J. Inst. Maths Applics (1967) **4**, 295–314

The Use of Pre-conditioning in Iterative Methods for Solving Linear Equations with Symmetric Positive Definite Matrices

D. J. EVANS

University Computing Laboratory,
Department of Applied Mathematics, University of Sheffield

[Received 14 January, 1966 and in revised form 20 June, 1967]

The asymptotic convergence rates of many standard iterative methods for the solution of linear equations can be shown to depend inversely on the *P*-condition number of the coefficient matrix. The notion of minimizing the *P*-condition number and hence maximizing the convergence rate by the introduction of a new pre-conditioning factor is shown to be computationally feasible. The application of this idea to the method of Simultaneous Displacement, Richardson's method and other iterative methods, are discussed and numerical examples given to illustrate its effectiveness.

1. Introduction

IN the numerical solution of boundary value problems involving discrete approximations to elliptic partial differential equations the problem of solving iteratively large systems of linear equations is important. We now consider the application of some of these iterative methods applied to the system

$$\mathbf{Ax} = \mathbf{b}, \tag{1}$$

i.e.

$$\sum_i a_{ij}x_j = b_i, \qquad 1 \leqslant i, \quad j \leqslant N,$$

where **A** is a real symmetric positive definite matrix, $b_i (i = 1,2,...,N)$ are known, and $x_i (i = 1,2,...,N)$ are unknown quantities.

We may, without loss of generality, set all the diagonal coefficients $a_{ii} = 1$; then

$$\mathbf{A} = -\mathbf{L}+\mathbf{I}-\mathbf{U} \tag{2}$$

where **L** and **U** are respectively lower and upper triangular matrices with null diagonals and **I** is the unit matrix of order N.

The fundamental equation for the basic iterative method, i.e. the Jacobi method is,

$$\mathbf{x}^{(n+1)} = (\mathbf{L}+\mathbf{U})\mathbf{x}^{(n)}+\mathbf{b}, \tag{3}$$

where the superscript n denotes the iteration cycle. Alternatively, it can be written in the form

$$\mathbf{x}^{(n+1)} = (\mathbf{I}-\mathbf{A})\mathbf{x}^{(n)}+\mathbf{b} = \mathbf{x}^{(n)}+\mathbf{r}^{(n)}, \tag{4}$$

showing that the change in each component is equal to the corresponding component of the residual vector.

Two forms of the above equation (4) in which a constant factor α or a different choice α_n for each iteration is multiplied by each component of the residual vector,

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

and then added to each component of the present iterate $\mathbf{x}^{(n)}$ give rise to the equations

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} + \alpha \mathbf{r}^{(n)} = (\mathbf{I} - \alpha \mathbf{A})\mathbf{x}^{(n)} + \alpha \mathbf{b}, \tag{5}$$

and

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} + \alpha_n \mathbf{r}^{(n)} = (\mathbf{I} - \alpha_n \mathbf{A})\mathbf{x}^{(n)} + \alpha_n \mathbf{b}, \tag{6}$$

which are the well known Simultaneous Displacement and Richardson's method respectively. (Young, 1954.)

Briefly, in the method of Simultaneous Displacement, i.e. equation (5), if we denote the error vector $\mathbf{e}^{(n)}$ by

$$\mathbf{e}^{(n)} = \mathbf{x}^{(n)} - \mathbf{A}^{-1}\mathbf{b},$$

then the error vector can be shown to satisfy the equation

$$\mathbf{e}^{(n+1)} = (\mathbf{I} - \alpha \mathbf{A})\mathbf{e}^{(n)} = (\mathbf{I} - \alpha \mathbf{A})^{n+1}\mathbf{e}^{(0)},$$

and, since the error operator $(\mathbf{I} - \alpha \mathbf{A})$ is kept constant throughout the iteration, then the method is classified as a stationary linear iterative process. For the iteration to converge, the criteria necessary is that the modulus of the largest eigenvalue, or spectral radius, of $(\mathbf{I} - \alpha \mathbf{A})$ is less than unity.

If we assume that the spectrum of real eigenvalues λ_i of $\mathbf{A}$ are bounded by the values a and b such that

$$0 < a \leqslant \lambda_i \leqslant b < \infty, \quad (i = 1, 2, \ldots N) \tag{7}$$

then the criteria for convergence makes it necessary for

$$|1 - \alpha \lambda_i| \leqslant 1, \tag{8}$$

which gives the permissible range for values of α to be

$$0 < \alpha < \frac{2}{b}. \tag{9}$$

The fastest convergence rate is obtained by choosing α so that the spectral radius of $(\mathbf{I} - \alpha \mathbf{A})$ is minimized. Clearly the best choice of α is the one for which

$$1 - \alpha a = -(1 - \alpha b),$$

i.e.

$$\alpha = \frac{2}{a + b}. \tag{10}$$

With this choice of α, the convergence factor or spectral radius is, for both $i = 1$ and N,

$$|1 - \alpha \lambda_i| \leqslant \frac{b - a}{b + a} = \frac{P - 1}{P + 1}, \tag{11}$$

where $P = b/a$ is the P-condition number of $\mathbf{A}$ and is defined as the ratio of the maximum eigenvalue to minimum eigenvalue of a positive definite matrix.

We further define the rate of convergence R by the formula

$$R = -\log_n |\gamma|, \tag{12}$$

where γ is the spectral radius of the error operator $(I - \alpha A)$ and is given by (11). For the method of Simultaneous Displacement, we obtain the rate of convergence from (11) and (12) and it is given by

$$R \doteq \frac{2}{P}. \tag{13}$$

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

Similarly, for Richardson's method, we have

$$\mathbf{e}^{(n+1)} = \mathbf{x}^{(n+1)} - \mathbf{x} = (\mathbf{I} - \alpha_n \mathbf{A})\mathbf{e}^{(n)} = \prod_{i=0}^{n} (\mathbf{I} - \alpha_i \mathbf{A})\mathbf{e}^{(0)} = \mathbf{Q}_{n+1}(\mathbf{A})\mathbf{e}^{(0)}, \tag{14}$$

and since the error operator changes for each iteration, this is a non-stationary linear iterative process.

If the N eigenvalues of $\mathbf{A}$, λ_i with corresponding eigenvectors V_i form a basis for the space, then since the eigenvalues and eigenvectors of $\mathbf{Q}_{n+1}(\mathbf{A})$ are $Q_{n+1}(\lambda_i)$ and V_i, respectively, we obtain the result

$$\mathbf{e}^{(n+1)} = \mathbf{Q}_{n+1}(\mathbf{A})\mathbf{e}^{(0)} = \sum_{i=1}^{N} a_i \mathbf{Q}_{n+1}(\mathbf{A})V_i = \sum_{i=1}^{N} a_i \mathbf{Q}_{n+1}(\lambda_i)V_i.$$

Clearly for the error vector $\mathbf{e}^{(n+1)}$ to be small we need the value of $|\mathbf{Q}_{n+1}(x)|$ to be small over the entire interval $[a,b]$, under the constraint $\mathbf{Q}_{n+1}(0) = 1$. Such a polynomial has been given by Markoff (1916) and is

$$|\mathbf{Q}_{n+1}(\mathbf{x})| = \frac{T_{n+1}\left(\dfrac{b+a-2x}{b-a}\right)}{T_{n+1}\left(\dfrac{b+a}{b-a}\right)}, \tag{15}$$

where $T_{n+1}(x) = \cos[(n+1)\cos^{-1}(x)]$, is the Chebyshev polynomial of degree $(n+1)$ over the interval $[-1, 1]$. The α_i are chosen so that the roots of $\prod_{i=0}^{n} (1-\alpha_i x)$ are coincident with the roots of $T_{n+1}\{(b+a-2x)/(b-a)\}$. The details concerning the choice of α_i are given by Forsythe & Wasow (1960).

