JOURNAL OF COMPUTATIONAL PHYSICS 63, 268–282 (1986)

Invariance Theorems Concerning Reflection at Numerical Boundaries

R. VICHNEVETSKY*

Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, New Jersey 08544

Received December 8, 1983; revised February 12, 1985

Downwind computational boundaries in the numerical approximation of hyperbolic equations are in general not transparent, and they create spurious reflection. A useful measure of this is given by the energy (or usual sum of squares) of the reflected solution in response to an arbitrary solution which originates from within the computing domain. We prove, in that respect, a somewhat unexpected property: namely, that for those full-discretizations which are obtained by applying to a space-discretization of the equations an energy conservative discrete time-marching method, the energy reflected at the boundary is independent of the value of Δt, and is strictly equal to the energy reflected in the semidiscrete case. This is verified in numerical experiments. Optimal boundary equations may be defined in the semidiscrete case of those which maximize the rate of convergence to zero of the reflected energy when $h \to 0$. A corollary of the preceding result is that those boundary equations remain optimal, in the same sense, when used in an energy conservative full discretization. Moreover, this convergence result continues to hold when a nonconservative but stable (i.e., dissipative) time-discretization is used. This is verified numerically with the first-order Adams predictor–corrector method. The results of this paper are derived with the mathematics of the simple three-point central difference discretization of spatial derivatives. Obvious generalizations to other cases are mentioned at the end of the paper.

1. INTRODUCTION

Consider the half space

$$D \equiv (-\infty, 0] \tag{1}$$

on which the simple advection equation

$$\frac{\partial U}{\partial t} + c\frac{\partial U}{\partial x} = 0 \qquad (c > 0) \tag{2}$$

is approximated by the central differences semidiscretization:

$$\frac{du_j}{dt} = -c\left(\frac{u_{j+1} - u_{j-1}}{2h}\right) \equiv \mathbf{A} \cdot u_j \qquad (u_j(t) \simeq U(jh, t)) \tag{3}$$

* Permanent address: Department of Computer Science, Rutgers University, New Brunswick, New Jersey 08903.

0021-9991/86 $3.00

in interior points, and by an equation of the form

$$\mathscr{B} \equiv \frac{du_0}{dt} - b_0 u_0 - b_1 u_{-1} - \cdots \equiv \frac{du_0}{dt} - \mathbf{B} \cdot u_0 = 0 \tag{4}$$

at the boundary $x = 0$. Fully discrete approximations are obtained when a discrete time-marching method is applied to (3), (4). This may be expressed in operator notations as

$$\mathbf{M}(\mathbf{Z}) \cdot u_j^n = \mathbf{A}^* \cdot u_j^n \qquad (u_j^n \simeq U(jh, n\Delta t)), \tag{5}$$

where $\mathbf{A}^*$ is the discrete spatial operator consisting of $\mathbf{A}$ in interior points and $\mathbf{B}$ at the boundary, and the operator $\mathbf{M}$, which approximates $\partial/\partial t$, contains the time-shift operator $\mathbf{Z}$ defined by the identity

$$\mathbf{Z} \cdot u_j^n \equiv u_j^{n+1}. \tag{6}$$

The theorems which shall be derived in this paper are concerned with certain properties of the spurious reflection which occurs at the numerical boundary. For simplicity, we shall establish those theorems in the particular case of time discretization with the Crank–Nicolson (or trapezoidal) method, and will describe thereafter which of the corresponding results may carry over to other cases.

The operator notation for the Crank-Nicolson method is

$$\mathbf{M}(\mathbf{Z}) = \frac{2}{\Delta t}\left(\frac{\mathbf{Z}-1}{\mathbf{Z}+1}\right) \tag{7}$$

and (5) may also be rewritten in that case as:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \mathbf{A}^*\left(\frac{u_j^{n+1} + u_j^n}{2}\right). \tag{5a}$$

2. Fourier Analysis

Spurious reflection at the numerical boundary may be analyzed by Fourier methods. We briefly recall the basic properties that shall be used:

The t-Fourier transform of $u_j(t)$ is defined, in the semidiscrete case

$$\hat{u}_j(\omega) = \int_{-\infty}^{\infty} u_j(t)\, e^{-i\omega t}\, dt \tag{8}$$

(it is assumed, here and in further occurrences of Fourier transforms, that the functions being transformed are of finite $\mathscr{L}_2$ or l_2 norm). If the set of functions $\{u_j(t)\}$ is a solution of (3) in D, then it may be decomposed in two components

$$\{u_j\} = \{p_j\} + \{q_j\} \tag{9}$$

whose Fourier transforms satisfy

$$\frac{\hat{p}_{j+1}(\omega)}{\hat{p}_j(\omega)} = -i\frac{\omega h}{c} + \sqrt{1-(\omega h/c)^2} \equiv \hat{E}_p(\omega) \tag{10}$$

and

$$\frac{\hat{q}_{j+1}(\omega)}{\hat{q}_j(\omega)} = -i\frac{\omega h}{c} - \sqrt{1-(\omega h/c)^2} \equiv \hat{E}_q(\omega), \tag{11}$$

respectively. This is obtained by taking the Fourier transform of Eq. (3), thus resulting in a difference equation for $\hat{u}_j(\omega)$. This equation is solved by solving the corresponding quadratic characteristic equation whose two roots are (10) and (11).

