Nonlinear Truncation Error Analysis of Finite Difference Schemes for the Euler Equations

Goetz H. Klopfer*
Nielsen Engineering & Research, Inc., Mountain View, California
and
David S. McRae†
NASA Ames Research Center, Moffett Field, California

Finite difference solutions to nonlinear equations governing fluid flow often exhibit large errors in the vicinity of discontinuities or steep gradients. The present study details a technique for analyzing these errors through use of the modified equation approach. Once the leading error terms are identified, a procedure is demonstrated whereby these terms can be removed as the solution proceeds. The resulting corrected solution is shown to be much improved at discontinuities. In order to illustrate this technique, the modified equation is developed to fourth order for both general classes of two-step Lax-Wendroff schemes and the implicit schemes due to Beam and Warming. Both are applied to the one-dimensional Euler equations. Solutions of the one-dimensional shock tube problem are obtained, both with and without removal of the leading error terms. Numerical evaluation of the errors for the Lax-Wendroff schemes reveals that a nearly third-order accurate scheme can be obtained very readily by removal of only the leading error term. Correction of the implicit schemes by explicit means does not result in a stable scheme. Results of implicit correction for the implicit scheme are shown to be stable only for CFL > 1 by linear analysis.

Introduction

IN general, dissipative finite difference integration schemes have been found to be quite robust when applied to the Euler equations of gas dynamics. When the equations are written in conservation form, solutions obtained by use of these schemes have the remarkable feature of modeling discontinuities of the flowfield (such as shock waves and contact surfaces). This property has been used to advantage in many studies.[1-3] However, large errors can be present in the vicinity of these discontinuities, resulting in oscillatory behavior of the solution. The magnitude of these oscillations is dependent on the strength of the discontinuity; for the nonlinear equations, the oscillations are ultimately destabilizing, thereby severely limiting the ability of the integration technique to compute strong shock waves and contact surfaces. In addition, an error propagating from the vicinity of the discontinuities may distort other important features of the flow. Previous efforts to stabilize these nonlinear instabilities (apart from the normal requirements for linear stability) have used either various forms of additional dissipation terms[4] or the (often laborious) fitting of each discontinuity as it appears in the solution.[5] In the first instance, nonphysical solutions may result from the use of these dissipation terms as the apparent viscosity of the equation set is usually increased. In the second instance, fitting of each discontinuity as it arises presents a very difficult coding problem and prevents its use in universally applicable computer codes.

The present study details a modified equation analysis of both implicit and explicit finite difference techniques as applied to the Euler equations. The analysis is used to identify those error terms which contribute most to the observed solution errors. A technique for analytically removing the dominant error terms is demonstrated, resulting in a greatly improved solution for the explicit Lax-Wendroff schemes.

Analysis

The Euler equations in conservative form can be written as a general nonlinear system:

$$w_t + [f(w)]_x = 0 \tag{1}$$

where

$$w = (\rho, m, e)^T$$

$$f = \left[m,\ \left((\gamma - 1)e + \frac{3-\gamma}{2}\frac{m^2}{\rho}\right), \left(\gamma e - \frac{\gamma - 1}{2}\frac{m^2}{\rho}\right)\frac{m}{\rho}\right]^T$$

m/ρ is the fluid velocity, ρ is the fluid density, and e is the total energy per unit volume. The numerical schemes which will be examined are the generalized Lax-Wendroff scheme[6] and the implicit scheme of Beam and Warming.[7]

Generalized Lax-Wendroff Scheme

The generalized Lax-Wendroff scheme as given by Lerat and Peyret[6] is as follows:

$$\tilde{w}_j = (1-\beta) w_j^n + \beta w_{j+1}^n - \alpha\sigma (f_{j+1}^n - f_j^n)$$

$$w_j^{n+1} = \tilde{w}_j^n - (\sigma/2\alpha)\,[\,(\alpha - \beta) f_{j+1}^n + (2\beta - 1) f_j^n$$

$$+ (1 - \alpha - \beta) f_{j-1}^n + \tilde{f}_j - \tilde{f}_{j-1}] \tag{2}$$

where α and β are arbitrary parameters with $\alpha \neq 0$ and $\sigma = \Delta t/\Delta x$. The predicted value $\tilde{w}_j$ approximates the solution at $x = (j+\beta)\Delta x$ and $t = (n+\alpha)\Delta t$. The w and f vectors of Eqs. (2) are those of Eq. (1) evaluated at (for example) the discrete points $j\Delta x$ and $n\Delta t$.

Modified Equation

The modified equation results when the vectors in Eqs. (2) are expanded in a Taylor series. All time derivatives of a

Presented as Paper 81-0193 at the AIAA 19th Aerospace Sciences Meeting, St. Louis, Mo., Jan. 12-15, 1981; submitted March 10, 1981; revision received March 22, 1982.

*Research Scientist. Member AIAA.

†Major, U.S. Air Force, AFFDL Flight Vehicle Technology Office; at present: Associate Professor, North Carolina State University. Member AIAA.

higher degree than 1 may be eliminated by a series of linear operations using the modified equation itself. The procedure for the Lax-Wendroff scheme with a fixed time step is as follows[8]:

1) First the quantity $\tilde{f}$ of Eqs. (2) is expanded about the location j,

$$\tilde{f}_j = f(\tilde{w})_j = f(w_j^n) + f'(w_j^n)(\tilde{w} - w_j^n) + (1/2!)f''(w_j^n)(\tilde{w} - w_j^n)^2 + (1/3!)f'''(w_j^n)(\tilde{w} - w_j^n)^3 + \ldots \tag{3}$$

The quantity $w - \tilde{w}_j^n$ is cleared using the first of equations (2), which yields

$$\tilde{f}_j = f(w_j^n) + f'(w_j^n)[\beta(w_{j+1}^n - w_j^n) - \alpha\sigma(f_{j+1}^n - f_j^n)] + (1/2!)f''(w_j^n)[\beta(w_{j+1}^n - w_j^n) - \alpha\sigma(f_{j+1}^n - f_j^n)]^2 + (1/3!)f'''(w_j^n)[\beta(w_{j+1}^n - w_j^n) - \alpha\sigma(f_{j+1}^n - f_j^n)]^3 + \ldots \tag{4}$$

where $f' = \partial f/\partial w$ is the Jacobian matrix and the higher-order Jacobian matrices are

$$f'' = \partial^2 f/\partial w^2, \quad f''' = \partial^3 f/\partial w^3, \text{ etc.}$$

A similar expression results for $\tilde{f}_{j-1}$. The two series for the flux vectors evaluated at the predictor level are then substituted into the second of Eqs. (2). The resulting expression contains terms evaluated at the time levels n and $n+1$ only, so straightforward expansions of f and w can now be performed. After the Taylor series expansions of f and w are carried out and substituted into Eqs. (2), considerable algebra yields an expression containing the original differential equation plus higher-order space and time derivatives. This expression is called the modified equation[8] in this paper. A series of linear operations is then carried out to eliminate the higher-order time derivatives. These operations involve differentiating the modified equation, multiplying by the proper coefficients, and then subtracting the result from the modified equation in order to eliminate a specified higher-order time derivative. This process is repeated, in each instance using derivatives of the modified equation, until the higher time derivatives have been eliminated to the desired order. After collection of terms and algebraic simplification, the resulting equation is known as the modified equation[8] without the higher order time derivatives.

