On the Operator Approach to the Problem of Stability
of Difference Schemes and the Convergence of Certain
Iteration Procedures

by

Arnold N. Lowan*

The University of California Radiation Laboratory
Livermore, California

(Work performed under the auspices of the Atomic Energy Commission)

In a number of articles which appeared in recent years** the authors have made use of the "operator approach" in dealing with the problem of stability of difference analogs of differential equations. The success of this method of approach depended on the fact that the eigenvalues of certain matrices could be determined and that an arbitrary M-dimensional vector could be expressed in terms of the eigenvectors of the matrices in question. More specifically whenever the eigenvalues could be shown to be smaller than unity in absolute value, the conclusion could be drawn that the difference scheme under consideration is stable.

Very frequently the desired aim may be reached without knowing the eigenvalues of the pertinent matrices.

It is the purpose of this article to discuss a number of problems dealing either with the stability of certain difference analogs of the differential equation of heat conduction or with the convergence of iteration

*On leave from Yeshiva University, New York, New York

**See "Bibliographic Notes" at the end of this article

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

procedures derived from implicit difference equations. We will make use of the operator approach in a manner which dispenses with the knowledge of the eigenvalues.

As a first illustration consider the explicit difference analog

$$(1) \quad u_{h,k+1} = u_{h,k} + r(u_{h-1,k} - 2u_{h,k} + u_{h+1,k})$$

of the differential equation

$$(2) \quad \frac{\partial T}{\partial t} = \sigma \frac{\partial^2 T}{\partial x^2}, \quad t > 0, \quad 0 \leq x \leq a$$

In (1), $u_{h,k} = u(h\Delta x, k\Delta t)$ and $r = \sigma \Delta t / (\Delta x)^2$. For the sake of concreteness assume that the solution of (2), and therefore that of (1) is required to assume prescribed values at $t=0$ and to vanish on the boundaries, so that $u_{0,k} = u_{M+1,k} = 0$ where $(M+1)\Delta x = a$. The system of equations derived from (1) by putting $h=1, 2, 3, \dots, M$ may be written in the compact form

$$(3) \quad \vec{u}_{k+1} = A \vec{u}_k$$

where $\vec{u}_k$ and $\vec{u}_{k+1}$ are the M -dimensional vectors whose components are $u_{h,k}$ and $u_{h,k+1}$ with $h=1, 2, 3, \dots, M$, respectively and where A is a matrix whose main diagonal elements are $= 1-2r$ and whose elements off the main diagonal in both directions are $= r$ all other elements being $= 0$.

From (3) we ultimately get

$$(4) \quad \vec{u}_n = A^n \vec{u}_0$$

where $\vec{u}_0$ is the "initial temperature vector". If the difference scheme (1) is to be stable $\vec{u}_n$ must tend to 0 as $n \rightarrow \infty$ and therefore $A^n \vec{u}_0$ must tend to 0 as $n \rightarrow \infty$ since, as is well known, the solution of the differential equation which assumes prescribed values for $t=0$ and vanishes for $x=0$ and $x=a$ tends to 0 as $t \rightarrow \infty$. It should be pointed out that this stability criterion coincides with the usual criterion based on the boundedness of the error in $\vec{u}_n$ induced by an initial error in $\vec{u}_0$. Indeed the error vector $\vec{E}$ evidently satisfies the equation $\vec{E}_n = A^n \vec{E}_0$ identical with (4).

Since the vector $\vec{u}_0$ does not vanish identically, we conclude that the condition of stability is that A^n tend to the zero matrix, Z , as $n \rightarrow \infty$. The question therefore arises: Under what conditions, and in particular for what values of the mesh-ratio r will A^n tend to Z as $n \rightarrow \infty$?

There are two important criteria, to be used in this article, for the asymptotic approach of A^n to Z as n approaches infinity.

I. Let S denote the largest in the sequence of $2n$ numbers $S_k = \sum_{h=1}^M |a_{hk}|$, $k=1,2,3,\dots,M$

and $S'_h = \sum_{k=1}^M |a_{hk}|$, $k=1,2,3,\dots,M$ where the a_{hk} 's are the elements of the matrix A . Let similarly $S^{(p)}$ denote the largest in the sequence of $2n$ number

$$S_k^{(p)} = \sum_{h=1}^M |a_{hk}^{(p)}| \quad \text{and} \quad S_h^{(p)} = \sum_{k=1}^M |a_{hk}^{(p)}| \quad \text{where the } a_{hk}^{(p)} \text{ are the elements}$$

of the p^{th} power of the matrix A . Then $S^{(p)} < S^p$. In the special case where $S \leq 1$, $S^{(p)} \rightarrow 0$ as $p \rightarrow \infty$ and the matrix A^p tends to Z as $p \rightarrow \infty$.

