On Difference Approximations with Wrong Boundary Values*

By Heinz-Otto Kreiss and Einar Lundqvist

1. Introduction. Consider the differential equation

$$\partial u/\partial t = \partial u/\partial x \tag{1.1}$$

in the quarter space $x \geqq 0,\ t > 0$. (1.1) has a unique solution if initial values

$$u(x, 0) = f(x)\,, \qquad 0 \leqq x < \infty\,, \tag{1.2}$$

are given. We want to solve this problem by difference approximation. Therefore, we introduce a time-step $k > 0$, a mesh-width $h = 1/N$, N a natural number, and gridpoints x_ν by $x_\nu = \nu h$, $\nu = 0, \pm 1, \pm 2, \cdots$. As usual, we assume that $k/h = \lambda$ where $\lambda > 0$ is a constant. Denoting by $v_\nu(t) = v(x_\nu, t)$ a function defined for all $x = x_\nu$ and $t = t_m = mk$, $m = 0, 1, 2, \cdots$ we approximate (1.1), (1.2) by

$$\begin{aligned} v_\nu(t + k) &= Qv_\nu(t)\,, \\ v_\nu(0) &= f(x_\nu)\,. \end{aligned} \qquad \nu = 0, 1, 2, \cdots . \tag{1.3}$$

Here Q is a difference operator which can be written under the form

$$Q = \sum_{j=-p}^{q} a_j E^j\,, \qquad Eg(x) = g(x + h)\,, \tag{1.4}$$

where a_j are constants.

In contrast to the continuous problem $v_\nu(t)$ is not uniquely determined by (1.3), because we cannot compute $v_0(t + k)$ without $v_{-p}(t), v_{-p+1}(t), \cdots, v_{-1}(t)$. We therefore introduce extra boundary conditions

$$v_\mu(t) = g_\mu(t)\,, \qquad \mu = -1, -2, -3, \cdots, -p\,, \tag{1.5a}$$

where $g_\mu(t)$ are any uniformly bounded functions, i.e.,

$$|g_\mu(t)| \leqq M\,, \qquad M \text{ constant.} \tag{1.5b}$$

Assume that the approximation (1.3) is stable. What can we say about the convergence of $v_\nu(t)$ towards $u(x, t)$, as $h \to 0$?

For two special cases this question has been answered in an interesting paper by S. Parter [1]. He has shown that the estimates of Theorem 1 hold for the Lax-Wendroff scheme and the Friedrichs scheme.

We want to generalize this result to general dissipative approximations, using a completely different technique. Furthermore, we shall give a fairly complete classification of all stable difference approximations according to the influence which the

Received May 29, 1967.

* The research reported in this paper was partially supported by the Swedish Technical Research Council and by the U. S. Atomic Energy Commission.

boundary conditions (1.5a) have on the solutions. To state our main results we need some definitions:

Definition 1. Let

$$\hat{Q}(\xi) = \sum_{j=-p}^{q} a_j e^{ij\xi} \tag{1.6}$$

be the Fourier transform of the difference operator Q. Then we call the approximation dissipative if

$$|\hat{Q}(\xi)| \leqq 1 - \delta|\xi|^{2s} \qquad \text{for } -\pi \leqq \xi \leqq \pi\,. \tag{1.7}$$

Here $\delta > 0$ is a constant and $s > 0$ a natural number.

Definition 2. We say that (1.3) is accurate of order m if

$$\hat{Q}(\xi) = e^{i\lambda\xi} + O(\xi^{m+1})\,, \qquad \lambda = k/h\,.$$

It is well known (see for example [2]) that Definition 2 is equivalent to the usual definition of the order of accuracy.

Definition 3. We say that $g_\nu = g(x_\nu)$ and $q_\nu = q(x_\nu)$ are grid functions if they are defined for all $x = x_\nu$, $\nu = 0, \pm 1, \pm 2, \cdots$ and

$$\sum_{\nu=-\infty}^{+\infty} |g_\nu|^2 < \infty\,, \qquad \sum_{\nu=-\infty}^{+\infty} |q_\nu|^2 < \infty\,.$$

Furthermore, we define scalar products and norms by

$$(g, q)_r = \sum_{\nu=r}^{\infty} \bar{g}_\nu q_\nu h\,, \qquad \|g\|_r^2 = (g, g)_r = \sum_{\nu=r}^{\infty} |g_\nu|^2 h\,.$$

If a function f_ν is only defined for $\nu \geqq l$, then we always extend the definition to all ν, $-\infty < \nu < \infty$, by assuming $f_\nu = 0$ for $\nu < l$.

We can now formulate our main result:

THEOREM 1. *Assume that the initial values $f(x) \in C^{m+1}(0, \infty)$*** and vanish for $x > R$, R some constant. Let the difference approximation be dissipative, accurate of order m, and assume that* (1.5b) *holds for the extra boundary conditions. Then there are constants $K_i > 0$, $i = 1, 2$, and $\alpha > 0$ such that we can write the solution $v_\nu(t)$ of the difference equation under the form*

$$v_\nu(t) = v_\nu^{(1)}(t) + v_\nu^{(2)}(t)\,, \qquad \nu = 0, 1, 2, \cdots,$$

and we have the estimates:

$$\begin{aligned} \|v_\nu^{(1)}(t) - u(x_\nu, t)\|_0 &\leqq tK_1 h^m\,, \\ |v_\nu^{(2)}(t)| &\leqq K_2\Big(M + \max_x |f(x)|\Big)e^{-\nu\alpha}\,. \end{aligned} \tag{1.8}$$

Therefore, the influence of the extra boundary conditions is present in an interval of length $\simeq$ const $h|\log h|$.

We want to formulate a more general result. For that reason we need

Definition 4. Let ξ and α be real, and consider the Fourier transform $\hat{Q}$ for com-

** $C^l(a, b)$ is the class of all functions which for $a \leqq x \leqq b$ are l times continuously differentiable.

plex arguments. The difference operator Q is called contractive if for all ξ with $|\xi| \leqq \pi$ and some $\alpha > 0$

$$|\hat{Q}(\xi)| = |\hat{Q}(\xi + i0)| \leqq 1 , \qquad |\hat{Q}(\xi + i\alpha)| \leqq e^{-\beta} . \tag{1.9}$$

Here $\beta > 0$ is a constant.

We are going to show:

THEOREM 2. *If Q is dissipative and accurate of order* (*at least*) *one, then Q is contractive.*

However, the converse is not true. If, for example, $\hat{Q}(\xi) = e^{i\xi}$, then $\hat{Q}(\xi)$ is contractive but not dissipative.

The more general result is stated in

THEOREM 3. *Replace in Theorem* 1 *the condition dissipative by contractive. Then the estimate* (1.8) *is still true.*

We can write $\hat{Q}(\xi + i\alpha)$ under the form

$$\hat{Q}(\xi + i\alpha) = \hat{Q}(\xi) + \alpha\hat{Q}_1(\xi) + O(\alpha^2) .$$

Therefore, a function $c(\xi)$ exists such that:

$$|\hat{Q}(\xi + i\alpha)| = |\hat{Q}(\xi)| \cdot e^{\lambda\alpha c(\xi)} + O(\alpha^2) \qquad \text{for } \hat{Q}(\xi) \neq 0 ,$$

and the condition (1.9) certainly holds if either $|\hat{Q}(\xi)| < 1$ or $|\hat{Q}(\xi)| = 1$ and $c(\xi) < 0$. This suggests the following definition

Definition 5. $\hat{Q}(\xi)$ is strictly noncontractive if $|\hat{Q}(\xi)| \leqq 1$ and for some ξ_0

$$|\hat{Q}(\xi_0 + i\alpha)| = \exp(\lambda c(\xi_0)\alpha) + O(\alpha^2) \qquad \text{with } c(\xi_0) > 0 . \tag{1.10}$$

The following theorem shows that Theorem 3 is almost the best possible result.

THEOREM 4. *Consider the difference approximation* (1.3) *with initial values $f(x) \equiv 0$ and assume that it is strictly noncontractive. Then we can find boundary functions $g_\mu(t)$ such that*:

$$y_\nu(t) = \exp(i\xi_0\nu + i\phi t/k)u(x_\nu, t) , \qquad e^{i\phi} = Q(\xi_0) ,$$

where $u(x, t)$ converges to the solution of the continuous problem:

$$\partial u/\partial t = -c(\xi_0)\partial u/\partial x , \qquad u(x, 0) = 0 , \qquad u(0, t) = 1 . \tag{1.11}$$

Therefore, the difference approximation does not converge to the solution $u(x, t) \equiv 0$ of the continuous problem (1.1), (1.2).

For simplicity we have only formulated the theorems for explicit difference approximations. However, our results hold also for implicit equations. Using the work of G. Strang [3], we get:

THEOREM 5. *Consider an implicit difference approximation*

$$Q_1 v_\nu(t + k) = Q_2 v_\nu(t) , \tag{1.12}$$

where

$$Q_1 = \sum_{j=-p}^{q} b_j E^j , \qquad Q_2 = \sum_{j=-p}^{q} a_j E^j . \tag{1.13}$$

Assume that $\hat{Q}_1(\xi) \neq 0$ for all ξ and that the index condition

$$(1.14)\qquad \int_{-\pi}^{\pi} d \arg \hat{Q}_1(e^{i\xi}) = 0$$

is fulfilled. Then the above theorems hold also for implicit difference approximations (1.12). ($\hat{Q}(\xi) = \hat{Q}_2(\xi)/\hat{Q}_1(\xi)$.)

It is not difficult to generalize the results to equations

$$\partial u/\partial t = d(x, t)\, \partial u/\partial x\,, \qquad d(0, t) \geqq d_0 > 0\,,$$

with variable coefficients, because all arguments used depend on L_2-estimates only. We get using a theorem of P. D. Lax and L. Nirenberg [7]:

THEOREM 6. *All results hold for equations with variable coefficients, provided the coefficients of the differential equation and of the difference approximation belong to C^2 (dissipative, contractive, etc., are defined in the usual way, i.e., pointwise).*

There are essentially two different types of difference approximations which are used in practice: dissipative methods and energy conserving methods. In the last chapter we investigate what properties the energy-conserving methods must have to be contractive.

The reason why we are interested in this problem comes from the following considerations: In applications one often has to determine solutions of hyperbolic differential equations which are only piecewise smooth, i.e., the solutions have contact discontinuities, travelling along the characteristics, and—for nonlinear equations— they have shocks. Thus we get in the x, t-plane discontinuity-lines which we can consider as internal boundaries. Now one often uses difference approximations without doing anything special along these lines of discontinuity. We can view the computation in the following way: When using the difference approximation along a discontinuity-line we in general get completely wrong values. We can consider these values as boundary values for the computation of the solution in those regions where the solution of the differential equation is smooth.

The question then is: What is the influence of the "wrong boundary values" on the solution? In a forthcoming paper by M. Apelkranz [5] precise estimates are given for contact discontinuities by a refinement of our technique. In another paper we shall consider conservation laws $\partial u/\partial t = \partial f(u)/\partial x$ and investigate convergence properties of difference approximations.

2. Contractive Difference Approximations. We start this paragraph by proving Theorem 2, i.e., if Q is dissipative and accurate of order (at least) one, then Q is contractive. By (1.6)

$$(2.1)\qquad \hat{Q}(\xi + i\alpha) = \sum_{j=-p}^{q} a_j \exp\,(ij(\xi + i\alpha))$$

is an analytic function of $z = \xi + i\alpha$. Therefore, there is a constant K such that for all sufficiently small $\alpha > 0$:

$$|\hat{Q}(\xi + i\alpha)| \leqq |\hat{Q}(\xi)| + K\alpha\,.$$

If Q is dissipative, then the last inequality and (1.7) imply

$$|\hat{Q}(\xi + i\alpha)| \leqq 1 - \delta|\xi|^{2s} + K\alpha\,.$$

Therefore, the theorem is proved if we can find real numbers $\xi_1 > 0$ and $\alpha_1 > 0$ such that for all α with $0 \leqq \alpha \leqq \alpha_1$ and all ξ with $|\xi| \leqq \xi_1$ the inequality

$$|\hat{Q}(\xi + i\alpha)| \leqq 1 - \tfrac{1}{2}\alpha\lambda \tag{2.2}$$

holds. By Definition 2 and (2.1) we can write $\hat{Q}(\xi + i\alpha)$ under the form ($R_j(\xi + i\alpha)$ being analytic functions of $z = \xi + i\alpha$):

$$\begin{aligned}\hat{Q}(\xi + i\alpha) &= e^{-\lambda\alpha + i\xi\lambda} + (\xi + i\alpha)^2 R_1(\xi + i\alpha) \\ &= e^{i\xi\lambda} + \xi^2 R_1(\xi) + e^{i\xi\lambda}(e^{-\lambda\alpha} - 1) + (\xi + i\alpha)^2 R_1(\xi + i\alpha) \\ &\quad - \xi^2 R_1(\xi) = \hat{Q}(\xi) - \lambda\alpha + O(\alpha\xi + \alpha^2) .\end{aligned}$$

Now $\hat{Q}(\xi) = 1 + \xi R_2(\xi)$ and by (1.7) $|\hat{Q}(\xi)| \leqq 1$. Therefore

$$|\hat{Q}(\xi + i\alpha)| \leqq 1 - \lambda\alpha + O(\alpha\xi + \alpha^2)$$

and (2.2) follows immediately.

We want to reduce the proof of Theorem 3 to the case where the initial values $f(x) \equiv 0$. Let us assume that $f(x) \in C^{m+1}(0, \infty)$. By extrapolation we can define $f(x)$ for $-ph \leqq x < 0$ in such a way that $f(x) \in C^{m+1}(-ph, \infty)$ and $\max_{-ph \leqq x} |f(x)| \leqq 2 \max_{0 \leqq x} |f(x)|$. Then it is well known that (1.1), (1.2) has the solution

$$u(x, t) = f(x + t) \in C^{m+1}(-ph \leqq x < \infty, t \geqq 0) . \tag{2.3}$$

We consider now the difference approximation

$$\begin{aligned} v_\nu^{(1)}(t + k) &= Qv_\nu^{(1)}(t) , \\ v_\nu^{(1)}(0) &= f(x_\nu) , \end{aligned} \qquad \nu = 0, 1, 2, \cdots , \tag{2.4}$$

with boundary conditions

$$v_\mu^{(1)}(t) = u(x_\mu, t) : \quad x_\mu = \mu h ; \qquad \mu = -1, -2, \cdots , -p . \tag{2.5}$$

Now, G. Strang [3] (see also [4, Lemma 2 and Theorem 2]) has shown:

Lemma 1. *If* $|\hat{Q}(\xi)| \leqq 1$, *then* (2.4), (2.5) *is stable with stability constant* 1, *i.e., if we consider the solution* $w_\nu(t)$ *of*

$$\begin{aligned} w_\nu(t + k) &= Qw_\nu(t) , \\ w_\nu(0) &= y(x_\nu) , \end{aligned} \qquad \nu = 0, 1, 2, \cdots , \tag{2.6}$$

with homogeneous boundary conditions

$$w_\mu(t) = 0 , \qquad \mu = -1, -2, \cdots, -p , \tag{2.7}$$

then

$$||w(t)||_0 \leqq ||w(0)||_0 . \tag{2.8}$$

If Q is contractive, then by definition $|\hat{Q}(\xi)| \leqq 1$. Therefore, we get from Lemma 1 in the usual way the first estimate (1.8), i.e.:

$$||v_\nu^{(1)}(t) - u(x_\nu, t)||_0 \leqq K_1 t h^m . \tag{2.9}$$

We consider now the difference approximation

$$\begin{aligned} v_\nu^{(2)}(t+k) &= Qv_\nu^{(2)}(t)\,, \\ v_\nu^{(2)}(0) &= 0\,, \end{aligned} \qquad \nu = 0, 1, 2, \cdots, \tag{2.10}$$

with boundary conditions

$$v_\mu^{(2)}(t) = g_\mu(t) - u(x_\mu, t)\,, \qquad \mu = -1, -2, \cdots. \tag{2.11}$$

By (2.3) we have

$$|v_\mu^{(2)}(t)| \leqq M + \max_{0 \leqq x < \infty} |f(x)|\,. \tag{2.12}$$

It is obvious that the solution $v(x, t)$ of (1.3), (1.4) can be written in the form

$$v(x, t) = v^{(1)}(x, t) + v^{(2)}(x, t)\,.$$

Therefore, we have proved Theorem 3 (and therefore, by Theorem 2, also Theorem 1), if we can show that for $v^{(2)}(x, t)$ the second estimate (1.8) holds.

Let $y^{(n)}(x_\nu) = y_\nu^{(n)}$ denote the functions ($n \geqq 0$ natural number).

$$\begin{aligned} y_\nu^{(n)} &= 0 \quad \text{for}\, \nu = 0, 1, 2, \cdots, \\ &= v_\nu^{(2)}(t) \quad \text{for}\, \nu = -1, -2, \cdots, -p \,\text{and}\, t = nk\,. \end{aligned} \tag{2.13}$$

Define functions $\omega_\nu^{(n)}(t)$ by

$$\begin{aligned} \omega^{(n)}(t) &\equiv 0 \quad \text{for}\, t = 0, k, \cdots, (n-1)k\,, \\ \omega_\nu^{(n)}(t+k) &= Q\omega_\nu^{(n)}(t)\,, \\ \omega_\nu^{(n)}(nk) &= Qy_\nu^{(n-1)}\,, \end{aligned} \qquad \nu = 0, 1, 2, \cdots, t \geqq nk\,, \tag{2.14}$$

with boundary conditions

$$\omega_\mu^{(n)}(t) \equiv 0 \qquad \text{for}\, \mu = -1, -2, \cdots, -p\,. \tag{2.15}$$

Then we can write the solution of (2.10), (2.11) for $\nu \geqq 0$ in the form:

$$v_\nu^{(2)}(t) = \sum_{j=0}^{t/k-1} \omega_\nu^{(j)}(t)\,, \qquad \nu = 0, 1, 2, \cdots. \tag{2.16}$$

We want to estimate the $\omega_\nu^{(j)}(t)$. By assumption, Q is contractive. Let α, β be positive numbers for which (1.9) holds. Introducing new variables $\tilde{\omega}_\nu^{(n)}(t)$ by

$$\omega_\nu^{(n)}(t) = \exp\left(-\alpha\nu - \beta(t/k - n)\right)\tilde{\omega}_\nu^{(n)}(t)$$

into (2.14), we get

$$\begin{aligned} \tilde{\omega}_\nu^{(n)}(t+k) &= Q_1\tilde{\omega}^{(n)}(t)\,, \\ \tilde{\omega}_\nu^{(n)}(nk) &= e^{\alpha\nu}Qy_\nu^{(n-1)}\,, \end{aligned} \qquad \nu = 0, 1, 2, \cdots, \tag{2.17}$$

with boundary conditions

$$\tilde{\omega}_\mu^{(n)} \equiv 0\,, \qquad \mu = -1, -2, \cdots, -p\,.$$

The Fourier transform of Q_1 has the form $\hat{Q}_1 = e^\beta\hat{Q}(\xi + i\alpha)$, and therefore by (1.9):

$$|\hat{Q}_1| \leqq 1\,. \tag{2.18}$$

Furthermore, we get from (2.13) that $Qy_\nu^{(n-1)} \equiv 0$ for $\nu \geqq p$. Therefore, there exists a constant K such that

$$\|\tilde{\omega}^{(n)}(nk)\|_0^2 \leqq K^2 h e^{2\alpha(p-1)} \sum_{\mu=-p}^{-1} |v_\mu^{(2)}(t)|^2 . \tag{2.19}$$

(2.18) says that we can apply Lemma 1, and (2.19) gives us

$$\|\tilde{\omega}^{(n)}(t)\|_0^2 \leqq K^2 h e^{2\alpha(p-1)} \sum_{\mu=-p}^{-1} |v_\mu^{(2)}(t)|^2 , \qquad t \geqq nk .$$

Therefore,

$$\begin{aligned} |\omega_\nu^{(n)}(t)| &= e^{-\alpha\nu-\beta(t/k-n)}|\tilde{\omega}_\nu^{(n)}(t)| \\ &\leqq e^{-\alpha\nu-\beta(t/k-n)}h^{-1/2}\|\tilde{\omega}^{(n)}(t)\|_0 \\ &\leqq Ke^{\alpha(p-1)}\left(\sum_{\mu=-p}^{-1} |v_\mu^{(2)}(t)|^2\right)^{1/2} \cdot e^{-\alpha\nu-\beta(t/k-n)} \\ &\leqq K'e^{\alpha(p-1)}\left(M + 2\max_x |f(x)|\right)e^{-\alpha\nu-\beta(t/k-n)} , \qquad t \geqq nk . \end{aligned} \tag{2.20}$$

Now (2.16) and (2.20) imply the second estimate (1.8) without difficulty. Therefore, Theorem 3 and Theorem 1 are proved.

The proofs of the first three theorems depend on algebraic manipulations performed with $\hat{Q}(\xi + i\alpha)$ and Lemma 1. For equations with variable coefficients and implicit difference approximations these manipulations can be done in the same way and Lemma 1 is still valid provided the index condition (1.14) is fulfilled. (See Strang [3].) Therefore, the first three theorems are also proved under these circumstances.

3. Noncontractive Difference Approximations. In this section we prove Theorem 4. We assume that the initial values $f(x) \equiv 0$. We consider also implicit equations:

$$\begin{aligned} Q_1 v_\nu(t+k) &= Q_2 v_\nu(t) , \\ v_\nu(0) &= 0 , \end{aligned} \qquad \nu = 0, 1, 2, \cdots , \tag{3.1}$$

where

$$Q_1 = \sum_{j=-p}^{q} b_j E^j , \quad Q_2 = \sum_{j=-p}^{q} a_j E^j , \qquad a_j, b_j \text{ real} .$$

Furthermore, extra boundary conditions are given by

$$v_\mu(t) = g_\mu(t) , \qquad \mu = -1, -2, \cdots, -p . \tag{3.2}$$

As G. Strang [3] has shown, (3.1) has a unique solution $v_\nu(t)$ with $||v_\nu(t)||_0 < \infty$ for every fixed t, if

$$\hat{Q}_1(\xi) \neq 0 \text{ for all (real) } \xi , \quad \text{and} \quad \int_{-\pi}^{\pi} d \arg \hat{Q}_1(\xi) = 0 . \tag{3.3}$$

Furthermore, the approximation is stable if

$$|\hat{Q}(\xi)| = |\hat{Q}_1^{-1}(\xi)\hat{Q}_2(\xi)| \leqq 1 . \tag{3.4}$$

In this section, we always assume that the conditions (3.3) and (3.4) are fulfilled. We return to the proof of Theorem 4. Let ξ_0 be such that (1.10) holds. Obviously, we have in that case $|\hat{Q}(\xi_0)| = 1$.

Let $\hat{Q}(\xi_0) = e^{i\phi}$; then we get in a neighbourhood of $z = \xi_0$:

$$e^{-i\phi}\hat{Q}(z) = 1 + ia_1(z - \xi_0)^{\cdot} + a_2(z - \xi_0)^2 + \cdots . \tag{3.5}$$

Now $|e^{i\phi}\hat{Q}(z)| \leqq 1$ implies that a_1 is real.

Furthermore (1.10) implies

$$a_1 = -\lambda c(\xi_0) \tag{3.6}$$

because

$$|\hat{Q}(\xi_0 + i\alpha)| = |1 - a_1\alpha + O(\alpha^2)| = 1 + \lambda c(\xi_0)\alpha + O(\alpha^2) .$$

Introduce now in (3.1) a new variable $y_\nu(t)$ by

$$v_\nu(t) = \exp\,(i\phi t/k + i\xi_0\nu)y_\nu(t) . \tag{3.7}$$

Then $y_\nu(t)$ is the solution of a similar difference equation

$$R_1 y_\nu(t + k) = R_2 y_\nu(t) , \qquad y_\nu(0) = 0 , \tag{3.8}$$

where R_1, R_2 are difference operators of the same kind as Q_1, Q_2. The boundary conditions get the form:

$$y_\mu(t) = \exp\,(-i\phi t/k - i\xi_0\mu)g_\mu(t) , \qquad \mu = 0, -1, -2, \cdots , -p . \tag{3.9}$$

Now

$$\hat{R}_1(\xi) = \hat{Q}_1(\xi + \xi_0) , \qquad \hat{R}_2(\xi) = e^{-i\phi}\hat{Q}_2(\xi + \xi_0) ,$$

and according to (3.5), (3.6) and (3.9)

$$\hat{R}(\xi) = \hat{R}_1^{-1}(\xi)\hat{R}_2(\xi) = e^{-i\phi}\hat{Q}(\xi + \xi_0) = 1 - ic(\xi_0)\lambda\xi + O(\xi^2) .$$

Therefore (3.8) is a difference approximation to the differential equation

$$\begin{aligned} \partial u/\partial t &= -c(\xi_0)\partial u/\partial x , \\ u(x, 0) &= 0 , \qquad x \geqq 0 . \end{aligned} \tag{3.10}$$

Furthermore, the approximation fulfills the conditions (3.3) and (3.4). If we therefore choose the boundary conditions (3.9) in such a way that $y_\mu(t) \equiv 1$, i.e. $g_\mu(t) = \exp\,(i\phi t/k + i\xi_0\mu)$, then the solution $y_\nu(t)$ of (3.7), (3.8) converges to the solution $u(x, t)$ of the differential equation (3.10) with boundary condition $u(0, t) = 1$.

From (3.7) Theorem 4 follows immediately. Furthermore, by the same argument as in the last section, Theorem 4 holds also for equations with variable coefficients. Therefore, we have also proved Theorems 5 and 6.

We consider now an example: Approximate the differential equation

$$\partial u/\partial t = 2a\,\partial u/\partial x , \qquad a > 0 ,$$

with initial values $f(x) \equiv 0$ by (Crank-Nicolson)

$$(I - k(aD_0 + bh^2D_0D_+D_-))v_\nu(t + k)$$
$$= (I + k(aD_0 + bh^2D_0D_+D_-)v_\nu(t)) \,, \qquad \nu = 0, 1, 2, \cdots \,,$$
$$v_\nu(0) \equiv 0 \,,$$

Here D_0, D_+, D_- are the difference operators

$$2hD_0 = E - E^{-1} \,, \qquad hD_+ = E - I \,, \qquad hD_- = I - E^{-1} \,, \qquad I = E^0 \,.$$

For simplicity, we assume that $b \geqq 0$. The Fourier transform can be written in the form

$$\hat{Q}(\xi) = \frac{1 + i\lambda(a - 4b \sin^2 \xi/2)\sin \xi}{1 - i\lambda(a - 4b \sin^2 \xi/2)\sin \xi} = e^{i\phi} \,.$$

By simple computations we get

$$\hat{Q}(\xi + i\alpha) = e^{i\phi}\left(1 + 2\lambda\alpha \frac{2b \sin^2 \xi - (a - 4b \sin^2 \xi/2)\cos \xi}{1 + \lambda^2\, (a - 4b \sin^2 \xi/2)^2 \sin^2 \xi} + O(\alpha)\right).$$

Therefore $\hat{Q}(\xi + i\alpha)$ fulfills the condition (1.10) for all $\xi = \xi_0$ with $2b \sin^2 \xi_0 - (a - 4b \sin^2 \xi_0/2)\cos \xi_0 > 0$. If we choose the boundary conditions for such $\xi = \xi_0$ to be

$$v_{-\mu}(t) = \exp\,(i\phi t/k - i\xi_0\mu) \,,$$

then the solution of (3.11) behaves as described in Theorem 4. If we are especially interested in boundary conditions independent of t (i.e. $\phi = 0$), we have to distinguish between two cases. Denoting by $g(x)$ the function $g(x) = 0$ for $x \leqq 0$, $g(x) = 1$ for $x > 0$ we get

(1) $a - 4b > 0$. Then $\phi = 0$ for $\xi = \pi$ and $v_\nu(t) \simeq (-1)^\nu g(2(a - 4b)t - x_\nu)$.

(2) $a - 4b < 0$. Then $\phi = 0$ for $\xi = \xi_0$ with $a = 4b \sin^2 \xi_0/2$ and $v_\nu(t) \simeq \exp\,(-i\xi_0\nu)g(4bt \sin^2 \xi_0 - x_\nu)$.

Furthermore, for increasing b, the value of ξ_0 becomes smaller and smaller, which means that the "wave length" of $v_\nu(t)$ becomes larger and larger. While in the first case it is easy to detect that $v_\nu(t)$ is a mere numerical effect, it is much more difficult to see this in the second case. It can only be detected by halving the step-size and doing the computation twice.

4. Energy Conserving Methods. There are essentially two different types of difference approximations which are used in practice: (1) the dissipative methods which we have discussed in Section 2; and (2) energy-conserving methods. The latter are defined by

Definition 5. The difference approximation (3.1) is called energy-conserving if the condition (3.3) is fulfilled and for all (real) ξ

$$(4.1) \qquad |\hat{Q}(\xi)| = |\hat{Q}_1^{-1}(\xi)\hat{Q}_2(\xi)| = 1 \,.$$

We want to investigate what properties the energy-conserving methods must have to be contractive. We start with some lemmata:

LEMMA 3. *For all energy-conserving methods $\hat{Q}(\xi)$ can be written in the form*:

$$(4.2) \qquad \hat{Q}(\xi) = \frac{s(z)}{r(z)} = \frac{c_0 z^n + c_1 z^{n-1} + \cdots + c_n}{c_n z^n + c_{n-1} z^{n-1} + \cdots + c_0} = f(z) \; ;$$

$z = e^{i\xi}, n = p + q$, i.e., $s(z) = z^n \cdot r(z^{-1})$.

We have in particular that an explicit difference approximation ($\hat{Q} \equiv 1$) is energy-conserving if and only if $\hat{Q}(\xi) = e^{in\xi}$.

Proof. From (3.3) we get that the function ($z = e^{i\xi}$)

$$(4.3) \qquad \frac{\hat{Q}_2(\xi)}{\hat{Q}_1(\xi)} = \frac{\sum_{j=-p}^{q} a_j z^j}{\sum_{j=-p}^{q} b_j z^j} = g(z)$$

is analytic in a neighbourhood of $|z| = 1$ and that $|g(z)| = 1$ for $|z| = 1$. Therefore, using analytic continuation, the relation

$$(4.4) \qquad 1/g(z) = g(1/z)$$

must hold. Introducing into (4.4) the expression (4.3) and identifying, we get (4.2).

LEMMA 4. *$f(z)$ can be written in the form*

$$(4.5) \qquad f(z) = \prod_{j=1}^{n} \frac{1 - \alpha_j}{1 - \bar{\alpha}_j} \frac{1 - \bar{\alpha}_j z}{z - \alpha_j}$$

where α_j are the zeros of $r(z)$.

Proof. If α_j is a zero of $r(z)$ then $\bar{\alpha}_j^{-1}$ is a zero of $s(z)$. (Observe that the a_j are real and therefore if α_j is a zero of $r(z)$, then $\bar{\alpha}_j$ is also a zero of $r(z)$.) Therefore

$$f(z) = \text{const} \prod_{j=1}^{n} \frac{1 - \bar{\alpha}_j z}{z - \alpha_j}$$

Observing that $f(1) = 1$ ($\hat{Q}(0) = 1$), we get (4.5).

LEMMA 5. *If m is the number of those zeros α_j of $r(z)$ with $|\alpha_j| < 1$, then the index condition is fulfilled if and only if $p = m$, i.e., in* (3.1) *the Q_j have the form*:

$$Q_1 = E^{-m} \sum_{j=0}^{n} c_{n-j} E^j \, , \qquad Q_2 = E^{-m} \sum_{j=0}^{n} c_j E^j \, .$$

Proof. $\hat{Q}_1(\xi) = e^{-mi\xi} r(e^{i\xi})$. Therefore,

$$\int_{-\pi}^{+\pi} d \arg \hat{Q}_1(\xi) = -2\pi m + \int_{-\pi}^{\pi} d \arg r(e^{i\xi}) = -2\pi m + 2\pi m = 0 \, .$$

When constructing difference approximations one is, in general, not interested in methods which just work for one differential equation. We require that the approximation (3.1) should work for all differential equations

$$(4.6) \qquad \partial u/\partial t = c \, \partial u/\partial x \, , \qquad 0 \leqq c \leqq M \, , \qquad M > 0 \text{ some constant} \, .$$

We assume that the coefficients a_j of our difference approximations are polynomials in c such that we can use the method for systems of differential equations. It is natural to make the following assumptions:

$$(4.7) \quad \begin{array}{ll} (1) & \hat{Q}_1 = \hat{Q}_1(\xi, c) \neq 0 \quad \text{for all (real) } \xi \text{ and all } c \, , \\ (2) & \hat{Q}_1(\xi, c)/\hat{Q}_1(\xi, c) = 1 \quad \text{for all } \xi \text{ and } c = 0 \, . \end{array}$$

LEMMA 6. *Under the conditions* (4.7), *we have* $p = q = m = n/2$ (*m is defined in Lemma* 5 *and p, q are defined in* (3.1)).

Proof. The number of zeros α_j of $r(z) = r(z, c)$ with $|\alpha_j| < 1$ must be independent of c, because otherwise $\hat{Q}_1(\xi, c) = 0$ for some c. For $c = 0$ we have by (4.7) that $f(z) \equiv 1$, (see (4.5)). This is possible only if with α_j also $\bar{\alpha}_j^{-1}$ is a zero. Therefore, the number m of α_j with $|\alpha_j| < 1$ is equal to the number of α_j with $|\alpha_j| > 1$; i.e., $n = 2m$ and by Lemma 5, $p = q = m = n/2$.

Now we can show

THEOREM 5. *All energy-conserving methods that approximate* (1.1) *with accuracy at least one and for which the number m of roots α_j with $|\alpha_j| < 1$ of $r(z) = 0$ fulfill the inequality $m \geqq n/2$ are strictly noncontractive, i.e., there is a ξ_0 such that* (1.10) *holds. Specifically, all methods for which* (4.7) *holds are strictly noncontractive.*

Proof. Observing that $s(e^{i\xi}) = e^{in\xi}r(e^{-i\xi})$, we get:

$$\begin{aligned}\hat{Q}(\xi + i\alpha) &= \frac{s(e^{i\xi-\alpha})}{r(e^{i\xi-\alpha})} = \frac{s(e^{i\xi}) + i\alpha\partial s(e^{i\xi})/\partial\xi}{r(e^{i\xi}) + i\alpha\partial r(e^{i\xi})/\partial\xi} + O(\alpha^2)\\ &= \frac{s(e^{i\xi})}{r(e^{i\xi})}\left(1 + i\alpha\left(\frac{\partial s/\partial\xi}{s} - \frac{\partial r/\partial\xi}{r}\right)\right) + O(\alpha^2)\\ &= \frac{s(e^{i\xi})}{r(e^{i\xi})}\left(1 - n\alpha - i\alpha\left(\frac{\partial r(e^{-i\xi})/\partial\xi}{r(e^{-i\xi})} + \frac{\partial r(e^{i\xi})/\partial\xi}{r(e^{i\xi})}\right)\right) + O(\alpha^2)\\ &= \frac{s(e^{i\xi})}{r(e^{i\xi})}\left(1 - n\alpha + 2\alpha \operatorname{Im}\frac{\partial}{\partial\xi}\log r(e^{i\xi})\right) + O(\alpha^2)\,.\end{aligned}$$

If the method is not strictly noncontractive, i.e., there is no ξ_0 such that (1.10) holds, then

$$I = n\alpha - 2\alpha \operatorname{Im}\{(\partial/\partial\xi)\log r(e^{i\xi})\} \geqq 0$$

and there is a neighbourhood of $\xi = 0$ where $I > 0$.† Therefore,

$$\int_{-\pi}^{+\pi} n\alpha d\xi - 2\alpha\int_{-\pi}^{+\pi} d\arg r(e^{i\xi}) = 2\pi\alpha(n - 2m) > 0\,,$$

i.e., $m < n/2$, which proves the theorem.

There are methods which are energy-conserving and contractive. For example, all methods

$$\sum_{j=0}^{n} a_jE^jv_\nu(t + k) = \sum_{j=0}^{n} a_{n-j}E^jv_\nu(t) \tag{4.8}$$

with $m = 0$ are of this type because no extra boundary conditions are needed. However, none of these methods is useful if one wants to integrate equations (4.6) for small c, because either they do not reduce to unity for $c = 0$ or $Q_1(\xi, 0)$ has a root on the unit circle. Furthermore, they are useless for differential equations (4.6) with $c < 0$ because the index condition cannot be fulfilled. If one, therefore, wants to use (4.8) for systems of differential equations $\partial u/\partial t = A\ \partial u/\partial x$, where the eigenvalues of A have different signs, then one has to transform u in such a way

† In a neighbourhood of $\xi = 0$, $\hat{Q}(\xi) = e^{i\lambda\xi} + O(\xi^2)$, i.e., $\hat{Q}(\xi + i\alpha) = e^{i\lambda\xi-\alpha} + O((\xi + i\alpha)^2)$.

that the new A is in diagonal form and one must use different formulas for different components. Methods of this type have been considered in [6].

Department of Computer Sciences
Uppsala University
Sturegatan 4 Sweden

1. S. V. Parter, "Stability, convergence and pseudo-stability of finite-difference equations for an over-determined problem," *Numer. Math.*, v. 4, 1962, pp. 277–292. MR **26** #5740.

2. P. D. Lax, "On the stability of difference approximations to solutions of hyperbolic equations with variable coefficients," *Comm. Pure Appl. Math.*, v. 14, 1961, pp. 497–520. MR **26** #3215.

3. W. G. Strang, "Wiener-Hopf difference equations," *J. Math. and Mech.*, v. 13, 1964, pp. 85–96. MR **28** #3548.

4. H.-O. Kreiss, "Difference approximations for the initial-boundary value problem for hyperbolic differential equations," *Numerical Solutions of Nonlinear Differential Equations*, Proceedings of a Symposium at the University of Wisconsin (May 1966), edited by Donald Greenspan, Wiley, New York, 1966.

5. M. Y. T. Apelkranz, "On difference schemes for hyperbolic equations with discontinuous initial values." (To appear.)

6. 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.

7. P. D. Lax & L. Nirenberg, "On stability for difference schemes; a sharp form of Gårdings inequality," *Comm. Pure Appl. Math.*, v. 19, 1966, pp. 473–492.