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 $Ay = b$ 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:

\begin{align*}
\beta_k &= u = r^2\beta_{k-1}, \quad \beta_1 = u_1; \quad \gamma_k = -r\beta^{-1}; \quad z_k = (b_k + rz_{k-1})\beta_k^{-1}, \quad z_1 = b_1u^{-1}; \\
y_k &= z_k - \gamma_ky_{k+1}, \quad y_M = z_M.
\end{align*}

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

\[ \frac{\partial T}{\partial t} = \sigma \frac{\partial^2 T}{\partial x^2}, \quad 0 \leq x \leq a, \quad t > 0 \]

is replaced by the “implicit” difference analog

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

\begin{align*}
m &= 1, 2, 3, \cdots M, \quad \Delta x = \frac{a}{M + 1}
\end{align*}

or

(1)

\[ (2 + 2r)T_{m,n+1} - r(T_{m-1,n+1} + T_{m+1,n+1}) = (2 - 2r)T_{m,n} + r(T_{m-1,n} + T_{m+1,n}) \]

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

\begin{equation}
AT_{n+1} = BT_n = b \quad \text{(say)}
\end{equation}

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 $b$ are slightly altered.
The system of equations (1) may be easily solved by the following algorithm [2]

(2) \[ \beta_k = u - \frac{r^2}{\beta_{k-1}} \quad k = 1, 2, 3, \ldots, M; \quad \beta_1 = u \]

(3) \[ \gamma_k = -\frac{r}{\beta_k} \quad k = 1, 2, 3, \ldots, M; \]

(4) \[ z_k = \frac{1}{\beta_k} (b_k + rz_{k-1}) \quad k = 1, 2, 3, \ldots, M; \quad z_1 = \frac{b_1}{u} \]

(5) \[ y_k = z_k - \gamma_k y_{k+1} \quad k = 1, 2, 3, \ldots, M; \quad y_M = z_M \]

where we have written \( u \) for \( 2 + 2r \) and we have denoted the components of \( T_{n+1} \) 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 \( \beta_k \)'s and upper bounds of \( \gamma_k, z_k \) and \( y_k \). In (2) we put \( k = 2, 3, \ldots \) we get

(6) \[ \beta_2 = u - \frac{r^2}{\beta_1} = \beta_1 - \frac{r^2}{\beta_1} \]

\[ \vdots \]

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

(7) \[ r^2 -(2+2r)x + r^2 = 0. \]

Accordingly

(8) \[ \beta_* = 1 + r + \sqrt{1+2r}. \]

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

(9) \[ \gamma^* = \frac{r}{\beta_*} = \frac{r}{1 + r + \sqrt{1+2r}}. \]

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

\[ z_k = \frac{b_k}{\beta_k} + \frac{rb_{k-1}}{\beta_k \beta_{k-1}} + \frac{r^2b_{k-2}}{\beta_k \beta_{k-1} \beta_{k-2}} + \cdots + \frac{r^{k-1}b_1}{\beta_k \beta_{k-1} \cdots \beta_1} \]
whence
\[ |z_k| \leq \frac{b^*}{\beta_*} \left[ 1 + \frac{r}{\beta_*} + \frac{r^2}{\beta_*^2} + \cdots + \left( \frac{r}{\beta_*} \right)^{k-1} \right] \leq \frac{b^*}{\beta_*} \cdot \frac{1}{1 - \frac{r}{\beta_*} - r} = \frac{b^*}{\beta_* - r} \]

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
\[(10) \quad z^* = \frac{b^*}{\beta_* - r}.
\]

Finally from (5) we readily get
\[
y_M = z_M
\]
\[
y_{M-1} = z_{M-1} - \gamma_{M-1}z_M
\]
\[
y_{M-2} = z_{M-2} - \gamma_{M-2}z_{M-1} + \gamma_{M-3}y_{M-1}z_M
\]
\[\vdots\]
\[
y_1 = z_1 - \gamma_1z_2 + \gamma_1\gamma_2z_3 - \cdots (-1)^{M-1}\gamma_1\gamma_2 \cdots \gamma_{M-1}z_1.
\]

