JOURNAL OF COMPUTATIONAL PHYSICS 19, 229-235 (1975)

Note

Nonlinear Transformations and the Numerical Treatment of Shocks

In solving different equations, changes of dependent variables are often made to simplify the problem, or, as in [8], to ensure the positivity of these variables. However, for nonlinear hyperbolic equations such changes lead to different solutions when shocks are formed. In a previous paper [9], an analysis was made of the effect of such nonlinear transformations on numerical solutions computed by the Lax-Friedrichs method, via the "modified equation" approach [1, 7]. For the L-F scheme, which is of first order accuracy, it was shown that, if the first term in the truncation error is properly taken into account when a nonlinear change of dependent variables is made, the weak solution of the original equations is preserved.

The essence of the modified equation method is the following.

One finds another differential equation which is solved by the difference scheme to a higher order of accuracy than the original equation. The modified equation then consists of the original equation plus terms involving increasing powers of the grid size.

Thus, for the Lax-Friedrichs scheme the first extra term gives a differential equation solved to second order accuracy. This term involves derivatives up to second order multiplied by Δt; when linearized, it becomes a linear parabolic term.

When investigating second order accurate schemes, such as Lax-Wendroff's [3], one finds that, in order to obtain an analogous parabolic term in a linearized analysis, it is necessary to include two truncation terms, one with Δt^2 and one with Δt^3. This would lead one to expect that both terms must be taken into account, if one wants to make a nonlinear transformation and still obtain the original weak solution. However, the investigation described here reveals that it is sufficient to use the first term. This indicates that the true effects of the terms are not completely revealed by a linearized analysis but are influenced strongly by their nonlinear structure. This point is further reinforced by an examination of the first term itself, before and after the transformation. The difference between the two consists of two nonlinear expressions of apparently dissimilar character, as specified below. It is shown that both of these significantly influence the behavior of the shocklike discontinuities.

As a typical problem, we consider the equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = 0, \tag{1}$$

with initial data of the form

$$u(x, 0) = \phi(x) = \begin{cases} c, & x \leqslant -1, \\ 1 - (c-1)\,x, & -1 \leqslant x \leqslant 0, \\ 1 + (c-1)\,x, & 0 \leqslant x \leqslant 1, \\ c, & 1 \leqslant x, \end{cases} \tag{2}$$

with constant $c > 1$.

These initial data were chosen because the solution can be obtained analytically and consists of a rarefaction wave (classical behavior) followed by a compression wave from which a shock is formed at $t = 1/(c-1)$.

Specifically, the analytic solution is given by

$$u(x, t) = \begin{cases} c, & x \leqslant ct - 1 \\ \dfrac{(c-1)\,x - 1}{(c-1)\,t - 1}, & ct - 1 \leqslant x \leqslant t \\ \dfrac{(c-1)\,x + 1}{(c-1)\,t + 1}, & t \leqslant x \leqslant ct + 1 \\ c, & ct + 1 \leqslant x \end{cases} \qquad \begin{pmatrix} 0 \leqslant t < \dfrac{1}{c-1} \\ \text{continuous} \end{pmatrix} \tag{3a}$$

and

$$u(x, t) = \begin{cases} c, & x \leqslant X_S(t) \\ \dfrac{(c-1)\,x + 1}{(c-1)\,t + 1}, & X_S(t) < x \leqslant ct + 1 \\ c, & ct + 1 \leqslant x \end{cases} \qquad \begin{pmatrix} \dfrac{1}{c-1} \leqslant t \\ \text{discontinuous} \end{pmatrix}, \tag{3b}$$

where $X_s(t)$ is the trajectory of the u-shock which is obtained from the Rankine–Hugoniot relation and is given by

$$X_s(t) = ct + 1 - [2(c-1)\,t + 2]^{1/2}. \tag{3c}$$

As our nonlinear transformation, we chose $v = u^2$, which, when applied to (1), leads to

$$\frac{\partial v}{\partial t} + \frac{\partial}{\partial x}\left(\frac{2}{3}\, v^{3/2}\right) = 0, \qquad v((x, 0) = [\phi(x)]^2. \tag{4}$$

