

SIAM J. NUMER. ANAL.
Vol. 5, No. 3, September 1968
*Printed in U.S.A.*

# ON THE CONSTRUCTION AND COMPARISON OF DIFFERENCE SCHEMES*

GILBERT STRANG†

In this note we propose a new device for the construction of accurate difference schemes. The most natural applications are to nonlinear initial value problems in two space variables. For these problems, methods which are accurate only to first order are often too crude, and third order methods too complicated. The computations are thus made expensive either by the fine mesh required by a first order scheme in order to provide enough detail, or else by the delicate differencing which maintains a high order of accuracy. Second order schemes are the obvious compromise, and several have already been proposed for the equations of fluid dynamics; to compare them with our current suggestion, which is an alternating direction scheme with the half-steps ordered for maximum accuracy, we begin with the linear hyperbolic model problem

$$u_t = Au_x + Bu_y\,, \qquad u(0) = u_0\,. \tag{1}$$

The coefficients $A$ and $B$ are symmetric matrices, constant but not necessarily commuting. The wave equation, reduced to a first order system, is a familiar example.

Surprisingly, there seem to be no recognized rules for the comparison of alternative difference schemes. Clearly there are three fundamental criteria —accuracy, simplicity, and stability—and we shall evaluate each of the competing schemes in these terms. In the Appendix we attempt to combine these three requirements into a single measure of effectiveness, the *normalized error*. For families of ordinary difference schemes, such as the Runge-Kutta methods discussed by Rosser [1], as well as for constant coefficient partial difference methods in one space variable or one unknown, this measure should be directly applicable. For the more complex system (1), however, a linear ordering of difference analogues is impossible, and we are able to extend the theory only to the point where precise properties of $A$, $B$ and $u_0$ enter the analysis.

We consider difference analogues of the form

$$U(t + k\,, k) = S_k U(t\,, k), \tag{2}$$

* Received by the editors November 24, 1967.

† Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. This research was supported in part by the Office of Naval Research, the National Science Foundation under Grant GP-7477 and the United States Air Force under Contract F-44620-67-C-0007.

where the linear difference operator $S_k$ is typically a weighted sum of translations

$$(3) \qquad (S_k f)(x, y) = \sum C_{ij} f(x + ih, y + jh).$$

The matrix coefficients $C_{ij}$ depend on $A$, $B$ and the constant ratio $r = k/h$. The choice of equal mesh widths $h$ in the $x$- and $y$-directions is simply a matter of scaling.

To find the order of accuracy, one compares $S_k u$ with the Taylor expansion of $u(t + k)$. In particular, second order accuracy requires that for smooth solutions $u$,

$$(4) \qquad S_k u = u + k u_t + \frac{k^2}{2} u_{tt} + O(k^3).$$

Writing $u(t) = f$ and determining $u_t$ and $u_{tt}$ from the differential equation, we therefore require that, for all smooth functions $f = f(x, y)$,

$$(5) \quad S_k f \approx f + k(Af_x + Bf_y) + \frac{k^2}{2}(A^2 f_{xx} + (AB + BA) f_{xy} + B^2 f_{yy}),$$

where the symbol $\approx$ indicates equality up to $O(k^3)$.

First we review very briefly the case of one space variable: $B = 0$. There the celebrated Lax-Wendroff operator

$$(6) \qquad L_k^x = I + rA\Delta_0 + \frac{r^2}{2} A^2 \Delta_{+-}$$

is easily seen to be second order accurate, given

$$\Delta_0 f(x) = \tfrac{1}{2}(f(x + h) - f(x - h)) \approx h f_x(x),$$
$$\Delta_{+-} f(x) = f(x + h) - 2f(x) + f(x - h) \approx h^2 f_{xx}(x).$$