Sinusoidal wave propagation may exist for frequencies less than the cut-off frequency ω_c given by

$$\frac{\omega_c h}{c} = 1 \qquad \text{or} \qquad \omega_c = \frac{c}{h}. \tag{12}$$

The group velocity of solutions of p type (in $|\omega| < \omega_c$) is positive: these are rightgoing solutions. The group velocity of solutions of q type is negative: these are leftgoing solutions.

When a rightgoing solution arrives from D at the boundary, it is partially reflected toward D as a leftgoing solution. An expression of this is given by the *amplitude reflection ratio* $\rho(\omega)$, obtained by taking the t-Fourier transform of the boundary equation (4), and then solving for $\hat{q}_0(\omega)/\hat{p}_0(\omega)$.

To obtain the corresponding relations in the fully discrete case, one has to work with discrete t-Fourier transforms

$$\overline{u_j}(\omega) = \Delta t \sum_{n=-\infty}^{\infty} u_j^n e^{-i\omega n \Delta t}. \tag{13}$$

One then takes the discrete Fourier transform of Eq. (5). Solving for $\overline{u_{j+1}}(\omega)/\overline{u_j}(\omega)$ in interior points results in a quadratic characteristic equation which is identical to that of the semi-discrete case, save for the replacement of ω by $\mu(\omega)$, the spectral function or symbol of the operator **M**, defined as

$$i\mu(\omega) \equiv \frac{\mathbf{M}(\mathbf{Z}) \cdot e^{i\omega n \Delta t}}{e^{i\omega n \Delta t}} = \mathbf{M}(e^{i\omega \Delta t}). \tag{14}$$

Accordingly, we have in the fully discrete case

$$\{u_j^n\} = \{p_j^n\} + \{q_j^n\}, \tag{9a}$$

$$\frac{\overline{p}_{j+1}(\omega)}{\overline{p}_j(\omega)} = -i\frac{\mu(\omega)\,h}{c} + \sqrt{1-\left(\frac{\mu(\omega)\,h}{c}\right)^2} = \hat{E}_p(\mu(\omega)), \tag{10a}$$

$$\frac{\overline{q}_{j+1}(\omega)}{\overline{q}_j(\omega)} = -i\frac{\mu(\omega)\,h}{c} - \sqrt{1-\left(\frac{\mu(\omega)\,h}{c}\right)^2} = \hat{E}_q(\mu(\omega)), \tag{11a}$$

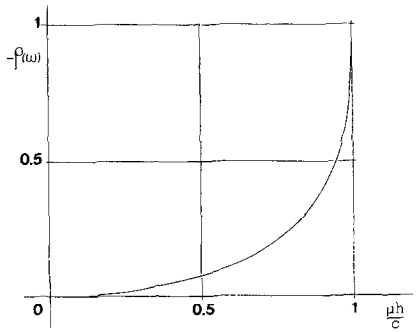

FIG. 1. Amplitude reflection ratio for the 2-point equation (40):
$\rho(\mu(\omega)) = (\sqrt{1-(\mu(\omega)(h/c))^2} - 1/(\sqrt{1-(\mu(\omega)(h/c))^2} + 1)$.

where $\hat{E}_p$ and $\hat{E}_q$ are the same operators as in (10) and (11), except for the difference in arguments.

In the particular case of the Crank–Nicolson method we shall have

$$\mu(\omega) = \frac{2}{\Delta t} \tan\left(\frac{\omega \Delta t}{2}\right). \tag{15}$$

The cut-off frequency is given in the fully discrete case by

$$\mu(\omega_c)\frac{h}{c} = 1 \tag{16}$$

or (Crank–Nicolson case)

$$\omega_c = \frac{2}{\Delta t} \arctan\left(\frac{c\Delta t}{2h}\right) \tag{16a}$$

3. ENERGY

The energy (or square of the l_2 norm) of $\{u_j\}$ on D is defined as:

$$\varepsilon_u \equiv h \sum_{j<0} |u_j|^2. \tag{17}$$

It may be expressed in Fourier space by the appropriate form of Parseval's relation

$$\varepsilon_u \equiv \int_{-\pi/h}^{\pi/h} |\overline{u}(\xi)|^2 \frac{d\xi}{2\pi} = \int_0^{\pi/h} |\overline{u}(\xi)|^2 \frac{d\xi}{\pi}, \tag{18}$$

where $\overline{u}(\xi)$ is the discrete x-Fourier transform of the set $\{u_n\}$ defined as

$$\overline{u}(\xi) = h \sum_{j<0} u_j e^{-i\xi jh}. \tag{19}$$

We note that those relations are generally given for functions defined on $(-\infty, \infty)$. That they also apply in the case of a semi-infinite domain such as D may be verified by simply assuming that the domain of definition is still $(-\infty, \infty)$, but that the numerical value of u is zero outside of D. (This remark shall also apply to the t-Fourier transforms used in (27), (28).)