The modified equation with all terms retained is the differential form of the finite difference equation.

When this procedure is followed for the generalized Lax-Wendroff system as applied to the Euler equations, the resulting modified equation is

$$w_t + f_x + \frac{\Delta x^2}{6}\frac{\partial}{\partial x}\left\{f_{xx} + \frac{3}{2\alpha}\beta(\beta-1)f''(w_x,w_x) - \frac{3\sigma}{2}(2\beta-1)f''(w_x,f_x) + \sigma^2\left[\left(\frac{3\alpha}{2}-1\right)f''(f_x,f_x) - f'(f'f_{xx}) - f'f''(w_x,f_x)\right]\right\} + \frac{\Delta x^3}{24}\frac{\partial EX}{\partial x} + O(\Delta^4) = 0, \tag{5}$$

The third-order terms denoted by EX are given in the Appendix.

Previous studies have used the modified equation approach to generate new numerical schemes or to examine the shock wave structure for Lax-Wendroff-type schemes. Lerat[9] and Majda and Osher[10] have examined the scalar conservation equation (inviscid Burgers equation). Lerat added correction terms to obtain monotonic shock profiles and Majda and Osher have determined the general form of the stabilizing term to be included with the scheme to guarantee nonlinear stability. Lerat and Peyret[6] have considered the full system of nonlinear equations (one-dimensional Euler equations) to determine the optimum values of α and β. Warming et al.[4] have used the modified equation of the linear convection equation to minimize the dissipation or dispersive error for the third-order MacCormack-type schemes. However, none have used the full system of Euler equations to determine the correction terms necessary for improved accuracy.

In the present study the behavior of the higher-order terms (Δ^2 and Δ^3) in the modified equation is examined. This is accomplished by numerically evaluating the terms in Eq. (5) which do not appear in the original equation. This means that the derivatives appearing in these terms are approximated by second-order accurate finite differences. It should be noted that the numerical approximation of the derivatives in the additional terms does introduce error into the modified equation. However, the use of second-order accurate approximations for these terms places the error in fourth order and higher terms (at least for second-order accurate schemes such as the Lax-Wendroff schemes) and therefore should not affect the conclusions significantly. The terms in the modified equation which are contributing most to the inaccuracies of the solution can now be identified at any time step.

Since the Lax-Wendroff schemes can also be used to obtain steady-state solutions, it is of interest to obtain the truncation errors for this case. The modified equation is obtained by setting $w^{n+1} = w^n$ in Eqs. (2). It is up to third order as follows:

$$f_x - \frac{\Delta t}{2}\frac{\partial}{\partial x}[f'(f_x)] + \frac{\Delta x^2}{3!}\frac{\partial}{\partial x}\left(f_{xx} + \frac{3\beta(\beta-1)}{2\alpha}f''(w_x,w_x)\right) + \frac{3\Delta t\Delta x}{4}(1-2\beta)\frac{\partial}{\partial x}[f''(w_x,f_x)] + \frac{3\alpha\Delta t^2}{4}\frac{\partial}{\partial x}[f''(f_x,f_x)] = O(\Delta^3)$$

We note that the steady-state solution depends on the time step and furthermore that it is only first-order accurate. The first-order term is dissipative (and dispersive) provided the eigenvalues of the Jacobian matrix f' do not vanish. If they do, as for example at a stagnation or sonic point, then artificial dissipation must be appended to the scheme to guarantee stability.

Correction of the Original Differential Equation

Since the additional terms in the modified equation do not appear in the original differential equation, contributions from these terms will lead to inaccuracies when compared to an analytic solution of the original differential equation. In the present study a technique is developed for altering the original differential equation in order to remove the dominant error terms of the modified equation. The cancellation of these terms (if all those of a given order are removed) will result in a higher-order integration scheme.

As an example, assume that the $(\Delta x^2/6)\cdot f_{xx}$ term appearing in Eq. (5) has been observed to contribute substantially to the oscillatory behavior. This term can be removed from the modified equations by changing the original system of equations as follows:

$$w_t + [f_0(w)]_x = 0 \tag{6}$$

where $f_0 = f - (\Delta x^2/6)\cdot f_{xx}$. As noted above, additional terms in the modified equations result from this approach. However, they are of higher order (Δ^4 and above) for the Lax-Wendroff system.

Implicit Scheme

The implicit scheme considered in this paper is the scheme developed by Beam and Warming[7] with smoothing parameters added as by Desideri.[11] The scheme is a four-parameter scheme and is given by

$$\left(I+\frac{\bar{\theta}\Delta t}{1+\xi}\delta_x f'^{\,n}-\epsilon_i\nabla_x\Delta_x\right)(w^{n+1}-w^n)$$

$$=\frac{\xi}{1+\xi}(w^n-w^{n-1})-\frac{\Delta t}{1+\xi}\delta_x f^n-\epsilon_e(\nabla_x\Delta_x)^2 w^n \qquad (7)$$

where δ_x and ∇_x define the usual difference operators:

$$\delta_x F_k^n=(F_{k+1}^n-F_{k-1}^n)/2\Delta x$$

$$\Delta_x\nabla_x F_k^n=F_{k+1}^n-2F_k^n+F_{k-1}^n$$

$$(\Delta_x\nabla_x)^2 F_k^n=F_{k+2}^n-4F_{k+1}^n+6F_k^n-4F_{k-1}^n+F_{k-2}^n$$

The two smoothing parameters are ϵ_e and ϵ_i for the explicit and implicit smoothing, respectively. Scheme (7) encompasses the three well-known formulas for the time integration:

$\bar{\theta}=½$, $\xi=0$ trapezoidal rule

$\bar{\theta}=1$, $\xi=0$ Euler implicit

$\bar{\theta}=1$, $\xi=½$ three-point backward

The scheme is second-order accurate in time if $\bar{\theta}=\xi+½$. The trapezoidal rule and the three-point backward scheme are two examples of second-order time accurate implicit schemes. Note that the implicit smoothing consists of a second spatial derivative instead of a fourth spatial derivative, as for the explicit smoothing. This limit to second derivative is required to maintain the block tridiagonal structure of the implicit scheme.

