ON THE THEORY OF RELAXATION

by A. R. MITCHELL *and* D. E. RUTHERFORD

(Received 4th March, 1952)

§ 1. When a numerical method of obtaining an approximate solution of a linear differential equation is employed, the process involves two distinct types of approximation. The region of integration having been covered with a regular net, the differential equation and the appropriate boundary conditions are replaced by finite difference equations which are linear equations in the values of the dependent variable at the nodes of the net.

The exact solution V of these finite difference equations is, at best, only an approximation to the values taken at the nodes by the true solution D of the differential equation. Whether or not V converges to D as the mesh becomes infinitely fine, may depend upon the finite difference approximations employed and also upon the shape of the mesh used. The conditions for the convergence of V to D, in certain important cases, have been discussed by Courant, Friedrichs and Lewy (**1**), by O'Brien, Hyman and Kaplan (**2**), and by Thomas (**3**). Assuming that the conditions for convergence are satisfied, an approximation to D of any required degree of accuracy can, in principle, be obtained by choosing the mesh of the net sufficiently small. The computation of V may, however, be so laborious as to render this procedure impracticable. In such a case a second type of approximation must be employed. This is an iterative process for finding an approximate solution N of the finite difference equations. When N, which is an approximation to V, is obtained by a step by step process it may happen that small round-off errors, introduced at any stage, give rise to large errors in the final solution N. In this case the finite difference equations are said to be unstable. Rutishauser (**4**) has recently given a criterion for the stability of ordinary differential equations, while some partial differential equations are examined for stability in reference (**2**). The relaxation method, on the other hand, is one which forces N to tend to V. Consequently, instability will not entirely vitiate this method unless the round-off errors are introduced before the relaxation commences, although it may result in the slow convergence of N to V. In the present paper we are concerned with the relaxation method of obtaining good approximations N to V rather than with the convergence of V to D.

The fact is sometimes overlooked that many finite difference approximations of varying degrees of accuracy are available for any given derivative of a dependent variable u. Many of these, including the following

$$u'(0) = \frac{-u(0)+u(a)}{a} + a(\quad) \qquad \ldots\ldots(1)$$

$$= \frac{-u(-a)+u(0)}{a} + a(\quad) \qquad \ldots\ldots(2)$$

$$= \frac{-3u(0)+4u(a)-u(2a)}{2a} + a^2(\quad) \qquad \ldots\ldots(3)$$

$$= \frac{-u(-a)+u(a)}{2a} + a^2(\quad) \qquad \ldots\ldots(4)$$

$$= \frac{+u(-2a)-4u(-a)+3u(0)}{2a} + a^2(\quad) \qquad \ldots\ldots(5)$$

$$= \frac{-11u(0)+18u(a)-9u(2a)+2u(3a)}{6a} + a^3(\quad) \qquad \ldots\ldots(6)$$

have been tabulated by Bickley (**5**, **6**). Other approximations such as

$$u''(0) = \frac{2}{3} \cdot \frac{u(-a) - u(0) + au'(a)}{a^2} + a(\quad) \quad \text{.....(7)}$$

$$= \frac{2}{11} \cdot \frac{u(-2a) + u(-a) - 2u(0) + 3au'(a)}{a^2} + a^2(\quad) \quad \text{.....(8)}$$

may be used at nodes on or near a boundary on which certain derivatives of u are specified.

In the light of the foregoing discussion it will now be apparent that if the relaxation net has n nodes at which the values of u are to be determined, and if the boundary conditions are sufficient to guarantee a unique solution, there will be no inherent difficulty in obtaining n finite difference equations in the values taken by u at the n nodes. It is assumed, of course, that the mesh has been chosen fine enough for the order of the differential equation and the degree of accuracy desired. These equations can therefore be written in the form

$$\mathbf{Av} = \mathbf{h}$$

where $\mathbf{A}$ is a square matrix of order n with known elements, $\mathbf{h}$ is a column vector of known elements and $\mathbf{v}$ is a column vector whose components $v_1, v_2, \ldots v_n$ are approximations to the values of u taken at the nodes of the net.

In constructing the matrix $\mathbf{A}$ it is, of course, desirable to use as few types of finite difference approximation to the differential equation as possible. This will ensure that $\mathbf{A}$ has a regular structure and facilitates subsequent calculations. One must be careful also to ensure that all the given boundary conditions are incorporated and that the resulting matrix $\mathbf{A}$ is not the direct sum of two or more submatrices. Certain other precautions should be taken to ensure that the equations are convergent and, if, possible, also stable.