The importance of $L_k^x$ derives not only from its accuracy (in this respect it is the optimal combination of $f(x), f(x \pm h)$) but also from its *stability*. To test for stability one applies the difference operator to exponentials:

$$L_k^x e^{i\xi x} v = \left(I + irA \sin \xi h + \frac{r^2}{2} A^2 (1 - \cos \xi h)\right) e^{i\xi x} v$$
$$= G_x(\xi h) e^{i\xi x} v.$$

Then stability requires that these amplification matrices $G$ have uniformly bounded powers:

$$(7) \qquad |(G_x(\xi h))^n v| \leqq \text{const.} |v|$$

for all real $\xi h$, all $n > 0$, and all vectors $v$. Behind these definitions lies a

very substantial literature; we recommend above all the book of Richtmyer and Morton [2] and the résumé by Lax and Wendroff [3]. Here we simply recall the stability condition on the particular operator $L_k^x$. If the eigenvalues of $A$ satisfy

$$\max |\lambda_j(A)| \leqq \frac{h}{k} = \frac{1}{r}, \tag{8}$$

then $L_k^x$ dissipates energy:

$$\int |L_k^x f(x)|^2 \, dx \leqq \int |f(x)|^2 \, dx.$$

Equivalently, (7) holds with the constant equal to one; this property is called *strong stability*.

The physical basis for this restriction (8) on the mesh ratio $r$ was pointed out by Courant, Friedrichs and Lewy. The left side, $\max |\lambda_j(A)|$, represents the maximal signal speed in the differential equation $u_t = Au_x$, and the right side is the corresponding velocity in the difference equation $U(t + k) = L_k^x U(t)$. If (8) is violated, so that the flow of energy in the approximate solution cannot keep up with the true flow, then convergence must fail. And the only alternative to convergence is instability. We shall call a scheme *optimally stable* if it is stable whenever this Courant-Friedrichs-Lewy condition is met. For many schemes (including all those of accuracy $\geqq 3$) the condition for numerical stability is more severe than this physical argument would suggest; dropping the $\Delta_{+-}$ term in $L_k^x$, for example, leaves a first order scheme which is unstable for every $r > 0$.

Three of the schemes which we discuss for the two-dimensional system (1) use the same nine indices, where $i$ and $j$ equal $-1$, 0, or 1, in (3). The Courant-Friedrichs-Lewy condition

$$|\lambda_j(A)| \frac{k}{h} \leqq 1, \qquad |\mu_j(B)| \frac{k}{h} \leqq 1 \tag{9}$$

relating the eigenvalues $\lambda$ and $\mu$ to the mesh ratio, is therefore common to them all. The difficulty is, by a suitable choice of the weights $C_{ij}$, to provide the best combination of accuracy, simplicity and stability. Our model (1) illustrates this area of numerical analysis almost perfectly; if $AB = BA$ or if we settle for first order accuracy, most of the complications disappear, while the demands made by third order accuracy or by three space variables would create problems too difficult to be very illuminating.

Lax and Wendroff [3] proposed two schemes for the model problem, the simpler being

$$S_k^{(1)} = I + r(A\Delta_0^x + B\Delta_0^y) + \frac{r^2}{2}(A^2\Delta_{+-}^x + (AB + BA)\Delta_0^x\Delta_0^y + B^2\Delta_{+-}^y).$$

The superscripts indicate the direction in which to take differences.

Obviously $S_k^{(1)}$ satisfies (5). It is known to be stable if the ratio $r$ permitted by (9) is reduced by $\sqrt{8}$. Numerical experiments by Burstein [4], [5] suggest that stability holds under weaker restrictions, although (9) itself is definitely too weak [3]. (In establishing this $\sqrt{8}$ result, Lax and Wendroff were led to an important theorem on the field of values of a matrix; its refinement to the "Halmos inequality" [6] presented a beautiful problem in operator theory.)