The relationship of the wave number ξ to the frequency ω is given by the dispersion relation which is

$$\omega = -\frac{c}{h}\sin(\xi h) \tag{20}$$

in the semidiscrete case, and

$$\mu(\omega) = -\frac{c}{h}\sin(\xi h) \tag{21}$$

in the fully discrete case. That $\mu(\omega)$ is real for all ω in the Crank–Nicolson case results in the fact that energy is conserved in D: the fully discrete scheme is *energy conservative*, and the only changes in ε_u are those which may result from energy flow across the boundary.

The x-Fourier transform of rightgoing solutions has its support in $|\xi_p| \in [0, \pi/2h)$ and that of leftgoing solutions in $|\xi_q| \in (\pi/2h, \pi/h]$. This results in the important property that the leftgoing and rightgoing components of the energy are separated in Fourier space as

$$\begin{aligned}\varepsilon_u &= \int_0^{\pi/2h} |\overline{u}(\xi)|^2 \frac{d\xi}{\pi} + \int_{\pi/2h}^{\pi/h} |\overline{u}(\xi)|^2 \frac{d\xi}{\pi} \\ &= \varepsilon_p + \varepsilon_q.\end{aligned} \tag{22}$$

As for ε_u, both ε_p and ε_q are constant except for changes which may result from energy flow across the boundary.

4. Invariance Theorems

We may now state the energy reflection invariance theorems which are the main results of this paper:

Consider the semidiscretization (3) of (2) on the semi-infinite domain (1) with the semidiscrete equation (4) used at the outflow boundary $x = 0$. Consider also the

numerical integration of these semi-discrete equations by the Crank–Nicolson method.

When initial data $\{u_j(0)\}$ are imposed on D, the subsequent solution consists of a leftgoing and a rightgoing component. The rightgoing component of $\{u_j(0)\}$ is partially reflected at the boundary as a leftgoing solution and one may measure the performance of $\mathscr{B}$ by the smallness of the reflected energy

$$\varepsilon_R^\infty = \lim_{n\to\infty} h \sum_{j<0} |q_{j,R}^n|^2 \tag{23a}$$

or by the smallness of the energy remaining in D,

$$\varepsilon_u^\infty = \varepsilon_q^0 + \varepsilon_R^\infty = \lim_{n\to\infty} h \sum_{j<0} |u_j^n|^2, \tag{23b}$$

where ε_q^0 is the energy in the leftgoing component of $\{u_j(0)\}$. In that respect, we shall prove that:

THEOREM 1. *The total energy reflected by $\mathscr{B}$ is, for a constant h, independent of the value of Δt, and is strictly equal to ε_R^∞ for the semidiscrete model.*

A second theorem concerns convergence rates. Let $\{u_j(0)\}$ be obtained by sampling an initial $u(x, 0)$ at the mesh points $x_j = jh$ and let $h \to 0$. We may define an optimal boundary equation $\mathscr{B}$ as one which maximizes the rate of convergence to zero of the reflected energy in the semidiscrete case (3), (4). Then,

THEOREM 2. *The convergence rates of $\mathscr{B}$ when $h \to 0$ in the energy norm of the reflected solution for the semidiscrete and the fully discrete schemes are identical (and is, in the fully discrete case, independent of values taken by the ratio $\Delta t/h$). A boundary formula $\mathscr{B}$ which is optimal for the semidiscrete equations remains optimal for the corresponding full-discretizations.*

5. PROOF OF THEOREM 1

Reflection at the Boundary in the Semidiscrete Case

Consider an initial solution $\{u_j(0)\}$ in D. Its rightgoing component will, for large t, have passed entirely through the boundary point $x = 0$, where it will have been partially reflected into D as a leftgoing solution. We consider the semi-discrete case at first:

If $\hat{p}_0(\omega)$ and $\hat{q}_0(\omega)$ are the t-Fourier transforms of the incident and reflected solutions in $x = 0$, then the amplitude reflection ratio—obtained by taking the t-Fourier transform of (4) and using (10) and (11)—is found to be

$$\rho(\omega) \equiv \frac{\hat{q}_0(\omega)}{\hat{p}_0(\omega)} = -\frac{i\omega - b_0 - b_1 \hat{E}_p^{-1}(\omega) - \cdots}{i\omega - b_0 - b_1 \hat{E}_q^{-1}(\omega) - \cdots}. \tag{24}$$

During reflection, each rightgoing sinusoidal component of wavenumber $|\xi_p|$ in $[0, \pi/2h)$ generates a leftgoing sinusoidal component of wavenumber $|\xi_q|$ in $(\pi/2h, \pi/h]$ corresponding to the same value of ω through the dispersion relation (20), i.e.,

$$|\xi_q| = \frac{\pi}{h} - |\xi_p|, \tag{25}$$

where ξ_p and ξ_q have the same sign.

