# A novel finite-difference approximation to the biharmonic operator

By G. J. Tee

The two-dimensional biharmonic operator is approximated by a finite-difference operator over a square  $(h \times h)$  net, which connects each node with 16 neighbouring nodes in such a manner that the resulting matrix has "Young's Property A," for simple boundary conditions. It is shown that the local truncation error, the convergence rate of S.O.R. for solving the finite-difference (or "net") equations, and the truncation error of the solution of the net equations are each of order  $O(h^2)$  as  $h \to 0$ . Comparisons are made with the conventional 13-node approximation to the biharmonic operator: in particular, numerical experiments indicate that the convergence rate of S.O.R. for solving net equations based on the conventional 13-node operator is considerably less than that for the novel 17-node operator.

#### **Contents**

1.	Introduc	ction							PAGE 177	3.8 Convergence rate of S.O.R		AGE 187
2.	Constru	ction of	molec	ules	• •				178	4. Truncation errors		187
	2.1 Ta	aylor se	ries ex	pansio	n				179		•	107
	2.2 Finite-difference approximation to $\nabla^4$								180	4.1 Effect of errors in representing boundary		
	2.3 D	iagonal	domir	nance					180			188
		amily of							181	4.2 Truncation error of the solution		188
		ptimiza					• • •		181	4.3 Errors of 13-node molecule	•	189
3.	Eigenval	lue anal	ysis						181	5. Numerical experiments		190
	3.1 Be	oundary	condi	tions					182	-		191
	3.2 C	omplete	set of	equati	ions				182			191
	3.3 E	xtended	rectan	ıgular i	net				183			
		igenvalu		~					183	5.3 Conclusions	•	192
	3.5 Ei 3.6 Ei	igenvalu igenvalu	es of o	compos circular	und-cii nt mat:	rculant гіх	matrix	• •	184 185	6. Acknowledgements		192
		laximun equatio		mınım 	um ei	genvalı 	ies of i		x 186	References		192

#### 1. Introduction

The conventional finite-difference operator which, on a square net  $(h \times h)$ , approximates to the biharmonic operator  $\nabla^4$ , is symbolized by the "molecule" of Fig. 1. This mnemonic scheme represents the following relation between  $\nabla^4 \phi$  at node (j, k) of a square net and the finite-difference approximation thereto (including the first term of the local truncation error)

$$\nabla^{4} \phi \Big|_{y=kh}^{x=jh} = \frac{1}{h^{4}} \{ \phi_{j-2,k} + \phi_{j,k-2} + \phi_{j,k+2} + \phi_{j+2,k} + 2(\phi_{j-1,k-1} + \phi_{j-1,k+1} + \phi_{j+1,k-1} + \phi_{j+1,k-1} + \phi_{j+1,k+1}) - 8(\phi_{j-1,k} + \phi_{j,k-1} + \phi_{j,k+1} + \phi_{j+1,k}) + 20\phi_{j,k} \} - \frac{h^{2}}{6} (\zeta_{0,6} + \zeta_{2,4}) + \dots$$
(1.1)

Here and subsequently, in the truncation terms we use the abbreviation

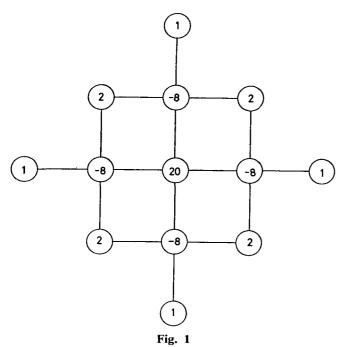
$$\zeta_{r,s} = \frac{\partial^{r+s}\phi}{\partial x^r \partial y^s} + \frac{\partial^{r+s}\phi}{\partial x^r \partial y^r}$$
 (1.2)

where the values of the derivatives are taken at the central node (j, k) (Collatz, 1960, p. 543).

The expression for the local truncation error is valid provided that all 6th derivatives of  $\phi$  are assumed to exist over the region covered by the molecule. If these 6th derivatives are to exist, then it is clearly necessary that all second derivatives of  $\nabla^4 \phi$  itself must exist. But if  $\nabla^4 \phi$  varies so irregularly that all 5th derivatives of  $\phi$  exist but not all 6th derivatives, then the local truncation error will be of order O(h) rather than  $O(h^2)$ .

Thus, for sufficiently smooth  $\phi$ , the 13-node molecule of Fig. 1 represents the biharmonic operator with a local truncation error of order  $O(h^2)$ .

Ignoring this local truncation error, we may equate



the algebraic expression on the right of (1.1) to the known value of  $\nabla^4 \phi$  at every internal node of a net covering a region R. If the boundary conditions can be approximated in a simple manner by assigning fixed values to  $\phi$  at each node of a double layer enclosing the boundary  $\Gamma$  of R (Fox, 1950), then it can be shown that the resulting system of linear algebraic equations in the (unknown) values of  $\phi$  at all the internal nodes has a positive-definite matrix (cf. §4.3). Hence the system of net equations may be solved by S.O.R. (i.e. Successive Over-Relaxation), and the process will always converge provided that the over-relaxation parameter  $\omega$ remains between 0 and 2 (Ostrowski, 1954). But it is exceedingly difficult to analyse the effect of S.O.R. in more detail, since the matrix of net equations does not possess "Young's Property A" (Young, 1954; Forsythe and Wasow, 1960, §22). On the other hand, if a matrix does have "Property A" then the effects of S.O.R. applied to the matrix (when it is "consistently ordered") can be analysed in considerable detail. (Young, 1954; Engeli, 1959; Forsythe and Wasow, 1960).

Therefore we shall construct (§2.2) a novel finite-difference approximation to the biharmonic operator, such that the system of linear algebraic equations resulting when this novel operator is applied at all internal nodes of a net covering  $R + \Gamma$  (with simple boundary conditions) has a matrix with "Property A." Indeed we shall construct (§2.4) a one-parameter family of such operators with local truncation error of order  $O(h^2)$ , and shall consider in detail a 17-node operator (from this family) which appears best-suited to iterative methods of solution (§2.5).

The matrix  $\vec{F}$  resulting from the application of this operator at all internal nodes of a net covering R is proved to be positive-definite (for simple boundary

conditions). Furthermore, by regarding F as a principal minor of a compound-circulant matrix, an upper bound is attained for its eigenvalues. From this and from the fact that F has "Property A," a positive lower bound is deduced for the eigenvalues of F (cf. §3.7).

When the system of net equations (whose matrix is F) is solved by S.O.R., an under-estimate of the convergence rate (for optimum  $\omega$ ) is deduced from the eigenvalue bounds for F (cf. §3.8): this convergence rate is shown to be of the order  $O(h^2)$  as  $h \to 0$ . The truncation error of the solution (i.e. the difference between the exact solution of the net equations and the exact solution of the differential equation at the nodes) is shown to be of the order  $O(h^2)$  for the simplest boundary conditions: other simple boundary conditions give a truncation error of order O(h) (cf. §4.2). The truncation errors of the solution are also examined (in §4.3) for the 13-node molecule.

Finally, (in §5), we describe the results of numerical experiments in which the 17-node molecule is applied to test cases and the empirical convergence rates of S.O.R. are compared with the theoretical underestimates. Comparisons are also made with the convergence rates of another molecule of the family (cf. §2.2), and with the convergence rate of S.O.R. for solving the net equations based on the 13-node molecule (cf. §5.2).

### 2. Construction of molecules

Consider a square  $(h \times h)$  net, with the nodes segregated into two classes (X and O) in "chessboard" fashion:



Fig. 2

If each molecule associated with a node of one class (e.g. X) connects that node only to nodes of the opposite class (i.e. O), then the matrix of the equations will have "Property A" (Engeli, 1959, p. 86, Fig. 2b). Furthermore, if the equations are so ordered that in each cycle of S.O.R. all nodes of one class are adjusted before all nodes of the other class, then this is a "consistent ordering" known as the  $\sigma_1$ -ordering (Young, 1954, p. 108). For any consistent ordering, the eigenvalues of the S.O.R. operator will be known functions of  $\omega$  and of the eigenvalues of the matrix (Engeli, 1959, p. 87).

Accordingly, we shall construct linear algebraic expressions containing the values of a sufficiently smooth function  $\phi$  at a node of one class (say X) and at nearby nodes of the other class, and we shall choose the coefficients in such a manner that the algebraic expression approximates to the value of  $\nabla^4 \phi$  at the associated X-node, with a local truncation error of order  $O(h^2)$ .

Take the origin of coordinates at a typical X-node, and

label typical 0-nodes at distances h,  $h\sqrt{5}$ , 3h and  $h\sqrt{13}$  by  $\alpha$ ,  $\beta$ ,  $\gamma$  and  $\delta$  respectively.

#### 2.1 Taylor series expansion

We shall apply the two-dimensional form of Taylor series, given formally by

$$f(x, y) = e^{x\frac{\delta}{\delta x} + y\frac{\delta}{\delta y}} f(0, 0). \tag{2.1}$$

Denote the sum of values of  $\phi$  at the four  $\alpha$ -nodes by A, the sum of values of  $\phi$  at the eight  $\beta$ -nodes by B, the sum of values of  $\phi$  at the four  $\gamma$ -nodes by C, and the sum of values of  $\phi$  at the eight  $\delta$ -nodes by D.

We shall express each of these sums in terms of  $\phi$  and its derivatives at the central node. Unless otherwise stated,  $\phi$  and its derivatives will be evaluated at the associated X-node (i.e. the origin in Fig. 3). Then the Taylor series expansion (2.1) shows that for the  $\alpha$ -nodes

$$A = \left(e^{h\frac{\delta}{\delta x}} + e^{-h\frac{\delta}{\delta x}} + e^{h\frac{\delta}{\delta y}} + e^{-h\frac{\delta}{\delta y}}\right) \phi$$

$$= 2\left(1 + \frac{h^2}{2!} \cdot \frac{\partial^2}{\partial x^2} + \frac{h^4}{4!} \cdot \frac{\partial^4}{\partial x^4} + \frac{h^6}{6!} \cdot \frac{\partial^6}{\partial x^6} + \dots + 1 + \frac{h^2}{2!} \cdot \frac{\partial^2}{\partial y^2} + \frac{h^4}{4!} \cdot \frac{\partial^4}{\partial y^4} + \frac{h^6}{6!} \cdot \frac{\partial^6}{\partial y^6} + \dots\right) \phi$$

$$= 4\phi + h^2 \nabla^2 \phi + \frac{h^4}{12} \left(\frac{\partial^4}{\partial x^4} + \frac{\partial^4}{\partial y^4}\right) \phi$$

$$+ \frac{h^6}{360} \zeta_{0,6} + \dots$$
 (2.2)

where we have used the 6th-order term of the Taylor series expansion to express the local truncation error, on the assumption that the 6th derivatives of  $\phi$  do exist throughout the region being considered.