II. If the eigenvalues of a matrix are known to be numerically smaller than unity, then $A^p \rightarrow Z$ as $p \rightarrow \infty$.

The criterion under I follows from the fact that

$$\begin{aligned} \sum_{q=1}^M |a_{pq}^{(2)}| &= \sum_{q=1}^M \sum_{\alpha=1}^M |a_{p\alpha}| |a_{\alpha q}| \\ &= \sum_{\alpha=1}^M \left\{ |a_{p\alpha}| \sum_{q=1}^M |a_{\alpha q}| \right\} < S \sum_{\alpha=1}^M |a_{p\alpha}| < S^2 \end{aligned}$$

*See footnote on Page 16 of "Methoden der Mathematischen Physik", Vol. I (1931)

The criterion under II may readily be proved on the assumption that the eigenvalues are simple. Indeed let Λ denote the diagonal matrix whose diagonal elements are the eigenvalues of A . Let further U denote the matrix whose columns consist of the components of the eigenvectors corresponding to the various eigenvalues of A . Then it is obvious that $AU = U\Lambda$ so that $A = U\Lambda U^{-1}$ and therefore $A^n = U\Lambda^n U^{-1}$. If the eigenvalues have moduli smaller than unity it is evident that $\Lambda^n \rightarrow Z$ as $n \rightarrow \infty$ and therefore $A^n \rightarrow Z$ as $n \rightarrow \infty$.

If the eigenvalues of A are not all simple then it may be shown* that the counterpart of the above relation $A = U\Lambda U^{-1}$ is a relation of the form $A = W\Lambda^* W^{-1}$ where Λ^* is a quasi-diagonal matrix whose successive powers may be shown to tend to Z . Since $A^n = W\Lambda^{*n} W^{-1}$ it follows that the high powers of A equally approach the zero matrix.

Returning to (4) it is clear that if we take $r \leq 1/2$ then $S_1 = S_M = 1 - r < 1$ and $S_k = r + 1 - 2r + r = 1$ for $k = 2, 3, \dots, M-1$. Thus $S = 1$ and therefore by criterion (I) we have $A^n \rightarrow Z$ as $n \rightarrow \infty$. The stability of (1) is thus proven for $r \leq 1/2$.

If instead of requiring the temperature to vanish at the boundaries, we require its gradient to vanish then $u_{0,k} = u_{1,k}$ and $u_{M,k} = u_{M+1,k}$ and the counterpart of (3) is

$$\vec{u}_{k+1} = B \vec{u}_k$$

where the matrix B may be obtained from the matrix A by replacing the first and last elements of the main diagonal by $1-r$ while leaving all other elements unchanged. For the matrix B , the quantity S of criterion I is readily seen to be $= 1$. The stability of the difference scheme is thus proven for the case when the temperature gradient vanishes at the boundaries.

Finally if the boundary conditions for the solution of the differential equation (2) are $\frac{\partial T}{\partial x} = hT$ for $x=0$ and $\frac{\partial T}{\partial x} + hT = 0$ for $x=a$, the difference counterparts of these boundary conditions are $u_1 - u_0 = hu_0 \Delta x$ and $u_{M+1} - u_M = hu_{M+1} \Delta x$

*For details see chapter 10 of "The Mathematics of Physics and Chemistry" by H. Margenau and G. M. Murphy

and the counterpart of (3) is

$$\vec{u}_{k+1} = C \vec{u}_k$$

where the matrix C may be obtained from the matrix A by replacing the first and last elements on the main diagonal by $1-2r+\alpha r$ where $\alpha = 1/(1+h\Delta x) < 1$

For the matrix C it is once more seen that the quantity S of criterion I is $=1$ which proves the stability of the difference scheme (1) for the case of the last mentioned boundary conditions which obtain when the boundaries $x=0$ and $x=a$ radiate into a medium at 0° .

