Difference Methods for Nonlinear First-Order Hyperbolic Systems of Equations

By L. F. Shampine and R. J. Thompson*

Abstract. Two difference methods for approximating some first-order nonlinear hyperbolic differential equations are considered. The methods apply to problems arising in a number of physical applications. Each of the methods is explicit and can be implemented on a computer easily. It is proved that the methods are first-order convergent in the maximum norm. For one of the methods in order to obtain convergence it is necessary to monitor, and perhaps change, the size of the time step as the computation proceeds. The other method is unconditionally convergent.

1. Introduction. We shall present two finite-difference schemes for the solution of some initial-boundary value problems for systems of nonlinear hyperbolic partial differential equations. Our schemes are first-order, explicit, one-level schemes, and we shall show that they are convergent in the maximum norm. One is unconditionally stable and is very easy to study and use. The other is conditionally stable, but it applies to pure initial-value problems too.

The kinds of problems we treat arise in a wide variety of applications. In particular, some problems in chemical engineering provide excellent motivation for our study. Koenig [1] has shown that the application of invariant imbedding to certain design problems in chemical engineering leads to equations of the form

$$\frac{\partial u}{\partial t} + g(x, u)\frac{\partial u}{\partial x} = f(x, u)$$

$$u(x, 0) \equiv 0 \equiv u(0, t) .$$

For these applications the solutions are smooth and a premium is placed on simplicity and convenience rather than on high accuracy. The function g will not generally be nonnegative for all values of its arguments, but, on physical grounds, one expects $g(x, u(x, t)) \geqq 0$. Our difference schemes provide fast, simple methods for solving such a problem numerically.

Chemical engineering applications are by no means the only source of problems like the preceding. Wing [2] applies invariant imbedding to some nonlinear models of particle transport in a rod to derive similar equations. Systems of partial differential equations also arise naturally in some transport models [2], [3]. The systems have many properties in common with the single equations, but they are pure initial-value problems. For this reason we emphasize that our conditionally stable scheme copes with such problems although we shall not give the details.

Courant, Isaacson, and Rees have given in [4] a method for solving nonlinear

Received March 3, 1969, revised May 5, 1969.

AMS Subject Classifications. Primary 6567, 6570.

Key Words and Phrases. Difference approximations, numerical methods, partial differential equations, hyperbolic systems, stability, convergence.

* This work was supported by the U. S. Atomic Energy Commission.

systems of hyperbolic equations. Although the method is only conditionally stable, the analysis involves the unknown solution so that no computable stability criterion is available. As Forsythe and Wasow point out in their discussion of the method [5, p. 51], for nonlinear problems one must often be content with this unsatisfactory state of affairs. Kowalski [6] has obtained a usable stability condition by making very restrictive additional assumptions. Our modification of the method yields a computable stability condition with quite reasonable hypotheses, indeed precisely those ordinarily available in the chemical engineering context cited above. In addition our difference scheme does not depend on the problem and its solution so that it is easier to program and use.

A method more closely related to our unconditionally stable scheme was discussed by Keller and Thomée [7]. Their method is restricted to problems with two independent variables and ours is not, but they can handle more general boundary conditions. They prove the scheme is unconditionally stable though their analysis applies only if the mesh spacing is sufficiently small—how small depending on the unknown solution. Our scheme is also unconditionally stable and our analysis applies for all mesh spacings. Since we make a rather different application of the maximum principle, our analysis is somewhat simpler.

In Section 2 the general class of differential equations to be treated is introduced. The next two sections deal with the numerical schemes. In Section 3 a conditionally stable approximation which uses forward time differences is discussed. The scheme discussed in Section 4 uses backward time differences and is unconditionally stable. In Section 5 some numerical examples are discussed.

2. The Differential Equations. In what follows x will denote $(x_1, \cdots, x_p)$ and u will denote $(u_1, \cdots, u_s)$. We will consider difference methods for approximating the system:

for $i = 1, \cdots, s$

$$\frac{\partial u_i}{\partial t} + \sum_{j=1}^{p} g_{ij}(x, t, u) \frac{\partial u_i}{\partial x_j} = f_i(x, t, u) , \tag{1}$$

$$u_i(x, 0) = \alpha_i(x) \text{ on the set } 0 \leqq x_j \leqq a_j; j = 1, \cdots, p$$

for $l = 1, \cdots, p$

$$u_i(x, t) = \beta_{il}(x, t) \text{ on the set } 0 \leqq t \leqq T ;$$

$$x_l = 0; 0 \leqq x_j \leqq a_j, j = 1, \cdots, p, j \neq l .$$

Here the functions g_{ij}, f_i, α_i and β_{il} are assumed given and are defined on the appropriate sets.

