PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Matrix and other Direct Methods for the Solution of Systems of Linear Difference Equations

W. G. Bickley and J. McNamee

Phil. Trans. R. Soc. Lond. A 1960 **252**, 69-131
doi: 10.1098/rsta.1960.0001

MATRIX AND OTHER DIRECT METHODS FOR THE SOLUTION OF SYSTEMS OF LINEAR DIFFERENCE EQUATIONS

By W. G. BICKLEY and J. McNAMEE

(*Communicated by G. Temple, F.R.S.—Received 26 November 1958*)

CONTENTS

VOL. 252. A. 1005. (Price 20s.) 9 [Published 21 January 1960

The investigations described in this paper were initiated in an attempt to replace by direct methods the successive approximation methods such as those of Southwell and Thom for the solution of systems of difference equations arising in the approximate solutions of linear partial differential equations. Boundary problems of this type form the subject of part III, which is the kernel of the paper. As the work progressed it was found that the methods evolved were applicable, and capable of extension, to step-by-step solutions also, and to ordinary as well as partial differential equations. These topics are presented in parts I, II and IV.

Matrix methods naturally predominate. The methods are illustrated by small-scale examples worked on desk machines, but the operations involved are, we believe, capable of being handled efficiently and simply by modern high-speed digital computers.

Prefatory note

When we began this work two years ago, our main interest was in the numerical solution of partial differential equations of elliptic type and we envisaged publication of part III as a separate and complete paper. Subsequently, it proved not difficult to extend some of the methods of part III to other types of partial differential equation and the results we have obtained here are given in part IV.

One does not proceed very far in the numerical study of partial differential equations without encountering difficulties whose full elucidation is best achieved in the simpler context of ordinary differential equations. As instances of such difficulties, we cite the numerical treatment of non-linear equations and the problem of control, i.e. the exclusion of unwanted solutions; we consider such questions in part I. Part II is concerned with boundary-value problems of ordinary differential equations and is intended mainly as a simple introduction to the concepts and notation of part III.

The four parts of the paper are unified by their subject-matter and treatment, since the entire paper is concerned with methods for the numerical solution of differential equations. The compact notation of matrix algebra has been employed extensively. In writing the paper, we thought it preferable to make each part more or less self-contained, even at the cost of some repetition.

Our main method of attack is to replace the given differential system by a difference system and the techniques we propose are for the most part methods for solving difference equations. Current discrete techniques are more closely related to familiar methods of continuous analysis than might at first sight appear; indeed, some of the techniques presented in the paper are direct borrowings from continuous analysis. None the less, we have consistently adopted the point of view that the difference system is worthy of study in its own right. Sometimes we can improve the approximation (to the differential system), by including higher differences—though even then we are still, in practice, concerned with a difference equation.

It is worth emphasizing that the goodness or badness of the approximation depends not merely on the tabular interval chosen and the order of differences included, but also on the nature of the desired solution—in particular, on the distribution of its singularities and the rapidity of its component Fourier oscillations.

In general, we may hope that the difference equation will yield a solution which is at least co-tabular with the desired solution for the interval chosen and the prescribed ranges of the variables. It is intuitively evident that the property of co-tabularity is not in general sufficient to define a unique function when the variables range continuously over the

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

prescribed domain. Some method of using the tabular values to interpolate to non-tabular points is also required, and by imposing suitable restrictions, we may be able to define a unique interpolating function whose relationship to the desired solution can be precisely stated. Whittaker (1915) has shown that a unique interpolating function of a single variable can be defined by a set of tabular values uniformly spaced along the real axis; so far as we are aware, the properties of a similar interpolating function of two or more variables have not been investigated.

From another point of view, the discrete solution of a continuous linear problem can be regarded as an attempt to approximate the solution by means of discrete base functions. If the base functions can be interpolated in a unique way, the problem of assessing the value of a discrete solution can be treated as a particular instance of the approximation of continuous functions by continuous base functions.

Some of the material in the paper was presented in colloquia at Farnborough, Teddington, and Newcastle upon Tyne. The discussion which followed these colloquia has helped us to eliminate some faults of presentation; in particular, we are indebted to Dr E. T. Goodwin and Dr H. I. Scoins who read and commented on the draft of part III.

I. THE USE OF MATRIX OPERATORS IN THE NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS: INITIAL-VALUE PROBLEMS

1. Introduction

Systematic use of matrix arrays in the formulation of difference operations and difference equations is of fairly recent growth. In this paper we give some elementary examples of matrix operators which produce the successive differences of a univariate table; these operators are then used to elucidate certain features of methods for solving linear difference equations. We shall be concerned in parts I and II with those difference equations which arise in the numerical treatment of ordinary differential equations by discrete techniques. Initial-value problems are treated in § 3 of part I; boundary-value problems in part II. Our illustrations of matrix formulations are confined almost entirely to second-order differential equations; but the concepts and techniques can be applied more widely.

The presentation of difference schemes in the form of matrix operators acting on vector operands is prima facie a mere notational device and some of the results recorded here can be obtained otherwise. The matrix presentation has however two advantages. It codifies in a compact notation the essential elements of computing procedures and provides algorithms for carrying out the computation; and it is suggestive: sometimes, the achievement of a perspicuous notation brings computational shortcuts to our notice.

1·1. *Matrix notation*

We introduce here four simple matrices which are frequently employed throughout the paper: the unit matrix I and the matrix I_{ij} which has a unit in the (i, j)th location and zeros elsewhere; the auxiliary unit matrices, S and S^* (transpose of S). The only non-zero elements of S are units in the first superdiagonal; S^r has units in the rth superdiagonal. Zero elements of a matrix are frequently indicated by dots; occasionally, a dot is used to indicate a zero matrix.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Unless otherwise specified, square matrices employed are assumed to be of order n. The operand (generally a column or row vector, occasionally a rectangular $m \times n$ matrix) is distinguished throughout part I by a bold-face symbol. There appears to be no agreed symbolism for matrix operators and vector equations of the type considered in this paper, and the notation adopted here and later is tentative.

2. Difference and summation operators

The familiar difference operators can be represented by linear combinations of the matrices I, S and S^*, and we illustrate this here for backward and forward differences. We can regard S and S^* as operators, and it will be convenient to recall their operational properties.

The action of S and S^* on a matrix $\mathbf{Z}$ is a shift, up or down, right or left, and can be conveniently exhibited by a mnemonic due to Turnbull & Aitken (1945):

$$\left.\begin{aligned} S\mathbf{Z} &= \begin{bmatrix} z_{i+1,j} \\ \cdot \end{bmatrix}, \qquad S^*\mathbf{Z} = \begin{bmatrix} \cdot \\ z_{i-1,j} \end{bmatrix}, \\ \mathbf{Z}S &= [\cdot \quad z_{i,j-1}], \quad \mathbf{Z}S^* = [z_{i,j+1} \quad \cdot]. \end{aligned}\right\} \tag{1}$$

The shift gives rise to a vacant row or column (as indicated by the dot) and suppresses a row or column; for example, the first row of $\mathbf{Z}$ is suppressed when $S\mathbf{Z}$ is formed.

We now consider a sequence of function values $\ldots, z_{-2}, z_{-1}, z_0, z_1, z_2, \ldots$, and select from it the sequence $z_1, z_2, \ldots, z_n$ which we array in a column $\mathbf{z}$. To obtain the backward differences of the tabular values in $\mathbf{z}$, we operate on $\mathbf{z}$ with $(I-S^*)$:

$$(I-S^*)\,\mathbf{z} = \mathbf{z}'.$$

The difference ∇z_i $(i+1)$ now occupies the position previously occupied by z_i; i.e. $(I-S^*)$ is a representation of the difference operator ∇. The difference column contains a redundant element z_1, since the first element of $\mathbf{z}$ is unaltered by the operation. Operating again with $(I-S^*)$, we obtain

$$(I-S^*)^2\,\mathbf{z} = \mathbf{z}'',$$

i.e. a column of second difference with two redundant elements. Noting, however, that $(I-S^*)$ possesses an inverse, we have

$$\mathbf{z} = (I-S^*)^{-1}\,\mathbf{z}' = (I-S^*)^{-2}\,\mathbf{z}'',$$

and it is seen that the redundant elements are precisely those needed to build the tabular values from the differences. The inverse $(I-S^*)^{-1}$ is a lower-triangular matrix all of whose elements are units; it is a summation operator. It is sometimes desirable to suppress the redundant element: this can be achieved by operating, not on $\mathbf{z}$, but on $\mathbf{z}-\mathbf{z}_1$, where $\mathbf{z}_1$ is a vector all of whose elements are z_1.

We may notice here an interpretation of initial-value processes which is suggested by the Heaviside operational theory. To this end, we introduce the notation $\mathbf{z}^{(m)}$ for a vector whose entries are $\nabla^m z_i$ $(i = 1, \ldots, n)$ and $\mathbf{z}_j^{(m)}$ for a vector whose elements have the constant value $\nabla^m z_j$. Noting, for example, that

$$(I-S^*)\,(\mathbf{z}-\mathbf{z}_0) - \mathbf{z}_0^{(1)} = \mathbf{z}^{(1)} - \mathbf{z}_0^{(1)},$$

we have

$$(I-S^*)\,[(I-S^*)\,(\mathbf{z}-\mathbf{z}_0) - \mathbf{z}_0^{(1)}] = \mathbf{z}^{(2)}. \tag{2}$$

From this we deduce that the difference equation

$$\nabla^2 z_i = f_i, \tag{3}$$

with the initial values z_0 and ∇z_0, has the solution

$$\mathbf{z} = \mathbf{z}_0 + (I - S^*)^{-1}\mathbf{z}_0^{(1)} + (I - S^*)^{-2}\mathbf{f}. \tag{4}$$

This interpretation can be extended to more general applications; but it is perhaps less well adapted to step-by-step computing routines than is the alternative method of § 3·1. In anticipation of § 3·1, we employ the above interpretation to determine the result of operating with $(I - S^*)$ on a column whose elements are (written row-wise):

$$0, \quad 0, \quad \delta^{2m} f_2, \quad \delta^{2m} f_3, \quad \delta^{2m} f_4, \quad \ldots.$$

It is easily seen from the above reasoning that

$$(I - S^*)^{-2}\begin{bmatrix} 0 \\ 0 \\ \delta^{2m} f_2 \\ \delta^{2m} f_3 \\ \delta^{2m} f_4 \end{bmatrix} = \begin{bmatrix} \delta^{2m-2}(f_1 - f_1) \\ \delta^{2m-2}(f_2 - f_1) \\ \delta^{2m-2}(f_3 - f_1) \\ \delta^{2m-2}(f_4 - f_1) \\ \delta^{2m-2}(f_5 - f_1) \end{bmatrix} - (I - S^*)^{-1}\begin{bmatrix} 0 \\ \delta^{2m-1} f_{1\frac{1}{2}} \\ \delta^{2m-1} f_{1\frac{1}{2}} \\ \delta^{2m-1} f_{1\frac{1}{2}} \\ \delta^{2m-1} f_{1\frac{1}{2}} \end{bmatrix}$$

and we may express this result by the notation

$$(I - S^*)^{-2}\mathbf{f}^{[2m]} = \mathbf{f}^{[2m-2]} - \mathbf{f}_1^{[2m-2]} - (I - S^*)^{-1} S^* \mathbf{f}_{1\frac{1}{2}}^{[2m-1]}. \tag{5}$$

This result enables us to determine the column of values produced by double summation of differences of order $2m$ without explicitly operating with $(I - S^*)^{-2}$. The operation automatically interprets the constants of summation by securing that the first two elements of the sum of the columns on the right side of (5) are zero.

Hitherto we have employed the operator $(I - S^*)$ and its powers to represent backward differences. If—as is most frequent in practical applications of difference equations—we wish to advance from data at $x_1, x_2, \ldots$, this appears to be the natural interpretation.† We can obtain equivalent representations for forward differences by using the operator $(S - I)$ and its powers. For example, a complete forward difference table can be presented in matrix notation as

$$I\mathbf{Z} + (S - I)\,\mathbf{Z}S + (S - I)^2\,\mathbf{Z}S^2 + \ldots.$$

$\mathbf{Z}$ here denotes a matrix of indefinite lateral extent with $\mathbf{z}$ as its first column and zeros elsewhere; the post-multipliers S^r execute the requisite right shift. The differences appear as an upper-triangular table terminated by the backward diagonal, the elements below this diagonal being redundant. A lower-triangular table of backward differences—such

† The designation of the entries in a difference table as backward, forward, or central, is somewhat misleading. The entries in the table are independent of the labelling; when we wish to use the entries, we select from them a sequence appropriate to our purpose, and it is more correct to attribute the designation backward, forward, or central, to this sequence.

as is produced by the National Accounting machine—can be obtained if the fore-operators $(S-I)^r$ are replaced by $(I-S^*)^r$.

Another form of difference table—with no redundant elements—can be obtained if the function values are located along the principal diagonal of a matrix which we call $\mathbf{Y}$ to distinguish it from $\mathbf{Z}$ above. From the properties of S it is clear that

$$S\mathbf{Y}-\mathbf{Y}S \tag{6}$$

is a superdiagonal matrix whose elements are the first forward differences of the tabular values. Repeating the operation,

$$S^2\mathbf{Y}-2S\mathbf{Y}S+\mathbf{Y}S^2$$

is a matrix with second differences in the second super-diagonal, $\Delta^2 y_i$ being in the ith row. When the units in the super-diagonal of S are replaced by $1/h_1, 1/h_2, \ldots$, the entries in the matrix array after m operations are

$$\frac{\Delta^m y_i}{\prod_{r=0}^{m-1} h_{i+r}}.$$

Conventional divided differences can be obtained by a simple modification of the successive operators. It may be noted that (6) above is reminiscent of the definition of a derivative in matrix mechanics.

We shall make considerable use of the operator $(I-S^*)^2$ and its inverse in § 3. A significant feature of these operators is that they are lower-triangular and can be extended indefinitely downward. If the computation is interrupted after n rows, we can extend it to additional rows without modification of the computation already completed; the operators are, then, appropriate to step-by-step arithmetic.

3. Matrix formulation of initial-value problems

In this section we formulate some algorithms for the numerical solution of second-order differential equations. We follow current practice in replacing differential operators by central-difference expansions truncated at a suitable order of difference. The matrix presentation of this truncated expansion assumes different forms according as the data is of initial-value or boundary-value type. A further distinction arises from the nature of the equation to be solved. If this equation is (i) linear or (ii) formally soluble for the highest derivative, initial-value problems are generally tractable. Computing routines for equations of the type (ii) are usually unsophisticated and in general they do not make use of any simplifying features in the form of the equation to be solved; the conceptual simplicity of such routines renders them widely applicable and easy to comprehend. The relative sophistication of matrix methods may appear to be an unnecessary refinement in the numerical treatment of these equations; we have, however, explained in some detail the routine of § 3·1, since it may be of use in electronic computing and it appears to differ considerably from routines currently employed. Section 3·2 gives methods for linear equations of initial-value type; these methods depend essentially on simplifying features which are frequently present in the second-order differential equations of mathematical physics.

3·1. *Equations formally soluble for the highest derivative*

We consider first the equation

$$\mathrm{d}^2z/\mathrm{d}x^2 = f(x, z) \tag{7}$$

(the modifications required when the derivative $\mathrm{d}z/\mathrm{d}x$ is explicitly present in f are considered later). This equation can be written in the central-difference form

$$\delta^2\left(1-\frac{\delta^2}{12}+\frac{\delta^4}{90}-\ldots\right)z = h^2 f, \tag{8}$$

h being the tabular interval. Dividing by the expression in brackets, this becomes

$$\delta^2 z = h^2\left(1+\frac{\delta^2}{12}-\frac{\delta^4}{240}+\frac{31\delta^6}{60480}-\ldots\right)f. \tag{9}$$

We may use the initial data (which invariably involves a derivative) at, say, z_1 to determine two starting values z_1, z_2. Equation (9) and the initial conditions then determine a set of equations of which the first four are

$$\left.\begin{aligned} z_1 &= z_1,\\ -2z_1+z_2 &= -2z_1+z_2,\\ z_1-2z_2+z_3 &= h^2f_2+\frac{h^2}{12}\delta^2f_2-\frac{h^2}{240}\delta^4f_2+\ldots,\\ z_2-2z_3+z_4 &= h^2f_3+\frac{h^2}{12}\delta^2f_3-\frac{h^2}{240}\delta^4f_3+\ldots. \end{aligned}\right\} \tag{10}$$

The matrix on the left of (10) is $(I-S^*)^2$. If this operator is inverted, the right side of (10) can be summed row by row as the differences become available. The double sum of the column $\mathbf{v}$ whose elements are

$$z_1, \quad -2z_1+z_2, \quad h^2f_2, \quad h^2f_3, \quad \ldots$$

can be expressed as a recurrence relation. Using the notation

$$\mathbf{w} = (I-S^*)^{-2}\,\mathbf{v},$$

then

$$w_i = 2w_{i-1}-w_{i-2}+h^2f_{i-1} \quad (i>2). \tag{11}$$

Explicit operation with $(I-S^*)^{-2}$ on columns involving differences of f_i can be evaded by using (5). As illustration, we write out the first four rows of the summed equations:

$$\begin{bmatrix} z_1\\ z_2\\ z_3\\ z_4 \end{bmatrix} = \begin{bmatrix} w_1\\ w_2\\ w_3\\ w_4 \end{bmatrix} + \tfrac{1}{12}h^2 \begin{bmatrix} f_1-f_1\\ f_2-f_1\\ f_3-f_1\\ f_4-f_1 \end{bmatrix} - (I-S)^{-1} \begin{bmatrix} \cdot\\ \delta f_{1\frac{1}{2}}\\ \delta f_{1\frac{1}{2}}\\ \delta f_{1\frac{1}{2}} \end{bmatrix} + \ldots.$$

We may write these equations in the symbolic form

$$z_i = w_i + h^2\Sigma^2\left\{\frac{\delta^2 f_i}{12}-\frac{\delta^4 f_i}{240}+\ldots\right\}, \tag{12}$$

where $\Sigma^2\delta^{2m}$ symbolizes the matrix summation of (5). In the practical use of this equation, we must be able to build $f(x_i, z_i)$ from its estimated or computed differences before we know z_i; (12) is, then, an equation to determine z_i.

This routine differs considerably from current summation procedures (e.g. Herrick 1951; Milne 1953; I.A.A.T. 1956). The main deviation is in the choice of summation constants: starting values z_1 and z_2 are computed in the normal way from the Taylor series but each column of summed differences carries its own summation constants which ensure that the pre-set values z_1 and z_2 are not disturbed by the addition of higher-order differences.

Minor deviations are:

(i) the Taylor series is converted into a factorial series (see example 3·11 below) and used to compute even-order differences $\delta^{2m}f_1$ and odd-order differences $\delta^{2m-1}f_{1\frac{1}{2}}$ as far as necessary or practicable;

(ii) the factorial series is used to choose and compute a difference of moderately high order which varies only slowly throughout the range (it is sometimes assumed that a difference of some order is constant and any error resulting from this assumption is corrected in a trial-and-error process);

(iii) the elements of the column $\mathbf{w}$ are of the order of magnitude of the initial values [in a currently used method (Herrick 1951), they are of the order h^{-2} (initial values)];

(iv) the summed differences which we have denoted by $\Sigma^2\delta^{2m}$ are in general considerably smaller than the differences δ^{2m-2} but this does not necessarily enhance convergence since the elements of the column $\mathbf{w}$ are here of smaller magnitude;

(v) odd-order differences (save for the starting differences) are omitted, but this convenience may be lost if the derivative $\mathrm{d}z/\mathrm{d}x$ is explicitly present in the function f of (7).

It is convenient to interpolate here a note on first-order summation—a process required in the numerical solution of second-order equations of the type

$$\mathrm{d}^2z/\mathrm{d}x^2 = f(x, z, \mathrm{d}z/\mathrm{d}x).$$

The above method of solution then requires a subroutine to determine $\mathrm{d}z/\mathrm{d}x$. We can exhibit the subroutine most simply by considering the first-order equation

$$\mathrm{d}z/\mathrm{d}x = f(x, z), \tag{13}$$

taking (for simplicity of explanation) the initial value z_1 to be zero. We replace (13) by the finite-difference equivalent

$$\delta z_{i-\frac{1}{2}} = \nabla z_i = h\mu(1 - \tfrac{1}{12}\delta^2 + \tfrac{11}{720}\delta^4 - \ldots)f_{i-\frac{1}{2}}, \tag{14}$$

the backward-difference operator being introduced, since we shall use (14) as an equation to determine z_i. Suppressing for the moment, the difference $\mu\delta^2 f$ and higher-order differences, (14) can be written in matrix form as

$$(I - S^*)\,\mathbf{z} = \tfrac{1}{2}h(I + S^* - I_{11})\,\mathbf{f},$$

i.e.

$$\mathbf{z} = \tfrac{1}{2}h(I - S^*)^{-1}(I + S^* - I_{11})\,\mathbf{f}.$$

The operator $T = \frac{1}{2}(I - S^*)^{-1}(I + S^* - I_{11})$ is the trapezoidal-sum operator since the pth element of $T\mathbf{f}$ is

$$\tfrac{1}{2}f_1 + f_2 + \ldots + \tfrac{1}{2}f_p.$$

If the difference columns are now restored, (14) may be treated by operating directly with $(I-S^*)^{-1}$ on even-order difference columns; alternatively, the values yielded by summation can be expressed as odd-order differences by an interpretation similar to (5). The first element of each difference column is zero, and the first element of the corresponding summed column is zero.

A linear example which has been studied in detail by Herrick (1951) permits a simple presentation of the numerical solution of (7) and at the same time affords a direct numerical comparison with an alternative method. The first four rows of the computation are given in example 3·11.

Example 3·11
$$\frac{\mathrm{d}^2z}{\mathrm{d}x^2}+z=0, \quad z_1=z(0)=0, \quad \frac{\mathrm{d}z}{\mathrm{d}x}=1 \quad \text{for} \quad x=0.$$

A second starting value can be obtained from the Taylor series. The starting differences can also be obtained if the Taylor series is converted into a central factorial series, using the identities
$$\frac{x^r}{h^r r!}=m_{\{r\}}+(r+1)\,[a_{r2}m_{\{r-2\}}+a_{r4}m_{\{r-4\}}+\ldots],$$
where $x=mh$, and the central factorials and reduced central factorials are defined by
$$m^{\{r\}}=\prod_{p=0}^{r-1}\{m-p+\tfrac{1}{2}(r-1)\}, \quad m_{\{r\}}=\frac{m^{\{r\}}}{r!}, \quad \delta^p m_{\{r\}}=m_{\{r-p\}}.$$

The notation for central factorials and reduced central factorials is that of Aitken (1932). The first five of the coefficients $a_{r,\,2p}$ are
$$a_{r2}=\frac{1}{24}, \quad a_{r4}=\frac{3+5(r-4)}{5760},$$
$$a_{r6}=\frac{9+7(r-6)\,(5r-26)}{29\,03040},$$
$$a_{r8}=\frac{15+(r-8)\,(175r^2-2695r+10449)}{13934\,59200},$$
$$a_{r,\,10}=\frac{9+(r-10)^2\,(385r^2-7700r+38874)}{36\,78732\,28800}.$$

The above coefficients suffice for the conversion of the first twelve powers, save for the constant term in the factorial polynomial for the reduced twelfth power; this term is $(2^{12}\,12!)^{-1}$.