Modified Equation

The modified equation for scheme (7) can be derived in a manner similar to that for the explicit scheme. The modified equation for the four-parameter implicit scheme is given to fourth order by

$$w_t+f_x-\Delta t\left(\bar{\theta}-\frac{1}{2}-\xi\right)\frac{\partial}{\partial x}(f'f'w_x)+\Delta x^2\frac{\partial}{\partial x}$$

$$\times\left\{\left(\frac{1}{6}+(1+\xi)\epsilon_i\right)f_{xx}-\sigma^2\left[\frac{1}{6}+\left(\frac{1}{2}+\xi\right)\left(\bar{\theta}-\frac{1}{2}-\xi\right)\right]f''f_x^2\right.$$

$$\left.+\sigma^2\left[\bar{\theta}^2+\frac{\bar{\theta}}{2}-\frac{1}{6}+\left(\frac{1}{2}+\xi\right)(1+2\xi-3\bar{\theta})\right]f'(f'f_x)_x\right\}$$

$$+\Delta x^3\frac{\partial}{\partial x}IM+O(\Delta^4)=0 \qquad (8)$$

The term *IM* includes the third-order truncation errors and is given in the Appendix. We note that the leading truncation error term is first order in time and is removed if $(\bar{\theta}-½-\xi)$ vanishes. For a linear differential equation this truncation error behaves like a second-order smoothing term, i.e., it is highly dissipative. For a nonlinear differential equation, the first-order truncation error behaves again as a second-order dissipation term with some dispersion added; i.e., the term $(\partial/\partial x)(f'f'w_x)$ expands to the following form:

$$\frac{\partial}{\partial x}(f'f'w_x)=(f''w_xf'+f'f''w_x)w_x+f'f'w_{xx} \qquad (9)$$

The first term is strictly dispersive with the variable coefficient in parentheses and the second term is strictly dissipative with the variable coefficient $f'f'$, which is a positive definite matrix. It is this dissipative term which accounts for the improved stability properties of the Euler implicit scheme. But the fact that this truncation error is only first-order accurate in time is a major drawback for unsteady problems.

The modified equation for this scheme at steady state $(w^{n+1}=w^n=w^{n-1})$ is

$$f_x+\frac{\Delta x^2}{6}f_{xxx}+\frac{\epsilon_e(1+\xi)}{\Delta t}\Delta x^4 w_{xxxx}=O(\Delta^4)$$

For stability, additional dissipation in the form of the last term on the left-hand side has to be added. If the steady-state solution is to be independent of the time step, ϵ_e should be proportional to Δt.

Correction of the Original Differential Equation

The scheme can be made second-order time accurate either by the appropriate choice of the parameters $\bar{\theta}$ and ξ or by correcting the original differential equation as was done for the explicit scheme. If we form the new flux vector f_0

$$f_0=f+\Delta t\left(\bar{\theta}-\frac{1}{2}-\xi\right)\frac{\partial}{\partial x}(f'f'w_x)$$

and solve the equation

$$w+(f_0)_x=0 \qquad (10)$$

in place of the original equation, we obtain a second-order time accurate scheme. The corrected portion of the new flux term may be treated either explicitly or implicitly. However, the procedure is unstable regardless of whether the correction is solved explicitly or implicitly. The reason for this will become obvious later. Before proceeding further it is necessary to do at least a linear stability analysis for both schemes applied to the corrected and uncorrected equations.

Stability Analysis

It is of interest to examine the stability of both schemes as applied to the linear form of Eq. (5). Use of the von Neumann stability procedure[13] obtains the amplification factor of the finite difference scheme denoted by $g(k)$. The necessary and sufficient condition for the stability of the scheme is that

$$|g(k)|\leq 1 \qquad (11)$$

for all values of k, where k is the kth Fourier component of a harmonic decomposition.

Lax-Wendroff Scheme

The amplification factor for the Lax-Wendroff scheme is given by

$$g(k)=a+ib+\gamma(c+id) \qquad (12)$$

where

$$i=\sqrt{-1},\quad a=1-2\nu^2 z,\quad b=-\nu\sin\theta$$

$$c=-\nu^3\left[8z^2-\frac{\gamma}{2}\left(\frac{5}{6}+z(1+8z-32z^2)\right)\right]\Big/3$$

$$d=-2\nu z\sin\theta/3$$

where $\nu=c\sigma$ is the Courant number, $z=\sin^2(\theta/2)$, and $\theta=k\Delta x$. The parameter $\gamma=0$ if the scheme is applied to the

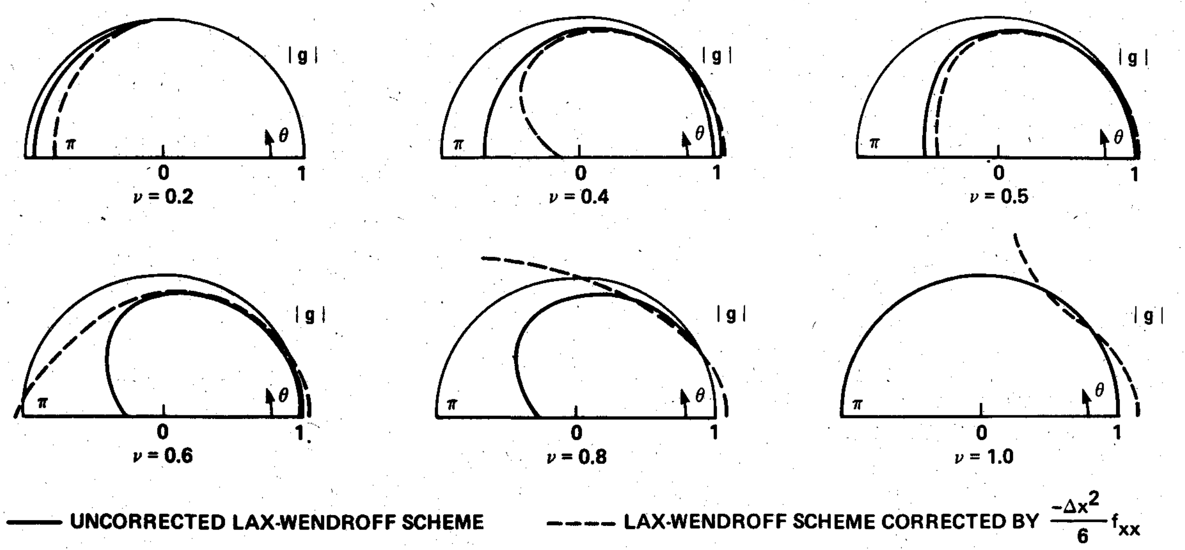


Fig. 1 Polar plot of linear amplification factor for the second-order corrected and uncorrected Lax-Wendroff scheme.