It may be briefly stated that if the boundary conditions for the solution of the differential equation (1) are $\frac{\partial T}{\partial x} - h_1 T = 0$ for $x=0$ and $\frac{\partial T}{\partial x} + h_2 T = 0$ for $x=a$ where $h_1 \neq h_2$ the difference counterparts of these boundary conditions are $u_1 - u_0 = h_1 u_0 \Delta x$ and $u_{m+1} - u_m = h_2 u_{m+1} \Delta x$ and the counterpart of (3) is

$$\vec{u}_{k+1} = D \vec{u}_k$$

where the matrix D may be obtained from the matrix A by replacing the first

(upper left) element on the main diagonal by $1-2r+\alpha_1 r$ where $\alpha_1 = 1/(1+h_1 \Delta x) < 1$ and the last (lower right) element of the main diagonal by $1-2r+\alpha_2 r$ where $\alpha_2 = 1/(1+h_2 \Delta x) < 1$. For the matrix D it is then easily seen that the quantity S of criterion I is $=1$ thus proving the stability of the difference scheme (1) for the boundary conditions under discussion.

The above analysis can be readily extended to the case of two-dimensional problems. If we put $\alpha = \sigma \Delta t / (\Delta x)^2$ and $\beta = \sigma \Delta t / (\Delta y)^2$, the counterpart of (1) is

$$(5): u_{h,k,n+1} = (1-\alpha-2\beta)u_{h,k,n} + \alpha(u_{h-1,k,n} + u_{h+1,k,n}) + \beta(u_{h,k-1,n} + u_{h,k+1,n})$$

where now $u_{h,k,n} = u(h\Delta x, k\Delta y, n\Delta t)$. For the sake of concreteness let it be

desired to obtain a solution of (5) vanishing on the boundary of the rectangle.

$0 \leq x \leq a, 0 \leq y \leq b$ and assuming prescribed values for $n=0$ (i.e. at time $t=0$)

at the mesh-points $x = h\Delta x, y = k\Delta y$ with $h=1,2,3,\dots,p, k=1,2,3,\dots,q$. It will be

convenient to consider the ordered set of pq quantities $u_{1,k,n}, u_{2,k,n}, \dots, u_{p,k,n}$

with $k=1,2,3,\dots,q$ as the coordinates of a pq -dimensional vector to be denoted

by $\vec{u}_n$. Similarly $\vec{u}_{n+1}$ will denote the $p \times q$ -dimensional vector with components $u_{h,k,n+1}$, the ordering of the components being identical with that for the vector $\vec{u}_n$, i.e. the first p components correspond to $k=1$, the second set of p components correspond to $k=2$, etc. Consistently with this ordering of the components, we rewrite the above equation in the form:

$$(5*) \quad u_{h,k,n+1} = \beta u_{h,k-1,n} + \alpha u_{h-1,k,n} + (1-2\alpha-2\beta) u_{h,k,n} + \alpha u_{h+1,k,n} + \beta u_{h,k+1,n}$$

If in the last equation we put in succession $k=1: h=1, 2, 3, \dots, p; k=2: h=1, 2, 3, \dots, p$ and so on up to $k=q: h=1, 2, 3, \dots, p$ we obtain a system of pq linear equations which may be written in the form $\vec{u}_{n+1} = A \vec{u}_n$ where A is the matrix of coefficients. It is important to note that the elements on the main diagonal of A are $= 1-2\alpha-2\beta$ and that each row of A contains in addition to its diagonal elements, two, three or four other non-vanishing elements which are either $= \alpha$ or $= \beta$.

The equation $\vec{u}_{n+1} = A \vec{u}_n$ ultimately yields $\vec{u}_n = A^n \vec{u}_0$ and the criterion for the stability of the above two-dimensional difference equation is that $A^n \rightarrow Z$ as $n \rightarrow \infty$. It is readily seen that if $\alpha + \beta < \frac{1}{2}$ the quantity S of criterion I is $= 1$. This proves the stability of the difference equation under consideration, in conformity with criterion I.

The extension of the above analysis to the explicit difference analog of the three-dimensional differential equation of heat conduction is obvious; the condition of stability is $\alpha + \beta + \delta < \frac{1}{2}$ where α and β have the same significance as above and $\delta = \sigma \Delta t / (\Delta z)^2$.