§ 2. We now consider the second process of approximation encountered in the relaxation method, that of approximating to the solution of the matrix equation

$$\mathbf{Av} = \mathbf{h}.$$

At best, the exact solution of this equation only gives approximations to the values of u at the several nodes. Only in a very few simple cases, e.g., when $\mathbf{A}$ is one of the matrices $P_n(x)$ or $T_n(z, a)$ of reference (**7**) can the inverse matrix $\mathbf{A}^{-1}$ be written down explicitly for any given size of mesh, but an exact step by step solution can be obtained if the matrix $\mathbf{A}$ is triangular, having all its elements above (below) the leading diagonal zero. Failing these two contingencies, we can use the relaxation method to obtain an approximate numerical solution. We shall find it convenient to denote successive approximations by upper suffixes. Let $\mathbf{v}^1$ be a trial solution and set

$$\mathbf{h} - \mathbf{Av}^1 = \mathbf{c}^1.$$

If column vectors $\mathbf{y}^1, \mathbf{y}^2, \ldots$ can be found such that the lengths of the column vectors $\mathbf{c}^1, \mathbf{c}^2, \ldots$ defined by

$$\mathbf{c}^1 - \mathbf{Ay}^1 = \mathbf{c}^2,$$
$$\mathbf{c}^2 - \mathbf{Ay}^2 = \mathbf{c}^3,$$
$$\ldots\ldots\ldots\ldots$$

tend monotonically to zero, then the vectors

$$\mathbf{v}^1, \mathbf{v}^2 = \mathbf{v}^1 + \mathbf{y}^1, \mathbf{v}^3 = \mathbf{v}^2 + \mathbf{y}^2, \ldots$$

tend to the solution $\mathbf{v}$ of $\mathbf{Av} = \mathbf{h}$. For clearly,

$$\mathbf{h} - \mathbf{Av}^2 = \mathbf{h} - \mathbf{Av}^1 - \mathbf{Ay}^1 = \mathbf{c}^1 + (\mathbf{c}^2 - \mathbf{c}^1) = \mathbf{c}^2,$$
$$\mathbf{h} - \mathbf{Av}^3 = \mathbf{h} - \mathbf{Av}^2 - \mathbf{Ay}^2 = \mathbf{c}^2 + (\mathbf{c}^3 - \mathbf{c}^2) = \mathbf{c}^3,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

In the usual relaxation method each vector $\mathbf{y}^i$ is chosen to have only one non-zero element, say, that in the jth row. That is to say, employing the Kronecker delta, the n components $y_r{}^i$ of $\mathbf{y}^i$ are given by

$$y_r{}^i = \delta_{rj} k_j{}^i. \quad \ldots\ldots(9)$$

We shall determine the best value to be chosen for the number $k_j{}^i$. The components $c_r{}^{i+1}$ of the $(i+1)$th residual vector $\mathbf{c}^{i+1}$ are found to be

$$c_r{}^{i+1} = c_r{}^i - a_{rj} k_j{}^i.$$

Denoting the length of the vector $\mathbf{c}^i$ by $|\,\mathbf{c}^i\,|$, we find that

$$\begin{aligned} |\,\mathbf{c}^{i+1}\,|^2 &= \sum_r (c_r{}^i - a_{rj} k_j{}^i)^2 \\ &= |\,\mathbf{c}^i\,|^2 - 2k_j{}^i \sum_r c_r{}^i a_{rj} + (k_j{}^i)^2 \sum_r (a_{rj})^2. \end{aligned}$$

Thus $|\,\mathbf{c}^{i+1}\,| < |\,\mathbf{c}^i\,|$ if

$$2k_j{}^i \sum_r c_r{}^i a_{rj} - (k_j{}^i)^2 \sum_r (a_{rj})^2 > 0\,; \quad \ldots\ldots(10)$$

that is to say, if

$$0 \leqslant k_j{}^i \leqslant 2 \sum_r c_r{}^i a_{rj} / \sum_r (a_{rj})^2. \quad \ldots\ldots(11)$$

The left-hand side of (10) has its maximum when

$$k_j{}^i = \sum_r c_r{}^i a_{rj} / \sum_r (a_{rj})^2 \quad \ldots\ldots(12)$$

and this maximum has the value

$$\left(\sum_r c_r{}^i a_{rj}\right)^2 / \sum_r (a_{rj})^2 \quad \ldots\ldots(13)$$

which may be written

$$(k_j{}^i)^2 \sum_r (a_{rj})^2, \quad \ldots\ldots(14)$$

where $k_j{}^i$ now has the value determined by (12). Thus, by giving $k_j{}^i$ any value satisfying (11) we can ensure that $|\,\mathbf{c}^{i+1}\,| < |\,\mathbf{c}^i\,|$. To ensure the maximum reduction in the length of the residual vector, $k_j{}^i$ must be given the value stated in (12). Since in most relaxation problems no column of $\mathbf{A}$ has more than a few non-zero elements, the calculation of $\sum_r c_r{}^i a_{rj}$ and of $\sum_r (a_{rj})^2$ will not be difficult. The labour, however, would be considerable if this method were employed in the case of a matrix $\mathbf{A}$ of large order possessing very few vanishing elements.

