LA-2314

c.3

LOS ALAMOS SCIENTIFIC LABORATORY
OF THE UNIVERSITY OF CALIFORNIA ○ LOS ALAMOS NEW MEXICO

NUMERICAL METHODS FOR SOLVING LINEAR SYSTEMS AND APPLICATIONS TO ELLIPTIC DIFFERENCE EQUATIONS

LEGAL NOTICE

Printed in USA. Price $2.50. Available from the

Office of Technical Services
U. S. Department of Commerce
Washington 25, D. C.

LA-2314
PHYSICS AND MATHEMATICS
(TID-4500, 15th Ed.)

LOS ALAMOS SCIENTIFIC LABORATORY
OF THE UNIVERSITY OF CALIFORNIA LOS ALAMOS NEW MEXICO

REPORT WRITTEN: September 1958

REPORT DISTRIBUTED: October 20, 1959

NUMERICAL METHODS FOR SOLVING LINEAR SYSTEMS AND APPLICATIONS TO ELLIPTIC DIFFERENCE EQUATIONS

by

C. E. Lee and P. M Stone

This report expresses the opinions of the author or authors and does not necessarily reflect the opinions or views of the Los Alamos Scientific Laboratory.

Contract W-7405-ENG. 36 with the U. S. Atomic Energy Commission

ABSTRACT

Iterative numerical methods for solving independent, simultaneous, inhomogeneous linear equations are surveyed. Application of the methods to elliptic difference equations as arise in neutron diffusion, heat conduction, and potential problems is discussed.

ACKNOWLEDGMENTS

We wish to express our gratitude to G. I. Bell, B. G. Carlson, S. M. Ulam, R. R. Johnston, C. L. Longmire, and G. H. Pimbley for their comments and criticisms.

CONTENTS

1. INTRODUCTION

This paper presents a survey of the methods of solving independent, simultaneous, inhomogeneous linear equations. Here we are concerned with systems too extensive to be handled conveniently with a desk calculator; therefore we restrict the discussion to iterative-type methods performable upon high speed digital computers.

We assume that we have a system of equations of the form

$$\sum_{j=1}^{N} a_{ij} x_j + s_i = 0 \qquad i = 1, 2, \ldots, N \tag{1.1}$$

where the x_j are the unknowns, and the coefficients a_{ij} and the inhomogeneous term s_i are given. In matrix notation (1.1) is

$$A\underline{x} + \underline{s} = 0 \tag{1.2}$$

where A is the N x N non-singular matrix of the coefficients a_{ij}, $\underline{x}$ is the N-dimensional vector of unknowns x_j, and $\underline{s}$ is the N-dimensional vector of the inhomogeneous term.

We assume a correspondence between unknowns and equations, i.e., the ith equation will be used to solve for the ith unknown. The order of improving the unknowns in an iterative procedure may vary according to a particular method, but a given equation will always be solved for the same unknown.

We assume that the main diagonal elements of A are dominant.* Thus, the solutions of Eq. (1.1) are not altered if each equation is multiplied by a normalization constant $1/a_{ii}$. The matrix A then takes the form

$$A = I + L + U \tag{1.3}$$

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

The methods for solving Eq. (1.1) are direct, iterative, and combinations thereof. Direct methods, such as Gauss elimination are those for which the exact solution is obtained (assuming no round-off errors) in a finite number of steps. For a general system of N equations, these methods require the order of N^3 operations.** They can be used for any set of independent equations (in contrast to iterative-type methods), but they suffer from the occurrence of round-off errors which, in some instances, can be large.

*Iterative methods can be used if the matrix A is positive definite (Sections 2, 3, and 6), it has real eigenvalues (Section 4), or if the main diagonal elements dominate the off-diagonal elements (Sections 2 and 3).

**For a multidiagonal matrix, the number of operations approximately equals the number of equations times the maximum row width between non-zero row entries times the maximum column width between non-zero column entries.

On the other hand, iterative methods (Sections 2 and 3) generate only an approximate solution in a finite number of steps: the exact solution would usually be obtained only in the limit of an infinite number of steps.* They are iterative in the sense that an approximate solution is successively improved in a stepwise manner. Round-off errors are generally small, do not tend to propagate, and are reduced by succeeding iterations. Iterative methods have an advantage over direct methods in that if a reasonable approximation is available initially, the effort involved in obtaining a "solution" will be reduced. Although the exact solution cannot be obtained generally, a convergent solution can be made to approximate the true solution of the equations as accurately as desired. However, the iterative methods are applicable only under certain circumstances, as discussed below.

If the system of equations is very nearly dependent, the number of iterations required to reduce an initial error of the approximate solution can, in some instances, involve more numerical operations than if the system were solved by a direct method.

The inclusion in the iterative method of a dependence of an extrapolation parameter, ω, on the iteration cycle number is discussed in Section 4. In Section 5, the application of iterative methods to blocks of unknowns is considered. Some variational methods are discussed in

*The methods of conjugate gradients and conjugate directions (Section 6) are exceptions. These methods obtain the exact solution in N steps by an iterative-direct method.

Section 6. And finally, the particular example of a system of difference equations originating from a second order linear elliptic differential operator, such as arises in neutron diffusion, heat conduction, and potential problems, is considered in Section 7.

We generally state results without proof, referring to the literature for the details. The bibliography at the end is meant to be only representative, not exhaustive. References to extensive bibliographies are listed.

2. BASIC ITERATIVE METHODS

Consider the system of N independent, simultaneous, inhomogeneous linear equations given by Eq. (1.1). We assume a matrix equation of the form

$$A\underline{x} + \underline{s} = 0 \tag{2.1}$$

where

$$A = I + L + U \tag{2.2}$$

as in Section 1. The independence assumption guarantees that A is nonsingular.

The fundamental equation for the basic iterative methods is obtained by rewriting Eq. (2.1)

$$\underline{x} = -(L + U)\underline{x} - \underline{s} \tag{2.3}$$

Simultaneous Iteration

Gauss was the first to develop Eq. (2.3) into an iteration scheme by forming

$$\underline{x}^{(p+1)} = -(L + U)\underline{x}^{(p)} - \underline{s} \tag{2.4}$$

where the superscript denotes the iteration cycle number. This equation defines Simultaneous Iteration.

The newest estimate of the solution, $\underline{x}^{(p+1)}$, is obtained from all the components of $\underline{x}^{(p)}$, connected and weighted through $-(L + U)$, and the inhomogeneous term. All iterates of $\underline{x}$ are advanced simultaneously and are entirely dependent on the previous iterate.

This procedure has various designations in the literature. Although frequently called *Gauss Iteration*, such a term is misleading because Gauss did not solve the equations in a fixed order as indicated in Eq. (2.4). *Schmidt*, who used the method extensively, often has had his name associated with it. When it has been used to solve elliptic difference equations, the label *Richardson Method* has been applied. Many later authors refer to it as *Method I*. In this paper, we shall refer to it by its descriptive title: *Simultaneous Iteration*.

It seems reasonable that the Simultaneous Iteration would converge more rapidly if a component of $\underline{x}^{(p+1)}$, when determined, were used in determining a component of $\underline{x}^{(p+1)}$ still unknown. Such is indeed the case, and this defines the next procedure.

Successive Iteration

Solving Eq. (2.1) in the order 1, 2, ..., N, we define

$$\underline{x}^{(p+1)} = -L\underline{x}^{(p+1)} - U\underline{x}^{(p)} - \underline{s} \qquad (2.5)$$

or

$$\underline{x}^{(p+1)} = -(I + L)^{-1}\, U\underline{x}^{(p)} - (I + L)^{-1}\, \underline{s} \qquad (2.6)$$

One notices from the form of Eq. (2.5) that the formation of the ith component of $\underline{x}^{(p+1)}$ involves the components x_k^{p+1}, $k < i$, and $x_{k'}^{(p)}$, $k' > i$. Thus, $x_i^{(p+1)}$ depends upon the advanced iterates already computed and the previous iterates yet unadvanced. Thus Eq. (2.6) defines a Successive Iteration procedure.

There are many different designations for this method in the literature. The most frequent are *Seidel's Method* and *Gauss-Seidel Method*. This is misleading, for Seidel did not use the equations in a fixed order.* *Liebmann*, who used it extensively as applied to Laplace's equation, has had his name associated with it. The term *Method II* is popular among many later authors. We shall adopt the descriptive title, *Successive Iteration*.

*Seidel's Method is essentially identical to Southwell's Relaxation Method of 50 years later.

Convergence Properties

The convergence rates of the above procedures are not easily evaluated for a general matrix A; however, some conditions guaranteeing convergence and relative convergence rates can be given.

Consider an error vector $\underline{E}^{(p)}$, defined by

$$\underline{E}^{(p)} = \underline{x} - \underline{x}^{(p)} \tag{2.7}$$

giving the difference of the pth approximation, $\underline{x}^{(p)}$, from the true solution, $\underline{x}$, which satisfies

$$A\underline{x} + \underline{s} = 0 \tag{2.8}$$

Substituting Eq. (2.7) into the defining equations for Simultaneous and Successive Iteration, we obtain the error equations

$$\text{Simultaneous Iteration:} \quad \underline{E}^{(p+1)} = -(L + U)\underline{E}^{(p)} \tag{2.9}$$

$$\text{Successive Iteration:} \quad \underline{E}^{(p+1)} = -(I + L)^{-1}\, U\underline{E}^{(p)} \tag{2.10}$$

If the matrices on the right-hand sides of Eqs. (2.9) and (2.10) have eigenvalues $|\lambda_i| < 1$, then the iterated error vector, $\underline{E}^{(p+1)}$, will tend to a null vector as $p \to \infty$. In that instance the Simultaneous and Successive Iteration methods are said to be convergent. The characteristic determinantal equations for the eigenvalues are, respectively:

Simultaneous Iteration: $\det \left| L + U + \lambda I \right| = 0$ (2.11)

Successive Iteration: $\det \left| U + \lambda(I + L) \right| = 0$ (2.12)

or, in terms of the coefficients of Eq. (2.1):

Simultaneous Iteration:

$$\begin{vmatrix} \lambda & a_{12} & a_{13} & \cdot & \cdot & \cdot & a_{1n} \\ a_{21} & \lambda & a_{23} & \cdot & \cdot & \cdot & a_{2n} \\ a_{31} & a_{32} & \lambda & & & & \cdot \\ \cdot & \cdot & & & & & \cdot \\ \cdot & \cdot & & & & & \cdot \\ \cdot & \cdot & & & & & \cdot \\ a_{n1} & a_{n2} & \cdot & \cdot & \cdot & \cdot & \lambda \end{vmatrix} = 0 \qquad (2.13)$$

Successive Iteration:

$$\begin{vmatrix} \lambda & a_{12} & a_{13} & \cdot & \cdot & \cdot & a_{1n} \\ \lambda a_{21} & \lambda & a_{23} & \cdot & \cdot & \cdot & a_{2n} \\ \lambda a_{31} & \lambda a_{32} & \lambda & & & & \cdot \\ \cdot & \cdot & & & & & \cdot \\ \cdot & \cdot & & & & & \cdot \\ \cdot & \cdot & & & & & \cdot \\ \lambda a_{n1} & \lambda a_{n2} & \cdot & \cdot & \cdot & \cdot & \lambda \end{vmatrix} = 0 \qquad (2.14)$$

The order of solving the Simultaneous Iteration equations, Eq. (2.4), has no effect upon the convergence rate. However, the order of solving the Successive Iteration equations, Eq. (2.5), does affect the convergence rate.

In general, the eigenvalues of the matrix A are not known in advance. We now consider theorems determining the convergence of these elementary procedures, which depend only upon a knowledge of the coefficients of A. Proofs of the theorems are sketched, and a reference in which the details can be found is given.

Theorem I (Ref. 30, p. 141, and Ref. 1)

Given a non-singular matrix

$$A = I + L + U$$

the methods of Simultaneous and Successive Iteration both converge for an arbitrary starting point and for any order of solving the equations if the following conditions* are satisfied**

$$\begin{aligned} &\text{Simultaneous Iteration:} \quad \max_i {\sum_j}' |a_{ij}| \leq 1 \\ &\text{Successive Iteration:} \quad \max_i {\sum_j}' |a_{ij}| < 1 \end{aligned} \qquad (2.15)$$

where the a_{ij} are the coefficients of the matrix A.

*The prime indicates the absence of i = j in the sum.

**That is, the diagonal elements of A are dominant.

The proof of this theorem for Simultaneous Iteration relies upon a theorem of Hadamard, concerning eigenvalues, that states: all eigenvalues of a matrix lie in or on circles centered at a_{ii} and of radius $\sum_j' |a_{ij}|$. Here we are concerned with the matrix $-(L + U)$, which has zeros along the main diagonal. Hence, all the eigenvalues will lie within circles of radius $\sum_j' |a_{ij}|$ centered at the origin. Convergence is guaranteed if the radius is less than 1. Now $-(L + U)$ cannot have eigenvalues equal to 1 since then we would have $A = I + L + U = 0$, which is forbidden, since A is non-singular by definition.

The proof of this theorem for Successive Iteration follows from observing the error at each step and showing that it decreases to zero.

Note that in Theorem I, convergence is obtained if the diagonal of the original matrix A (in our case I) is greater than the sum of the off-diagonal terms in a row. In other words, convergence is obtained if the diagonal term dominates. If the original set of equations does not have 1's along the diagonal, the matrix A can be obtained by dividing each equation by its diagonal term (if non-zero). The sums of the resultant off-diagonal terms will be less than 1 if the original diagonal terms were dominant.

Theorem II (Ref. 30, p. 141)

Given a matrix

$$A = I + L + U$$

the method of Simultaneous Iteration converges, if the norm of -(L + U) is

$$\|-(L + U)\| = \sqrt{\sum_{i,j}{}' |a_{ij}|^2} \leq 1 \tag{2.16}$$

The proof of this theorem follows from Theorem I [since all eigenvalues of -(L + U) lie inside a circle of unit radius].

A matrix is defined to be positive definite if

$$\sum_{i,j} a_{ij} x_i x_j > 0 \tag{2.17}$$

for all $\underline{x} \neq 0$. This brings us to Theorem III.

Theorem III (Ref. 30, p. 142)

Given a real symmetric matrix

$$A = I + L + U$$

the method of Successive Iteration converges if, and only if, A is positive definite.

The proof of this theorem follows by forming the quadratic function

$$F(\underline{x}) = \frac{1}{2} \underline{x} \cdot A\underline{x} + \underline{x} \cdot \underline{s} \tag{2.18}$$

and noting that the solution to the system of equations

$$A\underline{x} + \underline{s} = 0$$

minimizes Eq. (2.18). The proof then amounts to showing that $F(\underline{x})$ is diminished at each step of the iteration, if and only if, A is positive definite.

Stein and Rosenberg (Ref. 23) have proved some very valuable theorems related to the two methods under discussion. One of the most useful, whose proof relies upon the properties of non-negative matrices, is the following.

