JOURNAL OF COMPUTATIONAL PHYSICS **11**, 531–549 (1973)

A Class of Second-Order Accurate Methods for the Solution of Systems of Conservation Laws

G. R. McGUIRE

Department of Mathematics, University of Aberdeen

AND

J. Ll. MORRIS

Department of Mathematics, University of Dundee

Received May 4, 1972

A generalization of the two-step Richtmyer [7] method is derived for the solution of first-order systems of conservation laws. It is analysed with respect to stability and dissipation for the purpose of giving good solutions to shock problems. Applications to smooth, discontinuous, and physical problems are described briefly. The extension of the method to systems of equations in two space dimensions is achieved through Strang's formulation [19].

1. INTRODUCTION

In this paper we consider finite difference methods for the solution of systems of conservation laws

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = 0, \tag{1.1}$$

where $f = f(u)$ and $u = u(x, t)$ is a vector function $(u_1, u_2, ..., u_n)^T$. We consider the equation to be defined in $(0 < x \leqslant \alpha) \times (0 < t \leqslant T)$ and assume initial conditions

$$u(x, 0) = u_0(x) \tag{1.2}$$

are given on $0 \leqslant x \leqslant \alpha$.

We further assume that $A(u) = \partial f/\partial u$, the Jacobian of f with respect to u, has positive eigenvalues, so that boundary conditions

$$u(0, t) = u_1(t) \tag{1.3}$$

are given. Under these assumptions (1.1) with the extra conditions (1.2) and (1.3) is a well posed problem in the given region.

If the differentiation of f is carried out in (1.1) we obtain

$$\frac{\partial u}{\partial t} + A(u)\frac{\partial u}{\partial x} = 0. \tag{1.4}$$

For problems of the form (1.4) which have smooth solutions Strang [1] showed that, as regards convergence of difference schemes, there was very little difference between the nonlinear case (1.4) and the variable coefficient case in which A depends on x and t but not on u. The case in which $u(x, t)$ has discontinuities is much more difficult to analyze. The best point of view from which this problem can be attacked seems to be that given by Kreiss and Widlund [2] where they consider the internal discontinuities as internal boundaries on which no information is given. Hence they reduce the problem to one in which wrong or no boundary data are given but is in all other respects a problem with a smooth solution. In [3] Kreiss and Lundquist considered this problem for both the constant and variable coefficient scalar cases and found that for a dissipative difference scheme the influence of the wrong boundary data could be confined to a narrow band near the boundary. In [4] Apelkrans treats the scalar variable coefficient problem with discontinuous initial data for the situation when one uses a difference scheme throughout incorporating no special treatment when a discontinuity is met and shows that again errors introduced by the discontinuity can be confined to a narrow band around the discontinuity provided the scheme is dissipative. However when the case of a system of hyperbolic differential equations is considered Apelkrans [5] gives some rather discouraging results proving that the effect of the discontinuity is in fact spread out over a fairly large region. From his results it thus seems, at least for systems, that it is probably best to do something special near a discontinuity like "shock-fitting," in order to reduce the spread in the discontinuous region. More recently in [6], for variable coefficient hyperbolic systems, Gustafsson has shown that if a finite difference method of order m is used at internal points and one of order $m - 1$ at the boundaries, then under certain assumptions the overall accuracy of order m is maintained. Thus considering the discontinuity as an internal boundary, the use of a second-order method throughout the region means that at this boundary we only have a first order method. However Gustafsson's theory shows that under certain assumptions, the scheme may still be second-order accurate away from the discontinuity.

In this paper we consider the use of Lax-Wendroff (LW) type methods for the solution of systems of conservation laws (1.1). We describe a generalization of the two-step Richtmyer [7] version of the Lax-Wendroff method in Section 2. The dissipative properties both linear and nonlinear are discussed along with the

stability of the generalized scheme in Section 3. Numerical experiments were conducted using the subject methods. The relevant results are reported briefly in Section 4. The extension of the one-dimensional scheme to systems of conservation laws in two space dimensions is considered in Section 5. The paper is concluded in Section 6.

2. Generalized LW Scheme

Second order schemes approximate, correct to second order, the Taylor series expansion of[1] $u_i^{m+1} = u(ih, (m+1)K)$; that is

$$u_i^{m+1} = u_i^m + K\frac{\partial u_i^m}{\partial t} + \frac{K^2}{2}\frac{\partial^2 u_i^m}{\partial t^2} + O(K^3). \tag{2.1}$$

Different ways of replacing the time derivatives by space derivatives using the differential equations (1.1) give different schemes. In replacing the time derivatives care must be taken not to introduce the Jacobian $A(u) = \partial f/\partial u$ into (2.1) since then inefficient algorithms like the original LW scheme will result.

In this section, a generalization of the two step Richtmyer scheme [7] is introduced by using an intermediate approximation to $u(ih, (m+2a)K)$ as was done in [10]. We write (2.1) as

$$u_i^{m+1} = u_i^m + K\left(1 - \frac{1}{4a}\right)\frac{\partial u_i^m}{\partial t} + \frac{K}{4a}\frac{\partial}{\partial t}\left(u + 2aK\frac{\partial u}{\partial t}\right)_i^m + O(K^3) \tag{2.2}$$

$$= u_i^m + K\left(1 - \frac{1}{4a}\right)\frac{\partial u_i^m}{\partial t} + \frac{K}{4a}\frac{\partial u_i^{m+2a}}{\partial t} + O(K^3). \tag{2.3}$$

Using (1.1) in (2.3) then gives

$$u_i^{m+1} = u_i^m - K\left(1 - \frac{1}{4a}\right)\frac{\partial f_i^m}{\partial x} - \frac{K}{4a}\frac{\partial f_i^{m+2a}}{\partial x} + O(K^3), \tag{2.4}$$

with

$$u_i^{m+2a} = u_i^m - 2aK\frac{\partial f_i^m}{\partial x} + O(K^3), \tag{2.5}$$

and where f_i^m means $f(u_i^m)$.

