Direct methods for the solution of finite-difference approximations to separable partial differential equations

By M. R. Osborne*

Methods suggested by Bickley and McNamee are reviewed. The use of their semi-rational technique for solving problems in non-rectangular regions is illustrated, and an application to the solution of eigenvalue problems is explained and exemplified.

1. Introduction

This paper is concerned with the methods suggested by Bickley and McNamee (1960) for the solution of finite-difference approximations to separable elliptic partial differential equations. These methods have the advantage of being direct, and they bypass the problems of organization and storage usually associated with direct methods. In the author's opinion they provide the best numerical method for problems for which they are suitable. In particular they do not require the matrix of the difference equations (this will be referred to later as the *big matrix*) to be definite, so that they can be applied to solve forced-vibration and eigenvalue problems.

The methods of Bickley and McNamee are considered in Section 3, but first their formalism is introduced (Section 2) by considering a particular problem. In Section 4 the semi-rational method is used to derive expressions for influence matrices which show the effect of the boundary values on the solution of the difference equation. In Section 5 use is made of the results of Section 4 to solve more general boundary-value problems. In Section 6 an application to an eigenvalue problem is considered, and some numerical results are presented.

There are two possible ways in which boundary-value problems in more general regions can be tackled. If the region can be subdivided into rectangles then very efficient and well-conditioned computational schemes are possible. This is exemplified here for a mixed boundary-value problem. This provides perhaps the simplest application of the influence matrices. An application to the solution of the Dirichlet problem for Laplace's equation in a T-shaped region is indicated in Bickley and McNamee (1960). In Wilson (1962) similar calculations are given (for example for the Dirichlet problem for Laplace's equation in a rectangular annulus). Wilson's method is essentially the irrational method of Bickley and McNamee. Both methods are suitable only for equations with constant coefficients. Similar results appear also to have been obtained by G. N. Polozhii in the Soviet Union. The author is indebted to Mr. G. J. Tee of Lancaster University for this reference. Mr. Tee is engaged in preparing a translation of some of Polozhii's work.

The second possibility for handling a more general region is to embed it in a rectangle. There is considerable freedom in the way in which the problem can be defined in the extended region, and it is by no means obvious how to proceed to obtain a computational procedure which is both efficient and well conditioned. This difficulty is illustrated in Section 5 by considering a triangular region embedded in a rectangle, but no satisfactory solution is presented.

To provide a basis for assessing the efficiency of the procedures discussed here they are compared with the corresponding optimized line-overrelaxation procedures. The asymptotic rate of convergence for Laplace's equation in a square (Varga, 1962, p. 204) is used to estimate the number of iterations required for the overrelaxation computation.

It is stressed that the aim of this paper is to show how certain sets of algebraic equations, which occur in the solution by finite-difference methods of a class of partial differential equations, can be solved efficiently by the direct methods of Bickley and McNamee. No attempt has been made to estimate the accuracy of the results obtained as solutions of the partial differential equations.

The following notation is used:

$\rho_i(\boldsymbol{Q})$ for the ith row of the matrix $\boldsymbol{Q}$,
$\kappa_i(\boldsymbol{Q})$ for the ith column of the matrix $\boldsymbol{Q}$, and
$\boldsymbol{e}_j$ for the vector with 1 in the jth place and zeros elsewhere.

2. The formalism of Bickley and McNamee

The effectiveness of the techniques developed by Bickley and McNamee is largely due to the convenient and elegant formalism they use for representing the difference equation. In this section this formalism is developed for the standard five-point finite-difference approximation to the partial differential equation

$$\frac{\partial^2\phi}{\partial r^2}+\frac{1}{r}\frac{\partial\phi}{\partial r}+\frac{\partial^2\phi}{\partial z^2}=f(r, z) \tag{2.1}$$

subject to the boundary conditions

$$\left.\begin{array}{ll}\text{(i)} & \phi \text{ regular, } r=0,\\ \text{(ii)} & \phi(r, 0)=a(r)\\ \text{(iii)} & \dfrac{\partial\phi}{\partial r}(1, z)=0,\\ \text{(iv)} & \dfrac{\partial\phi}{\partial z}(r, l)=0.\end{array}\right\} \tag{2.2}$$

* *Computer Unit, Edinburgh University,* 7 *Buccleuch Place, Edinburgh,* 8.

This problem (with $l = 2$, and $f(r, z) = 0$) is the subject of the numerical experiments reported in Section 6.

To set up the finite-difference equations the region $0 \leqslant r \leqslant 1$, $0 \leqslant z \leqslant l$ is covered by families of lines parallel to the r and z axes respectively. The convention is adopted that lines parallel to the z axis correspond to constant values of r with

$$r_i = (i - 1)h, \quad i = 1, \ldots, m; \quad h = 1/(m - 1),$$

and that lines parallel to the r axis correspond to constant values of z with

$$z_i = ik, \quad i = 0, 1, \ldots, n; \quad k = l/n.$$

For any function W defined on the mesh points, and for any values of i and j, the value $W(r_i, z_j)$ is written W_{ij}.

