GENERALISED GALERKIN METHODS FOR STEADY AND UNSTEADY PROBLEMS

K. W. MORTON

NUMERICAL ANALYSIS REPORT 3/82

Talk presented at IMA Conference on Numerical Methods for Fluid Dynamics, Reading, March 1982

GENERALISED GALERKIN METHODS FOR STEADY AND UNSTEADY PROBLEMS

K. W. Morton
(Department of Mathematics, University of Reading)

1. Introduction

It is almost certainly true that the majority of practical fluid flow calculations are presently carried out using finite difference methods. In meteorology, aerodynamics, hydraulics, heat transfer and many other fields there is a large investment of experience, effort and expense in their use and they usually perform well enough. Finite element methods have as yet made a practical impact only in relatively few instances, see for example Hirsch & Warzée (1976), Kawahara (1978) and Jameson (1982). On the other hand, there is a very large literature covering their theory and their development for model problems. They have inherent advantages for equilibrium problems governed by quadratic extremal principles, that is, where the equations of motion are linear, elliptic and self-adjoint. But in fluid flow problems it is generally true that at least one of these properties is far from being satisfied.

In this paper we shall consider two particular developments of finite element methods to enable them to deal successfully with the wider classes of problems occurring in fluid flow. One is directed towards the solution of steady, linear diffusion-convection problems, which epitomise the effects of losing self-adjointness and whose successful solution is a necessary preliminary to tackling the Navier-Stokes equations at moderate to high Reynolds numbers. The other is concerned with evolutionary problems governed by hyperbolic equations. Both involve generalisations to the Galerkin formulation, from which the finite element method has drawn many of its advantageous properties.

Consider the following extremal problem for functions v defined in a region Ω,

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

the solution u satisfying the equation

$$T^*Tu = f, \tag{1.1b}$$

where T is a linear differential operator of order m, T^* is its adjoint, f is a given function and $\langle\cdot,\cdot\rangle$, $\|\cdot\|$ denote respectively the L^2 inner product and norm over Ω: in the minimisation v is to lie in $H^m(\Omega)$, the space of functions with square integrable m^{th} derivatives. Leaving aside the important but rather technical issues of how the boundary and boundary conditions are approximated - for which we refer the reader to standard texts such as Strang & Fix (1973) - suppose u is approximated from the conforming finite element space $S^h \subset H^m(\Omega)$, spanned by basis functions $\phi_j(x)$, that is,

$$S^h = \{V \in H^m(\Omega) \mid V(x) = \sum_{(j)} V_j\phi_j(x)\}. \tag{1.2}$$

Then carrying out the minimisation in (1.1a) over S^h gives the approximation U which satisfies the Galerkin equations

$$\langle TU, T\phi_i\rangle = \langle f,\phi_i\rangle \qquad \forall\, \phi_i \in S^h . \tag{1.3}$$

Since u also satisfies these equations we have

$$\langle T(u-U), T\phi_i\rangle = 0; \tag{1.4}$$

and from this it follows that

$$\|T(u-U)\| = \min_{V\in S^h} \|T(u-V)\| . \tag{1.5}$$

This equation expresses the crucial optimal approximation property of the Galerkin method: U is the best approximation to u from the trial space S^h in the energy norm determined by T. Both the theoretical error analysis and the practically important superconvergence properties follow from this equation.

Similarly, for the hyperbolic evolutionary problem

$$\frac{\partial u}{\partial t} + Lu = 0, \tag{1.6}$$

where L is a first order spatial operator, suppose u at each time t is approximated from S^h. Then the ordinary differential equations for the nodal parameters $U_j(t)$ may be determined by Galerkin equations

$$\langle \frac{\partial U}{\partial t} + LU, \phi_i \rangle = 0 \qquad \forall \phi_i \in S^h. \tag{1.7}$$

In many cases these again have important superconvergence properties: for example, with piecewise linear elements on a uniform mesh one obtains fourth order accuracy. Moreover, by multiplying each equation by U_i and summing, one obtains

$$\frac{d}{dt} \tfrac{1}{2} \|U\|^2 + \langle LU, U \rangle = 0. \tag{1.8}$$

Thus, just as for the exact solution u, the L^2 energy of the approximation is conserved or dissipated according to whether L is conservative or dissipative, i.e. $\langle Lv, v \rangle = $ or ≥ 0. By similar arguments one can show that the method has valuable non-linear stability properties.

In the next section we consider diffusion-convection problems and the Petrov-Galerkin methods which have been developed over recent years for their solution. We shall show that the widely used upwind schemes, including the streamline diffusion method, can be placed in a unified framework which provides a useful basis of comparison and sharp error bounds. This framework is based on symmetrizing the bilinear form associated with each problem, which can be done in two natural and distinct ways and yields approximations which are therefore optimal in alternative norms. Upwind schemes can generally be regarded as approximations to that symmetrization which leads to optimality in the Dirichlet norm, i.e. that arising from Poisson's equation: the alternative corresponds to that used by Barrett & Morton (1980) which leads to optimal approximations in a near least squares sense.

Petrov-Galerkin methods have also been developed by several authors for hyperbolic equations. However, in section 3 we present a generalisation of the Galerkin method which is based more directly on the use of characteristics.

With piecewise linear basis functions it is shown to be extremely accurate for smooth advection problems and to lead to several practically useful approximate schemes both in one and two space dimensions: these require little more computation than Galerkin methods and give large gains in stability and accuracy. With piecewise constant basis functions, the approach leads to methods for shock problems which are closely related to the difference methods of Engquist & Osher (1981).

2. Diffusion-convection problems

Consider the following problem in two or three dimensions:

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

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

Here a is a positive diffusion coefficient, $\underline{b}$ a convective velocity and f a given source. We shall assume that the convective medium is incompressible, $\underline{\nabla}\cdot\underline{b} = 0$, and that of the boundary $\Gamma_D \cup \Gamma_N$ of Ω, the Dirichlet part Γ_D includes all the inflow boundary; that is, if $\underline{n}$ is the outward normal then $\underline{b}\cdot\underline{n} \geq 0$ on Γ_N. A typical configuration is indicated on Fig. 1 and the non-dimensional Peclet number bL/a, where L is a characteristic length, may have values ranging from 10^2 for pollutant dispersal in a river to 10^4 and higher in heat transfer problems.

As is well-known, the Galerkin method in these situations often leads to wildly oscillatory solutions: in effect, central difference approximations to the convective term are generated and the discrete equations become almost singular. The remedy with difference methods has long been to use upwind differencing for this term and, in order to avoid the excessive damping associated with a wholly upwind scheme, to adopt a mixed strategy such as that advocated by Allen & Southwell (1955): in one dimension, for $-au'' + bu' = 0$ on a uniform mesh, this gives

$$-a\delta^2 U_j + bh[(1-\xi)\Delta_0 + \xi\Delta_-]U_j = 0 \qquad (2.2a)$$

$$\text{i.e.} \quad -(a + \tfrac{1}{2}\xi bh)\delta^2 U_j + bh\Delta_0 U_j = 0, \qquad (2.2b)$$

where h is the mesh spacing, $\Delta_0 U_j := \frac{1}{2}(U_{j+1} - U_{j-1})$ and $\delta^2 U_j := U_{j+1} - 2U_j + U_{j-1}$; with the exponentially fitted choice of mixing parameter,

$$\xi = \coth(\tfrac{1}{2}bh/a) - (\tfrac{1}{2}bh/a)^{-1}, \qquad (2.3)$$

this scheme gives exact nodal values for this simple model equation. Problems of accuracy still occur in practical situations through excessive crosswind diffusion and with variable flow fields and source terms.

Zienkiewicz (1975) seems to have been the first to recognise that various upwind schemes could be generated with finite elements if the weighting or test functions ϕ_i in (1.3) were modified. There is now a large literature on such Petrov-Galerkin methods and the reader is referred to the review by Heinrich & Zienkiewicz (1979) together with the other articles in the conference proceedings edited by Hughes (1979). Most methods aim to reduce to the Allen & Southwell scheme in the form (2.2a) by an appropriate choice of parameters: Hughes & Brooks (1979, 1981) on the other hand develop their streamline diffusion method from the form (2.2b) by adding an extra tensor diffusivity to the problem before using the Galerkin method, though this can also be regarded as a Petrov-Galerkin method.

A seemingly alternative approach was taken by Barrett & Morton (1980, 1981). Their aim was to choose a test space in such a way that the bilinear form associated with (2.1) was symmetrized and thus the optimal approximation property (1.5) restored to the method. However, as Morton (1981) has pointed out, both classes of method can be viewed from this objective of symmetrization, the difference being in the resulting symmetric form which is aimed at.

2.1 Alternative symmetrizations

Introducing the first order operators

$$T_1 v := a^{\frac{1}{2}} \underline{\nabla} v \ , \quad T_2 v := a^{\frac{1}{2}} \underline{\nabla} v - (\underline{b}/a^{\frac{1}{2}}) v, \tag{2.4}$$

equation (2.1a) may be written as

$$T_1^* T_2 u = f \quad \text{in } \Omega \ , \tag{2.5}$$

demonstrating the lack of self-adjointness. Similarly, by introducing the bilinear form

$$B(v,w) := \langle a\underline{\nabla} v, \ \underline{\nabla} w \rangle + \langle \underline{\nabla} \cdot (\underline{b} v), \ w \rangle \tag{2.6}$$

we obtain the weak form of (2.1) for $u \in H_E^1$ as

$$B(u,w) = \langle f, w \rangle \qquad \forall w \in H_0^1, \tag{2.7}$$

where $$H_E^1 := \{ v \in H^1(\Omega) \mid v = g \quad \text{on} \quad \Gamma_D \} \tag{2.8a}$$

and $$H_0^1 := \{ w \in H^1(\Omega) \mid w = 0 \quad \text{on} \quad \Gamma_D \}; \tag{2.8b}$$

then this may be written as

$$\langle T_2 u, \ T_1 w \rangle + \int_{\Gamma_N} \underline{b} \cdot \underline{n} u w \, dS = \langle f, w \rangle \qquad \forall w \in H_0^1. \tag{2.9}$$

There are two obvious symmetric forms related to this: one is the symmetric part of $B(\cdot,\cdot)$ and can be written in terms of T_1 as

$$B_1(v,w) := \langle T_1 v, T_1 w \rangle + \tfrac{1}{2} \int_{\Gamma_N} \underline{b} \cdot \underline{n} v w \, dS \tag{2.10a}$$

$$= \tfrac{1}{2}[B(v,w) + B(w,v)]; \tag{2.10b}$$

the other is that used by Barrett & Morton (1980) and based on T_2,

$$B_2(v,w) := \langle T_2 v, T_2 w \rangle + \int_{\Gamma_N} \underline{b} \cdot \underline{n} v w \, dS \tag{2.11a}$$

$$= \langle a\underline{\nabla} v, \ \underline{\nabla} w \rangle + \langle (|\underline{b}|^2/a) v, w \rangle. \tag{2.11b}$$

(Barrett & Morton actually introduce a weighting function ρ in their definition of B_S which we have taken as a^{-1}). The objective of approximately symmetrizing $B(\cdot,\cdot)$ can be sought through either form.