Before shock formation, the exact solution for v is $u^2(x, t)$, with u given by (3a), and, after shock formation,

$$v = \begin{cases} c^2, & x \leqslant X_S(t) \\ \left[\dfrac{(c-1)\,x + 1}{(c-1)\,t + 1}\right]^2, & X_S(t) < x \leqslant ct + 1 \\ c^2, & ct + 1 \leqslant x \end{cases} \qquad \begin{pmatrix} \dfrac{1}{c-1} \leqslant t \\ \text{discontinuous} \end{pmatrix}, \tag{5a}$$

where X_s is the v-shock trajectory.

From the Rankine–Hugoniot relation, $X_s(t)$ is given by

$$X_s = czt + (cz - 1)/(c - 1), \tag{5b}$$

where z is the middle root of the cubic equation,

$$z^3 - 3z + 2 - (2/c^3) \cdot (2c + 1)(c - 1)^2/[(c - 1)\,t + 1] = 0, \tag{5c}$$

obtained via Cardan's formula.

For a second order accurate numerical scheme, we chose the Lax–Wendroff method (see [5], p. 302),

$$w_j^{n+1} = L_\Delta w_j^{\,n}, \tag{6}$$

for the differential equation $w_t + f(w)_x = 0$. If $\tilde{w}$ denotes the exact solution of the differential equation, then

$$\tilde{w}(x_j\,, t_{n+1}) = L_\Delta \tilde{w}(x_j\,, t_n) + \Delta t \cdot O(\Delta t^r), \tag{7}$$

where $r = 2$, since the scheme is of second order.

The modified equation for $\tilde{w}$ can be obtained by adding terms to the differential equation which, when combined with (6), will give an equation analogous to (7) with $r \geqslant 3$. For $r = 3$, the modified equation turns out to be

$$w_t + [f(w)]_x = -(\Delta t^2/6\lambda^2)[(1 - \lambda^2 a^2) f_x]_{xx}\,. \tag{8}$$

This formula is valid for the scalar case and for special systems in which a is a matrix which commutes with its derivatives. Otherwise, the right-hand side of (8) contains extra terms which are also $O(\Delta t^2)$ (see [6]).

Application of (8) to the u and v equations yields

$$u_t + (u^2/2)_x = (\Delta t^2/6\lambda^2)[u(\lambda^2 u^2 - 1)\, u_{xxx} + 3(3\lambda^2 u^2 - 1)\, u_x u_{xx} + 6\lambda^2 u u_x{}^3], \tag{9}$$

and

$$\begin{aligned} v_t + (\tfrac{2}{3} v^{3/2})_x = (\Delta t^2/24\lambda^2)[&4v^{1/2}(\lambda^2 v - 1)\, v_{xxx} + 6v^{-1/2}(3\lambda^2 v - 1)\, v_x v_{xx} \\ &+ v^{-3/2}(3\lambda^2 v + 1)\, v_x{}^3]. \end{aligned} \tag{10}$$

However, if the transformation $v = u^2$ is now applied to (9), the result is the equation

$$v_t + (\tfrac{2}{3} v^{3/2})_x = (\Delta t^2/24\lambda^2)[4v^{1/2}(\lambda^2 v - 1)\, v_{xxx} + 12\lambda^2 v^{1/2} v_x v_{xx}]. \tag{11}$$

Note that the right sides of (10) and (11) are *not* the same. They agree with respect to the v_{xxx} term (which would have been predicted by a linearized analysis), but differ in the nonlinear lower derivative terms. Equations (9) and (10) are the equations whose solutions the L–W scheme approximates to one higher degree of accuracy than the original ones. When dealing with solutions which include shocks, the two will not agree, after the time of shock formation. When computing with schemes which can handle shocks, such as Lax–Wendroff's, the numerical procedure gives, in every case, the corresponding but different shock solution.

Now suppose a correction term is added to the numerical scheme for v such that (11), rather than (10), becomes the equation which is approximated to the next higher order of accuracy. Then, as will be demonstrated, the square root of the numerical solution for the equation with this correction term agrees with the solution for u, *even after a shock is formed.* The correction term is the difference between the right sides of (11) and (10), i.e.,

$$D = (\Delta t^2/24\lambda^2)(v_x/v^{3/2})[6(1 - \lambda v^2)\, v v_{xx} - (1 + 3\lambda^2 v)\, v_x{}^2]. \tag{12}$$

The significant fact here is that a proper correction is achieved with the use of only the first term in the truncation error, while one might have supposed that two terms would have been necessary.

The reason for at first believing that two truncation terms would be needed is that, in linear stability analysis, both terms are needed to obtain an expression which may be considered dissipative (see [5, pp. 330–332]). In the Lax–Friedrichs scheme, the first term is already dissipative [9].

The fact that a proper correction will be obtained here with only one term indicates that the effects of transformations on the computation of solutions with shocks are genuinely nonlinear, since $D = 0$ in a linearized analysis. Moreover, the nature of the linearized terms in the truncation error (dissipative, dispersive), while of importance in stability considerations, does not seem to be important here. This can be seen by the fact that in the Lax–Friedrichs case, the first term is dissipative, while for Lax–Wendroff, it is $(-\Delta t^2/6\lambda^2)(1 - \lambda^2 a^2)\, a w_{xxx}$, which is dispersive.

It is worth noting that the same expression is obtained as the linearized first truncation term in every explicit three-point second order accurate scheme (see [4]).

The following results were obtained on the computer.

(i) The Lax–Wendroff numerical solution for the u-problem (1) with (2), and a computation of the analytical solution (3) for comparison;

(ii) The square root of the numerical solution for the v-problem (4), and also the square root of the corresponding analytical solution;

(iii) The square root of the numerical solution for the v-problem, with the correction term D given in (12), properly discretized.

These programs were run with various values of $c > 1$.

It was found, for (i) and (ii), that the respective numerical and analytical solutions gave excellent agreement, before and after shock formation. Also $v^{1/2}$ and u agreed up to the time of formation of the shock, i.e., $t_s = 1/(c - 1)$. Then, as expected for $t > t_s$, the v-shock traveled faster than the u-shock. Thus, the scheme in each case produced the solution appropriate to the dependent variable used, in accordance with the theory of weak solutions.

In (iii), the numerical solution agree with (i) and (ii) up to the time of shock formation, t_s. For $t > t_s$, this numerical solution agree with the numerical solution for u in (i), thereby demonstrating that a full correction is obtained in the manner described above (compare Fig. 1 with Fig. 3).

A few typical plots illustrating these results are shown in Figs. 1–3. In these plots c was taken to be 6, $\Delta x = 0.04$, and $\Delta t = 0.006$, which is 0.9 times the maximal stable time step.

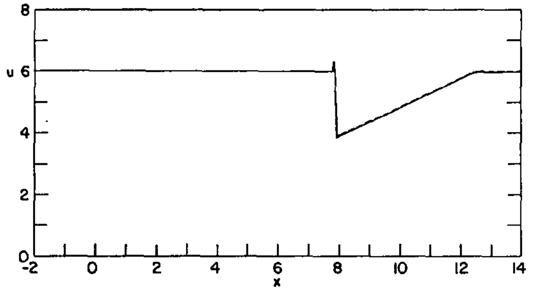

FIG. 1. The numerical solution (solid line) and the analytical solution (dashed line) for u, after 320 time steps. The shock is at $X_s = 7.9$.

