MATHEMATICS OF COMPUTATION, VOLUME 33, NUMBER 145
JANUARY 1979, PAGES 11–35

Error Analysis of Finite Difference Schemes Applied to Hyperbolic Initial Boundary Value Problems*

By Gunilla Sköllermo

Abstract. A technique for the complete error analysis of finite difference schemes for hyperbolic initial boundary value problems is developed. The error analysis is split into subproblems so that different boundary approximations or different initial approximations easily can be compared for a given interior scheme. The steps of the theoretical analysis are demonstrated on the leapfrog scheme for a simple model equation. The technique is applied to several choices of initial approximations and boundary conditions for leapfrog with second and fourth order accuracy in space. A comparison with two implicit schemes is also made. The theoretical error estimates are shown to agree very well with computational results.

1. Introduction. We will study finite difference methods for the numerical solution of mixed initial boundary value problems for hyperbolic equations. The purpose of this paper is to develop a technique for the total error analysis of a finite difference scheme taking initial approximations as well as boundary conditions and the interior approximation into account. The influence of the different sources of error is clearly exhibited, and the analysis makes it possible to compare different choices of initial approximations or boundary conditions for a given scheme in the interior.

The comparison of different methods for the numerical solution of partial differential equations is a challenging but difficult task. The methods may be applied to some set of test examples which is thought to be representative. The outcome of such tests may depend heavily on the chosen examples and the implementation of the methods. Another approach is to analyze the properties of the different methods in detail on a simple example. Such a study will need the complement of some more complicated test-runs to be complete, but a careful analysis should give valuable guidelines also for more general problems. The behavior of the error is also more easily understood and illustrated on simple examples and good model problems are of great value in the study of numerical phenomena.

We will formulate our problem as follows. Consider the equation $u_t = cu_x$, $c > 0$ constant, in a strip $0 \leqslant x \leqslant 1$, $t \geqslant 0$ with initial values given on the x-axis and boundary values given at the right boundary. We examine each Fourier component of

Received May 3, 1977.
AMS (MOS) subject classifications (1970). Primary 65M15, 65N15.
Key words and phrases. Error analysis, mixed initial boundary value problem.
*This work was sponsored by the Swedish Institute for Applied Mathematics and by NASA–Ames Research Center, Moffett Field, California, under Interchange NCA2-OR745-702.

0025-5718/79/0000-0002/$07.50

the solution. The efficiency is measured as the number of meshpoints per wavelength which is needed to obtain some preassigned accuracy. This information should also be relevant in more complicated situations, since the model problem can be used to describe the local behavior in most cases. First, we study the pure Cauchy problem and describe the influence of the initial approximation and the interior scheme. Secondly, we consider the influence of the boundary approximations separated from the other errors. Finally, we take into account how the error from the Cauchy problem is reflected in the boundaries. The general technique is illustrated throughout the paper on a simple example with leapfrog. Several examples and comparisons are made in the last sections.

A similar analysis for the Cauchy problem has been undertaken by Kreiss and Oliger [4] and by Swartz and Wendroff [7]. Kreiss and Oliger discuss only the discretization in time. In both papers the main interest is on the interior approximation. Here the boundary conditions and the initial approximations are also included in the analysis.

Acknowledgement. Many inspiring discussions with Professor H. O. Kreiss are gratefully acknowledged. I would also like to thank Dr. A. Sundström and Professor J. Oliger for valuable criticism and suggestions. This work was completed during a visit to Stanford University and many thanks are due to Professor G. H. Golub and Professor J. Oliger for their kind hospitality.

2. Background and Notations.

We consider the model equation

$$u_t = cu_x, \qquad 0 \leqslant x \leqslant 1,\ t \geqslant 0, \tag{2.1}$$

for a positive constant c. Initial values are given as $u(x, 0) = F(x)$, and we assume that $F(x)$ can be defined for every x such that $\int_{-\infty}^{\infty} |F(x)|^2\, dx < \infty$. Boundary values are given as $u(1, t) = F(1 + ct)$ and the solution is $u(x, t) = F(x + ct)$. We want to solve this problem by a finite difference approximation, which we apply at discrete meshpoints, $x_m = mh$, $h = 1/N$ for some integer N, and at discrete time levels $t_n = nk$. The ratio $k/h = \lambda$ is kept constant and m and n are integer numbers. Let v'_{mn} denote the approximation to $u_{mn} = u(mh, nk)$. The differential equation (2.1) is approximated for $m = r, r + 1, \ldots, N - q$, $n = s, s + 1, \ldots$ by a consistent multistep method

$$Qv_{mn} = 0. \tag{2.2}$$

The difference operator $Q = \Sigma_{\sigma=-1}^{s} Q_\sigma E_n^{-\sigma}$, where $Q_\sigma = \Sigma_{j=-r}^{q} A_{j\sigma}(h) E_m^j$, $E_n^\sigma v_{mn} = v_{m,n+\sigma}$, $E_m^j v_{mn} = v_{m+j,n}$, depend smoothly on the stepsize h. For the first few steps the scheme is modified to

$$S_n v_{mn} = s_{mn}, \qquad m = r, r + 1, \ldots, N - q,\ n = 0, 1, \ldots, s. \tag{2.3}$$

Here S_n are smooth operators of the same kind as Q and s_{mn} are given initial values. Usually, $S_0 = I$ and $s_{m0} = F(mh)$ for example. We must also define special approximations near the boundaries.

$$B_m v_{mn} = b_{mn}, \qquad m = 0, 1, \ldots, r - 1, N - q + 1, \ldots, N,\ n \geqslant s, \tag{2.4}$$

and

$$S_{mn}v_{mn} = b_{mn}, \qquad m = 0, 1, \ldots, r-1, N-q+1, \ldots, N, \quad n = 0, 1, \ldots, s. \tag{2.5}$$

Here $B_m = \Sigma_{\sigma=-1}^{s} B_{\sigma m}E_n^{-\sigma}$ with $B_{\sigma m} = \Sigma_{j=-m}^{q} C_{j\sigma}(h)E_m^j$ and similarly for S_{mn}. We make the following assumptions on the schemes.

Assumption 1. The equations (2.2)–(2.5) can be solved boundedly for v_{mn+1}, i.e. there exists a constant $K_1 > 0$ such that for every G there exists a unique solution w to

$$Q_{-1}w_m = G_m, \qquad m = r, r+1, \ldots, N-q,$$

$$B_{-1m}w_m = g_m, \qquad m = 0, 1, \ldots, r-1, n-q+1, \ldots, N,$$

with

$$\|w\|_x^2 \leqslant K_1\left(\|G\|_x^2 + h\left(\sum_{m=0}^{r-1}|g_m|^2 + \sum_{m=N-q+1}^{n}|g_m|^2\right)\right).$$

The discrete norms are defined as

$$\|v\|_x^2 = \sum_{m=0}^{N}|v_m|^2 h \quad \text{and} \quad \|v\|_{x,t}^2 = \sum_{n=0}^{\infty}\sum_{m=0}^{N}|v_{mn}|^2 hk.$$

Assumption 2. The scheme defined by (2.2)–(2.5) is strongly stable in the sense of Kreiss, see Gustafsson, Kreiss and Sundström [3].

Consider the case with homogeneous initial values. It has been shown that the scheme is strongly stable if and only if the resolvent equation connected with (2.2) and boundary condition (2.4) has a unique solution, which can be bounded in terms of the boundary values [3]. More precisely, if v_{mn} is replaced by the test-solution $v_{mn} = \hat{v}_m z^n$, the corresponding equation for $\hat{v}_m$ should have a unique bounded solution for all $|z| > 1$; and there should exist a constant K_2 such that

$$\sum_{\mu=0}^{r-1}|\hat{v}_\mu|^2 + \sum_{\mu=N-q+1}^{N}|\hat{v}_\mu|^2 \leqslant K_2\left(\sum_{\mu=0}^{r-1}|\hat{g}_\mu|^2 + \sum_{\mu=N-q+1}^{N}|\hat{g}_\mu|^2\right), \qquad |z| > 1.$$

Here, $\hat{g}_\mu$ is related to $g_{\mu n}$ via $g_{\mu n} = \hat{g}_\mu z^n$.

The main interest in this paper lies on the error analysis and especially on how the initial approximations and boundary conditions shall be taken into account. The discrete error function e_{mn} is defined by $e_{mn} = u_{mn} - v_{mn}$. It satisfies a set of difference equations where the right-hand sides represent the different truncation errors.

$$\text{(a)} \quad Qe_{mn} = Qu_{mn}, \qquad m = r, r+1, \ldots, N-q, n \geqslant s,$$

$$\text{(b)} \quad S_n e_{mn} = S_n u_{mn} - s_{mn}, \qquad m = r, r+1, \ldots, N-q, n = 0, 1, \ldots, s,$$

$$\text{(c)} \quad B_m e_{mn} = B_m u_{mn} - b_{mn}, \qquad m = 0, 1, \ldots, r-1, N-q+1, \ldots, N, \quad n \geqslant s, \tag{2.6}$$

$$\text{(d)} \quad S_{mn}e_{mn} = S_{mn}u_{mn} - b_{mn}, \qquad m = 0, 1, \ldots, r-1, N-q+1, \ldots, N, \quad n = 0, 1, \ldots, s.$$

Gustafsson [2] has shown that if the right-hand side of (2.6a) is $O(h^{\nu+1})$ and the other three right-hand sides are $O(h^{\nu})$, then the solution e_{mn} can be estimated by $\|e\|_{x,t} = O(h^{\nu})$. This means that the boundary approximation and initial approximation may be one order of accuracy lower than the interior approximation without decreasing the overall accuracy. Here we are interested in more precise error estimates so that different schemes and different choices of initial or boundary approximations may be compared.

To facilitate the analysis we make a partition of the error corresponding to the different sources of error.

(i) Consider the pure Cauchy problem with a suitable extension of the initial values. Let the corresponding error function, e^{I}, satisfy (2.6a) and (2.6b) for every m. Using Fourier analysis in space, we can easily obtain estimates of e^{I}. This error shows the influence of the starting procedure and the interior truncation error.

(ii) Consider (2.6a) with right-hand side zero and the boundary conditions (2.6c). After a suitable extension of the region of definition we can use Fourier analysis with respect to time and obtain an estimate of the corresponding error function, e^{II}, which describes the influence of the boundary approximation.

(iii) In general, e^{I} fails to satisfy the homogeneous boundary approximation and e^{II} fails to satisfy the homogeneous initial approximation. To account for what happens as e^{I} is reflected in the boundaries *etc* we will introduce a third error function e^{III}, which is designed so that $e^{\mathrm{I}} + e^{\mathrm{II}} + e^{\mathrm{III}}$ satisfy all the equations of (2.6).

In this way the different sources of error can be discussed more or less separately. This technique makes it possible to compare different choices of initial approximations or different choices of boundary approximations for a given interior scheme.

3. The Cauchy Problem and the Initial Approximation. Let us consider the pure initial value problem

$$u_t = cu_x, \qquad -\infty \leqslant x \leqslant \infty,\ t \geqslant 0,$$

where c is a positive constant. The initial values are $u(x, 0) = F(x)$ and the solution for $0 \leqslant x \leqslant 1$, $t \geqslant 0$ is exactly the same as for the problem in a strip. Let us use the finite difference scheme (2.2) with the initial approximations (2.3) to compute an approximation to $u_{mn} = u(mh, nk)$ for all m and $n > 0$. We denote the discrete error by e^{I}_{mn}. The error must satisfy the following equations

$$\begin{aligned} Qe^{\mathrm{I}}_{mn} &= Qu_{mn} \quad \text{all } m,\ n \geqslant s, \\ S_n e^{\mathrm{I}}_{mn} &= S_n u_{mn} - s_{mn} \quad \text{all } m,\ n = 0, 1, \ldots, s. \end{aligned} \tag{3.1}$$

The functions s_{mn} are extended smoothly for $m < 0$, $m > N$ such that the functions belong to l_2 $(-\infty \leqslant m \leqslant \infty)$. We can interpret (3.1) to be valid for any $x = mh$, so that $e^{\mathrm{I}}_n(x)$ is defined by the equations above for all x. The initial data are chosen to be square integrable and the scheme is assumed to be stable for the Cauchy problem. Therefore, we can define the Fourier transform of $e^{\mathrm{I}}_n(x)$ with respect to x,

$$\hat{e}^{\mathrm{I}}_n(\omega) = \int_{-\infty}^{\infty} e^{\mathrm{I}}_n(x)\exp(-2\pi i\omega x)\,dx \quad \text{and} \quad e^{\mathrm{I}}_n(x) = \int_{-\infty}^{\infty} \hat{e}^{\mathrm{I}}_n(\omega)\exp(2\pi i\omega x)\,d\omega.$$

Multiplying (3.1) by $\exp(-2\pi i\omega x)$ and integrating, we obtain

$$\begin{aligned} &\hat{Q}\hat{e}_n^{\mathrm{I}}(\omega) = \hat{Q}\hat{u}_n(\omega), \qquad n \geqslant s, \\ &\hat{S}_n\hat{e}_n^{\mathrm{I}}(\omega) = \hat{S}_n\hat{u}_n(\omega) - \hat{s}_n, \qquad n = 0, 1, \ldots, s. \end{aligned} \tag{3.2}$$

The notation $\hat{Q}$ is the Fourier transform of the operator Q, i.e. $\hat{Q} = \exp(-2\pi i\omega mh)Qr$, $r_m = \exp(2\pi i\omega mh)$. Since $u(x, t) = F(x + ct)$, the Fourier transform $\hat{u}_n(\omega)$ is related to $\hat{u}_0(\omega)$ through $\hat{u}_n(\omega) = \exp(2\pi i\omega nck)\hat{u}_0(\omega)$.

We want to determine how small the stepsize h must be in order to ensure that the relative error $\hat{e}_n^{\mathrm{I}}(\omega)/\hat{u}_n(\omega)$ is smaller than some predetermined tolerance. More precisely, we will determine the number of points per wavelength, $M = 1/\omega h$, that are needed to obtain a certain accuracy.

So let us solve (3.2). The homogeneous equation $\hat{Q}r_n = 0$ has $s + 1$ characteristic roots $z_0, \ldots, z_s$. The scheme is stable and, therefore, $|z_l| \leqslant 1$, $l = 0, 1, \ldots, s$; and there are no multiple roots on the unit circle. Furthermore, the scheme is consistent; and therefore, one root, say $z_0 = \exp[2\pi i\omega ck + \omega ckO(h\omega)^{\nu}]$, for some $\nu \geqslant 1$. The interior scheme is of order ν.

The other roots $z_1, \ldots, z_s$ give rise to spurious solutions which cannot be interpreted as discretizations of any continuous functions. The initial conditions determine how much influence these extraneous roots will have on the computed solution.

One particular solution to (3.2) is given by $\hat{u}_n(\omega)$. Let us assume that there are no multiple roots z_l, $l = 0, \ldots, s$, to avoid cumbersome notations. Then the general solution to (3.2) can be written

$$\hat{e}_n^{\mathrm{I}}(\omega) = \sum_{l=0}^{s} C_l z_l^n + \hat{u}_n(\omega).$$

The coefficients C_l are determined by the initial conditions

$$\sum_{l=0}^{s} C_l P_n(z_l) = -\hat{s}_n(\omega), \qquad n = 0, 1, \ldots, s,$$

where $P_n(z) = \hat{S}_n\{z^n\}$ are polynomials in z. The typical solution to this system of equations is

$$C_0 = (-1 + O((h\omega)^{\alpha} + (h\omega)^{\nu}))\hat{u}_0(\omega),$$

$$C_l = O((h\omega)^{\alpha} + (h\omega)^{\nu}), \qquad l = 1, \ldots, s,$$

where ν is the global order of the interior approximation and α is the local order of the initial approximation. Thus, we get with $\exp(2\pi i\omega ck) = z$,

$$\hat{e}_n(\omega) = \hat{u}_0(\omega)z^n(1 - z_0^n/z^n) + O((h\omega)^{\alpha}) + O((h\omega)^{\nu}).$$

For small values of ωh the upper bound for the relative error $\hat{e}_R = \hat{e}_n^{\mathrm{I}}(\omega)/\hat{u}_n(\omega)$ for $t = nk$ is

$$\hat{e}_R = \omega ct \cdot (\omega h)^{\nu}\mathrm{const}_1 + (\omega h)^{\nu}\mathrm{const}_2 + (\omega h)^{\alpha}\mathrm{const}_3.$$

We can insert $\hat{e}_n^{\mathrm{I}}(\omega)$ into the Fourier transform for $e_n^{\mathrm{I}}(x)$. The part $\hat{u}_0(\omega)z^n(1 - z_0^n/z^n)$

$+ O((h\omega)^{\alpha})z_0^n$ corresponds to a smooth function

$$tF^{(\nu+1)}(x + ct)O(h^{\nu}) + O(h^{\alpha}) \cdot F^{(\alpha)}(x + ct).$$

The spurious roots give nonsmooth solutions. We will use $\hat{e}_R^{\mathrm{I}}$ to compare different initial approximations. Let p be the number of periods we want to compute in time, i.e. $\omega ct < p$. To keep $\hat{e}_R^{\mathrm{I}} < \epsilon$ we must choose $M = 1/\omega h$ such that

$$p \cdot \mathrm{const}_1/M^{\nu} + \mathrm{const}_2/M^{\nu} + \mathrm{const}_3/M^{\alpha} < \epsilon.$$

The choice of the stepsize h is also determined by how many of the leading frequencies we wish to represent accurately. We must, therefore, also consider how fast $\hat{u}_0(\omega)$ decay with increasing ω. If $u(x, 0)$ has $q - 1$ continuous derivatives in L_2 ($-\infty \leqslant x \leqslant \infty$), then $\hat{u}_0(\omega)$ decay as $1/(1 + \omega^q)$. Thus, we need only consider the first few frequencies if the initial data are smooth.

Example 1. Let us analyze leapfrog

$$v_{mn+1} - v_{mn-1} - \lambda c(v_{m+1n} - v_{m-1n}) = 0$$

with initial approximations

$$v_{m0} = u_{m0},$$

$$v_{m1} - v_{m0} - \tfrac{1}{2}\lambda c(v_{m+1n} - v_{m-1n}) = 0.$$

Fourier transformation with respect to x gives an inhomogeneous difference equation for $\hat{e}_n^{\mathrm{I}}(\omega)$. The corresponding characteristic equation is

$$z^2 - 2i\lambda c \sin(2\pi\omega h)z - 1 = 0.$$

The characteristic roots are

$$z_0 = \exp\left[2\pi i\omega ck - \frac{4}{3}\, ick\pi^3\omega^3h^2(1 - \lambda^2c^2) + O((\omega h)^4)\right], \qquad z_1 = -1/z_0.$$

The general solution is

$$\hat{e}_n^{\mathrm{I}}(\omega) = C_0z_0^n + C_1z_1^n + \hat{u}_0(\omega) \cdot \exp(2\pi i\omega cnk).$$

The initial conditions give

$$C_0 + C_1 = -\hat{u}_0(\omega), \qquad C_0P_1(z_0) + C_1P_1(z_1) = 0,$$

where $P_1(z) = z - 1 - \lambda ci \sin(2\pi\omega h)$. Thus,

$$C_0 = -\hat{u}_0(\omega)/(1 - P_1(z_0)/P_1(z_1)), \qquad C_1 = -C_0P_1(z_0)/P_1(z_1).$$

For ωh small $P_1(z_0) = -2\pi i\omega^2\lambda^2c^2h^2 + O((\omega h)^3)$ and $P_1(z_1) = -2 + O(\omega h)$. Thus,

$$\hat{e}_n^{\mathrm{I}}(\omega) = \hat{u}_0(\omega) \cdot z^n(1 - z_0^n/z^n) + \hat{u}_0(\omega)(z_1^n - z_0^n)P_1(z_0)/P_1(z_1) + O((\omega h)^3),$$

where $z = \exp(2\pi i\omega ck)$. For ωh small an upper bound on the relative error is given by

$$\hat{e}_R^{\mathrm{I}} = \frac{4}{3}\, \omega ct \cdot \omega^2h^2 \cdot \pi^3(1 - \lambda^2c^2) + 2\pi^2\omega^2h^2\lambda^2c^2$$

with $t = nk$. Here we have used the triangle inequality and disregarded the fact that $|z_1^n - z_0^n| = |2\cos(2\pi\omega ct)|$ for n even and $|z_1^n - z_0^n| = |2\sin(2\pi\omega ct)|$ for n odd. This oscillatory behavior is clearly seen in Figure 1, where the theoretical estimate of the error and the exact error are shown for $u_t = u_x$, $u(x, t) = \sin(2\pi(x + t))$ with periodic boundary conditions. In Section 5 several other initial approximations are compared. Inserting $\hat{e}_n^{\mathrm{I}}(\omega)$ in the Fourier transform, we obtain

$$e_n^{\mathrm{I}}(x) = \frac{1}{6} cth^2(1 - \lambda^2 c^2)F^{(3)}(x + ct)$$

$$+ \frac{1}{2}\lambda^2 c^2 h^2 [(-1)^n F^{(2)}(x - ct) - F^{(2)}(x + ct)] + O(h^3).$$

For $\omega ct > 1.5\ \lambda^2 c^2/(1 - \lambda^2 c^2)$ the error from the interior approximation dominates the error from the initial approximation.

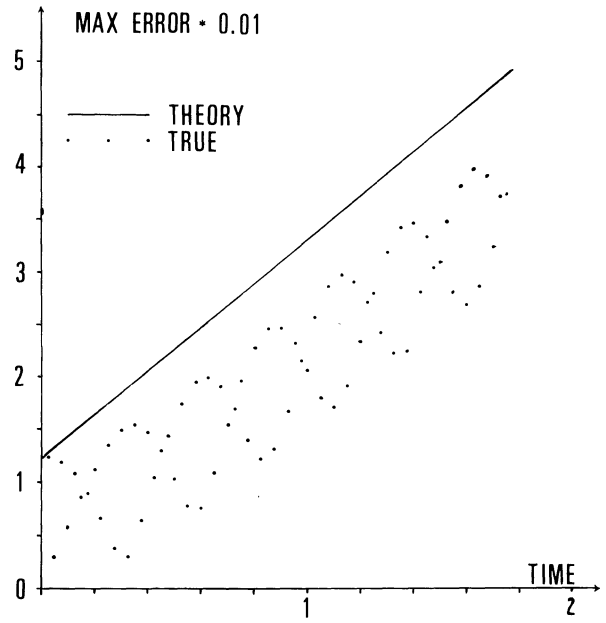

FIGURE 1

The theoretical error estimate and the true error for the model problem $u_t = u_x$, $u(x, 0) = \sin(2\pi x)$, $u(0, t) = u(1, t)$ are shown for two periods of time. The interior scheme is second order leapfrog and the initial approximation is centered Euler. The value of λc is 0.75 and we have used 30 points per wavelength in space.

4. The Influence of the Boundary Approximations. In the previous section we discussed how the interior truncation error and the errors in the initial approximations propagate in space and time. We will now study how the truncation error in the boundary approximation is propaged into the interior.

4.1. *The Homogenous Problem in an Infinite Strip with Inhomogeneous Boundary Data.* Let us for a moment disregard e_{I} and the presence of the initial level. Con-

sider the problem defined by

$$
\begin{aligned}
&Qe^{\mathrm{II}}_{mn} = 0,\\
&B_m e^{\mathrm{II}}_{mn} = B_m u_{mn} - b_{mn}, \quad m = 0, 1, \ldots, r-1, N-q+1, \ldots, N,\ n \geqslant 0.
\end{aligned} \tag{4.1}
$$

Let us extend the region of definition for u_{mn}, b_{mn} and e^{II}_{mn} to include negative values of n. If $u \in L_2$ $(0 \leqslant t < \infty)$, this extension can be done so that the functions belong to l_2 $(-\infty \leqslant n \leqslant \infty)$. If we wish to allow solutions $u(x, t)$ that grow with time, such that $ue^{-\alpha t} \in L_2$ $(0 \leqslant t < \infty)$ for some positive constant α, we introduce new variables $w = e^{-\alpha t}u$ and similarly for e before the extension is made. All the conclusions that we will draw in the following discussion are valid also in that case.

We interpret (4.1) to be valid for every time t so that $e^{\mathrm{II}}_m(t)$ is defined for all t by the relations above. We apply the Fourier transform with respect to time and obtain

$$
\tilde{e}^{\mathrm{II}}_m(\gamma) = \int_{-\infty}^{\infty} e^{\mathrm{II}}_m(t) \cdot \exp(-2\pi i\gamma t)\, dt,
$$

$$
e^{\mathrm{II}}_m(t) = \int_{-\infty}^{\infty} \tilde{e}^{\mathrm{II}}_m(\gamma) \cdot \exp(2\pi i\gamma t)\, d\gamma.
$$

Multiplying (4.1) by $\exp(-2\pi i\gamma t)$ and integrating, we obtain

$$
\begin{aligned}
&\tilde{Q}\tilde{e}^{\mathrm{II}}_m(\gamma) = 0, \quad r \leqslant m \leqslant N-q,\\
&\tilde{B}_m \tilde{e}^{\mathrm{II}}_m(\gamma) = \tilde{B}_m \tilde{u}_m(\gamma) - \tilde{b}_m, \quad m = 0, 1, \ldots, r-1, N-q+1, \ldots, N.
\end{aligned} \tag{4.2}
$$

Here $\tilde{Q}$ and $\tilde{B}_m$ are the Fourier transforms of the operators Q and B_m with respect to time, $\tilde{Q} = \exp(-2\pi i\gamma nk)Qr_n$, $r_n = \exp(2\pi i\gamma nk)$. Furthermore, $\tilde{u}_m(\gamma) = \exp(2\pi i\gamma/cmh) \cdot \tilde{u}_0(\gamma)$. The difference equation (4.2) is the resolvent equation with $z = \exp(2\pi i\gamma k)$. The corresponding characteristic equation has $r+q$ characteristic roots $\kappa_j(\gamma)$, $j = 1, \ldots, r+q$. If none of these are of multiplicity greater than one, the general solution is

$$
\tilde{e}^{\mathrm{II}}_m(\gamma) = \sum_{j=1}^{r+q} A_j \kappa_j^m(\gamma).
$$

The boundary conditions $\tilde{B}_m \tilde{e}^{\mathrm{II}}_m = \tilde{B}_m \tilde{u}_m - \tilde{b}_m$ give a system of equations for the coefficients A_j, $j = 1, \ldots, r+q$, $\sum_{j=1}^{r+q} A_j R_m(\kappa_j) = \tilde{u}_0 \cdot R_m(\kappa) - \tilde{b}_m$, $m = 0, \ldots, r-1, N-q+1, \ldots, N$, where $R_m(x)$ are polynomials in x and x^{-1} that correspond to the difference operators B_m and $\kappa = \exp(2\pi i\gamma h/c)$. This system has a unique solution.

In general, we get $\tilde{e}^{\mathrm{II}}_m(\gamma) = \tilde{u}_0(\gamma) \cdot O((\gamma h)^{\beta})$ for some exponent β, which is the local truncation error for the boundary approximations. We use the relative error $\tilde{e}^{\mathrm{II}}_R(\gamma) = \tilde{e}^{\mathrm{II}}_m(\gamma)/\tilde{u}^{\mathrm{II}}_m(\gamma)$ as a convenient measure of the influence of the boundary approximations when comparing different schemes.

Example 2. Let us demonstrate the technique on leapfrog with boundary conditions

$$
v_{0n+1} - v_{0n} - \lambda\alpha(v_{1n} - v_{0n}) = 0, \qquad v_{Nn} = u_{Nn}.
$$

The resolvent equation for leapfrog is

$$(z - z^{-1})\tilde{e}^{\mathrm{II}}_m - \lambda c(\tilde{e}^{\mathrm{II}}_{m+1} - \tilde{e}^{\mathrm{II}}_{m-1}) = 0.$$

The characteristic equation

$$(z - z^{-1})\kappa - \lambda c(\kappa^2 - 1) = 0$$

has the roots $\kappa_1 = \exp(2\pi i\gamma/ch + O((\gamma h)^3))$, $\kappa_2 = -1/\kappa_1$ for $z = \exp(2\pi i\gamma k)$. We obtain $\tilde{e}^{\mathrm{II}}_m = A_1\kappa_1^m + A_2\kappa_2^m$. The polynomials corresponding to the boundary operators are

$$R_0(\kappa) = z - 1 - \lambda c(\kappa - 1), \qquad R_N(\kappa) = \kappa^N.$$

Thus,

$$A_1R_0(\kappa_1) + A_2R_0(\kappa_2) = \tilde{u}_0 \cdot R_0(\kappa), \qquad A_1\kappa_1^N + A_2\kappa_2^N = 0,$$

where $\kappa = \exp(2\pi i\gamma h/c)$ and $h = 1/N$.

$$\tilde{e}^{\mathrm{II}}_m = \tilde{u}_0 \cdot \left(\kappa_1^m\left(\frac{\kappa_2}{\kappa_1}\right)^N - \kappa_2^m\right) \cdot R_0(\kappa)/R_0(\kappa_2) + \text{higher order terms.}$$

With $z = \exp(2\pi i\gamma k)$ we get $R_0(\kappa) = 2\lambda c(1 - \lambda c)\pi^2\gamma^2h^2/c^2 + O(\gamma^3h^3)$ and $R_0(\kappa_2) = 2\lambda c + O(\gamma h)$. Thus,

$$\tilde{e}^{\mathrm{II}}_m = \tilde{u}_0(\omega) \cdot (1 - \lambda c)\pi^2\gamma^2h^2/c^2\,[(-1)^N\exp(2\pi i\gamma/c(-2 + mh)) - (-1)^m\exp(-2\pi i\gamma mh/c) + O(\gamma^3h^3).$$

The magnitude of the relative error is estimated by

$$\tilde{e}^{\mathrm{II}}_R = 2(1 - \lambda c)\pi^2\gamma^2h^2/c^2 = 2(1 - \lambda c)\pi^2/M^2,$$

where $M = c/\gamma h = 1/\omega h$ is the number of points per wavelength in space. Inserting $\tilde{e}^{\mathrm{II}}_m(\omega)$ into the Fourier transform, we get

$$e^{\mathrm{II}}(mh, t) = \tfrac{1}{4}(1 - \lambda c)h^2((-1)^NF^{(2)}(mh + ct - 2) - (-1)^mF^{(2)}(ct - mh)) + O(h^3).$$

4.2. *The Reflection of Errors in the Boundaries.* We must now consider how the result from the Cauchy problem and from the pure boundary value problem must be patched together to describe the error propagation for the mixed initial boundary value problem. The sum of e^{I} and e^{II} satisfies (2.6a) and almost satisfies (2.6b) and (2.6c). We need to estimate the remaining error $e^{\mathrm{III}} = e - e^{\mathrm{I}} - e^{\mathrm{II}}$, which is defined by the equations

$$Qe^{\mathrm{III}}_{mn} = 0, \qquad r \leqslant m \leqslant N - q,\ n \geqslant s,$$

$$B_me^{\mathrm{III}}_{mn} = -B_me^{\mathrm{I}}_{mn}, \qquad m = 0, 1, \ldots, r - 1, N - q + 1, \ldots, N,\ n \geqslant s,$$

$$S_ne^{\mathrm{III}}_{mn} = -S_ne^{\mathrm{II}}_{mn}, \qquad r \leqslant m \leqslant N - q,\ n = 0, 1, \ldots, s,$$

$$S_{mn}e^{\mathrm{III}}_{mn} = S_{mn}u_{mn} - b_{mn} - S_{mn}(e^{\mathrm{I}}_{mn} + e^{\mathrm{II}}_{mn}),$$

$$m = 0, 1, \ldots, r - 1, N - q + 1, \ldots, N,\ n = 0, 1, \ldots, s.$$

There is one source of error from the reflection of e^{I} in the boundaries, another from the reflection of e^{II} in the initial level; and finally, we have to take into account the special boundary approximations that may be needed at the first few levels.

We will discuss the error e^{III} in terms of our example with leapfrog and then generalize from this example. We use S_n as in Example 1, B_m as in Example 2, and the initial boundary conditions are $S_{00} = S_{N0} = I$, $b_{00} = u(0, 0)$, $b_{N0} = u(1, 0)$ while $S_{01} = B_0$, $b_{01} = 0$ and $S_{N1} = I$, $b_{N1} = U(1, k)$. Recall that

$$e^{\mathrm{I}}(s, t) = h^2 t F_1(x + ct) + h^2(F_2(x + ct) - (-1)^{t/k}F_2(x - ct)) + O(h^3),$$

$$e^{\mathrm{II}}(x, t) = h^2((-1)^N F_3(x + ct - 2) - (-1)^{x/h}F_3(ct - x)) + O(h^3),$$

for some functions F_1, F_2 and F_3.

Let us consider the function $tF_1(x + ct)$. We introduce a discrete function d_{mn} which shall satisfy

$$Qd_{mn} = 0, \qquad 1 \leqslant m \leqslant N - 1,\ n \geqslant 1,$$

$$B_m d_{mn} = -B[tF_1(x + ct)], \qquad m = 0, N,\ n \geqslant 1,$$

$$S_n d_{mn} = 0, \qquad 1 \leqslant m \leqslant N - 1,\ n = 0, 1,$$

$$S_{mn} d_{mn} = 0, \qquad m = 0, N,\ n = 0, 1.$$

In our example

$$B_0[tF_1(x + ct)] = (t + k)F_1(c(t + k)) - tF_1(ct) + \frac{kc}{h}(F_1(h + ct) - F_1(ct)) = O(k)$$

and $B_N[tF_1(x + ct)] = tF_1(x + ct)$. Thus, the difference equation above is a stable and consistent approximation to $d_t = cd_x$, $d(x, 0) = 0$, $d(1, t) = tF_1(ct)$. Therefore, $d_{mn} = d(mh, nk) + O(h)$ and

$$nkF_1(mh + cnk) + d_{mn} = \begin{cases} nkF_1(mh + cnk) + O(h), & mh + cnk \leqslant 1, \\ \dfrac{1 - mh}{c} \cdot F_1(mh + cnk) + O(h), & mh + cnk > 1. \end{cases}$$

This means that the error does not continue to grow with time indefinitely but rather depends on the distance from the closest boundary with given boundary values. This is a general conclusion that pertains also for schemes other than leapfrog. If the operators B_m are extrapolation formulas or if they are derived from the differential equations, we get

$$B_m[tF_1(x + ct)] = \begin{cases} O(h), & m = 0, \ldots, r - 1, N - q + 1, \ldots, N - 1, \\ tF_1(ct) + O(h), & m = N. \end{cases}$$

If we overspecify the boundary values, this may no longer be true. In that case we would have to consider the equation for d_{mn} with specially chosen values for the overspecified equations to obtain the same result as above. The influence of the overspecification must then be treated separately.

Let us try to account for all the other errors by introducing a discrete function g_{mn} satisfying

$$Qg_{mn} = 0,$$

$$B_m g_{mn} = -B_m[F_2(mh + cnk) - (-1)^n F_2(mh - cnk)],$$

$$S_n g_{mn} = -S_n[(-1)^N F_3(mh + cnk - 2) - (-1)^m F_3(-mh + cnk)],$$

$$S_{mn} g_{mn} = S_{mn} F(mh + cnk) - b_{mn} - S_{mn}(F_2(mh + cnk) - (-1)^n F_2(mh - cnk)) - S_{mn}((-1)^N F_3(mh + cnk - 2) - (-1)^m F_3(-mh + cnk)).$$

We will express the leading part of g_{mn} as a test solution of the form

$$g_{mn} = G_1(mh + cnk) + (-1)^n G_2(mh - cnk) + (-1)^m G_3(-mh + cnk) + (-1)^n G_4(mh + cnk) + O(h).$$

We get $Qg_{mn} = O(h^2)$. The boundary conditions give

$$-2(-1)^n G_2(-cnk) + 2\lambda c G_3(cnk) - 2(-1)^n G_4(cnk) = -2(-1)^n F_2(-cnk) + O(h) \quad \text{for } cnk \geqslant 0.$$

$$G_1(1 + cnk) + (-1)^n G_2(1 - cnk) + (-1)^N G_3(-1 + cnk) + (-1)^n G_4(1 + cnk) = -F_2(1 + cnk) + (-1)^n F_2(1 - cnk) + O(h), \quad cnk \geqslant 0.$$

$$-2G_2(mh) - 2G_4(mh) = O(h), \qquad 0 \leqslant mh \leqslant 1.$$

$$G_1(mh) + G_2(mh) + (-1)^m G_3(-mh) + G_4(mh) = -(-1)^N F_3(mh - 2) + (-1)^m F_3(-mh) + O(h), \qquad 0 \leqslant mh \leqslant 1.$$

Let $G_3(-x) = F_3(-x)$, $0 \leqslant x \leqslant 1$, and $G_3(x) = 0$, $x \geqslant 0$. Let

$$G_1(x) = \begin{cases} -(-1)^N F_3(x - 2), & 0 \leqslant x \leqslant 1, \\ -F_2(x) - (-1)^N F_3(x - 2), & 1 \leqslant x \leqslant 2, \\ -F_2(x), & x \geqslant 2. \end{cases}$$

Let $G_2(x) = F_2(x) + H(x)$; then we get

$$H(x) = \begin{cases} 0, & -2k \leqslant x \leqslant -2k + 1, \ k = 0, 1, 2, \ldots, \\ F_2(-2k - x), & -2k - 1 \leqslant x \leqslant -2k, \end{cases}$$

$$G_4(x) = \begin{cases} -F_2(x - 2k), & 2k \leqslant x \leqslant 2k + 1, \ k = 0, 1, 2, \ldots, \\ 0, & 2k + 1 \leqslant x \leqslant 2k + 2, \end{cases}$$

and

$$G_1(x+ct)+F_2(x+ct)+(-1)^N F_3(x+ct-2)=\begin{cases} F_2(x+ct), & 0\leqslant x+ct\leqslant 1,\\ 0, & 1\leqslant x+ct\leqslant 2,\\ (-1)^N F_3(x+ct-2), & x+ct\geqslant 2,\end{cases}$$

$$G_2(x-ct)-F_2(x-ct)=H(x-ct),$$

$$G_3(-x+ct)-F_3(-x+ct)=\begin{cases} 0, & 0\leqslant -x+ct\leqslant 1,\\ -F_3(-x+ct), & ct\geqslant x.\end{cases}$$

We can write $e^{\mathrm{III}}_{mn}=h^2(d_{mn}+g_{mn})+O(h^3)$, and the total error is $e=e^{\mathrm{I}}+e^{\mathrm{II}}+e^{\mathrm{III}}$. If we only consider the $O(h^2)$-terms, we find that the error from the boundary approximation is not present close to the initial level and that the smooth part of the initial error disappears for $x+ct\geqslant 1$. That part of the initial error which is not smooth is trapped between the two boundaries and bounces back and forth, neither increasing nor decreasing in magnitude. For a dissipative scheme this error would quickly decrease. For an implicit scheme it is not as obvious that the boundary errors will only propagate gradually into the interior but the same kind of analysis undertaken, e.g. for Crank-Nicolson's scheme with a suitable boundary approximation shows the same characteristic features. In general, we use a test solution which contains solutions corresponding to the different characteristic roots in space and time. The general pattern is the same; parts of the initial error and the boundary error are annihilated in certain regions, and certain solutions travel back and forth between the boundaries. For the quarterplane problem $0\leqslant t$, $0\leqslant x$ the pattern is somewhat simpler since we need never take any reflection in the right boundary into account.

5. Numerical Examples.

5.1. *Comparison of Initial Approximations for Leapfrog with Second and Fourth Order Approximation in Space.* Let us consider the second order leapfrog scheme which we have used for illustration in the previous examples. We will investigate a few different initial approximations and demonstrate their influence on the total error for the Cauchy or the periodic problem. The following schemes may be used as initial approximations.

(a) $v_{m1}=u_{m1}$,

(b) $v_{m1}=v_{m0}+0.5\lambda c(v_{m+1\,0}-v_{m-1\,0})$, all m,

(c) $v_{m1}=v_{m0}+0.5\lambda c(v_{m+1\,0}-v_{m-1\,0})+0.5\lambda^2c^2(v_{m+1\,0}-2v_{m0}+v_{m-1\,0})$.

The list is by no means exhaustive, but we think that these schemes are representative of common choices. Let us call them exact, centered Euler and Lax-Wendroff, respectively. After Fourier transformation with respect to x the error can be written (cf. Example 1)

$$\tilde{e}^{\mathrm{I}}_n(\omega)=\tilde{u}_0(\omega)(\exp(2\pi i\omega cnk)-z_0^n)+\frac{P_1(z_0)}{P_1(z_1)}\tilde{u}_0(\omega)(z_1^n-z_0^n)$$
$$+O((P_1(z_0)/P_1(z_1))^2),$$

where

$$z_0 = \exp\left(2\pi i\omega ck - \frac{4}{3}\, ick\pi^3\omega^3h^2(1 - \lambda^2c^2) + O((\omega h)^4)\right)$$

and $z_1 = -1/z_0$. The function $P_1(z)$ is defined from the initial condition after Fourier transformation. In case (a) we get

$$P_1(z_0)/P_1(z_1) = (z_0 - \exp(2\pi i\omega ck))/(z_1 - z_0),$$

while for the other two cases we obtain (b) $P_1(z) = z - 1 - \lambda ci \sin(2\pi\omega h)$ and (c) $P_1(z) = z - 1 - \lambda ci \sin(2\pi\omega h) + 2\lambda^2c^2\sin^2(\pi\omega h)$. An upper bound for the relative error is given by

$$\tilde{e}_R^I = \frac{4}{3}\, \omega ct\pi^3(1 - \lambda^2c^2)/M^2 + 2|P_1(z_0)/P_1(z_1)|,$$

where $M = 1/(\omega h)$ is the number of points per wavelength. In Table 1 the leading term of $P_1(z_0)/P_1(z_1)$ is listed together with an estimate of how large M must be chosen to make $\tilde{e}_R^I$ less than 10% or less than 1% after one period in time. The figures are given for $\lambda c = 0.75$.

Scheme	$P_1(z_0)/P_1(z_1)$	10%	1%
Exact	$\frac{2}{3}\,\pi^3\lambda c(1 - \lambda^2c^2)/M^3$	14	43
Centered Euler	$\pi^2\lambda^2c^2/M^2$	18	55
Lax-Wendroff	$2\pi^4\lambda^2c^2(1 - \lambda^2c^2)/M^4$	14	43

TABLE 1

The leading term of $P_1(z_0)/P_1(z_1)$ for different initial approximations and the second order leapfrog scheme is given in column 1 while columns 2 and 3 contain the number of points per wavelength that are needed to obtain a relative error of less than 10% or less than 1% respectively. The value of λc is 0.75.

In practice the exact values are seldom known, but we see that the Lax-Wendroff scheme gives at least the same accuracy. For complicated problems in several space dimensions Lax-Wendroff may be difficult to implement, and we may have to use centered Euler. However, the interior error dominates the initial error after a very short period of time, for Lax-Wendroff already at the first step and for centered Euler the two errors are of the same size for $\omega ct \geqslant 3\lambda^2c^2/(2\pi(1 - \lambda^2c^2))$, (e.g. $\omega ct = 0.614$ for $\lambda c = 0.75$). From the expression for $\tilde{e}_n^I(\omega)$ we expect the error to oscillate with time. The true error and the estimate $\tilde{e}_R^I$ are plotted in Figure 1 (p. 17) over two periods of time using $M = 30$, $\lambda c = 0.75$ and centered Euler. To avoid mixture with the influence of the boundary conditions we have chosen the periodic problem $u_t = u_x$, $u(x, 0) = \sin(2\pi x)$ and $u(0, t) = u(1, t)$. The oscillations are clearly seen, and the theory is seen to give only a very small overestimate of the true error.

We will also consider an approximation which is fourth order in space and second

order in time, namely

$$v_{mn+1} = v_{mn-1} + 2\lambda c\left(\frac{2}{3}(v_{m+1n} - v_{m-1n}) - \frac{1}{12}(v_{m+2n} - v_{m-2n})\right),$$

which is stable for $\lambda c < 0.7287$. The local truncation error is $(1/6)k^3u_{ttt} + (1/30)kh^4u_{xxxxt} + O(k^5)$. The characteristic roots corresponding to the Cauchy problem are

$$z_0 = \exp\left(2\pi i\omega ck + \frac{4}{3}i\pi^3\omega^3c^3k^3 - \frac{16}{15}ickh^4\pi^5\omega^5 + O(k^5)\right)$$

and $z_1 = -1/z_0$. We will also include a scheme which is fourth order accurate in space

(d) $$v_{m1} = v_{m0} + \lambda c\left(\frac{2}{3}(v_{m+1\,0} - v_{m-1\,0}) - \frac{1}{12}(v_{m+2\,0} - v_{m-2\,0})\right)$$

with the corresponding function

$$P_1(z) = z - 1 - \lambda ci\left(\frac{4}{3}\sin(2\pi\omega h) - \frac{1}{6}\sin(4\pi\omega h)\right).$$

The relative error is given by

$$\hat{e}_R^{\mathrm{I}} = \left|\frac{4}{3}\pi^3\omega^2c^2k^2 - \frac{16}{15}\pi^5\omega^4h^4\right|\omega ct + 2|P_1(z_0)/P_1(z_1)|.$$

Due to the minus sign in the expression of the interior error there are certain choices of λc and M, which are more favorable than others, and for which the error emanating from the initial approximation has an important influence on the total error for quite a long time. The interior error after one period in time is plotted as a function of the number of points per wavelength in Figure 2 with $\lambda c = 0.2$. We see that for $M = 14$ the error is less than 10^{-4}, while for M in the interval 15–32 the interior error is greater than 10^{-3}. The initial error for centered Euler and the theoretical estimate of the total error as well as the true total error after one period in time are also plotted in Figure 2. We have used the same periodic problem as above. In Table 2 we give the expression for the leading term of $P_1(z_0)/P_1(z_1)$ and an estimate of how large M must be chosen to obtain errors less than 1% or less than 0.1% after one period in time. The results are given both for $\lambda c = 0.2$ and $\lambda c = 0.02$.

These results show that if we are satisfied with an error of about 1% we should not choose λc too small, since nothing is gained from the increase in work. If we require an error of about 0.1%, smaller λc seems to be favorable. However, the computational work is proportional to the total number of points. For $\lambda c = 0.2$ the work is proportional to $48^2 \cdot 5$ in one space dimension or $48^3 \cdot 5$ or $48^4 \cdot 5$ in two or three space dimensions. These figures should be compared to $24^2 \cdot 50, 24^3 \cdot 50$ and $24^4 \cdot 50$ for $\lambda c = 0.02$, respectively. Thus, $\lambda c = 0.2$ gives a more efficient scheme unless we require even better accuracy or work with a three-dimensional problem. The choice of initial approximation is not very important, but for high accuracy and not so small λc the Lax-Wendroff scheme seems to be preferable, whereas for high accuracy and small λc the fourth order approximation should be recommended.

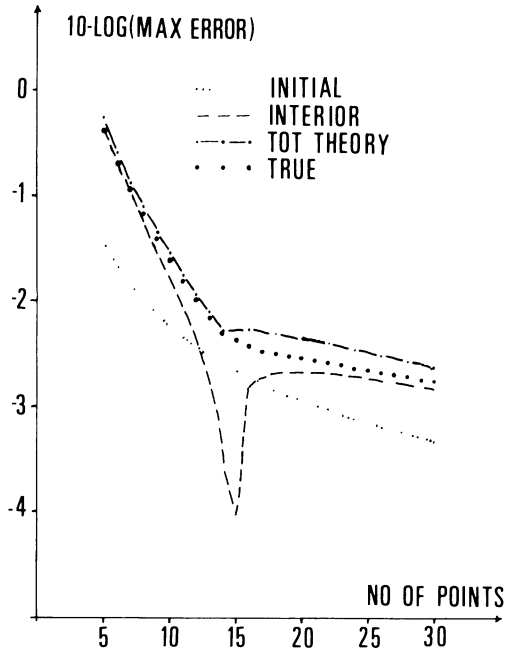

FIGURE 2

The interior error, the initial error $P_1(z_0)/P_1(z_1)$ and the total theoretical error estimates are shown together with the true error for the model problem $u_t = u_x$, $u(x, 0) = \sin(2\pi x)$, $u(0, t) = u(1, t)$. The interior scheme is leapfrog with fourth order approximation in space and the initial approximation is centered Euler. The value of λc is 0.2 and the error is shown after one period in time for different numbers of points per wavelength.

Scheme	$P_1(z_0)/P_1(z_1)$	1%		0.1%	
		0.2	0.02	0.2	0.02
Exact	$0.5\lambda c/M\left\lvert\frac{4}{3}\pi^3\lambda^2c^2/M^2 - \frac{16}{15}\pi^5/M^4\right\rvert$	11	14	38	24
Centered Euler	$\lvert\pi^2\lambda^2c^2/M^2 + i\frac{2}{3}\lambda c\pi^3/M^3\rvert$	12	14	48	28
Lax-Wendroff	$\left\lvert\frac{4}{3}\pi^4\lambda^2c^2/M^4 + i\frac{2}{3}\lambda c\pi^3/M^3\right\rvert$	12	14	41	28
Fourth order	$\pi^2\lambda^2c^2/M^2$	12	14	48	24

TABLE 2

The leading term of $P_1(z_0)/P_1(z_1)$ is listed for different initial approximations and the fourth order leapfrog scheme in column 1. In columns 2 and 3 theoretical estimates are given of the number of points per wavelength which are needed to obtain a relative error of less than 1% for $\lambda c = 0.2$ and 0.02, respectively. In columns 4 and 5 the corresponding estimates are given for 0.1% relative error.

5.2. *Comparison of Boundary Conditions for Second Order Leapfrog.* Let us consider the usual leapfrog scheme with the following boundary approximations at the outflow boundary:

(a) $v_{0n+1} = v_{0n} + \lambda c(v_{1n} - v_{0n})$,

(b) $v_{0n+1} = v_{0n-1} + 2\lambda c(v_{1n} - \frac{1}{2}(v_{0n+1} + v_{0n-1}))$

(c) $v_{0n+1} + v_{1n+1} - \lambda c(v_{1n+1} - v_{0n+1}) = v_{0n} + v_{1n} + \lambda c(v_{1n} - v_{0n})$.

We call these approximations explicit, weighted in time and the box scheme. They have all been shown to give stable total schemes, (a) and (c) in Gustafsson et al. [3] and (b) in Elvius and Sundström [1]. The relative error $\tilde{e}_R^{II}$, which was introduced in Section 4.1, can be expressed as $(\kappa_1^m(\kappa_2/\kappa_1)^N - \kappa_2^m) \cdot R_0(\kappa)/R_0(\kappa_2)$, where $\kappa = \exp(2\pi i\gamma h/c)$. The interior root κ_2 approaches -1 as h and k go to zero. We obtain the following expressions for $R_0(\kappa)$:

(a) $R_0(\kappa) = z - 1 - \lambda c(\kappa - 1)$,

(b) $R_0(\kappa) = z - z^{-1} - 2\lambda c(\kappa - \frac{1}{2}(z + z^{-1}))$,

(c) $R_0(\kappa) = (z - 1)(1 + \kappa) - \lambda c(z + 1)(\kappa - 1)$,

where $z = \exp(2\pi i\gamma k)$. The total error for the problem in a strip $0 \leqslant x \leqslant 1$ is given by

$$e_R = (x - 1)\frac{4}{3}\pi(1 - \lambda^2 c^2)/M^2 + 2P_1(z_0)/P_1(z_1) + 2R_0(\kappa)/R_0(\kappa_2).$$

If we choose the Lax-Wendroff scheme for the initial approximation, the term $P_1(z_0)/P_1(z_1)$ is $O(1/M^4)$ and, therefore, negligible in comparison with the other terms. In Table 3 the leading term of $R_0(\kappa)/R_0(\kappa_2)$ is listed together with estimates of how many points per wavelength that are needed to obtain a relative error e_R of less than 10% or less than 1%. For comparison we also give the true error for the problem $u_t = u_x$, $u(x, 0) = \sin(2\pi x)$, $u(1, t) = \sin(2\pi t)$ using meshes with 16 or 48 points per wavelength.

Boundary approximation	$R_0(\kappa)/R_0(\kappa_2)$	10%	1%	16 pts.	48 pts.
Explicit	$\pi^2(1 - \lambda c)/M^2$	16	48	9.00E-2	9.90E-3
Weighted	$\pi^2(1 - \lambda^2 c^2)/M^2$	19	60	1.04E-1	1.16E-2
Box Scheme	$(1/3)\pi^3(1 - \lambda^2 c^2)/M^3$	14	43	7.30E-2	7.90E-3

TABLE 3

The leading term of the boundary error is listed for different boundary conditions for second order leapfrog in column 1. In columns 2 and 3 the theoretical estimates of the number of points needed to obtain an error of less than 10% or less than 1% are given. In columns 4 and 5 the true maximum error for the model problem are given for 16 and 48 points, respectively. The value of λc is 0.75, and we have computed for 5 periods of time.

The figures in the table show that the theory agrees very well with the computational results. This is also seen in Figure 3, where the true error and the theoretical estimate

are plotted as a function of time for the weighted approximation. The same model problem as above was used.

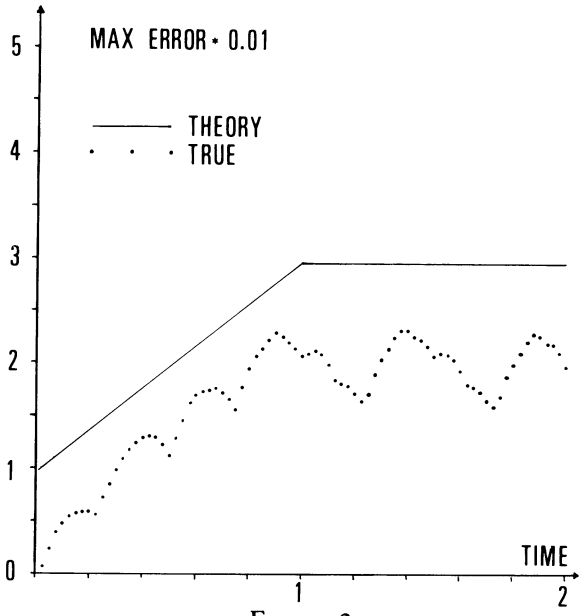

FIGURE 3

The theoretical error estimate and the true error for the model problem $u_t = u_x$, $u(x, 0) = \sin(2\pi x)$, $u(1, t) = \sin(2\pi t)$ are shown for two periods of time. The interior scheme is second order leapfrog, the initial approximation is Lax-Wendroff and the boundary approximation is weighted in time. The value of λc is 0.75 and we have used 30 points per wavelength in space.

In this example the box scheme is of "unnecessarily" high order—we only need a scheme with the local second order accuracy to ensure the overall second order accuracy. As we see from the table, however, there is certainly a substantial gain of accuracy in using the box scheme instead of the weighted scheme.

Let us briefly discuss two different boundary approximations for the leapfrog scheme with a fourth order accurate difference operator in space. The box scheme applied twice at the left boundary and once at the right boundary can be shown to give a stable approximation for both the right and the left quarter plane problems. The "correct" boundary scheme should be such that the *function values* are approximated with second order accuracy in time and fourth order accuracy in space. Such a scheme was proposed and shown to be stable by Oliger in [5]. The box scheme on the other hand gives only third order accurate function values (in both time and space). The error constants for the box scheme are much smaller than those for the extrapolation scheme. Therefore, the results using the box scheme are actually somewhat better, when the number of points per wavelength is relatively small although the box scheme is of lower order of accuracy. This is true even for a very small timestep, $\lambda c = 0.02$. Theoretically, we can see this by computing the solution to the relevant boundary value problem for the resolvent equation and compare the resulting coefficients for the two schemes. In Table 4 we show the results of applying the two schemes to our test problem for different choices of λc.

Boundary Approximation	$\lambda c = 0.2$			$\lambda c = 0.02$	
	No. of points			No. of points	
	8	16	32	8	16
Box Scheme	5.1E-2	1.6E-2	3.2E-3	5.1E-2	8.2E-3
Oliger's scheme	1.4E-1	9.7E-3	1.3E-3	1.4E-1	1.1E-2

TABLE 4
The maximum error obtained in computations using the leapfrog scheme with fourth order approximation in space and exact initial approximation.

We have included this result to demonstrate the importance of having some knowledge of the error constant when we compare different schemes. For a fixed stepsize the limiting order of the scheme does not provide sufficient information for us to make the proper choice.

In the final choice of difference approximation we must also include an estimate of the work and storage requirements. This is discussed in more detail in Section 5.4. We only want to point out here that in several dimensions the box scheme is usually not a good choice together with an explicit scheme. It will connect all the points along one boundary. Therefore, we will need to solve a system of equations at each step, which may increase the operation count drastically. On the other hand the box scheme is a very good choice together with some implicit schemes as we will see below.

5.3. *Comparison of Boundary Conditions for Two Implicit Schemes.* In this section we will discuss the choice of boundary conditions for Crank-Nicolson and a modified version which is of fourth order in space. The interior scheme is defined by

$$v_{m\,n+1} - \frac{1}{4}\lambda c(v_{m+1\,n+1} - v_{m-1\,n+1}) + \beta(v_{m+1\,n+1} - 2v_{m\,n+1} + v_{m-1\,n+1})$$

$$= v_{mn} + \frac{1}{4}\lambda c(v_{m+1\,n} - v_{m-1\,n}) + \beta(v_{m+1\,n} - 2v_{mn} + v_{m-1\,n})$$

with $\beta = 0$ for Crank-Nicolson and $\beta = 1/6$ for the modified scheme. The local truncation errors for these schemes are

$$-\frac{1}{12}k^3u_{ttt} - \frac{1}{6}kh^2u_{xxt} + O(h^4)$$

and

$$-\frac{1}{12}k^3u_{ttt} + kh^4\frac{1}{180}u_{xxxxt} + O(k^4 + h^2k^3).$$

The relative error for the pure initial value problem is thus given by

$$\hat{e}_R^{\mathrm{I}} = \omega ct\frac{4}{3}\left(1 + \frac{1}{2}\lambda^2c^2\right)\pi^3/M^2$$

and

$$\hat{e}_R^{\mathrm{I}} = \omega ct\left(\frac{2\pi^3\lambda^2c^2}{3M^2} + \frac{8\pi^5}{45M^4}\right),$$

respectively, where M is the number of points per wavelength.

The schemes are unconditionally stable for the Cauchy problem. They both need one additional boundary approximation. This boundary condition must be chosen with care since it may affect both the stability and the accuracy of the scheme. In Sköllermo [6] a few different choices of boundary conditions are shown to give unconditionally stable schemes, but there is also one example of an explicit boundary approximation which makes the total scheme only conditionally stable.

We will consider the following boundary approximations and study their influence on the accuracy of the schemes:

(a) $(hD_+)^j v_{0\,n+1} = 0, \quad j = 1, 2, 3$, where $D_+ v_m = (v_{m+1} - v_m)/h$,

(b) $v_{0\,n+1} - \lambda c(v_{1\,n+1} - v_{0\,n+1}) = v_{0n}$,

(c) $v_{0\,n+1} - \tfrac{1}{2}\lambda c(v_{1\,n+1} - v_{0\,n+1}) = v_{0n} + \tfrac{1}{2}\lambda c(v_{1n} - v_{0n})$,

(d) $v_{0\,n+1} + v_{1\,n+1} - \lambda c(v_{1\,n+1} - v_{0\,n+1}) = v_{0\,n} + v_{1\,n} + \lambda c(v_{1\,n} - v_{0\,n})$.

We will call these schemes extrapolation of order j, fully implicit, half implicit and the box scheme. They give unconditionally stable schemes, which was shown in [6] for (a), (b) and (d), and which is easy to show also for (c). In Section 4.1 we showed that the influence of the boundary approximation on the error is described by a quotient $R_0(\kappa)/R_0(\kappa_2)$, where $z = \exp(2\pi i\gamma k)$, $\kappa = \exp(2\pi i\gamma h/c)$ and κ_2 corresponds to the interior root to the resolvent equation. This root approaches -1 as h and k go to zero both for Crank-Nicolson and for the modified scheme. The functions $R_0(\kappa)$ are defined by

(a) $R_0(\kappa) = (\kappa - 1)^j$,

(b) $R_0(\kappa) = z - 1 - z\lambda c(\kappa - 1)$,

(c) $R_0(\kappa) = z(1 - \tfrac{1}{2}\lambda c(\kappa - 1)) - (1 + \tfrac{1}{2}\lambda c(\kappa - 1))$,

(d) $R_0(\kappa) = z(1 + \kappa - \lambda c(\kappa - 1)) - (1 + \kappa + \lambda c(\kappa - 1))$.

The leading terms of the quotients are tabulated in Table 5 together with estimates of how many points per wavelength are needed to obtain a relative error of less than 10% or less than 1%. The true maximum error is also given for 16 and 48 points per wavelength. We have used the same model problem as was used previously.

From this table we can see that the first order boundary approximation indeed destroys the total accuracy while the box scheme and the third order extrapolation are equivalent. The box scheme is preferable, however, since it is somewhat easier to implement efficiently.

The same boundary conditions can be used together with the modified scheme. Since $\kappa_2 = -1$ for $h = 0$ also in this case, the leading term of the boundary error is the same as above. In Table 6 we list the number of points per wavelength that are needed to obtain a relative error of less than 1% or less than 0.1%. We also give the true maximum error for the model problem with 8, 16 and 32 points per wavelength in Table 7.

Boundary Approximation	$R_0(\kappa)/R_0(\kappa_2)$	10%	1%	16	48
Extrapolation					
$j = 1$	π/M	71	637	5.9E-1	1.5E-1
$j = 2$	π^2/M^2	27	86	2.4E-1	2.6E-2
$j = 3$	π^3/M^3	24	74	2.2E-1	2.3E-2
Fully implicit	$\pi^2(1 + \lambda c)/M^2$	30	94	3.3E-1	3.0E-2
Half implicit	π^2/M^2	27	86	2.8E-1	2.6E-2
Box scheme	$\frac{1}{3}\pi^3(1 - \lambda^2 c^2)M^3$	24	73	2.3E-1	2.3E-2

TABLE 5

The leading term of the boundary error is tabulated in column 1. Columns 2 and 3 contain the number of points per wavelength that are needed to obtain a relative error of less than 10% or less than 1% according to the theory. In columns 4 and 5 the true maximum error has been listed for the model problem. The value of λc is 0.75.

Boundary	1%		0.1%	
Approximation	0.2	0.02	0.2	0.02
Extrapolation				
$j = 2$	46	45	144	141
$j = 3$	21	19	47	40
Box scheme	16	14	38	29

TABLE 6

The estimated number of points per wavelength needed to obtain a relative error less than 1% or less than 0.1% for $\lambda c = 0.2$ and $\lambda c = 0.02$.

The box scheme is seen to be the best choice in all cases. Its superiority is more pronounced for the fourth order scheme than it was for the second order scheme. Extrapolation of order 2 was chosen as representative for the second order schemes. We see that the boundary approximation plays a dominant role in the error in this case. The theoretical estimates of Table 6 are seen to agree very well with the figures of Table 7.

5.4. *Discussion.* In this section we will make an attempt to compare the schemes which we have considered in the previous sections.

Let us discuss the pure initial value problem. We have considered four schemes, namely leapfrog and Crank-Nicolson of second order in both time and space and two versions which are of fourth order in space. We compare the number of points per wavelength which are needed to obtain a relative error of less than 10% or less than 1% in Table 8.

	$\lambda c = 0.2$			$\lambda c = 0.02$	
Boundary	No. of points			No. of points	
Approximation	8	16	32	8	16
Extrapolation					
$j = 2$	2.61E-1	7.40E-2	1.91E-2	2.74E-1	7.87E-2
$j = 3$	1.03E-1	1.92E-2	2.92E-3	1.05E-1	2.37E-2
Box scheme	5.31E-2	6.59E-3	1.14E-3	4.10E-2	4.75E-3

TABLE 7

The true maximum error for the modified scheme with different boundary approximations and for different numbers of points.

Scheme	10%	1%
Leapfrog with Lax-Wendroff	18	56
4th order leapfrog with Lax-Wendroff	10	33
Crank-Nicolson	21	68
Modified Crank-Nicolson	7	23

TABLE 8

The number of points per wavelength which are needed to obtain a relative error of less than 10% or less than 1% are listed for $\lambda c = 0.50$.

We have chosen $\lambda c = 0.5$ so that all four schemes can be compared. The leapfrog scheme is stable for $\lambda c \leqslant 1$ but the fourth order explicit scheme is stable only for $\lambda c \leqslant 0.7287$. The ordinary leapfrog scheme involves four gridpoints and the fourth order approximation six gridpoints to determine a new point. The increase in work is thus about 50% for the fourth order scheme and the corresponding figures in Table 8 should be multiplied by 1.5 before they are compared to those for the second order leapfrog scheme. We then find that the fourth order scheme is more efficient even for this relatively large value of λc. The choice between the second and fourth order versions of the Crank-Nicolson scheme is easily made in favor of the higher order scheme, which requires very little extra work since exactly the same gridpoints are involved in both schemes.

The explicit schemes are clearly preferable unless the problem is very stiff which we discuss below. The fourth order schemes show the same result both for the pure initial value problem and for the problem in a strip when they are compared for smaller values of λc and for better accuracy. In Table 9 we list the number of meshpoints per wavelength that are needed for the Cauchy problem.

Scheme	1%		0.1%	
	0.2	0.02	0.2	0.02
4th order leapfrog with Lax-Wendroff	12	14	41	28
Modified Crank-Nicolson	7	90	28	16

TABLE 9

The number of points per wavelength which are needed to obtain a relative error of less than 1% or less than 0.1% are listed for $\lambda c = 0.02$.

The importance of the implicit schemes lies, however, in the fact that they are unconditionally stable. When we work with stiff problems with widely varying values of the constant λc the conditional stability of the explicit schemes may force us to use unnecessarily small timesteps, while for the implicit schemes we can choose λc such that the important part of the solution is accurately described. Let us try to estimate how stiff a problem should be for the implicit schemes to be competitive.

We consider the one-dimensional Cauchy problem for a system of equations where the moduli of the eigenvalues range from c_{min} to c_{max}. Let c_{int} be the largest value for which we are interested in an accurate solution. Let λ_I denote the ratio k/h for the implicit scheme and let λ_E be the ratio k/h for an explicit scheme. Similarly, we let M_I and M_E denote the number of points per wavelength for the implicit and explicit schemes, respectively.

The relative error for Crank-Nicolson second order implicit scheme is

$$\omega c_{int}\, t \cdot \frac{4}{3}\pi^3 \left(1 + \frac{1}{2}\lambda_I^2 c_{int}^2\right) \Big/ M_I^2$$

and the error for the second order explicit leapfrog scheme is

$$\omega c_{int}\, t \cdot \frac{4}{3}\pi^3 (1 - \lambda_E^2 c_{int}^2) M_E^2 .$$

The number of operations required to advance the solution from the initial time 0 to time t are

$$14\alpha M_I^2/\lambda_I \quad \text{and} \quad 4\alpha M_E^2/\lambda_E$$

for Crank-Nicolson's scheme and the leapfrog scheme, respectively, where α is a certain constant depending on t and 14 and 4 reflect the number of operations per point needed to advance the solution one step in time.

Suppose we choose M_I and M_E so that the relative errors are equal. At which ratio c_{int}/c_{max} is the Crank-Nicolson scheme more efficient than the leapfrog scheme? If we decide to pick $\lambda_E c_{max} = 1$ to ensure stability for the explicit leapfrog scheme, we get the following table for different choices of $\lambda_I \cdot c_{int}$.

$\lambda_I c_{int}$	c_{int}/c_{max}
0.99	0.18
0.75	0.21
0.50	0.23

TABLE 10

The stiffness ratio c_{int}/c_{max} tabulated against different values of $\lambda_I c_{int}$. The implicit scheme is second order Crank-Nicolson and the explicit scheme is second order leapfrog in time and space.

The conclusion is that the usual Crank-Nicolson scheme is more efficient than the second order leapfrog scheme if the stiffness ratio c_{int}/c_{max} is less than approximately 1/5. If the problem at hand involves much overhead common to both methods this ratio may be increased.

We can compare the second and fourth order schemes in a similar way but the expressions get more complicated so we need not only to pick $\lambda_I c_{int}$ but also a certain error level. The modified fourth order Crank-Nicolson scheme and the second order leapfrog scheme give the same error but the implicit scheme requires less work if

$$14\left(\frac{\lambda_I^2 c_{int}^2}{3}+\frac{4\pi^2}{45M_I^2}\right)\cdot\frac{c_{int}}{c_{max}} < \frac{8}{3}\left(1-\frac{c_{int}^2}{c_{max}^2}\right)\cdot \lambda_I c_{int}.$$

We list M_E, the error level e_R, and the ratio c_{int}/c_{max} for some combinations of $\lambda_I c_{int}$ and M_I in Table 11.

$\lambda_I c_{int}$	M_I	M_E	e_R	c_{int}/c_{max}
0.5	10	21	2.86-2	0.62
0.5	15	32	1.20-2	0.64
0.2	10	37	6.85-3	0.75
0.2	15	56	2.37-3	0.79

TABLE 11

Comparison of the fourth order Crank-Nicolson scheme and the second order leapfrog scheme for different choices of $\lambda_I c_{int}$ and M_I.

Thus, the fourth order implicit scheme is more efficient than the second order explicit scheme if the stiffness ratio is less than 0.6 to 0.8 and if we require an error level of less than 3%. The smaller error we require the larger the stiffness ratio can be.

Finally, the leapfrog scheme with fourth order accuracy in space can be compared to the modified Crank-Nicolson scheme which is also fourth order accurate in space. If we, e.g., choose $\lambda_I c_{int} = 0.2$ and pick $M_I = M_E = 10$ we get the same

error, 6.85E-3, if $c_{int}/c_{max} = 0.31$, in which case the implicit scheme requires less work. The nonlinear relation between M_E^2 and M_I^2 makes a strict analysis very complicated without adding any substantial new knowledge.

For problems in several space dimensions the implicit schemes can compare favorably to the explicit ones only for much smaller stiffness ratios, except maybe in special cases where the resulting block-tridiagonal systems can be solved very efficiently.

6. Summary. We have developed a technique for the error analysis of finite difference approximations to hyperbolic mixed initial boundary value problems. The errors emanating from the interior scheme, the initial approximation and the boundary conditions can be discussed more or less separately.

We know that the initial approximation should have a local truncation error at least of the same order as the global error for the interior scheme to keep the overall accuracy at the desired level. From Table 1 we see that it may be quite profitable to use higher order schemes if possible. In our example centered Euler needs about 30% more points per wavelength than Lax-Wendroff to guarantee a relative error of less than 1%. Formally, the local truncation error in the boundary approximation should also be of the same order as the global error of the interior scheme. A small error constant may, however, make a scheme competitive which is formally not of the "right" order–at least for a small number of points. However, we also notice, in Tables 5 and 6, how a boundary condition of too low accuracy dominates the total error (extrapolation of order one and two, respectively). Neither the box scheme nor the third order extrapolation are formally of the "right" order to use together with the fourth order implicit scheme, but we see from Table 7 that at least the box scheme works very well. As was pointed out in Section 5.4, the implicit schemes are not competitive unless the problem is fairly stiff.

The information in the tables in Section 5 should give some insight into how different choices of initial or boundary approximations can be expected to influence the accuracy of the total scheme. The relative merits of the different schemes should hold also in more complicated situations, although the number of points needed to obtain a certain accuracy can only be used as a guideline. The technique of the analysis has been demonstrated in the examples of Sections 3 and 4 and its usefulness and applications should be evident from Section 5. It is our hope that this paper will be useful in the comparison and choice of difference approximations in various situations.

Department of Computer Sciences
Sturegatan 4B 2tr
S-752-23 Uppsala, Sweden

1. T. ELVIUS & A. SUNDSTRÖM, "Computationally efficient schemes and boundary conditions for a fine-mesh barotropic model based on the shallow-water equations," *Tellus*, v. 25, 1973, pp. 132–156.

2. B. GUSTAFSSON, "The convergence rate for difference approximations to mixed initial boundary value problems," *Math. Comp.*, v. 29, 1975, pp. 396–406.

3. B. GUSTAFSSON, H.-O. KREISS & A. SUNDSTRÖM, "Stability theory of difference approximations for mixed initial boundary value problems. II," *Math. Comp.*, v. 26, 1972, pp. 649–686.

4. H.-O. KREISS & J. OLIGER, "Comparison of accurate methods for the integration of hyperbolic equations," *Tellus*, v. 24, 1972, pp. 199–215.

5. J. OLIGER, "Fourth order difference methods for the initial boundary value problem for hyperbolic equations," *Math. Comp.*, v. 28, 1974, pp. 15–25.

6. G. SKÖLLERMO, *How the Boundary Conditions Affect the Stability and Accuracy of Some Implicit Methods for Hyperbolic Equations*, Report No. 62, Dept. of Comput. Sci., Uppsala University, 1975.

7. B. SWARTZ & B. WENDROFF, "The relative efficiency of finite difference and finite element methods. I: Hyperbolic problems and splines," *SIAM J. Numer. Anal.*, v. 11, 1974, pp. 979–993.