On the Propagation of Round-Off Errors in the Numerical Treatment of the Wave Equation

By Arnold N. Lowan

Abstract. An upper bound of the norm of the error vector after n time steps is $\frac{1}{2}(n + 1)(n + 2) \| \boldsymbol{\delta}^* \|$. For the explicit scheme $\delta^* = \| \boldsymbol{\delta}^* \| = 3 \times \frac{1}{2} \times 10^{-p}$ where p is the number of decimals carried in the computations. For the implicit scheme $\delta^* = \| \boldsymbol{\delta}^* \|$ is an upper bound of the errors which arise both from using approximations to A^{-1} and $A^{-1}B$ in the determination of $\mathbf{u}_{k+1}$ from equation (6*) and from rounding off the values of the products and quotients involved in the computation of the components of $\mathbf{u}_{k+1}$.

Consider the numerical treatment of the differential equation of wave motion

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \qquad 0 \leqq x \leqq a, \quad t > 0 \tag{1}$$

the solution of which is required to satisfy the following initial and boundary conditions

$$u(x, 0) = f(x) \tag{2}$$

$$u_t(x, 0) = g(x) \tag{3}$$

$$u(0, t) = u(a, t) = 0. \tag{4}$$

With the differential equation (1) we will associate either of the following two difference analogs [1]

$$u_{h,k+1} - 2u_{h,k} + u_{h,k-1} = R^2(u_{h-1,k} - 2u_{h,k} + u_{h+1,k}) \tag{5}$$

$$\begin{aligned} u_{h,k+1} - 2u_{h,k} + u_{h,k-1} = \frac{R^2}{2} (u_{h-1,k+1} &- 2u_{h,k+1} + u_{h+1,k+1} \\ &+ u_{h-1,k-1} - 2u_{h,k-1} + u_{h+1,k-1}) \end{aligned} \tag{6}$$

where $R = c\Delta t/\Delta x$ and $u_{h,k} = u(h\Delta x, k\Delta t)$ with $(M + 1)\Delta x = a$.

The difference counterpart of (3) will be taken in the form

$$\frac{u_{h,1} - u_{h,0}}{\Delta t} = g(h\Delta x);$$

whence

$$u_{h,1} = u_{h,0} + g(h\Delta x)\Delta t = f(h\Delta x) + g(h\Delta x)\Delta t. \tag{7}$$

The difference equations (5) and (6) may be written in the compact forms

$$\mathbf{u}_{k+1} = A\mathbf{u}_k - \mathbf{u}_{k-1} \tag{5*}$$

$$A\mathbf{u}_{k+1} = 4\mathbf{u}_k + B\mathbf{u}_{k-1}. \tag{6*}$$

Received August 17, 1959; in revised form, December 24, 1959.

In (5*) A is a tridiagonal matrix whose elements on the principal diagonal are $= 2(1 - R^2)$ and whose elements off the principal diagonal are $= R^2$ and $\mathbf{u}_k$ is the vector whose components are the values of $u(x, t)$ at time $t = k\Delta t$ at the lattice points $x = h\Delta x, h = 1, 2, 3 \cdots$. In (6*) A is a tridiagonal matrix whose elements on the principal diagonal are $= 2(1 + R^2)$ while the elements off the principal diagonal are $= -R^2$ and B is a tridiagonal matrix whose elements on the principal diagonal are $= -2(1 + R^2)$ while the elements off the principal diagonal are $= R^2$.

Consider first the explicit difference scheme (5*). Since both $\mathbf{u}_0$ and $\mathbf{u}_1$ are known, (5*) will yield in succession $\mathbf{u}_2$, $\mathbf{u}_3 \cdots$. Specifically,

$$\text{(8)} \qquad \begin{cases} \mathbf{u}_2 = A\mathbf{u}_1 - \mathbf{u}_0 \\ \mathbf{u}_3 = A\mathbf{u}_2 - \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n = A\mathbf{u}_{n-1} - \mathbf{u}_{n-2} . \end{cases}$$

It is reasonable to assume that the components of $\mathbf{u}_0$ are exact while those of $\mathbf{u}_1$, obtained from (7), have been rounded off to the number of decimal places to be carried in the computations. Let $\mathbf{u}_1^*$ denote the vector whose components are the rounded off values of the components of $\mathbf{u}_1$. It is then easily seen that we introduce two types of errors in the evaluation of $\mathbf{u}_2$. A first error is due to using $\mathbf{u}_1^*$ in lieu of $\mathbf{u}_1$. A second error is introduced as a result of rounding off of the values of the products involved in the expression of $u_{h,k+1}$ obtained from (5) to the number of decimal places carried in the computations. Thus, in lieu of the exact vector $\mathbf{u}_2$, the first step in the sequence of operations (8) yields the vector $\mathbf{u}_2^* = A\mathbf{u}_1^* - \mathbf{u}_0 + \boldsymbol{\delta}_2$ where $\boldsymbol{\delta}_2$ is the error vector whose components are the round-off errors just discussed. Similarly, error vectors are introduced in each of the successive steps in the sequence of operations (8). Thus

