COMMENTS ON 'THE GROUP VELOCITY OF SOME NUMERICAL SCHEMES'

PHILIP M. GRESHO AND ROBERT L. LEE

Lawrence Livermore National Laboratory, University of California, Livermore, California 94550, U.S.A.

SUMMARY

The proper phase and group speeds when quadratic finite elements are applied to the one-dimensional pure advection equation are presented and the myth of a spurious computational mode is dispelled.

KEY WORDS Quadratic Elements FEM Phase Speed Spurious Modes

In a very interesting and generally well done paper,[1] Cathers and O'Connor (hereafter referred to as COC) fell into a common trap regarding the analysis of a finite element method in which quadratic basis functions are employed. Before presenting any details, let us first make clear the major point of this paper: *there is no spurious computational mode associated with the quadratic finite element*. In this paper we explain why this is so and also demonstrate the final result numerically.

First we mention two other references[2,3] in which the authors misinterpreted the final equations resulting from a phase speed (and sometimes group speed) analysis when quadratic finite elements are used to model pure advection—note that the first of these refers to our own work!

The scalar problem addressed herein is

$$\frac{\partial u}{\partial t}+V\frac{\partial u}{\partial x}=0;\quad 0\leqslant x\leqslant 1, \tag{1}$$

where V is a constant advecting velocity. Periodic boundary conditions are employed (for analytical simplicity) and we seek the approximate solution via the (Galerkin) finite element method using quadratic basis functions on a uniform mesh and the trapezoid rule (TR) for time integration. (The TR is the best of the 'θ-family' considered by COC, in that it is second-order accurate and dissipation-free.) We denote $h\equiv 1/(N+1)$ as the mesh spacing, where there are $(N+1)/2$ quadratic elements; there are thus $N+1$ nodes and $N+1$ unknowns. (Note that $N+1$ is even.)

The semi-discretized equations can be written as

$$M\dot{u}+Ku=0, \tag{2}$$

where M is the mass matrix, K the advection matrix, and u is now the $(N+1)$-vector of nodal values. The detailed expansion of (2) is given in equations 42 (for end nodes) and 43 (for mid-side nodes) in COC when the TR

$$M\frac{(u_{m+1}-u_m)}{\Delta t}+K\frac{(u_{m+1}+u_m)}{2}=0 \tag{3}$$

is employed for time integration ($\theta=1/2$ in their equations).

0271–2091/87/121357–06$05.00

Received 14 November 1985
Revised 29 January 1987

We now present a summary of the analysis of (2) and (3) for the numerical phase speed and group velocity (both of which are V in the continuum). Seeking a solution to (2) of the form $u = z\exp(\mathrm{i}\omega t)$ leads to the generalized eigenvalue problem

$$Kz = -\mathrm{i}\omega Mz, \tag{4}$$

where the eigenvalues $\{\omega\}$ are real since M is symmetric positive definite and K is skew-symmetric (note that equation (43) in COC needs to be multiplied by 2 to display these symmetries). The eigenvectors of (4) are assumed to exhibit the form

$$\begin{aligned} z_j &= \exp(\mathrm{i}j\gamma), && \text{at the end nodes} && j = 2, 4, \ldots, N+1, \\ z_j &= \beta\exp(\mathrm{i}j\gamma), && \text{at the mid-side nodes} && j = 1, 3, \ldots, N, \end{aligned} \tag{5}$$

where $\gamma = \sigma h = 2\pi nh = 2\pi n/(N+1)$; $n = 1, 2, \ldots, N+1$ is the mode number, $h = \Delta x$ is the mesh spacing and $\sigma = 2\pi n$ is the wave number.

It is important to note that an amplitude coefficient β (which is a function of the dimensionless wave number γ) must be employed in the eigenvectors for the mid-side nodes. After substituting the trial eigenvectors (5) into (4), we obtain a non-linear pair of equations relating the frequency ω, and β, to γ. The solution of the equations is

$$\omega^{\pm} = \frac{(N+1)V}{(3-\cos 2\gamma)}\left[-2\sin 2\gamma \pm \surd(19 - 20\cos 2\gamma + \cos^2 2\gamma)\right], \tag{6a}$$

$$\beta^{\pm} = \frac{1}{4}\left[\frac{5(N+1)\sin\gamma}{\omega^{\pm}} - \cos\gamma\right], \tag{6b}$$

where both ωs and βs are real.