The wanted function in this example has no singularity except at infinity and the differences at $x=0$ and $x=\frac{1}{2}$ can be calculated easily. We can go further and compute the fourth differences from $x=0{\cdot}0$ to $x=1{\cdot}0$, using an interval $h=0{\cdot}1$. The computed fourth differences of z from the first two terms of the expression for $\delta^4\mathrm{d}^2z/\mathrm{d}x^2$, i.e.
$$-h^5m+h^7[m_{\{3\}}+\tfrac{1}{3}m]$$
are

m	1	2	3	4	5	6	7	8	9	10
$-h^{-7}\delta^4\mathrm{d}^2z/\mathrm{d}x^2$	100	198	295	389	478 (1)	563 (1)	642 (1)	713 (3)	775 (5)	832 (8)

The errors in the computed values are recorded in brackets. Remembering the overall factor h^2 in (12), it can be seen that conversion of the first four terms of the Taylor series enables us to determine z to 8 or 9 decimals in the range 0·0 to 1·0.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

The starting differences, multiplied by 10^7, are

$$\delta\frac{\mathrm{d}^2z(\frac{1}{2})}{\mathrm{d}x^2} = -9\,98334, \quad \delta^3\frac{\mathrm{d}^2z(\frac{1}{2})}{\mathrm{d}x^2} = 9975, \quad \delta^5\frac{\mathrm{d}^2z(\frac{1}{2})}{\mathrm{d}x^2} = -100,$$

and the first four rows of the computation are

x	z	w	$\frac{\mathrm{d}^2z}{\mathrm{d}x^2}$	$\delta^2\frac{\mathrm{d}^2z}{\mathrm{d}x^2}$	$\delta^4\frac{\mathrm{d}^2z}{\mathrm{d}x^2}$	$\Sigma^2\delta^2\frac{\mathrm{d}^2z}{\mathrm{d}x^2}$	$\Sigma^2\delta^4\frac{\mathrm{d}^2z}{\mathrm{d}x^2}$
0·0	0000 00000	0000 00000	00 00000	00000	000	00000	000
0·1	0998 33417	0998 33417	−09 98334	9975	−100	00000	000
0·2	1986 69331	1986 68500	−19 86693	19850	−198	9975	−100
0·3	2955 20209	2955 16890	−29 55202	29527	−295	39800	−398

e.g. using (12), $$z(0{\cdot}3) = 2955\,16890 + \frac{39800}{12} + \frac{398}{240}.$$

The use of the converted Taylor series made it unnecessary to estimate ahead in this example. The Taylor series is, however, quite rapidly convergent here and the higher difference columns taper off satisfactorily (about two digits from one column to the next). Less satisfactory tapering will frequently be encountered in solving non-linear equations.

As a counter-example, we cite the non-linear equation

$$\frac{\mathrm{d}^2z}{\mathrm{d}x^2} + z^3 = 0, \quad \text{where} \quad z(0) = 1, \quad z'(0) = 0.$$

This equation is satisfied by the Jacobian elliptic function $\mathrm{cn}\,(x, k)$ when the modulus k has the value $1/\sqrt{2}$; the function is meromorphic, the pole nearest to the origin being at $x = \mathrm{i}K'$, where $K' = 1{\cdot}85\ldots$. The Taylor series of $\mathrm{d}^2z/\mathrm{d}x^2$ is

$$-\frac{\mathrm{d}^2z}{\mathrm{d}x^2} = 1 - \frac{3}{2!}x^2 + \frac{27}{4!}x^4 - \frac{441}{6!}x^6 + \frac{11529}{8!}x^8 - \frac{4\,42827}{10!}x^{10} + \ldots.$$

Near the origin, the tenth difference appears to be of the order $4\,42827h^{10}$, and evaluation or estimation of the differences will be difficult even with an interval of $h = 0{\cdot}1$. Suppose, for example, we attempt to evaluate sixth differences from the first three relevant terms of the factorial series, i.e. from the three terms

$$h^6[441 - 11529h^2(m_{\{2\}} + \tfrac{3}{8}) + 4\,42827h^4(m_{\{4\}} + \tfrac{11}{24}m_{\{2\}} + \tfrac{121}{1920})].$$

With $m = 0, 1, 2$, the three terms sum to

$$413 \qquad 363 \qquad 235$$

and for larger values of m the truncation error of the factorial series is apparent. Since the Taylor series is convergent up to $x = 1{\cdot}85$, it may appear surprising that the truncated factorial series fails so drastically to evaluate sixth differences in the range $x = 0{\cdot}0$ to $x = 1{\cdot}0$. In this instance, the differences can be evaluated from existing tables of the function cn, or directly from the Fourier series (in standard notation)

$$\delta^6\frac{\mathrm{d}^2}{\mathrm{d}x^2}\,\mathrm{cn}\,(x, k) = \frac{16\pi}{Kk}\sum_{n=0}^{\infty}\frac{a_n^2 q^{n+\frac{1}{2}}(1-\cos a_n h)^3\cos a_n x}{1+q^{2n+1}}, \quad a_n = (n+\tfrac{1}{2})\frac{\pi}{K}.$$

The sixth differences for $x = 0{\cdot}0\,(0{\cdot}1)$ are found to be

x	0·0	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9	1·0
$10^6\delta^6$	413	362	228	55	−105	−212	−250	−230	−173	−103	−39

and the rapid variation of the differences explains the failure of the polynomial representation; it is also clear that estimating ahead would be very difficult if this interval were used. With $h = 0{\cdot}05$, the sixth differences are reduced to $O(7 \times 10^{-6})$ and a step-by-step computation becomes practicable. Even with this interval, the factorial series (truncated as above) fails to evaluate the sixth differences for values of m greater than 4 ($x > 0{\cdot}2$); the differences, though now small, still vary rapidly.

Clearly, step-by-step computation compares unfavourably here with the alternative computation by means of the rapidly convergent Fourier series. The point we wish to emphasize in example 3·11 and the counter-example is that the factorial series helps us to decide whether a step-by-step solution of a given equation is practicable; if it is, and if no better method is available, the series enables us to choose a tabular interval appropriate to the accuracy required; sometimes, we can go further and evaluate (in advance of the main computation) a chosen small difference throughout the required range or a part of the range.

3·2. *Linear initial-value problems of second-order equations*

A very simple example will serve to illustrate the main point with which this section is concerned. Let us suppose that we wish to determine the decreasing function e^{-x} from the differential system

$$\frac{\mathrm{d}^2 z}{\mathrm{d}x^2} - z = 0, \quad z(0) = 1, \quad z'(0) = -1. \tag{15}$$

If we approximate to the differential equation by the second-order difference equation (cf. (7) and (9))

$$\delta^2(1 - \tfrac{1}{12}\delta^2)\, z_i - h^2 z_i = 0, \tag{16}$$

and if we then write the difference equation in the equivalent form

$$\left(E^2 - \frac{24 + 10h^2}{12 - h^2} E + 1\right) z_i = 0 \quad (E = \mathrm{e}^{h\mathrm{d}/\mathrm{d}x}),$$

it is clear that (16) is equivalent to the two equations

$$z_{i+1} - \rho z_i = 0 \quad (a), \qquad z_{i+1} - (1/\rho)\, z_i = 0 \quad (b), \tag{17}$$

ρ being the larger root of the above quadratic. The decreasing solution of (16) can be computed readily by means of (17 b); (16) itself is not a practical equation for the decreasing solution, since the inevitable round-off errors will be magnified by the increasing factor in the operator.

The solutions of (16) are $z_r = \mathrm{e}^{\pm r \ln \rho}$

and the value of $h\,\mathrm{d}/\mathrm{d}x$ at $x = 0$ for those solutions is $\pm \ln \rho$; e.g. with $h = \frac{1}{2}$, $\ln \rho = 0{\cdot}50006\ldots$. Hence, the decreasing solution of the *difference* equation cannot be obtained from the starting data appropriate to the decreasing solution of the differential equation.

This illustration brings to light two practical points which will guide us in the numerical solution of more complicated equations: (i) before we begin computation, we seek to exclude the increasing solution by factorizing the difference operator; (ii) we ignore, at the outset, the initial conditions appropriate to the decreasing solution of the differential equation and accept the initial conditions appropriate to the decreasing solution of the difference equation.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

If the decreasing solution of the difference equation is an acceptable first approximation to the solution of the differential system, and if we then seek to improve this approximation by subsequent inclusion of higher differences, the increasing solution of the difference equation is admitted and must be controlled. Control is difficult, and will occupy us later; for the moment, we shall ignore higher differences and consider only second-order *difference* processes.

The basic operator for initial-value difference systems is of the form $I-S^*A$, where A is a diagonal matrix and the simplest forms of first- and second-order problems are

$$(I-S^*A)\,\mathbf{z} = \mathbf{f} \tag{18}$$

and

$$(I-2S^*A+S^{*2}B)\,\mathbf{z} = \mathbf{f}, \tag{19}$$

where $\mathbf{f}$ is a known column and A and B are diagonal; in (18), $\mathbf{f}$ has a single non-zero element, f_1, and in (19) $\mathbf{f}$ has two non-zero elements, f_1 and f_2.

It will be convenient to express the operator in (19) as a product of two non-commuting factors; we write

$$I-2S^*A+S^{*2}B = (I-S^*C)\,(I-S^*D), \tag{20}$$

where C and D are diagonal. Denoting the non-zero elements of A, B, C, D, by a_i, b_i, c_i, d_i $(i = 1, 2, \ldots)$, we have

$$c_i+d_i = 2a_i, \quad c_i d_{i-1} = b_i. \tag{21}$$

There are $2n-3$ equations to determine the relevant $2n-2$ elements c_i, d_i, and this leaves some freedom of choice in the determination. We shall suppose that the successive quadratics

$$\rho^2-2a_i\rho+b_i = 0 \tag{21a}$$

have real roots, ρ_i' and ρ_i'' $(\rho_i' > \rho_i'')$. One convenient choice is

$$c_1 = \rho_1'', \quad d_1 = \rho_1', \tag{22}$$

the remaining values being determined in order from (21). If the a_i and b_i do not vary too rapidly in the range considered, each element d_i is greater in absolute value than the corresponding element c_i.

A second choice is

$$c_{n-1} = \rho_{n-1}', \quad d_{n-1} = \rho_{n-1}'', \tag{23}$$

the matrices of (19) and succeeding equations being of order n; the c_i are now greater than the corresponding d_i.

(*Note*. The choice of coefficients in (22) and (23) can be motivated by consideration of a similar but simpler process, namely, a sequence of approximations to the roots of the quadratic equation $x^2-2ax+b = 0$. We suppose that the roots of the quadratic are real.

Let some value c_1 be chosen and let c_{i+1} be determined by the forward process

$$c_{i+1} = \frac{b}{2a-c_i}.$$

If ρ is one root of the quadratic, then

$$\rho-c_{i+1} = \rho-\frac{b}{2a-c_i} = \frac{\rho(\rho-c_i)}{2a-c_i},$$

i.e.

$$\frac{\rho-c_{i+1}}{\rho-c_i} = \frac{\rho}{b}\,c_{i+1}.$$

Since b is the product of the roots, the coefficient of c_{i+1} on the right is $1/\rho'$, where ρ' is the second root of the quadratic. Convergence of the sequence c_i to ρ requires that ρ' should be the greater root and ρ the lesser root.

Alternatively, let some value d_{n-1} be chosen and let d_{i-1} be determined by the backward process

$$d_{i-1} = 2a - b/d_i;$$

then

$$\rho - d_{i-1} = \rho - 2a + b/d_i,$$

i.e.

$$\frac{\rho - d_{i-1}}{\rho - d_i} = \frac{b}{\rho d_i}$$

and convergence of the sequence d_i to ρ requires that ρ should be the larger root.

If one starts with an approximation to the wrong root, the sequence oscillates violently and eventually crosses over to the correct root.

Similar considerations apply to the matrix-factorization process of the text if the a_i and b_i do not vary too rapidly. The choice of coefficients in (22) or (23) coupled with the equations (21) ensures that at each step the process is held stable.)

We ignore the initial values on the right of (19) and consider instead two specialized problems. Writing

$$\mathbf{z} = (I - S^*D)\,\mathbf{w} \tag{24}$$

in (19), we have

$$(I - S^*C)\,\mathbf{w} = \mathbf{f}. \tag{25}$$

We choose the components of $\mathbf{f}$ in (25) to be $1, -c_1, 0, 0, 0, \ldots$; the components of $\mathbf{w}$ are then $1, 0, 0, 0, \ldots$ and $\mathbf{z}$ is determined by (24). By means of the first factorization, the larger (variable) root is isolated. The second factorization permits the isolation of the smaller root, and the solution of the original initial-value problem can be achieved as the sum of two simpler initial-value problems. It is easily verified that the choice of the initial-value vector $\mathbf{f}$ in (25) determines the initial values of $\mathbf{z}$ to be $1, d_1$.

The backward factorization implicit in (23) is equivalent to a reverse computation of (19) from starting data at x_n and x_{n-1}; but the factorization process may be more convenient in practice in that it can achieve a more complete isolation of the decreasing solution and requires only one starting value.

We illustrate the factorization technique by means of the differential equation

$$h^2\,\mathrm{d}^2z/\mathrm{d}x^2 + h^2q(x)\,z = h^2f(x) \tag{26}$$

and its difference equivalent

$$\delta^2 z + h^2\left(1 + \frac{\delta^2}{12} - \frac{\delta^4}{240} + \ldots\right) q(x)\,z = h^2\left(1 + \frac{\delta^2}{12} - \frac{\delta^4}{240} + \ldots\right) f(x)$$

which we write in the form

$$\nabla^2\left[\left(1 + \frac{h^2q}{12}\right)z\right]_{i+1} + h^2q_iz_i = h^2\left(1 + \frac{\delta^2}{12} - \frac{\delta^4}{240} + \ldots\right)f_i + \frac{h^2}{240}\delta^4 q_i z_i - \ldots, \tag{27}$$

the first term on the left being written as a backward difference since the equation will be used to determine z_{i+1}. Equation (27) can be expressed in matrix form using the notation: Q for the diagonal matrix whose elements are h^2q_i, $\mathbf{v}$ for an initial column whose elements are determined from the initial values and the f_i (the mode of formation will be clear from

the numerical example below), and $\mathbf{t}$ for a truncation or correction vector which contains fourth and higher differences of qz. In this notation, (27) becomes

$$(I-S^*)^2\,(I+Q/12)\,\mathbf{z}+S^*Q\mathbf{z}=\mathbf{v}+\mathbf{t}. \tag{28}$$

The correction vector cannot be taken into account until a first approximation has been obtained and we temporarily ignore it. We can then obtain by inversion

$$\mathbf{z}=(I+Q/12)^{-1}\,(I-S^*)^{-2}\,[\mathbf{v}-S^*Q\mathbf{z}]$$

which is a variation on the summation method. Alternatively, we can write (28) in the form

$$[I-2S^*(I-5Q/12)\,(I+Q/12)^{-1}+S^{*2}]\,(I+Q/12)\,\mathbf{z}=\mathbf{v}+\mathbf{t}. \tag{29}$$

Equation (29) is a matrix formulation of the Numerov (1933) process and is a form suitable for use with the factorization algorithm above. This is illustrated numerically by example 3·21.

Example 3·21

To determine the decreasing function $x^{\frac{1}{2}}K_0(x)$ from the Bessel equation

$$\frac{\mathrm{d}^2z}{\mathrm{d}x^2}-\left(1-\frac{1}{4x^2}\right)z=0$$

in the range $x=2{\cdot}0$ to $x=5{\cdot}0$, the initial data being

$$z(2)=0{\cdot}16107\,03,\quad z'(2)=-0{\cdot}15753\,27.$$

We take the tabular interval to be $h=0{\cdot}5$. This interval would generally be considered large for a function of this type, since we are attempting to evaluate a decreasing solution in the presence of an increasing solution of the same equation and the singularity of the decreasing solution is only a few intervals away from the starting point. A large interval is, however, desirable in an illustrative example, since it enables us to exhibit clearly the truncation error and the disturbing influence of the singularity.

The component matrices encountered in the computation are diagonal matrices or column matrices, and to economize space the elements of these matrices are recorded in columns in tables 1 and 2.

The backward factorization of (23) is used, since we wish to obtain the decreasing solution. Comparing (19) and (29) and regarding $(1+\frac{1}{12}Q)\,\mathbf{z}$ as the variable in (29), we have

$$2A=2(I-5Q/12)\,(I+Q/12)^{-1}\quad(B=I),$$

and the relevant quadratic of (21 *a*) is

$$\rho^2-2{\cdot}25210\,08\rho+1=0,$$

i.e.

$$\rho=1{\cdot}64372\,74,\quad 0{\cdot}60837\,34.$$

The elements of the diagonal matrices C and D in the factors $I-S^*C$ and $I-S^*D$ can now be obtained from (21), and they are recorded in the second and third columns of table 1. If the truncation vector $\mathbf{t}$ is neglected, we need only the factor $I-S^*D$; we use (24) and then multiply the solution of (24) by the diagonal matrix $(I+Q/12)^{-1}$. Since the initial

value $z(2)$ is to be 0·16107 03, we multiply by a suitable constant to obtain the solution of the difference equation—recorded as solution 1 in table 2. The discrepancies between the solution of the differential equation and the solution of the difference equation are recorded in units of the seventh decimal.

TABLE 1. COMPONENT MATRICES

$2A$	C	D	$I+\frac{1}{12}Q$
—	—	—	0·9804688
2·2390437	1·6273105	0·6117332	0·9800000
2·2448980	1·6346995	0·6101985	0·9797454
2·2480803	1·6388109	0·6092694	0·9795918
2·2500000	1·6413101	0·6086899	0·9794922
2·2512462	1·6428728	0·6083734	0·9794239
2·2521008	1·6437274	0·6083734	0·9793750

TABLE 2. SOLUTION 1 AND SOLUTION 2

x	z solution 1	$10^7\times x^{\frac{1}{2}}K_0-z$	z solution 2	$10^7\times x^{\frac{1}{2}}K_0-z$	$10^7\times\frac{h^2}{240}\delta^4 qz$	$10^7\times\frac{h^2}{240}\delta^4 qx^{\frac{1}{2}}K_0$
1·5	—	—	0·2618414	159	—	—
2·0	0·1610703	0	0·1610703	0	—	—
2·5	0·0985792	9	0·0985787	14	16	16
3·0	0·0601685	21	0·0601673	33	−18	−18
3·5	0·0366648	14	0·0366625	37	−17	−17
4·0	0·0223198	− 4	0·0223160	34	−12	−12
4·5	0·0135797	−35	0·0135786	−24	− 8	− 8
5·0	0·0082619	−83	0·0082550	−14	− 5	− 5
5·5	—	—	0·0050172	−15	—	—
6·0	—	—	0·0030493	−22	—	—

$q=-1+1/4x^2$

It is difficult to improve this solution by inclusion of higher differences since the truncation error near the starting point cannot easily be obtained by difference techniques. We can, however, verify that the discrepancy is of the order of the fourth differences of $h^2qz/240$. To this end, we compute a second solution in the range $x=1{\cdot}5$ to $x=6{\cdot}0$ with the same tabular interval. It is unnecessary to record the computation; the solution obtained is recorded as solution 2 in table 2, and from this solution we compute the fourth differences of qz. The large discrepancy at $x=1{\cdot}5$ is probably to be attributed to the influence of the singularity.

The accuracy achieved here is due mainly to two factors: (i) the matrix factorization achieves a complete separation of the increasing and decreasing solutions of the difference equation; (ii) the Numerov difference operator on the left side of (29) is a very good approximation to the differential operator, and the two solutions (increasing and decreasing) of the difference equation are quite close to the corresponding solutions of the differential equation. The efficiency of the Numerov approximation is made more evident by comparison with the method of § 3·3 in which the Numerov operator is replaced by a simpler but less accurate operator.

We shall indicate in the next section how the solution of a truncated difference equation can be brought closer to the corresponding solution of the associated differential equation, provided that the singularities of the equation are a reasonable number of tabular intervals away from the segment in which the solution is to be determined.

3·3. *Linear equations with nearly constant coefficients*

We consider the typical equation

$$\mathrm{d}^2z/\mathrm{d}x^2+q(x)\,z=0 \quad (a\leqslant x\leqslant b), \tag{30}$$

where $$q(x)=-1+r(x)$$

and $r(x)$ is small compared with unity in the range $a\leqslant x\leqslant b$. This equation can be written in the difference form

$$(\delta^2-h'^2)\,z_i=-h'^2\left[1+\frac{\delta^2}{12}-\frac{\delta^4}{240}\right](rz)_i-\frac{h'^2}{240}\delta^4 z_i \quad \left(h'^2=\frac{h^2}{1-h^2/12}\right), \tag{31}$$

sixth and higher differences being omitted.

The operator on the left of (31) has constant coefficients; if the terms on the right are neglected, the equation possesses solutions which are analogous to the positive and negative exponentials. This suggests that we employ a change of variable which is familiar in the theory of asymptotic solutions. The simplicity of the operator in (30) enables us to dispense with matrix notation; but it is desirable to explain first a notation which is sufficiently general to embrace (27) and (29) as well as (31).

The term $\delta^2 z_i$ on the left of (31) may be replaced by $\nabla^2 z_{i+1}$ and thus is represented by $(I-S^*)^2\,\mathbf{z}$; the term $h'^2(rz)_i$ is represented by $R\mathbf{z}$, R being a diagonal matrix whose elements are $h'^2 r(x_i)$. We now introduce the change of variable

$$\mathbf{z}=(I+Z)\,\mathbf{w}, \tag{32}$$

where $$(\delta^2-h'^2)\,w_i=0 \tag{33}$$

and Z is a diagonal matrix whose elements are small compared with unity. Making use of (33), we can write equation (31) as

$$(I-S^*)^2\,Z\mathbf{w}-h'^2S^*Z\mathbf{w}=\mathbf{v}-h'^2[S^*+\tfrac{1}{12}(I-S^*)^2]\,R(I+Z)\,\mathbf{w}+\mathbf{t}, \tag{34}$$

where $\mathbf{v}$ is the starting vector which ensures that the first two equations are identities, and the truncation vector $\mathbf{t}$ embraces terms of the order $\delta^4 z_i$.

Equation (34) is merely another way of writing (29), and the device of using a diagonal matrix Z as an auxiliary variable can obviously be used in conjunction with the earlier equation. The discrepancy between the variable $\mathbf{w}$ of (34) and the solution of the differential equation is greater than in the earlier treatment which completely incorporated second differences in the first approximation. This disadvantage is offset by the advantage that the solution $\mathbf{w}$ can be written down at once; we have

$$w_i=c\rho^{-(i-1)} \quad (i=1,2,\ldots), \tag{35}$$

where c is the prescribed initial value at the starting point and ρ is the larger root of the quadratic

$$\rho^2-(2+h'^2)\,\rho+1=0. \tag{36}$$

This simplification enables us to write (31) in the recursive form

$$\delta'^2\left[\left(1+\frac{h'^2}{12}r\right)Z\right]_i+h'^2\rho(qZ)_i=-h'^2\left(\rho+\frac{\delta'^2}{12}\right)r_i+\frac{h'^2}{240\rho}\delta'^4[q(1+Z)]_i+O(\delta^6 z_i), \tag{37}$$

where $$\delta'^2 f_i=f_{i+1}-2\rho f_i+\rho^2 f_{i-1}.$$

We illustrate this device by using again the Bessel equation of example 3·21.

