THE NUMERICAL SOLUTION OF FINITE DIFFERENCE EQUATIONS WITH APPLICATIONS TO PROBLEMS IN FLUID DYNAMICS

DECLARATION

The following thesis is based on unsupervised research work carried out at the University of St. Andrews during the period 1949-1955. No part of the thesis has previously been presented for a higher degree. Where work was carried out in collaboration with others, an indication is given if possible of the individual contributions. In particular, the method of relaxation outlined on p. 31 was used by the present author although no contribution was made towards its discovery.

SUMMARY OF CONTENTS

INTRODUCTION

In the numerical solution of differential equations much attention has recently been given to the replacement of the differential equation by an equivalent finite difference equation. This technique has received a boost in recent years with the rapid development of computing machines.

The exact solution V of the difference equation depends on the mesh spacing, and as the mesh becomes infinitely fine, V may converge to the exact solution D of the differential equation. The conditions under which $V \rightarrow D$ is the problem of convergence. The computation of V may, however, be so laborious as to render this procedure impracticable. In such a case an approximate solution N of the finite difference equation is obtained. The conditions under which $N \rightarrow V$ is the problem of stability. The numerical methods used in obtaining approximate solutions of difference equations separate into two main types;

(1) explicit methods where the numerical solution is calculated step by step from the difference equation and its initial boundary conditions, and

(2) implicit methods where the unknown values of the function are given by a system of simultaneous equations.

The present dissertation is concerned with the numerical solution and stability of difference replacements. Few references are made to the problem of convergence. It is not

intended to duplicate material which can be found in recent text books on numerical analysis (1), but rather to give an account of the particular topics which have interested the author in recent years.

First and second derivatives of a dependent variable u are constantly being replaced, and so from Bickley (2), the following finite difference approximations of varying degrees of accuracy are listed with obvious notation

$$u'(0) = \frac{-u(0) + u(h)}{h} - \tfrac{1}{2} h u''(0) \ldots\ldots\ldots \quad (1)$$

$$= \frac{-u(-h) + u(0)}{h} + \tfrac{1}{2} h u''(0) \ldots\ldots\ldots \quad (2)$$

$$= \frac{-3u(0) + 4u(h) - u(2h)}{2h} + \tfrac{1}{3} h^2 u'''(0) \ldots\ldots \quad (3)$$

$$= \frac{-u(-h) + u(h)}{2h} - \tfrac{1}{6} h^2 u'''(0) \ldots\ldots\ldots \quad (4)$$

$$= \frac{u(-2h) - 4u(-h) + 3u(0)}{2h} + \tfrac{1}{3} h^2 u'''(0) \ldots\ldots \quad (5)$$

$$= \frac{-11u(0) + 18u(h) - 9u(2h) + 2u(3h)}{6h} - \tfrac{1}{4} h^3 u''''(0) \ldots\ldots \quad (6)$$

$$= \frac{-2u(-h) - 3u(0) + 6u(h) - u(2h)}{6h} + \tfrac{1}{12} h^3 u''''(0) \ldots\ldots \quad (7)$$

$$= \frac{u(-2h) - 6u(-h) + 3u(0) + 2u(h)}{6h} - \tfrac{1}{12} h^3 u''''(0) \ldots\ldots \quad (8)$$

$$= \frac{-2u(-3h) + 9u(-2h) - 18u(-h) + 11u(0)}{6h} + \tfrac{1}{4} h^3 u''''(0) \ldots\ldots \quad (9)$$

$$u''(0) = \frac{u(0) - 2u(h) + u(2h)}{h^2} - h\,u'''(0) \quad \ldots\ldots \quad (10)$$

$$= \frac{u(-h) - 2u(0) + u(h)}{h^2} - \frac{1}{12}h^2 u''''(0) \quad \ldots\ldots \quad (11)$$

$$= \frac{u(-2h) - 2u(-h) + u(0)}{h^2} + h\,u'''(0) \quad \ldots\ldots \quad (12)$$

$$= \frac{6u(0) - 15u(h) + 12u(2h) - 3u(3h)}{3h^2} + \frac{11}{24}h^2 u''''(0) \quad \ldots (13)$$

$$= \frac{-3u(-3h) + 12u(-2h) - 15u(-h) + 6u(0)}{3h^2} \quad \frac{11}{24}h^2 u''''(0) \quad \ldots\ldots (14)$$

$$= \frac{35u(0) - 104u(h) + 114u(2h) - 56u(3h) + 11(4h)}{12h^3} - \frac{5}{12}h^3 u'''''(0) \quad \ldots (15)$$

$$= \frac{11u(-h) - 20u(0) + 6u(h) + 4u(2h) - u(3h)}{12h^3} + \frac{1}{24}h^3 u'''''(0) \quad \ldots\ldots (16)$$

$$= \frac{-u(-2h) + 16u(-h) - 30u(0) + 16u(h) - u(2h)}{12h^3} + \frac{1}{180}h^4 u''''''(0) \quad \ldots (17)$$

$$= \frac{-u(-3h) + 4u(-2h) + 6u(-h) - 20u(0) + 11u(h)}{12h^3} \quad \frac{1}{24}h^3 u'''''(0) \quad \ldots (18)$$

$$= \frac{11u(-4h) - 56u(-3h) + 114u(-2h) - 104u(-h) + 35u(0)}{12h^3} + \frac{5}{12}h^3 u'''''(0) \quad \ldots$$

(19)

CHAPTER 1

ORDINARY DIFFERENCE EQUATIONS

In this chapter y will be taken as the dependent variable and x as the independent variable. The extent of the problem, if finite, will be from $x = 0$ to $x = L$. The number of internal nodes will be N and so $(N + 1)h = L$, where h is the mesh length.

If at least one boundary condition is given at each end of the extent of the problem, the difference equation is solved implicitly by relaxation methods. This technique forces N to tend to V, and instability does not vitiate this method unless round-off errors are introduced before the relaxation commences. On the other hand, if no boundary condition is given at one end of the extent, the difference equation is solved explicitly by step-by-step methods and round-off-errors introduced at any stage of the calculation may give rise to large errors in the final solution N. Consequently, stability is essential in a step by step technique.

LINEAR DIFFERENCE EQUATIONS WITH CONSTANT COEFFICIENTS

1. Stability Condition for Step-by-step Methods.

Let a linear differential equation with constant coefficients be replaced by a finite difference approximation of order p (i.e., one involving $p + 1$ tabular values). Then the n^{th} tabular entry is calculated

from

$$A_0 y_n + A_1 y_{n-1} + A_2 y_{n-2} + \cdots\cdots + A_p y_{n-p} = \phi_n, \qquad (20)$$

where $A_0, A_1, A_2, \ldots\ldots A_p$ are functions of the mesh length h, and ϕ_n is a known function of n and h. Now suppose the errors existing in the entries $y_{n-p}, y_{n-p+1}, \ldots\ldots, y_{n-1}$ are $\epsilon_{n-p}, \epsilon_{n-p+1}, \cdots, \epsilon_{n-1}$ respectively, then the consequent error in y_n is ϵ_n where

$$A_0 \epsilon_n + A_1 \epsilon_{n-1} + A_2 \epsilon_{n-2} + \cdots\cdots + A_p \epsilon_{n-p} = 0, \qquad (21)$$

provided ϕ_n requires no rounding-off. If solutions of (21) exist of the form

$$\epsilon_n = \lambda^n, \qquad (22)$$

then

$$A_0 \lambda^p + A_1 \lambda^{p-1} + A_2 \lambda^{p-2} + \cdots\cdots + A_p = 0, \qquad (23)$$

and so the general error given by (21) is

$$\epsilon_n = a_1 \lambda_1^n + a_2 \lambda_2^n + \cdots\cdots + a_p \lambda_p^n, \quad (n > p) \qquad (24)$$

where $a_1, a_2, \ldots\ldots a_p$ are constants and $\lambda_1, \lambda_2, \cdots, \lambda_p$ are the roots of (23). The condition for stability is that all the roots of (23) lie inside or on the unit circle. (3)

2. Replacement of a Differential Equation by a Higher Order D Difference Equation.

In a recent paper (4), Todd demonstrated the danger

of replacing a differential equation, for the purposes of step-by-step computation, by a higher order difference equation. Using (17), he replaced the second order differential equation

$$y'' + y = 0 \tag{25}$$

subject to the boundary conditions $y(0) = 0$, $y'(0) = 1$ by the fourth order difference equation

$$y_{n+2} - 16y_{n+1} + 30y_n - 16y_{n-1} + y_{n-2} = 12h^2 y_n , \quad (n \geq 2) \tag{26}$$

where y_n is written for $y(n\,h)$. The values of y_0, y_1, are given by the boundary conditions, y_2, y_3, by using a second order difference replacement,

$$y_{n+1} - 2y_n + y_{n-1} + h^2 y_n = 0, \tag{27}$$

obtained from (11), and y_{n+2} $(n \geq 2)$ by using (26). Todd expected the above step-by-step procedure using the fourth order replacement to give superior results to a step-by-step calculation using the second order replacement (27) at all nodes $n \geq 1$. This expectation was based on the orders of the terms neglected in the second and fourth order central difference replacements of y''.

The latter are given respectively by

$$y_n'' = \frac{1}{h^2}(y_{n+1} - 2y_n + y_{n-1}) - \frac{1}{12}h^2 y_n'''' \ \ldots\ldots$$

and

$$y_n'' = \frac{1}{12h^2}(-y_{n+2} + 16y_{n+1} - 30y_n + 16y_{n-1} - y_{n-2}) + \frac{1}{90}h^4 y_n'''' \ \ldots$$

The principal part of the truncation error is thus much smaller in the case of a fourth order replacement, and so the solution of the difference equation (V) should be correspondingly nearer the solution of the differential equation (D) for a prescribed mesh length h. This was not so, however, because of the instability inherent in the fourth order replacement. Equation (25) becomes

$$\lambda^4 - 16\lambda^3 + (30 - 12k^2)\lambda^2 - 16\lambda + 1 = 0,$$

and as h approaches zero the roots of this equation tend to 1, 1, $7 - \sqrt{48}$, and $7 + \sqrt{48}$. The last root quoted lies outside the unit circle and is responsible for the instability found by Todd. In the case of the second order replacement (27), equation (25) becomes

$$\lambda^2 - (2 - k^2)\lambda + 1 = 0,$$

the roots of which tend to 1, 1, as h approaches zero, and so the procedure is stable. The additional roots $7 \pm \sqrt{48}$ introduced by the fourth order replacement are called

"smuggled" (eingeschleppt) by Rutishauser (5). It should be noted that the larger "smuggled" root will take charge of the true (error free) difference solution in the same way as the error, and so the discrepancies in corresponding values in columns (1) and (4) of table II are due to divergence as well as to instability.

3. Stability of Step-by-step Methods based on Backward Differences. (Mitchell and Craggs (3)).

The use of a difference equation of higher order than the differential equation does not necessarily lead to instability. If, for example, equation (25) is replaced using (19) by the fourth order backward difference formula

$$35y_{n+2} - 104y_{n+1} + 114y_n - 56y_{n-1} + 11y_{n-2} + 12h^2y_{n+2} = 0, \quad (n \geq 2) \qquad (27)$$

equation (23) becomes

$$(35 + 12h^2)\lambda^4 - 104\lambda^3 + 114\lambda^2 - 56\lambda + 11 = 0, \qquad (28)$$

and as h approaches zero, the roots of this equation tend to 1, 1, $\sqrt{(11/35)}$, and $\sqrt{(11/35)}$. A step-by-step calculation based on (27) is thus stable since the roots of the auxiliary equation (28) all lie within or on the unit circle.

This is illustrated in table I which shows values of sin x to five places of decimals at decimal intervals. Column (2) was constructed using fourth order backward differences. Columns (1), (3), and (4) are respectively

the correct value of sin x, and the values computed by Todd using the second and fourth order central difference formulae.

TABLE 1

x	(1)	(2)	(3)	(4)
0	0•00000	0•00000	0•00000	0•00000
0•1	0·09983	0·09983	0•09983	0•09983
0·2	0·19867	0·19867	0·19866	0•19867
0•3	0•29552	0·29552	0·29550	0·29552
0·4	0·38942	0·38941	0·38939	0·38934
0·5	0·47945	0·47941	0·47939	0·47819
0·6	0·56464	0·56462	0·56460	0·54721
0·7	0·64422	0·64419	0•64416	0•40096
0·8	0·71736	0·71733	0·71728	-2·67357
0·9	0•78333	0·78331	0·78323	
1·0	0·84147	0·84147	0·84135	
1·1	0•89121	0·89122	0·89106	
1·2	0·93204	0·93206	0·93186	
1·3	0·96356	0·96358	0·96334	
1·4	0·98545	0·98546	0·98519	
1·5	0·99750	0·99748	0·99719	
1·6	0·99957	0·99954	0·99922	

It might be expected from the results just quoted that methods based on the use of backward difference formulae will always be stable. This is, however, not the case. Consider for example the differential equation

$$y' + y = 0, \tag{29}$$

Write

$$h y_n' = \nabla y_n + \tfrac{1}{2}\nabla^2 y_n + \tfrac{1}{3}\nabla^3 y_n + \cdots + \tfrac{1}{m}\nabla^m y_n ,$$

where $\nabla y_n = y_n - y_{n-1}$, $\nabla^2 y_n = y_n - 2y_{n-1} + y_{n-2}$ etc..

This leads to the auxiliary equation

$$h + \left(1 - \tfrac{1}{\lambda}\right) + \tfrac{1}{2}\left(1 - \tfrac{1}{\lambda}\right)^2 + \tfrac{1}{3}\left(1 - \tfrac{1}{\lambda}\right)^3 + \cdots + \tfrac{1}{m}\left(1 - \tfrac{1}{\lambda}\right)^m = 0. \tag{30}$$

It can be shown that for h = 0 the roots of (30) other than $\lambda = 1$ have modulus less than unity for $m \leqslant 6$. For m = 7, there is a pair of conjugate complex roots approximately equal to $\pm i$. For $m \geqslant 8$ there will always be at least one pair of conjugate roots of modulus greater than unity.

Table II demonstrates the instability of the twelfth order backward difference formula applied to equation (29). The values of e^{-x} at decimal intervals from 0 to 1·1 required to start the computations, were taken from five figure tables. The theoretical and computed values from 1·2 to 2·0 are given in rows (1) and (2) respectively in

table II. Again the differences in corresponding values in rows (1) and (2) are due to divergence as well as to instability.

TABLE II

x	1·2	1·3	1·4	1·5	1·6	1·7	1·8	1·9	2·0
(1)	·30199	·27253	·24660	·22313	·20190	·18268	·16530	·14957	·13534
(2)	·30148	·27541	·24708	·22244	·20337	·18665	·16017	·14168	·16065

Next consider Adams' method based on the formula

$$\frac{1}{h}(y_{n+1} - y_n) = y_n' + \tfrac{1}{2}\nabla y_n' + \tfrac{5}{12}\nabla^2 y_n' + \tfrac{3}{8}\nabla^3 y_n' + \tfrac{251}{720}\nabla^4 y_n' \cdots \quad (30)$$

where a sufficient number of starting values for y, y' is supposed computed by an independent method. (e.g. by Taylor series). Consider again the first order equation (29). Adams' formula leads to the auxiliary equation

$$F(\lambda) \equiv \lambda - 1 + h\left\{1 + \tfrac{1}{2}\left(1 - \tfrac{1}{\lambda}\right) + \tfrac{5}{12}\left(1 - \tfrac{1}{\lambda}\right)^2 + \tfrac{3}{8}\left(1 - \tfrac{1}{\lambda}\right)^3 + \tfrac{251}{720}\left(1 - \tfrac{1}{\lambda}\right)^4\right\} = 0, \quad (31)$$

where only fourth differences are retained and it is clear that there is stability as h tends to zero. Now from (31), $F(-\infty) < 0$, and $F(-1) = -2 + 551h/45$. There is therefore a root of modulus greater than unity when h exceeds 90/551, and the method is stable only for sufficiently small tabular interval. Moreover if higher order differences are retained, the maximum value of h for which the method is stable is decreased.

Similar arguments show that Moulton's method based on the formula

$$\frac{1}{h}(y_n - y_{n-1}) = y_n' - \frac{1}{2}\nabla y_n' - \frac{1}{12}\nabla^2 y_n' - \frac{1}{24}\nabla^3 y_n' - \frac{19}{720}\nabla^4 y_n' \cdots$$

is also unstable for large values of the tabular interval when differences higher than the first are retained. The upper limit on h for stability for a given number of differences is very much higher than in Adams' method.

4. Application of Relaxation Methods to the Solution of "Marching" Problems.

A "marching" problem is one in which the numerical solution is usually constructed by a step-by-step integration process. Allen and Severn (6) considered the solution of the first order differential equation

$$y' - y = x, \tag{32}$$

over the range $0 \leq x \leq 1$ with the end condition $y = 1$ at $x = 0$. This is a "marching" problem, and the lack of any given end condition at $x = 1$ apparently seemed to the above authors an insuperable difficulty to the direct use of the relaxation technique. Accordingly by doubling the order of the differential equation by the substitution

$$y = Y' + Y, \tag{33}$$

and imposing an arbitary boundary condition $Y = 0$ at $x = 1$, they were able to solve the second order equation

$$Y'' - Y = x,$$

by relaxation. Dividing the range of x into ten equal intervals ($h = 0{\cdot}1$), the values of Y at the node points (o, 1, 2, ... 9, 10) are given by the matrix equation

$$\begin{bmatrix} .905 & -1 & & & & & & & & \\ 1 & -2{\cdot}01 & 1 & & & & & & & \\ & 1 & -2{\cdot}01 & 1 & & & & & & \\ & & \ddots & \ddots & \ddots & & & & & \\ & & & & & & 1 & -2{\cdot}01 & 1 & \\ & & & & & & & 1 & -2{\cdot}01 & \end{bmatrix} \begin{bmatrix} Y_0 \\ Y_1 \\ Y_2 \\ \\ \\ Y_8 \\ Y_9 \end{bmatrix} = \begin{bmatrix} -{\cdot}1 \\ {\cdot}001 \\ {\cdot}002 \\ \\ \\ {\cdot}008 \\ {\cdot}009 \end{bmatrix} \qquad (34)$$

together with $Y_{10} = 0$ at $x = 1$. The values of y at the node points are then obtained from

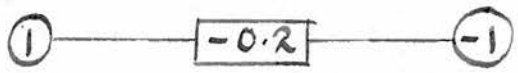
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_8 \\ y_9 \\ y_{10}-.05 \end{bmatrix} = \begin{bmatrix} -5 & 1 & 5 & & & & & & \\ & -5 & 1 & 5 & & & & & \\ & & -5 & 1 & 5 & & & & \\ & & & \ddots & \ddots & \ddots & & & \\ & & & & & & -5 & 1 & 5 \\ & & & & & & & -5 & 1 \\ & & & & & & & & -10 \end{bmatrix} \begin{bmatrix} Y_0 \\ Y_1 \\ Y_2 \\ \vdots \\ Y_7 \\ Y_8 \\ Y_9 \end{bmatrix} \tag{35}$$

together with $y_0 = 1$ at $x = 0$.

As pointed out by Mitchell (7), Allen and Severn need not have doubled the order of the differential equation in order to use relaxation in the present problem. Using equation (5) at $x = 1$ and (4) at all other nodes, the matrix equation becomes

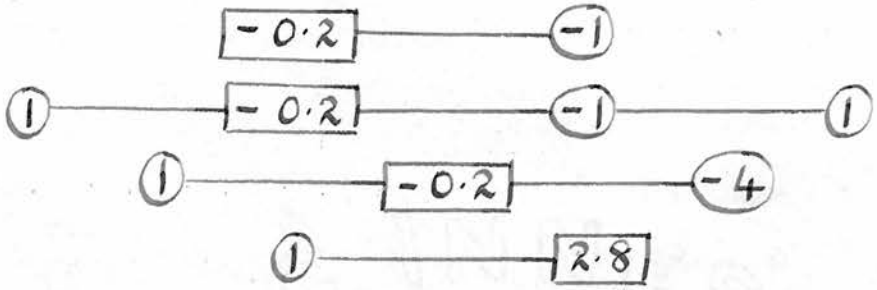
$$\begin{bmatrix} -.2 & 1 & & & & & & & & \\ -1 & -.2 & 1 & & & & & & & \\ & -1 & -.2 & 1 & & & & & & \\ & & \ddots & \ddots & \ddots & & & & & \\ & & & & & & & & & \\ & & & & & -1 & -.2 & 1 & & \\ & & & & & & -1 & -.2 & 1 & \\ & & & & & & & 1 & -4 & 2.8 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_8 \\ y_9 \\ y_{10} \end{bmatrix} = \begin{bmatrix} 1.02 \\ .04 \\ .06 \\ \vdots \\ .16 \\ .18 \\ .20 \end{bmatrix}, \tag{56}$$

from which the values of y can be obtained directly by relaxation. It is not suggested that in any problem the matrix equation be written out in full. This is particularly the case in a two dimensional problem. Sufficient elements should, however, be sketched in to indicate the structure of the matrix. The relaxation pattern at any node can then easily be read off from the columns of the matrix. For example in the present problem the pattern for the nodes at x = 0·2, 0·3, 0·4, 0·5, 0·6, 0·7 would be



in each case, and the patterns for the nodes at x = 0·1,

0·8, 0·9, 1·0 would be respectively



where the rectangle denotes the node which is relaxed.

The method of Allen and Severn is of course a feasible one and in certain cases it may happen that it requires less laborious calculation than the direct method. It should, however, be remembered that it involves an additional approximation, the finite difference approximation for (33), by means of which the values of y are obtained from those of Y. It will be observed that the diagonal elements do not dominate the matrix in equation (36). Accordingly the relaxation was carried out according to a scheme devised by Rutherford (7) details of which will be given later.

In order to provide a basis for comparison the problem described by Allen and Severn has been worked out using a direct step-by-step method. The difference approximation to (32) used is

$$\frac{1}{h}(y_{n+1} - y_n) = \frac{1}{2}(y_{n+1} + y_n + x_{n+1} + x_n). \qquad (37)$$

With h = 0·1, this becomes

$$y_{n+1} = \frac{21}{19} y_n + \frac{1}{19} (x_n + x_{n+1}), \qquad (37)$$

$$(n = 0, 1, 2, \ldots 9)$$

which together with the end condition $y_0 = 1$ gives the values $y_1, y_2, \ldots y_{10}$ directly in turn. It should be realized that the difference equation (37) is actually unstable, the only root of the auxiliary equation being $\lambda = 21/19$. Since y_0 requires no rounding-off, round-off errors are introduced by the coefficients 21/19 and 1/19 ($h = 0{\cdot}1$). The effect of these errors appears to be negligible in the restricted range of this calculation.

The values of y as calculated by the different methods at the nodes are shown in table III together with the theoretical values. The calculations are all accurate to three decimal places. There seems little doubt that in this "marching" problem the use of relaxation, as described by Allen and Severn, does not prevent the solution for y having cumulative errors. Imposing an arbitrary condition on Y at x = 1 has failed to "tether" the value of y at x = 1. (compare y_{10} from Allen and Severn for $h = 0{\cdot}1, 0{\cdot}2$). In view of the greater ease of calculation and in this case greater accuracy of final solution, step-by-step methods appear to be superior to relaxation methods in dealing with "marching" problems

of this type.

TABLE III

	Relaxation			Step-by-Step	Theoretical
	Allen and Severn ($Y_{10} = 0$)		Mitchell		
	h = 0·1	h = 0·2	h = 0·1	h = 0·1	
y_0	1·000	1·000	1·000	1·000	1·000
y_1	1·109		1·108	1·110	1·110
y_2	1·241	1·236	1·241	1·243	1·243
y_3	1·397		1·396	1·400	1·400
y_4	1·579	1·575	1·580	1·585	1·584
y_5	1·792		1·791	1·799	1·797
y_6	2·037	2·031	2·038	2·046	2·044
y_7	2·319		2·318	2·330	2·328
y_8	2·641	2·629	2·642	2·654	2·651
y_9	3·007		3·006	3·023	3·019
y_{10}	3·422	3·405	3·423	3·441	3·436

In a recent communication (8), Fox makes a further contribution to the relaxation - step-by-step controversary in solving "marching" type problems. He considers the first order linear differential equation

$$y' - 12y + 11e^x = 0 \qquad (38)$$

subject to the condition $y = 1$ at $x = 0$. Differentiating (38) and eliminating y' by using (38), the second order equation

$$y'' - 144y + 143e^x = 0 \qquad (39)$$

is obtained. Fox considers the range $0 \leqq x \leqq 1$, and with $h = 0.2$, the nodes are numbered 0, 1, ... 5. At a general node r, equation (39) is replaced by

$$(1-12h^2)y_{r+1}-(2+120h^2)y_r+(1-12h^2)y_{r-1}=-\frac{143}{12}h^2(10e^{x_r}+e^{x_{r+1}}+e^{x_{r-1}}) \qquad (40)$$

and equation (38) by

$$(1-4h)h_{r+1}-16hy_r-(1+4h)y_{r-1}=-\frac{11}{3}h(4e^{x_r}+e^{x_{r+1}}+e^{x_{r-1}}) \qquad (41)$$

Equation (40) is applied at nodes 1, 2, 3, 4, together with (41) at node 1 in a step-by-step solution, and together with (41) at node 4 in a relaxation solution. The values of y given by Fox correct to four places of decimals are

	Step-by-Step	Relaxation	Theoretical
y_0	1·0000	1·0000	1·0000
y_1	1·2214	1·2214	1·2214
y_2	1·4918	1·4918	1·4918
y_3	1·8218	1·8220	1·8221
y_4	2·2214	2·2255	2·2255
y_5	2·6644	2·7184	2·7183

This example is used by Fox to illustrate the statement that "in certain cases relaxation will give better precision in the computed pivotal values" than step-by-step methods. The present author (9) now extends this remark considerably by showing that there is a large class of initial value problems, involving ordinary differential equations of all orders, which can only be solved satisfactorily by relaxation.

The error equations corresponding to (38) and (39) are

$$\epsilon' = 12\epsilon$$

and

$$\epsilon'' = 144\epsilon$$

respectively, with solutions

$$\epsilon(x) = A e^{12x}$$

and

$$\epsilon(x) = A_1 e^{12x} + A_2 e^{-12x}$$

where ϵ is the error in y. The factor e^{12x} will cause the

the error to increase rapidly with x. In a step-by-step calculation using a difference replacement of (38) or (39), the error will grow according to the factor e^{12rh}, and so a round-off error, introduced at any stage of the calculation, will grow alarmingly. For example, the roots of the auxiliary equation (see p. 6) corresponding to the difference equation (40) are

$$\lambda = 1 \pm 12h + 72h^2 \ldots\ldots,$$

and so the error equation corresponding to (40) has solution

$$\epsilon_r = a_1(1 + 12h + 72h^2\ldots)^r + a_2(1-12h + 72h^2 \ldots.)^r$$

$$= a_1 e^{12rh} + a_2 e^{-12rh}.$$

To illustrate the instability of a step-by-step calculation using (40) together with (41) at node 1, the present author has worked out Fox's problem as far as x = 1·4 (r=7 when h = 0·2). For comparison the same problem has been worked out using relaxation as described by Fox ((40) together with (41) at node 6), and using relaxation as described by Allen and Severn (p. 13). It should be pointedout that the outline of the method of relaxation used by Fox was also described by Allen and Severn (6), although no calculations using this method were carried out by these authors. In all calculations, the working is correct to three places of decimals, and the results are given in table IV.

TABLE IV (h = 0·2)

	Relaxation		Step-by-Step	Theoretical
	Fox	Allen and Severn ($Y_7 = 0$)		
y_0	1·000	1·000	1·000	1·000
y_1	1·221	1·222	1·221	1·221
y_2	1·492	1·493	1·490	1·492
y_3	1·822	1·821	1·798	1·822
y_4	2·226	2·205	1·914	2·226
y_5	2·718	2·554	-1·292	2·718
y_6	3·320	2·065	-49·395	3·320
y_7	4·055	-1·636	-681·237	4·055

The most surprising feature of the table is the apparent instability of Allen and Severn's method of relaxation. Accordingly this method of relaxation will now be studied rather more carefully.

In the present problem, using the method of Allen and Severn, the order of equation (38) is doubled by using the substitution.

$$y = Y' + 12Y, \tag{42}$$

resulting in the second order equation

$$Y'' - 144Y + 11e^{x} = 0, \qquad (43)$$

subject to the boundary condition $Y' + 12Y = 1$ at $x = 0$. In order to solve a second order difference replacement of (43) using relaxation, a boundary condition on Y is necessary at $x = 1{\cdot}4$. Allen and Severn pointed out that the theoretical solution y of equation (38) is independent of Y_7, and so Y_7 can be chosen quite arbitrarily. In numerical calculations, however, it does not follow that the final values of y will be completely independent of Y_7, and so in the following calculations Y_7 is retained as a general quantity to which an arbitrary value can be assigned at any time. Equation (43) is now replaced by

$$Y_{r+1} + Y_{r-1} - (2 + 144h^2)Y_r + 11h^2e^{x_r} = 0$$

at nodes 1, 2, ... 6, and the boundary conditions by

$$Y_1 - (1 - 12h + 72h^2)Y_0 - (h - \tfrac{11}{2}h^2) = 0$$

and

$$Y_7 = Y_7$$

at nodes 0 and 7 respectively. A relaxation is carried out according to the above scheme with $h = 0{\cdot}2$ and the following values of Y obtained

$$Y_0 = {\cdot}00000\, Y_7 + {\cdot}0770$$

$$Y_1 = \cdot00001\, Y_7 + \cdot0940$$

$$Y_2 = \cdot00004\, Y_7 + \cdot1148$$

$$Y_3 = \cdot00029\, Y_7 + \cdot1401$$

$$Y_4 = \cdot00225\, Y_7 + \cdot1705$$

$$Y_5 = \cdot01718\, Y_7 + \cdot2037$$

$$Y_6 = \cdot13108\, Y_7 + \cdot2145$$

$$Y_7 = 1\cdot00000\, Y_7 + \cdot0000 .$$

These values give zero residuals at all nodes to four places of decimals in the coefficient of Y_7 and to three places of decimals in the term not involving Y_7. Equation (42) is now replaced by

$$y_r = (Y_{r+1} - Y_{r-1}) + 12Y_r$$

at nodes r = 1, 2, ... 6, and by

$$y_7 = \frac{1}{24}\left[Y_5 - 4Y_6 + 3(1+8h)\, Y_7\right]$$

at node 7. Putting h = 0·2, the following values of y are obtained

$$y_0 = 1\cdot0000$$

$$y_1 = 1\cdot2225 + \cdot0002\ Y_7$$

$$y_2 = 1\cdot4928 + \cdot0012\ Y_7$$

$$y_3 = 1{\cdot}8205 + {\cdot}0090 \quad Y_7$$

$$y_4 = 2{\cdot}2050 + {\cdot}0693 \quad Y_7$$

$$y_5 = 2{\cdot}5556 + {\cdot}5284 \quad Y_7$$

$$y_6 = 2{\cdot}0647 + 4{\cdot}0302 \quad Y_7$$

$$y_7 = -1{\cdot}6358 + 18{\cdot}2321 \quad Y_7$$

The values quoted in table IV are the above with $Y_7 = 0$. Two points arise from the values of $y_0 \ldots\ldots y_7$;

(1) the values of y are not independent of Y_7, and

(2) the method is unstable, although a fortunate choice of Y_7 may prevent serious instability. (e.g. $Y_7 = 0{\cdot}312$ will give values of y close to the theoretical values).

It is the arbitrary nature of Y_7 in Allen and Severn's method which prevents it being a true method of relaxation in y. A true relaxation method requires a condition on y at the node r = 7. Examples of true relaxation methods for solving first order "marching" problems are given by Mitchell and Rutherford (7) and Fox (8), although the method of the former authors, as pointed out by Fox, is only accurate to two decimal places. Allen and Severn's method of course will give reasonable results in problems where the error growth, governed by the original differential

equation, is small. This is the case in the problem governed by equation (32) where an error grows like e^x. Over the limited range of the problem $0 \leqslant x \leqslant 1$, this will not cause any appreciable error in the numerical calculations.

So far only first order linear differential equations have been discussed. Now consider the second order equation

$$y'' - 11y' - 12y + 22e^x = 0 \qquad (44)$$

subject to the boundary conditions $y = y' = 1$ at $x = 0$. A numerical solution is required in the range $0 \leqslant x \leqslant 0{\cdot}8$ and once again the problem is of "marching" type. The simplest difference approximation to (44) is

$$\frac{1}{h^2}(y_{r+1}-2y_r+y_{r-1}) - \frac{11}{2h}(y_{r+1}-y_{r-1}) - 12y_r+22e^{x_r} = 0,$$

which for $h = 0{\cdot}1$ becomes

$$0{\cdot}45y_{r+1} = 2{\cdot}12y_r - 1{\cdot}55y_{r-1} - 0{\cdot}22e^{x_r}. \qquad (45)$$

This formula is applicable at nodes $r = 1, 2, \ldots 7$. It is not claimed that (45) is the most accurate three point difference replacement of (44), but that it is a simple formula of reasonable accuracy. To start a step-by-step calculation using (45), suppose for convenience that y_0 and y_1 are known from the boundary conditions as

$y_0 = 1$, $y_1 = e^{0.1}$. The values of y_2 ..... y_8, correct to three places of decimals, are then obtained in turn from (45), and compared with the theoretical values as follows;

x	0·0	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8
y	1·000	1·105	1·222	1·353	1·506	1·703	2·036	2·808	5·254
e^x	1·000	1·105	1·221	1·350	1·492	1·649	1·822	2·014	2·226

From these figures, the step-by-step method of solving this problem appears to be unstable. This instability is again explained by considering the error equation

$$\epsilon'' - 11\epsilon' - 12\epsilon = 0,$$

corresponding to (44), with solution

$$\epsilon(x) = A_1 e^{12x} + A_2 e^{-x},$$

and the error equation corresponding to (45) for general h, with solution

$$\epsilon_r = a_1(1 + 12h + 72h^2 \ldots)^r + a_2 (1 - h + \tfrac{1}{2}h^2 \ldots..)^r$$

$$= a_1 e^{12rh} + a_2 e^{-rh}.$$

The factor e^{12rh} governing part of the error growth is again responsible for the instability of the step-by-step solution.

A relaxation solution of (44) subject to the given boundary conditions can be obtained without increasing the order of the differential equation. This is accomplished

by using a third order difference replacement of (44), which allows a boundary condition to be imposed on y at x = 0·8. The boundary condition of course must not violate the conditions of the problem. The difference replacement used is

$$\frac{1}{h^2}\left(y_{r+1}-2y_r+y_{r-1}\right) - \frac{11}{6h}\left(2y_{r+1}-3y_r-6y_{r-1}+y_{r-2}\right) - 12y_r + 22e^{x_r} = 0,$$

which for h = 0·1 reduces to

$$\frac{1\cdot 9}{5}y_{r+1} - 2\cdot 67y_r + 2\cdot 1y_{r-1} - \frac{1\cdot 1}{6}y_{r-2} + \cdot 22e^{x_r} = 0. \qquad (46)$$

This formula holds at nodes r = 2, 3,..... 7. As in the step-by-step calculation, y_0 and y_1 are supposed known from the boundary conditions $y = y' = 1$ at $x = 0$ as $y_0 = 1$, $y_1 = e^{0\cdot 1}$. This is merely for convenience as the boundary condition $y' = 1$ at $x = 0$ should be replaced by a forward difference formula involving two or more nodes, depending on the accuracy required. The extra "boundary" condition imposed at x = 0·8 is a backward difference replacement of (44). A simple formula of reasonable accuracy is

$$\frac{1}{h^2}\left(y_8-2y_7+y_6\right)-\frac{11}{2h}\left(3y_8-4y_7+y_6\right) - 12y_8 + 22e^{x_8} = 0,$$

which for h = 0·1 becomes

$$0\cdot 77y_8 - 0\cdot 20y_7 - 0\cdot 45y_6 - 0\cdot 22e^{0\cdot 8} = 0. \qquad (47)$$

With y known at r = 0, 1, equation (46) at nodes r = 2, 3,...7

and equation (47) at $r = 8$ constitute seven equations in seven unknowns. Sufficient to say that the relaxation solution obtained, which gives zero residuals to three places of decimals, is within 0·001 of the theoretical solution at all nodes. A relaxation solution of this problem can also be obtained by increasing the order of the differential equation (44), replacing the higher order equation by a difference equation and adding the appropriate number of "boundary" conditions at $x = 0{\cdot}8$ in the manner of Fox. This will certainly require more labour, but may well give greater accuracy. At present, however, the author is only concerned with establishing the fact that relaxation methods are essential in dealing with "marching" problems of this type.

Without considering differential equations of order higher than the second, the following comments can be made about the methods of numerical solution of initial value (or "marching") problems involving differential equations of any order.

(1) If a stable finite difference approximation to the differential equation can be found, step-by-step methods of solution should be used over any range of x.

(2) If a difference approximation is mildly unstable, step-by-step methods of solution may be used over a small range of x. This is the case in the solution of the first order equation $y' - y = 0$ subject to the boundary condition $y = 1$

at $x = 0$ over the range $0 \leq x \leq 1$. This is actually the problem solved by Allen and Severn to illustrate their method of relaxation.

(3) If all difference replacements of the differential equation are unstable, relaxation as described by Fox (increasing the order of the differential equation) or as described by the present author (replacing the differential equation by a higher order difference equation) must be used.

(4) Relaxation as described by Allen and Severn (using a higher order differential equation in a new dependent variable) can only be used in problems where step-by-step methods are possible.

5. A Theory of Relaxation. (Suitable when the Diagonal Elements do not dominate the Matrix).

Equations (34) and (36) can be written in the form

$$Av = h \qquad (38)$$

with obvious notation. Since the matrix A is not triangular, an approximate numerical solution N of (38) can be found by relaxation methods. The following is an account of a theory of relaxation (Rutherford (7)), specially suitable when the diagonal elements do not dominate the matrix A.

Denoting successive approximations by upper suffices, let v^1 be a trial solution and put

$$h - Av^1 = c^1$$

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

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

tend monotonically to zero, then the vectors

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

tend to the solution v of $Av = h$. For clearly,

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

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

$$y_r{}^i = \delta_{rj} k_j{}^i \ldots \qquad (39)$$

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

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

Denoting the length of the vector c^i by $|c^i|$, it is found that

$$|c^{i+1}|^2 = \sum_r (c_r^i - a_{rj}k_j^i)^2$$

$$= |c^i|^2 - 2k_j^i \sum_r c_r^i a_{rj} + (k_j^i)^2 \sum_r (a_{rj})^2.$$

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

$$2k_j^i \sum_r c_r^i a_{rj} - (k_j^i)^2 \sum_r (a_{rj})^2 > 0 \; ; \tag{40}$$

that is to say, if

$$0 \leqslant k_j^i \leqslant 2 \sum_r c_r^i a_{rj} \Big/ \sum_r (a_{rj})^2. \tag{41}$$

The left-hand side of (40) has its maximum when

$$k_j^i = \sum_r c_r^i a_{rj} \Big/ \sum_r (a_{rj})^2 \tag{42}$$

and this maximum has the value

$$\left(\sum_r c_r^i a_{rj}\right)^2 \Big/ \sum_r (a_{rj})^2 \tag{43}$$

which may be written

$$(k_j^i)^2 \sum_r (a_{rj})^2, \tag{44}$$

where k_j^i now has the value determined by (42). Thus, by giving k_j^i any value satisfying (41) it is ensured that $|c^{i+1}| < |c^i|$. To ensure the maximum reduction in the length of the residual vector, k_j^i must be given the value stated in (42). Since in most relaxation problems no column of A has more than a few non-zero elements, the calculation of $\sum_r c_r^i a_{rj}$ and of $\sum_r (a_{rj})^2$ will not be difficult. The labour,

however, would be considerable if this method were employed in the case of a matrix A of large order possessing very few vanishing elements.

If it happens that $\sum_r c_r^i a_{rj} = 0$, then y^i is the zero vector and no reduction in c^i is possible by means of a vector y^i by (39). At any stage, however, there are n possible choices of j and at least one of these will yield a vector y^i which reduces the length of the residual vector, unless for all j

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

Assuming, however, that A is non-singular, these questions imply that c^i is the zero vector and that the exact solution $v = v^i$ has already been reached. It follows that the length of the residual vector can always be reduced by the foregoing method provided the exact solution has not already been reached. The value of j which gives the greatest reduction at the ith step will also give (43) the largest value. In practice however it will be sufficient to choose j so that (42) has a large, but not necessarily the largest, value. Even this provision may be dispensed with if each j is chosen in turn according to some plan.