Now, as it stands, (6a) would generate $2(N+1)$ values for ω. But it is clear from (4), and from the properties of K and M, that there should be exactly $(N+1)$ eigenvalues and eigenvectors. The apparent anomaly can be explained by noting that

$$\omega^{\pm}(\gamma+\pi) = \omega^{\pm}(\gamma),$$

$$\beta^{\pm}(\gamma+\pi) = -\beta^{\pm}(\gamma)$$

and

$$z^{\pm}(\gamma+\pi) = z^{\pm}(\gamma); \tag{7}$$

therefore every eigenvalue and eigenvector would be *duplicated* as n ranges over the integers $1, 2, \ldots., N+1$.

However, since the matrices K and M are both real, these eigenvalues (viewed now as $\lambda = \mathrm{i}\omega$) must occur in complex conjugate pairs. This redundancy issue can be resolved by realizing that the only way the solutions can be consistent with the symmetry properties of (7) is to take the '+' sign for the first half of the spectrum ($n = 1, 2, \ldots, (N+1)/2$), which gives positive values of ω, and the '−' sign for the second half ($n = (N+1)/2 + 1, \ldots, N+1$), which gives negative values of ω. It can be shown that this choice of eigenmodes also leads to a positive value of β (which ranges between 1/2 and 1) over the entire spectrum, and correctly eliminates the occurrence of 'spurious modes' (which were merely the extraneous roots of a quadratic equation). It also causes λ to 'behave like' simpler centred schemes, such as second-order finite differences, for which $\lambda = \mathrm{i}NV\sin 2\pi n/N$ for $n = 1, 2, \ldots, N$.

We have also verified these results numerically with a generalized eigenvalue routine.

Having resolved the spurious mode issue, we can now move on to phase and group speeds. For these it turns out, almost ironically, that the upper half of the spectrum (and the concomitant minus sign) is never even used—owing to aliasing (the upper half corresponds to waves whose length would be between h and $2h$—and only waves of length $\geqslant 2h$ are resolvable). The phase speed is given by

$$P = \omega/\sigma \tag{8}$$

and the group speed by

$$G = \mathrm{d}\omega/\mathrm{d}\sigma, \tag{9}$$

which leads to

$$P = V\frac{\sqrt{(19 - 20\cos 2\gamma + \cos^2 2\gamma)} - 2\sin 2\gamma}{\gamma(3 - \cos 2\gamma)} \tag{10}$$

and

$$G = \frac{2V}{3 - \cos 2\gamma}\left[-2\cos 2\gamma + \frac{(10 - \cos 2\gamma)\sin 2\gamma}{\sqrt{(19 - 20\cos 2\gamma + \cos^2 2\gamma)}} - \frac{\sin 2\gamma}{3 - \cos 2\gamma}(\sqrt{(19 - 20\cos 2\gamma + \cos^2 2\gamma)} - 2\sin 2\gamma)\right], \tag{11}$$

and γ, treated for convenience as a continuous variable, ranges from 0 to π (hence $n_{\max} = (N + 1)/2$). These results apply to the semi-discretized equations (2) and do not include any effects of time truncation error.

To obtain the analogous results when the TR is used for time integration, i.e. from (3), we proceed as follows: seek a solution to (3) in the form $u_m = y\mathrm{e}^{\mathrm{i}m\phi}$ which leads to

$$\frac{(\mathrm{e}^{\mathrm{i}\phi} - 1)}{\Delta t}My + \frac{(\mathrm{e}^{\mathrm{i}\phi} + 1)}{2}Ky = 0,$$

or

$$Ky = -\left(\frac{2\mathrm{i}}{\Delta t}\tan\phi/2\right)My, \tag{12}$$

which, when compared to (4), yields $y = z$ (the eigenvectors are unchanged) and

$$\omega = \frac{2}{\Delta t}\tan\phi/2, \tag{13}$$

where ω is given by (6a); i.e. (13) is to be solved for ϕ. In this case, since $m = t/\Delta t$, the phase speed is

$$\tilde{P} = \phi/\sigma\,\Delta t = V\phi/c\gamma, \tag{14}$$

where $c \equiv V\Delta t/h$ is the Courant number; and the group speed is

$$\tilde{G} = \frac{1}{\Delta t}\frac{\mathrm{d}\phi}{\mathrm{d}\sigma} = \frac{1}{\Delta t}\frac{\mathrm{d}\phi}{\mathrm{d}\omega}\frac{\mathrm{d}\omega}{\mathrm{d}\sigma} = \frac{G}{\Delta t}\frac{\mathrm{d}\phi}{\mathrm{d}\omega}. \tag{15}$$