The maximum value of $\mathbf{Q}_{n+1}(\mathbf{x})$ as given by (15) for $n \to \infty$ (since the maximum absolute value of the numerator is unity), is

$$\max_{a \leqslant x \leqslant b} |\mathbf{Q}_{n+1}(\mathbf{x})| = [T_{n+1}(z)]^{-1}, \quad \text{where} \quad z = \frac{b+a}{b-a}, \tag{16}$$

and using the relationship for $T_{n+1}(z)$, $z > 1$ given by

$$2T_{n+1}(z) = [z + \sqrt{(z^2-1)}]^{n+1} + [z - \sqrt{(z^2-1)}]^{n+1},$$

we see that (16) simplifies to

$$\max_{a \leqslant x \leqslant b} |\mathbf{Q}_{n+1}(\mathbf{x})| \leqslant \frac{2}{[z + \sqrt{(z^2-1)}]^{n+1}},$$

and it follows that the eigenvalues $|\mathbf{Q}_{n+1}(\lambda_i)|$ are uniformly bounded as $n \to \infty$. The asymptotic bound to the convergence factor is

$$[z - \sqrt{(z^2-1)}]$$

and hence the rate of convergence as $n \to \infty$ is

$$R = \log [z + \sqrt{(z^2-1)}]. \tag{17}$$

For substantially large problems, we have

$$z \doteq 1 + \frac{2}{P},$$

and the final result for the rate of convergence is

$$R \doteq \frac{2}{\sqrt{P}}. \tag{18}$$

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

Methods involving two previous iterates by way of a three term recurrence relationship are well known. The second order Richardson's method, for example, is defined by the formula

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} + \alpha(\mathbf{b} - \mathbf{A}\mathbf{x}^{(n)}) + \beta(\mathbf{x}^{(n)} - \mathbf{x}^{(n-1)}), \tag{19}$$

where α and β remain constant throughout the iteration and are chosen to provide maximum convergence to the solution.

The theoretical justification for the choice of α and β has been given by Frankel (1950). For brevity, we state the final result. The optimum values are

$$\alpha = \left[\frac{2}{\sqrt{a}+\sqrt{b}}\right]^2 \quad \text{and} \quad \beta = \left[\frac{\sqrt{a}-\sqrt{b}}{\sqrt{a}+\sqrt{b}}\right]^2, \tag{20}$$

the spectral radius is $\sqrt{\beta}$ and the rate of convergence is again

$$R \doteq \frac{2}{\sqrt{P}}. \tag{21}$$

The Chebyshev acceleration of (6) has the form

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} + \alpha_n(\mathbf{b} - \mathbf{A}\mathbf{x}^{(n)}) + \beta_n(\mathbf{x}^{(n)} - \mathbf{x}^{(n-1)}), \tag{22}$$

where the parameters α_n and β_n vary with each iteration and are defined by Stiefel (1958) to be of the form

$$\alpha_n = \frac{4T_n(z)}{(b-a)T_{n+1}(z)} \quad \text{and} \quad \beta_n = \frac{T_{n-1}(z)}{T_{n+1}(z)}. \tag{23}$$

Since the coefficients α_n and β_n are less than unity, round off difficulties do not arise as they do in (6) due to the later values of α_i ($i = 1, 2, \ldots m$) when the number of parameters m is large and hence the method given by equations (22) is to be preferred over equation (6). Other advantages such as the elimination of the need to preselect m, the number of parameters such that after m applications of (15) the desired accuracy is attained and the precalculation of the α_i outweigh some of the disadvantages of the necessity to retain two iterates. Since the minimizing polynomial for the iteration (22) is identical to that given by (15), it follows immediately that the rate of convergence of this method is equal to the result obtained in (18). (Stiefel, 1958.)

Finally, the Chebychev semi-iterative method (Golub & Varga, 1961) can be derived from (22) if we assume that a and b have values $1-\mu$, $1+\mu$, respectively, where μ is the spectral radius of the error operator of the Jacobi method. The recurrence relationship for the Chebychev polynomials then becomes

$$\alpha_n = 1 + \beta_n, \tag{24}$$

and the iteration takes the particularly simple form

$$\mathbf{x}^{(n+1)} = \alpha_n[\mathbf{b} + (\mathbf{L}+\mathbf{U})\mathbf{x}^{(n)} - \mathbf{x}^{(n-1)}] + \mathbf{x}^{(n-1)},$$

or in standard form

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} + \alpha_n(\mathbf{b} - \mathbf{A}\mathbf{x}^{(n)}) + \alpha_{n-1}(\mathbf{x}^{(n)} - \mathbf{x}^{(n-1)}). \tag{25}$$

The parameters α_n are calculated from the formulae

$$\alpha_1 = 1, \qquad \alpha_2 = \frac{2}{2-\mu^2}, \qquad \alpha_{n+1} = \frac{1}{(1-\mu^2\alpha_{n/4})}, \tag{26}$$

for $n = 2, 3, \ldots$, whilst the rate of convergence is generally the same.

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

2. Minimization of *P*-condition Number of Coefficient Matrix

For the two classes of methods discussed in the previous sections it is clear that the optimal speed of convergence even with the correct choice of parameters α or α_i is limited by the value of P, i.e. the P-condition number of the coefficient matrix $\mathbf{A}$. Any attempt to improve these basic fundamental methods must clearly apply some form of pre-conditioning to the original equations, in order to minimize the P-condition number and hence increase the rate of convergence. We attempt to do this in the following manner.

Let y be an intermediate transformation vector given by

$$\mathbf{y} = (\mathbf{I}-\omega\mathbf{U})\mathbf{x}, \tag{27}$$

where ω is an acceleration parameter to be defined later. By pre-multiplying equation (1) by $(\mathbf{I}-\omega\mathbf{L})^{-1}$, we obtain the equation

$$(\mathbf{I}-\omega\mathbf{L})^{-1}\mathbf{A}(\mathbf{I}-\omega\mathbf{U})^{-1}[(\mathbf{I}-\omega\mathbf{U})\mathbf{x}] = (\mathbf{I}-\omega\mathbf{L})^{-1}\mathbf{b},$$

which can be written as

$$\mathbf{G}^T\mathbf{A}\mathbf{G}\mathbf{y} = \mathbf{d}, \tag{28a}$$

or

$$\mathbf{B}\mathbf{y} = \mathbf{d}, \tag{28b}$$

where $\mathbf{G} = (\mathbf{I}-\omega\mathbf{U})^{-1}$, $\mathbf{d} = (\mathbf{I}-\omega\mathbf{L})^{-1}\mathbf{b}$ and the superscript T denotes the transpose of a matrix in the usual way.

It can be noticed that the matrix $\mathbf{G}$ is in a form which is readily inverted. By this we mean that the evaluation of products like $\mathbf{z} = \mathbf{G}\mathbf{y}$ where $\mathbf{y}$ is a known vector can be obtained by the solution of sets of equations like $(\mathbf{I}-\omega\mathbf{U})\mathbf{z} = \mathbf{y}$ which because of the form of the matrix is simply a back substitution process. Similarly, the evaluation of $\mathbf{z} = \mathbf{G}^T\mathbf{y}$ is a forward substitution process involving only the terms in the lower triangular part of the matrix $\mathbf{A}$. Hence it can be noticed that the sparseness of $\mathbf{A}$ is retained in $(\mathbf{G}^T)^{-1}$ and $(\mathbf{G})^{-1}$ for they are simply the lower and upper triangular matrices as contained in the original matrix $\mathbf{A}$.

Furthermore, since $\mathbf{G}$ is a non-singular matrix and $\mathbf{A}$ is positive definite and symmetric by definition, it possesses a positive definite square root matrix $\mathbf{A}^{\frac{1}{2}}$ which is also symmetric. Then

$$\mathbf{G}^T\mathbf{A}\mathbf{G} = (\mathbf{G}^T\mathbf{A}^{\frac{1}{2}})(\mathbf{A}^{\frac{1}{2}}\mathbf{G}), \tag{29}$$

from which it follows immediately that $\mathbf{G}^T\mathbf{A}\mathbf{G}$ is a positive definite, symmetric matrix.