Example 3·31

To find the decreasing solution $x^{\frac{1}{2}}K_0(x)$ from the equation

$$\frac{d^2z}{dx^2}+\left(-1+\frac{1}{4x^2}\right)z=0$$

in the range $x=2{\cdot}0$ to $x=4{\cdot}0$, using the initial data of example 3·21.

We use a smaller tabular interval here, namely $h=0{\cdot}2$, since the first approximation w does not incorporate the complete second differences of the differential equation. From the run of the fourth differences in the preceding example, we expect that the fourth differences of $h^2qz/240$ with the smaller interval $h=0{\cdot}2$ will be about $20/(2{\cdot}5)^4$, i.e. about 0·5 in units of the seventh decimal; hence we should be able to obtain about 6 or 7 decimal accuracy if fourth differences are neglected. The initial data can be used to obtain a second starting value at $x=2{\cdot}2$:

$$z_1=0{\cdot}16107\,03, \quad z_2=0{\cdot}13240\,73.$$

If we insert the value of h'^2, the difference equation (37) becomes

$$\left[1+\frac{r_{i+1}}{299}\right]Z_{i+1}-2\rho\left[\frac{305}{299}-\frac{5r_i}{299}\right]Z_i+\rho^2\left[1+\frac{r_{i-1}}{299}\right]Z_{i-1}=-\frac{12}{299}\left[\rho+\frac{\delta'^2}{12}\right]r_i+O(\delta'^4Z_i),$$

where $\rho=1{\cdot}22140\,36$.

The equations to determine the Z_i can be sufficiently illustrated by displaying the first two and last two equations:

$$\begin{aligned}
1{\cdot}00014\,52Z_3-2{\cdot}48971\,66Z_2+1{\cdot}49213\,85Z_1&=-0{\cdot}00256\,70,\\
1{\cdot}00012\,37Z_4-2{\cdot}49005\,37Z_3+1{\cdot}49208\,45Z_2&=-0{\cdot}00215\,44,\\
&\cdots\cdots\cdots\cdots\\
1{\cdot}00005\,79Z_{10}-2{\cdot}49103\,86Z_9+1{\cdot}49193\,46Z_8&=-0{\cdot}00095\,38,\\
1{\cdot}00005\,23Z_{11}-2{\cdot}49115\,95Z_{10}+1{\cdot}49192\,29Z_9&=-0{\cdot}00085\,57,
\end{aligned}$$

where

$$Z_1=0, \quad 1+Z_2=\frac{0{\cdot}13240\,73}{c/\rho}=1{\cdot}00405\,12.$$

The values of Z and of the solution z are given in table 3.

The main interest in the numerical solution is the problem of control in passing from the decreasing solution of the difference equation (33) to the decreasing solution of the

TABLE 3. COMPUTED VALUES OF Z AND z

(All function values have been multiplied by 10^7.)

x	2·2	2·4	2·6	2·8	3·0
Z	405 12	751 82	1052 03	1314 61	1546 28
z	1324073	1087802	893271	733247	601704
$x^{\frac{1}{2}}K_0-z$	0	0	0	2	2
x	3·2	3·4	3·6	3·8	4·0
Z	1752 23	1936 55	2102 49	2252 67	2389 31
z	493633	404885	332031	272244	223191
$x^{\frac{1}{2}}K_0-z$	1	1	1	1	3

$$z_i=c(1+Z_i)\,\rho^{i-1}, \quad c=0{\cdot}16107\,03.$$

difference equation (37). The problem does not appear to be acute here, but it exists. When $r(x)$ is small, a satisfactory approximation to the left side of (37) is

$$Z_{i+1} - \rho(2 + h'^2)\, Z + \rho^2 Z_{i-1}$$

and the complementary function of the equations which determine the Z_i can be estimated by equating the above expression to zero; we find

$$Z_i \sim \sigma^i, \quad \text{where} \quad \sigma = 1, \rho^2,$$

and one of the complementary functions increases more rapidly than the solution for Z_i which has been recorded above ($\rho^2 \sim 1{\cdot}5$ when $h = 0{\cdot}2$). The simplest test for control is to treat two of the computed values of z as exact and start a new computation from this point. Suppose, for example, a new computation is started, treating the computed $z(2{\cdot}4)$ and $z(2{\cdot}6)$ as exact. The starting values of Z are

$$Z_1 = 0, \quad 1 + Z_2 = \frac{8932\,71}{10878\,02}\rho = \frac{8932\,71}{8906\,16} = 1{\cdot}00298\,11.$$

A significant figure has been lost in computing Z_2 (which should be $0{\cdot}00298\,00$) and we cannot expect more than six-decimal agreement between the two computations; but we retain seven decimals to test whether the round-off error increases. Denoting the new computed values of the solution by z_i, the discrepancy between the two computations at subsequent tabular points is

x	2·8	3·0	3·2	3·4	3·6	3·8	4·0
$10^7(z_i' - z_i)$	4	5	6	7	9	11	15

The agreement between the two computations is satisfactory; but the discrepancy appears to be increasing, and this increase is probably to be attributed to the complementary solution.

In the present calculation, we were able to ensure that the contribution from fourth differences was negligible for the interval chosen. Without this control on the computation, it appears likely that catastrophic failure of the type familiar in the misuse of asymptotic series may supervene.

If the device of an auxiliary variable is used in conjunction with the factorization method of the previous section, the task of controlling undue increase or decrease of the solution is considerably eased, since the auxiliary variable is not required until the fourth and higher differences are to be incorporated.

II. BOUNDARY-VALUE PROBLEMS OF LINEAR ORDINARY DIFFERENTIAL EQUATIONS

1. Introduction

Part II is complementary to part III: some concepts and techniques are introduced here which will be employed in a wider context in part III.

We shall make considerable use of the matrix operator

$$\mathbf{D}^2 = 2I - S - S^*,$$

the bold symbol being used to distinguish the matrix operator from the differential operator $\mathrm{D} = \mathrm{d}/\mathrm{d}x$. With this single exception, we abandon—here and in later parts—the bold notation which was used in part I.

The matrix $\mathbf{D}^2$ is ideally adapted to the solution of difference equations when the boundary values are prescribed, and we shall employ it here to treat boundary-value problems in a strip of $n+1$ intervals, the points 0 and $n+1$ being boundary points [$h = 1/(n+1)$]. If the terminal values z_0 and z_{n+1} are zero, the action of $\mathbf{D}^2$ on a vector z is to replace each element z_i ($i = 1, \ldots, n$) by $-\delta^2 z_i$, i.e. by its second difference with sign changed. No loss of generality is involved here; in the solution of difference equations, we can always reduce the terminal values z_0 and z_{n+1} to zero by adding suitable terms to the right side of the equation.

For other types of boundary conditions, the matrix $\mathbf{D}^2$ is not directly useful; but it is sometimes possible to take more general conditions into account by modifying the border elements of $\mathbf{D}^2$ (cf. § 2·1 of part III). In the solution of differential equations by difference methods, the replacement of differential boundary conditions by difference expressions introduces an element of approximation which is always difficult to assess; this is particularly true in higher-difference approximations. It is, however, apparent from general considerations that a difference approximation to the solution of a linear differential system is equivalent to a representation of the solution in terms of some set of discrete base functions. From this point of view, it is often preferable to by-pass the difference approximation completely and to represent the solution in terms of continuous base functions, and recent work has emphasized the value of this type of approximation; for example, the utility of Chebyshev polynomials has been amply demonstrated, notably by Lanczos in a number of papers and in his book (1957) and by Clenshaw (1957).

In this part, we illustrate both types of approximation. If the representational aspect is borne in mind, the two types are seen to be more closely related than is sometimes suspected; but they differ considerably in technique, and technique is of no small importance in numerical analysis.

1·1. *Simple difference approximations to differential equations*

The properties of matrix difference operators have been explored in considerable detail by Rutherford (1947, 1952) and Todd (1950). For our present purpose, it is sufficient to record that the eigenvalues and eigenfunctions of the nth-order matrix $2I - S - S^*$ are

$$\lambda_m = 4\sin^2(\tfrac{1}{2}m\pi h), \quad r_{pm} = (2h)^{\frac{1}{2}}\sin(pm\pi h);$$

here, p and m run over the values $1 \ldots n$, and the matrix r_{pm} is its own inverse.

The eigenvalues of $\mathbf{D}^2$ are confined to the range $0 < \lambda_m < 4$, the smallest being

$$4\sin^2(\tfrac{1}{2}\pi h) \sim h^2\pi^2.$$

The matrix $\mathbf{D}^2$ is not itself ill-conditioned unless very small values of h are used. Ill-conditioning may, however, arise in a less obvious form in the solution of difference equations. Consider, for example, the equation

$$-h^2 \mathrm{d}^2 z/\mathrm{d}x^2 + h^2 q z = h^2 f(x), \quad z(0) = 0 = z(1),$$

where q is constant and the boundary values z_0 and z_{n+1} are zero. The simplest discrete approximation to this equation is

$$[\mathbf{D}^2 + h^2 q \mathbf{I}]\, z = h^2 f,$$

the components of z and f being $z(x_i)$ and $f(x_i)$; we use the same symbols for the continuous functions z and f and their discrete representations. The matrix on the left of this equation and the inverse of the matrix can be written in the spectral form

$$\sum_m (\lambda_m + h^2q)\, r_m r_m^* \quad \text{and} \quad \sum_m (\lambda_m + h^2q)^{-1} r_m r_m^*;$$

r_m is here an eigencolumn and r_m^* its row transpose. The matrix is singular if, for any m,

$$-h^2q = 4\sin^2(\tfrac{1}{2}m\pi h).$$

If q is positive, no difficulties arise, the equation being of exponential type. If q is negative, the severe conditions

$$-q < \pi^2 \quad \text{or} \quad -q > 4/h^2$$

exclude the possibility of singularity or ill-conditioning. If, however, $0 < -h^2q < 4$, ill-conditioning need not, in general, be feared for practical values of h, provided that $-h^2q$ is not close to an eigenvalue of $\mathbf{D}^2$. When small values of h are used, the density of the eigenvalues on the strip 0 to 4 increases, and the point h^2q is more closely enclosed.

The above conditions indicate qualitatively the phenomena to be anticipated in the discrete analogue of the more general equation

$$-\mathrm{d}^2z/\mathrm{d}x^2 + q(x)\, z = f(x),$$

but we can obtain more precise guidance from the comparison theorems of the Sturm–Liouville theory; the conditions derived above are in fact analogous to the exclusion theorems of the Sturmian theory (see, for example, Ince (1944)).

1·2. *Higher-difference approximations*

Higher-difference approximations can be obtained by using a power series in $\mathbf{D}^2$; in effect, this amounts to using eigenvalues which are better approximations to the eigenvalues of the differential operator $\mathrm{d}^2/\mathrm{d}x^2$, the eigenfunctions remaining unchanged. Alternatively, we can discard the matrix formulation and employ any convenient basis of discrete or continuous functions.

The powers of $\mathbf{D}^2$ or any similar matrix are closely related to difference operators; they may even yield exact differences of all even orders when the operand is an eigenvector of the operator, and it is sometimes practicable to make use of this property by expanding an arbitrary operand in terms of the eigenvectors. In general, however, the end differences of $\mathbf{D}^{2r}z$ are incomplete when z is an arbitrary vector. For example, the square of $\mathbf{D}^2$ is

$$\mathbf{D}^4 = (2I - S - S^*)^2 = 6I - 4(S + S^*) + S^2 + S^{*2} - I_{11} - I_{nn}.$$

If this matrix operates on a column z, the fourth differences of z_1, z_2, and z_{n-1}, z_n, are incomplete. For example, the missing terms in the fourth differences of z_1 and z_2 are

$$z_1 - 4z_0 + z_{-1} \quad \text{and} \quad z_0.$$

The missing terms (some of which will in general be unknown) can be incorporated in a correction vector and taken up iteratively; we shall ignore this correction here since it involves merely a trivial modification of the procedure.

We now consider the equation

$$-h^2\,\mathrm{d}^2z/\mathrm{d}x^2 + q(x)\, z = f(x)$$

which may be written in either of the finite-difference forms

$$-\left(\delta^2-\frac{\delta^4}{12}+\frac{\delta^6}{90}-\ldots\right)z+q(x)\,z=f(x),$$

$$-\delta^2 z+\left(1+\frac{\delta^2}{12}-\frac{\delta^4}{240}+\ldots\right)q(x)\,z=\left(1+\frac{\delta^2}{12}-\ldots\right)f(x).$$

If the first form is used, the corresponding matrix equation is

$$\left(\mathbf{D}^2+\frac{\mathbf{D}^4}{12}+\frac{\mathbf{D}^6}{90}+\ldots\right)z+Qz=f,$$

Q being a diagonal matrix whose elements are $q(x_i)$ and z, f, are column vectors with elements $z(x_i), f(x_i)$. If z' is the solution of the equation

$$(\mathbf{D}^2+Q)\,z'=f,$$

and if $z=z'+\eta$, we find

$$(\mathbf{D}^2+Q)\,\eta=-\left(\frac{\mathbf{D}^2}{12}+\frac{\mathbf{D}^4}{90}+\ldots\right)(f-Qz')-\left(\frac{\mathbf{D}^4}{12}+\frac{\mathbf{D}^6}{90}+\ldots\right)\eta.$$

Neglecting the term in η on the right side, we find

$$z=[\mathbf{D}^2+Q]^{-1}\left[f-\left(\frac{\mathbf{D}^2}{12}+\frac{\mathbf{D}^4}{90}+\ldots\right)(f-Qz')\right].$$

If the second finite-difference form is used, the corresponding matrix equation is

$$\mathbf{D}^2z+\left(1-\frac{\mathbf{D}^2}{12}-\frac{\mathbf{D}^4}{240}-\ldots\right)Qz=\left(1-\frac{\mathbf{D}^2}{12}-\ldots\right)f.$$

Writing $z=z''+\epsilon$, where

$$\mathbf{D}^2z''+\left(1-\frac{\mathbf{D}^2}{12}\right)Qz''=\left(1-\frac{\mathbf{D}^2}{12}\right)f,$$

we find

$$\left[\mathbf{D}^2\left(1-\frac{Q}{12}\right)+Q\right]\epsilon=-\left(\frac{\mathbf{D}^4}{240}+\ldots\right)(f-Qz'')+\left(\frac{\mathbf{D}^4}{240}+\ldots\right)Q\epsilon.$$

Neglecting the term in ϵ on the right side, we obtain

$$z=\left[\mathbf{D}^2\left(1-\frac{Q}{12}\right)+Q\right]^{-1}\left[\left(1-\frac{\mathbf{D}^2}{12}\right)f-\left(\frac{\mathbf{D}^4}{240}+\ldots\right)(f-Qz'')\right].$$

The second approximation has ostensibly a smaller difference correction, but the essential *a priori* uncertainty in both these forms of difference correction is the neglected terms in η or ϵ; this term can, however, be estimated *a posteriori* and used in a further step.

An alternative technique is to use any convenient and admissible set of base vectors. Suppose, for example, we use a discrete basis. It is clearly possible to express any column vector in terms of this basis and of the reciprocal or adjoint basis, and it is shown below that a term of the type Qz (Q, a diagonal matrix; z, a column vector) can be treated without difficulty. In the solution of differential equations, we also require expressions for a sequence of difference operations on any vector of the basis, for example, the truncated sequence

$$\delta^2-\frac{\delta^4}{12}+\frac{\delta^6}{90}$$

yields an approximation to the second derivative of a base function at tabular points. The use of such a sequence implies some assumption defining the continuation of the base functions outside the segment in which the differential equation is to be solved. The simplest assumption we can make is that the continuation is periodic and it is sometimes possible to ensure this by a change of variable in the original differential equation.

The admissible sets of base functions are restricted by the requirement that the sum of components defining the solution z must satisfy the prescribed boundary conditions.

It is convenient to introduce here some notation which we shall use frequently in the sequel. We denote the basis functions by r_i $(i = 1, \ldots, n)$ and the adjoint basis by ρ_i. The r_i may be supposed to be displayed in columns and the vectors ρ_i in rows; and both sets are normalized so that the scalar $\rho_i r_j$ is unity for $i = j$, i.e.

$$\rho_i r_j = \delta_{ij}.$$

The components of the vector z in the basis r_i are denoted by z_i and the sum

$$\sum_{i=1}^{n} z_i r_i$$

denotes a vector. To avoid confusion with a similar notation in tensor analysis, we indicate sums by a summation sign; we do not use the more convenient convention of summing over the repeated index, since this convention is apt to suggest a scalar. To obtain a similar representation of the vector Qz when Q is diagonal, we notice that we can write

$$Qz = \sum_i z_i Q r_i = \sum_i \beta_i r_i.$$

The coefficients β_j are evaluated by forming the scalar product of each side of the equations with the row ρ_j, i.e.

$$\sum_i z_i \rho_j Q r_i = \beta_j = \sum_i z_i q_{ji}.$$

The matrix whose elements are q_{ij} define the representation of Q in the basis r_i, ρ_j and we can write

$$Q = \sum_{ij} q_{ij} r_i \rho_{ji}.$$

This sum denotes a matrix; each component matrix $r_i \rho_j$ (of order $n \times n$) is multiplied by the appropriate coefficient z_{ij} and the summation embraces n^2 component matrices.

Similar notation is employed in part III.

To illustrate these points, we repeat briefly here an example which is to be considered in more detail in another paper, namely, the solution of the equation

$$-\mathrm{d}^2 z/\mathrm{d}x^2 + q(x)\, z = f(x), \quad z(0) = z(1) = 0,$$

in terms of the eigenvectors r_i of the matrix $\mathbf{D}^2$. Equivalently, we may say that the basis is the set of continuous functions

$$\sin(m\pi x) \quad (1 \leqslant m \leqslant n),$$

where $h = 1/(n+1)$, but the components of a vector in this basis are defined by a summation and not by an integration, i.e. the coefficient z_m is defined by the scalar product

$$\rho_m z.$$

The equation becomes

$$-\sum_m z_m \frac{\mathrm{d}^2 r_m}{\mathrm{d}x^2} + \sum_{ij} q_{ij} r_i \rho_j \sum_m z_m r_m = \sum_m f_m r_m.$$

We may write

$$-\frac{\mathrm{d}^2 r_m}{\mathrm{d}x^2} = -\frac{\mathrm{d}^2}{\mathrm{d}x^2}\sin(m\pi x) = m^2\pi^2 r_m$$

and on pre-multiplying the equation by ρ_m we have

$$m^2\pi^2 z_m + \sum_p q_{mp} z_p = f_m.$$

Example 1·21

As a numerical illustration of the above method, we seek the lowest eigenvalue b_0 of the even periodic solutions of the Mathieu equation

$$\mathrm{d}^2z/\mathrm{d}x^2 + (b - t\cos^2 x)\, z = 0, \quad z(0) = z(2\pi).$$

The analytical solution of this problem is studied in Morse & Feshbach (1953) and it is shown there that the eigenvalues b can be expressed as a power series in t, the series for the even periodic solutions being (p. 1017)

$$b_0 = \tfrac{1}{2}t - \frac{t^2}{\tfrac{1}{8} + \tfrac{1}{2}t - b_0} + O(t^4),$$

$$b_n = n^2 + \tfrac{1}{2}t - \frac{t^2}{16}\left[\frac{1}{(n-2)^2 + \tfrac{1}{2}t - b_n} + \frac{1}{(n+2)^2 + \tfrac{1}{2}t - b_n}\right] + O(t^4) \quad (n \neq 0).$$

The above expressions are obtained by treating x as a continuous variable in the range 0 to 2π. The method of § 1·2 is convenient here and may be expected to give quite accurate results if the even periodic basis

$$r_{pm} = \cos(pmh) \quad (ph = 0, h, 2h, \ldots, 2\pi;\ m = 0, 1, \ldots, \pi/h)$$

is used. The eigenvalues b are given by the determinantal equation

$$|(p^2 - b)\,\delta_{mp} + tq_{mp}| = 0,$$

where q_{mp} are the elements of the diagonal matrix $\cos^2(ph)$ in the basis r_{pm}.

We shall simplify the exposition by choosing large values of h in order to keep the order of the determinantal equation small. Taking first $h = \pi$, we have two base functions

$$\cos(pm\pi) \quad (m = 0, 1;\ p = 0, 1, 2).$$

Since $\cos(m\pi - mx) = \cos(m\pi + mx)$, it is sufficient to consider only the range 0 to π. The basis can be made self-adjoint by adopting the trapezoidal definition of the scalar product of two functions f and g, namely

$$T(f.g) = \tfrac{1}{2}f_0g_0 + f_1g_1 + f_2g_2 + \ldots + \tfrac{1}{2}f_ng_n.$$

With this definition of scalar product, the normalized half-range vectors are (reading by columns)

$$B = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

and the matrix Q in this basis is

$$Q = \begin{bmatrix} 1 & . \\ . & 1 \end{bmatrix},$$

the dots denoting zeros. The determinantal equation is

$$\begin{vmatrix} t-b & . \\ . & 1+t-b \end{vmatrix} = 0, \quad \text{i.e.} \quad b = t,\ 1+t.$$

Using now the interval $h = \frac{1}{2}\pi$, we have three eigenfunctions, and the half-range basis is

$$B = \begin{bmatrix} \frac{1}{2}\sqrt{2} & 1 & \frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & . & -\frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & -1 & \frac{1}{2}\sqrt{2} \end{bmatrix}.$$

The representation of Q is

$$Q = \begin{bmatrix} \frac{1}{2} & . & \frac{1}{2} \\ . & 1 & . \\ \frac{1}{2} & . & \frac{1}{2} \end{bmatrix}$$

and b is determined by

$$(1+t-b)\,[(\tfrac{1}{2}t-b)\,(4+\tfrac{1}{2}t-b)-\tfrac{1}{4}t^2] = 0,$$

the smallest eigenvalue satisfying

$$b_0 = \tfrac{1}{2}t - \frac{1}{4}\frac{t^2}{4+\frac{1}{2}t-b_0}.$$

We shall employ two finer intervals, namely $h = \frac{1}{3}\pi, \frac{1}{4}\pi$, since the above approximations, though of the right order, are fairly crude. It will be sufficient to record only the Q matrices for these intervals.

$$Q_{\frac{1}{3}\pi} = \begin{bmatrix} \frac{1}{2} & . & \frac{1}{4}\sqrt{2} & . \\ . & \frac{3}{4} & . & \frac{1}{4}\sqrt{2} \\ \frac{1}{4}\sqrt{2} & . & \frac{3}{4} & . \\ . & \frac{1}{4}\sqrt{2} & . & \frac{1}{2} \end{bmatrix},\dagger \quad Q_{\frac{1}{4}\pi} = \begin{bmatrix} \frac{1}{2} & . & \frac{1}{4}\sqrt{2} & . & . \\ . & \frac{3}{4} & . & \frac{1}{4} & . \\ \frac{1}{4}\sqrt{2} & . & \frac{1}{2} & . & \frac{1}{4}\sqrt{2} \\ . & \frac{1}{4} & . & \frac{3}{4} & . \\ . & . & \frac{1}{4}\sqrt{2} & . & \frac{1}{2} \end{bmatrix}.$$

We can now write down two further approximations to b_0:

$$h = \tfrac{1}{3}\pi\colon\ b_0 = \tfrac{1}{2}t - \frac{t^2}{8}\left[\frac{1}{4+\frac{3}{4}t-b_0} + \frac{\frac{1}{2}t-b_0}{(1+\frac{3}{4}t-b_0)\,(9+\frac{1}{2}t-b_0)}\right] + \frac{t^4}{64}\frac{1}{(1+\frac{3}{4}t-b_0)\,(4+\frac{3}{4}t-b_0)\,(9+\frac{1}{2}t-b_0)},$$

$$h = \tfrac{1}{4}\pi\colon\ b_0 = \tfrac{1}{2}t - \frac{t^2}{8}\left[\frac{1}{4+\frac{1}{2}t-b_0} + \frac{\frac{1}{2}t-b}{(4+\frac{1}{2}t-b_0)\,(16+\frac{1}{2}t-b_0)} + \frac{\frac{1}{2}(\frac{1}{2}t-b_0)}{(1+\frac{3}{4}t-b_0)\,(9+\frac{3}{4}t-b_0)}\right] + \frac{t^4}{128}\frac{1}{(1+\frac{3}{4}t-b_0)\,(4+\frac{1}{2}t-b_0)\,(9+\frac{3}{4}t-b_0)} + \frac{t^4}{128}\frac{\frac{1}{2}t-b_0}{(1+\frac{3}{4}t-b_0)\,(4+\frac{1}{2}t-b_0)\,(9+\frac{3}{4}t-b_0)\,(16+\frac{1}{2}t-b_0)}.$$

The above polynomial expressions for b_0 are of the same form as the series expressions quoted earlier but there are discrepancies in detail. The following table shows the values of b_0 obtained for the four intervals chosen when $t = 4$:

interval	π	$\frac{1}{2}\pi$	$\frac{1}{3}\pi$	$\frac{1}{4}\pi$
b_0	4	1·17	1·63	1·539

† We may note here that $q_{ij} = 0$ unless i and j are both odd or both even. This simplification occurs also in the analytical solution (cf. Morse & Feshbach).