Since $B(v,w)$ is continuous on $H_0^1 \times H_0^1$ and an equivalent norm on H_0^1 may be defined from either $B_1(v,v)$ or $B_2(v,v)$, one may deduce from the Riesz representation theorem that operators R_1 and R_2 exist such that

$$\text{for } m = 1, 2 \qquad B(v,w) = B_m(v, R_m w) \qquad \forall v, w \in H_0^1 . \tag{2.12}$$

Leaving aside for the moment the problem of explicitly representing R_m, or its inverse, consider the Petrov-Galerkin method for U in a trial space S_E^h with basis functions ϕ_i, the subscript denoting that the essential boundary condition $U = g$ on Γ_D is satisfied, and based on test functions ψ_i spanning a test space $T_0^h \subset H_0^1$:

$$B(U,\psi_i) = \langle f,\psi_i \rangle \qquad \forall \psi_i \in T_0^h. \tag{2.13}$$

Since u satisfies the same equations we have the projection property for the error

$$B(u-U,\psi_i) = 0 \qquad \forall \psi_i \in T_0^h. \tag{2.14}$$

Now suppose we were able to choose the test functions so that, for $m = 1$ or 2,

$$\text{span } \{R_m \psi_i^*\} = \text{span } \{\phi_i \in H_0^1\} = S_0^h := S^h \cap H_0^1 . \tag{2.16}$$

Then denoting the corresponding Petrov-Galerkin approximation by U_m^* and noting that $u - U_m^* \in H_0^1$, we have from (2.12)

$$B_m(u-U_m^*,\phi_i) = 0 \qquad \forall \phi_i \in S_0^h \tag{2.17}$$

and hence the optimal approximation property holds,

$$\| u-U_m^* \|_{B_m} = \min_{V \in S_E^h} \| u-V \|_{B_m} . \tag{2.18}$$

In fact consider any test space T_0^h which has the same dimension as S_0^h, and for which the positive definiteness of $B(v,v)$ ensures the non-singularity of the stiffness matrix in (2.13): and suppose the closeness with which S_0^h can be approximated by $R_m T_0^h$ is described by the constant Δ such that

$$\min_{W \in T_0^h} \| V-R_m W \|_{B_m} \le \Delta \| V \|_{B_m} \qquad \forall\, V \in S_0^h. \tag{2.19}$$

Then one can shown for the corresponding Petrov-Galerkin approximation U,

$$\| u-U \|_{B_m} \leq (1-\Delta^2)^{-\frac{1}{2}} \| u-U_m^* \|_{B_m} . \tag{2.20}$$

In particular, one can deduce that the Galerkin approximation falls short of the optimal approximation by a factor dominated by the mesh Peclet number $|\underline{b}| h/a$.

2.2 Schemes derivable from $B_1(v,w)$ symmetrization

Now let us consider how closely we can in practice approximate either of the ideal test spaces given by (2.16). For $m = 1$, the relation (2.12) can be regarded as an equation for w with $R_1 w$ given, which takes the form

$$\langle a\underline{\nabla}(w-R_1 w) + \underline{b}w, \underline{\nabla}v \rangle - \tfrac{1}{2} \int_{\Gamma_N} \underline{b}\cdot\underline{n}(R_1 w)v\,ds = 0 \qquad \forall v \in H_0^1 . \tag{2.21}$$

In one dimension, on the unit interval with a and b positive constants and Dirichlet boundary conditions at $x = 1$ as well as $x = 0$, it becomes

$$aw' + bw = a(R_1 w)' + \text{const.}, \quad w(0) = w(1) = 0, \tag{2.22a}$$

the constant being determined by the two boundary conditions; and with a Neumann boundary condition at $x = 1$ it becomes

$$aw' + bw = a(R_1 w)' + \tfrac{1}{2}b(R_1 w)(1), \quad w(0) = 0. \tag{2.22b}$$

When the trial functions are piecewise linear, one obtains from these equations as the ideal test functions the negative exponential functions used successfully by Hemker (1977): on a uniform mesh, they are proportional to $\psi_j(x) = \psi(h^{-1}x - j)$ where

$$\psi(t) = \begin{cases} 1 - e^{-\beta(t+1)} & -1 \leq t \leq 0 \\ e^{-\beta t} - e^{-\beta} & 0 \leq t \leq 1, \end{cases} \tag{2.23}$$

and $\beta = bh/a$. These not only give the Allen & Southwell scheme but for $-au'' + bu' = f$ they give exact nodal values for any source function f. This is consistent with (2.18) for $m = 1$ because optimality in this norm for piecewise linear approximations corresponds to linear interpolation between

the nodes.

For large values of β the exponentials in (2.23) are awkward to deal with and in any case it is difficult to see how (2.21) could be solved in two dimensions. However, simpler functions can be used which reproduce the Allen & Southwell scheme and approximate (2.23) sufficiently well as regards modelling the effect of the source function. They include the quadratic functions used by Christie et al.(1976) and Heinrich et al. (1977) which, again on a uniform mesh, become

$$\psi(t) = \phi(t) + \alpha\sigma(t) \tag{2.24a}$$

with

$$\sigma(t) = \begin{cases} -3t(1-|t|) & |t| \le 1 \\ 0 & |t| > 1 \ : \end{cases} \tag{2.24b}$$

choosing the parameter $\alpha > 1 - 2/\beta$ ensures no oscillation in the solution, while taking $\alpha = \xi$, given by (2.3), gives the Allen & Southwell scheme. Moreover it is easy to extend this scheme into two dimensions using bilinear elements on rectangles. The trial basis functions are given by the product $\phi_{ij}(x,y) = \phi_i(x)\phi_j(y)$ and the test functions can be taken as

$$\psi_{ij}(x,y) = [\phi_i(x) + \alpha_1\sigma_i(x)]\,[\phi_j(y) + \alpha_2\sigma_j(y)] \tag{2.25}$$

with α_1, α_2 based on the two components of $\underline{b}$.

As noted above, the methods of Hughes & Brooks (1979, 1981) are prompted by the form (2.2b): the oscillations produced by the central differencing for the convective term $\underline{b}\cdot\underline{\nabla}u$ are damped by introducing extra diffusion, while still retaining the Galerkin test functions. In two dimensions the convective differencing is approximately in the streamline direction and thus they add the extra diffusion only in this direction: that is, the diffusion a in the term $-\underline{\nabla}\cdot(a\underline{\nabla}u)$ is replaced by a tensor diffusivity,

$$-\underline{\nabla}\cdot(\underline{\underline{A}}\underline{\nabla}u), \quad \text{where} \quad A_{\ell m} = a\delta_{\ell m} + \tilde{a}\, b_\ell\, b_m/|\underline{b}|^2. \tag{2.26}$$

On a uniform rectangular mesh with spacings h_1, h_2 the suggested choice of the parameter $\tilde{a}$ is

$$\tilde{a} = \tfrac{1}{2}(\xi_1 b_1 h_1 + \xi_2 b_2 h_2) \tag{2.27}$$

with $$\xi_m = \coth(\tfrac{1}{2}b_m h_m/a) - (\tfrac{1}{2}b_m h_m/a)^{-1}, \quad m = 1, 2 .$$

This scheme seems to work well in many practical situations. To place it in our general framework, we note that in their more recent paper Hughes & Brooks (1981) show that it is, in part at least, a Petrov-Galerkin method since

$$\langle \underline{\underline{A}}\nabla v, \underline{\nabla}\phi\rangle = \langle a\underline{\nabla}v, \underline{\nabla}\phi\rangle + \langle \underline{b}\cdot\underline{\nabla}v, (\tilde{a}/|b|^2)\underline{b}\cdot\underline{\nabla}\phi\rangle \ . \tag{2.28}$$

That is, it is equivalent to using test functions

$$\psi_{ij} = \phi_{ij} + (\tilde{a}/|\underline{b}|^2)\underline{b}\cdot\underline{\nabla}\phi_{ij}, \tag{2.29}$$

on just the convection term. For linear or bilinear trial functions these are discontinuous test functions and lie outside our theoretical framework: but if that part of $B(U,\psi_{ij})$ arising from the diffusion term and the $\underline{b}\cdot\underline{\nabla}\phi_{ij}$ term in (2.29) is evaluated element by element, it gives no contribution since $\nabla^2 U = 0$ on each element. In this way (2.29) can be regarded as a test function used in the whole of the bilinear form. Although Hughes & Brooks did not originally use the test function (2.29) on the source term, Johnson & Nävert (1981) in their analysis of the streamline diffusion method for the singular case $a = 0$ do in fact modify the source function in a way that is consistent with applying (2.29) and Hughes & Brooks (1981) now apply (2.29) consistently on source and time derivative terms.

Thus to summarise the $m = 1$ case, in one dimension the exact test functions (2.23) of Hemker, those advocated by Heinrich et al. (1977) and typified by (2.24) and those in effect used by Hughes & Brooks (1981) and given by (2.29) all reproduce the Allen & Southwell difference operator but differ in the way in which they sample the source function. These three test functions ψ_i for typical values of the mesh Peclet number β are plotted in Fig. 2a. The last two appear to differ significantly from the first:

yet from Fig. 2b, which shows in each case the extent to which $R_1\psi_i$ can reproduce the trial basis function ϕ_i, we see that either of them is very effective. These figures give a rough pictorial representation of the approximation properties of each scheme as defined by (2.19) and (2.20). Actual calculations of the parameter Δ in each case and in the Galerkin case give the results in Table I. Though there is little to choose between the two main practical methods in this simple case, they differ much more markedly, and of course much more importantly, in the way in which they extend to 2D.

Table I Ratios of Petrov-Galerkin error to optimal error, given by $(1-\Delta^2)^{-\frac{1}{2}}$ from (2.19), (2.20) with $m = 1$.

β	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

2.3 Symmetrization using $B_2(v,w)$

Turning now to the case $m = 2$ adopted by Barrett & Morton (1980, 1981), the equation corresponding to (2.12) and (2.21) becomes

$$\langle \underline{\nabla} w - [\underline{\nabla}(R_2 w) - a^{-1}\underline{b}(R_2 w)],\ a\underline{\nabla} v - \underline{b} v \rangle + \int_{\Gamma_N} \underline{b}\cdot\underline{n}(w - R_2 w)\, v\, dS = 0$$

$$\forall v \in H_0^1. \qquad (2.31)$$

In one dimension, on the unit interval with a and b positive constants and Dirichlet boundary conditions, this becomes

$$w' = (R_2 w)' - a^{-1} b (R_2 w) + \text{const.}\, e^{-bx/a}, \quad w(0) = w(1) = 0, \qquad (2.32)$$

the constant again being determined by the boundary conditions. Clearly the negative exponential plays a less important role in this case and it is

a straightforward matter to compute the ideal test functions defined by (2.16) for any choice of trial functions. It is, however, unnecessary to do this computation and it is preferable to move directly to the symmetrized form of the equations. Assuming for simplicity that the Dirichlet boundary conditions are homogeneous, the Petrov-Galerkin equations (2.13) with the test functions (2.16) can be written, using (2.12) as

$$B_2(U_2^*, \phi_i) = \langle f, \psi_i^* \rangle \qquad \forall\ \phi_i \in S_0^h. \tag{2.33}$$

Now suppose that instead of solving (2.16) we solve

$$R_2^*\ \tilde{f} = f \tag{2.34}$$

where R_2^* is the adjoint operator to R_2. Then (2.33) becomes

$$B_2(U_2^*, \phi_i) = \langle \tilde{f}, \phi_i \rangle \qquad \forall\ \phi_i \in S_0^h, \tag{2.35}$$

a symmetric system of Galerkin equations using a transformed source function. For the simple one-dimensional problem we obtain from the equation corresponding to (2.32)

$$f(\tilde{x}) = f(x) + a^{-1}b\,[F(x) - \bar{F}] \tag{2.36}$$

where

$$F(x) = \int_0^x f(y)\,dy, \quad \bar{F} = \int_0^1 e^{-bx/a} F dx \Big/ \int_0^1 e^{-bx/a} dx.$$

This is easily generalised to variable a, b and to general boundary conditions. The main two points to notice are that, firstly, the discrete operator on the left of (2.35) which one obtains with linear basis functions is no longer the Allen & Southwell operator but instead

$$-a\delta^2 U + bh\beta(1 + \tfrac{1}{6}\delta^2)U, \tag{2.37}$$

which corresponds to the self-adjoint differential form $-au'' + a^{-1}b^2u$; secondly, approximations to this ideal scheme are obtained, just as in the $m = 1$ case, by approximating the source function term - for instance, by omitting $\bar{F}$ in (2.36).