To improve the stability, Lax and Wendroff added a dissipative term, forming

$$S_k^{(2)} = S_k^{(1)} - \frac{r^4}{8}(A^2 + B^2)\Delta_{+-}^x \Delta_{+-}^y.$$

The accuracy is unchanged, and now stability follows from

$$2r^2\lambda_j(A^2 + B^2) \leqq 1. \tag{10}$$

Again the scheme is not optimally stable; (9) is not sufficient for stability, though (10) may be more than sufficient. Burstein's preliminary verdict was that the improved stability of $S_k^{(2)}$ did not compensate for its increased complexity.

A third scheme, recommended very recently by Crowley [7], may be written as

$$S_k^{(3)} = I + r(A\Delta_0^x + B\Delta_0^y)\left(I + \frac{r}{2}A\Delta_0^x + c\left(\frac{r}{2}\right)^2 A^2\Delta_{+-}^x\right) \cdot \left(I + \frac{r}{2}B\Delta_0^y + c\left(\frac{r}{2}\right)^2 B^2\Delta_{+-}^y\right).$$

This is second order for any $c$, which may therefore be chosen to improve stability; $c = 1$ in [7], while $c = 0$ leads to instability. The stability analysis of $S_k^{(3)}$ appears to be difficult, even when $A$ and $B$ are scalar. In this case, a partial analysis led Crowley to conjecture that $r^2(A^2 + B^2) \leqq 9/5$ would ensure stability. Since $r^2A^2 > 2$, $B = 0$ is definitely unstable with $c = 1$, and the Courant-Friedrichs-Lewy condition for $S_k^{(3)}$ is weaker than $r^2(A^2 + B^2) \leqq 4$, the stability is not optimal.

Two of the expressions in $S_k^{(3)}$ are identical in complexity with $L_k^x$ and $L_k^y$, and the factor $A\Delta_0^x + B\Delta_0^y$ essentially contributes a third such unit. This means a factor 3/2 in complexity in comparison with $S_k^{(1)}$, and it is not yet clear whether there are compensating advantages. (The same is true in one dimension, where Crowley alternates Lax-Wendroff with the leapfrog scheme $\Delta_0^t U = rA\Delta_0^x U$.)

A fourth method, suggested by the present author in [8], is the symmetric product of the one-dimensional Lax-Wendroff operators:

$$S_k^{(4)} = \tfrac{1}{2}(L_k^x L_k^y + L_k^y L_k^x).$$

Optimal stability is easy to prove, since the corresponding amplification matrix is just

$$G^{(4)} = \tfrac{1}{2}(G_x(\xi h)G_y(\eta h) + G_y(\eta h)G_x(\xi h)).$$

If (9) holds, then both $L_k^x$ and $L_k^y$ are strongly stable, so $S_k^{(4)}$ is, too:

$$|G^{(4)}v| \leqq \tfrac{1}{2}(|G_xG_yv| + |G_yG_xv|) \leqq |v|.$$

By choosing the symmetric product, rather than simply alternating directions by means of $S_k = L_k^xL_k^y$, we see that second order accuracy is retained. Gourlay and Morris [9] have shown how $S_k^{(4)}$ ought to be organized in practice, for nonlinear problems as well; in the numerical experiments [4] it seems to have been implemented less efficiently. They estimate that $S_k^{(1)}$ can be applied about twice as fast as $S_k^{(4)}$, and since the admissible ratios $r$ are roughly in the opposite ratio 1:2, this leaves a rather delicate balance. Our new suggestion, however, is intended to reduce this factor two in complexity nearly to one; if this is borne out in actual computations, $S_k^{(4)}$ cannot survive.

To introduce our proposed scheme, we begin with

$$S_k^{(5)} = L_{k/2}^xL_k^yL_{k/2}^x.$$

The first question is whether this alternation of one-dimensional operators retains second order accuracy. This can be decided only by a computation:

$$S_k^{(5)}f \approx \left(I + \frac{k}{2}A\partial_x + \frac{k^2}{8}A^2\partial_x^2\right)\left(I + kB\partial_y + \frac{k^2}{2}B^2\partial_y^2\right)$$
$$\cdot\left(I + \frac{k}{2}A\partial_x + \frac{k^2}{8}A^2\partial_x^2\right)f$$
$$\approx f + k(Af_x + Bf_y) + \frac{k^2}{2}(A^2f_{xx} + (AB + BA)f_{xy} + B^2f_{yy}),$$

which is precisely the requirement (5). We remark that the asymmetry of $S_k^{(5)}$ is actually an asset in case $u$ is known to vary more rapidly in one direction than in the other. By working twice as hard, with a second order scheme, the accuracy in the $x$-direction is roughly four times better.

