On the Propagation of Errors in the Inversion of Certain Tridiagonal Matrices

By Arnold N. Lowan

Abstract. When the differential equation of heat conduction is replaced by the implicit difference analog, one is led to the solution of $A\mathbf{y} = \mathbf{b}$ where A is a tridiagonal matrix whose elements on the principal diagonal are $=2 + 2r$ and whose elements off the principal diagonal are $= -r$.

The system of equations may be solved by the following algorithm:

$$\beta_k = u = r^2\beta_{k-1}^{-1}, \quad \beta_1 = u_1; \qquad \gamma_k = -r\beta^{-1}; \qquad z_k = (b_k + rz_{k-1})\beta_k^{-1}, \quad z_1 = b_1u^{-1};$$

$$y_k = z_k - \gamma_k y_{k+1}, \qquad y_M = z_M.$$

An upper bound of the round-off errors in the computed values of the y_k's is obtained. An actual test case showed that the theoretical upper bound is about four times larger than the true round-off error. Moreover, the theoretical upper bound does not seem to vary appreciably with r.

When the differential equation of heat conduction

$$\frac{\partial T}{\partial t} = \sigma \frac{\partial^2 T}{\partial x^2}, \qquad 0 \leqq x \leqq a, \quad t > 0$$

is replaced by the "implicit" difference analog

$$\frac{T_{m,n+1} - T_{m,n}}{\Delta t} = \frac{\sigma}{2(\Delta x)^2}[T_{m-1,n+1} - 2T_{m,n+1} + T_{m+1,n+1} + T_{m-1,n} - 2T_{m,n} + T_{m+1,n}]$$

$$m = 1, 2, 3, \cdots M, \quad \Delta x = \frac{a}{M+1}$$

or

$$(1) \qquad \begin{aligned}(2 + 2r)T_{m,n+1} - r(T_{m-1,n+1} + T_{m+1,n+1}) &= (2 - 2r)T_{m,n} \\ &\quad + r(T_{m-1,n} + T_{m+1,n})\end{aligned}$$

where $T_{m,n} = T(m\Delta x, n\Delta t)$ and $r = \sigma\Delta t/(\Delta x)^2$ it is a known fact that the difference scheme (1) is unconditionally stable [1]. If the desired solution is required to vanish on the boundaries $x = 0$ and $x = a$, the system of equations (1) may be written in the compact form†

$$(1^*) \qquad A\mathbf{T}_{n+1} = B\mathbf{T}_n = \mathbf{b} \quad \text{(say)}$$

where A is a tridiagonal matrix whose elements on the principal diagonal are $= 2 + 2r$ while the elements off the principal diagonal are $= -r$.

Received July 31, 1959; revised form February 9, 1960

† When the temperature is prescribed on the boundaries, equation (1*) is essentially unchanged except for the fact that the first and last components of $\mathbf{b}$ are slightly altered.

The system of equations (1) may be easily solved by the following algorithm [2]

$$\beta_k = u - \frac{r^2}{\beta_{k-1}} \qquad k = 1, 2, 3, \cdots, M; \quad \beta_1 = u \tag{2}$$

$$\gamma_k = -\frac{r}{\beta_k} \qquad k = 1, 2, 3, \cdots, M; \tag{3}$$

$$z_k = \frac{1}{\beta_k}(b_k + rz_{k-1}) \qquad k = 1, 2, 3, \cdots, M; \quad z_1 = \frac{b_1}{u} \tag{4}$$

$$y_k = z_k - \gamma_k y_{k+1} \qquad k = 1, 2, 3, \cdots, M; \quad y_M = z_M \tag{5}$$

where we have written u for $2 + 2r$ and we have denoted the components of $\mathbf{T}_{n+1}$ by y_k. The question arises: if the computations involved in the above algorithm are carried to p decimals (i.e., if products and ratios are rounded to p decimals) what is the upper bound of the round-off errors in the computed values of y_k?

In the derivation of the desired upper bound we shall require a lower bound of the β_k's and upper bounds of γ_k, z_k and y_k. If in (2) we put $k = 2, 3, \cdots$ we get

$$\begin{aligned} \beta_2 &= u - \frac{r^2}{\beta_1} = \beta_1 - \frac{r^2}{\beta_1} \\ \beta_3 &= u - \frac{r^2}{\beta_2} \\ &\vdots \end{aligned} \tag{6}$$

From the first of the above equations it is clear that $\beta_2 < \beta_1$. From the first two equations it follows that

$$\beta_3 - \beta_2 = r^2\left(\frac{1}{\beta_1} - \frac{1}{\beta_2}\right) < 0.$$