As a second illustration of the use of the operator approach in conjunction with criteria I and II, consider the two-dimensional counterpart of the Crank-Nicolson implicit difference scheme

$$(6) \quad u_{h,k,n+1} = u_{h,k,n} + \frac{\alpha}{2} (u_{h-1,k,n+1} - 2u_{h,k,n+1} + u_{h+1,k,n+1} + u_{h-1,k,n} - 2u_{h,k,n} + u_{h+1,k,n}) \\ + \frac{\beta}{2} (u_{h,k-1,n+1} - 2u_{h,k,n+1} + u_{h,k+1,n+1} + u_{h,k-1,n} - 2u_{h,k,n} + u_{h,k+1,n})$$

Consistently with the adopted ordering of the components, this equation is rewritten in the form

$$(7) \quad -\beta u_{h,k-1,n+1} - \alpha u_{h-1,k,n+1} + (2+2\alpha+2\beta) u_{h,k,n+1} - \alpha u_{h+1,k,n+1} - \beta u_{h,k+1,n+1}$$

$$= \beta u_{h,k-1,n} + \alpha u_{h-1,k,n} + (2-2\alpha-2\beta) u_{h,k,n} + \alpha u_{h+1,k,n} + \beta u_{h,k+1,n}$$

The last equation may be written in the matrix-vector form $A \vec{u}_{n+1} = B \vec{u}_n$ whence

ultimately $\vec{u}_{n+1} = (A^{-1}B)^n \vec{u}_0 = (4A^{-1} - I)^n \vec{u}_0 = A^{*n} \vec{u}_0$ (say). Let λ denote the

eigenvalues of A . Then the eigenvalues of A^* are $\frac{4}{\lambda} - 1$ and in conformity with criterion II, A^{*n} will converge to the zero matrix as $n \rightarrow \infty$ if $|\frac{4}{\lambda} - 1| < 1$ or

$\lambda > 2$. The proof that the eigenvalues of A are > 2 follows readily from a theorem by Alfred Brauer* which states that the eigenvalues of a matrix must satisfy the inequalities $|\lambda - a_{kk}| < P_k$, $k=1,2,3,\dots,n$ where n is the order of the matrix and $P_k = \sum_{\substack{h=1 \\ h \neq k}}^n |a_{hk}|$. In the present case it will be observed that

the elements on the main diagonal of A are $= 2+2\alpha+2\beta$ and that each row of A contains two, three or four non-vanishing elements which are either $=\alpha$ or $=\beta$, in addition to the diagonal element. Thus, in accordance with Brauer's theorem $|\lambda - 2-2\alpha-2\beta| < 2\alpha+2\beta$ whence $\lambda > 2$. The unconditional stability of the implicit scheme under consideration is thus proven.

Consider now the "high order accuracy" implicit scheme:

$$(8) \quad (10+12\alpha) u_{h,k+1} + (1-6\alpha)(u_{h-1,k+1} + u_{h+1,k+1}) = (10-12\alpha) u_{h,k} + (1+6\alpha)(u_{h-1,k} + u_{h+1,k})$$

or

$$(9) \quad A \vec{u}_{k+1} = B \vec{u}_k$$

*Limits for the characteristic roots of a matrix. Duke Math. Jour. Vol. 13 pp. 387-395 (1946)

whence

$$(10) \quad \vec{U}_{k+1} = A^{-1} B \vec{U}_k = A^* \vec{U}_k \quad (\text{say})$$

As before we shall prove that the eigenvalues of the matrix A^* are numerically smaller than unity. For this purpose we observe that

$$(11) \quad \begin{aligned} A &= (1-6\nu)C + (9+18\nu)I \\ B &= (1+6\nu)C + (9-18\nu)I \end{aligned}$$

where C is a matrix whose elements on the main diagonal are $= -1$, while the elements off the main diagonal are $= 1$, all other elements being $= 0$. If the eigenvalues of C are denoted by λ , those of A and B are $(1-6\nu)\lambda + 9 + 18\nu$ and $(1+6\nu)\lambda + 9 - 18\nu$ respectively and the condition that the eigenvalues of A^* in (10) be numerically smaller than unity leads to the double inequality

$$-1 < \frac{(1+6\nu)\lambda + 9 - 18\nu}{(1-6\nu)\lambda + 9 + 18\nu} < 1$$

whence $-9 < \lambda < 3$. In this case Brauer's theorem yields $|\lambda + 1| < 2$ or $-2 < \lambda + 1 < 2$ whence $-3 < \lambda < 1$ and a fortiori $-9 < \lambda < 3$. The stability of the difference scheme (8) is thus proven.

As a further illustration of the use of the operator approach in conjunction with the above criteria I and II we shall investigate the convergence of two iteration procedures derived from the implicit difference equation (5).