The individual operator

$$L_{k/2}^x = I + \frac{k}{2h}A\Delta_0^x + \frac{1}{2}\left(\frac{k}{2h}\right)^2A^2\Delta_{+-}^x$$

is strongly stable as long as every $|\lambda_j(A)| \leqq 2h/k$. Since $L_k^y$ is strongly stable if $|\mu_j(B)| \leqq h/k$, these conditions combine to ensure the strong stability of $S_k^{(5)}$. In fact, this combination coincides with the Courant-Friedrichs-Lewy condition, so the scheme is optimally stable. To see this,

imagine $S_k^{(5)}$ expanded as in (3). Because $L^x$ is applied twice, the index $i$ ranges from $-2$ to 2. Therefore the signal speed in the $x$-direction is raised to $2h/k$ in the difference equation, and consequently $|\lambda_j(A)| \leqq 2h/k$ is optimal. In the $y$-direction nothing is new.

As it stands, $S_k^{(5)}$ would consume perhaps half again as much time as $S_k^{(1)}$, at least if one adapts them both to nonlinear systems in conservation form (see below) and counts the nonlinear evaluations. A simple alteration, however, makes the comparison more nearly even. Whenever no printout is demanded, the combination of $L^x_{k/2}$ at the end of one step and again at the beginning of the next can be replaced by the single operator $L_k^x$. The order of accuracy is still two, and if this takes place most of the time, the complexity is reduced nearly to the minimum. The use of $L_k^x$ returns the stability condition to (9), and the scheme is very nearly an alternation of $L_k^x$ and $L_k^y$. Only at $t = 0$ and around printout do the half-steps enter; the programming for this option should present no difficulty.

For nonlinear systems, we first reconsider the question of accuracy. We admit the very general system

$$u_t = c(D^\alpha u, x, t) = a(D^\alpha u, x, t) + b(D^\alpha u, x, t),$$

where the splitting $c = a + b$ is arbitrary. $D^\alpha u$ runs over any derivatives $u, u_{x_1}, \cdots, u_{x_d}, u_{x_1x_1}, \cdots$ of the unknown vector $u$ with respect to the space variables $x = (x_1, \cdots, x_d)$.

We assume that for the separate problems

$$v_t = a(D^\alpha v, x, t), \qquad w_t = b(D^\alpha w, x, t),$$

the nonlinear difference operators $M_k(t)$ and $N_k(t)$ provide second order accuracy. To see what this requires of $N_k$, we differentiate the last equation and commute each $D^\alpha$ with $\partial/\partial t$:

$$\begin{aligned} w_{tt} &= \sum_\alpha B_\alpha \frac{\partial}{\partial t}(D^\alpha w) + b_t \\ &= \sum_\alpha B_\alpha D^\alpha[b(D^\alpha w, x, t)] + b_t(D^\alpha w, x, t), \end{aligned}$$

where $B_\alpha = B_\alpha(D^\alpha w, x, t)$ is the Jacobian of $b$ with respect to $D^\alpha w$. Then the requirement on $N_k(t)$ is that for any smooth vector function $f = f(x)$,

$$N_k(t)f \approx f + kb + \frac{k^2}{2}(\textstyle\sum B_\alpha D^\alpha[b(D^\alpha f, x, t)] + b_t), \tag{11}$$

where $b$, $b_t$, and $B_\alpha$ are evaluated at $(D^\alpha f, x, t)$.