Moreover (also through the dispersion relation), the corresponding group velocities may be verified to satisfy

$$G(\xi_p) = -G(\xi_q). \tag{26}$$

If $\overline{p}(\xi_p, 0)$ is the x-Fourier transform of the rightgoing component of $\{u_j(0)\}$, then its amplitude is related to that of the t-Fourier transform of the resulting $p_0(t)$ as the boundary by the relation which is a consequence of energy conservation [7]

$$|\hat{p}_0(\omega)| = |\overline{p}(\xi_p, 0)/G(\xi_p)|. \tag{27}$$

Likewise, if $\hat{q}_0(\omega)$ is the t-Fourier transform of the reflected solution at the boundary, then its amplitude is related to that of the resulting $\overline{q}_R(\xi_q, t)$ by [7]

$$\lim_{t \to \infty} |\overline{q}_R(\xi_q, t)/G(\xi_q)| = |\hat{q}_0(\omega)|. \tag{28}$$

We may combine these results to give

$$\begin{aligned} \lim_{t \to \infty} |\overline{q}_R(\xi_q, t)| &= |\rho(\xi_p)\overline{p}(\xi_p, 0)\, G(\xi_q)/G(\xi_p)| \\ &= |\rho(\xi_p)\overline{p}(\xi_p, 0)|, \end{aligned} \tag{29}$$

where $\rho(\xi)$ (instead of $\rho(\omega)$) is derived with the dispersion relation (20), i.e., by substituting $-(c/h)\sin(\xi h)$ for ω in the expression (24) of the reflection ratio,

$$\rho(\xi) = \rho\left(\omega = -\frac{c}{h}\sin(\xi h)\right) = \rho_s(\xi) \tag{30}$$

(the subscript s refers here to the semidiscrete case). What this establishes is that the amplitude of the reflection ratio also applies to the amplitude of x-Fourier transform in $t = 0$ and $t \to \infty$, respectively. And the total energy reflected at the boundary may therefore be expressed as

$$\begin{aligned} \varepsilon_{R,S}^{\infty} &= \lim_{t \to \infty} h \sum_{j<0} |q_{j,R}(t)|^2 = \lim_{t \to \infty} \int_{\pi/2h}^{\pi/h} |\overline{q}_R(\xi, t)|^2 \frac{d\xi}{\pi} \\ &= \int_0^{\pi/2h} |\rho_s(\xi)\overline{p}(\xi, 0)|^2 \frac{d\xi}{\pi}. \end{aligned} \tag{31}$$

Reflection at the Boundary in the Fully Discrete Case

The mathematics leading to the derivation of (30), (31) may be repeated identically in the fully discrete case, with $\mu(\omega)$ replacing ω, and we obtain for the amplitude reflection ratio (the subscript F refers here to the fully discrete case)

$$\begin{aligned}\rho_F(\omega) &= \rho_s(\mu(\omega)) \\ &= -\frac{i\mu(\omega) - b_0 - b_1 \hat{E}_p^{-1}(\mu(\omega)) - \cdots}{i\mu(\omega) - b_0 - b_1 \hat{E}_q^{-1}(\mu(\omega)) - \cdots},\end{aligned} \tag{32}$$

which is identical to (24), except for the difference in arguments. We may express the energy reflected at the boundary by the equation which corresponds to (31),

$$\begin{aligned}\varepsilon_{R,F}^{\infty} &= \lim_{n\to\infty} h \sum_{j<0} |q_{j,R}^n|^2 = \lim_{n\to\infty} \int_{\pi/2h}^{\pi/h} |\overline{q_R^n}(\xi)|^2 \frac{d\xi}{\pi} \\ &= \int_0^{\pi/2h} |\rho_F(\xi)\,\overline{p^0}(\xi)|^2 \frac{d\xi}{\pi},\end{aligned} \tag{33}$$

where $\rho_F(\xi)$ is to be derived from $\rho_F(\omega)$ by use of the dispersion relation (21). That is, $\mu(\omega)$ is to be replaced by $-(c/h)\sin(\xi h)$. But we may observe that the corresponding expression is identical to (30) obtained in the semidiscrete case,

$$\begin{aligned}\rho_F(\xi) &= \rho_F\left(\mu(\omega) = -\frac{c}{h}\sin(\xi h)\right) \\ &= \rho_s(\xi).\end{aligned} \tag{34}$$

That is, *the amplitude reflection ratio (in x-Fourier space) for the full- and semidiscretizations are identical.* Thus, the expression of the reflected energy is (31) in the fully discrete as well as in the discrete cases, and, therefore

$$\varepsilon_{R,F}^{\infty} = \varepsilon_{R,S}^{\infty}, \tag{35}$$

which proves the theorem.

6. Convergence Rates and Proof of Theorem 2

The proof of Theorem 2 is of course contained in (35), but deserves some amplification:

Consider an initial $U(x, 0)$ in D, and the discrete set obtained by its sampling

$$\{u_j(0)\} = U(jh, 0).$$

In general, the Fourier transforms $\overline{u}(\xi, 0)$ and $\hat{U}(\xi, 0)$ will differ: components of

$\hat{U}(\xi, 0)$ beyond the sampling wavenumber $|\xi_s| = \pi/h$ are folded into $|\xi| \in |0, \pi/h|$ by the sampling relation. But the difference, measured in an energy norm, becomes negligible when $h \to 0$ for any well-behaved $U(x, 0)$.