These procedures are:

$$(12) \quad u_{h,k+1}^{(p+1)} = u_{h,k} + \frac{\nu}{2} \left(u_{h-1,k+1}^{(p)} - 2u_{h,k+1}^{(p)} + u_{h+1,k+1}^{(p)} \right) + \frac{\nu}{2} \left(u_{h-1,k} - 2u_{h,k} + u_{h+1,k} \right)$$

$$(13) \quad u_{h,k+1}^{(p+1)} = u_{h,k} + \frac{\nu}{2} \left(u_{h-1,k+1}^{(p)} - 2u_{h,k+1}^{(p+1)} + u_{h+1,k+1}^{(p)} \right) + \frac{\nu}{2} \left(u_{h-1,k} - 2u_{h,k} + u_{h+1,k} \right)$$

Consider first the case of the iteration procedure (12) which may be written in the matrix-vector form

$$(14) \quad \vec{u}_{k+1}^{(p+1)} = A \vec{u}_{k+1}^{(p)} + B \vec{u}_k$$

where A and B are the matrices derived from (12) in the usual manner. With (14) we associate the original difference equation (5) which we now conveniently rewrite in the form

$$(15) \quad \vec{u}_{k+1} = A \vec{u}_{k+1} + B \vec{u}_k$$

From (14) and (15) we get

$$\vec{u}_{k+1} - \vec{u}_{k+1}^{(p+1)} = A (\vec{u}_{k+1} - \vec{u}_{k+1}^{(p)})$$

whence ultimately

$$(16) \quad \vec{u}_{k+1} - \vec{u}_{k+1}^{(p+1)} = A^{p+1} (\vec{u}_{k+1} - \vec{u}_{k+1}^{(0)})$$

It is clear from (16) that $\vec{u}_{k+1}^{(p+1)}$ will converge to $\vec{u}_{k+1}$ if it can be proven that $A^n \rightarrow Z$ as $n \rightarrow \infty$. In the present case the quantity S of criterion I is evidently $\lambda + \frac{\lambda}{2} + \frac{\lambda}{2} = 2\lambda$. Thus if $\lambda \leq 1/2$, $S \leq 1$ and A^n tends to the zero matrix as $n \rightarrow \infty$, in conformity with criterion I, thus proving the convergence of the iteration procedure (12).

Consider now the iteration procedure (13). The counterpart of (16) is now

$$\vec{u}_{k+1} - \vec{u}_{k+1}^{(p+1)} = A^{*p+1} (\vec{u}_{k+1} - \vec{u}_{k+1}^0)$$

where now, as is apparent from (13), A^* is a matrix whose elements on the

main diagonal are $\neq 0$ and whose elements off the main diagonal in both directions are $= \frac{\nu}{2+2\nu}$, all other elements being $= 0$. Thus in this case $S = \frac{\nu}{2+2\nu} + \frac{\nu}{2+2\nu} = \frac{\nu}{\nu+1} < 1$. In conformity with criterion I we conclude that $A^n \rightarrow Z$ as $n \rightarrow \infty$, thus proving the unconditional convergence of the iteration procedure (13).

The above analysis may be extended to the following two iteration procedures which may be called procedures of Liebmann type:

$$(17) \quad u_{h,k+1}^{p+1} = u_{h,k} + \frac{\nu}{2} \left(u_{h-1,k+1}^{p+1} - 2u_{h,k+1}^{p+1} + u_{h+1,k+1}^p + u_{h-1,k} - 2u_{h,k} + u_{h+1,k} \right)$$

$$(18) \quad u_{h,k+1}^{p+1} = u_{h,k} + \frac{\nu}{2} \left(u_{h-1,k+1}^{p+1} - u_{h,k+1}^p - u_{h,k+1}^{p+1} + u_{h+1,k+1}^p + u_{h-1,k} - 2u_{h,k} + u_{h+1,k} \right)$$

In order to appreciate the distinction between (17) and (18) on the one hand and (12) and (13) on the other hand, it is necessary to stipulate the order in which the successive approximations to the components of the vector $\vec{u}_{k+1}$ are computed. It is reasonable to assume that the order is that corresponding to $h=1, h=2, \dots, h=M$. It is then clear from (12) and (13) that each "new" value $u_{h,k+1}^{p+1}$ is expressed in terms of "old" values of $u_{h,k+1}$ (i.e. values characterized by the superscript p). In (17) and (18) however by the time the new value of $u_{h,k+1}$ is computed, the new value of $u_{h-1,k+1}$ is known and therefore this new value is used in the evaluation of $u_{h,k+1}^{p+1}$. This is the significance of using $u_{h-1,k+1}^{p+1}$ in lieu of $u_{h-1,k+1}^p$ in (17) and (18).