The difference approximation is considered first on the line $z = z_j$, $2 \leqslant j \leqslant n - 1$. There are three cases to consider. The appropriate difference equations are given by

(i) $r = r_1(=0)$,

$$4h^{-2}(\phi_{2j} - \phi_{1j}) + k^{-2}(\phi_{1(j-1)} - 2\phi_{1j} + \phi_{1(j+1)}) = f_{1j}. \tag{2.3}$$

(ii) $r = r_i, 2 \leqslant i \leqslant m - 1$,

$$\begin{aligned} h^{-2}\{&(1 - 1/2(i-1))\phi_{(i-1)j} - 2\phi_{ij} \\ &+ (1 + 1/2(i - 1))\phi_{(i+1)j}\} \\ &+ k^{-2}\{\phi_{i(j-1)} - 2\phi_{ij} + \phi_{i(j+1)}\} = f_{ij}. \end{aligned} \tag{2.4}$$

(iii) $r = r_m$,

$$\begin{aligned} h^{-2}\{&2\phi_{(m-1)j} - 2\phi_{mj}\} \\ &+ k^{-2}\{\phi_{m(j-1)} - 2\phi_{mj} + \phi_{m(j+1)}\} = f_{mj}. \end{aligned} \tag{2.5}$$

By introducing the vectors $\boldsymbol{\phi}_j$ whose components are ϕ_{ij}, $i = 1, \ldots, m$, the equations (2.3) — (2.5) are combined to give

$$\boldsymbol{A\phi}_j + k^{-2}\{\boldsymbol{\phi}_{j-1} - 2\boldsymbol{\phi}_j + \boldsymbol{\phi}_{j+1}\} = \boldsymbol{f}_j \tag{2.6}$$

where $\boldsymbol{A}$ is the $(m \times m)$ matrix

$$h^{-2}\begin{bmatrix} -4 & 4 & & & & \\ \frac{1}{2} & -2 & \frac{3}{2} & & & \\ & \frac{3}{4} & -2 & \frac{5}{4} & & \\ & \cdots & \cdots & \cdots & \cdots & \\ & & & & 2 & -2 \end{bmatrix}. \tag{2.7}$$

Equation (2.6) holds for all lines $z = z_j$, $j = 1, 2, \ldots, n$ provided that for $\boldsymbol{\phi}_0$ is substituted a vector $\boldsymbol{a}$ evaluated from the boundary condition $\phi(r, 0) = a(r)$, and provided that $\boldsymbol{\phi}_{n-1}$ is substituted for $\boldsymbol{\phi}_{n+1}$. That the latter substitution is appropriate follows from the boundary condition $\frac{\partial\phi}{\partial z}(r, l) = 0$.

It will be seen that equation (2.6) consists of an operation $\boldsymbol{A}$ independent of j on each of the vectors $\boldsymbol{\phi}_j$, plus an operation which combines together, $\boldsymbol{\phi}_{j-1}$, $\boldsymbol{\phi}_j$, $\boldsymbol{\phi}_{j+1}$. If $\boldsymbol{\Phi}$* is written for the $(m \times n)$ matrix whose ith column is $\boldsymbol{\phi}_i$, then the equations (2.6) for $j = 1, 2. \ldots, n$ can be combined to give

$$\boldsymbol{A\Phi} + \boldsymbol{\Phi B} = \boldsymbol{F} - h^{-2}[\boldsymbol{a}|0] \tag{2.8}$$

where $\boldsymbol{B}$ is the $(n \times n)$ matrix

$$k^{-2}\begin{bmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ & & & 1 & -2 & 2 \\ & & & & 1 & -2 \end{bmatrix}. \tag{2.9}$$

Equation (2.8) is in the form considered by Bickley and McNamee. The standard finite-difference approximation (whereby derivatives are replaced by the first terms of their central difference or averaged central difference representations) to a separable partial differential equation can always be put into this form provided appropriate boundary conditions are given on a rectangle whose sides are parallel to axes of the separable variables. For second-order equations, for example, boundary conditions of the form $J\phi + K\partial\phi/\partial n$ can be prescribed provided J and K are constant on any one side. The separation of operations on the rows and columns of the solution matrix in equation (2.8) provides an elegant analogue to the separability of the original partial differential equation.

3. On the methods suggested by Bickley and McNamee

Bickley and McNamee give three methods for solving the equation

$$\boldsymbol{AX} + \boldsymbol{XB} = \boldsymbol{F}. \tag{3.1}$$

These they designate as the *irrational*, *semi-rational*, and *rational* methods respectively (pages 99–109).

The semi-rational method appears to be the most generally useful of the three. It requires a knowledge of the similarity normal form of either $\boldsymbol{A}$ or $\boldsymbol{B}$. Let us assume that that of $\boldsymbol{B}$ is known. Then

$$\boldsymbol{B} = \boldsymbol{T\Lambda T}^{-1}. \tag{3.2}$$

Assume first that $\boldsymbol{\Lambda}$ is diagonal. Substituting for $\boldsymbol{B}$ in Equation (3.1) gives

$$\boldsymbol{AXT} + \boldsymbol{XT\Lambda} = \boldsymbol{FT}.$$

Let $\quad \overline{\boldsymbol{X}} = \boldsymbol{XT}, \overline{\boldsymbol{F}} = \boldsymbol{FT} \quad$ then

$$\boldsymbol{A}\overline{\boldsymbol{X}} + \overline{\boldsymbol{X}}\boldsymbol{\Lambda} = \overline{\boldsymbol{F}} \tag{3.3}$$

so that

$$(\boldsymbol{A} + \lambda_i\boldsymbol{I})\kappa_i(\overline{\boldsymbol{X}}) = \kappa_i(\overline{\boldsymbol{F}}) \qquad i = 1, 2, \ldots, n. \tag{3.4}$$

Each column of $\overline{\boldsymbol{X}}$ can be found from equation (3.4) by solving a set of linear equations. The solution of equation (3.1) is then completed by a matrix multiplication.