Next, for the  $\beta$ -nodes,

$$B = \left(e^{2h\frac{\partial}{\partial x} + h\frac{\partial}{\partial y}} + e^{2h\frac{\partial}{\partial x} - h\frac{\partial}{\partial y}} + e^{-2h\frac{\partial}{\partial x} + h\frac{\partial}{\partial y}}\right)$$

$$+ e^{-2h\frac{\partial}{\partial x} - h\frac{\partial}{\partial y}} + e^{h\frac{\partial}{\partial x} + 2h\frac{\partial}{\partial y}} + e^{h\frac{\partial}{\partial x} - 2h\frac{\partial}{\partial y}}$$

$$+ e^{-h\frac{\partial}{\partial x} + 2h\frac{\partial}{\partial y}} + e^{-h\frac{\partial}{\partial x} - 2h\frac{\partial}{\partial y}}\right)\phi$$

$$= 2\left\{1 + \frac{\left(2h\frac{\partial}{\partial x} + h\frac{\partial}{\partial y}\right)^2}{2!} + \frac{\left(2h\frac{\partial}{\partial x} + h\frac{\partial}{\partial y}\right)^4}{4!}\right\}$$

$$+\frac{\left(2h\frac{\partial}{\partial x}+h\frac{\partial}{\partial y}\right)^{6}}{6!}+\dots$$

$$+1+\frac{\left(2h\frac{\partial}{\partial x}-h\frac{\partial}{\partial y}\right)^{2}}{2!}+\frac{\left(2h\frac{\partial}{\partial x}-h\frac{\partial}{\partial y}\right)^{4}}{4!}$$

$$+\frac{\left(2h\frac{\partial}{\partial x}-h\frac{\partial}{\partial y}\right)^{6}}{6!}+\dots$$

$$+1+\frac{\left(h\frac{\partial}{\partial x}+2h\frac{\partial}{\partial y}\right)^{2}}{2!}+\frac{\left(h\frac{\partial}{\partial x}+2h\frac{\partial}{\partial y}\right)^{4}}{4!}$$

$$+\frac{\left(h\frac{\partial}{\partial x}+2h\frac{\partial}{\partial y}\right)^{6}}{6!}+\dots$$

$$+1+\frac{\left(h\frac{\partial}{\partial x}-2h\frac{\partial}{\partial y}\right)^{2}}{2!}+\frac{\left(h\frac{\partial}{\partial x}-2h\frac{\partial}{\partial y}\right)^{4}}{4!}$$

$$+\frac{\left(h\frac{\partial}{\partial x}-2h\frac{\partial}{\partial y}\right)^{6}}{6!}+\dots\right\}\phi$$

$$=8\phi+10h^{2}\nabla^{2}\phi+\frac{17h^{4}}{6}\left(\frac{\partial^{4}}{\partial x^{4}}+\frac{\partial^{4}}{\partial y^{4}}\right)\phi$$

$$+8h^{4}\frac{\partial^{4}\phi}{\partial x^{2}\partial y^{2}}+\frac{h^{6}}{360}(65\zeta_{0,6}+300\zeta_{2,4})+\dots$$
(2.3)

Then for the  $\gamma$ -nodes, by analogy with (2.2) we get

$$C = \left(e^{3h\frac{\delta}{\delta x}} + e^{-3h\frac{\delta}{\delta x}} + e^{3h\frac{\delta}{\delta y}} + e^{-3h\frac{\delta}{\delta y}}\right)$$

$$= 4\phi + 9h^2\nabla^2\phi + \frac{81}{12}h^4\left(\frac{\delta^4}{\delta x^4} + \frac{\delta^4}{\delta y^4}\right)\phi$$

$$+ \frac{729}{260}h^6\zeta_{0,6} + \dots \quad (2.4)$$

Finally, for the  $\delta$ -nodes we get

$$D = \left(e^{3h\frac{\partial}{\partial x} + 2h\frac{\partial}{\partial y}} + e^{3h\frac{\partial}{\partial x} - 2h\frac{\partial}{\partial y}} + e^{-3h\frac{\partial}{\partial x} + 2h\frac{\partial}{\partial y}}\right)$$

$$+ e^{-3h\frac{\partial}{\partial x} - 2h\frac{\partial}{\partial y}} + e^{2h\frac{\partial}{\partial x} + 3h\frac{\partial}{\partial y}} + e^{2h\frac{\partial}{\partial x} - 3h\frac{\partial}{\partial y}}$$

$$+ e^{-2h\frac{\partial}{\partial x} + 3h\frac{\partial}{\partial y}} + e^{-2h\frac{\partial}{\partial x} - 3h\frac{\partial}{\partial y}}\right)\phi$$

$$= 2\left\{1 + \frac{\left(3h\frac{\partial}{\partial x} + 2h\frac{\partial}{\partial y}\right)^2}{2!} + \frac{\left(3h\frac{\partial}{\partial x} + 2h\frac{\partial}{\partial y}\right)^4}{4!} + \frac{\left(3h\frac{\partial}{\partial x} + 2h\frac{\partial}{\partial y}\right)^6}{6!} + \dots\right\}$$

$$+1 + \frac{\left(3h\frac{\delta}{\delta x} - 2h\frac{\delta}{\delta y}\right)^{2}}{2!} + \frac{\left(3h\frac{\delta}{\delta x} - 2h\frac{\delta}{\delta y}\right)^{4}}{4!} + \frac{\left(3h\frac{\delta}{\delta x} - 2h\frac{\delta}{\delta y}\right)^{6}}{6!} + \dots$$

$$+1 + \frac{\left(2h\frac{\delta}{\delta x} + 3h\frac{\delta}{\delta y}\right)^{2}}{2!} + \frac{\left(2h\frac{\delta}{\delta x} + 3h\frac{\delta}{\delta y}\right)^{4}}{4!} + \frac{\left(2h\frac{\delta}{\delta x} + 3h\frac{\delta}{\delta y}\right)^{6}}{6!} + \dots$$

$$+1 + \frac{\left(2h\frac{\delta}{\delta x} - 3h\frac{\delta}{\delta y}\right)^{2}}{2!} + \frac{\left(2h\frac{\delta}{\delta x} - 3h\frac{\delta}{\delta y}\right)^{6}}{4!} + \dots$$

$$+\frac{\left(2h\frac{\delta}{\delta x} - 3h\frac{\delta}{\delta y}\right)^{2}}{2!} + \frac{\left(2h\frac{\delta}{\delta x} - 3h\frac{\delta}{\delta y}\right)^{6}}{4!} + \dots\right\}\phi$$

$$= 8\phi + 26h^{2}\nabla^{2}\phi + \frac{97}{6}h^{4}\left(\frac{\delta^{4}}{\delta x^{4}} + \frac{\delta^{4}}{\delta y^{4}}\right)\phi + 72h^{4}\frac{\delta^{4}\phi}{\delta x^{2}\delta y^{2}} + \frac{h^{6}}{360}(793\zeta_{0,6} + 4,860\zeta_{2,4}) + \dots$$
 (2.5)

# 2.2 Finite-difference approximation to $\nabla^4$

The three equations (2.2), (2.3) and (2.4) may be solved to give the quantities

$$\nabla^2 \phi$$
,  $\left(\frac{\partial^4}{\partial x^4} + \frac{\partial^4}{\partial y^4}\right) \phi$  and  $\frac{\partial^4 \phi}{\partial x^2 \partial y^2}$ 

in terms of A, B, C and truncation terms. It may readily be verified that

$$\nabla^{2}\phi = \frac{81A - 320\phi - C}{72h^{2}} + \frac{h^{4}}{40}\zeta_{0,6} + \dots (2.6)$$

$$\frac{\partial^{4}\phi}{\partial x^{4}} + \frac{\partial^{4}\phi}{\partial y^{4}} = \frac{32\phi - 9A + C}{6h^{4}} - \frac{h^{2}}{3}\zeta_{0,6} + \dots (2.7)$$

$$\frac{\partial^{4}\phi}{\partial x^{2}\partial y^{2}} = \frac{64\phi - 21A + 3B - C}{24h^{4}}$$

$$+ \frac{h^{2}}{576}(37\zeta_{0,6} - 60\zeta_{2,4}) + \dots (2.8)$$

Combining (2.7) and (2.8), we get

$$\nabla^{4}\phi = \left(\frac{\partial^{4}}{\partial x^{4}} + 2\frac{\partial^{4}}{\partial x^{2}\partial y^{2}} + \frac{\partial^{4}}{\partial y^{4}}\right)\phi$$

$$= \frac{128\phi - 39A + 3B + C}{12h^{4}}$$

$$-\frac{h^{2}}{288}(59\zeta_{0,6} + 60\zeta_{2,4}) + \dots (2.9)$$

This equation may be symbolized by the molecule:

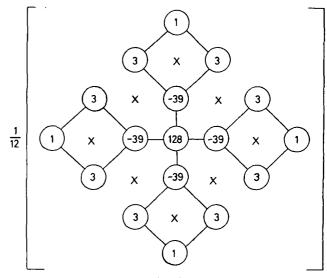


Fig. 4

However, this molecule is not particularly convenient for the application of iterative methods of solution to a system of such equations, inasmuch as the equations (2.9) do not give "diagonal dominance" (q.v.).

### 2.3 Diagonal dominance

The matrix of a set of equations is said to possess the property of "diagonal dominance" if in each row the diagonal term is not less than the sum of the moduli of the non-diagonal terms, and in at least one row we can replace "not less than" by "strictly greater than." (Fox, 1962, p. 288). It can be shown that the Simultaneous and the Successive Displacement iterative methods will each converge when applied to such a matrix. Roughly speaking, the greater the degree of diagonal dominance in a matrix, the more rapidly will any iterative process converge.

Any algebraic approximation to a differential expression (such as (1.1) or (2.9)) must have the sum of its coefficients equal to zero; for when the finite-difference operator is applied to a function  $\phi$  which is constant over the region involved (say,  $\phi \equiv 1$ ), then all derivatives of  $\phi$  are zero so that the truncation term will be strictly zero, and hence the sum of the coefficients must be exactly zero. Therefore the matrix of a system of finite-difference approximations to differential expressions can have diagonal dominance (for suitable boundary conditions), if and only if in each row of the matrix every non-zero non-diagonal coefficient has sign opposite to that of the diagonal coefficient. Thus, neither the conventional molecule (1.1) nor the novel molecule can give diagonal dominance, and the convergence of the Simultaneous and Successive Displacement methods cannot be guaranteed a priori. Indeed, it can be proved (Heller, 1960) that the Simultaneous method diverges when it is applied to the conventional 13-node molecule of (1.1).