Theorem IV (Ref. 23)

Given a matrix A with coefficients

$$a_{ii} = 1$$

and

$$a_{ij} \leq 0 \qquad \text{for all } i \neq j$$

then the methods of Simultaneous and Successive Iteration converge and diverge simultaneously. In the case of convergence (A positive definite), Successive Iteration is more rapid.

3. EXTRAPOLATED ITERATIVE METHODS

The Simultaneous and Successive Iteration convergence rates can be increased in various ways by extrapolation (using iterates at previous cycles).

Assume that the matrix A has real eigenvalues, μ_i, and a complete set of eigenvectors,[*] η_i, such that[**]

$$A\underline{\eta}_i = \mu_i \, \underline{\eta}_i \tag{3.1}$$

where $\mu_i = 1 - \lambda_i$. The λ_i are the eigenvalues of the matrix $-(L + U)$ of Section 2. Throughout this section, we assume that Simultaneous Iteration converges; i.e., $|\lambda_i| < 1$.

Extrapolated Simultaneous Iteration

In a straightforward manner, Simultaneous Iteration, Eq. (2.4), can be extrapolated by

[*] That is, each eigenvalue is of index 1 or simple.

[**] It is not implied that extrapolation procedures will not follow without these assumptions; however, the choice of extrapolation parameters might be more difficult.

$$\underline{x}^{(p+1)} = \underline{x}^{(p)} - \omega \left[A\underline{x}^{(p)} + \underline{s} \right] \tag{3.2}$$

where ω is the extrapolation parameter.

The error vector, $E^{(p)}$, of Eq. (3.2) is transformed by

$$\underline{E}^{(p+1)} = (I - \omega A)\underline{E}^{(p)} \tag{3.3}$$

Expanding the error in the eigenvectors of A, $\underline{\eta}_i$, we have

$$\underline{E}^{(p)} = \sum_{i=1}^{N} \alpha_i \, \underline{\eta}_i = \sum_{i=1}^{N} \underline{E}_i^{(p)} \tag{3.4}$$

where each mode is transformed by

$$\underline{E}_i^{(p+1)} = \left[1 - \omega(1 - \lambda_i) \right] \underline{E}_i^{(p)} \tag{3.5}$$

Any individual error mode $\underline{E}_i^{(p)}$ can be eliminated by the parameter choice

$$\omega = \frac{1}{1 - \lambda_i} \tag{3.6}$$

Thus, if all the eigenvalues were known, we could eliminate all error modes and obtain an exact solution in N iterations by choosing ω differently for each iteration. Unfortunately, the eigenvalues are not known in general.

For a fixed constant ω, we must choose it so that all modes will be decreased simultaneously; i.e.,

$$\left|1 - \omega(1 - \lambda_i)\right| < 1 \qquad \lambda_o \leq \lambda_i \leq \lambda_m \tag{3.7}$$

where λ_o and λ_m are the minimum and maximum eigenvalues, respectively. If Simultaneous Iteration converges ($|\lambda_i| < 1$), then ω is in the range

$$0 < \omega \leq 1$$

To select an ω that will minimize Eq. (3.7) requires knowledge of λ_o and λ_m. However, if we assume $\lambda_o = -\lambda_m$, then the optimum ω is

$$\omega = 1$$

Thus, the optimum Extrapolated Simultaneous Iteration becomes simply Simultaneous Iteration and nothing has been gained.

On the other hand, improvement might be expected by forming a second order extrapolation scheme

$$\underline{x}^{(p+1)} = \underline{x}^{(p)} + \alpha\left[\underline{x}^{(p)} - \underline{x}^{(p-1)}\right] - \omega\left[A\underline{x}^{(p)} + \underline{s}\right] \tag{3.8}$$

where the new estimate, $\underline{x}^{(p+1)}$, depends upon the two previous iterates, $\underline{x}^{(p)}$ and $\underline{x}^{(p-1)}$. The form of the equation is determined by requiring the equality to hold in the limit of convergence. Two parameters, α and

ω, must be chosen. As before, consider each error mode, $\underline{E}_i^{(p+1)}$, where

$$\underline{E}_i^{(p+1)} = \left[1 + \alpha - \omega(1 - \lambda_i)\right] \underline{E}_i^{(p)} - \alpha\, \underline{E}_i^{(p-1)} \qquad (3.9)$$

Assuming the relation*

$$\underline{E}_i^{(p+1)} = \Theta_i\, \underline{E}_i^{(p)} = \Theta_i^2\, \underline{E}_i^{(p-1)} \qquad (3.10)$$

we have

$$\underline{E}_i^{(p+1)} = \left(1 + \alpha - \omega + \omega\lambda_i - \frac{\alpha}{\Theta_i}\right) \underline{E}_i^{(p)} \qquad (3.11)$$

or

$$\Theta_i^2 - (1 + \alpha - \omega + \omega\lambda_i)\Theta_i + \alpha = 0$$

which has the solution

$$\Theta_i = \frac{1}{2}\left[(1 + \alpha - \omega + \omega\lambda_i) \pm \sqrt{(1 + \alpha - \omega + \omega\lambda_i)^2 - 4\alpha}\,\right] (3.12)$$

If α and ω are chosen so that the expression under the square root is negative for <u>all</u> λ_i, then all Θ_i will be complex and, more importantly, all $|\Theta_i|$ will be identical.

*We are assuming that the second order procedure of (3.9) is equivalent to applying some matrix with eigenvalues Θ_i to only the previous iterate. We then choose the parameters such that this is true.

In particular, the choices

$$\begin{aligned} 1 + \alpha - \omega + \omega\lambda_m &= 2\sqrt{\alpha} \\ 1 + \alpha - \omega + \omega\lambda_o &= -2\sqrt{\alpha} \end{aligned} \tag{3.13}$$

give α and ω

$$\begin{aligned} \alpha &= \left(\frac{\sqrt{1 - \lambda_o} - \sqrt{1 - \lambda_m}}{\sqrt{1 - \lambda_o} + \sqrt{1 - \lambda_m}}\right)^2 \\ \omega &= \left(\frac{2}{\sqrt{1 - \lambda_o} + \sqrt{1 - \lambda_m}}\right)^2 \end{aligned} \tag{3.14}$$

This choice of parameters, which makes the square root of (3.12) negative, or zero, yields

$$|\Theta_i| = \sqrt{\alpha} \tag{3.15}$$

The eigenvalues Θ_i are complex, but all have the same absolute value. Thus, all error modes are decreased at the same rate. Note that $|\Theta_i|$ is always smaller than the maximum of $|\lambda_o|$ and $|\lambda_m|$. Consequently, the procedure is always more rapidly convergent than the Simultaneous Iteration.

If estimates of the minimum and maximum eigenvalues (λ_o, λ_m) for Eq. (3.14) are not exact, it is best to estimate λ_o lower and λ_m higher to assure the Θ_i being complex and therefore all modes decaying at the same rate.

Actually, the situation of Eq. (3.10) is never achieved (except asymptotically) because of the first mesh sweep. We estimate the initial error $\underline{E}^{(o)}$ but require a second iterate $\underline{E}^{(-1)}$ in Eq. (3.9) differing from $\underline{E}^{(o)}$ by Eq. (3.10). As only the norm of Θ_i is known, the initial iterates cannot be picked appropriately. Thus the convergence rate indicated by Eq. (3.15) is approached asymptotically only as $p \to \infty$.

In the literature, this method has generally been called Second Order Richardson. We will refer to it here as Second Order Simultaneous Iteration.

(The authors were unable to find an advantage in third order extrapolation. The optimum scheme would seem to have been obtained using a second order extrapolation.)

Extrapolated Successive Iteration

The extension of Successive Iteration, Eq. (2.5), can be achieved in the same way as for the Simultaneous Iteration, by

$$\underline{x}^{(p+1)} = \underline{x}^{(p)} - \omega\left[\underline{x}^{(p)} + L\underline{x}^{(p+1)} + U\underline{x}^{(p)} + \underline{s}\right] , \qquad (3.16)$$

where ω is the extrapolation parameter. As for the basic iteration, Eq. (3.16) is obtained when the latest estimates of the unknowns are used in the Extrapolated Simultaneous Iteration, Eq. (3.2). Successive Iteration is obtained whenever $\omega = 1$, and the terms Under- or Over-relaxation apply as ω is less than or greater than 1. Equation (3.16) can be rewritten as

$$\underline{x}^{(p+1)} = (I + \omega L)^{-1} \left[(1 - \omega)I - \omega U\right]\underline{x}^{(p)} - (I + \omega L)^{-1} \omega\underline{s} \quad (3.17)$$

For the error we have

$$\underline{E}^{(p+1)} = (I + \omega L)^{-1} \left[(1 - \omega)I - \omega U\right]\underline{E}^{(p)} \quad (3.18)$$

There seems to be no clear procedure for writing (3.18) in terms of the eigenvalues λ_i. Consequently, it is almost impossible to optimize the iteration procedure in the general case. However, for an important particular case, when the matrix A results from using a five-point difference equation, the matrix of (3.18) is such that ω can be optimized. In that instance, Extrapolated Successive Iteration becomes a powerful method, as will be seen in Section 7.

4. THE METHOD OF TSCHEBYSCHEFF POLYNOMIALS

In Simultaneous Iteration, the repeated application of a matrix M to an error vector results in the application of M^p to the initial error vector. We now consider the application of a general polynomal $G_p(M)$ and find that Tschebyscheff Polynomals are optimum for convergence.

In Simultaneous Iteration, the iteration matrix M = -(L + U) is applied to each iterate

$$\underline{x}^{(p)} = M\underline{x}^{(p-1)} - \underline{s} \tag{4.1}$$

The true solution, $\underline{x}$, satisfies Eq. (4.1) exactly; i.e.,

$$\underline{x} = M\underline{x} - \underline{s}$$

As in Section 2, the error, $\underline{E}^{(p)}$, associated with the iterate, $\underline{x}^{(p)}$, is defined by

$$\underline{E}^{(p)} = x - x^{(p)}$$

and satisfies

$$\underline{E}^{(p)} = M\underline{E}^{(p-1)} \tag{4.2}$$

Writing $\underline{x}^{(p)}$ and $\underline{E}^{(p)}$ in terms of the initial approximation $\underline{x}^{(o)}$ and $\underline{E}^{(o)}$ gives

$$\underline{x}^{(p)} = M^p \, \underline{x}^{(o)} - (1 + M + M^2 + \ldots + M^{p-1})\underline{s} \tag{4.3}$$

and

$$\underline{E}^{(p)} = M^p \, \underline{E}^{(o)} \tag{4.4}$$

For convergence, all eigenvalues of M must be less than 1.

Consider the case where the pth degree polynomial in M, $G_p(M)$, is applied to the initial error, rather than the particular polynomial, M^p of Eq. (4.4). Then we would have

$$\underline{E}^{(p)} = G_p(M) \, \underline{E}^{(o)} \tag{4.5}$$

This will give an improvement in convergence rate if the polynomial $G_p(M)$ reduces the slowly decaying modes of M more rapidly than does M^p. It is rather well known that such a polynomial exists and is given in terms of Tschebyscheff's polynomials (Refs. 5 and 7). First, however, we discuss methods whereby a general polynomial can be applied.

Methods

A semi-iterative procedure for applying a polynomial to an initial approximation consists of generating the Simultaneous Iteration iterates and forming a polynomial as a linear combination of them. Considering only the error, denote the Simultaneous Iteration iterates by $\tilde{\underline{E}}^{(p)}$, and the resultant approximation by $\underline{E}^{(p)}$. We have

$$\tilde{\underline{E}}^{(p)} = M\tilde{\underline{E}}^{(p-1)} = M^{p}\,\tilde{\underline{E}}^{(o)} \qquad (4.6)$$

and

$$\underline{E}^{(p)} = \sum_{i=0}^{p} a_i\,\tilde{\underline{E}}^{(i)} = \sum_{i=1}^{p} a_i\,M^{i}\,\underline{E}^{(o)} \qquad (4.7)$$

$$= G_p(M)\,\underline{E}^{(o)} \qquad (4.8)$$

where the a_i are real coefficients. The polynomial obtained depends upon the choice of coefficients a_i.

A second procedure for generating the polynomial is purely iterative in nature; namely,

$$E^{(p)} = \underline{E}^{(p-1)} + \omega_p \left[M\underline{E}^{(p-1)} - \underline{E}^{(p-1)} \right] \tag{4.9}$$

$$= (1 - \omega_p \, M - \omega_p) \, \underline{E}^{(p-1)} \tag{4.10}$$

$$= \prod_{i=1}^{p} (1 - \omega_i \, M - \omega_i) \, \underline{E}^{(o)} \tag{4.11}$$

$$= G_p(M) \, \underline{E}^{(o)} \tag{4.12}$$

Again, the $G_p(M)$ is completely dependent upon the choice of parameters ω_i.

In both of the above cases, the generating polynomial has the property

$$G_p(I) = 1 \tag{4.13}$$

In actual computation we do not work with the error, but with the function itself. Using similar notation, the semi-iterative method has the form

$$\underline{\tilde{x}}^{(p)} = M\underline{\tilde{x}}^{(p-1)} - \underline{s} \tag{4.14}$$

$$\underline{x}^{(p)} = \sum_{i=0}^{p} a_i \, \underline{\tilde{x}}^{(i)} \tag{4.15}$$

and the iterative procedure*

$$\underline{x}^{(p)} = \underline{x}^{(p-1)} + \omega_p \left[M\underline{x}^{(p-1)} - \underline{x}^{(p-1)} - s \right] \tag{4.16}$$

There are other similar iterative procedures which lead to a polynomial in M operating on an initial error.

Tschebyscheff Polynomials

Flanders and Shortley (Ref. 5) prove the following theorem:

Theorem: The pth degree polynomial in ω, $S_p(\omega)$, having the minimum-maximum absolute value for ω real and in the range $-1 < \omega < +1$, and normalized such that $S_p(\omega_o) = 1$ for some $\omega_o > 1$ is given by

$$S_p(\omega) = \frac{T_p(\omega)}{T_p(\omega_o)} \tag{4.17}$$

where $T_p(\omega)$ is the Tschebyscheff Polynomial. The denominator of Eq. (4.17) is constant and achieves the desired normalization.

The Tschebyscheff Polynomials, $T_p(\omega)$, are defined by (Ref. 10)

$$\begin{aligned} T_p(\omega) &= \cos\,(p \cos^{-1} \omega) \qquad |\omega| \leq 1 \\ T_p(\omega) &= \cosh\,(p \cosh^{-1} \omega) \quad |\omega| \geq 1 \end{aligned} \tag{4.18}$$

*Essentially Extrapolated Simultaneous Iteration, Eq. (3.4) with variable ω.