If in (17) we put $h=1, 2, 3, \dots, M$ the resulting system of M linear equations may be solved for the unknowns $u_{h,k+1}^{p+1}$ in terms of the quantities $u_{h,k+1}^p$ and $u_{h,k}$. Specifically

$$(19) \quad \vec{u}_{k+1}^{p+1} = A \vec{u}_{k+1}^p + B \vec{u}_k$$

where

$$(20) \quad A = \begin{bmatrix} 0 & \rho & & & & \\ 0 & \rho^2 & \rho & & & \\ 0 & \rho^3 & \rho^2 & \rho & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \rho^{M-1} & \rho^{M-2} & \rho^{M-3} & \dots & \rho^2 & \rho \\ 0 & \rho^M & \rho^{M-1} & \rho^{M-2} & \dots & \rho^3 & \rho^2 \end{bmatrix} \quad \rho = \frac{\nu}{2+2N}$$

The expression for the matrix B in (19) may be found but is not required in the subsequent developments. Indeed, if in (19) we omit the superscripts p and p+1, the resulting equation is the original Crank-Nicolson difference equation. Its matrix-vector form must be

$$(21) \quad \vec{u}_{k+1} = A \vec{u}_{k+1} + B \vec{u}_k$$

derived from (19) by omitting the superscripts p and p+1. From (19) and (21) we obtain

$$\vec{u}_{k+1}^{p+1} - \vec{u}_{k+1}^p = A (\vec{u}_{k+1}^p - \vec{u}_{k+1}^0)$$

whence ultimately

$$(22) \quad \vec{u}_{k+1}^p - \vec{u}_{k+1}^0 = A^p (\vec{u}_{k+1}^0 - \vec{u}_{k+1}^0)$$

where $\vec{u}_{k+1}^0$ is a first approximation to $\vec{u}_{k+1}$. Since $\rho < 1$ it is clear that the quantity S of criterion I is given by

$$(23) \quad S = \rho^{M-1} + \rho^{M-2} + \dots + \rho = \frac{\rho - \rho^M}{1 - \rho} \approx \frac{\rho}{1 - \rho} = \frac{\nu}{2 + \nu} < 1$$

Since $S < 1$ it follows that $A^n \rightarrow Z$ as $n \rightarrow \infty$ in conformity with criterion I. We conclude that the iteration procedure (17) is unconditionally convergent. Moreover since the value of S appropriate to (17) is smaller than the value of S appropriate to (13) it may be anticipated that the convergence of (17) is faster than that of (13).

We now turn to iteration procedure (18). Proceeding as above we ultimately

get:

$$\vec{u}_{k+1}^{p+1} = A^* \vec{u}_{k+1}^p + B^* \vec{u}_k^p$$

where

$$(20^*) \quad A^* = \begin{bmatrix} -\rho & \rho & & & \\ -\rho^2 & \rho^2 & -\rho & \rho & \\ -\rho^3 & \rho^3 & -\rho^2 & \rho^2 & -\rho & \rho \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -\rho^{M-1} & \rho^{M-1} & -\rho^{M-2} & \rho^{M-2} & -\rho^{M-3} & \dots & \rho \\ -\rho^M & \rho^M & -\rho^{M-1} & \rho^{M-1} & -\rho^{M-2} & \dots & \rho^2 & -\rho \end{bmatrix} \quad \rho = \frac{\nu}{2+2\nu}$$

Since $\rho < 1$, the quantities $\rho^2 - \rho, \rho^3 - \rho^2, \dots$ etc. are all negative. It is then readily seen that the sum of the absolute values of the elements of the first $M-1$ rows is $= 2\rho$ while the corresponding sum of the last row is $= \rho$. Thus S of criterion I is $= 2\rho = \frac{\nu}{2+\nu} < 1$, which proves the convergence of (18). It may be anticipated that the convergence of (18) is slower than that of (17) since the value of S appropriate to (18) is larger than that for (17).

