A stable explicit method for the finite-difference solution of a fourth-order parabolic partial differential equation

By D. J. Evans*

Adaptation of the Du Fort–Frankel explicit scheme to a fourth order linear parabolic equation is shown to possess unrestricted stability for any choice of mesh spacings. Convergence and compatibility considerations are discussed and the theoretical results confirmed by numerical experiments for a chosen model problem.

1. Introduction

The non-dimensional equation

$$\frac{\partial^4 y}{\partial x^4} + \frac{\partial^2 y}{\partial t^2} = 0 \quad 0 \leqslant x \leqslant 1, t > 0 \tag{1a}$$

subject to the prescribed initial conditions

$$\left.\begin{aligned} y(x, 0) &= g_0(x) \\ \text{and} \quad \frac{\partial y}{\partial t}(x, 0) &= g_1(x) \end{aligned}\right\} \text{ for } 0 \leqslant x \leqslant 1, \tag{1b}$$

and with boundary conditions at $x = 0$ and 1 of the form

$$\begin{gathered} y(0, t) = f_0(t), \qquad y(1, t) = f_1(t), \quad \text{and} \\ \frac{\partial^2 y}{\partial x^2}(0, t) = p_0(t), \; \frac{\partial^2 y}{\partial x^2}(1, t) = p_1(t). \end{gathered} \tag{1c}$$

occurs in the study of the transverse vibrations of a uniform flexible beam of unit length hinged at both ends. Here, y represents the transverse displacement of the beam and x and t the distance and time variables, respectively.

To obtain numerical solutions of (1), the usual procedure is to cover the specified domain by a rectangular network with spacing Δx and Δt, and to replace the differential equation by a finite-difference approximation which is to be evaluated at each point of the mesh. Both explicit and implicit methods have been successfully proposed by Collatz (1951), Crandall (1954), Conte and Royster (1956), Conte (1957), and Albrecht (1957).

The explicit method given by Collatz possesses great simplicity but has the disadvantage that the mesh ratio $\Delta t/(\Delta x)^2$ must be less than or equal to $\frac{1}{2}$ to ensure stability against growth of round-off error. This results in a very large number of time steps that need to be computed for even the most modest problem. The formula given by Albrecht is explicit and overcomes the stability problem but uses the values of the solution on four lines to compute the solution on a fifth line, which is slightly disadvantageous. On the other hand, the implicit methods have superior stability properties but suffer from the disadvantage that they involve the solution of a system of linear equations at each time step. Although each equation of this system involves only 5 or 3 unknowns, the application of these methods in general requires more work than the explicit methods.

The purpose of this paper is to introduce an alternative explicit method for the solution of (1) derived from the Du Fort–Frankel approach to second-order parabolic equations (Du Fort and Frankel, 1953) which possesses unrestricted stability whilst preserving the ease and simplicity which one associates with explicit methods.

2. A new finite difference scheme

By the introduction of two additional variables Φ and Ψ defined by the following equations

$$\frac{\partial y}{\partial t} = \Phi \text{ and } \frac{\partial^2 y}{\partial x^2} = \Psi \tag{2}$$

then we can rewrite equation (1) as two simultaneous partial differential equations of the form

$$\frac{\partial \Phi}{\partial t} = \frac{-\partial^2 \Psi}{\partial x^2} \tag{3a}$$

and

$$\frac{\partial \Psi}{\partial t} = \frac{\partial^2 \Phi}{\partial x^2} \tag{3b}$$

after Richtmyer (1957).

The Φ and Ψ domains are covered by rectangular networks with spacing Δx and Δt, and we distinguish the discrete variables $\phi(i\Delta x, j\Delta t)$ and $\psi(i\Delta x, j\Delta t)$ at the point (i, j) on each network from the continuous functions $\Phi(x, t)$ and $\Psi(x, t)$ specified by equations (3).

Now, if Φ and Ψ are functions defined as in (2), then they certainly satisfy (3a) and (3b) if y satisfies (1). Also, it can be noticed that the differential equations (3) are invariant under the transformation $-x$ for x. We shall use this fact for various edge conditions. We shall use the given initial and boundary conditions in (1), to derive the initial values on the ψ network in the following manner. From the initial condition $y(x, 0)$ given by (1b) $\frac{\partial^2 y}{\partial x^2}$ (or ψ) is calculated. Also since $\frac{\partial y}{\partial t}$

[*] *University Computing Laboratory, University of Sheffield, Sheffield, Yorks.*

Downloaded from https://academic.oup.com/comjnl/article/8/3/280/336153 by guest on 19 April 2024

(or ϕ) is prescribed in (1*b*), values of $\dfrac{\partial^2 \phi}{\partial x^2}$ are similarly obtained and immediately from (3*b*), values of $\dfrac{\partial \psi}{\partial t}$ are known on the line $t = 0$ which together with the previous knowledge of ψ on $t = 0$ is sufficient for us to determine the ψ values on the second line $t = \Delta t$. Similarly, on the ϕ network, values of ϕ are known on line $t = 0$ for $\dfrac{\partial y}{\partial t}$ (or ϕ) is prescribed initially by (1*b*) and, from a knowledge of ψ, $\dfrac{\partial^2 \psi}{\partial x^2}$ is obtained which by (3*a*) gives the values of $\dfrac{\partial \phi}{\partial t}$ on line $t = 0$. Hence, values of ϕ can be obtained on both lines $t = 0$ and $t = \Delta t$.