We define the optimum acceleration parameter $\bar{\omega}$ as that value of ω in which the ratio of largest to smallest eigenvalue $\Lambda_1(\omega)/\Lambda_N(\omega)$ of matrix $\mathbf{G}^T\mathbf{A}\mathbf{G}$ is minimized.

Further we define the matrix $\mathbf{T} \equiv (\mathbf{I}-\omega\mathbf{U})^{-1}(\mathbf{I}-\omega\mathbf{L})^{-1}\mathbf{A}$ and show that by the similarity transformation

$$(\mathbf{I}-\omega\mathbf{U})\mathbf{T}(\mathbf{I}-\omega\mathbf{U})^{-1} \equiv \mathbf{G}^T\mathbf{A}\mathbf{G}, \tag{30}$$

from which it follows that $\mathbf{T}$ possesses the same eigenvalues $\Lambda_i(\omega)$ $(i = 1,2,...,N)$ as the matrix $\mathbf{G}^T\mathbf{A}\mathbf{G}$ but different eigenvectors. Hence we have

$$(\mathbf{I}-\omega\mathbf{U})^{-1}(\mathbf{I}-\omega\mathbf{L})^{-1}\mathbf{A}\mathbf{u}_i = \Lambda_i(\omega)\mathbf{u}_i, \tag{31}$$

where each $\mathbf{u}_i$ $(i = 1,2,...,N)$ is an eigenvector of $\mathbf{T}$. By premultiplying equation (31) by $\mathbf{u}_i^T(\mathbf{I}-\omega\mathbf{L})(\mathbf{I}-\omega\mathbf{U})$, we obtain the result

$$\Lambda_i(\omega) = \frac{\tau_i}{(1-\omega+\omega\tau_i+\omega^2 k_i)}, \tag{32}$$

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

where $\mathbf{u}_i^T\mathbf{A}\mathbf{u}_i = \boldsymbol{\tau}_i$ and $\mathbf{u}_i^T\mathbf{L}\mathbf{U}\mathbf{u}_i = \mathbf{k}_i$. Hence the P-condition number of the matrix $\mathbf{G}^T\mathbf{A}\mathbf{G}$ is defined as

$$P(\omega) = \frac{\Lambda_1(\omega)}{\Lambda_N(\omega)} = \frac{\tau_1(1-\omega+\omega\tau_N+\omega^2 k_N)}{\tau_N(1-\omega+\omega\tau_1+\omega^2 k_1)}. \tag{33}$$

For the function $P(\omega)$ to possess a minimum value, we must have

$$\frac{d[P(\omega)]}{d\omega} = 0. \tag{34}$$

On differentiating (33) and clearing terms, we finally obtain the quadratic equation

$$c\omega^2 + d\omega + f = 0. \tag{35}$$

where $c = k_N\tau_1 - k_1\tau_N - k_N + k_1 d = 2(k_N - k_1)$ and $f = \tau_N - \tau_1$, for the optimum pre-conditioning parameter $\bar{\omega}$.

The optimum pre-conditioning parameter $\bar{\omega}$ is thus given by

$$\bar{\omega} = \frac{(\tau_N - \tau_1)}{(k_1 - k_N) - \{(k_1 - k_N)^2 - (\tau_N - \tau_1)(k_N\tau_1 - k_1\tau_N - k_N + k_1)\}^{\frac{1}{2}}} \tag{36}$$

and the minimum P-condition number by

$$\bar{P} = \frac{\tau_1(1-\bar{\omega}+\bar{\omega}\tau_N+\bar{\omega}^2 k_N)}{\tau_N(1-\bar{\omega}+\bar{\omega}\tau_1+\bar{\omega}^2 k_1)}. \tag{37}$$

The eigenvalue spectrum of the matrix $\mathbf{G}^T\mathbf{A}\mathbf{G}$ is bounded by the values $\bar{a}$ and $\bar{b}$ such that

$$0 < \bar{a} \leqslant \Lambda_i(\bar{\omega}) \leqslant \bar{b} \quad (i = 1,2,\ldots,N) \tag{38}$$

where

$$\bar{a} \leqslant \frac{\tau_N}{(1-\bar{\omega}+\bar{\omega}\tau_N+\bar{\omega}^2 k_N)}, \tag{39a}$$

and

$$\bar{b} \geqslant \frac{\tau_1}{(1-\bar{\omega}+\bar{\omega}\tau_1+\bar{\omega}^2 k_1)}. \tag{39b}$$

Thus we have derived expressions for the eigenvalue spectrum, minimum P-condition number, and the optimum pre-conditioning factor in terms of the quantities τ_1, τ_N, k_1 and k_N.

3. Theoretical Considerations on the Minimization of the *P*-condition Number

In this section we shall present some theoretical considerations to justify the procedures proposed in Section 2. Why do we expect the P-condition number to be reduced and minimized for some values of $\omega > 0$?

We now assume that the linear system (1) has been derived from simple finite difference approximations to an elliptic partial differential equation based on a rectangular mesh of grid lines, the model example being considered is the solution of the boundary value problem

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

over the unit square with the boundary conditions $u(0,y) = u(1,y) = 0$, $u(x,1) = 0$, $u(x,0) = 1$.

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

We choose a square grid of p rows and p columns parallel to the boundaries of the domain as our mesh with spacings $\Delta x = \Delta y = h$, and with the aid of central difference operators the finite difference equation at the point (i, j) on the mesh of $(p+1)(p+1)$ points is

$$-l_{i,j}u_{i-1,j} - r_{i,j}u_{i+1,j} - t_{i,j}u_{i,j+1} - b_{i,j}u_{i,j-1} + u_{i,j} = d_{i,j}, \tag{40}$$

for $i = 1(1)p$ and $j = 1(1)p$.

For a columnwise ordering of the grid points the coefficient matrix $\mathbf{A}$ for the finite difference equations has been shown to be positive definite and symmetric with the following tridiagonal partitioned form of order p with submatrix elements of order p (Varga, 1962), i.e.

$$\mathbf{A} = \begin{bmatrix} D_1 & -U_1 & & & \\ -L_1 & D_2 & -U_2 & 0 & \\ & \ddots & \ddots & \ddots & \\ & 0 & -L_{p-2} & D_{p-1} & -U_{p-1} \\ & & & -L_{p-1} & D_p \end{bmatrix}, \tag{41}$$

where

$$\mathbf{D}_i = \begin{bmatrix} 1 & -t_{i,1} & & & \\ -b_{i,2} & 1 & -t_{i,2} & 0 & \\ & \ddots & \ddots & \ddots & \\ & 0 & -b_{i,p-1} & 1 & -t_{i,p-1} \\ & & & -b_{i,p} & 1 \end{bmatrix},$$

$$\mathbf{L}_{i-1} = \begin{bmatrix} l_{i,1} & & & \\ & l_{i,2} & 0 & \\ & 0 & \ddots & \\ & & & l_{i,p} \end{bmatrix} \quad \text{and} \quad \mathbf{U}_i = \begin{bmatrix} r_{i,1} & & & \\ & r_{i,2} & 0 & \\ & & \ddots & \\ & 0 & & r_{i,p} \end{bmatrix}.$$

Such matrices are derived from equations said to have σ_2-ordering and property A. Young (1954) has shown that the finite difference equations can be re-ordered by choosing all the points in which $(i+j)$ is even, and then all the points in which $(i+j)$ is odd so that the coefficient matrix $\mathbf{A}$ has the simple partitioned form

$$\begin{bmatrix} \mathbf{I}_1 & -\mathbf{U}^* \\ -\mathbf{L}^* & \mathbf{I}_2 \end{bmatrix}, \tag{42}$$

where $\mathbf{L}^* = \mathbf{U}^{*T}$.

Such matrices are said to possess property A and the equations to have σ_1-ordering.

