On Convergence Criteria for the Method of Successive Over-Relaxation

By C. G. Broyden

1. Introduction. The solution of a set of n simultaneous linear equations,

$$Mx = c \tag{1.1}$$

where M is non-singular and both M and c are real is often attempted by iterative methods, which rely on the systematic improvement of an approximate solution. In order to simplify the algebra, it is convenient to consider the system.

$$Ax = b \tag{1.2}$$

where

$$A = DM, \qquad b = Dc \tag{1.3}$$

and D is a diagonal matrix chosen so that the elements of the principal diagonal of A are unity.

A may then be expressed as the sum of three matrices

$$A = I + L + U \tag{1.4}$$

where I is the unit matrix and L, U are lower and upper triangular matrices respectively.

The type of iterative process considered here is that known as the extrapolated Gauss-Seidel method, or the method of successive over-relaxation (SOR).

Convergence criteria have been established for this method by Ostrowski [3] for the case where M is symmetric. This paper derives sufficient conditions for the convergence of the method when applied to problems involving non-symmetric matrices.

2. The SOR Method. Let x_{i+1} and x_i be successive approximate solutions to the equation

$$Ax = b$$

The SOR method is defined by

$$x_{i+1} = x_i - \omega[(I + U)x_i + Lx_{i+1} - b]. \tag{2.1}$$

(See, e.g. [2], where, however, a different notation is used to that employed here). Eliminating U from the above equation by substitution from equation (1.4) gives

$$(I + \omega L)(x_{i+1} - x_i) = -\omega(Ax_i - b). \tag{2.2}$$

Now the residual ϵ_i corresponding to the approximate solution x_i is defined by

$$\epsilon_i = Ax_i - b.$$

Received February 19, 1963.

Therefore

$$\epsilon_{i+1} = Ax_{i+1} - b$$

and

$$A^{-1}(\epsilon_{i+1} - \epsilon_i) = x_{i+1} - x_i .$$

Hence equation (2.2) becomes

$$(I + \omega L)A^{-1}(\epsilon_{i+1} - \epsilon_i) = -\omega\epsilon_i$$

which gives the recurrence relation

$$\epsilon_{i+1} = [I - \omega A(I + \omega L)^{-1}]\epsilon_i . \tag{2.3}$$

This has the form

$$\epsilon_{i+1} = (I - B)\epsilon_i \tag{2.4}$$

and a sufficient condition for the convergence of processes of this type will now be derived.

THEOREM 1. *A sufficient condition for the convergence of processes where the residual vectors obey the equation*

$$\epsilon_{i+1} = (I - B)\epsilon_i$$

is that there exist matrices S and G such that

$$S > 0$$

and

$$G = B^TS + SB - B^TSB > 0.$$

(The notation $S > 0$ means that S is symmetric and positive definite, and the superscript T indicates transposition.)

Proof. Let $f_i = \epsilon_i{}^TS\epsilon_i$. Then since, by hypothesis, $S > 0$, $f_i > 0$ for $\epsilon_i \neq 0$. Hence, $f_i \to 0$ as $i \to \infty$ is a necessary and sufficient condition for convergence. Now

$$\begin{aligned} f_{i+1} &= \epsilon_{i+1}^TS\epsilon_{i+1} \\ &= \epsilon_i{}^T(I - B^T)S(I - B)\epsilon_i \\ &= \epsilon_i{}^T(S - G)\epsilon_i . \end{aligned}$$

Let $\phi_i = \epsilon_i{}^TG\epsilon_i$. Hence, $f_{i+1} = f_i - \phi_i$. A sufficient condition for f_i to tend to zero with increasing i is that there exists a positive constant k such that

$$\phi_i \geqq kf_i \tag{2.5}$$

for then $f_{i+1} \leqq (1 - k)f_i$ and the sequence f_i converges.

Since $S > 0$, all its eigenvalues are real and positive, and if the largest is λ_{max} then

$$f_i \leqq \lambda_{max}\epsilon_i{}^T\epsilon_i . \tag{2.6}$$

(See, e.g. [1], p. 65). Now since $S = S^T$, the matrix G is symmetric and its eigenvalues are real. Denote the smallest by $\mu_{\min}$. Then $\phi_i \geqq \mu_{\min}\epsilon_i^T\epsilon_i$ ([1], p. 65). Hence, from equation (2.6)

$$\phi_i \geqq \frac{\mu_{\min}}{\lambda_{\max}} f_i .$$

Now if $G > 0$, $\mu_{\min} > 0$ and equation (2.5) will be satisfied. This proves Theorem 1.

COROLLARY. *If $S > 0$ and $G \leqq 0$ i.e., G is negative semidefinite, then the process will never converge.*

Gal. 4 MT 8437 p. 5 Take 10-23-14 Gx 71 10-11-63

Proof. If $G \leqq 0$, $\phi_i \leqq 0$, and $f_{i+1} \geqq f_i$. Hence f_i can never be reduced to zero, and the corollary is proved.