Let R be the set

$$[0, a_1] \times [0, a_2] \times \cdots \times [0, a_p] \times [0, T] .$$

We are interested in approximating a solution of (1) at mesh points in R. The hypotheses regarding (1) will differ slightly for the two difference methods to be considered. However, it will always be assumed that:

(2) On R the system (1) has a solution $u = (u_1, \cdots, u_s)$ for which each component has continuous derivatives through order two. For $i = 1, \cdots, s$ and $j = 1, \cdots, p$, $g_{ij}(x, t, u(x, t)) \geqq 0$ everywhere on R. Finally, there are Lipschitz constants L_1 and L_2 such that for any two vectors $v = (v_1, \cdots, v_s)$ and $w = (w_1, \cdots, w_s)$ it is true that $\sup |g_{ij}(x, t, v) - g_{ij}(x, t, w)| \leqq L_1 \max |v_l - w_l|$ and $\sup |f_i(x, t, v) - f_i(x, t, w)| \leqq L_2 \max |v_l - w_l|$ where the sups are taken over $i = 1, \cdots, s; j = 1, \cdots, p$ and all the points of R.

Note that it is not assumed that the functions g_{ij} are nonnegative. Thus, for example, the results to be presented apply to the single equation $u_t + uu_x = 0$ for appropriate initial and boundary conditions.

Pure initial-value problems for equations like (1) have been studied by Courant, Isaacson, and Rees [4]. They deal with only two independent variables and use forward differences in the time (t) variable. By allowing the signs of the g_{ij} to influence the difference approximation in the x variable they avoid our requirement that $g_{ij}(x, t, u(x, t)) \geqq 0$. However, for initial-boundary value problems the signs of the g_{ij} are intimately connected with the type of boundary conditions that are imposed, and for boundary conditions like those in (1) one expects the g_{ij} to be nonnegative. The Courant, Isaacson and Rees scheme has a stability condition on the step size in t which involves the unknown solution $u(x, t)$. Recently Kowalski [6] has studied a similar scheme for initial-boundary value problems and for more general systems of equations. His stability condition does not involve knowledge of the solution, but it does require, in our notation, the knowledge of constants γ_{ij} such that $0 \leqq g_{ij}(x, t, u) \leqq \gamma_{ij}$. These inequalities are to hold for *all* (x, t, u) in the domain of the g_{ij}. This is very considerably more stringent than our requirements and we obtain essentially the same results. We shall show how these difficulties are avoided by using a slightly different scheme and a computable stability condition. Our conditionally stable scheme also works for pure initial-value problems which are important too, but their treatment being little different from the initial-boundary value case, we give only the latter in detail.

We shall also give an unconditionally stable scheme using backward time differences. Keller and Thomée [7] have given a similar scheme for some initial-boundary value problems involving only two independent variables. Their boundary conditions are more general than ours and by changing the difference scheme according to the sign of the g_{ij}, they avoid our sign requirement. They also adjust the difference scheme according to the magnitude of the g_{ij}. They find it necessary to require the step sizes to be sufficiently small in order to apply their analysis. Although the condition on step sizes depends on the solution, it is not as troublesome as a stability condition stated in terms of the unknown solution. Our simple analysis applies for all step sizes and is quite different in approach. They bound the error at one time level in terms of the error at the preceding level. Our approach proceeds by diagonal lines.

Both of our difference schemes will first be considered for linear systems, and the results will then be used to obtain convergence proofs for the nonlinear system (1). Thus we shall be considering the linear system:

for $i = 1, \cdots, s$

$$\frac{\partial u_i}{\partial t} + \sum_{j=1}^{p} g_{ij}(x, t) \frac{\partial u_i}{\partial x_j} = f_i(x, t) , \tag{3}$$

$$u_i(x, 0) = \alpha_i(x) \text{ on the set } 0 \leqq x_j \leqq a_j; j = 1, \cdots, p$$

for $l = 1, \cdots, p$

$$u_i(x, t) = \beta_{il}(x, t) \text{ on the set } 0 \leqq t \leqq T ;$$

$$x_l = 0; 0 \leqq x_j \leqq a_j, j = 1, \cdots, p, j \neq l .$$

Whenever such a linear system is considered it will be assumed that

(4) On R the system (3) has a solution $u = (u_1, \cdots, u_s)$ for which each component has continuous derivatives through order two, and each of the functions g_{ij}, $i = 1, \cdots, s; j = 1, \cdots, p$, is nonnegative on R.