Thus $\beta_3 < \beta_2$. Similarly it may be shown that $\beta_4 < \beta_3, \cdots, \beta_k < \beta_{k-1}$. Thus the β_k's form a monotonically decreasing sequence. It may be readily shown that the lower limit of the sequence, to be denoted by β_*, is the larger of the two roots of the quadratic equation

$$x^2 - (2 + 2r)x + r^2 = 0. \tag{7}$$

Accordingly

$$\beta_* = 1 + r + \sqrt{1 + 2r}. \tag{8}$$

From (3) it follows that $|\gamma_k| < r/\beta_k$. If then γ^* denotes an upper bound of $|\gamma_k|$ we may put

$$\gamma^* = \frac{r}{\beta_*} = \frac{r}{1 + r + \sqrt{1 + 2r}}. \tag{9}$$

If in (4) we put $k = 2, 3, \cdots$ and subsequently eliminate $z_2, z_3, \cdots, z_{k-1}$ we ultimately get

$$z_k = \frac{b_k}{\beta_k} + \frac{rb_{k-1}}{\beta_k\beta_{k-1}} + \frac{r^2 b_{k-2}}{\beta_k\beta_{k-1}\beta_{k-2}} + \cdots \frac{r^{k-1} b_1}{\beta_k\beta_{k-1}\cdots\beta_1}$$

whence

$$|z_k| \leqq \frac{b^*}{\beta_*}\left[1 + \frac{r}{\beta_*} + \frac{r^2}{\beta_*^2} + \cdots \left(\frac{r}{\beta_*}\right)^{k-1}\right] \cong \frac{b^*}{\beta_*} \cdot \frac{1}{1 - \frac{r}{\beta_*}} = \frac{b^*}{\beta_* - r}$$

where b^* is the largest of the absolute values of b_k. If then z^* denotes an upper bound of $|z_k|$ we may put

$$z^* = \frac{b^*}{\beta_* - r}. \tag{10}$$

Finally from (5) we readily get

$$\begin{aligned}
y_M &= z_M \\
y_{M-1} &= z_{M-1} - \gamma_{M-1} z_M \\
y_{M-2} &= z_{M-2} - \gamma_{M-2} z_{M-1} + \gamma_{M-2}\gamma_{M-1} z_M \\
&\vdots \\
y_1 &= z_1 - \gamma_1 z_2 + \gamma_1\gamma_2 z_3 - \cdots (-1)^{M-1}\gamma_1\gamma_2 \cdots \gamma_{M-1} z_1 .
\end{aligned}$$

From the above system of equations it is clear that

$$\begin{aligned}
y^* &= z^*(1 + \gamma^* + \gamma^{*2} \cdots + \gamma^{*M-1}) \\
&\cong \frac{z^*}{1 - \gamma^*} = \frac{b^*}{\beta_* - r} \cdot \frac{1}{r - \frac{r}{\beta_*}} = \frac{b^*\beta_*}{(\beta_* - r)^2}
\end{aligned} \tag{11}$$

is an upper bound of the absolute values of the y_k's.

We now turn to the evaluation of upper bounds of the errors in the β_k's, γ_k's, z_k's and y_k's. It will be convenient to denote by $E(\beta_k)$ the absolute value of the error in β_k and by $E^*(\beta)$ an upper bound of the errors in the β_k's. A similar notation will be used for the γ_k's, z_k's and y_k's. From (2) we have

$$E(\beta_2) = \frac{r^2}{\beta_1^2} E(\beta_1) + \delta \leqq \frac{r^2}{\beta_*^2} E(\beta_1) + o$$

where $\delta = \frac{1}{2} \times 10^{-p}$ is the maximum round-off error. Similarly

$$\begin{aligned}
E(\beta_3) &= \frac{r^2}{\beta_2^2} E(\beta_2) + \delta \leqq \frac{r^2}{\beta_*^2} E(\beta_2) + \delta \\
&= \frac{r^2}{\beta_*^2}\left[\frac{r^2}{\beta_*^2} E(\beta_1) + \delta\right] + \delta \\
&= \left(1 + \frac{r^2}{\beta_*^2}\right)\delta + \left(\frac{r^2}{\beta_*^2}\right)^2 E(\beta_1).
\end{aligned}$$

Proceeding in this manner we ultimately get

$$\begin{aligned}
E(\beta_n) &\leqq \left[1 + \left(\frac{r^2}{\beta_*^2}\right) + \cdots \left(\frac{r^2}{\beta_*^2}\right)^{M-2}\right]\delta + \left(\frac{r^2}{\beta_*^2}\right)^{M-1} E(\beta_1) \\
&\cong \frac{1}{1 - \frac{r^2}{\beta_*^2}}\,\delta = \frac{\beta_*^2}{\beta_*^2 - r^2}\,\delta
\end{aligned}$$

where we have neglected the second term of the above inequality since $r < \beta_*$. Thus

$$E^*(\beta) = \frac{\beta_*^2}{\beta_*^2 - r^2}\,\delta \tag{12}$$

is an upper bound of the absolute values of the $E(\beta_k)$'s.