Now we define the composite operator $S_k(t)$ by

$$S_k(t) = M_{k/2}\left(t + \frac{k}{2}\right) N_k(t) M_{k/2}(t)$$

and verify its accuracy. The details are customarily omitted, and therefore took us some time to supply; we hope by including them to speed up future calculations of the same sort. According to (11),

$$\begin{aligned} g = N_k(t)M_{k/2}(t)f &\approx M_{k/2}\,f + kb(D^\alpha M_{k/2}\,f, x, t) \\ &\quad + \frac{k^2}{2}\,(\sum B_\alpha\, D^\alpha[b(D^\alpha M_{k/2}\,f, x, t)] + b_t) \\ &\approx f + \frac{k}{2}\,a + \frac{k^2}{8}\,(\sum A_\alpha\, D^\alpha a + a_t) \\ &\quad + k\left(b + \sum B_\alpha \frac{k}{2}\, D^\alpha a\right) + \frac{k^2}{2}\,(\sum B_\alpha\, D^\alpha b + b_t), \end{aligned}$$

where the evaluations are now at $(D^\alpha f\,, x\,, t)$. Applying $M_{k/2}(t + k/2)$ to this vector, we get

$$\begin{aligned} S_k(t)f &\approx g + \frac{k}{2}\,a\left(D^\alpha g, x, t + \frac{k}{2}\right) + \frac{k^2}{8}\,(\sum A_\alpha\, D^\alpha a + a_t) \\ &\approx g + \frac{k}{2}\left[a + \sum A_\alpha\left(\frac{k}{2}\,D^\alpha a + kD^\alpha b\right) + \frac{k}{2}\,a_t\right] \\ &\quad + \frac{k^2}{8}\,(\sum A_\alpha\, D^\alpha a + a_t) \\ &\approx f + k(a + b) + \frac{k^2}{2}\,[\sum (A_\alpha + B_\alpha) D^\alpha (a + b) + (a + b)_t], \end{aligned}$$

again bringing the point of evaluation back to $(D^\alpha f\,, x\,, t)$. This establishes the second order accuracy of $S_k(t)$; the final result is identical with the right side of (11), if we replace $b$ by $c = a + b$.

As in the linear case, the product $S$ will be strongly stable (on linearization) whenever the factors $M$ and $N$ are. And as before, the single operator $M_k(t + k/2)$ can be substituted for the pair made up of $M_{k/2}(t + k/2)$ at the end of one step and $M_{k/2}(t + k)$ at the beginning of the next.

We shall discuss one specific application, to time-dependent inviscid flow in two dimensions, in more detail. The equations representing conservation of mass, the two momenta, and energy can be written in the divergence-free form

$$u_t = \frac{\partial}{\partial x} f(u) + \frac{\partial}{\partial y}\, g(u). \tag{12}$$

The vectors $f$ and $g$ are nonlinear functions of the four unknowns; they are given explicitly, for example, in [4].

The right side of (12) is already split, in a natural way, into a sum

$a + b$ of relatively simple expressions. Therefore we require second order (nonlinear) operators $M$ and $N$ for the separate equations

$$v_t = \frac{\partial}{\partial x} f(v), \qquad w_t = \frac{\partial}{\partial y} g(w).$$

The one-dimensional Lax-Wendroff operator was nonlinearized in [10], and later Richtmyer [11] found an equivalent form which is more convenient for computation. He forms $N_{2k} = N_{2k}(t)$ in two stages:

$$w^*(y) = \frac{1}{2}(w(y+h) - w(y-h)) + \frac{k}{2h}[g(w(y+h)) - g(w(y-h))], \tag{13}$$

$$N_{2k}\, w(y) = w(y) + \frac{k}{h}[g(w^*(y+h)) - g(w^*(y-h))].$$

This operator is applied on each line $x = \text{const}$. For the difference operator $M_k$ in the $x$-direction, one replaces $g$ by $f$ and $k$ by $k/2$; then $S_{2k} = M_k N_{2k} M_k$.

The intermediate operation yielding $w^*$ is a single step in another well-known scheme, due to Friedrichs and Lax; by itself it is accurate only to first order. The second stage raises the overall accuracy to two and yields $N_{2k} = L^x_{2k}$ when $g$ is linear.

Notice that $g$ is evaluated four times in each application of $N$. In Richtmyer's two-stage nonlinearization of $S_k^{(1)}$, $f$ is also evaluated four times at each step, while our operator requires four extra evaluations in the first step and adjacent to printout. The compensation, of course, is the removal of the factor $\sqrt{8}$ in the stability condition. Assuming $M_{2k}$ replaces $M_k^2$ away from printout, we have as our condition

$$r\left|\lambda_j\left(\frac{\partial f}{\partial u}(u)\right)\right| < 1, \qquad r\left|\mu_j\left(\frac{\partial g}{\partial u}(u)\right)\right| < 1.$$

This is the condition on the linearized system; we have shown elsewhere [12], [2] that for smooth solutions, the global error will then be $O(k^2)$ as in the linear case.

For problems with shocks, Lax and Wendroff [10] added pseudoviscous terms in order to provide additional stability in regions where the solution is undergoing rapid change. Burstein [5] found such terms essential in his two-dimensional calculations. In our scheme, as in the formulation of Gourlay and Morris, artificial viscosity is introduced one-dimensionally into the separate factors $M$ and $N$.

We hope that as progress is made on the outstanding theoretical question in one space variable (the proof of convergence when the solution is dis-

continuous) it will be possible to obtain corresponding results for our composite operator in two variables.

**Appendix: The comparison of difference schemes.** Suppose we are given a linear initial value problem

$$u_t = Lu, \qquad u(0) = u_0 . \tag{A1}$$

We think of $L$ as either a partial differential operator in $d$ space variables or, when $d = 0$ and (A1) is a system of ordinary differential equations, simply a matrix multiplication. For convenience let $L$ be independent of $t$.

The problem is to find an appropriate measure of the effectiveness of the difference analogue

$$U(t + k , k) = S_k U(t , k), \qquad U(0 , k) = u_0 . \tag{A2}$$

We assume for the present that $S_k$ is stable and that all relationships $h = h(k)$ between mesh widths are fixed. In the hyperbolic problems discussed above, this means that $r = k/h$ is fixed.

We regard one difference scheme as better than another if, with an equal number of arithmetical operations, it yields a closer approximation to $u$. The analysis will be *asymptotic*, in that we assume smooth solutions and estimate the error by computing the leading term $k^p\phi(t)$ in its asymptotic expansion

$$U(t , k) = u(t) + k^p\phi(t) + o(k^p). \tag{A3}$$

Our comparison is therefore nontrivial only for schemes with a common order of accuracy $p > 0$; a stable scheme with larger $p$ automatically wins as $k \to 0$, although not perhaps with finite $k$.

We recall how the *principal error function* $\phi$ is defined (cf. Henrici [13]). The order of accuracy enters the expansion

$$S_k = e^{kL} + k^{p+1}D + \cdots , \tag{A4}$$

which we carry out explicitly for the one-dimensional Lax-Wendroff operator $S_k = L_k^x$. Since $p = 2$, $D$ is the discrepancy between the coefficients of $k^3$ in $L_k^x$ and $e^{kL}$; $k^3Du$ is the leading term in the so-called *local truncation error*. Substituting the individual expansions

$$\Delta_0 = h\frac{\partial}{\partial x} + \frac{h^3}{6}\frac{\partial^3}{\partial x^3} + O(h^5), \qquad \Delta_{+-} = h^2\frac{\partial^2}{\partial x^2} + O(h^4)$$

into the expression (6) for $L_k^x$ leads directly to

$$D = \frac{1}{6}\left(\frac{A}{r^2} - A^3\right)\frac{\partial^3}{\partial x^3}.$$