Thus,

$$\tilde{P} = \frac{2V}{c\gamma}\tan^{-1}\frac{\omega\,\Delta t}{2} = \frac{2V}{c\gamma}\tan^{-1}\left(\frac{c\gamma}{2}\frac{P}{V}\right) \tag{16}$$

and

$$\tilde{G} = \frac{G}{1 + \left(\frac{\omega\,\Delta t}{2}\right)^2} = \frac{G}{1 + \left(\frac{c\gamma}{2}\frac{P}{V}\right)^2} \tag{17}$$

are the phase and group velocities for the fully discrete scheme. It is noteworthy (although rather well known) that $\tilde{P} \leqslant P$ and $\tilde{G} \leqslant G$; the TR (always) retards the wave-forms. Also, for $c \to 0$ (i.e. $\Delta t \to 0$), $\tilde{P} \to P$ and $\tilde{G} \to G$ and all of the error is spatial.

Finally, we remark that (16) and (17) are general in that they relate the continuous-in-time results [e.g. from (10) and (11)] to those when the TR is used for time integration; i.e. P and G could just as well correspond to any (non-dissipative) spatial discretization [by changing M and K—and thus ω—in (4)].

Figure 1 shows $\tilde{P}/V$ and $\tilde{G}/V$ vs $2\pi/\gamma$ for several values of c. Although the group speed is < 0 for short waves (and -5 for the $2\,\Delta x$ wave), it is important to notice that the phase speed is always $\geqslant 0$. These results are to be compared with Figures 10 and 11 in COC; note, however, that they used $\theta = 0{\cdot}6$, a dissipative first-order scheme, in Figure 10. Finally, their Figure 12 (as well as the dashed curves in Figure 10) is extraneous.

To complete the story, we performed our own version of their experiment shown in their Figure 13. The 'numerical' wave packet is predicted to move leftward at predominantly the group speed of a '$2\,\Delta x$' wave, namely at $\tilde{G} = -5$. Figure 2 shows this to be essentially the case. In fact, the results shown (for $c = 1$) are close to those for $c = 0{\cdot}1$ (not shown), showing the good accuracy of the TR integrator. The latter result (i.e. one with negligible time truncation error) was also verified via an analytic solution using an eigenvector expansion, the details of which are described by Rowley and Gresho.[5] We, like they, have no explanation for the 'odd behaviour' computed in their

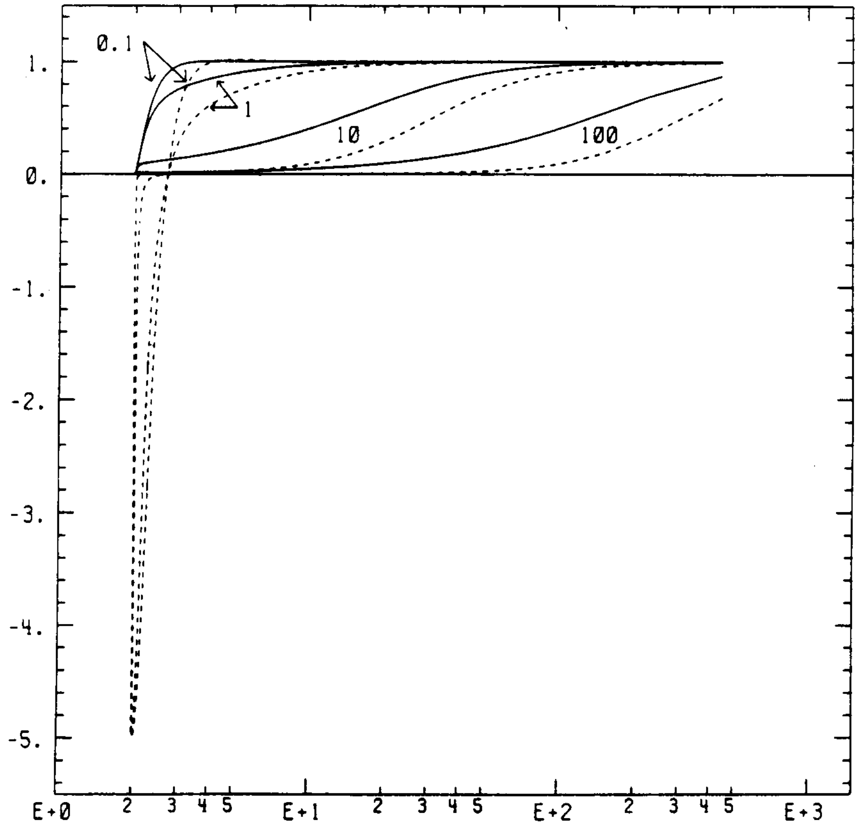

Figure 1. Phase speed (solid) and group speed (dashed) vs dimensionless wavelength ($2\pi/\gamma$) for several values of c