Finally, since ψ is specified at $x = 0$ and 1 for $t > 0$, and from (3*a*) $\dfrac{\partial \phi}{\partial t}$ is specified there also, it follows immediately that ϕ and ψ are specified along $x = 0$ and 1 for $t > 0$. Hence the values of ϕ and ψ for the first two lines $t = 0$ and $t = \Delta t$ of our networks and at $x = 0$ and 1 for all $t > 0$ can be obtained from equations (1*b*) and (1*c*); we are now concerned with the problem of proceeding with our solution for increasing t within the domain $0 < x < 1$.

We now apply directly the Du Fort–Frankel scheme to the ϕ and ψ networks simultaneously. This scheme is based on the four-point stencil on each network as indicated in **Fig. 1** and involves the use of the central-difference expression

$$[\phi(x, t + \Delta t) - \phi(x, t - \Delta t)](2\Delta t)^{-1}$$

to replace the left-hand side of equation (3*a*). The second derivative with respect to x is replaced by the approximate expression

$$[\psi(x + \Delta x, t) - \psi(x, t + \Delta t) - \psi(x, t - \Delta t) + \psi(x - \Delta x, t)](\Delta x^{-2})$$

with similar approximations for equation (3*b*).

If we let $\phi_{i,j} \equiv \phi(i\Delta x, j\Delta t)$, the finite-difference equations at the point (i, j) on each mesh can be represented as follows:

$$\frac{[\phi_{i,j+1} - \phi_{i,j-1}]}{2k} = -\frac{[\psi_{i+1,j} + \psi_{i-1,j} - \psi_{i,j-1} - \psi_{i,j+1}]}{h^2} \qquad (4a)$$

and

$$\frac{[\psi_{i,j+1} - \psi_{i,j-1}]}{2k} = \frac{[\phi_{i+1,j} + \phi_{i-1,j} - \phi_{i,j-1} - \phi_{i,j+1}]}{h^2} \qquad (4b)$$

where $h = \Delta x$ and $k = \Delta t$.

These equations can further be simplified by solving for the two unknown quantities $\psi_{i,j+1}$ and $\phi_{i,j+1}$ in (4*a*) and (4*b*) and rearranging to give the final result

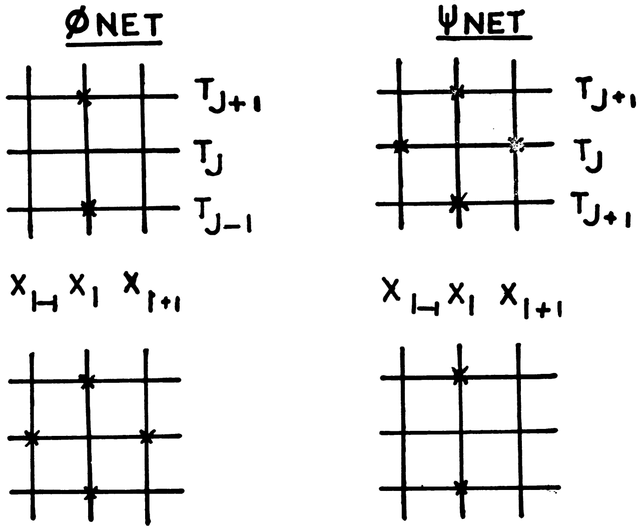

Fig. 1

$$\psi_{i,j+1} = a\psi_{i,j-1} - b(2\phi_{i,j-1} - \phi_{i+1,j} - \phi_{i-1,j}) + c(\psi_{i+1,j} + \psi_{i-1,j}) \qquad (5a)$$

and

$$\phi_{i,j+1} = a\phi_{i,j-1} - b(\psi_{i+1,j} + \psi_{i-1,j} - 2\psi_{i,j-1}) + c(\phi_{i+1,j} + \phi_{i-1,j}) \qquad (5b)$$

where

$$a = \frac{1 - \alpha^2}{1 + \alpha^2}, b = \frac{\alpha}{1 + \alpha^2}, c = \frac{\alpha^2}{1 + \alpha^2} \text{ and } \alpha = \frac{2k}{h^2}. \qquad (5c)$$

Thus, the equations (5*a*) and (5*b*) are now explicit in form and provide a single ψ and ϕ value in the $(j + 1)$th row when the ψ and ϕ values are known on the jth and $(j - 1)$th rows. Thus the solution can be obtained by an extremely simple repetitive computation performed from the operational stencils illustrated in **Fig. 2**.

3. Stability of the proposed scheme

To discuss the stability considerations of equations (5), we denote by θ_j, the vector of pivotal values along the line $j\Delta t$ for both ϕ and ψ values. Then,

$$\theta_j = \begin{bmatrix} v_j \\ \\ u_j \end{bmatrix} \text{ where } u_j \begin{bmatrix} \psi_{1,j} \\ \psi_{2,j} \\ \psi_{3,j} \\ \\ \psi_{N-1,j} \end{bmatrix} \qquad (6)$$

and v_j is similarly defined for $\phi_{i,j}$ with $i = 1, 2, \ldots N-1$ and $N\Delta x = 1$. The equations (5) can now be expressed in the matrix form

$$\theta_{j+1} = H\theta_j + K\theta_{j-1} \qquad (7)$$

Downloaded from https://academic.oup.com/comjnl/article/8/3/280/336153 by guest on 19 April 2024

where

$$\left[\begin{array}{cccc:cccc} o & c & & 0 & o & -b & & 0\\ c & \ddots & \ddots & & -b & \ddots & \ddots & \\ & \ddots & \ddots & c & & \ddots & \ddots & -b \\ 0 & & c & \cdot o & 0 & & -b & \cdot o \\ \hdashline o & b & & 0 & o & c & & 0\\ b & \ddots & \ddots & & c & \ddots & \ddots & \\ & \ddots & \ddots & b & & \ddots & \ddots & c\\ 0 & & b & o & 0 & & c & o \end{array}\right]$$

$$= \left[\begin{array}{c:c} cT & -bT \\ \hdashline bT & cT\end{array}\right] \tag{8}$$

and $K =$

$$\left[\begin{array}{cccc:cccc} a & & & 0 & 2b & & & 0\\ & \ddots & & & & \ddots & & \\ & & \ddots & & & & \ddots & \\ 0 & & & a & 0 & & & 2b \\ \hdashline -2b & & & 0 & a & & & 0\\ & \ddots & & & & \ddots & & \\ & & \ddots & & & & \ddots & \\ 0 & & & -2b & 0 & & & a \end{array}\right]$$

$$= \left[\begin{array}{c:c} aI & 2bI \\ \hdashline -2bI & aI\end{array}\right]; \tag{9}$$