Replacement of the space derivatives by differences which do not destroy the orders of accuracy of the first terms on the right hand sides of (2.4) and (2.5),

[1] h is the grid spacing, K the step size and $p = K/h$ is the mesh ratio.

and which give a scheme in which values at only (i, m), $(i \pm 1, m)$ are involved, results in the scheme

$$\overset{*}{\omega}_i^{m+1} = \tfrac{1}{2}(\omega_{i+1/2}^m + \omega_{i-1/2}^m) - 2ap(f_{i+1/2}^m - f_{i-1/2}^m), \tag{2.6}$$

$$\omega_i^{m+1} = \omega_i^m - \frac{p}{2}\left[\left(1 - \frac{1}{4a}\right)(f_{i+1}^m - f_{i-1}^m) + \frac{1}{2a}(\overset{*}{f}_{i+1/2}^{m+1} - \overset{*}{f}_{i-1/2}^{m+1})\right], \tag{2.7}$$

where ω_i^m is an approximation to u_i^m and f_i^m means $f(\omega_i^m)$. In (2.6), (2.7), as in [10], $\overset{*}{\omega}_i^{m+1}$ is a first-order approximation to $u(ih, (m + 2a)K)$; $a = \frac{1}{4}$ gives the two-step Richtmyer scheme; and $a = \frac{1}{2}$ gives the scheme introduced by Rubin and Burstein in [9]. Applied to linear problems with constant coefficients (2.6), (2.7) reduces to the LW method and hence has optimal linear stability (see [8], p. 304) as allowed under the Courant–Friedrichs–Lewy condition [11], that is

$$p \mid \lambda \mid \leqslant 1, \tag{2.8}$$

where $\mid \lambda \mid$ is the maximum modulus eigenvalue of A. In [10], the difference operators were chosen in such a way that points other than (i, m), $(i \pm 1, m)$ were used in the evaluation of ω_i^{m+1} so that a more restrictive (linear) stability condition than (2.8) was obtained.

In their study of third-order methods, Burstein and Mirin [12] considered equations which were constituent steps of a third-order method and which were similar to Eqs (2.6) and (2.7).

3. Dissipative Properties of the Generalized Scheme

In this section we look at the dissipative properties of (2.6) and (2.7), and as a result try to indicate how one should choose the parameter a.

Performing the usual Fourier analysis on the linearized version of (2.6) and (2.7), which reduces to the LW method, shows that (2.8) is sufficient as well as necessary for stability [8, p. 84]. Also, (see [8, p. 109]) the scheme is, in the linearized case, dissipative of order 4 provided

$$0 < p \mid \lambda \mid < 1. \tag{3.1}$$

Since the linearization of (2.6) and (2.7) eliminates a, we will turn to an analysis of the truncation error terms to decide what effect a particular value of a has on the values given by the scheme. To examine the truncation error we must assume that u and f have as many bounded derivatives as are involved in the expression for the truncation error.

Applying the operator of (2.6) at $x = ih$ and denoting it by $L_x{}^*$ we have

$$L_x{}^* u_i{}^n = u_i{}^n - 2aK \frac{\partial f_i{}^n}{\partial x} + \frac{1}{8} h^2 \frac{\partial^2 u_i{}^n}{\partial x^2} - \frac{2ap}{24} h^3 \frac{\partial^3 f_i{}^n}{\partial x^3} + O(h^4), \tag{3.2}$$

where $f_i{}^n$ means $f(u_i{}^n)$. Thus since the coefficient of h^3 in (3.2) is smooth,

$$f(L_x{}^* u_{i+1/2}^n) - f(L_x{}^* u_{i-1/2}^n) = h \frac{\partial f(u_i^{n+2a})}{\partial x} + h^3 \frac{\partial f(u_i{}^n)}{\partial u} \left\{ \frac{1}{8} \frac{\partial^2 u_i{}^n}{\partial x^2} - 2a^2 p^2 \frac{\partial^2 u_i{}^n}{\partial t^2} \right\} + \frac{h^3}{24} \frac{\partial^3 f(u_i{}^n)}{\partial x^3} + O(h^4), \tag{3.3}$$

where we have used (1.1) and where $\partial f / \partial u$ means the Jacobian of f. Hence, applying the operator of (2.7) at $x = ih$ and denoting it by L_x, we have

$$\begin{aligned} L_x u_i{}^n = u_i{}^n &- K \frac{\partial f_i{}^n}{\partial x} - \frac{K^2}{2} \frac{\partial^2 f_i{}^n}{\partial t \, \partial x} - \frac{Kh^2}{6} \frac{\partial^3 f_i{}^n}{\partial x^3} \\ &+ \frac{Kh^2}{32a} \frac{\partial}{\partial x} \left\{ \frac{\partial^2 f_i{}^n}{\partial x^2} - \frac{\partial f_i{}^n}{\partial u} \cdot \frac{\partial^2 u_i{}^n}{\partial x^2} \right\} \\ &+ \frac{aK^3}{2} \frac{\partial}{\partial x} \left\{ \frac{\partial^2 f_i{}^n}{\partial t^2} - \frac{\partial f_i{}^n}{\partial u} \cdot \frac{\partial^2 u_i{}^n}{\partial t^2} \right\} + O(h^4), \end{aligned} \tag{3.4}$$

where again (1.1) has been used.