The above method for obtaining N resembles, and is related to, that systematised by Temple (10) and, more recently, by Stiefel (11). The original relaxation method of Southwell was based upon the analogy of an elastic

framework subjected to prescribed loads at the joints. At each stage in the approximation the displacement at one joint is modified in such a way that the potential energy of the system is diminished. Temple showed that this procedure is equivalent to altering one component of the vector v at a time in such a way as to reduce the value of $v'(\frac{1}{2}Av - h)$, v' being the transpose of the vector v. This method can, in fact, be applied to any system of equations $Av - h = 0$, provided the matrix A is symmetric and positive definite. If A does not satisfy these conditions, Temple's method can still be applied to the equivalent equations $A'Av - A'h = 0$, which have been "prepared" by premultiplying the original equations by the transposed matrix A'. The modifications are then made with a view to minimising the expression $v'A(\frac{1}{2}Av - h)$. In contrast, the method which has been described minimises the expression $(v'A' - h')(Av-h)$, which is the square of the length of the residual vector. Since $(v'A' - h')(Av - h) = 2v'A'(\frac{1}{2}Av - h) + h'h$, the method here advocated must be equivalent to Temple's method applied to the prepared equation in the sense that the modification of v at any stage will be the same. The intermediate numerical calculations, however, will be different in the two cases and the present method has the advantage that it can be applied directly without any preparation, whenever the matrix A is non-singular.

Furthermore, A is usually simpler than the prepared matrix A'A in the sense that the former is likely to have more zero elements than the latter.

In many relaxation problems the diagonal elements of the matrix A dominate the others in the sense that a diagonal element has a much larger absolute value that the other elements in its column. If this is the case then c_j^1/a_{jj} will be a good approximation to the optimum value of k_j given by (42). This, indeed, coincides with the value suggested by Fox (12) and by Temple. Fox also suggests selecting the value of j which makes c_j^1/a_{jj} a maximum; but it would appear from (44) that a better guess at the best j in the case under consideration would be that which makes $|c_j^1|$ itself a maximum, as originally suggested by Southwell (1, p.47 and p. 65).

If, however, the diagonal elements of A are of smaller magnitudes than some other neighbouring ones, the choice $k_j^1 = c_j^1/a_{jj}$ may be a bad one for the method, as may also both Fox's and Southwell's suggestions for the choice of j. In such cases it would be wise to use the formulae (42) and (43). These formulae were in fact used in solving the relaxation problem given by equation (56) where the diagonal elements do not dominate the matrix.

6. Round-off Errors

The discrepancy between the numerical solution N and the theoretical solution V of a difference equation may be

considered as entirely due to round-off errors. Because of the many different places (coefficients, pivotal values, etc), where rounding-off may be necessary in a calculation, no general expressions for round-off errors can be obtained. Instead the following particular examples will be used to illustrate the errors dur to different types of rounding in both relaxation (Mitchell (13)) and step-by-step solutions.

Consider initially the ordinary differential equation

$$y'' + ky = f, \tag{45}$$

where f(x) is a known function. When k = 0, (45) reduces to Poisson's equation in one dimension. The boundary conditions consist of a knowledge of y at x = 0, L. The range $0 \leqq x \leqq L$ is divided equally by N internal points, and for convenience suppose that N is odd.

The conventional finite difference approximation to (45) is

$$\frac{1}{h^2}(y_{n-1} - 2y_n + y_{n-1}) + ky_n = f.$$

If the residuals at any stage of the relaxation solution using this difference equation are R_n (n = 1, 2, ... N), then the errors ϵ_n in y_n satisfy the N equations

$$\epsilon_{n+1} - \left\{2 - k\frac{L^2}{(N+1)^2}\right\}\epsilon_n + \epsilon_{n-1} = R_n, \tag{46}$$

$$(n = 1, 2, \cdots N)$$

Assuming for convenience $\epsilon_0 = \epsilon_{N+1} = 0$, the N error equations can be written as the single matrix equation

$$A\underline{\epsilon} = R, \tag{47}$$

where $\underline{\epsilon}$ and R are column vectors and A is a square matrix of order N. At any stage of the relaxations process, the residuals lie in a range $-R \leqslant R_n \leqslant +R$, where R decreases as the accuracy of the calculation increases.

Now from (47) it follows that

$$\underline{\epsilon} = A^{-1} R, \tag{48}$$

where from Rutherford (14) the p, q^{th} element $\nu_{p,q}$ of A^{-1} is

$$\nu_{p,q} = \nu_{q,p} = (-1)^{p+q} \frac{\phi_{p-1}\, \phi_{N-q}}{\phi_N} \qquad (p \leq q)$$

where

$$\phi_t = \frac{\sin(t+1)\,\theta}{\sin\theta} \qquad (t = 0, 1, 2, \ldots N)$$

and

$$\theta = \cos^{-1}\left[-\tfrac{1}{2}\left\{2 - k\,\frac{L^2}{(N+1)^2}\right\}\right].$$

Thus the round-off error at an internal node is given by

$$\epsilon_n = \frac{\sin n\theta \sin (N+1-n)\theta}{\sin\theta \sin(N+1)\theta}.$$

$$\left[\sum_{r=1}^{n-1} (-1)^{r+n} \frac{\sin r\theta}{\sin n\theta} R_r + R_n + \sum_{r=n+1}^{N} \frac{\sin (N+1-r)\theta}{\sin (N+1-n)\theta} R_r\right]. \qquad (49)$$

$$(n = 1, 2, \ldots N)$$

The worst possible round-off errors occur when

$$R_n = R, \qquad (n = 1, 2, \ldots N)$$

Making this simplification in (49) the errors become

$$\epsilon_n = \frac{\cos \frac{n}{2}\theta \cos \frac{N+1-n}{2}\theta}{2\cos^2 \tfrac{1}{2}\theta \cos \frac{N+1}{2}\theta} R \qquad (50a)$$

if n is odd, and

$$\epsilon_n = \frac{\sin \frac{n}{2}\theta \sin \frac{N+1-n}{2}\theta}{2\cos^2 \tfrac{1}{2}\theta \cos \frac{N+1}{2}\theta} R \qquad (50b)$$

if n is even.

A residual distribution much more likely to be obtained in practice is the rectangular distribution $-R \leq R_n \leq +R$. From (49) the maximum standard mean deviation in the error which occurs at the middle node is given by

$$\sigma^2_{\frac{N+1}{2}} = \frac{(N+1) - \sin(N+1)\theta \cot\theta}{24 \sin^2\theta \cos^2 \frac{N+1}{2}\theta}. \qquad (51)$$

When $k = 0$, $\theta = \pi$ and (49), (50) and (51) become

$$\epsilon_n = -\frac{n(N+1-n)}{(N+1)}\left[\sum_{r=1}^{n-1}\frac{r}{n}R_r + R_n + \sum_{r=n+1}^{N}\frac{N+1-r}{N+1-n}R_r\right], \quad (52)$$

$$\epsilon_n = -\tfrac{1}{2}n(N+1-n)R, \quad (53)$$

and

$$\sigma^2_{\frac{N+1}{2}} = \frac{(N+1)(N^2+2N+3)}{144} \quad (54)$$

respectively. When $k<0$, θ is imaginary and from (50) ϵ_n is bounded above as N approaches infinity. When $k>0$, θ is real and from (50) ϵ_n becomes infinitely large when $(N+1)\theta = \pi$.

Consider next equation (45) together with a knowledge of y and y' at $x=0$. This time the finite difference approximation is used to obtain a step-by-step solution, and the appropriate error equation is

$$\epsilon_{n+1} = (2-kh^2)\epsilon_n - \epsilon_{n-1} \qquad (n \geq 1) \quad (55)$$

where ϵ_0 and ϵ_1 are known. The solution of (55) is

$$\epsilon_n = a_1\lambda_1^n + a_2\lambda_2^n, \quad (56)$$

where

$$a_1 = \frac{\epsilon_1 - \lambda_2\epsilon_0}{\lambda_1-\lambda_2} \text{ and } a_2 = \frac{\lambda_1\epsilon_0 - \epsilon_1}{\lambda_1-\lambda_2}, \text{ and}$$

λ_1, λ_2 are the roots of

$$\lambda^2 - (2 - kh^2)\lambda + 1 = 0.$$

This time, although ϵ_0 may be zero, ϵ_1 will almost always have a value other than zero, and so round-off-errors will develop according to (56). These errors will of course be serious if the modulus of either λ_1 or λ_2 exceeds unity.

In the step-by-step solution described, rounding-off has been considered only at nodes where the value of y has been given by the boundary conditions. The growth of error due to rounding off any pivotal value is of course given by (56) with the appropriate values of a_1 and a_2. In such cases, expressions for the general error are relatively easily obtained. This is not so, however, when rounding errors are introduced into the coefficients of the difference equation. As a simple example of this, consider a step-by-step solution of the differential equation $y' - y = 0$ subject to the boundary condition $y = 1$ at $x = 0$. The difference equation used is (37),

$$y_{n+1} = (21/19)y_n = (1.10526 \ldots\ldots)\, y_n. \qquad (n \geq 0)$$

The coefficient of y_n is expressed as $(C+E)$ where E is the rounding error. It is easily shown that the error ϵ_n in y_n is given by

$$\epsilon_n = (C+E)^n - C^n, \qquad (n \geq 0)$$

provided the calculated value of y is not rounded off at any stage. Next consider the equation $y'' - y = 0$ subject to the boundary conditions of given y and y' at $x = 0$. Using the backward difference replacement (12) for y'', the difference equation becomes

$$y_n = \frac{2}{1-h^2} y_{n-1} - \frac{1}{1-h^2} y_{n-2} \qquad (n \geq 2)$$

$$= 2(C+E) y_{n-1} - (C+E) y_{n-2} \qquad (57)$$

where again E is the rounding error in the coefficient. The solution of (57) is

$$y_n = a_1 \lambda_1^n + a_2 \lambda_2^n \qquad (n \geq 2)$$

where λ_1, λ_2 are the roots of

$$\lambda^2 - 2(C+E)\lambda + (C+E) = 0,$$

and a_1, a_2 are obtained from the boundary conditions. The error ϵ_n in y_n due to rounding the coefficients is given by

$$\epsilon_n = a_1 (\lambda_1^n - \mu_1^n) + a_2 (\lambda_2^n - \mu_2^n) \qquad (n \geq 2)$$

where μ_1, μ_2 are the roots of

$$\lambda^2 - 2C\lambda + C = 0,$$

and again the calculated value of y is not rounded off at any stage. This procedure for finding the error when the constant coefficients are rounded in a second order linear

difference equation can be extended in an obvious manner to a linear difference equation of any order.

Finally, if in a calculation involving a difference equation, rounding-off is necessary in the values of y at each stage as well as in the values of the coefficients, the estimation of the round-off error becomes rather troublesome, and only a rough estimate of the overall error can be obtained by combining the appropriate methods of this section.

LINEAR DIFFERENCE EQUATIONS WITH VARIABLE COEFFICIENTS.

Consider the linear difference equation of the first order

$$y_{n+1} = A(n)y_n + f(n), \qquad (n \geq 0) \qquad (58)$$

where $A(n)$ and $f(n)$ are known functions involving the mesh length h and y_0 is given.

If $f(n)$ and $A(n)$ require no rounding, the error equation is

$$\epsilon_{n+1} = A(n)\epsilon_n, \qquad (n \geq 0)$$

which has solution

$$\epsilon_n = A(n-1)\,A(n-2)\,\ldots\,A(0)\epsilon_0$$

$$= \left[\prod_{k=0}^{n-1} A(k)\right]\epsilon_0. \qquad (n \geq 1) \qquad (59)$$

The error at any stage of a step-by-step calculation using

(58) is thus given by (59). If f(n) is zero and A(n) requires rounding, say $A(n) = C(n) + E(n)$ $(n \geq 0)$ where E(n) is the rounding error, the error at any stage is given by

$$\epsilon_n = \left[\prod_{k=0}^{n-1} \{ C(k) + E(k) \} - \prod_{k=0}^{n-1} \{ C(k) \} \right] y_0 \quad (n \geq 1)$$

No general solutions exist for linear difference equations with variable coefficients of order higher than the first although infinite series solutions can be obtained in some cases. (Milne-Thomson (1))

DIFFERENCE APPROXIMATIONS TO NON-LINEAR DIFFERENTIAL EQUATIONS.

In this section difference approximations to non-linear differential equations of the form

$$y^r = f(x, y, y^1, \ldots, y^{r-1}) \tag{60}$$

are examined. Consider initially the first order differential equation

$$y' = f(x,y) \tag{61}$$

where y_0 is given and y is required for $x > 0$.
From (1),

$$y_n' = \frac{1}{h}(y_{n+1} - y_n) - \frac{hB}{2dx}(y_n')$$

$$= \frac{1}{h}(y_{n+1} - y_n) - \tfrac{1}{2}(y'_{n+1} - y'_n),$$

and so

$$y_{n+1} = y_n + \frac{h}{2}(y'_n + y'_{n+1}). \qquad (n \geq 0) \qquad (62)$$

A solution to the problem can now be computed from (61) with $x = nh$, (62), and the boundary condition. The error equation corresponding to (61) is

$$\epsilon' = f_y \epsilon, \qquad (63)$$

and so from (62) the error in the numerical solution satisfies the difference equation

$$\epsilon_{n+1}(1 - \frac{h}{2} f_{y,n+1}) = \epsilon_n (1 + \frac{h}{2} f_{y,n}). \qquad (64)$$

This equation has solution

$$\epsilon_n = \lambda_1^n \epsilon_0, \qquad (n \geq 1)$$

provided f_y is constant, where

$$\lambda_1 = \frac{(1 + \frac{h}{2}fy)}{1 - \frac{h}{2}fy} = 1 + hf_y + \frac{h^2}{2}f_y + \ldots\ldots = e^{hf_y}.$$

Now λ_1^n approximates to the solution of (63) when f_y is constant and the procedure outlined is stable if $f_y < 0$. If f_y is variable, (64) will require fresh examination in

order to obtain details of the stability of the procedure.

Rutishanser (5) used the second order difference equation

$$y_{n+1} = y_{n-1} + \frac{h}{3}\left(y'_{n+1} + 4y'_n + y'_{n-1}\right) \qquad (n \geq 1)$$

to obtain a numerical solution of (61). This leads to an error difference equation

$$\epsilon_{n+1}\left(1 - \frac{h}{3} f_{y,n+1}\right) - \epsilon_n \frac{4h}{3} f_{y,n} - \epsilon_{n-1}\left(1 + \frac{h}{3} f_{y,n-1}\right) = 0 \qquad (65)$$

with solution

$$\epsilon_n = a_1 \lambda_1^{\,n} + a_2 \lambda_2^{\,n}, \qquad (n \geq 2)$$

provided f_y is constant, where $\lambda_1^{\,n} = e^{nhf_y}$ and $\lambda_2^{\,n} = -e^{-\frac{1}{3}nhf_y}$. The former approximates to the solution of (63), whilst the latter has slipped in due to the higher order of the difference equation. As pointed out by Rutishanser this extra "smuggled" root of the error difference equation will cause trouble if $f_y < 0$. (compare section 2, p. 9). Again if f_y is variable, (65) will have to be re-examined in order to obtain stability details for Rutishanser's procedure.

A numerical step-by-step method of solution is now outlined for the second order non-linear differential equation

$$y'' = f(x, y, y') \tag{66}$$

subject to the boundary conditions of given y_0 and y_0'.

From (1)

$$y_n' = \frac{1}{h}(y_{n+1} - y_n) - \tfrac{1}{2} h y_n''$$

and so

$$y_{n+1} = y_n + h y_n' + \tfrac{1}{2} h^2 y_n''. \tag{67}$$

Also from (2)

$$y_{n+1}' = \frac{1}{h}(y_{n+1} - y_n) + \tfrac{1}{2} h y_{n+1}''. \tag{68}$$

A numerical solution to the problem can now be computed from (66), (67), (68) and the boundary conditions.

The error equation corresponding to (66) is

$$\epsilon'' = f_y \epsilon + f_{y'} \epsilon', \tag{69}$$

and so from (67) and (68) respectively, the error in the numerical solution and its derivative satisfy the difference equations

$$\epsilon_{n+1} = \epsilon_n \left(1 + \tfrac{1}{2} h^2 f_{y,n}\right) + \epsilon_n' h \left(1 + \tfrac{1}{2} h f_{y',n}\right) \tag{70}$$

and

$$\epsilon_{n+1}'\left(1-\tfrac{1}{2}hf_{y',n+1}\right)=-\epsilon_n\tfrac{1}{h}+\epsilon_{n+1}\left(\tfrac{1}{h}+\tfrac{1}{2}hf_{y,n+1}\right). \qquad (71)$$

If f_y and f_y' are constant, (70) and (71) can be solved by putting $\epsilon_n = p\lambda^n$ and $\epsilon_n{}^1 = q\lambda^n$. The two resulting values of λ are

$$\lambda_1,\lambda_2 = 1 + \tfrac{1}{2}\left[f_{y'} \pm (f_y{}'^2 + 4f_y)^{1/2}\right]h + \ldots\ldots \qquad (72)$$

where the solution of (70) and (71) is

$$\epsilon_n = a_1\lambda_1{}^n + a_2\lambda_2{}^n. \qquad (n \geq 2)$$

Now $\lambda_1{}^n$ and $\lambda_2{}^n$ approximate to the solutions of (69) with f_y and f_y' constant, and when both $|\lambda_1|$ and $|\lambda_2|$ are ≤ 1, the numerical procedure is stable. If f_y and f_y' are not constant, (70) and (71) will have to be solved in an approximate manner. The procedure will depend on the values of f_y and f_y'.

For example, suppose (67) and (68) are used to obtain a numerical solution of Bessel's equation

$$xy'' + y' + xy = 0 \qquad (73)$$

subject to the conditions $y_0 = 1$ and $y_0' = 0$. If (73) is put in the form of (66), $f_y = -1$ and $f_y' = -\frac{1}{nh}$. Considering this latter value to have small variation over the range of

x, $nh \leq x \leq (n+1)h$, (70) and (71) are solved with $f_{y,n} = f_{y,n+1} = -1$ and $f'_{y',n} = f'_{y',n+1} = -\frac{1}{nh}$, to give

$$\lambda_1, \lambda_2 = \frac{(2-h^2) \pm (n^2 - 4h^2 + h^4)^{\frac{1}{2}}}{2 + \frac{1}{n}} \qquad (74)$$

The numerical procedure will be stable if the value of n at each step gives $|\lambda_1|$ and $|\lambda_2| \leq 1$. During the calculation, n may pass through the following three stages.

(1) $n^2 < \frac{1}{h^2(4-h^2)}$ leading to a stable procedure if $|\lambda_1|$ and $|\lambda_2|$, as given by (74), are less than unity.

(2) $n^2 = \frac{1}{h^2(4-h^2)}$, $\lambda_1 = \lambda_2 = \frac{2 - h^2}{2+h(4-h^2)^{\frac{1}{2}}} < 1$ for all h, and so the procedure is stable.

(3) $n^2 > \frac{1}{h^2(4-h^2)}$, $|\lambda_1| = |\lambda_2| = (\frac{2n-1}{2n+1})^{\frac{1}{2}}$, and so the procedure is stable.

Rutishanser used the central difference formulae (Collatz (|) p.80)

$$y_{n+1} = 2y_n - y_{n-1} + \frac{1}{12}h^2 \left(y''_{n+1} + 10y''_n + y''_{n-1}\right)$$

$$y_{n+1} = y_{n-1} + \frac{1}{3}h \left(y''_{n+1} + 4y''_n + y''_{n-1}\right) \qquad (75)$$

to obtain a numerical solution of (66). This is a fourth order difference procedure, and again provided f_y and $f_{y'}$ are constant the fourth order error difference equations have solutions

$$\lambda_1, \lambda_2, \lambda_3, \lambda_4 = \lambda_1, \lambda_2, 1, -\left(1 - \tfrac{1}{3} f_{y'} h \cdots\right),$$

where λ_1 and λ_2 are given by (72). The "smuggled" solutions λ_3 and λ_4 will cause trouble in the numerical solution if $f_{y'} < 0$. Rutishanser demonstrated this "smuggled" instability by solving numerically the differential equation $y'' = -(y' + \frac{5}{4} y)$ using (75). This equation will give no trouble if solved numerically using (67) and (68) since from (72) λ_1 and λ_2 have moduli less than unity.

The method used to solve equations (61) and (66) can be extended in an obvious manner to obtain numerical solutions of (60) with $r \geq 3$. Provided the r difference equations used in the solution constitute a difference procedure of order r, no "smuggled" instability as described by Rutishanser will be present. The r error equations corresponding to the r

difference equations can be solved provided f_y, $f_{y'}$, $f_{y''}$, ...$f_{y^{r-1}}$ are constant, and the stability or instability of the method determined. If at least one of these derivatives is not constant, no general method of solution of the error equations is available, and an approximate technique, similar to that used with Bessel's equation, will have to be employed.

In conclusion, before leaving the numerical solution of ordinary equations, mention must be made of a very recent paper by Lotkin (15) where several standard methods of numerical integration are investigated for stability and error propagation.

CHAPTER II

PARTIAL DIFFERENCE EQUATIONS

Most of the partial differential equations of mathematical physics are special cases of the second-order linear equation

$$a\phi_{xx} + b\phi_{xy} + c\phi_{yy} + d\phi_x + e\phi_y + f = g \qquad (1)$$

where a, b, c, d, e, and f are constants or functions of x and y. Equation (1) is said to be elliptic if $b^2 - 4ac < 0$, parabolic if $b^2 - 4ac = 0$, and hyperbolic if $b^2 - 4ac > 0$. The study of the numerical solutions of difference replacements of partial differential equations is still in its infancy and replacements of Laplace's and Poisson's equations (elliptic), the heat conduction equation (parabolic), the wave equation (hyperbolic), and associated equations only will be considered. Scant reference will be made to non-linear equations in this chapter as no general theory is available in the non-linear case.

Two main types of problem arise in connection with partial equations;

(1) boundary value problems governed by equations of elliptic type where conditions are given on a closed boundary, and

(2) initial value problems generally governed by equations of parabolic or hyperbolic type where the boundary conditions are given on an open boundary.

Examples of (1) are problems governed by Laplace's and Poisson's equations whose finite difference replacements are solved implicitly by relaxation methods. As with ordinary difference equations, instability does not vitiate this method. Examples of (2) are problems governed by the heat conduction and wave

equations whose finite difference replacements are solved either by explicit or implicit methods. This time the step-by-step nature of the solution permits round-off errors introduced at any stage of the calculation to grow and render the final solution N worthless. Consequently, stability is essential in the numerical solution of problems of this type.

ELLIPTIC EQUATIONS.

1. Poisson's Equation (including Laplace).

Poisson's equation in two dimensions is

$$\phi_{xx} + \phi_{yy} = f, \qquad (2)$$

where $f(x,y)$ is a known function. When $f = 0$, (2) reduces to Laplace's equation. For convenience suppose a numerical solution of Poisson's equation is required over a rectangle where the boundary conditions consist of a knowledge of ϕ along the rectangle sides $x = \pm a$, $y = \pm b$. The interior of the rectangle is covered by a rectangular net of $(2M-1)$ columns $(-M < m < +M)$ and $(2N - 1)$ rows $(-N < n < +N)$.

The conventional finite difference approximation to (2) is

$$\frac{1}{(\Delta x)^2}\left(\phi_{m+1,n} - 2\phi_{m,n} + \phi_{m-1,n}\right) + \frac{1}{(\Delta y)^2}\left(\phi_{m,n+1} - 2\phi_{m,n} + \phi_{m,n-1}\right) = f,$$

$$\left(-M < m < +M\,;\ -N < n < +N\right) \qquad (3)$$

where Δx and Δy are the mesh lengths in the x- and y- directions respectively. Considering the case $\Delta x = \Delta y = h$, the $(2M-1)(2N-1)$ equations (3) can be written as the single matrix equation

$$\begin{bmatrix} A & I & & & & \\ I & A & I & & & \\ & I & A & I & & \\ & & \ddots & \ddots & \ddots & \\ & & & I & A & I \\ & & & & I & A \end{bmatrix} \begin{bmatrix} \Phi_{N-1} \\ \Phi_{N-2} \\ \Phi_{N-3} \\ \vdots \\ \Phi_{-N+2} \\ \Phi_{-N+1} \end{bmatrix} = \begin{bmatrix} G_{N-1} \\ G_{N-2} \\ G_{N-3} \\ \vdots \\ G_{-N+2} \\ G_{-N+1} \end{bmatrix}, \tag{4}$$

where

$$A = \begin{bmatrix} -4 & 1 & & & & \\ 1 & -4 & 1 & & & \\ & 1 & -4 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -4 & 1 \\ & & & & 1 & -4 \end{bmatrix}, \quad \Phi_n = \begin{bmatrix} \phi_{-M+1,n} \\ \phi_{-M+2,n} \\ \\ \\ \phi_{M-2,n} \\ \phi_{M-1,n} \end{bmatrix} \quad (-N<n<N),$$

I is the unit matrix of order $(2M-1)$, and the column vectors G_n $(-N < n < N)$ are known in terms of f, h, and the given boundary values of ϕ. With Laplace's equation, of course, only the boundary values of ϕ appear on the right hand side of (4).

The (2M-1)(2N-1) values of ϕ at the internal node points can be obtained numerically from (4) by the method of relaxation. This process forces the numerical solution towards the theoretical solution of (4) if sufficient number of places are retained after the decimal point. An estimation of the round-off errors arising from the numerical solution will be made in the next paragraph. Various authors have shown that the solution V of the difference equation (3) converges to the solution D of the differential equation (2) as $\Delta x, \Delta y$ tend independently to zero. In particular in a recent paper, Batschelet (16) shows that the solution of a finite difference approximation to a general elliptic equation converges to the exact solution of the differential equation as the mesh size tends to zero. The boundary conditions may involve derivatives and are given on a curved boundary.

Karlquist (17) and Cornock (18) solved (4) by numerically inverting the matrix on the left hand side of (4). Both solutions, like the method of relaxation, fail to give the answer in closed form. The present author has made several unsuccessful attempts at obtaining the inverse of the matrix in (4) in closed form, by extrapolating from known inverses when N and M are small. For example when N = M = 2 the matrix and its inverse are respectively

$$\left[\begin{array}{ccc|ccc|ccc}
-4 & 1 & & 1 & & & & & \\
1 & -4 & 1 & & 1 & & & & \\
 & 1 & -4 & & & 1 & & & \\
\hline
1 & & & -4 & 1 & & 1 & & \\
 & 1 & & 1 & -4 & 1 & & 1 & \\
 & & 1 & & 1 & -4 & & & 1 \\
\hline
 & & & 1 & & & -4 & 1 & \\
 & & & & 1 & & 1 & -4 & 1 \\
 & & & & & 1 & & 1 & -4
\end{array}\right]$$

and $-\frac{1}{224}$

$$\left[\begin{array}{ccc|ccc|ccc}
67 & 22 & 7 & 22 & 14 & 6 & 7 & 6 & 3 \\
22 & 74 & 22 & 14 & 28 & 14 & 6 & 10 & 6 \\
7 & 22 & 67 & 6 & 14 & 22 & 3 & 6 & 7 \\
\hline
22 & 14 & 6 & 74 & 28 & 10 & 22 & 14 & 6 \\
14 & 28 & 14 & 28 & 84 & 28 & 14 & 28 & 14 \\
6 & 14 & 22 & 10 & 28 & 74 & 6 & 14 & 22 \\
\hline
7 & 6 & 3 & 22 & 14 & 6 & 67 & 22 & 7 \\
6 & 10 & 6 & 14 & 28 & 14 & 22 & 74 & 22 \\
3 & 6 & 7 & 6 & 14 & 22 & 7 & 22 & 67
\end{array}\right]$$

Although this example and others give the general shape of the inverse, the author can see no pattern in the actual numbers.

2. Round-off Errors in Relaxation.

As stated before, the difference between the relaxation solution N and the theoretical solution V of (4) is entirely due to round-off errors. At any stage of the relaxation, the residuals at the internal nodes must lie in a range $-R \leqslant R_{m,n} \leqslant +R$ $(-M < m < +M;\ -N < n < +N)$, where R decreases as the accuracy of the calculation increases. Assuming that the elements of the vectors $G_n (-N < n < N)$ require no rounding off, the errors $\epsilon_{m,n}$ in $\phi_{m,n}$ satisfy the equation

$$\begin{bmatrix} A & I & & & & \\ I & A & I & & & \\ & I & A & I & & \\ & & \ddots & \ddots & \ddots & \\ & & & I & A & I \\ & & & & I & A \end{bmatrix} \begin{bmatrix} E_{N-1} \\ E_{N-2} \\ E_{N-3} \\ \vdots \\ E_{-N+2} \\ E_{-N+1} \end{bmatrix} = \begin{bmatrix} R_{N-1} \\ R_{N-2} \\ R_{N-3} \\ \vdots \\ R_{-N+2} \\ R_{-N+1} \end{bmatrix}, \tag{5}$$

where

$$E_n = \begin{bmatrix} \epsilon_{-M+1,n} \\ \epsilon_{-M+2,n} \\ \\ \\ \\ \epsilon_{M-2,n} \\ \epsilon_{M-1,n} \end{bmatrix} \quad \text{and} \quad R_n = \begin{bmatrix} R_{-M+1,n} \\ R_{-M+2,n} \\ \\ \\ \\ R_{M-2,n} \\ R_{M-1,n} \end{bmatrix} \qquad (-N < n < N).$$

Attempts to invert the matrix in (5) have been unsuccessful and so the following approximate method (Mitchell (13)) is outlined. The conditions governing the solution are;

(1) ϕ is given at nodes on the boundary in such a way that no rounding off is required.

(2) f is given at internal nodes in such a way that no rounding off is required.

(3) Δx and Δy satisfy the relation $k = \frac{(\Delta x)^2}{(\Delta y)^2} = \left(\frac{a}{b}\right)^2 \left(\frac{N}{M}\right)^2$.

(4) $R_{m,n} = R$ for all internal nodes (m,n), which distribution of residuals gives rise to maximum round-off errors for a given accuracy of calculation.

The errors $\epsilon_{m,n}$ in $\phi_{m,n}$ thus satisfy the $(2M - 1)(2N - 1)$ error equations

$$(\epsilon_{m+1,n} - 2\epsilon_{m,n} + \epsilon_{m-1,n}) + k(\epsilon_{m,n+1} - 2\epsilon_{m,n} + \epsilon_{m,n-1}) = R$$
$$(-M < m < +M \; ; \; -N < n < +N). \qquad (6)$$

By making the substitution

$$\epsilon_{m,n} = \left[E_{m,n} + \left(\frac{m^2 - M^2}{2}\right)\right] R, \qquad (7)$$

(6) is simplified to

$$(E_{m+1,n} - 2E_{m,n} + E_{m-1,n}) + k(E_{m,n+1} - 2E_{m,n} + E_{m,n+1}) = 0 \qquad (8)$$

with boundary conditions $E_{+M,n} = E_{-M,n} = 0$ for all n, and $E_{m,+N} = E_{m,-N} = \frac{1}{2}(M^2 - m^2)\,R$ for all m. Using the method of separation of variables, (8) is now solved for $E_{m,n}$ subject to the prescribed boundary conditions. The solution obtained is substituted in (7) and the errors found to be

$$\epsilon_{m,n} = \left[\frac{m^2-M^2}{2} + \frac{1}{4M}\sum_{r=0}^{M-1} \frac{(-1)^r \cos\,(r+\tfrac{1}{2})\frac{\pi}{2M}}{\sin^3(r+\tfrac{1}{2})\frac{\pi}{2M}\,\cosh N\beta} \cdot \cos\frac{(r+\tfrac{1}{2})\pi m}{M}\,\cosh\beta n\right] R. \quad (9)$$

$$(-M < m < +M;\ -N < n < +N)$$

where

$$\cosh\beta = 1 + \frac{2}{k}\sin^2\frac{(r+\tfrac{1}{2})\pi}{2M} \quad (10)$$

$$(r = 0, 1, 2, \ldots M-1)$$

The symmetry of the error in the four quadrants makes it only necessary to consider (9) over the restricted range of nodes $0 \leqslant m < M;\ 0 \leqslant n < N$. Setting

$$MN = A^2, \quad (11)$$

the results

$$M = A\left(\frac{a^2}{kb^2}\right)^{1/4} \qquad N = A\left(\frac{kb^2}{a^2}\right)^{1/4} \quad (12)$$

are obtained. The round-off error as given by (9) can now be expressed in terms of the two parameters, k the mesh ratio, and A^2 the number of internal nodes.

The maximum round-off error occurs at the origin and from (9) is given by

$$\epsilon_{0,0} = \left[-\frac{1}{2}\frac{A^2}{k^{1/2}}\frac{a}{b} + \frac{k^{1/4}}{4A}\left(\frac{b}{a}\right)^{1/2} \cdot \sum_{r=0}^{\frac{A}{k^{1/4}}\left(\frac{a}{b}\right)^{1/2}-1} \frac{(-1)^r \cos\left(r+\frac{1}{2}\right)\frac{k^{1/4}}{A}\left(\frac{b}{a}\right)^{1/2}\frac{\pi}{2}}{\sin^3\left(r+\frac{1}{2}\right)\frac{k^{1/4}}{A}\left(\frac{b}{a}\right)^{1/2}\frac{\pi}{2} \cdot \cosh Ak^{1/4}\left(\frac{b}{a}\right)^{1/2}\beta}\right] R \quad (13)$$

where

$$\cosh\beta = 1 + \frac{2}{k}\sin^2(r + \tfrac{1}{2})\,\frac{k^{1/4}}{A}\,(\tfrac{b}{a})^{1/2}\,\frac{\pi}{2}\,, \qquad (14)$$

$$(r = 0, 1, 2, \ldots, \frac{A}{k^{1/4}}\,(\tfrac{a}{b})^{1/2} - 1)$$

Two calculations involving (13) and (14) will now be carried out keeping the mesh ratio and the number of internal nodes constant in turn.

(a) Variation of round-off error. (Constant mesh ratio). The number of internal nodes is increased indefinitely, the mesh ratio being kept constant. In the limit as A^2 approaches infinity, the last term in the series expansion of (13) approaches zero. For the important terms at the beginning of the expansion $A\beta$ tends to

$$(\frac{b}{a})^{1/2}\,\frac{(r + \tfrac{1}{2})\pi}{k^{1/4}}\,,$$

and so from (13), $\epsilon_{0,0}$ approaches

$$-\tfrac{1}{2}(\tfrac{a}{b})\,\frac{A^2}{k^{1/2}}\left[1 - \frac{4}{\pi^3}\sum_{r=0}^{\frac{A}{k^{1/4}}(\frac{a}{b})^{1/2}-1}\frac{(-1)^r}{(r + \tfrac{1}{2})^3\cosh(\tfrac{b}{a})(r + \tfrac{1}{2})\pi}\right]R.$$

Successive terms in the series diminish so rapidly that a very good approximation to the series sum is given by the first term and so approximately

$$\epsilon_{0,0} = -\tfrac{1}{2}\left(\frac{a}{b}\right)\frac{A^2}{k^2}\left[1 - \frac{32}{\pi^3 \cosh\left(\frac{b}{a}\right)\frac{\pi}{2}}\right]R, \qquad (15)$$

as A^2 approaches infinity.

For a square region (a = b) with unit mesh ratio (k = 1), (15) gives $\epsilon_{0,0}$ approaching $-0.294\,A^2R$ as A^2 tends to infinity. This agrees with the round-off error given by Thom (19) for this problem using the method of squares. In addition an exact calculation, consistent with (20) and (21) modified for a = b and k = 1, gives values for the error at the origin as shown in table V

TABLE V

A^2	1	4	9
$-\epsilon_{0,0}/A^2R$	0.250	0.281	0.289

It would thus seem that in problems involving constant (b/a) and k, (15) together with a few exact calculations for small A^2 will give the variation of the round-off error with the number of internal nodes.

(b) Variation of round-off error. (Constant number of internal nodes). In this section, the number of internal nodes is kept constant, the mesh ratio being varied over its

allowable range. The latter extends from $(a/b)^2 1/A^4 (M = A^2, N = 1)$ to $(a/b)^2 A^4 (M = 1, N = A^2)$. Three values of k will be considered:

(1) $k = (a/b)^2 1/A^4$. Substituting in (13) and (14) the error at the origin becomes

$$\epsilon_{o,o} = \left[-\tfrac{1}{2} A^4 + \frac{1}{4A^2} \sum_{r=0}^{A^2-1} \frac{(-1)^r \cos(r + \tfrac{1}{2}) \frac{\pi}{2A}}{\sin^3 (r + \tfrac{1}{2}) \frac{\pi}{2A^2}} \cdot \right.$$

$$\left. \frac{1}{\left\{ 1 + 2 \left(\frac{b}{a}\right)^2 A^4 \sin^2 \frac{(r + \tfrac{1}{2})\pi}{2A^2} \right\}} \right] R.$$

Except for very small A^2 (up to about 4) a very good approximation to the error is given by

$$\epsilon_{o,o} = -\frac{A^4}{2} \left[1 - \frac{32}{\pi^5 \left\{ 1 + \frac{\pi^2}{8} \left(\frac{b}{a}\right)^2 \right\}} \right] R. \qquad (16)$$

2) $k = 1$. This value of the mesh ratio simplifies (13) and (14) and again except for very small A^2 the error is accurately given by

$$\epsilon_{o,o} = -\frac{a}{b} \frac{A^2}{2} \left[1 - \frac{32}{\pi^5 \cosh \frac{b}{a} \frac{\pi}{2}} \right] R. \qquad (17)$$

3) $k = (a/b)^2 A^4$. Equations (13) and (14) reduce

considerably to $\epsilon_{0,0} = -\frac{1}{2}\left[1 - 1/\cosh A^2\beta\right]R$, where $\cosh\beta = 1 + (b/a)^2\, 1/A^4$. Again except for very small A^2, the error at the origin is given by

$$\epsilon_{0,0} = -\frac{1}{2}\left[1 - \frac{1}{\cosh\left(\frac{b}{a}\right)2^{1/2}}\right] R. \qquad (18)$$

Equations (16), (17) and (18) thus give the variation of the round-off error over the mesh ratio range for a given number of internal nodes. Table VI shows this variation in the case of the square (a = b) for three values of A^2.

TABLE VI

$A^2 = 16$	$K^{1/2}$	0·06	1	16
	$-\epsilon_{0,0}/R$	0·27	4·7	69
$A^2 = 36$	$K^{1/2}$	0·03	1	36
	$-\epsilon_{0,0}/R$	0·27	10·6	350
$A^2 = 100$	$K^{1/2}$	0·01	1	100
	$-\epsilon_{0,0}/R$	0·27	29·5	2700

For convenience the inverse mesh ratio $K = 1/k$ has been used. It should be kept in mind that although the smallest permissible K for each A^2 produces the smallest round-off errors, the nodes are not distributed in an even manner over the square, and so factors other than minimising the round-off error will influence the choice of mesh ratio.

It will be emphasized again that the round-off errors evaluated here are in excess of the errors which arise in actual relaxation calculations. Nevertheless, there seems no doubt that round-off errors will be important if the number of nodes is large, unless a sufficient number of places is retained after the decimal point. A safe guide to the number of places necessary is given by the above results. If ϕ requires rounding off at nodes on the boundary, the results obtained are still substantially correct. If f requires rounding off at internal nodes, adjustments will have to be made. These, of course, will be unnecessary with Laplace's Equation.

8. Solution of Laplace's Equation by Step-by-Step Methods.

Hyman (20) replaced Laplace's equation by the conventional difference equation written in the form

$$\phi_{m,n+1} = 2(1+r^2)\phi_{m,n} - r^2(\phi_{m-1,n} + \phi_{m+1,n}) - \phi_{m,n-1} \qquad (19)$$

where $r = \Delta y/\Delta x$. A knowledge of ϕ at nodes on the $(n-1)^{th}$ and n^{th} rows enables ϕ at nodes on the $(n+1)^{th}$ row to be calculated explicitly using (19). In order to start the calculation, ϕ must be given at nodes on the rows $n = -N$, $-N+1$, and for the calculation to continue and cover the complete rectangle ϕ must be given at nodes on the columns $m = -M, +M$. The boundary conditions of the problem,

however, consist of a knowledge of ϕ at nodes on the boundaries m = - M, + M, and n = - N, + N, and so somehow the boundary data must be transferred to the row n = - N + 1. Hyman did this by solving (19) in the form of a series and computing the values of ϕ at nodes in the row n = - N + 1. The solution is then stepped off using (19).

A more serious difficulty in the step-by-step procedure is the essential instability of step-by-step methods when used to solve elliptic equations. Von Neumann's criterion for stability of step-by-step methods, as outlined by O'Brien, Hyman, and Kaplan (21), will be discussed in detail in the sections dealing with parabolic and hyperbolic equations, where the quantities used will be explained. Corresponding to (19) is the auxiliary equation

$$\xi^2 - 2\left(1 + 2r^2 \sin^2 \frac{\beta \Delta x}{2}\right)\xi + 1 = 0, \qquad (\beta > 0)$$

from which

$$|\xi| = 1 + 2r^2 \sin^2 \frac{\beta \Delta x}{2},$$

and so stability criterion $|\xi| \leq 1$ cannot be satisfied for any value of r. Other second-order finite difference approximations to Laplace's equation are the backward difference replacement

$$r^2\left(\phi_{m-1,n} + \phi_{m+1,n}\right) + (1-2r^2)\,\phi_{m,n} = 2\phi_{m,n-1} - \phi_{m,n-2},$$

and the forward difference replacement

$$\phi_{m,n} = 2\phi_{m,n-1} + (2r^2-1)\phi_{m,n-2} - r^2(\phi_{m-1,n-2} + \phi_{m+1,n-2}).$$

The auxiliary equations of these replacements give

$$\xi = \frac{1}{1 \pm 2r \sin \frac{\beta \Delta x}{2}}$$

and

$$\xi = 1 \pm 2r \sin \frac{\beta \Delta x}{2}$$

respectively, leading to instability in both cases for all r. In order to keep this instability in check, Hyman computed from the series solution of (19), the values of ϕ at nodes on several pairs of adjacent rows spaced out between n = - N and n = + N. Using (19) he then stepped off the whole region piecewise. It has already been stated that various authors, most recently Rosenbloom (22) have demonstrated the convergence of the theoretical solution of the difference replacement of Laplace's equation as the mesh size is diminished. As a result the instability of the step-by-step procedure is not accompanied by divergence.