I is the unit matrix of order $N-1$ and T is the $(N-1) \times (N-1)$ matrix.

$$T = \begin{bmatrix} 0 & 1 & & & 0\\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ 0 & & & 1 & 0 \end{bmatrix} \tag{10}$$

It can easily be noticed that both H and K are $(N-1) \times (N-1)$ block skew symmetric matrices of order $2(N-1)$.

We can reduce the two-level formula (7) by combining θ_{j-1} and θ_j into a single vector w_j, so that

$$w_{j+1} = Mw_j \tag{11}$$

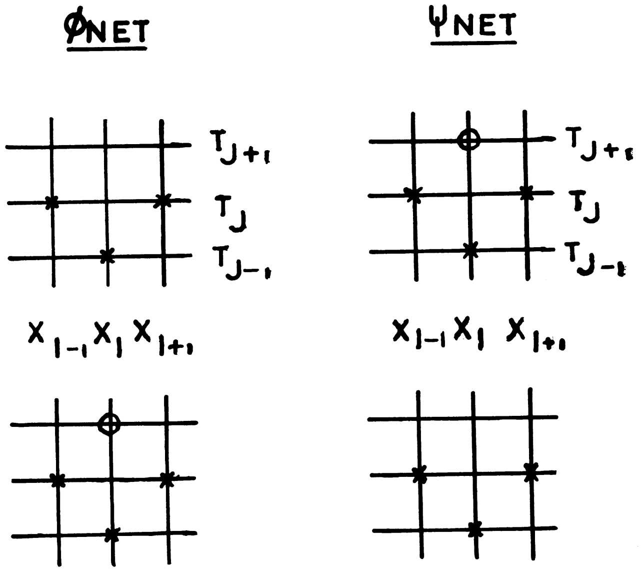

Fig. 2

where

$$M = \begin{bmatrix} H & K \\ I & 0 \end{bmatrix} \text{ and } w_j \begin{bmatrix} \theta_j \\ \theta_{j-1} \end{bmatrix}. \tag{12}$$

Since M is unsymmetric, we assume that it possesses $4(N-1)$ distinct eigenvalues λ_r, and hence the normalized eigenvectors $z^{(r)}$ of M form a complete set. Then we can write w_0 as a linear combination of these eigenvectors in the form

$$w_0 = \sum_1^{4N-4} \beta_r z^{(r)} \tag{13}$$

and $$w_{j+1} = M^{j+1}w_0 = \sum_1^{4N-4} \beta_r \lambda_r^{j+1} z^{(r)} \tag{14}$$

where β_r, $(r = 1, 2, 3, \ldots 4\,N-4)$ are chosen constants given by the relationship

$$\beta_r = (w_0, z_r). \tag{14b}$$

The stability analysis of the finite-difference scheme represented by equation (7) has been discussed adequately by Lowan (1957). A brief outline of this will now be given.

Initially the matrices H and K can be verified to be commutative; hence they possess simultaneous eigenvectors. From (7) we can obtain the results

$$\theta_2 = H\theta_1 + K\theta_0$$
$$\theta_3 = H(H\theta_1 + K\theta_0) + K\theta_1$$

and $$\theta_4 = (H^3 + 2HK)\theta_1 + (H^2 + K)K\theta_0.$$

By continuation of this sequence the following result

$$\theta_{j+1} = Q_j(H, K)\theta_1 + Q_{j-1}(H, K)K\theta_0 \tag{15}$$

is obtained where

$$Q_m(H, K) = H^m + \binom{m-1}{1} H^{m-2}K$$

Downloaded from https://academic.oup.com/comjnl/article/8/3/280/336153 by guest on 19 April 2024

$$+\binom{m-2}{2}H^{m-4}K^{2}\ldots+\binom{m-p}{p}H^{m-2p}K^{p} \quad (16)$$

and $\binom{m}{r}$ denotes the binomial coefficient $\frac{m!}{r!(m-r)!}$.

(17)

Now let us denote by θ^*_{j+1} the true accurate solution of the finite-difference equation (15). Assuming no round-off error to be present in the calculation, then

$$\theta^*_{j+1} = Q_j\theta^*_1 + Q_{j-1}K\theta^*_0. \quad (18)$$

Then, denoting the error vector $\epsilon_{j+1} = \theta_{j+1} - \theta^*_{j+1}$ as the difference between the numerical solution of (15) and the true accurate solution of (15), we can write

$$\epsilon_{j+1} = Q_j\epsilon_1 + Q_{j-1}K\epsilon_0. \quad (19)$$

Expanding ϵ_1 and ϵ_0 in terms of the normalized eigenvectors $\eta_{r,}$ $(r = 1, 2, 3, \ldots 2N-2)$ of H and K we have

$$\epsilon_0 = \sum_{r=1}^{2(N-1)} \alpha_r\eta_r \text{ and } \epsilon_1 = \sum_{r=1}^{2(N-1)} \beta_r\eta_r \quad (20)$$

where the α_r and β_r are chosen constants given by the relationships $\alpha_r = (\epsilon_0, \eta_r)$ and $\beta_r = (\epsilon_1, \eta_r)$. Further, if we denote by γ_r and $\mu_r\,(r = 1, 2, 3, \ldots 2N-2)$ the eigenvalues of H and K, then it follows that we can write

$$\epsilon_{j+1} = \sum_{r=1}^{2(N-1)} \beta_r Q_j(\gamma_r, \mu_r)\eta_r + \sum_{r=1}^{2(N-1)} \alpha_r Q_{j-1}(\gamma_r, \mu_r)\mu_r\eta_r. \quad (21)$$

