ and a contribution to U_{k+1}^n results. Thus the general formulae will be like (2.21) for the piecewise linear discontinuous case: as there, ϕ_j^n is generally discontinuous at x_S^n and needs to be defined as an average of limits from the left and right. To sum up we give the formula for the simplest case, dropping the superscripts:

$$\begin{aligned} a_{k-1}, a_{k+1} \geq 0 :\ & [f_{k+1} - f_{k-1}]\phi_k(x_S) \quad \text{to} \quad U_k \\ & [f_{k+1} - f_{k-1}]\phi_{k+1}(x_S) \quad \text{to} \quad U_{k+1}. \end{aligned} \tag{2.25}$$

This is sufficient to show that in this case, in the situation envisaged in the derivation of (2.24) and for small time steps, the recovered shock moves with the correct speed as given by the Rankine-Hugoniot condition

$$\frac{x_S^{n+1} - x_S^n}{\Delta t} = \frac{f_{k+1} - f_{k-1}}{U_{k+1} - U_{k-1}}. \tag{2.26}$$

The superscripts to f and U have been omitted here as $U_{k\pm1}$ are not changed in this simple case. If the CFL condition is satisfied with this shock speed one can show that the monotonicity of U is preserved.

We conclude this section with a case in which shock recovery is clearly necessary, the continuous piecewise linear approximation on a uniform mesh. It is relatively easy to recognise shocks which are at least four mesh widths apart by either (2.22) or (2.23): in contrast to the piecewise constant case, a shock between x_{k-1} and

x_k typically leads to an overshoot of U_{k-1} and an undershoot of U_k so that (2.23) will be satisfied just for $j = k$. One may then obtain the shock parameters as part of a global recovery process as given by (1.6) with $\tilde{u}$ consisting of shocks joined by cubic splines with knots at the original nodes, and this works well. However, an alternative is to use a simple explicit local recovery formula of the form

$$\tilde{u}_L = \langle U, \phi_{k-2} + \phi_{k-1}\rangle/2h, \quad \tilde{u}_R = \langle U, \phi_k + \phi_{k+1}\rangle/2h \tag{2.27a}$$

$$\eta = [h^{-1}\langle U, \phi_{k-1} + \phi_k\rangle - \tfrac{1}{2}\tilde{u}_L - \tfrac{3}{2}\tilde{u}_R]/[\tilde{u}_L - \tilde{u}_R]. \tag{2.27b}$$

To incorporate the recovered shock into (2.17) one can either make use of the results obtained with the discontinuous linear elements or, more straightforwardly, introduce two extra nodes at $x_S \pm \frac{1}{2}\varepsilon$, $\varepsilon \ll h$, and give u a linear variation between them. Then unless recovery is used in the smooth flow, which is hardly necessary, the formula (2.17) is little changed from the standard treatment of continuous piecewise linears: one has only to note that, in the definition of $\tilde{\Phi}_j$ for instance, $\tilde{a}$ has extra nodes compared with ϕ_j.

2.5 Extensions to two dimensions, systems, etc.

It is a simple matter in principle to extend (2.3) and (2.17) into more dimensions with Φ_j defined as in (2.4) by an upwind average of the basis function ϕ_j:

$$\Phi_j^n(\underline{x}) = \frac{1}{|\underline{a}^n(\underline{x})|\Delta t}\int_{\underline{x}}^{\underline{x}+\underline{a}\Delta t} \phi_j(\underline{y})d\ell\,. \tag{2.28}$$

For continuous linear elements on triangles the details together with a practical approximate scheme are given in (Morton, 1982). The derivation of formulae for piecewise constants corresponding to (2.14) is perhaps more interesting. From a flux vector $\underline{f}$ one obtains contributions from $\underline{\nabla}\cdot\underline{f}$ just along the edges of each triangle and clearly it is only $\partial_n f_n$, the normal derivative of the normal component of $\underline{f}$, which plays a rôle. Just as in one dimension, one can spread the discontinuity in u and $\underline{f}$ across a thin strip either side of the edge and take limits as the strip shrinks to zero. However, the more complicated geometry now means that several Φ_j may be non-zero along the edge and correspondingly several U_j be affected in the update. A typical situation is shown in Fig. 3, and the update formulae take more the form of those in (2.21) and (2.25). A typical contribution to $U_j^{n+1} - U_j^n$ will be