These approximations oscillate about the true value 1·5448 ... but the amplitudes of the oscillations appear to diminish as the interval decreases.

We should expect less accuracy in computing the higher eigenvalues by this technique. The approximations to the third eigenvalue when $t = 1$ are:

interval	$\frac{1}{2}\pi$	$\frac{1}{3}\pi$	$\frac{1}{4}\pi$
b_2	4·56	4·78	4·52

The true value is about 4·51.

III. CLOSED METHODS FOR THE SOLUTION OF LINEAR PARTIAL-DIFFERENCE EQUATIONS

1. Introduction

The remarkable success of methods of successive approximation in the solution of partial difference equations of elliptic type has perhaps obscured the need for closed solutions. It is *a priori* probable that closed rational solutions of elliptic partial-difference equations should exist, though it is not to be expected that these solutions can be determined by the methods of elementary algebra.

In part III we explore a number of techniques which lead to closed solutions of simple elliptic-difference equations in a bounded plane region; we touch briefly on other types of equation but our main concern is with the elliptic. All of these techniques save one use concepts of matrix algebra and they are applicable only if the region of existence of the governing equations is bounded by curves of an orthogonal net. The remaining technique can be regarded as a finite-difference analogue of Bergman's kernel function method (Bergman & Schiffer 1953) and was in fact suggested by it; it depends on our ability to find a number of independent solutions of the partial-difference equations which may be termed complementary solutions. The complementary solutions are not required to fulfil the prescribed boundary conditions, but the sum of a sufficient number of them can be made to satisfy the boundary conditions. This method is the most flexible of the methods described in part III but the least susceptible of reduction to an invariable routine of operations.

The concepts of which we make use here are drawn partly from analysis and partly from algebra and we have endeavoured throughout to acknowledge our conscious debt to earlier workers. We have been unable to make a complete survey of the relevant literature which is scattered over a wide range of journals, but the references we have consulted suggest that interest in closed solutions has been sporadic. An early paper by Courant, Friedrichs & Lewy (1928), on the partial-difference equations of mathematical physics showed that many of the fundamental theorems of analysis could be directly transformed into finite-difference analogues and thus laid the foundations for a rational attack on the associated difference problems. Further reference may be found in subsequent papers by Stöhr (1950), Hyman (1952) and Stiefel (1952). It seems probable that the hypercircle method developed by Synge (1947) for the solutions of boundary-value problems of differential equations could form the starting point for the development of analogous methods for difference equations but we have not explored this point in any detail.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

We have not been able to trace any work which attacks the matrix equations considered in § 3 below. A superficially similar equation, i.e. the matrix equation $AX = XB$, was considered by Turnbull & Aitken (1945) who give references to earlier work.

For the most part we shall be concerned with finite-difference analogues of the equations of Laplace and Poisson and of the biharmonic equation and we introduce in the next section some matrix operations which permit a concise formulation of the problems to be considered. The properties and some applications of these matrices have already been explored in the literature but mainly in problems of a single variable; see, for example, Rayleigh (1926), Lennard-Jones (1937), Rutherford (1947, 1952), Todd (1950), Bolton & Scoins (1957).

The emphasis in part III is on *closed* methods but we give some illustrations in which a potentially closed method is used iteratively. Moreover, we restrict our considerations mainly to solutions of partial-*difference* equations, even though the motive behind our work is that these solutions are approximations to the solutions of differential equations. It appears desirable to distinguish two techniques here, namely, the technique of finding the solution of the difference equation in a given mesh and the technique of using this solution to obtain a better approximation to the solution of the differential equation at the nodal points. Since the present work was begun, we have found, however, that the two techniques are more closely related than might at first sight be supposed and that it is sometimes practicable to by-pass the intermediate step of solving the difference equation; this point is discussed in § 3 below and more fully in a companion paper.

The simplest type of boundary condition in difference or differential equations (i.e. the prescription of boundary values of the wanted function) is basic in the numerical methods of succeeding sections. We give some illustrations of the incorporation of other types of boundary conditions for difference equations; but we have not pursued this point in great detail, since the practical problem usually originates from a differential equation and the incorporation of adequate approximations to differential boundary conditions in a difference treatment is not readily reducible to a generalized formulation such as we envisage in this part of the paper.

2. Formulation of partial-difference equations as matrix equations

2·1. *Basic types of bivariate equations*

We shall discuss very briefly a compact formulation of the Poisson difference equation and the biharmonic difference equation as matrix equations. Similar considerations apply to the formulation of other types of elliptic difference equation.

When the boundary values of the wanted function are prescribed, the operator $\mathbf{D}^2 = 2I - S - S^*$ of part II may be used. If Z is a rectangular matrix of n rows and m columns, the operation $\mathbf{D}^2 Z$ replaces each element z_{ij} ($i \neq 1$ or n) by its column-wise difference $-\delta_x^2 z_{ij}$; similarly, the operation $Z\mathbf{D}^2$ replaces each z_{ij} ($j \neq 1$ or m) by its row-wise difference $-\delta_y^2 z_{ij}$. We have chosen the axes so that the x variation is indicated by the row suffix i and the y variation by the column suffix j, i.e. x is measured vertically downward, y horizontally to the right.

The sum of the two operations, namely

$$\mathbf{D}_n^2 Z + Z\mathbf{D}_m^2,$$

yield the complete second difference (with sign changed) of all elements z_{ij} which do not lie in the border of the matrix Z, i.e. in the first or last row or column. It may be noted that—as earlier—we choose to symbolize the central-difference operator $-\delta^2$ rather than δ^2 in order that $\mathbf{D}^2$ may have positive eigenvalues; suffixes such as n or m above may be used to indicate the order of the matrix $\mathbf{D}^2$, but we generally omit the suffix since there is little danger of ambiguity on this score.

The missing entries in the incomplete differences of the border elements are the prescribed boundary values which will be entered on the right side of an eventual equation, the left side being reserved for operations on unknown quantities. It follows easily from the above that the central-difference equation

$$-(\delta_x^2+\delta_y^2)\, z_{ij} = 0 \quad (i = 1, \ldots, n; j = 1, \ldots, m) \tag{1}$$

and the prescribed values at the nodal points z_{0j}, $z_{n+1,j}$, z_{i0}, $z_{i,m+1}$ can be formulated in the matrix equation

$$\mathbf{D}^2 Z + Z\mathbf{D}^2 = F, \tag{2}$$

where the border elements of F are the prescribed values and the interior elements are zero. If the inhomogeneous term $f(x_i, y_j)$ is added to the right side of (1), the nodal values of f are entered in the appropriate locations in the matrix F of equation (2).

Somewhat more general forms of (1) are

$$\left.\begin{aligned} -[\delta_x^2 a z_{ij} + \delta_y^2 b z_{ij}] &= f(x_i, y_j) \\ \text{and} \qquad -[a\delta_x^2 + b\delta_y^2]\, z_{ij} &= f(x_i, y_j), \end{aligned}\right\} \tag{3}$$

where $a = a(x_i)$ and $b = b(y_j)$; when the region of definition is a rectangle, the equations (3) can be replaced by the matrix equations

$$\left.\begin{aligned} \mathbf{D}^2 A Z + Z B \mathbf{D}^2 &= F \\ \text{and} \qquad A\mathbf{D}^2 Z + Z\mathbf{D}^2 B &= F, \end{aligned}\right\} \tag{4}$$

the matrices A and B being diagonal. If the coefficients a and b in (3) are functions of both x and y, concise formulation is difficult; we may then write out all the equations in some suitable order (as in the big matrix of § 3·1) or use the more general method of § 6.

The incorporation of other types of boundary condition necessitates modification of the basic operator $\mathbf{D}^2$. For example, if the boundary conditions at z_{0j} and $z_{n+1,j}$ are of the form

$$\left.\begin{aligned} z_{0j} &= a_1 z_{1j} + a_2 z_{2j} + a_0, \\ z_{n+1,j} &= a_n z_{nj} + a_{n-1} z_{n-1,j} + a_{n+1}, \end{aligned}\right\} \tag{5}$$

the fore-operator $\mathbf{D}^2$ must be replaced by

$$2I - S - S^* - a_1 I_{11} - a_2 I_{12} - a_{n-1} I_{n-1,n} - a_n I_{nn} \tag{6}$$

and the inhomogeneous terms a_0 and a_{n+1} are entered on the right of the matrix equation. Clearly, the matrix formulation implies some restriction on the admissible types of boundary condition.

These illustrations bring to light some points of general interest. It is important to observe that a matrix-difference equation represents not merely a difference equation but a difference system consisting of the difference equation and its appropriate boundary

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

conditions; we cannot, indeed, formulate the matrix equation without including the boundary conditions. We may also note that row-wise differencing, i.e. differencing with respect to the horizontal co-ordinate, is effected by aft-operators, and column-wise differencing, i.e. with respect to the vertical co-ordinate, is effected by fore-operators. A minor point is that a uniform difference interval, h, say, can easily be incorporated in equations such as (2).

From the considerations already advanced, it is clear that the matrix originating from the fourth-order central difference $\delta^4 = (E^{\frac{1}{2}} - E^{-\frac{1}{2}})^4$ has the structure

$$6I - 4(S + S^*) + S^2 + S^{*2},$$

but the boundary conditions are now more cumbersome and less easily reducible to a simple general statement. The following conditions on the boundary element z_0 and on the element z_{-1} (just outside the boundary) typify conditions which may arise in practical problems:

$$z_0 = a_1 z_1 + a_2 z_2 + a_3 z_3 + a_0,$$
$$z_{-1} = a'_1 z_1 + a'_2 z_2 + a'_3 z_3 + a_{-1}.$$

With these boundary conditions and with similar conditions at the opposite boundary, the matrix originating from the fourth-order central difference δ^4 is of the form†

$$\mathbf{D}^4 = 6I - 4(S + S^*) + S^2 + S^{*2} + (a'_1 - 4a_1) I_{11} + (a'_2 - 4a_2) I_{12} + (a'_3 - 4a_3) I_{13} + a_1 I_{21} + a_2 I_{22} + a_3 I_{23} + \ldots, \quad (7)$$

the ... indicating similar terms in the last two rows of the matrix. The matrix (7) is used when the operand is a column and the transpose $\mathbf{D}^{*4}$ is appropriate when the operand is a row. This representation enables us to transform the difference equation

$$(\delta_x^4 + 2\delta_{xy}^4 + \delta_y^4) z_{ij} = f(x_i, y_j), \quad (8)$$

together with its boundary conditions, into the matrix equation

$$\mathbf{D}^4 Z + 2\mathbf{D}^2 Z \mathbf{D}^2 + Z\mathbf{D}^4 = F \quad (9)$$

(again we suppose a rectangular domain). The matrix $\mathbf{D}^4$ may be of the general form (7), but the fore and aft matrices need not be identical, and the matrices $\mathbf{D}^2$ in (9) may be somewhat more complicated than those encountered earlier.

The obvious method of attacking (9) is to attempt to factorize it, i.e. to write

$$W = A_1 Z + Z B_1$$

and seek to determine matrices A_1, A_2, B_1, B_2, such that

$$A_2 W + W B_2 = A_2 A_1 Z + A_2 Z B_1 + A_1 Z B_2 + Z B_1 B_2$$

is identical with the left-hand side of (9). Two simple factorizations are suggested by the natural boundary conditions of analysis. In the first, we take

$$A_1 = A_2 = 2I - S - S^*;$$

† It is convenient to employ invariable symbols $\mathbf{D}^2$ and $\mathbf{D}^4$ for the second- and fourth-difference operators, even though the precise meaning to be attributed to these symbols will always depend on the boundary conditions. The notation does not imply that $\mathbf{D}^4$ is always the square of $\mathbf{D}^2$.

and we take B_1 and B_2 to be of the same form (but not in general of the same order); it is easy to verify that this factorization is possible when z and $\delta^2 z$ are specified on a rectangular boundary. The matrix A_1^2 has the form

$$A_1^2 = 6I - 4(S+S^*) + S^2 + S^{*2} - I_{11} - I_{nn}.$$

In the second factorization, we take A_1 and A_2 to be of the form given above, but

$$B_1 = 2I - S - S^* - I_{21} - I_{n-1,n} = B_2;$$

z and $\delta^2 z$ must then be specified on the horizontal edges, and $\mu\delta z$ and $\mu\delta^3 z$ on the vertical edges. Obviously, we can achieve other simple factorizations by combining these two cases, but the number of simple factorizations appears to be very limited.

2·2. *Properties of matrix-difference operators*

The preceding section served to elucidate the importance of the boundary conditions in the formulation of matrix-difference equations. This suggests that we may, as in analysis, seek to obtain the solution of an inhomogeneous equation as a sum of components which are individually solutions of the allied homogeneous equation. For example, in solving the Poisson-type equation [(3) or (4) above], we may employ vectors z whose components z_p satisfy the difference equation

$$\delta^2 z_p = -\lambda z_p, \tag{10}$$

together with homogeneous boundary conditions of the type (5). The Poisson equation involves two operators and we may expect that its solution will depend on the homogeneous solutions for both operators. The use of homogeneous components or eigenfunctions in practical problems depends mainly on the ease with which they can be obtained; it will be seen that the operators of § 2·1 possess eigenfunctions which can be expressed in simple analytical form. The properties of these and similar operators have been studied by Rutherford (1947, 1952) and by Todd (1950); we record some of their results here and add some minor results which are pertinent to our present purpose.

If we write

$$\lambda = 2(1-\cos\theta)$$

in (10), we find

$$z_p = A\,\mathrm{e}^{\mathrm{i}p\theta} + B\,\mathrm{e}^{-\mathrm{i}p\theta}.$$

On using the boundary conditions (5), θ is found to be determined by the equation

$$\sin(n+1)\theta - (a_1+a_n)\sin n\theta + (a_1 a_n - a_2 - a_{n-1})\sin(n-1)\theta + (a_1 a_{n-1} + a_2 a_n)\sin(n-2)\theta + a_2 a_{n-1}\sin(n-3)\theta = 0; \tag{11}$$

from this, a number of special results can be deduced. For example, if the constants a_i are all zero,

$$\left.\begin{aligned} \lambda_r &= 2\left[1-\cos\left(\frac{r\pi}{n+1}\right)\right] \quad (r=1,\ldots,n),\\ z_p &= \frac{2}{n+1}\sin\left(\frac{pr\pi}{n+1}\right) \quad (p=1,\ldots,n). \end{aligned}\right\} \tag{11'}$$

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

These eigenvalues and eigenvectors are appropriate when the matrix $\mathbf{D}^2$ is of the form $2I-S-S^*$; the inverse of $\mathbf{D}^2$ has then a particularly simple structure:

$$(n+1)\,[(\mathbf{D}^2)^{-1}]_{kj} = \left.\begin{matrix} k(n+1-j) & (k \leqslant j) \\ j(n+1-k) & (k \geqslant j) \end{matrix}\right\}; \tag{12}$$

i.e. the upper triangle of the nth order symmetric matrix $(2I-S-S^*)^{-1}$ is obtained by writing the natural numbers $n, n-1, \ldots, 1$ in the top row and doubling, trebling, quadrupling ... this row downward. This result is due to Todd.

It may be useful at this point to give two elementary examples of the operator $\mathbf{D}^2$ in numerical work.

(i) It is obvious from equation (12) that the matrix $\mathbf{D}^2$ will play a fundamental role in a numerical solution of the classical integral equation

$$z(x) = f(x) + \int_0^1 K(x,y)\, z(y)\, \mathrm{d}y,$$

where

$$K(x,y) = \begin{cases} x(1-y) & (x \leqslant y), \\ y(1-x) & (x \geqslant y). \end{cases}$$

(ii) Consider the numerical solution of the differential equation

$$\frac{\partial^2 z}{\partial x^2} = \frac{\partial z}{\partial t}, \tag{13}$$

where boundary conditions are prescribed at the two ends of the range in x and the initial state is zero.

A solution might proceed thus. We replace the differential operator $\partial^2/\partial x^2$ by the finite difference approximation δ_x^2/h^2 (the tabular interval being h), and we represent the values of z at the tabular points of the x-range by a vector z; simultaneously we eliminate the t variable by a Laplace transform. For simplicity we may suppose that the boundary conditions are not time-dependent and that they are represented by a vector f; equation (13) becomes

$$(\mathbf{D}^2 + p)\,\bar{z} = f/p. \tag{14}$$

To solve (14), we make use of the spectral resolution of a matrix: if the eigenrow and eigencolumn vectors of $\mathbf{D}^2$ are ρ_i, r_j $(\rho_i r_j = \delta_{ij})$, and if the latent roots are λ_i, then

$$\mathbf{D}^2 = \sum_i \lambda_i r_i \rho_i,$$

$$(\mathbf{D}^2)^{-1} = \sum_i \frac{1}{\lambda_i} r_i \rho_i.$$

The solution of (14) is then

$$\bar{z} = \sum_i \frac{1}{p(\lambda_i + p)} r_i \rho_i f;$$

or, inverting,

$$z = \left[(\mathbf{D}^2)^{-1} - \sum_i \mathrm{e}^{-\lambda_i t} \frac{r_i \rho_i}{\lambda_i}\right] f. \tag{15}$$

This solution is clearly the discrete analogue of the familiar solution

$$z = z_s - \sum_0^\infty a_n \mathrm{e}^{-\beta_n^2 t} \cos \beta_n x,$$

z_s denoting the steady state. Similar treatment of the equation

$$\frac{\partial^2 z}{\partial x^2} = \frac{\partial^2 z}{\partial t^2} \quad (0 \leqslant x \leqslant x_0)$$

yields a solution which is analogous to the double Fourier series solution for the vibration of strings. The fourth-order vibration equation

$$\left(a\frac{\partial^4}{\partial x^4} - b\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial t^2}\right) z = 0 \quad (a, b > 0)$$

can be treated by the same technique.

3. Methods of solving matrix-difference equations

In the preceding section, we formulated a number of difference equations together with their boundary conditions as matrix equations. In this section, we present some methods of solving these equations, in particular equations (2) and (4), which are special cases of the more general equation

$$AZ + ZB = F. \tag{16}$$

The matrices A, B and F in (16) are known matrices of orders $n \times n$, $m \times m$ and $n \times m$; Z is an unknown matrix of order $n \times m$. The matrices A and B are not of course completely arbitrary, but the restrictions which we shall presently be obliged to impose concern only the existence of the solution, and the methods explained in §§ 3·2 to 3·4 do not depend on any special properties or simplicity of structure which the matrices A or B may possess.

It is of course clear that the matrix formulation presented in § 2 is not unique nor even the most obvious method of treating problems of this type. Before considering equation (16), we shall outline a method of attack which has been exploited in the literature (see, for example, Karlqvist 1952; Burgerhout 1954; Cornock 1954).

3·1. *The big matrix*

The big matrix—we use the term for want of a better description—is obtained by writing down in consecutive order all the difference equations which are to be solved. The values of the unknown function at the nodal points are united in a column vector which is operated on by the big matrix of difference coefficients and equated to a known column vector. There is no predetermined method of ordering the successive equations, but it is natural in dealing with a rectangular domain to order by rows or columns. For example, suppose we wish to solve Laplace's difference equation for a mesh of 12 interior points arranged in three rows and four columns, the values of the function being specified on the rectangular boundary of the domain. If the equations for the points z_{i1} $(i = 1, 2, 3)$ are written in order, followed by the equations for z_{i2}, z_{i3}, z_{i4}, we obtain the operator matrix $\mathscr{B}$ of order 12:

$$\mathscr{B} = \begin{bmatrix} P & -I & 0 & 0 \\ -I & P & -I & 0 \\ 0 & -I & P & -I \\ 0 & 0 & -I & P \end{bmatrix}, \quad P = \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix},$$

I being the unit matrix of order three. This continuant structure persists for any rectangular array of points, provided that the unknown function is specified at the boundary.

To invert a matrix of the type $\mathscr{B}$, we notice that any polynomial in a matrix P commutes with any other polynomial or with the reciprocal of any other polynomial in P. We may, then, invert $\mathscr{B}$, treating P as a real number in arithmetical operations. If P is of order n, and if $\mathscr{B}$ is partitioned into m^2 submatrices each of order n, the inverse $\mathscr{B}^{-1}$ is given partitioned form by

$$B_m(\mathscr{B}^{-1})_{kj} = \begin{cases} B_{m-k}B_{j-1} & (k \geqslant j) \\ B_{m-j}B_{k-1} & (k \leqslant j), \end{cases}$$

where $$B_k = \sinh(k+1)\,\theta/\sinh\theta$$

and $$\cosh\theta = \tfrac{1}{2}P.$$

An equivalent result was given by Burgerhout (1954).