Clearing terms, we have the expression

$$\begin{aligned}\epsilon_{j+1} = \sum_{r=1}^{2(N-2)} \beta_r\Big\{\gamma_r^j &+ \binom{j-1}{1}\gamma_r^{j-2}\mu_r \\ &+ \binom{j-2}{2}\gamma_r^{j-4}\mu_r^2 + \ldots\Big\}\eta_r \\ &+ \sum_{r=1}^{2(N-1)} \alpha_r\Big\{\gamma_r^{j-1} + \binom{j-2}{1}\gamma_r^{j-3}\mu_r \\ &+ \binom{j-3}{2}\gamma_r^{j-5}\mu_r^2 + \ldots\Big\}\mu_r\eta_r. \quad (22)\end{aligned}$$

By the introduction of the quantities $\lambda_{1,\,r}$ and $\lambda_{2,\,r}$ defined as

$$\left.\begin{aligned}\gamma_r &= \lambda_{1,r} + \lambda_{2,r} \\ \text{and} \qquad \mu_r &= -\lambda_{1,r}\lambda_{2,r}\end{aligned}\right\} \quad (23)$$

where $\lambda_{1,r}$ and $\lambda_{2,r}$ are the roots of the quadratic equation

$$\lambda^2 - \gamma_r\lambda - \mu_r = 0, \quad (24)$$

derived from (7), then equation (22) simplifies to the expression

$$\begin{aligned}\epsilon_{\cdot+1} = \sum_{r=1}^{2(N-1)} \beta_r\left\{\frac{\lambda_{1,r}^{j+1} - \lambda_{2,r}^{j+1}}{\lambda_{1,r} - \lambda_{2,r}}\right\}\eta_r \\ + \sum_{r=1}^{2(N-1)} \alpha_r\left\{\frac{\lambda_{1,r}^{j} - \lambda_{2,r}^{j}}{\lambda_{1,r} - \lambda_{2,r}}\right\}\mu_r\eta_r. \quad (25)\end{aligned}$$

In order for the proposed finite-difference scheme to be stable, it is necessary and sufficient for the norm of ϵ_j i.e., $||\epsilon_j||$ to remain bounded as $j \to \infty$. If the roots of the quadratic equation (24) are real, then from (25) it is evident that for the term $(\lambda_{1,r}^{j+1} - \lambda_{2,r}^{j+1})/(\lambda_{1,r} - \lambda_{2,r})$ to remain bounded the roots must be numerically smaller than unity. However, in general, the roots of (24) are complex, and if we denote them by $\lambda_r e^{iw_r}$ and $\lambda_r e^{-iw_r}$, then the term $(\lambda_{1,r}^{j+1} - \lambda_{2,r}^{j+1})/(\lambda_{1,r} - \lambda_{2,r})$ is $\lambda_r^j \sin j\omega_r/\sin \omega_r$, and the necessary and sufficient condition for the term to remain bounded is that $|\lambda| \leqslant 1$. Hence, the condition which is both necessary and sufficient for the stability of equation (7) is that the eigenvalues of (7) be numerically smaller than unity if real, or have modulus not greater than unity, if complex.

The eigenvalues λ_r of M are given by the determinantal equation

$$|M - \lambda I| = 0 \quad (26)$$

which in partitioned matrix form is

$$\begin{vmatrix} H - \lambda I & K \\ I & -\lambda K \end{vmatrix} = 0. \quad (27)$$

Partitioning $z^{(r)}$ into the form $\begin{bmatrix} p \\ q \end{bmatrix}$,

we obtain $\begin{bmatrix} H - \lambda I & K \\ I & -\lambda I \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ (28)

from which it follows that $p = \lambda q$ (29)

and $[I\lambda^2 - H\lambda - K]p = 0.$ (30)

Further partitioning of p into the form $\begin{bmatrix} s \\ t \end{bmatrix}$ and using the simpler representations of H and K, we obtain

$$\begin{bmatrix} (\lambda^2 - a)I - \lambda cT & \lambda bT - 2bI \\ -bT\lambda + 2bI & (\lambda^2 - a)I - \lambda cT \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (31)$$

Elimination of s is now possible owing to the commutativity of the sub-matrices involved, and we finally get the equation

$$\begin{aligned}&[(\lambda^2 - a)I - \lambda cT + ib(\lambda T - 2I)] \\ &\qquad [(\lambda^2 - a)I - \lambda cT - ib(\lambda T - 2I)]t = 0 \quad (32)\end{aligned}$$

which simplifies to

$$[(\lambda^2 - a \mp i2b)I + (-\lambda c \pm ib\lambda)T]t = 0. \quad (33)$$

It follows that t must be an eigenvector of T whose eigenvalues are known to be of the form $-2\cos\left(\frac{k\pi}{N}\right)$, $k = 1, 2, \ldots N-1$. Using this result, we get the equation

$$(\lambda^2 - a \mp 2ib)t + (-\lambda c \pm ib\lambda)\left(-2\cos\frac{k\pi}{N}\right)t = 0. \quad (34)$$

Downloaded from https://academic.oup.com/comjnl/article/8/3/280/336153 by guest on 19 April 2024

Equating the coefficient of t to zero, since $t \neq 0$, gives the quadratic equation

$$(1 + \alpha^2)\lambda^2 + 2\alpha(\alpha \mp i)\cos\frac{k\pi}{N}\lambda - (1 - \alpha^2 \pm i2\alpha) = 0 \tag{35}$$

on substitution of the values of a, b, and c from (5c) for the eigenvalues λ of the matrix M.