If it happens that $\sum_r c_r{}^i a_{rj} = 0$, then $\mathbf{y}^i$ is the zero vector and no reduction in $\mathbf{c}^i$ is possible by means of a vector $\mathbf{y}^i$ given by (9). At any stage, however, there are n possible choices of j and at least one of these will yield a vector $\mathbf{y}^i$ which reduces the length of the residual vector, unless for all j

$$\sum_r c_r{}^i a_{rj} = 0.$$

Assuming, however, that $\mathbf{A}$ is non-singular, these equations imply that $\mathbf{c}^i$ is the zero vector and that we have already reached the exact solution $\mathbf{v} = \mathbf{v}^i$. It follows that the length of the residual vector can always be reduced by the foregoing method provided the exact solution has not already been reached. If we wish to choose the value of j which gives the greatest reduction at the ith step we must choose the j which gives (13) the largest value. In practice however it will be sufficient to choose j so that (13) has a large, but not necessarily the largest, value. Even this provision may be dispensed with if each j is chosen in turn according to some plan.

https://doi.org/10.1017/S2040618500035541 Published online by Cambridge University Press

The above method for obtaining N resembles, and is related to, that systematised by Temple (**8**) and, more recently, by Stiefel (**9**). The original relaxation method of Southwell was based upon the analogy of an elastic framework subjected to prescribed loads at the joints. At each stage in the approximation the displacement at one joint is modified in such a way that the potential energy of the system is diminished. Temple showed that this procedure is equivalent to altering one component of the vector $\mathbf{v}$ at a time in such a way as to reduce the value of $\mathbf{v}'(\frac{1}{2}\mathbf{A}\mathbf{v}-\mathbf{h})$, $\mathbf{v}'$ being the transpose of the vector $\mathbf{v}$. This method can, in fact, be applied to any system of equations $\mathbf{A}\mathbf{v}-\mathbf{h}=0$, provided the matrix $\mathbf{A}$ is symmetric and positive definite. If $\mathbf{A}$ does not satisfy these conditions, Temple's method can still be applied to the equivalent equations $\mathbf{A}'\mathbf{A}\mathbf{v}-\mathbf{A}'\mathbf{h}=0$, which have been "prepared" by premultiplying the original equations by the transposed matrix $\mathbf{A}'$. The modifications are then made with a view to minimising the expression $\mathbf{v}'\mathbf{A}'(\frac{1}{2}\mathbf{A}\mathbf{v}-\mathbf{h})$. In contrast, the method which we have described in the present paper minimises the expression $(\mathbf{v}'\mathbf{A}'-\mathbf{h}')(\mathbf{A}\mathbf{v}-\mathbf{h})$, which is the square of the length of the residual vector. Since $(\mathbf{v}'\mathbf{A}'-\mathbf{h}')(\mathbf{A}\mathbf{v}-\mathbf{h})=2\mathbf{v}'\mathbf{A}'(\frac{1}{2}\mathbf{A}\mathbf{v}-\mathbf{h})+\mathbf{h}'\mathbf{h}$, the method here advocated must be equivalent to Temple's method applied to the prepared equation in the sense that the modification of $\mathbf{v}$ at any stage will be the same. The intermediate numerical calculations, however, will be different in the two cases and the present method has the advantage that it can be applied directly without any preparation, whenever the matrix $\mathbf{A}$ is non-singular. Furthermore, $\mathbf{A}$ is usually simpler than the prepared matrix $\mathbf{A}'\mathbf{A}$ in the sense that the former is likely to have more zero elements than the latter.

In many relaxation problems the diagonal elements of the matrix $\mathbf{A}$ dominate the others in the sense that a diagonal element has a much larger absolute value than the other elements in its column. If this is the case then c_j^i/a_{jj} will be a good approximation to the optimum value of k_j given by (12). This, indeed, coincides with the value suggested by Fox (10) and by Temple (8). Fox also suggests selecting the value of j which makes $|c_j^i/a_{jj}|$ a maximum; but it would appear from (14) that a better guess at the best j in the case under consideration would be that which makes $|c_j^i|$ itself a maximum, as originally suggested by Southwell (**11**, p. 47 and p. 63).

If, however, the diagonal elements of $\mathbf{A}$ are of smaller magnitudes than some other neighbouring ones, the choice $k_j^i=c_j^i/a_{jj}$ may be a bad one for our method, as may also both Fox's and Southwell's suggestions for the choice of j. In such cases it would be wise to use the formulae (12) and (13).

We have seen that at any intermediate stage, at which the residual vector is $\mathbf{c}^i$, an improvement can be effected so that $|\mathbf{c}^{i+1}|<|\mathbf{c}^i|$. We now verify that, with the choice of $\mathbf{y}^i$ mentioned, $\lim_{i\to\infty}|\mathbf{c}^i|=0$.