The structure of the matrix $\mathscr{B}$ is invariably of the form given above, but the submatrices will be altered if the boundary conditions are changed. For example, suppose in the 12-point example that the function is prescribed on the north, east, and west boundaries and that the reflexion condition holds at the points z_{3i} ($i = 1$ to 4). The matrix has the same shape, but the submatrix P is replaced by

$$P_1 = \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -2 & 4 \end{bmatrix}.$$

Again, if the reflexion condition holds east and south, the function being prescribed north and west, the submatrix P is replaced by P_1, and $-I$ in the bottom row of $\mathscr{B}$ is replaced by $-2I$. $\mathscr{B}$ is now asymmetric but its inverse can be determined by the method given above.

3·2. *The irrational solution of* $AZ+ZB = F$

The analysis given in Turnbull & Aitken (1945) for the solution of the homogeneous equation $AZ = ZB$ can readily be adapted to yield a solution of the inhomogeneous equation (16). The method requires that the eigenvalues and eigenvectors of both A and B be available; we suppose that A and B are non-defective, i.e. that A possesses n and B possesses m independent eigenvectors. Writing

$$A = R\Lambda R^{-1}, \quad B = SNS^{-1},$$

where Λ and N are diagonal matrices, we premultiply (16) by R^{-1} and post-multiply by S; then

$$\Lambda W + WN = F_1, \tag{17}$$

where $$W = R^{-1}ZS, \quad F_1 = R^{-1}FS,$$

or $$(\lambda_i+\nu_j)\,w_{ij} = (F_1)_{ij}. \tag{18}$$

It is clear from (18) that a necessary condition for the existence of a unique solution is

$$\lambda_i+\nu_j \neq 0 \quad (i = 1, \ldots, n; j = 1, \ldots, m). \tag{19}$$

Equation (18) can also be used to indicate whether ill-conditioned equations may be expected when using the harmonic operator. The ratio of the largest and smallest eigenvalues is a useful measure of ill-conditioning and is convenient to use here. When boundary values are prescribed, it is clear from (11′) that $\mathbf{D}^2$ is a positive-definite operator, and (18)

shows that the harmonic operator is then no more ill-conditioned than the operator $\mathbf{D}^2$. In practical applications, operators of the type $\mathbf{D}^2$ are in general non-negative, but either λ_i or ν_j in (18) may be zero. For example, if the reflexion condition is prescribed on two opposite boundaries, it is seen from (11) on putting $a_1 = 0 = a_n$, $a_2 = 1 = a_{n-1}$ that one eigenvalue of the corresponding operator is zero. The harmonic operator is then more ill-conditioned than if the boundary values were prescribed on all four edges. Similar considerations apply to fourth-order difference operators.

The solution (17) or (18) is not usually the most apt method of numerical computation, but it possesses the useful property of exhibiting the influence of each boundary point or each nodal load on every internal point of the domain. For example, suppose that f_{pq} is the only non-zero element of the matrix F and that the elements of R, R^{-1}, S, S^{-1}, are

$$r_{ij} \quad \rho_{ij} \quad s_{ij} \quad \sigma_{ij},$$

then,

$$z_{ij} = f_{pq} \sum_{\alpha,\beta} r_{i\alpha} \frac{\rho_{\alpha p} s_{q\beta}}{\lambda_\alpha + \nu_\beta} \sigma_{\beta j}. \tag{20}$$

An immediate application of (20) is in the solution of Laplace's or Poisson's difference equation for a region composed of a number of rectangles. For example, suppose that the region is T-shaped. To simplify the explanation, we may suppose that the two rectangles composing the T have only one internal nodal point in common. We choose some convenient numerical value c for the function at this point and solve for the two rectangles separately. If we add to c a constant c', the influence of c' on all the internal points of the T-domain is determined by (20) and c' itself is determined by the satisfaction of the difference equation at the common nodal point. The reasoning here is perfectly general and it is obvious that p connecting points will give rise to equations for the constants $c'_1, c'_2, \ldots, c'_p$.

Two other methods for solving (16) are given in §§3·3 and 3·4; numerical illustrations of the three methods are given at the end of §3·4.

3·3. *Semi-rational solution of* $AZ+ZB = F$

In the preceding solution, the eigenvalues and eigenvectors of the two matrices A *and* B were required. An alternative solution can be expressed in terms of the eigenvalues and eigenvectors of A *or* B.

Suppose we choose B. Denoting the eigenrow vectors of B by σ_i, i.e.

$$\sigma_i B = \nu_i \sigma_i,$$

we seek a solution of the form

$$Z = \sum_{i=1}^{m} z_i \sigma_i,$$

where z_i denotes a column vector of order n and $z_i \sigma_i$ is a rectangular $n \times m$ matrix; since there are m columns z, nm values of the coefficients in z are at our disposal and it is thus always possible to express any $n \times m$ matrix in this form. Equation (16) becomes

$$A \sum_i z_i \sigma_i + \sum_i \nu_i z_i \sigma_i = F. \tag{21}$$

Denoting the eigencolumns of B by s_j, and choosing the norm so that

$$\sigma_i s_j = \delta_{ij},$$

we have, on multiplying (21) by s_j,

$$(A+\nu_j)\, z_j = Fs_j. \tag{22}$$

When the matrix B is of order m, the complete solution is obtained by solving m sets of equations of the type (22). Again, we encounter the condition $\lambda_i+\nu_j \neq 0$.

This method of solving the matrix equation is clearly analogous to Fourier series resolution. We may obviously extend the technique to yield an analogue of double Fourier series resolution, i.e. we write

$$Z = \sum_{i,j} z_{ij} r_i \sigma_j, \tag{23}$$

where the z_{ij} are now scalars and the r_i are eigencolumns of A. Equation (16) becomes

$$\sum_{i,j} (\lambda_i+\nu_j)\, z_{ij} r_i \sigma_j = F.$$

The z_{ij} can be determined by reducing each side of the above equation to a scalar product. Pre-multiplying by ρ_p and post-multiplying by s_q, we find

$$(\lambda_p+\nu_q)\, z_{pq} = \rho_p F s_q. \tag{24}$$

The double resolution is of course equivalent to the irrational method of § 3·2; it will be seen in § 6 that a resolution of the form (23) can be approached from a different standpoint and that it can be used in a wider range of applications.

3·4. *A rational solution of* $AZ+ZB = F$

In dealing with the matrix equations $AX = XA$ and $AX = XB$, Turnbull & Aitken (1945) remark 'It is clear that the solutions to problems I and II [i.e. the equations $AX = XA$, $AX = XB$] are not quite final in that the problems must, by their nature, possess *rational* solutions, whereas the classical forms which have been utilized are irrational. We leave to the reader the consideration of the rational solution'.† This remark stimulated the search for the method given below.

We can introduce most simply the concepts on which the method is based by considering the special case in which B is of order two; the illustration will also enable us to trace the connexion between the matrix equation and the vector equation of § 3·1 in which the big matrix is employed.

If the matrix Z is partitioned into two columns z_1 and z_2, and if F is partitioned into f_1 and f_2, equation (16) can—in this special case—be written as

$$\begin{bmatrix} A+b_{11}I & b_{21}I \\ b_{12}I & A+b_{22}I \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}.$$

If we write the characteristic equation of B as

$$B^2 - p_1 B + p_2 = 0$$

and introduce the notation

$$C = A^2 + p_1 A + p_2,$$

the solution is

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} A+b_{22}I & -b_{21}I \\ -b_{12}I & A+b_{11}I \end{bmatrix} C^{-1} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix},$$

or

$$Z = AC^{-1}F - C^{-1}F(B-p_1 I).$$

† A rational solution of $AX = XB$ was given by Rutherford (1932), Weitzenböck (1932) and also later by Foulkes (1949).

To generalize this result when B is of order m, we write the characteristic equation of B as

$$B_{(m)} = B^m - p_1 B^{m-1} + p_2 B^{m-2} - \ldots + (-)^m p_m = 0, \tag{25}$$

where p_m is an abbreviation for $p_m I$, I being a unit matrix of order m; similar abbreviations are used in the succeeding formulae. We also write

$$\left.\begin{aligned} F_1 &= C^{-1}F, \\ \text{where} \quad C &= A^m + p_1 A^{m-1} + p_2 A^{m-2} + \ldots + p_m. \end{aligned}\right\} \tag{26}$$

We now seek a solution of the form

$$\left.\begin{aligned} Z &= Z_1 - Z_2 + Z_3 - \ldots + (-)^{m+1} Z_m \\ \text{with} \quad Z_1 &= A^{m-1}F_1, \end{aligned}\right\} \tag{27}$$

the leading term being suggested by the illustrative example above. We easily verify that

$$AZ_1 + Z_1 B = A^m F_1 + A^{m-1}F_1 B = F + A^{m-1}F_1(B - p_1 I) - \sum_{i=2}^{m} p_i A^{m-i}F_1.$$

The form of the remainder terms on the right suggests that we define Z_2 by

$$Z_2 = A^{m-2}F_1 B_{(1)},$$

where the partial polynomials $B_{(r)}$ are defined by

$$B_{(r)} = B^r - p_1 B^{r-1} + \ldots + (-)^r p_r. \tag{28}$$

It follows that

$$A(Z_1 - Z_2) + (Z_1 - Z_2) B = F - A^{m-2}F_1 B_{(2)} - \sum_{i=3}^{m} p_i A^{m-i}F_1.$$

The next term of the sequence is

$$Z_3 = A^{m-3}F_1 B_{(2)}$$

and

$$A(Z_1 - Z_2 + Z_3) + (Z_1 - Z_2 + Z_3) B = F + A^{m-3}F_1 B_{(3)} - \sum_{i=4}^{m} p_i A^{m-i}F_1.$$

Proceeding in this way, we can prove inductively that

$$Z = C^{-1} \sum_{r=1}^{m} (-)^{r-1} A^{m-r} F B_{(r-1)}. \tag{29}$$

We have termed this solution the rational solution since it depends only on a knowledge of the coefficients of the characteristic equation of B and these coefficients can be determined by purely rational calculations. Clearly, there is a similar solution which depends on the characteristic equation of A, and in practice one will choose the matrix of lower order. A computer would probably remark at this point that the present distinction between rational and irrational solutions is somewhat artificial, since eigenvalues and eigenvectors are in practice determined by purely rational operations and the determination of the coefficients of the characteristic equation is not in general a trivial task. It is, however, important to observe that the irrational solution of § 3·2 requires that *A and B* should be non-defective but the semi-rational solution of § 3·3 requires that *A or B* should be non-defective. The solution (29) is free from these restrictions and is thus more general than the two preceding solutions.

We conclude this section by sketching an alternative route to the solution. This time we proceed by systematically deflating the matrix B. If ν_1 is any constant, we can write (16) in the form

$$(A+\nu_1)\,Z = F+Z(\nu_1-B). \tag{30}$$

Again, if ν_2 is any constant, we have, after a little algebra,

$$(A+\nu_2)\,(A+\nu_1)\,Z = \nu_2 F+\nu_2 Z(\nu_1-B)+AF+\nu_1 AZ-AZB;$$

the terms $\nu_1 AZ$ and $-AZB$ can be transformed by using (30) and we find

$$(A+\nu_2)\,(A+\nu_1)\,Z = (\nu_1+\nu_2)\,F+(AF-FB)+Z(\nu_2-B)\,(\nu_1-B).$$

Similarly

$$(A+\nu_3)\,(A+\nu_2)\,(A+\nu_1)\,Z = (\nu_1\nu_2+\nu_2\nu_3+\nu_3\nu_1)\,F+(\nu_1+\nu_2+\nu_3)\,(AF-FB) \\ +(A^2F-AFB+FB^2)+Z(\nu_3-B)\,(\nu_2-B)\,(\nu_1-B).$$

These results can be continued to as many factors as desired and it follows from the Cayley–Hamilton theorem that the term in Z on the right-hand side is eventually annihilated when the constants ν_i are the eigenvalues of B. The solution can be written in the alternative form

$$Z = C^{-1}\sum_{r=0}^{m-1} p_r F_{(m-r-1)}, \tag{31}$$

where

$$F_{(r)} = A^rF-A^{r-1}FB+A^{r-2}FB^2-\ldots+(-)^r FB^r,$$

$$F_{(0)} = F.$$

Example 3·31

$$\begin{matrix} 0 & 0 & 0 & 0 \\ 0 & z_{11} & z_{12} & 4 \\ 0 & z_{21} & z_{22} & 6 \\ 0 & z_{31} & z_{32} & 4 \\ 0 & 0 & 0 & 0 \end{matrix}$$

The methods are illustrated with reference to the solution of the equation

$$(\delta_x^2+\delta_y^2)\,z = 0$$

for the region and boundary values shown on the diagram above. The example has been chosen so that the arithmetic can be carried out mentally.

If the matrix of the elements z_{ij} is partitioned with two columns z_1 and z_2, the solution can be determined by the vector equation

$$\begin{bmatrix} P & -I \\ -I & P \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix},$$

where

$$P = \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}, \quad f_2 = \begin{bmatrix} 4 \\ 6 \\ 4 \end{bmatrix}$$

and the elements of f_1 are all zeros. The concise formulation is

$$\mathbf{D}^2Z+Z\mathbf{D}^2 = F, \quad F = \begin{bmatrix} 0 & 4 \\ 0 & 6 \\ 0 & 4 \end{bmatrix}$$

and $\mathbf{D}^2 = 2I - S - S^*$, the fore-operator being of order three and the aft-operator of order two.

From the results of § 3·1, the inverse of the big matrix is

$$(P^2-I)^{-1}\begin{bmatrix} P & I \\ I & P \end{bmatrix}, \quad (P^2-I)^{-1} = \frac{1}{2415}\begin{bmatrix} 208 & 120 & 47 \\ 120 & 255 & 120 \\ 47 & 120 & 208 \end{bmatrix},$$

$$(P^2-I)^{-1}P = \frac{1}{2415}\begin{bmatrix} 712 & 225 & 68 \\ 225 & 780 & 225 \\ 68 & 225 & 712 \end{bmatrix}.$$

For the remaining three methods, we require the eigenvalues and eigenvectors of $\mathbf{D}^2$;

$$\lambda_p = 2(1-\cos \tfrac{1}{4}p\pi) \quad r_{pq} = \sqrt{\tfrac{1}{2}} \sin(\tfrac{1}{4}pq\pi) \quad (p, q = 1, 2, 3),$$
$$\nu_p = 2(1-\cos \tfrac{1}{3}p\pi) \quad s_{pq} = \sqrt{\tfrac{2}{3}} \sin(\tfrac{1}{3}pq\pi) \quad (p, q = 1, 2).$$

The irrational solution

$$Z = \Sigma z_{pq} r_p \sigma_q,$$

$$z_{pq} = \frac{\rho_p F s_q}{\lambda_p + \nu_q}.$$

The scalars z_{pq} (i.e. the elements of Z in the double basis $(r, \rho; s, \sigma)$ are recorded below in matrix form:

$$\begin{bmatrix} \dfrac{3+2\sqrt{2}}{3-\sqrt{2}} & -\dfrac{3+2\sqrt{2}}{5-\sqrt{2}} \\ 0 & 0 \\ -\dfrac{3-2\sqrt{2}}{3+\sqrt{2}} & \dfrac{3-2\sqrt{2}}{5+\sqrt{2}} \end{bmatrix}$$

and the relevant base matrices $r_p \sigma_q$ are

$$r_1\sigma_1 = \tfrac{1}{4}\begin{bmatrix} \sqrt{2} & \sqrt{2} \\ 2 & 2 \\ \sqrt{2} & \sqrt{2} \end{bmatrix}, \quad r_1\sigma_2 = \tfrac{1}{4}\begin{bmatrix} \sqrt{2} & -\sqrt{2} \\ 2 & -2 \\ \sqrt{2} & -\sqrt{2} \end{bmatrix},$$

$$r_3\sigma_1 = \tfrac{1}{4}\begin{bmatrix} \sqrt{2} & \sqrt{2} \\ -2 & -2 \\ \sqrt{2} & \sqrt{2} \end{bmatrix}, \quad r_3\sigma_2 = \tfrac{1}{4}\begin{bmatrix} \sqrt{2} & -\sqrt{2} \\ -2 & 2 \\ \sqrt{2} & -\sqrt{2} \end{bmatrix}.$$

The semi-rational solution

$$Z = \sum_{p=1}^{2} z_p \sigma_p$$

and the equations to determine the columns z_p are

$$(\mathbf{D}^2 + \nu_p I)\, z_p = f_p, \quad \nu_1 = 1, \quad \nu_2 = 3,$$

$$f_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 4 \\ 6 \\ 4 \end{bmatrix}, \quad f_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 4 \\ 6 \\ 4 \end{bmatrix}.$$

The rational solution

$$[(\mathbf{D}^2)^2+4\mathbf{D}^2+3I]\,Z=\mathbf{D}^2F-F\mathbf{D}^2+4F,$$

$$(\mathbf{D}^2)^2+4\mathbf{D}^2+3I=\begin{bmatrix}16&-8&1\\-8&17&-8\\1&-8&16\end{bmatrix},$$

$$\mathbf{D}^2F-F\mathbf{D}^2+4F=\begin{bmatrix}4&10\\6&16\\4&10\end{bmatrix}.$$

The solution

By any of these methods, the solution is

$$Z=\frac{1}{161}\begin{bmatrix}116&298\\166&432\\116&298\end{bmatrix}.$$

Example 3·32

The finite difference approximation to the solution of the equation

$$h^2\left(-\frac{\partial^2}{\partial r^2}+\frac{3\partial}{r\partial r}-\frac{\partial^2}{\partial y^2}\right)z=256h^2 \tag{32}$$

affords a more practical illustration of the methods developed in § 3. The function z is to be determined within the region $1<r<2$, $-\frac{1}{2}<y<\frac{1}{2}$, given that z vanishes on the boundary of the domain. A solution of this problem by relaxation methods was given earlier by Fox (1947).

We replace (32) by a second-difference approximation with a tabular interval. The operators $-h^2\,\partial^2/\partial r^2$ and $-h^2\,\partial^2/\partial y^2$ are then replaced by the matrices $\mathbf{D}^2$. The operator $h\,\partial/\partial r$ can be approximated by the difference operator $\frac{1}{2}(E-E^{-1})$, i.e. by the matrix $\frac{1}{2}(S-S^*)$. Taking $h=\frac{1}{8}$ the matrix equivalent of the term $3h^2\,\partial/r\,\partial r$ is

$$-\tfrac{3}{2}\begin{bmatrix}0&\frac{1}{15}&0&0&0&0&0\\-\frac{1}{14}&0&\frac{1}{14}&0&0&0&0\\0&-\frac{1}{13}&0&\frac{1}{13}&0&0&0\\.&.&.&.&.&.&.\\.&.&.&.&.&.&.\\0&0&0&0&-\frac{1}{10}&0&\frac{1}{10}\\0&0&0&0&0&-\frac{1}{9}&.\end{bmatrix}$$

and the top row of the matrix corresponds to $r=1\frac{7}{8}$, the bottom row to $r=1\frac{1}{8}$. Denoting the above matrix by C, and the values of z at the tabular points by z_{ij}, the determining equation is

$$\mathbf{D}^2Z-CZ+Z\mathbf{D}^2=F, \tag{33}$$

where $f_{ij}=4$. The matrices $\mathbf{D}^2$ are of order 7.

We shall illustrate here by using the irrational method. If the first two terms in (33) are added together, the equation is of the form (16) but the eigenvalues and eigenvectors of

$\mathbf{D}^2 - C$ are not readily accessible. We may as an alternative use the irrational method in an iterative routine. Since

$$\mathbf{D}^2 = R\Lambda R^*$$

we can write (33) in the form

$$\Lambda W + W\Lambda - \alpha W = R^* F R = F_1,$$

$$W = R^* Z R, \quad \alpha = R^* C R.$$

If we take $$W = W_0 + W_1,$$

where $$\Lambda W_0 + W_0 \Lambda = F_1 \tag{34}$$

and $$\Lambda W_1 + W_1 \Lambda = \alpha(W_0 + W_1), \tag{35}$$

we can find W_1 from (35) by iterating. Correspondingly, we have

$$Z = Z_0 + Z_1, \quad Z_1 = Z_{11} + Z_{12} + Z_{13} + \ldots.$$

We record in table 4 the terms Z_0, $Z_0 + Z_{11}$, $Z_0 + \sum_1^4 Z_{1j}$, rounded off to five figures; the solution is symmetric in the co-ordinate y and only half the solution need be recorded.

TABLE 4

	$y = -\frac{3}{8}$	$y = -\frac{1}{4}$	$y = -\frac{1}{8}$	$y = 0$
		Z_0		
$r = 1\frac{7}{8}$	4·551	7·103	8·426	8·838
	7·103	11·434	13·765	14·500
	8·426	13·765	16·699	17·632
	8·838	14·500	17·632	18·632
	8·426	13·765	16·699	17·632
	7·103	11·434	13·765	14·500
$r = 1\frac{1}{8}$	4·551	7·103	8·426	8·838
		$Z_0 + Z_{11}$		
$r = 1\frac{7}{8}$	5·118	8·094	9·671	10·166
	7·668	12·442	15·047	15·873
	8·695	14·247	17·314	18·292
	8·666	14·184	17·223	18·190
	7·785	12·601	15·203	16·023
	6·110	9·652	11·491	12·060
$r = 1\frac{1}{8}$	3·583	5·404	6·289	6·555
		$Z_0 + \sum_1^4 Z_{1j}$		
$r = 1\frac{7}{8}$	5·068	8·002	9·550	10·036
	7·520	12·172	14·695	15·494
	8·476	13·845	16·791	17·728
	8·434	13·759	16·671	17·595
	7·610	12·281	14·789	15·577
	6·048	9·541	11·349	11·908
$r = 1\frac{1}{8}$	3·635	5·499	6·411	6·687

It can be seen that the sum $Z_0 + Z_{11}$ gives a fair approximation to the solution. Taking the absolute-value norms of $Z_0, Z_{11}, \ldots$ we obtain a rough measure of the convergence of the iteration:

$\frac{N(Z_{11})}{N(Z_0)}$	$\frac{N(Z_{12})}{N(Z_{11})}$	$\frac{N(Z_{13})}{N(Z_{12})}$	$\frac{N(Z_{14})}{N(Z_{13})}$	$\frac{N(Z_{15})}{N(Z_{14})}$
0·096	0·248	0·180	0·232	0·200

$$N(Z) = \sum_{ij} |z_{ij}|.$$

The next term of the series is (rounded off)

		Z_{15}	
0·001	0·001	0·002	0·002
0·001	0·002	0·002	0·001
0·000	0·001	0·001	0·001
−0·001	−0·002	−0·002	−0·002
−0·002	−0·004	−0·005	−0·005
−0·002	−0·004	−0·006	−0·006
−0·001	−0·002	−0·003	−0·003

Since successive terms of the series are alternatively positive and negative, we can accept $Z_0+\sum_1^4 Z_{1j}$ as a reliable solution of the difference equation.