The most important distinction, however, between this $m = 2$ case and the $m = 1$ case is that U_2^* is a best fit in the norm given by (2.11b), which for large Peclet numbers becomes the L^2 norm, while U_1^* was exact at the nodes. This means that in a sharp boundary layer U_2^* exhibits oscillations, but of a controlled kind: indeed the extent of the overshoot is a valuable measure of the thickness of the boundary layer. The general problem of recovering information about u given its L^2 best fit has now been studied by many authors - see, for instance, Barrett, Moore & Morton (1982) and the references therein. The formulae correspond in some sense to the interpolation formulae that are needed to recover u from its nodal values: one of the best known gives accurate nodal values from the nodal parameters of a best linear fit on uniform mesh,

$$\left| u(jh) - \frac{1}{12}\,(U_{j-1} + 10U_j + U_{j+1}) \right| \le \frac{1}{360} h^4 \| u^{(iv)} \|_\infty \,. \qquad (2.38)$$

We shall see that L^2 best fits play an important rôle in the next section too. Thus although it is inappropriate to dwell at length on the recovery problem here, it is important to note that there are few disadvantages and often some advantage (as in the boundary layer case) in a method yielding an L^2 best fit rather than nodal values.

To continue the discussion of the $m = 2$ case, there are several ways in which two dimensional problems may be treated. In Morton & Barrett (1980) tensor products of the one-dimensional ideal test functions were used, as with the method of Heinrich & Zienkiewicz (1979). These are both somewhat awkward to use and less successful with strongly curved streamlines than the mixed method used in Barrett & Morton (1981) and forming the natural extension of the form (2.35). Without attempting to solve the equation for f, we introduce the flux function

$$\underline{v} = \underline{b}u - a\underline{\nabla}u \qquad (2.39)$$

and can then write the equation (2.7) for u as

$$B_2(u,w) = \langle f,w\rangle + \langle a^{-1}\underline{b}\cdot\underline{v},w\rangle \qquad \forall\, w \in H_0^1. \qquad (2.40)$$

This is approximated by

$$B_2(U,\phi_i) = \langle f,\phi_i\rangle + \langle a^{-1}\underline{b}\cdot\underline{V},\phi_i\rangle \qquad \forall\ \phi_i \in S_0^h. \tag{2.41}$$

with $\underline{V}$ obtained by approximating the equation $\underline{\nabla}\cdot\underline{v} = f$. Then one finds that, if S^* is the best L^2 fit to $a^{-1}\underline{b}\cdot\underline{v}$ from S_0^h, one has

$$\|U-U_2^*\|_{B_2} \leq \| |\underline{b}|^{-1}(a^{-1}\underline{b}\cdot\underline{V} - S^*)\| . \tag{2.42}$$

Although it is far from clear that the best procedures for approximating $\underline{V}$, or rather $\underline{b}\cdot\underline{V}$, have yet been found the results obtained so far are encouraging.

2.4 A test problem

We end this section with a few numerical results for a test problem which is a modification of one put forward by Hutton (1981). The flow field $\underline{b}$ is indicated in Fig. 3 and is derived from a stream function $(1-x^2)(1-y^2)$. In Hutton's test problem a tanh input profile for u was specified on $y = 0$, $-1 \leq x \leq 0$ with Dirichlet conditions on the tangential boundary consistent with pure convection: the main test was for the output profile for various values of the Peclet number. We have tested the Heinrich et al. scheme using (2.25), the Hughes & Brooks scheme using (2.29) and the Barrett & Morton (1981) scheme and all performed reasonably well on this problem with the Hughes & Brooks scheme giving the best results, presumably because of its small crosswind diffusion: a two dimensional version of the Allen & Southwell scheme by contrast gave very poor results. Our modification to the problem is to specify $u = 0$ on the input and all the tangential boundaries except $x = 1$, where we put $u = 100$: this models a situation where a cold fluid is channelled past a hot plate.

The interesting profiles are those for fixed values of y. In Fig. 4 we show the profiles for each scheme at $y = 0.9$, $y = 0.5$ and $y = 0$, when the mesh Peclet number $\beta = 20$; Fig. 5 shows corresponding results for $\beta = 100$. Despite their objective of non-oscillatory solutions both the Heinrich et al.

scheme and that of Hughes & Brooks show considerable oscillation, particularly the latter. The Barrett & Morton scheme, on the other hand, is aimed at a best fit in the norm defined by (2.11b) and would be expected to be oscillatory for these values of β. Indeed the variation of the boundary layer thickness with y, which is most obvious in Fig. 4, can be calculated from these results and each thickness is within a few per cent of that calculated from an asymptotic analysis.

More analysis is clearly required to fully explain the behaviour in this two dimensional example of the two schemes which were motivated by the Allen & Southwell difference method and which we have associated with the B_1 symmetrization. However, when the results are combined with those for the Barrett & Morton scheme and viewed in the context of the general analysis given above, they add to the growing evidence that generalised Galerkin methods can successfully handle a wide class of diffusion-convection problems: the important point is that their output must not be viewed as if it came from just another fancy difference scheme.

3. Hyperbolic equations

Any method for approximating hyperbolic equations sacrifices a good deal if it takes no account of the presence of characteristics. The semi-discrete Galerkin equations (1.7) yield such methods: thus as soon as a standard time discretisation is introduced, disadvantages to the Galerkin formulation appear. First of all, a reduced stability range for explicit schemes is generally obtained. For example that for the leapfrog method is reduced by a factor $\sqrt{3}$: while Euler's method applied to $u_t + au_x = 0$ gives the central difference scheme which is well-known to be stable only for $\Delta t = O(h^2)$. Indeed the phenomenon in this latter case is very similar

to that in the previous section and some form of upwinding is strongly indicated. One of the consequences of this loss of stability is that the schemes cannot be used when the CFL number is unity, while most common difference schemes are exact in this limit, a fact which improves their accuracy over the whole stability range.

Many authors have sought means to remedy these defects and several of them are based on a Petrov- Galerkin approach. Thus suppose a one-step method in time is used and $u(x,t)$ is approximated at $n\Delta t$ by

$$U^n(x) = \sum_{(j)} U_j^n \phi_j(x). \tag{3.1}$$

Then test functions $\psi_i(x)$ are sought for the equations

$$< \frac{U^{n+1} - U^n}{t} + L(\theta U^{n+1} + \overline{1-\theta} U^n), \psi_i > = 0. \tag{3.2}$$

Morton & Parrott (1980) introduced special test functions $\chi_i(x)$, corresponding to the use of linear trial functions ϕ_j for the model scalar problem $u_t + au_x = 0$, which have the property of giving exact results when the CFL number $\mu = a\Delta t/h$ is unity (the unit CFL property): because the Galerkin method is highly accurate for small μ they therefore used in the general case

$$\psi_i = (1 - \nu)\phi_i + \nu\chi_i \tag{3.3}$$

with $\nu = \mu$ or $\nu = \mu^2$, determined by a Fourier analysis. Highly accurate schemes which are closely related to well-known finite difference schemes result for either Euler's method, $\theta = 0$, or Crank-Nicolson, $\theta = \frac{1}{2}$. A similar scheme was given for leap-frog time differencing and both this and the Crank-Nicolson scheme retain conservation of U though not of U^2: the leap-frog scheme is 4th order accurate in both Δt and h. Convenient generalisations were given for hyperbolic systems and the use of bilinear elements allowed Morton & Stokes (1981) to extend the methods to two dimensions. However, limitations were found with the Petrov-Galerkin formulation when triangular elements were used and an approach even more closely based on the characteristics was introduced.

3.1 Euler characteristic Galerkin method (ECG Method) in one dimension

Consider the scalar conservation law in one dimension

$$\partial_t u + \partial_x f(u) = 0, \tag{3.4a}$$

$$\text{or} \qquad \partial_t u + a(u)\partial_x u = 0, \tag{3.4b}$$

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 similarly for a and f, we have for smooth flows

$$u^{n+1}(y) = u^n(x) \quad \text{where} \quad y = x + a^n(x)\Delta t. \tag{3.5}$$

Thus the L^2 projection of u^{n+1} onto the trial space S^h spanned by $\{\phi_j\}$ is related to that of u^n by

$$\begin{aligned} \langle u^{n+1} - u^n, \phi_j \rangle &= \int_{-\infty}^{\infty} u^n(x) \left[\phi_j(y) \frac{dy}{dx} - \phi_j(x)\right] dx \\ &= \int_{-\infty}^{\infty} u^n(x) \left[\frac{d}{dx} \int_x^y \phi_j(z)dz\right] dx \\ &= -\int_{-\infty}^{\infty} \partial_x u^n(x) \left[\int_x^y \phi_j(z)dz\right] dx. \end{aligned} \tag{3.6}$$

On a uniform mesh with $\phi_j(x) = \phi(h^{-1} x - j)$, we introduce the upwind-averaged test function

$$\Phi(s,\mu) = \frac{1}{\mu} \int_s^{s+\mu} \phi(\sigma)d\sigma \tag{3.7}$$

and set

$$\begin{aligned} \Phi_j^n(x) &= \Phi(h^{-1} x - j, h^{-1} a^n(x)\Delta t) \\ &= \frac{1}{a^n(x)\Delta t} \int_x^{x+a^n(x)\Delta t} \phi_j(z)dz. \end{aligned} \tag{3.8}$$

This test function $\Phi(s,\mu)$ is plotted for various values of μ in Fig. 6, where it is seen to have its maximum at $-\frac{1}{2}\mu$: this corresponds to Φ_j^n peaking midway between jh and the foot of the characteristic drawn back from $(jh, \overline{n+1}\Delta t)$ to time level n. From (3.6) and (3.8) we then get

$$\langle u^{n+1} - u^n, \phi_j \rangle + \Delta t \langle a^n \partial_x u^n, \Phi_j^n \rangle = 0. \tag{3.9}$$

This exact relationship does not of course allow complete tracking of the evolution of u^n, as does (3.5), since only the projection quantities $\langle u^n, \phi_j \rangle$ are obtained at each level and these are insufficient to calculate the second term of (3.9). But several approximation schemes can be based on this relation.