3. Forward Time Differences. Suppose each interval $[0, a_j]$ is subdivided into intervals of length h_j. A solution of the system (1) is to be approximated on a mesh in R. The points of the mesh will have coordinates of the form $(\nu_1 h_1, \cdots, \nu_p h_p, t_n)$ where the ν_j are integers, $t_0 = 0$ and, for $n \geqq 1$, $t_n = k_0 + \cdots + k_{n-1}$. The k's are the time steps, and the way they are chosen will be discussed below. For brevity a mesh point will be denoted by (ν, n). The backward shift operator B_j is defined by $B_j\nu = (\nu_1, \cdots, \nu_{j-1}, \nu_j - 1, \nu_{j+1}, \cdots, \nu_p)$. As an approximation for (1) we will consider the following system of difference equations:

for $i = 1, \cdots, s$

$$\frac{U_i(\nu, n+1) - U_i(\nu, n)}{k_n} + \sum_{j=1}^{p} g_{ij}(\nu, n, U(\nu, n)) \frac{U_i(\nu, n) - U_i(B_j\nu, n)}{h_j} \tag{5}$$

$$= f_i(\nu, n, U(\nu, n)) .$$

Here $U = (U_1, \cdots, U_s)$ and $\bar{g}_{ij}(\nu, n, U(\nu, n))$ denotes the larger of 0 and $g_{ij}(\nu, n, U(\nu, n))$. The system (5) is, of course, to be solved subject to initial and boundary conditions determined by those specified in (1).

The device of replacing $g_{ij}(\nu, n, U(\nu, n))$ by $\bar{g}_{ij}(\nu, n, U(\nu, n))$ is important. We are only assuming the g_{ij} are nonnegative when evaluated at the true solution $u(x, t)$. The nonnegativity is needed in our proof so we simply make the coefficients satisfy this requirement. As it turns out, this global constraint can only improve the computed values.

The system (5) can be solved explicitly at the mesh points of R once the h's and k's have been specified. The remainder of this section will be devoted to considering the question of the convergence of the approximations to the solution of (1). First we shall deal with the special case when the system is linear. (5) can then be written in the form

for $i = 1, \cdots, s$

$$U_i(\nu, n+1) = \left(1 - k_n \sum_{j=1}^{p} \frac{g_{ij}(\nu, n)}{h_j}\right) U_i(\nu, n) \tag{6}$$

$$+ k_n \sum_{j=1}^{p} \frac{g_{ij}(\nu, n)}{h_j} U_i(B_j\nu, n) + k_n f_i(\nu, n) .$$

It is well known that even in the very special case when (3) is the single homogeneous equation $\partial u/\partial t + a(\partial u/\partial x) = 0$, $a > 0$, this difference approximation is not convergent unless an appropriate restriction is placed on the choice of the mesh lengths. The restrictions will be stated as conditions on the choice of the time steps k_n. Suppose then that the $h_j, j = 1, \cdots, p$, are specified, and let $\lambda_n = k_n \sum_{j=1}^{p} 1/h_j$. The λ_n (and so the k_n) are chosen subject to the following conditions:

(7) Suppose t_n is known and $0 \leqq t_n \leqq T$. (Here, again, $t_0 = 0$ and, for $n \geqq 1$, $t_n = k_0 + \cdots + k_{n-1}$.) Let $g_n = \max g_{ij}(\nu, n)$, $i = 1, \cdots, s; j = 1, \cdots, p$ and (ν, n) in R. Then if $g_n \leqq 1$, $\lambda_n (= k_n \sum_{j=1}^{p} 1/h_j)$ is chosen to satisfy $0 < \lambda_n \leqq 1$; otherwise λ_n is chosen to satisfy $0 < \lambda_n \leqq 1/g_n$.

(7) can be regarded as a stability condition for the difference equations (6). For the single equation $\partial u/\partial t + a(\partial u/\partial x) = 0$, $a > 0$, (7) requires that for the corresponding difference equation the time step divided by the x mesh length must be less than or equal to $1/a$. This is, of course, the familiar stability criterion.

LEMMA 1. *Let $u = (u_1, \cdots, u_s)$ be a solution for the linear system of equations* (3), *and suppose the conditions in* (4) *are satisfied and that the functions g_{ij} are bounded on R. For $j = 1, \cdots, p$ suppose the interval $[0, a_j]$ is subdivided into intervals of length c_j. Then there exists a constant K such that the following is true: if the intervals $[0, a_j]$ are subdivided into intervals of length h_j such that $h_j = hc_j$ for some number h, and if $U = (U_1, \cdots, U_s)$ is the solution for the difference equations* (6) *where the k_n have been chosen so that the conditions in* (7) *are satisfied, then for $i = 1, \cdots, s$ and (ν, n) in R, $|u_i(\nu, n) - U_i(\nu, n)| \leqq Kh$.*

Proof. The proof is standard, and the details will be omitted. By expanding the u_i in Taylor's series about the mesh points in R one is able to derive a system of difference equations satisfied by the $u_i - U_i$. A maximum principle holds for these equations when the conditions in (7) are satisfied. The inequality to be proved then follows immediately.

We now turn to a consideration of the difference equations (5) as an approximation for the nonlinear differential equations (1). The choice of λ_n for this case is expressed in terms of a parameter r which does not appear in (7). r is any number which satisfies $0 < r < 1$, but it remains fixed as the mesh is refined.