We can write (12) in the form

$$\sum_r c_r^i a_{rj}=k_j^i\sum_r (a_{rj})^2, \qquad j=1,\ldots,n,$$

and consequently, denoting the transpose of $\mathbf{A}$ by $\mathbf{A}'$, $\mathbf{A}'\mathbf{c}^i$ is the column vector whose jth component is $k_j^i\sum_r (a_{rj})^2$. On solving these equations for the component residuals c_s^i we find that

$$c_s^i=\sum_j A_{js}k_j^i\sum_r (a_{rj})^2$$

where A_{js} is the (j,s)th element of $\mathbf{A}^{-1}$. Hence, if $|k_j^i|<\eta$ for all j, it follows that

$$|c_s^i|<\eta\sum_j|A_{js}|\,\Big|\sum_r (a_{rj})^2\Big|$$

https://doi.org/10.1017/S2040618500035541 Published online by Cambridge University Press

for all s. That is to say,

$$|\, c_s^{\,i} \,| < \eta\alpha$$

where α depends upon $\mathbf{A}$ only. An immediate consequence is that

$$|\, \mathbf{c}^i \,| < \eta\alpha\sqrt{n}.$$

Now $|\, \mathbf{c}^i \,|$ decreases monotonically and tends to a lower limit l. Consequently, for any $\epsilon > 0$ we can find an i such that

$$|\, \mathbf{c}^i \,|^2 - l^2 < \epsilon.$$

But we could reduce the value of $|\, \mathbf{c}^i \,|^2$ by ϵ and so contradict the statement that l is the lower limit of $|\, \mathbf{c}^i \,|$ unless, by (14),

$$(k_i^{\,i})^2 \Sigma\, (a_{rj})^2 < \epsilon$$

for all j; that is to say, unless

$$(k_j^{\,i})^2 < \epsilon / \min_j \Sigma_r\, (a_{rj})^2.$$

for all j. In this case

$$l^2 \leqslant |\, \mathbf{c}^i \,|^2 < \frac{n\alpha^2\epsilon}{\min_j \Sigma_r\, (a_{rj})^2}$$

for any $\epsilon > 0$. Thus $l = 0$ and $|\, \mathbf{c}^i \,| \to 0$.

§ 3. Although the approximation process which we have described is certainly convergent, in some cases this convergence may be very slow. This is no doubt associated with the instability of the finite difference approximation used. It has already been pointed out by Fox (**10**), that this is a consequence of the matrix $\mathbf{A}$ being afflicted by the rather vague malady called ill-conditioning. To study this phenomenon we consider {the maximum possible reduction of $|\, \mathbf{c}^i \,|^2$} divided by $|\, \mathbf{c}^i \,|^2$. The expression for this measure of convergence at the ith stage is

$$q_i = \max_j \frac{(\Sigma_r\, c_r^{\,i} a_{rj})^2}{\Sigma_r\, (a_{rj})^2 \Sigma_s\, (c_s^{\,i})^2}$$

$$= \max_j \frac{(\mathbf{c}^i \,.\, \mathbf{a}^j)^2}{|\, \mathbf{a}^j \,|^2 \,|\, \mathbf{c}^i \,|^2}$$

where $\mathbf{a}^j$ denotes the vector, constructed from the jth column of $\mathbf{A}$, and the dot denotes a scalar product. We may therefore write

$$q_i = \max_j \cos^2 \theta_{ij},$$

where θ_{ij} is the angle between the ith residual vector $\mathbf{c}^i$ and the n-dimensional vector $\mathbf{a}^j$. Since $\mathbf{c}^i$ is a consequence of an initial guess, we must assume that it is as unfavourable as possible, that is to say, that the minimum value of θ_{ij} is as large as possible. An orthogonal matrix $\mathbf{A}$ would be the best possible, for in this case it would be impossible to obtain a residual vector $\mathbf{c}^i$ for which the minimum value of θ_{ij} was greater than $\cos^{-1}(1/\sqrt{n})$ and q_i was less than $1/n$. On the other hand, if the several vectors $\mathbf{a}^j$ are each of them almost perpendicular to some direction in n-dimensional space, then if $\mathbf{c}^i$ happened to lie in this direction, the minimum value of θ_{ij} would be nearly $\pi/2$ and q_i would be small. The extreme case in which each of the n-vectors $\mathbf{a}^j$ lies in an n-dimensional prime is excluded since this would imply that $\mathbf{A}$ was singular, but the most unfavourable cases for convergence would be those in which these vectors most nearly lay in a prime. A consequence of such an unsatisfactory state of affairs would be that

https://doi.org/10.1017/S2040618500035541 Published online by Cambridge University Press

the determinant of **A** would be very small in comparison with the magnitudes of the elements of **A**. In the light of the foregoing, we would suggest