The initial $\overline{u}(\xi, 0)$ is separated into its leftgoing and rightgoing components by the wave number $|\xi_c| = \pi/2h$. When $h \to 0$ then $|\xi_c| \to \infty$ and here also the difference between $\{u_j^0\}$ and $\{p_j(0)\}$ becomes negligible. We may thus write in the energy norm:

$$\|\hat{U}(\xi, 0) - \overline{p}(\xi, 0)\|_2 = O(h^k), \tag{37}$$

where k is a large number for any well-behaved $U(x, 0)$. Therefore, we may analyze convergence rates of the reflected solutions at the boundary when $h \to 0$ by (31) or (33) with $\hat{u}(\xi, 0)$ replacing $\overline{p}(\xi, 0)$ or $\overline{p^0}(\xi)$. It may then be verified algebraically that the rate of convergence to zero of the reflected energy at the boundary is twice that of $\rho(\xi)$ when $\xi h \to 0$, which is the same as twice that of $\rho(\omega)$ when $\omega h \to 0$.

One may define optimal boundary equations as those which, for a given number of terms in (4), maximize the order of the corresponding reflection ratio (examples of such high-order boundary equations may be found in [9]). From (35) it follows that both rates of convergence and optimally are preserved when one goes from the semidiscrete to the corresponding fully discrete case. Which proves Theorem 2.

7. Numerical Experiments

Smooth solutions may be considered as wave packets of wavenumber $|\xi_p| \to 0$. To these correspond reflected solutions of wave number $|\xi_q| \to \pi/h$. These reflected solutions are modulated sine functions, of wavelength $\lambda \to 2h$, with a characteristic sawtoothed appearance. This case is treated in the following numerical example:

The Gaussian initial function

$$U(x, 0) = e^{-(1/2)[(x - x_0)/\sigma]^2}, \qquad x_0 = -50, \sigma = 10 \tag{38}$$

was prescribed on the domain

$$D \equiv [-100, 0] \tag{39}$$

on which $U_t + cU_x = 0$ is approximated by (3) with $h = 1$ in D, and the two-point equation:

$$\frac{du_0}{dt} + c\left(\frac{u_0 - u_{-1}}{h}\right) = 0 \tag{40}$$

at the downwind boundary. Time marching is implemented with the Crank–Nicolson method. Initial conditions for the numerical calculation are obtained by sampling $U(x, 0)$ at the mesh points.

Some comments are in order. First, it may be verified that the energy of the Gaussian (38) which lies outside of the finite domain D is less than 10^{-10} times its total energy. Second, the energy of the Fourier transform of $U(x, 0)$:

$$|\hat{U}(\xi, 0)| = \sqrt{2\pi}\,\sigma\, e^{-(1/2)\sigma^2\xi^2} \tag{41}$$

which lies outside of the p band $|\xi_p| \in [0, \pi/2h)$ is less than 10^{-11} times this total energy. The asymptotic approximations

$$\hat{U}(\xi, 0) = \overline{u}(\xi, 0) = \overline{p}(\xi, 0), \tag{42}$$

$$\overline{q}(\xi, 0) = 0, \tag{43}$$

thus hold to within the accuracy of the calculation (8 significant digits). The time evolution of the numerical solution is illustrated in Fig. 2.

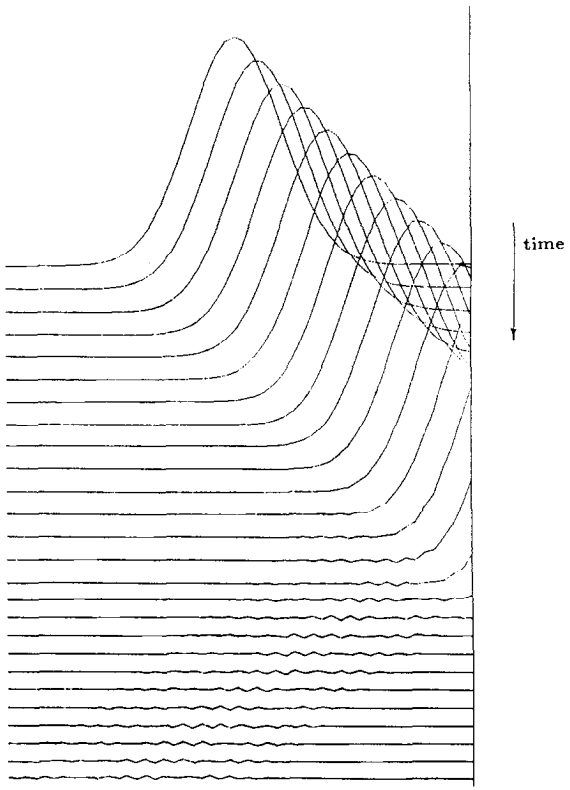

FIG. 2. Partial reflection of a smooth solution at the boundary (two point boundary equation (40), Courant number $R = 0.1$). Both incident and reflected solutions are wave packets, of wave number near $\xi_p = 0$, and $\xi_q = \pi/h$, respectively. The group velocity of the reflected solution is $G_q(\omega = 0) = -C$. Its phase velocity is zero.