Now using the fact that

$$\begin{aligned} \frac{\partial^2 f}{\partial x^2} &= \frac{\partial}{\partial x} \left(\frac{\partial f}{\partial u} \cdot \frac{\partial u}{\partial x} \right) \\ &= \frac{\partial}{\partial x} \left(\frac{\partial f}{\partial u} \right) \cdot \frac{\partial u}{\partial x} + \frac{\partial f}{\partial u} \cdot \frac{\partial^2 u}{\partial x^2} \end{aligned} \tag{3.5}$$

and similarly for differentiation with respect to t, we have that

$$\begin{aligned} L_x u_i{}^n = u_i{}^n &- K \frac{\partial f_i{}^n}{\partial x} - \frac{K^2}{2} \frac{\partial^2 f_i{}^n}{\partial t \, \partial x} - \frac{Kh^2}{6} \frac{\partial^3 f_i{}^n}{\partial x^3} + \frac{Kh^2}{32a} \frac{\partial}{\partial x} \left\{ \frac{\partial}{\partial x} \left(\frac{\partial f}{\partial u} \right) \cdot \frac{\partial u}{\partial x} \right\}_i^n \\ &- \frac{aK^3}{2} \frac{\partial}{\partial x} \left\{ \frac{\partial}{\partial t} \left(\frac{\partial f}{\partial u} \right) \cdot \frac{\partial u}{\partial t} \right\}_i^n + O(h^4). \end{aligned} \tag{3.6}$$

Thus

$$\begin{aligned} u_i^{n+1} - L_x u_i{}^n = &\frac{Kh^2}{6} \frac{\partial^3 f_i{}^n}{\partial x^3} - \frac{Kh^2}{32a} \frac{\partial}{\partial x} \left\{ \frac{\partial}{\partial x} \left(\frac{\partial f}{\partial u} \right) \cdot \frac{\partial u}{\partial x} \right\}_i^n \\ &+ \frac{aK^3}{2} \frac{\partial}{\partial x} \left\{ \frac{\partial}{\partial t} \left(\frac{\partial f}{\partial u} \right) \cdot \frac{\partial u}{\partial t} \right\}_i^n + \frac{K^3}{6} \frac{\partial^3 u_i{}^n}{\partial t^3} + O(h^4). \end{aligned} \tag{3.7}$$

If

$$Q(u) = -\frac{p^2}{6}\frac{\partial^3 u}{\partial t^3} + \frac{1}{6}\frac{\partial^3 f}{\partial x^3} + \frac{1}{32a}\frac{\partial}{\partial x}\left\{\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial u}\right)\cdot\frac{\partial u}{\partial x}\right\}$$
$$-\frac{ap^2}{2}\frac{\partial}{\partial x}\left\{\frac{\partial}{\partial t}\left(\frac{\partial f}{\partial u}\right)\cdot\frac{\partial u}{\partial t}\right\}, \tag{3.8}$$

then the scheme (2.6), (2.7) is third order accurate for solutions to

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = h^2 Q(u). \tag{3.9}$$

It is to be noted that truncation error is usually defined in terms of K and h as $h \to 0$ and thus the above is only meaningful if we consider some small fixed value of h. However in the implementation of any finite difference method this is exactly what is done.

Thus what we have established is that the solutions of the difference scheme (2.6, 2.7) are closer approximations to the solutions of (3.9) than (1.1).

It is noted that the truncation error of the scheme (2.6), (2.7) applied to (1.1) is in fact given by

$$-Kh^2 \| Q(u)\| + O(Kh^3). \tag{3.10}$$

In the linear case, of course, $Q(u)$ reduces to

$$\frac{1}{6}A(p^2A^2 - \mathrm{I})\frac{\partial^3 u}{\partial x^3}, \tag{3.11}$$

and so the differential equation (3.9) reduces to one in which only odd ordered derivatives appear and is hence a dispersive (nondissipative) system of differential equations. The dissipation in the Lax-Wendroff method in this case comes from the $O(Kh^3)$ terms in the truncation error of (3.10). If we write the $KO(h^3)$ terms of L_x applied at $x = ih$ as $Q'u$ then from [8, p. 332], we have

$$Q'u = \frac{1}{8}pA^2(p^2A^2 - \mathrm{I})\frac{\partial^4 u}{\partial x^4}. \tag{3.12}$$

In [13], Vreugdenhill has analysed the exact solutions of equations such as (3.9) and (3.9) with $Q'u$ added to its right hand side for the case of a scalar equation with a step function as initial data. He found that the numerical solution obtained from a second order method was closer to the solution of such equations, with $Q'u$ added than to the solution with only Qu.

It is extremely difficult to determine exactly what effect the nonlinear term in (3.9) has on the solutions of (3.9). Consequently we must make an approximation.

We have already discussed the case of taking $\partial f/\partial u = A$, constant. This however meant elimination of the parameter a.

Considering the form of (3.8), we go one stage further and consider a local linearization of A by taking

$$\begin{aligned} \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial u}\right) &= B, \quad \text{constant;} \\ \frac{\partial}{\partial t}\left(\frac{\partial f}{\partial u}\right) &= C, \quad \text{constant.} \end{aligned} \tag{3.13}$$

Equation (3.13) means that we take

$$\frac{\partial f}{\partial u} = A + B(x - x_e) + C(t - t_e) \tag{3.14}$$

locally around the point (x_e, t_e) where A is $\partial f/\partial u$ evaluated at (x_e, t_e). The approximation (3.14) is assumed to hold in a small area surrounding (x_e, t_e) just enough to include all the points involved in the finite difference scheme. Thus in applying the linearization (3.14) it is assumed that (x, t) is close to (x_e, t_e) so that

$$x - x_e = O(h), \qquad t - t_e = O(K). \tag{3.15}$$

Then $Q(u)$ from (3.8) becomes

$$Q(u) = -\frac{p^2}{6}\frac{\partial^3 u}{\partial t^3} - \frac{1}{6}\frac{\partial^3 f}{\partial x^3} + \frac{1}{32a}B\frac{\partial^2 u}{\partial x^2} - \frac{ap^2}{2}C\frac{\partial^2 u}{\partial x\,\partial t}. \tag{3.16}$$