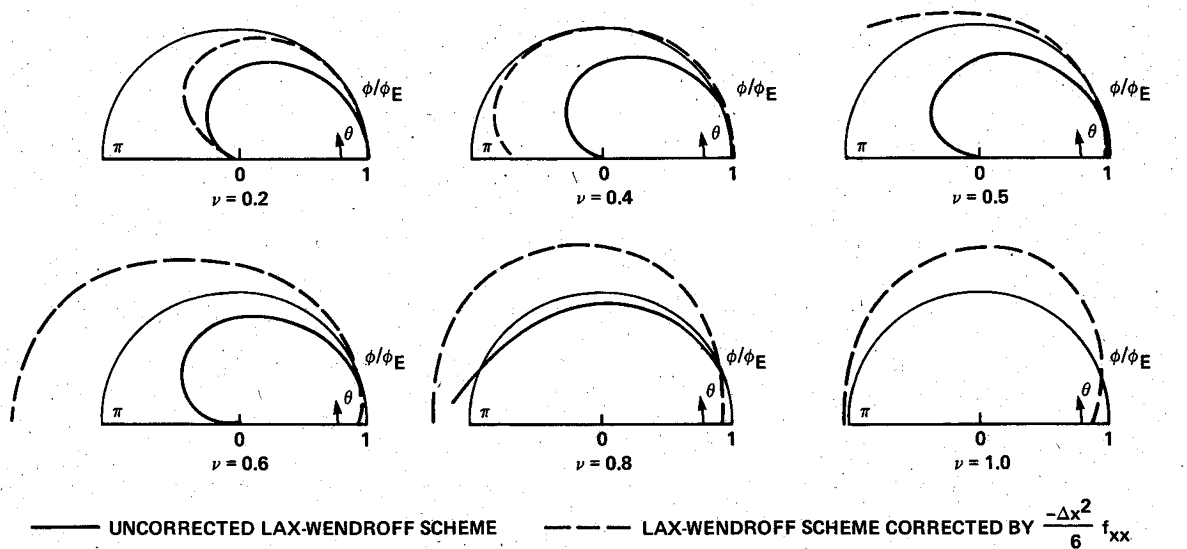


Fig. 2 Polar plot of phase error for the second-order corrected and uncorrected Lax-Wendroff scheme.

uncorrected Eq. (1) or equals 1 for the corrected Eq. (5). Polar plots for various Courant numbers ν and both values of the parameter are shown in Figs. 1 and 2. Only one survey of the stability is necessary, as all of the Lax-Wendroff schemes reduce to the same scheme for a linear equation. The dissipative and dispersive errors of the scheme are compared in the polar plots of $|g|$ and ϕ/ϕ_e. Since $g(k)$ is complex, we may write

$$g(k) = |g| e^{i\phi}$$

The phase shift per time step of an exact spatially periodic solution of the scalar equation

$$w_t + cw_x = 0$$

is given by

$$\phi_e = -ck\Delta t = -\nu k\Delta x$$

The finite difference scheme propagates the Fourier components of the solution with a speed of $(\phi/\phi_e) \cdot c$ instead of the correct speed c.

The first figure depicts the modulus of g for various Courant numbers of both the uncorrected scheme ($\gamma = 0$) and the scheme corrected by $-\Delta x^2/6 \cdot f_{xx}$, i.e., $\gamma = 1$. The uncorrected scheme is stable, i.e., $|g| \leq 1$, for all $\nu \leq 1$. The corrected scheme is slightly unstable, however, for all ν. This is shown by the dotted line in Fig. 1. For small θ, $|g| \approx 1 + \epsilon$, where $0 < \epsilon \ll 1$. For $\theta \approx \pi$, $|g| < 1$ only for $\nu \leq 0.55$. Numerical experiments, however, indicate that all of the corrected schemes are stable for $\nu < 0.6$. This indicates that the slight linear instability for small θ can be tolerated. The phase errors for the Lax-Wendroff scheme are shown in Fig. 2. It is interesting to note that the phase error for the corrected scheme is very small at a Courant number $\nu = 0.4$. For larger Courant numbers the phase error becomes larger than unity (leading phase error).

Implicit Schemes

The amplification factor for the four-parameter implicit scheme is given by

$$(a + ib)g^2 + (c + id)g + e = 0 \tag{13}$$

where

$$a = 1 + \xi + 2z[\gamma\bar{\gamma}\nu^2(\bar{\theta} - ½ - \xi) - 2\epsilon_i], \qquad b = \bar{\theta}\nu\sin\theta$$

$$c = -1 - 2\xi - 4z\epsilon_i + 16z^2\epsilon_e + 4z\gamma(\bar{\gamma} - 1)\nu^2(\bar{\theta} - ½ - \xi)$$

$$d = (1 - \bar{\theta})\nu\sin\theta, \qquad e = \xi$$

As before, ν is the Courant number and $z = \sin^2(\theta/2)$. The two parameters γ and $\bar{\gamma}$ determine whether the scheme is corrected for the first-order truncation error. If $\gamma = 0$, the scheme is uncorrected. If $\gamma = 1$, the scheme is corrected to second-order time accuracy. If $\bar{\gamma} = 1$, the correction is fully implicit and if $\bar{\gamma} = 0$, then the correction is explicit.

There are two roots to Eq. (13). Both roots must satisfy condition (11) for the implicit scheme to be stable. The uncorrected Euler implicit scheme ($\theta = 1$, $\xi = 0$) is stable for all Courant numbers. However, the scheme corrected for the first-order truncation error is unstable whether the correction is treated explicitly or implicitly. The reason for this is quite apparent from Eq. (9). The second term of this first-order truncation error is dissipative and thus stabilizing. However, when the original differential equation is corrected for the first-order truncation error terms, as in Eq. (10), this equation now becomes ill-posed.

In other words, we are now solving the system

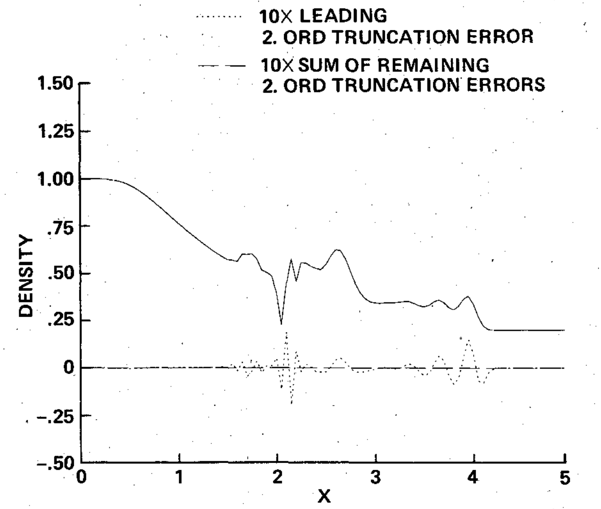
$$w_t + (f_0)_x = w_t + f_x + \Delta t(\bar{\theta} - \tfrac{1}{2} - \xi)(f'' w_x f'$$

$$+ f' f'' w_x) w_x + \Delta t(\bar{\theta} - \tfrac{1}{2} - \xi)(f' f') w_{xx} = 0$$

which is no longer well-posed due to the last term on the left-hand side. As is well known,[13] no stable numerical scheme can solve an ill-posed differential equation.