(8) Suppose t_n is known and $0 \leqq t_n \leqq T$. ($t_0 = 0$ and, for $n \geqq 1$, $t_n = k_0 + \cdots + k_{n-1}$.) Let $\bar{g}_n = \max \bar{g}_{ij}(\nu, n, U(\nu, n))$, $i = 1, \cdots, s; j = 1, \cdots, p$ and (ν, n) in R. Then $\lambda_n (= k_n \sum_{j=1}^{p} 1/h_j)$ is chosen as follows: if $\bar{g}_n \leqq r$, then $\lambda_n = 1$; otherwise, $\lambda_n = r/\bar{g}_n$.

THEOREM 1. *Let $u = (u_1, \cdots, u_s)$ be a solution for the system of equations* (1), *and suppose the conditions in* (2) *are satisfied. In addition, suppose there is a bound B such that for $i = 1, \cdots, s; j = 1, \cdots, p$ and (x, t) in R, $g_{ij}(x, t, u(x, t)) \leqq B$. Let r satisfy $0 < r < 1$, and for $j = 1, \cdots, p$ suppose the interval $[0, a_j]$ is subdivided into intervals of length c_j. Then there exists a constant K such that the following is true: if the intervals $[0, a_j]$ are subdivided into intervals of length h_j such that $h_j = hc_j$ for some number h sufficiently small, and if $U = (U_1, \cdots, U_s)$ is the solution for the difference equations* (5) *where the k_n have been chosen to satisfy the conditions in* (8),

then for $i = 1, \cdots, s$ *and* (ν, n) *in* R, $|u_i(\nu, n) - U_i(\nu, n)| \leqq Kh$. *Finally, there is a number* $M > 0$ *such that the* λ_n $(= k_n \sum_{j=1}^{p} 1/h_j)$ *satisfy* $M \leqq \lambda_n \leqq 1$.

Remark. Since h can be made arbitrarily small, the theorem shows that the solution for the difference equations can be made arbitrarily close to the solution for the differential equations at the mesh points of R. The bounds on the λ_n guarantee that any point in R can be made arbitrarily close to a mesh point. The bound B is used in showing the existence of M so that this last statement can be made. However, we emphasize that only the existence of B is required. B is not used in the computational procedure, and we do not need to know its magnitude.

Proof of Theorem 1. We first note that (8) implies that the λ_n satisfy $\lambda_n \leqq 1$. Now for $i = 1, \cdots, s; j = 1, \cdots, p$ and (x, t) in R, let $G_{ij}(x, t)$ denote $g_{ij}(x, t, u(x, t))$. On R each of the functions G_{ij} is defined and satisfies $0 \leqq G_{ij} \leqq B$. Similarly, let $F_i(x, t)$ denote $f_i(x, t, u(x, t))$. Then u satisfies the linear system:

for $i = 1, \cdots, s$

$$\frac{\partial u_i}{\partial t} + \sum_{j=1}^{p} G_{ij}(x, t) \frac{\partial u_i}{\partial x_j} = F_i(x, t) .$$

For each n let $G_n = \max G_{ij}(\nu, n)$, $i = 1, \cdots, s; j = 1, \cdots, p$ and (ν, n) in R. It is easy to show:

(9) Let n be fixed and suppose there is a number P such that, for $i = 1, \cdots, s$; $j = 1, \cdots, p$ and (ν, n) in R, $|g_{ij}(\nu, n, U(\nu, n)) - G_{ij}(\nu, n)| \leqq P$. Then $|\bar{g}_n - G_n| \leqq P$.

Now let $e_i(\nu, n) = u_i(\nu, n) - U_i(\nu, n)$ and, for each n, $\|e_n\| = \max |e_i(\nu, n)|$, $i = 1, \cdots, s$ and (ν, n) in R. We shall prove:

(10) There is a δ such that if $\|e_n\| \leqq \delta$ then λ_n satisfies the following: if $G_n \leqq 1$ then $0 < \lambda_n \leqq 1$; otherwise $\lambda_n \leqq 1/G_n$. In addition, there is a number $M > 0$ such that $M \leqq \lambda_n \leqq 1$.

By (2), $|g_{ij}(\nu, n, U(\nu, n)) - G_{ij}(\nu, n)| \leqq L_1\|e_n\|$. Let $\delta = r(1 - r)/L_1$. Then if $\|e_n\| \leqq \delta$, $|g_{ij}(\nu, n, U(\nu, n)) - G_{ij}(\nu, n)| \leqq r(1 - r)$. By (9), $|\bar{g}_n - G_n| \leqq r(1 - r)$. Since $G_n \leqq B$, $\bar{g}_n \leqq B + r(1 - r)$, and so M can be taken to be the smaller of 1 and $r/[B + r(1 - r)]$.