Returning to the difference equation (A2) and using (A3) and (A4) to expand both sides in powers of $k$, the coefficient of $k^{p+1}$ yields the principal error equation

$$\text{(A5)} \qquad \phi_t = L\phi + Du(t), \qquad \phi(0) = 0.$$

If $L$ has constant coefficients and either $d \leqq 1$ (as for $L_k^x$) or else the unknown $u$ is a scalar, we may expect that $D$ commutes with $L$. In this case, (A5) has the explicit solution

$$\phi(t) = tDu(t).$$

In general the result is

$$\phi(t) = \int_0^t e^{(t-\tau)L} Du(\tau)\, d\tau = \int_0^t e^{(t-\tau)L} De^{\tau L} u_0\, d\tau.$$

Stability is required to establish that (A3) *is* asymptotically correct, i.e., that $U - u - k^p\phi = o(k^p)$. For proof in the nonlinear case, we refer to Henrici [13] when $d = 0$ and Strang [12] when $d > 0$.

This function $\phi$ measures the *accuracy* of the scheme. We have now to take into account the computing time $T_k(t)$ required to achieve this accuracy, using the step $k$ between the initial time $t = 0$ and the time $t$. For many schemes, $T$ will be simply the product of the number of steps, $t/k$, and the computing time per step, say $\tau_k$. In an ordinary difference scheme, $\tau_k$ is essentially a constant $\sigma$, which can be estimated either by faithfully counting all arithmetical operations, or, especially in the extension to nonlinear problems $u_t = F(t, u)$, only the number of evaluations of the derivative $F$ at each step.

For partial difference schemes, the time $\tau_k$ required for a single step also depends on the width $h = h(k)$ of the space mesh. Given an explicit operator in $d$ space variables, $\tau_k$ is proportional to $\sigma V/h^d$, where now $\sigma$ is the computing time per meshpoint and $Vh^{-d}$ is the number of meshpoints in the volume $V$ in $x$-space over which the computation extends. Assuming a simple relation $h = k^\alpha/r$, with $r$ constant, we have altogether that

$$T_k(t) \sim \sigma V r^d t k^{-(1+d\alpha)}.$$

The object is now to normalize the error $k^p\phi$ by means of the computing time $T_k$, in order to have a single measure of efficiency. For given $t$, suppose we fix the computing time by

$$\sigma V r^d t k^{-(1+d\alpha)} = 1,$$

say, in appropriate units. This determines $k$, and substituting into $k^p\phi(t)$, we see that the normalized error is then

$$E(t) = (\sigma V r^d t)^{p/(1+d\alpha)} \phi(t).$$

Notice that $E$ is independent of $k$. This is an essential property, because although we have spoken freely about the time step $k$, it has been left undefined. In fact, *it has no intrinsic definition*; what constitutes a time step is simply a matter of convention. In many schemes with fractional steps (Runge-Kutta and alternating direction methods are typical) the appropriate convention is by no means obvious. The same is true for the schemes discussed in this paper; in Richtmyer's two-stage definition (13) of $N$, it was notationally convenient to use $2k$ rather than $k$.

The normalized error $E$ provides a basis of comparison for schemes with $p$, $d$ and $\alpha$ in common. The dependence on $r$ and on the data $u_0$ remains; for example, $L_k^x$ has $p = 2$, $d = 1$, $\alpha = 1$, and

$$E(t) = \frac{1}{6}\sigma V r t^2 \left(\frac{A}{r^2} - A^3\right)\frac{\partial^3}{\partial x^3}u(t).$$

Notice that the dimension of $E$ is that of $\sigma u$, where $\sigma$ is a computing time and $u$ has the same dimension as the error $U - u$.

For a multistep ordinary difference method, $d = 0$ and

$$E(t) = (\sigma t)^p C \frac{\partial^{p+1} u}{\partial t^{p+1}}.$$

The constant $C$ is discussed by Henrici, and $\sigma$ is computed from the number of steps in the method, the number of applications of corrector formulas, and so forth. Our conclusion in this context is simply that the constant $\sigma^p C$ furnishes a basis for the comparison of such schemes.