Both incident and reflected wave packets have, to within computing accuracy, a finite support in D. This fact is illustrated in Fig. 3 giving the measured energy

$$\varepsilon_u^n = h \sum_{j<0} |u_j^n|^2 \tag{44}$$

versus time (see also Table I), which verifies the zero time-derivative of ε_u near $t=0$ (when the initial solution has not yet reached the boundary) and after $t=80$ when reflection has been completed.

The asymptotic value of the reflected energy (measured by (44) for large n) agrees to within arithmetical accuracy with the integral (31) evaluated by numerical quadrature ($\varepsilon_R^\infty = 1.37509...10^{-3}$).

Verification of Theorem 1 is obtained by repeating the calculation with values of Δt corresponding to a Courant number

$$R \equiv \frac{c\Delta t}{h} \tag{45}$$

varying from $R=0.05$ to $R=7.0$. The variation of the energy with time measured with (44) (see Fig. 4 and Table I) shows that, as expected, the nature of the numerical solution is affected by changes in Δt. *But the asymptotic value of the reflected energy* ε_R^∞ *is indeed an invariant.*

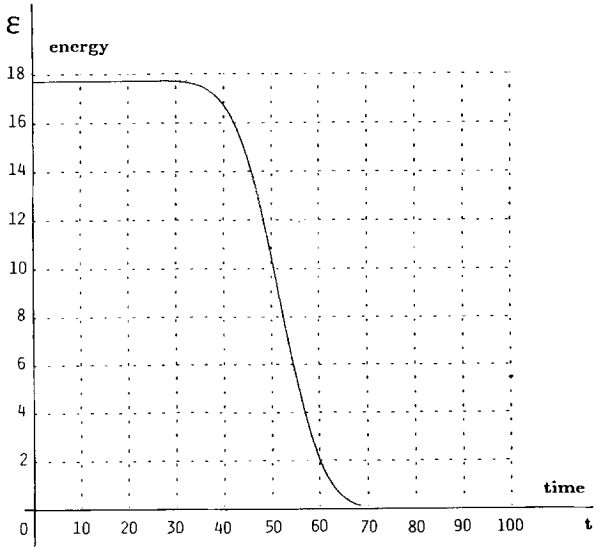

FIG. 3. Energy versus time for the example illustrated in Fig. 2. The final value of ε (which is the reflected energy) does not show on this figure (linear scale) but appears clearly on Fig. 4 (logarithmic scale).

TABLE I

Illustration of the Invariance of the Reflected Energy to Time-Discretization

TIME	R=0.05	R=0.1	R=0.25	R=0.5	R=1.0	R=2.5	R=5.0
0	17.7245385090	17.7245385090	17.7245385090	17.7245385090	17.7245385090	17.7245385090	17.7245385090
10	17.7245384519	17.7245384516	17.7245384500	17.7245384435	17.7245384114	17.7245377037	17.7245172986
20	17.7243907194	17.7243903958	17.7243881192	17.7243798356	17.7243443577	17.7239996857	17.7218880673
30	17.6942372434	17.6942171350	17.6940765826	17.6935774971	17.6916249278	17.6795483815	17.6499573633
40	16.7225046028	16.7226600849	16.7237429103	16.7275326079	16.7415916659	16.8094617482	16.9136909156
50	10.4718970857	10.4740035843	10.4887087043	10.5406533943	10.7396944302	11.8139470243	13.6081147130
60	2.0632558711	2.0644855210	2.0731317144	2.1045531011	2.2380626428	3.3634704798	6.8904879258
70	0.0454293388	0.0453789107	0.0450323667	0.0438906746	0.0410381092	0.1245621968	1.5600331854
80	0.0013977403	0.0013987422	0.0014065798	0.0014473863	0.0018908281	0.0520874615	0.8034780424
90	0.0013753155	0.0013753248	0.0013753994	0.0013758453	0.0013850574	0.0049727478	0.5669547926
100	0.0013750998	0.0013751000	0.0013751014	0.0013751115	0.0013754591	0.0021635676	0.1483119093
110	0.0013750964	0.0013750964	0.0013750964	0.0013750965	0.0013751130	0.0016504968	0.1513534968
120	0.0013750963	0.0013750963	0.0013750963	0.0013750963	0.0013750970	0.0014350113	0.0413352158
130	0.0013750963	0.0013750963	0.0013750963	0.0013750963	0.0013750964	0.0013855910	0.0429712909
140	0.0013750963	0.0013750963	0.0013750963	0.0013750963	0.0013750963	0.0013770067	0.0132338356
150	0.0013750963	0.0013750963	0.0013750963	0.0013750963	0.0013750963	0.0013753360	0.0135306622
160	0.0013750963	0.0013750963	0.0013750963	0.0013750963	0.0013750963	0.0013752239	0.0054693823
170	0.0013750963	0.0013750963	0.0013750963	0.0013750963	0.0013750963	0.0013751383	0.0047606622
180	0.0013750963	0.0013750963	0.0013750963	0.0013750963	0.0013750963	0.0013751095	0.0031063124
190	0.0013750963	0.0013750963	0.0013750963	0.0013750963	0.0013750963	0.0013751020	0.0022183865

Note. These numbers, which contain those used to obtain Fig. 4, have been obtained by numerical integration of (3)–(40) with the Crank–Nicolson method, $h = \text{constant}$ and variable Δt or R. The final value $\varepsilon_R^\infty = 1.375...10^{-3}$ is also obtained to within arithmetical accuracy by numerical quadrature of (31–(41).

