JOURNAL OF COMPUTATIONAL PHYSICS 36, 249–270 (1980)

Generalised Galerkin Methods for First-Order Hyperbolic Equations

K. W. MORTON AND A. K. PARROTT*

Department of Mathematics, University of Reading, Whiteknights, Reading, Berkshire, England RG6 2AX

Received April 16, 1979; revised September 24, 1979

The Petrov–Galerkin technique is used to derive finite-element methods for hyperbolic equations in one space dimension. Special test functions are developed to ensure that the "unit CFL property" is satisfied, first for scalar equations and then for linear first-order systems. With linear elements and Euler, leap-frog, or Crank–Nicolson time-stepping, several interesting relations with standard finite-difference methods emerge and the schemes exhibit particularly high phase accuracy on coarse grids as well as improved stability properties compared with Galerkin methods.

1. INTRODUCTION

During the last several years considerable interest has developed in the potential advantages of finite-element methods for the solution of initial-value problems, firstly for parabolic equations and more recently for hyperbolic: individual references are too numerous to list but see the recent reviews by Mitchell [7] and Morton [9]. The optimal global approximation properties which are crucial to the success of these methods for self-adjoint elliptic problems may not at first sight seem so relevant to initial-value problems, particularly in the hyperbolic case: there the phenomena are more local and the Taylor series expansion which lies at the heart of finite-difference methods may be more appropriate. However, the high accuracy attainable with spline-Galerkin methods in a semidiscrete formulation is now well established and is very encouraging–Cullen [1], Cullen and Morton [2], Swartz and Wendroff [12, 13], Thomée and Wendroff [14]. The spatial approximation provided by finite-element methods can be very satisfactory therefore, but the difficulty comes in coupling this to the time discretisation. A direct application of the Galerkin principle in space-time can lead to a very inefficient algorithm and poor convergence; see, for example, Lesaint [4]. In this paper we combine finite-difference approximations in time with finite-elements in space (one dimension) in a formulation which leads to schemes of high accuracy right up to unit CFL number.

* Now at British Gas, Mathematics Division, London Research Station, Michael Road, London SW6 2AD.

0021-9991/80/080249-22$02.00/0

In the semidiscrete Galerkin formulation, an approximation to the equation $u_t = Lu$ is sought of the form

$$U(x, t) = \sum_j U_j(t)\, \phi_j(x) \tag{1.1}$$

by setting

$$\langle U_t - LU, \phi_i \rangle = 0, \tag{1.2}$$

where $\langle u, v \rangle$ denotes the L_2 inner product $\int u\, v\, dx$ over the domain of the problem: that is, the "test" functions $\phi_i(x)$ in (1.2) are the same as the "trial" functions in (1.1). This can be justified or motivated as follows: if (1.1) represents a discrete approximation to $u(x, t)$ at time t, then the best least-squares approximation to LU of the same form $V = \sum V_j \phi_j(x)$ is given by

$$\underset{V_j}{\text{minimise}} \| V - LU \|^2, \tag{1.3}$$

where $\| u \|^2 = \langle u, u \rangle$. This should therefore be taken for U_t and gives Eq. (1.2). If the ϕ_j are linear basis functions on a uniform mesh of size h, these equations are accurate to $O(h^4)$ when L represents a first-order hyperbolic system, either linear or quasi-linear. But when the system of ordinary differential equations is approximated by finite differences in time, this fourth-order accuracy is very quickly eroded as Δt is increased. Indeed, with forward time differencing the Galerkin scheme becomes unstable unless $\Delta t = O(h^2)$.

Our approach in this paper is to use generalised Galerkin or Petrov–Galerkin methods in which the test functions are different from the trial functions and are chosen so that waves are propagated without distortion for the special case when characteristics pass through mesh points or nodes at each time level. As a simple example, consider the scalar convection equation in one dimension $u_t = au_x$ with $a > 0$. The Petrov–Galerkin equations with forward time-differencing take the form

$$\langle U^{n+1} - (U^n + a\Delta t\, \partial_x U^n), \psi_i \rangle = 0 \tag{1.4}$$

for some set of test functions $\psi_i(x)$. When $a\Delta t = h$ it should be possible for U^{n+1} to exactly represent the shift of U^n through one mesh-length: we call this the *unit CFL property*, which is satisfied by virtually all practical finite-difference schemes. It turns out that a simple choice of test function $\chi_i^+(x)$ will give this property. Then in general we will take $\psi_i = (1 - \mu)\, \phi_i + \mu \chi_i^+$, where $\mu = a\Delta t/h$, so as to match the high accuracy of Galerkin as $\mu \to 0$ with the new property at $\mu = 1$. The resulting scheme is closely related to the single-step Lax–Wendroff method, but a considerable improvement over it. Such relations between generalised Galerkin methods and finite difference methods have been noted previously by Mock [8] among others.

In the next section we give an analysis of the properties of the method in terms of phase errors and damping of Fourier modes. Then, as a first step towards extending this basic scheme to practical problems, we deal with variable coefficients and linear systems of equations, although still in one space dimension. The results of numerical experiments are also given for a variable coefficient problem and a standard reflection problem involving the wave equation. To conclude this section an initial attempt at dealing with the nonlinear equation $u_t + uu_x = 0$ is given. From this it is clear that one will not be able to reproduce the correct propagation speed of a shock with this scheme. The next section is therefore devoted to similar developments of methods based upon a theta time-stepping scheme and on a leap-frog scheme. Both approaches lead to interesting and accurate methods: the former gives, for $\theta = \frac{1}{2}$, methods which are intermediate between the Crank–Nicolson–Galerkin method and the Keller box scheme; the latter gives a scheme which is fourth order in space and time, has an improved stability range compared with leap-frog Galerkin, and has a particularly simple form.

2. Methods Based on Forward (or Euler) Time-Stepping

At time level $n\Delta t$ and on a uniform mesh of size h we seek approximations of the form

$$U^n(x) = \sum_{(j)} U_j{}^n \phi_j(x), \tag{2.1}$$

and we confine our studies to the case of linear basis functions $\phi_j(x) = \phi(x/h - j)$, where $\phi(s)$ is the usual "roof" function

$$\begin{aligned} \phi(s) &= 1 + s, \qquad s \in [-1, 0], \\ &= 1 - s, \qquad s \in [0, 1] \end{aligned} \tag{2.2}$$

and $\phi(s) = 0$ otherwise. We consider first the scalar convection problem

$$\begin{aligned} u_t &= au_x, \qquad a = \text{positive constant}, \\ u(x, 0) &= u_0(x), \end{aligned} \tag{2.3}$$

the exact solution of which is given by $u(x, t) = u_0(x + at)$.

2.1. *Derivation of Special Test Function*

The exact solution may be exactly reproduced by an Euler–Petrov–Galerkin method (1.4) when $u_0(x) = U^0(x)$ is of the form (2.1) and $a\Delta t = h$, provided that the test functions (denoted for this special case by $\chi_i{}^+(x)$) satisfy the relations

$$\sum_{(j)} U_{j+n}^0 \langle \phi_{j-1} - (\phi_j + h\, \partial_x \phi_j), \chi_i{}^+ \rangle = 0, \qquad \text{all } i. \tag{2.4}$$

This has to be true for all coefficients $U_j{}^0$ so that each of the inner products has to be zero. To simplify further, we introduce the characteristic function $\phi_j^0(x) = \phi^0(x/h - j)$ of the interval $[jh, (j+1)h]$, where

$$\phi^0(s) = 1, \qquad s \in [0, 1] \tag{2.5}$$

and $\phi^0(s) = 0$ elsewhere. Then it is clear that $\partial_x \phi_j = h^{-1}(\phi_{j-1}^0 - \phi_j^0)$ so that (2.4) reduces to

$$\langle \phi_j - \phi_j^0, \chi_i^+ \rangle = \text{const.}, \qquad \text{all } i, j; \tag{2.6}$$

and if χ_i^+ is to be localised near $x = ih$, it follows that the constant must be zero. Furthermore, $\phi_j - \phi_j^0$ is itself the difference between two triangular-shaped functions $r_{j-1}(x) - r_j(x)$, where $r_j(x) = x/h - j$ in $[jh, (j+1)h]$ and zero elsewhere. So for the same localisation reasons, we have

$$\langle r_j, \chi_i^+ \rangle = 0, \qquad \text{all } i, j. \tag{2.7}$$

If the χ_j^+ are all to be translates of a single basis function $\chi_j^+(x) = \chi(x/h - j)$, we obtain the defining relation

$$\int_0^1 s\chi(s)\, ds = 0. \tag{2.8}$$

The simplest such function, normalised to give $\int \chi\, ds = 1$, is

$$\chi(s) = 4 - 6s, \qquad s \in [0, 1], \tag{2.9}$$

with $\chi(s) = 0$ elsewhere: we shall use this in all of the following, with χ_j^+ defined in terms of it.

2.2. *The Euler–Petrov–Galerkin* (*EPG*) *Method*

The scheme which results from using test functions in (1.4) which are a linear combination of the functions (2.9) and the trial functions is as follows:

$$\langle U^{n+1} - (U^n + a\Delta t \partial_x U^n), (1 - \nu)\, \phi_i + \nu \chi_i^+ \rangle = 0, \qquad \text{all } i. \tag{2.10}$$

We call this the Euler–Petrov–Galerkin (or EPG) method and in the Appendix it is shown to be stable for

$$0 \leqslant a\Delta t/h \leqslant \nu \leqslant 1. \tag{2.11}$$

In terms of the nodal parameters U_j^n, it may be written as

$$[1 + \tfrac{1}{6}(1 - \nu)\, \delta^2](U_j^{n+1} - U_j^n) = \mu \Delta_0 U_j^n + \tfrac{1}{2}\mu\nu\delta^2 U_j^n, \tag{2.12}$$

where, using the usual notation, $\Delta_0 U_j \equiv \frac{1}{2}(U_{j+1} - U_{j-1})$ and $\delta^2 U_j \equiv U_{j+1} - 2U_j + U_{j-1}$ with $\mu = a\Delta t/h$. Substituting a Fourier mode e^{ikx} into (2.12) and setting $\xi = kh$ gives an amplification in one time step of

$$\begin{aligned}\lambda(\xi) &= 1 + [1 - \tfrac{2}{3}(1 - \nu)\sin^2 \tfrac{1}{2}\xi]^{-1}\,[i\mu \sin \xi - 2\mu\nu \sin^2 \tfrac{1}{2}\xi] \\ &\sim 1 + i\mu\xi - \tfrac{1}{2}\mu\nu\xi^2 - \tfrac{1}{6}i\mu\nu\xi^3 + \cdots, \qquad \text{as} \quad \xi \to 0,\end{aligned} \tag{2.13}$$

to be compared with $e^{i\mu\xi}$ for the differential equation. Thus the scheme is first-order accurate for $\nu > \mu$; but for $\nu = \mu$ it is second-order accurate and the right-hand side of (2.12) is the same as that which appears in the Lax–Wendroff method—see Richtmyer and Morton [10]. However, the presence of the mass matrix modifies the leading term of the truncation error from $\frac{1}{6}\mu(1 - \mu^2)\,\xi^2$ for Lax–Wendroff to $\frac{1}{6}\mu^2(1 - \mu)\,\xi^2$ here, a useful reduction by $\mu/(1 + \mu)$.

To treat a problem like (2.3) but with $a < 0$, it is merely necessary to reflect everything in the plane $x = 0$ and to introduce test functions χ_j^- still based on (2.9) but with

$$\chi_j^-(x) = \chi(j - x/h). \tag{2.14}$$

2.3. *Dissipation and Dispersion Errors*

It is particularly important in assessing finite-element methods to calculate the accuracy outside the asymptotic limits. So in Table I we give the phase error Δ and

TABLE I

Phase Errors Δ in Units of Wavelength, and Amplification Factors κ after One Oscillation, for Euler Time-stepping[a]

		Lax–Wendroff		Euler–Petrov–Galerkin		Euler–Galerkin		Euler–Petrov–Galerkin II	
μ	ξ	Δ	κ	Δ	κ	Δ	κ	Δ	κ
0.2	$\pi/4$	−0.0951	0.936	−0.0182	0.966	−0.0103	1.624	−0.0018	0.973
	$\pi/2$	−0.3462	0.675	−0.1080	0.724	−0.0722	2.367	−0.0358	0.755
	$3\pi/4$	−0.6803	0.462	−0.3877	0.303	−0.3271	1.944	−0.2537	0.272
0.55	$\pi/4$	−0.0663	0.879	−0.0243	0.923	−0.0579	3.29	0.0003	0.952
	$\pi/2$	−0.2271	0.436	−0.0882	0.528	−0.2015	6.15	0.0058	0.641
	$3\pi/4$	−0.4773	0.148	−0.1963	0.134	−0.4334	3.29	0.0428	0.191
0.9	$\pi/4$	−0.0163	0.948	−0.0073	0.965	−0.1310	5.02	0.0008	0.982
	$\pi/2$	−0.0360	0.715	−0.0107	0.797	−0.3398	7.96	0.0122	0.886
	$3\pi/4$	−0.0039	0.551	+0.0258	0.672	−0.5399	3.18	0.0505	0.812

[a] The fourth set of values is included here for convenience and is referred to later in the text, in Section 2.5.

amplification factor κ for Fourier modes corresponding to $\xi = \pi/4, \pi/2$, and $3\pi/4$ and $3\pi/4$ and for mesh ratios given by $\mu = 0.2$, 0.55, and 0.9: in each case, Δ and κ are calculated after one complete oscillation so that in terms of λ given by (2.13) we have

$$\kappa = |\lambda|^n, \; \Delta = n(\arg\lambda - \mu\xi)/2\pi, \qquad \text{where } n = 2\pi/\mu\xi. \tag{2.15}$$

These are all for the second-order scheme with $\nu = \mu$. Comparison is made with both Lax–Wendroff and Euler–Galerkin in the table. There is clearly a valuable improvement compared with Lax–Wendroff, especially for small ξ and μ and in phase speeds. The Euler–Galerkin scheme is of course unstable for any fixed μ, giving $\kappa > 1$ in the table. But apart from restoring stability, the new choice of test functions has almost as good a phase accuracy as Euler–Galerkin at $\mu = 0.2$ and retains this high accuracy over the whole stable range of μ. This was the main aim of the construction. These conclusions are confirmed by numerical experiments convecting a Gaussian profile, typical results of which are given in Fig. 1.

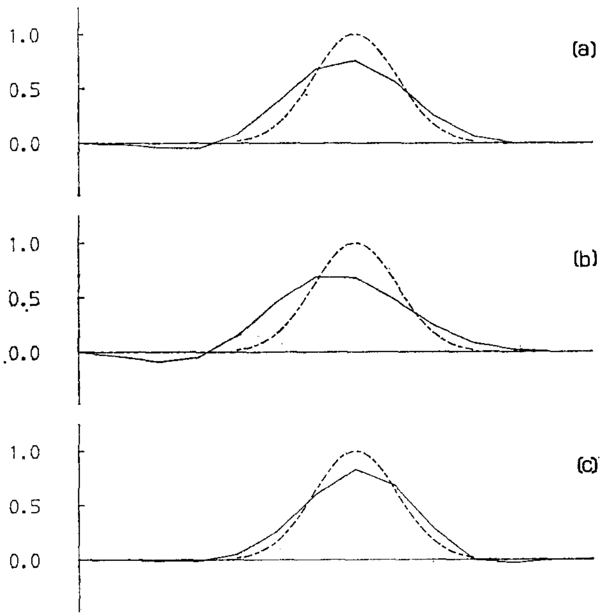

FIG. 1. Solutions of $u_t = au_x$ by (a) EPG, (b) Lax–Wendroff, (c) EPGII.

2.4. *Variable Coefficient Problems*

The choice of $\nu = \mu$ for the scalar convection equation has the more general interpretation of choosing ν to be the ratio of the time-step Δt to the time Δt_c for a characteristic to travel one mesh length. For the variable coefficient equation $u_t = a(x)\, u_x$ or $u_t = \partial_x[a(x)\, u]$, Δt_c will differ for each mesh interval so it follows that ν should

also vary. To aim for second-order accuracy with a smooth $a(x)$, it would seem to be adequate to put $\Delta t_c \simeq h/a(x_i)$ and so we take $\nu_i = a(x_i)\,\Delta t/h$. (The complete achievement of second-order accuracy will depend on the form of the equation and the quadrature rules used as well as the choice of ν_i.)
The Euler–Petrov–Galerkin equations are then

$$\langle U^{n+1} - (U^n + \Delta t\, \partial_x[a(x)\, U^n]), (1 - \nu_i)\, \phi_i + \nu_i \chi_i^+ \rangle = 0, \qquad \text{all } i, \tag{2.16}$$

leading to difference equations for the nodal parameters which again resemble Lax–Wendroff although the schemes differ now by more than just the mass matrix. The EPG scheme has been tested on the problem

$$u_t + \partial_x \left(\frac{u}{1 + 2x}\right) = 0, \qquad x \in [0, \pi], \tag{2.17}$$

with the initial-boundary data

$$\begin{aligned} u(0, t) = 0, \qquad u(x, 0) &= (1 + 2x) \sin 9x, \qquad && x \in [0, \pi/3], \\ &= 0, && x \in [\pi/3, \pi]. \end{aligned} \tag{2.18}$$

The true solution is

$$\begin{aligned} u(x, t) &= (1 + 2x) \sin 9\{[x^2 + x - t + \tfrac{1}{4}]^{1/2} - \tfrac{1}{2}\} \\ &\text{for} \quad 0 \leqslant x^2 + x - t \leqslant [\pi/3(1 + \pi/3)]. \end{aligned} \tag{2.19}$$

and zero elsewhere: approximate solutions using Lax–Wendroff and the EPG scheme are displayed in Fig. 2. They show that a clear improvement in phase error and dissipation is still obtained by this scheme despite the approximation used for ν.

2.5. *Linear Hyperbolic Systems*

To be of any practical use the properties of the EPG methods must be taken over to systems of equations, and this can be done, in part at least, for linear symmetric hyperbolic systems. Consider the equations

$$\mathbf{u}_t = A\mathbf{u}_x, \tag{2.20}$$

where A is a real symmetric $m \times m$ matrix and $\mathbf{u} = [u_1, u_2, \ldots, u_m]^{\mathrm{T}}$. It is straightforward to extend (2.10) to system (2.20) when A is diagonal, so suppose the diagonalised form of (2.20) is

$$\mathbf{v}_t = \Lambda \mathbf{v}_x, \tag{2.21}$$

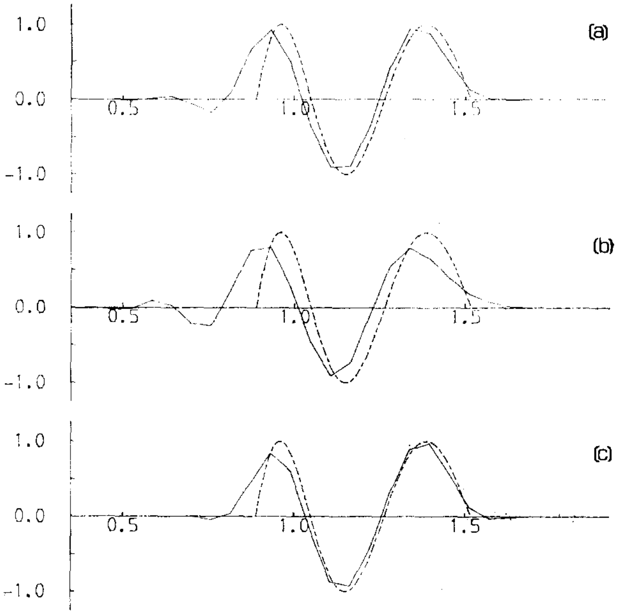

FIG. 2. Solutions of Eqs. (2.17) and (2.18) by (a) EPG, (b) Lax-Wendroff, (c) EPGII. The quantity plotted is $u/(1+2x)$ at $t = 1.6755$.

where

$$\mathbf{v} = S^{\mathrm{T}}\mathbf{u} \qquad \text{and} \qquad \Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_m \end{bmatrix} = S^{\mathrm{T}}AS; \tag{2.22}$$

the orthogonal matrix S is composed of the orthonormalised eigenvectors of A (assumed to be linearly independent). Since the signs of the eigenvalues are not known we write the generalization of (2.10) to the rth component of (2.21) in the form

$$\psi_{i(r)} = (1 - |\mu_r|)\,\phi_i + \tfrac{1}{2}|\mu_r|(\chi_i^+ + \chi_i^-) + \tfrac{1}{2}\mu_r(\chi_i^+ - \chi_i^-) \qquad \text{for all } i, \tag{2.23}$$

where $\mu_r = \lambda_r\,\Delta t/h$. This rather complicated expression can be simplified by introducing σ_i and τ_i, where

$$\sigma_i = \tfrac{1}{2}(\chi_i^+ + \chi_i^-) - \phi_i\,,\ \tau_i = \tfrac{1}{2}(\chi_i^+ - \chi_i^-), \qquad \text{for all } i. \tag{2.24}$$

Then for $V_{(r)} = \sum V_{j(r)}\phi_i$, a piecewise linear approximation to the rth component of $\mathbf{v}$, we have

$$\langle V_{(r)}^{n+1} - (V_{(r)}^{n} + \lambda_r\,\Delta t\,\partial_x V_{(r)}^{n}),\ \phi_i + (\Delta t/h)[|\lambda_r|\,\sigma_i + \lambda_r\tau_i]\rangle = 0, \quad \text{for all } i. \tag{2.25}$$

Let $\mathbf{V} = \sum \mathbf{V}_j \phi_j$, where $\mathbf{V}_j = [V_{j(1)}, V_{j(2)}, ..., V_{j(m)}]$, and let $\mathbf{e}_{(r)}$ denote the rth unit vector of $\mathbb{R}^m$. Then the approximation to the system can be written compactly as

$$\langle \mathbf{V}^{(n+1)} - (\mathbf{V}^n + \Delta t\, \Lambda \partial_x \mathbf{V}^n), \{\phi_i + (\Delta t/h)[|\Lambda| \sigma_i + \Lambda \tau_i]\} \mathbf{e}_{(r)} \rangle = 0 \quad \text{for } r = 1,..., m \text{ and for all } i. \tag{2.26}$$

We can now transform (2.26) into an approximation of (2.20) by using the fact that $\langle \mathbf{a}, \mathbf{b} \rangle = \langle S\mathbf{a}, S\mathbf{b} \rangle$ and defining $\mathbf{U} = S\mathbf{V}$; we obtain

$$\langle \mathbf{U}^{n+1} - (\mathbf{U}^n + \Delta t\, A \partial_x \mathbf{U}^n), \{\phi_i + (\Delta t/h)[S |\Lambda| S^{\mathrm{T}} \sigma_i + A \tau_i]\} S\mathbf{e}_{(r)} \rangle = 0, \quad r = 1,..., m \text{ and for all } i. \tag{2.27}$$

Since S is nonsingular we can replace $S\mathbf{e}_{(r)}$ by $\mathbf{e}_{(r)}$ in (2.27) and it can easily be shown that

$$\langle \phi_j, \tau_i \rangle = 0 \text{ and } \langle \partial_x \phi_i, \sigma_j \rangle = 0, \qquad \text{for all } i \text{ and } j. \tag{2.28}$$

Thus (2.27) simplifies to

$$\langle \mathbf{U}^{n+1} - (\mathbf{U}^n + \Delta t\, A \partial_x \mathbf{U}^n), \phi_i \mathbf{e}_{(r)} \rangle + (\Delta t/h) \langle \mathbf{U}^{n+1} - \mathbf{U}^n, \sigma_i \tilde{A} \mathbf{e}_{(r)} \rangle = (\Delta t^2/h) \langle A \partial_x \mathbf{U}^n, \tau_i A \mathbf{e}_{(r)} \rangle, \qquad r = 1,..., m, \text{ all } i, \tag{2.29}$$

where $\tilde{A} = S |\Lambda| S^{\mathrm{T}}$.

The structure of this approximation to (2.20) is quite illuminating: the first inner product is just the Euler–Galerkin approximation; the second inner product acts to increase the diagonal dominance of the Galerkin mass matrix since $\tilde{A}$ is positive definite, its diagonal entries are necessarily positive, and

$$\langle \phi_j, \sigma_i \rangle = h\delta_{ij} - \langle \phi_i, \phi_j \rangle; \tag{2.30}$$

finally, the third inner product is the appropriate second-order correction term to an Euler time difference, as in the one-step form of the Lax–Wendroff method. However, there is the disadvantage that $\tilde{A}$ is not in general diagonal and so there will be some coupling between components of $\mathbf{U}^{n+1}$, and also there may be considerable effort involved in calculating $\tilde{A}$. For both reasons it may be desirable to approximate the second term in practical computations.

The full equations (2.29) are a direct extension of the Euler–Petrov–Galerkin approximation to the scalar equation, so they retain its good phase and dissipation properties. We have applied the method to the wave equation

$$u_t = v_x, \; v_t = u_x \tag{2.31}$$

(v now refers to the 2nd component) for which $\tilde{A} = I$ and $A^2 = I$. The EPG equations are now much simpler namely,

$$\langle \mathbf{U}^{n+1} - (\mathbf{U}^n + \Delta t\, A\partial_x \mathbf{U}^n), \phi_i e_{(r)}\rangle + (\Delta t/h)\langle \mathbf{U}^{n+1} - \mathbf{U}^{n-1}, \sigma_i \mathbf{e}_{(r)}\rangle = (\Delta t^2/h)\langle \partial_x U^n, \tau_i \mathbf{e}_{(r)}\rangle, \qquad r = 1, 2, \text{ for all } i, \tag{2.32}$$

and the equations for the nodal parameters become, with $\mu = \Delta t/h$,

$$\begin{aligned} [1 + \tfrac{1}{6}(1 - \mu)\,\delta^2](U_i^{n+1} - U_i^n) &= \mu\,\Delta_0 V_i^n + \tfrac{1}{2}\mu^2\,\delta^2 U_i^n, \\ [1 + \tfrac{1}{6}(1 - \mu)\,\delta^2](V_i^{n+1} - V_i^n) &= \mu\,\Delta_0 U_i^n + \tfrac{1}{2}\mu^2\,\delta^2 V_i^n. \end{aligned} \tag{2.33}$$

The right-hand sides of (2.33) are identical with those of the single-step Lax–Wendroff method, as is true for general A, while the mass matrix again improves the accuracy.

Results are given below for the following initial-boundary-value data on $[0, \pi]$.

$$\begin{aligned} v(x, 0) = 2 - u(x, 0),\ u(x, 0) &= 1, && 0 \leqslant x \leqslant \pi/3, \\ &= \tfrac{1}{2}(1 - \cos 3x), && \pi/3 \leqslant x \leqslant 2\pi/3, \\ &= 0, && 2\pi/3 \leqslant x \leqslant \pi, \end{aligned}$$

$$u(0, t) = 1, \qquad u(\pi, t) = 0,$$

where the solution involves a reflection at $x = \pi$ for $t \in [\pi/3, 2\pi/3]$. Maximum nodal errors for the EPG approximation (2.33) are given in Table II together with comparison values for Lax–Wendroff. Two different boundary conditions on v at $x = 0$ and $x = \pi$ were used with Lax–Wendroff, namely, a one-sided difference and the Matsuno scheme; cf. May and Morton [6], Matsuno [5]. For the EPG method, the most obvious treatment of the boundary conditions was used: that is, the boundary condi-

TABLE II

Maximum Nodal Errors for the Reflection Problem with $h = \pi/27$ and $\Delta t = \pi/30$[a]

t	Euler–Petrov–Galerkin	Lax–Wendroff (a)	Lax–Wendroff (b)
$\pi/6$	0.0040	0.0062	0.0062
$\pi/3$	0.0060	0.0089	0.0089
$\pi/2$	0.0110	0.0499	0.0123
$2\pi/3$	0.0140	0.0555	0.0220
$5\pi/6$	0.0110	0.0670	0.0171
π	0.0120	0.0650	0.0197
$7\pi/6$	0.0140	0.0708	0.0197

[a] Lax–Wendroff results are for two boundary conditions on v: (a) one-sided difference, (b) Matsuno scheme.

tions in (2.34) were imposed on U so that the first component of (2.32) was used for all except the first and last values of i and the second component for all i. These results illustrate both the improved accuracy of the Petrov–Galerkin method and the suitability of its so-called "natural" boundary conditions for reflection problems.

The dependence of (2.29) upon the matrix S is clearly a disadvantage to the method. This prompted the idea of modifying the basic test function so as to avoid such a dependence. One possible modification would be to choose

$$\psi_{i(r)} = (1 - \mu_r^2)\, \phi_i + \tfrac{1}{2}\mu_r^2(\chi_i^+ + \chi_i^-) + \tfrac{1}{2}\mu_r(\chi_i^+ - \chi_i^-) \qquad \text{for all } i. \quad (2.35)$$

The "unit CFL property" is retained as $| \mu_r | \to 1$ and (2.29) is replaced by

$$\begin{aligned}\langle \mathbf{U}^{n+1} - (\mathbf{U}^n + \Delta t\, A \partial_x \mathbf{U}^n), \phi_i \mathbf{e}_{(r)} \rangle + (\Delta t/h)^2 \langle \mathbf{U}^{n+1} - \mathbf{U}^n, \sigma_i A^2 \mathbf{e}_{(r)} \rangle \\ = (\Delta t^2/h) \langle A \partial_x \mathbf{U}^n, \tau_i A \mathbf{e}_{(r)} \rangle, \qquad r = 1, 2, \ldots, m, \text{ for all } i. \quad (2.36)\end{aligned}$$

This scheme does not depend on S and furthermore appears to be more accurate than (2.29). The Fourier analysis of the truncation error for the scalar equation shows that the leading order truncation term of Lax–Wendroff has now been cancelled and the method is third order. Compared with the expansion in (2.13), we have now introduced two parameters instead of ν so as to match the first four terms of the amplification factor to $e^{i\mu\xi}$. This is not a scheme which could readily have come to mind for solving the scalar equation (2.3) but for completeness we have tabulated its phase and dissipation properties in Table I and have given results obtained with it in Figures 1 and 2, where we have called it EPG II.

2.6. *Nonlinear Equations*

As a final test of the EPG scheme, we have considered briefly its application to the standard test proboem: $u_t + \frac{1}{2}\partial_x(u^2) = 0$ with initial data $u(x, 0) = 1$ for $x \leqslant 0$, $u(x, 0) = 0$ for $x > 0$. To accurately reproduce the speed of propagation of the discontinuity, any scheme must satisfy a discrete analogue of the conservation law

$$2 \int_A^B u_t \, dx = u_A^2 - u_B^2. \qquad (2.37)$$

For a Petrov–Galerkin method, this means that the test functions must span the constant function: then (2.37) is reproduced with u_t replaced by a time difference. Unfortunately the EPG test functions do not have this property. Nevertheless we have applied the scheme to this problem to test the effect of setting $\nu_i = U_i^n \Delta t/h$: as predicted, the shock speed was incorrect but the shock remained very sharp with virtually no oscillations.

3. Methods Based on "Theta" and Leapfrog Time-Stepping

The principles of Section 2 are applied in this section to obtain Petrov–Galerkin approximations for both theta-method time-stepping and leapfrog time-stepping. Special test functions are derived having the unit CFL property for the scalar convection equation, and these are taken in linear combination with the Galerkin test functions. The Petrov–Galerkin method using theta time-stepping is given first.

3.1. *"Theta" Time-Stepping*

We replace the time derivative in $u_t = au_x$, $a > 0$, with a "theta method" time discretisation so that the Petrov–Galerkin equations, for some approximate solution U and test functions ψ_i, are

$$\langle U^{n+1} - a\Delta t\theta\partial_x U^{n+1} - (U^n + a\Delta t(1-\theta)\,\partial_x U^n), \psi_i\rangle = 0 \qquad \text{for all } i. \tag{3.1}$$

Consider again the special case of piecewise linear initial data $u_0 = U^0(x)$ with $a\Delta t = h$, for which Eqs. (3.1) give the exact solution if $\psi_i \equiv \chi_i^+$ and the χ_i^+ satisfy

$$\sum_{(j)} U^0_{n+j}\langle \phi_{j-1} - \theta h\,\partial_x\phi_{j-1} - \phi_j - (1-\theta)\,h\,\partial_x\phi_j\,, \chi_i^+\rangle = 0 \qquad \text{for all } i. \tag{3.2}$$

These constraints simplify to

$$\langle \phi_j - (1-\theta)\,\phi_j^{\,0} - \theta\phi^0_{j-1}\,, \chi_i^+\rangle = 0, \qquad \text{for all } i \text{ and } j, \tag{3.3}$$

which further reduce to

$$\langle r_j - \theta\phi_j^{\,0}, \chi_i^+\rangle = 0 \qquad \text{for all } i \text{ and } j \tag{3.4}$$

in the notation of subsection 2.1, where (2.7) corresponds to $\theta = 0$. As before we put $\chi_i^+ = \chi(x/h - i)$ for all i, where $\chi(s)$ is nonzero only in $[0, 1]$. Equations (3.4) then imply that $\chi(s)$ satisfies

$$\int_0^1 s\chi(s)\,ds = \theta\int_0^1 \chi(s)\,ds \tag{3.5}$$

and the simplest such χ with the appropriate normalisation $\int_0^1 \chi(s)\,ds = 1$ is given by

$$\chi(s) = 2(2-3\theta) - 6(1-2\theta)\,s, \qquad s \in [0, 1]. \tag{3.6}$$

We use a linear combination of these functions with Galerkin test functions in (3.1), i.e.,

$$\psi_i = (1-\nu)\,\phi_i + \nu\chi_i^+ \tag{3.7}$$

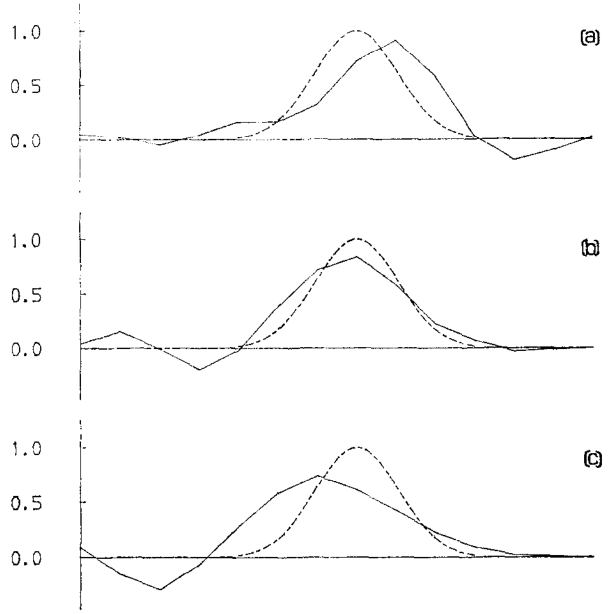

FIG. 5. Solutions of $u_t = au_x$ by (a) Leap-frog Galerkin, (b) LPG, (c) leap-frog finite-difference scheme. Note that (a) is calculated with $\mu = 0.45$, whereas (b) and (c) are at $\mu = 0.9$.

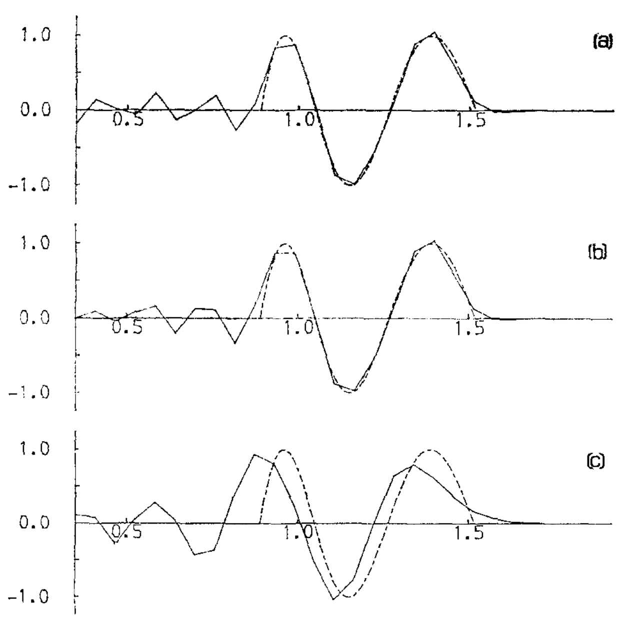

FIG. 6. Solutions of Eqs. (2.17) and (2.18) by (a) Leap-frog Galerkin, (b) LPG, (c) leap-frog finite difference, all with $\Delta t/\Delta x = 0.45$.

These equations have been tested on the wave reflection problem given in Section 2.5, for which case they have the simpler form

$$\langle \mathbf{U}^{n+1} - \mathbf{U}^{n-1} - 2\,\Delta t\, A\partial_x \mathbf{U}^n, \phi_i \mathbf{e}_{(r)}\rangle + (\Delta t/h)^2 \langle \mathbf{U}^{n+1} - \mathbf{U}^{n-1}, \sigma_i \mathbf{e}_{(r)}\rangle = 0, \quad r = 1, 2 \text{ for all } i. \tag{3.24}$$

Table V presents the maximum nodal errors for this problem using the same format as Table II. Natural boundary conditions are again used and the results show an encouraging improvement over Galerkin and are even better than those with leap-frog on a staggered mesh at later times.

TABLE V

Maximum Nodal Errors for the Wave Reflection Problem $\mu = 0.45$ and 0.9

	$\mu = 0.45$		$\mu = 0.9$	
Time	Petrov–Galerkin	Galerkin	Petrov–Galerkin	Staggered leap-frog
$\pi/6$	0.0040	0.0047	0.0030	0.0031
$\pi/3$	0.0060	0.0076	0.0070	0.0035
$\pi/2$	0.0110	0.0074	0.0104	0.0052
$2\pi/3$	0.0140	0.0168	0.0095	0.0055
$5\pi/6$	0.0110	0.0248	0.0074	0.0106
π	0.0120	0.0255	0.0060	0.0078
$7\pi/6$	0.0140	0.0319	0.0097	0.0122

4. Concluding Remarks

The schemes which have been devised in the preceding sections have demonstrated the value of the extra flexibility afforded by the Petrov–Galerkin technique as compared with the purely Galerkin approach. Stability has been markedly improved and in several cases computation has been simplified by making the mass matrix more diagonally dominant: but, most importantly, accuracy has been greatly increased for practical time-steps. It is not claimed that either the best test functions have been found for the limiting case of unit CFL number or that the proposed linear combinations of these functions with the trial functions give the optimal schemes for intermediate time-steps: but the results tabulated in the foregoing do show the potential of this approach.

Much still remains to be done, however. Our studies have been largely confined to linear problems and always to simple systems of equations and one space dimension. Studies are in progress for the shock-tube problem used by Sod [11] in his survey article and work has started on multidimensional problems. But we expect that success with these more demanding problems may well depend upon devising multistep versions of the algorithms given above, as, for example, in the two-step Lax–Wendroff method or used by Douglas and Dupont [3] for Crank–Nicolson–Galerkin.

Appendix

THEOREM. *The Euler–Petrov–Galerkin method* (2.10) for $u_t = au_x$, $a > 0$, *is stable for* $0 \leqslant a\,\Delta t/h \leqslant \nu \leqslant 1$.

Proof. We first note that the basis functions ϕ_j^0, ϕ_j, and χ_j^+ satisfy

$$\langle \chi_i^+, \phi_j^0\rangle = \langle \chi_i^+, \phi_j\rangle = \langle \phi_i^0, \phi_j^0\rangle = \tfrac{1}{4}\langle \chi_i^+, \chi_j^+\rangle = h\delta_{ij}\,. \tag{A1}$$

Then we introduce the function $V(x) = \sum V_j\phi_j(x)$ with nodal parameters given by

$$\langle V - a\partial_x U, \psi_i\rangle = 0, \qquad \text{for all } i, \tag{A2}$$

where ψ_i are the test functions $\psi_i = (1-\nu)\,\phi_i + \nu\chi_i^+$. Multiplication of each Petrov–Galerkin equation (2.10) by the coefficient $(U_i^{n+1} + U_i^n + \Delta t\, V_i)$ and summing over i thus lead to the equation

$$\left\langle U^{n+1} - (U^n + a\,\Delta t\,\partial_x U^n), \sum_{(i)} (U_i^{n+1} + U_i^n + \Delta t\, V_i)\,\psi_i\right\rangle = 0. \tag{A3}$$

From (A1), the inner products between U^n and U^{n+1} are symmetric and therefore cancel. If we introduce the discrete norm $|\,U\,|^2$ for $h\sum_{(i)} U_i^2$, (A3) reduces therefore to

$$\begin{aligned}
&[(1-\nu)\,\|\,U^{n+1}\,\|^2 + \nu\,|\,U^{n+1}\,|^2] - [(1-\nu)\,\|\,U^n\,\|^2 + \nu\,|\,U^n\,|^2]\\
&\quad = \left\langle a\,\Delta t\,\partial_x U^n, \sum_{(i)} (U_i^{n+1} + U_i^n)\,\psi_i\right\rangle - \left\langle U^{n+1} - U^n, \Delta t \sum_{(i)} V_i\psi_i\right\rangle\\
&\qquad + \left\langle a\,\Delta t\,\partial_x U^n, \Delta t \sum_{(i)} V_i\psi_i\right\rangle.
\end{aligned} \tag{A4}$$

But, by the construction (A2), the second inner product can be replaced by $\langle a\,\Delta t\, \partial_x U^n, \sum_{(i)}(U_i^{n+1} - U_i^n)\,\psi_i\rangle$ and combined with the first so that the right-hand side of (A4) reduces to

$$\left\langle a\,\Delta t\,\partial_x U^n, \sum_{(i)} (2U_i^n + \Delta t\, V_i)\,\psi_i\right\rangle. \tag{A5}$$

We need to bound this in order to apply the energy method and establish stability. Now $\langle \partial_x U^n, U^n \rangle = 0$ and also

$$\left\| h\,\partial_x U + \sum_{(i)} U_i^n \chi_i^+ \right\|^2 = \left\| \sum_{(i)} U_i^n(\phi_{i-1}^0 - \phi_i^0 + \chi_i^+) \right\|^2$$
$$= 4 \mid U^n \mid^2 = \left\| \sum_{(i)} U_i^n \chi_i^+ \right\|^2. \tag{A6}$$

Hence, multiplying out the square, we have

$$2\left\langle \partial_x U^n, \sum_{(i)} U_i^n \psi_i \right\rangle = 2\nu \left\langle \partial_x U^n, \sum_{(i)} U_i^n \chi_i^+ \right\rangle = -\nu h \,\| \partial_x U^n \|^2. \tag{A7}$$

For the other term in (A5) we have, from (A2),

$$\left\langle a\,\partial_x U^n, \sum_{(i)} V_i \psi_i \right\rangle = \left\langle V, \sum_{(i)} V_i \psi_i \right\rangle = (1-\nu)\,\| V \|^2 + \nu \mid V \mid^2. \tag{A8}$$

In addition,

$$2\left\langle a\,\partial_x U^n, \sum_{(i)} V_i \psi_i \right\rangle = 2(1-\nu)\,\langle a\,\partial_x U^n, V \rangle + 2\nu \left\langle \sum_{(j)} aD_+ U_j^n \phi_j^0, \sum_{(i)} V_i \chi_i^+ \right\rangle$$
$$\leqslant (1-\nu)[\| a\,\partial_x U^n \|^2 + \| V \|^2] + \nu[\| a\,\partial_x U^n \|^2 + \mid V \mid^2]. \tag{A9}$$

Subtracting (A8) from (A9) gives

$$\left\langle a\,\partial_x U^n, \sum_{(i)} V_i \psi_i \right\rangle \leqslant \| a\,\partial_x U^n \|^2; \tag{A10}$$

and, combining this with (A7), gives as a bound for the expression in (A5)

$$(\text{A5}) \leqslant [(a\,\Delta t)^2 - \nu h a\,\Delta t]\, \| \partial_x U^n \|^2. \tag{A11}$$

This establishes stability of the EPG method for the pure initial-value problem in the energy norm given on the left-hand side of (A4).

ACKNOWLEDGMENTS

This work was started while one of the authors (K.W.M.) was supported by the U.S. Army under Contract DAA 929-75-C-0024 at the University of Wisconsin, Madison. The second author has been supported by the Science Research Council under Research Grant GR/A 07706.

REFERENCES

1. M. J. P. CULLEN, "Application of the Finite Element Method to Numerical Weather Prediction," Ph. D. thesis, Reading University, 1975.

2. M. J. P. CULLEN AND K. W. MORTON, "Analysis of Evolutionary Error in Finite Element and Other Methods," Report NA 5/77, University of Reading Mathematics Department, 1977, *J. Comput. Phys.* **34** (1980), 245–267.
3. J. DOUGLAS AND T. DUPONT, *SIAM J. Numer. Anal.* **4** (1970), 575–626.
4. P. LESAINT, *Numer. Math.* **21** (1973), 244–255.
5. T. MATSUNO, *J. Met. Soc. Japan, Ser.2* **44** (1966), 145–157.
6. T. MAY AND K. W. MORTON, "Numerical Boundary Conditions for the Wave Equation in One Dimension," Report NA 4/76, University of Reading Mathematics Department, 1976.
7. A. R. MITCHELL, *in* "The State of the Art in Numerical Analysis" (D. A. H. Jacobs, Ed.), pp. 671–697, Academic Press New York, 1977.
8. M. S. MOCK, *Math. Comput.* **30** (1976), 400–416.
9. K. W. MORTON, *in* "The State of the Art in Numerical Analysis" (D. A. H. Jacobs, Ed.), pp. 699–756, Academic Press New York, 1977.
10. R. D. RICHTMYER AND K. W. MORTON, "Difference Methods for Initial Value Problems," Wiley, New York, 1967.
11. G. A. SOD, *J. Comput. Phys.* **27** (1978), 1–31.
12. B. SWARTZ AND B. WENDROFF, *Math. Comput.* **23** (1969), 37–50.
13. B. SWARTZ AND B. WENDROFF, *SIAM J. Numer. Anal.* **11** (1974), 979–993.
14. V. THOMÉE AND B. WENDROFF, *SIAM J. Numer. Anal.* **11** (1974), 1059–1068.