From the above system of equations it is clear that
\[(11) \quad y^* = z^*(1 + \gamma^* + \gamma^*2 \cdots + \gamma^{*M-1})
\]
\[
\geq \frac{z^*}{1 - \gamma^*} = \frac{b^*}{\beta_* - r} \cdot \frac{1}{r - \frac{r}{\beta_*}} = \frac{b^*\beta_*}{(\beta_* - r)^2}
\]
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 \( \beta_k \)'s, \( \gamma_k \)'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 \( \beta_k \) and by \( E^*(\beta) \) an upper bound of the errors in the \( \beta_k \)'s. A similar notation will be used for the \( \gamma_k \)'s, \( z_k \)'s and \( y_k \)'s. From (2) we have
\[
E(\beta_k) = \frac{r^2}{\beta_*^2} E(\beta_1) + \delta \leq \frac{r^2}{\beta_*^2} E(\beta_1) + o
\]
where \( \delta = \frac{1}{2} \times 10^{-p} \) is the maximum round-off error. Similarly
\[
E(\beta_k) = \frac{r^2}{\beta_*^2} E(\beta_2) + \delta \leq \frac{r^2}{\beta_*^2} E(\beta_2) + \delta
\]
\[
= \frac{r^2}{\beta_*^2} \left[ \frac{r^2}{\beta_*^2} E(\beta_1) + \delta \right] + \delta
\]
\[
= \left( 1 + \frac{r^2}{\beta_*^2} \right) \delta + \left( \frac{r^2}{\beta_*^2} \right)^2 E(\beta_1)
\]
Proceeding in this manner we ultimately get
\[
E(\beta_k) \leq \left[ 1 + \left( \frac{r^2}{\beta_*^2} \right) + \cdots + \left( \frac{r^2}{\beta_*^2} \right)^{M-1} \right] \delta + \left( \frac{r^2}{\beta_*^2} \right)^M E(\beta_1)
\]
\[
\leq \frac{1}{1 - \frac{r^2}{\beta_*^2}} \delta = \frac{\beta_*^2}{\beta_*^2 - r^2} \delta
\]
where we have neglected the second term of the above inequality since \( r < \beta_\ast \).

Thus

\[
E^\ast(\beta) = \frac{\beta_\ast^2}{\beta_\ast^2 - r^2} \delta
\]

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

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

\[
E(\gamma_k) = \frac{r}{\beta_\ast^2} E(\beta_k) + \delta
\]

\[
< \frac{r}{\beta_\ast^2} E^\ast(\beta) + \delta
\]

whence

\[
E^\ast(\gamma) = \frac{r}{\beta_\ast^2} E^\ast(\beta) + \delta = \frac{r}{\beta_\ast^2} \cdot \frac{\beta_\ast^2}{\beta_\ast^2 - r^2} \delta + \delta
\]

(13)

\[
= \left( 1 + \frac{r}{\beta_\ast^2 - r^2} \right) \delta.
\]

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

\[
E(z_k) = \frac{1}{\beta_\ast^2} \left\{ (b_k + rz_{k-1})E(\beta_k) + \beta_\ast [E(b_k) + rE(z_{k-1})] \right\} + \delta
\]

\[
\leq \frac{1}{\beta_\ast^2} (b^* + rz^*)E^\ast(\beta) + \frac{1}{\beta_\ast} E^\ast(b) + \frac{r}{\beta_\ast} E(z_{k-1}) + \delta
\]

\[
= \frac{1}{\beta_\ast^2} (b^* + rz^*) \delta + \frac{1}{\beta_\ast} E^\ast(b) + \frac{r}{\beta_\ast} E(z_{k-1}).
\]

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

(14)

\[
E^\ast(z) = \frac{\beta_\ast}{\beta_\ast - r} \left( 1 + \frac{b^* + rz^*}{\beta_\ast^2 - r^2} \right) \delta + \frac{E^\ast(b)}{\beta_\ast - r}.
\]

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

(15)