Results

The model problem chosen for this study is the one-dimensional shock tube. The tube is initialized with a perfect gas at constant temperature and a pressure ratio of 5.0 across the diaphragm. At time $t = 0$, the diaphragm is "broken" and the solution proceeds, resulting in an expansion wave, a contact surface, and a shock wave. The solution proceeds until the waves nearly reach the ends of the computational domain at a nondimensional $t = 1.5$. This problem has the advantage of combining the three flow situations of primary interest in this study (shock waves, expansion waves, and contact surfaces) and requires little computation time. Identical initial conditions are used for both the explicit Lax-Wendroff scheme and the implicit Beam-Warming scheme to allow comparison of the relative accuracies of the two integration techniques for this model problem. Since the solution was not allowed to reach the end "walls," boundary conditions are not considered in the present study.



α = 1.000 β = 0.000 UNCORRECTED SCHEME
CFL = 0.40 ITERATION NO. = 152 SPATIAL STEP = 0.05
TIME STEP = 0.01000 TIME = 1.500

Fig. 3 Density distribution in a one-dimensional shock tube as solved by MacCormack's method.

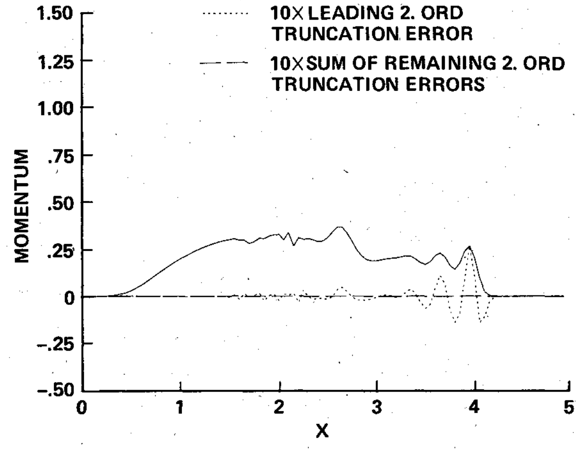
Lax-Wendroff Schemes

Results of applying the Lax-Wendroff system to the model problem are shown in Fig. 3. For illustrative purposes, results for the forward predictor MacCormack's scheme ($\alpha = 1.0$, $\beta = 0.0$) only are shown with differences noted in the discussion for the other Lax-Wendroff schemes. The spatial and temporal steps were maintained constant at 0.05 and 0.01, respectively, during the computation.

As can be noted in Fig. 3, the solution for density (solid line) is characterized by large oscillations in the vicinity of the shock wave and contact surface. The very large oscillation at $x = 2.0$ is near the initial position of the diaphragm and appears to be the remains of an initial instability which is not damped as the solution proceeds.

The second-order terms in the modified equation are most likely to contribute to this oscillatory behavior. These terms are shown by the dotted line for $\Delta x^2/6 \cdot f_{xx}$ and the dash-dot line for the sum of the remainder of Δ^2 terms in the lower part of Fig. 3. Note that these terms are multiplied by a factor of 10 for plotting clarity. In the vicinity of the shock wave, the oscillations in the $\Delta x^2/6 \cdot f_{xx}$ term are in phase with those present in the solution and are of virtually identical frequency. The same is true of the oscillations present at the contact surface. The term $\Delta x^2/6 \cdot f_{xx}$ is clearly the leading second-order term, as the individual magnitudes of the remaining second-order terms are at least one order less than the f_{xx} term. When all of the second-order errors are summed, the resultant error is essentially unchanged from that shown for $\Delta x^2/6 \cdot f_{xx}$ except for a slight reduction in the amplitude. The third-order errors are much smaller and thus not plottable on the scale of Fig. 3. Figures 4 and 5 give plots of momentum and energy for this computation. Note that both the characteristics of the oscillatory behavior in the solutions and the distribution of the second-order truncation error are different for these two quantities. Note also the great relative amplitude of error at the shock waves in both instances. The implication is clear that the nonlinear modified equation error terms must be examind for each element in the solution vector before an overall picture of the error in the solution can be obtained.

For the conditions of this computation, the backward predictor MacCormack's scheme ($\alpha = 1.0$, $\beta = 1.0$) is unstable. Examining the eigenvalue history reveals that the "u-c" eigenvalue changes sign very early in computation for all of the Lax-Wendroff techniques. The nonlinear instability which this sign change denotes is not damped by MacCormack's backwards predictor technique for the right running shock



α = 1.000 β = 0.000 UNCORRECTED SCHEME CFL = 0.40
ITERATION NO. = 152 SPATIAL STEP = 0.05
TIME STEP = 0.01000 TIME = 1.500

Fig. 4 Momentum distribution in a one-dimensional shock tube as solved by MacCormack's method.

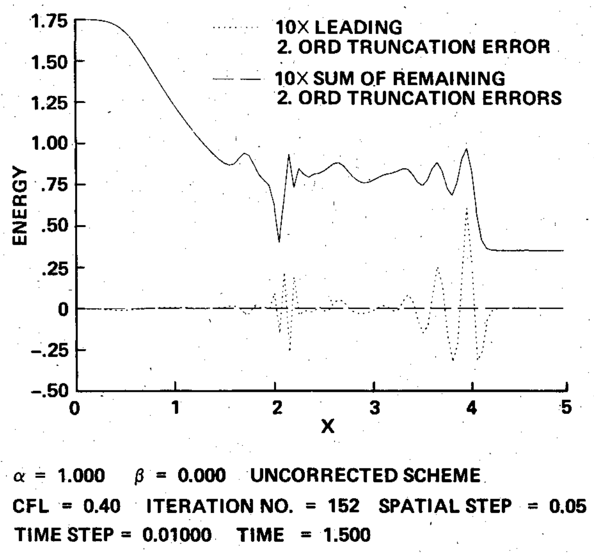


Fig. 5 Energy distribution in a one-dimensional shock tube as solved by MacCormack's method.

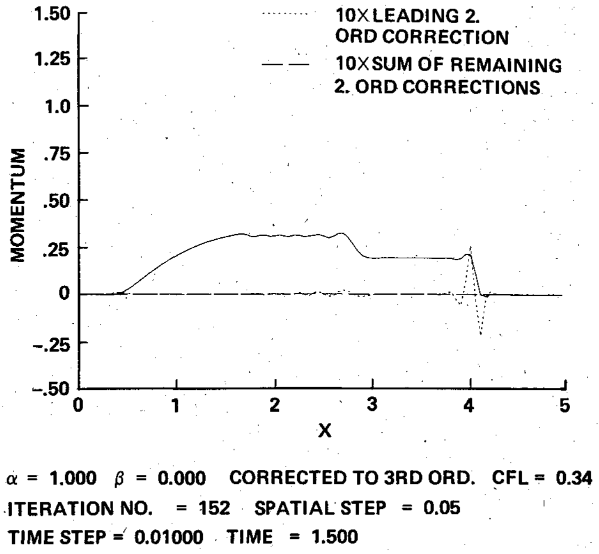


Fig. 7 Momentum distribution in a one-dimensional shock tube as solved by MacCormack's method corrected to third order.