We now proceed to the following theorem concerning the minimization of the P-condition number of the matrix $\mathbf{B} = (\mathbf{I}-\omega\mathbf{L})^{-1}\mathbf{A}(\mathbf{I}-\omega\mathbf{U})^{-1}$ where $\mathbf{A}$ is defined to have the form $\mathbf{I}-\mathbf{L}-\mathbf{U}$ and ω to be the pre-conditioning factor.

THEOREM 1. *The P-condition number of the matrix* $\mathbf{B} = (\mathbf{I}-\omega\mathbf{L})^{-1}\mathbf{A}(\mathbf{I}-\omega\mathbf{U})^{-1}$ *is minimized when the value of the pre-condition factor* ω *is unity for all matrices* $\mathbf{A}$ *possessing property A and* σ_1 *ordering.*

Proof. Let us assume that $\mathbf{A}$ has the form

$$\mathbf{A} = \mathbf{I}-\mathbf{L}-\mathbf{U},$$

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

and by virtue of the σ_1 ordering of the equations,

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}_1 & -\mathbf{U}^* \\ -\mathbf{L}^* & \mathbf{I}_2 \end{bmatrix}, \tag{43}$$

where

$$\mathbf{L} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{L}^* & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \mathbf{U} = \begin{bmatrix} \mathbf{0} & \mathbf{U}^* \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{44}$$

Suppose $\mathbf{U}^*$ is a $(m \times r)$ submatrix, $\mathbf{L}^*$ is a $(r \times m)$ submatrix, $\mathbf{I}_1$, $\mathbf{I}_2$ are $(m \times m)$ and $(r \times r)$ identity submatrices, respectively and $m+r = p^2$. Earlier in Section 2, the transformation

$$\mathbf{B} = (\mathbf{I}-\omega\mathbf{L})^{-1}\mathbf{A}(\mathbf{I}-\omega\mathbf{U})^{-1}$$

was proposed, and when matrix $\mathbf{A}$ has property A and σ_1 ordering we have immediately

$$\mathbf{B} = (\mathbf{I}+\omega\mathbf{L})\mathbf{A}(\mathbf{I}+\omega\mathbf{U}), \tag{45}$$

which on simplification gives the result

$$\begin{aligned} \mathbf{B} &= \begin{bmatrix} \mathbf{I}_1 & \mathbf{0} \\ \omega\mathbf{L}^* & \mathbf{I}_2 \end{bmatrix} \begin{bmatrix} \mathbf{I}_1 & -\mathbf{U}^* \\ -\mathbf{L}^* & \mathbf{I}_2 \end{bmatrix} \begin{bmatrix} \mathbf{I}_1 & \omega\mathbf{U}^* \\ \mathbf{0} & \mathbf{I}_2 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I}_1 & \omega\mathbf{U}^*-\mathbf{U}^* \\ \omega\mathbf{L}^*-\mathbf{L}^*, & \mathbf{I}_2+\omega^2\mathbf{L}^*\mathbf{U}^*-2\omega\mathbf{L}^*\mathbf{U}^* \end{bmatrix}. \end{aligned} \tag{46}$$

If Λ is an eigenvalue of $\mathbf{B}$ and $\mathbf{y} = \begin{vmatrix} \mathbf{c} \\ \mathbf{d} \end{vmatrix}$ is the corresponding eigenvector, the partitions of $\mathbf{y}$ corresponding to the partitions of $\mathbf{A}$ in (43), then from the statement of the eigenvalue problem, we have

$$\begin{bmatrix} \mathbf{I}_1 & \omega\mathbf{U}^*-\mathbf{U}^* \\ \omega\mathbf{L}^*-\mathbf{L}^* & \mathbf{I}_2+\omega^2\mathbf{L}^*\mathbf{U}^*-2\omega\mathbf{L}^*\mathbf{U}^* \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix} = \Lambda\mathbf{I}\begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}, \tag{47}$$

which simplifies to

$$\mathbf{I}_1\mathbf{c}+(\omega\mathbf{U}^*-\mathbf{U}^*)\mathbf{d} = \Lambda\mathbf{I}_1\mathbf{c} \tag{48a}$$

$$(\omega\mathbf{L}^*-\mathbf{L}^*)\mathbf{c}+(\mathbf{I}_2+\omega^2\mathbf{L}^*\mathbf{U}^*-2\omega\mathbf{L}^*\mathbf{U}^*)\mathbf{d} = \Lambda\mathbf{I}_2\mathbf{d}. \tag{48b}$$

Eliminating $\mathbf{c}$ from (48a) and (48b) gives the result

$$[(1-\Lambda)(\mathbf{I}_2-2\omega\mathbf{L}^*\mathbf{U}^*+\omega^2\mathbf{L}^*\mathbf{U}^*-\Lambda\mathbf{I}_2)-\mathbf{L}^*\mathbf{U}^*(\omega-1)^2]\mathbf{d} = 0. \tag{49}$$

It is easily shown that the non-zero eigenvalues of $(\mathbf{L}+\mathbf{U})$ occur in pairs $\pm\mu_i$ $(i = 1,2,...,r)$, where r is less than or equal to the number of rows in $\mathbf{L}^*$ or $\mathbf{U}^*$. Furthermore, the eigenvalues of $\mathbf{L}^*\mathbf{U}^*$ are precisely μ_i^2 $(i = 1,2,...,r)$ or zero.

Therefore, since $\mathbf{d} \neq 0$, we must have

$$\Lambda^2+\Lambda(2\omega\mu_i^2-2-\omega^2\mu_i^2)+(1-\mu_i^2) = 0, \tag{50}$$

and

$$\Lambda = (1+0{\cdot}5\omega^2\mu_i^2-\omega\mu_i^2)\pm[(1+0{\cdot}5\omega^2\mu_i^2-\omega\mu_i^2)^2-(1-\mu_i^2)]^{\frac{1}{2}} \tag{51}$$

for $(i = 1,2,...,r)$.

Now, the P-condition number of matrix $\mathbf{B}$ is given by the expression

$$P(\omega) = \frac{\Lambda_1}{\Lambda_N}, \tag{52}$$

and is obtained by taking $i = 1$ and both choices of sign in (51) to give the result

$$P(\omega) = \frac{(1+0{\cdot}5\omega^2\mu_1^2-\omega\mu_1^2)+[(1+0{\cdot}5\omega^2\mu_1^2-\omega\mu_1^2)^2-(1-\mu_1^2)]^{\frac{1}{2}}}{(1+0{\cdot}5\omega^2\mu_1^2-\omega\mu_1^2)-[(1+0{\cdot}5\omega^2\mu_1^2-\omega\mu_1^2)^2-(1-\mu_1^2)]^{\frac{1}{2}}}, \tag{53}$$

where $\mu_r < \mu_{r-1} < ... < \mu_2 < \mu_1$.

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

The quantity $P(\omega)$ has a stationary value when $dP/d\omega = 0$, i.e. when

$$4(\bar{\omega}-1)[\mu_1^2\bar{\omega}(\bar{\omega}-2)+2] = 0, \tag{54}$$

which gives the value $\bar{\omega} = 1$ or two complex values of ω which we shall consider inadmissible.

Finally, from (54) it can be easily ascertained that the value of $dP/d\omega$ for $\omega = 1-\varepsilon$ is

$$-4\varepsilon[2-(1-\varepsilon^2)\mu_1^2], \tag{55}$$

and for $\omega = 1+\varepsilon$ is

$$4\varepsilon[2-(1-\varepsilon^2)\mu_1^2]. \tag{56}$$

Thus the sign of $dP/d\omega$ goes from negative to positive as ω goes from $1-\varepsilon$ to $1+\varepsilon$, and proves conclusively that P has a minimum value at $\omega = 1$.

The minimum value of P for the optimum pre-conditioning factor $\omega = 1$ is then

$$P_{\omega=1} = \frac{1}{1-\mu_1^2}, \tag{57}$$

and hence we have proved the theorem for σ_1 ordering.