Consider now the convergence of the so-called Richardson iteration procedure for the following difference analog of the Laplace differential equation:

$$(24) \quad u_{h,k-1} + u_{h-1,k} - 4u_{h,k} + u_{h+1,k} + u_{h,k+1} = 0$$

The Richardson iteration procedure appropriate to (24) is represented by the formula

$$(25) \quad u_{h,k}^{n+1} = u_{h,k}^n + \alpha (u_{h,k-1}^n + u_{h-1,k}^n - 4u_{h,k}^n + u_{h+1,k}^n + u_{h,k+1}^n)$$

where α is an undetermined multiplier. As in the case of the above two-dimensional problem in heat conduction, we shall confine our attention to finding a solution of (24) for the rectangle $0 \leq x \leq a, 0 \leq y \leq b$, the value of u being prescribed on the boundary.

From (24) and (25) it follows that

$$(26) \quad E_{h,k}^{n+1} = E_{h,k}^n + \alpha (E_{h,k-1}^n + E_{h-1,k}^n - 4E_{h,k}^n + E_{h+1,k}^n + E_{h,k+1}^n)$$

where $E_{h,k}^n = u_{h,k}^n - u_{h,k}$ and where the $u_{h,k}$'s are the exact solutions of (24). The last equation may be written in the matrix vector form $\vec{E}_{n+1} = (I + \alpha L) \vec{E}_n$ and therefore ultimately $\vec{E}_n = (I + \alpha L)^n \vec{E}_0$. It is readily ascertained that in the matrix $I + \alpha L$ the diagonal elements are $= 1 - 4\alpha$ and that each row contains two, three or four additional non-vanishing elements, all equal to α , the remaining elements being $= 0$. If then $\alpha \leq 1/4$ the quantity S of criterion I is $= 1$. We conclude that $(I + \alpha L)^n \rightarrow Z$ as $n \rightarrow \infty$. The convergence of the Richardson process is thus proven.

Consider now the "Liebmann iteration procedure" represented by the formula

$$(27) \quad u_{h,k}^{n+1} = (1 - 4\alpha) u_{h,k}^n + \alpha (u_{h,k-1}^{n+1} + u_{h-1,k}^{n+1} - 4u_{h,k}^n + u_{h+1,k}^n + u_{h,k+1}^n)$$

which bears the same relation to (25) as (17) bears to (13). As above if $E_{h,k}^n = u_{h,k}^n - u_{h,k}$ where the $u_{h,k}$'s are the exact solutions of (24), then the $E_{h,k}^n$'s satisfy equation (27). When equation (27) written out for the $E_{h,k}^n$'s is applied systematically in the order $\kappa=1: h=1, 2, 3, \dots, p; \kappa=2: h=1, 2, 3, \dots, p$ and so on up to $\kappa=q: h=1, 2, 3, \dots, p$ we obtain a system of pq linear equations expressing the "new coordinates" $E_{h,k}^{n+1}$ defining the pq -dimensional vector $\vec{E}_{n+1}$ in terms of the "old coordinates" $E_{h,k}^n$ defining the pq -dimensional vector $\vec{E}_n$. Let K be the matrix of the system Ω . Then $\vec{E}_{n+1} = K \vec{E}_n$ whence ultimately $\vec{E}_n = K^n \vec{E}_0$. The criterion for the convergence of the Liebmann iteration procedure (27) is that K^n tend to the zero matrix as $n \rightarrow \infty$.

Let $S_{h,k}$ denote the sum of the absolute values of the coefficients in the expression of $U_{h,k}^{n+1}$ (or $E_{h,k}^{n+1}$) in terms of the $U_{h,k}^n$'s (or $E_{h,k}^n$'s). Then, starting with (27), it can be shown that, if $1-4\alpha \geq 0$ or $\alpha \leq 1/4$, the $S_{h,k}$'s satisfy the equation

$$(28) \quad S_{h,k} = \alpha S_{h-1,k} + \alpha S_{h,k-1} + 1-2\alpha$$

If the pq equations in the system Ω are ordered in q consecutive groups corresponding to $k=1,2,3,\dots,q$, each group consisting of p equations corresponding to $h=1,2,3,\dots,p$, then (28), states that the value of S associated with the h -th equation in the k -th group is $= \alpha$ times the value of S associated with the preceding equation in the same group $+ \alpha$ times the value of S for the h -th equation in the preceding group, augmented by the positive quantity $1-2\alpha$. The quantity $S_{h,k}$ above defined is evidently the sum of the absolute values of the elements in the $((k-1)p+h)$ -th row of the matrix K .

By applying (27) to the derivation of the first p equations in the system Ω , it is readily ascertained that the largest S is that for the $(p-1)$ st equation and that

$$\begin{aligned} S_{p-1,1} &= (1-4\alpha)(1+\alpha+\alpha^2+\dots+\alpha^{p-2}) + 2(\alpha+\alpha^2+\dots+\alpha^{p-1}) \\ &= \frac{1-4\alpha}{1-\alpha} + \frac{2\alpha}{1-\alpha} = \frac{1-2\alpha}{1-\alpha} \end{aligned}$$

If in (28) we put $h=1,2,3,\dots,p-1$ we get

$$\begin{aligned} S_{1,k} &= \alpha S_{1,k-1} + 1-2\alpha \\ S_{2,k} &= \alpha S_{1,k} + \alpha S_{2,k-1} + 1-2\alpha \\ S_{3,k} &= \alpha S_{2,k} + \alpha S_{3,k-1} + 1-2\alpha \\ &\vdots \\ S_{p-1,k} &= \alpha S_{p-2,k} + \alpha S_{p-1,k-1} + 1-2\alpha \end{aligned}$$

whence

$$S_{p-1,k} = \alpha S_{p-1,k-1} + \alpha^2 S_{p-2,k-1} + \dots + \alpha^{p-1} S_{1,k-1} + (1-2\alpha)(1+\alpha+\alpha^2+\dots+\alpha^{p-2})$$

We have seen that in the first group of p equations the largest S is that associated with the $(p-1)$ st equation. It may be shown that this is true of each of the successive groups of p consecutive equations. Accordingly the last equation yields

$$S_{p-1} < (\alpha + \alpha^2 + \dots + \alpha^{p-1}) S_{p-1, k-1} + (1-2\alpha)(1 + \alpha + \alpha^2 + \dots + \alpha^{p-2})$$

and a fortiori

$$S_k < \frac{\alpha}{1-\alpha} S_{k-1} + \frac{1-2\alpha}{1-\alpha}$$

where we have written S_k and S_{k-1} for $S_{p-1, k}$ and $S_{p-1, k-1}$ respectively. If we put $A = \frac{\alpha}{1-\alpha}$ and $B = \frac{1-2\alpha}{1-\alpha}$ the last inequality yields

$$S_2 < AS_1 + B$$

$$S_3 < AS_2 + B \leq A(AS_1 + B) + B = A^2 S_1 + AB + B$$

and in general

$$S_q \leq A^{q-1} S_1 + B(1 + A + A^2 + \dots + A^{q-2})$$

For $\alpha < 1/4$ we have $A = \frac{\alpha}{1-\alpha} < 1$, so that $A^{q-1} \rightarrow 0$ as q gets larger and larger and the last inequality yields

$$S_q \leq \frac{B}{1-A} = \left(\frac{1-2\alpha}{1-\alpha} \right) / \left(1 - \frac{\alpha}{1-\alpha} \right) = 1$$

We have thus proved that for the matrix K the value of S in criterion I is smaller than unity. Accordingly K^n tends to the zero matrix as $n \rightarrow \infty$ in conformity with criterion I. The convergence of the Liebmann iteration procedure is thus proven. The last inequality suggests in fact that the convergence of the Liebmann process is faster than that of the Richardson process for which the corresponding value of S is equal to unity.

Radiation Laboratory
Livermore, California
March, 1957

Bibliographic Notes

- S. P. Frankel: Convergence rates of iterative treatments of partial differential equations, MTAC, 4, 1950 pp. 65-75.
- A. R. Mitchell: Round-off errors in the solution of the heat conduction equation by relaxational solution Appl. Sci. Res. A4, 1953 pp. 109-119.
- Round-off errors in relaxational solutions of Poisson's equation. Appl. Sci. Res. B3, 1954 pp. 456-464.
- Round-off errors in implicit finite difference methods Quart Jour. of Mech. & Appl. Math. Vol IX Part I Mar. 1956.
- Jim Douglas, Jr.: On the relation between stability and convergence in the numerical solution of linear parabolic and hyperbolic differential equations. Jour. Soc. Indust. Appl. Math. Vol 4 No. 1 March 1956.
- P. D. Lax and R. D. Richtmyer: Survey of the stability of linear finite difference equations. Communications on pure and applied mathematics. Vol IX No. 2 May 1956.
- John Todd: A direct approach to the problem of stability in the numerical solution of partial differential equations. Communications on pure and applied Mathematics Aug. 1956.