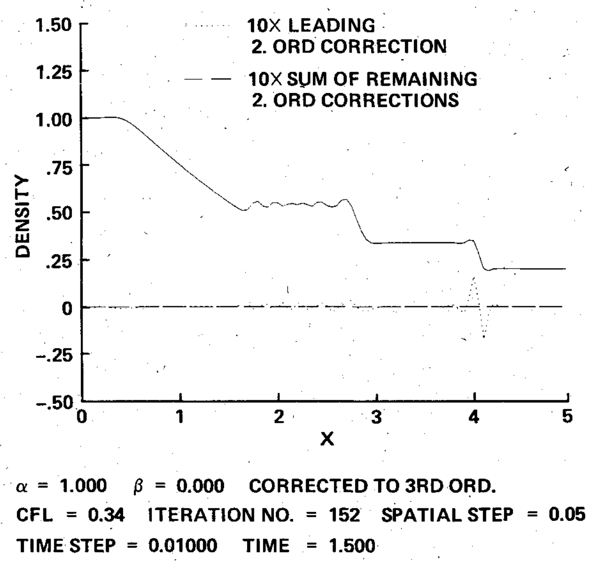


Fig. 6 Density distribution in a one-dimensional shock tube as solved by MacCormack's method corrected to third order.

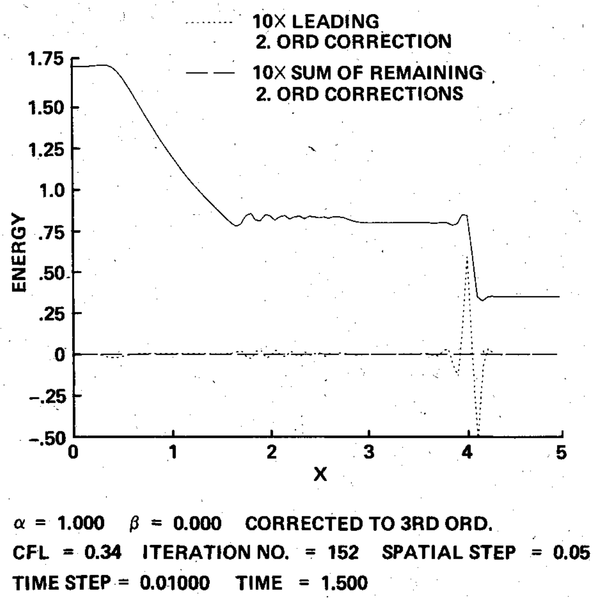


Fig. 8 Energy distribution in a one-dimensional shock tube as solved by MacCormack's method corrected to third order.

wave. (Note that the forward predictor MacCormack's scheme would be unstable for a left running shock wave.) The remainder of the Lax-Wendroff schemes for which $\beta=0.5$ have large oscillations in the vicinity of the shock wave and contact surface for these conditions but do not contain as large as oscillation at $x=2.0$. This is apparently due to the presence of the second-order term, $3/2\alpha\,\beta(\beta-1)f''(w_x,w_x)$, in the modified equation and the absence of the third term, $-3\sigma/2(2\beta-1)f''(w_x,f_x)$, for $\beta=0.5$. The second term is basically dissipative in nature and is much larger (although still one to two orders of magnitude smaller than the leading term) than the third error term which appears in both of the MacCormack schemes.

The oscillations noted in the previous figures are greatly reduced when the original differential system is corrected to remove the second-order terms which appear in the modified equation. Figure 6 gives the density distribution for conditions identical to those previously noted. In this instance, the term f_0 of Eq. (6) contains all of the second-order modified equation terms with the sign changed on each one, thereby removing them at each step of the solution. It should be noted that removal of all of the second-order error terms results in a third-order accurate method. Although small oscillations remain between the shock wave and contact surface, the solution is much improved, with no additional smearing of the discontinuities. In this instance, the leading second-order correction term plotted in the lower part of Fig. 6 is that used to produce the solution for density as shown. This term $(\Delta x^2/6\cdot f_{xx})$ is in reality a measure of the curvature of the vector f and will not be zero at a given time increment even though the term has been removed from the equation solved at the previous time increment. Figures 7 and 8 give the solutions for momentum and energy obtained by solving the corrected equation. The most striking improvement is apparent in the energy solution.

Since the leading truncation error term is much larger than the sum of the remaining terms, solutions were obtained in which only the leading second-order term was used to correct the original equation. These solutions produced essentially the same results as correcting for all of the second-order terms and are not included in the figures. It was also noted that the net effect of the summation of the remaining terms was to reduce the amplitude of the f_{xx} slightly. Solutions were obtained by correcting with 0.9 $\Delta x^2/6\cdot f_{xx}$ that were virtually identical with those produced by correcting with the full second-order set.

The significance of the use of the leading second-order term only for correction is that a two-step essentially third-order accurate method can be obtained by the subtraction of $\Delta x^2/6 \cdot f_{xx}$ from the original f vector in the equation. The finite difference method is then applied as it would be normally. In the present study, central-difference approximations were used for the f_{xx} derivative at each mesh location.

While there are other numerical schemes such as the "artificial compression method"[12] which give better results for the shock tube problem than our present procedure, these other schemes are much more expensive to run. For the present procedure only f_{xx} needs to be computed along with f from the given w.

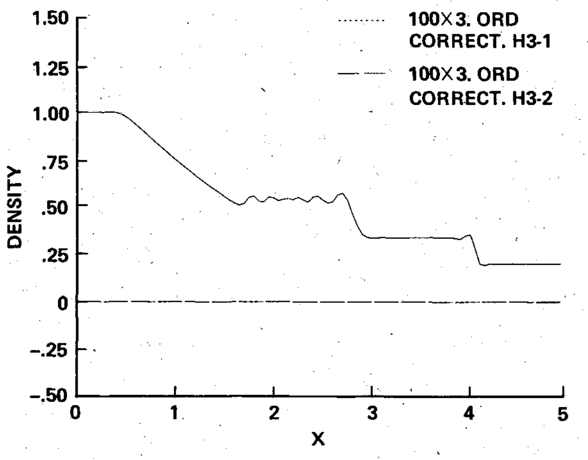
To illustrate the effect of correcting the original equation to remove both second- and third-order truncation error terms, Fig. 9 gives the solution obtained for the density. Very little difference is noted between Fig. 9 (a fourth-order accurate solution) and Fig. 6 (a third-order accurate solution). Examination of the amplitudes of the third-order terms show most to be of order 10^{-12} or smaller. This is a pleasing result, as considerable computational work is required to correct for the third-order terms.

Implicit Schemes

The results for the shock tube problem solved by the implicit schemes are shown in Figs. 10-12. For brevity we will limit our attention to the Euler implicit scheme ($\theta = 1.0$, $\xi = 0.0$). The spatial and temporal steps are identical to those used for the results of the explicit scheme. Following the rules similar to those given by Desideri et al.,[11] the explicit and implicit smoothing parameters were set as follows: $\epsilon_e = \Delta t$, $\epsilon_i = 4\epsilon_e$.