$$-\Delta t V_j^{-1}[f_n]_e \int_e \tfrac{1}{2}[\Phi_j^n(\underline{x}_+) + \Phi_j^n(\underline{x}_-)]d\ell, \tag{2.29}$$

where V_j is the area of element j, $[f_n]_e$ is the jump in f_n across the edge and the integral of Φ_j^n along either side of the edge, using corresponding characteristic speeds $\underline{a}_+^n$ and $\underline{a}_-^n$, can be calculated as indicated in Fig. 3: a trapezium

is constructed along the edge using the vector $\underline{a}_{\pm}\Delta t$ and the length of the edge is allocated to each element in proportion to the area of the trapezium lying in each element. Recalling that, in one dimension and for convex f, (2.14) is identical with the Engquist-Osher algorithm, it is interesting to compare the formulae resulting from (2.29) with those obtained from their scheme in two dimensions: thus in Osher (1981) only the edge lengths are used rather than the trapezia which reduce to them when $\Delta t \to 0$; and also the normal component a_n of $\underline{a}$ rather than $|\underline{a}|$ is used and zeros in this lead to a splitting of the flux difference as in (2.14c).

Extensions of the ECG schemes to systems of equations is of course both more important and more difficult. From the derivation leading to (2.17), one sees that in effect one is constructing the exact solution at time level $n+1$ corresponding to the approximate solution at level n using the characteristic relation (2.5) and then projecting it. Thus for piecewise constants the correct generalisation is provided by the method of Godunov (1959) in which the Riemann problem is solved for each discontinuity in $\underline{U}^n$ and the result projected back onto the piecewise constants: or, alternatively and more in line with the use of (2.5), one could say that the Engquist-Osher algorithm, which resolves each discontinuity by using the full set of rarefaction and compression waves for the system on the overturned manifolds created by the crossing characteristics, is the most appropriate generalisation. Furthermore, work is in progress on developing these approaches to piecewise linear discontinuous elements.

However, early experiments indicate that more direct generalisations of (2.17) could form practical and effective alternatives to these established approaches. From considering the characteristic normal form for the system, it is clear that $\Phi_j(\lambda_i)$ must in principle be constructed for each characteristic speed λ_i: these then form a diagonal matrix so that transforming back to the original variables suggests approximation schemes of the form

$$\langle \underline{U}^{n+1} - \underline{U}^n, \phi_j \rangle + \Delta t \langle \partial_x \underline{f}(\underline{U}^n), \underline{\phi}_j^n \rangle = 0 \qquad (2.30a)$$

where

$$S^{-1}\,\underline{\Phi}_j^T S = \text{diag.}\,\{\Phi_j(\lambda_i)\},\quad S^{-1}AS = \text{diag.}\,\{\lambda_i\} \qquad (2.30b)$$

and A is the Jacobian matrix of the system. To calculate this exactly is clearly expensive: but recall that in (2.14) with piecewise constants only the signs of λ_i are important and that we have ensured that all the Φ_j are necessarily positive. Thus there is considerable scope for effective approximation, just as the Engquist-Osher and the Roe schemes owe much of their success to approximate solution of the Riemann problem.

Finally, it should be mentioned that though Euler time-stepping is perfectly adequate for the scalar problem (2.1) because the characteristics are straight, for

systems it may be desirable to use other time-stepping algorithms: for in (2.30b) not only are the λ_i changing but so is the matrix S. Thus unlike (2.3), (2.30) is only an approximation to evolution over one time step followed by projection, both because of the x-variation of S and the t-variation of both S and λ_i. Some of this might be improved by alternative time-stepping and it is not difficult to calculate what the corresponding Φ_j should be for some of the common schemes.

2.6 Numerical tests

Each of the schemes has been tested with the inviscid Burger's equation and tests continue with the Sod (1978) problem. Fig. 4 shows initial data and its exact evolution for the Burger test: the $\cos^2\frac{1}{2}\pi x$ data forms a shock at $t = 2/\pi$; the ramp tests resolution of a rarefaction shock. Fig. 5 shows that for piecewise constants simple shock recovery is extremely effective in improving accuracy. For continuous linears it works well (Fig. 6) but the shock recognition test used needs improving: Fig. 7 shows that recovery is essential with discontinuous linears to control the overshoots and get useful results.