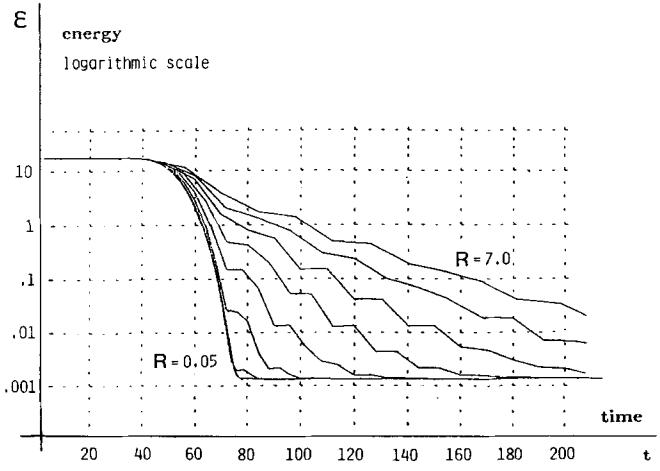

FIG. 4. Energy vs. time for the same numerical experiment repeated for increasing values of Δt (or Courant number R).

8. Discussion of Other Cases

While the invariance theorems have been established for the 3 point central difference Crank–Nicolson method, their applicability extends obviously beyond that simple case. A first case of interest is the *leapfrog method* of time marching which is also of the form (5) with **M** given by

$$\mathbf{M} = \frac{\mathbf{Z} - \mathbf{Z}^{-1}}{2\Delta t}. \tag{46}$$

Its spectral function is

$$\mu(\omega) = \frac{\sin(\omega \Delta t)}{\Delta t}. \tag{47}$$

This method, used with the central difference discretization (3) is known to be stable when $R \leqslant 1$. That $\mu(\omega)$ is real (when stability exists) indicates that this method is also energy conservative in D. But to each ξ there corresponds now through the dispersion relation two values of the frequency ω. One is in

$$|\omega| \leqslant \omega_c = \frac{1}{\Delta t} \operatorname{arc\,sin}(R) \tag{48}$$

(which may be called the consistent band) and the other is in

$$|\omega| \in \left[\frac{\pi}{\Delta t} - \omega_c, \frac{\pi}{\Delta t} \right] \tag{49}$$

(which may be called the spurious band). There are four types of fundamental solutions: two with positive group velocities, and two with negative group velocities. Rightgoing solutions in the consistent band are reflected as leftgoing solutions with an amplitude reflection ratio $\rho_F(\omega)$ given by (32). Rightgoing solutions in the spurious band corresponding to the same pair $\xi_p - \xi_q$ are reflected with an amplitude reflection ratio equal to $1/\rho_F(\omega)$.

In particular rightgoing solutions in the spurious band corresponding to $\xi_p = 0$ have an infinite reflection ratio at the boundary: This results in numerical instability [3, 4]. While one could argue that Theorems 1 and 2 continue to apply to solutions restricted to the consistent band (and one can, indeed, construct such solutions), numerical instability in the other band makes the argument somewhat academic.

But one may find other energy conservative time marching methods for which the theorems continue to apply strictly. A sufficient condition is that the corresponding $\mu(\omega)$ be a single-valued real monotonic function of ω. (consistency requires also that $\mu(0) = 0$; $\mu'(0) = 1$.)

An example is given by the method which consists in dividing Δt in two nonequal intervals

$$\beta \Delta t \quad \text{and} \quad (1 - \beta)\, \Delta t \quad (0 < \beta < 0.5) \tag{50}$$

and using the Crank–Nicolson algorithm over each subinterval, resulting in

$$\mathbf{M}^2 = \frac{4}{\beta(1-\beta)\,\Delta t^2}\left(\frac{\mathbf{Z}^\beta - 1}{\mathbf{Z}^\beta + 1}\right)\left(\frac{\mathbf{Z}^{1-\beta} - 1}{\mathbf{Z}^{1-\beta} + 1}\right),$$

$$\mu(\omega) = \left[\frac{4}{\beta(1-\beta)\,\Delta t^2}\tan\left(\frac{\beta\omega\Delta t}{2}\right)\tan\left(\frac{(1-\beta)\,\omega\Delta t}{2}\right)\right]^{1/2}. \tag{51}$$

The case of *nonenergy conservative methods* is also of interest. An example is given by the first-order Adams method:

$$\begin{aligned} u_{j,p}^{n+1} &= u_j^n + \Delta t\mathbf{A}^* \cdot u_j^n, \\ u_j^{n+1} &= u_j^n + \Delta t\mathbf{A}^* \cdot u_{j,p}^{n+1}, \end{aligned} \tag{52}$$

which may also be expressed as

$$u_j^{n+1} - u_j^n = [\Delta t\mathbf{A}^* + (\Delta t\mathbf{A}^*)^2] \cdot u_j^n. \tag{53}$$