$$L(\mathbf{A}) \equiv 1 - n \min_{\mathbf{c}} \left\{ \max_{j} \frac{(\mathbf{c} \cdot \mathbf{a}^j)^2}{|\mathbf{c}|^2 |\mathbf{a}^j|^2} \right\}$$

as a measure of the ill-conditioning of the matrix **A**. Turing (**12**) has discussed ill-conditioning at some length, and suggests alternative M- and N-condition numbers for **A**. Our L-condition number shows why his equations (8·1) are better conditioned than his (8·2) and confirms his statement that orthogonal matrices are the best conditioned of all. The M- and N-condition numbers are usually increased when a row or column is multiplied by a very large or a very small number. The same is true for the L-condition number when a row of **A** is multiplied, but $L(\mathbf{A})$ is unaltered when a column of **A** is multiplied by any constant. While the M- and N-condition numbers and the P-condition number, so named by Todd (**13**) but suggested by von Neumann and Goldstine (**14**) are the same for a matrix **A** and for the transposed matrix **A′**, this is not so for the L-condition number. The truth is that a matrix can suffer from several allied diseases each of which has its own condition number. The L-condition number measures the slowness of convergence of the relaxation process and it is defined in such a way that $1 > L(\mathbf{A}) \geqslant 0$.

As Fox (**10**) has pointed out, another kind of ill-conditioning can arise in which it is possible for small values of $|\mathbf{c}^i|$ to be associated with large errors in the approximation $\mathbf{v}^i$ to the exact solution $\mathbf{v}$. This trouble will be pronounced when

$$\max_{\mathbf{c}^i} \left\{ \frac{|\mathbf{v} - \mathbf{v}^i|}{|\mathbf{c}^i|} \right\}$$

is large. Since

$$\mathbf{v} - \mathbf{v}^i = \mathbf{A}^{-1}\mathbf{h} - \mathbf{A}^{-1}(\mathbf{h} - \mathbf{c}^i) = \mathbf{A}^{-1}\mathbf{c}^i,$$

this maximum may be written

$$\max_{\mathbf{c}} \frac{|\mathbf{A}^{-1}\mathbf{c}|}{|\mathbf{c}|} = B(\mathbf{A}^{-1}),$$

where $B(\mathbf{A})$ is the maximum expansion of **A**. (Cf. (**12**), p. 297.) Now,

$$\max_{\mathbf{c}} \frac{|\mathbf{A}^{-1}\mathbf{c}|}{|\mathbf{c}|} = \max_{\mathbf{c}} \frac{|\mathbf{c}|}{|\mathbf{A}\mathbf{c}|},$$

so we may choose

$$\min_{\mathbf{c}} \frac{|\mathbf{A}\mathbf{c}|^2}{|\mathbf{c}|^2} = (D(\mathbf{A}))^2 \qquad \text{.....(15)}$$

as a measure of this type of conditioning. The condition will be good if the minimum expansion $D(\mathbf{A})$ of **A** is large.

On the other hand,

$$\begin{aligned} n \min_{\mathbf{c}} \left\{ \max_{j} \frac{(\mathbf{c} \cdot \mathbf{a}^j)^2}{|\mathbf{c}|^2 |\mathbf{a}^j|^2} \right\} &\geqslant \min_{\mathbf{c}} \left\{ n \max_{j} \frac{(\mathbf{c} \cdot \mathbf{a}^j)^2}{|\mathbf{c}|^2} \cdot \frac{1}{\max_j |\mathbf{a}^j|^2} \right\} \\ &\geqslant \frac{1}{\max_j |\mathbf{a}^j|^2} \cdot \min_{\mathbf{c}} \frac{|\mathbf{A}'\mathbf{c}|^2}{|\mathbf{c}|^2} \\ &= \frac{1}{\max_j |\mathbf{a}^j|^2} (D(\mathbf{A}'))^2 \\ &= \frac{1}{\max_j |\mathbf{a}^j|^2} (D(\mathbf{A}))^2. \qquad \text{.....(16)} \end{aligned}$$

The convergence of the residual vectors will therefore be good when this expression is large.

https://doi.org/10.1017/S2040618500035541 Published online by Cambridge University Press

The fact that $D(\mathbf{A})$ appears in both (15) and (16) suggests that the two types of ill-conditioning encountered in the relaxation process are interrelated.