Note that when $\boldsymbol{A}$ is a band matrix, as in equation (2.8), the matrix multiplications necessary to calculate $\overline{\boldsymbol{F}}$ and $\boldsymbol{\Phi}$ can be a large part of the computation. However, in

* Note that $(\boldsymbol{\Phi})_{ij} = \phi_{ij}$ so that the notation introduced at the beginning of this section is compatible with the usual matrix notation.

many cases these can be avoided (see, for example, equations (4.3) and (4.6) in the next section). For this reason the setting up of these transformation matrices is included in the estimates of the work involved in the computation; however, the matrix multiplications are not.

If $\boldsymbol{\Lambda}$ is not diagonal, then at least one of the eigenvalues of $\boldsymbol{B}$ is associated with a principal vector of grade 2. Assume that λ_i is a twice repeated eigenvalue, and that $(\boldsymbol{B} - \lambda_i \boldsymbol{I})$ has rank $(n - 1)$. Then $\boldsymbol{\Lambda}$ will have the form

$$\begin{array}{cccc} \cdots & \cdots & \cdots & \cdots \\ \lambda_{i-1} & & & \\ & \boxed{\begin{matrix}\lambda_i & 1 \\ 0 & \lambda_i\end{matrix}} & & \\ & & \lambda_{i+1} & \\ \cdots & \cdots & \cdots & \cdots \end{array}$$

In this case the ith column of $\overline{\boldsymbol{X}}$ satisfies equation (3.4), while the $(i + 1)$th column satisfies

$$(\boldsymbol{A} + \lambda_i \boldsymbol{I})\kappa_{i+1}(\overline{\boldsymbol{X}}) + \kappa_i(\overline{\boldsymbol{X}}) = \kappa_{i+1}(\overline{\boldsymbol{F}}). \qquad (3.5)$$

The solution of this equation is readily found once $\kappa_i(\overline{\boldsymbol{X}})$ has been computed.

The semi-rational method has very considerable advantages over direct methods which involve factorization of what Bickley and McNamee call "the big matrix,"* for it bypasses all the very considerable problems of organization and storage which these methods encounter. Again it is in general better than the over-relaxation iterative techniques currently in use even when the optimum values of the over-relaxation parameters are known (for the problems considered here these can be readily calculated, see Osborne (1963)). This is so even if it is necessary to solve the eigenvalue problem for either $\boldsymbol{A}$ or $\boldsymbol{B}$ to compute the similarity normal form.

To see why this should be, note that the calculation of $\overline{\boldsymbol{X}}$ involves about the same amount of work as one sweep through the mesh in a typical line iteration. Therefore, unless a similarity normal form has to be calculated, the semi-rational method must be vastly superior. If the orders of $\boldsymbol{A}$ and $\boldsymbol{B}$ are similar then, for problems of the type considered in the previous section, the calculation of all the eigenvalues and eigenvectors of $\boldsymbol{B}$ involves of the order of $10n$ solutions of sets of linear equations with matrix $\boldsymbol{A}$†, or of the order of 10 sweeps of a typical line iteration by columns.

This number of sweeps would be sufficient for the convergence of a line iteration only for very small problems.

Equation (2.8) is clearly one which should be solved using the semi-rational method for here the eigenvalue problem for the matrix $\boldsymbol{B}$ (given by equation (2.9)) is readily solved explicitly in terms of trigonometric functions.

The irrational method requires a knowledge of the similarity normal forms of both $\boldsymbol{A}$ and $\boldsymbol{B}$. Assume that

$$\boldsymbol{A} = \boldsymbol{SMS}^{-1} \qquad (3.6)$$

where $\boldsymbol{M}$ is diagonal. Then equation (3.4) can be written

$$(\boldsymbol{M} + \lambda_i \boldsymbol{I})\kappa_i(\boldsymbol{S}^{-1}\overline{\boldsymbol{X}}) = \kappa_i(\boldsymbol{S}^{-1}\overline{\boldsymbol{F}}). \qquad (3.7)$$

As the matrix on the left-hand side is diagonal the solution of equation (3.7) is trivial. The solution of equation (3.1) can now be obtained by two matrix multiplications.

In the applications envisaged in this paper both similarity normal forms are available only when the partial differential equation has constant coefficients. In all other cases the semi-rational method is to be preferred. It is interesting that the semi-rational method is very nearly competitive with the irrational method, even in the example most favourable to the latter—the case of Laplace's equation in a rectangle with Dirichlet boundary conditions. Here a count of multiplications (counting division as two multiplications) for computational schemes that were probably pretty efficient for both methods found a factor of 5/7 in favour of the irrational method.

The rational method requires a knowledge of the characteristic equation of either $\boldsymbol{A}$ or $\boldsymbol{B}$. For tridiagonal $\boldsymbol{A}$ and $\boldsymbol{B}$ this can be found using the obvious three-term recurrence. The rational method is most easily derived by making use of the identity

$$\boldsymbol{A}^r\boldsymbol{X} = \boldsymbol{X}(-\boldsymbol{B})^r + \sum_{i=0}^{r-1} (-1)^i \boldsymbol{A}^{r-i-1}\boldsymbol{F}\boldsymbol{B}^i. \qquad (3.8)$$

If the characteristic equation for $\boldsymbol{A}$ is $\psi(\boldsymbol{A}) = 0$ then, by equation (3.8),

$$0 = \psi(\boldsymbol{A})\boldsymbol{X} = \boldsymbol{X}\psi(-\boldsymbol{B}) + R(\boldsymbol{A}, \boldsymbol{F}, \boldsymbol{B}). \qquad (3.9)$$

Equation (3.9) is a set of linear equations for the rows of $\boldsymbol{X}$. R has been written as a shorthand for the terms independent of $\boldsymbol{X}$.

The author does not know of any application of the rational method other than the one given by Bickley and McNamee. The rational method does not appear to offer any particular advantages in economy over the other two methods. It may be numerically unstable when the big matrix is not definite.

4. Dependence of the solution on the boundary conditions

In this section expressions are derived which show explicitly the dependence of the solution of equation (2.8) on the vector $\boldsymbol{a}$ of boundary values given on $z = 0$. The semi-rational method is used to solve equation (2.8).

For convenience the solution matrix is broken into two parts $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Phi}_2$ where

$$\boldsymbol{A\Phi}_1 + \boldsymbol{\Phi}_1\boldsymbol{B} = \boldsymbol{F} \qquad (4.1)$$

and

$$\boldsymbol{A\Phi}_2 + \boldsymbol{\Phi}_2\boldsymbol{B} = -h^{-2}[\boldsymbol{a}|0]. \qquad (4.2)$$

* This is of order mn. It is the matrix of the difference equations when the ϕ_{ij} are ordered as the components of a vector.

† The eigenvalues of tridiagonal matrices can be found very efficiently. Here the principal minors form a Sturm sequence.

As $\boldsymbol{\Phi}_1$ is independent of $\boldsymbol{a}$ attention is now concentrated on $\boldsymbol{\Phi}_2$. In equation (4.2) only the first column of the right-hand side is non zero. However, the analysis is no more complicated if it is the rth column which is non zero. In this slightly more general case it is the dependence of the solution on the rth column of the right-hand side that is computed, and the equation to be solved now takes the form (omitting the $-h^{-2}$ which is relevant only to the boundary condition)

$$\boldsymbol{A}\boldsymbol{\Phi}_2 + \boldsymbol{\Phi}_2\boldsymbol{B} = [0|\boldsymbol{a}_r|0]. \tag{4.2a}$$

There are two possibilities for the calculation of $\boldsymbol{\Phi}_2$ depending on which of the matrices $\boldsymbol{A}$ or $\boldsymbol{B}$ is resolved into its similarity normal form. Both cases will be treated in detail.

(i) $\boldsymbol{B} = \boldsymbol{T\Lambda T}^{-1}$.

In this case

$$\kappa_i(\overline{\boldsymbol{F}}) = \kappa_i([0|\boldsymbol{a}_r|0]\boldsymbol{T}) = T_{ri}\boldsymbol{a}_r. \tag{4.3}$$

The semi-rational method gives

$$\begin{aligned}\kappa_j(\boldsymbol{\Phi}_2) &= [\ldots, T_{ri}(\boldsymbol{A} + \lambda_i\boldsymbol{I})^{-1}\boldsymbol{a}_r, \ldots]\kappa_j(\boldsymbol{T}^{-1}) \qquad (4.4)\\ &= [\sum_{i=1}^{n} (\boldsymbol{T}^{-1})_{ij}\boldsymbol{T}_{ri}(\boldsymbol{A} + \lambda_i\boldsymbol{I})^{-1}]\boldsymbol{a}_r\\ &= \boldsymbol{Q}_j^r\boldsymbol{a}_r. \qquad (4.5)\end{aligned}$$

Equation (4.5) shows explicitly the dependence of the solution on the boundary values (or the right-hand side). The matrices $\boldsymbol{Q}_j^r$ will be called *influence matrices.*

(ii) $\boldsymbol{A} = \boldsymbol{SMS}^{-1}$.

Here

$$\begin{aligned}\rho_i(\overline{\boldsymbol{F}}) &= \rho_i(\boldsymbol{S}^{-1}[0|\boldsymbol{a}_r|0])\\ &= (\rho_i(\boldsymbol{S}^{-1}).\boldsymbol{a}_r)\boldsymbol{e}_r^T. \qquad (4.6)\end{aligned}$$

The semi-rational method gives

$$\boldsymbol{\Phi}_2 = \boldsymbol{S}\begin{bmatrix}\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\\ (\rho_i(\boldsymbol{S}^{-1}).\boldsymbol{a}_r)\boldsymbol{e}_r^T(\boldsymbol{B} + \mu_i\boldsymbol{I})^{-1}\\ \cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot\end{bmatrix}. \tag{4.7}$$

Whence

$$\begin{aligned}\kappa_j(\boldsymbol{\Phi}_2) &= \left[\sum_{i=1}^{m} (\boldsymbol{B} + \mu_i\boldsymbol{I})_{rj}^{-1}\kappa_i(\boldsymbol{S})\rho_i(\boldsymbol{S}^{-1})\right]\boldsymbol{a}_r\\ &= \boldsymbol{Q}_j^r\boldsymbol{a}_r. \qquad (4.8)\end{aligned}$$

Comparing the different forms of $\boldsymbol{Q}_j^r$ given by equations (4.5) and (4.8) it is seen that the former expression requires the calculation of inverse matrices while the latter requires only single elements of the inverses. Each of these elements can be found from the solution of a single set of linear equations. The calculation from equation (4.5) requires the solution of $O(n^2)$ sets of linear equations, and this is of the same order of magnitude as the number required for the convergence of an optimized line iteration.

Methods for the solution of finite-difference approximations to elliptic partial differential equations which make use of influence matrices (or matrices derived from them) appear to be successful only when the influence matrices are well conditioned. This will be illustrated in the next section. This ill-conditioning of the influence matrices derives from the fact that the influence on the solution of an elliptic difference equation of a disturbance at an isolated point of the mesh falls away rapidly with distance from the point.

This section concludes with a discussion of the conditioning of the influence matrices in a special case. From equation (4.8) it is seen that the influence matrices have as eigenvectors the $\kappa_i(\boldsymbol{S})$, and as eigenvalues $(\boldsymbol{B} + \mu_i\boldsymbol{I})_{rj}^{-1}$ $i = 1, 2, \ldots, m$. In many applications a measure of the condition of $\boldsymbol{Q}_j^r$ is provided by the ratio of its eigenvalue of greatest to that of least modulus. This ratio will be called the *condition number.* It is readily calculated for the special case of finite-difference approximation to Laplace's equation in a rectangle with Dirichlet type boundary conditions. If, for simplicity, it is assumed that $h = k$ then

$$\begin{aligned}-[h^2(\boldsymbol{B} + \mu_i\boldsymbol{I})]_{rs}^{-1} &= \frac{\sinh(n-s)\sigma_i \sinh r\sigma_i}{\sinh\sigma_i \sinh n\sigma_i}\; r \leqslant s\\ &= \frac{\sinh s\sigma_i \sinh(n-r)\sigma_i}{\sinh\sigma_i \sinh n\sigma_i}\; r \geqslant s \qquad (4.9)\end{aligned}$$

where $2 - h^2\mu_i = 2 + 4\sin^2 i\pi/2n = 2\cosh\sigma_i$.

From equation (4.9) it follows (for example) that the condition number of $\boldsymbol{Q}_1^1$ is approximately 6 for large values of n while that for $\boldsymbol{Q}_n^1$ is $O(e^{\alpha n}/n)$ for a fixed value of α between 1 and 2. Thus the conditioning of $\boldsymbol{Q}_1^1$ is excellent while that of $\boldsymbol{Q}_n^1$ is catastrophically bad.

5. Application to more general boundary-value problems

The formulae developed in Section 4 can be used to extend the range of problems which can be solved by the methods suggested by Bickley and McNamee. First the problem discussed in Section 2 is considered with the mixed boundary condition

$$\left.\begin{aligned}\frac{\partial\phi}{\partial z}(r, 0) &= 1, \quad 0 \leqslant r < R\\ \phi(r, 0) &= 0, \quad R \leqslant r \leqslant 1\end{aligned}\right\}. \tag{5.1}$$

It is assumed that the other boundary conditions are as before, and that $R = (t + 1)h$ is a mesh point. The difference approximation to this problem is solved by finding a vector of boundary values $\boldsymbol{\phi}_0$ having the property that the solution for it (which is readily obtained by the methods of Section 3) also satisfies the standard finite-difference approximation to the mixed condition (5.1). The results of Section 4 are used in finding $\boldsymbol{\phi}_0$.

Let the vector $\boldsymbol{\phi}^{(t)}$ be made up of the first t components of $\boldsymbol{\phi}$, and let $\boldsymbol{\phi}^{(m-t)}$ stand for the vector consisting of the remaining $(m - t)$ components. Then the finite-difference approximation to equation (5.1) can be written

$$\begin{aligned}\boldsymbol{\phi}_{-1}^{(t)} &= \boldsymbol{\phi}_1^{(t)} + 2k\boldsymbol{e}^{(t)}\\ \boldsymbol{\phi}_0^{(m-t)} &= 0\end{aligned} \tag{5.2}$$

where $\boldsymbol{e}$ is the vector each component of which is 1. The column $\boldsymbol{\phi}_{-1}$ can be found by applying the difference equation on the line $z = 0$. This gives

$$\boldsymbol{\phi}_{-1} = -k^2\boldsymbol{A}\boldsymbol{\phi}_0 + 2\boldsymbol{\phi}_0 - \boldsymbol{\phi}_1 + k^2\boldsymbol{f}_0. \tag{5.3}$$

Let the leading $t \times t$ principal submatrix of $\boldsymbol{A}$ be denoted by $\boldsymbol{A}^{(t)}$, then, combining equations (5.2) and (5.3), the boundary conditions (5.2) become

$$\begin{aligned} 2\boldsymbol{\phi}_1^{(t)} - (2\boldsymbol{I}^{(t)} - k^2\boldsymbol{A}^{(t)})\boldsymbol{\phi}_0^{(t)} &= k^2\boldsymbol{f}_0^{(t)} - 2k\boldsymbol{e}^{(t)} \\ \boldsymbol{\phi}_0^{(m-t)} &= 0. \end{aligned} \tag{5.4}$$

The vector $\boldsymbol{\phi}^{(t)}$ can be eliminated from equation (5.4) by using equations (4.1), (4.2), and (4.8). These give

$$\boldsymbol{\phi}_1 = -h^{-2}\boldsymbol{Q}_1^1\boldsymbol{\phi}_0 + \kappa_1(\boldsymbol{\Phi}_1) \tag{5.5}$$

so that the appropriate vector of boundary values can be found by solving the set of linear equations

$$\begin{aligned} &[-2h^{-2}\boldsymbol{Q}_1^{1(t)} - 2\boldsymbol{I}^{(t)} + k^2\boldsymbol{A}^{(t)}]\boldsymbol{\phi}_0^{(t)} \\ &\qquad = k^2\boldsymbol{f}_0^{(t)} - 2k\boldsymbol{e}^{(t)} - 2\kappa_1(\boldsymbol{\Phi}_1). \end{aligned} \tag{5.6}$$

Once equation (5.6) has been solved for $\boldsymbol{\phi}_0^{(t)}$ the solution of the difference equation can readily be completed. The solution of a related problem is described in Section 6.

The second problem considered is that of solving equation (2.1) subject to the boundary conditions

$$\left.\begin{aligned} &\text{(i)}\ \ \phi(0, z)\ \text{regular} \\ &\text{(ii)}\ \ \frac{\partial\phi}{\partial z} = 0, \quad z = l \\ &\text{(iii)}\ \ \phi(r, z) = a(r, z), \quad r = z/l. \end{aligned}\right\} \tag{5.7}$$

One possibility for solving this problem is to embed the triangular region specified in equation (5.7) in the rectangle $0 \leqslant r \leqslant 1$, $0 \leqslant z \leqslant l$. To apply the methods of Section 3 boundary conditions must be given on the sides $r = 1$ and $z = l$, and the right-hand side of the differential equation must be defined in $z/l \leqslant r < 1$. These conditions must be adjusted so that the condition (5.7) (iii) is satisfied.

If there are the same number of mesh points on each line $r =$ constant and $z =$ constant then each intersection of a grid line with the line $r = z/l$ occurs at a mesh point. In this case there are the same number of mesh points on $r = z/l$ as there are on each of the new boundary lines. As values of the right-hand side in the extended region are also disposable there are many more variables to be adjusted than there are conditions to be satisfied ((5.7) (iii)) becomes

$$\boldsymbol{\Phi}_{j+1,j} = a(r_{j+1}, z_j), \quad j = 1, \ldots, n-1).$$

One may proceed by arbitrarily fixing all but $n - 1$ of the disposable quantities and then adjusting these so that (5.7) (iii) is satisfied. However, not all such methods are satisfactory. For example if the $(n - 1)$ quantities to be adjusted are the values on one of the new boundaries then it is not difficult to derive the equations which these values must satisfy, and the cost in terms of the number of operations involved is the same as that for setting up one of the $\boldsymbol{Q}_s^r$. However, the matrix of this set of equations rapidly becomes extremely ill-conditioned as n is increased. For example a program written to implement this procedure on the Atlas computer worked successfully for $n = 4$, but for $n = 10$ the components of the vector of boundary values were $O(10^{12})$ and the solution was nearly meaningless.

Another possibility is based on Tee's description of Polozhii's work. In this case the quantities to be adjusted are the right-hand side values on $r = z/l$. Assume that all other disposable quantities can be set equal to zero, then, using the influence matrices, the conditions to be satisfied are

$$\begin{aligned} a(r_{j+1}, z_j) &= \sum_{k=1}^{n-1} f_k\boldsymbol{e}_{j+1}^T\boldsymbol{Q}_j^k\boldsymbol{e}_{k+1}, \quad j = 1, \ldots, n-1 \\ &= \sum_{k=1}^{n-1} f_k(\boldsymbol{Q}_j^k)_{(j+1)(k+1)} \end{aligned} \tag{5.8}$$

where f_k is the right-hand side value at the point r_{k+1}, z_k. The set of equations for the f_k has the possibility of being better conditioned. However, this procedure has the serious disadvantage that the setting up of the matrix of equation (5.8) requires $(n - 1)^2$ triangulations and forward and back substitutions. This is an amount of computation comparable with that required to solve the problem by an optimized overrelaxation technique.

6. An eigenvalue problem

The procedure described in the previous section for the solution of the mixed boundary-value problem fails:

(*a*) if any of the matrices which have to be inverted in the calculation of $\boldsymbol{Q}_1^1$ are singular, or

(*b*) if the matrix

$$\boldsymbol{H} = -2h^{-2}\boldsymbol{Q}_1^{1(t)} - 2\boldsymbol{I}^{(t)} + k^2\boldsymbol{A}^{(t)}$$

is singular.

In case (*a*) the matrix of the difference equation for the problem with a vector of boundary values $\boldsymbol{\phi}_0$ prescribed is singular. In case (*b*), however, it is the matrix of the difference equation for the problem with the mixed boundary condition which is singular, and a solution of the homogeneous mixed boundary-value problem can readily be constructed by using the Bickley-McNamee technique to solve the difference eqaation for a vector of boundary values $\boldsymbol{\phi}_0$ where $\boldsymbol{\phi}_0^{(t)}$ is a non-trivial solution of

$$\boldsymbol{H}\boldsymbol{\phi}^{(t)} = 0$$

and

$$\boldsymbol{\phi}_0^{(m-t)} = 0.$$

Use can be made of the above observation to solve the eigenvalue problem associated with equation (2.1)

$$\frac{\partial^2\phi}{\partial r^2} + \frac{1}{r}\frac{\partial\phi}{\partial r} + \frac{\partial^2\phi}{\partial z^2} = -\sigma\phi \tag{6.1}$$

subject to a homogeneous mixed boundary condition on $z = 0$. The term on the right-hand side of equation (6.1) can readily be incorporated into the difference equation. In particular the forms for the influence matrices become

$$\begin{aligned} \boldsymbol{Q}_s^r &= \sum_{i=1}^{n} (\boldsymbol{T}^{-1})_{ij}\boldsymbol{T}_{ri}(\boldsymbol{A} + (\sigma + \lambda_i)\boldsymbol{I})^{-1} \\ &= \sum_{i=1}^{m} (\boldsymbol{B} + (\sigma + \mu_i)\boldsymbol{I})_{rj}^{-1}\kappa_i(\boldsymbol{S})\rho_i(\boldsymbol{S}^{-1}) \end{aligned} \tag{6.2}$$

while the equation for determining the vector of boundary values $\boldsymbol{\phi}_0$ appropriate to the inhomogeneous mixed boundary conditions (5.1) becomes

$$[-2h^{-2}\boldsymbol{Q}_1^{1(t)} - (2 - k^2\sigma)\boldsymbol{I}^{(t)} + k^2\boldsymbol{A}^{(t)}]\boldsymbol{\phi}_0^{(t)} = -\ 2k\boldsymbol{e}^{(t)} \tag{6.3}$$

where $\boldsymbol{Q}_1^1$ is defined by setting $r = s = 1$ in equation (6.2).

The eigenvalues can now be characterized as the values of σ for which the matrix

$$\boldsymbol{H}(\sigma) = -\ 2h^{-2}\boldsymbol{Q}_1^{1(t)} - (2 - k^2\sigma)\boldsymbol{I}^{(t)} + k^2\boldsymbol{A}^{(t)} \tag{6.4}$$

is singular. In equation (6.4) $\boldsymbol{H}(\sigma)$ is only a $(t \times t)$ matrix so that this characterization of the eigenvalues has the advantage of compactness. However, equation (6.4) is nonlinear in the eigenvalue parameter.

The numerical solution of eigenvalue problems in which the eigenvalue parameter appears nonlinearly has been considered by Osborne and Michaelson (1964), and by the author (1964). In Osborne and Michaelson is derived the iteration

$$\begin{aligned} \boldsymbol{H}(\sigma_i)\boldsymbol{v}_{i+1} &= \boldsymbol{x}_i/(\boldsymbol{x}_i)_{p_i} \\ \boldsymbol{H}(\sigma_i)\boldsymbol{x}_{i+1} &= \frac{d\boldsymbol{H}}{d\sigma}(\sigma_i)\boldsymbol{v}_{i+1} \\ \sigma_{i+1} &= \sigma_i - \frac{(\boldsymbol{v}_{i+1})_{p_{i+1}}}{(\boldsymbol{x}_{i+1})_{p_{i+1}}} \end{aligned} \tag{6.5}$$

where p_i is the index of the component of maximum modulus in $\boldsymbol{x}_i$. The author (1964) has shown that this iteration is of second order, and that this is true also of the iteration

$$\begin{aligned} \boldsymbol{H}(\sigma_i)\boldsymbol{x}_{i+1} &= \frac{d\boldsymbol{H}}{d\sigma}(\sigma_i)\boldsymbol{x}_i/(\boldsymbol{x}_i)_{p_i} \\ \sigma_{i+1} &= \sigma_i - (\boldsymbol{x}_i)_{p_{i+1}}/(\boldsymbol{x}_{i+1})_{p_{i+1}}. \end{aligned} \tag{6.6}$$

In equations (6.5) and (6.6) $d\boldsymbol{H}/d\sigma$ is given by

$$\begin{aligned} \frac{d\boldsymbol{H}}{d\sigma} &= k^2\boldsymbol{I}^{(t)} - 2h^{-2}\frac{d\boldsymbol{Q}_1^{1(t)}}{d\sigma} \\ &= k^2\boldsymbol{I}^{(t)} + 2\{h^{-2}\sum_{i=1}^{m}(\boldsymbol{B} + (\sigma + \mu_i)\boldsymbol{I})_{11}^{-2} \\ &\qquad \kappa_i(\boldsymbol{S})\rho_i(\boldsymbol{S}^{-1})\}^{(t)}. \end{aligned} \tag{6.7}$$

These two iterations applied to $\boldsymbol{H}(\sigma)$ are compared in **Table 6.1**. It will be seen that the number of iterations required is very similar in each case, so that it would seem that (6.6) is to be preferred as it requires only one forward and back substitution per step. In every case the iteration is terminated when $|\sigma_{i+1} - \sigma_i| < 10^{-6}$. With the exception of the case discussed in the next paragraph, the location of the eigenvalues presented no great difficulty. The values of σ_0 were obtained from preliminary calculations made with slightly different values of m, n, and t to those given in Table 6.1. In these preliminary calculations the starting values were found by interpolating between the known eigenvalues of the separable problems with the boundary conditions $\phi(r, 0) = 0$ and $\partial\phi/\partial z(r, 0) = 0$. This technique was successful for all but the smallest eigenvalues, and for these it was necessary to try several different values of σ_0. However, this was perfectly feasible as the largest matrix $\boldsymbol{H}$ considered was only 10×10 (and in this case the big matrix was 480×480). This clearly demonstrates the compactness of our method.

Difficulty is to be expected if σ at any stage comes close to making any of the matrices $[B + (\sigma + \mu_i)I]$ singular. If this happens then an eigenvalue of the mixed boundary problem is close to an eigenvalue of the problem with $\boldsymbol{\phi}_0 = 0$. This difficulty was encountered in applying the iteration (6.5) both with $m = 10$, $t = 3$, and $\sigma_0 = 0{\cdot}6$, and $m = 16$, $t = 5$ $\sigma_0 = 0{\cdot}6$. In the first case after 10 iterations the iteration converged to $\sigma \simeq 41$, and in the second case after 13 iterations it converged to $\sigma \simeq 28$. In these cases the iteration (6.6) was used with the initial value of $\sigma = 0{\cdot}605$ and it converged in the first case after 5 iterations to $\sigma = 0{\cdot}60964$, and in the second case after 4 iterations to $\sigma = 0{\cdot}60891$. In both cases $\sigma = 0{\cdot}6167$ is an eigenvalue of the problem for the boundary condition $\boldsymbol{\phi}_0 = 0$. In all cases it was observed that the ultimate convergence to an eigenvalue σ was from above (so that the final corrections were negative), so that starting the iterations below the desired value of σ can be expected to aggravate the difficulties inherent in there being a close root of the problem with $\boldsymbol{\phi}_0 = 0$.

Table 6.1
Summary of numerical results

Mesh $n = 30$, $m = 10$

t	σ_0	σ	ITERATION (6.5)	ITERATION (6.6)
3	5·5	5·4724	5	5
3	15	14·987	4	4
6	0·52	0·54805	5	3
6	5	4·8547	5	5
6	12·5	12·581	3	4

Mesh $n = 30$, $m = 16$

t	σ_0	σ	ITERATION (6.5)	ITERATION (6.6)
5	5·5	5·4653	5	6
5	15	15·008	3	4
10	0·52	0·54501	5	4
10	5	4·8285	5	5
10	12·5	12·545	3	4

7. Acknowledgement

The author is indebted both to Professor Bickley and Mr. Sidney Michaelson for drawing his attention to some of the extensions of the Bickley-McNamee techniques that are given here. The advantages to be gained by using the second form for the influence matrices (equation (4.8)) were pointed out to the author by Mr. Sidney Michaelson.

References

Bickley, W. S., and McNamee, J. (1960). "Matrix and other Direct Methods for the Solution of Systems of Linear Difference Equations," *Phil. Trans.*, 1005, Vol. 252, pp. 69–131.

Osborne, M. R., and Michaelson, S. (1964). "The numerical solution of eigenvalue problems in which the eigenvalue parameter appears nonlinearly, with an application to differential equations," *The Computer Journal*, Vol. 7, pp. 66–71.

Osborne, M. R. (1963). "Iterative procedures for solving finite-difference approximations to separable partial differential equations," *The Computer Journal*, Vol. 6, pp. 93–99.

Osborne, M. R. (1964). "A new method for the solution of eigenvalue problems," *The Computer Journal*, Vol. 7, pp. 228–232.

Varga, R. S. (1962). *Matrix Iterative Analysis*, Prentice-Hall.

Wilson, S. B. L. (1962). "A Difference Method for the Numerical Solution of Boundary Value Problems." Thesis. Glasgow University.

Book Review

Integration of Equations of Parabolic Type by the Method of Nets, by V. K. Saul'yev, 1964; 346 pages. (Oxford: *Pergamon Press Ltd.*, 80s.)

This useful book is essentially a practical guide to the numerical solution of parabolic equations, and incidentally of elliptic equations also, by finite-difference methods. The first part of the book, accounting for rather more than half its length, is devoted to the finite-difference approximation of parabolic equations; that is, to the replacement of a parabolic equation by a system of algebraic equations. Systems of explicit and mixed types are derived, and while the emphasis is naturally on uniform nets, the non-uniform net is also considered, as is the net with fictitious nodes (to deal with irregular boundaries). Most of the formulae concern problems with one or two space variables, but the three-dimensional case is discussed; indeed most situations of common practical occurrence are covered. There is a section on equations of order greater than two, and one on nonlinear equations. Many contributions of the author and his associates appear in this part of the book; some will probably be new to Western readers. The stability of all formulae is considered, and the truncation error is examined at all relevant stages.

Part II is concerned with the solution of the systems of algebraic equations derived in Part I. Since explicit formulae are solved trivially, the methods discussed are those for solving an implicit system; in the important two-dimensional case they are essentially methods for solving, at each interval of time, the system arising from an *elliptic* equation. Direct methods are dismissed early in the discussion on the ground that they usually demand too much arithmetic. A comparison of various iterative methods therefore dominates this part of the book. There are accounts of the Jacobi and Gauss–Seidel methods, successive over-relaxation, variational methods (including the method of steepest descent and the method of conjugate gradients), methods using Chebyshev polynomials, and methods using block iteration (including alternating-direction methods). There are also sections on iterative methods of second and higher degrees (a method of nth degree being one in which the evaluation of the mth approximation $u^{(m)}$ to the solution requires knowledge of $u^{(m-1)}, u^{(m-2)}, \ldots, u^{(m-n)}$.) Questions of error, convergence, amount of computation and computer storage space required are discussed fully.

There are many references, particularly in Part I, to papers dealing more fully with the theoretical background to some of the methods. Many of these references are, as might be expected, to Russian sources, but several of the more important Western works have been added to this translation.

Part I covers its subject-matter admirably, and if Part II is not quite up to date, this is perhaps understandable in a survey of a subject which is still developing rapidly. The reader might consult R. S. Varga, *Matrix Iterative Analysis* (Prentice-Hall, 1962) for a more complete account. The translation is generally very good; seldom does the style remind the reader that it *is* a translation. The title of the book might be misleading, in that "method of nets" will be a term unfamiliar to many people, and there may be a suspicion that a basically *new* method is being proposed. The author's preference to restrict the term "method of finite differences" to the corresponding method for *ordinary* differential equations is difficult to understand, and hardly likely to find general acceptance.

The author has not set out to make his book completely comprehensive. He mentions topics which are *not* covered; they include problems with general boundary conditions and problems with moving boundaries. He says "The present book is designed for a wide class of readers, having direct or indirect contact with the numerical solution of parabolic and elliptic net equations (particularly heat conduction and Laplace's equation)." The reader who is often confronted with such problems will be very grateful for this volume, which provides the desired information with all the necessary warnings.

C. W. Clenshaw