We refer to the following as the (exact) Euler Characteristic Galerkin (ECG) Method.

$$\langle U^{n+1} - U^n, \phi_j \rangle + \Delta t \langle a(U^n) \partial_x U^n, \Phi_j^n \rangle = 0, \tag{3.10}$$

where U^n is given by (3.1) and is assumed continuous and Φ_j^n is given by (3.8) with $a^n(x)$ taken as $a(U^n)$. We leave the second term in the form $a(U)\partial_x U$ because $a(U)$ will cancel: but the evaluation of this inner product still involves considerable computation and various approximate schemes will be considered below. The merit of (3.10), however, is that the only error involved is that due to the projection at each time step: if the initial data is projected into S^h then this is carried forward exactly through the first time step before being projected again, and so on. Thus if the objective is to carry forward the L^2 projection of $u(t)$ onto S^h this is the very best that can be achieved by a one-step algorithm using the time step Δt.

3.2 Approximate ECG schemes

We confine ourselves here to piecewise linear basis functions ϕ_j and to schemes which are exact when $a(U)$ is constant and the CFL number $\mu = a\Delta t/h$ lies in $(0,1)$. In this case, (3.10) involves only three neighbouring nodes and their coefficients can be correctly reproduced by either of the following replacements for Φ_j:

$$\begin{aligned}\Phi_j \approx \Phi_j^T := (1-\tfrac{1}{2}\mu)\phi_j + \tfrac{1}{2}\mu\phi_{j-1} + \tfrac{1}{2}\mu(\tfrac{1}{2}-\tfrac{1}{3}-\mu)(\phi_j'-\phi_{j-1}') \\ + M[(\phi_j-\phi_{j-1}) + \tfrac{1}{2}(\phi_j'-\phi_{j-1}')];\end{aligned} \tag{3.11}$$

$$\Phi_j \approx \Phi_j^S(x) := \frac{1}{6}[\phi_j(x) + 4\phi_j(x + \tfrac{1}{2}\mu h) + \phi_j(x + \mu h)]. \tag{3.12}$$

In the first alternative, where the superscript T denotes that the integral (3.7) has been approximated by considering Taylor series expansions, M may be any smooth function of μ which tends to zero at either limit $\mu \to 0$ or 1: one possibility is $M = 0$ and another, $M = \frac{1}{2}\mu(1-\mu)^2$, makes Φ_j^T the best L^2 fit to Φ_j by a linear fit in each interval; the latter is shown in Fig. 7. For Φ^T only one extra set of inner products needs to be evaluated as compared with the Galerkin method, $\langle\phi_i', \phi_j'\rangle$ as well as $\langle\phi_i', \phi_j\rangle$. In the second alternative, given by (3.12) where the superscript S stands for "shifted" or for Schoombie (1982) who introduced such test functions in a Petrov-Galerkin setting, two extra sets of inner products need to be evaluated: moreover, these also depend on the value of μ so that in more general cases they cannot be evaluated once and for all.

There is a third possibility besides (3.11) and (3.12) which approximates directly the inner product $\langle\phi_i', \Phi_j\rangle$ needed in the linear case of (3.10):

$$\langle\phi_i', \Phi_j\rangle \approx (1-\mu)^2 \langle\phi_i, \phi_j\rangle - \mu(1-\mu)\langle\phi_i', \phi_{j-1}\rangle + \mu(3-2\mu)\langle\phi_i, \phi_j - \phi_{j-1}\rangle. \tag{3.13a}$$

Here there are no extra inner products needed at all. Moreover, integrating by parts we see that this scheme is equivalent to using a test function

$$\Phi_j \approx \Phi_j^I(x) := (1-\mu)^2\phi_j - \mu(1-\mu)\phi_{j-1} + \mu(3-2\mu)\int_{-\infty}^{x}(\phi_{j-1}-\phi_j)dy, \tag{3.13b}$$

which is shown in Fig. 8 where we see that it is an exact match at $\mu = 1$ and extremely accurate at $\mu = \frac{3}{4}$.

There is clearly no stability limit on (3.10) while there will be such limits on the approximate schemes given above. For (3.11) this is independent of M and is actually $-\frac{1}{2} \le \mu \le \frac{3}{2}$ although for reasons of accuracy one would wish to keep μ in $(0,1)$.

The application of either (3.11), (3.12) or (3.13) to a general scalar conservation law is straightforward: one merely replaces μ in the formulae by the local CFL number $a(U_j^n)\Delta t/h$. However, because $a(U^n)$ does not cancel

as it does in (3.10), either some of the speed advantage has been lost through having to integrate inner products which contain $a(U^n)$ explicitly, or some accuracy is lost by replacing $a(U^n)$ by the constant $a(U_j^n)$ and absorbing it into the coefficients in (3.11), (3.12) and (3.13). An alternative is to use the product approximation strongly advocated by Christie et al. (1981): that is, to use the approximation

$$\partial_x f(u) \approx \sum_{(j)} f(U_j^n)\phi_j' \,. \tag{3.14}$$

Such an approximation is of particular advantage when coupled to (3.13a), the amount of computation then differing little from that for the Galerkin method applied to the simple advection equation $u_t + au_x = 0$.

3.3 Test results in one dimension

We will show a sample of results obtained with these schemes: a more complete account will be found in Morton & Stokes (1982). The first tests are for pure advection. Fig. 9 shows the results for a Gaussian profile and for a ramp function as compared with Gadd's modification of the Lax-Wendroff method (Gadd, 1978). A Fourier analysis shows that the error generated in the L^2 projection at each time step is given by

$$\text{error} \sim \frac{1}{24}\mu^2(1-\mu^2)\xi^4 \tag{3.15}$$

where $\xi = kh$ and k is the wave number.

Fig. 10 gives results obtained for the non-linear advection equation $u_t + uu_x = 0$ with an initial isolated cosine wave. Each picture shows the leading edge of the wave at $t = \frac{1}{2}$ where $t = 2/\pi$ is the time to first breaking: on the left are results for Crank-Nicolson-Galerkin (the most reliable second order accurate time-stepping for Galerkin), the exact ECG scheme (3.10) and the approximate ECG scheme (3.11); on the right the same schemes are used but coupled with the product approximation (3.14).

Taken together these two sets of results demonstrate impressive accuracy for the ECG methods and this is confirmed by results obtained by other authors experimenting with similar schemes - Bercovier & Pironneau (1981) and Benqué & Ronat (1982). Note that all the results presented for the finite element schemes should be interpreted as approximations to best L^2 fits. The particular interpretation that is used is unimportant for linear constant coefficient problems but is very important for non-linear problems - see Cullen & Morton (1980). Finally we should point out that similar test functions can be developed for other basis functions and for other time-differencing schemes and these will be found in Morton & Stokes (1982). In particular, though one has to generalise the development of the algorithm to deal with discontinuous basis functions, it can then lead to methods related to the upwind schemes presently used in shock modelling: thus piecewise constant elements plus an exact Riemann solver to replace (3.5) gives the method of Godunov (1959) while simpler extensions to (3.5) can lead to the basic method of Engquist & Osher (1981).

3.4 ECG schemes in two dimensions

For the linear advection equation

$$\partial_t u + \underline{a}\cdot\underline{\nabla} u = 0 \tag{3.16}$$

the exact Euler Characteristic Galerkin Method uses an upwind-averaged test function completely analgous to (3.7) and (3.8). However, for piecewise linear elements over triangles, the computation of the test function, or of its inner products with $\underline{a}\cdot\underline{\nabla}U$, is clearly a considerable task. Fortunately either (3.11) or (3.13) extend in a natural and economic manner to give very accurate results. The vector $-\underline{a}\Delta t$ extends from the node j into a triangle for which this is one vertex and defines the foot of the characteristic drawn back from node j at time level $n+1$ to time level n: the approximation to Φ_j uses the basis functions, and with (3.11) their gradients, corresponding to all the vertices of this triangle - see Fig. 11. Using the notation of this

figure, the generalisation of (3.11) is most simply given in the local co-ordinate system based on node j in which the triangle becomes a canonical right triangle as shown in Fig. 11 and $\underline{a}\Delta t$ becomes (μ_1,μ_2). No choice of coefficients gives an exact match to the perfect test function but there is a two parameter family of methods corresponding to (3.11) which give third order accuracy: the simplest, corresponding to $M = 0$ in (3.11), takes the form

$$\begin{aligned}\Phi_A \approx \Phi_A^T := {} & (1-\tfrac{1}{2}\mu_1-\tfrac{1}{2}\mu_2)\phi_A + \tfrac{1}{2}\mu_1\phi_B + \tfrac{1}{2}\mu_2\phi_C \\ & + \tfrac{1}{2}\mu_1[(\tfrac{1}{2}-\tfrac{1}{3}\mu_1)\partial_\xi(\phi_A-\phi_B) + \tfrac{1}{3}\mu_2\partial_\xi\phi_C] \\ & + \tfrac{1}{2}\mu_2[(\tfrac{1}{2}-\tfrac{1}{3}\mu_2)\partial_\eta(\phi_A-\phi_C) + \tfrac{1}{3}\mu_1\partial_\eta\phi_B]. \end{aligned} \qquad (3.17)$$

Stability also depends on the choice of parameters and the stability region of (3.17) has been shown to include $\mu_1^2 + \mu_2^2 \le 1$, and therefore all cases where $-\underline{a}\Delta t$ lies in a triangle with node j as a vertex.

Numerical tests on this and related schemes have so far been carried out with advection of a Gaussian, both on straight line tracks and around a circle, with excellent accuracy. Fig. 12 shows radial cross sections for a Gaussian carried round a circular trajectory after travelling a quarter, a half, three-quarters and a complete revolution. The problem is solved on $(-1,1)\times(-1,1)$ with right-triangular elements, $\Delta x = \Delta y = \frac{1}{16}$, $\Delta t = 0.8\Delta x$, circle radius $\frac{1}{2}$ and standard deviation $\frac{1}{4}\sqrt{2}$: after 160 time-steps the phase error is about one mesh length.

The present objective is to apply the method to the shallow water equations. The test function (3.17) is applied to just the advection terms and the Galerkin formulation used for the remainder: it may be desirable however to use the alternative time-stepping schemes referred to briefly at the end of the previous section. The immediate aim is to improve the stability range and accuracy reported for model problems in Cullen & Morton (1980) and obtained with schemes which generalised the purely Galerkin approach only in regard to modelling the non-linear terms by a two-stage Galerkin

procedure.

4. Conclusions

Diffusion-convection problems and hyperbolic equations provide two related but distinct problem areas where the deficiencies of the Galerkin derivation of finite element methods are most apparent. We have shown that generalised Galerkin methods can be formulated that can, in the one case completely and in the other very largely, restore the optimal approximation properties that make the Galerkin approach so successful with self-adjoint equilibrium problems. These ideal methods are not completely practical but we have also shown how a number of existing methods can be viewed as approximations to them and also how they provide guidelines to the development of new practical algorithms.

A few such algorithms have been presented along with results for model problems. But, as indicated in the introduction, much work has yet to be done to develop them into competitive techniques for typical practical fluid flow problems.

Acknowledgements

I am indebted to Bryan Scotney for the computations in Section 2 and to Alan Stokes for those in Section 3.

References

Allen, D., & Southwell, R., 1955. Relaxation methods applied to determining the motion, in two dimensions, of a viscous fluid past a fixed cylinder. Quart. J. Mech. and Appl. Math. VIII, 129-145.