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

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

(15*)

\[
E^\ast(z) = \frac{\beta_\ast}{\beta_\ast - r} \left( 1 + \frac{b^* + rZ}{\beta_\ast^2 - r^2} \right) \delta + \frac{E^\ast(b)}{\beta_\ast - r}.
\]

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 \( E^\ast(y) \). From (5) we get

\[
E(y_k) = E(z_k) + \gamma_k E(y_{k+1}) + y_{k+1} E(\gamma_k) + \delta
\]
whence
\[ E(y_k) \leq E^*(z) + \gamma^* E(y_{k+1}) + y^* E^*(\gamma) + \delta. \]
Substituting for \( E^*(z), \gamma^*, y^* \) and \( E^*(\gamma) \) their expressions from (15), (9), (11), and (13) the last inequality becomes
\[
E(y_k) \leq \frac{\beta_*}{\beta_* - r} \left[ 1 + \frac{b^* \beta_*}{(\beta_* - r)(\beta_*^2 - r^2)} \right] \delta + \frac{E^*(b)}{\beta_* - r} \\
+ \frac{b^* \beta_*}{(\beta_* - r)^2} \left( 1 + \frac{r}{\beta_*^2 - r^2} \right) \delta + \delta + \frac{r}{\beta_*} E(y_{k+1}).
\]
Proceeding again as in the evaluation of \( E^*(\beta) \), the last inequality ultimately yields:
\[
E^*(y) = \frac{\beta_*}{\beta_* - r} \left[ 1 + \frac{b^* \beta_*}{(\beta_* - r)(\beta_*^2 - r^2)} \right] \delta + \frac{E^*(b)}{\beta_* - r} \\
+ \frac{b^* \beta_*}{(\beta_* - r)^2} \left( 1 + \frac{r}{\beta_*^2 - r^2} \right) + 1 \delta + \frac{r}{\beta_*} E(y_{k+1}).
\]
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 \( \gamma^* \) and \( E^*(\gamma) \) are replaced by their expressions from (9) and (13), we obtain as the counterpart of (17)
\[
E(y_k) \leq \frac{\beta_*}{\beta_* - r} \left[ 1 + \frac{b^* + rZ}{\beta_*^2 - r^2} \right] \delta + \frac{E^*(b)}{\beta_* - r} \\
+ Y \left( 1 + \frac{r}{\beta_*^2 - r^2} \right) \delta + \frac{r}{\beta_*} E(y_{k+1}).
\]
Proceeding again as in the evaluation of \( E^*(\beta) \), the last inequality ultimately yields:
\[
E^*(y) = \frac{\beta_*}{\beta_* - r} \left[ 1 + \frac{b^* + rZ}{\beta_*^2 - r^2} \right] \delta + \frac{Y \left( 1 + \frac{r}{\beta_*^2 - r^2} \right) + 1 \delta}{\beta_* - r} \\
+ \frac{\beta_*}{(\beta_* - r)^2} E^*(b).
\]
It will be convenient to rewrite the last equation in the form
\[
E^*(y) = S_0(r) + b^* S_1(r) + Y S_2(r) + Z S_3(r) + E^*(b) S_4(r)
\]
where
\[
\begin{align*}
S_0(r) &= \frac{\beta_*}{\beta_* - r} + \frac{\beta_*^2}{(\beta_* - r)^2} \\
S_1(r) &= \frac{\beta_*^2}{(\beta_* - r)^2(\beta_*^2 - r^2)} \\
S_2(r) &= \frac{\beta_*}{\beta_* - r} \left( 1 + \frac{r}{\beta_*^2 - r^2} \right) \\
S_3(r) &= \frac{r \beta_*}{(\beta_* - r)^2(\beta_*^2 - r^2)} = r S_1(r) \\
S_4(r) &= \frac{\beta_*}{(\beta_* - r)^2}.
\end{align*}
\]
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 $Ay = b$. 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 $b$ were computed from $Ay = b$ where $A$ is a $20 \times 20$ tridiagonal matrix of the type above considered with $r = 1$ and $r = 2$ and the 20 components of $y$ 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