From (3.9) we have

$$\frac{\partial^2 u}{\partial x\,\partial t} + \frac{\partial^2 f}{\partial x^2} = h^2\frac{\partial}{\partial x}Q(u);$$

$$\therefore \quad \frac{\partial^2 u}{\partial x\,\partial t} + \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial u}\cdot\frac{\partial u}{\partial x}\right) = h^2\frac{\partial}{\partial x}Q(u);$$

$$\therefore \quad \frac{\partial^2 u}{\partial x\,\partial t} + \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial u}\right)\frac{\partial u}{\partial x} + \frac{\partial f}{\partial u}\cdot\frac{\partial^2 u}{\partial x^2} = h^2\frac{\partial}{\partial x}Q(u). \tag{3.17}$$

Neglecting terms of $O(h)$ we obtain

$$\frac{\partial^2 u}{\partial x\,\partial t} = -B\frac{\partial u}{\partial x} - A\frac{\partial^2 u}{\partial x^2}. \tag{3.18}$$

Substitution of (3.18) in (3.16) gives

$$\begin{aligned} Q(u) &= -\frac{p^2}{6}\frac{\partial^3 u}{\partial t^3} - \frac{1}{6}\frac{\partial^3 f}{\partial x^3} + \frac{1}{32a}B\frac{\partial^2 u}{\partial x^2} + \frac{ap^2}{2}C\left(B\frac{\partial u}{\partial x} + A\frac{\partial^2 u}{\partial x^2}\right) \\ &= -\frac{p^2}{6}\frac{\partial^3 u}{\partial t^3} - \frac{1}{6}\frac{\partial^3 f}{\partial x^3} + \left(\frac{B}{32a} + \frac{ap^2}{2}CA\right)\frac{\partial^2 u}{\partial x^2} + \frac{ap^2}{2}CB\frac{\partial u}{\partial x}. \end{aligned} \tag{3.19}$$

Now only the last two terms of (3.19) can be adjusted by the parameter a. Also only the first of these can be dissipative. Let us write

$$D(a)u = \left\{\frac{1}{32a} B + \frac{ap^2}{2} CA\right\} \frac{\partial^2 u}{\partial x^2}. \tag{3.20}$$

Thus $D(a)u$ represents the only terms of $Q(u)$ which are dissipative to our degree of approximation and can be adjusted by use of the parameter a. Unless the matrices of (3.20) are of a specially simple form, it is still extremely difficult to tell how one should choose a to give a dissipative term for (3.20). When u is a scalar it is of course an easy problem to decide what effect a will have.

If (3.20) is to be a parabolic term in the sense of Petrowskii [see 14, p. 130] then the eigenvalues of the matrix

$$\frac{1}{32a} B + \frac{ap^2}{2} CA \tag{3.21}$$

must be positive. Further the term (3.20) will have a stronger dissipative effect the larger are the eigenvalues of (3.21).

To find the eigenvalues analytically of (3.21) is an impossible task except in very simple cases. The scalar case is analysed in Section 4. The only possibility that one has is to adjust a so that the matrix (3.21) is diagonally dominant, which will then automatically ensure positive eigenvalues. One must also remember that the truncation error of the scheme (2.6) and (2.7) applied to (1.1) depends on the parameter a. Thus one must be careful not to choose values of a which will introduce large errors and hence nonlinear instabilities. A rough guide for this can come from the fact that ω_i^{*n+1} is a first-order approximation to u_i^{n+2a} of equation (1.1).

Hence overall a suitable criterion for choosing a is difficult to find. Some analysis in this direction may be carried out for simple cases. This analysis along with computational experiments will help to indicate how a should be chosen in more difficult problems.

4. Some Numerical Experiments in the One Dimensional Case

Experiment 1.

The first problem to be considered was Eq. (1.1) with $f = \frac{1}{2}u^2$ and $u(x, 0) = x$, which gives a problem with the smooth solution

$$u(x, t) = x/(1 + t). \tag{4.1}$$

Exact boundary conditions were used on $x = 0$ and an extrapolation formula of

the type used in Gourlay and Morris [15] was used for values on $x = 1$. The solution of the problem was sought in the strip

$$(0 < x \leqslant 1) \times (t > 0).$$

The local truncation error of the scheme (2.6) and (2.7) as given by (3.7) can be explicitly written as the norm of $E_1(x, t)$ where

$$E_1(x, t) = \frac{K^3 x}{(1 + t)^4}(a - 1) + KO(h^3). \tag{4.2}$$

It is obvious from (4.2) that to minimize the truncation error one should choose a close to unity. It is impossible to consider the damping effect of terms involving $\partial^2 u/\partial x^2$ in the case of u given by (4.1), as we require to do when using our approximation (3.13), as in this case $\partial^2 u/\partial x^2$ itself is zero.