$T_p(\omega)$ can be expressed as a polynomial

$$T_p(\omega) = \omega^p - \binom{p}{2}\omega^{p-2}(1 - \omega^2) + \binom{p}{4}\omega^{p-4}(1 - \omega^2)^2$$

$$- \binom{p}{6}\omega^{p-6}(1 - \omega^2)^3 + \ldots \qquad (4.19)$$

or, using simple trigonometric identities, can be written in the following recursion relation!

$$\begin{aligned} T_p(\omega) &= 2\omega T_{p-1}(\omega) - T_{p-2}(\omega) \\ T_o(\omega) &= 1 \qquad\qquad (4.20) \\ T_1(\omega) &= \omega \end{aligned}$$

From (4.18) we have

$$\begin{aligned} T_p(-1) &= +1 \qquad p \text{ even} \\ T_p(-1) &= -1 \qquad p \text{ odd} \qquad\qquad (4.21) \\ T_p(1) &= +1 \qquad p \text{ even or odd} \end{aligned}$$

Optimum Polynomial

Using the previous theorem, the optimum polynomial can be determined. The matrix M has eigenvalues bounded absolutely by 1 (otherwise the Simultaneous Iteration would diverge).

If, further, the assumption is made that all the eigenvalues of M are real, then

$$-1 < b \leq \lambda \leq a < 1 \tag{4.22}$$

for all λ. Thus the conditions of the theorem can be obtained by shifting the range of the interval (a,b) to (-1,1), using the transformation

$$\begin{aligned} U(M) &= \frac{2M - a - b}{a - b} \\ \mu(\lambda) &= \frac{2\lambda - a - b}{a - b} \end{aligned} \tag{4.23}$$

where $\mu(\lambda)$ are the eigenvalues of the matrix U(M) corresponding to λ of M. From Eq. (4.23) one notes that

$$\begin{aligned} \mu(a) &= +1 \\ \mu(b) &= -1 \\ \mu(1) &= \frac{2 - a - b}{a - b} > 1 \end{aligned} \tag{4.24}$$

Using the theorem and the normalization of Eq. (4.13), the polynomial to be determined is

$$G_p(M) = \frac{T_p[U(M)]}{T_p[U(I)]} = \frac{T_p\left[\frac{2M - (a + b)I}{a - b}\right]}{T_p\left[\frac{2 - a - b}{a - b}\right]} \tag{4.25}$$

and the error for the pth iterate, from Eq. (4.5), is

$$\underline{E}^{(p)} = \frac{T_p\left(\frac{2M - a - b}{a - b}\right)}{T_p\left(\frac{2 - a - b}{a - b}\right)} \underline{E}^{(o)} \tag{4.26}$$

The maximum absolute value of the numerator of Eq. (4.26) is 1. Thus, convergence is governed by the denominator, and

$$\underline{E}^{(p)} \leq \left[T_p\left(\frac{2 - a - b}{a - b}\right)\right]^{-1} \underline{E}^{(o)} \tag{4.27}$$

In comparison, Simultaneous Iteration gives

$$\underline{E}^{(p)} \leq a^p \, \underline{E}^{(o)} \tag{4.28}$$

D. Young (Ref. 6) established that the convergence rate of (4.27) is <u>at least</u> a factor $\sqrt{2}$ better than that of (4.28). Actually, it might be much better.

The polynomial method, on the other hand, requires more work because the coefficients, as well as the iterates, must be calculated. This is not very significant. Shortley (Ref. 7) estimates that the computing time is of the order of $N^{1/2}$, where N is the number of equations of the system. This tends to be optimistic, but is somewhat more qualitative.

If the matrix A, Eq. (2.2), is symmetric (or can be made so by premultiplication by a positive definite matrix), then the M obtained for Simultaneous Iteration (2.4) has real eigenvalues. If the Successive Iteration matrix has real eigenvalues, the application of Tschebyscheff Polynomials results in an increased convergence rate. In general, however, it is not clear that the eigenvalues of the Successive Iteration matrix are real.

Choice of Parameters

In order to generate the Tschebyscheff Polynomials, the iteration parameters of Eq. (4.7) and Eq. (4.9) must be determined.

The coefficients of the semi-iterative procedure are obtained by expressing the Tschebyscheff Polynomials in polynomial form and matching coefficients.

For the iterative procedure, Eq. (4.9), the roots of the Tschebyscheff Polynomial and the generating polynomial, Eq. (4.11), are matched.

The roots of $T_p(\mu)$, $|\mu| < 1$, are

$$\mu = \cos \frac{2k + 1}{2p} \pi \qquad k = 0, 1, 2, \ldots, p - 1 \qquad (4.29)$$

Thus, for the transformed roots, Eq. (4.23), we have

$$\mu = \frac{2\lambda - a - b}{a - b} = \cos \frac{2k + 1}{2p} \pi$$

$$k = 0, 1, 2, \ldots, p - 1 \qquad (4.30)$$

or

$$\lambda = \frac{1}{2} \left[a + b + (a - b) \cos \frac{2k + 1}{2p} \pi \right]$$

$$k = 0, 1, 2, \ldots, p - 1 \qquad (4.31)$$

Similarly, the roots of the general polynomial, Eq. (4.11), occur whenever

$$1 + \omega_k \lambda - \omega_k = 0 \qquad (4.32)$$

or

$$\lambda = \frac{\omega_k - 1}{\omega_k} \qquad (4.33)$$

Matching Eq. (4.31) and Eq. (4.33) gives

$$\omega_k = \left\{ 1 - \frac{1}{2} \left[a + b + (a - b) \cos \frac{2k + 1}{2p} \right] \right\}^{-1}$$

$$k = 0, 1, 2, \ldots, p - 1 \qquad (4.34)$$

Both procedures have different coefficients for different orders p of the polynomial. The usual procedure is to estimate in advance the required p and then calculate the appropriate coefficients. If it turns out that the estimated p is not large enough to give the required error reduction, then the cycle can be repeated. This is not, however, as good as generating the larger polynomial initially, and is clearly a disadvantage of the method.

A second disadvantage is the susceptibility of the method to large round-off error (Ref. 7). This occurs because the coefficients ω_i and a_i may be very large.

One means of generating the Tschebyscheff Polynomials which eliminates both of these problems is to use a second order iteration scheme (Ref. 4, p. 9, and Ref. 24), such as

$$\underline{x}^{(p)} = \underline{x}^{(p-1)} + \alpha_p\left[M\underline{x}^{(p-1)} - \underline{x}^{(p-1)} - \underline{s}\right] + \beta_p\left[\underline{x}^{(p-1)} - \underline{x}^{(p-2)}\right] \qquad (4.35)$$

Such a form is plausible because of the known recursion relation $T_p(\omega)$, Eq. (4.20). Again the error reduction is taken as in Eq. (4.26)

$$\underline{E}^{(p)} = \frac{T_p[U(M)]}{T_p[U(I)]}\,\underline{E}^{(o)} \qquad (4.36)$$

which, when substituted into the error equation resulting from Eq. (4.35), yields, upon rearranging,

$$T_p[U(M)] = (1 + \alpha_p M - \alpha_p + \beta_p) \frac{T_p[U(I)]}{T_{p-1}[U(I)]} \; T_{p-1}[U(M)]$$

$$- \beta_p \frac{T_p[U(I)]}{T_{p-2}[U(I)]} \; T_{p-2}[U(M)] \tag{4.37}$$

Equating the coefficients of Eq. (4.37) with those of the recursion relation gives

$$\alpha_p = \frac{4T_{p-1}[U(I)]}{(a - b)T_p[U(I)]} = \frac{4T_{p-1}\left(\frac{2 - a - b}{a - b}\right)}{(a - b)T_p\left(\frac{2 - a - b}{a - b}\right)} \tag{4.38}$$

and

$$\beta_p = \frac{T_{p-2}[U(I)]}{T_p[U(I)]} = \frac{T_{p-2}\left(\frac{2 - a - b}{a - b}\right)}{T_p\left(\frac{2 - a - b}{a - b}\right)} \tag{4.39}$$

The first two values of the Tschebyscheff Polynomials (4.20) are satisfied if the following choices are made

$$\alpha_1 = \frac{2}{2 - a - b}$$

$$\tag{4.40}$$

$$\beta_1 = 0$$

Using the values of α_p and β_p given by Eqs. (4.38), (4.39), and (4.40), a set of iterates $\underline{x}^{(p)}$ is generated in Eq. (4.35) such that Eq. (4.36) is obtained at each step. Round-off difficulties are greatly reduced, as the coefficients are order 1.

5. BLOCK OR IMPLICIT METHODS

It is sometimes advantageous to advance a set or block of unknowns simultaneously. The equations are implicit if each of the unknowns of the block depends upon the other unknowns of the block. The method of solving these implicit equations depends, of course, upon the structure of the matrix. Usually they must be solved simultaneously. In that case, the blocks should be chosen so that the solution is simple and the round-off errors are small.

The ideal procedure would be to obtain the exact solution by solving all the equations simultaneously as a single block. However, this is impractical for a general system of N^2 non-zero coefficients because of round-off errors. On the other hand, if there are of the order of only N non-zero coefficients, such a procedure might prove feasible if round-off is not large.

If each block consists of just one unknown, then the procedure is identical to the "point" methods of the previous sections. Also, if the equations of a block are uncoupled (i.e., each unknown does not depend upon the other unknowns in the block), then each unknown can be written explicitly in terms of previous iterates and the method is identical to the corresponding point method (Ref. 17).

One can have simultaneous and successive block methods as discussed in Section 2, as well as extrapolation of these basic procedures as in Section 3.

As an example, assume that each block involves k unknowns. Order Eq. (2.1) such that the first k equations have the coefficients of k unknowns of the first block on the main diagonal,* and the second set of k equations has the coefficients of the k unknowns of the second block on the main diagonal, etc. Then Eq. (2.2) can be written (Ref. 17) as

$$A\underline{x} + \underline{s} = (D + \tilde{L} + \tilde{U})\underline{x} + \underline{s} = 0 \tag{5.1}$$

where the matrix D is formed from the (k x k) matrices of coefficients of unknowns in the blocks. The matrices $\tilde{L}$ and $\tilde{U}$ are the remaining lower and upper triangular matrices which couple unknowns within a block to unknowns outside the block.

Corresponding to Eq. (2.3) for the basic form, we have from Eq. (5.1)

$$D\underline{x} = -(\tilde{L} + \tilde{U})\underline{x} - \underline{s} \tag{5.2}$$

Analogous to Eq. (2.4) for Simultaneous Iteration, we have Simultaneous Block Iteration

$$D\underline{x}^{(p+1)} = -(\tilde{L} + \tilde{U})\underline{x}^{(p)} - \underline{s} \tag{5.3}$$

*As discussed in Section 1, each equation distinguishes an unknown, the coefficient of which appears on the main diagonal of the matrix A of Eq. (2.2).

or

$$\underline{x}^{(p+1)} = -D^{-1}(\tilde{L} + \tilde{U})\underline{x}^{(p)} - D^{-1}\,\underline{s} \tag{5.4}$$

Similarly, from Eq. (2.5) for Successive Iteration, we have for Successive Block Iteration

$$D\underline{x}^{(p+1)} = -\tilde{L}\underline{x}^{(p+1)} - \tilde{U}\underline{x}^{(p)} - \underline{s} \tag{5.5}$$

or

$$\underline{x}^{(p+1)} = -(D + \tilde{L})^{-1}\,\tilde{U}\underline{x}^{(p)} - (D + \tilde{L})^{-1}\,\underline{s} \tag{5.6}$$

As before, the error obeys these same equations with the source term absent. The characteristic equations for the error matrices become

Simultaneous Block Iteration: $$\det\left|\tilde{L} + \tilde{U} + \lambda D\right| = 0 \tag{5.7}$$

Successive Block Iteration: $$\det\left|\tilde{U} + \lambda(D + \tilde{L})\right| = 0 \tag{5.8}$$

with arrays corresponding to Eqs. (2.13) and (2.14). In particular, if $k = 2$ (still assuming $a_{ii} = 1$)

Simultaneous Block Iteration:

$$\begin{vmatrix} \lambda & \lambda a_{12} & a_{13} & a_{14} & \cdot & \cdot & \cdot & \cdot & a_{1n} \\ \lambda a_{21} & \lambda & a_{23} & a_{24} & \cdot & \cdot & \cdot & \cdot & a_{2n} \\ a_{31} & a_{32} & \lambda & \lambda a_{34} & \cdot & \cdot & \cdot & \cdot & \\ a_{41} & a_{42} & \lambda a_{43} & \lambda & & & & & \\ \cdot & \cdot & & & & & & & \lambda a_{n-1,n} \\ \cdot & \cdot & & & & & \lambda & & \\ a_{n1} & a_{n2} & & & & & \lambda a_{n,n-1} & & \lambda \end{vmatrix} = 0 \quad (5.9)$$

Successive Block Iteration:

$$\begin{vmatrix} \lambda & \lambda a_{12} & a_{13} & a_{14} & \cdot & \cdot & a_{1n} \\ \lambda a_{21} & \lambda & a_{23} & a_{24} & \cdot & \cdot & a_{2n} \\ \lambda a_{31} & a_{32} & \lambda & \lambda a_{34} & & & \\ \lambda a_{41} & \lambda a_{42} & \lambda a_{43} & & & & \\ \cdot & \cdot & & & & & \lambda a_{n-1,n} \\ \cdot & \cdot & & & & \lambda & \\ \lambda a_{n1} & \lambda a_{n2} & & & & \lambda a_{n,n-1} & \lambda \end{vmatrix} = 0 \quad (5.10)$$

The basic block methods can be extrapolated as in Section 3 for the point methods. Thus for Extrapolated Simultaneous Block Iteration, we have

$$D\underline{x}^{(p+1)} = D\underline{x}^{(p)} - \omega \left[A\underline{x}^{(p)} + \underline{s} \right] \tag{5.11}$$

or

$$\underline{x}^{(p+1)} = \underline{x}^{(p)} - \omega D^{-1} \left[A\underline{x}^{(p)} + \underline{s} \right] \tag{5.12}$$

and for the Second Order Simultaneous Block Iteration

$$\underline{x}^{(p+1)} = \underline{x}^{(p)} + \alpha \left[\underline{x}^{(p)} - \underline{x}^{(p-1)} \right] - \omega D^{-1} \left[A\underline{x}^{(p)} + \underline{s} \right] \tag{5.13}$$

Similarly, for Extrapolated Successive Block Iteration

$$\underline{x}^{(p+1)} = \underline{x}^{(p)} - \omega D^{-1} \left[D\underline{x}^{(p)} + \tilde{L}\underline{x}^{(p+1)} + \tilde{U}\underline{x}^{(p)} + \underline{s} \right] \tag{5.14}$$

The parameters and eigenvalues and the extrapolated procedures of Section 3 were dependent on the eigenvalues of the Simultaneous Iteration matrix equation (2.13). Similarly, the parameters of the extrapolated implicit methods are dependent on the eigenvalues of the Simultaneous Block Iteration matrix equation (5.9). In fact, the parameters and eigenvalues are the same functions of the Simultaneous Block Iteration eigenvalues as the corresponding parameters and eigenvalues of the point extrapolations are of the Simultaneous Iteration eigenvalues.