Example 3·33

The semi-rational method provides a somewhat quicker route to the solution of example 3·32. We write (33) in the form

$$(\mathbf{D}^2-C)\,Z+Z\mathbf{D}^2=F \tag{36}$$

and seek a solution based on the eigenrows of the aft-operator $\mathbf{D}^2$. Since F is symmetric about $y=0$, the solution is of the form

$$\sum_1^4 z_{2p-1}\sigma_{2p-1},$$

where $$\nu_p=2(1-\cos\tfrac{1}{8}p\pi),\quad \sigma_p=\tfrac{1}{2}\sin\tfrac{1}{8}pq\pi.$$

The remaining three eigenrows of $\mathbf{D}^2$ are anti-symmetric about $y=0$ and make no contribution to the solution; hence, the solution is obtained by solving four sets of equations of order seven. The matrix of these equations is a continuant matrix of the form

$$[(2+\nu_{2p-1})\,I-S-S^*-C]$$

and the equations are easily solved by the well-known factorization method.

TABLE 5

	$y=-\frac{3}{8}$	$y=-\frac{1}{4}$	$y=-\frac{1}{8}$	$y=0$
		sum of first two harmonics		
$r=1\frac{7}{8}$	4·873	8·170	9·585	9·887
	7·287	12·372	14·739	15·314
	8·235	14·051	16·835	17·540
	8·191	13·963	16·712	17·403
	7·368	12·482	14·827	15·385
	5·818	9·732	11·385	11·726
$r=1\frac{1}{8}$	3·452	5·654	6·440	6·545
		sum of first three harmonics		
$r=1\frac{7}{8}$	5·052	8·033	9·511	10·080
	7·501	12·208	14·650	15·546
	8·456	13·881	16·744	17·779
	8·413	13·793	16·620	17·643
	7·588	12·314	14·736	15·623
	6·027	9·572	11·298	11·952
$r=1\frac{1}{8}$	3·618	5·526	6·371	6·725
		contribution from fourth harmonic		
$r=1\frac{7}{8}$	0·017	−0·030	0·040	−0·043
	0·019	−0·036	0·047	−0·050
	0·020	−0·036	0·048	−0·051
	0·020	−0·036	0·048	−0·052
	0·020	−0·036	0·047	−0·051
	0·019	−0·035	0·046	−0·050
$r=1\frac{1}{8}$	0·016	−0·029	0·038	−0·041

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

The four terms of the solution can be regarded as successive harmonics and we should normally expect that the higher harmonics will make only a small contribution to the solution; if their contribution is of the same magnitude as the contribution from the lower harmonics, the solution may be a poor approximation to the solution of the differential equation (32). The constituent terms of the solution are recorded in table 5.

4. Solutions generated from factorial polynomials

In preceding sections, it was necessary to introduce the boundary conditions at the very outset of the numerical solution. Here, we temporarily ignore the boundary conditions and seek first solutions which satisfy the difference equation; the boundary conditions are subsequently fulfilled by forming a suitable linear functional of these functions. In the language of function space, we regard the true solution as a vector whose components individually satisfy the difference equation.

The polynomial solutions developed below are examples of a class of functions which satisfy Laplace's difference equation; this class is called preharmonic by Allen & Murdoch† (1953) who give references to earlier work.

We use a Cartesian lattice and consider first Laplace's equation defined over a rectangular area, the function being prescribed on the boundary. There is no inherent difficulty in extending the technique to domains bounded by curved lines, but in doing so, we shall be forced to consider more closely the uniqueness of the solution.

The reduced factorial polynomials which we have already used in part I are obviously suited to our purpose here. We use the notation of Aitken (1932).

$$\left.\begin{aligned} &\text{Reduced descending factorials:} \quad x_{(p)} = \frac{x(x-1)\,(x-2)\,\ldots\,(x-p+1)}{p!}, \\ &\text{Reduced central factorials:} \quad x_{\{p\}} = (x+\tfrac{1}{2}(p-1))\,(p). \end{aligned}\right\} \tag{37}$$

These polynomials satisfy the difference equations

$$\Delta x_{(p)} = x_{(p-1)}, \quad \delta x_{\{p\}} = x_{\{p-1\}}. \tag{38}$$

From these definitions it is clear that the functions

$$\left.\begin{aligned} U_p &= x_{\{p\}} - x_{\{p-2\}}\, y_{\{2\}} + x_{\{p-4\}}\, y_{\{4\}} - \ldots, \\ V_p &= x_{\{p-1\}}\, y_{\{1\}} - x_{\{p-3\}}\, y_{\{3\}} + x_{\{p-5\}}\, y_{\{5\}} \end{aligned}\right\} \tag{39}$$

satisfy the Cauchy–Riemann equations

$$\delta_x U_p = \delta_y V_p, \quad \delta_y U_p = -\delta_x V_p \tag{40}$$

and that both satisfy Laplace's difference equation. More generally, the function

$$z(x, y) = \Sigma(a_p U_p + b_p V_p) \tag{41}$$

satisfies the finite-difference equation at all lattice-points in the domain. The constants a_p, b_p are at our disposal to satisfy the boundary conditions.

If there are $n \times m$ points in the interior of the domain, the boundary will contain $2(n+m)$ points. Consequently, there will be a reduction in the number of unknowns if mn is greater than $2(m+n)$, i.e. $(m-2)\,(n-2) > 4$. The limiting cases are 4×4 and 6×3.

† We are indebted to Dr H. I. Scoins for this reference.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

4·1. *Suitable sets of polynomials*

There are two infinite sets of functions at our disposal and they are all linearly independent. It appears that we may make any suitable choice of these functions and obtain a solution satisfying the boundary conditions; if so, however, we may make a second choice and obtain another solution. From the considerations of § 3 there is a unique solution. Consequently, the difference of our two choices is a function which is zero at all interior lattice-points and at the boundary points—*but* not necessarily zero at intermediate points.

We illustrate by some very simple examples—chosen, not to show the method in action where it is efficient, but to elucidate the apparent non-uniqueness of the solution and the considerations which must govern the choice of component functions.

Consider first the simplest possible case of four boundary points and one interior point. The result is trivial and well-known; the central value is the average of its four neighbours. But to apply the suggested method.

The possible functions starting with those of lowest degree, are

$$U_0 = 1, \quad V_0 = 0,$$
$$U_1 = x, \quad V_1 = y,$$
$$U_2 = \tfrac{1}{2}(x^2-\tfrac{1}{4}) - \tfrac{1}{2}(y^2-\tfrac{1}{4}) = \tfrac{1}{2}(x^2-y^2),$$
$$V_2 = xy \ldots.$$

Taking the origin at the centre of the square, we choose four functions of the lowest degree available; V_2 is *not* suitable since it vanishes at all boundary points. Putting

$$z = a+bx+cy+d(x^2-y^2),$$

and, using the notation familiar in 'relaxation', we have

$$z_1 = a+b \quad +d,$$
$$z_2 = a \quad +c-d,$$
$$z_3 = a-b \quad +d,$$
$$z_4 = a \quad -c-d$$

and

$$z_0 = a.$$

Eliminating a, b, c, d, we have

$$z_0 = \tfrac{1}{4}(z_1+z_2+z_3+z_4);$$

but we possess an unexpected second-degree interpolation polynomial for the values at intermediate points—unless $z_1+z_3 = z_2+z_4$, when the polynomial is of first degree.

It is not, however, necessary that the polynomials be centred at the centre of the square. Let the origin be taken at the south-west corner. This time V_2 is admissible and we take

$$z = a+bx+cy+dxy,$$

then

$$z_1 = a+2b+ \ c+2d,$$
$$z_2 = a+ \ b+2c+2d,$$
$$z_3 = a+ \quad + \ c,$$
$$z_4 = a+ \ b$$

and

$$z_0 = a+b+c+d = \tfrac{1}{4}(z_1+z_2+z_3+z_4).$$

It may be verified that the same value for z_0 is obtained if (x^2-y^2) replaces xy.

Similar results are to be expected for larger domains and any *suitable* choice of solving functions will lead to the unique solution at the lattice points of a rectangular domain. The boundary conditions determine a matrix of order $2(m+n)$ which operates on the coefficients a_p, b_p; if this matrix is inverted, the influence of each boundary value upon the values at the lattice points is obtained.

It is natural in seeking a solution to choose functions of lowest order—if only for case in computation; but there is no compelling reason to do so and it is meaningless to describe one choice of polynomials as more exact than any other choice. That the results of the different choices for interpolation would be very different, although true, is really not relevant to the problem in hand: the solution is defined only at the lattice points and any attempt to interpolate can be based only on lattice-point values. We may borrow Whittaker's term 'cotabularity' to describe these solutions which are identical at the lattice points and boundary points.

In applying this method to regions bounded by curved lines, two courses are open. We may adhere to a net of lattice points which fit the boundary as closely as possible or we may use broken lattice lines which terminate on the boundary. It is convenient here to introduce the following terminology for lattice points of the domain: *inner points*, all of whose lattice neighbours lie in the region, *outer points*, one or more of whose lattice neighbours lie outside the region,† and *boundary points* which lie in the boundary curve. These distinctions are commonly accepted in numerical work though the terminology may not be. If the values of the function at outer points are estimated from the given boundary conditions, the values at inner points can be obtained straightforwardly by using p suitable polynomials if p is the number of outer points. If boundary points are used and if they are q in number, it is possible to use q polynomials with arbitrary constants a_k, b_k equating the value of $\Sigma(a_k U_k + b_k V_k)$ at each boundary point to the value prescribed at that point; but the solution obtained in this way may not be free from ambiguity, since different choices of polynomials will in general interpolate differently at boundary points. An unambiguous solution can, however, be obtained by using p polynomials to obtain the values of the function at the outer points in terms of p arbitrary constants. The constants are evaluated by setting up approximate finite-difference equations relating the outer points to the prescribed boundary points. Similar considerations apply if the boundary conditions are of a more general form.

The elucidation of polynomial solutions of the difference equations of mathematical physics would be of practical value in computation. It seems likely that many trial-and-error methods of solving difference equations are equivalent to the use of polynomial solutions. If polynomial solutions of an adequate range of difference equations were available, they would serve to uncover some of the hidden assumptions in trial-and-error methods and they would reveal the discrepancy to be expected between the solution of a differential system and the solution of an approximating difference system on meshes of practical size.

† Courant *et al.* (1928), distinguish two sets of points, Randpunkte and Innere Punkte which appear to be identical with our outer and inner points.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

4·2. *Some practical considerations*

In equation (39), the polynomials U_p and V_p are defined for unit tabular interval. If the tabular interval is h, we must replace the fundamental definitions (37) by

$$\left.\begin{aligned} x'_{(p)} &= \frac{x'(x'-h)\,(x'-2h)\,\ldots\,(x'-(p-1)\,h)}{p!}, \\ x'_{\{p\}} &= (x'+\tfrac{1}{2}(p-1)\,h)_{(p)}. \end{aligned}\right\} \qquad (37')$$

If $x' = xh$ (x, integral), equation (37′) are identical with (37) save for the multiplicative factor h^p. Similarly, the definitions of U_p and V_p in (39) may be used for any tabular interval if a factor h^p is inserted.

The use of these polynomials in desk computation can be provided for if they are tabulated for integral x and y over an adequate range. When a problem is solved within an automatic computing machine, the accessibility of the polynomial values in the computation presents a very different problem. Both these requirements encounter the practical difficulty that the tabular values of the polynomials are spread over the range $\pm 10^6$ for $0 \leqslant x,\ y \leqslant 10$, $0 \leqslant p \leqslant 10$. A suitable norm must be found for each polynomial and practical use requires that the norm should originate naturally from the problem to be solved; one method of normalizing is given in example 4·21 and a second method is given in § 5.

The solution of a difference equation in a given domain can always be obtained as accurately as we may desire by using a sufficient number of polynomials; but this accuracy may be unnecessary and even misleading if our ultimate aim is to obtain a reliable approximation to the solution of a differential equation, the solution of the difference equation being merely an intermediate step. For example, the polynomials U_p and V_p are identical with the real and imaginary parts of $(x+iy)^p/p!$ for $p \leqslant 2$; for $p > 2$, the discrepancy between the polynomial solutions of Laplace's differential and difference equations becomes progressively greater as p increases. The practical question is 'how many polynomials are really needed in the approximate solution of a given problem?' This is equivalent to the requirement that an estimate of the error committed in adopting a given solution should be readily available.

It is also pertinent to inquire whether the intermediate step of replacing a differential system by a difference system is profitable here, since the considerations of the present section can be readily employed in generating solutions whose components are functions of a *continuous* variable. No general answer can be given to this question, since the relative merits of discrete and continuous approximations depend very largely on the nature of the problem to be solved.

The use of the polynomials U_p and V_p in a conformal transform is exemplified in example 4·21.

Example 4·21

The conformal transformation of the quadrant bounded by the x and y axes and by the curves

$$\frac{x^2}{2{\cdot}38110} + \frac{y^2}{1{\cdot}38110} = 1, \qquad (42)$$

$$2(x^2 - y^2) = 1 \qquad (43)$$

into a rectangle provides a simple illustration of the practical considerations mentioned in § 4·2. We denote the transforming function by

$$w = u(x, y) + \mathrm{i}v(x, y)$$

and seek a discrete solution of the form

$$w = \Sigma a_p (U_p + \mathrm{i}V_p) \tag{44}$$

and we may take v to be unity on the curve (43). The value of u on the curve (42) is constant, but this constant is initially unknown and we temporarily denote it by c (the analytical value of c is $\frac{1}{4}\pi$). The function u is odd in x and even in y; v is even in x and odd in y.

We choose $h = \frac{1}{4}$; with this interval, the value of u or v is available at nine boundary points. These values can be used to determine the coefficients a_p in terms of the constant c and nine components suffice to solve the difference equation exactly. If a smaller number of components is employed, the determining equations can be satisfied in a least-square sense. The boundary data will not then be satisfied exactly and the constant c can be chosen to minimize the sum of the squared errors at the boundary; this is the procedure followed here, using two, three or four complex components. Despite the cautionary remarks in § 4·1, we interpolate freely in obtaining the boundary values of the polynomials since we are not seeking high accuracy here.

A suitable norm for the components can be formed thus. For a fixed value of p, we write down in order the numerical values of V_p on the curve $v = 1$ and the numerical values of U_p on the curve $u = c$. These values form a vector b_p of order 9, and the norm of the vector can be defined as $(b_p^* b_p)^{\frac{1}{2}}$; the same norms can be used for the components. It is now relatively easy to set up equations to determine the coefficients a_p, or rather, the coefficients $\bar{a}_p$, the bar being used to distinguish the coefficients appropriate to normalized components.

The determining equations for four components is given below. The matrix on the left of the equations is symmetric and only the lower half is recorded.

$$\begin{bmatrix} 1{\cdot}00000 & . & . & . \\ -0{\cdot}19628 & 1{\cdot}00000 & . & . \\ -0{\cdot}46791 & -0{\cdot}41926 & 1{\cdot}00000 & . \\ -0{\cdot}20083 & -0{\cdot}49250 & 0{\cdot}17721 & 1{\cdot}00000 \end{bmatrix} \begin{bmatrix} \bar{a}_1 \\ \bar{a}_2 \\ \bar{a}_3 \\ \bar{a}_4 \end{bmatrix} = \begin{bmatrix} 1{\cdot}82092 + 1{\cdot}13754c \\ -0{\cdot}25945 - 0{\cdot}01996c \\ -0{\cdot}73469 - 0{\cdot}71870c \\ -0{\cdot}50285 - 0{\cdot}21219c \end{bmatrix}.$$

The corresponding equations for two or three components are obtained by truncating the above set.

The value of the constants $\bar{a}_p$ are

two components	three components	four components
$\bar{a}_1 = 1{\cdot}84091 + 1{\cdot}17905c$	$\bar{a}_1 = 2{\cdot}03122 + 1{\cdot}09102c$	$\bar{a}_1 = 2{\cdot}01612 + 1{\cdot}18010c$
$\bar{a}_3 = 0{\cdot}10187 + 0{\cdot}21146c$	$\bar{a}_3 = 0{\cdot}27866 + 0{\cdot}12969c$	$\bar{a}_3 = 0{\cdot}25711 + 0{\cdot}25682c$
	$\bar{a}_5 = 0{\cdot}33257 - 0{\cdot}15383c$	$\bar{a}_5 = 0{\cdot}32148 - 0{\cdot}08843c$
		$\bar{a}_7 = -0{\cdot}02830 + 0{\cdot}16696c$

These values of the coefficients can now be used to evaluate the error at each of the nine boundary points. If we choose c to minimize the sum of the squared errors, the resultant values of c using two, three and four components are

$$0{\cdot}7856 \quad 0{\cdot}7794 \quad 0{\cdot}7831 \quad (c \to \tfrac{1}{4}\pi,\ h \to 0).$$

The numerical values of the errors at the boundary points are recorded in table 6, rounded off to three decimals.

These results suggest that reliable approximations can be obtained using a small number of components; but further trial on a range of examples is desirable.

TABLE 6

two components	three components	four components
−0·036	0·020	−0·006
−0·026	0·002	0·002
−0·002	−0·032	0·009
0·022	−0·035	−0·009
0·011	0·038	0·003
−0·092	−0·042	−0·021
−0·058	−0·016	0·000
0·026	0·032	0·030
0·125	0·025	−0·009

5. VARIATIONAL METHODS

The concepts developed in this section are complementary to those of the preceding section and we shall employ the terminology introduced there to distinguish different sets of points; we suppose that the set of outer points approximate to the true boundary with sufficient accuracy and we shall make no use of boundary points.

The application of variational concepts to finite-difference equations is due in the main to the fundamental paper by Courant *et al.* which we have cited several times; but the range and depth of these concepts have been considerably enriched in recent years. The use of variational methods depends mainly on the choice of a suitable scalar which induces a positive-definite metric in the domain. If there are p outer points in a closed domain of any shape, the metric determines a function space of p dimensions consisting of functions which satisfy the difference equation and any such function in the domain can be expressed in terms of p orthogonal components; we may, in fact, regard the function as a vector in the function space. The sense in which we use the terms 'function', 'orthogonal component' and 'vector' will appear presently. The nominal finitude of the function space may be of little interest to a computer if p is large, but the concept is none the less important.

We shall typify these concepts by applying them to Laplace's difference equation and by reviewing rapidly certain fundamental results of Bergman's kernel function method (Bergman 1953) and of the hypercircle method (Prager & Synge 1947). These results are developed here for a finite-difference equation, but it will be apparent from the preceding considerations, or by reference to the sources, that they are direct translations of results obtained for *differential* equations by the authors cited.

We consider two functions U and V defined at the lattice points of a closed domain of any shape and form a scalar $E(U, V)$ from their divided first differences. The scalar is formed thus: the forward difference between each inner point and its four neighbours (the neighbours may be inner points or outer points) is divided by the mesh length h and the divided difference is denoted by the symbol Δ and corresponding differences of U and V are multiplied together and summed; no difference is formed between an outer point and a neighbouring point. By definition,

$$E(U, V) = h^2 \sum_{I} \Delta U_i \Delta V_i, \tag{45}$$

the suffix I indicating that differences between outer points are excluded. Transforming the sum and ordering with respect to V, we have

$$E(U, V) = -h^2 \sum_I V_i(\delta_{xx}^2 + \delta_{yy}^2)\, U_i + h \sum_0 V_i U_i^\nu. \tag{46}$$

The first summation on the right is effected over inner points only and the second over outer points; x and y are the co-ordinate directions and the symbol U_i^ν denotes the *outward* differences between the outer point U_i and each of its neighbouring inner points, divided by h: the metric of the domain is determined by the coefficients of the bilinear form. A function U can be regarded as a vector of unit length in the function space defined by the metric if

$$E(U, U) = 1$$

and two functions U and V can be regarded as orthogonal vectors if

$$E(U, V) = 0.$$

We may then employ the vector notation

$$U.V \quad \text{or} \quad (U.V) \quad \text{instead of} \quad E(U, V).$$

From the definition of E it is clear that

$$E(U+U', U+U') = E(U, U) + 2E(U, U') + E(U', U'). \tag{47}$$

If U is a function which satisfies the difference equation at inner points and U' is any function which satisfies zero boundary conditions at outer points, then from (47)

$$\left.\begin{aligned} E(U+U', U+U') &= E(U, U) + E(U', U') \\ &\geqslant E(U, U), \end{aligned}\right\} \tag{48}$$

i.e. U possesses an extremum property with respect to all functions which satisfy the same boundary conditions as U. It is easy to show by the method of Lagrangian multipliers that the function which satisfies the difference equation and the prescribed boundary conditions possesses a minimum property with respect to all functions which satisfy the same boundary conditions.

We now return to the factorial polynomials of the preceding section which satisfy the Laplace difference equation at inner points, and we now use the vector notation $U_1, U_2, \ldots$ for these functions making no distinction between the U set and the V set of functions. If W is any other function of the space, (46) reduces to

$$E(W, U_i) = h \sum_0 W U_i^\nu = W.U_i \tag{49}$$

and the right-hand side of (49) can be evaluated if the boundary values of W are prescribed. If W satisfies the difference equation, the right-hand side of (49) can be written in either of the forms

$$h \sum_0 W U_i^\nu = h \sum_0 U_i W^\nu \tag{50}$$

and can be evaluated if W or W^ν is prescribed. The solution U which satisfies the difference equation and the boundary conditions can now be determined by a routine slightly different from that of § 4. We write

$$U = c_1 U_1 + c_2 U_2 + \ldots + c_p U_p$$

and

$$U.U_1 = c_1 U_1^2 + c_2(U_1.U_2) + \ldots + c_p(U_1.U_p).$$

The scalar products on the left side of these equations can be determined from the prescribed boundary conditions, and the symmetric matrix on the right has the elements $(U_i . U_j)$. If any of the components U_i satisfied zero boundary conditions, the corresponding row and column of the matrix would be zero; we may suppose that any such components have been discarded from the sequence. More generally, we must suppose that the matrix $(U_i . U_j)$ is non-singular; the component vectors are then independent in the domain. To prove that p independent vectors exhaust the space, it is sufficient to observe that any prescribed boundary values at p outer points can be obtained by superposition of p independent vectors.

A more elegant method of reaching the solution depends on the construction of a set of orthogonal and normed vectors $I_1, I_2, \ldots, I_p$ from the set $U_1, U_2, \ldots$ by the Gram–Schmidt process. The method of construction of the solution ensures that the vectors I_k are independent.

The vector I_1 is obtained from U_1 by choosing a norming constant such that

$$I_1 . I_1 = 1.$$

If U_2 is not orthogonal to I_1, then

$$I_1 . (U_2 - \alpha I_1) = 0, \quad \text{where} \quad I_1 . U_2 = \alpha$$

and we define I_2 as $N(U_2 - \alpha I_1)$, N being the norming constant. The remaining vectors I_k are similarly defined. The components I_k can be regarded as projection operators and the function

$$K = I_1 + I_2 + \ldots + I_p$$

can be regarded as the kernel function or the unit multiplier for the domain in the sense that the operation of K on any function W defined in the domain and satisfying the difference equations, yields the function W again. If W does not satisfy the difference equation, the operator K selects from it those components which satisfy the difference equation and rejects the remainder:

$$K . W = (W . I_1) I_1 + (W . I_2) I_2 + \ldots + (W . I_p) I_p.$$

When the kernel function for a given domain has been constructed it is a trivial task to determine the function U which satisfies the difference equation and boundary conditions; for

$$K . U = \sum_{k=1}^{p} (U . I_k) I_k = U$$

and the coefficients $(U . I_k)$ are determined by (49) or (50).