$$\text{(9)} \qquad \begin{cases} \mathbf{u}_2^* = A\mathbf{u}_1^* - \mathbf{u}_0 + \boldsymbol{\delta}_2 \\ \mathbf{u}_3^* = A\mathbf{u}_2^* - \mathbf{u}_1^* + \boldsymbol{\delta}_3 \\ \vdots \\ \mathbf{u}_n^* = A\overset{*}{\mathbf{u}}_{n-1} - \overset{*}{\mathbf{u}}_{n-2} + \boldsymbol{\delta}_n . \end{cases}$$

If we put

$$\text{(10)} \qquad \mathbf{E}_n = \mathbf{u}_n^* - \mathbf{u}_n$$

then from (8) and (9) it follows that

$$\text{(11)} \qquad \mathbf{E}_n = A\mathbf{E}_{n-1} - \mathbf{E}_{n-2} + \boldsymbol{\delta}_n .$$

In entirely similar manner it may be shown that the counterpart of (11) for the implicit scheme (6*) is

$$\text{(12)} \qquad \mathbf{E}_n = 4A^{-1}\mathbf{E}_{n-1} + A^{-1}B\mathbf{E}_{n-2} + \boldsymbol{\delta}_n .$$

There is, however an important distinction between (11) and (12); whereas in (11) the components of $\boldsymbol{\delta}_n$ are round-off errors as above explained, in (12) the components of $\boldsymbol{\delta}_n$ are the aggregate of the errors arising both from using approximations to A^{-1} and $A^{-1}B$ in the determination of $\mathbf{u}_{k+1}$ and the round-off errors introduced as a result of rounding-off the values of the products and quotients involved in the computation of the components of $\mathbf{u}_{k+1}$.

The error equations (11) and (12) are of the form

$$\mathbf{E}_n = M\mathbf{E}_{n-1} + N\mathbf{E}_{n-2} + \boldsymbol{\delta}_n . \tag{13}$$

If in (13) we put in succession $n = 2, 3, 4, \cdots$ and write $\boldsymbol{\delta}_1$ for $\mathbf{E}_1$, it may be shown by induction that

$$\mathbf{E}_n = P_{n-1}(M, N)\boldsymbol{\delta}_1 + P_{n-2}(M, N)\boldsymbol{\delta}_2 + \cdots \boldsymbol{\delta}_n \tag{14}$$

or

$$\mathbf{E}_n = \sum_{p=0}^{n} P_p(M, N)\boldsymbol{\delta}_{n-p} \tag{14*}$$

where

$$P_n(M, N) = M^n + C_{n-1}^1 M^{n-2}N + C_{n-2}^2 M^{n-4}N^2 + \cdots C_{n-s}^s M^{n-s}N^s + \cdots \tag{15}$$

or

$$P_n(M, N) = \sum_{s=0}^{(n/2)} C_{n-s}^s M^{n-2s}N^s \tag{15*}$$

where $(n/2)$ denotes the largest integer in $n/2$, where $C_n{}^0 = 1$ and $C_m{}^n$ denote the binomial coefficient $m(m-1)(m-2)\cdots(m-n+1)/n!$.

We shall prove that if M and N have the same eigenvectors, then

$$\| P_p(M, N)\boldsymbol{\delta}_{n-p} \| \leqq \| \boldsymbol{\delta}_{n-p} \| \cdot (p+1) \tag{16}$$

where for any M-dimensional vector $\boldsymbol{\phi}$, its norm $\| \boldsymbol{\phi} \|$ is defined by

$$\| \boldsymbol{\phi} \| = \sqrt{(\boldsymbol{\phi}, \boldsymbol{\phi})} = \sqrt{\frac{1}{M}\sum_{h=1}^{M} (\phi_h)^2} \tag{17}$$

the ϕ_h's being the components of $\boldsymbol{\phi}$,
provided that the roots of the quadratic equation

$$x^2 - \lambda_r x - \mu_r = 0 \tag{18}$$

where the λ_r's and μ_r's, the eigenvalues of M and N respectively, are either numerically equal to or smaller than unity (if real) or have a modulus equal to or smaller than unity (if complex). Indeed, let

$$\boldsymbol{\delta}_{n-p} = \sum_{r=1}^{M} \alpha_r^{(n-p)}\mathbf{w}_r \tag{19}$$

where the $\mathbf{w}_r$'s are the normalized eigenvectors of the matrices M and N. From (19) and (14*) we get

$$P_p(M, N)\boldsymbol{\delta}_{n-p} = \sum_{s=0}^{(p/2)} \sum_{r=1}^{M} C_{p-s}^s\, \alpha_r^{(n-p)} M^{p-2s}N^s\mathbf{w}_r . \tag{20}$$

But

$$M^{p-2s}N^s\, \mathbf{w}_r = M^{p-2s}\mu_r{}^s\mathbf{w}_r = \lambda_r^{p-2s}\mu_r{}^s\mathbf{w}_r\ ;$$

whence

$$\begin{aligned} P_p(M,N)\boldsymbol{\delta}_{n-p} &= \sum_{r=1}^{M} \alpha^{(n-p)}\mathbf{w}_r \sum_{s=0}^{(p/2)} C_{p-s}^{s}\,\lambda_r^{p-2s}\mu_r{}^s \\ &= \sum_{r=1}^{M} \beta_r(p)\alpha_r^{(n-p)}\mathbf{w}_r \qquad \text{(say)}. \end{aligned} \tag{21}$$

It may be proved by induction that

$$\beta_r(p) = \sum_{s=0}^{(p/2)} C_{p-s}^{s}\lambda_r^{p-2s}\mu_r{}^s = \frac{x_{1,r}^{p+1} - x_{2,r}^{p+1}}{x_{1,r} - x_{2,r}} = \sum_{\sigma=0}^{p} x_{1,r}^{\sigma}\, x_{2,r}^{p-\sigma} \tag{22}$$

where $x_{1,r}$ and $x_{2,r}$ are the roots of the quadratic equation (18). From (22) it is clear that if these roots are numerically smaller than unity then

$$|\beta_r(p)| < p + 1; \tag{23}$$

and furthermore $\beta_r(p) \to 0$ as $p \to \infty$. In view of (23), (21) yields

$$\| P_p(M,N)\boldsymbol{\delta}_{n-p} \| = \sqrt{\sum_{r=1}^{M} [\beta_r(p)]^2[\alpha_r^{(n-p)}]^2} \leqq (p+1)\sqrt{\sum_{r=1}^{M} [\alpha_r^{(n-p)}]^2}\ ;$$

or

$$\| P_p(M,N)\boldsymbol{\delta}_{n-p} \| \leqq (p+1)\| \boldsymbol{\delta}_{n-p} \|. \tag{16}$$

From (14*) the Minkowski inequality yields

$$\| \mathbf{E}_n \| \leqq \sum_{p=0}^{n} \| P_p(M,N)\boldsymbol{\delta}_{n-p} \|, \tag{24}$$

whence, in view of (16)

$$\| \mathbf{E}_n \| \leqq \sum_{p=0}^{n} (p+1)\| \boldsymbol{\delta}_{n-p} \|\ ; \tag{25}$$

and a fortiori

$$\| \mathbf{E}_n \| \leqq \| \boldsymbol{\delta}^* \| \sum_{p=0}^{n} (p+1) = \frac{(n+1)(n+2)}{2}\| \boldsymbol{\delta}^* \|, \tag{25*}$$

where $\| \delta^* \|$ is the largest of the sequence $\| \boldsymbol{\delta}_1 \|, \| \boldsymbol{\delta}_2 \| \cdots \| \boldsymbol{\delta}_n \|$. If δ^* denotes an upper bound of the components of all the vectors $\boldsymbol{\delta}_p$, it is readily seen that

$$\| \boldsymbol{\delta}^* \| \leqq \delta^*.$$

Furthermore, since

$$\| \mathbf{E}_n \| = \left\{ \frac{1}{M} \sum_{h=1}^{M} (E_{nh})^2 \right\}^{1/2}$$

where the E_{nh}'s are the components of $\mathbf{E}_n$, it is clear that the maximum of any of the components is obtained by assuming that all but one of the components are $= 0$.

Calling the maximum value of the components E_n^* we finally get

$$E_n^* \leqq \tfrac{1}{2}(n + 1)(n + 2)\sqrt{M}\delta^*. \tag{26}$$

The second member of (26) is an upper bound of the round-off errors for both the explicit analog (5) and the implicit analog (6).

In the case of the explicit scheme (5) the matrix M of (13) is the matrix A appropriate to (5) while the matrix N of (13) is $= -I$ where I is the $M \times M$ identity matrix. The eigenvalues of A are known [2] to be

$$\lambda_r = 2 - 4R^2 \cos \frac{r\pi}{2(M + 1)} \tag{27}$$

The eigenvalues of $-I$ are clearly $= -1$. Thus the quadratic equation (18) becomes

$$x^2 - \left[2 - 4R^2 \cos \frac{r\pi}{2(M + 1)}\right] x + 1 = 0. \tag{28}$$

It is clear that if the roots of (28) were real, one would have to be larger than unity, since the products of the roots is $= 1$. Under these conditions $\beta_r(p)$ as defined in (22) would not be bounded as $p \rightarrow \infty$ and the difference scheme (5) could not be stable. Thus the roots of (28) must be complex, in which case the modulus of the roots is $= 1$ and $\beta_r(p) \leqq p + 1$.

An upper bound of the round-off errors after n time steps is then given by

$$E_n^* = \tfrac{1}{2}(n + 1)(n + 2)\sqrt{M}\delta^*$$

where $\delta^* = 3 \times \frac{1}{2} \times 10^{-p}$ if the computations are carried to p decimal places. In the case of the implicit scheme (6) matrices M and N of (13) are A^{-1} and $A^{-1}B$ respectively where the matrices A and B appropriate to (6) have been defined earlier.

It can be easily shown that the matrices A^{-1} and $A^{-1}B$ have the same eigenvectors, as required in the above developments [2, p. 20], and that their eigenvalues are

$$\lambda_r = 2/\left(1 + 2R^2 \cos^2 \frac{r\pi}{2(M + 1)}\right); \qquad \mu_r = -1 \tag{29}$$

Thus the quadratic equation (18) becomes

$$x^2 - \frac{2}{1 + 2R^2 \cos^2 \dfrac{r\pi}{2(M + 1)}} x + 1 = 0. \tag{30}$$

Clearly the roots of (30) must be complex. This leads to the condition

$$1/\{1 + 2R^2 \cos [r\pi/2(M + 1)]\} < 1$$

which is evidently satisfied for any value of R. Thus the difference scheme (6) is unconditionally stable. Furthermore, and for the same reason as above,

$$\beta_r(p) \leqq p + 1.$$

An upper bound of the round-off errors after n time steps is, therefore, once more given by

$$E^* \leqq \tfrac{1}{2}(n + 1)(n + 2)\sqrt{M}\delta^*. \tag{26*}$$

In this case, however, as previously mentioned δ^* is an upper bound of the errors which arise both from the use of approximations to A^{-1} and $A^{-1}B$ in lieu of exact matrices and from the process of rounding-off the values of the products and quotients involved in the evaluation of the components of $\mathbf{u}_{k+1}$. Clearly δ^* depends on the specific scheme for solving the system of equations (6) with $h = 1, 2, 3, \cdots M$ for the $u_{h,k+1}$'s.

In order to estimate δ^* for the implicit scheme we note that the counterpart of the typical equation (9) is

$$\mathbf{u}_k^* = 4A^{-1}\mathbf{u}_{k-1}^* + A^{-1}B\mathbf{u}_{k-2}^* + \boldsymbol{\delta}_k$$

whence

$$A\mathbf{u}_k^* = 4\mathbf{u}_{k-1}^* + B\mathbf{u}_{k-2}^* + A\boldsymbol{\delta}_k .$$

Let $\mathbf{R}_k$ denote the known vectors $A\mathbf{u}_k^* - 4\mathbf{u}_{k-1}^* - B\mathbf{u}_{k-2}^*$. Then $A\boldsymbol{\delta}_k = \mathbf{R}_k$ and therefore $\boldsymbol{\delta}_k = A^{-1}\mathbf{R}_k$. Since the eigenvalues of A are known to be larger than 2, it follows that the eigenvalues of A^{-1} are smaller than unity and therefore

$$\| \boldsymbol{\delta}_k \| = \| A^{-1}\mathbf{R}_k \| \leqq \| \mathbf{R}_k \| .$$

We conclude that δ^* in equation (26*) is the largest of the norms of the n "residual vectors" $\mathbf{R}_k = A\mathbf{u}_k^* - 4\mathbf{u}_{k-1}^* - B\mathbf{u}_{k-2}^*$. These vectors will depend, of course, on the specific method of computing the $\mathbf{u}_{k+1}$'s from (6).

A discussion of two alternative schemes for solving implicit systems of equations of the type (6) is contained in [3].

Yeshiva University
New York, New York; and
AVCO Corporation
Wilmington, Massachusetts

1. R. D. Richtmeyer, *Difference Methods for Initial-Value Problems*, Interscience Publishers, Inc., New York, 1957.
2. A. N. Lowan, *The Operator Approach to Problems of Stability and Convergence*, Scripta Mathematica, Yeshiva University, New York, 1957, p. 55, p. 82–86.
3. A. N. Lowan, "On the propagation of round-off errors in the numerical integration of the heat equation," *Math. Comp.*, v. 14, 1960, p. 139–146.