In conclusion, although this step-by-step method of solution of elliptic equations seems feasible, it is unlikely to be used in practise. If a series solution of the difference equation can be obtained, it is most likely to be used to compute ϕ at all the nodes throughout the region.

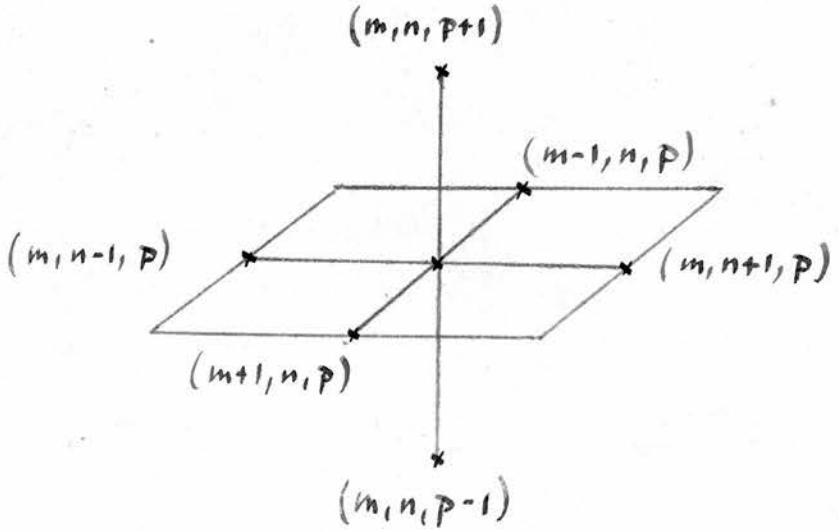
If a theoretical solution of the difference equation can not be obtained, the step-by-step method is impossible and relaxation will be used.

4. Poisson's Equation in Three Dimensions.

Poisson's equation in three dimensions is

$$\phi_{xx} + \phi_{yy} + \phi_{zz} = f, \tag{20}$$

where f(x, y, z) is a known function. Suppose a solution is required inside a rectangular block whose faces are the planes $x = \pm a$, $y = \pm b$, $z = \pm c$, where the boundary conditions consist of a knowledge of ϕ at nodes on these faces. The block is divided up by (2P - 1) planes, $(-P < p < +P)$ parallel to the x, y plane, neighbouring planes being distance Δz apart. The interior of each rectangular region parallel to the x, y plane is covered by a rectangular net of (2M - 1) columns $(-M < m < +M)$ parallel to the y- axis and (2N - 1) rows $(-N < n < +N)$ parallel to the x - axis.



The conventional finite difference approximation to (20) is

$$\frac{1}{(\Delta x)^2}\left(\phi_{m+1,n,p} + \phi_{m-1,n,p} - 2\phi_{m,n,p}\right) + \frac{1}{(\Delta y)^2}\left(\phi_{m,n+1,p} + \phi_{m,n-1,p} - 2\phi_{m,n,p}\right) + \frac{1}{(\Delta z)^2}\left(\phi_{m,n,p+1} + \phi_{m,n,p-1} - 2\phi_{m,n,p}\right) = f, \qquad (21)$$

$$(-M < m < +M\ ;\ -N < n < +N\ ;\ -P < p < +P)$$

where Δx, Δy, and Δz are the mesh lengths in the x, y, and z directions respectively. Considering the case $\Delta x = \Delta y = \Delta z = h$, $M = N$, $a = b$, the $(2N-1)^2(2P-1)$ equations (21) become in matrix form

$$\begin{bmatrix} A & I_1 & & & & \\ I_1 & A & I_1 & & & \\ & I_1 & A & I_1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & I_1 & A & I_1 \\ & & & & I_1 & A \end{bmatrix} \begin{bmatrix} F_{P-1} \\ F_{P-2} \\ F_{P-3} \\ \vdots \\ F_{-P+2} \\ F_{-P+1} \end{bmatrix} = \begin{bmatrix} G_{P-1} \\ G_{P-2} \\ G_{P-3} \\ \vdots \\ G_{-P+2} \\ G_{-P+1} \end{bmatrix}, \qquad (22)$$

where

$$A = \begin{bmatrix} B & I_2 & & & \\ I_2 & B & I_2 & & \\ & \ddots & \ddots & \ddots & \\ & & I_2 & B & I_2 \\ & & & I_2 & B \end{bmatrix}, \quad F_p = \begin{bmatrix} \Phi_{N-1} \\ \Phi_{N-2} \\ \vdots \\ \Phi_{-N+2} \\ \Phi_{-N+1} \end{bmatrix}, \quad G_p = \begin{bmatrix} \gamma_{N-1} \\ \gamma_{N-2} \\ \vdots \\ \gamma_{-N+2} \\ \gamma_{-N+1} \end{bmatrix}$$

$(-P < p < +P)$

with

$$B = \begin{bmatrix} -6 & 1 & & & \\ 1 & -6 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -6 & 1 \\ & & & 1 & -6 \end{bmatrix}, \quad \Phi_n = \begin{bmatrix} \phi_{n,N-1,p} \\ \phi_{n,N-2,p} \\ \vdots \\ \phi_{n,-N+2,p} \\ \phi_{n,-N+1,p} \end{bmatrix}, \quad (-N < n < +N)$$

I_1 and I_2 are unit matrices of orders $(2N-1)^2$ and $(2N-1)$ respectively, and the column vectors γ_n $(-N < n < +N)$ are known in terms of f, h and the given boundary values of ϕ.

The values of ϕ at the internal node points can be obtained numerically from (22) by the method of relaxation. This is illustrated for a comparatively small number of nodes by Allen and Dennis (23). It is unlikely that (22) will be solved by matrix inversion due to the extremely complicated nature of the matrix on the left hand side of (22).

5. The Biharmonic Equation.

The biharmonic equation is

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^2 \phi = 0, \tag{23}$$

and suppose a solution is required over the rectangle $x = \pm a$, $y = \pm b$, the interior of which is covered by a rectangular net of $(2M - 1)$ columns $(-M < m < +M)$ and $(2N - 1)$ rows $(-N < n < +N)$. The conventional finite difference approximation to (23) with $\Delta x = \Delta y = h$ is represented with obvious notation by the block

$$\begin{array}{ccccc} & & 1 & & \\ & 2 & -8 & 2 & \\ - & -8 & 20 & -8 & 1 \\ & 2 & -8 & 2 & \\ & & 1 & & \end{array} \tag{24}$$

If the boundary conditions consist of a knowledge of ϕ and ϕ_x on $x = \pm a$ and of ϕ and ϕ_y on $y = \pm b$, the values of the unknown function ϕ are given by the matrix equation

$$\begin{bmatrix} B & A & I & & & & & \\ A & B & A & I & & & & \\ I & A & B & A & I & & & \\ & I & A & B & A & I & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & I & A & B & A & \\ & & & & I & A & B & \end{bmatrix} \begin{bmatrix} \Phi_{N-2} \\ \Phi_{N-1} \\ \vdots \\ \Phi_{-N+3} \\ \Phi_{-N+2} \end{bmatrix} = \begin{bmatrix} G_{N-2} \\ G_{N-1} \\ \vdots \\ G_{-N+3} \\ G_{-N+2} \end{bmatrix}, \tag{25}$$

where

$$B = \begin{bmatrix} 20 & -8 & & & & & \\ -8 & 20 & -8 & 1 & & & \\ 1 & -8 & 20 & -8 & 1 & & \\ & 1 & -8 & 20 & -8 & 1 & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & 1 & -8 & 20 & -8 \\ & & & & 1 & -8 & 20 \end{bmatrix},$$

$$A = 2\begin{bmatrix} -4 & 1 & & & & \\ 1 & -4 & 1 & & & \\ & 1 & -4 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -4 & 1 \\ & & & & 1 & -4 \end{bmatrix}, \quad \Phi_n = \begin{bmatrix} \phi_{-M+2,n} \\ \phi_{-M+3,n} \\ \\ \\ \\ \phi_{M-3,n} \\ \phi_{M-2,n} \end{bmatrix}, \quad (-N+1 < n < N-1)$$

I is the unit matrix of order (2M-3), and the column vectors $G_n (-N+1 < n < N-1)$ are known in terms of h, and the given boundary values of ϕ at nodes on m = - M, -M + 1, M - 1, M and n = - N, -N + 1, N - 1, N. The (2M-3)(2N-3) unknown values of ϕ can be obtained from (25) by relaxation. Cornock (18) has solved (25) by numerically inverting the matrix on the left hand side of (25) for the case of seven internal rows and columns.

If the boundary conditions consist of a knowledge of ϕ and ϕ_{xx} on $x = \pm a$ and ϕ and ϕ_{yy} on $y = \pm b$, the solution can be obtained simply in two steps. First let

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) \phi = W. \tag{26}$$

Since ϕ_{xx} and ϕ_{yy} are known on the boundary, W is also known there, and since

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) W = 0 \tag{27}$$

at (2M-1)(2N-1) interior points, W is obtained by relaxation in

the usual manner. Second, having found W, Poisson's equation (26) is solved by relaxation for ϕ. This technique is used in solving numerically the equation for very slow flow of a viscous fluid and details will be given in the section on fluid dynamics.

PARABOLIC EQUATIONS.

1. The Heat Conduction Equation.

The parabolic heat conduction equation in one dimension is

$$\phi_{xx} - \phi_t = f, \tag{20}$$

where x and t are the distance and time co-ordinates respectively, $\phi(x, t)$ is the temperature and f(x,t) is a known function, usually zero. The boundary conditions consist of a knowledge of ϕ along $x = 0$, L and $t = 0$. The solution is required in the region $0 \leq x \leq L$, $t \geq 0$.

Suppose this region is covered by a rectangular net, the mesh lengths being Δx and Δt in the x - and t - directions respectively. If j and k are the row and column numbers respectively, a general finite difference replacement of (20) is

$$a\left[\phi_{j,k-1} - 2\phi_{j,k} + \phi_{j,k+1}\right] + (1-a)\left[\phi_{j-1,k-1} - 2\phi_{j-1,k} + \phi_{j-1,k+1}\right]$$

$$= \frac{(\Delta x)^2}{\Delta t}\left[\phi_{j,k} - \phi_{j-1,k}\right] + (\Delta x)^2 f, \qquad (21)$$

$$(j = 1,2, \cdots ; k = 1,2, \cdots N)$$

where $a > 0$. This difference equation is first order in the time co-ordinate and second order in the distance co-ordinate. When $a = 0$, (21) reduces to the four point formula

$$\phi_{j,k} = \left(1 - \frac{2}{s}\right)\phi_{j-1,k} + \frac{1}{s}\left(\phi_{j-1,k-1} + \phi_{j-1,k+1}\right) - \Delta t\, f, \qquad (22)$$

where $s = (\Delta x)^2/\Delta t$. Using (22), the solution can be stepped-off explicitly from the boundary values of ϕ. For any other value of a, ϕ is given implicitly and a relaxation solution is necessary for each row in turn.

2. Stability Criterion for Step-by-Step Methods.

The danger of instability in calculations involving ordinary difference equations has already been adequately illustrated. Accordingly von Neumann's method (21) for examining the stability of step-by-step finite difference methods applied to initial value problems with two independent variables is now outlined and applied to (21). At each step of a step-by-step calculation, a row of round-off errors is introduced which propagates forward, in the case of (21), according to the error difference equation

$$a\left[\epsilon_{j,k-1} - 2\epsilon_{j,k} + \epsilon_{j,k+1}\right] + (1-a)\left[\epsilon_{j-1,k-1} - 2\epsilon_{j-1,k} + \epsilon_{j-1,k+1}\right]$$

$$= s\left[\epsilon_{j,k} - \epsilon_{j-1,k}\right], \qquad (23)$$

with obvious notation. Suppose a harmonic decomposition is made of a row of errors $E(x)$ introduced at any step; say

$$E(x) = \sum_n A_n e^{i\beta_n x},$$

where in general the frequencies $|\beta_n|$ and n are somewhat arbitrary. It is necessary only to consider the single term $e^{i\beta x}$ where β is any real number. For convenience, suppose that the row being considered corresponds to $t = 0$. A solution of (23) which reduces to $e^{i\beta x}$ when $t = 0$ is

$$e^{\alpha t}\ e^{i\beta x} \qquad (24)$$

where $\alpha = \alpha(\beta)$. More generally, if the difference equation is N^{th} order in t, there are N α's for each β. The original error $e^{i\beta x}$ will not grow with time if

$$\left| e^{\alpha \Delta t} \right| \leq 1 \qquad (25)$$

for all α, which is von Neumann's criterion for stability.

Substituting $e^{\alpha t}\ e^{i\beta x}$ in (23), and simplifying, the

result

$$e^{\alpha \Delta t} = 1 - \frac{4 \sin^2 \beta \Delta x/2}{s+4a\sin^2 \beta \Delta x/2}$$

is obtained. The condition (25) for stability yields

$$-1 \leqslant 1 - \frac{4 \sin^2 \beta \Delta x/2}{s+4a\sin^2 \beta \Delta x/2} \leqslant 1.$$

The right inequality is satisfied for all a and s, whilst the left inequality is satisfied if

$$s \geqslant 2(1-2a) \sin^2 \beta \Delta x/2.$$

Thus (23) is stable for all s if $a \geqslant \frac{1}{2}$, and for $s \geqslant 2(1-2a)$ if $0 \leqslant a < \frac{1}{2}$. The four point formula (22) is stable for $s \geqslant 2$.

For problems limited in the x-direction to an extent L with N internal nodes, the above criterion for stability is too severe. For example, (21) is shown to be stable by von Neumann's criterion provided $s \geqslant 2(1-2a)$, but as will be illustrated in section 6, the maximum value of $\sin^2 \frac{\beta \Delta x}{2}$ is not unity but $\sin^2 \frac{N}{N+1} \frac{\pi}{2}$, and so the condition for stability can be relaxed to give

$$s \geqslant 2(1-2a) \sin^2 \frac{N}{N+1} \frac{\pi}{2},$$

where N is the number of internal nodes. The stability criterion of von Neumann will not be discussed further at this

point, but will be referred to throughout the remainder of the chapter.

5. Round-off Errors in Difference Replacements of the Heat-Conduction Equation.

Summing up the results of the previous sections, there are two ways of solving the heat conduction equation by finite difference methods:

(1) explicitly using the four point formula (22). This method has the great advantage of ease of calculation. Unfortunately the four point formula is only stable provided $S \geq 2$, and so if a solution is required over a large time range the number of steps of calculation necessary, even with $S = 2$, might be prohibitively large. On the credit side, this method is free from serious round-off errors.

(2) implicitly using the six point formula (21) with $a \geq \frac{1}{2}$. This method has the disadvantage of requiring a relaxation solution at each step. On the credit side, it is stable for all s and so Δt can be chosen at the computer's convenience. This enables problems involving large time ranges to be solved in a reasonable number of steps, provided the truncation errors (Section 6) are not large. The relaxation solution at each step, however, introduces round-off errors and the remainder of this section will be taken up with assessing the magnitudes of these errors, and determining the value of the parameter a which will keep these errors to a minimum. (Mitchell (24)).

For convenience, suppose that (21) has been divided through by a, and so the error equation (23) becomes

$$\left[\epsilon_{j,k-1} - \left(2 + \frac{s}{a}\right)\epsilon_{j,k} + \epsilon_{j,k+1}\right] + \frac{1-a}{a}\left[\epsilon_{j-1,k-1} - \left\{2 - \frac{s}{1-a}\right\}\epsilon_{j-1,k} + \epsilon_{j-1,k+1}\right] = r_{j,k}, \qquad (36)$$

where $r_{j\,k}$ ($j = 1, 2, \ldots.; k = 1, 2, \ldots. N$) are the residuals in the relaxation solutions. The relaxation process is continued at each step until all the residuals vanish to a prescribed degree of accuracy. If r is the modulus of the maximum residual neglected, then all other residuals lie in the range $-r \leq r_{j\,k} \leq +r$. For example, working correctly to two places of decimals, $r = 0.005$. It is assumed that no rounding-off is required at nodes on the boundaries, and so $\epsilon_{j,o} = \epsilon_{j,N+1} = \epsilon_{o,k} = 0$ for all j and k.

The error equations (26) for the case of M rows can be written as the matrix equation

$$\begin{bmatrix} A & & & & & \\ B & A & & & & \\ & B & A & & & \\ & & \ddots & \ddots & & \\ & & & B & A & \\ & & & & B & A \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \vdots \\ E_{M-2} \\ E_{M-1} \\ E_M \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_{M-2} \\ R_{M-1} \\ R_M \end{bmatrix}, \qquad (27)$$

where, if P(x) denotes the square matrix

$$\begin{bmatrix} -x & 1 & & & \\ 1 & -x & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -x & 1 \\ & & & 1 & -x \end{bmatrix}$$

of order N, $A = P(2 + \frac{s}{a})$, $B = \frac{1-a}{a} P\left(2 - \frac{s}{1-a}\right)$,

$$E_j = \begin{bmatrix} \epsilon_{j,1} \\ \epsilon_{j,2} \\ \vdots \\ \epsilon_{j,N-1} \\ \epsilon_{j,N} \end{bmatrix} \quad \text{and} \quad R_j = \begin{bmatrix} r_{j,1} \\ r_{j,2} \\ \vdots \\ r_{j,N-1} \\ r_{j,N} \end{bmatrix}, \quad (j = 1, \ldots. M)$$

The solution of (27) is

$$E_1 = A^{-1} R_1$$

together with

$$E_j = A^{-1} R_j - A^{-1} B E_{j-1}, \quad (j = 2 \ldots\ldots M)$$

which lead immediately to

$$E_M = \sum_{j=1}^{M} (-A^{-1}B)^{M-j} A^{-1} R_j. \qquad (28)$$

The magnitude of the round-off errors will of course depend on the residual distribution. A good indication of the variation of the errors with s and a can be obtained by considering the variation of the maximum errors with these parameters.

Consider first of all the residual distribution

$$R_j = R_1. \qquad (j = 2, 3, \ldots M)$$

This modifies (28) to give

$$E_M = \left[\sum_{j=1}^{M} (-A^{-1}B)^{M-j} \right] E_1 ,$$

where the latent roots of $A^{-1}B$ are

$$-1 + \frac{4 \cos^2 \frac{\alpha\pi}{2(N+1)}}{s+4a \cos^2 \frac{\alpha\pi}{2(N+1)}} , \qquad (\alpha = 1, \ldots N)$$

all of which have modulus less than or equal to unity for $s \geq 2(1 - 2a)$. Now E_M can be simplified further to give

$$E_M = \left[I + (-1)^{M-1} (A \quad B)^M \right] (A+B)^{-1} R_1,$$

where the p, q^{th} element of $(A^{-1}B)^M$ is

$$\eta_{p,q} = \frac{2}{N+1} \sum_{\alpha=1}^{N} \sin \frac{p\alpha\pi}{N+1} \sin \frac{q\alpha\pi}{N+1} \left[-1 + \frac{4 \cos^2 \frac{\alpha\pi}{2(N+1)}}{S + 4a \cos^2 \frac{\alpha\pi}{2(N+1)}} \right]^M .$$

Since the latent roots of $A^{-1}B$ all have modulus less than unity, it follows that all the elements of $(A^{-1}B)^M$ tend to zero as M approaches infinity, and so the limiting value of the error vector is given by

$$e_1 = (A + B)^{-1} R_1 .$$

It should be noted that $(A + B)$ is independent of the mesh ratio S and in fact

$$e_1 = a\left[P(2)\right]^{-1} R_1 , \tag{29}$$

where the p, qth element η_{pq} of $\left[P(2)\right]^{-1}$ is

$$\eta_{p,q} = \eta_{q,p} = \frac{-p(N+1-q)}{N+1} \qquad (p \leq q)$$

Next consider the residual distribution

$$R_j = (-1)^{j+1} R_1 . \qquad (j = 2, 3, \ldots).$$

Proceeding as before, this leads to the new limiting value for the error vector

$$e_2 = (A-B)^{-1} R_1 = \begin{cases} \frac{a}{2a-1}\left[P\left\{2\left(1 + \frac{S}{2a-1}\right)\right\}\right]^{-1} R_1 & (a \neq \tfrac{1}{2}) \\ \frac{1}{4S} R_1 , & (a = \tfrac{1}{2}) \end{cases} \tag{30}$$

where no importance is attached to the sign of e_2. The values of $\left[P(x)\right]^{-1}$ are given by Rutherford (14). The maximum possible error vector is given by (29) or (30).

For $a \geq 1$, the maximum error vector is given by e_1 which is independent of S. From (29) it is easily seen that for this range of the parameter, the smallest maximum error is given by $a = 1$. The value of this error is $\left[P(2)\right]^{-1} R_1$, which is independent of the mesh ratio s, but is a function of the number of internal columns N. For $\frac{1}{2} \leq a < 1$, the maximum error vector is given by either e_1 or e_2 depending on the values of a, s, and N. For $0 < a < \frac{1}{2}$, the maximum error is given by e_2 for all s and N.

Without carrying out detailed calculations in the range $0 < a < 1$, it can be said that if stability is required for all s, the value of a giving rise to minimum round-off errors always lies in the range $\frac{1}{2} \leq a \leq 1$. When $a = \frac{1}{2}$, (21) becomes the six point formula suggested by von Neumann, the round-off errors of which have been studied in detail by the present author (25). When $a = 1$, (21) becomes the convenient backward four point formula with the simple result that the maximum round-off error is $\left[P(2)\right]^{-1} R_1$ for all values of the mesh ratio s.

4. Von Neumann's Six Point Formula.

From the previous section it appears that if Δt is to be chosen at the computer's convenience implicit formulae will have to be used. These formulae give rise to the smallest

round-off errors when $\frac{1}{2} \leq a \leq 1$, and so the errors arising from van Neumann's difference approximation, (21) with $a = \frac{1}{2}$, will now be studied in detail.

From (27) with $a = \frac{1}{2}$, the errors in the first row are given by the matrix equation.

$$E_1 = A^{-1} R_1 , \tag{31}$$

where the p, q^{th} element $\eta_{p,q}$ of A^{-1} is

$$\eta_{p,q} = \eta_{q,p} = (-1)^{p+q} \frac{\phi_{p-1}\, \phi_{N-q}}{\phi_N} \qquad (p \leq q), \tag{32}$$

where

$$\phi_t = \frac{\sin (t+1)\theta}{\sin \theta} \quad (t = 0, 1, \ldots N), \tag{33}$$

and

$$\theta = \cos^{-1}\left[-(1+s)\right].$$

For all n,

$$\cos n\theta = \tfrac{1}{2} (-1)^n \left[U^n + V^n\right] \tag{34}$$

where $U = (1+s) + (2s+s^2)^{\frac{1}{2}}$ and $V = (1+s) - (2s+s)^{\frac{1}{2}}$.

Consider initially a constant residual distribution in the first row $r_{1,k} = r$ $(1 \leq k \leq N)$. Although this distribution is unlikely to arise in practise, it leads to the worst possible round-off errors in the first row and so provides a useful upper limit. From (31), (32), (33) and (34), after a little manipulation, the largest error in the first row, which occurs

at the central node, can be shown to be

$$\epsilon_{1,\frac{N+1}{2}} = -\frac{1}{2s}(1 - C)r, \tag{35}$$

where $1/C = \frac{1}{2}\left[U^{\frac{1}{2}(N+1)} + V^{\frac{1}{2}(N+1)}\right]$. In a similar manner the smallest error is

$$\epsilon_{1,1} = \epsilon_{1,N} = -\frac{1}{2s}(1 - D)r, \tag{36}$$

where

$$D = \frac{U^{\frac{1}{2}(N-1)} + V^{\frac{1}{2}(N-1)}}{U^{\frac{1}{2}(N+1)} + V^{\frac{1}{2}(N+1)}} .$$

The errors at all other nodes of the first row lie between these values. For $N \geq 1$, $s > 0$, it may be shown that $0 < C < 1$ and $0 < D < 1$. In addition, for large N, (35) and (36) become approximately,

$$\epsilon_{1,\frac{N+1}{2}} = -\frac{r}{2s}, \tag{37}$$

and

$$\epsilon_{1,1} = \epsilon_{1,N} = \frac{-r}{2s}\left[1 - 1/U\right]. \tag{38}$$

Next consider a rectangular residual distribution in the first row $-r \leq r_{1,k} \leq +r$ $(1 \leq k \leq N)$. This represents much more accurately a residual distribution likely to be obtained in practise. Returning to (31) and (32) the error at the

central node of the first row is

$$\epsilon_{1,\frac{N+1}{2}} = (-1)^{\frac{N-1}{2}} \frac{\phi_{\frac{N-1}{2}}}{\phi_N} \left[\phi_0 (r_{1,1} + r_{1,N}) - \phi_1 (r_{1,2} + r_{1,N-1}) \cdots \right.$$

$$\left. \cdots - \phi_{\frac{N-3}{2}} (r_{1,\frac{N-1}{2}} + r_{1,\frac{N+3}{2}}) + \phi_{\frac{N-1}{2}} r_{1,\frac{N+1}{2}} \right] .$$

and so the standard mean deviation σ in the error at this node is given by

$$\sigma^2_{1,\frac{N+1}{2}} = \frac{\phi^2_{\frac{N-1}{2}}}{3\phi^2_N} \left[2\left(\phi_0^2 + \phi_1^2 + \cdots + \phi^2_{\frac{N-3}{2}} \right) + \phi^2_{\frac{N-1}{2}} \right] r^2 .$$

After some manipulation using (33), the right-hand side becomes

$$\frac{1}{12 \cos^2 \frac{N+1}{2}\theta \sin^2\theta} \left[\frac{N+1}{2} + \frac{\cos (N+3)\theta - \cos (N-1)\theta}{8 \sin^2 \theta} \right] r^2 , \qquad (39)$$

which, with the help of (34) gives approximately for large N,

$$\sigma^2_{1,\frac{N+1}{2}} = \frac{U^4 - 1}{48 s^2 U^2 (s+2)^2} r^2 . \qquad (40)$$

Similarly, the square of the standard mean deviation at the end nodes is given approximately by

$$\sigma^2_{1,1} = \sigma^2_{N,1} = \frac{U^2 - 1}{12 U^2 (s^2 + 2s)} r^2 \qquad (41)$$

for large N. Again the squares of the standard mean deviation at intermediate nodes lie between the values given by (40) and (41).

In order to obtain the values of the round-off errors in subsequent rows, (29) with $a = \frac{1}{2}$ is first of all used together with the constant residual distribution $r_{1,k} = r$ $(1 \leq k \leq N)$ in the first row. The limiting values of the maximum and minimum errors in the M^{th} row are then given by

$$e_{\frac{N+1}{2}} = -\frac{1}{16}(N+1)^2 r \tag{42}$$

and

$$e_1 = e_N = -\frac{1}{4} N r \tag{43}$$

respectively as M approaches infinity. If, however, the rectangular distribution of residuals $-r \leq r_{1,k} \leq +r$ $(1 \leq k \leq N)$ is considered in the first row, the limiting values of the maximum and minimum standard mean deviations in the errors in the M^{th} row are given by

$$\rho^2_{\frac{N+1}{2}} = \frac{1}{576}(N+1)(N^2+2N+3) r^2 \tag{44}$$

and

$$\rho_1^2 = \rho_N^2 = \frac{1}{72} \frac{N(2N+1)}{N+1} r^2 \tag{45}$$

respectively.

Next consider (30) with $a = \frac{1}{2}$ together with the constant residual distribution in the first row. The errors in the

M^{th} row approach the values

$$e_k = \frac{(-1)^M}{4s} r, \qquad (k = 1, 2, \ldots N) \tag{46}$$

and for a rectangular distribution of residuals in the first row, the standard mean deviations in the errors in the Mth row approach the values

$$\rho_k^2 = \frac{1}{48s^2} r^2 \qquad (k = 1, 2, \ldots N) \tag{47}$$

as M tends to infinity.

Using a constant residual distribution in the first row along with $R_j = R_1$ $(j = 2, \ldots M)$, it follows from (35) and (42) that in the limit as s approaches zero

$$\epsilon_{1, \frac{N+1}{2}} = 2 e_{\frac{N+1}{2}}$$

for $N \geq 1$. Consequently, provided s is sufficiently small $|\epsilon_{1, \frac{N+1}{2}}|$ may be larger than $|e_{\frac{N+1}{2}}|$ by a factor not exceeding 2. For any $N \geq 3$, the values of s_{crit} for which $\epsilon_{1, \frac{N+1}{2}} = e_{\frac{N+1}{2}}$ are shown in fig. 1 from which it follows that $|\epsilon_{1, \frac{N+1}{2}}| \leq |e_{\frac{N+1}{2}}|$ for $s \geq s_{crit}$. For all s, the maximum error occurs in the first or last row, and as the row number increases, the error at the mid-point either oscillates with decreasing amplitude about, or tends asymptotically to, $e_{\frac{N+1}{2}}$.

For N = 3, ($s_{crit.} = 0.5$) fig. 3 illustrates the change in the mid-point error with increasing row number for s = 0.05 and s = 0.5.

If a rectangular distribution of residuals is used in the first row together with $R_j = R_1$ (j = 2, ... M) it can be shown from (40) and (44) that in the limit, as s approaches zero,

$$\sigma^2_{1,\frac{N+1}{2}} = 4\rho^2_{\frac{N+1}{2}} \quad \ldots\ldots$$

for $N \geq 1$. Again for any $N \geq 3$, the values of $s_{crit.}$ are shown in fig. 1, for which $\sigma^2_{1,\frac{N+1}{2}} = \rho^2_{\frac{N+1}{2}}$. For $s \gtrless s_{crit.}$, $\sigma^2_{\frac{N+1}{2}} \lesseqgtr \rho^2_{\frac{N+1}{2}}$.

When a constant residual distribution is assumed in the first row together with $R_j = (-1)^{j+1} R_1$ (j = 2, ... M), it follows from (35) and (46) that in the limit as N approaches infinity

$$|\epsilon_{1,\frac{N+1}{2}}| = 2|\epsilon_k|, \quad (k = 1, 2, \ldots. N)$$

for $s > 0$. The values of $N_{crit.}$ are shown in fig. 1 for which $|\epsilon_{1,\frac{N+1}{2}}| = |\epsilon_k|$, from which it follows that $|\epsilon_{1,\frac{N+1}{2}}| \gtrless |\epsilon_k|$ for $N \gtrless N_{crit.}$ (k = 1, 2, ...N).

The results obtained using the rectangular residual distribution in the first row along with $R_j = (-1)^{j+1} R_1$ (j=2,....M), are shown in (40) and (46) from which $\sigma^2_{1,\frac{N+1}{2}} < \rho^2_k$ (k=1,2, ... N)

for $s < 0.435$ for all $N \geq 1$. The values of N and s for which $\sigma^2_{1,\frac{N+1}{2}} = \rho^2_R$ are illustrated in fig. 2.

In all cases, the maximum error or standard mean deviation in the error occurs in the first or last row and not in an intermediate row. It can be determined from figs. 1 and 2 where the maximum occurs for given values of N and s. If the maximum occurs in the first row, it is obtained immediately from (35) or (40) depending on the residual distribution assumed in the first row. If the maximum occurs in the last row, the results of (42) and (46) or (44) and (47) must be considered, depending again on the residual distribution assumed in the first row. Now, in comparison with other residual row distributions, $R_j = R_1$ and $R_j = (-1)^{j+1}R_j$ ($j = 2, \ldots. M$) appear to give rise to maximum and minimum error values or vice versa in the last row, depending on the values of N and s. Consequently for a residual distribution likely to be obtained in practise, it is sufficient, to consider the required value of ρ^2 as lying somewhere between the values given by (44) and (47).

As an illustration, suppose that it is required to estimate the maximum rounding error when $N = 200$ and $s = 0.0001$. A rectangular distribution of residuals will be assumed in the first row, and r is the maximum residual neglected anywhere. From figs. 1 and 2, it follows that the square of the maximum standard mean deviation in the error occurs in the last row, and from (44) and (47), its bounds are shown to be $1.4 \times 10^4 r^2$ and

$8.3 \times 10^6 r^2$.

5. A Stable Explicit Difference Approximation.

In a recent paper, Du Fort and Frankel (26) introduced a difference approximation of the heat conduction equation

$$\phi_{xx} = \phi_t \,, \tag{48}$$

which is both explicit and stable for all values of the mesh ratio. It is obtained by considering initially the replacement

$$\frac{1}{2\Delta t}\left(\phi_{j+1,k} - \phi_{j-1,k}\right) = \frac{1}{(\Delta x)^2}\left(\phi_{j,k-1} - 2\phi_{j,k} + \phi_{j,k+1}\right), \tag{49}$$

together with the condition

$$\phi_{j,k} = \tfrac{1}{2}\left(\phi_{j+1,k} + \phi_{j-1,k}\right).$$

The latter is the difference replacement of $\phi_{tt} = 0$, and (49) is modified to give the "diamond" formula

$$\phi_{j+1,k} = \frac{2}{s+2}\left(\phi_{j,k-1} + \phi_{j,k+1}\right) + \frac{s-2}{s+2}\phi_{j-1,k}. \tag{50}$$

Substituting $e^{\alpha t}\, e^{i\beta x}$ in (50), and simplifying, the result

$$e^{\alpha\Delta t} = \frac{\frac{2}{5}\cos\beta\Delta x \pm \left(1 - \frac{4}{5^2}\sin^2\beta\Delta x\right)^{1/2}}{1 + \frac{2}{5}}$$

is obtained. The condition (25) gives stability for all s.

The difference equation (50) is second order in t, and so ϕ must be known at nodes in the two initial rows in order to carry out the calculation. The value of ϕ at nodes in the first row is given as a boundary condition and ϕ at nodes in the second row is obtained from it by using a conventional difference equation, first order in t. The difference approximation (49) is really the conventional replacement of the hyperbolic equation

$$\frac{1}{c^2}\phi_{tt} + \phi_t = \phi_{xx},$$

where Δx and Δt have approached zero with constant ratio $c = \Delta x/\Delta t$. In the next section, the "diamond" difference formula, although stable for all values of the mesh ratio, is shown to give inaccurate solutions for small s due to truncation errors.

6. Truncation Errors.

If Φ is the exact solution of the differential equation (48), then substituting in (50), the truncation error $T_{j,k}$ in the "diamond" formula is given by

$$\Phi_{j+1,k} = \frac{2}{s+2}\left(\Phi_{j,k-1} + \Phi_{j,k+1}\right) + \frac{s-2}{s+2}\Phi_{j-1,k} + T_{j,k}. \quad (51)$$

Subtracting (50) from (51), the error $(\Phi-\phi)$ due to truncation is denoted by e and satisfies the difference equation

$$e_{j+1,k} = \frac{2}{s+2}\,(e_{j,k-1} + e_{j,k+1}) + \frac{s-2}{s+2}\,e_{j-1,k} + T_{j,k}\,,$$

where e is of course zero at nodes on the boundary of the region. Now the dominant terms in $T_{j,k}$ are given by

$$T_{j,k} = \frac{2}{2+s}\left[\,(\Delta t)^2\,\Phi_{tt} - \frac{1}{12}(\Delta x)^4\,\Phi_{xxxx}\,\right]$$

$$= \frac{12-s^2}{6(2+s)}\,(\Delta t)^2\,\Phi_{tt}$$

$$= \frac{12-s^2}{6s^2(2+s)}\,(\Delta x)^4\,\Phi_{tt}\,,$$

since the heat conduction equation implies $\Phi_{tt} = \Phi_{xxxx}$. Thus the errors due to truncation will be smallest when $s = 12^{1/2}$.

To examine the errors due to truncation for small s (large Δt) it is necessary to consider the forms of solution of (48) and (50). Using the method of separation of variables, the differential equation (48) is solved to give

$$\Phi = \sum_{\beta} e^{-\beta^2 t}\,(A_\beta e^{i\beta x} + B_\beta e^{-i\beta x}). \qquad (52)$$

If there are no boundaries of the problem in the x-direction, then β takes all values. If, however, the extent of the problem in the x-direction is L and $\Phi = 0$ at $x = 0, L$, the frequencies β of period 2L are given by

$$\beta = \frac{r\pi}{L}\,, \qquad (r = 0, 1, 2, \ldots\ldots)$$

and the solution becomes

$$\Phi(x,t) = \sum_{r = 1,2,\ldots} C_r e^{-\frac{r^2\pi^2}{L^2}t} \sin\frac{r\pi}{L}x. \qquad (53)$$

The difference equation (50) is also solved by separation of variables to give

$$\phi_{j,k} = \sum_{\alpha}\left(A_\alpha e^{ik\alpha} + B_\alpha e^{-ik\alpha}\right)\left(C_\alpha \mu_1^{\,j} + D_\alpha \mu_2^{\,j}\right) \qquad (54)$$

where

$$\mu_1, \mu_2 = \frac{2\sigma\cos\alpha \pm \left(1 - 4\sigma^2\sin^2\alpha\right)^{1/2}}{1+2\sigma}$$

with $\sigma = 1/s$. Again if the problem has no boundaries in the x - direction, α takes all real values. If the problem is limited in the x- direction to an extent L with N internal nodes and $\phi_{j,o} = \phi_{j,N+1} = 0$, then

$$\alpha = \frac{r\pi}{N+1}, \qquad (r = 1, 2, \ldots N)$$

and (54) becomes

$$\phi_{j,k} = \sum_{r=1}^{N} \left(C_r\mu_1^{\,j} + D_r\mu_2^{\,j}\right) \sin\frac{kr\pi}{N+1} \qquad (55)$$

with

$$\mu_1, \mu_2 = \frac{2\sigma \cos \frac{r\pi}{N+1} \pm (1 - 4\sigma^2 \sin^2 \frac{r\pi}{N+1})^{1/2}}{1 + 2\sigma} \qquad (56)$$

In a problem of finite extent in the x - direction, the error due to truncation is the value of $\Phi - \phi$ obtained from (53) and (55). The constants C_r in (53) are obtained from $\Phi(x,o)$, whilst the constants C_r and D_r in (55) are obtained from $\phi_{o,k}$ and $\phi_{1,k}$. Evaluation of these constants is a tedious business, and as a new set is required for each set of boundary conditions, a rough guide to the effect of the truncation error will now be given without evaluating the constants.

In (53), the term $e^{-\frac{r^2\pi^2}{L^2} t}$ can be expressed as

$$e^{-\frac{r^2\pi^2}{L^2} t} = e^{-\frac{r^2\pi^2}{L^2} j\Delta t} = (e^{-\frac{r^2\pi^2}{L^2} \sigma (\Delta x)^2})^j = K^j$$

where

$$K = e^{-\frac{r^2\pi^2}{(N+1)^2} \sigma} \qquad (57)$$

This factor K is the amount by which $\sin \frac{r\pi}{L} x$ is attenuated in time Δt by the differential equation. In table VII, K is compared with μ_1, and μ_2 as given by (56), for three values of σ. It should be noticed from (56) that μ_1 and μ_2 are complex for $\sin \frac{r\pi}{N+1} > \frac{1}{2\sigma}$ and are of common modulus $\left[(2\sigma - 1)/(2\sigma + 1)\right]^{1/2}$. When μ_1 and μ_2 are complex, the

$\sigma = 0.10$

$x/_{N+1}$	μ_1	μ_2	K
0	1	-0.67	1
0.25	0.945	-0.71	0.94
0.50	0.82	-0.82	0.78
0.75	0.71	-0.945	0.58
1	0.67	-1	0.37

$\sigma = 1.0$

$x/_{N+1}$	μ_1	μ_2	K
0	1	0.33	1
0.1	0.90	0.37	0.91
0.2	0.58	0.58	0.67
0.3	0.58	0.58	0.41
0.4	0.58	0.58	0.21
0.5	0.58	0.58	0.085

$\sigma = 10.0$

$x/_{N+1}$	μ_1	μ_2	K
0	1	0.91	1
0.03	0.95	0.95	0.92
0.06	0.95	0.95	0.70
0.09	0.95	0.95	0.45
0.12	0.95	0.95	0.24
0.15	0.95	0.95	0.11

TABLE VII

modulus is the value included in the table, and a value zero in the table means a value less than 10^{-3}. Although the higher order of the "diamond" formula, resulting in two attenuation factors μ_1 and μ_2, complicates comparison with the differential equation, it is obvious from table VII that the difference replacement (50) will be of unacceptable accuracy except for small σ. (cf. errors due to truncation are smallest when $\sigma = 1/12^{\frac{1}{2}}$.) Du Fort and Frankel pointed out that for sufficiently large j, μ_1^j, μ_2^j, and K^j are essentially zero. This of course does not compensate for the deficiencies of the "diagonal" formula already mentioned.

The errors due to truncation will now be examined for two other stable difference formulae, von Neumann's six point replacement and the backward four point formula. Again if Φ is the exact solution of (48), the truncation error $T_{j,k}$ in von Neumann's formula is given by

$$\Phi_{j,k-1} - 2(1+s)\Phi_{j,k} + \Phi_{j,k+1} = -\Phi_{j-1,k-1} + 2(1-s)\Phi_{j-1,k} - \Phi_{j-1,k+1} + T_{j,k} \quad (58)$$

where the dominant terms in $T_{j,k}$ are given by

$$T_{j,k} = \tfrac{1}{4}(\Delta x)^2(\Delta t)^2\left[\Phi_{xxtt} - \tfrac{1}{3}\Phi_{ttt}\right] + \tfrac{1}{6}(\Delta x)^4\Phi_{xxxx}$$

$$= \tfrac{s}{6}(\Delta t)^3\Phi_{ttt} + \tfrac{1}{6}(\Delta x)^4\Phi_{tt}$$

$$= \frac{1}{6s^2}(\Delta x)^6\Phi_{xxxxxx} + \tfrac{1}{6}(\Delta x)^4\Phi_{xxxx}.$$

This expression involves a term in Φ_{xxxx} which depends only on Δx, and so the errors due to truncation may be small when Δx is small. To verify this von Neumann's difference equation is solved to give

$$\phi_{j,k} = \sum_{\alpha} (A_\alpha e^{ik\alpha} + B_\alpha e^{-ik\alpha}) \left(\frac{s - 1 + \cos\alpha}{s + 1 - \cos\alpha}\right)^j ,$$

which, if the problem is restricted in the x-direction to an extent L with N internal nodes and $\phi_{j,o} = \phi_{j,N+1} = 0$, becomes

$$\phi_{j,k} = \sum_{r=1}^{N} C_r K_1^j \sin \frac{kr\pi}{N+1} , \tag{69}$$

where

$$K_1 = \frac{s - 1 + \cos\frac{r\pi}{N+1}}{s + 1 - \cos\frac{r\pi}{N+1}} . \tag{70}$$

Again the factor K_1 is the amount by which $\sin\frac{kr\pi}{N+1}$ is attenuated in time Δt by the difference equation.

With the four point backward difference formula the truncation error $T_{j,k}$ is given by

$$\Phi_{j,k-1} - (2 + s)\Phi_{j,k} + \Phi_{j,k+1} = -s\Phi_{j-1,k} + T_{j,k}$$

where the dominant terms in $T_{j,k}$ are given by

$$T_{j,k} = (\Delta x)^2 \left[\tfrac{1}{12}(\Delta x)^2 \Phi_{xxxx} + \tfrac{1}{2}(\Delta t)\Phi_{tt}\right]$$

$$= \frac{s(s+6)}{12} (\Delta t)^2 \Phi_{tt}$$

$$= \frac{s+6}{12s} (\Delta x)^4 \Phi_{xxxx} .$$

It appears from this expression that the errors due to truncation may be large when s is small. To test this statement, the backward difference formula is solved to give

$$\phi_{j,k} = \sum_{\alpha} (A_\alpha e^{ik\alpha} + B_\alpha e^{-ik\alpha}) \left(\frac{s}{s + 2 - 2\cos\alpha}\right)^j , \qquad (71)$$

which, if the problem is restricted in the x-direction to an extent L with N internal nodes and $\phi_{j,o} = \phi_{j,N+1} = 0$, becomes

$$\phi_{j,k} = \sum_{r=1}^{n} C_r K_2^{\,j} \sin \frac{kr\pi}{N+1} , \qquad (72)$$

where

$$K_2 = \frac{s}{s + 2 - 2\cos\frac{r\pi}{N+1}} . \qquad (73)$$

In table VIII, K_1 and K_2 are compared with K. With $\sigma = 1.0$ and 10.0, $r/_{N+1}$ is considered only as far as 0.5 and 0.15 respectively. The values of K_1, K_2, and K given in the table are the important terms in the series (69), (72) and (63) respectively. The table shows that the errors due to truncation are smaller with von Neumann's replacement than with the backward difference formula for the three values of σ considered. This is in agreement with the preliminary survey of the truncation

errors where only the dominant terms in the truncation are considered.

To sum up the results of this section so far, it appears that although von Newmann's replacement, the "diamond" formula introduced by Du Fort and Frankel, and the four point backward difference formula are stable replacements of the heat conduction equation for all values of s, only the first and to a lesser degree the third can take advantage of this stability for small s. (large σ). As shown in table VII, truncation errors are serious in the "diamond" formula even at s = 1. Again it will be emphasized that this approximate investigation will be most accurate for large values of N. (small Δx).

The study of truncation errors in this section is a very approximate one, and there is need for an exact solution of the error ($\Phi - \phi$) due to truncation for say von Neumann's difference approximation to the flow of heat equation. If this exact solution were known, it would be theoretically possible to study the magnitude of the error due to truncation for various choices of s and Δx. Without this exact solution it can only be presumed (see table VIII) that the use of von Neumann's six point formula for small Δx and with a value of s, say as low as 0.1, will not introduce truncation errors of unacceptable magnitude. This value of 0.1 compares with the value of 2.0 for the four point forward difference replacement, and so, for the same Δx, the time interval using the implicit

σ = 0·10 (s = 10·0)

r/N+1	K_1	K_2	K
0	1	1	1
0·25	0·95	0·95	0·94
0·50	0·82	0·84	0·78
0·75	0·71	0·75	0·58
1	·67	0·72	0·37

σ = 1·0 (s = 1·0)

r/N+1	K_1	K_2	K
0	1	1	1
0·1	0·91	0·91	0·91
0·2	0·68	0·72	0·67
0·3	0·42	0·55	0·41
0·4	0·18	0·42	0·21
0·5	0	0·33	0·09

σ = 10·0 (s =0·1)

r/N+1	K_1	K_2	K
0	1	1	1
0·03	0·91	0·91	0·92
0·06	0·69	0·72	0·70
0·09	0·43	0·55	0·45
0·12	0·18	0·42	0·24
0·15	0·03	0·33	0·11

TABLE VIII

formula is twenty times the time interval using the explicit formula. For a moderate number of nodes at each step, this increased time interval may well offset the additional labour required in obtaining a solution by relaxation, and so present a case for the use of implicit formulae.

It has been assumed in this section that for a given value of Δx the most desirable value of s for a given approximation formula is the smallest stable value of s that leads to a preassigned upper bound of error due to truncation. Blanch (27) shows that in certain cases the smallest stable value of s consistent with a preassigned upper bound of truncation error is not necessarily most economical and that a great deal depends on the form of the differential system and the boundary conditions. The difference replacements of the heat conduction equation studied by Blanch are the four point forward difference equation ((30) with $f = 0$), and a six point forward difference replacement (given by (75)), both explicit and stable over a restricted range of s. Blanch also concludes from the study that higher order difference approximations are worth while if solutions are required consistent with a preassigned upper bound of error due to truncation.

So far no mention has been made of the convergence of the exact solution of a difference replacement of the heat conduction equation to the solution of the differential equation. Courant, Friedricks, and Lewy (28) found that the solution of the difference equation (30) with $f = 0$ converges to the required

solution in the half plane $t > 0$ provided $s = 2$. Leutert (29) constructed a solution of (30) which satisfies the boundary conditions $\phi = 0$ at $x = 0,1$ for $t > 0$ and which for all values of s, converges to the solution of the differential equation, with the correct boundary condition on $t = 0$ $(0 < x < 1)$, as the mesh size diminishes. However when $s < 2$, convergence cannot be realized in practise, since in numerical calculation round-off errors will grow due to the instability of the difference approximation. In another recent paper, Hildebrand (30) establishes the convergence of (30) for $\phi = 0$ on $x = 0,1$ $(t > 0)$, $\phi = f(x)$ on $t = 0$ $(0 < x < 1)$, for the mesh ratio range $s \geq 2$, the function $f(x)$ being severly restricted when $s = 2$. Finally John (31) gives sufficient conditions under which the solution of the difference equation

$$\phi_{j+1,k} = \sum_{r} C^{r}_{j,k} \phi_{j,k+r} + \Delta t \, \alpha_{j,k}$$

convergences to the solution of the linear differential equation

$$\phi_t = a_0(x,t)\phi_{xx} + 2a_1(x,t)\phi_x + a_2(x,t)\phi + d(x,t),$$

where the C^r are suitable coefficients and r may go from $-\infty$ to $+\infty$.

This section will be concluded with a statement concerning the stability of a certain type of difference approximation to the flow of heat equation. John defines the solution $\phi_{j,k}$

of an approximating difference equation to be stable if it is bounded for all finite t, independently of Δx. Using this definition, he establishes that a sufficient though not necessary condition for the stability of ϕ in the difference equation

$$\phi_{j+1,k} = \sum_{w=-p}^{p} a_w \phi_{j,k+w} \qquad (74)$$

is that all the coefficients a_w should be non negative. When $p = 1$, (74) becomes

$$\phi_{j+1,k} = \frac{1}{s}\left(\phi_{j,k-1} + \phi_{j,k+1}\right) + \left(1 - \frac{2}{s}\right)\phi_{j,k}$$

and so the above condition gives the well known result $s \geq 2$ for stability. When $p = 2$, (74) becomes

$$\phi_{j+1,k} = \frac{1}{s^2}\left[\left(s^2 - \frac{5}{2}s + 3\right)\phi_{j,k} + \left(\frac{4}{3}s - 2\right)\left(\phi_{j,k-1} + \phi_{j,k+1}\right)\right.$$

$$\left. + \left(\frac{1}{12}s - \frac{1}{12}s\right)\left(\phi_{j,k-2} + \phi_{j,k+2}\right)\right], \qquad (75)$$

where all the coefficients are positive if $\frac{5}{2} \leq s \leq 6$. For general p, the coefficients a_w in (74) are obtained as follows;

(1) expand in a Taylor series $\phi_{j+1,k} = \phi_{j,k} + \sum_{r=1}^{p} \frac{(\Delta t)^r}{r!}\left(\frac{\partial^r \phi}{\partial t^r}\right)_{j,k}$

(2) put $\left(\frac{\partial^r \phi}{\partial t^r}\right)_{j,k} = \left(\frac{\partial^{2r} \phi}{\partial x^{2r}}\right)_{j,k}$ for $2 \leq r \leq p$, giving

$$\left(\frac{\partial \phi}{\partial t}\right)_{j,k} = \frac{1}{\Delta t}\left(\phi_{j+1,k} - \phi_{j,k}\right) - \sum_{r=2}^{p} \frac{(\Delta t)^{r-1}}{r!}\left(\frac{\partial^{2r}\phi}{\partial x^{2r}}\right)_{j,k} .$$

(5) substitute the above value in the flow of heat equation

$$\left(\frac{\partial \phi}{\partial t}\right)_{j,k} = \left(\frac{\partial^2}{\partial x^2}\right)_{j,k} , \text{ and replace } \left(\frac{\partial^{2r}\phi}{\partial x^{2r}}\right)_{j,k} \quad (r=1,2,\ldots..p)$$

by the appropriate central difference replacements of order 2p. It appears likely without doing an excessive amount of calculation that increasing the value of p in (74) gives a slight improvement in stability, a marked improvement in errors due to truncation (cf. Blanch), and a considerable increase in calculation time. The restricted range of stability will prevent full advantage being taken of the small truncation errors for high values of p, and so it is unlikely that values of p above 3 will be considered economical.

Another consequence of the definition of stability given by John is that the difference equation satisfied by the error $(\Phi - \phi)$ due to truncation can exist only if the finite difference approximation satisfied by ϕ is stable.

Most finite difference solutions of two variable parabolic problems, whether linear or non linear, use a difference solution of $\phi_{xx} = \phi_t$ in one form or another (see section I). Accordingly a summary will now be given of the properties of the more useful difference approximations to the flow of heat equation.

Summary of Difference Replacements of $\phi_t = \phi_{xx}$.

(A) $$\phi_{j+1,k} = (1 - \tfrac{2}{s})\,\phi_{j,k} + \tfrac{1}{s}(\phi_{j,k-1} + \phi_{j,k+1}).$$

(B) $$\phi_{j,k+1} - 2(1+s)\,\phi_{j,k} + \phi_{j,k-1} = -\phi_{j-1,k+1} + 2(1-s)\,\phi_{j-1,k} - \phi_{j-1,k-1}.$$

(C) $$\phi_{j,k+1} - (2+s)\,\phi_{j,k} + \phi_{j,k-1} = -s\,\phi_{j-1,k}$$

(D) $$\phi_{j+1,k} = \tfrac{1}{s^2}\left[(s^2 - \tfrac{5}{2}s + 3)\,\phi_{j,k} + (\tfrac{4}{3}s - 2)(\phi_{j,k-1} + \phi_{j,k+1}) + (\tfrac{1}{2} - \tfrac{1}{12}s)(\phi_{j,k-2} + \phi_{j,k+2})\right].$$

(E) $$\phi_{j+1,k} = \tfrac{2}{s+2}(\phi_{j,k-1} + \phi_{j,k+1}) + \tfrac{s-2}{s+2}\,\phi_{j-1,k}.$$

Replacement	Type	Stability	Principal Truncation Term	Round-off Errors
A	Explicit	$s \geq 2$	$\frac{6-s}{12s^2}(\Delta x)^4 \phi_{xxxx}$	Small
B	Implicit	$s > 0$	$\frac{1}{6}(\Delta x)^4 \phi_{xxxx}$	pp. 84-91
C	Implicit	$s > 0$	$\frac{6+s}{12s^2}(\Delta x)^4 \phi_{xxxx}$	pp. 82, 83
D	Explicit	$6 \geq s \geq \frac{3}{2}$	$\frac{6-s}{12s^2}(\Delta x)^4 \phi_{xxxx}$	Small
E	Explicit	$s > 0$	$\frac{12-s^2}{6s^2(2+s)}(\Delta x)^4 \phi_{xxxx}$	Small

7. Relaxation Methods (in the time co-ordinate) applied to Step-by-Step Problems.

All the difference replacements of the flow of heat equation so far outlined have been step-by-step in the time co-ordinate. This is not surprising considering the boundary conditions usually consist of a knowledge of ϕ or its normal derivative along three sides of a rectangle open in the t-direction. The natural procedure is to step off the solution from the boundary values. If the difference replacement used is implicit (good stability properties), a relaxation solution in x is necessary at each value of t. In this section the merits of difference solutions will be discussed which are relaxational in t.

Allen and Severn (6) considered relaxation methods of solving the equation $\phi_t = \phi_{xx}$ in the region $0 < x < L$, $0 < t < T$, subject to the boundary conditions $\phi = 0$ at $x = 0$, $t = 0$, and $\phi = 100$ at $x = L$, where for convenience T is taken to be $L^2/10$. By making the substitution

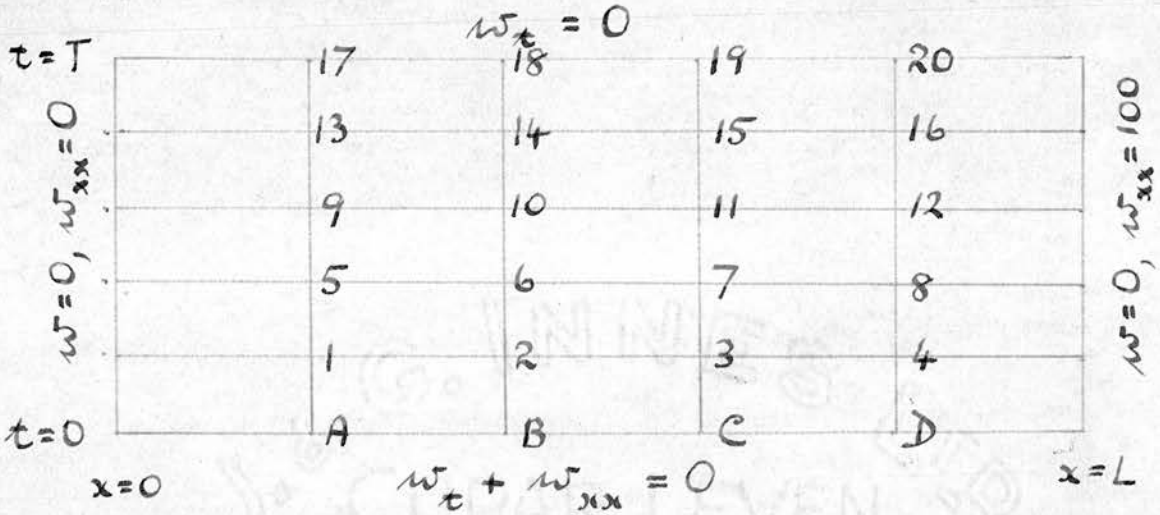
$$\phi = w_t + w_{xx} ,$$

the heat conduction equation becomes

$$w_{tt} = w_{xxxx} , \tag{76}$$

together with the boundary conditions $w_t + w_{xx} = 0$ on $x = 0$, $t = 0$, and $w_t + w_{xx} = 100$ on $x = L$. Since the equation for w is second order in t and fourth order in x, Allen and Severn introduced the additional arbitrary boundary conditions $w = 0$

on x = 0, L and $w_t = 0$ on t = T, and so were able to solve (76) by relaxation. The results of this section will be illustrated with respect to the coarse mesh illustrated below with s = 2. The approximation to



(76) used by Allen and Severn is

$$\frac{1}{(\Delta t)^2}(w_{j-1,k}-2w_{j,k}+w_{j+1,k}) = \frac{1}{(\Delta x)^4}(w_{j,k-2}-4w_{j,k-1}+6w_{j,k}-4w_{j,k+1}+w_{j,k+2})$$

and the values of w at the node points are given by the matrix equation

$$\begin{bmatrix} P & 8I & & & & \\ 4I & Q & 4I & & & \\ & 4I & Q & 4I & & \\ & & 4I & Q & 4I & \\ & & & 4I & Q & 4I \\ & & & & 8I & Q \end{bmatrix} \begin{bmatrix} W_0 \\ W_1 \\ W_2 \\ W_3 \\ W_4 \\ W_5 \end{bmatrix} = \begin{bmatrix} C \\ C \\ C \\ C \\ C \\ C \end{bmatrix}, \qquad (77)$$

where

$$P = \begin{bmatrix} -21 & 8 & -1 & \\ 8 & -22 & 8 & -1 \\ -1 & 8 & -22 & 8 \\ & -1 & 8 & -21 \end{bmatrix}, \qquad Q = \begin{bmatrix} -13 & 4 & -1 & -1 \\ 4 & -14 & 4 & -1 \\ -1 & 4 & -14 & 4 \\ & -1 & 4 & -13 \end{bmatrix},$$

$$I = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \quad C = 100\,(\Delta x)^2 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad \text{and} \quad W_o = \begin{bmatrix} w_A \\ w_B \\ w_C \\ w_D \end{bmatrix} \quad \text{etc.}$$

together with $w = 0$ at nodes on $x = 0$, L. The values of ϕ at the nodes are then obtained from

$$\begin{bmatrix} \Phi_1 \\ \Phi_2 \\ \Phi_3 \\ \Phi_4 \\ \Phi_5 \end{bmatrix} = \begin{bmatrix} -I & R & I & & \\ & -I & R & I & \\ & & -I & R & I \\ & & & -I & R \end{bmatrix} \begin{bmatrix} W_0 \\ W_1 \\ W_2 \\ W_3 \\ W_4 \\ W_5 \end{bmatrix} \tag{78}$$

where $R = \begin{bmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & 1 & -2 & 1 \\ & & 1 & -2 \end{bmatrix}$ $\qquad \Phi_1 = (\Delta x)^2 \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{bmatrix}$ etc.,

together with $\phi = 0$ at $x = 0$, $t = 0$ and $\phi = 100$ at $x = L$.

Mitchell and Rutherford (7) again pointed out that Allen and Severn need not have doubled the order of the differential equation in order to use relaxation. Using

$$\phi_{j-2,k} - 4\phi_{j-1,k} + 3\phi_{j,k} = \phi_{j,k-1} + \phi_{j,k+1} - 2\phi_{j,k}$$

at nodes 17, 18, 19, 20, and

$$\phi_{j+1,k} - \phi_{j-1,k} = \phi_{j,k-1} + \phi_{j,k+1} - 2\phi_{j,k} \qquad (79)$$

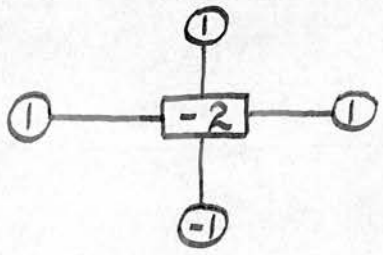
at all other nodes, the values of ϕ at the nodal points are given by the matrix equation

$$\begin{bmatrix} R & -I & & & \\ I & R & -I & & \\ & I & R & -I & \\ & & I & R & -I \\ & & -I & 4I & S \end{bmatrix} \begin{bmatrix} \Phi_1 \\ \Phi_2 \\ \Phi_3 \\ \Phi_4 \\ \Phi_5 \end{bmatrix} = \begin{bmatrix} D \\ D \\ D \\ D \\ D \end{bmatrix} \qquad (80)$$

where

$$S = \begin{bmatrix} -5 & 1 & & \\ 1 & -5 & 1 & \\ & 1 & -5 & 1 \\ & & 1 & -5 \end{bmatrix}, \quad D = -100 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad \text{and} \quad \Phi_1 = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{bmatrix} \quad \text{etc.}$$

Of the twenty nodes in this net only those numbered 6 and 7 are typical in that their pattern, illustrated below, is unaffected by the presence of the boundary.



When, however, the size of the mesh is decreased, the great majority of the nodes will have a pattern of this type. In all there will be fifteen types of relaxation pattern used, but eight of these will be used at one node only. Again the relaxation pattern at any node can easily be read off from the appropriate column of the matrix on the left hand side of (80). The difficulty mentioned by Allen and Severn of moving residuals in the t-direction using the illustrated pattern does not arise when the latter is used on a block of nodal points. To illustrate this, a rectangle of twenty five nodes is considered with a temperature and corresponding residual specified at each node. The pattern is used only on the inner rectangle of nine nodes in order to keep the total of the residuals at the twenty five nodal points constant. The initial and revised temperatures with their corresponding residuals, shown in fig. 5, illustrate that a considerable change in the residual distribution takes place in both the x- and t- directions, although the residual total over the twenty five nodes remains constant.

In order to provide a comparison, the above problem is worked out using the stable approximation

$$\phi_{j+1,k} - \phi_{j-1,k} = \tfrac{1}{2}\left[\left(\phi_{j+1,k-1} + \phi_{j+1,k+1} - 2\phi_{j+1,k}\right) + \left(\phi_{j-1,k-1} + \phi_{j-1,k+1} - 2\phi_{j-1,k}\right)\right] \tag{81}$$

The latter formula is of course step-by-step in the time co-ordinate and relaxational in distance co-ordinate. Of the results, sufficient to say that the calculations made using (80)

and (81) together with the appropriate boundary conditions, are in close agreement with the results of Allen and Severn. As with the corresponding problem in ordinary difference equations (p. 19), there seems no reason to suppose that relaxation is superior to stable step-by-step methods in dealing with "marching" problems. In parabolic problems, however, where only unstable step-by-step formulae are available, doubling the order of the equation and relaxing in the manner of Fox (8) or using a higher order difference equation and relaxing in the manner of the present author (9), may be the answer.

8. The Heat Conduction Equation in Two Dimensions.

In two dimensions, the parabolic heat conduction equation is

$$\phi_{xx} + \phi_{yy} - \phi_t = 0. \qquad (82)$$

In this section it will be assumed for convenience that ϕ is given without round-off error on $x = 0, L$, $y = 0, L$, and $t = 0$. The solution is required in the region $0 \leq x \leq L$, $0 \leq y \leq L$, $t \geq 0$. Suppose this region is covered by an orthogonal mesh, the mesh lengths being h in the x- and y- directions, and Δt in the t- direction. If j and k are the row and column numbers in the (x,y) plane, and l is the number of the square section parallel to the (x,y) plane, a simple difference replacement of (82) is

$$\frac{1}{h^2}\left(\phi_{j,k-1,\ell} + \phi_{j,k+1,\ell} + \phi_{j-1,k,\ell} + \phi_{j+1,k,\ell} - 4\phi_{j,k,\ell}\right)$$

$$= \frac{1}{\Delta t}\left(\phi_{j,k,\ell+1} - \phi_{j,k,\ell}\right). \qquad (83)$$

For the purpose of examining the stability of the six point forward difference formula (83), consider a distribution of errors $E(x,y)$ introduced at $t = 0$. A harmonic decomposition

$$E(x,y) = \sum_{p,q} A_{p,q}\, e^{i\beta_p x}\, e^{i\gamma_q y} \qquad (84)$$

is made of the errors, where the sum (84) must reduce to the correct error value at each mesh point on $t = 0$. The frequencies $|\beta_p|$, $|\gamma_q|$ and p,q are in general somewhat arbitrary. Considering only the single term $e^{i\beta x} e^{i\gamma y}$ where β and γ are real numbers, a solution of (83) which reduces to this value when $t = 0$ is

$$e^{\alpha t}\, e^{i\beta x}\, e^{i\gamma y}$$

where $\alpha = \alpha(\beta,\gamma)$. Again the original error $e^{i\beta x}\, e^{i\gamma y}$ will not grow if

$$\left|e^{\alpha \Delta t}\right| \leq 1$$

for all α. Substituting $e^{\alpha t}\, e^{i\beta x}\, e^{i\gamma y}$ in (83) and simplifying, the result

$$e^{\alpha \Delta t} = 1 - \frac{4}{s}\left(\sin^2 \frac{\beta \Delta x}{2} + \sin^2 \frac{\gamma \Delta y}{2}\right)$$

is obtained where $s = h^2/\Delta t$, and $\Delta x = \Delta y = h$. Thus the condition for stability yields

$$-1 \leqslant 1 - \frac{4}{s}\left(\sin^2 \frac{\beta \Delta x}{2} + \sin^2 \frac{\gamma \Delta y}{2}\right) \leqslant 1,$$

which simplifies to $s \geqslant 4$ for all β and γ. This is also the condition for convergence given by Courant, Friedrichs and Lewy (28).

If Φ is the exact solution of the differential equation (82), then substituting in (83), the truncation error $T_{j,k,\ell}$ in the six point forward difference formula is given by

$$\Phi_{j,k,\ell+1} = \frac{s-4}{s}\Phi_{j,k,\ell} + \frac{1}{s}\left(\Phi_{j,k+1,\ell} + \Phi_{j,k-1,\ell} + \Phi_{j+1,k,\ell} + \Phi_{j-1,k,\ell}\right) + T_{j,k,\ell} \quad (85)$$

where the dominant terms in $T_{j,k,\ell}$ are given by

$$T_{j,k,\ell} = \tfrac{1}{2}(\Delta t)^2\left[\Phi_{tt} - \frac{s}{6}\left(\Phi_{xxxx} + \Phi_{yyyy}\right)\right]$$

$$= \frac{1}{2s^2} h^4\left[\frac{6-s}{6}\Phi_{tt} + \frac{s}{3}\Phi_{xxyy}\right].$$

This truncation error will be specially large when s is small, but of course there is no danger of a large truncation error since the condition for stability is $s \geqslant 4$. When $s = 4$, neglecting $T_{j,k,\ell} = \frac{h^4}{96}\left[\Phi_{xt} + 4\Phi_{xxyy}\right]$, (85) becomes the simple stable approximation

$$\phi_{j,k,\ell+1} = \frac{1}{4}\left(\phi_{j,k+1,\ell} + \phi_{j,k-1,\ell} + \phi_{j+1,k,\ell} + \phi_{j-1,k,\ell}\right). \quad (86)$$

A useful forward difference formula given by Milne (1) is

$$\phi_{j,k,\ell+1} = \frac{1}{36}\Big[\left(\phi_{j+1,k+1,\ell} + \phi_{j+1,k-1,\ell} + \phi_{j-1,k+1,\ell} + \phi_{j-1,k-1,\ell}\right)$$

$$+ 4\left(\phi_{j+1,k,\ell} + \phi_{j-1,k,\ell} + \phi_{j,k+1,\ell} + \phi_{j,k-1,\ell}\right) + 16\phi_{j,k,\ell}\Big].$$

This approximation implies s = 6, and although requiring more calculation than (86), it has smaller truncation errors and so gives greater accuracy.

Now if Δt is to be chosen at the computer's convenience, the difference approximation will have to be stable for all s. Accordingly an implicit formula, corresponding to von Neumann's six point formula, will now be described. It is

$$\frac{1}{\Delta t}\left(\phi_{j,k,\ell} - \phi_{j,k,\ell-1}\right) = \frac{1}{2h^2}\Big[\phi_{j+1,k,\ell} + \phi_{j-1,k,\ell} + \phi_{j,k+1,\ell} + \phi_{j,k-1,\ell}$$

$$- 4\phi_{j,k,\ell} + \phi_{j+1,k,\ell-1} + \phi_{j-1,k,\ell-1} + \phi_{j,k+1,\ell-1} + \phi_{j,k-1,\ell-1} - 4\phi_{j,k,\ell-1}\Big]. \quad (87)$$

Substituting $e^{\alpha t}\, e^{i\beta x}\, e^{i\gamma y}$ in (87) and simplifying,

$$e^{\alpha\Delta t} = \frac{s - 2\left[\sin^2\frac{\beta\Delta x}{2} + \sin^2\frac{\gamma\Delta y}{2}\right]}{s + 2\left[\sin^2\frac{\beta\Delta x}{2} + \sin^2\frac{\gamma\Delta y}{2}\right]}$$

is obtained, and so (87) is stable for all s. Using (87), a

relaxation solution is necessary at each time step, and so round-off errors of considerable magnitude may arise during the calculation. Taking the results for the von Neumann formula in one dimension as a guide, however, the round-off errors using (87) are unlikely to be large if the number of mesh points in the x, y plane is small.

In order to estimate these errors, the error equations

$$\epsilon_{j,k-1,\ell} + \epsilon_{j-1,k,\ell} + \epsilon_{j,k+1,\ell} + \epsilon_{j+1,k,\ell} - 2(2+s)\epsilon_{j,k,\ell}$$
$$+ \epsilon_{j,k-1,\ell-1} + \epsilon_{j-1,k,\ell-1} + \epsilon_{j,k+1,\ell-1} + \epsilon_{j+1,k,\ell-1} - 2(2+s)\epsilon_{j,k,\ell-1} = r_{j,k,\ell}, \tag{88}$$

are considered where $r_{j,k,l}$ ($j = 1,2, \ldots N$; $k = 1,2,\ldots N$; $l = 1,2,\ldots$) are the residuals in the relaxation solutions. For the case of M time intervals, the error equations (88) can be written as the matrix equation

$$\begin{bmatrix} A & & & \\ B & A & & \\ & \ddots & \ddots & \\ & & B & A \\ & & & B \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_{M-1} \\ E_M \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_{M-1} \\ R_M \end{bmatrix}, \tag{89}$$

where

$$A = \begin{bmatrix} A & I & & & \\ I & A & I & & \\ & \ddots & \ddots & \ddots & \\ & & I & A & I \\ & & & I & A \end{bmatrix}, \quad B = \begin{bmatrix} B & I & & & \\ I & B & I & & \\ & \ddots & \ddots & \ddots & \\ & & I & B & I \\ & & & I & B \end{bmatrix},$$

$$E_\ell = \begin{bmatrix} E_{1,\ell} \\ E_{2,\ell} \\ \vdots \\ E_{N,\ell} \end{bmatrix}, \quad R_\ell = \begin{bmatrix} R_{1,\ell} \\ R_{2,\ell} \\ \vdots \\ R_{N,\ell} \end{bmatrix} \quad (\ell = 1 \ldots M), \quad P(x) = \begin{bmatrix} -x & 1 & & & \\ 1 & -x & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -x & 1 \\ & & & 1 & -x \end{bmatrix}$$

$$E_{j,\ell} = \begin{bmatrix} \epsilon_{j,1,} \\ \epsilon_{j,2,} \\ \vdots \\ \epsilon_{j,N,} \end{bmatrix} \qquad R_{j,\ell} = \begin{bmatrix} r_{j,1,} \\ r_{j,2,} \\ \vdots \\ r_{j,N,} \end{bmatrix} \quad (j = 1 \ldots . N), \quad A = P\left[2(2+s)\right],$$

and $B = P\left[2(2-s)\right]$. From (89), the errors in the first time step are given by

$$E_1 = A^{-1} R_1,$$

an approximate solution of which is given in section 2 of Elliptic Equations. The errors in subsequent time steps are

obtained in the manner of section 3 of Parabolic Equations. Unless N is small, the calculation of the round-off errors is a tedious business.

The truncation error $T_{j,k,\ell}$ in the ten point implicit formula is given by

$$\Phi_{j,k+1,\ell} + \Phi_{j,k-1,\ell} + \Phi_{j+1,k,\ell} + \Phi_{j-1,k,\ell} - 2(2+s)\Phi_{j,k,\ell}$$

$$= -\Phi_{j,k+1,\ell-1} - \Phi_{j,k-1,\ell-1} - \Phi_{j+1,k,\ell-1} - \Phi_{j-1,k,\ell-1} + 2(2-s)\Phi_{j,k,\ell-1} + T_{j,k,\ell}$$

where the dominant terms of $T_{j,k,\ell}$ are given by

$$T_{j,k,\ell} = \tfrac{1}{6}h^4\left[\Phi_{tt} - 2\Phi_{xxyy}\right].$$

This principal part of the truncation error is negligible for sufficiently small h irrespective of the value of the mesh ratio s.

An explicit approximation of (82), stable for all values of the mesh ratio, will now be derived. Consider the replacement

$$\frac{1}{2\Delta t}\left[\phi_{j,k,\ell+1} - \phi_{j,k,\ell-1}\right] = \frac{1}{h^2}\left[\phi_{j,k-1,\ell} + \phi_{j,k+1,\ell} + \phi_{j-1,k,\ell} + \phi_{j+1,k,\ell} - 4\phi_{j,k,\ell}\right], \tag{90}$$

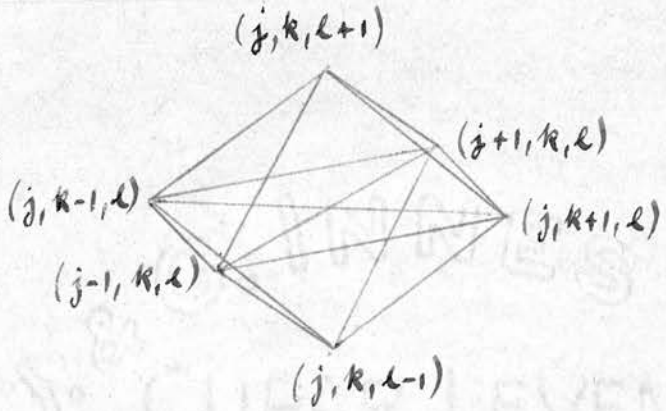
together with the condition

$$\phi_{j,k,\ell} = \tfrac{1}{2}\left[\phi_{j,k,\ell+1} + \phi_{j,k,\ell-1}\right].$$

The latter is the difference replacement of $\phi_{tt} = 0$, and substituting it in (90), the six point formula

$$\phi_{j,k,\ell+1} = \frac{s-4}{s+4}\phi_{j,k,\ell-1} + \frac{2}{s+4}\left[\phi_{j,k-1,\ell} + \phi_{j,k+1,\ell} + \phi_{j-1,k,\ell} + \phi_{j+1,k,\ell}\right] \quad (91)$$

is obtained. Substituting $e^{\alpha t}\, e^{i\beta x}\, e^{i\gamma y}$ in (91) and



simplifying the result

$$e^{\alpha\Delta t} = \frac{(\cos\beta h + \cos\gamma h) \mp \left[(\cos\beta h + \cos\gamma h)^2 - 4 + s^2/4\right]^{1/2}}{2 + s/2}$$

is derived, which shows the explicit approximation (91) to be stable for all s. This explicit stable formula is comparable to the "diamond" replacement (60) of the one dimensional heat conduction equation. As with the "diamond" formula it is expected that (91) despite its stability, will give inaccurate solutions for small s due to truncation errors. With the usual notation, the truncation error in (91) is given by

$$\Phi_{j,k,\ell+1} = \frac{s-4}{s+4}\Phi_{j,k,\ell-1} + \frac{2}{s+4}\left[\Phi_{j,k-1,\ell} + \Phi_{j,k+1,\ell} + \Phi_{j-1,k,\ell} + \Phi_{j+1,k,\ell}\right] + T_{j,k,\ell}$$

where the dominant terms in $T_{j,k,}$ are given by

$$T_{j,k,} = -\frac{1}{s+4}\left[4(\Delta x)^2 \Phi_{tt} - \frac{1}{6}h^4 \left(\Phi_{xxxx} + \Phi_{yyyy}\right)\right]$$

$$= \frac{h^4}{s+4}\left[\left(\frac{4}{s^2} - \frac{1}{6}\right)\Phi_{tt} + \frac{1}{3}\Phi_{xxyy}\right].$$

For moderate values of h, this error will be large when the mesh ratio s is small.

A summary will now be given of the properties of the most useful difference approximations to the heat conduction equation in two dimensions.

Summary of Difference Replacements of $\phi_t = \phi_{xx} + \phi_{yy}$

(A) $$\phi_{j,k,l+1} = \frac{s-4}{s}\phi_{j,k,l} + \frac{1}{s}\left[\phi_{j,k+1,l} + \phi_{j,k-1,l} + \phi_{j+1,k,l} + \phi_{j-1,k,l}\right]$$

(B) $$\phi_{j,k,l+1} = \frac{1}{56}\left[\left(\phi_{j+1,k+1,l} + \phi_{j+1,k-1,l} + \phi_{j-1,k+1,l} + \phi_{j-1,k-1,l}\right)\right.$$

$$\left. + 4\left(\phi_{j+1,k,l} + \phi_{j-1,k,l} + \phi_{j,k+1,l} + \phi_{j,k-1,l}\right) + 16\,\phi_{j,k,l}\right].$$

(C) $$\phi_{j,k+1,l} + \phi_{j,k-1,l} + \phi_{j+1,k,l} + \phi_{j-1,k,l} - 2(2+s)\,\phi_{j,k,l}$$

$$= -\phi_{j,k+1,l-1} - \phi_{j,k-1,l-1} - \phi_{j+1,k,l-1} - \phi_{j-1,k,l-1} + 2(2-s)\phi_{j,k,l-1}$$

(D) $$\phi_{j,k,l+1} = \frac{s-4}{s+4}\phi_{j,k,l-1} + \frac{2}{s+4}\left[\phi_{j,k+1,l} + \phi_{j,k-1,l} + \phi_{j+1,k,l} + \phi_{j-1,k,l}\right.$$

Replacement	Type	Stability	Principal Truncation Term	Round-off Errors
A	Explicit	$s \geq 4$	$\frac{h^4}{2s^2}\left[\frac{6-s}{6}\phi_{xxt}+\frac{s}{3}\phi_{xxyy}\right]$	Small
B	Explicit	Stable	$< \frac{h^4}{56}\phi_{xxyy}$	Small
C	Implicit	$s > 0$	$\frac{h^4}{6}\left[\phi_{xxt} - 2\phi_{xxyy}\right]$	Probably Small if N is small
D	Explicit	$s > 0$	$\frac{h^4}{s+4}\left[\left(\frac{4}{s^2}-\frac{1}{6}\right)\phi_{xxt}+\frac{1}{3}\phi_{xxyy}\right]$	Small

9. <u>The Linear Difference Equation with Variable Coefficients.</u>

A general parabolic linear equation in one dimension is

$$\phi_t = p(x,t)\phi_{xx} + q(x,t)\phi_x + r(x,t)\phi, \qquad (92)$$

where p, q, and r are known functions of x and t. A simple explicit finite difference replacement of (92) is, with the usual notation,

$$\frac{1}{\Delta t}\left(\phi_{j+1,k} - \phi_{j,k}\right) = \frac{p_{j,k}}{(\Delta x)^2}\left(\phi_{j,k+1} - 2\phi_{j,k} + \phi_{j,k-1}\right) + \frac{q_{j,k}}{2\Delta x}\left(\phi_{j,k+1} - \phi_{j,k-1}\right) + r_{j,k}\phi_{j,k},$$

which reduces to

$$\phi_{j+1,k} = \frac{\Delta t}{\Delta x}\left[\frac{p_{j,k}}{\Delta x} + \frac{q_{j,k}}{2}\right]\phi_{j,k+1} + \frac{\Delta t}{\Delta x}\left[\frac{p_{j,k}}{\Delta x} - \frac{q_{j,k}}{2}\right]\phi_{j,k-1}$$

$$+ \left[1 - \frac{2\Delta t}{(\Delta x)^2} p_{j,k} + \Delta t\, r_{j,k}\right]\phi_{j,k}, \qquad (93)$$

a linear difference equation with variable coefficients.

The solution of (93) may be expected to provide a reasonable approximation to the solution of (92) only if (93) is stable for the values of Δx and Δt chosen. No general criterion is available for determining the stability or otherwise of linear difference equations with variable coefficients. If, however, the coefficients are reasonably constant over parts of the region to be considered, von Neumann's criterion for stability (Section 2) can be applied within each part. If the coefficients are not reasonably constant over parts of the region, no general procedure is available, although Du Fort and Frankel (26) give necessary conditions for the stability of (93) as

(1) $p \geq 0$

(2) $|q\,\Delta t/\Delta x| \leq 1$, and

(3) $|r\,\Delta t| \ll 1$.

These conditions are of course not sufficient.

10. Non-linear Equations.

Consider the non-linear equation in one dimension

$$\phi_t = f(\phi, \phi_x, \phi_{xx}, x, t). \qquad (94)$$

A simple explicit finite difference replacement of (94) is

$$\frac{1}{\Delta t}(\phi_{j+1,k} - \phi_{j,k}) = f\left[\phi_{j,k}, \frac{1}{2\Delta x}(\phi_{j,k+1} - \phi_{j,k-1}), \frac{1}{(\Delta x)^2}(\phi_{j,k+1} + \phi_{j,k-1} - 2\phi_{j,k}), k\Delta x, j\Delta t\right] \quad (95)$$

which can be used to obtain a numerical solution of (94). This solution will only be reasonable if (95) is stable for the values of Δx and Δt chosen. No general criterion is available for determining the stability of non-linear equations, and each non-linear problem must be considered independently on its own merits. In some problems, however, the stability arguments for linear equations can be applied over limited regions of the field. Crank and Nicolson (32) and Blanch (27) have examined the stability of difference solutions of a two variable parabolic equation containing a non-linear term. When the non-linear term is absent, the equation reduces to the flow of heat equation in one dimension, and their stability arguments are really based on this reduced equation. Other instances of finite difference methods applied to non-linear parabolic problems will be given in Chapter 5.

HYPERBOLIC EQUATIONS.

1. The Wave Equation.

The hyperbolic wave equation in one dimension is

$$\phi_{xx} - \phi_{tt} = f, \quad (86)$$

where f(x,t) is a known function usually zero. The boundary

conditions consist of a knowledge of ϕ along $x = 0$, L and of ϕ and ϕ_t along $t = 0$. The solution is required in the region $0 \leq x \leq L$, $t \geq 0$.

Two general implicit finite difference replacements of (96) are considered. The first is the asymmetrical difference approximation

$$a\left[\phi_{j,k-1} - 2\phi_{j,k} + \phi_{j,k+1}\right] + (1-a)\left[\phi_{j-2,k-1} - 2\phi_{j-2,k} + \phi_{j-2,k+1}\right]$$

$$= \left(\frac{\Delta x}{\Delta t}\right)^2\left[\phi_{j,k} - 2\phi_{j-1,k} + \phi_{j-2,k}\right] + (\Delta x)^2 f, \quad (j = 2,3,\cdots;\ k = 1,2,\cdots N). \qquad (97)$$

and the second the symmetrical approximation

$$a\left[\phi_{j,k-1} - 2\phi_{j,k} + \phi_{j,k+1}\right] + (1-2a)\left[\phi_{j-1,k-1} - 2\phi_{j-1,k} + \phi_{j-1,k+1}\right]$$

$$+ a\left[\phi_{j-2,k-1} - 2\phi_{j-2,k} + \phi_{j-2,k+1}\right] = \left(\frac{\Delta x}{\Delta t}\right)^2\left[\phi_{j,k} - 2\phi_{j-1,k} + \phi_{j-2,k}\right] + (\Delta x)^2 f; \qquad (98)$$

$$(j = 1,2,\cdots;\ k = 1,2,\cdots N)$$

with $a > 0$ in both cases. The difference equations (97) and (98) are second order in both the time and distance co-ordinates. Except when $a = 0$, ϕ is given implicitly by both (97) and (98), and a relaxation solution is necessary for each row in turn. When $a = 0$, (98) reduces to the five point formula

$$\phi_{j,k} = 2\left(1 - \frac{1}{s^2}\right)\phi_{j-1,k} + \frac{1}{s^2}\left(\phi_{j-1,k-1} + \phi_{j-1,k+1}\right) - \phi_{j-2,k} - (\Delta t)^2 f, \qquad (99)$$

where $s = \Delta x/\Delta t$. Using (99) the solution can be stepped off explicitly from the boundary values of ϕ on the rows $j = 0$, 1.

Now for convenience dividing through (97) and (98) by a, the errors $\epsilon_{j,k}$ in $\phi_{j,k}$ are given by the equation

$$\left[\epsilon_{j,k-1} - \left(2 + \frac{s^2}{a}\right)\epsilon_{j,k} + \epsilon_{j,k+1}\right] + 2\frac{s^2}{a}\epsilon_{j-1,k}$$

$$+ \frac{1-a}{a}\left[\epsilon_{j-2,k-1} - \left(2 + \frac{s^2}{1-a}\right)\epsilon_{j-2,k} + \epsilon_{j-2,k+1}\right] = r_{j,k}, \quad (100)$$

$$(j = 2, 3, \cdots; \; k = 1, 2, \cdots N)$$

in the asymmetrical case, and by

$$\left[\epsilon_{j,k-1} - \left(2 + \frac{s^2}{a}\right)\epsilon_{j,k} + \epsilon_{j,k+1}\right] + \frac{1-2a}{a}\left[\epsilon_{j-1,k-1} - 2\left(1 - \frac{s^2}{1-2a}\right)\epsilon_{j-1,k} + \epsilon_{j-1,k+1}\right]$$

$$+ \left[\epsilon_{j-2,k-1} - \left(2 + \frac{s^2}{a}\right)\epsilon_{j-2,k} + \epsilon_{j-2,k+1}\right] = r_{j,k}, \quad (101)$$

$$(j = 2, 3, \cdots; \; k = 1, 2, \cdots N)$$

in the symmetrical case where $r_{j,k}$ is the residual at the node (j,k). From (100) with $r_{j,k} = 0$, it is easily shown using von Neumann's method of examining stability that (97) is stable for all s if $a \geq \frac{1}{2}$ and unstable for all s if $0 < a < \frac{1}{2}$. From (101) with $r_{j,k} = 0$, (98) is shown to be stable for all s if $a > \frac{1}{4}$ (excluding $a = \frac{1}{2}$) and stable for $s > (1 - 4a)^{\frac{1}{2}}$ if $0 < a \leq \frac{1}{4}$. The five point formula (99) is stable for $s \geq 1$, and with $s = 1$ it reduces to

$$\phi_{j,k} = \phi_{j-1,k-1} + \phi_{j-1,k+1} - \phi_{j-2,k}, \quad (102)$$

where $f = 0$. This formula is particularly easy to use, and it has the useful property that it is satisfied by

$$\phi = F(x - t) + G(x + t),$$

which is the solution of the wave equation (96) with $f = 0$. As

a result, there is no problem of convergence using (102), since V = D regardless of mesh size.

2. Round-off Errors.

In order that Δt can be chosen at the computer's convenience, the wave equation is solved implicitly using (97) with $a \geq \frac{1}{2}$ or using (98) with $a \geq \frac{1}{4}$. The relaxation solution at each step introduces round-off errors and in this section the magnitudes of these errors will be assessed and the value of the parameter a determined which will keep these errors to a minimum. (see Mitchell (24)). For convenience, suppose that no rounding off is required at nodes on x = 0, L and t = 0, Δt. Hence $\epsilon_{j,o} = \epsilon_{j,N+1} = \epsilon_{o,k} = \epsilon_{1,k} = 0$ for all j and k. If rounding is required at these nodes the errors neglected will have small effect on the maximum round-off error. The relaxation process is continued at each step until all the residuals vanish to a prescribed degree of accuracy. If r is the modulus of the maximum residual neglected, then all other residuals lie in the range $-r \leq r_{j,k} \leq +r$.

With the asymmetrical difference approximation (97), the error equations (100) for the row j = 2 can be written in matrix form with the usual notation

$$E_2 = A^{-1} R_2 \qquad (103)$$

where $A = P(2 + s^2/a)$. Now the latent roots of A^{-1}, given by $\frac{1}{4\cos^2 \frac{\alpha}{N+1}\frac{\pi}{2} + \frac{s^2}{a}}$ $(\alpha = 1, 2, \ldots N)$, are small when s^2/a is large.

Thus the value $a = \frac{1}{2}$ ensures stability for all s and gives rise to minimum round-off errors in the row $j = 2$.

Making this simplification $a = \frac{1}{2}$ in (100) and putting $R_2 = R$, $R_j = 0$, $(j = 3, 4, \ldots)$ the growth of error as far as $j = 12$ is given by table IX where $A = P\left[2(1 + s^2)\right]$ and $B = 4s^2I$, I being the unit matrix of order N. The main feature of the table is the occurence of the binomial coefficients in the diagonal columns.

	E_2	$(A^{-1}B)E_2$	$(A^{-1}B)^2E_2$	$(A^{-1}B)^3E_2$	$(A^{-1}B)^4E_2$	$(A^{-1}B)^5E_2$	$(A^{-1}B)^6E_2$	$(A^{-1}B)^7E_2$	$(A^{-1}B)^8E_2$	$(A^{-1}B)^9E_2$	$(A^{-1}B)^{10}E_2$
E_2	1										
E_3		-1									
E_4	-1		1								
E_5		2		-1							
E_6	1		-3		1						
E_7		-3		4		-1					
E_8	-1		6		-5		1				
E_9		4		-10		6		-1			
E_{10}	1		-10		15		-7		1		
E_{11}		-5		20		-21		8		-1	
E_{12}	-1		15		-35		28		-9		1

TABLE IX

Now the latent roots of $A^{-1}B$ are $\dfrac{2}{1 + \frac{2}{s^2}\cos^2 \frac{\alpha}{N+1}\frac{\pi}{2}}$ $(\alpha = 1, 2, \ldots N)$,

all of which tend to zero as the mesh ratio s tends to zero. Thus for s = 0, it is seen from tabel IX that $E_2 = -E_4 = -E_6 = -E_8 = \ldots A^{-1}R$ and $E_3 = E_5 = E_7 = \ldots\ldots = 0$. This value of the mesh ratio together with the residual distribution $R_j = (-1)^{j/2+1} R$ $(j = 2, 4, \ldots\ldots)$ gives rise to the error vectors

$$e_j = e_{j+1} = \frac{j}{2} (A^{-1}R), \qquad (j = 2, 4, \ldots\ldots) \tag{104}$$

where $A^{-1} = [P(2)]^{-1}$. The value s = 0 is of course quite artificial but the error vector obtained provides a useful guide to the maximum error possible for small s. This is illustrated in fig 5 where the growth of error ratio and the maximum error possible at the middle node of each row as far as j = 20 are shown for s = 1, 0·1, and 0 when N = 3. The growth of error ratio at any row is considered to be the ratio of the error at the middle node in that row to the magnitude of the error at the middle node in the row j = 2.

Another useful value of the parameter is a = 1. The modified equation (100) together with the residual distribution $R_2 = R$, $R_j = 0$ $(j = 2,4,\ldots\ldots)$ leads to an error growth which is illustrated in table X as far as j = 12, where $A = P(2 + s^2)$.

	E_2	$(s^2A^{-1})E_2$	$(s^2A^{-1})^2E_2$	$(s^2A^{-1})^3E_2$	$(s^2A^{-1})^4E_2$	$(s^2A^{-1})^5E_2$	$(s^2A^{-1})^6E_2$	$(s^2A^{-1})^7E_2$	$(s^2A^{-1})^8E_2$	$(s^2A^{-1})^9E_2$	$(s^2A^{-1})^{10}E_2$
E_2	1										
E_3		-2									
E_4		1	4								
E_5			-4	-8							
E_6			1	12	16						
E_7				-6	-32	-32					
E_8				1	24	80	64				
E_9					-8	-80	-192	-128			
E_{10}					1	40	240	448	256		
E_{11}						-10	-224	-512	-1024	-512	
E_{12}						1	124	400	1792	2304	1024

TABLE X

Now the latent roots of s^2A^{-1} are $\dfrac{1}{1 + \frac{4}{s^2}\cos^2\frac{\alpha}{N+1}\frac{\pi}{2}}$ $(\alpha = 1,2,\ldots N)$

all of which tend to zero as s tends to zero. Thus for s = 0 it is seen from table X that $E_3 = E_4 = E_5 = \ldots = 0$, and so irrespective of the values of the residuals at nodes for which $j > 2$, the error vectors are given by

$$e_j = A^{-1}R. \qquad (j = 2,3,4 \ldots\ldots) \qquad (105)$$

Once again the value s = 0 is completely artificial, but the

error vector given by (105) provides an indication of the size of the maximum error possible for small s. The growth of error ratio and the maximum error possible are shown in fig. 6 for s = 1, 0.1, and 0 when N = 5.

Comparing the errors given in figs. 5 and 6, it is seen that although a = ½ gives rise to minimum errors in the row j = 2, a = 1 gives a much smaller error growth. In fact when $s \leqslant 1$, no matter how many rows of calculation are considered, there is no chance of large round-off errors using the asymmetrical replacement (97) with a = 1, provided N is not excessively large.

With the symmetrical difference replacement (98), the growth of error is again given by table IX with $A = P(2 + s^2/a)$ and $B = \left(\frac{1-2a}{a}\right) P\left[2\left\{1 - \frac{s^2}{1-2a}\right\}\right]$. It should be pointed out at this stage that (97) and (98) reduce to the same expression when a = ½, and so this value of a will be excluded from the symmetrical case. The latent roots of $A^{-1}B$ are

$$\frac{\left(\frac{1-2a}{a}\right)\left[2\cos^2\frac{\alpha\pi}{2(N+1)} - \frac{s^2}{1-2a}\right]}{\left[2\cos^2\frac{\alpha\pi}{2(N+1)} + \frac{s^2}{2a}\right]} \quad (\alpha = 1, 2, \ldots\ldots N),$$

all of which tend to $\frac{1-2a}{a}$ as s tends to zero, and to -2 as s tends to infinity. In addition, it can easily be shown that the magnitude of a latent root exceeds unity for $s^2 \geqslant 0$ $(a \geqslant 1)$, $s^2 > 4(1-a)\cos^2\frac{\alpha\pi}{2(N+1)}$ $(\frac{1}{2} < a < 1)$, and $4(1-a)\cos^2\frac{\alpha\pi}{2(N+1)} < s^2 < \frac{4}{3}(1-3a)\cos^2\frac{\alpha\pi}{2(N+1)}$ $(a < \frac{1}{2})$.

The errors in the row $j = 2$ are given by (105) and so $a = \frac{1}{4}$ ensures stability for all s and gives minimum round-off errors in the row $j = 2$. With $a = \frac{1}{4}$ in (101), the growth of error is given as far as $j = 12$ by table IX with $A = P\left[2(1 + 2s^2)\right]$ and $B = 2P\left[2(1 - 2s^2)\right]$. Here the latent roots of $A^{-1}B$ tend to 2 as s tends to zero, and their magnitudes are all less than unity when s lies within the range $\frac{1}{6}(1 + \cos\frac{\pi}{N+1}) < s^2 < \frac{3}{2}(1+\cos\frac{\pi}{N+1})$. It seems likely that the minimum error growth will occur for a value of s within the above range. As an illustration, consider the simple case $N = 3$ where the range reduces to $\cdot 55 < s < \cdot 65$. The growth of error ratio and the maximum error possible at the middle node of each row as far as $j = 20$ are shown for $s = 1$, 0·6, and 0 in fig. 7. The artificial case $s = 0$ being unstable provides a limiting value for the error growth which is far in excess of the growth when s is small but not zero. When $s = 0$ it is seen from table IX that $E_j = (-1)^j(j-1)\,E_2$ $(j \geq 2)$ and so using the residual distribution $R_j = (-1)^j R$ $(j \geq 2)$, the maximum error vectors become

$$e_j = \frac{j(j-1)}{2}\,(A^{-1}R). \qquad (j = 2,3,\ldots..) \qquad (106)$$

Consider lastly the value of the parameter $a = 1$. The latent roots of $A^{-1}B$ tend to -1 as s tends to zero and to -2 as s tends to infinity, and so the minimum error growth will occur for small s. This time the case $s = 0$ is stable and so gives a useful guide to the errors for small s. With $s = 0$, $A = -B = P(2)$, and so from table IX

$$E_j \equiv \begin{cases} E_{3p-1} = (-1)^{p+1} E_2 \\ E_{3p} = (-1)^{p+1} E_2 \\ E_{3p+1} = 0 \end{cases} \qquad (p = 1, 2, \ldots\ldots)$$

Using the residual distribution

$$R_j \equiv \begin{cases} R_{3p-1} = (-1)^{p+1} R \\ R_{3p} = (-1)^{p+1} R \\ R_{3p+1} = 0 \end{cases} \qquad (P = 1, 2, \ldots\ldots)$$

the maximum error vectors become

$$e_j \equiv \begin{cases} e_{3p-1} = p(A^{-1}R) \\ e_{3p} = 2p(A^{-1}R) \\ e_{3p+1} = p(A^{-1}R) \end{cases} \qquad (p = 1, 2, \ldots\ldots) \qquad (107)$$

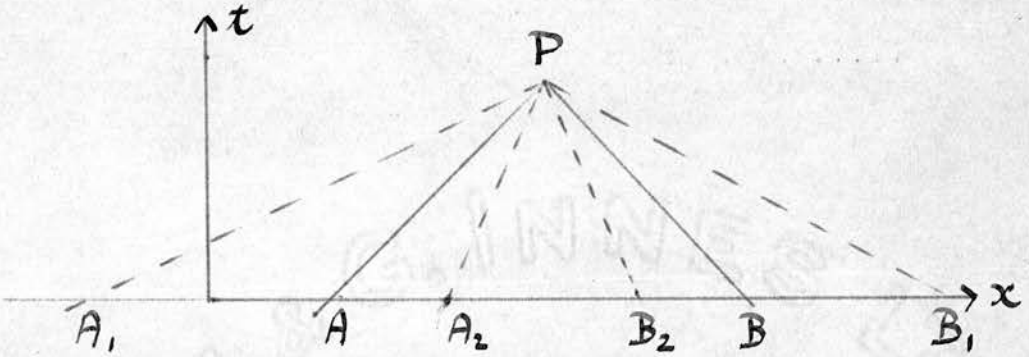
It should be realized that the round-off errors described in this section are in excess of the errors likely to be incurred in actual calculations. Nevertheless, the present study will give useful information concerning the values of the variable parameter and the mesh ratio, which give rise to minimum round-off errors for each type of implicit finite difference replacement considered.

Finally, for a given number of nodes along t = 0, the smaller the value of the mesh ratio s, the fewer will be the

number of rows necessary to solve the problem for a given period of time, and so from round-off error considerations alone there may be no lower limit to the optimum value of s. For small s, however, the solution of the difference equation may not be a good approximation to the solution of the differential equation, and so considerations other than minimising the round-off errors will impose a lower limit on the mesh ratio.

3. The Problem of Convergence.

The characteristics of the wave equation (96) are the two families of lines $x \mp t = \text{constant}$. In a problem of unrestricted extent in the x - direction, any point P in the



upper half plane has two characteristics passing through it. The characteristics have slopes $\pm \pi/4$, and they cut the x-axis in the points A and B. It is well known from the theory of characteristics in the solution of hyperbolic equations that only those values of ϕ on the x-axis in the interval AB influence the solution in the triangular region APB. Now using the explicit difference replacement (99) of the wave equation, the values of ϕ on the x-axis in the interval A_1B_1 (including the

interval AB) influence the solution in the region A_1PB_1 if $s \geq 1$, and the values of ϕ on the x-axis in the interval A_2B_2 (included in the interval AB) influence the solution in the region A_2PB_2 if $s < 1$. Courant, Friedrichs and Lewy (28) pointed out that if the region of determination A_2PB_2 of the difference equation lies completely inside the region of determination APB of the differential equation, then the solution of the difference equation at P does not depend on the boundary values of ϕ on the x-axis in the intervals AA_2 and BB_2. As the net spacings tend to zero with s remaining constant, the above situation continues to hold, and so the solution of the difference equation with $s < 1$ does not in general converge to the exact solution of the differential equation. If, however, the region APB lies inside or coincides with the region of determination A_1PB_1 of the difference equation, then the solution of the difference equation at P depends on the boundary values of ϕ on the x-axis in the intervals A_1A, AB, BB_1. As the net spacings tend to zero with s remaining constant, the influence on the value of ϕ at P of the boundary values of ϕ in the intervals A_1A and BB_1 diminishes to zero, and so the solution of the difference equation with $s \geq 1$ converges to the exact solution of the differential equation.

From the above considerations, it seems clear that explicit difference replacements of the wave equation are in

general divergent if $s < 1$. As a result there is no point in attempting to find explicit approximations stable for all values of s. (cf. the "diamond" formula for the heat conduction equation). In fact there seems no doubt that the best explicit replacement of the wave equation is the four point formula (102), ((99) with $s = 1$), which possesses the additional merit of having no truncation error. It should be pointed out that Leutert and O'Brien (33) obtained a solution of (99), convergent for all values of the mesh ratio s, which satisfies the boundary conditions $\phi = 0$ at $x = 0, 1$ $(t > 0)$, $\phi = f(x)$, $\phi_t = 0$ at $t = 0$ $(0 < x < 1)$. However when $s < 1$, convergence cannot be realized in numerical calculation using (99), since round-off errors will grow due to the instability of the difference approximation.

The region of dertermination of an implicit approximation always includes the region of determination of a hyperbolic differential equation and so implicit replacements of the wave equation are probably convergent for all values of s. Since many implicit formulae are also stable for all s, the lower limit on the mesh ratio is usually imposed by the allowable magnitude of the truncation error.

4. Truncation Errors.

If Φ is the exact solution of the differential equation (96), and ϕ is the exact solution of the difference equation (99), then

the error due to truncation is e ($= \Phi - \phi$) which satisfies the difference equation

$$e_{j+1,k} = 2(1-s^2)e_{j,k} + s^2(e_{j,k-1} + e_{j,k+1}) - e_{j-1,k} + T_{j,k}$$

where $T_{j,k}$ is the truncation error in the explicit five point formula (99). The dominant terms in $T_{j,k}$ are given by

$$T_{j,k} = \frac{1}{12}(\Delta x)^2\left[(\Delta t)^2 \Phi_{tttt} - (\Delta x)^2 \Phi_{xxxx}\right]$$

$$= \frac{(1-s^2)}{12s^4}(\Delta x)^4 \Phi_{xxxx} .$$

When $s = 1$, the dominant terms in $T_{j,k}$ vanish together with all the other terms and so the difference approximation (102) has no truncation error.

The truncation errors are now considered for two common implicit finite difference replacements of the wave equation, which are stable for all values of the mesh ratio. First the five point backward replacement (97) with $a = 1$) has an error due to truncation which satisfies the difference equation

$$e_{j,k-1} + e_{j,k+1} - (2+s^2)e_{j,k} = s^2(e_{j-2,k} - 2e_{j-1,k}) + T_{j,k}$$

where the dominant terms in $T_{j,k}$ are given by

$$T_{j,k} = \frac{1}{3}(\Delta x)^3 \Phi_{xxx} + \frac{s^2-7}{12s^2}(\Delta x)^4 \Phi_{xxxx} . \qquad (108)$$

Second the symmetrical seven point replacement ((98) with a = ½) has an error due to truncation which satisfies the difference equation

$$e_{j+1,k-1} - 2(1+s^2)e_{j+1,k} + e_{j+1,k+1} = -4s^2 e_{j,k} - \left[e_{j-1,k-1} - 2(1+s^2)e_{j-1,k} + e_{j-1,k+1} \right] + T_{j,k},$$

where the dominant terms in $T_{j,k}$ are given by

$$T_{j,k} = \frac{5+s^2}{6s^2} (\Delta x)^4 \Phi_{xxxx} . \qquad (109)$$

Without making any numerical assessment of the truncation errors (108) and (109), it can be said that the symmetrical seven point formula is generally more accurate than the backward five point formula. Also when s is sufficiently small for a given Δx, the errors due to truncation will be unacceptably large with both formulae, although the seven point approximation is still superior at any value of s. As in the study of the heat conduction equation, there is a great need for an exact solution of the error e due to truncation for say the symmetrical seven point formulae. Only from such an exact solution, can the magnitude of the error due to truncation be assessed for different values of s and Δx, and a lower limit attached to the mesh ratio consistent with a preassigned upper bound of error due to truncation. If the lower limit has a value 0·1, the time interval using the symmetrical seven point formula will be ten

times the time interval using the explicit formula (102) where the mesh ratio has a value 1·0. As with the heat conduction equation, this increased time interval may well offset the additional labour required in obtaining a solution by relaxation.

Summary of Difference Replacements of $\phi_{tt} = \phi_{xx}$.

(A) $$\phi_{j,k} = 2(1-\frac{1}{s^2})\phi_{j-1,k} + \frac{1}{s^2}(\phi_{j-1,k-1} + \phi_{j-1,k+1}) - \phi_{j-2,k}$$

(B) $$\phi_{j,k} = \phi_{j-1,k-1} + \phi_{j-1,k+1} - \phi_{j-2,k}$$

(C) $$\phi_{j,k-1} - (2+s^2)\phi_{j,k} + \phi_{j,k+1} = s^2(\phi_{j-2,k} - 2\phi_{j-1,k})$$

(D) $$\phi_{j,k-1} - 2(1+s^2)\phi_{j,k} + \phi_{j,k+1} = -4s^2\phi_{j-1,k} - (\phi_{j-2,k-1} - 2(1+s^2)\phi_{j-2,k} + \phi_{j-2,k+1}).$$

Replacement	Type	Stability	Principal Truncation Term	Round-off Errors
A	Explicit	$s \geq 1$	$\frac{1-s^2}{12s^4}(\Delta x)^4 \Phi_{xxxx}$	Small
B	Explicit	Stable	No Truncation Error	Small
C	Implicit	$s > 0$	$\frac{1}{s}(\Delta x)^3 \Phi_{xxx}$	pp. 126-133.
D	Implicit	$s > 0$	$\frac{5+s^2}{6s^2}(\Delta x)^4 \Phi_{xxxx}$	pp. 126-133.

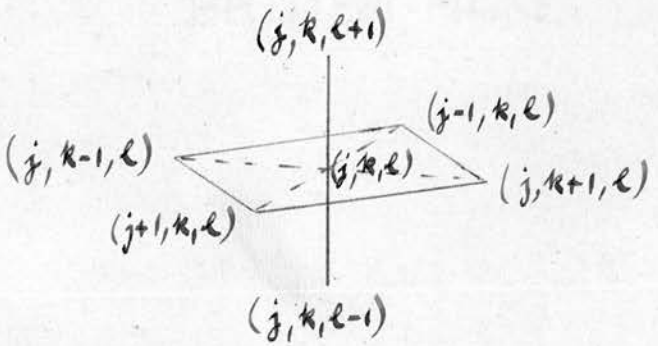
5. The Wave Equation in Two Dimensions.

In two dimensions, the hyperbolic wave equation is

$$\phi_{xx} + \phi_{yy} = \phi_{tt} . \tag{110}$$

The region to be examined is covered by an orthogonal mesh, where the mesh lengths are h in the x- and y- directions and Δt in the t- direction. If j and k are the row and column numbers in the (x,y) plane, and l is the number of the plane parallel to the (x,y) plane, a simple explicit difference replacement of (110) is

$$\left(\phi_{j,k-1,\ell} + \phi_{j,k+1,\ell} + \phi_{j-1,k,\ell} + \phi_{j+1,k,\ell} - 4\phi_{j,k,\ell}\right) = \frac{h^2}{(\Delta t)^2}\left(\phi_{j,k,\ell-1} - 2\phi_{j,k,\ell} + \phi_{j,k,\ell+1}\right). \tag{111}$$



Making the substitution

$$\phi_{j,k,l} = e^{\alpha t}\, e^{i\beta x}\, e^{i\gamma y}$$

in (110), and simplifying, the result

$$e^{\alpha\Delta t} + e^{-\alpha\Delta t} - 2 = -\frac{4}{s^2}\left(\sin^2\frac{\beta\Delta x}{2} + \sin^2\frac{\gamma\Delta y}{2}\right)$$

if obtained where $s = \frac{h}{\Delta t}$ and $\Delta x = \Delta y = h$. This leads to

$$e^{\alpha \Delta t} = A \pm (A^2 - 1)^{\frac{1}{2}}$$

where $A = 1 - \frac{2}{s^2}(\sin^2 \frac{\beta h}{2} + \sin^2 \frac{\gamma h}{2})$, and so the condition for stability yields

$$-1 \leqq 1 - \frac{2}{s^2} (\sin^2 \frac{\beta h}{2} + \sin^2 \frac{\gamma h}{2}) \leqq 1,$$

which simplifies to $s^2 \geqslant 2$ for all β and γ. Now the region of determination of the differential equation (110) for a point P on the plane $t = T$ is a cone with vertex at P, whose section by the plane $t = 0$ is a circle of radius $2^{\frac{1}{2}} T$. When $s^2 = 2$, the region of determination of the difference equation (111) is a pyramid with vertex at P, whose section by the plane $t = 0$ is a square of side $2^{3/2}T$. The pyramid just encloses the circular cone and so $s^2 \geqslant 2$ is also the condition for convergence. With $s^2 = 2$, (111) becomes the six point formula

$$\phi_{j,k-1,\ell} + \phi_{j,k+1,\ell} + \phi_{j-1,k,\ell} + \phi_{j+1,k,\ell} = 2(\phi_{j,k,\ell+1} + \phi_{j,k,\ell-1}). \quad (112)$$

If Φ is the exact solution of the wave equation (110) the dominant terms in the truncation error $T_{j,k,l}$ are given by

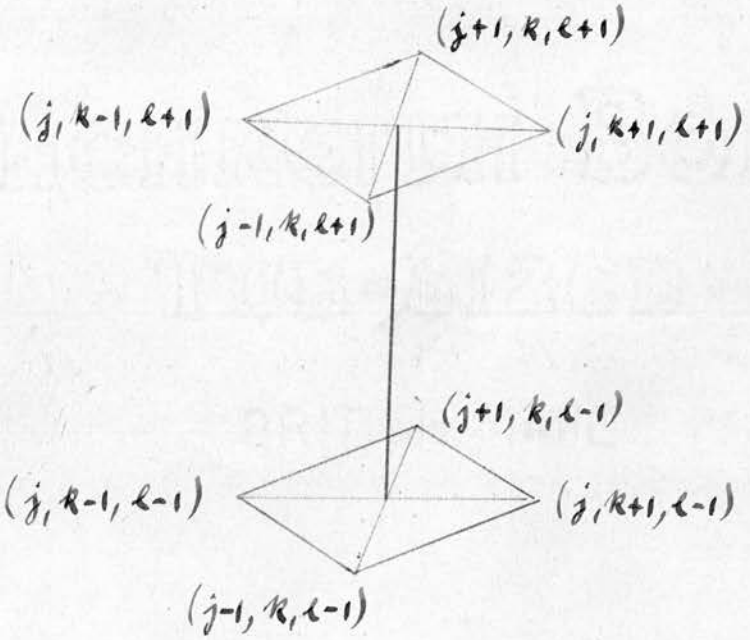
$$T_{j,k,l} = \frac{1}{24} h^4 \left(\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}\right)^2 \Phi .$$

This truncation error can be made negligible by choosing h sufficiently small.

The implicit difference replacement of (110), corresponding

to the symmetrical seven point formula in one dimension, is

$$(\phi_{j,k-1,\ell+1} + \phi_{j,k+1,\ell+1} + \phi_{j-1,k,\ell+1} + \phi_{j+1,k,\ell+1} - 4\phi_{j,k,\ell+1}$$
$$+ \phi_{j,k-1,\ell-1} + \phi_{j,k+1,\ell-1} + \phi_{j-1,k,\ell-1} + \phi_{j+1,k,\ell-1} - 4\phi_{j,k,\ell-1})$$
$$= \frac{2k^2}{(\Delta t)^2}(\phi_{j,k,\ell-1} + \phi_{j,k,\ell+1} - 2\phi_{j,k,\ell}). \qquad (113)$$



Substituting $e^{\alpha t} e^{i\beta x} e^{i\delta y}$ in (113) and simplifying, the result

$$e^{\alpha\Delta t} = B \pm (B^2 - 1)^{\frac{1}{2}},$$

where $B = \dfrac{1}{1 + \frac{2}{s^2}(\sin^2 \frac{\beta h}{2} + \sin^2 \frac{\delta h}{2}}$ is obtained and so (113) is stable for all s. When (113) is used, a relaxation solution is necessary at each time step and so errors due to rounding-off will arise during the calculation. The estimation of these

errors is a long and tedious business, and taking the results for the seven point symmetrical formula in one dimension as a guide, the round-off errors using (113) are unlikely to be large provided the number of mesh points in the x,y plane is small. The dominant terms in the truncation error $T_{j,k,l}$ are given by

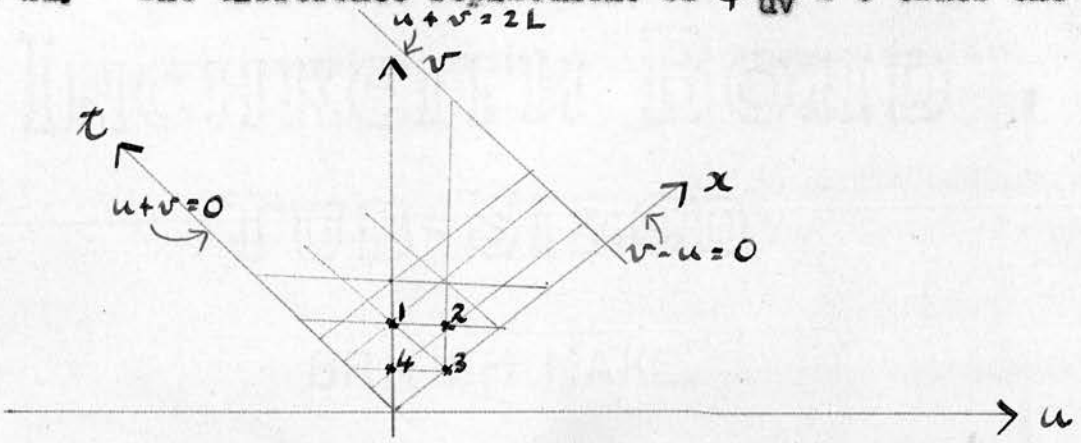
$$T_{j,k,l} = \frac{1}{6}h^4 \left[\left(\frac{5+s^2}{s^2}\right) \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^2 \Phi - 2 \frac{\partial^4 \Phi}{\partial x^2 \partial y^2} \right],$$

and so far a given value of h, the errors due to truncation will be unacceptably large for a sufficiently small value of s. Without an exact solution for the error due to truncation showing the variation of the error with s and h, no lower limit can be attached to the mesh ratio s consistent with a preassigned upper bound of truncation error. If this lower limit is considerably below $2\frac{1}{2}$, the value of s for the explicit formula (112), the increased time interval may justify the use of an implicit approximation.

6. General Remarks.

It is often adviseable when solving a hyperbolic equation by difference methods to ensure that the mesh lines coincide with the characteristics of the differential equation. This is achieved by changing the independent variables to u and v where u = constant and v = constant are the two families of characteristics of the differential equation. If a rectangular

net is now used in a plane where u and v are cartesian co-ordinates, the regions of determination for net points relevant to the difference equation can be made to coincide with the regions of determination for the differential equation. In the case of the wave equation, $u = x - t$ and $v = x + t$, and the wave equation reduces to $\phi_{uv} = 0$. The initial line $t = 0$ maps into the line $v - u = 0$, and the strip $(t > 0,\ 0 \leq x \leq L)$ maps into the diagonal strip bounded by the lines $v - u = 0$, $u + v = 0$, $u + v = 2L$. The difference replacement of $\phi_{uv} = 0$ takes the



form $\phi_1 = \phi_2 + \phi_4 - \phi_3$ (cf. (102)) for equal spacings in the u- and v- directions. It is obvious that nothing has been gained in this problem by changing the independent variables to u and v. The figure for the problem in x and t has merely been rotated anticlockwise through $\pi/4$. However in hyperbolic problems where the characteristics are curves rather than straight lines, the transformation to the characteristic variables is advantageous. In particular, it simplifies considerably the problem of convergence.

The general hyperbolic linear equation in two independent variables is given by (1) with $b^2 > 4ac$. The simplest finite

difference replacement of this equation using central differences is an implicit nine point formula. This reduces to an explicit five point formula if the term in ϕ_{xy} is absent. The numerical solutions of these linear difference equations will provide reasonable approximations to the solution of the differential equation only if the difference approximations are stable for the values of Δx and Δy chosen. No general criterion is available for determining the stability or otherwise of linear difference equations with variable coefficients. If the coefficients are reasonably constant over parts of the x,y plane, however, von Neumann's criterion for stability (section 2) can be applied within each part where the coefficients are approximately constant.

Nothing useful can be said at this stage about the numerical solution using difference methods of non-linear hyperbolic equations, although instances of difference methods applied to such problems will be given in Chapter 3. However, for a hyperbolic system of n quasi - linear first order partial differential equations in two independent variables x and y, Courant, Isaacson, and Rees (34) have shown that the mesh ratio $\Delta y/\Delta x$ must be chosen in such a way that the region of determination of any point in the net, as given by the difference equations, is not less than the region of determination as given by the corresponding differential equations.

CHAPTER III

APPLICATIONS IN FLUID DYNAMICS

So far no extensive use has been made of finite difference methods in the solution of physical problems. The three main reasons for this are

(1) A separate computation must be made for each set of values of the variable parameters in a problem, and so it is difficult to formulate general laws from finite difference solutions.

(2) In problems involving non-linear differential equations, it is difficult to establish the stability and convergence of finite difference replacements, and so in many cases there is no guarantee that the numerical solution obtained is a reasonable approximation to the exact solution of the problem.

(3) The time required to carry out the finite difference calculation is often prohibitively large, especially when the problem is three dimensional. However, the study of numerical solutions of finite difference equations is still in its infancy, and once more knowledge of the stability and convergence of difference solutions is obtained and more powerful computing machines become available, finite difference methods may well become invaluable in the solution of hitherto unsolved physical problems.

In the present chapter, an account will be given of the success so far achieved in the application of finite difference methods to the solution of problems in Fluid Dynamics. In this account, the shortcomings of the methods used will be pointed out, and suggestions made for further research in this field. To

facilitate the lay out of the chapter, the flow problems described will come under three main headings;

I compressible

II rotational but inviscid

III viscous,

where overlap is unavoidable in many cases. No mention will be made of the use of finite difference methods to the solution of problems in classical hydrodynamics. Finally, all problems considered are steady, and two-dimensional or axially symmetric.

A list of symbols used in this chapter will now be given. Any deviation from this list will be explained at the appropriate point, but where possible the notation will be kept standard.

LIST OF SYMBOLS

x, y	cartesian co-ordinates in the physical plane.
x, r	co-ordinates in axially symmetric flow.
ξ, η	elliptic co-ordinates.
ϕ	velocity potential.
ψ	stream function.
q	velocity.
u, v	velocity components parallel to the x- and y- axes.
p	pressure.
ρ	density.
c	speed of sound.
T	temperature.

S	entropy.
c_p	specific heat at constant pressure.
c_v	specific heat at constant volume.
γ	c_p/c_v.
R	perfect gas constant.
k	adiabatic gas constant, specific conductivity.
μ	coefficient of viscosity.
ν	coefficient of kinematic viscosity (μ/ρ).
U	standard velocity.
h	standard length.
ω	rotation, temperature index in viscosity law.
χ	$\rho^{-\frac{1}{2}}$
H	total head.
L	log $(1/q)$.
i	enthalpy (c_pT).
M	Mach number (q/c).
R	Reynolds number $(\rho Uh/\mu)$.
σ	Prandtl number $(\mu c_p/k)$.
N	non dimensional parameter $(U/\omega h)$.
G	mass flow per second in free stream.
ν	flow parameter $(G/\rho_s c_s h)$. (suffix s denotes stagnation conditions).
δ	mesh ratio.

I COMPRESSIBLE FLOW

Fundamental Equations

The equations of two dimensional steady motion for the irrotational flow of a non-viscous compressible fluid may be written in the form

$$\nabla^2(\chi\psi) - \psi\nabla^2\chi = 0, \tag{1}$$

$$\frac{2}{\gamma-1}\left\{1-\left(\frac{\chi}{\chi_s}\right)^{2(1-\gamma)}\right\} = \frac{\chi^4}{c_s^2}\left\{\left(\frac{\partial\psi}{\partial x}\right)^2+\left(\frac{\partial\psi}{\partial y}\right)^2\right\}, \tag{2}$$

where c_s, χ_s are the values of c, χ at a stagnation point and $\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$. It is advantageous to write (1) and (2) in the non-dimensional forms

$$\nabla^2(\chi\psi) - \psi\nabla^2\chi = 0, \tag{3}$$

$$\frac{2}{\gamma-1}\left\{1-\chi^{2(1-\gamma)}\right\} = \nu^2\chi^4\left\{\left(\frac{\partial\psi}{\partial x}\right)^2+\left(\frac{\partial\psi}{\partial y}\right)^2\right\}. \tag{4}$$

To do so, some significant linear dimension h pertaining to the specific problem under consideration is selected, and a non dimensional parameter ν introduced where

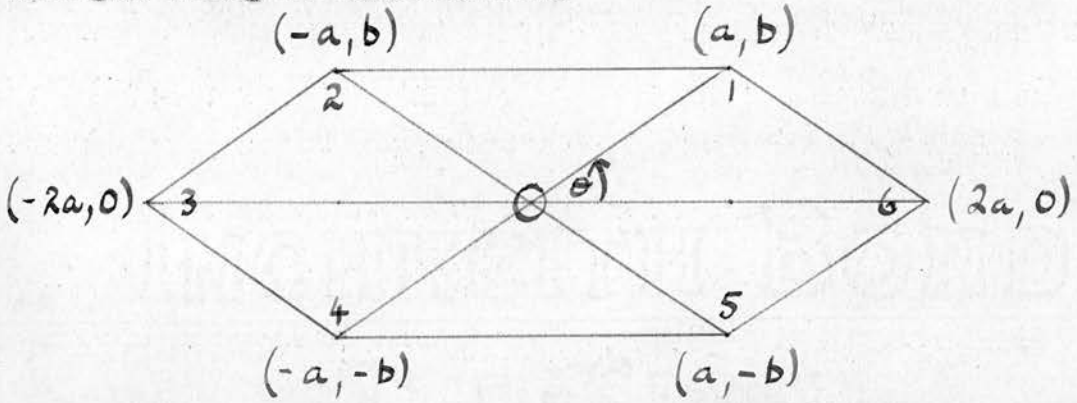
$$\nu = \frac{G}{\rho_s c_s h},$$

G being the mass flow per second under free stream conditions. Changing the notation slightly by writing χ, ψ, x, y for the

non dimensional quantities χ/χ_s, ψ/G, x/h, y/h respectively, (1) and (2) take the non dimensional forms (3) and (4).

Finite Difference Approximations.

To keep the network as flexible as possible the field to be examined is covered by a regular hexagonal network, a typical hexagon being as illustrated.



The square and triangular nets described by Southwell (1) are particular cases ($\theta = 45^\circ, 60^\circ$) of this more general net. The effectiveness of this network, however, may be impaired if θ has a value outside the range $45^\circ \leqq \theta \leqq 75^\circ$, since in such cases the six nodes of the hexagon are not the six nodes of the net which are closest to the centre of the hexagon.

The finite difference approximations to (3) and (4) applicable to the hexagonal network are (Mitchell and Rutherford (35))

$$F_o = \frac{1}{b^2} \sum_{1,2,4,5} \chi_1(\psi_1 - \psi_o) + \tfrac{1}{2}\left(\frac{1}{a^2} - \frac{1}{b^2}\right) \sum_{3,6} \chi_1(\psi_1 - \psi_o) = 0 \qquad (5)$$

and

$$\frac{2}{\gamma-1}\left(1-\chi_0^{2(1-\gamma)}\right) = \frac{\nu^2\chi_0^4}{b^2}\left\{\psi_0^2 + \frac{1}{4}\sum_{1,2,4,5}\psi_i(\psi_i-2\psi_0)\right\}$$

$$+\frac{1}{4}\left(\frac{1}{a^2}-\frac{1}{b^2}\right)\nu^2\chi_0^4\left\{\psi_0^2+\frac{1}{2}\sum_{3,6}\psi_i(\psi_i-2\psi_0)\right\}, \quad (6)$$

where χ_i, ψ_i denote the values of χ, ψ at the node labelled $i (i = 0,1,2,\ldots\ldots 6)$. To obtain an approximate solution for a compressible flow problem, (5) and (6) must be satisfied approximately throughout the field of flow. Initially, a value of ψ is assumed at each node of the net, and the appropriate ψ values substituted in (6). Thus χ is determined at each node, and (5) now determines the residual F at each node. The procedure thereafter is to modify repeatedly the ψ- distribution, either by trial and error or according to a pattern, so that the residuals in the field are made as small as possible.

Southwell describes a pattern as a systematic means of eliminating residuals at the nodes. A finite difference pattern for the regular hexagonal network ($b = \frac{1}{2}\sqrt{3}a$) is given by the formulae

$$\delta F_0 = -\left[\sum_{i=1}^{6}\frac{(\gamma-1)\nu^2}{12a^2}\{f'(R)\}_i(\psi_0-\psi_i)^2 + \sum_{i=1}^{6}\chi_i\right]\delta\psi_0, \quad (7)$$

$$\delta F_j = \left[\chi_0+\frac{(\gamma-1)\nu^2}{4a^2}\{f'(R)\}_0(\psi_0-\psi_i)(2\psi_0-\tfrac{1}{3}\sum_{i=1}^{6}\psi_i)\right]\delta\psi_0, \quad (8)$$

$$(j = 1,2,\cdots 6)$$

where δF_0, δF_1, ... δF_6 are the changes consequent on a modification $\delta\psi_0$ of ψ_0, and

$$R = \chi^{-4}\left(1 - \chi^{2(1-\gamma)}\right) \qquad (9)$$

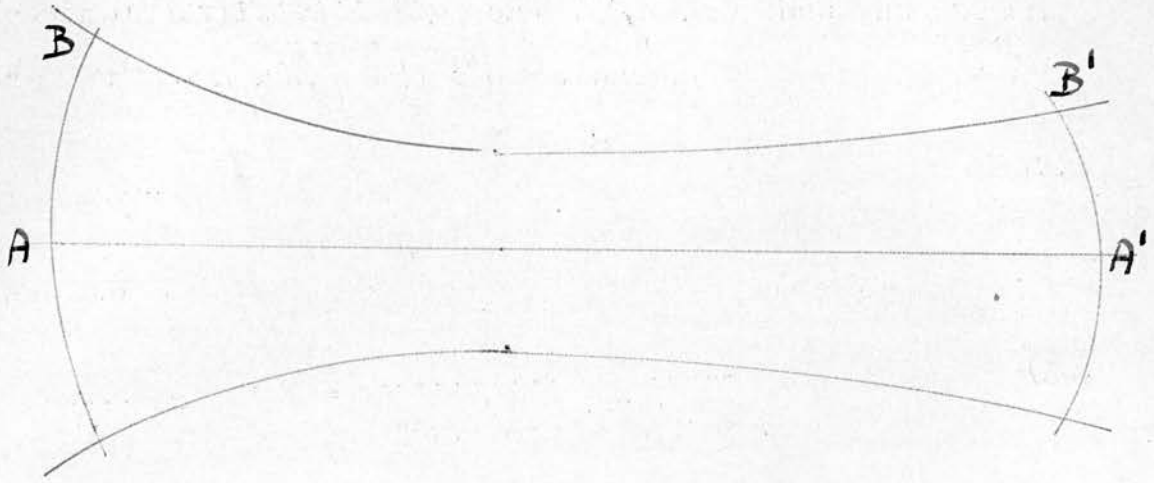
leading to

$$\frac{d\chi}{dR} = f'(R) = \frac{\chi^5}{2\left\{(1+\gamma)\chi^{2(1-\gamma)} - 2\right\}}, \qquad (10)$$

if (9) is expressed as $\chi = f(R)$. The pattern formulae (7) and (8) are used as follows to eliminate the residuals. From the original ψ distribution, χ is determined at each node using (6), and so $f'(R)$ is obtained using (10). From formulae (10) and (11), the modifications δF_0 ..... δF_6 arising out of an increment $\delta\psi_0$ in ψ_0 are then evaluated. A judicious choice of $\delta\psi_0$ will almost eliminate F_0 although this will in general be accompanied by a slight increase in the residuals at the surrounding nodes.

Application to Problems.

The flow of compressible fluid through a convergent-divergent nozzle as shown using finite difference methods was first considered by Green and Southwell (36). It was assumed by these authors that the fluid entered the nozzle radially



across AB and left it radially across $A'B'$. Thus the stream function ψ is known round the closed boundary $ABB'A'A$. Transforming the region $ABB'A'$ into a rectangle and using difference equations (5) and (6) with a = b (square network), a solution is obtained provided the mass flow through the nozzle is sufficiently small to ensure subsonic velocities everywhere in the region. However, once the mass flow through the nozzle is above the critical value (value when a sonic velocity first appears in the nozzle throat), the finite difference technique using (5) and (6) fails to yield definite results according to Green and Southwell. They ascribe this failure to the very large values taken by $f'(R)$ at and above the speed of sound, causing large modifications δF_j (j = 0,1,2,.....6) to arise from small increments $\delta\psi_0$. Thus under sonic and supersonic conditions, the finite difference technique using (5) and (6) diverges or at best converges very slowly. To meet this difficulty, Fox and Southwell (37) devised an alternative method

not involving finite difference equations which successfully dealt with compressible flow through a convergent divergent nozzle where the flow is everywhere supersonic beyond the throat.

In contrast to the results of Green and Southwell, the present author (38) showed that many problems in supersonic flow are amenable to solution by the finite difference technique. If relation (10) is tabulated for $\gamma = 1.4$ as under,

χ	1·00	1·15	1·20	1·25-	1·25+	1·50	2·00	2·70	3·50
M	0	0·75	0·88	1·0-	1·0+	1·4	1·9	2·4	3·0
$f'(R)$	1·25	6·70	15·5	+	-	-14	-26	-77	-236

it will be observed that for supersonic speeds ($M > 1$), $f'(R)$ has large negative values, large in comparison with the positive values which it has for subsonic speeds. The discontinuity in $f'(R)$ at $M = 1$ clearly indicates that the finite difference technique will break down at sonic velocity, but in several simple problems attempted using (5) and (6) no insuperable difficulty was found, except in the immediate neighbourhood of the sonic line, in applying the finite difference technique to shock free supersonic problems with $M < 3$. It may be impossible to use a central difference formula like (6) in dealing with supersonic regions where the boundary conditions are given on an open boundary. For the present, however, the point at issue is whether the magnitude of $f'(R)$ prevents formulae (5) and (6) being used when the flow is supersonic. In formula (10) there

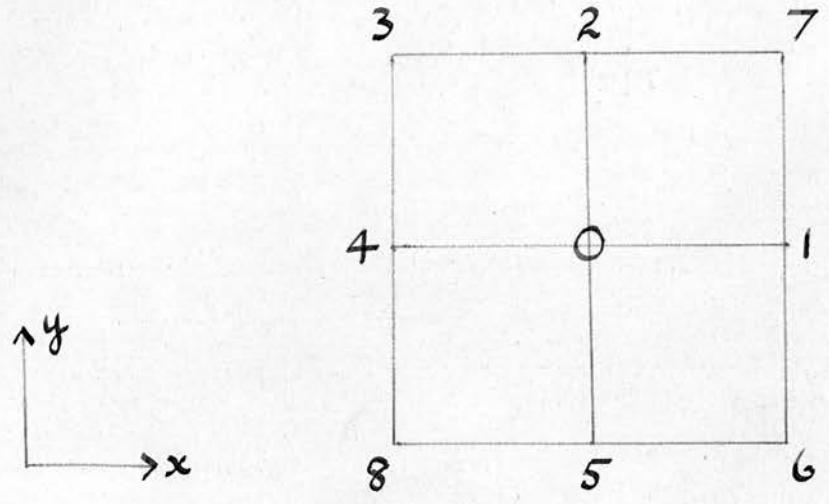
are two terms which contribute to the value of δF_0. So long as the term involving $f'(R)$ is not of a greater order of magnitude than the term $\sum_{i=1}^{6} \chi_i$, the order of magnitude of $f'(R)$ need not cause concern. In the supersonic problems examined with $M < 3$, nowhere did the term in $f'(R)$ outweigh the other term. To conclude, it would seem that although in certain supersonic problems the finite difference technique using (5) and (6) may become unworkable on account of very large values of $f'(R)$, yet in many cases this is not so.

It will be evident from what has been said above that the finite difference technique using (5) and (6) works for subsonic regions and for many supersonic regions, but breaks down in the neighbourhood of the sonic line. Thus mixed flows which are partly subsonic and partly supersonic are not directly amenable to solution by finite difference methods. If, however, the sonic line, whose position initially is unknown, is treated as a boundary between the subsonic and supersonic regions, the finite difference technique can be applied within each region. The position of the sonic line is determined by approaching it asymptotically from the subsonic side. Using this method, Mitchell and Rutherford (35), in the problem of subsonic compressible flow (free stream Mach Number of 0·2) past a symmetrical double wedge, found a small asymmetrical supersonic region round the wedge apex. However, too much reliance should not be placed on any procedure of determining the sonic line

when the latter encloses a localised supersonic region (a supersonic region embedded in a subsonic field). At the rear of such a region where the supersonic flow is slowing down, a shock wave is usually present whose strength depends on the size of the region. The presence of a shock wave together with an undetermined sonic line will be too much for the finite difference technique. The existence of shock free localised supersonic regions has been established mathematically for certain contours, but if any small part of the contour is straight, a continuous solution cannot be obtained. In the double wedge problem mentioned the supersonic region is so small that the shock wave is negligibly weak.

The position is much more hopeful when the supersonic region is not localised. This occurs in the problem of flow through a convergent divergent nozzle where the flow becomes sonic in the neighbourhood of the throat and is supersonic in the divergent part of the nozzle. Although the boundary conditions necessary for a unique solution are not certain, it is felt that the flow conditions at the entry of the nozzle will suffice. The computation of the subsonic region using the difference equations (5) and (6) will determine the position of the sonic line somewhere in the throat. From the sonic line the calculation will proceed stepwise through the supersonic region.

It should be realized that (6), being a central difference formula, is not directly applicable to step-by-step calculations. This is most easily illustrated in the case a = b where the general hexagonal network reduces to a square network with a mesh length



of $2^{1/2}$ a. For convenience a block of four squares is considered as shown which requires the introduction of two new nodes 7 and 8. It is supposed that ψ and χ are known at nodes on $x = X$, $X + 2^{1/2}$ a, where (X,Y) is node 8. It is required to find ψ and χ at nodes on $x = X + 8^{1/2}$ a, the calculation proceeding stepwise in the x-direction. A central difference replacement of (4) is not applicable in this problem and so (4) is replaced at node 1 by a backward difference formula. The latter is derived by writing

$$\left(\frac{\partial\psi}{\partial x}\right)^2 + \left(\frac{\partial\psi}{\partial y}\right)^2 = \frac{1}{2}\nabla^2(\psi^2) - \psi\nabla^2\psi$$

$$= \frac{1}{2}\left[\frac{(\psi^2)_6 + (\psi^2)_7 - 2(\psi^2)_1}{2a^2} + \frac{(\psi^2)_4 + (\psi^2)_1 - 2(\psi^2)_0}{2a^2}\right]$$

$$- \psi_1\left[\frac{\psi_6 + \psi_7 - 2\psi_1}{2a^2} + \frac{\psi_4 + \psi_1 - 2\psi_0}{2a^2}\right]$$

$$= \frac{1}{4a^2}\left[\psi_1^2 - 2\psi_0(\psi_0 - 2\psi_1) + \sum_{i=4,6,7} \psi_i(\psi_i - 2\psi_1)\right],$$

which leads to the required backward difference replacement

$$\frac{2}{\gamma-1}\left\{1-\chi_1^{2(1-\gamma)}\right\} = \frac{\nu^2\chi_1^4}{4a^2}\left[\psi_1^2 - 2\psi_0(\psi_0-2\psi_1) + \sum_{i=4,6,7}\psi_i(\psi_i-2\psi_1)\right]. \quad (11)$$

Using (11), χ is calculated at nodes on $x = X + 8^{1/2}\,a$ from the values of ψ at nodes on $x = X$, $X + 2^{1/2}\,a$ and an assumed distribution of ψ on $x = X + 8^{1/2}\,a$. With $a = b$, the central difference replacement (5) becomes

$$F_0 = \sum_{i=1,2,4,5}\chi_i(\psi_i-\psi_0) = 0. \quad (12)$$

The values of ψ at nodes on $x = X + 8^{1/2}\,a$ are altered until (5) is satisfied at all nodes on $x = X + 2^{1/2}\,a$. The alterations in ψ can be made by trial and error or according to a pattern. Once the values of ψ at nodes on $x = X + 8^{1/2}\,a$ are established, these values together with ψ at nodes on $x = X + 2^{1/2}\,a$ enable ψ on $x = X + 18^{1/2}\,a$ to be found by the procedure described. This step-by-step process progresses through the supersonic region in the x - direction. The pattern for this process is given by the formulae

$$\delta F_0 = \left[\frac{(\gamma-1)\nu^2}{4a^2}\{f'(R)\}_1\,\psi_1(\psi_1+2\psi_0-\sum_{i=4,6,7}\psi_i)+\chi_1\right]\delta\psi_1 \quad (13)$$

$$\delta F_2 = \frac{(\gamma-1)\nu^2}{4a^2}\{f'(R)\}_7\,\psi_7(\psi_1-\psi_7)\,\delta\psi_1 \quad (14)$$

$$\delta F_5 = \frac{(\gamma-1)\nu^2}{4a^2}\{f'(R)\}_6\,\psi_6(\psi_1-\psi_6)\,\delta\psi_1 \quad (15)$$

where δF_0, δF_2, δF_3 are the changes consequent on a modification $\delta\psi_1$ of ψ_1. Once again there is nothing to suggest that formulae (13), (14), and (15) cannot be employed in a supersonic region if the terms involving $f'(R)$ are of reasonable magnitude. In fact for shock free supersonic regions where the Mach number is everywhere less than 3 (a rather arbitrary upper limit) no trouble is expected using difference formulae (11) and (12) except in the neighbourhood of the sonic line.

The situation regarding the application of finite difference methods to compressible flow problems appears now to have resolved itself as follows;

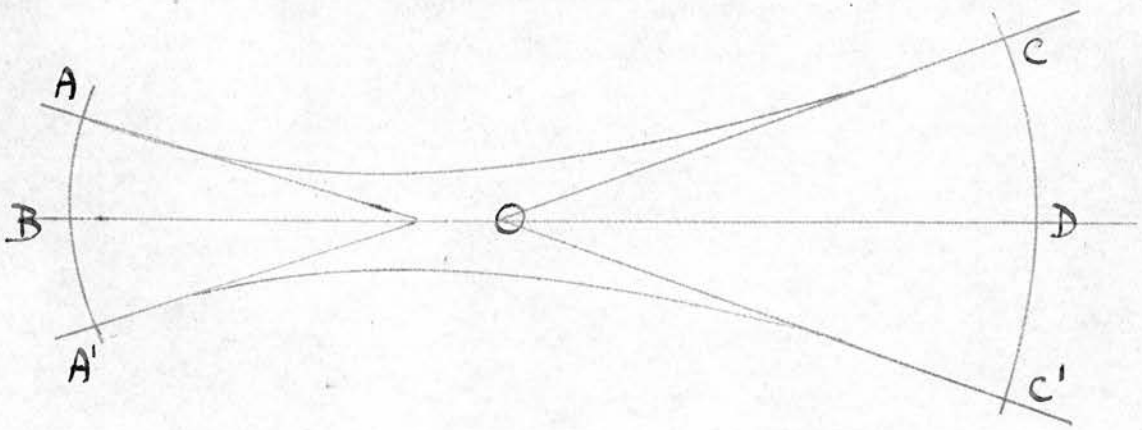
(1) if the problem is purely subsonic with conditions given on a closed boundary, a solution can be obtained using difference approximations (5) and (6).

(2) if the problem is purely supersonic with conditions given on an open boundary and no shock waves are present, a solution can be obtained using difference formulae (11) and (12).

(3) if the problem is mixed with a localised supersonic region embedded in a subsonic field, an attempt at solution can be made using (5) and (6) on either side of the initially unknown sonic line. Unless the supersonic region is very small, however, the shock wave present will upset the calculations.

(4) if the problem is mixed with a subsonic flow becoming

supersonic and remaining so, a solution can be obtained by using (5) and (6) in the subsonic region to determine the sonic line, and (11) and (12) in the supersonic part beyond it. So far, no mention has been made of boundary conditions in mixed flow problems. An attempt will be made to remedy this in the next section. It will be mentioned at this stage, however, that often conditions are known at the downstream end of a supersonic region, and so central formulae (5) and (6) can be used instead of step-by-step formulae (11) and (12) in the supersonic region. The use of the central formulae has the advantage that instability if present will not entirely vitiate the method, whereas instability makes a step-by-step process invalid. An example of a problem where conditions are known approximately at the downstream end of the supersonic region is the flow through a convergent-divergent nozzle where the flow changes from subsonic to supersonic in the throat and the nozzle walls are straight over a considerable part of the entrance and exit. The subsonic flow enters approximately



radially across the circular arc ABA' and the supersonic flow exits approximately radially across CDC', with the apparent source of the exiting fluid at O. From Bernoulli's law, the mass flow through the channel, and the adiabatic gas law, the flow is completely determined across CDC'. It should be appreciated, however, that this condition at the exit of the nozzle is not a boundary condition. The problem is completely specified by the mass flow through the channel together with the conditions at the entry, and the flow across CDC' is a consequence of these conditions. In computing a field using finite difference methods, it is advantageous to have as much knowledge of the solution as possible before starting the calculation. If part of the field can be evaluated analytically, the theoretical solution in this part will serve as a useful check on the finite difference solution which is intended to cover the complete field. In the problem of the convergent-divergent nozzle, the analytical solution of the flow across CDC' can either sit in judgment on the step-by-step calculation using (11) and (12) which is proceeding from the throat, or better, can be incorporated from the start in a difference solution using central formulae (5) and (6). As mentioned already, since nothing is known of the stability of the finite difference methods applied to compressible flow problems, it is much more satisfactory to use central formulae in a region with conditions given round a closed boundary, than

to employ step-by-step formulae with conditions given on an open boundary and the solution waiting at the end of the last step.

Boundary Conditions in Steady Flow.

In the previous section, no indication was given of the boundary conditions necessary for a unique solution of a mixed flow problem. The reasons for the author's reluctance to commit himself on this point will now be given.

The mathematical formulation in this chapter so far assumes the fluid to be inviscid and steady. Now steady flow does not exist in practise but is the limiting state of a flow changing under given boundary conditions from an initial state, with viscosity playing a leading part in the transition. It is therefore by no means obvious that the boundary conditions which govern the transient flow will be the proper boundary conditions for the idealised inviscid steady flow problem. In fact in many problems this is not so, and the search for steady flow boundary conditions has not been particularly successful. The main difficulty is that the boundary conditions required depend on the nature of the solution. They are different if the steady inviscid flow which they are to characterize is purely subsonic, purely supersonic, or mixed, and still different if the flow involves shock waves.

If the flow is everywhere subsonic, the differential

equation for the stream function is elliptic, and taking the potential equation as a guide, single data will be required round a closed boundary. If the flow is everywhere supersonic, the differential equation is hyperbolic, and the analogue is now the wave equation. Accordingly conditions will be required on an open boundary. For example, in supersonic shock free flow through a duct, the stream function will be required on the walls of the duct together with two initial conditions at the entrance and none at the exit. If the flow is subsonic with a small localised supersonic region, it is expected that the same boundary conditions may be imposed as for purely subsonic flow. (The existence of shock free localised supersonic regions is discussed by von Mises (39)). Finally in flow through a convergent divergent duct where the flow changes from subsonic to supersonic at the throat and then remains supersonic, it is expected that one boundary condition at the entrance and none at the exit will suffice.

The above conjectives concerning the boundary conditions necessary for unique solutions of steady continuous flow problems have little mathematical backing, and when shock waves are present the position regarding necessary boundary conditions is even less satisfactory. However, since it is not claimed that difference methods will in general determine the positions of shock waves, the circumstances under which a steady flow involving shocks will be uniquely determined by the boundary conditions will not be discussed.

The Hodograph Plane.

If q is the velocity and θ the angle between the velocity direction and the x - axis, the stream function ψ for a compressible fluid satisfies the differential equation

$$q^2 \frac{\partial^2 \psi}{\partial q^2} + q(1 + M^2)\frac{\partial \psi}{\partial q} + (1 - M^2)\frac{\partial^2 \psi}{\partial \theta^2} = 0, \qquad (16)$$

where $M = q/c$ is the Mach number, and the q, θ plane is the hodograph plane. Now Bernoulli's equation for a compressible fluid is

$$c^2 + \frac{\gamma-1}{2} q^2 = H, \qquad (17)$$

where H is constant, and so eliminating c between (16) and (17), the equation

$$q^2(1-q^2)\frac{\partial^2 \psi}{\partial q^2} + q(1- \frac{\gamma-3}{\gamma-1} q^2)\frac{\partial \psi}{\partial q} + (1 - \frac{\gamma+1}{\gamma-1}q^2) \frac{\partial^2 \psi}{\partial \theta^2} = 0 \qquad (18)$$

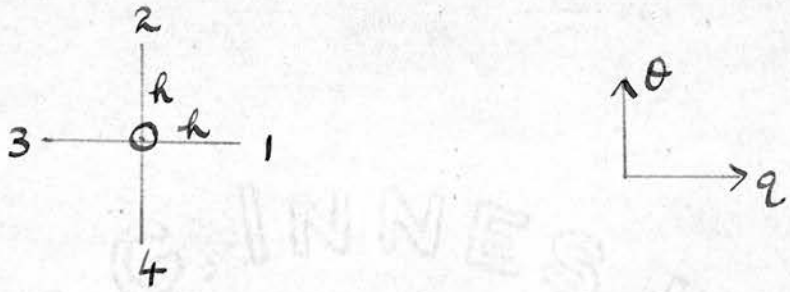
is obtained where, for convenience, the value $\frac{\gamma-1}{2}$ has been taken for H. If equations (16) and (17) are made non-dimensional by introducing a standard velocity c_s and a standard mass flow G, elimination of c/c_s once again gives (18) where q and ψ now stand for q/c_s and ψ/G respectively.

If the field to be examined in the hodograph plane is covered by a square network and the derivatives ψ_{qq}, ψ_q, $\psi_{\theta\theta}$ replaced by second order central difference formulae, (18) becomes at a typical node 0,

$$F = 2q_o^2(1-q_o^2)(\psi_1+\psi_3) + hq_o(1+4q_o^2)(\psi_1 - \psi_3)$$

$$+ 2(1-6q_o^2)(\psi_2+\psi_4) - 4(1-5q_o^2 - q_o^4)\psi_o = 0 \qquad (19)$$

for $\gamma = 1.4$, where h is the mesh length in the q- and θ-directions. To obtain a numerical solution of a compressible



flow problem, (19) must be satisfied approximately throughout the field of flow. In a relaxation solution using (19) the residuals F at nodes in the field are eliminated by altering the ψ-distribution either by trial and error or according to the pattern

$$\delta F_o = -4(1 - 5q_o^2 - q_o^4)\delta\psi_o$$

$$\delta F_1 = (-hq_1 + 2q_1^2 - 4hq_1^3 - 2q_1^4)\delta\psi_o$$

$$\delta F_3 = (hq_3 + 2q_3^2 + 4hq_3^3 - 2q_3^4)\delta\psi_o$$

$$\delta F_2 = 2(1 - 6q_2^2)\delta\psi_o$$

$$\delta F_4 = 2(1 - 6q_4^2)\delta\psi_o,$$

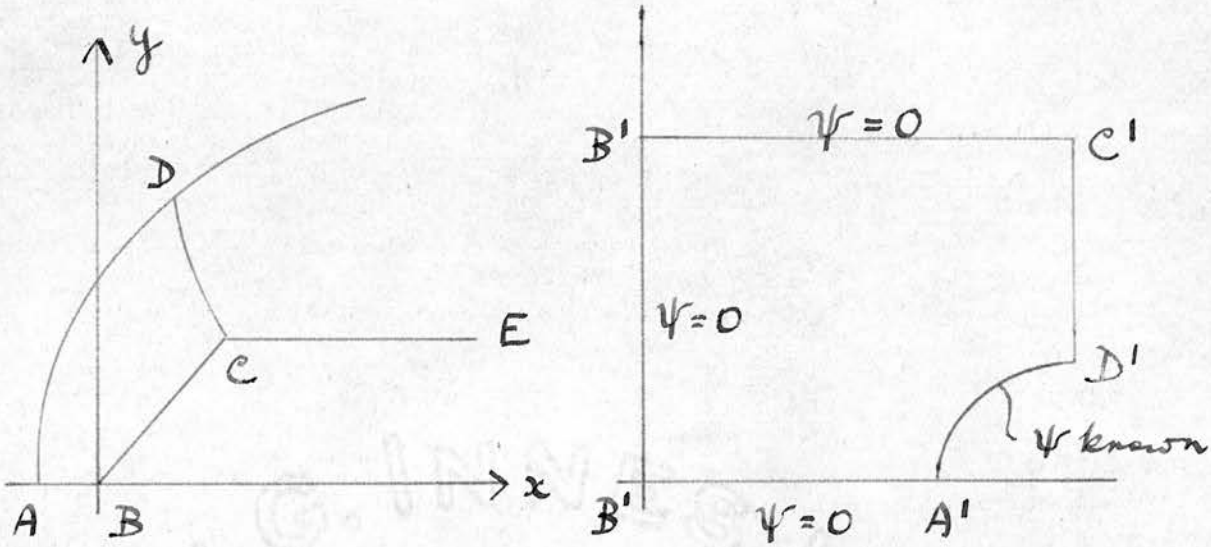
where $\delta F_o, \delta F_1, \ldots\ldots \delta F_4$ are the changes in the residuals

consequent on a modification $\delta\psi_0$ of ψ_0. In a step-by-step solution proceeding in the q-direction, (19) is used in the explicit form

$$\psi_1 = \frac{4(1-5q_0^2-q_0^4)\psi_0-2(1-6q_0^2)(\psi_2+\psi_4)-\left[2q_0^2(1-q_0^2)-hq_0(1+4q_0^2)\right]\psi_3}{2q_0^2(1-q_0^2)+hq_0(1+4q_0^2)} .$$

There is no doubt that numerical solutions, whether relaxation or step-by-step, will be obtained much more easily using (19) or (20), linear difference equations with variable coefficients, than using non-linear systems (5) and (6) or (11) or (12). This advantage, however, is more than offset in most compressible flow problems, by the difficulties of defining the boundaries of the problem in the hodograph plane and transferring a solution in the hodograph plane back to the physical plane. The former difficulty is illustrated in flow through a nozzle where although the flow direction θ is known on the nozzle walls, the velocity q is not known, and so the nozzle walls cannot be represented in the hodograph plane.

A mixed flow problem, where the subsonic part can be solved numerically in the hodograph plane, is the flow of air over a wedge with a detached shock wave. As shown, BCE represents the upper half of the wedge whose axis of symmetry is the x-axis,



AD is the detached shock, and CD the sonic line. In the hodograph plane the region $A'B'B'C'D'A'$ corresponding to the subsonic region ABCDA in the physical plane is shown, with $A'D'$ part of the shock polar. The subsonic region in the hodograph plane is covered by a square net and a numerical solution, using (19), is obtained at the nodal points from the boundary conditions, which consist of a knowledge of ψ along the open boundary $D'A'B'B'C'$. This problem in the hodograph plane has been attacked recently with computing machinery, in particular the Bell Relay Computer, at the Ballistic Research Laboratories, Aberdeen, U.S.A., but so far no details are available of the methods used or of the success achieved.

Methods Based on the Incompressible Velocity Potential and Stream Function.

It is a great advantage in carrying out a difference

solution of a flow problem if mesh points fall directly on solid boundaries. The flow of a compressible fluid, especially near the speed of sound, involves so many difficulties that it is desirable to avoid the boundary condition trouble. This is done by using as independent variables the velocity potential ξ and the stream function η of the corresponding incompressible flow problem. Channel walls of any shape in the physical x,y plane becomes straight lines parallel to the ξ axis in the ξ,η plane, and obstacles to the flow in the physical plane such as aerofoils, circular cylinders, etc. become slits parallel to the ξ-axis in the ξ,η plane. The incompressible solution of the problem is obtained either by theory or relaxation.

Several authors have carried out finite difference calculations of compressible flow problems on a grid in the ξ,η plane. In particular, Emmons (40) used the stream function ψ of the compressible flow as the dependent variable. According to Emmons the relaxation process "must be watched closely as the speed of sound is approached and becomes confusing for supersonic velocities." More recently, Thom and Woods (41) have developed a method involving $L = \log 1/q$ as the dependent variable. If the angle between the compressible and incompressible flow vectors is assumed to be negligible, L is shown to satisfy the comparatively simple equation

$$\frac{\partial^2 L}{\partial \xi^2} + \frac{\partial^2 L}{\partial \eta^2} = \frac{\partial}{\partial \xi}\left(M^2 \frac{\partial L}{\partial \xi}\right), \qquad (21)$$

where $M = q/c$, and c is again given in terms of q by Bernoulli's

equation. The difficulty of L taking infinite values at stagnation points and sharp corners is overcome by modifying the finite difference calculations in the neighbourhood of these singularities according to methods outlined by Thom (42) and Woods (43). These authors have used a finite difference approximation of (21) to solve some mixed flow problems, in particular, the flow past an aerofoil with circulation where a localised supersonic region appears on the upper surface of the aerofoil (44). In the latter reference it is stated that "relaxation in the supersonic patches is still possible, but somewhat less convergent than in the elliptical region of the differential equation." The present author feels very apprehensive about the accuracy of equation (21) in supersonic regions. In supersonic flow, increase of velocity is accompanied by divergence of the stream tubes, whereas in incompressible flow, increase of velocity is accompanied by convergence of the stream tubes. As a result in supersonic flow regions adjacent to solid boundaries with appreciable curvature, it seems unlikely that the angle between the compressible and incompressible flow vectors will be negligible, and so in such regions equation (21) will be a poor approximation to the compressible flow equation.

The methods of numerical solution of the compressible flow equation described in this chapter will now be gathered together,

and the merits and disadvantages of each outlined in brief.

(1) ψ (x,y). Solution in the physical plane seems possible except in the neighbourhood of the sonic line and in regions where the Mach number is high. Two difference equations are necessary, and the computation is rather heavy.

(2) ψ (q, θ). If the solid boundaries of a problem can be represented in the hodograph plane, a numerical solution in this plane requires only one difference equation and the calculation is much easier than in (1). An additional calculation, which may not always be possible, is necessary to transfer the solution back to the physical plane.

(3) ψ (ξ, η). The fundamental equations are comparable to those used in (1). A solution of the corresponding incompressible problem is required to start the calculation. On the credit side, solid boundaries in the physical plane are straight lines in the (ξ, η) plane, and so irregular nodes are avoided.

(4) L(ξ, η). One difference equation is sufficient and the calculations are relatively easy. The differential equation is an approximation to the compressible flow equation, and it is unlikely to be accurate for supersonic regions. As in (3), a solution of the corresponding incompressible problem is necessary to start the calculation, and irregular nodes are avoided.

Although tentative claims have been made that the position of a shock wave can be determined using methods (3) and (4), the experience of the present author is that the detection of shocks is still beyond the capabilities of finite difference techniques,

and any shock wave present must be inserted in the field before the finite difference calculation commences.

II ROTATIONAL FRICTIONLESS FLOW

Regions where rotation is present but viscosity effects can be neglected are found extensively in wakes and slipstreams and behind curved shock waves. Comparatively few exact solutions are available for problems involving a rotational frictionless fluid for the following reasons;

(1) except in some simple incompressible problems, the governing equations are non linear in the physical plane.

(2) insuperable difficulties are encountered in using the hodograph plane to obtain analytical solutions of flows with vorticity present. In the simple problem of incompressible uniform show flow past a circular cylinder, Goldstein and Lighthill (45) showed that the hodograph plane is a Riemann surface of six sheets.

In this section, rotational inviscid flow problems are classified as follows;

(1) incompressible (a) constant rotation
(b) variable rotation

(2) compressible (a) isentropic
(b) non-isentropic.

(3) axially symmetric.

Finite difference methods of solution are outlined for each type of problem, and results given for the rotational flow fields behind bow shock waves in two dimensional and axially symmetric flows.

(1) Incompressible Flow.

The stream function ψ for the rotational flow of an inviscid incompressible fluid in two dimensions satisfies the equation

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \omega = 0, \tag{22}$$

where the vorticity ω is given by

$$\omega = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}, \tag{23}$$

u and v being the velocity components parallel to the x- and y- axes respectively. The total head is constant along a streamline and is given by $H(\psi)$, from which it follows that

$$\omega = -\frac{dH}{d\psi} = f(\psi), \tag{24}$$

and so the vorticity is constant along a streamline. The pressure p anywhere in the field is given by

$$p = H(\psi) - \tfrac{1}{2}\left(\frac{\partial\psi}{\partial x}\right)^2 + \left(\frac{\partial\psi}{\partial y}\right)^2 \tag{25}$$

(a) Constant Rotation.

If the total head is given by

$$H = C - \omega_0\psi,$$

where C and ω_0 are constants, it follows from (24) that the rotation takes the constant value ω_0 everywhere in the field. Introducing some significant linear dimension h and velocity U pertaining to the specific problem under consideration,

equations (22) and (25) take the non dimensional forms

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{1}{N} = 0, \tag{26}$$

and

$$p = C - \frac{2}{N}\psi - \left(\frac{\partial\psi}{\partial x}\right)^2 + \left(\frac{\partial\psi}{\partial y}\right)^2 \tag{27}$$

where ψ, x, y, p, C now stand for ψ/hU, x/h, y/h, $p/\frac{1}{2}U^2$, $C/\frac{1}{2}U^2$ respectively and N is the non dimensional parameter $U/\omega_0 h$.

Exact solutions of (26) have been obtained for flow with constant shear past stationary cylinders with various cross sections. In particular, in the case of flow with constant shear past a general elliptic cylinder with centre at the origin, Mitchell and Murray (46) obtain the stream function in the dimensional form

$$\psi = \tfrac{1}{4}\omega_0 b^2 + U\left[c\sinh\xi - be^{(\xi_0-\xi)}\sin\eta\right.$$

$$\left. - \tfrac{1}{2}\omega_0\left[c^2\sinh^2\xi + \left\{b^2e^{2(\xi_0-\xi)} - c^2\sinh^2\xi\right\}\cos 2\eta\right]\right],$$

where the elliptic co-ordinates ξ, η are given by $x = c\cosh\xi\cos\eta$, $y = c\sinh\xi\sin\eta$, the elliptic cylinder is $\xi = \xi_0$ with b the length of the semi minor axis, and the free stream conditions at $x = -\infty$ consist of a velocity distribution $u = U - \omega_0 y$, $v = 0$ where U is constant. From this stream function it follows that the non-dimensional ordinate of the stagnation streamline at $x = -\infty$ is given by

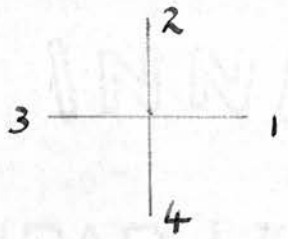
$$y = N - (N^2 + \tfrac{1}{2})^{\frac{1}{2}},$$

where $N = U/\omega_0 b$.

The finite difference approximation of (26) for a square net is

$$\sum_{r=1}^{4} \psi_r - 4\psi_0 + a^2/N = 0, \qquad (28)$$

where a is the non dimensional mesh length. In a problem involving a free stream of breadth h, ψ is restricted to the



range $0 \leq \psi \leq (1 - 1/2N)$. The region to be examined is covered by a square net with the solid boundaries of the problem passing through as many nodes as possible. An initial ψ- distribution is introduced and the values of ψ modified until (28) is satisfied at all the nodes of the field. This procedure at first sight appears to be comparable to that of obtaining a finite difference solution of Laplace's equation, and in flow through a channel this is the case. In the problem of flow past a cylinder, however, the value of ψ is not known initially on the cylinder, and this is a serious handicap to the use of finite difference methods in solving such problems. When a ψ- distribution satisfying (28) is obtained, the non-dimensional pressure at a node is given by

$$p = C - \frac{1}{4a^2}\Big[\sum_{r=1}^{4} \psi_r^{\,2} - 2(\psi_1\psi_3 + \psi_2\psi_4)\Big] - \frac{2}{N}\psi_0.$$

(b) Variable Rotation

If the total head is not a linear function of ψ, it follows from (24) that the rotation takes a different value on each streamline. Equation (22) takes the non-dimensional form

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \omega = 0 \qquad (29)$$

where ω now stands for $\omega h/U$, and dimensional ω is given by (24).

Murray and the present author, in work so far unpublished, have used (24) and (29) to obtain the stream function for incompressible flow with variable rotation past a circular cylinder, the vorticity distribution in the free stream being approximately linear in the neighbourhood of the x- axis. The stream function obtained is

$$\psi = \frac{Ur_0}{\cosh k}\sinh\left(\frac{y}{r_0} - k\right) + 2Ur_0 \tanh k \sum_{m=1}^{\infty}(-1)^m \frac{I_{2m}(1)}{K_{2m}(1)} K_{2m}\left(\frac{r}{r_0}\right)\cos 2m\theta$$

$$+ 2Ur_0 \sum_{m=0}^{\infty}(-1)^{m+1} \frac{I_{2m+1}(1)}{K_{2m+1}(1)} K_{2m+1}\left(\frac{r}{r_0}\right)\sin(2m+1)\theta, \qquad (30)$$

where the polar co-ordinates r, θ are given by $x = r\cos\theta$, $y = r\sin\theta$, the circular cylinder is $r = r_0$, the free stream conditions at $x = -\infty$ consist of a velocity distribution

$u = \frac{U}{\cosh k} \cosh\left(\frac{y}{r_0} - k\right)$, $v = 0$ where U and k are constants, and I, K are Bessel functions. In all cases considered, the stagnation streamline comes from a region where the velocity is higher than the upstream velocity on the x- axis. As far as the present author is aware, this is the only theoretical solution available in the case of variable rotation.

A finite difference solution of an incompressible problem with variable rotation can be carried out using

$$\sum_{r=1}^{4} (\psi_r) - 4\psi_0 + a^2\omega = 0, \qquad (31)$$

the difference replacement of (29), where a is the non dimensional mesh length and dimensional ω from (24) is a given function of ψ. The problem of flow past a cylinder, where the value of ψ is not known initially on the cylinder, again presents difficulties using the difference technique.

2. Compressible Flow.

The stream function ψ for the rotational flow of a frictionless compressible fluid in two dimensions satisfies the equation

$$\frac{\partial}{\partial x}\left(\frac{1}{\rho}\frac{\partial \psi}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{1}{\rho}\frac{\partial \psi}{\partial y}\right) + \omega = 0, \qquad (32)$$

where, if the entropy S and the total head H are constant along a streamline, the rotation ω is given by (47)

$$\omega = \frac{\rho}{R}\frac{\partial S}{\partial \psi} - \rho \frac{\partial H}{\partial \psi} \tag{33}$$

where R is the perfect gas constant.

(a) Isentropic Gas.

If the gas has constant entropy, the rotation ω from (34) becomes

$$\omega = -\rho \frac{\partial H}{\partial \psi} = \rho f(\psi), \tag{34}$$

from which it follows that ω/ρ is constant along a streamline. The pressure p anywhere in the field is given by Bernoulli's equation

$$\frac{\gamma}{\gamma-1}\frac{p}{\rho} + \frac{1}{2\rho^2}\left[\left(\frac{\partial \psi}{\partial x}\right)^2 + \left(\frac{\partial \psi}{\partial y}\right)^2\right] = H, \tag{35}$$

where p and ρ satisfy the adiabatic gas law

$$\frac{p}{\rho^\gamma} = k, \tag{36}$$

with k taking the same constant value everywhere in the field. Introducing a standard density ρ_0, a standard pressure p_0 and a significant linear dimension h, equations (32), (35), and (36) take the non dimensional forms

$$\nabla^2(\chi\psi) - \psi\nabla^2\chi - \frac{1}{\chi^3}\frac{\partial H}{\partial \psi} = 0, \tag{37}$$

$$\frac{p}{\rho} + \frac{\gamma-1}{2\rho^2}\left[\frac{1}{2}\nabla^2(\psi^2) - \psi\nabla^2\psi\right] = (\gamma-1)H, \tag{38}$$

and

$$p = \rho^\gamma, \tag{39}$$

where ψ, x, y, p, ρ, H now stand for $\psi/\rho_0 c_0 h$, x/h, y/h, p/p_0, ρ/ρ_0, H/c_0^2 respectively with $c_0^2 = \gamma p_0/\rho_0$ and $\gamma = c_p/c_v$ where c_p and c_v are the specific heats at constant pressure and volume respectively. Eliminating p between (38) and (39), the result

$$\chi^{2(1-\gamma)} + \frac{\gamma-1}{2}\chi^4\left[\tfrac{1}{2}\nabla^2(\psi^2) - \psi\nabla^2\psi\right] = (\gamma-1)H \qquad (40)$$

is obtained. Needless to say no exact solution has been obtained for any problem governed by equations (37) and (40).

The finite difference approximations of (37) and (40) for a square net are

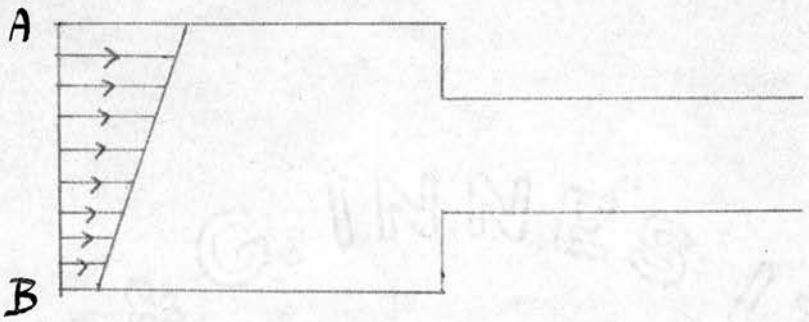
$$\sum_{i=1}^{4} \chi_i(\psi_i - \psi_0) - \frac{a^2}{\chi_0^3}\left(\frac{\partial H}{\partial \psi}\right)_0 = 0 \qquad (41)$$

and

$$\frac{1}{\gamma-1}\chi_0^{2(1-\gamma)} + \left[\psi_0^2 + \tfrac{1}{4}\sum_{i=1}^{4}(\psi_i - 2\psi_0)\psi_i\right]\frac{\chi_0^4}{a^2} = H_0 \qquad (42)$$

where a is the non dimensional mesh length. An initial ψ-distribution is introduced at the nodes of the network and χ calculated from (42) for a given total head distribution $H(\psi)$. Substituting the values of χ obtained in (41) the residuals are calculated at the nodes. The original ψ- distribution is then modified until (41) is satisfied at all the nodes of the field. This finite difference technique is unsuitable for flow past cylinders since the value of ψ is not known initially on the cylinder. As part of a general rotational flow investigation

carried out under the supervision of the present author, J.D. Murray (St. Andrews University), using the difference methods described, is calculating the incompressible and compressible fields of flow in a nozzle as shown where at the



entry AB, which is a large distance from the contraction, the flow is parallel to the axis of the nozzle, the velocity distribution is linear, and the pressure and density are of course constant. It is hoped that some of the effects of compressibility on rotational flow will become apparent from the study. Once again a big disadvantage of difference methods becomes apparent viz. that a large number of calculations involving a variety of entry conditions will be necessary before any general trends become apparent.

(b) Non-isentropic Gas.

In the flow behind a curved shock wave, the conditions although adiabatic are not isentropic. The total head H, however, has the same value on every streamline, and so from (32) and (33), the stream function satisfies the equation

$$\frac{\partial}{\partial x}\left(\frac{1}{\rho}\frac{\partial \psi}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{1}{\rho}\frac{\partial \psi}{\partial y}\right) + \frac{p}{R}\frac{\partial S}{\partial \psi} = 0. \qquad (43)$$

The entropy maintains a constant value along each stream-line, and so Bernoulli's equation for the r^{th} streamline downstream of the shock may be written

$$\left(\frac{q}{c_s}\right)^2 = \frac{2}{\gamma-1}\left\{1 - \frac{p}{\rho}\frac{(\rho_s)_r}{(p_s)_r}\right\},$$

where c_s is the speed of sound at a point of stagnation, and $(\rho_s)_r$, $(p_s)_r$ are the values of the stagnation density and pressure respectively on the r^{th} streamline downstream of the shock. The pressure may be eliminated by using the adiabatic gas law

$$\frac{p}{\rho^\gamma} = \frac{p_s}{\rho_s^\gamma} e^{(S_r - S_o)/c_v} = k_r = \frac{(p_s)_r}{(\rho_s)_r} \qquad (44)$$

where S_o, p_s, and ρ_s are the entropy, stagnation pressure, and density on all stream-lines upstream of the shock, and S_r and k_r are the entropy and adiabetic gas constant on the r^{th} stream-line downstream of the shock. Bernoulli's equation thus becomes

$$\left(\frac{q}{c_s}\right)^2 = \frac{2}{\gamma-1}\left\{1 - \left(\frac{\rho}{(\rho_s)_r}\right)^{\gamma-1}\right\},$$

which may be written in the form

$$\frac{2}{\gamma-1}\left\{1 - \left(\frac{\chi}{(\chi_s)_r}\right)^{2(1-\gamma)}\right\} = \frac{\chi^4}{c_s^2}\left\{\left(\frac{\partial\psi}{\partial x}\right)^2 + \left(\frac{\partial\psi}{\partial y}\right)^2\right\}. \qquad (45)$$

(43) and (45) are the fundamental equations for the flow behind a shock wave. It will be advantageous to write these equations

in the non-dimensional forms

$$\nabla^2(\chi\psi) - \psi\nabla^2\chi + \frac{R}{\kappa\chi}\frac{1}{\gamma v_r^2}\frac{\partial S}{\partial \psi} = 0 \qquad (46)$$

and $$\frac{2}{\gamma-1}\left\{1 - \chi^{2(1-\gamma)}\right\} = v_r^2\chi^4\left\{\left(\frac{\partial\psi}{\partial x}\right)^2 + \left(\frac{\partial\psi}{\partial y}\right)^2\right\}. \qquad (47)$$

To obtain these equations, some significant linear dimension h pertaining to the problem under consideration is chosen, and the flow parameter on the r^{th} stream-line behind the shock is expressed as

$$v_r = \frac{G}{(\rho_s)_r c_s h},$$

in which G is the mass flow per second under free stream conditions. With a slight change of notation, χ, ψ, p, x, y now stand for the non-dimensional quantities $\chi/(\chi_s)_r$, ψ/Φ, $p/(p_s)_r$, x/h, y/h. Eliminating the pressure by means of (44), (46) becomes

$$\nabla^2(\chi\psi) - \psi\nabla^2\chi + \frac{1}{\gamma v_r^2 \chi^{(2\gamma+1)}}\frac{1}{R}\frac{\partial S}{\partial \psi} = 0. \qquad (48)$$

Thus the fundamental equations (45) and (45) take the non-dimensional forms (47) and (48).

Before the last two equations can be used to evaluate the flow, it is necessary to obtain the values of v_r and $(\partial S/\partial\psi)/R$ on every stream-line downstream of the shock. As a consequence

of the fact that G/h and c_s have the same values on both sides of the shock, the flow parameter ratio for the r^{th} stream-line is given by

$$\frac{\nu_r^2}{\nu^2} = \frac{\rho_s^2}{(\rho_s)_r^2}$$

where ν is the flow parameter on all stream-lines upstream of the shock. Now using the gas laws

$$c_s^2 = \frac{\gamma p_s}{\rho_s} = \frac{\gamma (p_s)_r}{(\rho_s)_r}, \quad \frac{p_s}{\rho_s^\gamma} = k = \frac{p_1}{\rho_1^\gamma}$$

$$\frac{(p_s)_r}{(\rho_s)_r^\gamma} = k_r = \frac{p_2}{\rho_2^\gamma},$$

where the suffixes 1 and 2 indicate quantities at the shock on the upstream and downstream sides of the r^{th} stream-line respectively, the flow parameter ratio becomes

$$\frac{\nu_r^2}{\nu^2} = \left(\frac{p_2}{p_1}\right)^{2/(\gamma-1)} \left(\frac{\rho_1}{\rho_2}\right)^{2\gamma/(\gamma-1)}$$

or in terms of the ratio p_2/p_1 only,

$$\frac{\nu_r^2}{\nu^2} = \left(\frac{p_2}{p_1}\right)^{2/(\gamma-1)} \left[\frac{(\gamma-1)(p_2/p_1) + (\gamma+1)}{(\gamma+1)(p_2/p_1) + (\gamma-1)}\right]^{2\gamma/(\gamma-1)} \tag{49}$$

In order to use the method of finite differences, the field to be examined is covered with a square network. The central difference approximations to (47) and (48) are

$$R_o \equiv \chi_o^{-4} - \chi_o^{-2(1+\gamma)} = (\gamma-1)\frac{(\nu_r^2)_o}{a^2}\left\{\psi_o^2 + \frac{1}{4}\sum_{i=1}^{4}\psi_i(\psi_i - 2\psi_o)\right\} \quad (50)$$

and

$$F_o \equiv \sum_{i=1}^{4}\chi_i(\psi_i - \psi_o) + \frac{a^2}{\gamma(\nu_r^2)_o \chi_o^{(2\gamma+1)} R}\left(\frac{\partial S}{\partial \psi}\right)_o = 0 \quad (51)$$

where a is the non dimensional mesh size. Thus for a given ψ- distribution, using (50), R_o and hence χ_o can be calculated at every node of the network. Equation (51) will then enable the residual F_o to be evaluated at every node of the field. Modification of the original ψ- distribution with a view to eliminating these residuals is carried out either by trial and error or by a relaxation pattern. In the present case the changes in δF_j ($j = 0, 1, 2, 3, 4$) consequent on an assumed modification $\delta\psi_o$ of ψ_o are given by the relaxation pattern formulae

$$\delta F_o = -\left[\sum_{i=1}^{4}\frac{(\gamma-1)(\nu_r^2)_i}{2a^2}\{f'(R)\}_i(\psi_o - \psi_i)^2 + \sum_{i=1}^{4}\chi_i + \frac{(2\gamma+1)(\gamma-1)}{2\gamma\chi_o^{2(\gamma+1)}}\frac{1}{R}\left(\frac{\partial S}{\partial \psi}\right)_o\{f'(R)\}_o\left(4\psi_o - \sum_{i=1}^{4}\psi_i\right)\right]\delta\psi_o \quad (52)$$

and

$$\delta F_j = \left[\chi_o + \frac{(\gamma-1)(\nu_r^2)_o}{2a^2}\{f'(R)\}_o(\psi_o - \psi_j)\left(4\psi_o - \sum_{i=1}^{4}\psi_i\right) - \frac{(2\gamma+1)(\gamma-1)}{2\gamma\chi_j^{2(\gamma+1)}}\frac{1}{R}\left(\frac{\partial S}{\partial \psi}\right)_j\{f'(R)\}_j(\psi_o - \psi_j)\right]\delta\psi_o$$

$$(j = 1, 2, 3, 4)$$

where

$$f'(R) = \frac{\chi^5}{2\left\{(\gamma+1)\chi^{2(1-\delta)} - 2\right\}},$$

These formulae are only approximately true, as it is assumed in their derivation that $(\partial S/\partial \psi)_o$ and $(\nu_+^2)_o$ do not vary with the increment $\delta\psi_o$ of ψ_o. The consequent inaccuracy of the relaxation pattern formulae is unimportant except in the neighbourhood of the speed of sound, where a trial and error method of residual elimination is used on another account. (see pp. 154-5)

The present author (48) used the methods described to obtain a solution for the problem of a uniform parallel stream of Mach number M = 1·8 (ν = 0·4) flowing past a square nosed two dimensional obstacle. The shape of the bow shock wave was obtained from a photograph made in the N.P.L., and the extent of the field considered is from the axis of symmetry to the first streamline which is not appreciably deflected by the shock. At such a streamline the shock angle should be approximately the Mach angle $\sin^{-1}(1/1{\cdot}8)$. It is required to find the complete pattern downstream of the shock.

The significant linear dimension h, which is introduced in order to obtain non-dimensional quantities, is chosen to be five semi-widths of the square-nosed obstacle. The first stream-line not appreciably deflected by the shock is at a distance 10h from the axis of symmetry. The shock cuts this

streamline at A and the axis of symmetry at C. The sonic velocity point on the downstream side of the shock is B, the stagnation point is D, the obstacle corner is E, and the perpendicular from A meets the obstacle side at F. The methods of the previous section are now used to evaluate the field inside ABCDEF, where the rotational flow is mixed subsonic and supersonic. The problem is illustrated (Fig. 8).

The boundary conditions for this region consist of a knowledge of the stream-function ψ, the Mach number M, and the slope of the downstream streamlines along AC, and also of ψ along CDEF. It is also necessary at this stage to obtain ν_r^2 and $(\partial S/\partial \psi)_R$ at the shock on the downstream side. At every point of the shock wave the tangent makes the shock angle with the direction of the axis of symmetry. It is well known that for a given Mach number and deflexion angle there are two possible shock-wave angles. The larger angle corresponds to the strong shock and the smaller angle to the weak shock. For example, at points C and A on the bow shock wave where the angle of deflexion is zero, the shock conditions are strong and weak respectively. If G is the point on the bow shock where the streamline experiences its maximum deflexion, then all points from C to G correspond to strong shock conditions and all points from G to A to weak shock conditions. There is, of course, only one shock-wave angle possible for the maximum angle of deflexion, since the weak and strong shocks coincide at G. From an oblique shock

chart the pressure-ratio p_2/p_1 can be read off for an initial Mach number of 1·8. Substitution of this pressure ratio in (49) will then give the flow parameter ratio at every point of the shock wave. The entropy has a constant value S_0 everywhere in the field upstream of the shock, but takes a different value S_r on every streamline downstream. The entropy increase $(S_r - S_0)$ is obtained at all pointsof the shock from (57). Since the stream-function ψ is also known at every point of the shock, we may plot $(S_r - S_0)/R$ against ψ and read off the required quantity $(\partial S/\partial\psi)/R$ at all points on the shock. ν_r^2/ν^2 and $(\partial S/\partial\psi)/R$ are each plotted against ψ (Fig. 9).

The curves ν_r^2/ν^2 = constant and $(\partial S/\partial\psi)/R$ = constant are the streamlines, and so it follows that Fig. 9 may be used not only for points on the shock wave, but also for all points downstream of the shock in the field to be examined.

As already mentioned, the field inside ABCDEF is covered with a square net of non-dimensional mesh size a = 1/10. The only difficulty encountered in employing the relaxation method described to this mixed subsonic-supersonic problem occurs in the neighbourhood of the sonic line, which starts at B and ends at the obstacle corner E. As described in pp. 155-6, the position of the sonic line is obtained by approaching it from the subsonic side. The pattern formulae (52) and (53) do not yield definite results at nodes near the sonic line and so residuals are eliminated there by trial and error. Once the

position of the sonic line has been established, the supersonic region ABEF is evaluated. It might be argued that a step-by-step calculation is essential in evaluating a supersonic region. However, in this problem, the conditions are known at the downstream end of the supersonic region, (cf. pp.160-1) and so the central formulae (50) and (51) can be used. This relaxation solution has the merit that instability will not completely vitiate it, whereas instability in a step-by-step calculation will render it useless. The complete field of flow showing the lines of constant Mach number and vorticity is illustrated in (48).. Far downstream of the shock, where the flow is approximately parallel, the Mach number varies from 1·42 on the obstacle side to 1·80 on the first streamline which passes through the shock without appreciable deflexion. This illustrates the importance of taking into account the rotational nature of the flow behind the shock wave in the problem. If the flow behind the shock had been assumed irrotational, the Mach number of the flow far downstream would have been approximately 1·80 on every streamline.

Again at any point in the rotational field of flow behind the bow shock wave, the vorticity is given by

$$\omega = \frac{p}{R}\frac{\partial S}{\partial \psi}.$$

This equation can be written in the non-dimensional form

$$\omega = \frac{p}{\gamma \nu_r}\frac{1}{R}\frac{\partial S}{\partial \psi} \qquad (54)$$

where with a slight change of notation, ω, p, and ψ now stand for the non-dimensional quantities $(h/c_s)\omega$, $p/(p_s)_r$, and ψ/θ respectively.

Now, as mentioned earlier, ν_r^2/ν^2 and $(\partial S/\partial\psi)/R$ depend only on ψ and so the quantity $(\partial S/\partial\psi)/\gamma\nu_r R$ has a constant value along every streamline behind the shock. The non-dimensional quantity p, like the Mach number M, is a function of χ at the point in question, and so (54) enables the non-dimensional vorticity to be calculated everywhere in the field downstream of the shock.

Onthe shock wave itself, the value of the vorticity does not depend on the field downstream of the shock. Fig. 9 constructed from oblique shock-wave data, enables ω to be calculated from (54) without reference to the field downstream of the shock. Because p/ν_r decreases gradually with increasing ψ along the shock, the vorticity is found to be a maximum just before the point of inflexion on the graph of $(S_r - S_0)$ against ψ, which was obtained from (57). This point on the shock wave where the vorticity is a maximum is of some importance, since the stream-line starting there will always pass through regions of comparatively high vorticity.

A formula will now be derived for the angle between the streamline and the constant Mach number line through any point in a rotational compressible field. Consider the field covered by two orthogonal families of curves l = constant and n = constant, the latter family being the streamlines. The velocity q at a point is given by q = q(l,n). Thus along a line of constant

velocity,

$$\left(\frac{\partial q}{\partial l}\right)_n + \left(\frac{\partial q}{\partial n}\right)_l \left(\frac{dn}{dl}\right)_q = 0.$$

If ϕ is the angle between the constant velocity line and the streamline through a point in the field, it follows that

$$\tan\phi = \left(\frac{h_n}{h_l}\frac{dn}{dl}\right)_q = -\frac{h_n}{h_l}\frac{(\partial q/\partial l)_n}{(\partial q/\partial n)_l}$$

where $h_l dl$ and $h_n dn$ are the elementary arc lengths along the streamline and its orthogonal trajectory at the point. Now the vorticity at a point is given by

$$\omega = -\frac{1}{h_n}\left(\frac{\partial q}{\partial n}\right)_l - \frac{q}{L}$$

where L is the radius of curvature of the streamline at the point. Elimination of $(\partial q/\partial n)_l/h_n$ leads to the result

$$\tan\phi = \frac{L\frac{1}{q}\left(\frac{\partial q}{\partial s}\right)_n}{1 + \frac{L\omega}{q}},$$

where ds is the elementary arc length taken along a streamline.

Since c_s has the same value on every streamline behind a bow shock, it can be deduced from Bernoulli's equation that a line of constant velocity is also a line of constant Mach number. In addition, it is easily shown that

$$\frac{1}{q}\left(\frac{\partial q}{\partial s}\right)_n = \frac{1}{M}\left(\frac{\partial M}{\partial s}\right)_n \frac{1}{1 + (\gamma-1)M^2/2}.$$

The angle between the line of constant Mach number and the streamline through a point is therefore given by

$$\tan\phi = \frac{\left(\frac{\partial M}{\partial s}\right)_n}{\frac{M}{L}\left(1+\frac{\gamma-1}{2}M^2\right) + \omega\left(1+\frac{\gamma-1}{2}M^2\right)^{\frac{3}{2}}}, \qquad (55)$$

where all the quantities in (55) are non-dimensional. On the obstacle sides in the present problem, the radius of curvature is infinite and the vorticity zero. Thus the constant Mach number lines meet the obstacle everywhere at right angles.

It is worth while pointing out that in the rotational field of flow behind a bow shock, since c_s is constant everywhere, a line of constant dimensional velocity is a line of constant Mach number, constant non-dimensional pressure, and constant non-dimensional density. However, since $(p_s)_r$ and $(\rho_s)_r$ vary with r, lines of constant dimensional velocity are not lines of constant dimensional pressure or constant dimensional density.

(3) Axially Symmetric Flow.

Cylindrical polar co-ordinates x, r are used, x being measured along the axis of symmetry and r perpendicular to it. The components of the velocity q in these two directions are u, v respectively. The stream function ψ for the rotational axially symmetric flow of a frictionless compressible fluid satisfies the equation

$$\frac{\partial}{\partial x}\left(\frac{1}{\rho r}\frac{\partial \psi}{\partial x}\right) + \frac{\partial}{\partial r}\left(\frac{1}{\rho r}\frac{\partial \psi}{\partial r}\right) + \omega = 0, \qquad (56)$$

where if the entropy S and the total head H are constant along a streamline, the rotation is given by (47)

$$\omega = \frac{rp}{R}\frac{\partial S}{\partial \psi} - r\rho\frac{\partial H}{\partial \psi}. \qquad (57)$$

In this section, only the flow field behind a curved shock wave is considered, where the total head H has the same value on every streamline, and so from (56) and (57), the stream function satisfies

$$\frac{\partial}{\partial x}\left(\frac{1}{\rho r}\frac{\partial \psi}{\partial x}\right) + \frac{\partial}{\partial r}\left(\frac{1}{\rho r}\frac{\partial \psi}{\partial r}\right) + \frac{rp}{R}\frac{\partial S}{\partial \psi} = 0, \qquad (58)$$

which reduces to

$$\nabla^2(\chi\psi) - \psi\nabla^2\chi - \frac{\chi}{r}\frac{\partial \psi}{\partial r} + \frac{\rho r^2}{R\chi}\frac{\partial S}{\partial \psi} = 0 \qquad (58)$$

where ∇^2 denotes $\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial r^2}$.

The flow behind the shock wave although adiabatic is not isentropic. However, the entropy maintains a constant value along each streamline, and so in the manner of the corresponding two dimensional problem of the previous section, Bernoulli's equation for the r^{th} streamline behind the shock becomes

$$\frac{\chi}{\gamma-1}\left\{1-\left(\frac{\chi}{(\chi_s)_r}\right)^{2(1-\gamma)}\right\} = \frac{\chi^4}{r^2 c_s^2}\left\{\left(\frac{\partial \psi}{\partial x}\right)^2 + \left(\frac{\partial \psi}{\partial r}\right)^2\right\}. \qquad (59)$$

Equations (58) and (59) are the fundamental equations for the axially symmetric field downstream of the shock wave. In non-dimensional form, these equations become

$$\nabla^2(\chi\psi) - \psi\nabla^2\chi - \frac{\chi}{r}\frac{\partial\psi}{\partial r} + \frac{1}{\gamma\nu_r^2}\frac{p r^2}{\chi}\frac{1}{R}\frac{\partial S}{\partial\psi} = 0 \qquad (60)$$

and

$$\frac{2}{\gamma-1}\left\{1 - \chi^{2(1-\gamma)}\right\} = \frac{\nu_r^2\chi^4}{r^2}\left\{\left(\frac{\partial\psi}{\partial x}\right)^2 + \left(\frac{\partial\psi}{\partial r}\right)^2\right\} \qquad (61)$$

where χ, ψ, p, x, r now stand for the non dimensional quantities $\chi/(\chi_s)_r$, ψ/G, $p/(p_s)_r$, x/h, r/h. In these quantities, h is a significant length, $2\pi G$ is the mass flow per second under free stream conditions, $(\chi_s)_r$ and $(p_s)_r$ are the stagnation values of χ and p on the r^{th} streamline behind the shock, and the flow parameter on this streamline has been expressed as $\nu_r = G/(\rho_s)_r c_s h^2$, where c_s is the speed of sound at a point of stagnation. Eliminating the pressure p by using the adiabatic gas law, equation (60) becomes

$$\nabla^2(\chi\psi) - \psi\nabla^2(\chi) - \frac{\chi}{r}\frac{\partial\psi}{\partial r} + \frac{1}{\gamma\nu_r^2}\frac{r^2}{\chi^{2\gamma+1}}\frac{1}{R}\frac{\partial S}{\partial\psi} = 0, \qquad (62)$$

where the value of the flow parameter ν_r is given by (49).

In order to use the method of finite differences, the field to be examined is covered with a square network of non-dimensional mesh size a. The finite difference approximations to equations (61) and (62) are obtained as

$$R_o \equiv \chi_o^{-4} - \chi_o^{-2(1+\delta)} = \frac{(\gamma-1)(v_r^2)_o}{4a^2 r_o^2}\left[4\psi_o^2 + \sum_{i=1}^{4}\psi_i(\psi_i - 2\psi_o)\right] \quad (63)$$

and

$$F_o \equiv \sum_{i=1}^{4}\chi_i(\psi_i - \psi_o) - \frac{\chi_o a}{2r_o}(\psi_2 - \psi_4) + \frac{r_o^2 a^2}{\gamma(v_r)_o^2 \chi_o^{(2\delta+1)}}\frac{1}{R}\left(\frac{\partial S}{\partial \psi}\right)_o, \quad (64)$$

where the suffixes 0, 1, 2, 3, 4 are as shown. Thus for a given ψ-distribution, R_o, χ_o, and hence the residuals F_o can be calculated for each node of the net, using equations (63) and (64). The ψ-distribution is now modified either by using a relaxation pattern, or by trial and error, in order to reduce the residuals. The changes in the residuals $\delta F_j (j = 0, 1, 2, 3, 4)$ following a change $\delta\psi_o$ in ψ_o are given by the formulae

$$\delta F_o = -\left\{\sum_{i=1}^{4}\left[\chi_i + G_i(\psi_o - \psi_i)^2\right] + G_o\left[\frac{a}{2r_o}(\psi_2 - \psi_4) + \frac{(2\delta+1)r_o^2 a^2}{\gamma(v_r)_o^2 \chi_o^{2(\delta+1)}}\frac{1}{R}\left(\frac{\partial S}{\partial \psi}\right)_o\right]\left[4\psi_o - \sum_{i=1}^{4}\psi_i\right]\right\}\delta\psi_o, \quad (65)$$

and

$$\delta F_j = \left\{\chi_o + G_o(\psi_o - \psi_j)\left(4\psi_o - \sum_{i=1}^{4}\psi_i\right) - \frac{(2\delta+1)r_j^2 a^2}{\gamma(v_r)_j^2 \chi_j^{2(\delta+1)}} G_j(\psi_o - \psi_j)\frac{1}{R}\left(\frac{\partial S}{\partial \psi}\right)_j + Q_j\right\}\delta\psi_o, \quad (66)$$

where

$$C_0 = \frac{(\gamma-1)(v_r^2)_0}{2a^2 r_0^2}\left[f'(R)\right]_0,$$

$$C_j = \frac{(\gamma-1)(v_r^2)_j}{2a^2 r_j^2}\left[f'(R)\right]_j, \quad j = 1, 2, 3, 4,$$

$$f'(R) = \frac{\chi^5}{2(\gamma+1)\chi^{2(1-\delta)}-4}$$

and where Q_j has the following values for $j = 1, 2, 3, 4$:

$$Q_1 = -\frac{aG_1}{2r_1}(\psi_0 - \psi_1)(\psi_5 - \psi_8),$$

$$Q_2 = -\frac{aG_2}{2r_2}(\psi_0 - \psi_2)(\psi_{10} - \psi_0) + \frac{a\chi_2}{2r_2}$$

$$Q_3 = -\frac{aG_3}{2r_3}(\psi_0 - \psi_3)(\psi_6 - \psi_7),$$

$$Q_4 = -\frac{aG_4}{2r_4}(\psi_0 - \psi_4)(\psi_0 - \psi_{12}) - \frac{a\chi_4}{2r_4}.$$

In deriving these formulae, it has been assumed that $(v_r)_0^2$ and $\left(\frac{\partial S}{\partial \psi}\right)_0$ do not vary with the change $\delta\psi_0$ in ψ_0. The consequent inaccuracy in the relaxation pattern is unimportant.

Mitchell and McCall (49) used the methods described to

solve the problem of a uniform parallel air stream of Mach number 1·8 flowing past a square-nosed cylindrical obstacle. The position and shape of the bow shock wave formed in front of the obstacle was obtained from a photograph taken in the N.P.L. The part of the field examined downstream of the shock wave extends from the axis of symmetry to the first streamline not appreciably deflected by the shock wave. At this streamline the shock wave makes an angle $\sin^{-1}\frac{1}{1\cdot8}$ with the line of flow.

In this problem, the length h is chosen to be five times the semi-width of the obstacle, and $2\pi G$ to be the free stream flow per second through a circular cylinder of radius h. The first streamline not appreciably deflected by the shock wave is found to be a distance 2h from the axis of symmetry. The shock wave cuts this streamline in the point A, the sonic line in B and the axis of symmetry in C. The obstacle is bounded by DEF. The free stream conditions of this problem are $M = 1\cdot80$, $\rho = \cdot287$, $\nu = 1\cdot262$, $R = \cdot0324$, and $\psi = r^2$.

The field inside ABCDEF (fig.10) where the rotational flow is mixed subsonic and supersonic, is now evaluated by the methods of the previous section. A brief summary of the method used need only be given, as it is similar to that described fully in the evaluation of the similar two-dimensional problem. The boundary conditions again consist of a knowledge of the stream function ψ along ABCDEF together with the Mach number M and the

slope of the streamlines on the downstream side of the shock. In particular, knowledge of the pressure ratio $\frac{p_2}{p_1}$ across the shock enables the flow parameter ν_r to be calculated from equation (49) for every streamline. The quantity $\frac{1}{R}\frac{\partial S}{\partial \psi}$ on the downstream side of the shock wave is found from a knowledge of the entropy increase $\frac{1}{R}(S_r - S_0)$ across the shock wave at all points. $\frac{\nu_r^2}{\nu^2}$ and $\frac{1}{R}\frac{\partial S}{\partial \psi}$ are each plotted against ψ in fig. 11. These quantities are constant along each streamline, and hence fig. 11 can be used everywhere downstream of the shock. As in the two-dimensional problem, the ratio of the flow parameters on the two sides of the shock is as great as 1·6 for the streamline which crosses the shock normally.

Using an initial network of mesh size a = 1/40, it is now possible to start applying equations (63) and (64) to a distribution in the region ABCDEF. The value of R is calculated from (63) for each node, and from a graph of R against χ, the value of χ is found. The residuals F_0 are then calculated from (64). In the reduction of residuals at nodes away from the sonic line, ψ is modified using the relaxation pattern. At nodes near the sonic line, however, a trial and error method of altering the stream function is again employed. The part of the field containing the sonic line and the corner is examined in more detail using a finer net of mesh size a = 1/80. The complete field of flow

showing the lines of constant Mach number and vorticity is illustrated in (49). As in the corresponding two-dimensional problem, the Mach number well downstream of the shock wave varies from about 1·42 on the obstacle side to 1·80 on the first streamline not appreciably deflected by the shock, illustrating the importance of taking into account the rotational nature of the flow behind an axially symmetric bow shock wave. Due to the increased curvature of the shock wave near the axis, the point on the shock at which the vorticity is a maximum is much nearer the axis than in the two-dimensional case. Again the streamline starting there is a locus of high vorticity.

The non-dimensional vorticity in the rotational field of flow behind the bow shock wave is given by

$$\omega = \frac{pr}{\gamma \nu_r} \frac{1}{R} \frac{\partial S}{\partial \psi} \qquad (67)$$

where ω stands for $(h/c_s)\omega$. The quantity $(\frac{1}{\gamma \nu_r} \frac{1}{R} \frac{\partial S}{\partial \psi})$ has a constant value along each streamline behind the shock. The non dimensional vorticity is calculated from (67) at all nodes downstream of the shock. The value of the vorticity on the shock wave is of course independent of the field downstream of the shock, and since (pr/ν_r) increases with increasing ψ along the shock, the vorticity is found to be a maximum after the point of inflexion on the graph of $(S_r - S_0)$ against ψ.

In the flow fields behind bow shock waves, in both the two dimensional and axisymmetric cases, no attempt has been made to estimate the stability and convergence of the difference procedures employed. The complicated nature of the governing equations makes any worthwhile error investigation an impossible task at this stage. Sufficient to say that knowledge of the stream function and the velocity round the boundary of the region under examination will prevent serious instability in the calculated solution, whilst the truncation errors are unlikely to be serious for the mesh sizes used.

Other methods of obtaining finite difference solutions of axially symmetric compressible flow problems have been developed. Woods (50), with ϕ the velocity potential and ψ Stokes's stream function for the corresponding incompressible flow problem as independent variables, shows that L ($= \log(1/q)$), where q is the compressible velocity, satisfies the equation

$$\frac{\partial}{\partial\psi}\left(r\frac{\partial L}{\partial\psi}\right)+\frac{\partial}{\partial\phi}\left(\frac{1}{r}\frac{\partial L}{\partial\phi}\right)=\frac{\partial}{\partial\phi}\left[\frac{1}{r}\left\{M^2\frac{\partial L}{\partial\phi}+\frac{\sin\theta_i}{rq_i}\right\}\right]+\frac{\partial}{\partial\psi}\left(\frac{\omega}{qq_i}\right), \quad (68)$$

where the angle between the corresponding compressible and incompressible flow vectors has been neglected. The finite difference calculation involves a central difference replacement of (68). A solution of the corresponding incompressible problem is necessary to start the calculation, and from this solution, r, q_i, and θ_i are obtained as functions of ϕ and ψ where q_i and θ_i are the incompressible values of q and θ. In

the opinion of the present author, this method will be easy to use only if the flow is irrotational ($\omega = 0$). When rotation is present, it will be very difficult to apply unless ω is known initially at nodes on the (ϕ, ψ) grid. As with the corresponding two dimensional equation (21), equation (68) is unlikely to be accurate in supersonic regions.

At the Ballistic Research Laboratories, Aberdeen, U.S.A., Giese, Clippinger, and Carter (51) used the ENIAC to evaluate the axially symmetric rotational supersonic flow field behind a shock wave attached to the nose of a pointed body of revolution. These authors expressed the equations for steady axisymmetric flow in terms of the independent variables α and β, where $\alpha(x,r) = \text{constant}$ and $\beta(x,r) = \text{constant}$ are the two families of characteristic curves. To obtain a numerical solution, the above equations together with the appropriate boundary conditions are replaced by linear difference equations which are solved step-by-step on a square network in the characteristic plane. A measure of the truncation error in this difference procedure is obtained by calculating the expansion fan of a cone cylinder combination for several different values of the mesh length.

III VISCOUS FLOW

Due to the complicated nature of the Navier Stokes equation, no satisfactory theoretical solutions are available for viscous flow problems in two dimensions. In the case of a uniform incompressible viscous flow past a semi-infinite flat plate, lying in the direction of the stream, solutions obtained by Carrier and Linn (52), Dean (53), and others are not accurate over the complete field. For example, Carrier and Linn's solution is inaccurate in regions of the field where the velocity is relatively high, whereas Dean's solution is not applicable to the region round the leading-edge of the plate.

Introduction of the Prandtl boundary layer assumptions to the viscous flow equations has enabled theoretical solutions to be obtained in both the incompressible and compressible cases. In the problem of flow past a cylinder (including the flat plate), these solutions have two major disadvantages;
(1) in order to obtain the boundary layer solution it is necessary to postulate the distribution of pressure on the cylinder, a result which should emerge from the solution of the complete field.
(2) the boundary layer solution is unlikely to be accurate in the vicinity of the effective leading edge of the cylinder.

Theoretical solutions of viscous flow problems, valid over the complete field of flow, do not seem likely in the near future and so there is a great need for numerical solutions of

such problems. The former part of this section will be devoted to giving an account of the success so far achieved in applying finite difference methods to incompressible viscous flow problems. No reference is made to the equations governing the flow of a compressible viscous fluid, as it is felt that these equations will prove intractable even to finite difference methods, for some time to come. The latter part of the section will be concerned with the application of difference methods to the boundary layer equations.

Incompressible Viscous Flow.

The equations governing the two dimensional flow of an incompressible viscous fluid are

$$u \frac{\partial \omega}{\partial x} + v \frac{\partial \omega}{\partial y} = \nu \nabla^2 \omega, \tag{69}$$

$$\nabla^2 \psi + \omega = 0, \tag{70}$$

where ω is the vorticity and ν is the coeffient of kinematic viscosity. Introducing a basic length L and a basic velocity U, equations (69) and (70) take the non dimensional forms

$$u \frac{\partial \omega}{\partial x} + v \frac{\partial \omega}{\partial y} = \frac{1}{R} \nabla^2 \omega, \tag{71}$$

$$\nabla^2 \psi + \omega = 0, \tag{72}$$

where u, v, ω, x, y, ψ now stand for u/U, v/U, (Lω)/U, x/L, y/L, ψ/(UL) respectively, and R = UL/ν is the Reynolds Number. Several finite difference solutions of (71) and (72) have been

obtained for flow past a circular cylinder, with zero velocity on the cylinder, where L is the radius of the circle and U is the free stream velocity. A brief description of these methods will now be given.