The extension to Polynomial Implicit Methods is immediate if Eqs. (5.1) and (5.2) are used, rather than Eq. (4.1). The parameters obtained are those of Section 4 with the Simultaneous Block Iteration eigenvalues replacing those of the Simultaneous Iteration.

6. VARIATIONAL METHODS

In considering the variational methods for solving the system given by Eq. (2.1)

$$A\underline{x} + \underline{s} = 0 \tag{6.1}$$

we assume that the matrix A is real, symmetric, and positive definite; i.e.,

$$a_{ij} = \overline{a}_{ij} = a_{ji}$$

and for any $x \neq 0$

$$\sum_{i,j} a_{ij}\, x_i\, x_j > 0 \tag{6.2}$$

where the bar denotes complex conjugation. An arbitrary vector, $\underline{x}^{(p)}$, will not, in general, be an exact solution of (6.1) but will generate a residue, $\underline{r}^{(p)}$, defined as

$$\underline{r}^{(p)} = A\underline{x}^{(p)} + \underline{s} \tag{6.3}$$

If A is symmetric, the system of Eq. (6.1) can be written as the gradient of a quadratic function

$$F(\underline{x}) = \frac{1}{2}\,\underline{x} \cdot A\underline{x} + \underline{s} \cdot \underline{x} \tag{6.4}$$

Solving Eq. (6.1) is equivalent to minimizing $F(\underline{x})$. In particular, for an arbitrary trial vector, $\underline{x}^{(p)}$, the residue is

$$\underline{r}^{(p)} = A\underline{x}^{(p)} + \underline{s} = \underline{\text{Grad}}\; F\left[\underline{x}^{(p)}\right] \tag{6.5}$$

By Eq. (6.5), the residue, $\underline{r}^{(p)}$, is normal to the surface of the ellipsoid defined by Eq. (6.4) in the N-dimensional space of the elements of $\underline{x}$.

If $\underline{m}^{(p)}$ is some arbitrary direction and ω_p some arbitrary constant dependent upon p, then the iteration scheme for $\underline{x}^{(p+1)}$ can be defined as

$$\underline{x}^{(p+1)} = \underline{x}^{(p)} + \omega_p\, \underline{m}^{(p)} \tag{6.6}$$

Equation (6.6) states that the new iterate is given by the old iterate plus some "correction" vector.

The variational methods amount to choosing an ω_p such that the quadratic function $F\left[\underline{x}^{(p+1)}\right]$, given by Eq. (6.4), will be a minimum for a given direction $\underline{m}^{(p)}$. Consider Fig. 6.1, which represents the intersection of a plane defined by $\underline{m}^{(p)}$ and $\underline{r}^{(p)}$ and the surface $F(\underline{x})$ = constant.

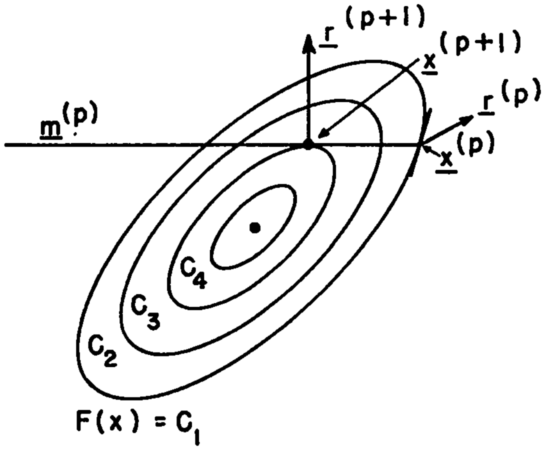

Fig. 6.1

The tip of the vector $\underline{x}^{(p)}$ appears as a point in this plane.

The minimum $F\left[x^{(p+1)}\right]$ is tangent to the direction $\underline{m}^{(p)}$ and the new residue, $\underline{r}^{(p+1)}$, is thus normal to $\underline{m}^{(p)}$. We have

$$\underline{r}^{(p+1)} - \underline{r}^{(p)} = A\underline{x}^{(p+1)} + \underline{s} - \left[A\underline{x}^{(p)} + \underline{s}\right] = A\left[\underline{x}^{(p+1)} - \underline{x}^{(p)}\right]$$

or by Eq. (6.6)

$$\underline{r}^{(p+1)} = \underline{r}^{(p)} + \omega_p \, A\underline{m}^{(p)} \tag{6.7}$$

Choosing $\underline{r}^{(p+1)}$ and $\underline{m}^{(p)}$ as above implies

$$\underline{r}^{(p+1)} \cdot \underline{m}^{(p)} = \underline{r}^{(p)} \cdot \underline{m}^{(p)} + \omega_p \, A\underline{m}^{(p)} \cdot \underline{m}^{(p)} = 0 \tag{6.8}$$

which determines

$$\omega_p = - \frac{\underline{r}^{(p)} \cdot \underline{m}^{(p)}}{\underline{m}^{(p)} \cdot A\underline{m}^{(p)}} \tag{6.9}$$

This choice of ω_p systematically reduces $F\left[x^{(p+1)}\right]$, and the method is convergent for any given $\underline{m}^{(p)}$.

The choice of $\underline{m}^{(p)}$ differentiates the methods; a few of them will now be discussed.

Southwell's Relaxation Method

If the direction $\underline{m}^{(p)}$ is chosen as one of the coordinates of the space (e_i) on which the matrix A defines a linear transformation, then

$$\underline{x}^{(p+1)} = \underline{x}^{(p)} - \frac{\left[\underline{r}_i^{(p)} \cdot e_i\right]}{a_{ii}} e_i \tag{6.10}$$

Thus, the residue of the ith unknown is eliminated and other values are undisturbed. Usually, the e_i with the largest residue is chosen, but the choice depends upon the individual performing the computation.

When the coordinates (e_i) are all used in a fixed order, the method is identioal to Successive Iteration of Eq. (2.5). Successive Iteration is, thus, just systematic relaxation, and is suitable for high speed computers, whereas ordinary relaxation is more appropriate for hand computation.

Method of Steepest Descent

Let direction $\underline{m}^{(p)}$ be chosen equal to $\underline{r}^{(p)}$. This is normal to the ellipsoid at the point $\underline{x}^{(p)}$ and, as such, is in the direction of steepest change of the function $F\left[\underline{x}^{(p)}\right]$ = constant. A useful diagram is obtained by passing a two-dimensional plane through the residue, $\underline{r}^{(p)}$. A manifold of ellipses with a common center is formed by the intersection of the plane and the surfaces $F(\underline{x})$ = constant (see Fig. 6.2).

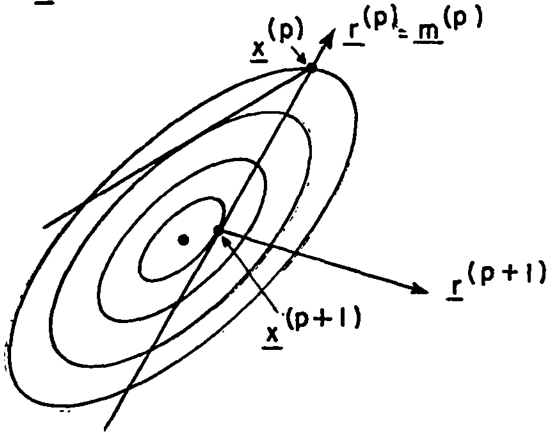

Fig. 6.2

From Eqs. (6.6), (6.7), and (6.9), we have

$$\underline{x}^{(p+1)} = \underline{x}^{(p)} + \omega_p \, \underline{r}^{(p)} \tag{6.11}$$

$$\omega_p = \frac{\underline{r}^{(p)} \cdot \underline{r}^{(p)}}{\underline{r}^{(p)} \cdot A\underline{r}^{(p)}} \tag{6.12}$$

$$\underline{r}^{(p+1)} = \underline{r}^{(p)} + \omega_p \, Ar^{(p)} \tag{6.13}$$

This choice of $\underline{m}^{(p)}$ defines the Simultaneous Extrapolated procedure of Section 3, with the parameter ω_p here variable rather than constant.

Method of Conjugate Gradients (C-G)

A better choice of the direction $m^{(p)}$ is $\underline{a}^{(p)}$, directed from a point on the ellipse towards the center of the ellipses. Then $\underline{a}^{(p)}$ and a vector $\underline{t}^{(p)}$, tangent to the ellipse at $\underline{x}^{(p)}$, define conjugate directions (Ref. 23); that is,

$$\underline{a}^{(p)} \cdot A\underline{t}^{(p)} = 0 \qquad (6.14)$$

They are orthogonal with respect to the matrix A. A useful diagram is obtained by passing a plane through $\underline{r}^{(p)}$ and $\underline{t}^{p}$, yielding a manifold of ellipses wherein the minimum $F\left[\underline{x}^{(p)}\right]$ occurs at the center of the ellipses in the plane.

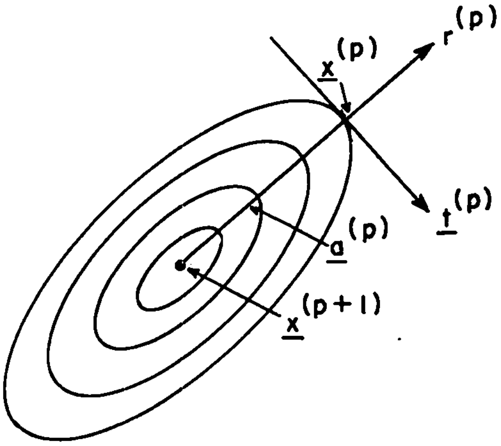

Fig. 6.3

The minimum $F\left[\underline{x}^{(p+1)}\right]$ occurs for an ω_p given from Eq. (6.9)

$$\omega_p = - \frac{\underline{r}^{(p)} \cdot \underline{a}^{(p)}}{\underline{a}^{(p)} \cdot A\underline{a}^{(p)}} \tag{6.15}$$

where, as before, the residue is given iteratively

$$\underline{r}^{(p+1)} = \underline{r}^{(p)} + \omega_p \, A\underline{a}^{(p)} \tag{6.16}$$

and the iteration scheme is

$$\underline{x}^{(p+1)} = \underline{x}^{(p)} + \omega_p \, \underline{a}^{(p)} \tag{6.17}$$

Thus, the new residue $\underline{r}^{(p+1)}$ is normal to $F\left[\underline{x}^{(p+1)}\right] = c$ at the point $\underline{x}^{(p+1)}$, the center of the plane at which point $F(\underline{x})$ is minimal. Since the plane is tangent to the surface at this point, the residue $\underline{r}^{(p+1)}$ is orthogonal to this plane and to the vectors $\underline{a}^{(p)}$, $\underline{r}^{(p)}$, and $\underline{t}^{(p)}$ which lie in the plane.

The vector $\underline{a}^{(p)}$ is easily found using Eq. (6.14). The plane is determined by the vectors $\underline{r}^{(p)}$ and $\underline{t}^{(p)}$. Thus, one could make the choice

$$\underline{a}^{(p)} = \underline{r}^{(p)} + \mu_p \, \underline{t}^{(p)} \tag{6.18}$$

Picking the parameter μ_p to satisfy Eq. (6.14), we have

$$\underline{a}^{(p)} \cdot A\underline{t}^{(p)} = \underline{r}^{(p)} \cdot A\underline{t}^{(p)} + \mu_p \, A\underline{t}^{(p)} = 0 \tag{6.19}$$

or, solving for μ_p,

$$\mu_p = - \frac{\underline{r}^{(p)} \cdot A\underline{t}^{(p)}}{\underline{t}^{(p)} \cdot A\underline{t}^{(p)}} \tag{6.20}$$

The resulting equations are simplified by realizing that $\underline{t}^{(p+1)}$, the new tangent vector, lies somewhere in the plane of $\underline{r}^{(p)}$ and $\underline{t}^{(p)}$. Thus, one might choose

$$\underline{t}^{(p+1)} = \underline{a}^{(p)} \tag{6.21}$$

with the result

$$\begin{aligned} \underline{a}^{(p)} &= \underline{r}^{(p)} + \mu_p \, \underline{a}^{(p-1)} \\ \mu_p &= - \frac{\underline{r}^{(p)} \cdot A\underline{a}^{(p-1)}}{\underline{a}^{(p-1)} \cdot A\underline{a}^{(p-1)}} \end{aligned} \tag{6.22a}$$

and the residue

$$\begin{aligned} \underline{r}^{(p+1)} &= \underline{r}^{(p)} + \omega_p \, A\underline{a}^{(p)} \\ \omega_p &= - \frac{\underline{r}^{(p)} \cdot \underline{a}^{(p)}}{\underline{a}^{(p)} \cdot A\underline{a}^{(p)}} \end{aligned} \tag{6.22b}$$

The iteration scheme is defined

$$\underline{x}^{(p+1)} = \underline{x}^{(p)} + \omega_p \, \underline{a}^{(p)} \tag{6.22c}$$

Equations (6.22a), (6.22b), and (6.22c) imply the orthogonality relations

$$\underline{r}^{(p+1)} \cdot \underline{r}^{(p)} = 0 \tag{6.23a}$$

$$\underline{r}^{(p+1)} \cdot \underline{a}^{(p)} = 0 \tag{6.23b}$$

$$\underline{r}^{(p+1)} \cdot \underline{a}^{(p-1)} = 0 \tag{6.23c}$$

$$\underline{a}^{(p)} \cdot A\underline{a}^{(p-1)} = 0 \tag{6.23d}$$

Likewise, it can be shown* that all $\underline{r}^{(p)}$ are mutually orthogonal and all $\underline{a}^{(p)}$ are mutually conjugate; i.e.

*Assume Eq. (6.24) holds through $\underline{r}^{(p)}$ and $\underline{a}^{(p-1)}$ and then show that $\underline{a}^{(p)}$ and $\underline{r}^{(p+1)}$ as defined by Eq. (6.22) satisfy (6.24) for each previous iterate.

$$\begin{aligned} &\underline{r}^{(p)} \cdot \underline{r}^{(j)} = 0 \\ & \qquad\qquad p \neq j \\ &\underline{a}^{(p)} \cdot A\underline{a}^{(j)} = 0 \end{aligned} \tag{6.24}$$

In performing the iteration defined by Eq. (6.22), one picks an arbitrary vector $\underline{x}^{(o)}$ and computes $\underline{r}^{(o)}$, but in order that Eq. (6.24) hold for all $\underline{r}^{(p)}$, the assignment $\underline{a}^{(o)} = \underline{r}^{(o)}$ must be made; i.e., $\mu_o = 0$. The method is exact in N steps and at each step

$$\left|r^{p+1}\right| < \left|r^{p}\right| \tag{6.25}$$

The procedure amounts to passing N mutually orthogonal planes through the ellipsoid $F(\underline{x})$ and finding the center of the resultant manifold of ellipses in each plane.