3. MOVING FINITE ELEMENT SCHEMES

Having explored in the previous section to what extent one can carry forward the best L_2 fit to the true solution on a fixed mesh, and having found it helpful to introduce extra nodes, it is a natural next step to include the node positions in the L_2 fitting. This is the starting point for the development of moving finite element methods such as that of Gelinas, Doss & Miller (1981). With continuous piecewise linear elements and using their notation, one seeks approximate solutions in the form

$$v(x,t) = \sum_{(j)} a_j(t)\, \alpha_j(x, \underset{\sim}{s}(t)), \tag{3.1}$$

where α_j are the usual hat-shaped linear basis functions but based on the set of nodes denoted by the vector $\underset{\sim}{s}$. One then has

$$\partial_t v = \sum_{(j)} [\, \dot{a}_j \alpha_j(x,\underset{\sim}{s}) + \dot{s}_j \beta_j(x,\underset{\sim}{a},\underset{\sim}{s})], \tag{3.2}$$

where the β_j are discontinuous basis functions depending on the amplitudes $\underset{\sim}{a}$ as well as the node positions $\underset{\sim}{s}$. Unlike U^n in (2.2), v in (3.1) lies not in a linear but in a non-linear manifold determined by the parameters $\underset{\sim}{a}$ and $\underset{\sim}{s}$, while $\partial_t v$ lies in its linear tangent space. Equations for $\underset{\sim}{a}$ and $\underset{\sim}{s}$ are obtained by taking an L_2 best fit for $\partial_t v$ in this tangent space: thus for the conservation law (2.1) one obtains

$$\langle \partial_t v + \partial_x f(v), \alpha_j \rangle = 0 = \langle \partial_t v + \partial_x f(v), \beta_j \rangle . \tag{3.3}$$

This gives a set of ordinary differential equations for $\underset{\sim}{a}$ and $\underset{\sim}{s}$ to be integrated by an appropriate ODE solver.

In the scalar case this last task is greatly simplified by the observation that if the flux function is quadratic then the nodes exactly follow the characteristics so that $\dot{\underline{\delta}} = 0$ and $\dot{\underline{s}} = \text{const}$. Thus even Euler's method is exact for any time step in this case and quite good enough for many problems.

To deal with shocks, Gelinas et al. introduced spring functions into the objective function, rather than just the L_2 norm, so as to prevent node overtaking. However with moving nodes it does seem natural to capture shocks explicitly. This has been done by Wathen (1982) in work on oil recovery problems. A shock is recognised by neighbouring nodes overtaking one another: and when this happens, two coincident nodes with differing amplitudes are followed by satisfying the Rankine-Hugoniot conditions. Results for a standard model problem using the Buckley-Leverett equations in which $f(v) = v^2/[v^2 + \frac{1}{2}(1-v)^2]$ are given in Fig. 8a.

For a system of equations, corresponding to more than two phases in the oil recovery problem, only one set of nodes is used so that clearly some compromise has to be struck as to which characteristics are followed most closely and so as to be able to recognise shocks by the phenomenon of node-crossing. Such a compromise is introduced by using a matrix weighting function W in the L_2 norm for $\partial_t \underline{v}$. Preliminary results obtained by Wathen for a three phase problem are shown in Fig. 8b.

ACKNOWLEDGEMENTS

I am greatly indebted to Stanley Osher, Phil Roe and Bram van Leer for lengthy discussions on the latest developments in finite difference methods for shock modelling, and for stimulating our application of ECG methods to these problems. Thanks are due to Alan Stokes for the calculations in Section 2 and to Andy Wathen for those in Section 3.

REFERENCES

Barrett, J.W., Moore, G. & Morton, K.W., 1982. Optimal recovery and defect correction in the finite element method. Univ. of Reading, Num. Anal. Report 7/82.

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.

Boris, J.P., & Book, D.L., 1973. Flux corrected transport.I SHASTA a fluid transport algorithm that works. J. Comp. Phys. 11, pp38-69.

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

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.

Fromm, J.E., 1968. A method for reducing dispersion in convective difference schemes. J. Comp. Phys. 3, 176-189.

Gelinas, R.J., Doss, S.K., & Miller, K., 1981. The moving finite element method: applications to general partial differential equations with multiple large gradients. J. Comp. Phys. 40, 202-249.

Godunov, S.K., 1959. A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics. Mat. Sb. 47, 271-290.

Morton, K.W., 1982. 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, (to appear).