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., Optimal finite element approximation for diffusion convection problems. To appear in Proc. MAFELAP 1981 Conf. (ed. J.R. Whiteman).

Barrett, J.W., Moore, G. & Morton, K.W., 1982. Optimal recovery and defect correction in the finite element method. In preparation.

Benqué, J.P., & Ronat, J., 1982. Quelques difficultes des modeles numerique en hydraulique. Presented at 5th Int. Symp. Computing Methods in Applied Sciences and Engng., Versailles 1981. To appear.

Bercovier, M., & Pironneau. Characteristics and finite element methods applied to the equation of fluids. To appear in Proc. MAFELAP 1981 Conf. (ed. J.R. Whiteman).

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. Numer. 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.

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

Gadd, A.J., 1978. A numerical advection scheme with small phase speed errors. Quart. J. Roy. Met. Soc. 104, 583-594.

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.

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. Engng. (New York), 105-136.

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

Hirsch, Ch., & Warzee, G., 1976. A finite element method for through-flow calculations in turbomachines. J. Fluids Engng., Trans. ASME 98, 403-421.

Hughes, T. (ed.) 1979. Finite Element Methods for Convection Dominated Flows AMD Vol. 34, Am. Soc. of Mech. Engng. (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. Engng. (New York), 19-35.

Hughes, T.J.R., & Brooks, A., 1981. A theoretical framework for Petrov-Galerkin methods with discontinuous weighting functions: application to the streamline-upwind procedure. To appear in Finite Elements in Fluids Vol. 4 (ed. R.H. Gallagher) J. Wiley & Sons (New York).

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

Jameson, A., 1982. Transonic aerofoil calculations using the Euler equations. Numerical Methods in Aeronautical Fluid Dynamics (ed. P.L. Roe), Academic Press.

Johnson, C., & Nävert, U., 1981. An analysis of some finite element methods for advection-diffusion problems. Conf. on Analytical and Numerical Approaches to Asymptotic Problems in Analysis (eds. O. Axelsson, L.S. Frank & A. van der Sluis) North-Holland.

Kawahara, M., 1978. Steady and unsteady finite element analysis of incompressible viscous fluid. Finite Elements in Fluids 3 (ed. R.H. Gallagher et al.), Wiley & Sons (New York), 23-54.

Morton, K.W., 1981. Finite element methods for non-self-adjoint problems. Univ. of Reading, Num. Anal. Rep. 3/81.

Morton, K.W., & Barrett, J.W., 1980. Optimal finite element methods for diffusion-convection problems. Proc. Conf. Boundary and Interior Layers - Computational and Asymptotic Methods (ed. J.J.H. Miller) Boole Press, Dublin, 134-148.

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., Generalised Galerkin methods for hyperbolic equations. To appear in Proc. MAFELAP 1981 Conf. (ed. J.R. Whiteman).

Morton, K.W. & Stokes, A., 1982. Characteristic Galerkin methods for hyperbolic equations. In preparation.

Schoombie, S.W., 1982. Spline Petrov-Galerkin methods for the numerical solution of The Kortweg-de Vries equation. IMA J. Numer. Anal. 2, 95-109.

Zienkiewicz, O.C., Gallagher, R.H., & Hood, P., 1975. Newtonian and non-Newtonian viscous incompressible flow, temperature induced flows: finite element solutions. 2nd Conf. Mathematics of Finite Elements and Applications (Ed. J.R. Whiteman), Academic Press (London).

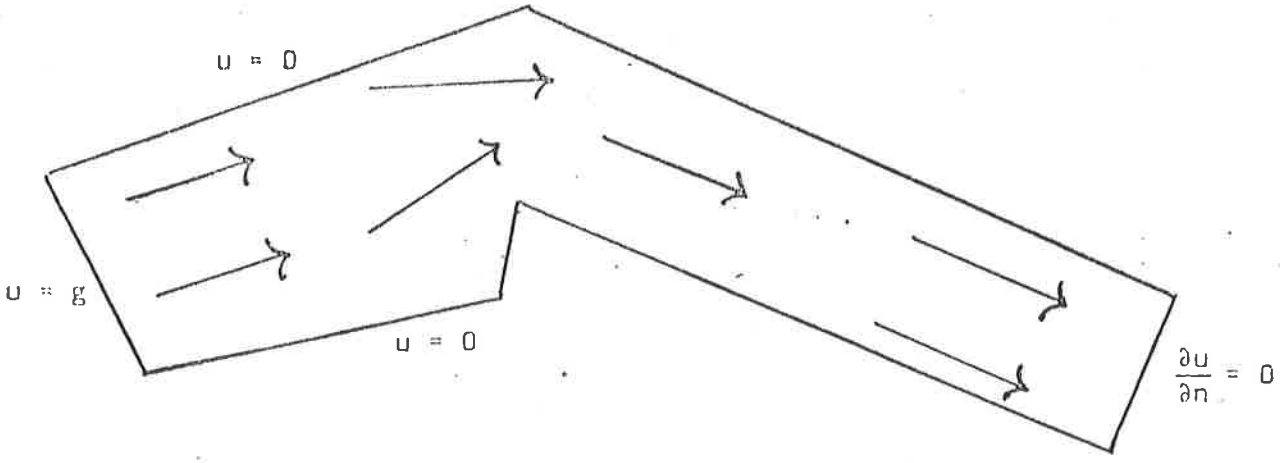
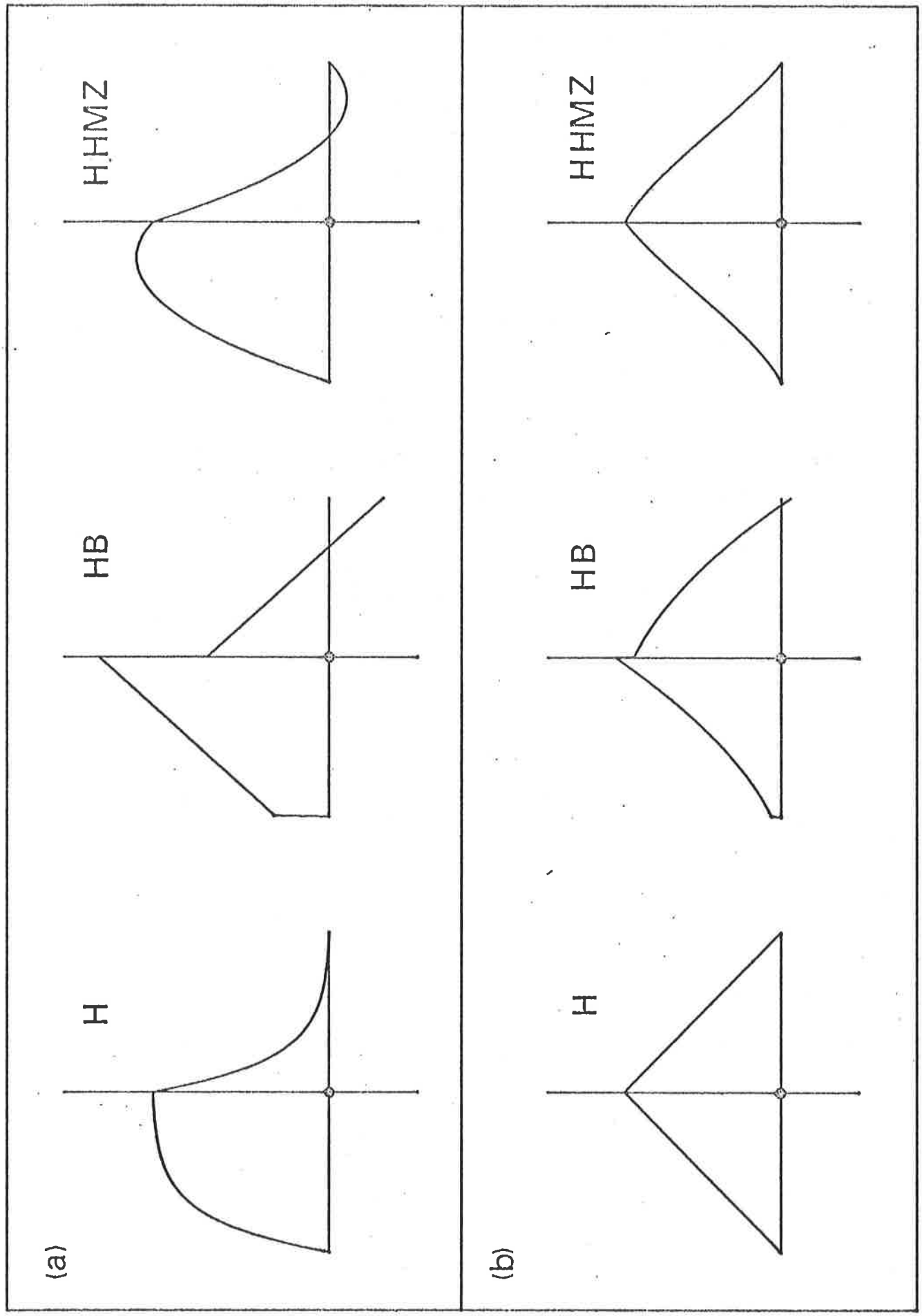


Figure 1 : A typical diffusion-convection problem in 2D.

Figure 2(a) : Test functions ψ_i used by Hemker (H), Hughes & Brooks (HB) and Heinrich et al. (HHMZ); and (b) corresponding approximations to the trial function ϕ_i constructed from $R_1\psi_i$. The mesh Peclet number $\beta = 5$



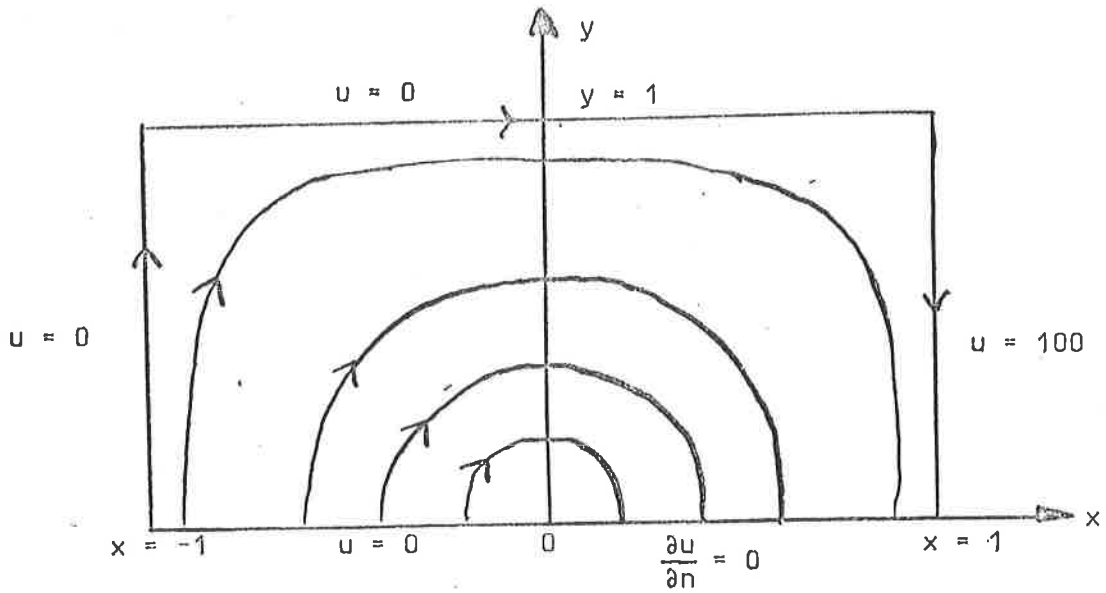


Figure 3 : A test problem modified from Hutton (1981).