The results of solving (1.1) for the above mentioned initial and boundary values are given in Table I for a series of values of a at a series of values of p. From the table of errors it is easily seen that in fact $a = 1$ gives the greatest accuracy with accuracy decreasing the greater the distance of the parameter a from 1. However,

TABLE I

Errors at the point $x = 0.5$, $h = 0.1$, the errors being given after 100 and 300 time steps as $[{}^{(100)}_{(300)}]$. All the entries are multiplied by 10^{-8}

p \ a	0.0125	0.125	0.25	0.5	1.0	2.0	4.0	8.0
0.3	*	1912 +ve 366	1633 +ve 312	1075 +ve 206	42 −ve 7	2272 −ve 433	6726 −ve 1282	15604 −ve 2978
0.5	*	2676 +ve 421	2280 +ve 359	1489 +ve 235	92 −ve 13	3247 −ve 509	9533 −ve 1497	22002 −ve 3460
0.4	*	3161 +ve 454	2687 +ve 386	1741 +ve 250	148 −ve 20	3913 −ve 559	11384 −ve 1630	26105 −ve 3746
1.0	*	3655 +ve 487	3096 +ve 413	1980 +ve 265	242 −ve 31	4655 −ve 618	13350 −ve 1777	30248 −ve 4040

Note that the sign of the errors changes as one passes across the value $a = 1$.

* denotes non-linear instability had occurred prior to this time.

the stability analysis (linear independence of a) is confirmed by the results in that the range of a for which stability is maintained is considerable.

Experiment 1 was also repeated for the smooth problems with

$$u(x, t) = \frac{1 + 2xt - \sqrt{1 + 4xt}}{2t^2} \tag{4.1b}$$

and

$$u(x, 0) = x^2, \tag{4.2b}$$

and also

$$u(x, t) = \frac{-t + \sqrt{t^2 + 4x}}{2} \tag{4.1c}$$

and

$$u(x, 0) = \sqrt{x}. \tag{4.2c}$$

Both problems were run for $a = 0.0125, 0.125, 0.25, 0.5, 1.0, 2.0, 4.0, 8.0$ at values of $p = 0.3, 0.5, 0.7, 1.0$. Non linear instability occurred in both problems for $a = 0.0125$. A best value of a was seen to exist between 0.125 and 0.25 for problem (4.1b, 2b) and around 0.25 for problem (4.1c, 2c). The best value was also dependent to some extent on the value of p used in the computation.

EXPERIMENT 2.

In this experiment problem (1.1) was solved with $f = \frac{1}{2}u^2$ but with discontinuous initial data

$$u(x, 0) = \begin{cases} 1 & 0 \leqslant x < 0.1, \\ 0 & x \geqslant 0.1. \end{cases} \tag{4.3}$$

The boundary datum

$$u(0, t) = 1 \tag{4.4}$$

was given on $x = 0$ and an extrapolation technique as in experiment 1 was used on $x = 1$. The solution to the problem was sought on the strip $(0 < x \leqslant 1) \times (t > 0)$. This problem gives a solution in which the initial discontinuity of (4.3) propogates into the field of solution along the line $x = 0.1 + 0.5t$. In this problem, the difference scheme does not recognise the discontinuity on the line $x = 0.1 + 0.5t$ but regards the solution as changing from 1 to 0 quickly as this line is crossed, in other words, we can regard the problem as one in which the solution is continuous but with a rapidly decreasing x derivative, at any given time, as the line $x = 0.1 + 0.5t$ is crossed. Hence in the region of interest, namely around the

discontinuity $\partial u/\partial x$ is always negative and also $\partial u/\partial t$ is always positive. Hence in this case for fixed x and t, (3.21) reduces to

$$\frac{1}{32a}\frac{\partial u}{\partial x} + \frac{ap^2}{2}u\frac{\partial u}{\partial t}. \tag{4.5}$$

Hence it is obvious that the larger the value of a, the larger will be (4.5) at any fixed x and t. Thus for the greatest damping effect a must be chosen as large as possible.

The problem stated was solved for a series of values of a for a series of values of p. The results were graphed and a typical example is shown in Fig. 1. The

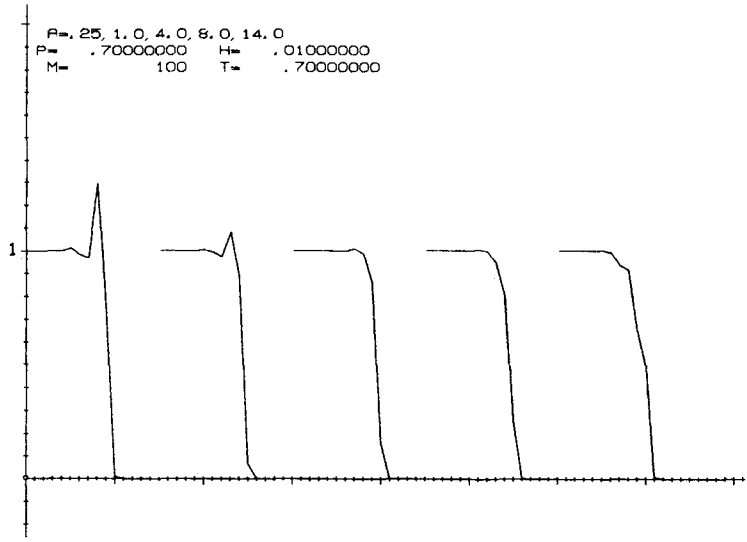

FIGURE 1

behaviour as regards damping was exactly as predicted away from values of p near the stability limit. Near the stability limit nonlinear instabilities disrupted the solution except for a very few values of a. However values of p near the stability limit with a reasonable value of a, a about 0.5 or 1, gave the best shock profiles.

EXPERIMENT 3.

The problem (1.1) was solved for the case of inviscid compressible flow of a gas as in [9]. Exactly the same equations as in [9] were used. A shock wave was set off along the x axis and its progress followed using the generalized LW scheme.

The problem was carried out for a series of values of a at a series of values of p from 0.5 times the CFL limit to the CFL limit itself. The shape of the shock front as given by the density after various numbers of time steps were graphed for each of the values of a and p.