With some simple algebraic manipulation, the relations of Eq. (6.22) can be rewritten as second order recursion formulas. We have

$$\underline{r}^{(p+1)} = \left(1 + \frac{\omega_p \mu_p}{\omega_{p-1}}\right) \underline{r}^{(p)} - \left(\frac{\omega_p \mu_p}{\omega_{p-1}}\right) \underline{r}^{(p-1)} + \omega_p Ar^{(p)} \tag{6.26a}$$

$$\underline{x}^{(p)} = \left(1 + \frac{\omega_p \mu_p}{\omega_{p-1}}\right) \underline{x}^{(p)} - \left(\frac{\omega_p \mu_p}{\omega_{p-1}}\right) \underline{x}^{(p-1)} + \omega_p(A\underline{x}^p + \underline{s}) \tag{6.26b}$$

$$\underline{a}^{(p+1)} = (1 + \mu_{p+1})\underline{a}^{(p)} - \mu_p \, \underline{a}^{(p-1)} + \omega_p \, A\underline{a}^{(p)} \qquad (6.26c)$$

We must still require, of course, that $\mu_o = 0$.

Note that Eq. (6.26b) is of the same form as the Second Order Simultaneous Iteration, Eq. (3.8), except coefficients are here variable. Also, it is of the same form as the second order procedure for generating the Tschebyscheff Polynomials, Eq. (4.35), where the coefficients are picked to generate the polynomial. Here they are picked to minimize the quadratic function $F(\underline{x})$.

Method of Conjugate Directions (C-D)

The method of Conjugate Directions is similar to the conjugate gradient method. It also gives the exact solution in N steps. The direction vectors are chosen just as before

$$\underline{a}^{(p)} \cdot A\underline{a}^{(j)} = 0 \qquad p \neq j \qquad (6.27)$$

However, Eq. (6.27) is the only restriction upon them. The direction vectors may be calculated at any time during the iteration (even before iteration is begun), which may often prove to be a convenience that the C-G method lacks. On the other hand, if fewer than N steps are used, the C-G method will probably give a more accurate result than the C-D method, even though both schemes improve the solution with each step and give the exact answer after N steps.

Although both the C-G and C-D methods theoretically give an exact answer in N steps, round-off errors prevent the residues from being truly orthogonal in actual practice. Thus, we will have $r^{(p-1)} \neq 0$, but one might expect it to be very close. If round-off is significant enough to cause the solution to become too inaccurate, then it is not clear what procedure should be followed. Iteration could be continued until $\underline{r}^{(p)}$ is sufficiently small or could be restarted with $\underline{x}^{(p+1)}$ as an initial guess.

For large systems of equations, the variational methods may not be best. This is because the scalar products needed for the calculation of the parameters require considerable time to form and may be severely in error due to round-off. Some computational experience should be acquired in order to make a fairer appraisal of the methods.

7. APPLICATIONS TO ELLIPTIC DIFFERENCE EQUATIONS

Elliptic difference equations result from differencing an elliptic differential equation at points of a mesh imposed upon the domain of the differential equation, associated with the imposed boundary conditions. A set of independent, simultaneous, inhomogeneous linear equations could result, one for each mesh point. The numerical solution of these equations represents a solution of the differential equation to the approximations inherent in the differencing used, and the imposed boundary conditions. The equations can be written in the form

$$\sum_{j=1}^{N} a_{ij} x_j + s_i = 0 \qquad i = 1, 2, \ldots, N \tag{7.1}$$

where N is the number of mesh points, and the source term, s_i, depends upon external sources in the problem as well as fixed non-zero boundary conditions. The coefficients a_{ij} are assumed to be real and to give the contribution of the point j to the point i in the differencing scheme. In addition to whatever physical conditions are imposed upon the a_{ij}, they must satisfy the following general conditions (Ref. 13):

(a) $a_{ii} \geq \sum_{j=1}^{N}{}' |a_{ij}|$, and for some i strict inequality holds.

(b) The matrix $A = (a_{ij})$ is irreducible, i.e., given any two non-empty, disjoint subsets S and T of the first N integers, W, such that $S + T = W$, then there exists some $a_{ij} \neq 0$ such that $i \epsilon S$ and $j \epsilon T$.*

(c) $a_{ii} \neq 0$ and for each equation can be chosen such that $a_{ii} > 0$.

It is easily shown (Ref. 1) that conditions (a) and (b) imply the non-singularity of the matrix A, i.e., non-vanishing determinant. Similarly, the matrix A must be positive definite. For, if λ is a negative real number, the matrix $A - \lambda I$ also satisfies conditions (a) and (b) above, and thus has non-zero determinant (Ref. 13). Hence, all eigenvalues of A are positive, and we have the additional property (assuming A is symmetric):**

(d) A is non-singular and positive definite.

As a simple example, consider Poisson's equation in a rectangular domain with zero boundary conditions along the edges

*This means that Eqs. (7.1) are coupled such that each unknown depends, perhaps indirectly, upon all others.

**The matrix A is symmetric if the coefficient in the difference equation, giving the contribution from the point i to the point j, is the same as the coefficient from the point j to the point i.

$$\frac{\partial^2\psi(x,y)}{\partial x^2} + \frac{\partial^2\psi(x,y)}{\partial y^2} = s(x,y) \qquad (7.2)$$

A mesh of k vertical lines and ℓ horizontal lines is constructed and Eq. (7.2) is differenced at each intersection (Fig. 7.1).

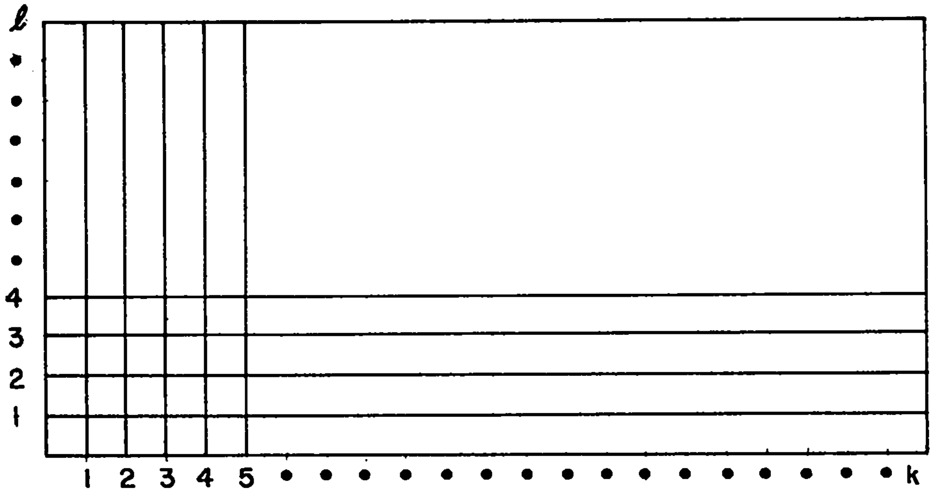

Fig. 7.1

Assuming equal spacing in both directions, define $(\Delta x)^2 = (\Delta y)^2 = h$, and for each interior mesh point we have

$$\frac{\psi_{k+1\ \ell} - 2\psi_{k\ell} + \psi_{k-1\ \ell}}{h} + \frac{\psi_{k\ \ell+1} - 2\psi_{k\ell} + \psi_{k\ \ell-1}}{h} = s_{k\ell} \qquad (7.3)$$

or

$$-\left(\frac{1}{h}\,\psi_{k+1\,\ell}+\frac{1}{h}\,\psi_{k-1\,\ell}+\frac{1}{h}\,\psi_{k\ell+1}+\frac{1}{h}\,\psi_{k\ell-1}\right)+\frac{4}{h}\,\psi_{k\ell}+s_{k\ell}=0 \qquad (7.4)$$

or, in terms of general coefficients

$$-\,(a_{k\ell}\,\psi_{k+1\,\ell}+b_{k\ell}\,\psi_{k-1\,\ell}+c_{k\ell}\,\psi_{k\,\ell+1}+d_{k\ell}\,\psi_{k\,\ell-1})$$

$$+\,e_{k\ell}\,\psi_{k\ell}+s_{k\ell}=0 \qquad (7.5)$$

where, for Eq. (7.4) the coefficients satisfy

$$a_{k\ell}=b_{k\ell}=c_{k\ell}=d_{k\ell}=\frac{1}{4}\,e_{k\ell}=\frac{1}{h} \qquad (7.6)$$

for the $y = (k - 2)(\ell - 2)$ internal points.

Labeling these points in some manner, 1, 2, ..., y, and writing Eq. (7.4) for each point, we would obtain an equation like Eq. (7.1). Condition (a) is satisfied, since the inequality holds for interior points adjacent to the boundaries. Since each equation is ultimately coupled to all the others, condition (b) is satisfied. Condition (c) is true from the form of Eq. (7.4).

For illustrative purposes, let us consider now a very small mesh ($k = 6$, $\ell = 5$) and write the equations in detail. There are twelve interior points, which could be labeled as in Fig. 7.2.

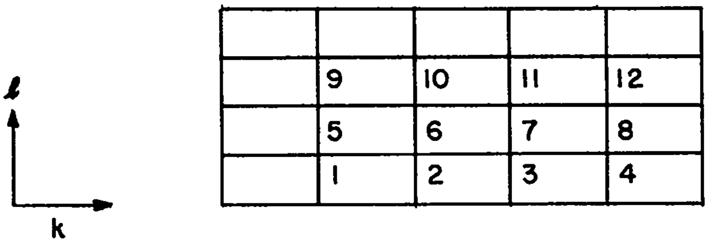

Fig. 7.2

Using this ordering of the points, the following matrix equation results

$$A\underline{\psi} + \underline{s} = 0 \tag{7.7}$$

Writing Eq. (7.7) in detail for Eq. (7.4), we have Eq. (7.8) (see separate page) or, in terms of Eq. (7.5), we have Eq. (7.9) (see separate page), where the coefficients satisfy Eq. (7.6). The form of Eq. (7.9) is typical of that obtained for the given ordering whenever a five-point differencing of an elliptic differential equation is made. There are two diagonals spaced k - 1 (interior points) to the right and left of the main diagonal, giving the effect of the points above and below, respectively, of the point considered in the difference equation. The matrix is symmetric, in the notation of Eq. (7.9), whenever

$$\begin{bmatrix}
4/h & -1/h & 0 & 0 & -1/h & & & & & & & \\
-1/h & 4/h & -1/h & 0 & 0 & -1/h & & & & & & \\
0 & -1/h & 4/h & -1/h & 0 & 0 & -1/h & & & & & \\
0 & 0 & -1/h & 4/h & 0 & 0 & 0 & -1/h & & & & \\
-1/h & 0 & 0 & 0 & 4/h & -1/h & 0 & 0 & -1/h & & & \\
 & -1/h & 0 & 0 & -1/h & 4/h & -1/h & 0 & 0 & -1/h & & \\
 & & -1/h & 0 & 0 & -1/h & 4/h & -1/h & 0 & 0 & -1/h & \\
 & & & -1/h & 0 & 0 & -1/h & 4/h & 0 & 0 & 0 & -1/h \\
 & & & & -1/h & 0 & 0 & 0 & 4/h & -1/h & 0 & 0 \\
 & & & & & -1/h & 0 & 0 & -1/h & 4/h & -1/h & 0 \\
 & & & & & & -1/h & 0 & 0 & -1/h & 4/h & -1/h \\
 & & & & & & & -1/h & 0 & 0 & -1/h & 4/h
\end{bmatrix}
\begin{bmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \\ \psi_4 \\ \psi_5 \\ \psi_6 \\ \psi_7 \\ \psi_8 \\ \psi_9 \\ \psi_{10} \\ \psi_{11} \\ \psi_{12} \end{bmatrix}
+
\begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \\ s_8 \\ s_9 \\ s_{10} \\ s_{11} \\ s_{12} \end{bmatrix}
= 0 \qquad (7.8)$$

$$\begin{bmatrix}
e_{11} & -a_{11} & 0 & 0 & -c_{11} & & & & & & & \\
-b_{21} & e_{21} & -a_{21} & 0 & 0 & -c_{21} & & & & & & \\
0 & -b_{31} & e_{31} & -a_{31} & 0 & 0 & -c_{31} & & & & & \\
0 & 0 & -b_{41} & e_{41} & 0 & 0 & 0 & -c_{41} & & & & \\
-d_{12} & 0 & 0 & 0 & e_{12} & -a_{12} & 0 & 0 & -c_{12} & & & \\
0 & -d_{22} & 0 & 0 & -b_{22} & e_{22} & -a_{22} & 0 & 0 & -c_{22} & & \\
 & & -d_{32} & 0 & 0 & -b_{32} & e_{32} & -a_{32} & 0 & 0 & -c_{32} & \\
 & & & -d_{42} & 0 & 0 & -b_{42} & e_{42} & 0 & 0 & 0 & -c_{42} \\
 & & & & -d_{13} & 0 & 0 & 0 & e_{13} & -a_{13} & 0 & 0 \\
 & & & & & -d_{23} & 0 & 0 & -b_{23} & e_{23} & -a_{23} & 0 \\
 & & & & & & -d_{33} & 0 & 0 & -b_{33} & e_{33} & -a_{33} \\
 & & & & & & & -d_{43} & 0 & 0 & -b_{43} & e_{43}
\end{bmatrix}
\begin{bmatrix}
\psi_{11} \\ \psi_{21} \\ \psi_{31} \\ \psi_{41} \\ \psi_{12} \\ \psi_{22} \\ \psi_{32} \\ \psi_{42} \\ \psi_{13} \\ \psi_{23} \\ \psi_{33} \\ \psi_{43}
\end{bmatrix}
+
\begin{bmatrix}
s_{11} \\ s_{21} \\ s_{31} \\ s_{41} \\ s_{12} \\ s_{22} \\ s_{32} \\ s_{42} \\ s_{13} \\ s_{23} \\ s_{33} \\ s_{43}
\end{bmatrix}
= 0 \tag{7.9}$$

$$\begin{aligned} &a_{k\ell} = b_{k+1\;\ell} \\ \text{and} \\ &c_{k\ell} = d_{k\;\ell+1} \end{aligned} \tag{7.10}$$

Or, in words, the coefficient of a point to a neighbor (in this case, one of the four surrounding points) is the same as the coefficient of the neighbor to the point. The effect of the fixed boundary condition is reflected in the absence of the terms $a_{K\ell}$, $b_{1\ell}$ in the matrix. Hence, again, condition (a) is satisfied.

In some instances, in other geometries, symmetry can be easily achieved, for instance, by premultiplication of the ith equation by the ith cell volume, but, of course, this depends upon the particular form chosen for the difference equations.

Young has shown that the matrix obtained by differencing an elliptic differential equation with a five-point differencing scheme possesses Property (A), which is defined (Ref. 13):

Definition: A matrix possesses Property (A) if there exist two disjoint subsets S and T of the first N integers W such that $S + T = W$, and if $a_{ij} \neq 0$, then either $i = j$ or $i \in S$ and $j \in T$ or $i \in T$ and $j \in S$.

With these properties of the matrix A, we consider the various methods described in the previous sections for solving the system $A\underline{x} + \underline{s} = 0$.