In general, the roots of this quadratic equation (35) are complex and a little analysis shows that

$$|\lambda_k| = \left[\left(\frac{1 - \alpha^2}{1 + \alpha^2}\right)^2 + \left(\frac{2\alpha}{1 + \alpha^2}\right)^2\right]^{1/2} = 1 \tag{36}$$

for all k.

Hence, it follows immediately that the eigenvalues of M are equal in modulus to unity, and the proposed explicit scheme is stable for all values of the mesh sizes.

4. Discretization error of the solution

It is now necessary to investigate the criteria for which solutions of the finite difference-equations (5) converge to the solution of the continuous partial differential equation (1) as the mesh sizes Δx and Δt tend to zero.

We now consider the expressions

$$\frac{(\Phi_{i,j+1} - \Phi_{i,j-1})}{2k} + \frac{(\Psi_{i+1,j} + \Psi_{i-1,j} - \Psi_{i,j-1} - \Psi_{i,j+1})}{h^2} + R_1 = 0 \tag{37a}$$

and

$$\frac{(\Psi_{i,j+1} - \Psi_{i,j-1})}{2k} - \frac{(\Phi_{i+1,j} + \Phi_{i-1,j} - \Phi_{i,j-1} - \Phi_{i,j+1})}{h^2} + R_2 = 0 \tag{37b}$$

where $\Phi_{i,j}$ and $\Psi_{i,j}$ represent the solutions of (3), and R_1, R_2 represent the local truncation errors. The terms R_1 and R_2 may be estimated by means of Taylor series expansions of $\Phi_{i,j}$ and $\Psi_{i,j}$ in the vicinity of the point (i, j) of the network, i.e.,

$$R_1 = \frac{k^2}{6}\left[\frac{\partial^3\Phi}{\partial t^3}\right]_{i,j} + \frac{h^2}{12}\left[\frac{\partial^4\Psi}{\partial x^4}\right]_{i,j} - \frac{k^2}{h^2}\left[\frac{\partial^2\Psi}{\partial t^2}\right]_{i,j} - \frac{k^4}{12h^2}\left[\frac{\partial^4\Psi}{\partial t^4}\right]_{i,j} + \ldots \tag{38a}$$

and

$$R_2 = \frac{k^2}{6}\left[\frac{\partial^3\Psi}{\partial t^3}\right]_{i,j} - \frac{h^2}{12}\left[\frac{\partial^4\Phi}{\partial x^4}\right]_{i,j} + \frac{k^2}{h^2}\left[\frac{\partial^2\Phi}{\partial t^2}\right]_{i,j} + \frac{k^4}{12h^2}\left[\frac{\partial^4\Phi}{\partial t^4}\right]_{i,j} + \ldots \tag{38b}$$

Immediately, it can be seen that the local truncation errors are of $O\left(h^2 + k^2 + \frac{k^2}{h^2}\right)$ for the proposed scheme, which compares favourably with the alternative finite-difference schemes mentioned in Section 1. Now it is evident that the finite-difference schemes (37) are compatible with the differential equations (3) only if the local truncation errors R_1 and $R_2 \to 0$ as h and $k \to 0$. By further inspection of equation (38) it is clearly seen that this is true only provided k goes to zero faster than h. If k and h tend to zero at the same rate, then denoting $k/h = c$, the finite-difference equations (37) will represent the solutions of the simultaneous partial differential equations

$$\frac{\partial\Phi}{\partial t} = \frac{-\partial^2\Psi}{\partial x^2} + c^2\frac{\partial^2\Psi}{\partial t^2} \tag{39a}$$

and

$$\frac{\partial\Psi}{\partial t} = \frac{\partial^2\Phi}{\partial x^2} - c^2\frac{\partial^2\Phi}{\partial t^2} \tag{39b}$$

which is the hyperbolic partial differential equation

$$\frac{\partial^2 y}{\partial t^2} + \frac{\partial^4 y}{\partial x^4} + c^4\frac{\partial^4 y}{\partial t^4} - 2c^2\frac{\partial^4 y}{\partial x^2\partial t^2} \tag{40}$$

in the original notation of (1).

Hence, for the equations (37) to be compatible with the differential equation it is necessary that as h and $k \to 0$, k must go to zero faster than h.

The discretization errors of the finite-difference equations (4) are defined as the differences between the solution Φ and Ψ of (3) and the solutions ϕ and ψ given by (4). If we denote by

$$e_{i,j} = \phi - \Phi \tag{41a}$$

and

$$f_{i,j} = \psi - \Psi, \tag{41b}$$

subtracting equations (37) from (4), we see that the discretization errors $e_{i,j}$ and $f_{i,j}$ satisfy the equations

$$\frac{(e_{i,j+1} - e_{i,j-1})}{2k} + \frac{(f_{i+1,j} + f_{i-1,j} - f_{i,j-1} - f_{i,j+1})}{h^2} = R_1 \tag{42a}$$

$$\frac{(f_{i,j+1} - f_{i,j-1})}{2k} - \frac{(e_{i+1,j} + e_{i-1,j} - e_{i,j-1} - e_{i,j+1})}{h^2} = R_2. \tag{42b}$$

