JOURNAL OF COMPUTATIONAL PHYSICS 14, 126–147 (1974)

A Class of Implicit, Second-Order Accurate, Dissipative Schemes for Solving Systems of Conservation Laws

G. R. McGUIRE

Department of Mathematics, Heriot-Watt University, Edinburgh, Scotland

AND

J. Ll. MORRIS

Department of Mathematics, University of Dundee, Scotland

Received April 11, 1973

All explicit difference schemes for solving systems of conservation laws are subject to the Courant–Friedrichs–Lewy [1] convergence condition. This condition manifests itself as a restrictive condition on the size of the time steps which can be used in the schemes. Implicit schemes, on the other hand, automatically satisfy the convergence condition. However, most implicit schemes used in the past have either only been first-order accurate or second-order accurate but nondissipative (Gary [2], Zwas and Abarbanel [17]). This paper develops a class of second-order-accurate implicit schemes which are dissipative. Some numerical results are presented which show their usefulness in solving problems involving discontinuities. These results appear promising for the case of a single equation. However, there appears to be some computational difficulties in the case of systems of equations which require further investigation.

1. INTRODUCTION

In recent years, many finite difference methods have been proposed for solving systems of conservation laws (Richtmyer [9], Gourlay and Morris [3], Rubin and Burstein [11], McGuire and Morris [7]). These methods have been explicit difference methods and as such are necessarily bound by the Courant–Friedrichs–Lewy (CFL) [1] convergence condition. This condition imposes a severe restriction on the size of the time steps.

We will consider the one-space dimensional system of conservation laws,

$$(\partial u/\partial t) + (\partial f/\partial x)(u) = 0. \tag{1.1}$$

In (1.1) u and f are n-vectors and the solution is sought in the region

$$R \equiv [0 < x \leqslant X] \times [t > 0]. \tag{1.2}$$

For the discussion of the physical nature of the solutions of this system and of the existence and uniqueness of solutions, see Jeffrey and Tanuiti [5].

The equation is considered to be hyperbolic over R which means that the Jacobian of f, namely, $A(u) = \partial f/\partial u$, has real eigenvalues and a complete linearly independent set of eigenvectors. It will also be assumed that these eigenvalues are positive throughout R so that, with initial conditions

$$u(x, 0) = u_0(x)$$

and boundary conditions

$$u(0, t) = u_1(t),$$

(1.1) gives a well-posed problem in R. The case when some of the eigenvalues are negative is easily dealt with (see [5]).

A grid of points with mesh spacing h in the x direction and time step k in the t direction is placed on R. It is assumed that $x = Nh$.

The value of the solution of the differential equation at the grid point $x = ih$, $t = mk$ is denoted by $u(ih, mk) \equiv u_i^m$ and any approximation to this value is denoted by w_i^m. One of the most used methods for solving (1.1) is the two-step Richtmyer Scheme [9]. A generalization of this method was presented in McGuire and Morris [7] and takes the following form:

$$w_i^{m+1} = w_i^m - \frac{p}{2}\left[\left(1 - \frac{1}{2a}\right)(f_{i+1}^m - f_{i-1}^m) + \frac{1}{a}(\overset{*}{f}{}_{i+\frac{1}{2}}^{m+a} - \overset{*}{f}{}_{i-\frac{1}{2}}^{m+a})\right], \tag{1.3}$$

where

$$\begin{aligned} &\overset{*}{f}{}_i^{m+a} = f(\overset{*}{w}{}_i^{m+a}), \qquad p = k/h = \text{constant}, \qquad a \neq 0, \\ &\overset{*}{w}{}_i^{m+a} = (w_{i+\frac{1}{2}}^m + w_{i-\frac{1}{2}}^m)/2 - ap(f_{i+\frac{1}{2}}^m - f_{i-\frac{1}{2}}^m). \end{aligned} \tag{1.4}$$

This method is second-order accurate.

In analyzing the linearized stability of the method, it is applied to a system

$$(\partial u/\partial t) + A(\partial u/\partial x) = 0 \tag{1.5}$$

where A is a constant matrix. This A is equivalent to locally constant values of $A(u) = \partial f/\partial u$. The amplification matrix is then derived using Fourier integrals as

$$G(\alpha) = I - \sqrt{-1}\, pA \sin \alpha + (p^2A^2/2)(2 \cos \alpha - 2),$$

where $\alpha = \beta h$, with β the variable in the Fourier space.

The matrix $G(\alpha)$ is uniformly diagonalizable by the assumption that the systems (1.5) are hyperbolic. Hence, the Von Neumann condition is sufficient as well as necessary for linearized stability.

The eigenvalues of $G(\alpha)$ have moduli

$$|g(\alpha)|^2 = 1 - p^2\lambda^2(1 - p^2\lambda^2)\sin^4(\alpha/2). \tag{1.6}$$

Hence, the method is stable under the condition

$$p|\lambda| \leqslant 1, \tag{1.7}$$

where $|\lambda|$ is the maximum modulus eigenvalue of A.

Also, it is easy to see that $\exists\delta > 0$ such that

$$|g(\alpha)| \leqslant 1 - \delta|\alpha|^4 \qquad (|\alpha| \leqslant \pi)$$

provided

$$0 < p|\lambda| < 1 \tag{1.8}$$

for all eigenvalues λ of A.

Thus, under condition (1.8), the scheme (1.3), (1.4) is dissipative (in the linearized sense) of order 4.

The stability condition (1.7) is the best possible for a scheme using the grid points (i, m), $(i \pm 1, m)$, $(i, m + 1)$. Other explicit methods on these points (Gourlay and Morris [3]) can be analyzed in exactly the same way. They are all, however, necessarily subject to the restriction (1.7) on the size of the time steps which can be used.

On the other hand, implicit methods have infinite domains of dependence and hence will automatically satisfy the CFL convergence criterion. The simplest implicit scheme

$$w_i^{m+1} + \frac{p}{2}[f_{i+1}^{m+1} - f_{i-1}^{m+1}] = w_i^m \tag{1.9}$$

is first-order accurate and unconditionally stable. The Crank–Nicolson method

$$w_i^{m+1} + \frac{p}{4}[f_{i+1}^{m+1} - f_{i-1}^{m+1}] = w_i^m - \frac{p}{4}[f_{i+1}^m - f_{i-1}^m] \tag{1.10}$$

is second-order accurate and has local amplification matrix

$$G(\alpha) = \left[I + \sqrt{-1}\,\frac{p}{2}A\sin\alpha\right]^{-1}\left[I - \sqrt{-1}\,\frac{p}{2}A\sin\alpha\right]. \tag{1.11}$$

The eigenvalues of this matrix all have modulus one. This means that the Crank–

Nicolson method is *not* dissipative and will thus be useless for the damping of high-frequency perturbations.

