JOURNAL OF COMPUTATIONAL PHYSICS 45, 221–245 (1982)

The Use of Quadratic Finite Element Methods and Irregular Grids in the Solution of Hyperbolic Problems

M. J. P. CULLEN

Forecasting Research Branch, Meteorological Office, Bracknell, Berkshire RG12 2SZ, *England*

Received June 18, 1981

The Galerkin method for first order hyperbolics using most higher order finite elements on any mesh, or using any type of element on an irregular mesh, is known to give a low order of accuracy. This is because the exact solutions become distorted, though the propagation speeds are handled to a higher order of accuracy. The distortion comes about because the method makes no smoothness assumptions in its formulation. Computations with these methods can show rapid growth of small scale noise. The algorithms can be improved by making smoothness assumptions. Many other techniques for obtaining higher order accuracy with such elements ignore the structure of the error and give worse results than standard Galerkin methods. This paper presents analysis and computational examples to support these statements.

1. INTRODUCTION

This paper studies the use of quadratic finite elements and irregular meshes in hyperbolic problems. There has been a divergence of experience with these methods between applications to meteorological problems, in particular integrations of the shallow water equations for long periods, and engineering applications with transient behavior. In the former case, good results have only been obtained using linear elements on a regular mesh, and the use of higher order elements or irregular meshes leads to rapid growth of noise (e.g., [7]). This experience has been explained in terms of superconvergence results for these cases [8, 28]. However, accurate results for transient problems in different contexts have been obtained using higher order elements, e.g., [29]. There is therefore a need to understand this difference. This is especially so because higher order elements and irregular grids are needed to be able to take advantage of the finite element method in situations that seem suited to it, for instance, fluid flow in irregular geometry. If the method has to be restricted to linear elements on a regular mesh there is no point in using it instead of finite difference and spectral methods. In particular, provided the geometry can be represented by a global coordinate transformation, the spectral method gives very good results including an accurate treatment of nonlinearity [3, 20, 24]. The potential advantage of finite

0021-9991/82/020221-25$02.00/0

element methods is that the correct treatment of nonlinearity resulting from the Galerkin method and demonstrated in [8] can be used in contexts where the spectral method is inconvenient.

A major difference between the meteorological and most engineering applications is the almost inviscid nature of the former problem. The absence of viscosity means that "optimal" order of accuracy is no longer obtained with the finite element Galerkin method [12]. The higher order of accuracy can be recovered by changing the algorithm, e.g., [9–11]. A recent analysis [15] has shown that applying the Galerkin method to the first order wave equation using either quadratic or Hermite cubic elements leads to an accurate solution with a spurious additional wave. It is the spurious wave which restricts the order of accuracy and leads to the growth of noise in the inviscid application. In other applications this effect may be damped by the viscosity.

In this paper the ideas of [15] are followed up, in particular to study what happens away from the limit of small mesh size. It is therefore necessary to carry out the analysis of modified Galerkin methods for all resolvable scales, as was done in [8]. With this viewpoint, it is demonstrated that the Galerkin method is difficult to improve on for a given finite element representation, because a technique which improves the asymptotic order of accuracy may be less accurate for other resolved scales. Only undamped schemes for the wave equation are analysed in this paper, for reasons of space. The use of schemes with some damping to eliminate the small scale noise is widespread [5, 10], but there is a danger that the detailed information carried by a higher order element may be lost as a result. If this information is not used there seems to be no advantage in using such an element which is very expensive to compute with.

The same sort of analysis is also applied to integrations on an irregular grid. Error analysis away from the asymptotic limit must now be carried out by a computer. It is again demonstrated that the Galerkin technique has advantages over modified methods, in that the error is almost entirely small scale noise which can be filtered easily. However, for transient problems, it still seems better to use as smoothly varying a grid as possible to avoid the need for filtering. With a smooth variation the method has proved successful [25].

2. Framework for Error Analysis

In order to extend the work of [15] to finite mesh size and analyse other possible modified schemes it is necessary to establish an appropriate framework. The spurious wave found in [15] was caused by an eigensolution of the discrete equation not corresponding to an eigensolution of the differential equation. The generalized error analysis of [8] shows that, in order to understand finite element methods, it is necessary to construct a restriction operator taking continuous data into a finite element representation, and to require that the same operator be used throughout the solution of a problem. In the case of higher order elements we consider three types of

restriction: a pure interpolation, a least squares fit, and a special restriction defined by taking eigenfunctions of the differential equation into those of the discrete equation. The difficulty comes about because we can usually only define eigenfunctions for certain parts of the differential equation, so that the special restriction will not be suitable for the rest of the calculation. The inconsistencies resulting are shown to cause errors. The advantage of the least squares restriction is that, for all resolvable scales, the restricted continuous eigenfunction can be shown to be very close to the discrete eigenfunction.

The error analysis is carried out by the same techniques as those in [8]. Consider pure initial value problems of the form

$$\begin{aligned} \partial_t u &= Lu \qquad \text{on} \quad [0, T] \times \mathbb{R}^d, \\ u &= u^0 \qquad \text{at} \quad t = 0, \end{aligned} \tag{2.1}$$

where u is vector-valued (possibly complex) and L is a differential operator on $\mathbb{R}^d$ which may be nonlinear but has real coefficients that do not depend explicitly on t. In many problems of practical interest L consists of a combination of sums, products and derivatives. It was demonstrated in [8] that the error made in approximating (2.1) by a multi-stage technique could be estimated by calculating the errors in approximating single derivatives and products. In this paper we calculate the errors made in approximating derivatives and products using quadratic elements.

The theoretical framework required is the same as in [8] but with an extension. Suppose that for each $t \in [0, T]$ the solution u of (2.1) lies in some Hilbert space V, called the solution space. Following Aubin [2] we associate the triplet (V_h, p_h, r_h) with any procedure for approximating members of V on a discrete mesh in $\mathbb{R}^d$ characterised by a positive mesh length h. V_h is the space of discrete parameter values defining an approximation, r_h a restriction operator associating such values with a given member of V, and p_h a prolongation which creates the approximation in V from the discrete parameter values. Three choices based on the quadratic finite element representation are as follows:

(i) Introduce mesh points $x_j = jh$, $j = -J, -J+1, \ldots, J$. On each mesh interval use the mesh point values and a midpoint value at $x = (j + \frac{1}{2})h$ to define a quadratic function. By construction this will give a C^0 piecewise representation over the interval $(-Jh, Jh)$. Write this function in the form

$$W(x) = \sum_{j=-J}^{J} W_j \theta_j(x) + \sum_{j=-J}^{J-1} W_{j+1/2} \chi_{j+1/2}(x), \tag{2.2}$$

where

$$\begin{aligned} \theta_j(x) &= \delta_{jk} \qquad \text{at} \quad x = kh \\ &= 0 \qquad \text{at} \quad x = (k + \tfrac{1}{2})h, \\ \chi_{j+1/2}(x) &= \delta_{jk} \qquad \text{at} \quad x = (k + \tfrac{1}{2})h \\ &= 0 \qquad \text{at} \quad x = kh. \end{aligned}$$

Define the best fit restriction r_h^q of a function w by setting

$$r_h^q w = (W_j^*, W_{j+1/2}^*), \qquad -J \leqslant j \leqslant J \qquad w \in V, \tag{2.3}$$

where W_j^* and $W_{j+1/2}^*$ are determined from

$$\int \left(w - \sum W_j^* \theta_j(x) - \sum W_{j+1/2}^* \chi_{j+1/2}(x) \right) \theta_j(x) = 0 \tag{2.4}$$

and

$$\int \left(w - \sum W_j^* \theta_j(x) - \sum W_{j+1/2}^* \chi_{j+1/2}(x) \right) \chi_{j+1/2}(x) = 0.$$

The corresponding prolongation operation p_h^q is

$$p_h^q(W_j^*) = \sum_{j=-J}^{J} W_j^* \theta_j(x) + \sum_{j=-J}^{J-1} W_{j+1/2}^* \chi_{j+1/2}(x). \tag{2.5}$$

(ii) Another restriction using the piecewice quadratic representation (2.2) is the Gauss-point collocation restriction r_h^c. This is defined by (2.3) with (2.4) replaced by the collocation

$$W = \sum W_j^* \theta_j(x) + \sum W_{j+1/2}^* \chi_{j+1/2}(x) \tag{2.6}$$

at the Gauss points. Two points are required for each mesh interval.

(iii) A grid-point restriction operator r_h^g is defined by satisfying (2.6) at the mesh points $x = jh$ and $x = (j + \frac{1}{2})h$. The finite element representation is now only being used as a method of interpolation.

Using this framework we consider a semi-discrete approximation u_h to the solution u of Eq. (2.1). For each $t \in [0, T]$ it is a member of the discrete parameter space V_h and is given by the system of ordinary differential equations

$$\partial_t u_h = L_h u_h, \tag{2.7}$$

where $L_h : V_h \to V_h$ is an operator which in some sense approximates L. Then, as in [8], define the evolutionary error e_h as $r_h u - u_h$. It satisfies the equation

$$\partial_t e_h - (L_h r_h u - L_h u_h) = (r_h L - L_h r_h) u, \tag{2.8}$$

where the term on the right is the truncation error (T.E.). Examples of the T.E. for various choices of finite element representation are given in [8]. In particular, the asymptotic behavior of the T.E. as $h \to 0$ is markedly different for spline Galerkin methods on a regular mesh, and other higher order finite element Galerkin methods. However, as noted in [8], and demonstrated extensively in [15], it may be possible to extract higher accuracy for general finite element methods, at least on a regular mesh.

Deterioration of the accuracy on an irregular mesh is observed in computations and it is to be expected that it will show up in the theory. If the problem on an irregular mesh is reinterpreted as a variable coefficient problem on a regular mesh, high order accuracy can be recovered at the cost of increased nonlinearity which, unfortunately, may be just as damaging as the irregular mesh.

To analyse the behavior of general finite elements on a regular mesh in a reasonably general context, consider the case where L is linear and has a complete set of normalised eigenfunctions ψ_n so that the exact solution of (2.1) can be written as $\sum u_n(t)\,\psi_n(x)$. This method can also be used for a local analysis of problems where the geometry or boundary conditions make it difficult to determine global eigenfunctions. Suppose that the approximation space V_h is N dimensional. Then construct an alternative approximation triplet (V_N, p_N, r_N) by setting

$$\begin{aligned} r_N w &= \{W_j^* : 1 \leqslant j \leqslant N\}, \\ p_N W^* &= \sum_{j=1}^{N} W_j^* \psi_j(x), \end{aligned} \tag{2.9}$$

where the ψ_j are ordered according to the modulus of the associated eigenvalues λ_j. Suppose that the subspace of V spanned by $p_N r_N u$, $u \in V$ is U. Now construct some operators between the spaces V, V_h, V_N and U. Since U, V_N and V_h all have the finite dimension N, the restriction $r_h : U \to V_h$ will in general be invertible. Define $q_h : V_h \to U$ as r_h^{-1}. The projection $q_h r_h : V \to U$ does not have the general approximation properties of the orthogonal projection $p_h r_h$, but q_h can be considered as a post-processing optimal recovery operator in the sense of [23] given a discrete solution $u_h \in V_h$.

It is clear that an approximate solution of Eq. (2.1) based on representation (2.9) would be exact for data in the subspace U. We use this to analyse the T.E. of the finite element solution in more detail. The semi-discrete finite element approximation is given by Eq. (2.7). Suppose that the approximate operator L_h has N independent normalized eigenfunctions ψ_{hn} in V_h with associated eigenvalues λ_{hn}.

Order these eigenfunctions by minimising $\|r_h \psi_n - \psi_{hn}\|_{V_h}$ over all possible orderings. Define a new restriction operator $r_L : V \to V_h$ by

$$r_L \psi_n = \psi_{hn}, \quad n \leqslant N; \qquad r_L \psi_n = 0, \quad n > N. \tag{2.10}$$

In the case of linear elements on a regular mesh and $L = \partial/\partial x$ we have $r_L = r_h$, but for general elements $r_L \neq r_h$. Define an operator $q_L : V_h \to U$ as $(r_L \mid U)^{-1}$. Define a composite operator $s_h = r_L q_h$, by construction it has an inverse s_h^{-1} which equals $r_h q_L$. Then

$$s_h r_h \psi_n = \psi_{hn}. \tag{2.11}$$

The operator s_h projects the finite element representation of eigenfunctions of the continuous equations into eigenfunctions of the discrete equations.

We now use these additional operators to study the T.E. in the semi-discrete

solution where L is approximated by L_h or, alternatively by $s_h^{-1}L_h s_h$. The error is analysed for solutions u in the subspace U. This is reasonable for problems where the solution remains smooth, but not for problems where shocks develop. Then write

$$u = \sum_{n=0}^{N} u_n \psi_n.$$

The T.E. for the operator L_h given by Eq. (2.8) is

$$\begin{aligned}
&\sum_{n=0}^{N} \lambda_n u_n r_h \psi_n - u_n L_h r_h \psi_n \\
&\quad = \sum_{n=0}^{N} \lambda_n u_n r_h \psi_n - u_n L_h \psi_{hn} + u_n L_h(\psi_{hn} - r_h \psi_n) \\
&\quad = \sum_{n=0}^{N} (\lambda_n - \lambda_{hn})\, u_n \psi_{hn} + (\lambda_n u_n - u_n L_h)(r_h \psi_n - \psi_{hn}). \qquad (2.12)
\end{aligned}$$

The limit of this as $h \to 0$ can be estimated from bounds on $(\lambda_n - \lambda_{hn})$ and $(r_h \psi_n - \psi_{hn})$. In the special case of the spline Galerkin method for $Lu = u_x$ on a regular mesh, $r_h \psi_n = \psi_{hn}$, and only $(\lambda_n - \lambda_{hn})$ need be estimated. This turns out to converge much more rapidly than the approximation error $(u - p_h r_h u)$. The analysis of [15] suggests that $(\lambda_n - \lambda_{hn})$ may be superconvergent for more general finite element approximations to this L, but the term $(r_h \psi_n - \psi_{hn})$ is of lower order and dominates the error estimate.

Now replace L_h by

$$s_h^{-1} L_h s_h. \qquad (2.13)$$

The expression $u_n L_h r_h \psi_n$ in (2.12) is replaced by

$$u_n s_h^{-1} L_h s_h r_h \psi_n.$$

Using (2.11), this becomes

$$u_n s_h^{-1} L_h \psi_{hn} = u_n s_h^{-1} \lambda_{hn} \psi_{hn} = u_n r_h \lambda_{hn} \psi_n.$$

The T. E. becomes

$$\sum_{n=0}^{N} (\lambda_n - \lambda_{hn})\, u_n r_h \psi_n. \qquad (2.14)$$

The only remaining term is often superconvergent. It is important to note that (2.13) is a different algorithm from L_h and in no sense has the estimate of the accuracy of the solution using L_h been altered.

This analysis shows that the nature of the restriction r_h is critical in the error estimates, because it can alter $\|r_h \psi_n - \psi_{hn}\|$. In [8] it was only important to consider

the nature of r_h for nonlinear problems. For linear problems on a regular mesh with one degree of freedom per element, $r_h\psi_n = \psi_{hn}$ for any reasonable choice of r_h. When other elements are used it is necessary to consider the effect of r_h on the algorithms for linear as well as nonlinear operators. The choice of r_h which allows $\|r_h\psi_n - \psi_{hn}\|$ to be minimised may be different from the choice which gives the best nonlinear properties and thus the overall choice will be problem dependent.

The analysis presented above includes the analysis of [15] in the case where r_h is the grid-point restriction. The filtering procedure proposed there is a local approximation to (2.11) such that $\|r_h\psi_n - \psi_{hn}\|$ is of the same order as $|\lambda_n - \lambda_{hn}|$. A similar type of eigenfunction analysis is used in [17] to understand the behavior of finite element approximations to a system of equations.

3. Review of Schemes for Linear Advection Using Quadratic Elements and Least Squares Representation of Data

Analysis of Galerkin Scheme

Assume that the boundary conditions are periodic with period D. Consider the restriction r_h^q (Eq. (2.3)) on a regular mesh applied to the initial condition $u = e^{ikx}$ with $k = 2n\pi/D$. Write $\xi = kh$, where h is the mesh length. Then we can define functions $\alpha_0(\xi)$, $\alpha_{1/2}(\xi)$ and write

$$\begin{aligned}(r_h^q e^{ikx})_j &= \alpha_0(\xi)\, e^{ij\xi},\\ (r_h^q e^{ikx})_{j+1/2} &= \alpha_{1/2}(\xi)\, e^{i(j+1/2)\xi}.\end{aligned} \tag{3.1}$$

The function of interest is $\alpha_0(\xi)/\alpha_{1/2}(\xi)$, which determines the shape of the restricted wave. The Galerkin approximation L_h to Eq. (2.7) with $Lu = u_x$ is given by

$$\begin{aligned}(u_t)_j + 8(u_t)_{j+1/2} + (u_t)_{j+1} &= 10(u_{j+1} - u_j)/h\\ -(u_t)_{j-1} + 2(u_t)_{j-1/2} + 8(u_t)_j + 2(u_t)_{j+1/2} - (u_t)_{j+1}&\\ = 10(2(u_{j+1/2} - u_{j-1/2}) - \tfrac{1}{2}(u_{j+1} - u_{j-1}))/h.&\end{aligned} \tag{3.2}$$

The calculations of [8] show that the data (3.1) are not an eigensolution of (3.2), and the calculations of [15] show the same thing for the grid-point restriction. Seek eigensolutions u_h, where

$$\begin{aligned}L_h u_h &= i\beta(\xi)\, u_h,\\ (u_h)_j &= \hat{\alpha}_0(\xi)\, e^{ij\xi},\\ (u_h)_{j+1/2} &= \hat{\alpha}_{1/2}(\xi)\, e^{i(j+1/2)\xi}.\end{aligned}$$

Then

$$\beta(\hat{a}_0 \cos \tfrac{1}{2}\xi + 4\hat{a}_{1/2}) = 10\hat{a}_0 \sin \tfrac{1}{2}\xi,$$
$$\beta\hat{a}_0(4 - \cos \xi) + 2\hat{a}_{1/2}\beta \cos \tfrac{1}{2}\xi = 20\hat{a}_{1/2} \sin \tfrac{1}{2}\xi - 5\hat{a}_0 \sin \xi. \tag{3.3}$$

This pair of nonlinear equations can be solved to give two values of $\hat{a}_0/\hat{a}_{1/2}$ and of β. This is to be expected, because substituting $\xi = 2\pi - \xi$ in (3.3) and constructing the quadratic equation for $\hat{a}_0/\hat{a}_{1/2}$ gives the same equation as that derived from (3.3) except that the sign of the linear term is reversed; so the solutions are reversed in sign. Therefore one of the two solutions corresponds to ξ and the other to $(2\pi - \xi)$ with a sign change. In no sense is either solution spurious. Since e^{ikx} is an eigenfunction of $L = \partial/\partial x$ with periodic boundary conditions, the difference between $\alpha_0/\alpha_{1/2}$ and $\hat{a}_0/\hat{a}_{1/2}$ is a measure of $(r_h\psi_n - \psi_{hn})$ in Eq. (2.12). Values of $\alpha_0/\alpha_{1/2}$, $\hat{a}_0/\hat{a}_{1/2}$ and β are shown in Table I for a full range of arguments. The eigenvalues are also plotted in Fig. 1 and the eigenfunctions in Fig. 2.

Inspection of Table 1 shows the second order accuracy of the approximation to $u_t = u_x$ resulting from a second order difference between $\alpha_0/\alpha_{1/2}$ and $\hat{a}_0/\hat{a}_{1/2}$. This results in an $O(\xi^2)$ term in $(r_h\psi_n - \psi_{hn})$ in the truncation error (Eq. (2.12)). The error in the eigenvalue is $O(\xi^4)$ with a very small coefficient. This analysis supersedes a similar analysis attempted in [8] where the two solutions were not properly separated. The results for $\hat{a}_0/\hat{a}_{1/2}$ and β for $\xi \to 0$ were presented in [15]; but the comparison with $\alpha_0/\alpha_{1/2}$ was not made there.

For large values of ξ, the agreement between $\alpha_0/\alpha_{1/2}$ and $\hat{a}_0/\hat{a}_{1/2}$ is very close. Since any practical use of an improved scheme like (2.13) would involve a local approximation to the extra projection s_h, it will be difficult to improve (3.2) effectively. We show this by illustrating two improvements based on (2.13).

Projection to Eigenspace of Galerkin Scheme

Seek a scheme of the form (2.13) with L_h defined by (3.2). The operator s_h must satisfy

$$(s_h r_h^q e^{ikx})_j = \hat{a}_0(\xi)\, e^{ij\xi},$$
$$(s_h r_h^q e^{ikx})_{j+1/2} = \hat{a}_{1/2}(\xi)\, e^{i(j+1/2)\xi}, \tag{3.4}$$

TABLE I

Projected Eigenfunctions of L, and Eigenfunctions and Eigenvalues of L_h

ξ	$\xi \to 0$	$\pi/8$	$\pi/4$	$\pi/2$	$3\pi/4$	π	$3\pi/2$	2π
$\alpha_0/\alpha_{1/2}$	$1 + 9\xi^4/896 + O(\xi^6)$	1.00023	1.00320	1.0348	1.1170	1.257	1.693	2.0
$\hat{a}_0/\hat{a}_{1/2}$	$1 + \xi^2/48 + O(\xi^4)$	1.00323	1.01312	1.0556	1.1358	1.265	1.684	2.0
$\beta(\xi)/i\xi$	$1 + \xi^4/4320 + O(\xi^6)$	1.000005	1.00008	1.0011	1.0043	1.007	0.900	0.0

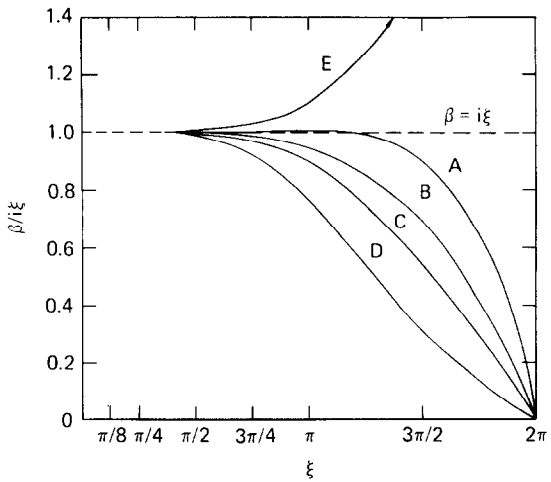

FIG. 1. Relative eigenvalues of various approximations to first derivative: (A) Quadratic Galerkin. (B) Linear Galerkin. (C) Quadratic Galerkin with mass lumping. (D) Taylor series. (E) Gauss-point collocation.

where r_h^q is defined by (3.1) and $\hat{\alpha}_0, \hat{\alpha}_{1/2}$ by (3.3). It is clear that s_h will not have a local definition. Seek a local approximation $\hat{s}_h$ to s_h such that the difference $(s_h r_h \psi_n - \psi_{hn})$ is small as $\xi \to 0$. The example here is chosen to remove the $O(\xi^2)$ term and to preserve the small error in the eigenfunction at $\xi = \pi$ shown in Table I. Alternatively, it could have been chosen to remove more terms in the asymptotic error. Define $\hat{s}_h$ by

$$\begin{aligned} \hat{s}_h u_{j+1/2} &= u_{j+1/2}, \\ \hat{s}_h u_j &= (22u_j + 2(u_{j+1/2} + u_{j-1/2}) - (u_{j+1} + u_{j-1}))/24. \end{aligned} \tag{3.5}$$

The analysis of the algorithm $\hat{s}_h^{-1} L_h \hat{s}_h$ is set out in Table II. The asymptotic error is now $O(\xi^4)$. For small ξ the error in the eigenfunction is reduced, at $\xi = \pi/8$ it is reduced by a factor of 20. For $\pi > \xi > \pi/2$ the error is similar for both schemes and for $\xi > \pi$ the error is greatly increased using (3.5) as can be seen in Fig. 2. For some applications, however, the larger errors in the eigenvalue for $\xi > \pi$ may make increased eigenfunction error unimportant, and the reduced errors for small ξ may reduce the generation of roughnesses sometimes observed with (3.2). A numerical experiment illustrating this statement is shown in Section 5.

Projection to Eigenspace of Finite Difference Operator

An alternative strategy is to calculate accurate grid-point values of u from u_h on a regular mesh, and then use a high order finite difference operator to estimate the point values of the derivative, for instance, using the schemes of [6]. These values are then projected back into a least squares representation. This can be done using the

techniques of [4] and [27] for optimal recovery of point values and derivatives from $r_h^q u$, given the least squares definition (2.4). For present purposes the scheme is derived at the mesh points by matching coefficients of the Taylor expansion. Define $\hat{s}_h$ by

$$\begin{aligned} \hat{s}_h u_{j+1/2} &= u_{j+1/2}, \\ \hat{s}_h u_j &= (74u_j + 36(u_{j+1/2} + u_{j-1/2}) - 9(u_{j+1} + u_{j-1}))/128. \end{aligned} \tag{3.6}$$

If L_h is now a finite difference operator on all the nodal values $u_j, u_{j+1/2}$, the eigenfunctions ψ_{hn} will satisfy

$$u_j = e^{ij\xi}, \qquad u_{j+1/2} = e^{i(j+1/2)\xi}$$

and the ratio $\hat{\alpha}_0/\hat{\alpha}_{1/2}$ will be unity. The analysis of eigenfunctions and eigenvalues of the composite operator $\hat{s}_h^{-1} L_h \hat{s}_h$, where $\hat{s}_h$ is given by (3.6), and L_h is the compact difference operator for Eq. (2.7) given by

$$\begin{aligned} (u_t)_j + 4(u_t)_{j+1/2} + (u_t)_{j+1} &= 6(u_{j+1} - u_j)/h, \\ (u_t)_{j-1/2} + 4(u_t)_j + (u_t)_{j+1/2} &= 6(u_{j+1/2} - u_{j-1/2})/h \end{aligned} \tag{3.7}$$

is shown in Table II, and plotted in Fig. 2.

The difficulty with this approach is seen from the form of (3.6). The projection error $(u - p_h^q r_h^q u)$ is $O(\xi^4)$ but with a large coefficient, and very strong averaging is required to cancel the distortion factor $\alpha_0/\alpha_{1/2}$ given by (3.1). The exact choice of s_h

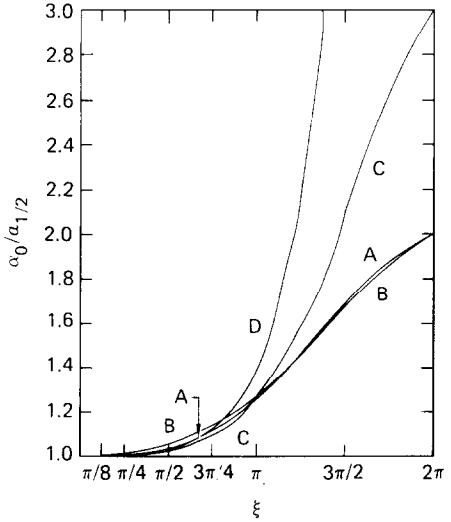

FIG. 2. Shape of eigenfunction $(\alpha_0/\alpha_{1/2})$ for various approximations to first derivative: (A) Least squares projection of sine wave. (B) Quadratic Galerkin eigenfunction. (C) Eigenfunction using extra projection (3.5). (D) Eigenfunction using extra projection (3.6).

TABLE II

Eigenfunctions and Eigenvalues of Modified Operators $\hat{s}_h^{-1} L_h \hat{s}_h$

ξ		$\xi \to 0$	$\pi/8$	$\pi/4$	$\pi/2$	$3\pi/4$	π	$3\pi/2$	2π
L_h is (3.2)	$\hat{\alpha}_0/\hat{\alpha}_{1/2}$	$1 + 13\xi^4/3456 + O(\xi^6)$	1.00009	1.00139	1.0204	1.0928	1.265	2.108	3.0
$\hat{s}_h$ is (3.5)	$\beta/i\xi$	$1 + \xi^4/4320 + O(\xi^6)$	1.000005	1.00008	1.0011	1.0043	1.007	0.900	0.0
L_h is (3.7)	$\hat{\alpha}_0/\hat{\alpha}_{1/2}$	$1 + 9\xi^4/896 + O(\xi^6)$	1.00010	1.00163	1.0237	1.1200	1.391	5.544	−8.0
$\hat{s}_h$ is (3.6)	$\beta/i\xi$	$1 - \xi^4/2880 + O(\xi^6)$	0.99999	0.99985	0.9977	0.9875	0.945	0.696	0.0

cannot be locally represented and it cannot be satisfactorily approximated locally for the full range of the argument ξ. Therefore this technique seems less appropriate as an alternative algorithm for L_h than for recovery of information at the end of a calculation, assuming that the solution is smooth.

For small ξ the values of $\hat{\alpha}_0/\hat{\alpha}_{1/2}$ given by (3.6) are closer to those of $\alpha_0/\alpha_{1/2}$ (Table I) than those given by (3.2). For ξ greater than $\pi/2$ the values of $\hat{\alpha}_0/\hat{\alpha}_{1/2}$ shown in Table I for the unmodified Galerkin scheme are just as close to $\alpha_0/\alpha_{1/2}$ as are those given by either (3.2) or (3.6). The bad effect of (3.6) for large ξ is easily seen. Both algorithms are tested computationally in Section 5.

Direct Construction of L_h by Taylor Series

The schemes discussed above allow the Galerkin algorithm (3.2) to be modified in regular geometry where the eigenfunctions are known. In general, these techniques could only be used on locally regular parts of a mesh. In this section we study the possibility of deriving an accurate approximation to $Lu = u_x$, assuming the least squares restriction r_h^q, which could be used on any mesh. This is a generalization of the Taylor series definition of finite difference schemes on a general mesh, e.g., [1]. The method will be derived here for a regular mesh.

The starting point is to regard the given data as the inner products $\int u\theta_j$, $\int u\chi_{j+1/2}$, where $\theta_j, \chi_{j+1/2}$ are the finite element basis functions defined in Eq. (2.2). Define an alternative restriction operator $\bar{r}_h$ by

$$v = \bar{r}_h u, \qquad v_j = 3\int u\theta_j \, dx, \qquad v_{j+1/2} = \tfrac{3}{2}\int u\chi_{j+1/2} \, dx. \tag{3.8}$$

Then, in order to solve $u_t = u_x$, we must estimate

$$\int u_x \theta_j \, dx, \qquad \int u_x \chi_{j+1/2} \, dx$$

in terms of $\{v_j, v_{j+1/2}\}$. On integrating by parts, this is equivalent to finding

$$-\int u(\theta_j)_x \, dx, \qquad -\int u(\chi_{j+1/2})_x \, dx.$$

Write

$$(v_j)_x = -3 \int u(\theta_j)_x \, dx,$$
$$(v_{j+1/2})_x = -\tfrac{3}{2} \int u(\chi_{j+1/2})_x \, dx. \tag{3.9}$$

Assume an approximation scheme of the form

$$\sum_{i=0}^{s} a_i((v_{j+i})_x + (v_{j-i})_x) = \sum_{i=0}^{s} b_i(v_{j+i} - v_{j-i}). \tag{3.10}$$

This formula has been simplified by symmetry, using the regularity of the mesh. The coefficients in (3.10) are calculated so as to make it exact for as high a degree polynomial u as possible. The equations for a_i and b_i are constructed using (3.8) and (3.9). Schemes of any asymptotic accuracy can be derived. The fourth order scheme is

$$10(v_0)_x + 7((v_{1/2})_x + (v_{-1/2})_x) = 24(v_{1/2} - v_{-1/2})/h,$$
$$5(v_{1/2})_x - (v_0)_x - (v_1)_x = 3(v_1 - v_0)/h. \tag{3.11}$$

This scheme can be analysed in the same way as the others. The eigenfunctions must be compared with the restricted eigenfunctions $\bar{r}_h e^{ikx}$ defined by (3.8), which have a shape defined by the parameter $\bar{\alpha}_0/\bar{\alpha}_{1/2}$. The results are shown in Table III, and Figs. 1 and 3.

This table shows that the scheme is accurate asymptotically but very much worse than (3.2) for large ξ. This is because the Taylor series matching assumes a smooth u, appropriate for small ξ while the Galerkin method takes all values of ξ into account in constructing a best fit. Thus (3.2) is less accurate for small ξ and much more accurate for large ξ. The results using (3.11) are so much worse at $\xi = \pi$ than the other schemes considered so far that the computational test was not worth carrying out. The poor performance of (3.11) as compared with compact difference schemes, which are derived in a similar way, is because of the uneven behavior of this finite element representation at endpoint and midpoint nodes. It represents a serious disadvantage of this representation, because this would have been a good way to derive algorithms on a general mesh.

TABLE III

ξ	$\xi \to 0$	$\pi/8$	$\pi/4$	$\pi/2$	$3\pi/4$	π	$3\pi/2$	2π
$\bar{\alpha}_0/\bar{\alpha}_{1/2}$	$1 + 3\xi^2/40 + O(\xi^4)$	1.01152	1.04554	1.1737	1.3604	1.571	1.911	2.0
$\hat{\alpha}_0/\hat{\alpha}_{1/2}$ (3.11)	$1 + 3\xi^2/40 + O(\xi^4)$	1.01164	1.04739	1.2032	1.5067	2.000	3.325	4.0
$\beta/i\xi$	$1 - \xi^4/450 + O(\xi^6)$	0.99995	0.99913	0.9852	0.9215	0.764	0.309	0.0

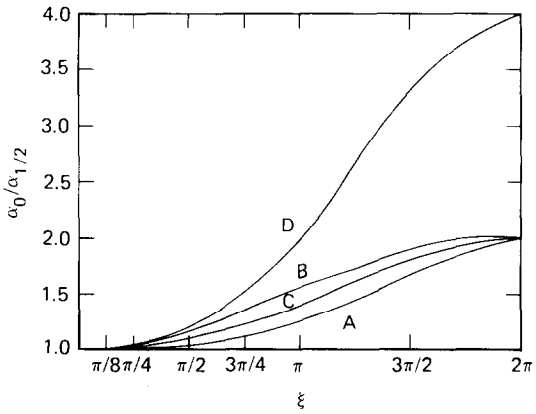

FIG. 3. Shape of eigenfunction $(\alpha_0/\alpha_{1/2})$ for various approximations to first derivative: (A) Least squares projection of sine wave. (B) Lumped mass projection of sine wave. (C) Eigenfunction using lumped mass. (D) Eigenfunction using Taylor series.

Scheme for Products

In order to complete the analysis of the quadratic element using the restriction r_h^q, we compute the error made in approximating products, using the Galerkin scheme. As shown in [8], approximations to general operators can then be derived. The approximation to $w = uv$, where $r_h u$ is written as $\{u_j, u_{j+1/2}\}$, etc., is given by

$$\begin{aligned} w_j + 8w_{j+1/2} + w_{j+1} &= ((48u_{j+1/2}v_{j+1/2} + 4(u_{j+1/2}v_j + v_{j+1/2}u_j \\ &\quad + u_{j+1/2}v_{j+1} + u_{j+1}v_{j+1/2}) + 5(u_jv_j + u_{j+1}v_{j+1}) - 2(u_jv_{j+1} + u_{j+1}v_j))/7 \\ &\quad - w_{j-1} + 2w_{j-1/2} + 8w_j + 2w_{j+1/2} - w_{j+1} = (39u_jv_j + 10(u_jv_{j+1/2} + v_ju_{j+1/2} \\ &\quad + u_jv_{j-1/2} + v_ju_{j-1/2}) - 3/2(u_jv_{j+1} + v_ju_{j+1} + u_jv_{j-1} + v_ju_{j-1}) \\ &\quad + 8(u_{j+1/2}v_{j+1/2} + u_{j-1/2}v_{j-1/2}) - 4(u_{j+1/2}v_{j+1} + v_{j+1/2}u_{j+1} + u_{j-1/2}v_{j-1} \\ &\quad + v_{j-1/2}u_{j-1}) - 3/2(u_{j+1}v_{j+1} + u_{j-1}v_{j-1}))/7. \end{aligned} \tag{3.12}$$

Simpler expressions could be obtained by collocation methods or numerical integration of the right hand side. Calculate the asymptotic error of this, using $u = e^{ikx}$, $v = e^{ilx}$ allowing for the effect of the restriction r_h^q. Then we find that

$$\begin{aligned} w_j &= \gamma_0(k, l)\, e^{ij(\xi+\eta)}, \\ w_{j+1/2} &= \gamma_{1/2}(k, l)\, e^{i(j+1/2)(\xi+\eta)} \qquad \text{with} \quad \xi = kh, \quad \eta = lh, \end{aligned}$$

where

$$\begin{aligned} \gamma_0(\xi, \eta) &= 1 + \frac{13}{1680}(\xi+\eta)^4 - \frac{3}{560}\xi\eta(\xi^2+\eta^2) + \cdots, \\ \gamma_{1/2}(\xi, \eta) &= 1 - \frac{31}{13{,}440}(\xi+\eta)^4 + \frac{19}{6720}\xi\eta(\xi^2+\eta^2) + \cdots. \end{aligned} \tag{3.13}$$

Fourier analysis of (3.1) gives

$$\alpha_0(\xi+\eta) = 1 + \frac{13}{1680}(\xi+\eta)^4 + \cdots,$$

$$\alpha_{1/2}(\xi+\eta) = 1 - \frac{31}{13{,}440}(\xi+\eta)^4 + \cdots.$$

The T.E. contribution is therefore fourth order and generated by the cross product terms in (3.13). These terms are multiplied by coefficients comparable to those multiplying the terms which appear in $\alpha_0(\xi+\eta)$. However, for large values of ξ and η the error does not grow excessively. For $\xi=\eta=\pi$, $\gamma_0(\pi,\pi)=1.94$, $\gamma_{1/2}(\pi,\pi)=0.91$, giving a ratio of 2.13 as compared with $\alpha_0(2\pi)/\alpha_{1/2}(2\pi)=2$.

4. Other Schemes for Advection Using the Quadratic Element

Gauss-Point Collocation

This method, as discussed in [9] and [11], gives fourth order accuracy for first order derivatives. The scheme for Eq. (2.7) with $Lu = u_x$ is

$$\begin{aligned}(u_t)_j + 4(u_t)_{j+1/2} + (u_t)_{j+1} &= 6(u_{j+1}-u_j)/h,\\ (1-\sqrt{3})((u_t)_{j-1}+(u_t)_{j+1}) + 4((u_t)_{j-1/2}+(u_t)_{j+1/2}) + (2+2\sqrt{3})(u_t)_j &\\ = 2(4\sqrt{3}(u_{j+1/2}-u_{j-1/2}) + (3-2\sqrt{3})(u_{j+1}-u_{j-1}))/h.&\end{aligned} \tag{4.1}$$

The eigenfunctions and eigenvalues of this scheme can be calculated as before, and the coefficients $\alpha_0/\alpha_{1/2}$ and $\beta/i\xi$ tabulated. Equation (4.1) can be viewed as a method for approximating derivatives as part of a multistage method based on the least squares representation. In this case the eigenfunction should be compared with $r_h^q e^{ikx}$ (Eq. (3.1)); as shown in Table I. Alternatively, the whole problem can be solved by collocation. In this case the initial data must be be derived using the restriction r_h^c (Eq. (2.6)) and the eigenfunctions of (4.1) compared with $r_h^c e^{ikx}$. These comparisons are made in Table IV and Figs. 1 and 4.

This table shows that the eigenfunctions of (4.1) are a better match to those given by r_h^q than r_h^c. The errors are large for large ξ, much greater than those given by the Galerkin scheme (3.2). For small ξ the errors are smaller than those of (3.2), as expected from the asymptotic analysis.

Use of Divided Differences

It is demonstrated in [13] and [18] that by using an mth order finite difference scheme to approximate L in Eq. (2.1) where finite elements including all polynomials up to degree $m-1$ are used to represent the data, an overall accuracy $O(h^m)$ is retained. For quadratic elements on a regular mesh we can use any fourth order finite difference scheme such as (3.7), to obtain fourth order accuracy. As with (4.1), a

TABLE IV

Eigenfunctions and Eigenvalues for Gauss-Point Collocation Scheme

ξ	$\xi \to 0$	$\pi/8$	$\pi/4$	$\pi/2$	$3\pi/4$	π	$3\pi/2$	2π
$\alpha_0/\alpha_{1/2}(r_h^q)$	$1 + 9\zeta^4/896 + O(\xi^6)$	1.00023	1.00320	1.0348	1.1170	1.257	1.693	2.0
$\alpha_0/\alpha_{1/2}(r_h^c)$	$1 + \xi^2/24 + O(\xi^4)$	1.00644	1.02599	1.1073	1.2536	1.476	2.065	2.0
$\hat{\alpha}_0/\hat{\alpha}_{1/2}$ (4.1)	$1 + \xi^4/390 + O(\xi^6)$	1.00006	1.00094	1.0133	1.0574	1.155	1.588	2.0
$\beta/i\xi$	$1 + \xi^4/710 + O(\xi^6)$	1.00003	1.00051	1.0075	1.0345	1.103	1.609	—

Note. The coefficients as $\xi \to 0$ are estimated from computed values.

finite difference scheme can be used to form part of a multistage scheme using a least squares representation of the data. Alternatively the whole problem can be solved with a finite difference algorithm, where the initial data and product terms are represented pointwise. The eigenfunctions in the finite difference representation have $\alpha_0(\xi) = \alpha_{1/2}(\xi) = 1$ for all ξ.

Use of Mass Lumping

This is a way of obtaining a less expensive rather than a more accurate scheme (e.g., [29]). The approximation to $Lu = u_x$ is obtained from (3.2) by replacing the left hand sides by $10(u_t)_{j+1/2}$ and $10(u_t)_j$. The eigenvectors and eigenvalues of the resulting scheme are shown in Table V and Figs. 1 and 3. The eigenvectors are compared with $r_h^q(e^{ikx})$ and with the "lumped mass" restriction $\bar{r}_h(e^{ikx})$ defined in Eq. (3.8).

This table shows that the eigenvectors of the lumped mass scheme are intermediate between those given by r_h^q and $\bar{r}_h$. The eigenvalue is still fourth order accurate, but with a larger coefficient than that given by (3.2). The errors in the eigenvector are also larger than those given by (3.2), especially near $\xi = \pi$. The eigenvalue errors resulting from mass lumping are much less than those from mass lumping with linear elements, as pointed out in [29]. However, the errors in matching the eigenfunction are substantially increased by lumping.

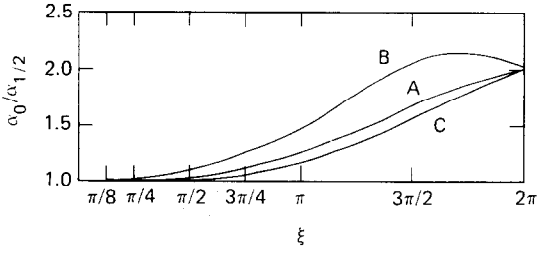

FIG. 4. Shape of eigenfunction $(\alpha_0/\alpha_{1/2})$ for various approximations to first derivative: (A) Least squares projection of sine wave. (B) Collocation projection of sine wave. (C) Eigenfunction using Gauss-point collocation.

TABLE V

Eigenvectors and Eigenvalues for Lumped Mass Scheme

ξ	$\xi \to 0$	$\pi/8$	$\pi/4$	$\pi/2$	$3\pi/4$	π	$3\pi/2$	2π
$\alpha_0/\alpha_{1/2}(r_h^q)$	$1+9\xi^4/896+O(\xi^6)$	1.00023	1.00320	1.0348	1.1170	1.257	1.693	2.0
$\alpha_0/\alpha_{1/2}(\bar{r}_h)$	$1+3\xi^2/40+O(\xi^4)$	1.01152	1.04554	1.1737	1.3604	1.571	1.911	2.0
$\hat{\alpha}_0/\hat{\alpha}_{1/2}$	$1+\xi^2/24+O(\xi^4)$	1.00643	1.02581	1.1042	1.2358	1.414	1.811	2.0
$\beta/i\xi$	$1-\xi^4/1080+O(\xi^6)$	0.99998	0.99964	0.9941	0.9691	0.900	0.544	0.0

Petrov–Galerkin Methods

These methods, where the trial and test functions are different, can be used in several ways. In [19] it is shown how to derive accurate schemes for the fully discrete approximation to (2.1) by this method. It is also possible to obtain asymptotically more accurate approximations to the semi-discrete problem this way. Two such techniques are discussed, one developed by Dendy [10] and the other using splines.

A Petrov–Galerkin approximation to (2.7) can be written for test functions η_n and trial functions θ_j in the form

$$\int (\Sigma(u_j)_t\,\theta_j - L(\Sigma u_j\theta_j))\,\eta_n\,dx = 0 \qquad \text{for all } m. \tag{4.2}$$

It is clear that if we choose $\eta_m = \phi_m$, where ϕ_m is an eigenfunction of L, and L is linear, then there will be no truncation error in (4.2) because we can integrate by parts and set $L\eta_m = \lambda_m\eta_m$. Such an algorithm will not be local. However, this suggests that the error can be reduced by using local test functions η_m which combine to give good approximations to eigenfunctions of L. For $L = \partial/\partial x$, the work described in [4] and [28] depends on splines giving such good approximations. A spline Petrov–Galerkin method can be defined for $u_t = u_x$. In order to construct a scheme which is as local as the standard Galerkin scheme (3.2), use as test functions:

(a) Linear splines, zero at all endpoint nodes, unity at one midpoint node.

(b) Quadratic splines, zero at all midpoint nodes, unity at one endpoint node.

The resulting Petrov–Galerkin scheme is

$$\begin{aligned}(u_t)_j + 4(u_t)_{j+1/2} + (u_t)_{j+1} &= 6(u_{j+1} - u_j)/h\\ -((u_t)_{j-1} + (u_t)_{j+1}) + 8(u_t)_{j-1/2} + (u_t)_{j+1/2}) + 18(u_t)_j &\\ = (20(u_{j+1/2} - u_{j-1/2}) + 2(u_{j+1} - u_{j-1}))/h.&\end{aligned} \tag{4.3}$$

This scheme is very similar to (4.1) and is also fourth order accurate. As with (4.1), the heavy averaging on the left side of the second equation leads to poor results for large values of ξ.

Another Petrov–Galerkin scheme is the second scheme of Dendy [10]. For test functions θ_n it can be written

$$\int \Sigma((u_t)_n \theta_n - u_n(\theta_n)_x)(\theta_n - (\theta_n)\, x)\, dx = 0. \tag{4.4}$$

The form of this scheme depends on the grid length h. It is therefore difficult to analyse in terms of $\xi = kh$ as has been done for other schemes. It is simpler to understand (4.4) with linear elements. On a regular mesh we obtain

$$\begin{aligned}(u_t)_{j-1} + 4(u_t)_j + (u_t)_{j+1} - 3((u_t)_{j+1} - (u_t)_{j-1})/h \\ = 3(u_{j+1} - u_{j-1})/h + 6(u_{j+1} - 2u_j + u_{j-1})/h^2.\end{aligned} \tag{4.5}$$

Substitute the shortest resolvable wave, $u = \exp(i\pi x/h)$ into (4.5). Then $u_j = -u_{j+1}$ and

$$(u_t)_j = -12u_j/h^2. \tag{4.6}$$

Thus instead of obtaining a stationary solution, $(u_t)_j = 0$, which would be given by the conventional Galerkin method, the solution is damped. Thus the scheme cannot readily be analysed in terms of travelling waves like the others considered earlier. Substituting a general wave $u = e^{ikx}$ into (4.5) gives

$$(u_t)_j = i\beta(\xi)\, e^{ij\xi},$$

where

$$\beta(2 + \cos\xi + 3\sin\xi/h) = (3h\sin\xi + 6(\cos\xi - 1))/h^2.$$

As $\xi \to 0$

$$\begin{aligned}(1 - \xi^2/6 + O(\xi^4))(1 + \xi/h)\beta = (\xi^2/6)\,\xi/h \\ + (1 - \xi^2/12)\,\xi^2/h^2 + O(\xi^6).\end{aligned} \tag{4.7}$$

If we take $h \to 0$ for fixed k in (4.7) the scheme is second order accurate, but if we take $k \to 0$ for fixed h it is third order. A similar analysis could be carried out for the quadratic element.

Petrov–Galerkin methods can also be used to give "upwind" advection schemes. These are also dissipative and not analysed here; detailed descriptions are given in [5] and elsewhere. Another "characteristic" type method using the finite element interpolation is described in [22].

5. Computational Results

Some of the different approximations to $Lu = u_x$ discussed above are tested using the same numerical example as was used in [8]. The equation $u_t + uu_x = 0$ is solved on the interval $[-1, 1]$ with periodic boundary conditions and initial data $u(x, 0) = \cos^2 \frac{1}{2}\pi x$. The exact solution is calculated at $t = \frac{1}{2}$. Most of the solutions given are for initial data calculated using the least squares restriction r_h^q, and verified against $r_h^q u(x, \frac{1}{2})$. In each case the algorithm is a two stage method, with different approximations to u_x combined with the Galerkin scheme (3.12) for the product. Two other solutions were obtained, one using the Gauss-point collocation restriction r_h^c throughout, and the other using a pointwise representation with the finite difference approximation (3.7) for u_x. In the latter case the finite element interpolation as used to derive verification solutions at points other than the grid points. For comparison a solution using linear elements with the same number of nodes as the quadratic element and a least squares representation of the data is also given. This used the two stage Galerkin scheme for the nonlinear term. All solutions are obtained by integrating with small time-steps and extrapolating to $\Delta t = 0$.

Consider first the results in Table VI. Using scheme (3.2) the results are good for $h = 1/3$, deteriorate at $h = 1/6$, and improve greatly for $h = 1/12$. This suggests that $h = 1/3$ is well short of adequate resolution and apparently good results may be obtained by accident. In the smoother part of the solution, for x between -1 and 0, the solution improves steadily from $h = 1/3$ to $h = 1/12$. At $t = 0$ there is a single wave $k = \pi$, so $\xi = \pi/3$ for $h = 1/3$. At $t = \frac{1}{2}$ the locally dominant k varies from about $\pi/2$ to 3π, giving $\xi = \pi$ for $h = 1/3$. The results using the extra projection (3.4) are similar for $h = 1/3$ and $1/6$. For $h = 1/12$ they are better in the smooth region ($\xi \sim \pi/24$) and worse where the wave is steepening ($\xi \sim \pi/4$). Figure 2 suggests that the change in relative behavior should occur for $\xi = \pi/2$. Thus the behavior of the schemes seems to be determined by the locally dominant wave in u_t rather than in u.

As would be expected, the pure finite difference operator (3.7) cannot follow the best fit restrictions r_h^q and large errors result. When this scheme is used with r_h^g the results are very accurate in the smooth region, but worse than the results using (3.2) where the wave is steepening (Table VII). When (3.7) is combined with the extra projection (3.6) the results are very good in the smooth region but worse than either result using (3.2) where the wave is steepening. This is consistent with the large errors shown in Fig. 2 for large ξ.

The results using linear elements are better than those using quadratics in the smooth region and worse elsewhere. This is quite consistent with the structure of the error, since there is no eigenfunction error with linear elements but the eigenvalue error of (3.7) is greater than that of (3.2) and the errors using the Galerkin product with linear and quadratic elements also have different structures (see analysis in [8] and Eq. (3.13)).

The results using the collocation scheme (4.1) are generally worse than those using (3.2). This is consistent with the error analysis. When the collocation scheme is used throughout, the calculation is almost unstable for the highest resolution. This is

TABLE VI A

Numerical Results for the Advection Equation, with $h = 1/3$, $1/6$ and $1/12$, and Various Approximations to u_x

	x	-1	$-13/24$	$-1/4$	$1/24$	$1/2$	$17/24$	$3/4$	$19/24$
$u(x,\frac{1}{2})$	h	0.0	0.25	0.5	0.75	1.0	0.75	0.5	0.25
$r_h^q u(x,\frac{1}{2})$	1/3	0.0087	0.2504	0.5007	0.7506	0.9951	0.6821	0.4670	0.2916
	1/6	−0.0037	0.2499	0.5000	0.7501	1.0037	0.7071	0.5000	0.2929
	1/12	0.0003	0.2500	0.5000	0.7500	0.9997	0.7495	0.5000	0.2505
Errors × 10^4	1/3	119	36	30	−49	−50	129	−26	−130
Scheme (3.2)	1/6	24	7	−9	−8	343	27	−153	−113
	1/12	−1	−2	8	−6	23	14	−20	10
Scheme (3.2)	1/3	126	31	51	−104	123	155	−3	−111
with (3.5)	1/6	−48	−1	4	−3	292	7	−146	−104
	1/12	−1	0	−1	0	−2	34	−39	13
Scheme (3.7)	1/3	137	3	35	−53	419	290	15	104
with (3.6)	1/6	−42	0	0	3	124	19	−251	−200
	1/12	−1	0	0	0	21	18	−75	31
Scheme (3.7)	1/3	94	18	84	73	516	−661	−514	−380
	1/6	−38	0	2	15	3	−84	−156	−56
	1/12	−1	0	0	1	8	138	−119	−15
Scheme (4.1)	1/3	−16	−14	−318	378	−316	4	0	−4
	1/6	9	−11	−6	−3	−50	−20	148	150
	1/12	−1	5	10	−8	31	−43	114	−31
Linear Galerkin	1/6	55	6	24	−47	482	−93	−122	−152
errors	1/12	10	0	0	3	62	28	−344	−107
	1/24	0	0	0	0	12	75	−82	34

Note. All schemes use r_h^q and the Galerkin product (3.12).

TABLE VI B

Numerical Results for the Advection Equation Using Gauss-Point Collocation Throughout

	x	-1	$-13/24$	$-1/4$	$1/24$	$1/2$	$17/24$	$3/4$	$19/24$
	h								
$r_h^c(x,\frac{1}{2})$	1/3	−0.0466	0.2581	0.4994	0.7460	0.9882	0.7251	0.5330	0.3682
	1/6	−0.0655	0.2569	0.5009	0.7436	1.0617	0.7357	0.5010	0.2648
	1/12	0.0051	0.2506	0.5001	0.7493	0.9943	0.7504	0.4994	0.2501
Error × 10^4	1/3	−19	336	−296	191	−149	−443	−420	−385
Scheme (4.1)	1/6	166	−81	−125	82	−624	−135	180	367
	1/12	32	30	−21	11	−39	−79	213	−109

because of the excessive averaging on the left side of (4.1), also reflected in the large positive eigenvalue errors in Table IV.

Overall this computation gives a reasonable confirmation of the error analysis. In this kind of integration the total error is dominated by the T.E. and the error growth

TABLE VII

Numerical Results for the Advection Equation Using Compact Finite Differences

	x	-1	$-13/24$	$-1/4$	$1/24$	$1/2$	$17/24$	$3/4$	$19/24$
$r_h^g u(x, \frac{1}{2})$	h	0.0	0.25	0.5	0.75	1.0	0.75	0.5	0.25
Errors $\times 10^4$	1/3	0	4	35	51	334	−128	−15	67
Scheme (3.7)	1/6	0	0	0	5	36	−55	−333	−208
	1/12	0	0	0	0	7	55	−56	14

term in Eq. (2.8) is not dominant. Under these circumstances the quadratic element on a regular mesh performs quite well compared to the linear element, presumably other high order elements would also do so. For long computations, e.g., meteorological problems, the error growth term is important and the additional errors in smooth regions induced by higher order elements would be damaging. These can be alleviated by a projection such as (3.5). In a linear problem it would be sufficient to apply this to the initial data as in [15], but in a nonlinear problem it would have to be applied in the algorithm at each step.

This analysis has shown how quadratic finite elements can be used effectively in regular geometry. In the next section the problems with irregular geometry are analysed.

6. Analysis of Advection on an Irregular Mesh

It is likely that finite element methods are most advantageous in practice in problems where the mesh has to be irregular and use of spectral methods requires piecewise coordinate transformations [21]. However, it is known that finite element methods for advective problems on an irregular mesh can give apparently poor results (e.g., [7]) and the asymptotic order of accuracy is low [16]. The Galerkin method leads to a conservative scheme which cannot therefore handle the propagation of a wave from a fine grid to a coarse grid which cannot resolve it. An "upwind" method may give a smoother but less accurate solution. The results of the earlier part of this paper suggests that the Galerkin method may give an accurate scheme for advection plus distortion errors which could be eliminated by extra projections, and that the low asymptotic order of accuracy is caused by the absence of a smoothness assumption in the formulation.

To investigate this possibility we carry out an error analysis away from the asymptotic limit. Two schemes are then applied to the computational example of Section 5. The error analysis is a calculation of the truncation error term $(r_h L - L_h r_h) u$ of Eq. (2.8). A fixed irregular grid is analysed. The limit $h \to 0$ is difficult to define for problems using irregular grids since it is not clear whether the ratio of adjacent grid lengths should be $O(h)$ or $O(1)$. Fourier analysis can be used if the irregular grid is extended periodically, but the asymptotic limit for long waves

cannot be taken as the periodicity is fixed by the dimensions of the problem. In this calculation a one-dimensional irregular grid is analysed by extending it periodically and calculating the truncation error for $u = \exp(2in\pi x/D)$, where D is the periodicity. The error must be calculated by computer by carrying out an integration of $u_t = u_x$ with small time-steps and extrapolating to $\Delta t = 0$. Because the solution distorts as it moves through the mesh the integration was carried out over a time interval D, after which the exact solution is identical to the initial data.

The grid was chosen to give a uniform representation of the solution of the example of Section 5 at $t = \frac{1}{2}$. This is the strategy used successfully in steady state problems. The endpoints of the quadratic elements were therefore at $x = \{-1, -13/24, -1/4, 1/24, 1/2, 17/24, 3/4, 19/24\}/2D$. The maximum ratio of element size is 11 to 1. The results are shown in Table VIII, for unit wave amplitude. For wave numbers greater than 3 the distortion becomes too great after time D for the speed to be calculated. The speed error is calculated from an average of the errors in values of x at which $u = 0$.

TABLE VIII

Truncation Error Calculation for Irregular Grid

Wavenumber n	1	2	3
Total L_2 error	0.00724	0.0955	0.2340
Speed error (%)	+0.037	+0.80	−1.32

The asymptotic rate of convergence of the L_2 error should be $O(h^2)$. This roughly agrees with Table VIII. For wavenumber 3 the L_2 error is 23% of the wave amplitude, while the error in wave speed is only 1%. Thus the existence of an accurate solution with added noise is confirmed, though it can no longer be easily expressed as a difference in order of accuracy.

We now illustrate the performance of the scheme on the computational example of Section 5. Two solutions were obtained and compared with $r_h^q u(x, \frac{1}{2})$. One was obtained by a two stage Galerkin approximation to $u_t = uu_x$ on the irregular mesh. The other was obtained by writing

$$u_x = u_\xi \xi_x, \tag{6.1}$$

where ξ is a new coordinate in which the mesh is uniform. ξ is a piecewise linear function of x. A direct differentiation of ξ gives a piecewise constant ξ_x, which when substituted into the Galerkin approximation to (6.1), leads to the standard algorithm. If, instead, ξ_x is approximated by a continous function by using $p_h^q r_h^q$, and then the product (6.1) evaluated, a non-conservative scheme results which can handle the transfer of information from fine to coarse mesh. Another method using (6.1) would be to form an algorithm to seek the best least squares fit to u integrated over ξ instead of x. This will be a different solution which may or may not be preferable. The solutions using the first two methods are shown in Table IX and Figs. 5 and 6.

TABLE IX

Solution of Advection Equation on Irregular Grid

x	-1	$-13/24$	$-1/4$	$1/24$	$1/2$	$17/24$	$3/4$	$19/24$
$r_h^q u(x, \frac{1}{2})$	−0.0013	0.2539	0.5000	0.7461	1.0013	0.7718	0.5000	0.2282
Error $\times 10^4$								
Galerkin	−55	−124	−174	−245	−408	−506	−162	−123
Non-conservative	−55	−82	89	−88	−4	90	375	287

The Galerkin scheme gives an accurate solution with, added to it, a large oscillation between endpoints and midpoints. The accurate solution could be obtained by post-processing, assuming the true solution to be smooth. This is best achieved by filtering $\int u_h \theta_n$, where θ_n are the normalized basis functions because $\int u\theta_n$ will vary smoothly for a smooth u, while the coefficients in $r_h^q u$ will not.

The second integration shows reduced errors, except between $x = 35/48$ and 1. However, the structure of the error is less well defined, it is not obvious how to recover a solution by post-processing, and there is clearly a lag in the wave speed. This is confirmed by the error analysis calculation for the second scheme. The extra nonlinearity introduced made it impossible to calculate the errors as shown in Table VIII, because the integration could not be continued to $t = D$. Thus truncation error is much larger in this scheme.

These results illustrate again that the finite element Galerkin algorithm contains an accurate scheme that can be reached by imposing extra smoothness requirements. However, they also emphasize that an irregular grid should not be used for this sort of problem, at least with an Eulerian scheme. This is true even when the grid is

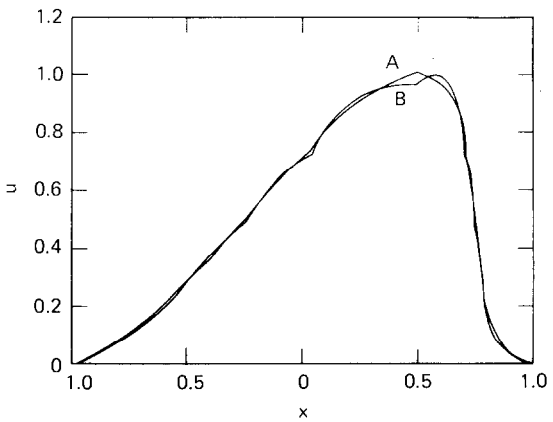

FIG. 5. Results using irregular grid for transport equation: (A) Least squares fit to exact solution. (B) Galerkin solution.

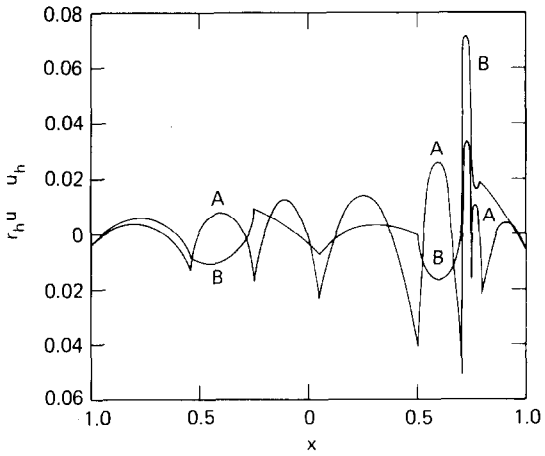

FIG. 6. Errors using irregular grid for transport equation: (A) Error in Galerkin solution. (B) Error in solution using non-conservative scheme.

designed to fit the exact solution as in this example. The errors tabulated here are greater than those for the Galerkin integrations with a regular grid, even the one with fewer nodes. If the mesh is allowed to move, as in [30], the situation is totally different and very accurate results can be obtained, at least in one dimension.

7. CONCLUSIONS

These results show a reasonable consistency between the behavior of the computed solutions and the error analysis. Quadratic finite element methods on regular and irregular grids are shown to contain accurate solutions with additional small scale errors. The latter errors only decrease slowly as the mesh size is reduced, leading to a low rate of convergence for smooth data. This is to be expected because the finite element Galerkin method attempts to optimise the solution for general data, not just smooth data. The small errors demonstrated here for data with significant variation on the grid scale confirms this.

In many applications this type of error structure is a disadvantage, because the small scale errors must be filtered out and the detailed information contained in the finite element solution will be lost. Thus for applications with little viscosity and long time integrations needed, only those finite element methods which satisfy special superconvergence properties are useful. In practice this means linear elements on a smoothly varying grid; because higher order splines are often inconvenient to use because of their non-local basis set. Under these restrictions it is difficult to see the advantage of finite elements over cheaper explicit finite differences, or spectral

methods. In the meteorological application it is found that all these methods give similar answers for similar work.

In transient applications where the viscosity is sufficient to control the noise, the analysis here suggests that the finite element Galerkin algorithm can be used safely and should be better than the modified algorithms discussed here. If time integration errors are important, the Petrov–Galerkin schemes discussed in [19] will be more appropriate. The decision as to whether the noise is damaging will have to be made separately for each problem, because it depends on the degree of nonlinearity and the flow normal to the axis of mesh refinement, as well as on the viscosity.

ACKNOWLEDGMENTS

This work was carried out while the author was visiting the Courant Institute of Mathematical Sciences, New York, the Department of Mathematics, University of California, Berkeley, and the Atmospheric Physics Division, Lawrence Livermore Laboratory.

REFERENCES

1. Y. ADAM, *J. Comput. Phys.* **24** (1977), 10–22.
2. J. P. AUBIN, "Approximation of Elliptic Boundary Value Problems," Wiley, New York, 1972.
3. W. BOURKE, B. MCAVANEY, K. PURI, AND P. THURLING, *Methods Comput. Phys.* **17** (1977), 267.
4. J. H. BRAMBLE AND A. H. SCHATZ, *Math. Comp.* **31** (1977), 94–111.
5. I. CHRISTIE, D. F. GRIFFITHS, A. R. MITCHELL, AND O. C. ZIENKIEWICZ, *Internat. J. Num. Methods Engrg.* **10** (1976), 1389–1396.
6. L. COLLATZ, "Numerical Treatment of Differential Equations," 3rd ed., Springer-Verlag, New York/Berlin, 1964.
7. M. J. P. CULLEN, *Quart. J. Roy. Meteor. Soc.* **102** (1976), 77–93.
8. M. J. P. CULLEN AND K. W. MORTON, *J. Comput. Phys.* **34** (1980), 245–267.
9. C. DE BOOR AND B. SWARTZ, *SIAM J. Numer. Anal.* **10** (1973), 582–606.
10. J. E. DENDY, JR., *SIAM J. Numer. Anal.* **11** (1974), 637–653.
11. J. DOUGLAS, JR., AND T. DUPONT, *Math. Comp.* **27** (1973), 17–28.
12. T. DUPONT, *SIAM J. Numer. Anal.* **10** (1973), 890–899.
13. G. J. FIX AND N. NASSIF, *Numer. Math.* **19** (1972), 127–135.
14. B. FORNBERG AND G. B. WHITHAM, *Philos. Trans. Roy. Soc. London Ser. A* **289** (1978), 373.
15. G. W. HEDSTROM, *SIAM J. Numer. Anal.* **16** (1979), 385–393.
16. P. LESAINT, *Numer. Math.* **21** (1973), 244–255.
17. M. LUSKIN, *Math. Comp.* **33** (1979), 493–521.
18. K. W. MORTON AND M. J. LONG, *J. Inst. Math. Appl.* **19** (1977), 307–323.
19. K. W. MORTON AND A. K. PARROTT, *J. Comput. Phys.* **36** (1980), 249–270.
20. S. A. ORSZAG, *J. Fluid Mech.* **49** (1971), 75–113.
21. S. A. ORSZAG, *J. Comput. Phys.* **37** (1980), 70–92.
22. P. A. RAVIART, "Lecture Notes in Physics No. 91," pp. 27–41, Springer-Verlag, Berlin/New York, 1979.
23. T. J. RIVLIN, C. A. MICCHELLI, AND S. WINOGRAD, *Numer. Math.* **26** (1976), 191–200.
24. H. SCHAMEL AND K. ELSAESSER, *J. Comput. Phys.* **22** (1976), 501–516.
25. A. STANIFORTH AND R. W. DALEY, *Monthly Weather Rev.* **107** (1979), 107–121.
26. G. STRANG AND G. J. FIX, "An Analysis of the Finite Element Method," Prentice–Hall, Englewood Cliffs, N. J., 1973.

27. V. THOMEE, *Math. Comp.* **31** (1977), 652–660.
28. V. THOMEE AND B. WENDROFF, *SIAM J. Numer. Anal.* **11** (1974), 1059–1068.
29. P. M. GRESHO, R. L. LEE, AND R. L. SANI, *in* "Finite Elements in Fluids," Vol. 3, Wiley, New York, 1977.
30. K. MILLER, *SIAM J. Numer. Anal.* **18** (1981), 1033–1057.