§ 4. In illustration of the foregoing discussion we should like to comment on a recent paper by Allen and Severn (**15**). These authors, by increasing the order of the differential equation to be solved, devise a method whereby certain problems, previously thought to be insoluble by the direct relaxation technique, can be solved by relaxation methods. We wish to point out that this device is unnecessary since the difficulties previously encountered yield to the methods advocated in the present paper. Following Allen and Severn, we shall, for the sake of simplicity, consider the solution of the first order equation

$$\frac{dv}{dx} = x + v \qquad \text{.....(17)}$$

over the range $0 \leqslant x \leqslant 1$ with the end condition $v_{x=0} = 1$. The lack of any given end condition at $x = 1$ apparently seemed to Allen and Severn an insuperable difficulty to the direct use of the relaxation technique. By doubling the order of the differential equation by the substitution

$$v = \frac{dw}{dx} + w \qquad \text{.....(18)}$$

and the imposition of an arbitrary condition on w at $x = 1$, however, they were able to solve the second order equation

$$\frac{d^2w}{dx^2} = w + x$$

by relaxation. These authors seem to have overlooked the possibility of employing equation (5) at $x = 1$ and were unable to use (4) in its place since $u(a)$ in that formula was not given by the boundary conditions. In fact, using (5) at $x = 1$ and (4) at all other nodes, the matrix equation becomes (we choose $a = 0{\cdot}1$)

$$\begin{bmatrix}
-\cdot2 & 1 & & & & & & & & \\
-1 & -\cdot2 & 1 & & & & & & & \\
 & -1 & -\cdot2 & 1 & & & & & & \\
 & & -1 & -\cdot2 & 1 & & & & & \\
 & & & -1 & -\cdot2 & 1 & & & & \\
 & & & & -1 & -\cdot2 & 1 & & & \\
 & & & & & -1 & -\cdot2 & 1 & & \\
 & & & & & & -1 & -\cdot2 & 1 & \\
 & & & & & & & -1 & -\cdot2 & 1 \\
 & & & & & & & & 1 & -4 \quad 2{\cdot}8
\end{bmatrix}
\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \\ v_8 \\ v_9 \\ v_{10} \end{bmatrix}
=
\begin{bmatrix} 1{\cdot}02 \\ {\cdot}04 \\ {\cdot}06 \\ {\cdot}08 \\ {\cdot}10 \\ {\cdot}12 \\ {\cdot}14 \\ {\cdot}16 \\ {\cdot}18 \\ {\cdot}20 \end{bmatrix}$$

Since the diagonal elements do not dominate the matrix **A**, formula (12) must now be employed for relaxing. The choice c_j/a_{jj} for k_j sometimes advocated will certainly be unsuitable in this case. Suffice it to say that one of us (Mitchell) has completed a relaxation solution of this problem by the above method and has obtained results in agreement with the solution of Allen and Severn and with the theoretical solution.

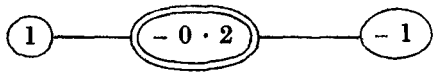
The method of Allen and Severn is, of course, a feasible one and in certain cases it may happen that it requires less laborious calculation than the direct method. It should, however, be remembered that it involves an additional approximation, the finite difference approximation for (18), by means of which the values for v_j are obtained from those of w_j. The fact that

https://doi.org/10.1017/S2040618500035541 Published online by Cambridge University Press

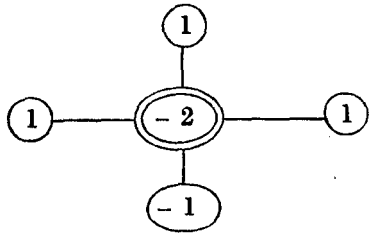
an arbitrary condition can be imposed on w at $x=1$ cannot, of course, prevent their final solution for v from having cumulative errors.

It is not suggested that in any problem the matrix equation be written out in full. In a two-dimensional problem with $n \times m$ nodes this would mean constructing a matrix of order nm. Sufficient elements should, however, be sketched in to indicate the structure of the matrix. The relaxation pattern at any node can then easily be read off from the columns of the matrix.

Thus, in solving the equation $\frac{dv}{dx}=x+v$ with $a=0{\cdot}1$, the pattern for the nodes at $x=0{\cdot}2$, $0{\cdot}3$, $0{\cdot}4$, $0{\cdot}5$, $0{\cdot}6$, $0{\cdot}7$ would be



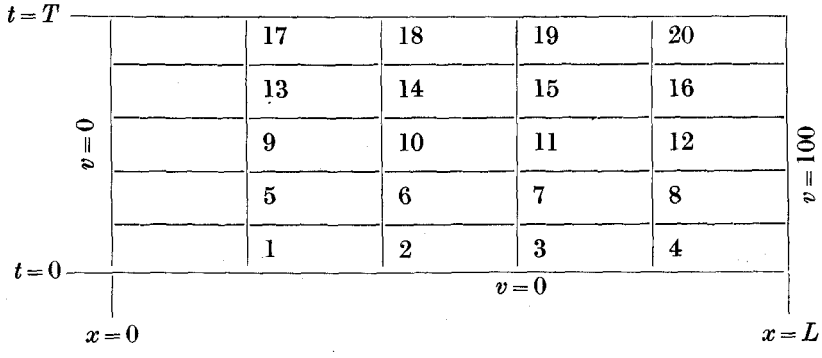
in each case, and the patterns for the nodes at $x=0{\cdot}1$, $0{\cdot}8$, $0{\cdot}9$, $1{\cdot}0$ would be respectively



In the above patterns the double ring indicates the node which is relaxed.