Other implicit schemes have been studied, see Gourlay and Morris [4], Gary [2], and Abarbanel and Zwas [17]. These methods however, have always been either first-order accurate or nondissipative and second-order accurate. We obtain in the next section a class of second-order-accurate schemes. In Section 3, the stability and dissipative properties are considered and it is shown that a suitable choice of parameters gives a locally dissipative method. The computational implementation of the scheme is considered in Section 4. Some numerical results are presented in Section 5, and extensions of the schemes to two space dimensions are considered in Section 6 where some numerical results are also given.

2. Second-Order-Accurate Implicit Schemes

Our aim is to develop a second-order-accurate implicit scheme which is dissipative and has good stability properties. The term which accounts for the dissipation in the two-step Richtmyer scheme is the term $p(\overset{*}{f}{}^{m+a}_{i+\frac{1}{2}} - \overset{*}{f}{}^{m+a}_{i-\frac{1}{2}})$. Thus, let us perturb the Crank–Nicolson scheme by adding a similar term and adjusting the parameters of the constituent terms to give second-order accuracy. Thus, we consider the scheme

$$w_i^{m+1} = w_i^m - p[b(f_{i+1}^{m+1} - f_{i-1}^{m+1}) + c\, f_{i+1}^m - f_{i-1}^m) + d(\overset{*}{f}{}^{m+a}_{i+\frac{1}{2}} - \overset{*}{f}{}^{m+a}_{i-\frac{1}{2}})], \quad (2.1)$$

where $\overset{*}{w}{}_i^{m+a}$ is given by (1.4). It is then an easy matter to show that the scheme is second-order accurate provided

$$\begin{aligned} 2b + 2c + d &= 1, \\ 2b + ad &= 1/2. \end{aligned} \quad (2.2)$$

Hence, we have obtained a two-parameter class of second-order-accurate implicit methods. (2.1) becomes the Crank–Nicolson method when $d = 0$, and it becomes the formulation (1.3), (1.4) when $b = 0$. In the next section, we consider the linearized stability and dissipative properties of this class of methods.

3. Analysis of Linearized Stability and Dissipation of (2.1)

Linearizing (2.1) and (1.4) and eliminating the starred values gives

$$\begin{aligned} w_i^{m+1} = w_i^m - pA\, b(w_{i+1}^{m+1} - w_{i-1}^{m+1}) - p\left(c + \frac{d}{2}\right) A(w_{i+1}^m - w_{i-1}^m) \\ + ad\, p^2 A^2(w_{i+1}^m - 2w_i^m + w_{i-1}^m). \end{aligned} \quad (3.1)$$

The linearized amplification matrix is thus

$$G(\alpha) = [I + \sqrt{-1}\, 2b\, pA \sin \alpha]^{-1} \left[I - \sqrt{-1}\, 2 \left(c + \frac{d}{2}\right) pA \sin \alpha + 2ad\, p^2 A^2 (\cos \alpha - 1)\right]. \tag{3.2}$$

The eigenvalues of $G(\alpha)$ are

$$g(\alpha) = \frac{1 - \sqrt{-1}\, 2 \left(c + \frac{d}{2}\right) p\lambda \sin \alpha + 2ad\, p^2 \lambda^2 (\cos \alpha - 1)}{1 + \sqrt{-1}\, 2b\, p\lambda \sin \alpha}, \tag{3.3}$$

where λ is an eigenvalue of A (λ is real).
Thus

$$\begin{aligned} |g(\alpha)|^2 &= \frac{\{1 + (1 - 4b)\, p^2\lambda^2(\cos \alpha - 1)\}^2 + \{(1 - 2b)\, p\lambda \sin \alpha\}^2}{1 + 4b^2 p^2 \lambda^2 \sin^2 \alpha} \\ &= 1 - \frac{2ad\, p^2\lambda^2(1 - 2ad\, p^2\lambda^2)(1 - \cos \alpha)^2}{1 + 4b^2 p^2 \lambda^2 \sin^2 \alpha}. \end{aligned} \tag{3.4}$$

The relations (2.2) were used in obtaining the form (3.4) for $|g(\alpha)|$.

From (3.2), the matrix $G(\alpha)$ is uniformly diagonalizable since A has a complete set of linearly independent eigenvectors. Thus the Von Neumann condition is sufficient as well as necessary for stability (see Richtmyer and Morton [10]).

From (3.4), stability (in the linearized sense) is thus assured provided

$$0 \leqslant 2ad\, p^2\lambda^2 \leqslant 1 \tag{3.5}$$

for all eigenvalues λ of A. Thus, unconditional stability is achieved when $ad = 0$; the method then reduces to the Crank–Nicolson method. When $ad > 0$, the scheme is stable provided

$$p\,|\lambda| \leqslant \frac{1}{(2ad)^{1/2}}, \tag{3.6}$$

where $|\lambda|$ now denotes the largest modulus eigenvalue of A.

Also, it is easy to see that, provided

$$0 < 2ad\, p^2\lambda^2 < 1, \tag{3.7}$$

there exists a constant $\delta > 0$ such that

$$|g(\alpha)| \leqslant 1 - \delta\,|\alpha|^4 \tag{3.8}$$

for all eigenvalues g of G and for all $|\alpha| \leqslant \pi$. (3.8) states that the scheme (2.1), (1.4) is dissipative (in the linearized sense) of order 4. Condition (3.7) is equivalent to

$$ad > 0,\ \lambda \neq 0,$$

$$p|\lambda| < \frac{1}{(2ad)^{1/2}}, \tag{3.9}$$

where $|\lambda|$ denotes the maximum modulus eigenvalue of A. Note that (3.9) requires that all the eigenvalues of A are nonzero.

Thus, under conditions (2.2) and (3.9), the scheme (2.1), (1.4) is a two-parameter class of implicit, dissipative, stable (in the linearized sense) methods for solving nonlinear conservation laws in one space dimension.

4. Computational Implementation of the Scheme

At each time level, Eq. (2.1) gives a nonlinear system of difference equations to solve. The obvious first approach is to define an iteration on the equations, namely

$$\begin{aligned} {}^{(j+1)}w_i^{m+1} = w_i^m - p[b({}^{(j)}f_{i+1}^{m+1} - {}^{(j)}f_{i-1}^{m+1}) + c(f_{i+1}^m - f_{i-1}^m) \\ + d(f_{i+\frac{1}{2}}^{*m+\alpha} - f_{i-\frac{1}{2}}^{*m+\alpha})] \qquad j = 0, 1, 2,... \end{aligned} \tag{4.1}$$

with some initial guess for ${}^{(0)}w_i^{m+1}$. Although iterative techniques like (4.1) would give the desired results, they will be lengthy, and difficulties with starting values and stopping criteria generally make such a process inefficient. Our technique will be to use a direct method which takes account of the tridiagonal structure of Eq. (4.1). This technique has already been exploited in Gourlay and Morris [4] for solving implicit nonlinear equations.