In all cases the computed shock front occurred at exactly the point it should occur as calculated from the Rankine–Hugoniot conditions. It was also seen that the best shock profiles were given by taking a value of p close to the CFL stability limit along with a fairly large value of a. The value $a = 1$ with $p = 1.0$ or 0.95 times the CFL stability limit gave the best shock profiles. It was noted from the graphs that above $a = 1.0$ there was virtually no more dissipation introduced into the scheme as displayed in the shape of the shock profile. This was probably due to the fact that one of the eigenvalues of the matrix (3.21) in this case, was zero and the others must not have changed much with increasing a above 1.0.

EXPERIMENT 4.

In this experiment we used the problem of [9] in which two shock waves travelling in opposite directions collided. The shocks used were those of [9]. Computations were carried out for a series of values of a and p and the configuration of the shock shapes for the density after 400 time steps was graphed. Again as in experiment 3 the best shock profiles were obtained by taking p close to the stability limit and choosing a fairly large value of a, around 1.0. Choosing values of a above 1.0 however gave virtually no improvement as regards damping of oscillations behind the shock fronts.

5. THE EXTENSION OF (2.6), (2.7) TO PROBLEMS IN TWO SPACE DIMENSIONS

In this section we consider how to extend the formulation (2.6) and (2.7) to systems of conservation laws in two space dimensions viz

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} = 0, \tag{5.1}$$

where f and g are functions of u only. Extending the method in the fashion of Gourlay and Morris [8] and Thommen [16] gives schemes which have restrictive stability conditions. The scheme used by Burstein in [17] is typical and has the stability (linearized) restrictions

$$p\,|\,\lambda(A)|, p\,|\,\lambda(B)| \leqslant \frac{1}{\sqrt{8}}, \tag{5.2}$$

where $|\,\lambda(A)|$ and $|\,\lambda(B)|$ are the maximum modulus eigenvalues of the matrices A

and B respectively. Burstein points out however that this bound is rather a poor one so that the method is probably stable for a larger range of p than is indicated by (5.2).

Methods with good linearized stability properties were introduced by Strang in [1] and [19]. In [1] he introduced the scheme

$$\omega^{n+1} = \tfrac{1}{2}(L_x L_y + L_y L_x)\,\omega^n, \tag{5.3}$$

which has optimal stability characteristics as allowed by the CFL condition

$$p\,|\,\lambda(A)|, \qquad p\,|\,\lambda(B)| \leqslant 1, \tag{5.4}$$

provided L_x, L_y are optimally stable operators in one space dimension. Its second-order accuracy is guaranteed if L_x, L_y are second-order one-space-dimensional operators. The method however was extremely inefficient in terms of function evaluations when compared with the methods of Burstein [17], Thommen [16], Gourlay and Morris [8]. Gourlay and Morris [19] showed how to implement the method (5.3) in practice. A more efficient method was introduced by Strang in [19], namely

$$\omega^{n+1} = L_{x/2} L_y L_{x/2} \omega^n, \tag{5.5}$$

where $L_{x/2}$ is L_x applied with a time step $\frac{1}{2}K$. Equation (5.5) is second-order accurate and optimally stable provided L_x and L_y are such operators in one space dimension. As it stands, however, (5.5) is still less efficient than the methods of Burstein *et al.* Strang however suggested combining operators $L_{x/2}$ at the end of a step and the beginning of the next as

$$L_{x/2} L_{x/2} = L_x + O(h^3). \tag{5.6}$$

By the relation (5.6) the accuracy is maintained and also the stability remains optimal. Also it is comparable in efficiency to any other method. Thus all its advantages having been pointed out, it is natural that we should extend our one-dimensional method in (2.6) and (2.7) to two space dimensions through the method (5.5) combining operators through (5.6). Gourlay and Morris [20] showed how (5.5) should be implemented in practice but had difficulty in using given boundary data when combining operators as in (5.6). Recently McGuire and Morris [21] indicated boundary techniques in which these difficulties could be overcome and produced workable algorithms, using (5.5) with (5.6) incorporating all given boundary data, for the case when L_x and L_y were two-step Richtmyer operators in x-space and y-space respectively. The extension of the paper of Gourlay and Morris [20] to cope with our new formulation in (2.6) and (2.7) is straightforward enough as is the extension of the boundary techniques of [21].

For reasons of space, we omit details of this procedure. For further details see [21].

In the Strang scheme (5.5) the one dimensional operators L_x and L_y solve the one-dimensional problems:

$$\frac{\partial v}{\partial t} + \frac{\partial f(v)}{\partial x} = 0, \tag{5.7}$$

$$\frac{\partial w}{\partial t} + \frac{\partial g(w)}{\partial y} = 0. \tag{5.8}$$

Hence in considering the dissipation of the overall scheme we can consider the dissipation of the one-dimensional operators applied to the one-dimensional problems (5.7) and (5.8). Using approximations like (3.13) for each of the Eqs. (5.7) and (5.8) a criterion similar to (3.21) can be devised for choosing a in the two-dimensional case. Namely, if we take

$$\frac{\partial f}{\partial u} \doteq A + A_1(x - x_e) + A_2(t - t_e), \tag{5.9}$$

and

$$\frac{\partial g}{\partial u} \doteq B + B_1(x - x_e) + B_2(t - t_e), \tag{5.10}$$

around the point (x_e, y_e, t_e) and where the matrices A, A_1, A_2 are evaluated at the fixed value of $y = y_e$ as are B, B_1, B_2 at $x = x_e$, then the criterion for choosing the parameter a to maximize the dissipation is to choose a so that the eigenvalue of

$$\frac{1}{32a} A_1 + \frac{ap^2}{2} A_2 A \tag{5.11}$$

and

$$\frac{1}{32a} B_1 + \frac{ap^2}{2} B_2 B \tag{5.12}$$

are as large as possible. In the scalar case with $f = g = \frac{1}{4}u^2$

$$\frac{\partial f}{\partial u} = \frac{\partial g}{\partial u} = \frac{1}{2} u.$$

Then

$$A = \frac{1}{2} u_e, \qquad A_1 = \frac{1}{2} \frac{\partial u_e}{\partial x}, \qquad A_2 = \frac{1}{2} \frac{\partial u_e}{\partial t},$$

and

$$B = \frac{1}{2} u_e, \qquad B_1 = \frac{1}{2} \frac{\partial u_e}{\partial x}, \qquad B_2 = \frac{1}{2} \frac{\partial u_e}{\partial t},$$

so that (5.11) and (5.12) become

$$\frac{1}{32a}\frac{1}{2}\frac{\partial u_e}{\partial x} + \frac{ap^2}{2}\frac{1}{4}u_e\frac{\partial u_e}{\partial t} \tag{5.13}$$

and

$$\frac{1}{32a}\frac{1}{2}\frac{\partial u_e}{\partial y} + \frac{ap^2}{2}\frac{1}{4}u_e\frac{\partial u_e}{\partial t}. \tag{5.14}$$

A rough idea of how the dissipation in the scheme is determined by the parameter a in the scalar case can thus be obtained by considering the sizes of the expressions in (5.13) and (5.14). In cases where f and g are vectors it is almost impossible to decide what effect a has on the amount of dissipation introduced into the method. An indication for the value in that case can only be given by what happens theoretically in the scalar case and from computational experiments performed on similar systems. It is also noted at this point that from Gustafsson's results, although the situation here is not quite the same, using the first-order boundary technique of [21] may in fact not destroy the second-order global accuracy of the method. It was also suggested in [21] that due to the way in which data are introduced using the first-order technique that this is the best technique to use.

Various numerical experiments were carried out on (5.1) using the scheme (5.5) with (5.6) and using generalised LW operators for L_x and L_y. In all the problems we chose $f = \frac{1}{2}u^2$ and $g = \frac{1}{2}u^2$.

EXPERIMENT 1.

In this problem we chose initial and boundary values for (5.1) so as to obtain a smooth solution. We took the smooth problem of [20] i.e. (5.1) with the smooth solution

$$u(x, y, t) = \{(1 - \sqrt{1 + (x + y)t})/t\}^2, \tag{5.15}$$

and we assumed that exact initial and lower x and y boundary conditions were given. The problem was run for a series of values of a at a series of values of p using the backward extrapolation boundary technique of [20] and the two boundary techniques of [21]. The errors were computed with varying numbers of steps between print outs.

From the results the stability problems of the Gourlay and Morris method of introducing given boundary data began to show up in runs with a less than 0.0125. The errors for the two boundary techniques of [22] were almost identical so that, as noted already in [22], it seems that no loss of accuracy occurs in using the first-order boundary technique of [21]. It was also evident that in this smooth

problem there existed a value of a, somewhere around 0.125 which would give the best results as regards errors.

EXPERIMENT 2.

In this experiment we used (5.1) with given discontinuous initial data. The data was such that a discontinuity moved in the region of solution parallel to the y axis or to the x axis. The problem was that used in [20]. Exact boundary data was given on the lower x and y boundaries. Since the boundary techniques of [20] limited the use that could be made of the relation (5.6) when using the scheme (5.5) we ran the problems only with the boundary techniques of [21]. L_x and L_y were taken as generalized LW operators and the problems were run for various values of a with $p = 1.0$ and with different numbers of steps between print outs.

From the results it was seen that for very low values of a nonlinearities disrupt the solution. For values just below $a = 0.25$ the shock front lagged about one half to a mesh width behind the true shock front. For values of a, equal to 0.25 and above, the shock front is at most one half a mesh width ahead of the true one. Also the larger the value of a the less are the oscillations behind the discontinuity as would be expected from (5.13) and (5.14) analysed in the same way as (4.5). Also it was noted that there was very little difference between the values for different numbers of steps between print outs. Further results indicated that in fact there were smaller oscillations behind the shock front when the first-order boundary technique was used than when the second-order technique was used. Thus as in [21] the first-order technique, which is simpler to apply anyway, is to be recommended for problems of the type used here.

6. CONCLUDING REMARKS

From the analysis in Section 3 and the results of experiments 1 in both sections 4 and 5 for the one- and two-space-dimensional problems respectively, it is seen that for smooth problems a best value of a can be found in terms of small errors, simply by analysing the truncation error of the method. Thus the best technique in solving smooth problems thus appears to be the following. Choose as large a time step as the stability (linearized) condition will allow and then choose that value of a which minimizes the principal error terms in the truncation error. This argument presupposes that an expression for the truncation error can be derived, and in fact the situation can be much more difficult to analyse than we have considered here.