Basic Iterative Methods

Young (Ref. 13) has shown two very important consequences of any matrix possessing Property (A). First, the eigenvalues of Simultaneous Iteration as applied to such a matrix occur in ± pairs. Thus, the maximum and minimum eigenvalues have the same absolute value. Second, there are certain orderings of the points, called consistent orderings, in which the eigenvalues of the matrices of Simultaneous and Successive Iteration are related in a simple manner. More precisely, the characteristic determinants of Simultaneous and Successive Iteration [Eqs. (2.13) and (2.14)] can be written, for a consistently ordered matrix A, as follows.

Simultaneous Iteration:

$$f(\lambda) = \left| \begin{array}{cccccc} \lambda B_1 & C_1 & & & & \\ D_2 & \lambda B_2 & C_2 & & & \\ & D_3 & \lambda B_3 & C_3 & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & \cdot & C_{k-1} \\ & & & & D_k & \lambda B_k \end{array} \right| = 0 \qquad (7.11)$$

Successive Iteration:

$$f(\lambda) = \begin{vmatrix} \bar{\lambda}B_1 & C_1 & & & & \\ \bar{\lambda}D_2 & \bar{\lambda}B_2 & C_2 & & & \\ & \bar{\lambda}D_3 & \bar{\lambda}B_3 & C_3 & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & \cdot & C_{k-1} \\ & & & & \bar{\lambda}D_k & \bar{\lambda}B_k \end{vmatrix} = 0 \qquad (7.12)$$

where B_1, B_2, ..., B_k are square matrices and λ and $\bar{\lambda}$ are the eigenvalues of the respective methods.

For the example, Eq. (7.8), we can order the equations along the diagonals of the mesh (without altering the value of the determinant) and obtain Eq. (7.13) (see separate page). Likewise, for Eq. (7.9) we would have Eq. (7.14) (see separate page). In Eqs. (7.13) and (7.14), the square matrices of Eqs. (7.11) and (7.12) have been blocked off. Clearly, the characteristic equations of Simultaneous and Successive Iteration applied to Eq. (7.13) or, to the more general form of Eq. (7.14), are of the forms of Eq. (7.11) and (7.12). This is an indication that the ordering used in Eqs. (7.8) and (7.9) is a consistent ordering.

The convergence rate of Simultaneous Iteration is independent of the ordering of the points. For Successive Iteration, however, this is not generally true. Young has proved that the convergence rate of Successive

$$\begin{bmatrix}
4/h & -1/h & -1/h & & & & & & & & & \\
-1/h & 4/h & 0 & -1/h & -1/h & & & & & & & \\
-1/h & 0 & 4/h & 0 & -1/h & -1/h & & & & & & \\
 & -1/h & 0 & 4/h & 0 & 0 & -1/h & -1/h & & & & \\
 & -1/h & -1/h & 0 & 4/h & 0 & 0 & -1/h & -1/h & & & \\
 & & -1/h & 0 & 0 & 4/h & 0 & 0 & -1/h & & & \\
 & & & -1/h & 0 & 0 & 4/h & 0 & 0 & -1/h & & \\
 & & & -1/h & -1/h & 0 & 0 & 4/h & 0 & -1/h & -1/h & \\
 & & & & -1/h & -1/h & 0 & 0 & 4/h & 0 & -1/h & \\
 & & & & & & -1/h & -1/h & 0 & 4/h & 0 & -1/h \\
 & & & & & & & -1/h & -1/h & 0 & 4/h & -1/h \\
 & & & & & & & & & -1/h & -1/h & 4/h
\end{bmatrix}
\begin{bmatrix} \psi_1 \\ \psi_2 \\ \psi_5 \\ \psi_3 \\ \psi_6 \\ \psi_9 \\ \psi_4 \\ \psi_7 \\ \psi_{10} \\ \psi_8 \\ \psi_{11} \\ \psi_{12} \end{bmatrix}
+
\begin{bmatrix} s_1 \\ s_2 \\ s_5 \\ s_3 \\ s_6 \\ s_9 \\ s_4 \\ s_7 \\ s_{10} \\ s_8 \\ s_{11} \\ s_{12} \end{bmatrix}
= 0 \qquad (7.13)$$

$$\begin{bmatrix}
e_{11} & -a_{11} & -c_{11} & & & & & & & & & \\
-b_{21} & e_{21} & 0 & -a_{21} & -c_{21} & 0 & & & & & & \\
-d_{12} & 0 & e_{12} & 0 & -a_{12} & -c_{12} & & & & & & \\
 & -b_{31} & 0 & e_{31} & 0 & 0 & -a_{31} & -c_{31} & & & & \\
 & -d_{22} & -b_{22} & 0 & e_{22} & 0 & 0 & -a_{22} & -c_{22} & & & \\
 & 0 & -d_{13} & 0 & 0 & e_{13} & 0 & 0 & -a_{13} & & & \\
 & & & -b_{41} & 0 & 0 & e_{41} & 0 & 0 & -c_{41} & 0 & \\
 & & & -d_{32} & -b_{32} & 0 & 0 & e_{32} & 0 & -a_{32} & -c_{32} & \\
 & & & 0 & -d_{23} & -b_{23} & 0 & 0 & e_{23} & 0 & -a_{23} & \\
 & & & & & & -d_{42} & -b_{42} & 0 & e_{42} & 0 & -c_{42} \\
 & & & & & & 0 & -d_{33} & -b_{33} & 0 & e_{33} & -a_{33} \\
 & & & & & & & & & -d_{43} & -b_{43} & e_{43}
\end{bmatrix}
\begin{bmatrix}
\psi_{11} \\ \psi_{21} \\ \psi_{12} \\ \psi_{31} \\ \psi_{22} \\ \psi_{13} \\ \psi_{41} \\ \psi_{32} \\ \psi_{23} \\ \psi_{42} \\ \psi_{33} \\ \psi_{43}
\end{bmatrix}
+
\begin{bmatrix}
s_{11} \\ s_{21} \\ s_{12} \\ s_{31} \\ s_{22} \\ s_{13} \\ s_{41} \\ s_{32} \\ s_{23} \\ s_{42} \\ s_{33} \\ s_{43}
\end{bmatrix}
= 0 \qquad (7.14)$$

Iteration is the same for all consistent orderings of the points. Moreover, the eigenvalues of Successive Iteration are the squares of those for Simultaneous Iteration. Thus the convergence rate of Successive Iteration is twice that of Simultaneous Iteration.

This last statement is demonstrated easily (Ref. 17). Consider any diagonal non-singular matrix Q. Then for a non-singular square matrix M

$$\det\left(Q^{-1} MQ\right) = \det\left(Q^{-1}\right) \det M \det\left(Q\right) = \det M \qquad (7.15)$$

If M is the matrix of Eqs. (7.12) and (7.15) and Q is defined as

$$Q = \begin{bmatrix} 1 & & & & & & \\ & \bar{\lambda}^{1/2} & & & & & \\ & & \bar{\lambda} & & & & \\ & & & \bar{\lambda}^{3/2} & & & \\ & & & & \cdot & & \\ & & & & & \cdot & \\ & & & & & & \cdot \\ & & & & & & & \bar{\lambda}^{k-1/2} \end{bmatrix} \qquad (7.16)$$

then from Eq. (7.15)

$$\det(Q^{-1}MQ) = \begin{vmatrix} \bar{\lambda} B_1 & \bar{\lambda}^{1/2} C_1 & & & & \\ \bar{\lambda}^{1/2} D_2 & \bar{\lambda} B_2 & \bar{\lambda}^{1/2} C_2 & & & \\ & \bar{\lambda}^{1/2} D_3 & \bar{\lambda} B_3 & \bar{\lambda}^{1/2} C_3 & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \bar{\lambda}^{1/2} C_{k-1} \\ & & & & \bar{\lambda}^{1/2} D_k & \bar{\lambda} B_k \end{vmatrix}$$

$$= \bar{\lambda}^{k/2} f(\bar{\lambda}^{1/2}) = 0 \qquad (7.17)$$

When Eq. (7.17) is compared with Eq. (7.11)

$$\bar{\lambda}^{1/2} = \lambda \qquad (7.18)$$

and the statement is proved. If the Simultaneous Iteration eigenvalues are real (e.g., the matrix A is symmetric), then the eigenvalues of Successive Iteration are also real.

When applied to difference equations, Simultaneous and Successive Iteration are often called Richardson's Method and Liebmann's Method, respectively.

Remarks on Consistent Orderings

Given a set of five-point difference equations on a mesh [i.e., a matrix with Property (A)], the consistent orderings for Successive Iteration are easily determined (Ref. 13). They are just those orderings for which the Successive Iteration equations can be solved in a consistent way, namely, those retaining the feature of Successive Iteration.

Let us assign an ordering vector $\underline{\alpha}$ to the mesh

$$\underline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N) \tag{7.19}$$

where the subscripts on the components α_i refer to the ith equation in the ordering, and the α_i are integers such that

$$|\alpha_i - \alpha_j| = 1 \tag{7.20}$$

if $a_{ij} \neq 0$ and $i \neq j$. Under these circumstances, one can then use the following definition to test for consistent orderings:

Definition: An ordering is consistent for Successive Iteration if, for $a_{ij} = 0$ and $\alpha_i > \alpha_j$, the ith equation in the ordering is solved for after the jth equation; and if, for $a_{ij} \neq 0$ and $\alpha_j > \alpha_i$, the jth equation in the ordering is solved for after the ith.

Given an ordering vector α, with the properties of Eq. (7.20), one form of consistent ordering is to arrange the component α_i in an increasing or decreasing sequence, corresponding to "forward" or "backward" mesh sweeping.

Now consider the example with twelve interior points. The ordering indicated in Fig. 7.2 already has been established as consistent; however, for illustrative purposes we demonstrate it using the ordering vector of Eq. (7.19) with N = 12.

If we assign the components* of $\underline{\alpha}$ as $\alpha_1 = 1$, $\alpha_2 = 2$, $\alpha_3 = 3$, and $\alpha_4 = 4$, then the choice $\alpha_5 = 2$, $\alpha_6 = 3$, $\alpha_7 = 4$, and $\alpha_8 = 5$; and $\alpha_9 = 3$, $\alpha_{10} = 4$, $\alpha_{11} = 5$, and $\alpha_{12} = 6$ will satisfy the definitions of the ordering vector. Thus, of the many possibilities, we have chosen

$$\begin{aligned}\alpha &= (\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5,\alpha_6,\alpha_7,\alpha_8,\alpha_9,\alpha_{10},\alpha_{11},\alpha_{12}) \\ &= (1,\ 2,\ 3,\ 4,\ 2,\ 3,\ 4,\ 5,\ 3,\ 4,\ \ 5,\ \ 6) \qquad (7.21)\end{aligned}$$

Equation (7.21) satisfies the consistency conditions, as can been seen easily. For i = 3, we have $\alpha_3 > \alpha_2$ (since 3 > 2), and the 3rd equation must be solved for, or follow, the 2nd equation. For i = 4, $\alpha_4 > \alpha_3$, and we should have the 4th equation following the 3rd equation. Now for i = 5, $\alpha_4 > \alpha_5$ (since 4 > 2), and we should have the 4th equation following the 5th equation except that the a_{ij} between these points is zero. Therefore, $\alpha_i - \alpha_j \neq 1$, and we could have the 5th equation following the 4th equation and still satisfy the consistency conditions. Again, for i = 6, $\alpha_6 > \alpha_5$, and the 6th equation should follow the 5th equation. In this way, we can see how the $\underline{\alpha}$ of Eq. (7.21) satisfies the consistency conditions with the ordering of the subscripts (of α_i) i = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12.

*This ordering corresponds to sweeping along the elements of a row (k = 1, 2, ..., K) for all rows (ℓ = 1, 2, ..., L).

The inverse ordering of subscripts i = 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1 is also consistent. These two examples correspond to a "forward" and "backward" mesh sweeping, respectively.

One could also have swept "up" or "down" by the ordering of the subscripts i = 4, 8, 12, 3, 7, 11, 2, 6, 10, 1, 5, 9 or i = 12, 8, 4, 11, 7, 3, 10, 6, 2, 9, 5, 1, giving the ordering vector

$$\underline{\alpha} = (6, 5, 4, 5, 4, 3, 4, 3, 2, 3, 2, 1) \tag{7.22}$$

for sweeping up (left or right), or

$$\underline{\alpha} = (4, 5, 6, 3, 4, 5, 2, 3, 4, 1, 2, 3) \tag{7.23}$$

for sweeping down (right to left).

Likewise, diagonal sweeping is consistent, corresponding to the arranging of the α_i of Eq. (7.21) in the increasing order

$$\underline{\alpha} = (1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 6) \tag{7.24}$$

with the subscript ordering i = 1, 2, 5, 3, 6, 9, 4, 7, 10, 8, 11, 12.

Extrapolation Methods

For the particular case of finite elliptic difference equations, Extrapolated Simultaneous Iteration carries over directly from the discussion of Section 3. The only alteration we could make is to utilize the fact that the eigenvalues occur in ± pairs, i.e., $\lambda_0 = -\lambda_m$, where λ_0 and λ_m are the minimum and maximum eigenvalues of the Simultaneous Iteration matrix (assumed real). The Second Order Extrapolated Simultaneous Iteration parameters, for this case, become

$$\left.\begin{aligned}
\omega &= \frac{2}{1 + \sqrt{1 - \lambda_m^2}} \qquad 1 < \omega < 2 \\
\alpha &= \frac{1 - \sqrt{1 - \lambda_m^2}}{1 + \sqrt{1 - \lambda_m^2}} = \omega - 1 \\
\left|\Theta_i\right| &= \sqrt{\omega - 1}
\end{aligned}\right\} \qquad (7.25)$$

where Θ_i are the eigenvalues of the procedure. We can then write the iteration procedure

$$\underline{X}^{(p+1)} = (1 - \omega)\, \underline{X}^{(p-1)} - \omega \left[L\underline{X}^{(p)} + U\underline{X}^{(p)} + \underline{s} \right] \qquad (7.26)$$

This procedure, as applied to difference equations, is usually called Second Order Richardson. Remember that the Θ, as given above, are approached asymptotically as $n \rightarrow \infty$, i.e., the error due to the first sweep (Section 3) becomes insignificant.