To complete the proof of (10) it remains to be shown that if $G_n \leqq 1$ then $\lambda_n \leqq 1$, or if $G_n > 1$ then $\lambda_n \leqq 1/G_n$. Since λ_n always satisfies $\lambda_n \leqq 1$, we need only consider the case $G_n > 1$—i.e., we will assume $G_n > 1$ and prove that $\lambda_n \leqq 1/G_n$. We first note that $\bar{g}_n \neq 0$ and $\lambda_n = r/\bar{g}_n$ since otherwise $\bar{g}_n \leqq r$ and so, since we know that $|\bar{g}_n - G_n| \leqq r(1 - r)$, $G_n \leqq \bar{g}_n + r(1 - r) \leqq r + r(1 - r) \leqq 1$—but this contradicts $G_n > 1$. Now, since $G_n > 1 > r$, $-r(1 - r) > -G_n(1 - r)$ and $G_n - r(1 - r) > G_n - G_n(1 - r) = rG_n$. Thus, again using $|\bar{g}_n - G_n| \leqq r(1 - r)$, it follows that $\bar{g}_n \geqq G_n - r(1 - r) > rG_n$, and so $r/\bar{g}_n\ (= \lambda_n) \leqq r/(rG_n) = 1/G_n$.

Now let $V_i(\nu, n)$ satisfy the difference equations:

for $i = 1, \cdots, s$

$$V_i(\nu, n+1) = \Big(1 - k_n \sum_{j=1}^{p} G_{ij}(\nu, n)/h_j\Big) V_i(\nu, n) + k_n \sum_{j=1}^{p} (G_{ij}(\nu, n)/h_j) V_i(B_j\nu, n) + k_n F_i(\nu, n) . \tag{11}$$

Let $E_i(\nu, n) = V_i(\nu, n) - U_i(\nu, n)$ and $\varepsilon_i(\nu, n) = u_i(\nu, n) - V_i(\nu, n)$. For each n, $\|E_n\|$ and $\|\varepsilon_n\|$ will denote max $|E_i(\nu, n)|$ and max $|\varepsilon_i(\nu, n)|$, respectively, where the max is taken over $i = 1, \cdots, s$ and the mesh points (ν, n) in R. Clearly

$$\|e_n\| \leqq \|\varepsilon_n\| + \|E_n\| . \tag{12}$$

By Lemma 1 and (10),

(13) There exist numbers K' and δ such that, for $n \geqq 1$, $\|\varepsilon_n\| \leqq K'h$ if $\|e_k\| \leqq \delta$ for $k = 0, 1, \cdots, n-1$.

Combining (5) and (11) it follows that

$$E_i(\nu, n+1) = \Big(1 - k_n \sum_{j=1}^{p} \bar{g}_{ij}(\nu, n, U(\nu, n))/h_j\Big) E_i(\nu, n) + k_n \sum_{j=1}^{p} (\bar{g}_{ij}(\nu, n, U(\nu, n))/h_j) E_i(B_j\nu, n) + k_n[F_i(\nu, n) - f_i(\nu, n, U(\nu, n))] + k_n \sum_{j=1}^{p} [\bar{g}_{ij}(\nu, n, U(\nu, n)) - G_{ij}(\nu, n)] \times [V_i(\nu, n) - V_i(B_j\nu, n)]/h_j .$$

By (2), $|F_i(\nu, n) - f_i(\nu, n, U(\nu, n))| \leqq L_2\|e_n\|$. Since $G_{ij} \geqq 0$,

$$|\bar{g}_{ij}(\nu, n, U(\nu, n)) - G_{ij}(\nu, n)| \leqq |g_{ij}(\nu, n, U(\nu, n)) - G_{ij}(\nu, n)| \leqq L_1\|e_n\| .$$

Now (8) implies that

$$\Big(1 - k_n \sum_{j=1}^{p} \bar{g}_{ij}(\nu, n, U(\nu, n))/h_j\Big) \geqq 0$$

and since $\bar{g}_{ij} \geqq 0$, it follows that

$$|E_i(\nu, n+1)| \leqq \|E_n\| + k_nL_2\|e_n\| + k_nL_1\|e_n\| \sum_{j=1}^{p} |V_i(\nu, n) - V_i(B_j\nu, n)|/h_j .$$

Now

$$|V_i(\nu, n) - V_i(B_j\nu, n)| \leqq 2\|\varepsilon_n\| + |u_i(\nu, n) - u_i(B_j\nu, n)| \leqq 2\|\varepsilon_n\| + K''h_j$$

where K'' is a bound on R for $|\partial u_k/\partial x_l|$, $k = 1, \cdots, s$; $l = 1, \cdots, p$. Thus

$$|E_i(\nu, n+1)| \leqq \|E_n\| + k_nL_2\|e_n\| + k_nL_1\|e_n\| \sum_{j=1}^{p} (2\|\varepsilon_n\| + K''h_j)/h_j$$

and so

$$\|E_{n+1}\| \leqq \|E_n\| + k_n\|e_n\|\left(L_2 + 2L_1\|\varepsilon_n\| \sum_{j=1}^{p} 1/h_j + L_1pK''\right). \tag{14}$$