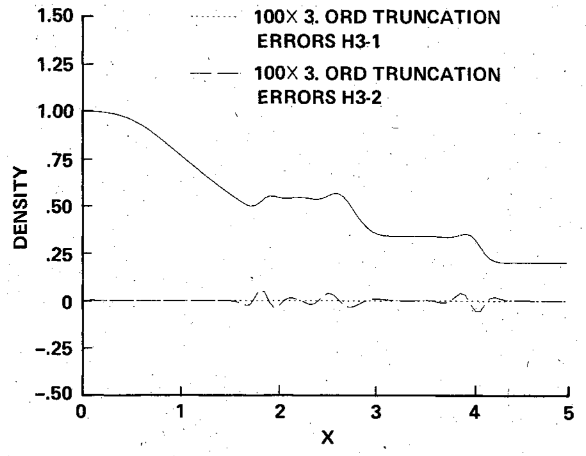
These values were used for all the implicit numerical results shown in this paper. Linear stability analysis predicts unconditional stability for the Euler implicit scheme for all wavenumbers k, i.e., $|g| \leq 1$. However, for the shortest wavelengths ($k\Delta x = \pi$) the scheme is not dissipative and thus some small amount of dissipation must be added to control the nonlinear instabilities.

The solution obtained by the uncorrected Euler implicit scheme is shown in Fig. 10 for the density. This solution was obtained for a constant $\sigma = \Delta t/\Delta x = 0.2$, exactly the same as for the explicit scheme. This figure also shows the first- and second-order truncation errors. These errors are quite large (on the order of 10-30% for the energy equation). Even more



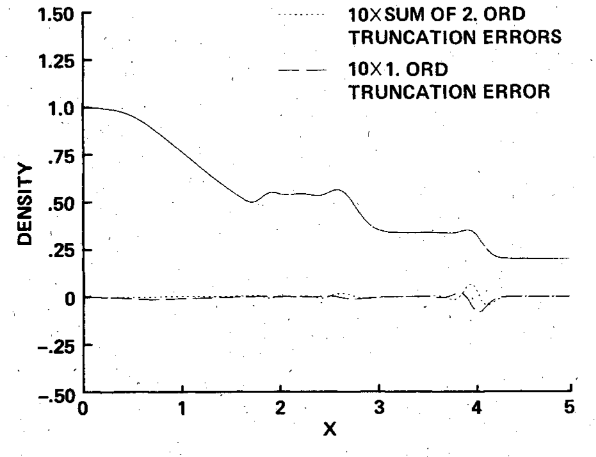
α = 1.000 β = 0.000 CORRECTED TO 4TH ORD. CFL = 0.35
ITERATION NO. = 152 SPATIAL STEP = 0.05
TIME STEP = 0.01000 TIME = 1.500

Fig. 9 Density distribution in a one-dimensional shock tube as solved by MacCormack's method corrected to fourth order.



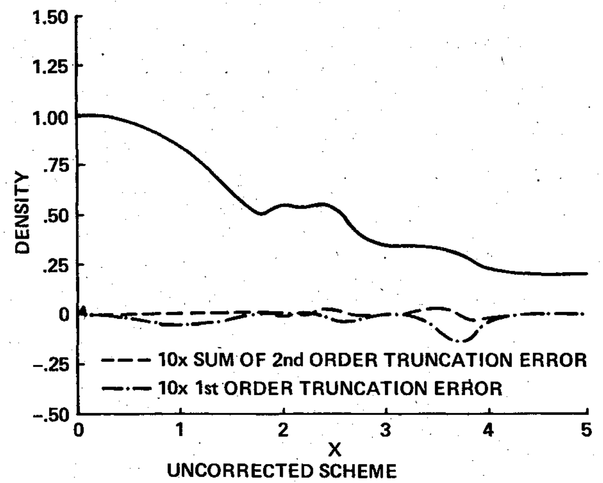
θ = 1.000 ξ = 0.000 UNCORRECTED SCHEME ϵ_e = 0.010
ϵ_i = 0.040 ITERATION NO. = 150 SPATIAL STEP = 0.05
DT/DX = 0.20 TIME = 1.500 CFL = 0.379

Fig. 11 Third-order truncation error for the Euler implicit scheme.



θ = 1.000 ξ = 0.000 UNCORRECTED SCHEME
ϵ_e = 0.010 ϵ_i = 0.040 ITERATION NO. = 150
SPATIAL STEP = 0.05 DT/DX = 0.20 TIME = 1.500
CFL = 0.379

Fig. 10 Density distribution in a one-dimensional shock tube as solved by the Euler implicit scheme.



UNCORRECTED SCHEME
ϵ_e = 0.050 ϵ_i = 0.200 θ = 1.000 ξ = 0.000
ITERATION NO. = 26 SPATIAL STEP = 0.05
DT/DX = 1.00 TIME = 1.300 CFL = 1.834

Fig. 12 Density distribution and first- and second-order truncation error in a one-dimensional shock tube solved by the Euler implicit scheme.

surprising are the third-order truncation errors, some of which are shown in Fig. 11 for the density truncation error. Even these third-order errors are of the order of a few percent of the solution. For this $\sigma = 0.2$, for which the CFL number is approximately 0.4, the basic features of the flow are fairly well resolved although some overshoots do occur. If the value of σ is increased to 1.0 (for which the CFL number increases to approximately 1.8), the solution shown in Fig. 12 is obtained. The errors are much larger (by a factor of from 4 to 5) and excessive smearing now takes place. The flow features such as the shock wave or the contact surface are no longer clearly distinguishable.

Conclusions

A nonlinear modified equation analysis has been demonstrated which allows detailed examination of both the linear and nonlinear truncation errors introduced by the finite difference approximation of a system of differential equations. The advantages of this procedure are that the full nonlinearities of the equations are considered and the full system of equations is included. In other words, no simplifications to linear, scalar, or model equations are necessary. The technique has been demonstrated for two popular types of finite difference schemes: the explicit generalized Lax-Wendroff scheme and the four-parameter implicit scheme. The system of differential equations considered is the one-dimensional unsteady Euler equation of gas dynamics.

The specific conclusions are as follows:

1) It was shown that the nonlinear truncation errors are quite large and distributed quite differently for each of the three conservation equations as applied to a one-dimensional shock tube problem.

2) A technique for removing these error terms from the solutions has been developed and demonstrated.

3) Removing the leading second-order error terms from the Lax-Wendroff modified equation resulted in greatly improved solutions.

4) The correction technique does not work for first-order accurate schemes such as the Euler implicit scheme. The cause of the difficulty for first-order schemes is that the corrected differential system becomes ill-posed.

5) The present technique is a powerful tool for analyzing finite difference techniques.

Appendix

The third-order truncation errors are given in this appendix for both the generalized Lax-Wendroff scheme and the four-parameter implicit scheme.