These equations can be combined to give finite-difference equations similar in structure to (5) i.e.,

$$e_{i,j+1} = ae_{i,j-1} - b(2f_{i,j-1} - f_{i+1,j} - f_{i-1,j}) + c(e_{i+1,j} + e_{i-1,j}) + b(h^2R_1 + 2kR_2)_{i,j} \tag{43a}$$

$$f_{i,j+1} = af_{i,j-1} - b(e_{i+1,j} + e_{i-1,j} - 2e_{i,j-1}) + c(f_{i+1,j} + f_{i-1,j}) + b(h^2R_2 - 2kR_1)_{i,j}. \tag{43b}$$

These two equations can now be combined to form the matrix equation

$$E_{j+1} = HE_j + KE_{j-1} + g_j \tag{44}$$

Downloaded from https://academic.oup.com/comjnl/article/8/3/280/336153 by guest on 19 April 2024

$$\text{where} \quad E_j = \begin{bmatrix} f_j \\ \\ e_j \end{bmatrix} \text{ and } e_j = \begin{bmatrix} e_{i,j} \\ e_{2,j} \\ \\ e_{N-1,j} \end{bmatrix}, \qquad (45)$$

and f_j is similarly defined;

H and K are defined as in (8) and (9),

$$\text{where } g_j = \begin{bmatrix} \xi_j \\ \\ \chi_j \end{bmatrix}, \; \xi_j = b \begin{bmatrix} (h^2R_1 + 2kR_2)_1 \\ (h^2R_1 + 2kR_2)_2 \\ \cdot\cdot \\ \cdot\cdot \\ (h^2R_1 + 2kR_2)_N{-}1 \end{bmatrix}$$

$$\text{and } \chi_j = b \begin{bmatrix} (h^2R_2 - 2kR_1)_1 \\ (h^2R_2 - 2kR_1)_2 \\ \cdot\cdot \\ \cdot\cdot \\ (h^2R_2 - 2kR_1)_N{-}1 \end{bmatrix}. \qquad (46)$$

It can be noticed that g_j is a vector composed of truncation error terms.

Now, initially it is reasonable to assume $\Phi = \phi$ and $\Psi = \psi$ for the line $j = 0$; hence $E_0 = 0$.

By a similar analysis to (15), we can derive the expression

$$\begin{aligned} E_2 &= HE_1 + g_1, \\ E_3 &= H(HE_1) + KE_1 + Hg_1 + g_2, \\ E_4 &= H(H^2 + K)E_1 + KHE_1 + Kg_1 + H^2g_1 + Hg_2 + g_3 \\ &= (H^3 + 2HK)E_1 + (H^2 + K)g_1 + Hg_2 + g_3, \end{aligned}$$

and finally

$$\begin{aligned} E_{j+1} = \Big[H^j &+ \binom{j-1}{1}H^{j-2}K \\ &+ \binom{j-2}{2}H^{j-4}K^2 + \ldots\Big]E_1 \\ &+ \sum_{s=1}^{j}\Big\{H^{j-s} + \binom{j-s-1}{1}H^{j-s-2}K \\ &+ \binom{j-s-2}{2}H^{j-s-4}K^2 + \ldots\Big\}g_s \qquad (47) \end{aligned}$$

where a similar notation to (15) has been used throughout.

Assuming that we can expand the vectors $E_1, g_1, g_2, \ldots g_j$ in terms of the simultaneous eigenvectors η_r, ($r = 1, 2, 3, \ldots, 2N - 2$) of H and K, suitably normalized, we have immediately

$$E_1 = \sum_{r=1}^{2(N-1)} \alpha_r \eta_r \text{ and } g_s = \sum_{r=1}^{2(N-1)} \beta_r{}^{(s)}\eta_r \qquad (48)$$

where the α_r and $\beta_r^{(s)}$ are chosen constants to be defined later. If now γ_r and μ_r, ($r = 1, 2, \ldots 2N - 2$) are the eigenvalues of H and K, equation (47) becomes

$$\begin{aligned} E_{j+1} = \sum_{r=1}^{2(N-1)} \alpha_r\Big\{\gamma^j &+ \binom{j-1}{1}\gamma_r^{j-2}\mu_r + \\ &+ \binom{j-2}{2}\gamma_r^{j-4}\mu_r + \ldots\Big\}\eta_r + \\ \sum_{s=1}^{j}\sum_{r=1}^{2(N-1)} \beta_r^{(s)}\Big\{\gamma_r^{j-s} &+ \binom{j-s-1}{1}\gamma_r^{j-s-2}\mu_r \\ &+ \binom{j-s-2}{2}\gamma_r^{j-s-4}\mu_r^2 + \ldots\Big\}\eta_r. \qquad (49) \end{aligned}$$

This expression can be written in the simplified form

$$\begin{aligned} E_{j+1} &= \sum_{r=1}^{2(N-1)} \alpha_r \frac{(\lambda_{1,r}^{j+1} - \lambda_{2,r}^{j+1})}{(\lambda_{1,r} - \lambda_{2,r})}\eta_r \\ &+ \sum_{s=1}^{j}\sum_{r=1}^{2(N-1)} \beta_r^{(s)} \frac{(\lambda_{1,r}^{j+1-s} - \lambda_{2,r}^{j+1-s})}{(\lambda_{1,r} - \lambda_{2,r})}\eta_r \qquad (50) \end{aligned}$$

using the abbreviation of (23).