Now let $c = L_2 + 2L_1K' \sum_{j=1}^{p} 1/c_j + L_1pK''$ and suppose that h is small enough that $K'h\, e^{cT} \leqq \delta$. $\|e_0\| = \|\varepsilon_0\| = \|E_0\| = 0$, so by (13) $\|\varepsilon_1\| \leqq K'h$, and from (12) and (14) it follows that $\|E_1\| = 0$ and $\|e_1\| \leqq K'h \leqq K'h\, e^{cT} \leqq \delta$. For $n \geqq 2$ it follows by induction, using (12), (13) and (14), that

$$\|\varepsilon_n\| \leqq K'h\,,$$

$$\|E_n\| \leqq K'h[(1 + ck_1) \cdots (1 + ck_{n-1}) - 1]$$

and

$$\|e_n\| \leqq K'h(1 + ck_1) \cdots (1 + ck_{n-1}) \leqq K'h \exp\,[c(k_1 + \cdots + k_{n-1})]$$
$$\leqq K'he^{cT} \leqq \delta\,.$$

By (10), $\lambda_n \geqq M$ and so, letting $K = K'e^{cT}$, the proof of the theorem is complete.

4. Backward Time Differences. For the backward difference scheme considered in this section no restrictions such as those imposed in (7) or (8) are necessary. The difference approximation is unconditionally stable in the maximum norm. Since the size of the time step need not be monitored, and perhaps changed, during the computation, a fixed time step is used. k will denote the time step. Mesh points in R will have coordinates of the form $(\nu_1h_1, \cdots, \nu_ph_p, nk)$—again denoted by (ν, n).

The system (1) is to be approximated by the difference equations:

$$\text{for } i = 1, \cdots, s$$

$$\frac{U_i(\nu, n+1) - U_i(\nu, n)}{k} + \sum_{j=1}^{p} \bar{g}_{ij}(\nu, n, U(\nu, n)) \frac{U_i(\nu, n+1) - U_i(B_j\nu, n+1)}{h_j} \tag{15}$$
$$= f_i(\nu, n, U(\nu, n))\,.$$

Once again, $\bar{g}_{ij}$ denotes the larger of g_{ij} and 0. These difference equations can be solved explicitly as easily as (5). Indeed, (15) is easier to implement on a computer since the size of the time step is fixed.

For the special case of the linear system (3), the difference equations can be written

$$U_i(\nu, n+1) = \frac{U_i(\nu, n) + \sum_{j=1}^{p} \rho_j g_{ij}(\nu, n) U_i(B_j\nu, n+1) + kf_i(\nu, n)}{1 + \sum_{j=1}^{p} \rho_j g_{ij}(\nu, n)} \tag{16}$$

where $\rho_j = k/h_j$.

LEMMA 2. *Let $u = (u_1, \cdots, u_s)$ be a solution for the linear system of equations* (3), *and suppose that the conditions in* (4) *are satisfied. In addition, suppose each of the functions g_{ij} and f_i has continuous first derivatives on R. For $j = 1, \cdots, p$ suppose the interval $[0, a_j]$ is subdivided into intervals of length c_j, and let k^* by any positive number. Then there exists a constant K such that the following is true: if the intervals $[0, a_j]$ are subdivided into intervals of length h_j such that $h_j = hc_j$ for some number h, and if*

$U = (U_1, \cdots, U_s)$ is the solution for the difference equations (16) *with $k = k^*h$, then for $i = 1, \cdots, s$ and (ν, n) in R, $|u_i(\nu, n) - U_i(\nu, n)| \leqq Kh$.*

Proof. Let $e_i(\nu, n) = u_i(\nu, n) - U_i(\nu, n)$. By expanding the functions u_i, f_i and g_{ij} in Taylor's series about the mesh points in R it can be shown, using the fact that $\rho_j g_{ij}(\nu, n) \geqq 0$, that

$$|e_i(\nu, n+1)| \leqq \frac{|e_i(\nu, n)| + \sum_{j=1}^{p} \rho_j g_{ij}(\nu, n)|e_i(B_j\nu, n+1)|}{1 + \sum_{j=1}^{p} \rho_j g_{ij}(\nu, n)} + k^2 B$$

for some constant B.