Morton, K.W., 1980. Petrov-Galerkin methods for non-self-adjoint problems. Proc. Dundee Conf. on Numerical Analysis, (ed. G.A. Watson), Lect. Notes Math. 773, Springer-Verlag, 110-116.

Morton, K.W., & Stokes, A. Generalised Galerkin methods for hyperbolic equations. Proc. MAFELAP 1981 Conf. (ed. J.R. Whiteman), (to appear).

Osher, S., 1981. Numerical solution of singular perturbation problems and hyperbolic systems of conservation laws. Conf. Proc., North-Holland Math. Studies 47, 179-205.

Roe, P.L., 1981. The use of the Riemann problem in finite difference schemes. Proc. VIIth Int. Conf. on Num. Meth. in Fluid Dynamics, Lect. Notes Phys. 141, Springer-Verlag, 354-9.

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

Rusanov, V.V., 1981. On the computation of discontinuous multi-dimensional gas flows. Proc. VIIth Int. Conf. Num. Meth. in Fluid Dynamics, Lect. Notes Phys. 141, Springer-Verlag, 31-43.

Sod, G.A., 1978. A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. J. Comp. Phys. 27, 1-31.

Van Leer, B., 1979. Towards the ultimate conservative differencing scheme V. A second order sequel to Godunov's method. J. Comp. Phys. 32, 101-136.

Wathen, A., 1982. Moving finite elements and applications to some problems in oil reservoir modelling. Univ. of Reading, Num. Anal. Report 4/82.

Zalesak, S., 1979. Fully multidimensional flux-corrected transport algorithms for fluids. J. Comp. Phys. 31, 335-362.

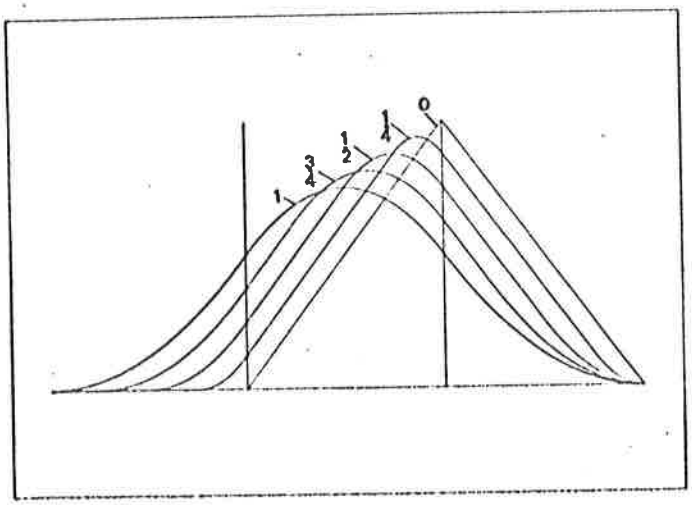


Fig. 1 Upwind averaged test functions for $\mu = 0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$.

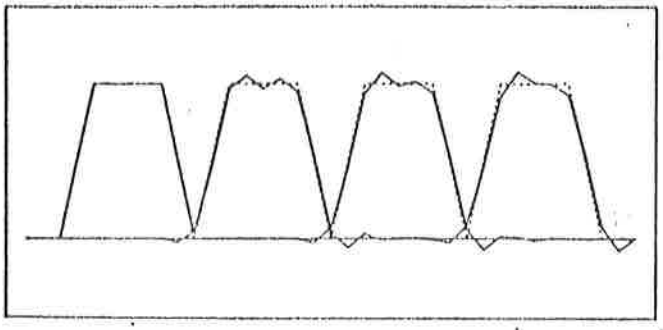


Fig. 2 Linear advection by ECG through 0, 20, 40 time steps with $\mu = 0.8$.

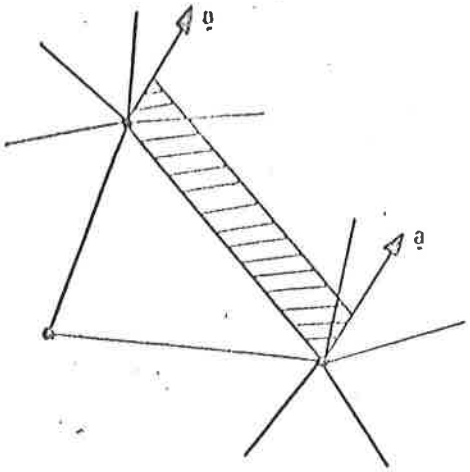


Fig. 3 Allocation of flux differences in 2D.

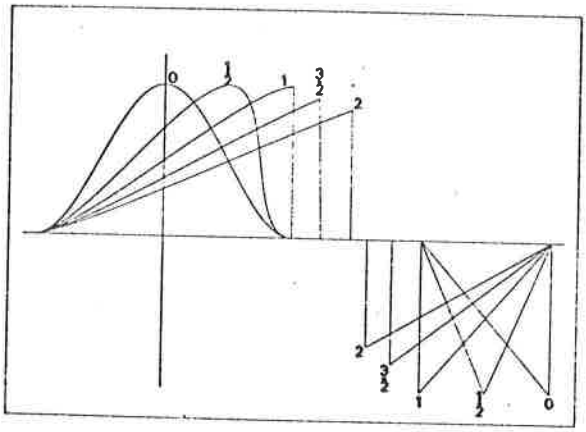
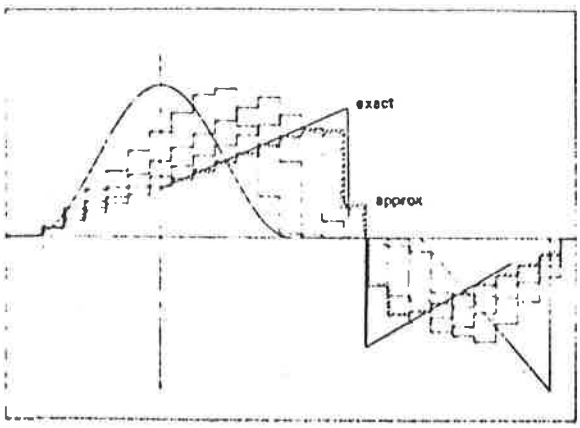
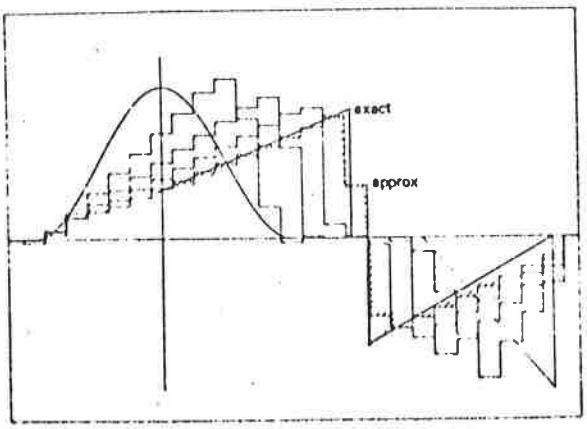


Fig. 4 Exact solution of model problem at $t = 0, \frac{1}{2}, 1, 1\frac{1}{2}, 2$.

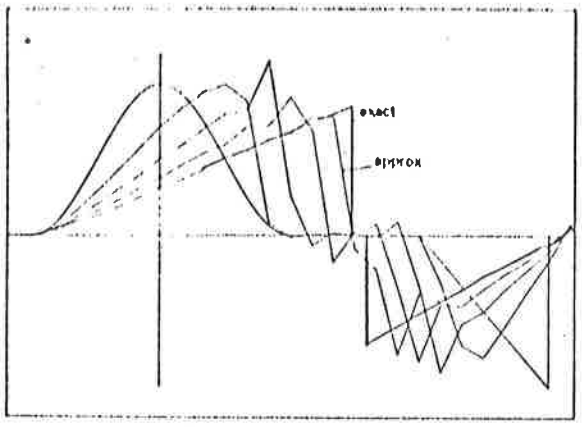


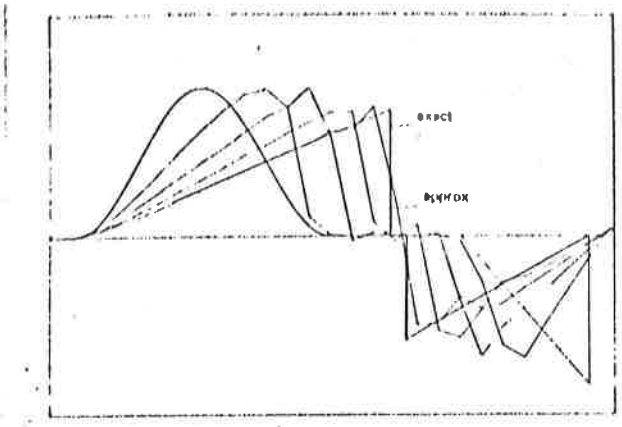
(a) without recovery



(b) with shock recovery

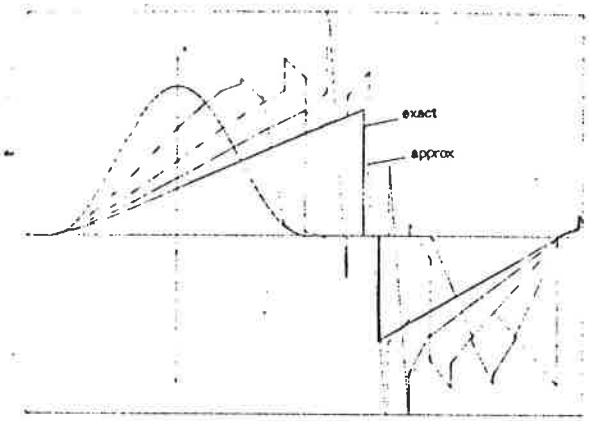
Fig. 5 Piecewise constant ECG approximation to model problem

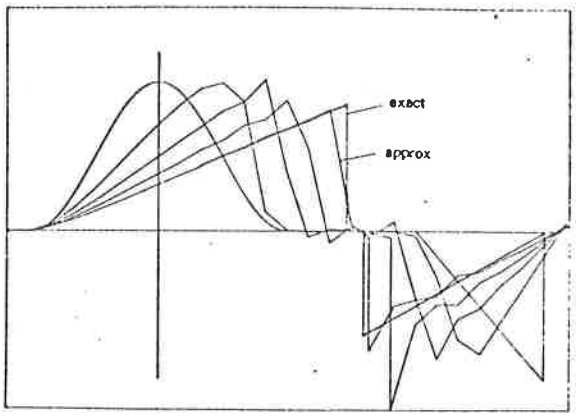




(a) without recovery (b) with shock recovery

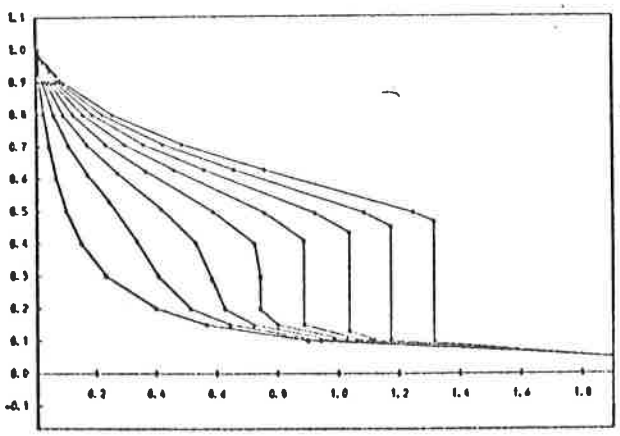
Fig. 6 Continuous piecewise linear ECG approximation to model problem.

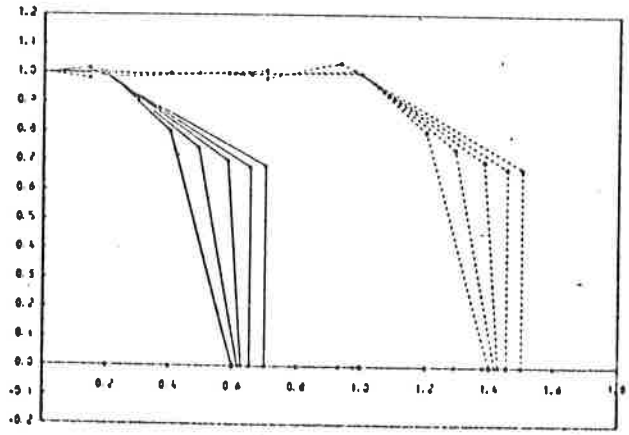




(a) without recovery (b) with shock recovery

Fig. 7 Discontinuous piecewise linear ECG approximation to model problem.





(a) two phase (b) three phase

Fig. 8 Moving finite element approximation to model oil recovery problems.