Now earlier we showed that the roots $\lambda_{1,r}$ and $\lambda_{2,r}$ are complex conjugate and of modulus unity, hence if we denote

$$\lambda_{1,r} = e^{i\omega_r} \text{ and } \lambda_{2,r} = e^{-i\omega_r} \qquad (51)$$

then (50) becomes

$$\begin{aligned} E_{j+1} &= \sum_{r=1}^{2(N-1)} \alpha_r \frac{\sin(j+1)\omega_r}{\sin\omega_r}\eta_r \\ &+ \sum_{r=1}^{2(N-1)}\sum_{s=1}^{j} \beta_r^{(s)} \frac{\sin(j+1-s)\omega_r}{\sin\omega_r}\eta_r. \qquad (52) \end{aligned}$$

$$\text{Now the expressions} \quad \frac{\sin(j+1)\omega_r}{\sin\omega_r} \qquad (53)$$

$$\text{and} \quad \sum_{s=1}^{j}\frac{\sin(j+1-s)\omega_r}{\sin\omega_r} \qquad (54)$$

can be verified to remain bounded as $j \longrightarrow \infty$.

Let T be an upper bound of both

$$\frac{\sin(j+1)\omega_r}{\sin\omega_r}$$

$$\text{and} \quad \sum_{s=1}^{j}\frac{\sin(j+1-s)\omega_r}{\sin\omega_r}$$

which on summation can be verified to be the quantity

$$\frac{\sin\omega_r - \sin j\omega_r + \sin(j-1)\omega_r}{4\sin\omega_r\sin^2(\omega_r/2)}. \qquad (55)$$

We further denote by $\bar{\alpha}$ and $\bar{\beta}$, the upper bounds of α_r and β_r defined by

$$\alpha_r = \frac{1}{2(N-1)}(E_1, \eta_r) \qquad (56)$$

$$\text{and} \quad \beta_r^{(s)} = \frac{1}{2(N-1)}(g_s, \eta_r). \qquad (57)$$

Finally, the components of the eigenvectors $\eta_r(r = 1, 2, \ldots 2N - 2)$ can be verified to be

$$\sin\frac{r\pi}{N-1}, -\sin\frac{2r\pi}{(N-1)}, \ldots (-1)^N \sin\frac{(N-2)r\pi}{(N-1)},$$

Downloaded from https://academic.oup.com/comjnl/article/8/3/280/336153 by guest on 19 April 2024

$$i\sin\frac{r\pi}{(N-1)},\ -i\sin\frac{2r\pi}{(N-1)},\ldots$$

$$\ldots(-1)^N i\sin\frac{(N-2)r\pi}{(N-1)}. \qquad (58)$$

An upper bound $\bar{\eta}$ of the components of the eigenvectors η_r can be obtained if we make use of the following identity:

$$\sum_{q=1}^{Q-1}\sin\frac{lq\pi}{Q}\sin\frac{kq\pi}{Q}=\begin{cases}0 & k\neq l\\ Q/2 & k=l\end{cases}. \qquad (59)$$

Hence, after a little analysis we obtain the result

$$\bar{\eta}=2. \qquad (60)$$

By substituting these quantities into equation (52), we have an upper bound for the discretization error at the point i on the line $j+1$,

$$|e_{i,j+1}|<4NT(\bar{\alpha}+\bar{\beta}) \qquad (61)$$

with a similar result for $f_{i,j+1}$.

If we assume that the components of E_1 are of the order of magnitude of $(\Delta x)^\sigma$ where $\sigma\geqslant 2$, then from (56) we have $\bar{\alpha}=O(2\{\Delta x\}^\sigma)$. Also, the components of g_s are given to the order of magnitude $\{(\Delta x)^4+(\Delta t)^2\}$ where we assume that the compatibility condition is obeyed and $\Delta t\to 0$ faster than $\Delta x\to 0$. Accordingly we have that $\bar{\beta}=O(2\{(\Delta x)^4+(\Delta t)^2\})$. Thus, when we substitute $\bar{\alpha}$ and $\bar{\beta}$ in equation (61) and use the result $N\Delta x=1$, the final result is

$$|e_{i,j+1}|<8T(A(\Delta x)^{\sigma-1}+B(\Delta x)^3+B\Delta t) \qquad (62)$$

where A and B are suitable constants, with a similar result true for $f_{i,j+1}$. Equation (61) clearly shows that $e_{i,j+1}\to 0$ as Δx and $\Delta t\to 0$ provided $\sigma\geqslant 2$.

Therefore, we now conclude this section with the statement that the solution of the difference equations (5) tends to the solution of the differential equation (1) for the conditions stated, whilst the maximum difference between the two solutions is given by equation (62).

5. Experimental verification

To test the validity of the proposed scheme and the derived stability and convergence criteria, the problem of the vibrating beam hinged at both ends was investigated with the following boundary conditions:

$$y(0,t)=\frac{\partial^2 y}{\partial x^2}(0,t)=y(1,t)=\frac{\partial^2 y}{\partial x^2}(1,t)=0,\ t>0.$$

The initial conditions were taken to be

$$y(x,0)=\frac{x}{12}(2x^2-x^3-1),\quad 0<x<1$$

and $$\frac{\partial y}{\partial t}(x,0)=0 \qquad 0<x<1.$$

The exact solution to the continuous problem is easily obtained by Fourier series analysis and is given by the expression

$$y(x,t)=\sum_{s=1}^{\infty}a_s\sin s\pi x\cos s^2\pi^2 t \qquad (63)$$

Table 1

$t=0\cdot02$

x	0	0·05	0·1	0·15	0·2	0·25	0·3	0·35	0·4	0·45	0·5
Exact solution equation (63)	0	0·03949	0·07802	0·11462	0·14840	0·17853	0·20426	0·22496	0·24012	0·24937	0·25248
Proposed explicit scheme equation (5)	0	0·03872	0·07676	0·11333	0·14766	0·17873	0·20555	0·22686	0·24141	0·24979	0·25258