Figure 4 : Results for the testproblem of Fig. 3, showing the boundary layer near $x = 1$ for various values of y. The mesh Peclet number $\beta = 20$ and the methods used are Heinrich et al. (HHMZ), Hughes & Brooks (HB) and Barrett & Morton (BM).

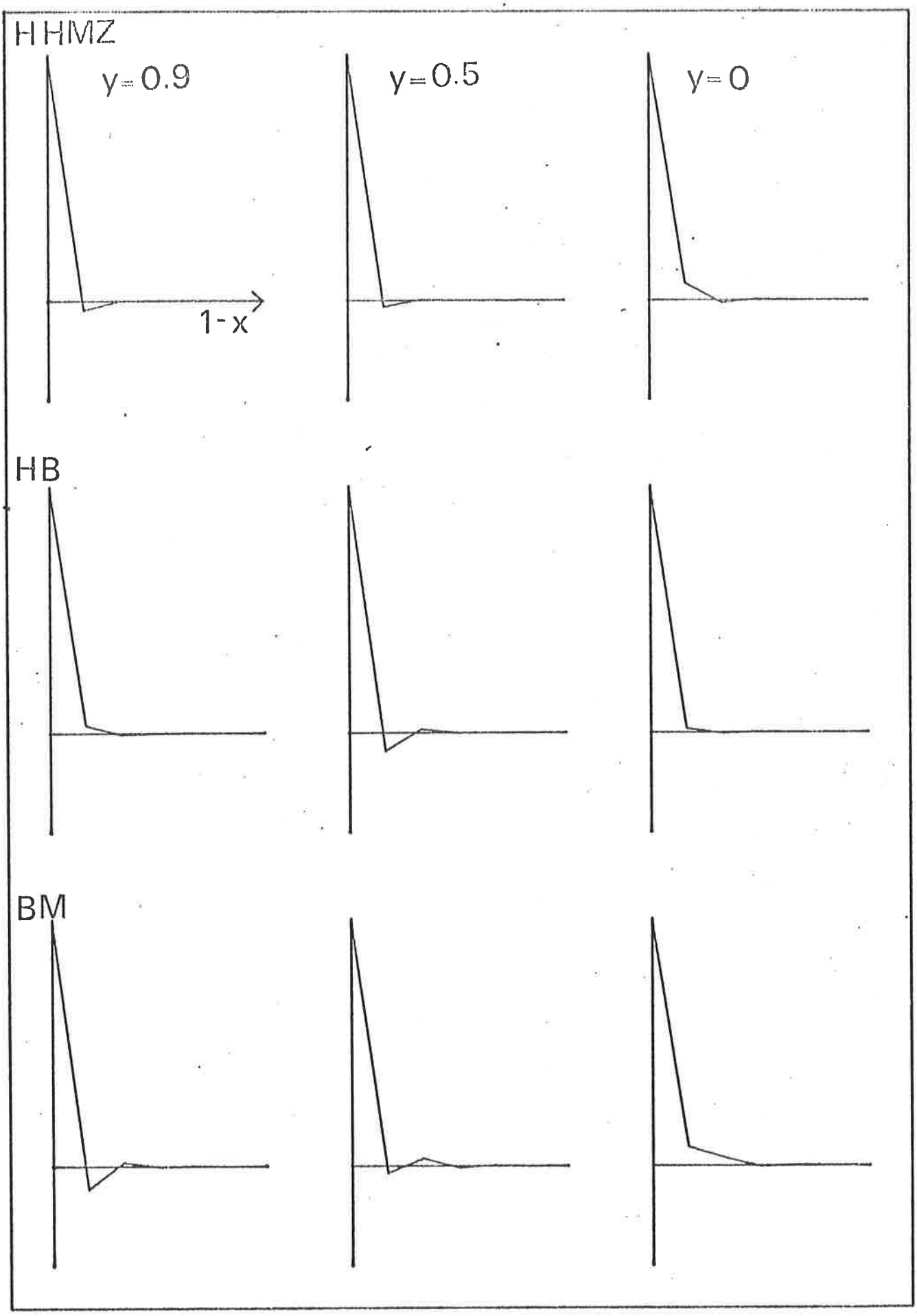
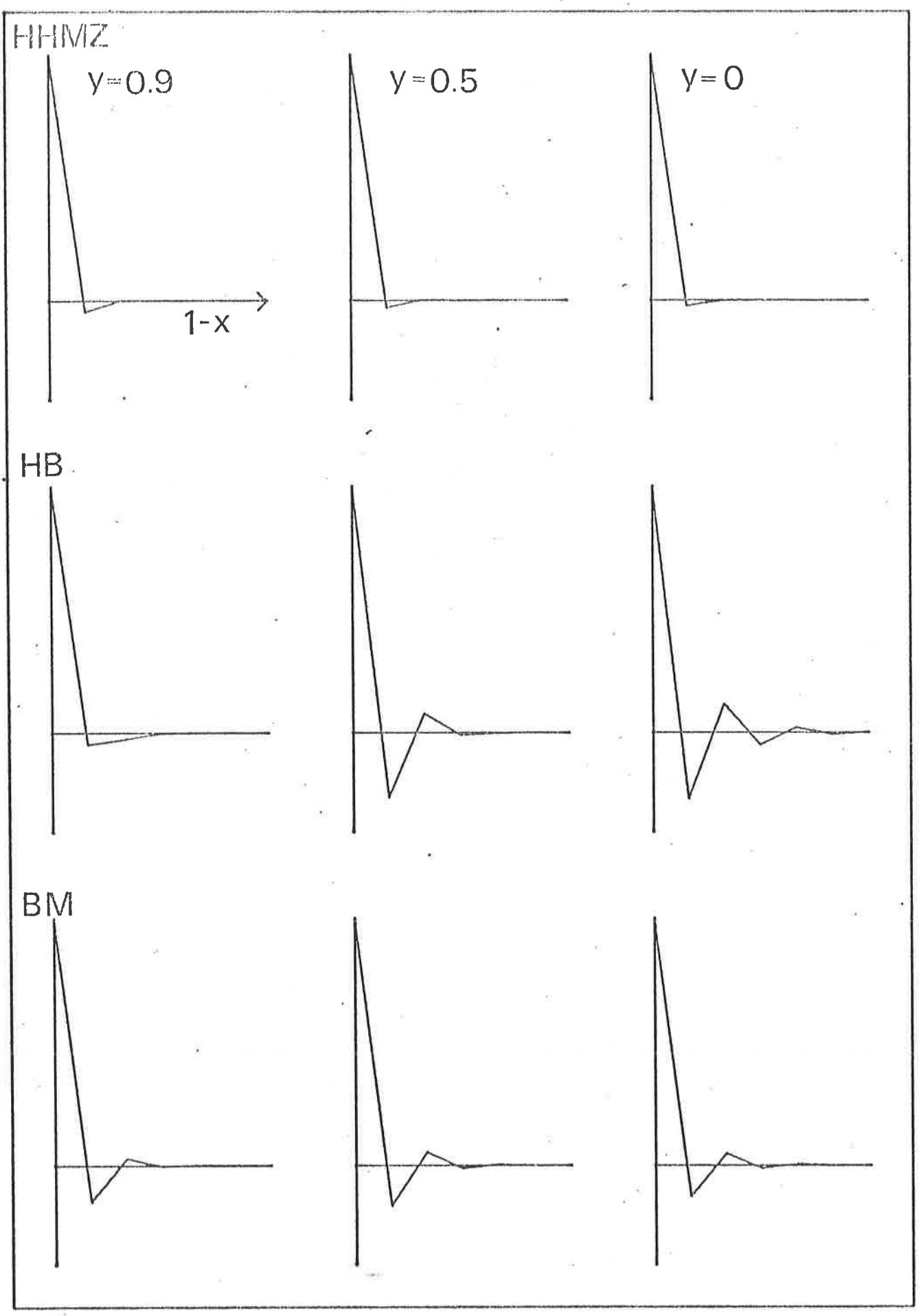


Figure 5 : Similar to Fig. 4 for $\beta = 100$



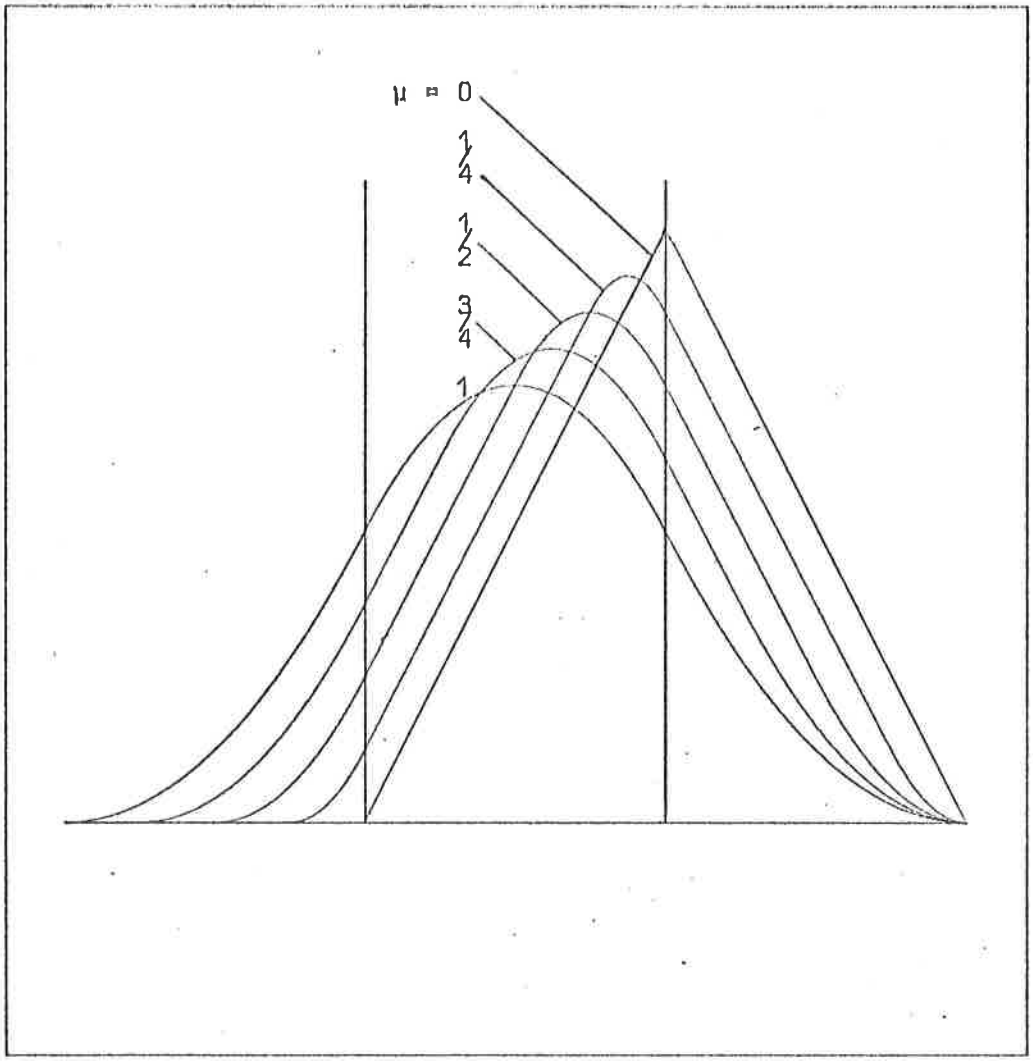


Figure 6 : Upwind-averaged test function $\Phi(s,\mu)$ for $\mu = 0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$

Figure 7 : The approximate test function $\Phi^T(s,\mu)$, with $M = \frac{1}{2}\mu(1-\mu)^2$, compared with the exact $\Phi(s,\mu)$.