Therefore we shall construct a molecule approximating to  $\nabla^4 \phi$  which produces a matrix which is more nearly diagonally dominant than that based on (2.9), by introducing also the nodes  $\delta$  at distance  $h\sqrt{13}$  from the origin. When S.O.R. is applied to the set of equations based on such a molecule, it is expected to converge more rapidly than when a molecule (e.g. (2.9)), which is less nearly diagonally dominant is used. Of course, the spectral radius of the matrix is a precise criterion of the convergence rate of S.O.R.; but this is more difficult to estimate.

#### 2.4 Family of biharmonic molecules

Inserting into (2.5) the expressions (2.6), (2.7) and (2.8), we get that

$$512\phi - 174A + 27B - 2C - 3D + h^{6} \left(\frac{94}{15}\zeta_{0.6} + 18\zeta_{2.4}\right) + \dots = 0. \quad (2.10)$$

Multiplying equation (2.10) by  $\frac{\epsilon}{12h^4}$  (where  $\epsilon$  is any number) and then subtracting it from the right-hand side of (2.9), we get

Table 1
Measure of Diagonal Dominance

E	<i>y</i> 1	4y2	8 <i>y</i> <sub>3</sub>	4 <i>y</i> <sub>4</sub>	8 <i>y</i> <sub>5</sub>	$\begin{array}{c} 4y_2 + 8y_3 + \\ 4y_4 + 8y_5 \end{array}$	$r(\epsilon)$
-1 -1 0 1 0 1 1 5 1 3 5 8 4	640 384 128 71 ½ 13 70 0	852 504 156 78 <del>3</del> 0 18 192	240 132 24 0 24 <sup>12</sup> / <sub>29</sub> 30 84	4 0 4 4 8 5 <sup>23</sup> / <sub>29</sub> 6 8	$ \begin{array}{c} 24 \\ 12 \\ 0 \\ 2\frac{2}{3} \\ 5^{\frac{1}{2}\frac{1}{9}} \\ 6 \\ 12 \end{array} $	1,120 648 184 86 <sup>2</sup> / <sub>9</sub> 35 <sup>1/2</sup> / <sub>0</sub> 60 296	0·571 0·592 0·696 0·825 0·372 0 0·433

Table 1 gives the values of  $r(\epsilon)$  for suitably chosen values of  $\epsilon$ , including those values for which

$$y_i = 0 (i = 1, 2, 3, 4, 5).$$

Each of the  $y_i(i=1,\ldots,5)$  is a linear function of  $\epsilon$  over any range of  $\epsilon$  for which that  $y_i$  remains non-zero. Therefore both the numerator and the denominator of (2.13) are linear functions of  $\epsilon$  over each such range of  $\epsilon$ .

$$\nabla^{4}\phi = \frac{(128 - 512\epsilon)\phi + (-39 + 174\epsilon)A + (3 - 27\epsilon)B + (1 + 2\epsilon)C + 3\epsilon D}{12h^{4}} - h^{2}\left\{\left(\frac{59}{288} + \frac{94\epsilon}{12 \times 15}\right)\zeta_{0,6} + \left(\frac{60}{288} + \frac{18\epsilon}{12}\right)\zeta_{2,4}\right\} + \dots$$
 (2.11)

This is the required expression for a one-parameter family of molecules, each approximating to  $\nabla^4 \phi$  with a local truncation error of order  $O(h^2)$ .

## 2.5 Optimization of molecule

As was indicated in §2.3, it is advisable to fix the parameter  $\epsilon$  at a value which will give the nearest approach to diagonal dominance in the equations constructed from the equation (2.11), in order to improve the likelihood that any iterative process will converge rapidly.

Define:

$$y_{1} = | 128 - 512\epsilon|,$$

$$y_{2} = |-39 + 174\epsilon|,$$

$$y_{3} = | 3 - 27\epsilon|,$$

$$y_{4} = | 1 + 2\epsilon|,$$

$$y_{5} = |3\epsilon|.$$
(2.12)

Then for equations based on (2.11) (ignoring the truncation terms), the ratio of the modulus of the diagonal coefficient to the sum of the moduli of the non-diagonal coefficients (which may be taken as a measure of diagonal dominance) is given by

$$r(\epsilon) = \frac{y_1}{4y_2 + 8y_3 + 4y_4 + 8y_5}.$$
 (2.13)

Hence  $r(\epsilon)$  varies monotonically within each such range of  $\epsilon$ , i.e. over the ranges

$$\epsilon = (-\infty, -\frac{1}{2}), (-\frac{1}{2}, 0), (0, \frac{1}{9}), (\frac{1}{9}, \frac{13}{58}), (\frac{13}{58}, \frac{1}{4}), (\frac{1}{4}, +\infty).$$

Accordingly, Table 1 shows that the measure  $r(\epsilon)$  of diagonal dominance assumes its maximum value when  $\epsilon = \frac{1}{9}$ . Hence we expect iterative methods of solution to converge the most rapidly (in general) when equation (2.11) is taken with  $\epsilon = \frac{1}{9}$ , when it assumes the form

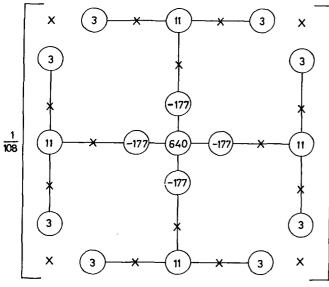
$$\nabla^4 \phi = \frac{640\phi - 177A + 11C + 3D}{108h^4} - h^2 \left(\frac{3,407}{12,960} \zeta_{0.6} + \frac{3}{8} \zeta_{0,4}\right) + \dots$$
 (2.14)

This may be represented by the 17-node molecule of Fig. 5.

Note that the  $\beta$ -nodes are absent from the 17-node molecule.

#### 3. Eigenvalue analysis

Molecules of the general type (2.11) may be applied at any internal node of the net over a region which is surrounded on each side by three or more rows and columns, so that each of the  $\alpha$ ,  $\beta$ ,  $\gamma$  and  $\delta$  nodes is actually included in the net. But some other types of



equations must be applied at nodes near the boundary  $\Gamma$  of the region R—these other equations must express the boundary conditions.

Fig. 5

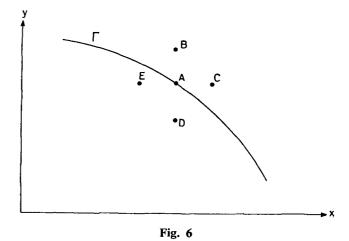
We shall examine the complete set of equations resulting from the application of the biharmonic molecule at each internal node, when the boundary conditions can be approximated by fixing  $\phi$  at a triple layer of nodes enclosing  $\Gamma$ . These equations may be solved by the S.O.R. process, but in order to estimate the convergence rate we shall need an estimate of the maximum eigenvalue of the matrix F of the complete set of equations (cf. §3.8). Further, a lower bound for the minimum eigenvalue of F is needed for estimating the accuracy with which the solution of the complete set of equations approximates to the solution of the differential biharmonic problem (cf. §4.2).

Bounds for the eigenvalues can be found with the aid of the theorem that the eigenvalues of a principal minor of a symmetric matrix lie within the range of eigenvalues of the matrix itself (cf. §3.7). If R is not a rectangle, we shall first add rows and columns to F so as to form a compound band matrix G of regular structure, containing F as a principal minor (cf. §3.3). The eigenvalues of G are shown to be those of a family of 7-band matrices (cf. §3.5), which can themselves be regarded as principal minors of a family of circulant matrices, whose eigenvalues are known (cf. §3.6). The maximum of all eigenvalues of this family cannot be less than the maximum eigenvalue of F, and hence we can find a lower bound for the convergence rate of S.O.R. (cf. §3.8).

## 3.1 Boundary conditions

We shall confine our attention to the simplest boundary condition, in which  $\phi$  and its normal derivative  $\frac{\partial \phi}{\partial \nu}$  are specified everywhere on  $\Gamma$ . This boundary condition may be approximated by fixing the values of  $\phi$  at each

node within a layer enclosing  $\Gamma$ . For example, if a node (j, k) lies on  $\Gamma$ , then whatever may be the direction of  $\Gamma$  at that node we can compute  $\frac{\partial \phi}{\partial x}$  and  $\frac{\partial \phi}{\partial y}$  there from the known values of  $\phi$  along the boundary and of  $\frac{\partial \phi}{\partial v}$  at the node itself.



Then, neglecting errors in  $\phi$  of order  $O(h^2)$ , we have

$$\phi_{B} = \phi_{A} + h \frac{\partial \phi}{\partial y} \Big|_{A},$$

$$\phi_{C} = \phi_{A} + h \frac{\partial \phi}{\partial x} \Big|_{A},$$

$$\phi_{D} = \phi_{A} - h \frac{\partial \phi}{\partial y} \Big|_{A},$$

$$\phi_{E} = \phi_{A} - h \frac{\partial \phi}{\partial x} \Big|_{A}.$$

$$(3.1)$$

Similarly, even if  $\Gamma$  passes between adjacent nodes, we may fix  $\phi$  at nodes close to  $\Gamma$ , with errors again of order  $O(h^2)$ . Roughly speaking, a triple layer of nodes enclosing  $\Gamma$  may have values of  $\phi$  fixed at them, so that the biharmonic molecule (2.11) can then be applied at all nodes inside this layer. We shall call these the "boundary nodes" and the "internal nodes," respectively.

# 3.2 Complete set of equations

Let  $\Phi$  be a vector of the values of  $\phi$  at the internal nodes, numbered in any order. Then if equations of the form (2.11) (neglecting the truncation terms) are written for each node (j, k) in the same order as for  $\Phi$ , and if all the values of  $\phi$  at boundary nodes are transferred to the right-hand sides of the equations, we shall have a set of equations (say, n in number) which may be written as

$$F\Phi = b. (3.2)$$

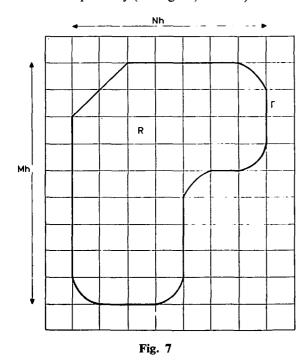
Here, F is an  $n \times n$  symmetric matrix (in view of the symmetry of the molecule based on (2.11)), whilst b is a function only of the values of  $\phi$  at the boundary nodes

and the values of  $\nabla^4 \phi$  at internal nodes, each of which is postulated to be known. Thus F has real eigenvalues, and of course it has "Property A" (cf. §2).

The eigenvalues of F are the same for all permutations  $\Pi F \Pi^T$  of F (where  $\Pi$  is a permutation matrix), and hence for any ordering  $\Pi \Phi$  of the values of  $\phi$  at internal nodes. Therefore, as far as the eigenvalues of F are concerned, without loss of generality we may assume that  $\Phi$  numbers the internal nodes in the so-called "pagewise" ordering, i.e. consecutively across each row, with the rows taken consecutively.