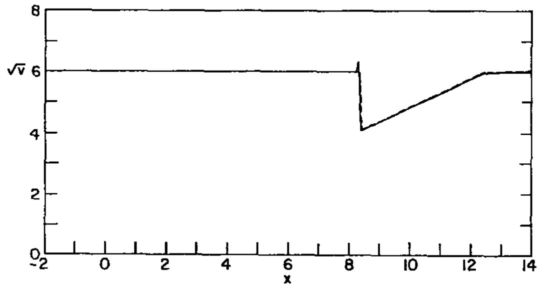

FIG. 2. The square root of the numerical solution (solid line) and of the analytical solution (dashed line) for v, after 320 time steps. The shock is at $X_s = 8.4$.

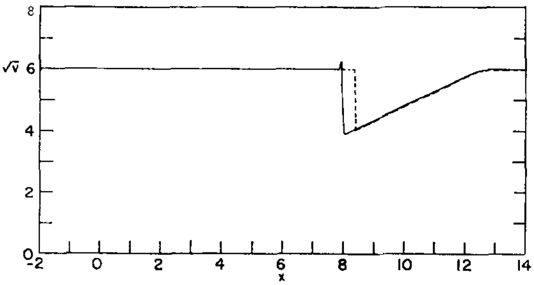

FIG. 3. The square root of the numerical solution (solid line) for v, obtained *with* the correction term, and the square root of the analytical solution (dashed line) for the original v-problem (given by (6)). The numerical shock here is at $X_s = 7.9$.

Writing $D = D_1 + D_2$, where

$$D_1 = (\Delta t^2/4\lambda^2)\, v^{-1/2}(1 - \lambda^2 v)\, v_x v_{xx}\,, \quad D_2 = -(\Delta t^2/24\lambda^2)\, v^{-3/2}(1 + 3\lambda^2 v)\, v_x^3, \tag{13}$$

we also investigated the effects produced by D_1 and D_2 above. It was found that the desired correction is obtained by the combination of the two and that both terms, D_1 and D_2, are significant.

On the basis of ideas obtained from linear analysis, D_1 and D_2 appear to be of quite different natures. One is tempted to call D_1 "dissipative" and D_2 "dispersive." However, the work here indicates that the true state of affairs is genuinely nonlinear and such descriptions may be too simplistic. In this regard, Kreiss and Oliger [2], in discussion numerical solutions of nonlinear equations, state "computational difficulties are apt to be wrongly ascribed. This can easily lead to a large and incorrect folklore which can steer future research in the wrong direction."

If some physical dissipation is included, then discontinuous solutions do not arise and a change of variables is permissible. However, if numerical diffusion-like terms are subtracted in order to sharpen up the results, an incorrect solution may arise since small alterations of the truncation errors can produce qualitative changes in the solution.

REFERENCES

1. C. W. HIRT, *J. Computational Phys.* **2** (1968), 339.
2. H. KREISS AND J. OLIGER, "Methods for the Approximate Solution of Time Dependent Problems," GARP Publications Series No. 10, Geneva, 1973.
3. P. LAX AND B. WENDROFF, *Comm. Pure Appl. Math.* **13** (1960), 217.
4. A. LERAT AND R. PEYRET, *C. R. Acad. Sci. Paris* **276** (1973), 759.
5. R. RICHTMYER AND W. MORTON, "Finite Difference Methods for Initial Value Problems," Interscience, New York, 1967.

6. B. VAN LEER, Ph.D. Thesis, Leiden University, Netherlands, 1970.
7. R. F. WARMING AND B. J. HYETT, *J. Computational Phys.* **14** (1974), 159.
8. J. P. WRIGHT, in "Proceedings of the Third International Conference on Numerical Methods in Fluid Mechanics" (July 1972), Vol. I, p. 169.
9. G. ZWAS AND J. ROSEMAN, *J. Computational Phys.* **12** (1973), 179.

RECEIVED: February 20, 1975; REVISED April 7, 1975

JOSEPH ROSEMAN
AND
GIDEON ZWAS

Department of Mathematical Sciences
Tel-Aviv University
Ramat-Aviv, Israel