What we have said concerning one-dimensional problems is also applicable, with suitable adjustments to two-dimensional problems. Mitchell has also computed a solution of the flow of heat problem treated by Allen and Severn, and has done so without increasing the order of the differential equation. This problem was the solution of $\frac{\partial v}{\partial t}=\kappa\frac{\partial^2 v}{\partial x^2}$ in the region $0<x<L$, $0<t<T$, subject to the boundary conditions $v_{t=0}=v_{x=0}=0$, $v_{x=L}=100$.

The coarse mesh illustrated below was chosen for the convenience of the printer and T was taken to be $L^2/10\kappa$.



Two computations were made, one using the stable approximation ((2), p. 231).

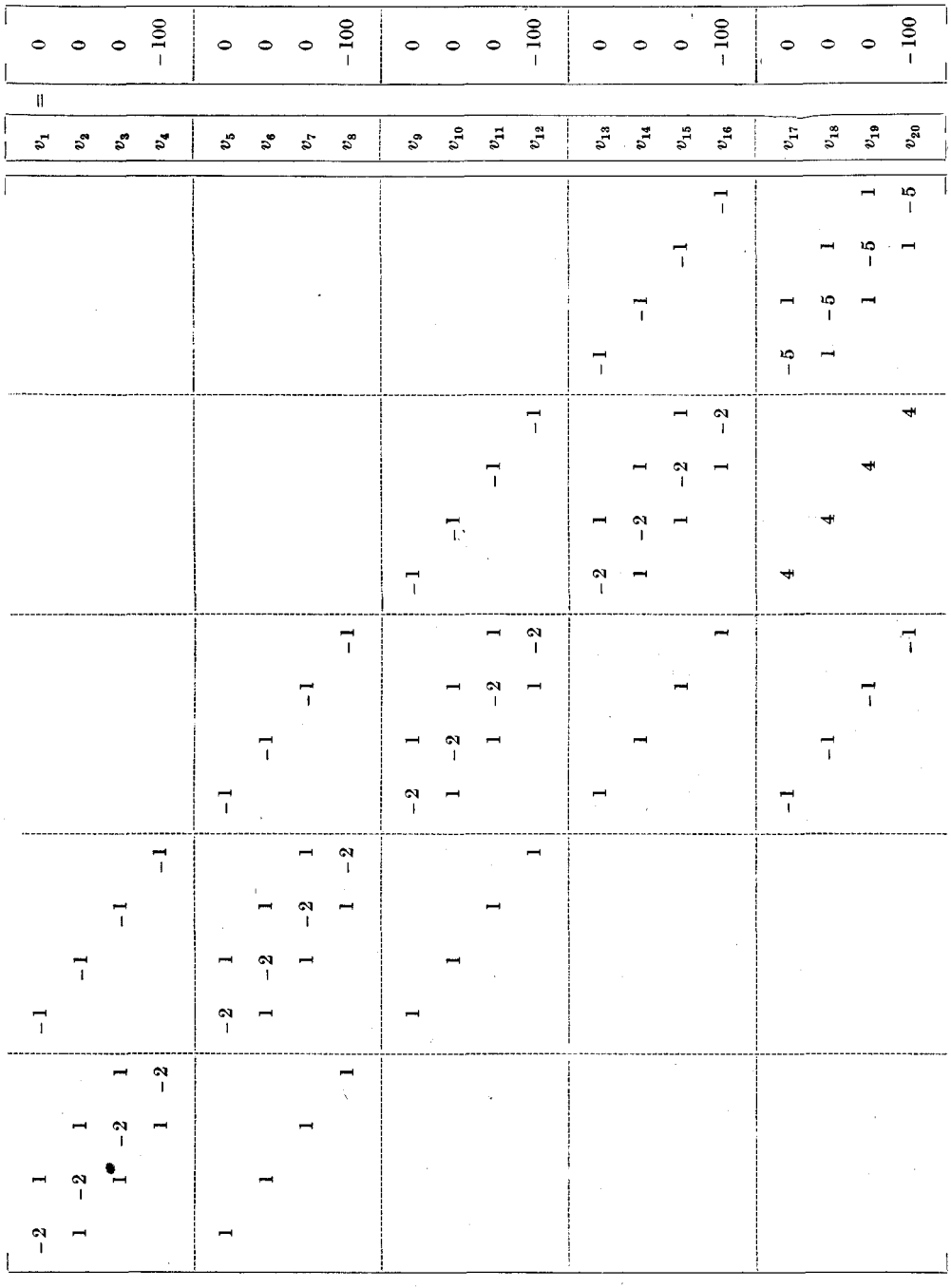
$$v_{r,s+1}-v_{r,s}=\frac{1}{2}\left(\frac{v_{r+1,s}-2v_{r,s}+v_{r-1,s}}{2}+\frac{v_{r+1,s+1}-2v_{r,s+1}+v_{r-1,s+1}}{2}\right) \quad \text{.....(19)}$$