# 3.3 Extended rectangular net

If the region R is not a rectangle, extend the net to become a rectangular net which includes all the internal and boundary nodes of  $R + \Gamma$ . Let the extended net have (M+3) rows and (N+3) columns, so that there are (M-3) rows and (N-3) columns containing internal nodes of the net. Thus if  $\Gamma$  passes through nodes and does not cut net lines between nodes, the maximum depth and width of the region are exactly Mh and Nh respectively (cf. Figs. 5, 6 and 7).



Consider the matrix G of order  $(M-3)(N-3) \times (M-3)(N-3)$ , which would result from applying the equation (2.11) at each internal node of the extended net, transferring to the right-hand side each term referring to a node distant not more than 2h from the outer boundary of the extended net, with the nodes

being taken in pagewise order. It is readily seen that the matrix F could be constructed by selecting from G those rows and columns which correspond to the internal nodes of R, so that F is a principal minor of G. We shall compute upper and lower bounds for the eigen-

values of G and deduce therefrom bounds for the eigenvalues of F, from which we shall obtain a lower bound for the optimized convergence rate of S.O.R. for solving the system of equations.

# 3.4 Eigenvalue analysis of extended matrix

From now on, we shall confine our attention to equations based on the optimized 17-molecule of Fig. 5.

By inspection of Figs. 5 and 7, it is seen that the extended matrix G has the following form:

$$G = \frac{1}{108h^4} \begin{bmatrix} a & b & c & d \\ b & a & b & c & d \\ c & b & a & b & c & d \\ d & c & b & a & b & c & d \\ d & c & b & a & b & c & d \\ & & & & & & & \\ & & & & & & \\ & & & & & & & \\ & & & & & \\ & & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & \\ & & & & & \\ & & & & \\ & & & & \\ & & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & \\ & & & & \\ & & & & \\ & & & \\ & & & & \\ & & & \\ & & & \\ & & & & \\ & & \\ & &$$

The compound matrix  $108h^4G$  is symmetric, and is composed of (M-3) rows and columns of  $(N-3)\times(N-3)$  submatrices. These submatrices are each symmetric band matrices, with structures given by

Each of a, b, c and d has half-bandwidth not greater than 3.

The matrix  $108h^4G$  is itself a principal minor of the following matrix:

$$H = \begin{bmatrix} a & b & c & d & & & d & c & b \\ b & . & & & & d & c & d \\ c & . & & & & d & d \\ & . & & . & & & d & d \\ & . & & . & & & & d & d \\ & & & & & & & & d & c & d \\ d & & & & & & & & d & c & b & a & b \\ b & c & d & & & & & & & & c & b & a & b \\ b & c & d & & & & & & & & & c & b & a \end{bmatrix}$$
(3.5)

H is a symmetric matrix with M compound rows and columns, and its structure may be called "compound-circulant."

# 3.5 Eigenvalues of compound-circulant matrix

By analogy with the eigenvalue analysis of circulant matrices (Marcus, 1960, p. 9), we shall require a compound vector w to be an eigenvector of H, where w is partitioned into M subvectors each with (N-3) elements, and  $\rho$  is a scalar.

$$w = \begin{bmatrix} v \\ \rho v \\ \rho^2 v \\ \vdots \\ \rho^{M-1} v \end{bmatrix}$$

$$(3.6)$$

If  $\lambda$  is the eigenvalue of H corresponding to w,

$$Hw = \lambda w. (3.7)$$

Then substitution of (3.5) and (3.6) into (3.7) shows that

$$[a + (\rho + \rho^{M-1})b + (\rho^{2} + \rho^{M-2})c + (\rho^{3} + \rho^{M-3})d]v = \lambda v$$

$$[a + (\rho + \rho^{-1})b + (\rho^{2} + \rho^{M-2})c + (\rho^{3} + \rho^{M-3})d]\rho v = \lambda \rho v$$

$$[a + (\rho + \rho^{-1})b + (\rho^{2} + \rho^{-2})c + (\rho^{3} + \rho^{M-3})d]\rho^{2}v = \lambda \rho^{2}v$$

$$[a + (\rho + \rho^{-1})b + (\rho^{2} + \rho^{-2})c + (\rho^{3} + \rho^{-3})d]\rho^{3}v = \lambda \rho^{3}v, \text{ etc.} (3.8)$$

All of the equations (3.8) will be satisfied, provided that

$$\rho^M = 1 \tag{3.9}$$

i.e.

$$\rho_r = e^{ir\theta}, \, \theta = \frac{2\pi}{M} \qquad (r = 0, 1, 2, ..., M - 1) \quad (3.10)$$

and provided that  $\lambda$  and v satisfy an equation of the form

$$[a + (\rho_r + \rho_r^{-1})b + (\rho_r^2 + \rho_r^{-2})c + (\rho_r^3 + \rho_r^{-3})d]v = \lambda v. \quad (3.11)$$

Substituting (3.10) into (3.11), and labelling  $\lambda$  and v with double suffixes, we get

$$[a + 2\cos r\theta.b + 2\cos 2r\theta.c + 2\cos 3r\theta.d]v_{r,s} = \lambda_{r,s}v_{r,s}$$
 (3.12)

$$(r = 0, 1, 2, ..., M - 1; s = 1, 2, ..., N - 3).$$

Thus,  $\lambda_{r,s}$  is the sth eigenvalue of the rth  $(N-3)\times(N-3)$  matrix  $J_r$ , where

$$J_r = a + 2\cos r\theta . b + 2\cos 2r\theta . c + 2\cos 3r\theta . d (3.13)$$

and  $v_{r,s}$  is the corresponding eigenvector. Each  $J_r$  is symmetric, in view of (3.4) and (3.13), so that all  $\lambda_{r,s}$  and  $v_{r,s}$  are real.

It has been shown above that each such  $\lambda_{r,s}$  is an eigenvalue of H: we shall now prove that all the eigenvalues of H are given by the  $\lambda_{r,s}$ , even if some of the  $\lambda_{r,s}$  happen to be multiple eigenvalues of H. Take the scalar product of two eigenvectors of the form w. Using  $w^*$  to denote the conjugate complex transpose of w, we have

$$w_{r,s}^* w_{t,u} = \begin{bmatrix} v_{r,s}^T \, \overline{
ho}_r v_{r,s}^T \, \overline{
ho}_r^2 v_{r,s}^T \dots \, \overline{
ho}_r^{M-1} v_{r,s}^T \end{bmatrix} \begin{bmatrix} v_{t,u} \\ 
ho_t \, v_{t,u} \\ 
ho_t^2 \, v_{t,u} \\ 
ho_t \\ 
ho_t \\ 
ho_t \\ 
ho_t \\ 
ho_t \end{bmatrix}$$

$$= \left\{1 + e^{\frac{2\pi i(t-r)}{M}} + e^{\frac{2\pi i(t-r)\times 2}{M}} + \dots + e^{\frac{2\pi i(t-r)(M-1)}{M}}\right\} v_{r,s}^{T} v_{t,u}$$

$$= \frac{1 - e^{2\pi i(t-r)}}{\frac{2\pi i(t-r)}{M}} v_{r,s}^{T} v_{t,u}. \tag{3.14}$$

The numerator of the coefficient of  $v_{r,s}^T v_{t,u}$  is always zero. Accordingly, the scalar product of  $w_{r,s}$  and  $w_{t,u}$  is certainly zero unless the denominator is also zero. But r and t are each confined to the range 0 to M-1, so that the denominator will be zero if and only if r=t. In that case

$$w_{r,s}^* w_{r,u} = (1+1+\ldots+1)v_{r,s}^T v_{r,u} = Mv_{r,s}^T v_{r,u}.$$
 (3.15)

But since the matrices  $J_r$  are symmetric, the (N-3) eigenvectors of  $J_r$  can be taken as being mutually orthogonal, whether or not any eigenvalues are multiple. Indeed the eigenvectors may be taken as orthonormal:

$$v_{r,s}^T v_{r,u} = \delta_{s,u} (3.16)$$

where  $\delta_{s,\mu}$  is the Kronecker delta, defined by

$$\begin{cases} \delta_{s, u} = 1 \text{ if } s = u \\ \delta_{s, u} = 0 \text{ if } s \neq u \end{cases}$$
(3.17)

(Faddeeva, 1959, p. 45).

Thus we have shown that

$$w_{r,s}^* w_{t,u} = M \delta_{r,t} \delta_{s,u} \tag{3.18}$$

i.e. each pair of vectors of the form (3.6), with  $\lambda$  and v satisfying (3.12), are mutually orthogonal, unless they are identically labelled. Hence the M(N-3) vectors  $w_{r,s}$  are mutually orthogonal, i.e. they form a complete set of eigenvectors of H, whether or not any eigenvalues happen to be multiple.

The argument from (3.6) to (3.18) is obviously applicable to a general symmetric compound-circulant matrix.

## 3.6 Eigenvalues of circulant matrix

The  $(N-3)\times(N-3)$  matrix  $J_r$  is a 7-band matrix (cf. (3.13) and (3.4)), whose first row may be written as  $(x_0, x_1, x_2, x_3, 0, \ldots, 0)$ ; where

$$x_0 = 640 - 354 \cos r\theta + 22 \cos 3r\theta,$$
  
 $x_1 = -177,$   
 $x_2 = 6 \cos 3r\theta,$   
 $x_3 = 11 + 6 \cos 2r\theta.$  (3.19)

It is readily seen that  $J_r$  is itself a principal minor of a symmetric  $N \times N$  circulant matrix K, whose first row contains the elements

$$x_0, x_1, x_2, x_3, 0, \ldots, 0, x_3, x_2, x_1.$$
 (3.20)

The eigenvalues of this circulant matrix are given (Marcus, 1960, p. 9) by

$$\mu_{r,s} = x_0 + 2x_1 \cos s\psi + 2x_2 \cos 2s\psi + 2x_3 \cos 3s\psi$$
(3.21)

$$\psi = \frac{2\pi}{N}, s = 0, 1, \dots, N - 1.$$
 (3.22)

Substituting (3.19) into (3.22), we get

$$\mu_{r,s} = 640 - 354 (\cos r\theta + \cos s\psi) + 22 (\cos 3r\theta + \cos 3s\psi) + 12 (\cos 2r\theta \cdot \cos 3s\psi + \cos 2s\psi \cdot \cos 3r\theta). (3.23)$$

Let

$$y_r = \cos r\theta,$$
  

$$z_r = \cos s\psi.$$
 (3.24)

Substituting (3.24) into (3.23), we get:

$$\mu_{r,s} = 640 - 354(y_r + z_s) + 22(4y_r^3 - 3y_r + 4z_s^3 - 3z_s) + 12\{(2y_r^2 - 1)(4z_s^3 - 3z_s) + (2z_s^2 - 1)(4y_s^3 - 3y_s)\}.$$
(3.25)

Note that  $\mu_{r,s}$  is a symmetric function of  $y_r$  and  $z_s$ .

In order to find the maximum  $\mu_{r,s}$  (for all r=0,  $1, \ldots, M-1$ , and  $s=0, 1, \ldots, N-1$ ), we shall first treat y and z as continuous variables, but subsequently we shall take account of the fact that y and z must assume the discrete values  $y_r$  and  $z_s$ . Differentiating (3.25) partially with respect to y whilst z is held constant, we get

$$\frac{\partial \mu}{\partial y}(y, z) = -354 + 264 y^2 - 66 + 48(4z^3 - 3z)y + 144(2z^2 - 1)y^2 - 36(2z^2 - 1)$$
$$= (264 + 144\cos 2s\psi)y^2 + (48\cos 3s\psi)y - (420 + 36\cos 2s\psi).$$
(3.26)

This is a quadratic expression in y, which is positive for  $y = \pm \infty$ . But when  $y = \pm 1$ , then according to (3.26)

$$\frac{\partial \mu}{\partial y} = 264 + (144 - 36)\cos 2s\psi - 420 \pm 48\cos 3s\psi$$
  

$$\leq 264 + 108 - 420 + 48 = 0.$$
 (3.27)

Since  $\frac{\partial \mu}{\partial y}$  is a quadratic expression in y, it follows that

$$\frac{\partial \mu}{\partial \nu} \leqslant 0 \text{ for all } -1 \leqslant \gamma \leqslant 1. \tag{3.28}$$

Similarly, in view of the symmetry of y and z in (3.25), we conclude that

$$\frac{\partial \mu}{\partial z} \leqslant 0 \text{ for all } -1 \leqslant z \leqslant 1. \tag{3.29}$$

Thus, when y and z are restricted to the discrete values  $y_r$  and  $z_s$ , the maximum value of  $\mu_{r,s}$  will be given by the minimum values of  $y_r$  and  $z_s$ . If M is even, then the minimum  $y_r$  is given by  $r = \frac{1}{2}M$ , in which event  $y_{min} = -1$ ; and similarly  $z_{min} = -1$  if N is even. But we shall first consider the case when both M and N are odd. Let

$$M = 2m + 1 N = 2n + 1$$
 (3.30)

(The cases where M and/or N are even are considered in §3.7.) Then the maximum value of  $\mu_{r,s}$  for all r and s will be equal to  $\mu_{m,n}$  where

$$y_m = \cos\left(\frac{2\pi m}{2m+1}\right) = \cos\left(\pi - \frac{\pi}{2m+1}\right)$$
$$= -\cos\left(\frac{\pi}{M}\right) = 2\sin^2\left(\frac{\pi}{2M}\right) - 1 \qquad (3.31)$$

and similarly

$$z_n = 2\sin^2\left(\frac{\pi}{2N}\right) - 1.$$
 (3.32)

Write 
$$Y = \sin\left(\frac{\pi}{2M}\right), Z = \sin\left(\frac{\pi}{2N}\right).$$
 (3.33)

Substituting (3.31) and (3.32) into (3.25) and using (3.33), after rearrangement we get

$$\mu_{m,n} = 1,280 - 1,728(Y^2 + Z^2)^2 + 1,088(Y^2 + Z^2)^3 + 3,072Y^2Z^2(Y^2 + Z^2 - Y^2Z^2)(1 - Y^2 - Z^2).$$
(3.34)

Let the mesh of the net covering the region R be refined, i.e.  $h \to 0$ . Then  $M = O(h^{-1})$  and  $N = O(h^{-1})$ ; and equations (3.33) and (3.34) show that

$$\mu_{m,n} = 1,280 - 108\pi^4(M^{-2} + N^{-2})^2 + O(h^6)$$
 (3.35)

so that

$$1,280 - \mu_{m,n} = O(h^4).$$
 (3.36)

# 3.7 Maximum and minimum eigenvalues of matrix of equations

Consider a symmetric matrix A, and a principal minor B of A. Then if x is any vector with the same number of rows as A, normalized so that  $x^Tx = 1$ , the maximum and minimum values (for all x such that  $x^Tx = 1$ ) of the "Rayleigh quotient"  $x^TAx$  are equal to the maximum and minimum eigenvalues of A (Fox, 1957, p. 162). Since B also is symmetric, its maximum and minimum eigenvalues will be the maximum and minimum values of  $v^TBv$ , for all vectors v with the same number of rows as B, normalized so that  $v^Tv = 1$ .

Expand v to form a vector u with the same number of rows as A, by the insertion of zero elements corresponding to those rows (and columns) of A which have been deleted to form B. Then

$$u^{T}Au = v^{T}Bv,$$
  

$$u^{T}u = v^{T}v = 1.$$
 (3.37)

But the class of vectors u cannot include all vectors x with the same number of rows as A, since certain elements of u are fixed at zero. Accordingly

$$\operatorname{Max}_{x^{T}x=1}(x^{T}Ax) \geqslant \operatorname{Max}_{u^{T}u=1}(u^{T}Au) = \operatorname{Max}_{v^{T}v=1}(v^{T}Bv) 
\geqslant \operatorname{Min}_{v^{T}v=1}(v^{T}Bv) = \operatorname{Min}_{u^{T}u=1}(u^{T}Au) \geqslant \operatorname{Min}_{x^{T}x=1}(x^{T}Ax). \quad (3.38)$$

Thus the spectrum of eigenvalues of B is contained

within the spectrum of eigenvalues of A; i.e. all eigenvalues of a principal minor of a symmetric matrix lie between the minimum and maximum eigenvalues of the original matrix.

We have seen in §3.6 that the circulant matrix K contains the 7-band matrix  $J_r$  as a principal minor, and hence all the eigenvalues  $\lambda_{r, s}$  of  $J_r$  are less than or equal to  $\mu_{m, n}$  (cf. (3.34)). The eigenvalues of the compound-circulant matrix H were shown to be the  $\lambda_{r, s}$  (cf. §3.5), and the compound 7-band matrix  $108h^4G$  is a principal minor of H (cf. §3.4). Finally, the original matrix F (cf. (3.2)) of the complete set of equations is itself a principal minor of G. Thus we conclude that all eigenvalues of  $108h^4F$  are less than or equal to  $\mu_{m, n}$ , i.e. if  $\lambda$  is any eigenvalue of F, then

$$108h^{4}\lambda \le 1,280 - 108\pi^{4}(M^{-2} + N^{-2})^{2} + O(h^{6});$$

$$(M = 2m + 1, N = 2n + 1). \tag{3.39}$$

Until now we have assumed that M and N are both odd (cf. (3.30)). If either M or N is even, let the extended net be further extended by a single row if M is even and by a single column if N is even. Then the extended net will have an odd number of rows and columns, and the above reasoning will be valid with M and/or N replaced by M+1 and/or N+1 respectively. On the other hand, if M, say, is odd then the expression on the right of (3.34) (which equals  $\mu_{m,n}$  if N also is odd) will be increased if M is replaced by M+1, in view of (3.27) and (3.33); and similarly if N is odd the expression on the right of (3.34) will be increased if N is replaced by N+1, in view of (3.29) and (3.33).

Hence, whether or not M and N are even or odd, the following inequality holds:

$$108h^4\lambda \le 1,280 - 108\pi^4((M+1)^{-2} + (N+1)^{-2})^2 + O(h^6).$$
 (3.40)

But 
$$M = O(h^{-1})$$
,  $N = O(h^{-1})$ , and accordingly
$$((M+1)^{-2} + (N+1)^{-2})^2$$

$$= (M^{-2}(1+M^{-1})^2 + N^{-2}(1+N^{-1})^2)^2$$

$$= (M^{-2}(1-2M^{-1}+O(h^2))$$

$$+ N^{-2}(1-2N^{-1}+O(h^2)))^2$$

$$= (M^{-2} + N^{-2} - 2M^{-3} - 2N^{-3} + O(h^4))^2$$

$$= (M^{-2} + N^{-2})^2 + O(h^5). \tag{3.41}$$

Thus the inequality (3.40) may be simplified to give  $108h^4\lambda < 1{,}280 - 108\pi^4(M^{-2} + N^{-2})^2 + O(h^5)$ . (3.42)

The diagonal elements of the matrix F are all equal to  $\frac{640}{108h^4}$  (cf. (3.3) and (3.4a)). Therefore, if as above  $\lambda$  is any eigenvalue of F, the Simultaneous Displacement Method (or S.D.M.; named variously after Gauss, Jacobi, and Richardson) will have an eigenvalue equal to  $\frac{108h^4\lambda}{640} - 1$  (Engeli, 1959, p. 90). Hence if  $\beta$  is the

maximum eigenvalue of S.D.M. applied to the complete set of equations  $F\Phi = b$ , then

$$\beta < 1 - \frac{108\pi^4}{640}(M^{-2} + N^{-2})^2 + O(h^5).$$
 (3.43)

But the matrix F has "Property A," from which it can be shown that all non-zero eigenvalues of S.D.M. occur in pairs with equal modulus and opposite sign (cf. Young, 1954, p. 98). Therefore the minimum eigenvalue of S.D.M. is  $-\beta$ , from which it follows that the minimum eigenvalue of F satisfies the inequality:

$$108h^4\lambda_{\min} > 108\pi^4(M^{-2} + N^{-2})^2 + O(h^5) \quad (3.44)$$

therefore 
$$\lambda_{\min} > \pi^4 \left\{ \frac{1}{(Mh)^2} + \frac{1}{(Nh)^2} \right\}^2 + O(h)$$
  
=  $\left\{ \left( \frac{\pi}{W_x} \right)^2 + \left( \frac{\pi}{W_y} \right)^2 \right\}^2 + O(h)$  (3.45)

where  $W_x$  and  $W_y$  are the maximum dimensions of the region R in the directions of the x and y axes, respectively (cf. Fig. 7)\*. The matrix F is symmetric, and since all of its eigenvalues have been shown to be strictly positive, then F is positive-definite. Therefore S.O.R. will converge when it is applied to F, for any  $0 < \omega < 2$  (cf. Ostrowski, 1954).

#### 3.8 Convergence rate of S.O.R.

When the equations  $F\Phi = b$  are solved by S.O.R., then for every fixed value of  $\omega$  (the relaxation parameter of the S.O.R. process), the convergence rate is maximized if the equations and the unknowns are permuted into a "consistent ordering" (Varga, 1959). Apparently the only type of consistent ordering which can be applied with the optimized 17-node molecule (cf. Fig. 5) is a " $\sigma_1$ -ordering" (Young, 1954, p. 108), in which during each cycle of S.O.R. all estimates of  $\phi$  at nodes of one class (say X, cf. Fig. 2) are adjusted before those at the nodes of the opposite class. The eigenvalues of F are unaffected by such a permutation into a consistent ordering, and hence the eigenvalues of S.D.M. (including  $\beta$ ) are also unchanged.

With the equations  $F\Phi = b$  permuted into a consistent ordering, the convergence rate of S.O.R. is maximized by using the optimal value  $\omega_0$  of  $\omega$ , where  $\omega_0$  is given by

$$\omega_0 = 1 + \frac{1 - \sqrt{(1 - \beta^2)}}{1 + \sqrt{(1 - \beta^2)}}$$
 (3.46)

(Young 1954, p. 95; Forsythe and Wasow 1960, p. 253). When  $\omega \geqslant \omega_0$ , all eigenvalues of the error operator have modulus equal to  $\omega - 1$ .

Substituting (3.43) into (3.46), after simplification we get:

$$\omega_0 - 1 < 1 - \pi^2 \sqrt{\left(\frac{27}{20}\right)} (M^{-2} + N^{-2}) + O(h^3).$$
 (3.47)

\* A positive lower bound for  $\lambda_{\min}$  cannot be deduced directly from the fact that the eigenvalues of  $108h^4F$  are contained within the range of values of  $\lambda_{r,s}$ ; for it is readily shown that in view of the fact that the sum of the coefficients of the basic molecule is zero (cf. (2.3)), then  $\lambda_{0,0} = 0$ .

Defining the convergence rate  $\rho$  as minus the natural logarithm of the spectral radius\* of the error operator (Young, 1954, p. 96), we see from (3.46) that the convergence rate of S.O.R. with optimal  $\omega$  satisfies the inequality

$$\rho = - \operatorname{Ln}(\omega_0 - 1)$$

$$= 2((1 - \beta^2)^{1/2} + \frac{1}{3}(1 - \beta^2)^{3/2} + \frac{1}{5}(1 - \beta^2)^{5/2} + \ldots)$$

$$> \pi^2 \sqrt{(\frac{27}{20})(M^{-2} + N^{-2}) + O(h^3)}.$$
(3.48)

Of course, (3.48) is a rather coarse under-estimate of the actual optimized convergence rate, in view of the chain of inequalities by means of which it has been established. However, it shows that for a fixed region R, the optimized convergence rate of S.O.R. is of the order  $O(h^2)$  as the mesh size  $h \to 0$ .

#### 4. Truncation errors

The principal term in the local truncation error for the 17-node molecule is (cf. (2.14))

$$-h^{2} \left[ \frac{3,407}{12,960} \left\{ \frac{\delta^{6}\phi}{\delta x^{6}} + \frac{\delta^{6}\phi}{\delta y^{6}} \right\} + \frac{3}{8} \left\{ \frac{\delta^{6}\phi}{\delta x^{4}\delta y^{2}} + \frac{\delta^{6}\phi}{\delta x^{2}\delta y^{4}} \right\} \right]$$

$$= h^{2} \left\{ \frac{-3,407}{12,960} \nabla^{6}\phi + \frac{1,787}{4,320} \zeta_{2,4} \right\}. \tag{4.1}$$

The principal term in the local truncation error for the conventional 13-node molecule is (cf. (1.1))

$$-\frac{h^2}{6} \left\{ \frac{\delta^6 \phi}{\delta x^6} + \frac{\delta^6 \phi}{\delta x^4 \delta y^2} + \frac{\delta^6 \phi}{\delta x^2 \delta y^4} + \frac{\delta^6 \phi}{\delta y^6} \right\}$$
$$= h^2 \left\{ -\frac{1}{6} \nabla^6 \phi + \frac{1}{3} \zeta_{2,4} \right\}. \tag{4.2}$$

In many problems,  $\nabla^4 \phi$  varies quite smoothly over R; indeed it is often true that  $\nabla^2(\nabla^4 \phi) \equiv 0$  over R. If this is the case, then the principal terms in the local truncation errors of the 17-node and 13-node molecules reduce to  $\frac{1,787}{4,320}h^2\zeta_{2,4}$  and  $\frac{1}{3}h^2\zeta_{2,4}$ , respectively. Thus despite the fact that the 17-node molecule connects the central node to nodes which are more distant (on the average) than does the 13-node molecule, the local truncation error (with which the molecule approximates to the differential equation at an internal node) is only about 1.24 times that for the 13-node molecule.

The equations (3.1) fix the values of  $\phi$  at a triple layer of nodes enclosing  $\Gamma$ , with errors in  $\phi$  of the order  $O(h^2)$ . Similar formulae may be applied when the conventional 13-node molecule is applied, but in that case only a double layer need be constructed. In the case of a straight boundary, this means that the net could be so drawn that  $\Gamma$  lay midway between mesh lines, in which event the errors in the values of  $\phi$  at nodes distant  $\frac{1}{2}h$  from  $\Gamma$  will be about  $\frac{1}{4}$  as large as those when the nodes

<sup>\*</sup> The spectral radius of a matrix is defined as the maximum modulus of any of its eigenvalues.

are distant h from  $\Gamma$ , as in Fig. 6. Thus the errors at the boundary nodes when the 17-node molecule is used will be only a few times (c4) as large as when the 13-node molecule is used.

## 4.1 Effect of errors in representing boundary condition

Let the differential biharmonic problem be

$$\nabla^{4} \phi = f(x, y) \text{ in } R$$

$$\phi = g(s), \frac{\partial \phi}{\partial \nu} = n(s) \text{ on } \Gamma$$
(4.3)

where s represents are length along  $\Gamma$  from some fixed point. Let accented Greek capitals ( $\Phi'$  and  $\Psi'$ ) be used to represent vectors of the exact nodal values of continuous functions, which are themselves denoted by small Greek letters ( $\phi$  and  $\psi$ ); whereas  $\Phi$  is a vector of approximations to the nodal values of  $\phi$ , satisfying equation (3.2), i.e.  $\Phi = F^{-1}b$ .

Instead of (3.1) we may write the exact equation:

$$\mathbf{F}\Phi' = b' + h^2r \tag{4.4}$$

where

$$b' = f + h^{-4}b''. (4.5)$$

Here f is a vector of the values of  $\nabla^4 \phi$  at internal nodes; b'' is a vector containing values of  $\phi$  at boundary nodes, multiplied by coefficients -3, -11 or +177; and  $h^2r$  represents the local truncation errors, so that each element of r represents the value of

$$\frac{3,407}{12,960}\,\zeta_{0,\,6}+\tfrac{3}{8}\zeta_{2,\,4}+\ldots$$

at an internal node (cf. (2.14)).

Almost never does it happen that the values of  $\phi$  are known exactly at all boundary nodes—usually they will be determined as in (3.1) with errors  $O(h^2)$ . In that case the elements of b will differ by  $O(h^{-2})$  from the elements of b' (cf. (4.5)). The errors of order  $O(h^2)$  in  $\phi$  occur at nodes which are distant from  $\Gamma$  by O(h). Let the exact solution  $\phi$  of (4.3) be perturbed to give a continuous function  $\psi$ , such that  $\nabla^4 \phi = \nabla^4 \psi$  in R and  $\phi = \psi$  on  $\Gamma$ , and such that  $\psi$  equals the values assigned to  $\phi$  at the boundary nodes. Then  $(\psi - \phi) \equiv 0$  on  $\Gamma$  and  $\nabla^4 (\psi - \phi) \equiv 0$  in R, but  $(\psi - \phi) = O(h^2)$  at nodes distant O(h) from  $\Gamma$ . Hence  $\frac{\delta}{\delta \nu} (\psi - \phi) = O(h)$  on  $\Gamma$ .

The fact that the biharmonic problem (4.3) is well-posed under the given boundary conditions may be expressed more precisely in the following manner. Let  $\Gamma$  possess a continually turning tangent, and let g(s), n(s) and  $\frac{dg}{ds}$  be continuous on  $\Gamma$  (cf. (4.3)). Then Miranda has shown (cf. Collatz, 1960, p. 407) that the solution of the biharmonic problem with  $f(x, y) \equiv 0$  (i.e.  $\nabla^4 \phi = 0$  in R) satisfies the inequalities

$$\sqrt{\left[\left(\frac{\partial\phi}{\partial x}\right)^{2} + \left(\frac{\partial\phi}{\partial y}\right)^{2}\right]} \leq Q$$

$$= K_{1} \left\{ \underset{\Gamma}{\text{Max}} |n| + \underset{\Gamma}{\text{Max}} \left|\frac{dg}{ds}\right| \right\} + K_{2} \underset{\Gamma}{\text{Max}} |g|$$

$$|\phi(P)| \leq Q\delta + \underset{\Gamma}{\text{Max}} |g|.$$
(4.6)

Here P is any internal point of R,  $\delta$  is the minimum distance from P to any point of  $\Gamma$ ; and  $K_1$ ,  $K_2$  are constants depending only on the geometry of R: the computation of numerical values for these constants is somewhat laborious.

Applying Miranda's Theorem to the function  $(\psi - \phi)$ , we see that

$$|(\psi - \phi)|_{P} \leqslant K_{1} \delta \max_{\Gamma} \left| \frac{\partial}{\partial \nu} (\psi - \phi) \right|$$

$$= O(h). \tag{4.7}$$

In particular, at each internal node  $\psi$  differs from  $\phi$  by O(h), i.e. the elements of the vector  $\Psi' - \Phi'$  are each O(h).

Note that this argument applies equally well to equations based on the conventional 13-node molecules, when the boundary conditions are approximated by fixing  $\phi$  (with errors  $O(h^2)$ ) at a double layer of nodes enclosing  $\Gamma$ .

#### 4.2 Truncation error of the solution

Applying (4.4) to the perturbed function  $\psi$ , we see that the vector  $\Psi'$  of its nodal values satisfies the equation

$$\mathbf{F}\Psi' = b + h^2 r \tag{4.8}$$

where the boundary values contained in b are the values computed for  $\phi$  at boundary nodes according to (3.1), but the 6th derivatives in the elements of r are those of  $\psi$  rather than  $\phi$ .

The solution of (4.8) is

$$\Psi' = F^{-1}b + h^{2}F^{-1}r$$

$$= \Phi + h^{2}F^{-1}r. \tag{4.9}$$

Now the matrix G is symmetric, and hence F is symmetric, and so also  $F^{-1}$  is symmetric. Therefore the spectral norm of  $F^{-1}$  is equal to its spectral radius, i.e. the maximum modulus of any of its eigenvalues.\*

Now we have seen (cf. (3.45)) that all eigenvalues of F are greater than the constant

$$\left[\left(\frac{\pi}{W_x}\right)^2 + \left(\frac{\pi}{W_y}\right)^2\right]^2$$

(ignoring O(h)). Therefore all eigenvalues of  $F^{-1}$  are positive and less than

$$\left[\left(\frac{\pi}{W_{x}}\right)^{2}+\left(\frac{\pi}{W_{y}}\right)^{2}\right]^{-2}.$$

\* The spectral norm ||A|| of a matrix A is defined as  $\sqrt{\lambda_1}$ , where  $\lambda_1$  is the spectral radius of A\*A. The corresponding norm ||x|| of a vector x is the Euclidean length of x, for which

$$||Ax|| \leqslant ||A|| \cdot ||x||$$

(Faddeeva, 1959, p. 59).

Accordingly

$$||\Psi' - \Phi|| = ||h^2 F^{-1} r|| \le |h^2 ||F^{-1}|| \cdot ||r||$$

$$< h^2 \left[ \left( \frac{\pi}{W_x} \right)^2 + \left( \frac{\pi}{W_y} \right)^2 \right]^{-2} ||r||. \quad (4.10)$$

From this, we see that the ratio of the root-meansquare of the elements of  $(\Psi' - \Phi)$  to the root-meansquare of the elements of r is of order  $O(h^2)$  as  $h \to 0$ . If it is assumed that  $\phi$  can be perturbed into  $\psi$  in such a manner that the sixth derivatives of  $\psi$  remain bounded as  $h \to 0$ , then the elements of r remain bounded. (On the other hand, if the exact values of  $\phi$  at the boundary nodes are known then  $\psi$  is simply  $\phi$ , and we need only assume that the sixth derivatives of  $\phi$  itself are bounded.)

The number n of internal nodes of the net is  $O(h^{-2})$ . Since the elements of r are assumed to be bounded, we have:

$$||r||^2 = r_1^2 + \dots r_n^2 = O(n) = O(h^{-2})$$
 (4.11)

therefore 
$$||r|| = O(h^{-1})$$
. (4.12)

Substituting (4.12) into (4.10), we get

$$||\Psi' - \Phi|| = h^2 O(h^{-1}) = O(h).$$
 (4.13)

Thus the vector  $\Phi$  which is actually found as the solution of (3.2) differs from the vector  $\Psi'$  of the exact nodal values of  $\psi$  by an error vector whose length is O(h). Hence the individual elements of  $(\Psi' - \Phi)$  are not greater than O(h), whilst their root-mean-square is  $O(h^2)$ . (cf. (4.10)).

We have seen in §4.1 that the elements of the vector  $(\Psi' - \Phi')$  are O(h). Combining this with the result of (4.13), we see that the elements of  $(\Phi - \Phi')$  are bounded by quantities of order O(h).

Note that the effects of errors in  $\phi$  at boundary nodes could not have been investigated profitably by using norms, since the elements of b will contain errors of order  $O(h^{-2})$  (cf. (4.5)).

On the other hand, the root-mean-square truncation error would be of order  $O(h^2)$  if the values at the boundary nodes were fixed with errors of order  $O(h^3)$ . This suggests that the accuracy of the results could probably be improved significantly by first solving the net equations with boundary values specified as in (3.1), estimating the second derivatives of  $\phi$  at nodes near the boundary by finite-difference techniques applied to the results, and then using these estimated values of

$$\frac{\partial^2 \phi}{\partial x^2}$$
,  $\frac{\partial^2 \phi}{\partial x \partial y}$  and  $\frac{\partial^2 \phi}{\partial y^2}$ 

extrapolated to  $\Gamma$  to correct the values of  $\phi$  assigned to the boundary nodes. The process could be repeated until the values computed at internal nodes cease to alter significantly. However, a more economical procedure would be to specify the values at boundary nodes as in (3.1), then partly solve the net equations (3.2) by applying S.O.R. until the estimated second derivatives of the current solution have settled down to steady

values. These may then be used to improve the values assigned to boundary nodes, and S.O.R. may then be resumed to continue the improvement of the values at internal nodes. The boundary values may be adjusted repeatedly in this manner during the process of solution by S.O.R.

#### 4.3 Errors of 13-node molecule

The local truncation errors of the 13-node and 17-node molecules have already been compared in (4.1) and (4.2). In order to compare the truncation error of the solution for the sets of equations based on the respective molecules, it is necessary to know also a lower bound for the eigenvalues of the matrix based on the 13-node molecule.

As in §3.4, the eigenvalues of the original matrix will lie within the range of eigenvalues for the matrix corresponding to an extended rectangular net containing the original net, with a double layer of fixed values of  $\phi$  around the boundary of the rectangle. Reasoning in a manner similar to that employed in §3.5–§3.8, we could deduce an upper bound for the eigenvalues of the matrix; but since the matrix does *not* have "Property A" we could not deduce therefrom a positive lower bound for the eigenvalues as in (3.44): the circulant matrix could only show that zero was a lower bound (cf. footnote below (3.45)).

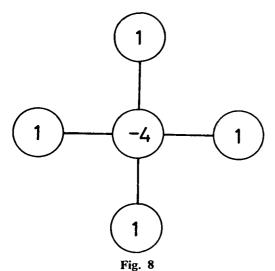
Consider in more detail the matrix L resulting from the application of the 13-node molecule at each internal node of the rectangular net as in Fig. 7, with each value of  $\phi$  at a boundary node transferred to the right-hand side as a known quantity. The outermost of the triple layer of nodes of the net has no effect on the values at internal nodes in this case (since the 13-node molecule only "reaches" over 2h), so that we need only to consider a rectangle of  $M \times N$  squares (each  $h \times h$ ) with  $\phi$  fixed at the outermost nodes and at the adjacent internal nodes. Thus L has (M-3)(N-3) rows and columns. Number the nodes in pagewise order.

Next, consider a matrix P with (M-1)(N-1) rows and columns, corresponding to the 5-node approximation to the Dirichlet problem over the rectangle with dimensions  $Mh \times Nh$ . This means that the values of  $\phi$  are specified on the rectangular boundary and the values of  $\nabla^2 \phi$  are known inside the rectangle. At each internal node of the rectangle, an equation is written of the form

$$\frac{1}{h^2} [\phi_{i-1,j} + \phi_{i,j-1} - 4\phi_{i,j} + \phi_{i,j+1} + \phi_{i+1,j}] 
= (\nabla^2 \phi)_{i,j} \quad (4.14)$$

which may be symbolized by the Laplace molecule of Fig. 8.

The 13-node molecule of Fig. 1 is, in a certain specific sense, the square of the harmonic operator of Fig. 8. By this it is meant that the elements of the matrix  $P^2$  in the rows and columns corresponding to the (M-3) net rows and the (N-3) net columns at which the 13-node molecule is applied are, as may be verified by inspection, identical with the elements of L in the corresponding



rows and columns of  $L^*$ . Thus the matrix L is a principal minor of  $P^2$ , and hence its eigenvalues lie within the range of the eigenvalues of  $P^2$ , which are of course the squares of the eigenvalues of P.

The eigenvalues of P may be found in the following manner (Frankel, 1950). Consider the set of (M + 1) (N + 1) quantities

$$e_{i,j}^{r,s} = \sin\left(\frac{r\pi i}{M}\right) \cdot \sin\left(\frac{s\pi j}{N}\right)$$

$$(0 \le i \le M)$$

$$(0 \le j \le N)$$

$$(4.15)$$

for fixed r and s, where 0 < r < M and 0 < s < N. Substituting  $e_{i,j}^{r,s}$  for  $\phi_{i,j}$  in each term on the left of (4.14), we find that at each internal node of the rectangle (0 < i < M, 0 < j < N)

$$\frac{1}{h^{2}} \left[ e_{i-1,j}^{r,s} + e_{i,j-1}^{r,s} - 4e_{i,j}^{r,s} + e_{i,j+1}^{r,s} + e_{i+1,j}^{r,s} \right] 
= \frac{1}{h^{2}} \left[ 2\cos\left(\frac{r\pi}{M}\right) + 2\cos\left(\frac{s\pi}{N}\right) - 4 \right] e_{i,j}^{r,s}. \quad (4.16)$$

Thus for any fixed r and s the set of  $e_{i,j}^{r,s}$  form an eigenvector of the matrix P of the set of equations (4.8), corresponding to the eigenvalue

$$\mu_{r, s} = \frac{1}{h^2} \left[ 2 \cos\left(\frac{r\pi}{M}\right) + 2 \cos\left(\frac{s\pi}{N}\right) - 4 \right]$$
 (4.17)  
 
$$(0 < r < M, 0 < s < N).$$

It is easy to show that (4.15) gives a complete set of eigenvectors of P, and hence all eigenvalues of P are given by (4.17). The maximum and minimum eigenvalues of P are therefore given by (r, s) = (1, 1) and (r, s) = (M - 1, N - 1), respectively. (Note that all

\*  $P^2$  is actually the matrix appropriate to a biharmonic problem over the rectangle of dimensions  $Mh \times Nh$ , when the boundary conditions specify the values of  $\phi$  and  $\nabla^2 \phi$  on the boundary; as occurs for example in the deflections of a simply-supported rectangular plate.

 $\mu_{r, s}$  are strictly negative). Hence the minimum eigenvalue of  $P^2$  equals  $\mu_{1,1}^2$ , so that all eigenvalues of L are greater than (or equal to)

$$\mu_{1.1}^{2} = \frac{4}{h^{4}} \left[ \left\{ \cos \left( \frac{\pi}{N} \right) - 1 \right\} + \left\{ \cos \left( \frac{\pi}{N} \right) - 1 \right\} \right]^{2}$$

$$= \frac{4}{h^{4}} \left[ -2 \sin^{2} \left( \frac{\pi}{2M} \right) - 2 \sin^{2} \left( \frac{\pi}{2N} \right)^{2} \right]^{2}$$

$$= \frac{16}{h^{4}} \left[ \left( \frac{\pi}{2M} \right)^{2} + \left( \frac{\pi}{2N} \right)^{2} \right]^{2} + O(h^{2})$$

$$= \left[ \left\{ \frac{\pi}{W_{x}} \right\}^{2} + \left\{ \frac{\pi}{W_{y}} \right\}^{2} \right]^{2} + O(h^{2}). \tag{4.18}$$

Comparing (3.45) with (4.18), we reach the remarkable conclusion that the lower bounds deduced for the eigenvalues of the matrices based on the 13-node and the 17-node molecules are *identical* (apart from quantities of order O(h) and  $O(h^2)$ ). Thus all eigenvalues of the matrix based on the 13-node molecule within the actual region R will be greater than or equal to the same lower bound (4.18) as was used for the 17-node molecule.\* In particular, the matrix is positive-definite.

Referring to §4.1, we see that when the exact values of  $\phi$  are used at the boundary nodes the norm of the truncation error of the solution will be related to the norm of the vector of local truncation errors by the same inequality (4.10) whether the 13-node or the 17-node molecule is used. The local truncation errors for the two molecules are compared in §4, in which it was shown that if  $\nabla^4 \phi$  is itself a harmonic function then the local truncation error of the 17-node molecule is only about 24% greater than the local truncation error for the 13-node molecule. We conclude that the bounds for the truncation error of the solution when the 17-node molecule is used will be only about 24% greater than when the 13-node molecule is applied over the same net.

However, the errors in fixing the values of  $\phi$  at nodes adjacent to  $\Gamma$  will introduce further errors of order O(h), in exactly the same manner for both molecules (cf. §4.2).

#### 5. Numerical experiments

In order to compare the under-estimate (3.48) of the S.O.R. convergence rate with the actual convergence rate for specimen problems, a program was written in TOPIC autocode to perform S.O.R. (with  $\sigma_1$ -ordering) upon the net equations resulting from applying the 17-node molecule at all internal nodes of a rectangular region surrounded by a triple layer of zero boundary values, i.e. the matrix F is G of (3.3).

The solution of the equations is of course zero at every node, so that if arbitrary initial values are taken at internal nodes the current estimate at any stage will equal the error at that stage. Accordingly, the decay

\* There would be little point in investigating further the upper bound  $\mu_{M-1, N-1}^2$  for the eigenvalues, since it is not possible to deduce therefrom the optimized convergence rate of S.O.R. for a matrix without "Property A."

rate for any value of  $\omega$  may be computed from observations of the decay rate of the current estimates, when all their elements decay uniformly.

#### 5.1 Test cases

If S.O.R. is applied with  $\omega < \omega_0$ , then the successive current estimates will ultimately decay uniformly by a decay factor  $\mu$ , where  $\omega_0$  is related to  $\mu$  by the formula (Carré, 1961):

$$\omega_0 = \frac{2}{1 + \sqrt{\left[1 - \frac{(\mu + \omega - 1)^2}{\omega^2 \mu}\right]}}.$$
 (5.1)

This value may be checked by applying S.O.R. with a value of  $\omega$  slightly less than the value predicted by (5.1), and re-estimating  $\omega_0$  from the observed decay factor with this new value of  $\omega$ .

The nets chosen for the test problems had (M, N) = (13, 13) and (17, 13); i.e. there were 10 (or 14) rows of internal nodes and 10 columns of internal nodes. The convergence rates predicted by (3.48) (neglecting  $O(h^3)$ ) are

(a) 
$$\rho_a > \pi^2 \sqrt{\left(\frac{27}{20}\right)} (13^{-2} + 13^{-2}) = 0.135$$
 (5.2a)

(b) 
$$\rho_b > \pi^2 \sqrt{\left(\frac{27}{20}\right)} (13^{-2} + 17^{-2}) = 0.108, \quad (5.2b)$$

and the optimal values of  $\omega$  predicted by (3.47) (neglecting  $O(h^3)$ ) are

(a) 
$$\omega_{0a} - 1 < 1 - \pi^2 \sqrt{\left(\frac{27}{20}\right)} (13^{-2} + 13^{-2}) = 0.865$$
 (5.3a)

(b) 
$$\omega_{0b} - 1 < 1 - \pi^2 \sqrt{\left(\frac{27}{20}\right) (13^{-2} + 17^{-2})} = 0.892.$$
 (5.3b)

In Table 2 there are given for each case a value of  $\omega$ , the value of  $\omega_0$  predicted by (5.1) from the observed decay factor  $\mu$  when S.O.R. is applied with that value of  $\omega$ , and the number  $n_1$  of cycles of S.O.R. which were found to be necessary for producing a reliable estimate of  $\mu$ . In each case the initial estimate was taken as unity at all internal nodes. The table also gives the estimate of the optimized convergence rate  $R = -\text{Ln}(\omega_0 - 1)$ ;

and it gives  $\nu = \frac{\text{Ln } 2}{R}$ , which is an estimate of the number of cycles of S.O.R. which ultimately are needed to halve

the errors.

Comparing Table 2 with (5.2), we see that in both of the test problems the actual optimized convergence rate for S.O.R. is nearly double the theoretical lower bound.

# 5.2 Other molecules

The molecule of Fig. 4 (containing only  $\alpha$ ,  $\beta$  and  $\gamma$  nodes) produces matrices with "Property A," but it was discarded in favour of the 17-node molecule of

Table 2 Estimation of  $\omega_0$ 

М	N	ω	$n_1$	$\omega_0$	R	ν
13 13 17	13 13 13	1·625 1·75 1·797	24 28 40	1·77 1·770 1·806	0·262 0·215	

Fig. 5, which is more nearly diagonally dominant. However, as a check this molecule of Fig. 4 was also applied to the test case with (M, N) = (13, 13); i.e. with  $10 \times 10$  internal nodes. A TOPIC program was written to perform S.O.R. upon the resulting equations, and it was found that the optimized  $\omega_0$  was 1.809, whereas the molecule of Fig. 5 gave  $\omega_0 = 1.770$  (cf. Table 2). This confirms our expectation that S.O.R. would converge more rapidly for equations based on the 17-node molecule of Fig. 5, rather than on that of Fig. 4.

In order to compare the 17-node molecule with the conventional 13-node molecule of Fig. 1, the 13-node molecule was applied to the test problem with  $10 \times 10$ internal nodes. Since the resulting matrix does not have "Property A," it is practically impossible to decide which ordering of the nodes is most advantageous for S.O.R. (Engeli, 1959, p. 92, and Heller, 1958) and for any particular ordering it is practically impossible to predict the optimum value of  $\omega$ . For simplicity the nodes were numbered pagewise, and the DEUCE IIA program BHM1 was employed to perform S.O.R. upon the 100 equations for various values of  $\omega$ . Table 3 gives the observed value of  $1 - \mu$  for each value of  $\omega$ , where  $\mu$  is the ultimate decay factor of the errors. It also gives the number  $n_2$  of cycles of S.O.R. which were necessary for estimating  $\mu$  correctly to three decimal places.

Table 3
Observed decay factors

ω	$n_2$	$1-\mu$
1	62	0.008
1.25	35	0.011
1.5	125	0.022
1.625	85	0.037
1.6875	59	0.057
1.75	63	0.062
1.8125	52	0.031

The asterisks indicate that the errors decayed in an oscillatory manner, and the values cited for  $\mu$  represent approximate mean decay factors (corresponding to the moduli of the complex dominant eigenvalues).

Table 3 indicates that several hundred cycles of S.O.R. were needed in order to determine the optimal value of

 $\omega$  (for which  $\mu$  was minimized) for the 13-node molecule. whereas Table 2 shows that 24 cycles were sufficient for  $\omega_0$  to be estimated with adequate accuracy for the 17-node molecule. Further, Table 3 shows that the minimum value of  $\mu$  is equal to 1 - 0.062 = 0.938. The corresponding optimized convergence rate  $\rho = - \text{Ln} (1 - \mu)$ = 0.062, whereas the optimized  $\rho = 0.262$  for the 17-node molecule. Thus the convergence for S.O.R. is 4.2 times as fast as with the 13-node molecule. However, during one cycle of S.O.R. each internal node requires 18 multiplications and 18 additions with the 17-node molecule, but only 14 multiplications and 14 additions with the 13-node molecule. Thus in terms of the number of arithmetic operations required, the 17-node molecule gives equations with which S.O.R. converges 3.3 times as rapidly as for the 13-node molecule  $\left(4\cdot2\times\frac{14}{18}\simeq3\cdot3\right).$ 

#### 5.3 Conclusions

The 17-node molecule has the following advantages over the conventional 13-node molecule, when net equations are constructed over a region with simple boundary conditions and the resulting equations are being solved by S.O.R.

- (a) An upper bound can be predicted for  $\omega_0$  and a lower bound for  $\rho$ .
- (b) The actual optimal  $\omega$  can be computed after a small number of cycles of S.O.R.
- (c) The optimized convergence rate is several times as

large as that for the 13-node molecule (in the test cases, at any rate). The number of arithmetic operations per cycle is only slightly greater than with the 13-node molecule.

The local truncation errors at internal nodes, and the corresponding part of the truncation error of the solution are each about \( \frac{5}{4} \) times those for the 13-node molecule. However, the major part of the truncation error of the solution in each case is due to errors in fixing values of  $\phi$  at boundary nodes, and these are about the same with the 17-node molecule as with the 13-node molecule.

It is intended that further test cases be computed, for comparing the accuracy of the solution and the S.O.R. rate of convergence of equations based on this 17-node molecule with those for the conventional 13-node molecule.

#### 6. Acknowledgements

The author wishes to thank Mr. K. L. Stewart for advice and frequent discussions on this work, Mr. C. Sheffield (now with Computer Usage Co. Inc., New York) who gave the first complete proof that the eigenvalues of F are non-negative; and Dr. D. M. Parkyn and Mr. M. Batty for assistance in writing the TOPIC programs.

This paper has been written as part of the research program of the Mechanical Engineering Laboratory, English Electric Co. Ltd., Whetstone. The paper is published by permission of the English Electric Co. Ltd.

#### References

CARRÉ, B. A. (1961). "The Determination of the Optimum Accelerating Factor for Successive Over-Relaxation," The Computer Journal, Vol. 4, pp. 73-78.

COLLATZ, L. (1960). The Numerical Treatment of Differential Equations (Translated by P. G. Williams), Third Edition, Berlin:

ENGELI, M., GINSBURG, TH., RUTISHAUSER, H., and STIEFEL, E. (1959). Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-Adjoint Boundary-Value Problems, Basle: Birkhauser.

FADDEEVA, V. N. (1959). Computational Methods of Linear Algebra (Translated by C. D. Benster), New York: Dover.

FORSYTHE, G. E., and WASOW, W. R. (1960). Finite-Difference Methods for Partial Differential Equations, New York: Wiley. Fox, L. (1950). "The Numerical Solution of Elliptic Differential Equations when the Boundary Conditions Involve a Derivative," Phil. Trans, A, Vol. 242, pp. 345-78.

Fox, L. (1957). The Numerical Solution of Two-Point Boundary Problems in Ordinary Differential Equations," Oxford: O.U.P.

Fox, L. (1962) (Editor). Numerical Solution of Ordinary and Partial Differential Equations, Oxford: Pergamon. Frankel, S. P. (1950). "Convergence Rates of Iterative Treatments of Partial Differential Equations," MTAC, Vol. 4, pp. 65-75. HELLER, J. (1960). "Simultaneous, Successive and Alternating Direction Schemes," J.S.I.A.M., Vol. 8, pp. 150-73.

MARCUS (1960). Basic Theorems in Matrix Theory, N.B.S., A.M.S. 57, Washington: N.B.S.

OSTROWSKI (1954). "On the Linear Iteration Procedures for Symmetric Matrices," Univ. Roma. Ist. Naz. Alta Mat. Rend. Mat. e Appl. (5), Vol. 13, pp. 140-63.

VARGA, R. S. (1959). "Orderings of the Successive Over-Relaxation Scheme," Pacific J. Math., Vol. 9, pp. 617-28. Young, D. (1954). "Iterative Methods for Solving Partial Difference Equations of Elliptic Type," Trans. Am. Math. Soc., Vol. 76, pp. 92-111.