This method is known to be numerically stable with (3) when $R \leqslant 1$. But it is energy dissipative. Moreover, we may observe that it is not of the separable form

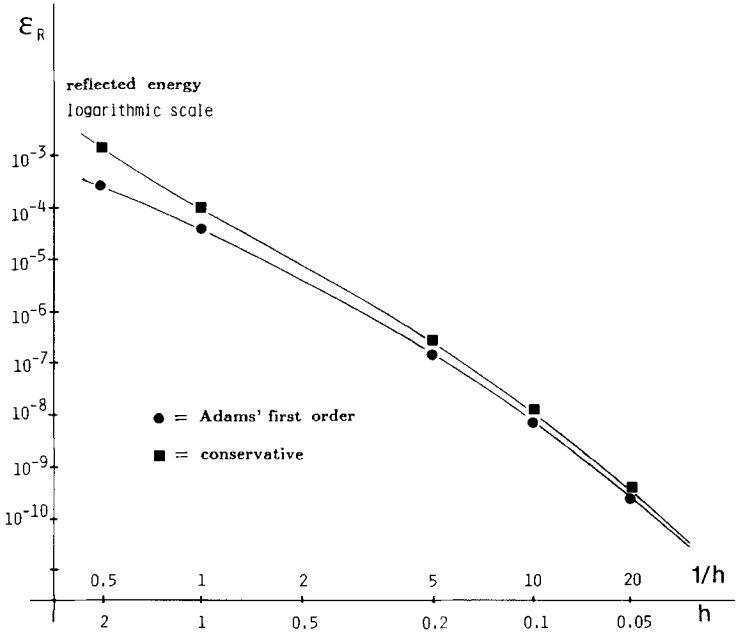

FIG. 5. Comparison of the reflected energy for the Adams first-order (dissipative) and the conservative (semi-discrete or Crank–Nicolson) methods. Initial conditions are those of Fig. 2 (or Eq. (38)). Courant number $R = 0.4$ in all cases. The difference of reflected energies goes to zero as $h \to 0$ verifying the same rate of convergence for both. ε_R is measured in $t = 100$ in the Adams case, and computed by numerical quadrature of (31)–(41) in the conservative case.

(5). Theorem 1 ceases to hold, but Theorem 2 may be shown to still apply modulo a minor modification of definition as follows:

Let $U(x, 0)$ be (save for possible remainders of negligible energy) of finite support near $x = 0$. in D and of finite support near $\xi = 0$ in x-Fourier space, i.e.,

$$\begin{aligned} \hat{U}(\xi, 0) &\simeq 0 \qquad \text{when} \quad |\xi| > \alpha, \\ U(x, 0) &\simeq 0 \qquad \text{when} \quad x < -\delta \end{aligned} \tag{54}$$

(α and δ positive and finite). For some small h, there is then a finite time τ at which reflection is completed (except for a possible remainder of negligible energy). Also, stability of (52) requires that $\Delta t \to 0$ as $h \to 0$, and solutions of that equation tend to solutions of (3), (4) in $t \in [0, \tau]$. Theorem 2, which is concerned with convergence rates, may be shown to hold, provided that ε_R be defined in $t = \tau$ instead of $t \to \infty$.

The proof is reasonably straightforward and shall be omitted. Given in Fig. 5, however, is an experimental verification of this interesting result. It is of course the case that these theorems may be extended to other spatial discretization schemes, but this may require a modification of the definition of energy. One may find in [10] the underlying mathematics which apply to the case of semidiscretizations obtained with linear finite elements, where such a modified form of conserved energy is found.

ACKNOWLEDGMENTS

Acknowledgments are due to E. C. Parriser who wrote the computer programs which resulted in several of the figures given in this paper.

REFERENCES

1. L. BRILLOUIN, "Wave Propagation and Group Velocity," Academic Press, New York, 1960.
2. B. ENGQUIST AND A. MAJDA, *Math. Comput.* **31** (1977), 629.
3. L. N. TREFETHEN, "Wave Propagation and Stability for Finite Difference Schemes," Ph.D. Thesis, Stanford University, 1982.
4. L. N. TREFETHEN, *J. Comput. Phys.* **49**(1983).
5. R. VICHNEVETSKY, *in* "Mathematics and Computers in Simulation," Vol. 23, pp. 333–343, North-Holland, Amsterdam, 1981.
6. R. VICHNEVETSKY, *J. Franklin Inst.* **315** (5/6) (1983), 307.
7. R. VICHNEVETSKY, *in* "Mathematics and Computers in Simulation," Vol. XVII, pp. 93–101, North-Holland, Amsterdam, 1984.
8. R. VICHNEVETSKY AND J. B. BOWLES, "Fourier Analysis of Numerical Approximations of Hyperbolic Equations," SIAM, Philadelphia, Penn., 1982.
9. R. VICHNEVETSKY AND E. C. PARISER, *J. Comput. Math. Appl.* **11** (1985), 67–78.
10. R. VICHNEVETSKY, *J. Comput Math. Appl.* **11** (1985), 733–746.