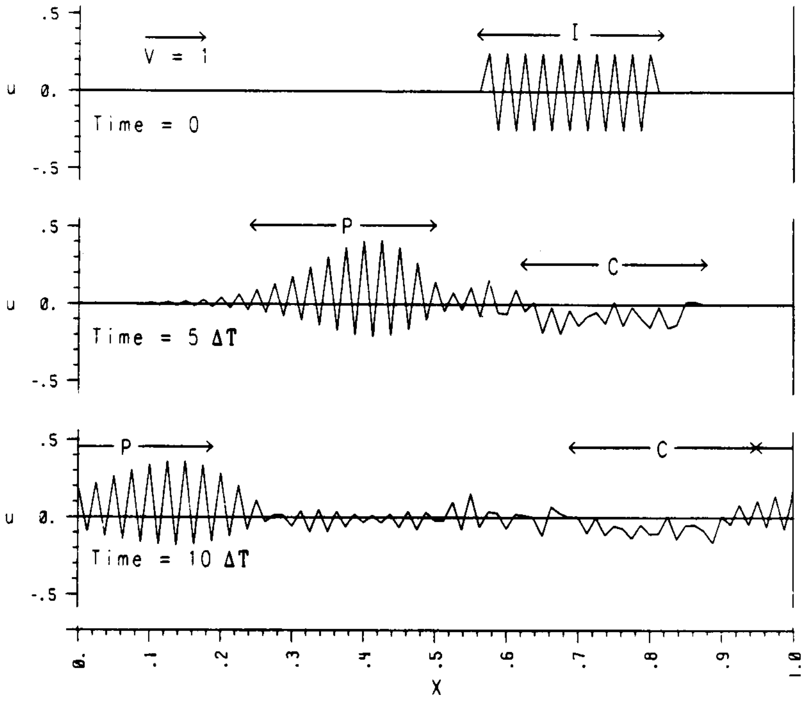

Figure 2. Time evolution of a '2 Δx' wave packet for $c = 1$: I = initial, P = predicted location and C = continuum location

experiment, in which the wave form seemed to move at a speed of about + 1. [Subsequent numerical experiments conducted by Cathers (private communication), who now also agrees with our analysis, confirmed our results.]

We end by remarking that anyone truly interested in group velocity of numerical schemes should become familiar with the excellent paper by Trefethen.[4]

ACKNOWLEDGEMENTS

We have benefited from the assistance provided by Mr. J. Rowley, Drs J. L. Leone and A. C. Hindmarsh of LLNL and Professor R. L. Sani of the University of Colorado. We also acknowledge the careful review of an earlier version of our manuscript and the referees' suggestions to clarify certain parts of the derivations.

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

REFERENCES

1. B. Cathers and B. A. O' Connor, 'The group velocity of some numerical schemes', *Int. j. numer. methods fluids*, **5** (3), 201–224 (1985).
2. P. M. Gresho, R. L. Lee and R. L. Sani, 'Advection-dominated flows, with emphasis on the consequences of mass lumping', in R. H. Gallagher, O. C. Zienkiewicz, J. T. Oden, M. Morandi Cecchi, and C. Taylor (eds), *Finite Elements in Fluids, Vol. 3*, Wiley, Chichester, 1978, p. 335.

3. G. Hedstrom, 'The Galerkin method based on Hermite cubics', *SIAM J. Num. Anal.*, **16** (3), 385 (1979).
4. L. Trefethen, 'Group velocity in finite difference schemes', *SIAM Review*, **24** (2), 113 (1982).
5. J. Rowley and P. Gresho, 'Some new results using quadratic finite elements for pure advection', *Proceedings, 6th IMACS Int. Symp. on Computer Methods for PDES*, Leheigh University; Bethlehem, PA, 23–27 June 1987. Also available as UCRL-96615.