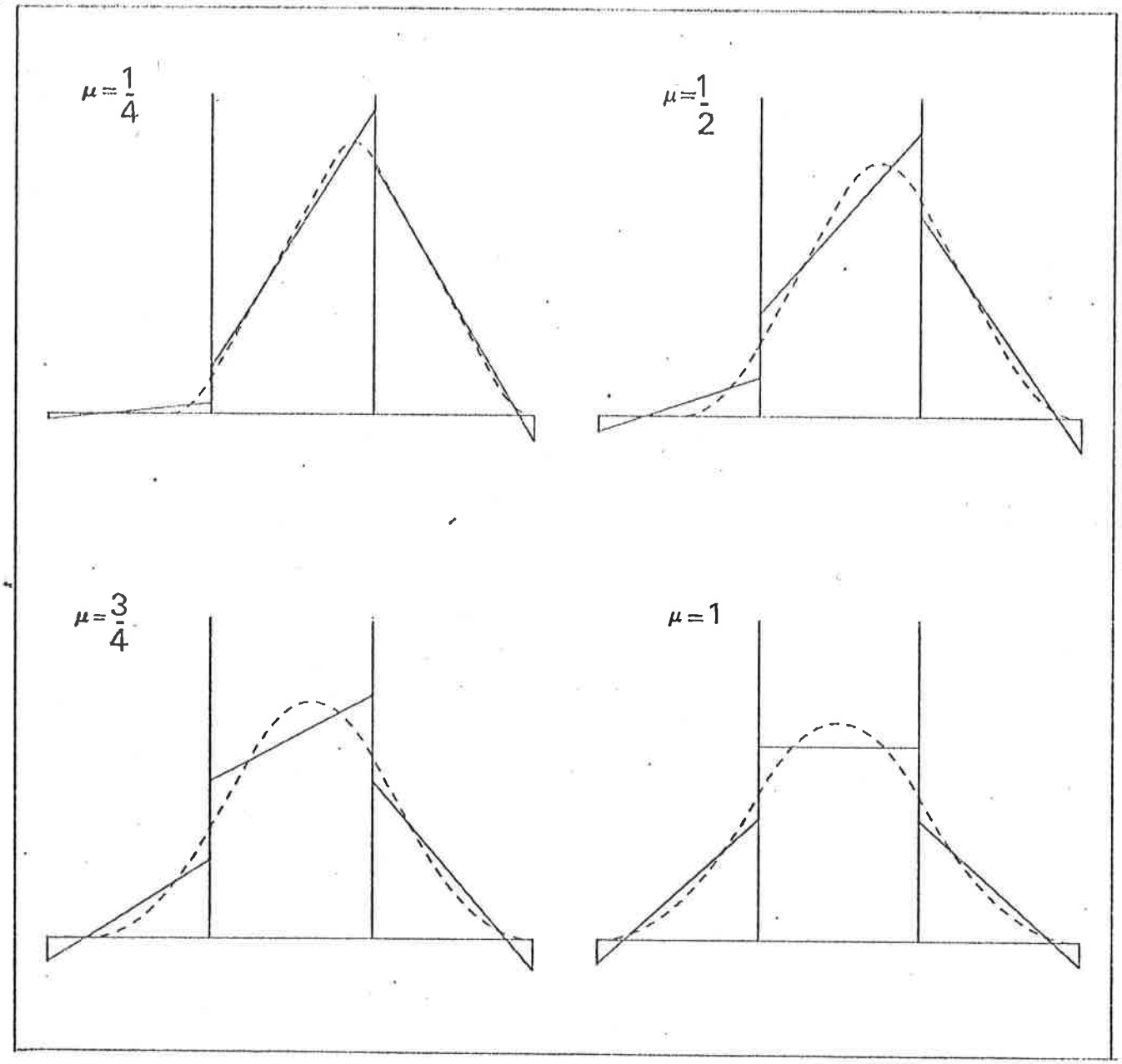


Figure 8 : The approximate test function $\Phi^I(s,\mu)$ compared with the exact $\Phi(s,\mu)$.

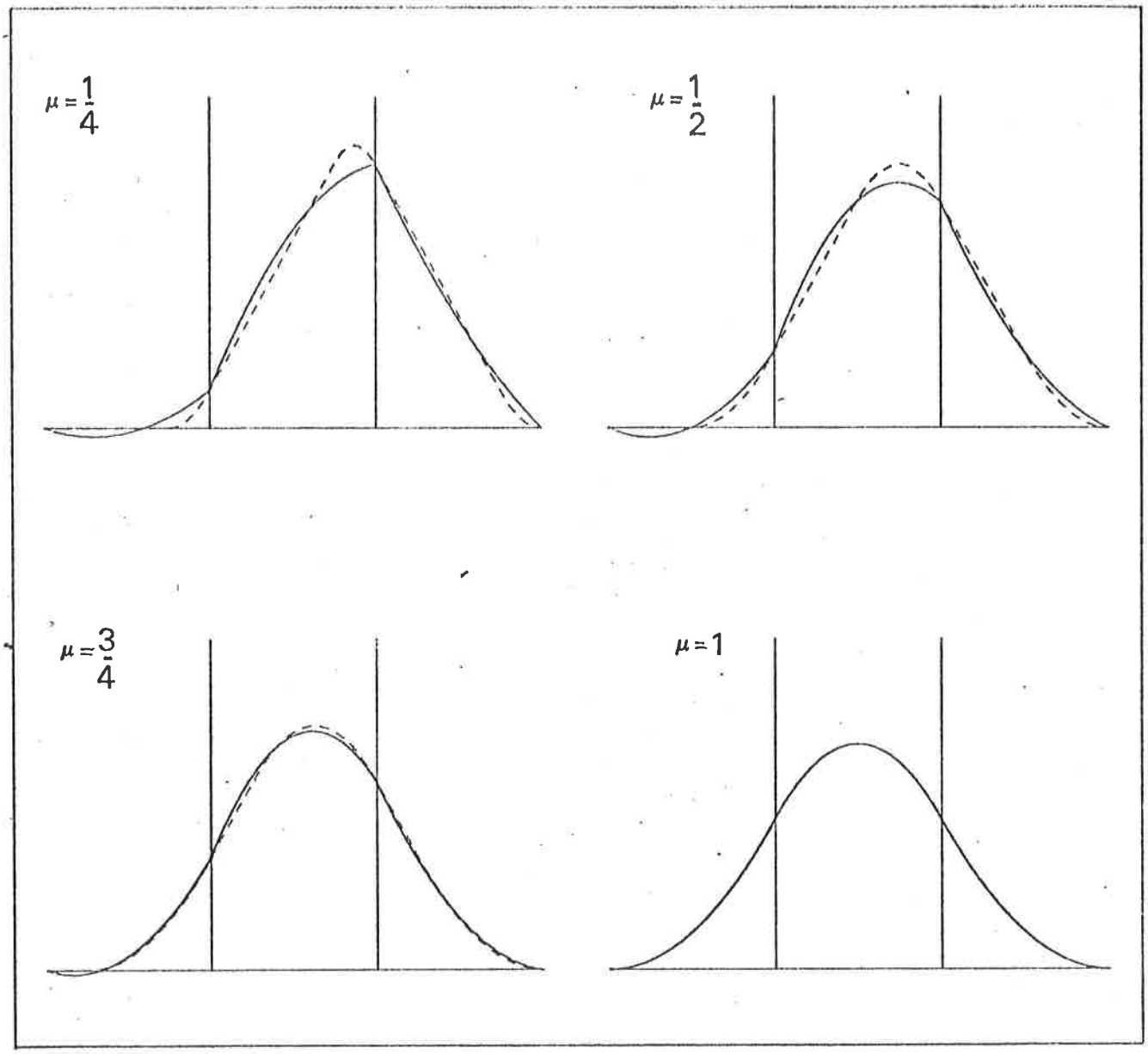


Figure 9 : Advection of a Gaussian profile and a ramp function by the ECG scheme and Gadd's scheme, for $\mu = 0.8$.