We would like to be able to apply a similar rigorous analysis to prove the existence of a minimum P-condition number for the matrix $\mathbf{B} = (\mathbf{I}-\omega\mathbf{L})^{-1}\mathbf{A}(\mathbf{I}-\omega\mathbf{U})^{-1}$, where the finite difference equations (40) possess σ_2 ordering the matrix $\mathbf{A}$ now assumes the tridiagonal partitioned form given in (41). However, the mathematical analysis becomes complicated even for the case $p = 3$. Alternatively, we can consider in detail the simple model problem and derive upper and lower bounds for the eigenvalues of $\mathbf{B}$ for $\omega > 0$. For this case, the coefficients of the finite difference equations are all constant so that for each point (i,j)

$$l_{i,j} = b_{i,j} = r_{i,j} = t_{i,j} = \tfrac{1}{4}.$$

Then the coefficient matrix $\mathbf{A}$ assumes the simple form after normalization, i.e.

$$\mathbf{A} = \begin{bmatrix} \mathbf{I} & -\mathbf{X} & & \\ -\mathbf{X} & \mathbf{I} & -\mathbf{X} & \mathbf{0} \\ & \ddots & \ddots & \ddots \\ \mathbf{0} & & -\mathbf{X} & \mathbf{I} \end{bmatrix} \tag{58}$$

where $\mathbf{I}$ is the identity matrix of order p and

$$\mathbf{X} = \mathbf{D}_i^{-\frac{1}{4}}\mathbf{L}_i\mathbf{D}_i^{-\frac{1}{4}} = \mathbf{D}_i^{-\frac{1}{4}}\mathbf{U}_i\mathbf{D}_i^{-\frac{1}{4}}, \quad \text{for } i = 1(1)p,$$

and from the simple definition of matrix $\mathbf{B}$ given by equation (28), we have

$$\mathbf{B} = \begin{bmatrix} \mathbf{I} & & & \\ \omega\mathbf{X} & \mathbf{I} & & \mathbf{0} \\ \omega^2\mathbf{X}^2 & \omega\mathbf{X} & \ddots & \\ \ddots & \ddots & \ddots & \\ (\omega\mathbf{X})^{p-1} & \omega^2\mathbf{X}^2 & \omega\mathbf{X} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{X} & & \\ -\mathbf{X} & \mathbf{I} & -\mathbf{X} & \\ & \ddots & \ddots & \ddots \\ & & \ddots & -\mathbf{X} \\ & & -\mathbf{X} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \omega\mathbf{X} & \omega^2\mathbf{X}^2 & (\omega\mathbf{X})^{p-1} \\ & \mathbf{I} & \omega\mathbf{X} & \ddots \\ & & \ddots & \ddots \\ & & & \omega^2\mathbf{X}^2 \\ & & & \omega\mathbf{X} \\ & & & \mathbf{I} \end{bmatrix},$$

which reduces to the form

$$\begin{bmatrix} \mathbf{I}, & (\omega-1)\mathbf{X}, & \omega(\omega-1)\mathbf{X}^2, & \dots \\ (\omega-1)\mathbf{X}, & \mathbf{I}+\omega(\omega-2)\mathbf{X}^2, & \dots & \dots \\ \omega(\omega-1)\mathbf{X}^2, & \dots & \dots & \dots \\ \vdots & & & \end{bmatrix}, \tag{59}$$

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

or

$$\begin{bmatrix} 1 & , & x_k(\omega-1) & , & x_k^2\omega(\omega-1) & , & x_k^3\omega^2(\omega-1) & , & \cdots \\ x_k(\omega-1) & , & 1+x_k^2\omega(\omega-2) & , & x_k(\omega-1)+x_k^3\omega^2(\omega-2) & , & x_k^2\omega(\omega-1)+x_k^4\omega^3(\omega-2) & , & \cdots \\ x_k^2\omega(\omega-1), & & x_k(\omega-1)+x_k^3\omega^2(\omega-2), & & 1+x_k^2\omega(\omega-2)+x_k^4\omega^3(\omega-2) & , & x_k(\omega-1)+x_k^3\omega^2(\omega-2)+x_k^5\omega^4(\omega-2) & , & \cdots \\ x_k^3\omega^2(\omega-1), & & x_k^2\omega(\omega-1)+x_k^4\omega^3(\omega-2), & & x_k(\omega-1)+x_k^3\omega^2(\omega-2)+x_k^5\omega^4(\omega-2), & & 1+x_k^2\omega(\omega-2)+x_k^4\omega^3(\omega-2)+x_k^6\omega^5(\omega-2), & & \cdots \\ \cdot & & \cdot & & \cdot & & \cdot & & \cdots \\ \cdot & & \cdot & & \cdot & & \cdot & & \cdots \end{bmatrix} \tag{60}$$

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

To obtain the eigenvalues of the matrix **B**, we use a result due to Afriat (1954) which is stated as follows: let a $(p^2 \times p^2)$ matrix **B** have the property that it can be partitioned into blocks $\mathbf{B}_{ij} = \mathbf{R}_{ij}(\mathbf{X})$, $i, j = 1, 2, \ldots, p$ where $\mathbf{R}_{ij}$ is a rational function of the $(p \times p)$ matrix **X**, and let x_k be an eigenvalue of **X**, $k = 1(1)p$, then the eigenvalues of the $(p \times p)$ matrix with elements $R_{ij}(x_k)$ are the eigenvalues of the matrix **B**. Thus we have immediately that the eigenvalues of the matrix **B** given in (59) are the eigenvalues of the matrix given by (60), where x_k, $k = 1(1)p$ denote the eigenvalues of the matrix **X**, which we know assumes the simple form

$$x_k = \frac{0{\cdot}25}{1+0{\cdot}5\cos k\pi/(p+1)} \qquad k = 1, 2, \ldots p \tag{61}$$

from the definition of **X** given in equation (58).

Since the form of matrix **B** is extremely complex, it is doubtful whether the eigenvalues can be determined exactly. However, simple estimates for the largest and smallest eigenvalues and hence the *P*-condition number of the matrix **B** in terms of ω, the preconditioning factor, can be obtained with the use of the following two expressions:

A simple upper bound to the largest eigenvalue of the matrix **B** can be obtained by the application of the Gerschgorin circle theorem and is given by the quantity

$$\Lambda_1 \leqslant \max_{1 \leqslant i \leqslant p} \sum_{j=1}^{p} |b_{i,j}|, \tag{62}$$

whilst an estimate to the smallest eigenvalue has been given by Collatz (1949), and is

$$\Lambda_N \geqslant \min_{1 \leqslant i \leqslant p} \left([b_{i,i}] - \sum_{j=1,\, j \neq i}^{p} |b_{i,j}| \right). \tag{63}$$

We now apply the Gerschgorin estimate (62) to the matrix **B** as given by (59) and, in the range $2 > \omega \geqslant 0$ and $p \geqslant 1$, the largest row sum is always given by the first row. The result is

$$\Lambda_1(\omega) \leqslant 1 + x_p\{1 + \omega x_p + \ldots + (\omega x_p)^{(p-2)}\}\{|\omega - 1|\}, \tag{64}$$

where x_p denotes the largest eigenvalue of the matrix **X**, which from (61) can be shown to satisfy the inequality

$$x_p < 0{\cdot}5. \tag{65}$$

Finally, the expression given by (64) can clearly be seen to be minimized when

$$\omega = 1, \tag{66}$$

at which the largest eigenvalue attains its minimum value. Also by further scrutiny it is readily verified that $\Lambda(\omega)_1$ is a monotonic increasing function for ω in the range $1 < \omega < 2$, and monotonically decreasing in the range $0 < \omega < 1$.

The application of the estimate (63) for the minimum eigenvalue of **B** gives results which are, however, not so consistent. Firstly, the minimum value of the Collatz estimate is not always given by the same row as the order p of the matrix increases. The only definite conclusion which can be stated is that in the range $2 > \omega \geqslant 0$ it is never given by the first row. Thus, for example, when $p = 3$, the Collatz sum gives the result

$$\Lambda_N(\omega) \geqslant [1 + \omega(\omega-2)x_p^2\{1 + \omega^2 x_p^2\}] - |x_p^2\omega(\omega-1)| - |x_p(\omega-1) + x_p^3\omega^2(\omega-2)|, \tag{67}$$

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

which is obtained from the third or last row, whilst for $p = 4$, the result

$$\Lambda_N(\omega) \geqslant [1+\omega x_p^2(\omega-2)\{1+\omega^2 x_p^2\}]- |x_p^2\omega(\omega-1)|-|x_p(\omega-1)+x_p^3\omega^2(\omega-2)|-|x_p(\omega-1)+x_p^3\omega^2(\omega-2)+x_p^5\omega^4(\omega-2)| \quad (68)$$

follows from the third row, where x_p is again the largest eigenvalue of the matrix **X**. In general, numerical experiments indicate that the required result is obtained from the row given by the largest integer not greater than $1+p/2$.

For the case given by $p = 4$, i.e. equation (68), we can make the following simple observations. When

$$\omega = 0 \qquad \Lambda_N(0) \geqslant 1-2x_p,$$
$$\omega = 1 \qquad \Lambda_N(1) \geqslant 1-x_p^2-x_p^4-2x_p^3-x_p^5,$$
$$\omega = 2 \qquad \Lambda_N(2) \geqslant 1-2x_p-2x_p^2,$$

and for the value of x_p given by equation (65), we obtain

$$\Lambda_N(0) < \Lambda_N(1),$$
$$\Lambda_N(2) < \Lambda_N(1),$$

from which we can tentatively infer that Λ_N is maximized somewhere in the range $0 < \omega < 2$.

Furthermore, for values of ω in the range $0 < \omega < 1+\delta_p$, where δ_p is small and dependent on the order p of the matrices **X** involved, the value of the terms in each modulus of the expression (68) are all negative, and hence the contribution of each term is in the same direction and subtracted out in full from the diagonal entry in the chosen row. Hence no maximum value of $\Lambda_N(\omega)$ is likely to occur in this range. Alternatively, if we examine the range $1+\delta_p < \omega < 2$, we can envisage that the value of the terms in each modulus in (68) are all positive and once again the contribution of each term is in the same direction and subtracted out in full from the diagonal entry.

Hence, in the range $0 < \omega < 2$, a value of ω exists such that the modulus terms in the expression (68) are zero.

We now rewrite (68) in the more compact form

$$\Lambda_N(\omega) \geqslant [1+\omega x_p^2(\omega-2)\{1+\omega^2 x_p^2\}]- |(\omega-1)x_p\{2+\omega x_p\}+(\omega-2)\omega^2 x_p^3(2+\omega x_p+\omega^2 x_p^2)|,$$

and in general for large p, we can write

$$\Lambda_N(\omega) \geqslant [1+\omega x_p^2(\omega-2)F_1(\omega x_p)]-|(\omega-1)F_2(\omega x_p)+(\omega-2)F_3(\omega x_p)|, \quad (69)$$

where the Collatz estimate is applied to the chosen row and $F_1(\omega x_p)$, $F_2(\omega x)_p$, $F_3(\omega x_p)$ are rational functions of ωx_p.

Now $\Lambda_N(\omega)$ is maximized when ω is so chosen that the quantity

$$(\omega-1)F_2(\omega x_p)+(\omega-2)F_3(\omega x_p) \quad (70)$$

derived from the latter half of equation (69) is identically zero.

The exact value of ω for which this expression is zero depends upon the actual form of the rational functions F_1, F_2 and F_3 and is extremely difficult to find for general p. However, by a simple application of Descartes' rule of signs, this equation can be verified to have at least one real root in the range $1 < \omega < 2$.

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

Thus we have the final result that the estimate to the maximum eigenvalue Λ_1 of the matrix **B** is minimized when $\omega = 1$ and the estimate to the minimum eigenvalue Λ_N is maximized for some value of ω in the range $1 < \omega < 2$.

Finally, an upper bound to the P-condition number of the matrix **B** can be obtained from (62) and (63) viz.,

$$P \leqslant \frac{\Lambda_1(\omega)}{\Lambda_N(\omega)}, \tag{71}$$

and from the analysis derived in (64) and (69) and since $dP/d\omega$ is negative for $\omega = 1$ and positive for $\omega = 2$, it follows that there is a value of ω in the range $1 < \omega < 2$ for which P achieves a minimum value. This value of ω we term the optimum pre-conditioning parameter $\bar{\omega}$.

4. Application of the Pre-conditioning Technique to the Basic Iterative Methods

We can now apply the results of the previous section to modify the original basic iterative methods discussed in Section 1. It follows immediately that if the techniques used in Section 3 are shown to apply to the methods given by equations (5) and (6), then an improvement in convergence rate must follow.

We now proceed from equation (28) and develop similar iterative processes to the method of Simultaneous Displacement, Richardson's method and the second order methods by working throughout in the transformed variable y.

Denoting the matrix $\mathbf{G}^T\mathbf{A}\mathbf{G}$ as before by **B**, then equations describing iterative processes similar to (3), (5) and (6) can be written down immediately. They are

$$\mathbf{y}^{(n+1)} = \mathbf{y}^{(n)} + (\mathbf{d} - \mathbf{B}\mathbf{y}^{(n)}), \tag{72}$$

$$\mathbf{y}^{(n+1)} = \mathbf{y}^{(n)} + \alpha(\mathbf{d} - \mathbf{B}\mathbf{y}^{(n)}), \tag{73}$$

$$\mathbf{y}^{(n+1)} = \mathbf{y}^{(n)} + \alpha_n(\mathbf{d} - \mathbf{B}\mathbf{y}^{(n)}). \tag{74}$$

The iterations proceed in the y variable until a specified degree of accuracy is achieved. The final solution is then obtained by one application of the formula

$$\mathbf{x} = (\mathbf{I} - \omega\mathbf{U})^{-1}\mathbf{y}. \tag{75}$$

The iteration (72) converges when

$$|1 - \Lambda_i(\omega)| < 1 \quad (i = 1,2,\ldots,p^2). \tag{76}$$

This gives the range $0 < \omega < \omega_f$ for which the iteration method is convergent. For equation (73) and the optimum $\bar{\omega}$ specified in Section 3, the iteration will converge for a range of values of α given by

$$0 < \alpha < \frac{2}{\bar{b}}. \tag{77}$$

It will further possess an optimum value of α given by

$$\alpha = \frac{2}{\bar{a} + \bar{b}}, \tag{78}$$

and have a convergence rate approximately equal to

$$\frac{2\bar{a}}{\bar{b}}. \tag{79}$$

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

Similarly, for equation (74) the optimum values of α_i are given when the zeros of $(1-\alpha_i x)$ coincide with the zeros of the polynomial

$$T_{n+1}\left(\frac{\bar{b}+\bar{a}-2x}{\bar{b}-\bar{a}}\right).$$

This gives for the parameters α_i, the equation

$$\alpha_i = \frac{2}{\left[(\bar{a}+\bar{b})-(\bar{b}-\bar{a})\cos\dfrac{(2i-1)\pi}{2m}\right]}, \quad (i=1,2,\ldots m), \tag{80}$$

where m is the number of Richardson parameters chosen to achieve the desired accuracy. This method will then have a rate of convergence approximately equal to

$$2\sqrt{\frac{\bar{a}}{\bar{b}}}. \tag{81}$$

The second order Richardson method can similarly be written as

$$\mathbf{y}^{(n+1)} = \mathbf{y}^{(n)} + \alpha(\mathbf{d}-\mathbf{B}\mathbf{y}^{(n)}) + \beta(\mathbf{y}^{(n)}-\mathbf{y}^{(n-1)}), \tag{82}$$

where

$$\alpha = \left(\frac{2}{\sqrt{\bar{a}}+\sqrt{\bar{b}}}\right)^2 \quad \text{and} \quad \beta = \left(\frac{\sqrt{\bar{a}}-\sqrt{\bar{b}}}{\sqrt{\bar{a}}+\sqrt{\bar{b}}}\right)^2. \tag{83}$$

The Chebychev acceleration of this method also has the form

$$\mathbf{y}^{(n+1)} = \mathbf{y}^{(n)} + \alpha_n(\mathbf{d}-\mathbf{B}\mathbf{y}^{(n)}) + \beta_n(\mathbf{y}^{(n)}-\mathbf{y}^{(n-1)}), \tag{84}$$

where

$$\bar{Z} = \frac{(\bar{b}+\bar{a})}{(\bar{b}-\bar{a})}, \tag{85}$$

and α_n and β_n are given by an equation similar to equation (23) (but with a and b replaced by $\bar{a}$ and $\bar{b}$). The rate of convergence of this method is given by (21) where

$$P(\bar{\omega}) = \frac{\bar{b}}{\bar{a}}. \tag{86}$$

Earlier in this section the iteration

$$\tilde{\mathbf{y}}^{(n+1)} = \mathbf{y}^{(n)} + \alpha(\mathbf{d}-\mathbf{B}\mathbf{y}^{(n)}) \tag{87}$$

was discussed. Now when α is given by (78), the error operator of (87) possesses equal and opposite eigenvalues of magnitude $\bar{Z}$ as given by (85). Since the eigenvalues are all real, this method can be further accelerated by Chebychev polynomials.

The result $\tilde{\mathbf{y}}^{(n+1)}$ from equation (87) is now used in the formula

$$\mathbf{y}^{(n+1)} = \mathbf{y}^{(n)} + \alpha_n(\tilde{\mathbf{y}}^{(n+1)}-\mathbf{y}^{(n)}) + \beta_n(\mathbf{y}^{(n)}-\mathbf{y}^{(n-1)}), \tag{88}$$

or in standard form

$$\mathbf{y}^{(n+1)} = \mathbf{y}^{(n)} + \alpha_n(\mathbf{d}-\mathbf{B}\mathbf{y}^{(n)}) + \beta_n(\mathbf{y}^{(n)}-\mathbf{y}^{(n-1)}), \tag{89}$$

where the coefficients α_n and β_n are determined from the relations

$$\alpha_n = \frac{4}{(\bar{b}-\bar{a})}\frac{T_n(\bar{Z})}{T_{n+1}(\bar{Z})} \quad \text{and} \quad \beta_n = \frac{T_{n-1}(\bar{Z})}{T_{n+1}(\bar{Z})}, \tag{90}$$

for $n = 1, 2,\ldots$ and initially we choose $\alpha_0 = 1$. The equations (87) and (89) constitute the pre-conditioned form of the Chebychev semi-iterative method described earlier in equation (25).

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

5. Numerical Results

The basic iterative methods discussed earlier are now compared with and without pre-conditioning by studying the chosen problem for the three mesh sizes chosen, i.e. $h^{-1} = 5$, 10, and 20 which leads to a linear system similar to (1) whose coefficient matrix is positive definite and very sparse of order 16, 81 and 361 respectively.

TABLE 1

5 × 5 Laplace model problem. Table of maximum and minimum eigenvalues of the coefficient matrix **B** *versus the pre-conditioning factor* ω

ω Pre-conditioning factor	Λ_1 Maximum eigenvalue	Λ_N Minimum eigenvalue	*P*-condition no.
0	1·80902	0·19098	9·47230
0·2	1·54700	0·22581	6·85089
0·4	1·33642	0·27023	4·94549
0·6	1·16648	0·32758	3·56090
0·8	1·04164	0·40213	2·59030
1·0	1·00000	0·49795	2·00823
1·1	1·01010	0·55359	1·82464
1·2	1·04167	0·61140	1·70374
1·25	1·06666	0·63907	1·66907
1·3	1·09882	0·66383	1·65529
1·35	1·13922	0·68306	1·66782
1·4	1·18903	0·69302	1·71572
1·45	1·24949	0·69009	1·81060
1·5	1·32178	0·67260	1·96518
1·6	1·50652	0·63891	2·35795
1·7	1·75353	0·60·28	2·88752
1·8	2·07694	0·57771	3·59512
1·9	2·49735	0·55010	4·53979

The plot of the maximum and minimum eigenvalues of the coefficient matrix B versus the pre-conditioning factor ω were determined by judicious use of the Power method and the values obtained for the case $h^{-1} = 5$, correct to five decimal places are given in Table 1 and Fig. 1. It can be seen that the maximum eigenvalue Λ_{max} is minimized when $\omega = 1$ and the minimum eigenvalue Λ_{min} is maximized for ω in the range $1 \leqslant \omega \leqslant 2$ and confirms the analytical results derived in Section 3. The plot of the *P*-condition number versus pre-conditioning parameter ω is shown in Fig. 2 and confirms experimentally that a minimum *P*-condition number exists.

The validity of equation (36) for the optimum pre-conditioning parameter ω and equation (37) for the minimum *P*-condition number for the matrix **B** was confirmed experimentally by determining the quantities k_1, k_N, τ_1 and τ_N for the optimum ω and then substituting into equation (35). This is a quadratic equation involving ω non-linearly and at present the only solution process for such an equation involves

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

guessing an initial ω and then determining the coefficients $c(\omega)$, $d(\omega)$ and $f(\omega)$ from which a better value of ω is obtained. The whole process is then repeated until the process converges and a final value ω is obtained. More refined methods of determining

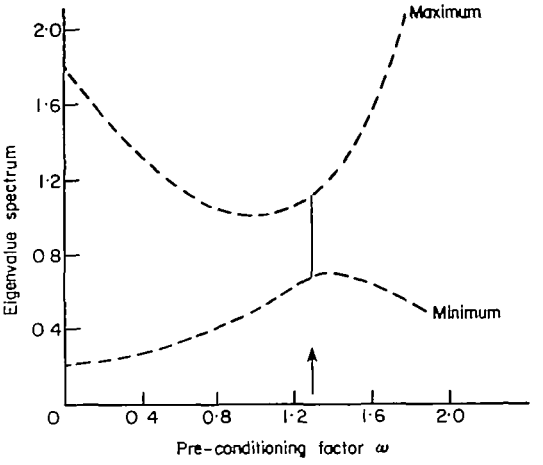

FIG. 1. Laplace model problem $h^{-1} = 5$. Plot of maximum and minimum eigenvalues of the coefficient matrix **B** versus pre-conditioning factor ω.

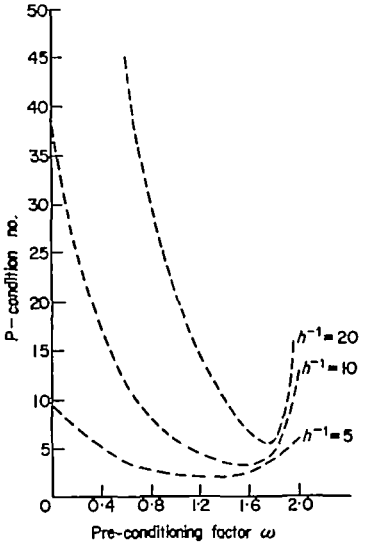

FIG. 2. Laplace model problem. Plot of *P*-conditioning number versus pre-conditioning factor ω.

ω are under investigation. To confirm the current results, the values of k_1, k_N, τ_1 and τ_N have been determined and the results given in Table 2. Excellent agreement was obtained between the theoretical results derived by use of equations (36) and (37) and the experimental results depicted in Fig. 2.

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

The iterative methods discussed, i.e. Simultaneous Displacement, Richardson's, Second order Richardson's, Chebychev second order Richardson and Chebychev semi-iterative methods were completed for the chosen problem with and without the use of the pre-conditioning factor and the results show a substantial saving of a large number of iterations by the use of pre-conditioning.

TABLE 2

Table showing agreement between experimental and theoretical results for the optimum pre-conditioning factor and P-condition number

Laplace model problem	k_1	τ_1	k_N	τ_N	Optimum pre-conditioning parameter		*P*-condition no.	
					Eqn (36)	Experimental Fig. 2	Eqn (37)	Experimental Fig. 2
$h^{-1} = 5$	0·0530	0·5396	0·2056	0·2301	1·32	1·32	1·655	1·65
$h^{-1} = 10$	0·1330	0·2708	0·2395	0·0573	1·578	1·575	2·847	2·85
$h^{-1} = 20$	0·1893	0·1297	0·2469	0·0147	1·769	1·775	4·915	5·03

TABLE 3

Method of Simultaneous Displacement

h^{-1} mesh size	Pre-condition parameter ω	Acceleration parameter α	a Lowest eigenvalue	b Largest eigenvalue	*P* *P*-condition number	*R* Rate of convergence	No. of iterations
5	0	1	0·19098	1·80902	9·473	0·2111	35
10	0	1	0·04894	1·95106	39·866	0·0502	120
20	0	1	0·01231	1·98769	161·470	0·0124	480
5	1·3	1·1	0·66383	1·09882	1·655	1·2083	8
10	1·6	0·95	0·55218	1·55512	2·816	0·7101	11
20	1·75	0·75	0·43322	2·28502	5·274	0·3792	21

The results are tabulated in Tables 3 to 7. All the iterations were initiated from the same initial approximation $u_{i,j}^{(0)}$ and continued until the convergence criteria

$$\max_{i,j} \frac{|u_{i,j}^{(n+1)} - u_{i,j}^{(n)}|}{|u_{i,j}^{(n)}|} < 5 \cdot 10^{-5} \tag{91}$$

was satisfied for $1 < i < p$ and $1 < j < p$.

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

The basic iterative methods described in Section 1 are usually performed with a generated matrix rather than with the original sparse matrix stored in the machine. Similarly, the non-zero coefficients arising from differential equations with non-constant coefficients are computed initially and stored in a compact way so that the same storage efficiency can be achieved. Hence, it is important that the application

TABLE 4

Richardson's Method

h^{-1} mesh size	Pre-condition parameter ω	m cycle of parameters	No. of iterations	Optimum pre-condition parameters	m cycle of parameters	No. of iterations
5	0	6	17	1·3	5	7
	0	7	14	1·3	6	6
	0	8	16	1·3	7	7
	0	9	17	1·3	8	7
10	0	6	41	1·6	4	12
	0	7	35	1·6	5	11
	0	8	40	1·6	6	12
	0	9	36	1·6	7	13
	0	10	39			
20	0	6	108	1·75	6	17
	0	7	98	1·75	7	14
	0	8	96	1·75	8	16
	0	9	90			
	0	12	72			
	0	15	75			

TABLE 5

Second order Richardson's Method

h^{-1} mesh size	Pre-condition parameter ω	Acceleration parameter α	β	Rate of convergence R	No. of iterations
5	0	1·25	0·25	0·6498	8
10	0	1·53	0·53	0·3167	32
20	0	1·73	0·73	0·1574	65
5	1·3	1·15	0·0175	1·5545	8
10	1·6	1·00	0·0641	1·1917	10
20	1·75	0·85	0·155	0·8708	13

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

of the pre-conditioning parameter ω can be completed computationally on a generated matrix. Earlier, the notation of **G** or $(\mathbf{I}-\omega\mathbf{U})^{-1}$ was introduced as representing a matrix which was readily inverted. Here we proceed further by depicting in Fig. 3 the stencils from which the vector **By** as used in Section 4 can be generated on the grid of mesh points.

TABLE 6

Chebychev second order Richardson's Method

h^{-1} mesh size	Pre-condition parameter ω	No. of iterations
5	0	15
10	0	30
20	0	59
5	1·3	7
10	1·6	8
20	1·75	14

TABLE 7

Chebychev semi-iterative method

h^{-1} mesh size	Pre-condition parameter ω	Spectral radius of the Jacobi method μ	No. of iterations
5	0	0·809	15
10	0	0·951	28
20	0	0·987	85
5	1·30	0·247	7
10	1·6	0·477	9
20	1·75	0·680	19

The individual terms of equation (40) which constitute the $(\mathbf{I}-\omega\mathbf{L})$ and $(\mathbf{I}-\omega\mathbf{U})$ matrices can be seen in finite difference form to be

$$-\frac{\omega}{4}u_{i-1,j}-\frac{\omega}{4}u_{i,j-1}+u_{i,j}, \tag{92}$$

and

$$u_{i,j}-\frac{\omega}{4}u_{i,j+1}-\frac{\omega}{4}u_{i+1,j}, \tag{93}$$

respectively.

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015

Thus the generation of the vector **By** on the net can be considered to be the application of (93) on the vector **y**, i.e. a back substitution process involving the operator given in Fig. 3(c); followed by the usual Laplace operator as given in Fig. 3(b) with a final application of (92) to the existing vector on the net, which is a forward substitution process as given by the operator in Fig. 3(a).

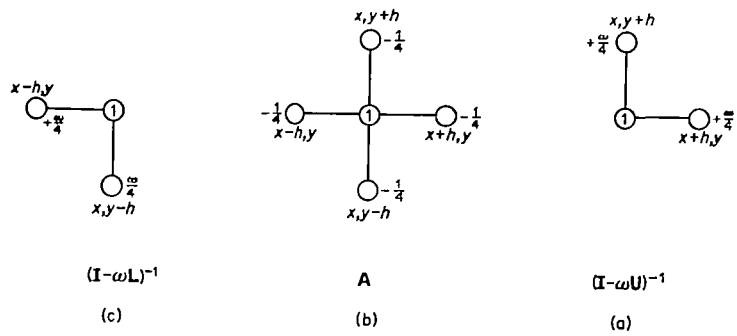

FIG. 3

Finally, the increase in the amount of computational labour involved in each iteration must be considered. Since the matrix **A** is sparse and each row involves only two off-diagonal elements in both lower and upper triangular constituents, the determination of the vector $\mathbf{y}^{(n+1)}$ from equation (73) involves at most four multiplications and ten additions for each iteration. When $\omega = 0$ and we revert to the basic iteration method, this reduces to the vector $\mathbf{x}^{(n+1)}$ from equation (5) and the amount of work involved is two multiplications and six additions. Since computer multiplication time is large compared to computer addition time, we see immediately that the application of pre-conditioning involves approximately a two-fold increase in work. To counteract this extra work, we have the greatly increased convergence rates as demonstrated in Tables 3 to 7.

The author is indebted to Mrs A. Fairburn for programming and computational assistance and the referee for valuable and constructive criticisms.

REFERENCES

AFRIAT, S. N. 1954 *Q. Jl Math.* **5**, (2), 81–98.
COLLATZ, L. 1949 *Eigenwertaufgaben mit technischen Anwendungen*. Leipzig.
FORSYTHE, G. E. & WASOW, W. 1960 *Finite Difference Methods for Partial Differential Equations*. 228–229. John Wiley & Sons.
FRANKEL, S. P. 1950 *M.T.A.C.*, **4**, 65–75.
GOLUB, G. H. & VARGA, R. S. 1961 *Num. Math.* **3**, 1+7.
MARKOFF, W. 1916 *Math. Annln*, **77**, 213–258.
STIEFEL, E. L. 1958 *Natn. Bur. Stand. App. Maths series*, no. 49, 1.
VARGA, R. S. 1962 *Matrix Iterative Analysis*. New York: Prentice Hall.
YOUNG, D. 1953 *J. Math. Phys.* **32**, 243–255.
YOUNG, D. 1954 *Trans. Am. math. Soc.* **76**, 92–111.

Downloaded from http://imamat.oxfordjournals.org/ at New York University on February 5, 2015