Consider now the evaluation of $E^*(\gamma)$. From (3) it follows that

$$\begin{aligned} E(\gamma_k) &= \frac{r}{\beta_k^2} E(\beta_k) + \delta \\ &< \frac{r}{\beta_*^2} E^*(\beta) + \delta \end{aligned}$$

whence

$$\begin{aligned} E^*(\gamma) &= \frac{r}{\beta_*^2} E^*(\beta) + \delta = \frac{r}{\beta_*^2} \cdot \frac{\beta_*^2}{\beta_*^2 - r^2}\,\delta + \delta \\ &= \left(1 + \frac{r}{\beta_*^2 - r^2}\right)\delta. \end{aligned} \tag{13}$$

Consider next the evaluation of $E^*(z)$. From (4) we get:

$$\begin{aligned} E(z_k) &= \frac{1}{\beta_k^2}\{(b_k + rz_{k-1})E(\beta_k) + \beta_k[E(b_k) + rE(z_{k-1})]\} + \delta \\ &\leqq \frac{1}{\beta_*^2}(b^* + rz^*)E^*(\beta) + \frac{1}{\beta_*}E^*(b) + \frac{r}{\beta_*}E(z_{k-1}) + \delta \\ &= \frac{1}{\beta_*^2}(b^* + rz^*)\frac{\beta_*^2}{\beta_*^2 - r^2} + \frac{1}{\beta_*}E^*(b) + \frac{r}{\beta_*}E(z_{k-1}) + \delta \\ &= \left(1 + \frac{b^* + rz^*}{\beta_*^2 - r^2}\right)\delta + \frac{1}{\beta_*}E^*(b) + \frac{r}{\beta_*}E(z_{k-1}). \end{aligned}$$

Proceeding as in the evaluation of $E^*(\beta)$ we ultimately get

$$E^*(z) = \frac{\beta_*}{\beta_* - r}\left(1 + \frac{b^* + rz^*}{\beta_*^2 - r^2}\right)\delta + \frac{E^*(b)}{\beta_* - r}. \tag{14}$$

If in the last equation we replace z^* by its expression from (10) we ultimately get

$$E^*(z) = \frac{\beta_*}{\beta_* - r}\left[1 + \frac{b^*\beta_*}{(\beta_* - r)(\beta_*^2 - r^2)}\right] + \frac{E^*(b)}{\beta_* - r}. \tag{15}$$

If on the other hand we replace z^* in (14) by Z, the largest absolute value of the z_k's we obtain

$$E^*(z) = \frac{\beta_*}{\beta_* - r}\left(1 + \frac{b^* + rZ}{\beta_*^2 - r^2}\right)\delta + \frac{E^*(b)}{\beta_* - r}. \tag{15*}$$

While the expression in (15) is an upper bound of the errors in the z_k's, it is reasonable to refer to the expression in (15*) as the least upper bound of the errors in the z_k's.

Finally, consider the evaluation of $E^*(y)$. From (5) we get

$$E(y_k) = E(z_k) + \gamma_k E(y_{k+1}) + y_{k+1}E(\gamma_k) + \delta$$

whence

$$E(y_k) \leqq E^*(z) + \gamma^* E(y_{k+1}) + y^* E^*(\gamma) + \delta. \tag{16}$$

Substituting for $E^*(z)$, γ^*, y^* and $E^*(\gamma)$ their expressions from (15), (9), (11), and (13) the last inequality becomes

$$\begin{aligned} E(y_k) \leqq \frac{\beta_*}{\beta_* - r}\left[1 + \frac{b^*\beta_*}{(\beta_* - r)(\beta_*^2 - r^2)}\right]\delta + \frac{E^*(b)}{\beta_* - r} \\ + \frac{b^*\beta_*}{(\beta_* - r)^2}\left(1 + \frac{r}{\beta_*^2 - r^2}\right)\delta + \delta + \frac{r}{\beta_*}E(y_{k+1}). \end{aligned} \tag{17}$$