The problem is less straightforward for difference analogues of the system $u_t = Au_x$, since there is a parameter $r$ to be chosen. Because the dimensions are correct, we may normalize $V = 1$, $t = 1$, and consider only the remaining coefficient $c = \sigma(rA - r^3A^3)/6r^2$. If $A$ were scalar, we could choose $rA = 1$, and the accuracy would be perfect: $U \equiv u$. In the matrix case, suppose $A$ has been diagonalized and normalized, say by $\| A \| = 1$. Then the above coefficient $c$ is a maximum at an eigenvalue for which $3r^2\lambda^2 = 1$, where it becomes $\sigma/9\sqrt{3}r^2$. (If $3r^2 < 1$, then the maximum is at $\lambda = 1$.) In this analysis, the largest stable value of $r$ is the best; combining the stability condition $r\,|\lambda| \leqq 1$ with the normalization $\| A \| = \max |\lambda| = 1$, we have that this value is $r = 1$. Finally, we take $\sigma = 4$ for Richtmyer's two-stage version, agreeing to count either the matrix-vector multiplications by $A$ in the linear case or the evaluations of the nonlinear $g$ in (13).

The resulting normalized error constant, in this case $4/9\sqrt{3}$, provides some indication of the effectiveness of competitive schemes. There remain significant features, such as the precise amount of dissipation $1 - |G(\xi h)|^2$, which cannot be taken into account by a single constant.

I wish it were possible, without gross distortion, to carry the correspond-

ing analysis even this far for the analogues $S_k^{(1)}, \cdots, S_k^{(5)}$ of the model problem $u_t = Au_x + Bu_y$. Since $p/(1 + d\alpha) = 2/3$, we do know the relative importance of simplicity and accuracy. The complications lie essentially in the operator $D$. All four derivatives $\partial^3 u/\partial x^i \partial y^{3-i}$ enter with rather involved matrix coefficients, so $\phi$ can be found explicitly only if $A$ and $B$ are assumed to commute. This destroys such an important element of the problem that we prefer the partial comparisons made earlier, based on separate estimates of simplicity and stability.

## REFERENCES


[1] J. B. Rosser, *A Runge-Kutta for all seasons*, SIAM Rev., 9 (1967), pp. 417–452.

[2] R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems*, Interscience, New York, 1967.

[3] P. Lax and B. Wendroff, *Difference schemes for hyperbolic equations with high order of accuracy*, Comm. Pure Appl. Math., 17 (1964), pp. 381–398.

[4] S. Z. Burstein, *Numerical methods in multidimensional shocked flows*, AIAA J., 2 (1964), pp. 2111–2117.

[5] ———, *Finite-difference calculations for hydrodynamic flows containing discontinuities*, J. Comp. Phys., 2 (1967), pp. 198–222.

[6] C. Berger, *On the numerical range of powers of an operator*, to appear.

[7] W. P. Crowley, *Second-order numerical advection*, J. Comp. Phys., 1 (1967), pp. 471–484.

[8] G. Strang, *Accurate partial difference methods I: Linear Cauchy problems*, Arch. Rational Mech. Anal., 12 (1963), pp. 392–402.

[9] A. R. Gourlay and J. Ll. Morris, *A multistep formulation of the optimized Lax-Wendroff method for nonlinear hyperbolic systems in two space variables*, to appear.

[10] P. Lax and B. Wendroff, *Systems of conservation laws*, Comm. Pure Appl. Math., 13 (1960), pp. 217–237.

[11] R. D. Richtmyer, *A survey of difference methods for nonsteady fluid dynamics*, Tech. Note 63-2, National Center for Atmospheric Research, Boulder, Colorado, 1962.

[12] G. Strang, *Accurate partial difference methods II. Nonlinear problems*, Numer. Math., 6 (1964), pp. 37–46.

[13] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley, New York, 1962.