First, a matrix $\tilde{A}$ is defined by

$$f(u) = \tilde{A}(u) \cdot u. \tag{4.2}$$

It is obvious from this equation that many choices exist for $\tilde{A}$. The choice will depend on the particular problem to be solved. Using (4.2) in (2.1) gives

$$\begin{aligned} w_i^{m+1} + pb[\tilde{A}(w_{i+1}^{m+1})\, w_{i+1}^{m+1} - \tilde{A}(w_{i-1}^{m+1})\, w_{i-1}^{m+1}] \\ = w_i^m - p[c(f_{i+1}^m - f_{i-1}^m) + d(f_{i+\frac{1}{2}}^{*m+\alpha} - f_{i-\frac{1}{2}}^{*m+\alpha})]. \end{aligned} \tag{4.3}$$

Now let w_i^{**m+1} be a first-order approximation to $u(ih, (m+1)k)$ which is smooth through second-order terms, namely

$$w_i^{**m+1} = u(ih, (m+1)k) + C_{ij}^m h^2 + O(h^3), \tag{4.4}$$

where C_{ij}^m is a smooth function. Then we have

$$\begin{aligned}
&\tilde{A}(\overset{**}{w}_{i+1}^{m+1})\, u((i+1)h, (m+1)k) - \tilde{A}(\overset{**}{w}_{i-1}^{m+1})\, u((i-1)h, (m+1)k) \\
&\quad = 2h(\partial/\partial x)\{\tilde{A}(\overset{**}{w}_{i}^{m+1})\, u(ih, (m+1)k)\} + O(h^3) \\
&\quad = 2h(\partial/\partial x)\{\tilde{A}(u(ih, (m+1)k))\, u(ih, (m+1)k)\} + O(h^3) \\
&\quad = 2h(\partial f/\partial x)(u(ih, (m+1)k)) + O(h^3). \qquad (4.5)
\end{aligned}$$

Thus, replacing $\tilde{A}(w_i^{m+1})$ in (4.3) by $\tilde{A}(\overset{**}{w}_i^{m+1})$ does not alter the accuracy of the method.

Equation (4.3) is a three block recurrence relation at each time level. To make these equations well posed at each time level, we require values of w_i^{m+1} at the boundaries, namely values for w_0^{m+1} and w_N^{m+1}.

At the lower boundary, we can simply take

$$w_0{}^m = u_1(mk) \qquad m = 1, 2, 3, \ldots. \qquad (4.6)$$

Supplying the values w_N^{m+1} is not quite as easy. Two techniques are considered. First, we can simply predict a second-order-accurate value for w_N^{m+1} by one of the many available explicit techniques. This approach gives a block tridiagonal system, $(N-1) \times (N-1)$ blocks, to be solved at each time level for $\{w_i^{m+1}\}_{i=1}^{N-1}$.

A second approach is to adjust Eqs. (2.1) and (1.4) to be forward difference equations of second-order accuracy at the upper boundary. (See McGuire and Morris [7].) Doing this in (2.1) and (1.4) at $i = N$ gives

$$w_N^{m+1} = w_N{}^m - p[b(2\nabla_x + \nabla_x{}^2) f_N^{m+1} + c(2\nabla_x + \nabla_x{}^2) f_N{}^m + a(\overset{*}{f}{}_{N+\frac{1}{2}}^{m+a} - \overset{*}{f}{}_{N-\frac{1}{2}}^{m+a})]^1 \qquad (4.7)$$

$$\overset{*}{w}{}_{N+\frac{1}{2}}^{m+a} = (2 + \nabla_x + \nabla_x{}^2)\, w_N^{m+1} - ap(\nabla_x + \nabla_x{}^2) f_N^{m+1}. \qquad (4.8)$$

Now, however, replacement of f_i^{m+1} in (2.1) and (4.7) by $\tilde{A}(\overset{**}{w}_i^{m+1})\, w_i^{m+1}$ no longer gives a block tridiagonal system at each time level since the term

$$(2\nabla_x + \nabla_x{}^2) f_N^{m+1}$$

involves values at $i = N, N-1, N-2$. However, by writing

$$\begin{aligned}
(2\nabla_x + \nabla_x{}^2) f_N^{m+1} &= \nabla_x(2 + \nabla_x) f_N^{m+1} = 3\nabla_x f_N^{m+1} - \nabla_x f_{N-1}^{m+1} \\
&\doteq 3\nabla_x A(\overset{**}{w}_N^{m+1})\, w_N^{m+1} - \nabla_x f(\overset{**}{w}_{N-1}^{m+1}) \qquad (4.9)
\end{aligned}$$

[1] $\nabla_x f_i^m = f_i^m - f_{i-1}^m$.

correct to order h^3 terms, it is then apparent that we will have a block tridiagonal system at each time level to be solved for $\{w_i^{m+1}\}_1^N$. Note that similar boundary techniques can be applied to be problem with nondefinite Jacobian $A(u)$; for the components of u corresponding to negative eigenvalues in A, boundary conditions are specified on the right-hand boundary with a boundary technique similar to that described in [6] and [7] used for values on the left-hand boundary.

These block tridiagonal systems are efficiently solved by the algorithm in Varga [12]. This algorithm reduces the systems to solving small matrix systems of the order of $\tilde{A}$. These small systems can be solved easily if $\tilde{A}$ is chosen suitably. The choice of $\tilde{A}$ will depend on the form of the function f.

It now only remains to indicate how values for $\overset{**}{w}_i^{m+1}$ may be determined. We have available values w_i^m and $\overset{*}{w}_i^{m+a}$. Hence, taking a linear combination of these values so as to be first-order accurate for w_i^{m+1} will satisfy our requirements. Thus we take

$$\overset{**}{w}_i^{m+1} = r w_i^m + s(\overset{*}{w}_{i+\frac{1}{2}}^{m+a} + \overset{*}{w}_{i-\frac{1}{2}}^{m+a}). \tag{4.10}$$

It is easily verified that first-order accuracy is achieved when

$$\begin{aligned} s &= 1/2a, \\ r &= 1 - 1/a. \end{aligned} \tag{4.11}$$

The computational details for the implementation of the class of methods is now complete.

5. Numerical Experiments in One Space Dimension

The results of some experiments using the schemes of Section 2 are tabulated and discussed in this section. The time steps used are in excess of those allowed by the CFL condition for explicit methods (particularly useful in the case of such problems as atmospheric models where wave speeds of widely differing magnitudes occur).