Proceeding again as in the evaluation of $E^*(\beta)$, the last inequality ultimately yields:

$$\begin{aligned} E^*(y) = \frac{\beta_*}{\beta_* - r}\left\{\frac{\beta_*}{\beta_* - r}\left[1 + \frac{b^*\beta_*}{(\beta_* - r)(\beta_*^2 - r^2)}\right]\right. \\ \left. + \frac{b^*\beta_*}{(\beta_* - r)^2}\left(1 + \frac{r}{(\beta_*^2 - r)^2}\right) + 1\right\}\delta + \frac{\beta_*}{(\beta_* - r)^2}E^*(b). \end{aligned} \tag{18}$$

If, on the other hand, we substitute for $E^*(z)$ in (16) its expression from (15*) and replace y^* by Y the largest of the absolute values of the y_k's, while γ^* and $E^*(\gamma)$ are replaced by their expressions from (9) and (13), we obtain as the counterpart of (17)

$$\begin{aligned} E(y_k) \leqq \frac{\beta_*}{\beta_* - r}\left[1 + \frac{b^* + rZ}{\beta_*^2 - r^2}\right]\delta + \frac{E^*(b)}{\beta_* - r} \\ + Y\left(1 + \frac{r}{\beta_*^2 - r^2}\right)\delta + \delta + \frac{r}{\beta_*}E(y_{k+1}). \end{aligned} \tag{17*}$$

Proceeding again as in the evaluation of $E^*(\beta)$, the last inequality ultimately yields:

$$\begin{aligned} E^*(y) = \frac{\beta_*}{\beta_* - r}\left\{\frac{\beta_*}{\beta_* - r}\left[1 + \frac{b^* + rZ}{\beta_*^2 - r^2}\right] + Y\left(1 + \frac{r}{\beta_*^2 - r^2}\right) + 1\right\}\delta \\ + \frac{\beta_*}{(\beta_* - r)^2}E^*(b). \end{aligned} \tag{18*}$$

It will be convenient to rewrite the last equation in the form

$$E^*(y) = S_0(r) + b^* S_1(r) + Y S_2(r) + Z S_3(r) + E^*(b) S_4(r) \tag{19}$$

where

$$\begin{cases} S_0(r) = \dfrac{\beta_*}{\beta_* - r} + \dfrac{\beta_*^2}{(\beta_* - r)^2} \\ S_1(r) = \dfrac{\beta_*^2}{(\beta_* - r)^2(\beta_*^2 - r^2)} \\ S_2(r) = \dfrac{\beta_*}{\beta_* - r}\left(1 + \dfrac{r}{\beta_*^2 - r^2}\right) \\ S_3(r) = \dfrac{r\beta_*^2}{(\beta_* - r)^2(\beta_*^2 - r^2)} = rS_1(r) \\ S_4(r) = \dfrac{\beta_*}{(\beta_* - r)^2}. \end{cases} \tag{20}$$

In (19) in conjunction with (20) we have an upper bound of the round-off errors in the values of the y_k's—the solutions of $A\mathbf{y} = \mathbf{b}$. In case the b_k's are exact we must of course put $E^*(b) = 0$.

To test the formula (18*) the exact components of $\mathbf{b}$ were computed from $A\mathbf{y} = \mathbf{b}$ where A is a 20×20 tridiagonal matrix of the type above considered with $r = 1$ and $r = 2$ and the 20 components of $\mathbf{y}$ were arbitrarily assigned; the values of the components were then calculated by the above algorithm in terms of the exact values of b_k. The computations were carried to eight decimals. The maximum discrepancy between the exact values of the y_k's and the corresponding computed values was two units in the last place. The upper bound of the round-off errors evaluated from (19) in conjunction with (20) was eight units in the last place in the case $r = 1$ and seven units in the last place in the case $r = 2$. The estimated upper bounds of the round-off errors must be considered as indeed very close to the actual round-off errors.

The writer wishes to express his appreciation to Roy Steeves who carried out the above-mentioned test.

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

1. The implicit scheme was first suggested by J. Crank and P. Nicolson in the paper entitled, "A practical method for the numerical evaluation of solutions of differential equations of the heat conduction type," Camb. Phil. Soc. *Proc.*, v. 43, 1943, p. 50–67. Its stability is discussed by G. G. O'Brien, Morton A. Hyman, and Sidney Kaplan in "A Study of the numerical solution of partial differential equations," *Jn. Math. and Phys.*, v. 29, 1950/51, p. 223–251. It is also discussed in the writer's book *The Operator Approach to Problems of Stability and Convergence*, Scripta Mathematica, 1957.

2. See for instance M. Lotkin's "The numerical integration of heat conduction equations," *Jn. Math. and Phys.*, v. 37, 1958, p. 178, and R. D. Richtmyer, *Difference Methods for Initial Value Problems*, Chap. VI., Interscience, N. Y., 1957.