Extrapolated Successive Iteration is treated much as the basic Successive Iteration. For any matrix possessing Property (A), the characteristic determinantal equation becomes

$$\begin{vmatrix} (1-\omega-\bar{\lambda})B_1 & -\omega C_1 & & & & \\ -\omega D_2 & (1-\omega-\bar{\lambda})B_2 & -\omega C_2 & & & \\ & -\omega D_3 & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & -\omega C_{k-1} \\ & & & & -\omega D_k & (1-\omega-\bar{\lambda})B_k \end{vmatrix} = 0 \qquad (7.27)$$

where $\bar{\lambda}$ represents the eigenvalues of the procedure. Factoring out the iteration parameter ω and using the diagonal Q matrix equation (7.16), we obtain the relation

$$(\bar{\lambda} + \omega - 1) = \omega \lambda \bar{\lambda}^{1/2} \qquad (7.28)$$

where λ is the basic Simultaneous Iteration eigenvalue equation (7.11). Solving Eq. (7.28), we have

$$\lambda = \left[\frac{\omega\lambda}{2} \pm \sqrt{\frac{\omega^2\lambda^2}{4} - (\omega - 1)} \right]^2 \qquad (7.29)$$

Since the convergence rate of the process depends upon the magnitude of its eigenvalues, we wish to minimize $\bar{\lambda}$. This is accomplished by forcing the square root to be zero for the largest eigenvalue, λ_m. Thus

$$\omega^2 \lambda_m^2 = 4(\omega - 1) \tag{7.30}$$

or

$$\omega = \frac{2}{1 + \sqrt{1 - \lambda_m^2}} \tag{7.31}$$

and Eq. (7.29) becomes

$$\bar{\lambda} = \omega - 1 \tag{7.32}$$

for <u>all</u> eigenvalues. These eigenvalues are the square of the asymptotic form for Second Order Extrapolated Iteration. One would expect that Extrapolated Successive Iteration would converge in, at most, half the number of iterations required by Second Order Extrapolated Simultaneous Iteration. Such is not the case! The iteration matrix defined by Extrapolated Successive Iteration does not have a complete set of eigenvectors. The Jordan normal form has one off-diagonal element, and thus the matrix lacks one eigenvector (Ref. 13, p. 103). Thus, instead of the error decaying as $(\omega - 1)^p$, we have

$$\underline{E}^{(p)} = p(\omega - 1)^p \, \underline{E}^{(o)} \tag{7.33}$$

The parameter ω, defined by Eq. (7.31), is greater than 1. This procedure is generally referred to as Successive Over-Relaxation (SOR). It is also often called Extrapolated Liebmann.

Tschebyscheff Polynomial Methods

Usually the matrix of the difference equations is symmetric,* and, in that case, all the eigenvalues of the Simultaneous and Successive Iteration are real. Then, Tschebyscheff Polynomials can be applied to either of these two basic techniques. Since the eigenvalues of the basic Simultaneous Iteration occur in ± pairs, the parameters of the Second Order Simultaneous Polynomial method are

$$\left.\begin{aligned}
& a = -b = \lambda_m \\
& \alpha_p = \frac{2}{\lambda_m}\frac{T_{p-1}(1/\lambda_m)}{T_p(1/\lambda_m)} \qquad p \geq 2 \\
& \beta_p = \frac{T_{p-1}(1/\lambda_m)}{T_p(1/\lambda_m)} \qquad p \geq 2 \\
& \alpha_1 = 1 \\
& \beta_1 = 0
\end{aligned}\right\} \qquad (7.34)$$

*The differential equations usually can be differenced so that a point has the same effect on its neighbors in the difference equations as the neighbors have on the point. This will yield a symmetric matrix.

where λ_m is the maximum eigenvalue of the Simultaneous Iteration matrix. The error reduction is

$$\underline{E}^{(p)} \leq \frac{1}{T_p(1/\lambda_m)} \underline{E}^{(o)} \tag{7.35}$$

The eigenvalues of Successive Iteration are the squares of the Simultaneous Iteration, and thus the replacement of λ_m by λ_m^2 in Eqs. (7.34) and (7.35) gives the appropriate values for the Second Order Successive Polynomial method.

The Extrapolated Successive Iteration for difference equations does not have real eigenvalues; therefore the convergence rate cannot be accelerated with Tschebyscheff Polynomials. Sheldon (Ref. 31) has shown, however, that if the mesh is swept first in one direction and then back in the other direction, using Extrapolated Successive Iteration for each sweep, then the two-step process has real eigenvalues. The eigenvalues are all positive, and the maximum is approximately the usual Extrapolated Successive Iteration eigenvalue. Accelerating this two-step process gives a gain in over-all convergence rate. Each iteration requires two mesh sweeps and the associated arithmetic of the polynomial generation. The Second Order Extrapolated Forward-Backward Successive Polynomial parameters are

$$\left.\begin{aligned}
a &\simeq \frac{2}{1 + \sqrt{1 - \lambda_m^2}} - 1 = \alpha - 1 \\
b &\simeq 0 \\
\alpha_p &= \frac{4}{\alpha - 1} \frac{T_{p-1}\left(\frac{3 - \alpha}{\alpha - 1}\right)}{T_p\left(\frac{3 - \alpha}{\alpha - 1}\right)} \\
\beta_p &= \frac{T_{p-2}\left(\frac{3 - \alpha}{\alpha - 1}\right)}{T_p\left(\frac{3 - \alpha}{\alpha - 1}\right)} \\
\alpha_1 &= \frac{2}{2 - \alpha - 1} \\
\beta_1 &= 0
\end{aligned}\right\} \qquad (7.36)$$

and the error reduction is

$$\underline{E}^{(p)} \leq \frac{1}{T_p\left(\frac{3 - \alpha}{1 - \alpha}\right)} \underline{E}^{(o)} \qquad (7.37)$$

If any eigenvalues exist that give a larger value of a than that estimated in Eq. (7.35), they must be damped out by some other procedure. Steifel (Ref. 4) suggests using a method essentially equivalent to Conjugate Gradients. Extrapolated Successive Iteration would probably be sufficient. We can, of course, apply the Tschebyscheff Polynomial method to Simultaneous or Successive line methods or even to a Forward-Backward Extrapolated line method (Ref. 25).

Block or Implicit Methods

Because the equations of this section arise from differencing an elliptic differential equation on a rectangular mesh, a logical block to be advanced simultaneously (Section 5) consists of all points on a row or column. Each such block is, in itself, a tridiagonal matrix which can be easily solved by Gauss Elimination (e.g., Ref. 18).*

Assume that the mesh has k rows and ℓ columns. If we advance each row as a block, then the diagonal matrix of Eq. (5.1) consists of k square matrices of size $\ell \times \ell$. Similarly, if we advance each column as a block, the matrix D consists of ℓ matrices of size $k \times k$.

Call the maximum eigenvalue of the Simultaneous Row Iteration σ_m, replacing λ_m. Thus, for the Second Order Extrapolated Simultaneous Row Iteration, we have

$$\omega = \frac{2}{1 + \sqrt{1 - \sigma_m^2}}$$

$$\alpha = \omega - 1 \qquad (7.38)$$

*Note added in proof: A particular form stable against round-off, apparently due to J. von Neumann, is discussed by J. Douglas (Ref. 20).

and for the Extrapolated Successive Row Iteration

$$\omega = \frac{2}{1 + \sqrt{1 - \sigma_m^2}} \qquad (7.39)$$

Since Property (A) holds for the five-point difference equations, one could take the logical block of unknowns to be the entire mesh, solve for the unknowns by direct inversion, and iterate the result with one of the methods discussed to reduce round-off effects. The matrix would be

$$A = \begin{bmatrix} B_1 & C_1 & & & & \\ D_2 & B_2 & C_2 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot & C_{n-1} \\ & & & & & D_n & B_n \end{bmatrix} \qquad (7.40)$$

where for orderings along rows or columns B_n is the ℓ x ℓ or k x k tri-diagonal matrix appearing in the row or column iteration, and C_n and D_n are ℓ x ℓ or k x k diagonal matrices representing the a_{ij} terms for columns or rows. For orderings along mesh diagonals B_n is an ℓ x ℓ or k x k diagonal matrix representing the a_{ii} terms of Eq. (7.1) and C_n and D_n are rectangular matrices, as in Eq. (7.14), representing the a_{ij} terms of Eq. (7.1)

In either case, direct inversion could be accomplished in a manner analogous to ordinary Gauss Elimination for tridiagonal matrices, only using matrix multiplication for products and matrix inversion for inverses. Either K $\ell \times \ell$ or L $k \times k$ inverses are involved in performing the direct inversion. If either K or L is small, then it should determine the maximum matrix size since round-off can be large in many instances. If the round-off does not render the answer ridiculous, the application of Simultaneous or Successive Iteration, as well as their generalizations, could be made to the solutions x_k.

Recently, Nohel and Timlake (Ref. 26) applied a procedure of this type to nine-point difference equations [possessing Block Property (A)] and, for the test cases considered, found no appreciable round-off. Oliphant and Baker (Ref. 29) have exhibited a factorization of the nine-point difference equations for the Heat Equation. Oliphant (Ref. 27) has proposed a factorization of the general case for nine-point difference equations using a Lagrangian multiplier technique to obtain consistency of the equations.

Peaceman and Rachford (Ref. 18) have introduced an alternating direction implicit method where first rows are swept as a block and then columns. Each sweep is similar to a Simultaneous Row (or Column) Iteration in that previous iterates of adjacent rows (or columns) are used.

However, the new values in each sweep also depend on old iterates of the row (or column). A single iteration of this method consists of sweeping first in one direction and then in another. Thus, the amount of work involved per iteration is equivalent to about two Simultaneous Row (Column) Iterations.

Peaceman and Rachford split the matrix A into the sum of two matrices H and V, where H results from differencing in one direction and V from differencing in the other direction. Referring to our example of Poisson's equation (7.2), we have

$$\begin{aligned} H\psi &= \frac{1}{h}\psi_{k+1\ \ell} - \frac{2}{h}\psi_{k\ell} + \frac{1}{h}\psi_{k-1\ \ell} \\ V\psi &= \frac{1}{h}\psi_{k\ \ell+1} - \frac{2}{h}\psi_{k\ell} + \frac{1}{h}\psi_{k\ \ell-1} \end{aligned} \tag{7.41}$$

The iteration procedure can be written (with $\underline{\psi} = \underline{x}$)

$$\begin{aligned} (H + \rho_p I)\ \underline{x}^* &= -(V - \rho_p I)\ \underline{x}^{p-1} - s \\ (V + \rho_p I)\ \underline{x}^{(p)} &= -(H - \rho_p I)\ \underline{x}^* - \underline{s} \end{aligned} \tag{7.42}$$

where ρ_p are parameters to be chosen for optimum convergence rate. We see that we sweep first to improve the terms corresponding to the H matrix and then sweep in the other direction to improve the V matrix terms. If $\underline{x}$ is the unique solution, i.e.,

$$(H + V)\, \underline{x} + \underline{s} = 0 \tag{7.43}$$

then the error, $\underline{E}^{(p)} = \underline{x} - \underline{x}^{(p)}$, satisfies Eq. (7.43) without the source terms. Combining the two equations of Eq. (7.42), we have

$$\underline{E}^{(p)} = (V + \rho_p I)^{-1} (H - \rho_p I)(H + \rho_p I)^{-1} (V - \rho_p I)\, \underline{E}^{(p-1)} \tag{7.44}$$

If H and V have a common set of eigenvectors, $\underline{\alpha}_i$, we can expand $\underline{E}^{(p-1)}$ in these eigenvalues. Thus,

$$\underline{E}^{(p-1)} = \sum_{i=1}^{n} c_i\, \underline{\alpha}_i \tag{7.45}$$

where

$$\begin{aligned} H\, \underline{\alpha}_i &= \sigma_i\, \underline{\alpha}_i \\ V\, \underline{\alpha}_i &= \gamma_i\, \underline{\alpha}_i \end{aligned} \tag{7.46}$$

and σ_i and γ_i are eigenvalues of H and V, respectively. Then from Eq. (7.44)

$$\underline{E}^{(p)} = \sum_{i} \left(\frac{\gamma_i - \rho_p}{\gamma_i + \rho_p}\right)\left(\frac{\sigma_i - \rho_p}{\sigma_i + \rho_p}\right) c_i \, \underline{\alpha}_i \tag{7.47}$$

In terms of the initial error we have

$$\underline{E}^{(p)} \leq \max_{i} \prod_{j=1}^{p} \left(\frac{\gamma_i - \rho_j}{\gamma_i + \rho_j}\right)\left(\frac{\sigma_i - \rho_j}{\sigma_i + \rho_j}\right) \underline{E}^{(o)} \tag{7.48}$$

The matrices H and V for our example (Poisson's equation) do in fact have a common basis of eigenvectors. They are just the sine or cosine functions. In general, any two matrices have a common set of eigenvectors if they commute, i.e., if

$$HV = VH \tag{7.49}$$

Keller (Ref. 16) has shown conditions under which H and V commute for elliptic difference equations. If the mesh region is rectangular, and H is derived from

$$K(x,y) \frac{\partial^2 \psi(x,y)}{\partial x^2} + L(x,y) \frac{\partial \psi(x,y)}{\partial x} + P(x,y) \, \psi(x,y) \tag{7.50}$$

and V is derived from

$$R(x,y) \frac{\partial^2 \psi(x,y)}{\partial y^2} + S(x,y) \frac{\partial \psi(x,y)}{\partial y} + Q(x,y) \, \psi(x,y) \tag{7.51}$$

by a three-point approximation for H and V, respectively, then Keller found that H and V will commute whenever the variable coefficients in the differential operators of Eqs. (7.51) and (7.52) are functions of x only when multiplying x derivatives and functions of y only when multiplying y derivatives.

A physical application of some interest is the steady state diffusion equation with absorption and an external source

$$\vec{\nabla} \cdot D(\underline{r}) \nabla\phi(\underline{r}) - \Sigma(\underline{r})\, \phi(\underline{r}) + s(\underline{r}) = 0 \qquad (7.52)$$

where $\underline{r}$ is the position vector. Defining H and V as for Laplace's equation, the coefficients must now include the diffusion constant D. We also have the vector $\Sigma(\underline{r})$ entering into Eqs. (7.42). With these additions, H and V will not, in general, commute. Moreover, if $\Sigma \neq 0$, then to expand as in Eq. (7.47), we must further require that Σ commute with both H and V. Varga and Birkhoff (Ref. 19) proved that when Σ and D are constant and the domain is rectangular, all the conditions are satisfied.

Wachspress (Ref. 24) has shown that expansions such as Eq. (7.47) can be carried out if a diagonal matrix F can be found such that FH, FV, and I satisfy the commuting requirements. Such an F will exist if Σ is constant and the coefficients of D vary in one direction only.

These two results are special cases of Keller's more recent results.

Assuming conditions are satisfied for the expansion of Eq. (7.47), we have yet to choose the parameters ρ_i. But usually the problem is so simple that the eigenvalues are known and the ρ_i can be picked to eliminate successively each error made. If the matrix is very large and there are many eigenvalues close together, one ρ_i can be chosen to reduce a group of eigenmodes to an insignificant contribution (Ref. 18).

Wachspress (Ref. 24) has obtained a set of parameters by minimizing the maximum absolute value of the polynomial in Eq. (7.48). He concludes that if

$$\begin{aligned} &\text{(a)} \quad a < \gamma_i \text{ and } \sigma_i < b \\ &\text{(b)} \quad \rho_i = bx^{i-1} \qquad i = 1, 2, \ldots, p \\ &\text{(c)} \quad \rho_p < a \text{ and } 0 < x < 1 \end{aligned} \tag{7.53}$$

then

$$\underline{E}^{(p)} = \left[\frac{1 - x}{1 + x} e^{-x^{3/2}/(1-x)}\right]^4 \underline{E}^{(o)} \tag{7.54}$$

where x is chosen to give the desired error reduction.