Theorem 1 will now be applied to the SOR method. Since, from equation (2.3) $B = \omega A(I + \omega L)^{-1}$, a sufficient condition for SOR to converge is that there exists a matrix S where $S > 0$ such that

$$\omega(I + \omega L^T)^{-1}A^TS + \omega SA(I + \omega L)^{-1} - \omega^2(I + \omega L^T)^{-1}A^TSA(I + \omega L)^{-1} > 0. \tag{2.7}$$

This condition may be simplified by using the following lemma.

LEMMA. *A necessary and sufficient condition for $P > 0$ is that $Q^TPQ > 0$ where Q is any non-singular matrix.*

Proof. Let $Qx = z$. Then $x^TQ^TPQx = z^TPz$. But since Q is non-singular, for every non-zero x there exists a non-zero z, and conversely. The lemma follows. Since $(I + \omega L)$ is non-singular, the sufficient condition 2.7 may be transformed by the lemma into

$$\omega[A^TS(I + \omega L) + (I + \omega L^T)SA - \omega A^TSA] > 0 \tag{2.8}$$

3. Symmetric Matrices. Suppose that M is symmetric. Put $S = D^{-1}M^{-1}D^{-1}$. Condition (2.8) becomes

$$\omega[D^{-1}(I + \omega L) + (I + \omega L^T)D^{-1} - \omega M] > 0 \tag{3.1}$$

Equations (1.3), (1.4), and the assumed symmetry of M, give

$$M = D^{-1}(I + L + U) = (I + L^T + U^T)D^{-1}.$$

Equating the upper triangular partitions of these two representations of M gives

$$D^{-1}U = L^TD^{-1} \tag{3.2}$$

Hence equation (3.1) reduces to

$$\omega(2 - \omega)D^{-1} > 0 \tag{3.3}$$

Now if $M > 0$ it follows from the lemma that $S > 0$ and since if $M > 0$ then $D > 0$, the condition (3.3) obtains. Hence SOR will converge, if $M > 0$ and $0 < \omega < 2$. If, however, $\omega < 0$ or $\omega < 2$ the matrix G becomes negative definite, so by the corollary to Theorem 1, SOR will not converge for ω lying outside the range $0 \rightarrow 2$.

4. Non-Symmetric Matrices. Take for S in equation (2.8) the matrix $(A^T)^{-1}A^{-1}$. Since A is non-singular this automatically fulfills the conditions of symmetry and positive definiteness. The condition for convergence becomes

$$\omega A^{-1}(I + \omega L) + \omega(I + \omega L^T)(A^T)^{-1} - \omega^2 I > 0$$

and since A is non-singular, the lemma gives

$$\omega(I + \omega L)A^T + \omega A(I + \omega L^T) - \omega^2 AA^T > 0.$$

Decomposing A by equation (1.4) and simplifying gives

$$\omega(A + A^T) - \omega^2[(I + U)(I + U^T) - LL^T] > 0. \tag{4.1}$$

A second condition, analogous to that given by equation (4.1) may be derived by putting $S = I$ in equation (2.8). This gives

$$\omega A^T(I + \omega L) + \omega(I + \omega L^T)A - \omega^2 A^T A > 0.$$

Decomposing A and simplifying gives the sufficient condition for convergence

$$\omega(A + A^T) - \omega^2[(I + U^T)(I + U) - L^T L] > 0. \tag{4.2}$$

It will now be shown that if $A + A^T > 0$ a positive ω may be found such that conditions 4.1 and 4.2 hold.

THEOREM 2. *If $P > 0$ and $Q = Q^T$ there exists a positive ω such that $P + \omega Q > 0$.*

Proof. Let $f_1 = x^T P x$. Since $P > 0$ all its eigenvalues are real and positive. Let the smallest be $\lambda_{\min}$

$$\therefore \quad f_1 \geqq \lambda_{\min} x^T x$$

Now Q is symmetric, hence all the eigenvalues are real. Denote the algebraically smallest by $\mu_{\min}$. If

$$f_2 = x^T Q x$$

$$f_2 \geqq \mu_{\min}\, x^T x$$

and

$$f = x^T(P + \omega Q)x \geqq (\lambda_{\min} + \omega\mu_{\min})x^T x.$$

Consider now the two cases

(a) $$\mu_{\min} \geqq 0$$

In this case $f > 0$ for all $\omega \geqq 0$.

(b) $$\mu_{\min} = -|\mu_{\min}| < 0$$

$$\therefore \quad f > (\lambda_{\min} - \omega\,|\,\mu_{\min}\,|)x^T x$$

$$\therefore \quad f > 0 \quad \text{for} \quad \omega < \frac{\lambda_{\min}}{|\,\mu_{\min}|} \quad \text{and} \quad x \neq 0$$

This proves Theorem 2.