A significant difference between the analysis of this scheme and that of Keller and Thomée (and our previous scheme) now arises. Let $z_r = \max |e_i(\nu, n)|$, where the max is taken over $n + \nu_1 + \cdots + \nu_p = r$, (ν, n) in R, and $i = 1, \cdots, s$. Thus the max is over "diagonals" rather than the next time level. The preceding inequality implies that $z_{r+1} \leqq z_r + k^2 B$ for $r \geqq 0$. Since $z_0 = 0$ (indeed $z_r = 0$ for $r = 0, \cdots, p$), by induction $z_r \leqq rk^2 B$. Since $n \leqq T/k$ and $\nu_j \leqq a_j/h_j = (k^* a_j)/(kc_j)$, $r \leqq (k^*/k)(T/k^* + a_1/c_1 + \cdots + a_p/c_p)$, and so

$$z_r \leqq k^*(T/k^* + a_1/c_1 + \cdots + a_p/c_p)kB = (k^*)^2(T/k^* + a_1/c_1 + \cdots + a_p/c_p)hB .$$

Thus K can be set equal to $(k^*)^2(T/k^* + a_1/c_1 + \cdots + a_p/c_p)B$ and the proof is complete.

THEOREM 2. *Let $u = (u_1, \cdots, u_s)$ be a solution for the system of equations* (1), *and suppose the conditions in* (2) *are satisfied. In addition, suppose that for each of the functions f_i and g_{ij} which appears in* (1) *the first partial derivatives exist and are continuous. For $j = 1, \cdots, p$ suppose the interval $[0, a_j]$ is subdivided into intervals of length c_j, and let k^* be any positive number. Then there exists a constant K such that the following is true: if the intervals $[0, a_j]$ are subdivided into intervals of length h_j such that $h_j = hc_j$ for some number h, and if $U = (U_1, \cdots, U_s)$ is the solution for the difference equations* (15) *with $k = k^*h$, then for $i = 1, \cdots, s$ and (ν, n) in R, $|u_i(\nu, n) - U_i(\nu, n)| \leqq Kh$.*

The proof is very similar in approach to that of Theorem 1 but less tedious. With Theorem 1 and Lemma 2 as guides it is straightforward enough to be omitted.

5. Numerical Examples. For simplicity, in this section we will consider only single equations with two independent variables. Thus each of the examples will be an equation of the form

$$\partial u/\partial t + g(x, t, u)\partial u/\partial x = f(x, t, u) .$$

We noted earlier that our difference methods are simple and very easy to use. This is particularly easy to see when they are applied to the preceding equation. One need only consider what equations (5) and (15) become in this case.

For convenience, in what follows we shall refer to the difference scheme represented by Eq. (5) as Method F (for forward) or simply F. Similarly, Eq. (15) shall be called Method B or sometimes B (for backward).

Example 1. Koenig [1] applies invariant imbedding to a boundary value problem in chemical engineering and obtains a partial differential equation which in our notation is

$$\partial u/\partial t + 2(x - 2u - .01u^2)(1 - x)\partial u/\partial x = 2(x - 2u - .01u^2)(1 - u) .$$

This is to be solved subject to the conditions $u(x, 0) \equiv 0 \equiv u(0, t)$. Here t denotes the length of a gas-absorption tower and $u(x, t)$ denotes the concentration in the outgoing liquid phase when the concentration in the incoming gas phase is x. Koenig indicates how the equation can be solved numerically by a scheme which is essentially our Method F, although he fails to note that the method is only conditionally stable. The following table shows some representative numbers computed using our difference schemes. For both methods $h = .1$ was used, and the tabulated values are at $x = 1/2$. For F the value $r = .95$ was used; $k = h$ was used for B.

t	0	1	2	3	4	5
F	0	.2380	.2485	.2496	.2497	.2497
B	0	.2367	.2484	.2495	.2497	.2497

As the table indicates, the solution tends to a "steady state"—in fact, for t greater than about 9, the approximations computed by the two schemes agreed to nine decimal places and the numbers were no longer changing with t.

The steady-state solution has a physical interpretation, and in addition the numerical results can be compared with the true steady-state solution. There is more than one steady-state solution. One is $u(x) = x$. The others are found by solving the quadratic $x - 2u - .01u^2 = 0$ for u as a function of x. The solution being found by the difference schemes is the one obtained by taking the positive root of the quadratic. Both schemes compute it to the accuracy of the machine. The expression $x - 2u - .01u^2$ incorporates the equilibrium data (see Koenig's derivation of the equation in [1])—i.e., when the gas phase concentration x and the liquid phase concentration u are in equilibrium, then $x - 2u - .01u^2 = 0$. Only positive values of u are physically meaningful, and one expects that if the tower is lengthened the concentration u should approach the value at which it is in equilibrium with the concentration x. As we have noted, the numerical approximations display this behavior.

Examples 2 *and* 3. Gourlay and Morris [8], [9] discuss several methods for solving problems like (1) although they deal with at most three independent variables. They are interested in second-order methods and their schemes are considerably more complicated than ours. They make only a linearized stability analysis and do not prove convergence. Our next two examples are from their papers. Both have known solutions, so it is easy to make comparisons. The results show that, as expected, their schemes usually do better than our first-order methods. Nevertheless, for the step sizes they used, their methods do not have a clear cut advantage over our much simpler schemes.