This definition of the kernel function as unit multiplier is due to Bergman who has applied the concept to a wide class of differential equations. In general, the construction of the kernel for a finite-difference problem would not be justified unless we wished to obtain solutions for several sets of boundary conditions, since the determination of the kernel components by orthogonalizing is in general a more laborious task than a straightforward application of the polynomial method; but the bare knowledge of the existence of the kernel function gives greater confidence in the use of polynomials.

What then becomes of the reputed independence of the factorial polynomials of § 4? The answer is that they are independent in a lattice space containing an infinite number of points but not all independent in a finite lattice space.

Hitherto, in this section, the solution has been constructed from functions which satisfied the difference equation. We might, however, have proceeded from satisfaction of the boundary conditions to satisfaction of the difference equation.

If U' is any function which satisfies zero boundary conditions in a Dirichlet problem and if V satisfies the difference equation, (49) shows that

$$U.V = 0,$$

i.e. the subspace of functions which satisfy zero boundary conditions is orthogonal to the subspace of functions which satisfy the difference equation. Starting from any function $\overline{U}$ which satisfies the boundary conditions, we may seek a solution by minimizing

$$E(W, W),$$

where

$$W = \overline{U} + \sum_1^N c'_k I'_k.$$

The superscript N denotes the number of inner points and I'_k is a set of orthonormal functions with zero boundary values. The coefficients c'_k are given by

$$c'_k = -I'_k.\overline{U}.$$

These coefficients are not easy to compute and the number of inner points usually exceeds the number of outer points; for these reasons the kernel function approach appears preferable if a closed solution is being sought, but it may happen in the solution of practical problems that a small number of components suffices to give a reliable approximation.

The variational concept can obviously be used in conjunction with difference equations of higher order or in conjunction with higher-difference correction terms. As elsewhere in analysis, the fundamental scalar (or metric) can be used to yield directly the difference equations and boundary conditions appropriate to the problem; or it can be used indirectly to motivate the choice of component functions.

6. Eigenvalues and eigenfunctions of partial-difference equations

The preceding sections discussed a number of techniques for the equilibrium problems of partial-difference equations. The rational matrix solution of § 3 is purely algebraic, but the irrational solutions of that section are clearly analogous to orthogonal function expansions in the theory of differential equations. Sections 4 and 5 discussed two types of solution; all the solutions of the first type individually satisfied the difference equation and all the solutions of the second type satisfied the boundary conditions.

We now seek the eigenfunctions of a partial difference equation and its boundary conditions. We suppose that the set of eigenfunctions is complete and spans the finite-difference domain under consideration; in other words, we suppose that an arbitrary function defined in the domain can be uniquely represented in terms of the eigenfunctions. Eigenfunction expansions can be used to solve partial-difference equations in much the same way as the one-dimensional equation $Ax = f$ can be solved by expanding f and x in terms of the eigenvectors of A.

We consider first a rectangular domain and seek solutions of the matrix equation

$$AU + UB = fU \tag{51}$$

and of the adjoint equation

$$VA + BV = fV. \tag{52}$$

The matrices U $(n \times m)$ and V $(m \times n)$ are rectangular, A and B being square matrices; f is a scalar. The problem is to determine the values of f for which non-trivial solutions exist and then to determine the corresponding arrays U and V.

Intuition suggests that there may well be some intimate connexion between the values of f and the eigenvalues of the matrices A and B. For example, if we put

$$f = \lambda + \nu$$

in equation (51) and bring the term on the right of the equation to the left, we have

$$(A - \lambda I)\, U + U(B - \nu I) = 0. \tag{53}$$

An obvious solution is obtained by making both terms in (53) vanish. We may choose λ to be an eigenvalue of A and ν to be an eigenvalue of B; the columns of U must then be eigencolumns of A and the rows of U eigenrows of B, i.e.

$$U = U_{ij} = r_i \sigma_j \quad \text{when} \quad f = \lambda_i + \nu_j; \tag{54}$$

the eigenmatrix U_{ij} is now completely determined apart from an arbitrary multiplicative factor. We follow here the notation established earlier, namely

A: eigenvalues, λ_i; eigencolumns, r_i; eigenrows, ρ_i.
B: eigenvalues, ν_j; eigencolumns, s_j; eigenrows, σ_j,

$$\rho_i r_j = \delta_{ij} = \sigma_i s_j.$$

Repeated eigenvalues of A or B will cause no difficulty provided that each root of multiplicity p has p independent eigenvectors. Similar considerations apply if

$$\lambda_i + \nu_j = \lambda_p + \nu_q \quad (i \neq p, j \neq q).$$

The requirement of completeness reduces, then, to the requirement that the eigenvectors of A should span a space of order n and that the eigenvectors of B should span a space of order m. Another line of reasoning follows from the observations that the matrix equation (51) can be written as a vector equation of order nm.

The matrices U_{ij} form a complete set of base matrices and any matrix Z of order $n \times m$ can be analyzed into a linear sum of the base matrices, i.e.

$$Z = \sum_{i,j} z_{ij} U_{ij}. \tag{55}$$

The z_{ij} are here the elements of Z in the basis U_{ij} and they can be determined by forming the scalar $\rho_p Z s_q$:

$$z_{pq} = \rho_p Z s_q, \quad \text{since} \quad \rho_p U_{ij} s_q = \delta_{pi}\delta_{jq}. \tag{56}$$

The adjoint equation (52) can be treated in a similar way. It can be shown that the eigenfunction which corresponds to the eigenvalue $\lambda_i + \nu_j$ of (52) is

$$V_{ij} = s_j \rho_i \quad \text{(note the transposition of suffixes)}.$$

This method of defining the components of the adjoint basis has the slight advantage that V_{ij} is the transpose of U_{ij} when the matrices A and B are symmetric.

It can also be shown that any matrix Z of order $m \times n$ can be analyzed in the form

$$Z = \sum_{i,j} z_{ij} V_{ij}, \tag{57}$$

where

$$z_{ij} = \sigma_j Z r_i. \tag{58}$$

6·1. *Another formulation of the eigenfunction expansions*

The following results are easily verified

$$V_{pq}\,U_{ij} = \delta_{pi} s_q \sigma_j,$$
$$U_{ij} V_{pq} = \delta_{jq} r_i \rho_p;$$

and, in particular, $$V_{iq}\,U_{ij} = s_q \sigma_j, \quad U_{ij} V_{pj} = r_i \rho_p.$$

It may be noticed that the last two products above yield projection operators and that the first commutes with B if $q = j$ and the second with A if $p = i$.

These results enable us to rephrase the expansion theorems (55) and (57) in a manner more consonant with the theory of orthogonal functions. We shall turn the well-established notation $V.U$ to a new use by employing it to denote the trace of a matrix product, i.e.

$$V.U = \text{trace}\,(VU). \tag{59}$$

Since the trace of a projection operator such as $s_q \sigma_j$ is identical with the scalar product $\sigma_j s_q$, we have

$$V_{pq}.U_{ij} = U_{ij}.V_{pq} = \begin{cases} 1 & (p = i \text{ and } q = j), \\ 0 & (p \neq i \text{ or } q \neq j). \end{cases} \tag{60}$$

In this sense, we may say that the adjoint eigenfunctions V_{pq} are orthogonal to the eigenfunctions U_{ij}.

The coefficients of the expansion

$$Z = \Sigma z_{ij} U_{ij} \tag{55 bis}$$

are now determined by $$z_{ij} = V_{ij}.Z. \tag{61}$$

This rephrasing of the expansion theorems opens the way to the familiar Cauchy–Schwarz and Bessel inequalities which are of importance in the metric theory of finite-dimensional spaces.

It would be possible to formulate analogous results for more general difference equations and more general domains; for example, it is always possible to write the determing equations as a single vector equation by using a big matrix similar to that of § 3·1. But it may not be easy to state the orthogonality relations in concise and simple form.

6·2. *A perturbation technique*

We shall illustrate the above results by describing a perturbation technique which is suggested by the Lennard-Jones (1930) method for the solution of quantum-mechanical eigenvalue problems.

We consider the eigenvalue problem

$$AZ + ZB + [\alpha Z] = gZ, \tag{62}$$

where $[\alpha Z]$ denotes a matrix whose (ij)th element is $\alpha_{ij} z_{ij}$ and g is the eigenvalue we wish to determine; the remaining symbols have their usual connotation. It is assumed that the eigenvalues of

$$AU + UB = fU \tag{63}$$

are known and complete; the term $[\alpha Z]$ can, then, be regarded as a perturbation.

We write $$Z = \Sigma z_{ij} U_{ij}$$

and we may also write $$[\alpha Z] = \Sigma z_{ij} [\alpha U_{ij}] = \Sigma \beta_{ij} U_{ij}.$$

The coefficients β_{pq} are given by

$$\begin{aligned}\beta_{pq} &= \Sigma z_{ij} V_{pq} . [\alpha U_{ij}] \\ &= \Sigma \alpha_{pq,ij} z_{ij},\end{aligned}$$

where

$$\alpha_{pq,ij} = V_{pq} . [\alpha U_{ij}]. \tag{64}$$

Equation (62) can now be written as

$$(\lambda_p + \nu_q - g)\, z_{pq} + \Sigma \alpha_{pq,ij} z_{ij} = 0. \tag{65}$$

The concepts of the present section are not restricted to those problems in which a difference equation is formulated as a matrix equation of the type considered in §§ 2, 3 and 6. The fundamental requirement is that a complete set of base functions should be available. It is shown in another paper that a finite set of trigonometric functions is a suitable *continuous* basis in one-dimensional problems and that this basis can be used to obtain continuous approximations to continuous functions which are defined by a differential system or by an integral equation; it was also shown that this method of approximation is sometimes equivalent to the use of a higher-difference correction. In two-dimensional problems, the set of products of a pair of trigonometric functions can be used as a *continuous* basis, and—as in the one-dimensional case—the components of a given continuous function in this basis are obtained by a *discrete* definition. An adjoint basis can usually be found and the resolution theorems of the present section require little modification; this technique is exemplified below in examples 6·21 and 6·22.

Example 6·21

The differential system

$$\left[-\frac{\partial^2}{\partial x^2} + \frac{3\partial}{x\partial x} - \frac{\partial^2}{\partial y^2}\right] Z = 256, \quad z = 0, \quad y = \pm\tfrac{1}{2}, \quad x = 1, 2$$

may be solved by using the continuous basis

$$U_{pq} = \left(\frac{2}{n+1}\right)^{\frac{1}{2}} \sin p\pi x \left(\frac{2}{n+1}\right)^{\frac{1}{2}} \sin q\pi(y+\tfrac{1}{2}) \quad (0 < p,\ q < n+1),$$

where p and q are integers and q assumes odd values only; the tabular interval is $h = 1/(n+1)$.

The component functions may be differentiated since they are continuous, but the coefficients z_{pq} of the representation

$$z = \Sigma z_{pq} U_{pq} \tag{66}$$

are determined as in § 6·1 by using the matrix determined by the tabular values of U_{pq}; the term $3\partial z/x\partial x$ can be resolved by using the device of § 6·2. The determining equations for the z_{pq} are of the form

$$(p^2 + q^2)\, \pi^2 z_{pq} + \sum_{i,j} \alpha_{pq,ij} z_{ij} = 256. \tag{67}$$

The order of magnitude of the discrepancy between the above continuous solution [obtained from (67)] and the solutions of examples 3·32 and 3·33 can be determined without detailed arithmetic. The iterative solution of example 3·32 shows that the term $3\partial z/x\partial x$ can be neglected in a first approximation. Denoting the coefficients of the

continuous solution by z_{pq} and the coefficients of the analogous difference equation by z'_{pq}, we have approximately

$$z'_{pq}-z_{pq}=256\left[\frac{h^2}{4[\sin^2(\frac{1}{2}p\pi h)+\sin^2(\frac{1}{2}q\pi h)]}-\frac{1}{(p^2+q^2)\pi^2}\right].$$

The discrepancy is due mainly to the lower harmonics and we can write

$$z'_{pq}-z_{pq}\sim\frac{256}{12}\frac{p^4+q^4}{(p^2+q^2)^2}h^2,$$

when p and q are small compared with $1/h$. Taking, for example, $p=q=1$, $h=\frac{1}{8}$,

$$z'_{pq}-z_{pq}\sim\tfrac{1}{6}.$$

By reason of the normalization factor $\sqrt{2}h$, the elements of the base functions U_{pq} are always less than $\frac{1}{4}$ for $h=\frac{1}{8}$. Hence, the discrepancy between any two corresponding elements of the two solutions is of the order $\frac{1}{24}$.

The analytical solution of this problem (Michell 1900) confirms this estimate (the tabular values of the analytical solution are recorded in Fox (1947)). It may, however, be noted that the solution (66) is a cardinal function approximation to the true solution; if the analytical solution is not available, the simplest method of assessing the value of an approximation of this type is to compare the results obtained on two different tabular intervals.

Example 6·22

It was shown by Bolton & Scoins (1957) that the determination of the energy E and the wave functions of the Schrödinger equation for two electrons enclosed in a sphere can be reduced to a solution of the eigenvalue equation

$$\left[-\frac{\partial^2}{\partial x^2}-\frac{\partial^2}{\partial y^2}+\frac{2R}{\max(x,y)}\right]z=gz,$$

where $g=R^2E$ and $z=0$ on $x=0,1$ and $y=0,1$.

The solution of this problem can be obtained by assuming for z a continuous representation of the type (66). For example, for $h=\frac{1}{3}$, $R=10$, the eigenvalues of the symmetric matrix

$$\begin{bmatrix} 2\pi^2+37{\cdot}5 & 7{\cdot}5 & 7{\cdot}5 & 7{\cdot}5 \\ . & 5\pi^2+37{\cdot}5 & 7{\cdot}5 & 7{\cdot}5 \\ . & . & 5\pi^2+37{\cdot}5 & 7{\cdot}5 \\ . & . & . & 8\pi^2+37{\cdot}5 \end{bmatrix}$$

(only the upper triangle is recorded)

are approximations to the eigenvalues g. The approximations for $h=\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$, are recorded below.

	λ_1	λ_2
$h=\frac{1}{2}$	56·00	—
$h=\frac{1}{3}$	54·07	79·35
$h=\frac{1}{4}$	53·54	78·82
	(symmetric mode)	(antisymmetric mode)

Bolton & Scoins obtained approximations to λ_1 and λ_2 by relaxations, using meshes from $h=\frac{1}{2}$ to $h=\frac{1}{8}$, and then extrapolated to $h=0$. Their final values were

$$52{\cdot}99\pm0{\cdot}01 \quad \text{and} \quad 78{\cdot}76\pm0{\cdot}01.$$

IV. BOUNDEDNESS OF SOME INITIAL-VALUE PROCEDURES FOR NUMERICAL SOLUTIONS OF PARTIAL-DIFFERENCE EQUATIONS

1. Introduction

We consider here a number of arithmetical procedures for linear bivarite problems involving a time-like variate and a space-like variate; it is shown that all of them can—under certain fairly general conditions—be represented by the matrix equation

$$AZ+CZB=F. \tag{1}$$

Z and F are here rectangular matrices of the same dimensions, and A, B, C, are square matrices; Z is to be determined, the remaining four matrices being supposed known. The matrices A, B, C, possess very simple structures in the examples considered, and in each example a solution of the appropriate form of (1) can be exhibited as an inverse matrix operating on a column of values. Inspection of the coefficients of this inverse matrix permits a ready assessment of the greatest possible growth of the matrix solution as the numerical procedure is carried forward from an initial time to a later time; equally, the greatest possible growth of a round-off error or blunder can be assessed simply. In this way, we can determine (i) the boundedness (with respect to the time co-ordinate) of the arithmetical result given by a numerical procedure, and (ii) upper bounds on the growth of an error or blunder.

The following partial-difference procedures are examined in § 3:

(i) elliptic-harmonic, stepping-ahead;
(ii) wave equation, (*a*) explicit, (*b*) implicit;
(iii) heat equation, (*a*) explicit, (*b*) implicit.

These procedures have been proposed as finite-difference techniques for the numerical solution of partial differential equations, and linear combinations of the implicit and explicit procedures in (ii) and (iii) are commonly employed. It is natural to ask whether the numerical answer provided by a given finite-difference procedure converges with diminishing mesh-length to the solution of the differential equation. A satisfactory treatment of convergence belongs to analysis rather than to algebra; but in a few instances the solution of the differential equation is of simple analytical form and it is then possible to treat the question of convergence by elementary considerations. We consider this question only briefly.

These topics have been treated extensively in the literature (see, for example, the bibliography in a recent paper by Todd (1956)); some of the solutions presented here have been given earlier in a somewhat different form by other workers and in particular by Todd (1956) who uses a method of resolution similar to that of § 2·3 below. Unification of a part of the theory in a single matrix equation, i.e. (1), may, nevertheless, be of interest.

2. Some prerequisite results

To avoid interruption of the arguments of §§ 3 and 4, or back-reference to parts I and III, it is convenient to repeat here some elementary results of matrix algebra which we shall require later. We follow notation used earlier in this paper, extending it where necessary.

2·1. The central-difference operator $-\delta^2$ is naturally represented in matrix algebra by the matrix operator $\mathbf{D}^2$, and the backward-difference operator ∇^2 is naturally represented by the matrix $\mathscr{D}^2$, where

$$\mathbf{D}^2 = 2I - S - S^*, \quad \mathscr{D}^2 = I - 2S^* + S^{*2},$$

I being the unit matrix and S a matrix with units in the super-diagonal and zeros elsewhere; S^* is the transpose of S. If necessary, the corner elements of $\mathbf{D}^2$ may be modified to take care of the boundary conditions. For example, the non-zero elements of the first column (or row), i.e. 2, -1, may be replaced by a_1, a_2; similarly, the non-zero elements of the last column (or row) of $\mathbf{D}^2$ may be replaced by a_{n-1}, a_n.

We posit (without attempting to prove) the distinction that space-like derivatives are naturally represented in matrix algebra by central-difference operators and time-like derivatives by backward-difference operators. The distinction arises from the fact that time processes are essentially open in the sense that advance to a new time does not disturb the steps already traversed. The corresponding matrix operator is naturally triangular. It is easy to see that the natural time-like operator for operations on a *column* is a lower-triangular matrix such as $\mathscr{D}^2$; for operations on a row, the appropriate operator is upper-triangular.

The utility of these distinctions will appear in the examples which follow.

2·2. To shift the rows of a matrix Z *up* one place, we pre-multiply Z by the superdiagonal matrix S; the first row of Z disappears when SZ is formed, and the elements of the last row of SZ are all zeros. Similarly, a down-shift of the rows of Z is effected on pre-multiplying by S^*; the elements of the first row of S^*Z are all zeros.

2·3. In solving the equation

$$AZ + CZB = F,$$

we can resolve Z into the components $\sum_1^n z_i \sigma_i$, where the z_i are columns of elements to be determined and the σ_i are eigenrows of B, i.e. $\sigma_i B = \nu_i \sigma_i$. Z and F are here rectangular $m \times n$ matrices, and B is square of order n. The $m \times n$ elements of Z can always be expressed as linear functions of the $m \times n$ elements in n columns z_i each containing m rows.

We can now analyze the matrix equation into vector equations as in § 3·3 of part III. Post-multiplying the equation by s_j $(\sigma_i s_j = \delta_{ij})$, we have

$$(A + \nu_j C)\, z_j = F s_j,$$

ν_j being an eigenvalue of B. This equation can be solved for z_j if the determinant of $(A + \nu_j C)$ is not zero.

2·4. The expression of the matrix $[I - \alpha S + \beta^2 S^2]^{-1}$ $(\alpha, \beta$, scalars) in powers of S is

$$\sum_0^\infty a_n S^n = I + \alpha S + (\alpha^2 - \beta^2)\, S^2 + \alpha(\alpha^2 - 2\beta^2)\, S^3 + (\alpha^4 - 3\alpha^2\beta^2 + \beta^4)\, S^4 + \ldots.$$

The coefficients a_n are related by the difference equation

$$a_n - \alpha a_{n-1} + \beta^2 a_{n-2} = 0,$$

and if we write

$$\alpha = 2\beta \cos\theta \quad \text{or} \quad \alpha = 2\beta \cosh\theta,$$

they can be expressed in the form

$$a_n = \beta^n \frac{\sin(n+1)\theta}{\sin\theta} \quad \text{or} \quad a_n = \beta^n \frac{\sinh(n+1)\theta}{\sinh\theta}.$$

The coefficients are certainly bounded if $\alpha = 2\beta\cos\theta$ and if $\beta \leqslant 1$; for $\beta = 1$, the sequence of coefficients is oscillatory. When $\alpha = 2\beta\cosh\theta$, the coefficients are in general unbounded.

3. Matrix formulation of numerical procedures

We are now in a position to consider the procedures mentioned in the Introduction. For definiteness, we shall keep in mind a block of four internal points in each row, the end points of the row having the column suffix 0 or 5 (table 7). We take the space-like co-ordinate x horizontal and the time-like co-ordinate y vertical. The mesh spacings are h_x, h_y, and

$$\kappa \begin{cases} = h_y^2/h_x^2 & \text{in the elliptic and wave equation examples,} \\ = h_y/h_x^2 & \text{in the heat equation examples.} \end{cases}$$

Table 7

00	01	02	03	04	05
.	.	.	.	.	.
10	11	12	13	14	15
.	+	+	+	+	.
20	21	22	23	24	25
.	+	+	+	+	.
30	31	32	33	34	35
.	+	+	+	+	.
40	41	42	43	44	45
.	+	+	+	+	.

$x \rightarrow$, $y \downarrow$

In each problem, the boundary values and initial values must be incorporated in the matrix formulation; but the genesis of the matrix equation is simple if the computer's steps in setting up the arithmetical procedure are visualized.

Example 3·1. *Elliptic-harmonic; stepping ahead*

.
. .
+
$z_{i+1,j}$

The difference analogue of the differential equation

$$\frac{\partial^2 z}{\partial y^2} + \frac{\partial^2 z}{\partial x^2} = 0$$

is

$$\frac{z_{i+1,j} + z_{i-1,j} - 2z_{ij}}{h_y^2} + \frac{z_{i,j+1} + z_{i,j-1} - 2z_{ij}}{h_x^2} = 0. \tag{2}$$

Treating the y variable as time-like, we regard (2) as an equation to determine $z_{i+1,j}$. The first two rows z_{0i} and z_{1i} $(i = 1, \ldots, n)$ must be assumed known in order to start the procedure, and boundary conditions at the ends of the range in x are required to continue.

We may set up the matrix representation by the transformations

$$h_y^2 \frac{\partial^2 z}{\partial y^2} \rightarrow \nabla^2 z_{i+1,j} \rightarrow \mathscr{D}^2 Z,$$

$$h_x^2 \frac{\partial^2 z}{\partial x^2} \rightarrow \delta^2 z_{ij} \quad \rightarrow -\mathbf{Z}\mathbf{D}^2,$$

but the second matrix term must be multiplied by the down-shift operator S^*, since in stepping ahead the x-operation is a row behind the z value which is currently being determined. The stepping-ahead matrix equation is then

$$\mathscr{D}^2 Z - \kappa S^* Z \mathbf{D}^2 = F_1, \tag{3}$$

where F_1 contains the boundary data.