A further sufficient condition for convergence may be derived from equation (4.2).

Define the matrices P and Q by

$$P = (I + L^T)(I + L) - U^T U$$
$$Q = (I + U^T)(I + U) - L^T L$$
$$\therefore \quad P + Q = A + A^T.$$

Equation (4.2) becomes

$$\omega^2 P - \omega(\omega - 1)Q > 0. \tag{4.3}$$

Hence SOR will converge for $\omega = 1$ if $P > 0$. A similar condition may be derived from equation (4.1) in the same way.

5. Conclusions. Theorem 2 indicates that conditions (4.1) and (4.2) will be satisfied if $A + A^T$ is positive definite, and a sufficiently small positive value of ω is used. Now any matrix A may be expressed as the sum of symmetric and anti-symmetric components, and the matrix $A + A^T$ is merely double the symmetric component. Thus if a matrix is decomposed in this way and the symmetric component is positive definite, it will always be possible to find an ω such that successive over-relaxation, or possibly successive under-relaxation, will converge. This is clearly a more general form of Ostrowski's criterion, to which it reduces in the limiting case when the anti-symmetric component becomes zero.

Condition (4.3) although derived from the same equation, is rather different in character. It shows that SOR will converge for $\omega = 1$ if

$$(I + L^T)(I + L) - U^T U > 0. \tag{5.1}$$

This leads to the conclusion that matrices exist for which an important factor in guaranteeing convergence of the method of successive over-relaxation is "lower triangular dominance". In the limiting case when U becomes zero, (5.1) holds, and if ω takes the value unity the equations are solved in one step.

That the conditions $P > 0$ and $A + A^T > 0$ are not equivalent is probably best demonstrated by examples.

$$A = \begin{bmatrix} 1 & -2 \\ 3 & 1 \end{bmatrix}. \tag{5.2}$$

In this case A is strongly anti-symmetric, and its symmetric component is positive definite although neither P nor Q possess this property. Successive relaxation will lead to a solution for ω sufficiently small, e.g., $\frac{1}{4}$.

$$A = \begin{bmatrix} 1 & \frac{1}{3} \\ 2 & 1 \end{bmatrix} \tag{5.3}$$

Here A is lower-triangularly dominant. P is positive definite, but $A + A^T$ and Q are not. Convergence is guaranteed for $\omega = 1$.

It should be emphasised here that although conditions (4.1), (4.2) and (4.3) are sufficient for convergence, they are not necessary. In particular, the largest value of ω for which (4.1) and (4.2) is valid may well be exceeded without the process diverging. They do, though, show that there exist two quite definite types

of non-symmetric matrix for which SOR will always converge provided that a suitable value of ω is chosen.

6. Acknowledgments. The author is grateful to the directors of the English Electric Co. for permission to publish this paper, and to his colleagues, F. Ford and J. M. Williamson, for their helpful comments and criticisms.

Atomic Power Division,
English Electric Co.,
Whetstone, Leicester.
England.

1. E. Bodewig, *Matrix Calculus*, North Holland Publishing Co. Amsterdam, 1959.
2. G. E. Forsythe & W. R. Wasow, *Finite Difference Methods for Partial Differential Equations*, John Wiley and Sons., Inc., New York. 1960.
3. A. M. Ostrowski, "On the linear iteration procedures for symmetric matrices," *Rend. Mat. e Appl.* v. 13, 1954, p. 140.

On Inverses of Finite Segments of the Generalized Hilbert Matrix

By Jean L. Lavoie

The purpose of this note is to show that two theorems given by Smith [1] on inverses of finite segments of the generalized Hilbert Matrix can be proved in a simple manner by using results from the theory of generalized hypergeometric series.

The usual notation for generalized hypergeometric functions will be used:

$$(1) \qquad {}_PF_Q(z) = {}_PF_Q\left(\begin{matrix} a_1 \cdots a_P \\ b_1 \cdots b_Q \end{matrix} \middle| z\right) = \sum_{K=0}^{\infty} \frac{\prod_{j=1}^{P} (a_j)_K}{\prod_{j=1}^{Q} (b_j)_K} \cdot \frac{z^K}{K!},$$

where

$$(\sigma)_\mu = \frac{\Gamma(\sigma + \mu)}{\Gamma(\sigma)}.$$

See Erdélyi [2], Chapters 2 and 4 for details.

Let H_n represent a finite segment of the generalized Hilbert matrix, i.e.,

$$(2) \qquad H_n = (h_{ij}), \qquad h_{ij} = (p + i + j - 1)^{-1}, \quad i, j = 1, 2, \cdots n.$$

Here n is the order of the segment and obviously

$$p \neq -1, -2, \cdots, -(2n - 1).$$

We shall assume that the above conditions on i, j, and p hold throughout this paper.

Received April 15, 1963. Revised July 24, 1963.