and the other using the unstable approximation

$$v_{r,s+1}-v_{r,s-1}=v_{r+1,s}-2v_{r,s}+v_{r-1,s}. \quad \text{.....(20)}$$

https://doi.org/10.1017/S2040618500035541 Published online by Cambridge University Press

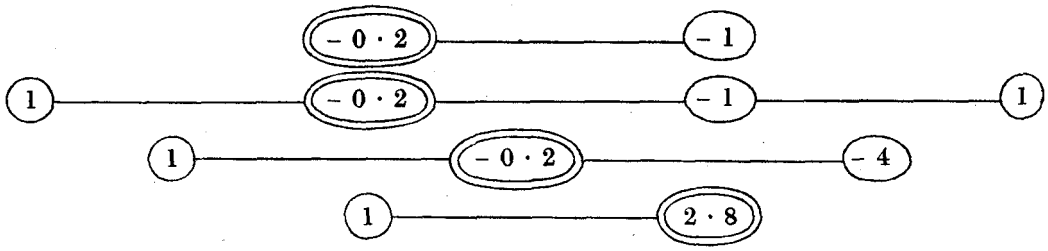
Allowing for the fact that our mesh is somewhat coarser than that used by Allen and Severn, the solutions obtained in both computations were in close agreement with that obtained by them. For the approximation (20) the following matrix equation was obtained, the rows and columns of which are arranged in the order in which nodes are numbered in the diagram.



Of the twenty nodes in this net only those numbered 6 and 7 are typical in that their

https://doi.org/10.1017/S2040618500035541 Published online by Cambridge University Press

pattern, illustrated below, is unaffected by the presence of the boundary. When, however, the size of the mesh is decreased, the great majority of the nodes will have a pattern of this type. In all there will be fifteen types of pattern used, but eight of these will be used at one node only.



In the light of the foregoing discussion it will now be apparent that an approximate solution by the methods of relaxation can be obtained for any linear differential equation with given boundary conditions, provided the following requirements are satisfied. (i) The boundary conditions must be sufficient to determine the uniqueness of the solution D of the differential equation. (ii) A convergent finite difference approximation of the differential equation and the appropriate boundary conditions must be available. (iii) Either this approximation must also be stable or else the problem must be such that no round-off errors need be introduced, e.g., by boundary conditions, into the finite difference equations which are to be relaxed. If these conditions are satisfied, relaxation methods can be applied, in principle to hyperbolic and parabolic, as well as to elliptic, differential equations of the second order.

REFERENCES

(1) Courant, R., Friedrichs, K., and Lewy, H., *Math. Annalen*, **100**, (1928), pp. 32–74.

(2) O'Brien, G. G., Hyman, M. A., and Kaplan, S., *Jour. Math. and Phys.*, **29** (1951), pp. 223–251.

(3) Thomas, L. H., " Symposium on Theoretical Compressible Flow, 1949," White Oak, Maryland.

(4) Rutishauser, H., *Zeit. f. angewandte Math. u. Phys.*, **3** (1952), pp. 65–74.

(5) Bickley, W. G., *Math. Gaz.*, **25** (1941), pp. 19–27.

(6) Bickley, W. G., *Q.J. Mech. App. Math.*, **1** (1948), pp. 35–42.

(7) Rutherford, D. E., *Proc. Roy. Soc. Edin.*, (A) **63** (1952), pp. 232–241.

(8) Temple, G., *Proc. Roy. Soc.*, (A) **169** (1939), pp. 476–500.

(9) Stiefel, E., *Zeit. f. angewandte Math. u. Phys.*, **3** (1952), pp. 1–33.

(10) Fox, L., *Q.J. Mech. App. Math.*, **1** (1948), pp. 253–280.

(11) Southwell, R. V., *Relaxation Methods in Theoretical Physics*, Oxford (1946).

(12) Turing, A. M., *Q.J. Mech. App. Math.*, **1** (1948), pp. 287–308.

(13) Todd, J., *Proc. Camb. Phil. Soc.*, **46** (1949), pp. 116–118.

(14) von Neumann, J. and Goldstine, H. H., *Bull. Amer. Math. Soc.*, **53** (1947), pp. 1021–1099.

(15) Allen, D. N. de G., and Severn, R. T., *Q.J. Mech. App. Math.*, **4** (1951), pp. 209–222.

United College
St. Andrews