The problem

$$\partial u/\partial t + u(\partial u/\partial x) = 0 , \qquad 0 \leqq x \leqq 1 , \quad t \geqq 0 ,$$
$$u(x, 0) = x, u(0, t) = 0$$

is used as an example in [8]. This problem has the solution $u(x, t) = x/(1 + t)$. The

following table shows some results of computations using our schemes to approximate this problem. The numbers tabulated are the differences between the solution and the approximation at $x = 1/2$ for the values of t shown. For both schemes $h = .1$ was used. $r = .95$ was used for F, and $h = k$ was used for B. Method B happens to be exact for this problem for any choice of h and k, so the row labelled B is roundoff.

t	0	2	4	6	8	10	12
F	0	6.2×10^{-3}	3.2×10^{-3}	2.0×10^{-3}	1.4×10^{-3}	9.9×10^{-4}	7.6×10^{-4}
B	0	1.5×10^{-11}	2.9×10^{-11}	2.7×10^{-11}	2.7×10^{-11}	1.5×10^{-11}	1.1×10^{-11}

Gourlay and Morris used several variations of their schemes on this problem for the mesh with $h = k = .1$. At the point $x = 1/2$, $t = 10$ they report errors of 1.1×10^{-3}, -2.9×10^{-2}, 1.1×10^{-3}, -2.5×10^{-4}, 5.7×10^{-4}, 9.2×10^{-3}, and 1.6×10^{-3}.

In [9] Gourlay and Morris treat inhomogeneous problems, and as an example use the problem

$$\frac{\partial u}{\partial t} + \frac{x^2u^2}{t}\frac{\partial u}{\partial x} = \left(\frac{2x^3u^2}{t} - 1\right)\cos(x^2 - t), \qquad 0 \leqq x \leqq 1, \quad t \geqq 1,$$

$$u(x, 1) = \sin(x^2 - 1), u(0, t) = -\sin t.$$

This problem has the solution $u(x, t) = \sin(x^2 - t)$. Some results of our computations are shown in the following table. This time the numbers tabulated are the differences between the solution and the approximations at $x = 0.7$. Again $h = .1$ was used for both schemes, $r = .95$ was used for F and $h = k$ for B.

t	1	3	5	7	9	11	13
F	0	7.4×10^{-2}	5.0×10^{-2}	-1.3×10^{-2}	6.3×10^{-2}	5.9×10^{-2}	-1.5×10^{-2}
B	0	4.8×10^{-2}	3.3×10^{-2}	-1.7×10^{-2}	5.1×10^{-2}	4.8×10^{-2}	-2.0×10^{-2}

The corresponding numbers obtained by Gourlay and Morris at $x = .7$, $t = 11$ using several variations of their methods with $h = k = .1$ were 1.6×10^{-2}, 5.4×10^{-3}, 2.3×10^{-3}, 1.2×10^{-3}, 2.6×10^{-4}, and 1.6×10^{-3}.

University of New Mexico
Albuquerque, New Mexico 87106

Sandia Laboratories
Albuquerque, New Mexico 87115

1. D. M. Koenig, "Invariant imbedding: new design method in unit operations," *Chem. Engrg.*, v. 74, no. 19, 1967, pp. 181–184.
2. G. M. Wing, *An Introduction to Transport Theory*, Wiley, New York, 1962. MR **27** #5580.
3. G. H. Meyer, "On a general theory of characteristics and the method of invariant imbedding," *SIAM J. Appl. Math.*, v. 16, 1968, pp. 488–509. MR **37** #5017.

R. COURANT, E. ISAACSON & M. REES, "On the solution of nonlinear hyperbolic differential equations by finite differences," *Comm. Pure Appl. Math.*, v. 5, 1952, pp. 243–255. MR **14**, 756.

5. G. E. FORSYTHE & W. R. WASOW, *Finite-Difference Methods for Partial Differential Equations*, Wiley, New York, 1960. MR **23** #B3156.

6. Z. KOWALSKI, "A difference method for certain hyperbolic systems of non-linear partial differential equations of the first order," *Ann. Polon. Math.*, v. 19, 1967, pp. 313–322. MR **36** #3517.

7. H. B. KELLER & V. THOMÉE, "Unconditionally stable difference methods for mixed problems for quasi-linear hyperbolic systems in two dimensions," *Comm. Pure Appl. Math.*, v. 15, 1962, pp. 63–73. MR **28** #1778.

8. A. R. GOURLAY & J. L. MORRIS, "Finite-difference methods for nonlinear hyperbolic systems," *Math. Comp.*, v. 22, 1968, pp. 28–39. MR **36** #6163.

9. A. R. GOURLAY & J. L. MORRIS, "Finite-difference methods for nonlinear hyperbolic systems. II," *Math. Comp.*, v. 22, 1968, pp. 549–556. MR **37** #3785.