Generalized Lax-Wendroff Scheme

The third-order truncation error for this scheme is given

$$\frac{\Delta x^3}{24}\frac{\partial}{\partial x}EX = \frac{\Delta x^3}{4!}\frac{\partial}{\partial x}\Big[3\sigma f'(1-\sigma^2 f'f')f_{xxx}$$

$$-3\sigma^3 f'[f'f''(w_{xx}f_x+2w_xf_{xx})+f'f'''w_x^2f_x+f''w_x(f''w_xf_x$$

$$+f'f_{xx})]+\sigma^3[3(\alpha-1)f'(2f''f_xf_{xx}+f'''w_xf_x^2)$$

$$-(2\alpha-1)(\alpha-1)f'''f_x^3+3(2\alpha-1)f''f_x(f'f_{xx}+f''w_xf_x)]$$

$$+3(2\beta-1)\sigma^2\{(\alpha-1)f'''w_xf_x^2-f''f_xf_{xx}-f'[f'''w_x^2f_x$$

$$+f''(f_{xx}w_x+f_xw_{xx})]-f''w_x(f''w_xf_x+f'f_{xx})\}$$

$$+\frac{\sigma}{\alpha}\{(1-2\alpha)3\beta(\beta-1)f'''f_xw_x^2+3[2\beta(\beta-1)+\alpha]$$

$$\times f''w_xf_{xx}+3\beta(\beta-1)f'(f'''w_x^3+2f''w_xw_{xx})\}$$

$$+\frac{\beta}{\alpha}(2\beta^2-3\beta+1)f'''w_x^3\Big]$$

Implicit Scheme

The third-order truncation error for the four-parameter implicit scheme is as follows:

$$\Delta x^3\frac{\partial}{\partial x}IM = +\Delta x^3\frac{\partial}{\partial x}\Big[\frac{(1+\xi)}{\sigma}\epsilon_e w_{xxx}$$

$$-\sigma\Big[\Big(\bar\theta-\frac{1}{2}-\xi\Big)\Big(\frac{1}{6}+(1+\xi)\epsilon_i\Big)\Big]f'f_{xxx}$$

$$+\sigma\Big[\Big(2\xi+\frac{1}{2}-\bar\theta\Big)(1+\xi)\epsilon_i-\frac{1}{6}\Big(\bar\theta-\frac{1}{2}-\xi\Big)\Big](f'f_x)_{xx}$$

$$+\sigma^3\Big\{4\Big(\bar\theta-\frac{1}{2}-\xi\Big)\Big[\frac{1}{6}-\Big(\frac{1}{2}+\xi\Big)^2\Big]-\Big(\bar\theta-\frac{1}{2}-\xi\Big)^3$$

$$+\frac{\bar\theta}{2}\Big(\frac{1}{6}+\xi\Big)-\frac{1}{12}\Big(\frac{1}{2}+\xi\Big)\Big\}f'[f'(f'f_x)_x]_x$$

$$+\sigma^3\Big\{\Big(\bar\theta-\frac{1}{2}-\xi\Big)\Big[\frac{1}{6}-\Big(\frac{1}{2}+\xi\Big)^2\Big]-\frac{1}{12}\Big(\frac{1}{2}+\xi\Big)\Big\}f'''f_x^3$$

$$+\sigma^3\Big\{\Big(\bar\theta-\frac{1}{2}-\xi\Big)\Big[\frac{1}{3}-3\Big(\frac{1}{2}+\xi\Big)^2+\bar\theta(1+2\xi)\Big]-\frac{\bar\theta}{6}\Big\}$$

$$\times f'(f''f_x^2)_x+\sigma^3\Big[(1+2\xi)\Big(\bar\theta(\bar\theta-\xi)-\frac{1}{4}\Big)$$

$$+5\Big(\bar\theta-\frac{1}{2}-\xi\Big)\Big(\frac{1}{6}-\frac{(1+2\xi)^2}{4}\Big)\Big]f''f_x(f'f_x)_x\Big]$$

Acknowledgments

The work by the first author reported herein was supported by the Office of Naval Research under Contract N00014-78-C-0490. Dr. Robert Whitehead was the contract monitor. The provision of computer time by Mr. H. Lomax, CFD Branch Chief, NASA Ames Research Center, is greatly appreciated.

References

[1] Lax, P. and Wendroff, B., "Systems of Conservation Laws," *Communications in Pure and Applied Mathematics*, Vol. 13, 1960, pp. 217-237.

[2] MacCormack, R., "The Effect of Viscosity in Hypervelocity Impact Cratering," AIAA Paper 69-354, 1969.

[3] Kutler, P. and Lomax, H., "Shock-Capturing, Finite Difference Approach to Supersonic Flows," *Journal of Spacecraft and Rockets*, Vol. 8, Dec. 1971, pp. 1175-1182.

[4] Warming, R. F., Kutler, P., and Lomax, H., "Second- and Third-Order Noncentered Difference Schemes for Nonlinear Hyperbolic Equations," *AIAA Journal*, Vol. 11, Feb. 1973, pp. 189-196.

[5] Moretti, G., "Experiments in Multi-Dimensional Floating Shock Fitting," Polytechnic Institute of Brooklyn, Aeronautical Laboratory, N.Y., Rept. 73-18, 1973.

[6] Lerat, A. and Peyret, R., "The Problem of Spurious Oscillations in the Numerical Solution of the Equations of Gas Dynamics," *Lecture Notes in Physics*, Vol. 35, Springer-Verlag, Berlin, 1975.

[7] Warming, R. F. and Beam, R. M., "On the Construction and Application of Implicit Factored Schemes for Conservation Laws," *Society for Industrial and Applied Mathematics—American Mathematical Society Proceedings*, Vol. XI, 1978, pp. 85-129.

[8] Warming, R. F. and Hyett, B. J., "The Modified Equation to the Stability and Accuracy Analysis of Finite-Difference Methods," *Journal of Computational Physics*, Vol. 14, 1974, pp. 159-179.

[9] Lerat, A., "Numerical Shock Structure and Nonlinear Correction for Difference Schemes in Conservation Forms," *Lecture Notes in Physics*, Vol. 90, Springer-Verlag, Berlin, 1979.

[10] Majda, A. and Osher, S., "A Systematic Approach for Correcting Nonlinear Instabilities," *Numerical Mathematics*, Vol. 30, 1978, pp. 429-452.

[11] Desideri, J. A., Steger, J. L., and Tannehill, J. C., "On Improving the Iterative Convergence Properties of an Implicit Approximate-Factorization Finite Difference Algorithm," NASA TM 78495, June 1978.

[12] Harten, A., "The Artificial Compression Method for the Computation of Shocks and Contact Discontinuities, III. Self-Adjusting Hybrid Schemes," *Mathematics of Computations*, Vol. 32, 1978, pp. 363-390.

[13] Richtmyer, R. D. and Morton, K. W., *Difference Methods for Initial-Value Problems*, United Interscience Publishers, 1967.