The Peaceman-Rachford iteration has been used to solve the general diffusion equation (7.52) even though convergence cannot be guaranteed (Ref. 24). When D and Σ are slowly varying functions in space, the commuting requirements are nearly satisfied and the error reduction of Eq. (7.54) is approximately correct.

If we choose

$$\rho_i = \rho > 0 \tag{7.53'}$$

then convergence is guaranteed in the general case.

Variational Methods

The methods of Section 6 can clearly be applied to elliptic difference equations whenever the matrix is positive definite. Though the method of Conjugate Gradients is exact in N steps, N may be so large for a reasonable mesh that N steps are too many. We know that for $p < N$ we have improved the initial approximation, but we have no accurate theorems for estimating the amount of this improvement. Hence, the variational methods have not been used a great deal for large systems.

In addition to the above drawbacks, round-off errors for large N can be very serious.

Convergence Properties

To compare the convergence rates of the various methods, assume we are solving Poisson's equation ($-\nabla^2\phi = s$) on a rectangular mesh. Then the scheme of Peaceman and Rachford can be rigorously applied. Moreover, we can easily calculate the eigenvalues for this simple problem and get estimates of the maximum number of iterations required to reduce the error by some given amount.

Assume the largest eigenvalue of the Simultaneous Iteration method is λ_m; then generally

$$\lambda_m \cong 1 - \epsilon \qquad \epsilon \ll 1 \tag{7.55}$$

If we further assume that the mesh spacing is constant and the same in the x and y directions, then the maximum eigenvalue of the Simultaneous Line Iteration, σ_m, is given by

$$\sigma_m \cong 1 - 2\epsilon \tag{7.56}$$

We can say nothing about the required iterations for the variational methods except that if round-off errors are insignificant, the conjugate direction methods* will converge to the exact solution in N steps, N being the number of mesh points and equations.

For the remaining methods, we can construct Table 7.1 (partly taken from Ref. 25). The quantity R is the error reduction

$$E^{(p)} = R\underline{E}^{(o)} \tag{7.57}$$

The simultaneous and successive iterations all require about the same amount of calculation per mesh sweep per point. If appropriate

*Conjugate Gradients is a method of conjugate directions.

coefficients are calculated in advance, the line methods require about the same effort as the point methods. The Tschebyscheff Polynomial method requires only slightly more calculation.

The alternating direction implicit method of Peaceman and Rachford requires about 50 per cent more calculation per point than the above methods. The variational methods require the calculation of three scalar products to get the necessary parameters, or approximately a 100 per cent increase in calculation per point.

Table 7.1

Approximate Number of Sweeps Required to Reduce the Initial Error to R of Its Initial Value for Given ε [Eqs. (7.55) and (7.56)]

	Approximate Number of Sweeps	R = 0.02		R = 0.002	
		$\epsilon = 10^{-2}$	$\epsilon = 10^{-4}$	$\epsilon = 10^{-2}$	$\epsilon = 10^{-4}$
Simultaneous Iteration	$-\frac{\ln R}{\epsilon}$	390	39,000	620	62,000
Successive Iteration	$-\frac{\ln R}{2\epsilon}$	195	19,500	310	31,000
Simultaneous Line Iteration	$-\frac{\ln R}{2\epsilon}$	195	19,500	310	31,000
Successive Line Iteration	$-\frac{\ln R}{4\epsilon}$	98	9,750	155	15,500
Second Order Simultaneous Iteration*	$-\frac{\ln R}{\sqrt{2\epsilon}}$	28	280	44	440
Second Order Simultaneous Line Iteration*	$-\frac{\ln R}{\sqrt{4\epsilon}}$	20	200	31	310
Extrapolated Successive Iteration	$-\frac{\ln(R/n)}{\sqrt{8\epsilon}}$	26	350	35	430
Extrapolated Successive Line Iteration	$-\frac{\ln(R/n)}{\sqrt{16\epsilon}}$	17	210	24	300
Simultaneous Tschebyscheff Polynomials	$-\frac{\ln(R/2)}{\sqrt{2\epsilon}}$	33	330	49	490
Successive Tschebyscheff Polynomials	$-\frac{\ln(R/2)}{\sqrt{4\epsilon}}$	24	240	35	350
Simultaneous Line Tschebyscheff Polynomials	$-\frac{\ln(R/2)}{\sqrt{4\epsilon}}$	24	240	35	350
Successive Line Tschebyscheff Polynomials	$-\frac{\ln(R/2)}{\sqrt{8\epsilon}}$	17	170	25	250
Forward-Backward Extrapolated Tschebyscheff Polynomials**	$-\frac{\ln(R/2)}{4\sqrt{16\epsilon}}$	12	27	15	46
Forward-Backward Extrapolated Line Tschebyscheff Polynomials	$-\frac{\ln(R/2)}{4\sqrt{32\epsilon}}$	8	20	10	32
Peaceman-Rachford***	$2 + \frac{2\,\ln \epsilon}{\ln x}$	8	14	10	18

*Assumes asymptotic eigenvalues.

**Assumes $a = 2/(1 + \sqrt{1 - \lambda_m^2})$.

***x is related to R by Eq. (7.54).

8. SUMMARY

Several of the rather well known methods for solving large systems of simultaneous equations have been presented. Both the general case and a specific application -- elliptic difference equations -- have been discussed.

It has been noted that simple restrictions on the coefficients of a given matrix are sufficient to guarantee convergence of the various methods. The Simultaneous Iteration and Second Order Simultaneous Iteration require that the diagonal terms of the matrix dominate or equal the off-diagonal terms. The Successive or Successive Extrapolated Iterations will converge if the diagonal terms dominate the off-diagonal terms. Successive Iteration, Extrapolated Successive Iteration, and Variational Methods require that the matrix be positive definite. The Tschebyscheff Polynomial method requires that the matrix used in the polynomial have real eigenvalues.

The basic iterations are usually very slowly convergent. Second Order Simultaneous Iteration gives an order of magnitude improvement while the Successive Extrapolated Iteration cannot be used optimally in the general case. The Tschebyscheff Polynomial method, when applicable, is generally the most rapidly convergent. Variational methods are rapid, but restricted to small matrices of less than a few hundred equations. They have the advantage of producing an exact solution in a finite number of steps.

The iteration and polynomial methods can be extended to implicit methods where blocks of unknowns are improved simultaneously. If the matrix of the block can be easily inverted, this procedure can save a great deal of computation time. The effort involved in solving for the unknowns in the block must be balanced against the improved convergence rate. This latter is difficult to estimate for a general matrix.

In the particular case of five-point finite difference equations, [more generally, when the matrix possesses Property (A)] more can be said about the iteration methods. It can be shown that the Successive Iteration converges twice as fast as the Simultaneous Iteration. The Successive Extrapolated Iteration is an order of magnitude improvement over the Successive Iteration and converges more rapidly than the Second Order Simultaneous Iteration. Moreover, both Successive and Simultaneous Iteration matrices have real eigenvalues and can be accelerated by Tschebyscheff Polynomials. Also, a Forward-Backward Successive Extrapolated Iteration (due to Sheldon, Ref. 31) which has real eigenvalues and can be accelerated with Tschebyscheff Polynomials can be used.

The entire mesh can be inverted simultaneously in block form where the blocks are diagonals of the mesh if convergence is slow and round-off low. The resultant solution can then be iterated as described in Section 7.

A logical block of unknowns for implicit methods consists of all the points along a row or column of the mesh. It is not always clear in which direction it is best to sweep. If the coefficients of the differential equation are constant over the mesh, the implicit equations should be obtained in the direction of fewest mesh points. Usually, no preferred direction can be determined and, in any case, the gain in convergence rate is often small.

The alternating direction implicit method of Peaceman and Rachford can be proved applicable only for the case in which the operator plus its domain is separable.

The convergence rate for this problem is so rapid, however, that the method has been used for more general problems with good success. It sometimes appears applicable if the operator is almost separable.

We close by proposing that for a general non-symmetric matrix, the Second Order Extrapolated Simultaneous Iteration is most applicable. If the matrix is symmetric (real eigenvalues), then the Tschebyscheff Polynomial applied to the Simultaneous Iteration is best. For a matrix possessing Property (A), the Extrapolated Successive Iteration is sufficient for problems having inherently rapid convergence rates (diagonal terms dominant), and the Forward-Backward Successive Extrapolation accelerated with Tschebyscheff Polynomials is best for poorly converging problems.

BIBLIOGRAPHY

A. Iterative Methods

1. Geiringer, H., Solutions of Linear Equations by Certain Iterative Methods, Reissne Anniversary Volume, Ann Arbor (1949), pp. 360-392

2. Young, D., "On the Solution of Linear Systems by Iteration," Proceedings of Symposia in Applied Mathematics, Vol. VI, McGraw-Hill, New York (1956), pp. 283-298.

3. Fox, L., "Practical Solution of Linear Equations and Inversion of Matrices," Contributions to the Solution of Systems of Linear Equations and the Determination of Eigenvalues, NBS Applied Mathematics Series, No. 39, U. S. Government Printing Office, Washington (1954), pp. 1-54.

B. Polynomial Methods

4. Stiefel, E., "Kernel Polynomials in Linear Algebra and Their Numerical Applications," Further Contributions to the Solution of Simultaneous Linear Equations and the Determination of Eigenvalues, NBS Applied Mathematics Series, No. 49, U. S. Government Printing Office, Washington (1956), pp. 1-22.

5. Flanders, D. A., and Shortley, G., "Numerical Determination of Fundamental Modes," J. Appl. Phys., 21, 1326-1332 (1950).

6. Young, D., "On Richardson's Method for Solving Linear Systems with Positive Definite Matrices," J. Math. and Phys., 32, 243-255 (1954).

7. Shortley, G., "Use of Tschebyscheff-Polynomial Operators in the Numerical Solution of Boundary-Value Problems," J. Appl. Phys., 24, 392-396 (1953).

C. Variational Methods

8. Stiefel, E., "Some Special Methods of Relaxation Techniques," Simultaneous Linear Equations and the Determination of Eigenvalues, NBS Applied Mathematics Series, No. 29, U. S. Government Printing Office, Washington (1953), pp. 43-48.

9. Hayes, R. M., "Iterative Methods of Solving Linear Problems on Hilbert Space," Contributions to the Solution of Systems of Linear Equations and the Determination of Eigenvalues, NBS Applied Mathematics Series, No. 39, U. S. Government Printing Office, Washington (1954), pp. 71-104.

10. Lanczos, C., "An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators," J. Research, Nat. Bur. Standards 45 (1950), 255-282.

11. Hestenes, M. R., and Stiefel, E., "Methods of Conjugate Gradients for Solving Linear Systems," J. Research, Nat. Bur. Standards 49 (1952), 409-436.

12. Rosenbloom, P. C., "The Method of Steepest Descent," Proceedings of Symposia in Applied Mathematics, Vol. VI, McGraw-Hill, New York (1956), pp. 127-176.

D. Applications to Elliptic Difference Equations

13. Young, D., "Iterative Methods for Solving Partial Difference Equations of Elliptic Type," Trans. Am. Math. Soc. 76, 92-111 (1954).

14. Frankel, S. P., "Convergence Rates of Iterative Treatments of Partial Differential Equations," Math. Tables Aids Computation, 4, 65-75 (1950).

15. Keller, H. B., "On Some Iterative Methods for Solving Elliptic Difference Equations," New York University, AEC Computing and Applied Mathematics Center Report NYO-7971 (1957).

16. Keller, J., "Simultaneous, Successive and Alternating Direction Schemes," New York University, AEC Computing and Applied Mathematics Center Report NYO-8675 (1958).

17. Friedman, B., "The Iterative Solution of Elliptic Difference Equations," New York University, AEC Computing and Applied Mathematics Center Report NYO-7698 (1957).

18. Peaceman, D. W., and Rachford, H. H., Jr., "The Numerical Solution of Parabolic and Elliptic Differential Equations," J. Soc. Indust. Appl. Math., 3, 28-41 (1955).

19. Birkhoff, G. B., and Varga, R. L., "Implicit Alternating Direction Methods," Westinghouse Electric Corporation Bettis Plant Report WAPD-T-650 (Rev.) (1958).

20. Douglas, J., "Round-off Error in the Numerical Solution of the Heat Equation," J. Assoc. Comp. Mach. 6, 48 (1959).

21. Turing, A. M., "Rounding-off Errors in Matrix Processes," Quart. J. Mech. App. Math. 1, 287 (1948).

E. Miscellaneous

22. Rosser, J. B., "Rapidly Converging Iterative Methods for Solving Linear Equations," Simultaneous Linear Equations and the Determination of Eigenvalues, NBS Applied Mathematics Series, No. 29, U. S. Government Printing Office, Washington (1953), pp. 59-64.

23. Stein, P., and Rosenberg, R. L., "On the Solution of Linear Simultaneous Equations by Iteration," J. London Math. Soc. 23, 111-118 (1948).

24. Wachspress, E. L., "Cure: A Generalized Two-Space-Dimension Multigroup Coding for the IBM-704," Knolls Atomic Power Laboratory Report KAPL-1724 (1957).

25. Wachspress, E. L., Stone, P. M., and Lee, C. E., "Mathematical Techniques in Two-Space-Dimension Multigroup Calculations," Paper 633, Second International Conference on the Peaceful Uses of Atomic Energy, Geneva, 1958.

26. Nohel, J. A., and Timlake, W. P., "Higher Order Differences in the Numerical Solution of the Two-Dimensional Neutron Diffusion Equations," Paper 634, Second International Conference on the Peaceful Uses of Atomic Energy, Geneva, 1958.

27. Oliphant, T. A., Jr., "A Direct Implicit Scheme for Solving Two-Dimensional Steady-State Diffusion Problems," to be published.

28. Baker, G. A., Jr., "Note on the Solution of the Neutron Diffusion Problem by an Implicit Numerical Method," to be published.

29. Baker, G. A., Jr., and Oliphant, T. A., Jr., "An Implicit, Numerical Method for Solving the Two-Dimensional Heat Equation," Los Alamos Scientific Laboratory Report LA-2232 (1958).

30. Bodewig, E., Matrix Calculus, Interscience, New York (1956).

31. Sheldon, J. W., "On the Numerical Solution of Elliptic Difference Equations," Math. Tables Aids Computation, 9, 101-112 (1955).

F. Extensive Bibliographies

32. Forsythe, G. E., "Tentative Classification of Methods and Bibliography on Solving Systems of Linear Equations," *Simultaneous Linear Equations and the Determination of Eigenvalues*, NBS Applied Mathematics Series, No. 29, U. S. Government Printing Office, Washington (1953), pp. 1-28.

33. Jensen, H., *An Attempt at a Systematic Classification of Some Methods for the Solution of Normal Equations*, Geodaetisk Institut, Meddelelse No. 18, Copenhagen (1944).