Experiment 1

The scalar equation

$$(\partial u/\partial t) + (\partial/\partial x)(\tfrac{1}{2}u^2) = 0 \tag{5.1}$$

was solved over the region $[0 < x \leqslant 1] \times [t > 0]$ for the following three sets of initial and boundary conditions;

(a)
$$\begin{aligned} u_0(x) &= x \\ u_1(t) &= 0, \end{aligned} \tag{5.2}$$

(b)
$$\begin{aligned} u_0(x) &= x^2 \\ u_1(t) &= 0, \end{aligned} \tag{5.3}$$

and

$$\text{(c)} \qquad \begin{aligned} u_0(x) &= \sqrt{x} \\ u_1(t) &= 0. \end{aligned} \tag{5.4}$$

The solutions of these three problems are

$$u(x, t) = x/(1 + t), \tag{5.5}$$

$$u(x, t) = (1 + 2xt - (1 + 4xt)^{1/2})/2t^2, \tag{5.6}$$

TABLE I

Errors (Multiplied by 10^6), After 300 Time Steps, at a Central Grid Point for the Problem with Initial Data Given by x

p	a \ d	0.125	0.25	0.5	1.0	2.0
(i)	0.25	−97	−31	−4	0	5
	0.5	−92	−27	0	4	8
1.0	1.0	44	11	3	6	12
	2.0	21	−0	0	−1	−3
	4.0	−6	−12	−26	−58	−156
	0.25	−433	−117	−11	2	6
	0.5	−241	−50	−3	5	10
2.0	1.0	−134	−11	3	6	*
	2.0	−14	−1	−1	−4	*
	4.0	−8	−16	−35	−86	*
(ii)	0.25	56	21	2	2	5
	0.5	7	3	2	4	9
1.0	1.0	−0	1	3	5	12
	2.0	−1	−0	−1	−1	*
	4.0	−6	−13	−26	*	*
	0.25	151	32	2	3	6
	0.5	−2	1	2	5	10
2.0	1.0	−4	1	3	6	*
	2.0	−1	−1	−2	*	*
	4.0	−8	−16	*	*	*

* Denotes nonlinear instability had occurred.

and

$$u(x, t) = (-t + (t^2 + 4x)^{1/2})/2, \tag{5.7}$$

respectively. It is noted that in (5.6)

$$\lim_{t \to 0} u(x, t) = u_0(x)$$

as given by (5.3).

The errors (the differences between w and u) for the above three problems are given in Tables I, II, and III, respectively. Each of these tables is divided into two

TABLE II

Errors (Multiplied by 10^6), After 300 Time Steps, at a Central Grid Point for the Problem with Initial Data x^2

p	a \ d	0.125	0.25	0.5	1.0	2.0
(i)	0.25	−607	−501	−440	−274	121
	0.5	−281	−423	−418	−209	187
1.0	1.0	−687	−550	−371	−185	200
	2.0	−645	−458	−366	−202	159
	4.0	−532	−458	−399	−262	10
	0.25	−998	−404	−271	−186	96
	0.5	−1443	−873	−329	−146	138
2.0	1.0	−1284	−446	−267	−137	138
	2.0	−457	−325	−271	−160	82
	4.0	−371	−328	−304	*	*
(ii)	0.25	−1150	−801	−472	−284	121
	0.5	−455	−476	−408	−209	188
1.0	1.0	−519	−471	−371	−187	201
	2.0	−534	−458	−371	−201	*
	4.0	−532	−464	−397	*	*
	0.25	−364	−320	−332	−190	96
	0.5	−384	−337	−281	−146	137
2.0	1.0	−347	−324	−267	−137	*
	2.0	−356	−325	−272	−160	*
	4.0	−371	−334	*	*	*

* Denotes nonlinear instability had occurred.

TABLE III

Errors (Multiplied by 10^6), After 300 Time Steps, at a Central Grid Point for the Problem with Initial Data $\sqrt{x}$

p	a \ d	0.125	0.25	0.5	1.0	2.0
(i)	0.25	1108	853	505	111	−104
	0.5	971	673	317	33	−104
1.0	1.0	718	485	235	33	−83
	2.0	585	421	235	61	*
	4.0	538	404	221	−17	*
	0.25	614	475	233	37	−24
	0.5	584	335	134	23	−14
2.0	1.0	355	225	111	37	*
	2.0	279	198	116	*	*
	4.0	249	176	49	*	*
(ii)	0.25	1373	961	521	112	−104
	0.5	981	679	319	33	−104
1.0	1.0	702	484	235	33	−83
	2.0	581	421	234	61	*
	4.0	538	403	221	*	*
	0.25	955	563	242	38	−24
	0.5	535	328	134	23	*
2.0	1.0	336	221	111	42	*
	2.0	273	198	116	*	*
	4.0	249	176	*	*	*

* Denotes nonlinear instability had occurred.

parts (i) and (ii) where (i) contains the results obtained using an explicit method to give values on $x = 1$, and (ii) contains those obtained using the formulation (4.9) at the upper boundary. In all cases, $h = 0.1$ and the errors are given after 300 time steps.

In each of the examples, the maximum value of the solution is 1 which occurs at $t = 0$ and the value of the solution decreases with increasing time. Thus, for stability in the linearized sense when $p = 1$, ad must lie in the interval $[0, \frac{1}{2}]$. From the results, however, it is seen that the schemes are stable (at least over 300 time steps) for a larger range of ad than is predicted by the linearized theory. The same

remark holds when $p = 2.0$. It is also to be noted that instability arises for smaller values of ad when $p = 2.0$ than when $p = 1.0$. Also, the errors for both $p = 1.0$ and 2.0 are of the same magnitude. Further, provided the stability condition was satisfied, best values of a and d for minimizing the truncation errors could be chosen from an analysis of the truncation error terms for both cases $p = 1.0$ and $p = 2.0$. However, because of greater efficiency, the largest value of p possible would be chosen in practice.

Also, comparison of parts (i) and (ii) of each of Tables I, II, and III shows that instability occurs for smaller values of ad in (ii) than in (i). Thus, it is recommended that, when using this method, the upper boundary values should be predicted by an explicit scheme.

Hence, for problems of the type shown here, p can be chosen greater than unity and still reasonable answers obtained. ad should then be chosen in the range allowed by the linearized stability theory and the conditions for a dissipative scheme. Further, smaller errors are given when ad is chosen near the top of this range. Finally, since $\overset{*}{w}_i^{m+a}$ is an approximation to u_i^{m+a}, a should be chosen so as to keep this approximation a reasonable one. For example, the choice $a = \frac{1}{2}$ gives an extremely centralized scheme involving values at (i, m), $(i \pm 1, m)$, $(i \pm \frac{1}{2}, m + \frac{1}{2})$, $(i \pm 1, m + 1)$, and $(i, m + 1)$.

Experiment 2

In this experiment, (5.1) was solved with discontinuous initial data,

$$u_0(x) = \begin{cases} 1 & 0 \leqslant x < 0.1, \\ 0 & x \geqslant 0.1, \end{cases} \tag{5.8}$$

and the boundary data,

$$u_1(t) = 1 \qquad t > 0. \tag{5.9}$$

The solution of this problem has a discontinuity travelling into the field of solution along the line $x = 0.1 + \frac{1}{2}t$ with velocity $dx/dt = 1/2$.