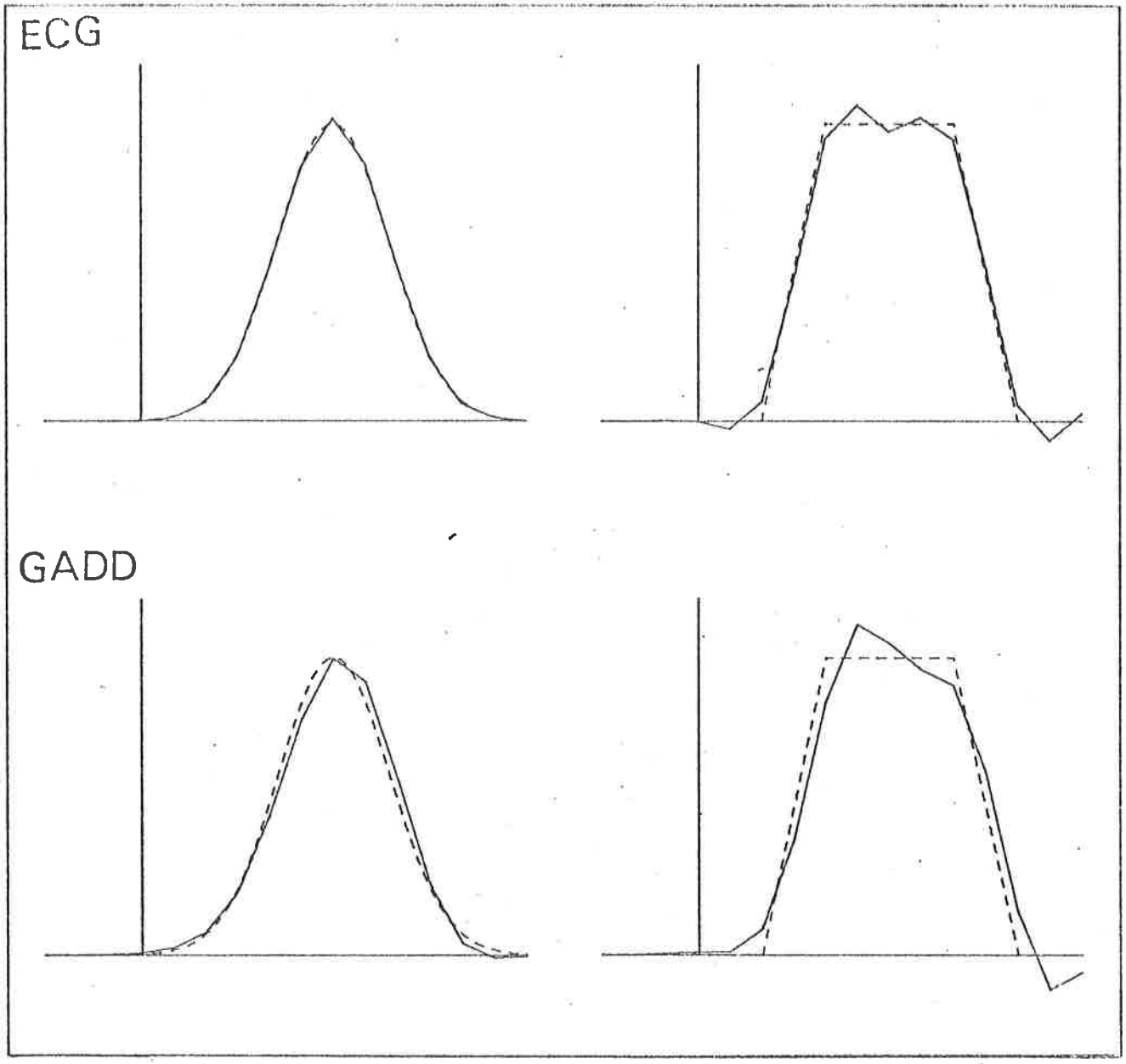
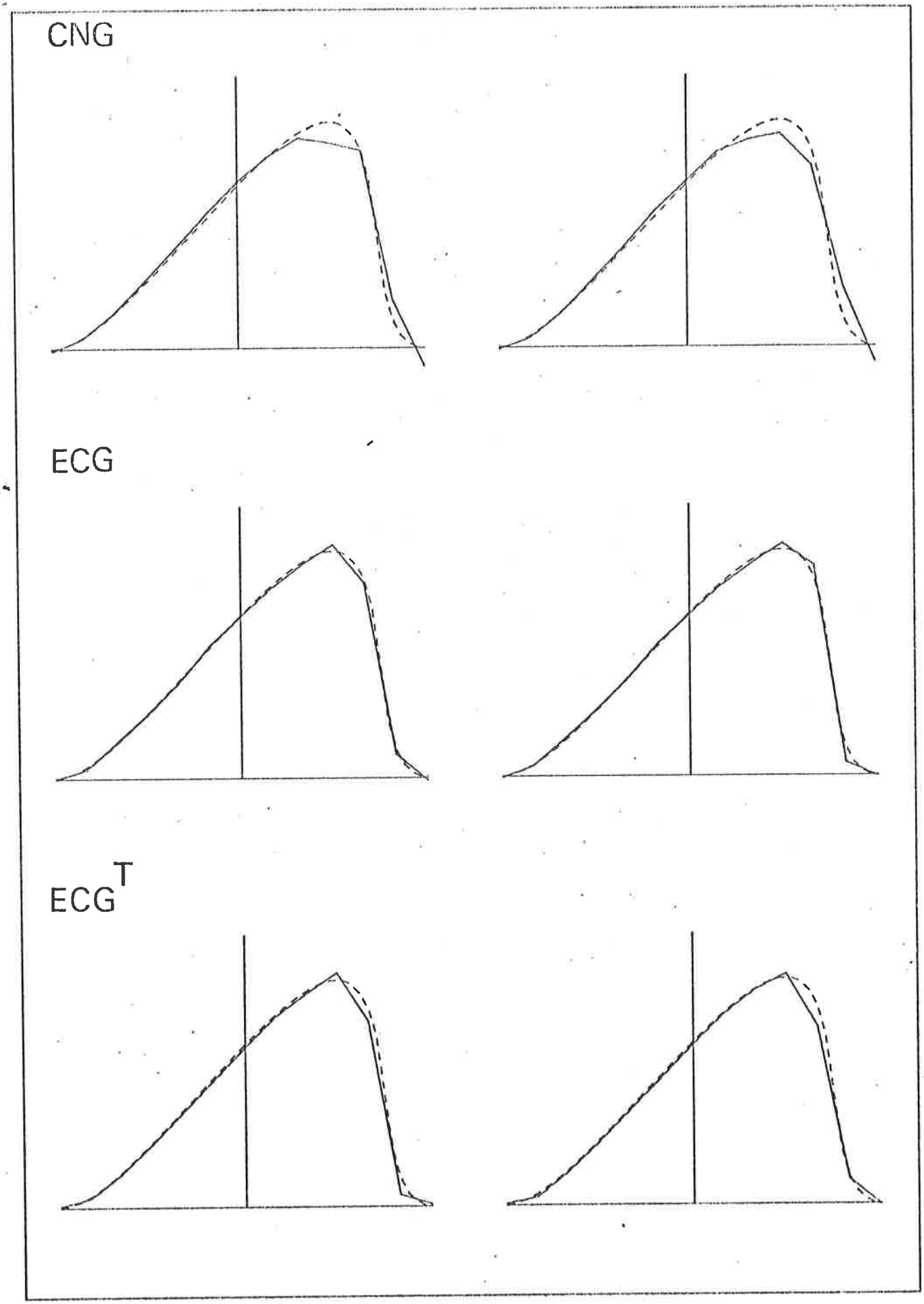


Figure 10 : Non-linear advection by Crank-Nicolson-Galerkin (CNG), exact ECG and approximate ECG (ECG^T) schemes : product approximation is used in the right-hand set.



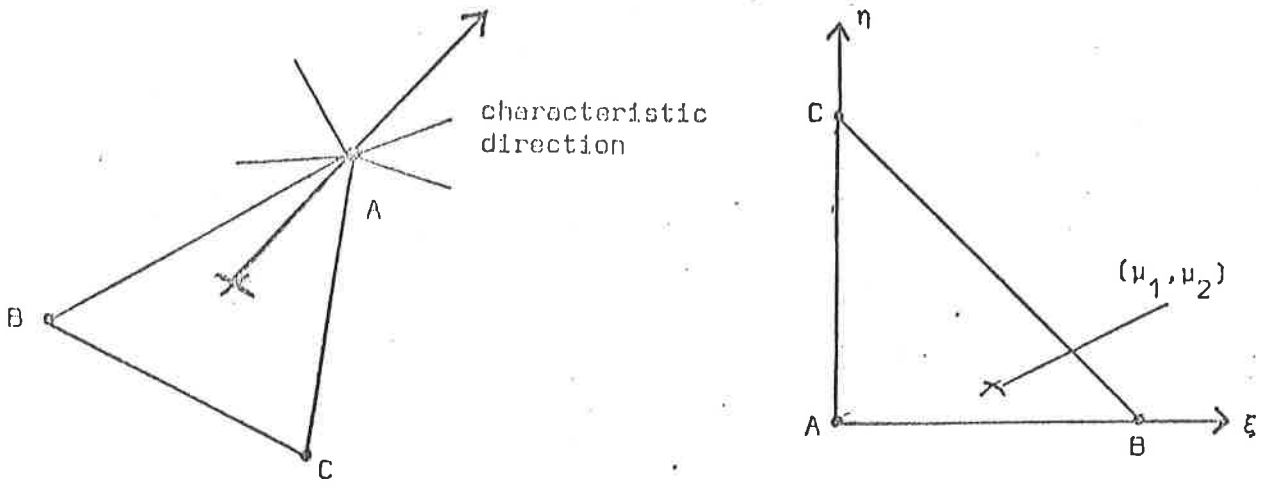


Figure 11 : Layout of triangle for calculation of approximate ECG test function from (3.17).

Figure 12 : Convection of a Gaussian after ¼, ½, ¾ and 1 revolution with approximate ECG scheme.

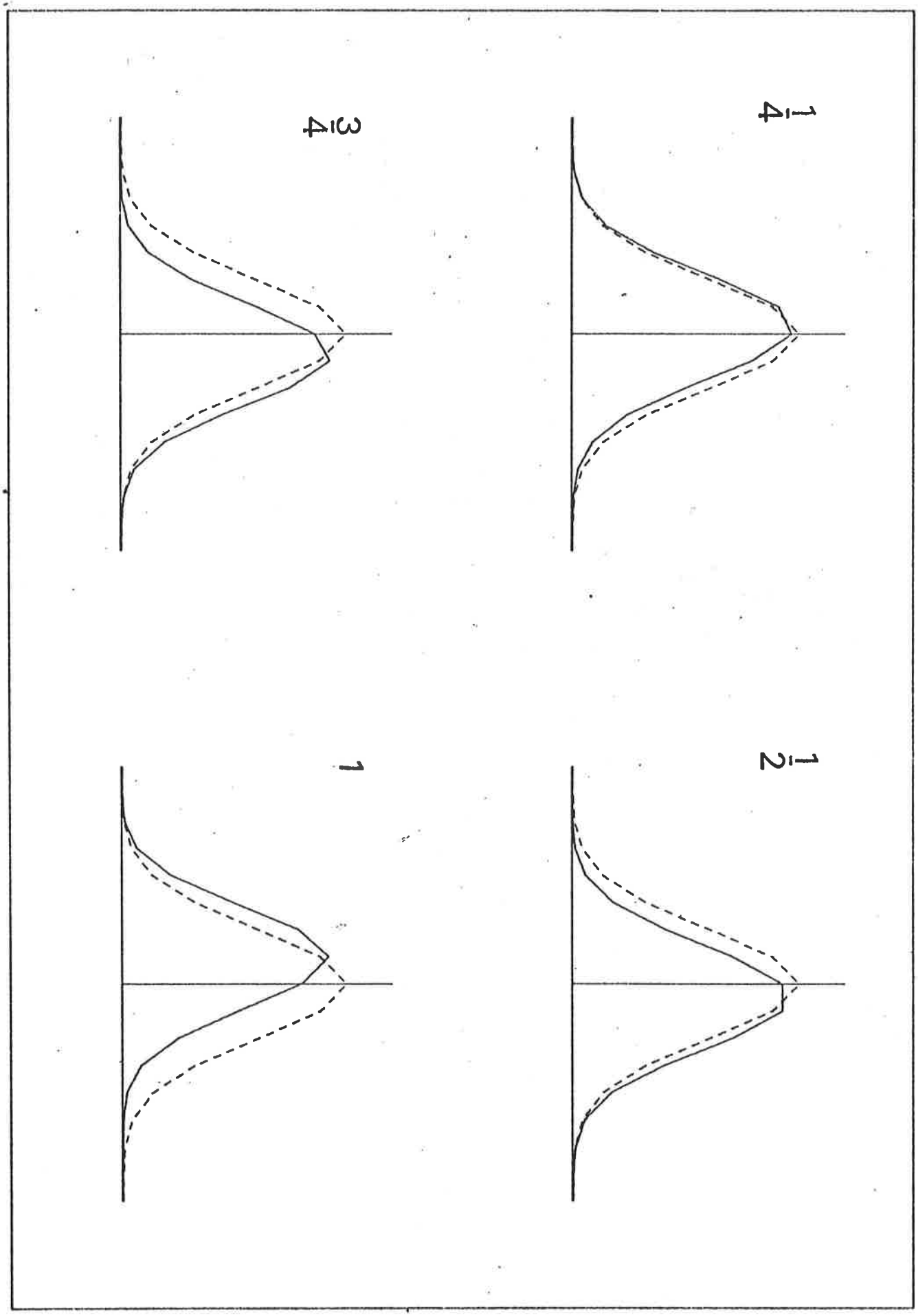


Fig. 12