In the case of problems with discontinuities the analysis of Section 3 shows that the scheme (2.6) and (2.7) solves the original system of differential equations with other terms added, the most important of these being terms which are dissipative

provided a certain criterion is satisfied. This criterion, that the eigenvalues of the matrix in (3.21) are positive, is based on a certain linearization of the truncation error terms and is difficult to use except in scalar cases or for very simple systems. The larger are these eigenvalues, the more dissipative are the terms governed by this matrix. In two space dimensions virtually the same analysis applies since Strang's scheme splits the two space dimensional problem into two one-dimensional problems. Thus in this case the criterion for the system of differential equations solved using Strang's scheme, and (2.6) and (2.7), is based on the eigenvalues of two matrices of the form (3.21), namely (5.13) and (5.14). For the scalar case of a discontinuity passing through the region of computation the analysis gives an indication of how, for a fixed value of p, a should be chosen to give dissipation. This analysis is borne out in the one space dimensional case by the numerical experiments reported in Section 4. Also as was remarked there, the choice of a value of p near the stability limit was more critical in giving a good shock profile than was the choice of a. Also as remarked the shock profile did not improve much for values of p near the stability limit for values of a above 1. For lower values of p it required a larger value of a, about 8 in the case of p equal to one half its allowed value, for a reasonable shock profile. Moreover since in pratice one wishes to progress as quickly as possible in time it follows that in these scalar cases one should choose a value of p fairly close to the stability limit then a value of a around unity. In Section 5 a similar experiment was carried out in the two-space-dimensional case with a value of p half the stability limit and from the results it was seen that not much more smoothing was gained by taking a any larger than about unity. Also in Section 5 it was deduced from the results of the experiments that the first order technique of [21] should be used in preference to the more complex second order boundary procedure as this gave better stability results in the discontinuous problem and as good error results in the smooth problem. When one came to the physical systems of the experiments in Section 4 the analysis of Section 3 was almost impossible to apply, although since one of the eigenvalues of the matrix (3.21) in these cases was zero, one could indicate that the effect of a might not be as marked as in the scalar case. In the experiments it was found that a value of p chosen near the stability limit was more critical in giving good shock profiles and that, in such cases, a value of a above unity gave virtually no improvement in the shock shape. Thus again the criterion for obtaining good shock shape seems to be to choose p close to the stability limit and then to choose a about unity. It was also to be noted that the shock position given by the schemes used in both the one- and two-dimensional cases was, within at most one-half of a grid spacing, where it ought to be as deduced from theoretical considerations.

Also, in [22], (in the case of the simple problem of Section 4, experiment 2) the scheme (2.6) and (2.7) was compared with other methods for a series of values of a, and was found to give as good shock profiles as any of the other methods

considered, provided p was chosen near the stability limit and a chosen around unity.

The extension of the scheme to higher dimensions is most easily achieved, as in [20], using Strang's formulations. For example we could use

$$\omega^{n+1} = L_{x/2}L_{y/2}L_zL_{y/2}L_{x/2}\omega^n$$

for three space dimensions and combine operators as in the two-space-dimensional case. The computational formulation and incorporation of given boundary data are easily deduced from the two-space-dimensional case.

ACKNOWLEDGMENT

One of the authors' (G. R. McGuire) share of the work was carried out whilst in receipt of a Science Research Council Grant at the University of Dundee.

REFERENCES

1. W. G. STRANG, *Numer. Math.* **6** (1964), 37–46.
2. H.-O. KREISS AND O. WIDLUND, Difference Approximations for Initial Value problems for Partial Differential Equations. Uppsala University Report NR. 7, 1967.
3. H.-O. KREISS AND E. LUNDQUIST, *Maths. Comp.* **22** (1968), 1–12.
4. M. Y. T. APELKRANS, *Math. Comp.* **22** (1968), 525–539.
5. M. Y. T. APELKRANS, Some Properties of Differences Schemes for Hyperbolic Equations with Discontinuities. Uppsala University Report NR. 15, 1968.
6. B. GUSTAFSSON, On the Convergence Rate for Difference Approximations to Mixed Initial Boundary Value Problems. Uppsala University Report No. 33, 1971.
7. R. D. RICHTMYER, *NCAR Tech. Notes* 63-2, 1962.
8. R. D. RICHTMYER AND K. W. MORTON, "Difference Methods for Initial Value Problems," Wiley, New York, 1967.
9. E. L. RUBIN AND S. Z. BURSTEIN, *J. Comp. Phys.* **2** (1967), 178–196.
10. A. R. GOURLAY AND J. LL. MORRIS, *Maths. Comp.* **22** (1968), 28–39.
11. R. COURANT, K. FRIEDRICHS, AND H. LEWY, *IBM. J. of Res. and Dev.* **11** (1967), 215–234.
12. S. Z. BURSTEIN AND A. A. MIRIN, *J. Comp. Phys.* **5** (1970), 547–571.
13. C. B. VREUGDENHILL, *J. Eng. Math.* **3** (1969), 285–288.
14. O. WIDLUND, Introduction to Finite Difference Approximations to Initial Value Problems for Partial Differential Equations, in "Symposium on the Theory of Numerical Analysis held in Dundee (1970)," Springer–Verlag, New York, 1970.
15. A. R. GOURLAY AND J. LL. MORRIS, *Maths. Comp.* **22** (1968), 549–556.
16. H. U. THOMMEN, *J. Appl. Math. Phys.* **17** (1966), 369–384.
17. S. Z. BURSTEIN, *J. Comp. Phys.* **1** (1966), 198–222.
18. A. R. GOURLAY AND J. LL. MORRIS, *Maths. Comp.* **22** (1968), 715–720.
19. G. STRANG, *SIAM J. Num. Anal.* **5** (1968), 506–517.
20. A. R. GOURLAY AND J. LL. MORRIS, *J. Comp. Phys.* **5** (1970), 229–243.

21. G. R. McGuire and J. Ll. Morris, Boundary Techniques for the Multistep Formulation of the Optimized Lax Wendroff Method for Non-linear Hyperbolic Systems in Two Space Dimensions, *J. IMA.* **10** (1972), 150–165.
22. A. R. Gourlay, G. R. McGuire and J. Ll. Morris, One Dimensional Methods for the Numerical Solution of Non-linear Hyperbolic Equations, Presented by J. Ll. Morris at a conference on "Applications of Numerical Analysis at the University of Dundee," Springer–Verlag, New York, 1971.