Since the solution of the schemes will be zero at all points above, and far from, the line of discontinuity, it does not matter what boundary technique is used at the upper boundary. The value at this boundary will always equal zero.

In Fig. 1, a selection of the results of this experiment are graphed. In each graph, the values of the solution given by the difference scheme after 50 time steps are plotted for grid points between $x = 25 * p * h$ and $x = 25 * p * h + 15 * h$. The theoretical shock position occurs at $x = 25 * p * h + 10 * h$. In all runs, h was taken equal to 0.01. The values of $d = 0.125, 0.25, 0.5, 1.0, 2.0, 4.0$ for a particular value of a where used as parameters on each of the runs. In those cases where some

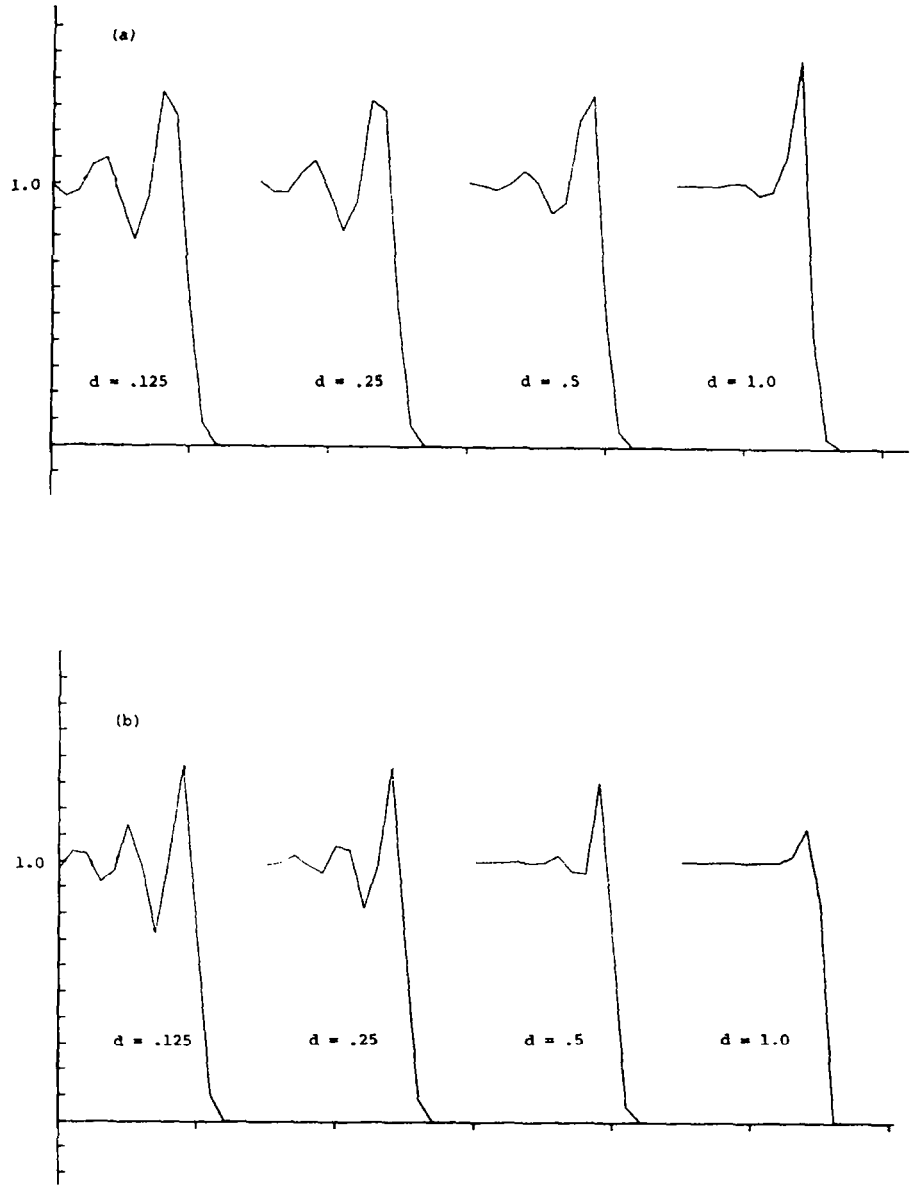
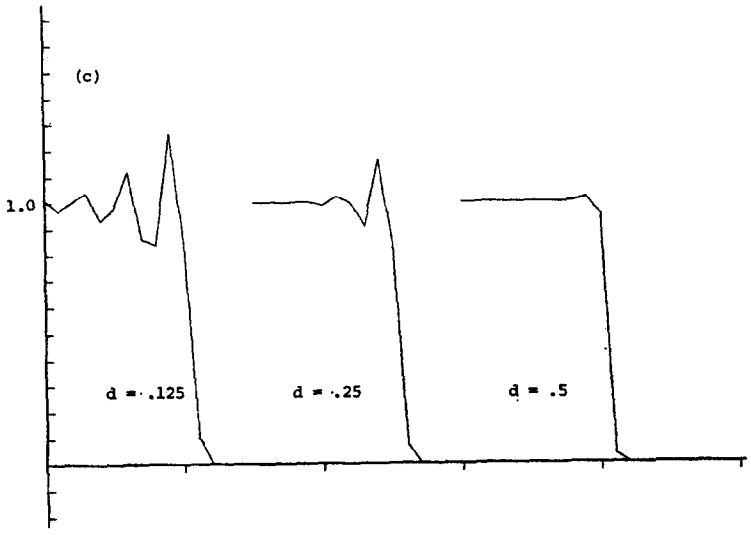


FIG. 1. Graphs of the solution, after 50 time steps, obtained using the generalized implicit scheme with $p = 1.0$ and $h = 0.01$, for the problem of experiment 2. (a) $a = 0.25$; $d = 2.0$ gave floating point overflow. (b) $a = 0.5$; $d = 2.0$ gave floating point overflow.



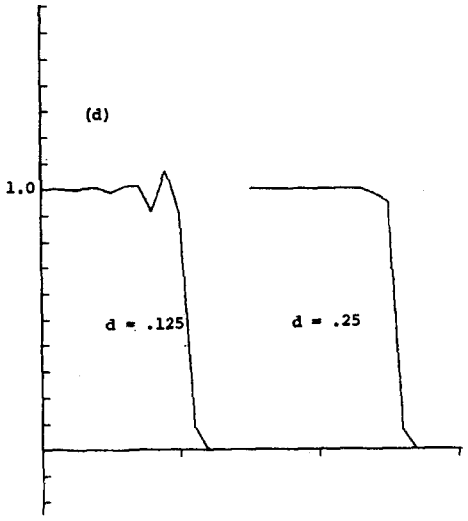


FIG. 1 (Continued). (c) $a = 1.0$; $d = 1.0$ gave floating point overflow. (d) $a = 2.0$; $d = .5$ gave floating point overflow.

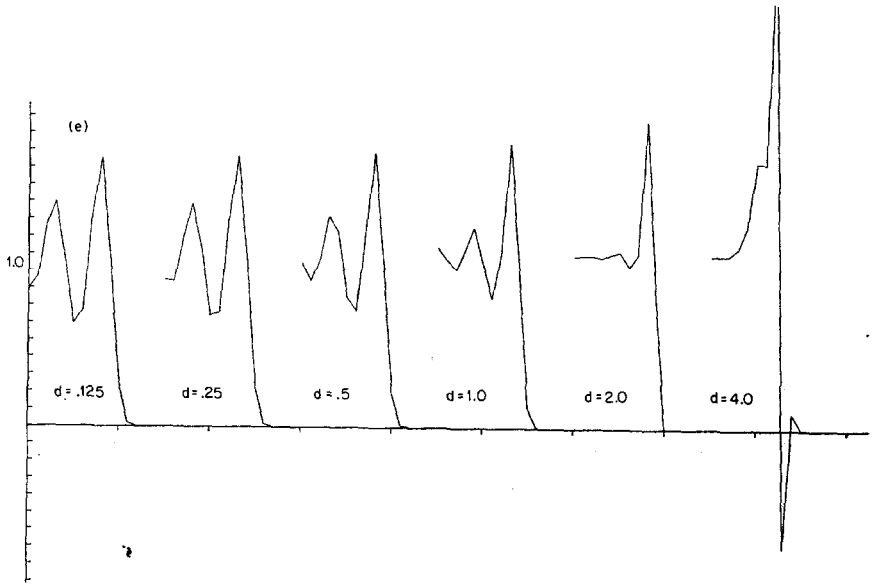


FIG. 1 (Continued). $p = 0.5$ and $h = 0.01$. (e) $a = 0.25$.

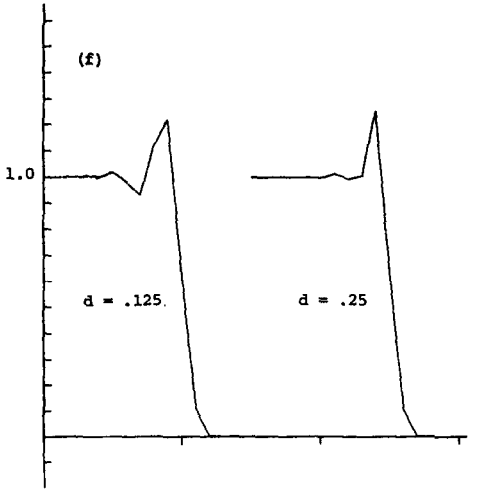


FIG. 1 (Continued). $p = 2.0$ and $h = 0.01$. (f) $a = 0.25$; $d = .5$ gave floating point overflow.

of the values of d have no graph, the values above the last which have a graph gave floating point overflow.

In this experiment, the maximum value of the theoretical solution is 1 and so the linearized stability condition is

$$p \leqslant 1/(2ad)^{1/2}.$$

When $p = 1$, the allowed range of ad is given by

$$0 \leqslant ad \leqslant 1/2.$$

This range of ad was fairly well adhered to in the praphs of Fig. 1. Also, for the graphs of $p = 0.5$ and $p = 2.0$, the stability range was adhered to. The best shock profiles were obtained by choosing p and ad close to the maximum values allowed by the stability conditions. Thus, for $p = 1$, the best profiles were given by taking $ad = 1/2$. Also, the larger the value of a (within reasonable limits), the better was the profile. Also, values of p with $p > 1$ gave excellent profiles with this method.

Thus, it is recommended that a fairly large time step be chosen; then, a large value of a, close to 2.0; and then a value of d to make the time step close to the stability limit.

It is also noted that the position of the discontinuity was in all cases within 1 grid point of the theoretical position.

Experiment 3

The physical system expressing the conservation of mass, momentum, and energy for an inviscid nonheat-conducting compressible ideal gas was solved by the schemes of Section 2 with the computational details of Section 4 incorporated into the formulation. The system was set up with a shock moving through the gas. This was exactly the same problem as used by Rubin and Burstein [11] and McGuire and Morris [7].

In the experiment, $\tilde{A}(u)$ was chosen to be a diagonal matrix. This led to an efficient solution of the block tridiagonal system of difference equation at each time level.

Graphs were obtained for the density for a series of values of a, d, and p. However, the profiles were poor in comparison to those obtained with the explicit schemes of McGuire and Morris [7]. Hence we omit these graphs.

We feel that $\tilde{A}$ requires to be chosen with greater care than we used in this experiment and that proper choice of $\tilde{A}$ will give good profiles. We are currently investigating the choice of $\tilde{A}$.

6. Extension of the Implicit Schemes to Problems in Two Space Dimensions

In this section, the scheme (2.1), (1.4), and (4.10) is extended to solving systems of conservation laws in two space dimensions, namely,

$$\begin{aligned}(\partial u/\partial t) + (\partial f/\partial x)(u) + (\partial g/\partial y)(u) &= 0, \\ (x, y) \in G \equiv [0, X] \times [0, Y],\ t &> 0,\end{aligned} \tag{6.1}$$

with appropriate initial and boundary conditions.

A uniform grid of spacing h is placed on G and a time step of size k is placed on the time axis. Without loss of generality, we let $X = N1 * h$ and $Y = N2 * h$. w_{ij}^m is used to denote an approximation to $u(ih, jh, mk)$.

Extending the one space dimensional scheme in the fashion of Richtmyer [9] gives the scheme

$$\begin{aligned}\overset{*}{w}{}_{ij}^{m+a} &= \tfrac{1}{4}(w_{i+\frac{1}{2},j}^m + w_{i-\frac{1}{2},j}^m + w_{i,j+\frac{1}{2}}^m + w_{i,j-\frac{1}{2}}^m) \\ &\quad - ap[(f_{i+\frac{1}{2},j}^m - f_{i-\frac{1}{2},j}^m) + (g_{i,j+\frac{1}{2}}^m - g_{i,j-\frac{1}{2}}^m)],\end{aligned} \tag{6.2}$$

$$\begin{aligned}w_{ij}^{m+1} &= w_{ij}^m - p[b(\tilde{A}_{i+1,j}^{**} w_{i+1,j}^{m+1} - \tilde{A}_{i-1,j}^{**} w_{i-1,j}^{m+1}) \\ &\quad + b(\tilde{B}_{i,j+1}^{**} w_{i,j+1}^{m+1} - \tilde{B}_{i,j-1}^{**} w_{i,j-1}^{m+1}) + c(f_{i+1,j}^m - f_{i-1,j}^m) \\ &\quad + c(g_{i,j+1}^m - g_{i,j-1}^m) + d(\overset{*}{f}{}_{i+\frac{1}{2},j}^{m+1} - \overset{*}{f}{}_{i-\frac{1}{2},j}^{m+1}) + d(\overset{*}{f}{}_{i,j+\frac{1}{2}}^{m+1} - \overset{*}{f}{}_{i,j-\frac{1}{2}}^{m+1})],\end{aligned} \tag{6.3}$$

$$f(u) = \tilde{A}(u) \cdot u, \qquad g(u) = \tilde{B}(u) \cdot u,$$

$$\tilde{A}_{i,j}^{**} = \tilde{A}(\overset{**}{w}{}_{i,j}^{m+1}), \qquad \tilde{B}_{i,j}^{**} = \tilde{B}(\overset{**}{w}{}_{i,j}^{m+1}),$$

$$\overset{**}{w}{}_{ij}^{m+1} = \frac{1}{4a}(\overset{*}{w}{}_{i+\frac{1}{2},j}^{m+a} + \overset{*}{w}{}_{i-\frac{1}{2},j}^{m+a} + \overset{*}{w}{}_{i,j+\frac{1}{2}}^{m+a} + \overset{*}{w}{}_{i,j-\frac{1}{2}}^{m+a}) + \left(1 - \frac{1}{a}\right) w_{ij}^m. \tag{6.4}$$

It is easily established that the scheme is second-order accurate provided

$$\begin{aligned}2b + 2c + d &= 1, \\ 2b + ad &= 1/2.\end{aligned} \tag{6.5}$$

Furthermore, $\overset{**}{w}{}_{ij}^{m+1}$ is a first-order approximation to u_{ij}^{m+1}.

Consideration of Eqs. (6.2) and (6.3) easily shows that this formulation cannot be used over a grid of size h, as "halfway" values $w_{i+\frac{1}{2},j\pm\frac{1}{2}}^m$, $w_{i\pm\frac{1}{2},j+\frac{1}{2}}^m$ are required. Hence, as explained in McGuire and Morris [7] for the explicit schemes, resort to some procedure like that of Thommen's paper [13] for predicting the starred values must be made. This makes the formulas for the starred values even more complicated. Also, (6.3) requires the solution of a 5-band block matrix system

at each time level. This really makes this scheme unworkable as no simple algorithms exist for solving such systems. Other Richtmyer-type extensions can also be made (see Wilson [14]) but each of them suffers from the disadvantage of the 5-band block matrix system at each time level.

An efficient extension to two space dimensions may be provided by resorting to Strang's formulation [15]

$$w^{m+1} = L_{x/2}L_yL_{x/2}w^m \tag{6.6}$$

with $L_{x/2}$ operators combined at the end of one step and the beginning of the next. In (6.6), L_x denotes the application of the one-dimensional scheme in the x direction and L_y denotes its application in the y direction. $L_{x/2}$ is simply L_x with p replaced by $p/2$. For further details, see Strang [15]. The computational procedure for this method when the operators L_x and L_y are defined in terms of two-step Richtmyer methods is given in Gourlay and Morris [16]. The extension of their ideas to the case when L_x and L_y are the implicit operators of earlier sections is straightforward. The application of this algorithm thus requires the solution of approximately two block tridiagonal systems per time step.

The linearized stability properties of the method are given by

$$p\,|\,\lambda\,| \leqslant 1/(2ad)^{1/2},$$
$$p\,|\,\mu\,| \leqslant 1/(2ad)^{1/2},$$

where $|\,\lambda\,|$, $|\,\mu\,|$ are the maximum modulus eigenvalues of A and B. It is noted that a and d in the operators L_x and L_y need not have the same values for both operators. Thus the stability properties can be

$$p\,|\,\lambda\,| \leqslant 1/(2a_xd_x)^{1/2},$$
$$p\,|\,\mu\,| \leqslant 1/(2a_yd_y)^{1/2},$$

using an obvious notation.

Explicit dissipation may also be added to the method in a one-dimensional manner in a similar fashion to that for the explicit methods of McGuire and Morris [7].

Some difficulties are experienced when using the scheme near the boundaries. Consider the first step of the algorithm; it is given by

$$w_{ij}^{(1)} = (w_{i+\frac{1}{2},j}^m + w_{i-\frac{1}{2},j}^m)/2 - ap(f_{i+\frac{1}{2},j} - f_{i-\frac{1}{2},j}^m), \tag{6.7}$$

$$\begin{aligned} w_{ij}^{(2)} = {} & w_{ij}^m - p[b(\bar{A}_{i+1,j}^{**}w_{i+1,j}^{(2)} - \bar{A}_{i-1,j}^{**}w_{i-1,j}^{(2)}) \\ & + c(f_{i+1,j}^m - f_{i-1,j}^m) + d(f_{i+\frac{1}{2},j}^{(1)} - f_{i-\frac{1}{2},j}^{(1)})], \end{aligned} \tag{6.8}$$

$$\bar{A}_{ij}^{**} = \bar{A}(\overset{**}{w}_{ij}),$$

$$\overset{**}{w}_{ij} = \frac{1}{2a}(w_{i+\frac{1}{2},j}^{(1)} + w_{i-\frac{1}{2},j}^{(1)}) + \left(1 - \frac{1}{a}\right) w_{ij}^m. \tag{6.9}$$

At the upper boundary of x, the values $w^{(2)}_{N1,j}$ are predicted by a second-order-accurate explicit formula which uses backward differences. For example,

$$w^{(1)}_{N1+\frac{1}{2},j} = (4w^m_{N1,j} - 3w^m_{N1-1,j} + w^m_{N1-2,j})/2 - ap(2f^m_{N1,j} - 3f^m_{N1-1,j} + f^m_{N1-2,j}), \tag{6.10}$$

$$w^{(2)}_{N1,j} = w^m_{N1,j} - \frac{p}{2}\left[\left(1 - \frac{1}{2a}\right)3f^m_{N1,j} - 4f^m_{N1-1,j} + f^m_{N1-2,j}) + \frac{1}{a}(f^{(1)}_{N1+\frac{1}{2},j} - f^{(1)}_{N1-\frac{1}{2},j})\right]. \tag{6.11}$$

can be used.

The values of $w^{(2)}_{0,j}$ are also required to be given in order to solve (6.8) along each y grid line. These values $w^{(2)}_{0,j}$ can be provided by the forward difference version of any second-order-accurate explicit formula. Ideally, it is inadvisable to use forward difference formulas at the lower boundary since they interpolate in a direction opposite to that of the characteristics of the differential equations. However, it is difficult to construct formulas for providing $w^{(2)}_{0,j}$ without using forward differences.

We denote the complete step of the algorithm for obtaining $w^{(2)}_{ij}$, viz., (6.7), (6.8), (6. 9) with (6. 10), (6. 11) and the equivalent formulas for $w^{(2)}_{0,j}$, by

$$w^{(2)} = L^I_{x/2}w^m. \tag{6.12}$$

This formula is applied along grid lines $j = 0, 1, 2,..., N2$. The second step is built up in exactly the same way and is denoted by

$$w^{(4)} = L^I_y w^{(2)}. \tag{6.13}$$

It is obvious how succeeding steps are built up.

The question of incorporating given boundary data is answered by the same arguments as used in McGuire and Morris [6]. To maintain overall second-order accuracy, the second-order-accurate boundary procedure of [6] should be used. However, due to the excellent results obtained with the first-order procedure of [6], we use the first-order boundary procedure in the numerical experiments.

Thus, we use the formula

$$w_0 = L^I_{x/2}u_0{}^m \qquad n = 1, 2,..., \tag{6.14}$$

where $u_0{}^m$ and w_0 denote values of u^m and w along the boundary $y = 0$, before each application of $L^{(I)}_y$. This formula provides data on the lower y boundary.

Similarly, data on the lower x boundary is provided by the formula

$$w_0 = L^I_{y/2}u_0^{m+1/2} \qquad m = 0, 1, 2,..., \tag{6.15}$$

where u_0^m and w_0 now denote values of u^m and w along the boundary $x = 0$, applied before each application of L_x^I.

Other procedures besides Strang's formulation (6.6) may be used to extend the one-dimensional schemes to two space dimensions. A procedure comparable in efficiency is given by

$$w^{m+2} = L_x L_y L_y L_x w^m. \tag{6.16}$$

This method is second-order accurate at even time levels. Its properties and implementation are considered in McGuire and Morris [8] for the case when L_x and L_y are explicit one-dimensional operators. The details in [8] are easily extended to the case when L_x and L_y are implicit.

7. Numerical Experiments in Two Space Dimensions

In this section, the scheme of Section 6 was applied to a simple example, namely, (6.1) with

$$f(u) = g(u) = \tfrac{1}{4}u^2. \tag{7.1}$$

Initial conditions and boundary conditions given were

$$u(x, y, 0) = \tfrac{1}{4}(x + y)^2, \tag{7.2}$$

$$u(0, y, t) = \left\{\frac{1 - (1 + yt)^{1/2}}{t}\right\}^2, \tag{7.3}$$

$$u(x, 0, t) = \left\{\frac{1 - (1 + xt)^{1/2}}{t}\right\}, \tag{7.4}$$

and these give the solution of (6.1) in

$$\{G = [0 \leqslant x \leqslant 1] \times [0 \leqslant y \leqslant 1]\} \times [t > 0] \tag{7.5}$$

to be

$$u(x, y, t) = \left\{\frac{1 - (1 + (x + y)\, t)^{1/2}}{t}\right\}^2. \tag{7.6}$$

The errors (differences between w and u) are given in Table (IV) for various values of a, d, and p. h was chosen equal to 0.1.

The first-order boundary procedure of McGuire and Morris [6] was used to incorporate the given boundary data. It is easily shown (see McGuire and Morris

[8]) that, in fact, for this equation (6.1) with (7.1), the first-order boundary procedure is second-order accurate. This emerges because certain derivatives in the error terms cancel. Thus for this problem, the scheme is overall second-order accurate.

The maximum value of the solution is 1 and this occurs at $t = 0$. Thereafter the solution decreases. Thus when $p = 1.0$, the stability condition requires that

TABLE IV

Errors (Multiplied by 10^6) at a Central Grid Point After 50 Time Steps

p	a \ d	0.125	0.25	0.5	1.0	2.0	4.0
(a)	0.25	−440	−412	−353	−223	57	627
	0.5	−406	−369	−296	−152	132	697
1.0	1.0	−358	−323	−254	−115	159	724
	2.0	−336	−306	−245	−122	131	665
	4.0	−336	−320	−286	−212	−61	218
(b)	0.25	−277	−253	−206	−116	54	407
	0.5	−221	−201	−159	−76	95	473
4.0	1.0	−209	−186	−145	−64	106	498
	2.0	−204	−185	−152	−89	50	*
	4.0	−213	−213	−213	−232	*	*

* Denotes nonlinear instability had occurred.

$0 \leqslant ad \leqslant 1/2$. From the results, it is seen that stability (at least for 50 time steps) does in fact hold over a larger range. The same remarks holds for $p = 4.0$.

As regards minimizing truncation errors, an explicit analysis would provide the best values of a and d to use. From Table IV, the best value of d seems to be between 1.0 and 2.0 for the case when $a = 0.25, 0.5, 1.0, 2.0$ and $p = 1.0$. A similar remark can be made for $p = 4.0$.

In using this method, it would seem that a large time step can be used in the case of problems like the one used here. Also, ad should be chosen so as to make the value of p close to the maximum allowed by the linear stability conditions. Then, in the absence of any other criteria, a could be chosen as in the one-dimensional case, that is, approximately unity. This choice of a has the effect of centralizing the one-space-dimensional operators and hence one hopes this will give small truncation errors.

REFERENCES

1. R. COURANT, K. FRIEDRICHS, AND M. LEWY, *IBM J. Res. Develop.* **11** (1967), 213–234.
2. J. GARY, *Math. Comp.* **18** (1964), 1–18.
3. A. R. GOURLAY AND J. LL. MORRIS, *Math. Comp.* **22** (1968), 28–39.
4. A. R. GOURLAY AND J. LL. MORRIS, *Math. Comp.* **22** (1968), 549–556.
5. A. JEFFREY AND T. TANUITI, *Nonlinear Wave Propogation*, Academic Press, New York, 1964.
6. G. R. MCGUIRE AND J. LL. MORRIS, *J. Inst. Math. Appl.* **10** (1972), 150–165.
7. G. R. MCGUIRE AND J. LL. MORRIS, A class of second-order accurate methods for the solution of systems of conservation laws, *J. Computational Phys.* **11** (1973), 531–549.
8. G. R. MCGUIRE AND J. LL. MORRIS, Restoring orders of accuracy for multilevel schemes for nonlinear hyperbolic systems in many space variables, manuscript.
9. R. D. RICHTMYER, A survey of difference methods for non-steady fluid dynamics, NCAR Tech. Notes 63-2, 1962.
10. R. D. RICHTMYER AND K. W. MORTON, *Difference Methods for Initial Value Problems*, Wiley, New York, 1967.
11. E. L. RUBIN AND S. Z. BURSTEIN, *J. Computational Phys.* **2** (1967), 178–196.
12. R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.
13. H. U. THOMMEN, *Z. Angew. Math. Phys.* **17** (1966), 369–384.
14. J. C. WILSON, *J. Inst. Math. Appl.* **10** (1972), 238–257.
15. G. STRANG, *Siam J. Numer. Anal.* **5** (1968), 506–517.
16. A. R. GOURLAY AND J. LL. MORRIS, *J. Computational Phys.* **5** (1970), 229–243.
17. S. ABARBANEL AND G. ZWAS, *Math. Comp.* **23** (1969), 549–565.