Table 2

$x=0\cdot5$

t	0	0·0387	0·0512	0·0637	0·0762	0·0887	0·1013
Exact solution equation (63)	0·25748	0·23888	0·22524	0·20817	0·18795	0·16486	0·13927
Proposed explicit scheme equation (5)	0·25000	0·24869	0·22878	0·20554	0·17832	0·16336	0·14421

Downloaded from https://academic.oup.com/comjnl/article/8/3/280/336153 by guest on 19 April 2024

$$\text{where} \quad a_s = 2\int_0^1 \frac{x}{12}(2x^2 - x^3 - 1)\sin(s\pi x)dx$$

$$= \frac{4}{s^5\pi^5}(\cos(s\pi) - 1).$$

Computation of the exact solution at the mesh points with spacings $\Delta x = 0{\cdot}05$, $\Delta t = 0{\cdot}00125$ from equation (63) correct to five decimal places was obtained and compared with the results obtained from the finite-difference equations (5). Both short-time and long-time studies of the exact and finite-difference solutions were compared, and agreement to within the derived error bound given by equation (62) was obtained. A typical sample of these comparisons are shown in **Table 1** for $t = 0{\cdot}02$ and $x = 0(0{\cdot}05)\ 0{\cdot}5$, and in **Table 2** for $x = 0{\cdot}5$ and larger values of t.

The stability criteria of (5) was checked experimentally by taking a large variety of mesh sizes Δx and Δt and obtaining the solutions on the Sheffield University Mercury computer. In each case, no evidence of numerical instability due to growth of round-off error was found.

Acknowledgement

The author wishes to express his gratitude to Mrs. A. Fairburn for expert programming assistance in verifying the theoretical results given in this paper.

References

Collatz, L. (1951). "Zur Stabilität des Differenzenverfahrens bei der Stabschwingungsgleichung," *Z.A. Math. Mech.*, Vol. 31, pp. 392–393.

Crandall, S. H. (1954). "Numerical Treatment of a Fourth Order Partial Differential Equation," *J. Assoc. Comp. Mach.*, Vol. 1, pp. 111–118.

Conte, S. D., and Royster, W. C. (1956). "Convergence of Finite Difference Solutions to a Solution of the Equation of the Vibrating Rod," *Proc. Amer. Math. Soc.*, Vol. 7, pp. 742–749.

Conte, S. D. (1957). "A Stable Implicit Finite Difference Approximation to a Fourth Order Parabolic Equation," *J. Assoc. Comp. Mach.*, Vol. 4, pp. 18–23.

Albrecht, J. (1957). "Zum Differenzenverfahren bei parabolischen Differentialgleichungen," *Z. Angew. Math and Mech.*, Vol. 37, No. 5–6, pp. 202–212.

Du Fort, E. C., and Frankel, S. P. (1953). "Stability Conditions in the Numerical Treatment of Parabolic Differential Equations," *Math. Tab., Wash.*, Vol. 7, pp. 135–152.

Richtmyer, R. C. (1957). "Difference Methods for Initial Value Problems," *Interscience Tracts. in Pure and Applied Mathematics*, No. 4, Interscience Pubs., N.Y.

Lowan, A. N. (1957). "The Operator Approach to Problems of Stability and Convergence," *Scripta Mathematica*, No. 8.

Book Review

Computing Methods (*Volumes I and II*), by I. S. Berezin and N. P. Zhidkov, 1965; 464 and 679 pages. (Oxford: *Pergamon Press Ltd.*, 100s. per volume.)

This enormous work—1143 pages in all—attempts to cover the whole field of what we would call numerical analysis. The first volume has six chapters, on approximate quantities and errors, interpolation, numerical integration and differentiation, general analytic methods of approximation to functions, least squares approximation: the second, on linear algebraic equations, non-linear algebraic equations and transcendental equations, eigenvalues and vectors of matrices, ordinary differential equations, partial differential equations, and integral equations. I found it all very laborious and uninspiring and I could not detect anything new in either the methods or the results. On the contrary, the treatment has a very old-fashioned air, with pages of heavy algebra, extensive displays of formulae—for example, for numerical integration and differentiation—which could have been put into appendices, or, better, left out altogether, and exhaustive pursuit of details with no great reward in the end, as in the 40 pages on the Runge-Kutta method. The contrast with the elegant writings of Henrici, for example, is very striking. The point of view is wholly that of the hand computer (there is a bare mention of electronic machines in the introduction to Volume I, which includes the statement that a modern high-speed computer operates at about 8,000 instructions per second). Even so, I got a strong impression all the way through that the authors had never done much actual computation, except perhaps a few academic exercises, and this was reinforced by particular points of detail. Thus the account of the Euler-Maclaurin formula relating an integral to a series does not say that the series is asymptotic, not convergent; and the brief note on page 294 on calculation of integrals with a variable upper limit suggests that they have never heard of the neat and efficient methods due I think to Comrie and given years ago in *Interpolation and Allied Tables*.

There is of course a great deal of information in the book, especially in the chapters on approximation; but for a work published in 1965 the omissions are unforgiveable. On quadrature and ordinary differential equations there is no mention of the modern European school—Dahlquist, Henrici, Rutishauser, Stiefel; on matrix calculations, Wilkinson's name appears only in a 1948 reference, and the methods of Jacobi, Givens and Householder for eigenvalue calculations are not mentioned; nor is the work of Young, Richtmyer, Varga on partial differential equations, nor indeed any of the modern American writers. The publishers should have given the date of the original Russian edition. The references, most of which are to work published between 1948 and 1954 with none later than 1958, suggest that it was written in 1955–56; if that is so, one can understand why the important modern work has been missed. It strikes me as yet another warning against trying to write everything down in one great work.

J. Howlett

Downloaded from https://academic.oup.com/comjnl/article/8/3/280/336153 by guest on 19 April 2024