Thom (54) avoided irregular nodes in the finite difference calculations by changing the independent variables from x and y to ξ and η the velocity potential and stream function for irrotational flow past the cylinder. Equations (71) and (72) now become

$$\frac{\partial \psi}{\partial \eta}\frac{\partial \omega}{\partial \xi} - \frac{\partial \psi}{\partial \xi}\frac{\partial \omega}{\partial \eta} = \frac{1}{R}\left(\frac{\partial^2 \omega}{\partial \xi^2} + \frac{\partial^2 \omega}{\partial \eta^2}\right) \tag{73}$$

$$Q^2\left(\frac{\partial^2 \psi}{\partial \xi^2} + \frac{\partial^2 \psi}{\partial \eta^2}\right) + \omega = 0, \tag{74}$$

where

$$Q^2 = \left(\frac{\partial \xi}{\partial x}\right)^2 + \left(\frac{\partial \xi}{\partial y}\right)^2 = \left(\frac{\partial \eta}{\partial x}\right)^2 + \left(\frac{\partial \eta}{\partial y}\right)^2 .$$

Equations (73) and (74) differ from (71) and (72) only by Q^2, where Q is the "velocity" of the transformation. The field of computation is the ξ, η plane in which the cylinder, irrespective of its shape, becomes a rectilinear slit. The shape of the cylinder influences Q only in equations (73) and (74). In the case of flow past a circular cylinder,

$$\xi = x\left(1 + \frac{1}{x^2+y^2}\right),$$

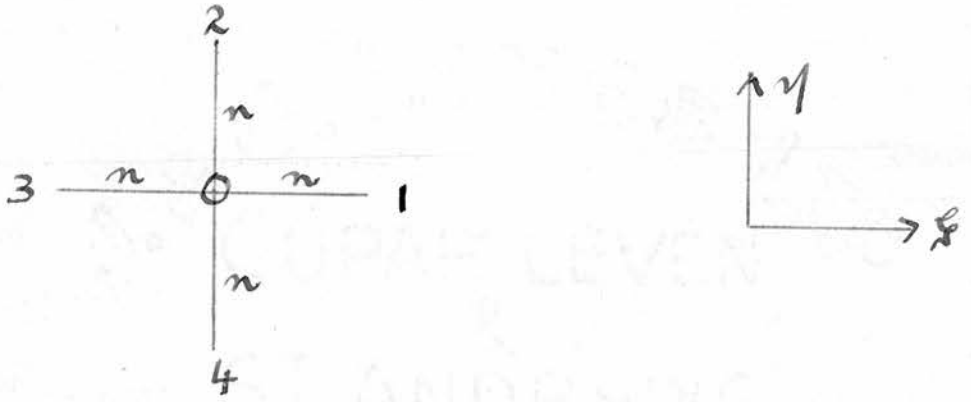
$$\eta = y\left(1 - \frac{1}{x^2+y^2}\right),$$

$$Q^2 = 1 + \frac{1-2(x^2-y^2)}{(x^2+y^2)^2}$$

The ξ, η plane in which the cylinder is a rectilinear slit is covered by a square net of mesh size n. The finite difference approximations of (73) and (74) are

$$(\psi_2 - \psi_4)(\omega_1 - \omega_3) - (\psi_1 - \psi_3)(\omega_2 - \omega_4) = \frac{4}{R}\left(\sum_{i=1}^{4} \omega_i - 4\omega_0\right), \quad (75)$$

$$\omega_0 = -\frac{Q_0^2}{n^2}\left(\sum_{i=1}^{4} \psi_i - 4\psi_0\right). \quad (76)$$



From an original ψ- distribution, ω is calculated at each non-boundary node using (76). At a node on the slit, ω is obtained from the values of ψ on the normal to the slit through the node. For a prescribed Reynold's number, the appropriate values of ω and ψ are then substituted in (75). The original ψ- distribution is altered by trial and error or according to a pattern until (75) is satisfied at all non-boundary nodes of the field. The boundary condition of zero velocity on the cylinder is of course replaced by a difference formula, which must be satisfied in the neighbourhood of the slit. The final values

of ψ and ω are transferred back to the x, y plane. Thom carried out calculations for R = 10 and R = 20 using his "Method of Squares". This technique differs from the relaxation technique described above by starting off with assumed values of both ψ and ω at the nodes of the square network. Instead of calculating residuals at the nodes, Thom uses rearranged forms of (75) and (76) to improve directly the values of ψ and ω. Fundamentally there is little difference between the "Method of Squares" and the "Method of Relaxation" as applied to boundary value problems.

Allen and Southwell (55), in the manner of Thom, used ξ and η as independent variables. In addition, a further change of variables enabled these authors to obtain fundamental equations which are substantially independent of Reynold's number. Thus relaxation solutions can be obtained for a range of values of R with a minimum of effort. Calculations were carried out for R = 0, 1, 10, 100, 1000. Now from experimental evidence (56, p.418) it is known that viscous flow past a circular cylinder changes from a steady to an unsteady state with an increase in Reynold's number. The transition value of R is uncertain but is probably about 50. Allen and Southwell explain their apparently steady solutions at R = 100, 1000 by asserting that these solutions would not be realizable in practise because the smallest disturbance would upset them.

Finally Kawaguti (57), using a transformation which

changes the infinite physical plane into a finite rectangular region, employs an iterative finite difference method inside the rectangle, and so obtains a solution for the viscous flow past a circular cylinder for R = 40. According to Kawaguti, the numerical integration took about one year and a half with twenty working hours every week.

A problem which so far has not been tackled by finite difference techniques is the flow of an incompressible viscous fluid past a semi-infinite flat plate lying in the direction of the stream. This problem, in which there is no fundamental length, appears to be simpler than that of flow past a circular cylinder, since the calculation can be carried out on a square net without irregular nodes in the physical x, y plane. An interesting comparison could be made between this solution and the corresponding theoretical boundary layer solution, and a measure of the error in the boundary layer solution near the leading edge of the plate, obtained.

In conclusion, by putting R = 0, equation (71) simplifies to

$$\nabla^2 \omega = 0, \tag{77}$$

Eliminating ω between (72) and (77), the stream function ψ for infinitely slow flow is seen to satisfy the biharmonic equation

$$\nabla^4 \psi = 0. \tag{78}$$

A "slow flow" solution to a viscous problem is thus obtained by solving (78) or (72) and (77), together with the appropriate boundary conditions. Thom used finite difference replacements of the latter pair of equations to solve the problem of slow

flow past a baffle in a channel (58) and at the mouth of a Stanton Pitot (59).

It will be emphasized again that all the problems considered so far in this section on viscous flow are boundary value problems, and so instability if present will not necessarily vitiate the solutions obtained.

The Incompressible Boundary Layer.

The equations for steady laminar boundary-layer flow of an incompressible fluid are

$$u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = u_1\frac{du_1}{dx} + \nu\frac{\partial^2 u}{\partial y^2}\ , \tag{79}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \tag{80}$$

where $u_1(x)$ is the mainstream velocity. The non-dimensional forms of these equations are

$$u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = u_1\frac{du_1}{dx} + \frac{\partial^2 u}{\partial y^2} \tag{81}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \tag{82}$$

where x, y, u, v, u_1, now stand for x/L, $(yR^{1/2})/L$, u/U, $(vR^{1/2})/U$, u_1/U, with R the Reynolds number and U, L the standard velocity and length respectively. Eliminating v between (81) and (82), the result

$$u\frac{\partial u}{\partial x} - \frac{\partial u}{\partial y}\int \frac{\partial u}{\partial x}\,dy = u_1\frac{du_1}{dx} + \frac{\partial^2 u}{\partial y^2} \qquad (83)$$

is obtained. Several numerical procedures for solving (83) have been outlined.

Hartree, replacing derivatives in the x-direction by finite differences and the other quantities by averages, related the velocity distribution through the boundary layer at one section to that at a neighbouring section. Using this step-by-step formula, he obtained solutions in the case of retarded flow, where the velocity in the main stream decreases linearly with the distance downstream (60), and also in the case where the pressure distribution in the free stream is taken to be Schubauer's observed pressure distribution for an elliptic cylinder (61). The errors in replacing the derivatives by finite differences are estimated by carrying out two integrations over the same range of x using intervals of different size. In a very recent paper, Leigh (62) outlined a method, based on Hartree's approximation of (83), which is particularly suitable for an automatic computer.

In Germany, Schröder (63), Görtler (64), and Witting (65) obtained numerical solutions of (83) for a variety of problems using the method of finite differences. In particular, the last two authors replaced the derivatives $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$, $\frac{\partial^2 u}{\partial y^2}$ in (83) by central difference formulae and the integral by a difference formula obtained by using the trapezium rule. The resulting finite difference replacement of (83) is second order

in x, an order higher than (83) itself. Since (83) is a parabolic type differential equation, it is essential that the difference replacement of (83) is stable. If this is not so, any step-by-step calculation using the difference formula will be of doubtful value. Accordingly the differential equation (83), or more conveniently (81) together with (82), and the higher order central difference replacement will now be examined for stability.

Suppose small errors ξ and η are present in the velocity components u and v respectively at a step in the calculation. Then from (81) and (82), correct to the first order in ξ and η, errors are propagated according to the system of linear differential equations

$$u\frac{\partial \xi}{\partial x} + \frac{\partial u}{\partial x}\xi + v\frac{\partial \xi}{\partial y} + \frac{\partial u}{\partial y}\eta = \frac{\partial^2 \xi}{\partial y^2} \tag{84}$$

$$\frac{\partial \xi}{\partial x} + \frac{\partial \eta}{\partial y} = 0. \tag{85}$$

In the usual manner, suppose solutions exist of the form

$$\xi = Ce^{\alpha x} e^{i\beta y}$$

$$\eta = De^{\gamma x} e^{i\delta y},$$

where β and δ are real. Substituting in (84) and (85), the results

$$\alpha = \gamma, \quad \beta = \delta, \quad \alpha C + i\delta D = 0,$$

$$\alpha\left(u + \frac{i}{\delta}\frac{\partial u}{\partial y}\right) + \left(\frac{\partial u}{\partial x} + \beta^2 + i\beta v\right) = 0, \tag{86}$$

are obtained. Now the condition for stability (p. 76) leads to

$$\text{Rl. } \alpha = -\frac{u\left(\frac{\partial u}{\partial x} + \beta^2\right) + v\frac{\partial u}{\partial y}}{u^2 + \frac{1}{\beta^2}\left(\frac{\partial u}{\partial y}\right)^2} \leqq 0, \qquad \text{(all } \beta\text{)}$$

which reduces to

$$u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} \geqq 0. \tag{87}$$

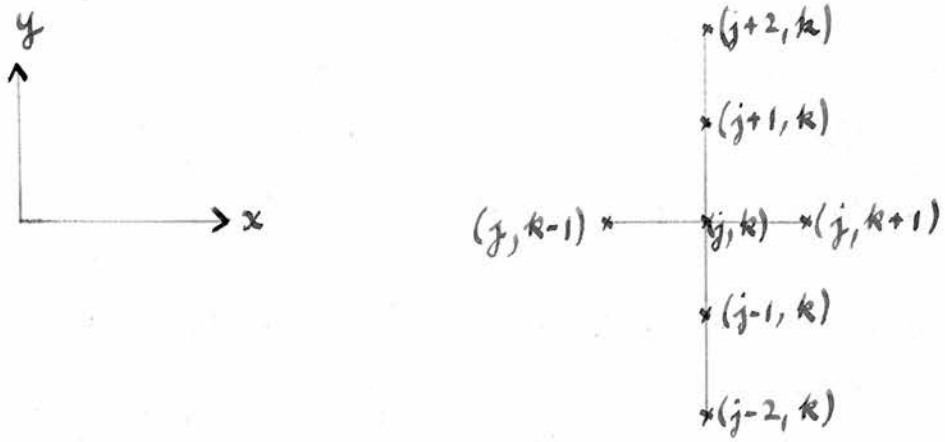
In the manner of the last paragraph, the error equations corresponding to the central difference replacements of (81) and (82) are

$$u_{j,k}\frac{1}{2\Delta x}\left(\xi_{j,k+1} - \xi_{j,k-1}\right) + \frac{1}{2\Delta x}\left(u_{j,k+1} - u_{j,k-1}\right)\xi_{j,k}$$

$$+ v_{j,k}\frac{1}{2\Delta y}\left(\xi_{j+1,k} - \xi_{j-1,k}\right) + \frac{1}{2\Delta y}\left(u_{j,k+1} - u_{j,k-1}\right)\eta_{j,k}$$

$$= \frac{1}{4(\Delta y)^2}\left(\xi_{j+2,k} + \xi_{j-2,k} - 2\xi_{j,k}\right), \tag{88}$$

$$\frac{1}{2\Delta x}\left(\xi_{j,k+1} - \xi_{j,k-1}\right) + \frac{1}{2\Delta y}\left(\eta_{j+1,k} - \eta_{j-1,k}\right) = 0. \tag{89}$$



Again if solutions exist of the form

$$\xi_{j,k} = Ce^{\alpha k \Delta x} e^{i\beta j \Delta y}$$

$$\eta_{j,k} = De^{\gamma k \Delta x} e^{i\delta j \Delta y},$$

substituting in (88) and (89), the results

$$\alpha = \gamma, \quad \beta = \delta, \quad \frac{1}{2\Delta x} C\left(e^{\alpha\Delta x} - e^{-\alpha\Delta x}\right) + \frac{i}{\Delta y} D \sin\beta\Delta y = 0,$$

$$\frac{1}{2\Delta x} u\left(e^{\alpha\Delta x} - e^{-\alpha\Delta x}\right) + \frac{\partial u}{\partial x} + \frac{i}{\Delta y} v \sin\theta + i\frac{\partial u}{\partial y} \frac{e^{\alpha\Delta x} - e^{-\alpha\Delta x}}{2\frac{\Delta x}{\Delta y}\sin\theta} + \frac{\sin^2\theta}{(\Delta y)^2} = 0 \quad (90)$$

are obtained where $\beta\Delta y = \theta$. Equation (90) can be simplified to give

$$A^2 + 2ZA - 1 = 0, \quad (91)$$

where

$$A = e^{\alpha\Delta x}, \quad Z = \Delta x \frac{\left(\frac{\partial u}{\partial x} + \frac{\sin^2\theta}{(\Delta y)^2}\right) - iv\frac{\sin\theta}{\Delta y}}{u + i\frac{\partial u}{\partial y}\frac{\Delta y}{\sin\theta}}. \quad (92)$$

Since the modulus of one of the roots of (91) exceeds unity, the procedure is always unstable. (p. 8) This illustrates again the danger of replacing a differential equation by a higher order difference equation.

Witting is now aware of the instability of the central difference replacement and in a paper in the press he explains this instability and also determines the stability condition for a difference equation of the same order as the differential equation. Such a difference equation is

$$u_{j,k}\frac{1}{\Delta x}(u_{j,k+1}-u_{j,k})+v_{j,k}\frac{1}{2\Delta y}(u_{j+1,k}-u_{j-1,k})$$

$$=\left(u_1\frac{du_1}{dx}\right)_{j,k}+\frac{1}{(\Delta y)^2}(u_{j+2,k}+u_{j-2,k}-2u_{j,k}), \tag{93}$$

together with

$$\frac{1}{\Delta x}(u_{j,k+1}-u_{j,k})+\frac{1}{2\Delta y}(v_{j+1,k}-v_{j-1,k})=0. \tag{94}$$

In an exactly similar manner to that explained in the previous paragraph, the result

$$A=1-Z \tag{95}$$

is obtained for (93) and (94), where A and Z are given by (92). After some manipulation, the condition for stability is obtained from (95) as

$$\Delta x \leqslant \frac{2\left[\left(u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}\right)+(u\sin^2\theta)/(\Delta y)^2\right]}{\left[\frac{\partial u}{\partial x}+(\sin^2\theta)/(\Delta y)^2\right]^2+(v^2\sin^2\theta)/(\Delta y)^2}. \tag{96}$$

Comparing (87) and (96) it is seen that provided the differential equation is stable, the difference scheme (93) and (94) is also stable provided the step forward Δx does not exceed the positive quantity on the right-hand side of (96).

It is surprising that no attempt has been made to obtain numerical solutions of the boundary layer equations in von Mises's form. Accordingly, J.Y. Thomson (St. Andrews University) and the present author are examining the possibility and so far the results obtained are quite encouraging. For an incompressible

boundary layer, with x and ψ as independent variables, the equation of motion in von Mises's form is

$$u \frac{\partial u}{\partial x} = u_1 \frac{du_1}{dx} + \nu u \frac{\partial}{\partial \psi} \left(u \frac{\partial u}{\partial \psi} \right) .$$

In non-dimensional notation, this becomes

$$\frac{\partial u}{\partial x} = \frac{u_1}{u} \frac{du_1}{dx} + \frac{1}{2} \frac{\partial^2 (u^2)}{\partial \psi^2} ,$$

where x, u, u_1, ψ now stand for (x/L), (u/U), (u_1/U), $(R^{1/2}\psi)/(UL)$ respectively with U, L the standard velocity and length, and R the Reynold's number. In the problem of flow along a flat plate with no adverse pressure gradient, (97) reduces to

$$u \frac{\partial u}{\partial x} = \frac{1}{2} \frac{\partial^2 (u^2)}{\partial \psi^2} . \qquad (98)$$

The error equation corresponding to (98), correct to the first order in ϵ, is

$$\frac{\partial \epsilon}{\partial x} = u \frac{\partial^2 \epsilon}{\partial \psi^2} + 2 \frac{\partial u}{\partial \psi} \frac{\partial \epsilon}{\partial \psi} + \frac{\partial^2 u}{\partial \psi^2} \epsilon$$

where ϵ is the error in u. Substituting

$$\epsilon = e^{\alpha x} e^{i \beta \psi} \qquad (\beta \text{ real})$$

in (99), the result

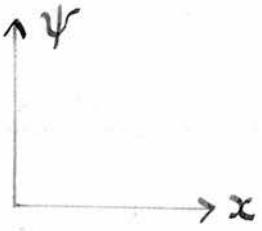
$$\alpha = - u \beta^2 + \frac{\partial^2 u}{\partial \psi^2} + 2 i \beta \frac{\partial u}{\partial \psi}$$

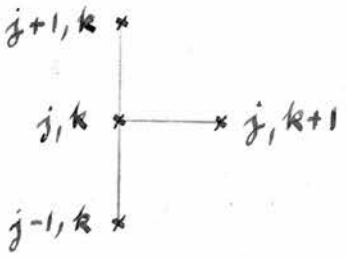
is obtained, and so (98) is stable if

$$\frac{\partial^2 u}{\partial \psi^2} \leq 0 .$$

Now a difference equation of the same order as (98) is

$$\frac{1}{\Delta x}\left(u_{j,k+1}-u_{j,k}\right)=\frac{1}{2(\Delta\psi)^2}\left[(u^2)_{j+1,k}+(u^2)_{j-1,k}-2(u^2)_{j,k}\right], \tag{101}$$





with the corresponding error equation

$$\frac{1}{\Delta x}\left(\epsilon_{j,k+1}-\epsilon_{j,k}\right)=\frac{1}{(\Delta\psi)^2}\left[u_{j+1,k}\epsilon_{j+1,k}+u_{j-1,k}\epsilon_{j-1,k}-2u_{j,k}\epsilon_{j,k}\right] \tag{102}$$

Substituting

$$\epsilon_{j,k}=e^{\alpha k\Delta x}\,e^{i\beta j\Delta\psi}$$

in (102), the result

$$e^{\alpha\Delta x}=1+\delta\left[(u_{j+1,k}+u_{j-1,k})\cos(\beta\Delta\psi)-2u_{j,k}+i(u_{j+1,k}-u_{j-1,k})\sin(\beta\Delta\psi)\right] \tag{103}$$

is obtained where $\delta=(\Delta x)/(\Delta\psi)^2$. The condition for stability $|e^{\alpha\Delta x}|\leq 1$ leads to

$$\delta\leq 2\,\frac{-(u_{j+1,k}+u_{j-1,k}-2u_{j,k})+2(u_{j+1,k}+u_{j-1,k})\sin^2\frac{\beta\Delta\psi}{2}}{\left\{(u_{j+1,k}+u_{j-1,k}-2u_{j,k})-2(u_{j+1,k}+u_{j-1,k})\sin^2\frac{\beta\Delta\psi}{2}\right\}^2+(u_{j+1,k}-u_{j-1,k})^2\sin^2\beta\Delta\psi} \tag{104}$$

From (104) it follows that the step-by-step procedure using (101) will be stable if

$$(u_{j+1,k} + u_{j-1,k} - 2u_{j,k}) \leqq 0, \tag{105}$$

and when this condition is satisfied, if

$$0 < \delta \leqq \tfrac{1}{2} . \tag{106}$$

Comparing (100) and (105) it is seen that the same condition is necessary for the stability of the differential and corresponding difference equation. Using (101), Thomson carried out a numerical calculation over a limited range of the boundary layer on a flat plate with constant velocity in the main stream. Throughout the calculation, conditions (105) and (106) were satisfied, and the results obtained are in agreement with Howarth's theoretical solution for the problem.

For flow in the boundary layer along a flat plate with adverse pressure gradient in the main stream it is advantageous to introduce

$$z = u_1^2 - u^2$$

as non-dimensional dependent variable.

Thus (97) becomes

$$\frac{\partial z}{\partial x} = (u_1^2 - z)^{\frac{1}{2}} \frac{\partial^2 z}{\partial \psi^2} , \tag{107}$$

with the corresponding linear error equation

$$\frac{\partial \epsilon}{\partial x} = (u_1^2 - z)^{\frac{1}{2}} \frac{\partial^2 \epsilon}{\partial \psi^2} - \frac{1}{2} \frac{\partial^2}{\partial \psi^2} \frac{\epsilon}{(u_1^2 - z)^{\frac{1}{2}}} .$$

If

$$\epsilon = e^{\alpha x} e^{i\beta\psi}, \qquad (\beta \text{ real})$$

the result

$$\alpha = -\left[\frac{1}{2(u_1^2 - z)^{1/2}} \frac{\partial^2 z}{\partial \psi^2} + \beta^2 (u_1^2 - z)^{1/2}\right]$$

is obtained, and so the condition for the stability of (107) is

$$\frac{\partial^2 z}{\partial \psi^2} \geq 0 . \tag{108}$$

The conventional four point difference replacement of (107) is

$$\frac{1}{\Delta x}(z_{j,k+1} - z_{j,k}) = (u_1^2 - z_{j,k})^{1/2} \frac{1}{(\Delta \psi)^2}(z_{j+1,k} + z_{j-1,k} - 2z_{j,k}) \tag{109}$$

with the corresponding linear error equation

$$\epsilon_{j,k+1} = \epsilon_{j,k} + \delta(u_1^2 - z_{j,k})^{1/2}\left[\epsilon_{j+1,k} + \epsilon_{j-1,k} - 2\epsilon_{j,k} - \frac{z_{j+1,k} + z_{j-1,k} - 2z_{j,k}}{2(u_1^2 - z_{j,k})}\epsilon_{j,k}\right]$$

where ϵ is the error in z. If

$$\epsilon_{j,k} = e^{\alpha k \Delta x} e^{i\beta j \Delta \psi}$$

it follows that

$$e^{\alpha \Delta x} = 1 - \frac{\delta}{2u}\left[8u^2 \sin^2 \frac{\beta \Delta \psi}{2} + (z_{j+1,k} + z_{j-1,k} - 2z_{j,k})\right] .$$

The condition for stability $|e^{\alpha \Delta x}| \leq 1$ gives

$$z_{j+1,k} + z_{j-1,k} - 2z_{j,k} \geq 0 \tag{110}$$

together with

$$0 < \delta \leq \frac{1}{2} . \tag{111}$$

As expected, (108) and (110) constitute the same condition. Thus a step-by-step calculation using (109) will be stable provided conditions (110) and (111) are satisfied during the

calculation. A calculation using this technique has not yet been carried out for flow in the boundary layer along a flat plate with adverse pressure gradient. It will be interesting to see how accurately the new technique locates the separation point, as most numerical procedures lose accuracy in the neighbourhood of the point of separation of the boundary layer. It can be stated at this stage, however, that the finite difference replacements of the boundary layer equations in von Mises's form are simpler and the conditions for stability easier to satisfy than the difference replacements and stability conditions of the boundary layer equations in their original form.

The Compressible Boundary Layer.

Theoretical solutions of the compressible boundary layer for the problem of flow along a flat plate are based on some or all of the following assumptions:

(1) no adverse pressure gradient along the plate.

(2) Prandtl number unity.

(3) ω unity, where $\mu = aT^{\omega}$, μ being the coefficient of viscosity, T the temperature and a constant.

Accordingly, there is need for a numerical solution of the above problem with an adverse pressure gradient in the main stream, and with σ (Prandtl number) and ω taking the values 0·72 and 0·89 respectively. These are the currently popular values of σ and ω for air, over the temperature range of most problems.

Attempts at numerical solution of the compressible boundary layer equations have already been made by Cope and Hartree (66) and Cadd (67). The former authors worked with the equations in their original form in the x, y plane, whereas the latter author used Crocco's transformed boundary layer equations with the viscous stress and enthalpy in terms of the independent variables x and u. Neither procedure was particularly successful, and so the present author is examining the possibility of obtaining finite difference solutions of the compressible boundary layer equations in von Mises's form, viz.

$$u \frac{\partial u}{\partial x} = \frac{\rho_1 u_1}{\rho} \frac{du_1}{dx} + u \frac{\partial}{\partial \psi} \left(\mu \rho u \frac{\partial u}{\partial \psi} \right) \tag{112}$$

$$\frac{\partial i}{\partial x} = - \frac{\rho_1 u_1}{\rho} \frac{du_1}{dx} + \frac{\partial}{\partial \psi} \left(\frac{\mu \rho u}{\sigma} \frac{\partial i}{\partial \psi} \right) + \mu \rho u \left(\frac{\partial u}{\partial \psi} \right)^2 , \tag{113}$$

where i (enthalpy) = $c_p T$ with c_p the specific heat at constant pressure. Equations (112) and (113) are non linear parabolic equations, and together with

$$\mu = \mu(i), \quad \rho = \frac{\gamma}{\gamma - 1} \frac{p}{i}, \quad p = p(x), \quad \sigma = \text{constant},$$

constitute two equations in the two unknowns u and i.

Making the substitution

$$q = u^2,$$

equations (112) and (113) become

$$\frac{\partial q}{\partial x} = \left[\frac{1}{i_1} \frac{dq_1}{dx} \right] i + \left[\frac{\gamma p}{\gamma - 1} \right] q^{\frac{1}{2}} \frac{\partial}{\partial \psi} \left(\frac{\mu(i)}{i} \frac{\partial q}{\partial \psi} \right) \tag{114}$$

$$\frac{\partial i}{\partial x} = -\left[\frac{1}{2I_1}\frac{dq_1}{dx}\right]i + \left[\frac{\gamma_p}{(\gamma-1)\sigma}\right]\frac{\partial}{\partial\psi}\left(q^{\frac{1}{2}}\frac{\mu(i)}{i}\frac{\partial i}{\partial\psi}\right) + \left[\frac{\gamma_p}{4(\gamma-1)}\right]\left(\frac{\mu(i)}{i}\frac{1}{q^{\frac{1}{2}}}\left(\frac{\partial q}{\partial\psi}\right)^2\right) \tag{115}$$

where the quantities in the square brackets depend only on the main stream conditions and so are known at the start of the calculation. Equations (114) and (115) can be approximated to by difference formulae in which first order forward differences take the place of $\frac{\partial q}{\partial x}$ and $\frac{\partial i}{\partial x}$, and central differences replace derivatives with respect to ψ. The boundary conditions necessary to carry out a step-by-step calculation using the difference replacements of (114) and (115) alternately are the main stream conditions together with

(1) $q = 0$ at $\psi = 0$

(2) i or $\frac{\partial i}{\partial\psi}$ known at $\psi = 0$.

(3) $q = q(\psi)$ at $x = x_1$.

(4) $i = i(\psi)$ at $x = x_1$.

Provided $q(\psi)$ and $i(\psi)$ can be obtained at a station $x = x_1$, the above step-by-step procedure looks reasonably straight forward. The accuracy of the answer, however, will depend on the stability of the procedure and so far no attempt has been made to examine the stability when an adverse pressure gradient is present and when $\mu(i)$ and σ take the values $aT^{0.89}$ and 0.72 respectively.

The finite difference procedure outlined above is simplified considerably if any of the assumptions, mentioned at the start of the section, are made. For example, if the main stream conditions are uniform, (112) reduces to

$$\frac{\partial u}{\partial x} = \frac{\partial}{\partial\psi}\left(\mu\rho u\frac{\partial u}{\partial\psi}\right). \tag{116}$$

If further, $\mu = aT$, (116) simplifies to

$$\frac{\partial u}{\partial x} = \frac{ap_1}{2R} \frac{\partial^2(u^2)}{\partial \psi^2} \qquad (117)$$

where p_1 is the constant value of the pressure everywhere in the boundary layer and R is the perfect gas constant. Except for the constant factor $(\frac{ap_1}{R})$, (117) is identical with (98) the equation in the incompressible case. Consequently the difference replacement and the stability conditions outlined for (98) will require only slight modification in order to be applicable to (117). The enthalpy is obtained from a modified version of (115) together with the calculated distribution of u. On the other hand if $\sigma = 1$ and $\mu = aT^{\omega}$, (116) reduces to

$$\frac{\partial u}{\partial x} = -\left(\frac{ap_1}{\omega R c_p^{\omega-1}}\right) \frac{\partial^2}{\partial \psi^2}\left[(A - \tfrac{1}{2}u^2)^{\omega}\right], \qquad (118)$$

where A is constant and is given by

$$A = i + \tfrac{1}{2}u^2. \qquad (119)$$

For a prescribed value of ω, say $\omega = 0.89$, (118) can be replaced by a simple four point difference formula, and a numerical solution of u obtained. The enthalpy i is then easily calculated using (119). The stability condition for the procedure, although complicated, can be obtained using the methods previously outlined in this section.

In conclusion, sufficient has now been said to justify the author's contention that von Mises's form of the boundary layer equations is particularly suitable for use in finite difference methods. Final judgment on the techniques described, however,

must await the results of the numerical calculations.

REFERENCES

(1) Milne Thomson, L.M.. The Calculus of Finite Differences. Macmillan and Co. (1933).

Southwell, R.V.. Relaxation Methods in Engineering Science. Oxford University Press (1940).

Southwell, R.V. Relaxation Methods in Theoretical Physics. Clarendon Press, Oxford (1946).

Milne, W.E. Numerical Calculus. Princeton University Press. (1949).

Collatz, L. Numerische Behandlung von Differentialgleichungen. Springer, Berlin. (1951).

Hartree, D.R., Numerical Analysis. Clarendon Press, Oxford. (1952).

Hildebrand, F.B., Methods of Applied Mathematics. Prentice Hall, New York. (1952).

Householder, A.S. Principles of Numerical Analysis. McGraw Hill, New York. (1953).

Milne, W.E. Numerical Solution of Differential Equations. John Wiley and Sons, New York. (1953).

Shaw, F.S. An Introduction to Relaxation Methods. Dover Publications, New York. (1953).

Allen, D.N. de G. Relaxation Methods. McGraw Hill, London. (1954).

Booth, A.D. Numerical Methods. Butterworth's Scientific Publications, London. (1955).

Kopal, Z. Numerical Analysis. (1955).

(2) Bickley, W.G. "Formulae for Numerical Differentiation." The Mathematical Gazette. 25, p.19. (1941).

(3) Mitchell, A.R., and Craggs, J.W. "Stability of Difference Relations in the Solution of Ordinary Differential Equations." Math. Tables and other Aids to Computation., 7, p.127. (1953).

(4) Todd, J. "Solution of Differential Equations by Recurrence Relations." Math. Tables and other Aids to Computation. 4, p.59 (1950).

(5) Rutishauser, H. "Uber die Instabilitat von Methoden zur Integration gewohulicher Differentialgleuchungen." Zeit. angew. Math. und Phys., 3, p.65. (1952).

(6) Allen, D.N. de G., and Severn, R.T. "The Application of Relaxation Methods to the Solution of Non-elliptic Partial Differential Equations." Quart. Journ. Mech. and App. Math., 4, p. 209 (1951).

(7) Mitchell, A.R., and Rutherford, D.E. "On the Theory of Relaxation." Proc. Glasgow Math. Assoc., 1, p. 101 (1953).

(8) Fox, L. "A note on the Numerical Integration of First-order Differential Equations." Quart. Journ. Mech. and App. Math., 7, p. 367 (1954).

(9) Mitchell, A.R. "A Note on the Application of Relaxation Methods to the Numerical Solution of Unstable Initial Value Problems."

(10) Temple, G. "The General Theory of Relaxation Methods Applied to Linear Systems." Proc. Roy. Soc. London (A), 169; p. 476, (1939).

(11) Stiefel, E. "Uber einige Methoden der Relaxationsrechnung." Zeit. angew. Math. und Phys., 3, p.1, (1952).

(12) Fox, L. "A Short Account of Relaxation Methods," Quart. Journ. Mech. and App. Math., 1, p. 253. (1948).

(13) Mitchell, A.R. "Round-off Errors in Relaxational Solutions of Poisson's Equation." Appl. Sci. Research, B, 3, p. 456. (1954).

(14) Rutherford, D.E. "Some Continuant Determinants arising in Physics and Chemistry", Proc. Roy. Soc. (Edinburgh), 63, p. 232. (1952).

(15) Lotkin, M. "The Propagation of Error in Numerical Integrations." Proc. Amer. Math. Soc. 5, 869, (1954).

(16) Batschelet, E. "Uber die numerische Auflosung von Randwertproblemen bei elliptischen partiellen Differentialgluchungen." Zeit. angew. Math. und Phys., 3, p. 165, (1952).

(17) Karlquist, O. "Numerical Solution of Elliptic Difference Equations by Matrix Methods." Tellus, 4, p. 374 (1952).

(18) Cornock, A.F. "The Numerical Solution of Poisson's and the Bi-Harmonic Equations by Matrices." Proc. Camb. Phil. Soc., 50, p. 524. (1954).

(19) Thom, A. "The Arithmetic of Field Equations. The Aeronautical Quarterly, 4, p. 205, (1953).

(20) Hyman, M.A. "Non-Iterative Numerical Solutions of Boundary-Value Problems." Appl. Sci. Research, B, 2, p. 325, (1952).

(21) O'Brien, G.G., Hyman, M.A., Kaplan, S. "A Study of the Numerical Solution of Partial Differential Equations," Journ. of Math. and Phys., 29, p. 223, (1951).

(22) Rosenbloom, P.C. "The Difference Equation Method for Solving the Dirichelet Problem."

(23) Allen, N.D. de G., and Dennis, S.C.R. "The Application of Relaxation Methods to the Solution of Differential Equations in Three Dimensions," Quart. Journ. Mech. and App. Math. 4, p. 199 (1951).

(24) Mitchell, A.R. "Round-off Errors in Implicit Finite Difference Methods." Quart. Journ. Mech. and App. Math. (in the press). (1955).

(25) Mitchell, A.R. "Round-off Errors in the Solution of the Heat Conduction Equation by Relaxation Methods." Appl. Sci. Research, A, 4, p. 109 (1953).

(26) Du Fort, E.C., and Frankel, S.P. "Stability Conditions in the Numerical Treatment of Parabolic Differential Equations." Math. Tables and other Aids to Computation, 7, p. 135 (1953).

(27) Blanch, J. "On the Numerical Solution of Parabolic Partial Differential Equations." Journ. of Research of the Nat. Bureau of Standards, 50, p. 345. (1953).

(28) Courant, R., Friedrichs, K., and Lewy, H. "Uber die

Partiellen Differenzengleichungen der Mathematischen Physik." Math. Ann., 100, p. 32 (1928).

(29) Leutert, W. "On the Convergence of Unstable Approximate Solutions of the Heat Equation to the Exact Solution." Journ. of Math. and Phys., 30, p. 245. (1952).

(30) Hildebrand, F.B. "On the Convergence of Numerical Solutions of the Heat Equation." Journ. of Math. and Phys., 31, p. 35 (1952).

(31) John, F., "On Integration of Parabolic Equations by Difference Methods." Comm. on Pure and App. Maths., 5, p. 155. (1952).

(32) Crank, J., and Nicolson, P. "A Practical Method for Numerical Evaluation of Solutions of Partial Differential Equations of the Heat Conduction Type." Proc. Camb. Phil. Soc., 43, p. 50 (1947).

(33) Leutert, W., and O'Brien, G.G.; "On the Convergence of Approximate Solutions of the Wave Equation to the Exact Solution." Journ. of Math. and Phys., 30, p. 252. (1952).

(34) Courant, R., Isaacson, E., and Rees, M. "On the Solution of Nonlinear Hyperbolic Differential Equations by Finite Differences." Comm. on Pure and App. Math., 5, p. 243. (1952).

(35) Mitchell, A.R., and Rutherford, D.E. "Application of Relaxation Methods to Compressible Flow past a Double Wedge." Proc. Roy. Soc. (Edinburgh), A. 63, p. 139. (1951).

(36) Green, J.R., and Southwell, R.V. "High Speed Flow of Compressible Fluid through a Two-Dimensional Nozzle.", Phil. Trans., A, 239, p. 367. (1943).

(37) Fox, L., and Southwell, R.V. "On the Flow of Gas through a Nozzle with Velocities exceeding the Speed of Sound," Proc. Roy. Soc. (London), A, 183, p. 38 (1944).

(38) Mitchell, A.R. "Relaxation Methods in Compressible Flow." Ph.D. Thesis (1949).

(39) Mises, R. von., "Discussion on Transonic Flow." Comm. on Pure and Applied Math., 7, p. 145. (1945).

(40) Emmons, H.W. "The Numerical Solution of Compressible Fluid Flow Problems." N.A.C.A. Tech. Note 932 (1944).

(41) Thom, A., and Woods, L.C. "The Design of an Aerofoil Profile to give a Specified Velocity Distribution at a given Stream Mach Number." Aeronautical Research Council, Report 12, 922. (1949).

(42) Thom, A. "Treatment of the Stagnation Point in Arithmetical Methods." Reports and Memoranda 2807. (1954).

(43) Woods, L.C. "The Numerical Solution of Two-Dimensional Fluid Motion in the Neighbourhood of Stagnation Points and Sharp Corners." Aeronautical Research Council, Report 12,887. (1949).

(44) Woods, L.C. "A New Relaxational Treatment of the Compressible Two-Dimensional Flow about an Aerofoil with Circulation." Aeronautical Research Council, Report 13,054. (1950).

(45) Goldstein, S. and Lighthill, M.J. "A Note on the Hodograph Transformation for the Two-Dimensional Vortex Flow of an Incompressible Fluid," Quart. Journ. Mech. and App. Math., 3, p. 297. (1950).

(46) Mitchell, A.R. "Rotational Flow past Cylinders." Proc. Inter. Math. Congress. p. 363. (1954).

Mitchell, A.R., and Murray, J.D. "Two Dimensional Flow with Constant Shear past Cylinders with Various Cross Sections." Zeit. angew. Math. und Phys., (in the press). (1955).

(47) Vazsonyi, A. "On Rotational Gas Flows." Quart. of App. Math., 3, p. 29. (1945).

(48) Mitchell, A.R. "Application of Relaxation to the Rotational Field of Flow behind a Bow Shock Wave." Quart. Journ. Mech. and App. Math., 4, p. 371. (1951).

(49) Mitchell, A.R., and McCall, Francis, "The Rotational Field Behind a Bow Shock Wave in Axially Symmetric Flow using Relaxation Methods." Proc. Roy. Soc. (Edinburgh), A, 63, p. 371. (1952).

(50) Woods, L.C. "A New Relaxation Treatment of Flow with Axial Symmetry." Quart. Journ. Mech. and App. Math., 4, p. 358 (1951).

(51) Giese, J.H. "Approximate Methods for Computing Flow Fields." Comm. on Pure and App. Math., 7, p. 65 (1954).

(52) Carrier, G.F., and Linn, C.C. "On the Nature of the Boundary Layer near the Leading Edge of a Flat Plate." Quart. of App. Math., 6, p. 65 (1948).

(53) Dean, W.R. "Note on the Motion of Viscous Liquid past a Parabolic Cylinder." Proc. Camb. Phil. Soc., 50, p. 125. (1954).

(54) Thom, A. "The Flow past Circular Cylinders at Low Speeds." Proc. Roy. Soc. (London), A, 141, p. 651. (1933).

(55) Allen, D.N.de G., and Southwell, R.V., "Relaxation Methods applied to determine the Motion in two Dimensions of a Viscous Fluid past a Fixed Cylinder." Quart. Journ. Mech. and App. Math., 8, p. 129. (1955).

(56) Goldstein, S. "Modern Developments in Fluid Dynamics." Oxford Press. (1938).

(57) Kawaguti, M. "Numerical Solution of the Navier-Stokes Equations for the Flow around a Circular Cylinder at Reynolds Number 40." Journ. Phys. Soc. (Japan), 8, p. 747. (1953).

(58) Thom, A. "Arithmetical Solution of Equations of the Type ^{4} = Constant." Reports and Memoranda 1604. (1933).

(59) Thom, A. "The Flow at the Mouth of a Stanton Pitot." Aeronautical Research Council Report 15,228. (1952)

(60) Hartree, D.R. "A Solution of the Laminar Boundary Layer Equation for Retarded Flow." Reports and Memoranda 2426. (1939).

(61) Hartree, D.R. "The Solution of the Equations of the Laminar Boundary Layer for Schubauer's Observed Pressure Distribution for an Elliptic Cylinder." Reports and Memoranda 2427. (1939).

(62) Leigh, D.C.F. "The Laminar Boundary Layer Equation; a Method of Solution by means of an Automatic Computer." Proc. Camb. Phil. Soc. 51, p. 320. (1955).

(63) Schroder, K. "Verwendung der Differenzenrechnung zur Berechnung der laminaren Grenzschicht." Math. Nachr., 4, p. 439. (1951).

(64) Gortler, H. "Uber die Losungen nightlinearer partieller Differentialgleichungen vom Reibungsschichttypus, Zeit. angew Math. und Mech., 30, p. 265 (1950).

(65) Witting, H. "Verbesserung des Differenzenverfahrens von H. Gortler zur Berechnung laminarer Grenzschichten." Zeit. angew. Math. und Phys. 4, p. 378. (1953).

(66) Cope, W.F., and Hartree, D.R. "The Laminar Boundary Layer in Compressible Flow." Phil. Trans., A, 241, 1. (1948).

(67) Gadd, G.E. "The Numerical Integration of the Laminar Compressible Boundary Layer Equations, with Special Reference to the Position of Separation when the Wall is Cooled." Aeronautical Research Council Report 15, 101. (1952).