To obtain a concrete picture of equation (3), we write out the first four rows of the illustrated scheme of table 7 (opposite):

$$\begin{bmatrix} 1 & . & . & . \\ -2 & 1 & . & . \\ 1 & -2 & 1 & . \\ . & 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} z_{11} & z_{12} & z_{13} & z_{14} \\ z_{21} & z_{22} & z_{23} & z_{24} \\ z_{31} & z_{32} & z_{33} & z_{34} \\ z_{41} & z_{42} & z_{43} & z_{44} \end{bmatrix} - \kappa \begin{bmatrix} . & . & . & . \\ z_{11} & z_{12} & z_{13} & z_{14} \\ z_{21} & z_{22} & z_{23} & z_{24} \\ z_{31} & z_{32} & z_{33} & z_{34} \end{bmatrix} \begin{bmatrix} 2 & -1 & . & . \\ -1 & 2 & -1 & . \\ . & -1 & 2 & -1 \\ . & . & -1 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} \bar{z}_{11} & \bar{z}_{12} & \bar{z}_{13} & \bar{z}_{14} \\ -\kappa\bar{z}_{10} - \bar{z}_{01} & -\bar{z}_{02} & -\bar{z}_{03} & -\bar{z}_{04} - \kappa\bar{z}_{15} \\ -\kappa\bar{z}_{20} & . & . & -\kappa\bar{z}_{25} \\ -\kappa\bar{z}_{30} & . & . & -\kappa\bar{z}_{35} \end{bmatrix}$$

The known values on the right side of the equation are barred, and the dots denote zeros. Here and elsewhere in the illustrative schemes we suppose that initial values (an arbitrary function of x) are prescribed on $y = 0$ and if necessary on $y = h_y$ and that boundary values (arbitrary functions of y) are prescribed at the ends of the range in x.

Returning to (3), we resolve Z into $\Sigma z_i \sigma_i$ and multiply by s_j:

$$(\mathscr{D}^2 - \kappa\nu_j S^*)\, z_j = F_1 s_j = f_j.$$

Since $\mathscr{D}^2 = I - 2S^* + S^{*2}$, we have

$$[I - \alpha S^* + S^{*2}]\, z_j = f_j,$$

i.e.
$$z_j = [I - \alpha S^* + S^{*2}]^{-1} f_j, \tag{4}$$

where
$$\alpha = 2 + \kappa\nu_j.$$

In the remaining examples we shall derive a number of equations similar to (4) and it is convenient to consider them all together.

Parenthetical note

The relation between the stepping-ahead equation (3) and the standard boundary value form of the elliptic-harmonic equation is easily demonstrated. We can in fact pass from (3) to the standard form on pre-multiplying across by S. Noting that

$$S\mathscr{D}^2 = \begin{bmatrix} -\mathbf{D}^2 \\ . \end{bmatrix}, \quad SS^* = \begin{bmatrix} I \\ . \end{bmatrix}, \quad SF_1 = \begin{bmatrix} F_2 \\ . \end{bmatrix},$$

we have from (3) on suppressing the zero bottom row

$$-\mathbf{D}^2 Z - \kappa Z \mathbf{D}^2 = F_2. \tag{5}$$

This equation is of standard boundary-value form.

Example 3·2 (*a*). *Wave equation; explicit form*

This is identical with the stepping-ahead form of the harmonic equation save for a change of sign in the second term of the matrix equation:

$$\mathscr{D}^2 Z + \kappa S^* Z \mathbf{D}^2 = F_1; \tag{6}$$

and, as before,

$$z_j = [I - \alpha S^* + S^{*2}]^{-1} f_j, \tag{7}$$

with

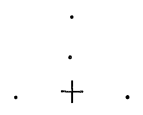
$$\alpha = 2 - \kappa \nu_j.$$

[*Note.* We can convert (7) to boundary-value form on pre-multiplying by S. The inverse matrix has then the form $[(2-\kappa\nu_j)\, I - S - S^*]^{-1}$. It can be seen, however, on purely algebraic grounds that this step is illegitimate. For example, if $\mathbf{D}^2$ is of order n and has the form $2I - S - S^*$, the roots ν_j are given by

$$\nu_j = 2[1 - \cos j\pi/(n+1)] \quad (j = 1, \ldots, n).$$

When n is odd and $j = \frac{1}{2}(n+1)$, $\nu_j = 2$ and the matrix $[(2-\kappa\nu_j)\, I - S - S^*]$ is singular for $\kappa = 1$.]

Example 3·2 (*b*). *Wave equation; implicit form*



The difference operation with respect to x is now acting in the same row as the quantity which is currently being determined, and the additional operator S^* is not required in formulating the matrix equation. The matrix Z of the two preceding examples is not, however, applicable here. The first backward difference with respect to y and the first central difference with respect to x operate on the row z_{2i}; hence, the rows z_{0i} and z_{1i} can be omitted from the matrix Z. The wave equation is then represented by the implicit form

$$\mathscr{D}^2 Z + \kappa Z \mathbf{D}^2 = F_1, \tag{8}$$

or

$$(\mathscr{D}^2 + \kappa\nu_j I)\, z = f;$$

i.e.

$$z_j = [(1 + \kappa\nu_j)\, I - 2S^* + S^{*2}]^{-1} f_j. \tag{9}$$

Example 3·3 (*a*, *b*). *Heat equation; explicit and implicit forms*

explicit,

implicit

These two illustrations are very similar to examples 3·2 (*a*) and (*b*) of the wave equation, and they may be treated together. The difference equations for the explicit and implicit forms of the finite-difference heat equation are

$$\frac{z_{i+1,j} - z_{ij}}{h_y} = \frac{z_{i,j+1} + z_{i,j-1} - 2z_{ij}}{h_x^2}$$

and

$$\frac{z_{i+1,j} - z_{ij}}{h_y} = \frac{z_{i+1,j+1} + z_{i+1,j-1} - 2z_{i+1,j}}{h_x^2}.$$

In each equation we seek to determine $z_{i+1,j}$. In the explicit case, the x-difference is operating on z_{ij}—a row behind the quantity being determined—and the pre-multiplier S^* is required with the space co-ordinate term; in the implicit case, the x-difference is operating on $z_{i+1,j}$. The backward difference with respect to the time-like term is represented by the matrix operator

$$\mathscr{D} = I - S^*.$$

A little care is needed in defining the operand Z. Reflexion shows that Z embodies the row z_{0i} $(i = 1, \ldots, n)$ and succeeding rows in the explicit case; in the implicit case, Z embodies z_{1i} and succeeding rows. The matrix equations are then

$$\left.\begin{aligned} \mathscr{D}Z + \kappa S^* Z \mathbf{D}^2 &= F_1 \quad \text{(explicit)} \\ \mathscr{D}Z + \kappa Z \mathbf{D}^2 &= F_2 \quad \text{(implicit)} \end{aligned}\right\} \quad (\kappa = h_y/h_x^2). \tag{10}$$

and (11)

Resolving Z, we obtain

$$z_j = (I - \alpha S^*)^{-1} f_j \quad \text{(explicit)}, \tag{12}$$

$$z_j = (\alpha I - S^*)^{-1} f_j \quad \text{(implicit)}, \tag{13}$$

where

$$\alpha = \begin{cases} 1 - \kappa\nu_j & \text{in (12)}, \\ 1 + \kappa\nu_j & \text{in (13)}. \end{cases}$$

The left-hand sides of (10) and (11) are sufficiently obvious, but it may be useful to write out the first few rows of the right-hand side in each equation:

$$F_1 = \begin{bmatrix} \bar{z}_{01} & \bar{z}_{02} & \bar{z}_{03} & \bar{z}_{04} \\ \kappa\bar{z}_{00} & . & . & \kappa\bar{z}_{05} \\ \kappa\bar{z}_{10} & . & . & \kappa\bar{z}_{15} \\ \kappa\bar{z}_{20} & . & . & \kappa\bar{z}_{25} \end{bmatrix}, \quad F_2 = \begin{bmatrix} \bar{z}_{01} + \kappa\bar{z}_{10} & \bar{z}_{02} & \bar{z}_{03} & \bar{z}_{04} + \kappa\bar{z}_{15} \\ \kappa\bar{z}_{20} & . & . & \kappa\bar{z}_{25} \\ \kappa\bar{z}_{30} & . & . & \kappa\bar{z}_{35} \\ \kappa\bar{z}_{40} & . & . & \kappa\bar{z}_{45} \end{bmatrix}.$$

4. Discussion of the inverse matrices

Collecting together the inverse matrices of the examples discussed in § 3, we have

$$\left.\begin{aligned} &\text{harmonic stepping-ahead:} && [I - (2 + \kappa\nu_j) S^* + S^{*2}]^{-1}, \\ &\text{wave equation} && \\ &\quad \text{explicit:} && [I - (2 - \kappa\nu_j) S^* + S^{*2}]^{-1}, \\ &\quad \text{implicit:} && [(1 + \kappa\nu_j) I - 2S^* + S^{*2}]^{-1}, \end{aligned}\right\} \quad (\kappa = h_y^2/h_x^2);$$

$$\left.\begin{aligned} &\text{heat equation} && \\ &\quad \text{explicit:} && [I - (1 - \kappa\nu_j) S^*]^{-1}, \\ &\quad \text{implicit:} && [(1 + \kappa\nu_j) I - S^*]^{-1}, \\ &\quad \text{averaged:} && [(1 + \tfrac{1}{2}\kappa\nu_j) I - (1 - \tfrac{1}{2}\kappa\nu_j) S^*]^{-1}, \end{aligned}\right\} \quad (\kappa = h_y/h_x^2).$$

The last line above is obtained by averaging the x-difference operation of the explicit and implicit procedures.

In the expansion of the typical matrix $(I - \alpha S^* + \beta^2 S^{*2})^{-1}$, the successive powers of S^* occupy successive subdiagonals and the coefficients of S^{*r}, S^{*r-1}, ..., appear in order (left to right) in the rth row of the inverse. The jth spectral component of the solution at any time is obtained on multiplying the appropriate row of the inverse by the column f_j. The contribution of the initial values to the solution is embodied in the first product or the first pair of products of the row-column multiplication; the contribution of the boundary values is embodied in the first or second product and succeeding products.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

The constitution of the column f_j is of interest. It is obtained on multiplication of the matrix F in equations such as (1) by the eigencolumns s_j. Each row of F can be written in the form

$$f_{r1}\sigma_1 + f_{r2}\sigma_2 + \ldots,$$

the σ_j being eigenrows of B. The product of the rth row of F by s_j yields the coefficient f_{rj}.

The conditions for boundedness have been discussed briefly in § 2·4, but we may note here the influence of the eigenvalues ν_j on the different procedures. It can be seen the step from explicit to implicit involves a strengthening of the main diagonal at the expense of off-diagonal terms—so long as the ν_j are positive and this is in general true in practical applications. In particular, the large positive roots which are apt to cause rapid growth in explicit procedures exercise a damping effect in implicit forms. The growth factor of the largest root is most clearly evident in the stepping-ahead form of the harmonic equation which has been included here on account of its formal similarity to the wave equation.

When boundary values of z are prescribed, the eigenvalues ν_j are given by

$$\nu_j = 4\sin^2\{\tfrac{1}{2}\pi j/(n+1)\}$$

and the upper and lower bounds of ν_j are 4 and 0. For other forms of boundary conditions, the eigenvalues can be determined by straightforward methods.

5. Convergence of the numerical procedures

The convergence of the procedures for the heat equation and wave equation is easily investigated when the boundary values are zero. The initial state may be analyzed into vector components and the convergence of the time factor for the jth component in the solution of the *heat* equation is determined by the coefficient of S^* in the inverse matrices above, i.e. by the expressions

$$(1+\kappa\nu_j)^{-1}\text{: implicit,} \quad (1-\kappa\nu_j)\text{: explicit.}$$

If we write

$$y = rh_y, \quad (n+1)\,h_x = 1, \quad \kappa = \frac{y}{r}(n+1)^2,$$

these become

$$\left[1+\frac{4y}{r}(n+1)^2\sin^2\left(\frac{\tfrac{1}{2}\pi j}{n+1}\right)\right]^{-r} \quad \text{and} \quad \left[1-\frac{4y}{r}(n+1)^2\sin^2\left(\frac{\tfrac{1}{2}\pi j}{n+1}\right)\right]^{r}.$$

In the limit $r\to\infty$, both these expressions approach $\exp(-\pi^2 j^2 y)$ for values of j which are small compared with n. When j is of the order of n, both expressions vanish rapidly with increasing r; i.e. the higher-order terms of the difference solution do not converge to terms of the differential solution but their contribution to the numerical solution is small. In the explicit case, the restriction $\kappa \leqslant \frac{1}{2}$ is needed to prevent the expression in square brackets becoming negative and less than minus one; this restriction is not required in the implicit case, but the freedom thus allowed may not be of much interest to a computer if he is seeking a reasonably accurate approximation to a solution of the allied *differential* equation.

In the *wave* equation with boundary values specified as zero, we anticipate a standing-wave type of solution and the fundamental quadratic encountered in § 2·4 should then have imaginary roots; these roots are

$$\text{implicit:} \quad 1 \pm \mathrm{i}(\kappa\nu_j)^{\frac{1}{2}}/(1+\kappa\nu_j),$$

$$\text{explicit:} \quad 2\,\mathrm{dn}^2\theta_j - 1 \pm 2\mathrm{i}k\,\mathrm{sn}\,\theta_j\,\mathrm{dn}\,\theta_j,$$

where

$$\mathrm{sn}\,(\theta_j, k) = \sin\{\tfrac{1}{2}\pi j/(n+1)\} \quad (\kappa = k^2 \leqslant 1).$$

The time variation is determined by

$$[1 \pm 2\mathrm{i}\sqrt{\kappa}\sin\{\tfrac{1}{2}\pi j/(n+1)\}]^{-r} \quad \text{and} \quad (2\,\mathrm{dn}^2\theta_j - 1 \pm 2\mathrm{i}k\,\mathrm{sn}\,\theta_j\,\mathrm{dn}\,\theta_j)^r$$

and convergence may be investigated as before if $\sqrt{\kappa}$ is replaced by $y(n+1)/r$. The absolute value of the roots in the explicit case is unity; in the implicit case, the absolute value is $(1+\kappa v_j)^{-\frac{1}{2}}$ and the solution of the differential equation is approached from below.

It may be remarked that it is always permissible to use quite small values of κ by decreasing h_y (keeping h_x fixed) and there is no apparent penalty—other than increased labour—for doing so. It is, however, clear that in the limit $\kappa \to 0$ we approach the solution of the differential-difference equation

$$h_x^2(\partial^r z_i/\partial t^r) = z_{i+1} + z_{i-1} - 2z_i,$$

the exponent r being 1 (heat equation) or 2 (wave equation). The numerical solution for unduly small values of κ may be a satisfactory approximation to the differential-difference solution, but it need not be a satisfactory approximation to the differential solution if the h_x mesh is coarse.

The implicit procedures examined above are always bounded but they may on occasion yield a poorer approximation than the corresponding explicit procedure. A simple example may point the warning against implicit reliance on implicit procedures.

TABLE 8. PROFILE OF THE STRING

κ	...	1	1	$\frac{1}{4}$	$\frac{1}{16}$	0	0	
h_x	...	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	0	
$x=\frac{1}{8}$		1	1·00	1·00	1·00	1·00	1	$y=\frac{1}{8}$
$\frac{1}{4}$		2	1·98	1·97	1·97	2·00	2	
$\frac{3}{8}$		3	2·87	2·82	2·77	2·93	3	
$\frac{1}{2}$		3	3·10	2·94	2·77	3·15	3	
$x=\frac{1}{8}$		0·91	0·94	0·91	0·96	0·99	1	$y=\frac{1}{4}$
$\frac{1}{4}$		1·74	1·76	1·66	1·79	1·88	2	
$\frac{3}{8}$		2·32	2·27	2·01	2·12	2·22	2	
$\frac{1}{2}$		2·21	2·52	1·92	1·90	1·82	2	
$x=\frac{1}{8}$		0·67	0·71	0·61	0·76	0·86	1	$y=\frac{3}{8}$
$\frac{1}{4}$		1·19	1·22	0·96	1·08	1·24	1	
$\frac{3}{8}$		1·42	1·45	0·99	0·98	0·90	1	
$\frac{1}{2}$		1·42	1·50	0·99	1·00	1·05	1	

(data)

A string of unit length is plucked into an isosceles triangle of height $4d$, the ends of the string being fixed. If the string is then released, the profile of the vibrating string is determined by two straight lines

$$z_1 = 8dx, \quad z_2 = 4d(1-2y) \quad (0 \leqslant x \leqslant \tfrac{1}{2},\ 0 \leqslant y \leqslant \tfrac{1}{2}),$$

which intersect on the characteristic $x = \frac{1}{2} - y$.

The explicit process with $\kappa = 1$ is exact here because the differences are formed symmetrically across the characteristic and any discontinuity in the second difference with respect to y is cancelled by an equal discontinuity in the x-difference. The implicit process is inexact; a few illustrative values expressed as multiples of d are given in table 8.

6. Concluding remarks

Procedures involving a larger number of points in the formulation of the governing difference equation may be treated by the methods given above. Similarly we may, as Todd (1956) has pointed out, use the technique to treat equations involving two space co-ordinates. Consider, for example, the equation

$$\frac{\partial^r z}{\partial t^r} = \frac{\partial^2 z}{\partial y^2} + \frac{\partial^2 z}{\partial x^2} \quad (r = 1, 2) \tag{14}$$

defined in a rectangular area of the (x, y)-plane. Forming differences with mesh length h in the x (horizontal) and y (vertical) directions, we replace z by a rectangular $n \times m$ matrix Z; the difference analogue of (14) is then

$$h^2(\partial^r Z/\partial t^r) + \mathbf{D}^2 Z + Z\mathbf{D}^2 = F, \tag{15}$$

the matrix F incorporating the boundary values. If the eigenvalues of $\mathbf{D}^2$ are λ_j and if the eigenrows and eigencolumns are ρ_i and r_j $(\rho_i r_j = \delta_{ij})$, we may replace Z by $\sum_1^n r_i z_i$, the z_i being *rows*; equation (15) can then be reduced to the row equation

$$h^2(\partial^r z_j/\partial t^r) + z_j(\mathbf{D}^2 + \lambda_j) = \rho_j F = f_j. \tag{16}$$

At this stage, the time derivative can be approximated by a difference, the row z_j becoming a matrix. At the same time, we may decide whether to be implicit or explicit, i.e. whether to incorporate or omit the shift operator S*.

The difference forms of the wave equation possess the simplifying feature that the roots of the fundamental quadratic (which determines the growth of the process) are complex and thus of equal absolute value. The difference forms of the heat equation possess only one real root. It is then possible to study in isolation the effect of replacing the differential operator by approximating difference operators, without reference to the nature of the initial values or the effect of an inhomogeneous term in the differential equation. This simplifying feature of the wave equation persists even when the operator possesses variable coefficients if the tabular interval is chosen so that the roots of the fundamental quadratic vary slowly; similar for the heat equation.

A very different problem exists when the solutions of a differential equation of the second or higher order increase (or decrease) at significantly different rates. If the equation is replaced by a difference equation, corresponding sets of discrete solutions (increasing and decreasing) of the difference equation will exist. It is then no longer possible to examine the solutions of the difference equation without consideration of the initial data.

The isolation of the increasing and decreasing solutions of a second-order ordinary differential equation has been examined in §§ 3·2, 3·3 of part I. The methods used there can be applied to partial differential equations involving initial-value conditions when a resolution of the type used in (16) above is practicable.

We are indebted to Mr R. S. Jenkins, who first brought to our attention the usefulness of matrix equations in formulating partial difference equations, and to Dr R. E. Gibson, especially for his invaluable assistance to one of us (W.G.B.) at the proof stage.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

References

Aitken, A. C. 1932 *Proc. Roy. Soc. Edinb.* A, **53**, 54.

Allen, A. C. & Murdoch, B. H. 1953 *Proc. Amer. Math. Soc.* **4**, 842.

Bergman, S. & Schiffer, M. 1953 *Kernel functions and elliptic differential equations in mathematical physics.* New York: Academic Press.

Bolton, H. C. & Scoins, H. I. 1957 *Proc. Camb. Phil. Soc.* **53**, 150.

Burgerhout, Th. J. 1954 *Appl. Sci. Res.* B, **4**, 161.

Clenshaw, C. W. 1957 *Proc. Camb. Phil. Soc.* **53**, 134.

Cornock, R. F. 1954 *Proc. Camb. Phil. Soc.* **50**, 524.

Courant, R., Friedrichs, K. & Lewy, H. 1928 *Math. Ann.* **100**, 32.

Foulkes, H. O. 1949 *Proc. Lond. Math. Soc.* (2) **50**, 196.

Fox, L. 1947 *Proc. Roy. Soc.* A, **190**, 31.

Herrick, S. 1951 *Math. Tab., Wash.*, **5**, 61.

Hyman, M. A. 1952 *Appl. Sci. Res.* B, **2**, 325.

I.A.A.T. 1956 *Interpolation and allied tables.* London: H.M.S.O.

Ince, E. L. 1944 *Ordinary differential equations.* New York: Dover.

Karlqvist, O. 1952 *Tellus*, **4**, 374.

Kosko, E. 1957 *Aero. Quart.* **8**, 157.

Lanzcos, C. 1957 *Applied analysis.* London: Pitman.

Lennard-Jones, J. E. 1930 *Proc. Roy. Soc.* A, **129**, 604.

Lennard-Jones, J. E. 1937 *Proc. Roy. Soc.* A, **158**, 208.

Michell, J. H. 1900 *Proc. Lond. Math. Soc.* **31**, 130.

Milne, W. E. 1953 *Numerical integration of differential equations.* New York: Wiley.

Morse, P. M. & Feshbach, H. 1953 *Methods of theoretical physics*, part 2, pp. 1008 *et seq.* New York: McGraw-Hill.

Numerov, B. 1933 *Publ. Observ. astrophys. central de Russe*, **2**, 188.

Prager, W. & Synge, J. L. 1947 *Quart. Appl. Math.* **5**, 241.

Rayleigh, Lord 1926 *The theory of sound*, vol. 1, p. 174. London: Macmillan.

Rutherford, D. E. 1932 *Proc. Akad. Wetenschap*, **35**, 54.

Rutherford, D. E. 1947 *Proc. Roy. Soc. Edinb.* A, **62**, 229.

Rutherford, D. E. 1952 *Proc. Roy. Soc. Edinb.* A, **63**, 232.

Stiefel, E. 1952 *Z. angew. Math. Phys.* **3**, 1.

Stöhr, A. 1950 *Math. Nachr.* **3**, 208.

Synge, J. L. 1947 *Proc. Roy. Soc.* A, **191**, 447.

Todd, J. 1950 *Proc. Camb. Phil. Soc.* **46**, 116.

Todd, J. 1956 *Commun. Pure Appl. Math.* **9**, 597.

Turnbull, H. W. & Aitken, A. C. 1945 *An introduction to the theory of canonical matrices.* London: Blackie.

Weitzenböck, R. 1932 *Proc. Akad. Wetenschap*, **35**, 60.

Whittaker, E. T. 1915 *Proc. Roy. Soc. Edinb.* **35**, 181.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES