The extrapolated modified Aitken iteration method for solving elliptic difference equations

By D. J. Evans

This paper describes a new approach to the extrapolation (over-relaxation) and linear acceleration of the Aitken iterative method, and presents theoretical and experimental evidence concerning its application to 5-point elliptic difference equations. Comparison is made between the performance of the new method and that of existing procedures on a Laplace problem, and on a boundary-value problem in one independent variable.

1. In this paper we shall be concerned with iterative methods for solving large systems of linear algebraic equations which arise in the finite-difference solutions of boundary-value problems associated with elliptic partial differential equations. We seek the solution vector x to the equation

$$Ax = d \tag{1.1}$$

where A is a given $(N \times N)$ real, symmetric and positive definite matrix, and d is a given column vector. Without loss of generality, we suppose A to have the form $I - L - U$, where L and U are respectively lower and upper triangular matrices with zero diagonal entries, and I is the identity matrix. Since A is symmetric, L is U^T.

The simplest form of iteration is that of the *stationary** methods, which involve the application of an unmodified computational cycle to the successive estimates $x^{(n)}$ of $x(n = 1, 2, \ldots)$. A simple example of this type of iteration is the Simultaneous or Jacobi iteration, defined by

$$x^{(n+1)} = (L + U)x^{(n)} + d. \tag{1.2}$$

Denoting by $e^{(n)}$, the *error vector*, $x^{(n)} - x$, we have

$$e^{(n+1)} = He^{(n)} \tag{1.3}$$

where the *error operator* H is $(L + U)$. In this *simultaneous displacement* iteration, no element of $x^{(n+1)}$ is used until every element has been calculated and consequently the elements of the *displacement vector*, $x^{(n+1)} - x^{(n)}$, are independent of the order of computation.

Another simple stationary method is the Gauss–Seidel iteration, in which the components of $x^{(n+1)}$ are used as soon as they have been calculated. In this case, we have

$$x^{(n+1)} = Lx^{(n+1)} + Ux^{(n)} + d$$

or $$(I - L)x^{(n+1)} = Ux^{(n)} + d. \tag{1.4}$$

The error vector again satisfies (1.3) but now the error operator is $(I - L)^{-1}U$. It is easy to see that the rate of convergence of this *successive displacement* method may depend upon the order in which the elements of $x^{(n)}$ are modified.

Aitken (1950) proposed a *symmetric* Gauss–Seidel process in which each iteration consists of two sweeps of Gauss–Seidel type, the second sweep adjusting the components of the approximate solution in an order reverse to the first. The first sweep is given by equation (1.4) with $\hat{x}^{(n+1)}$ written in place of $x^{(n+1)}$, whilst the second is defined by

$$x^{(n+1)} = L\hat{x}^{(n+1)} + Ux^{(n+1)} + d$$

or $$(I - U)x^{(n+1)} = L\hat{x}^{(n+1)} + d. \tag{1.5}$$

The error vector satisfies

$$e^{(n+1)} = (I - U)^{-1}L(I - L)^{-1}Ue^{(n)} \tag{1.6}$$

and Aitken showed that the error operator is non-negative definite with all its eigenvalues less than 1.

We may regard (1.4) as derived from (1.2) by the use of the superscript $n + 1$ in place of n for the term Lx in (1.2), and we note that this still permits the calculation of the successive elements of $x^{(n+1)}$ by means of a simple algorithm which is explicit in form. We propose here to extend this approach to include the term Ux, and to do so we use the forward and back substitution processes to solve the equation

$$x^{(n+1)} = Lx^{(n+1)} + Ux^{(n+1)} - LUx^{(n+1)} + LUx^{(n)} + d$$

by means of the equivalent form

$$(I - L)(I - U)x^{(n+1)} = LUx^{(n)} + d. \tag{1.7}$$

The error operator of iteration (1.7) is $(I - U)^{-1}(I - L)^{-1}LU$ and, since the two inner factors commute, this operator is identical with the operator in the Aitken iteration given by (1.6). However, since this iteration will be undertaken in a manner different from that envisaged by Aitken, we shall refer to it as the Modified Aitken method.

In this iteration, we replace the vector $x^{(n)}$ by a new vector $x^{(n+1)}$ after a series of operations involving the triangular matrices of the original matrix A. Hence, no rounding errors are involved in the coefficients, and any sparseness of the original matrix is retained. Furthermore, since the operations defined by $(I - U)^{-1}$ and $(I - L)^{-1}$ amount to no more than the processes of

* For a classification of iterative methods, based on the same terminology as that used here, we refer the reader to Martin and Tee (1961).

forward and back substitution, storage of only one vector is required for the method to operate. Finally, it will be shown in Section 4 that when A is derived from a 5-point finite-difference approximation to an elliptic partial differential equation, the matrix LU can be generated by a simple algorithm based on the mesh geometry. Thus, the proposed method has the possibilities of an efficient iterative method for use on a digital computer.

2. A standard technique for improving the convergence of the iterative methods mentioned in Section 1 is that of extrapolation or over-relaxation. Both the Extrapolated Simultaneous iteration,

$$x^{(n+1)} = x^{(n)} + \omega(Lx^{(n)} + Ux^{(n)} + d - x^{(n)}), \qquad (2.1)$$

and the Successive Over-relaxation (S.O.R.) method,

$$x^{(n+1)} = x^{(n)} + \omega(Lx^{(n+1)} + Ux^{(n)} + d - x^{(n)}) \qquad (2.2)$$

have been investigated and reported upon. (See, for example, Forsythe and Wasow, 1960). We shall initiate our investigations without discussion of these earlier methods, but overall comparison with them will be included in the numerical experiments.

Sheldon (1955) extended Aitken's method (1.4) and (1.5) by an analogue of S.O.R. and showed that the eigenvalues of the error operator are real, positive and less than unity for $0 < \omega < 2$. Accordingly, a further fundamental improvement to the method can be made by *linear acceleration* using Chebyshev polynomials. Each iteration of the accelerated Symmetric Successive Over-relaxation (S.S.O.R.) method is a three-step process given (Habetler and Wachspress, 1961) by the equations

$$\begin{aligned}
\hat{x}^{(n+1)} &= x^{(n)} + \omega(L\hat{x}^{(n+1)} + Ux^{(n)} + d - x^{(n)}) \\
\tilde{x}^{(n+1)} &= \hat{x}^{(n+1)} + \omega(Lx^{(n+1)} + U\tilde{x}^{(n+1)} + d - \hat{x}^{(n+1)}) \\
x^{(n+1)} &= x^{(n)} + \alpha_n(\tilde{x}^{(n+1)} - x^{(n)}) + \beta_n(x^{(n)} - x^{(n-1)})
\end{aligned} \qquad (2.3)$$

where α_n and β_n are defined below.

The error vector for this scheme satisfies

$$\tilde{e}^{(n+1)} = M_\omega e^{(n)} \text{ and } e^{(n+1)} = M_\omega{}^{(n)} e^{(n)}, \qquad (2.4)$$

where

$$M_\omega = I - \omega(2 - \omega)(I - \omega U)^{-1}(I - \omega L)^{-1} A$$

and

$$\begin{aligned}
M_\omega{}^{(n+1)} = M_\omega{}^{(n)} &+ \alpha_n(M_\omega M_\omega{}^{(n)} - M_\omega{}^{(n)}) \\
&+ \beta_n(M_\omega{}^{(n)} - M_\omega{}^{(n-1)}).
\end{aligned}$$

If the eigenvalues, λ, of M_ω satisfy $0 \leqslant \lambda \leqslant 1 - \eta$, we require that each contribution to $e^{(0)}$ from an eigenvector of M_ω shall be weighted in $e^{(n)}$ by the factor

$$\frac{T_n\left(\frac{2\lambda}{1-\eta} - 1\right)}{T_n\left(\frac{2}{1-\eta} - 1\right)},$$

where $T_n(x)$ is cosh $(n \cosh^{-1} x)$. This is achieved if α_n and β_n satisfy

$$\alpha_n = \frac{4}{(1-\eta)} \frac{T_n(\gamma)}{T_{n+1}(\gamma)} \left(\alpha_0 = \frac{2}{1+\eta}\right) \qquad (2.5)$$

and

$$\beta_n = \frac{T_{n-1}(\gamma)}{T_{n+1}(\gamma)}, \quad (n = 1, 2, \ldots)$$

where γ is $(1 + \eta)/(1 - \eta)$.

Frank (1960) has recently reported favourably on the advantages of using the three-term relationship outlined in (2.3). Briefly, for the small disadvantage of retaining a further vector, the problem of numerical instability is overcome and at each stage the iteration reduces the error in the 'best' minimax sense.

Although the iteration (2.3) will work for any value of ω in the range $0 < \omega < 2$, the best convergence rate results from the use of that value of ω for which the *spectral radius* (i.e., the modulus of the largest eigenvalue) of M_ω is minimized. Wachspress and Habetler (1961) have shown that the required value is

$$\omega_{opt} = \frac{2}{1 + \sqrt{(2\tau_1 - 1 + 4k_1)}}, \qquad (2.6)$$

where $\tau_1 = \psi_1 A \psi_1$, $k_1 = \psi_1 LU \psi_1$ and ψ_1 is the eigenvector corresponding to the largest eigenvalue of M_ω at the optimum ω (here, we have changed their notation to suit our purposes), and that the corresponding spectral radius $\bar{\mu}(M_\omega)$, denoted above by $1 - \eta$ satisfies

$$\bar{\mu}(M_\omega) = \left[1 - \frac{\tau_1}{\sqrt{(2\tau_1 - 1 + 4k_1)}}\right] \left[1 + \frac{\tau_1}{\sqrt{(2\tau_1 - 1 + 4k_1)}}\right]^{-1} \qquad (2.7)$$

The *average rate of convergence per iteration*, defined as the lower bound of

$$-\frac{1}{n} \log \left[\frac{T_n\left(\frac{2\lambda}{1-\eta} - 1\right)}{T_n\left(\frac{2}{1-\eta} - 1\right)} \right]$$

is not less than

$$R_1 = -\frac{1}{n} \log \left[\frac{1}{T_n(\gamma)}\right]. \qquad (2.8)$$

Since

$\lim_{n \to \infty} \log \{\cosh (n \cosh^{-1} y)\}^{1/n}$ is $\cosh^{-1} y$, and

$$\cosh^{-1}(1 + 2\eta) \simeq 2\sqrt{\eta} \text{ for } \eta \ll 1,$$

we have immediately for large n, optimum α_n, β_n, and $\bar{\mu}(M_\omega)$ close to 1,

$$R_1 = 2\sqrt{\eta}. \qquad (2.9)$$

Wachspress and Habetler investigated the dependence of η on ε, where $\varepsilon = 1 - \bar{\mu}(H)$ and $\bar{\mu}(H)$ is the spectral radius of H, the error operator of the Simultaneous Iteration (1.3) for some typical diffusion problems, and were able to compare the convergence rates of the S.O.R. and the accelerated S.S.O.R. methods. We now state some of their results which we need for comparison in a later section of this paper.

For a value of ω satisfying (2.6) and

$$\tau_1 \ll \sqrt{(2\tau_1 - 1 + 4k_1)},$$

we have

$$\eta \simeq \frac{2\tau_1}{\sqrt{(2\tau_1 - 1 + 4k_1)}}, \tag{2.10}$$

and since $k_1 \leqslant 1$ and $\tau_1 \geqslant \varepsilon = 1 - \bar{\mu}(H)$, we obtain as a lower bound for η, the result $\frac{2\varepsilon}{\sqrt{3}}$. Hence, the accelerated S.S.O.R. will have an average convergence rate at least as large as

$$R_1 \simeq \frac{2\sqrt{2}}{3^{1/4}}\varepsilon^{1/2} = 2{\cdot}14\sqrt{\varepsilon}. \tag{2.11}$$

3. We define our extrapolation of (1.7) by

$$x^{(n+1)} = x^{(n)} + \omega(Lx^{(n+1)} + Ux^{(n+1)} - \omega LUx^{(n+1)} + \omega LUx^{(n)} + d - x^{(n)}),$$

which can be written in the form

$$(I - \omega L)(I - \omega U)x^{(n+1)} = [\omega^2 LU + (1 - \omega)I]x^{(n)} + \omega d. \tag{3.1}$$

This iteration, which we call the Extrapolated Modified Aitken iteration, is the subject of the present investigation.

The error vectors in this iteration satisfy

$$e^{(n+1)} = Q_\omega e^{(n)}, \tag{3.2}$$

where $$Q_\omega = I - \omega(I - \omega U)^{-1}(I - \omega L)^{-1}A,$$

and if ω is the same for each iteration, we have the usual expression (1.3) for a stationary method.

We define the matrix T_ω by the similarity transformation

$$T_\omega = (I - \omega U)Q_\omega(I - \omega U)^{-1} = I - \omega(I - \omega L)^{-1}A(I - \omega U)^{-1}, \tag{3.4}$$

and we consider the matrix

$$I - T_\omega \equiv \omega(I - \omega L)^{-1}A(I - \omega U)^{-1}. \tag{3.5}$$

Since A is a positive definite symmetric matrix, it possesses a positive definite square root, $A^{\frac{1}{2}}$, defined as $Y\Lambda Y^{-1}$ where Y is the matrix of the eigenvectors of A and Λ is a diagonal matrix whose elements are the positive square roots of the eigenvalues of A. Thus, we may write

$$I - T_\omega \equiv \omega[(I - \omega L)^{-1}A^{\frac{1}{2}}][A^{\frac{1}{2}}(I - \omega L^T)^{-1}].$$

Hence, $\frac{1}{\omega}(I - T_\omega)$ is the product of a matrix times its transpose, and so has non-negative eigenvalues. Moreover, $I - \omega L$ is non singular, and A is positive definite. Consequently, the matrix $I - T_\omega$ is positive definite for $\omega > 0$ and so is $I - Q_\omega$, since its eigenvalues are those of $I - T_\omega$.

We shall now investigate the convergence of the iteration process (3.1) in order to determine the optimum value of the iteration parameter ω. Let us assume that, for a particular value of ω, the eigenvalues and eigenvectors of Q_ω are θ_s and ψ_s for $S = 1, 2, \ldots N$, respectively.

Then, by definition we have for $S = 1, 2, \ldots n$

$$\theta_s = \frac{\psi_s[(1 - \omega)I + \omega^2 LU]\psi_s}{\psi_s[(1 - \omega)I + \omega A + \omega^2 LU]\psi_s}, = \frac{\{2(\omega^{-1} - 1) + 2\omega k_s + \tau_s\} - \tau_s}{\{2(\omega^{-1} - 1) + 2\omega k_s + \tau_s\} + \tau_s}, = \frac{\rho_s - \tau_s}{\rho_s + \tau_s}, \tag{3.6}$$

where $$\rho_s = 2(\omega^{-1} - 1) + 2\omega k_s + \tau_s$$

$$k_s = \psi_s LU\psi_s$$

and $$\tau_s = \psi_s A\psi_s.$$

Now, the criterion for convergence is

$$-1 < \theta_s < 1 \text{ for } S = 1, 2, \ldots N.$$

The upper limit is satisfied when $\omega > 0$, whilst the lower limit is given when $\omega = \omega_f$ and $\rho_N = 0$, i.e., when

$$2(1 - \omega_f) + 2\omega_f^2 k_N + \omega_f \tau_N = 0. \tag{3.7}$$

Unfortunately, it is not easy to discern the form of variation of the largest eigenvalues of Q_ω with ω. Accordingly, we have proceeded empirically, and have calculated (see Section 4) the extreme values of θ for the lowest-order finite-difference representation of two model problems, namely

$$\text{Problem I: } \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} = 0$$

in the unit square with prescribed boundary values;

$$\text{Problem II: } \frac{d^2\phi}{dx^2} = 0$$

on a unit interval with prescribed terminal values.

The results are displayed in Fig. 1, while Fig. 2 shows the corresponding results for S.S.O.R. We observe that the extreme values of θ are monotonic in ω, and for convergence to be possible, ω must be in the range $0 < \omega < \omega_f$ where ω_f is given by (3.7). Furthermore, it is clear that the optimum value of ω is that for which the extreme positive and negative values of θ are of equal magnitude, that is, such that

$$\left(\frac{\rho_1 - \tau_1}{\rho_1 + \tau_1}\right) = -\left(\frac{\rho_N - \tau_N}{\rho_N + \tau_N}\right). \tag{3.8}$$

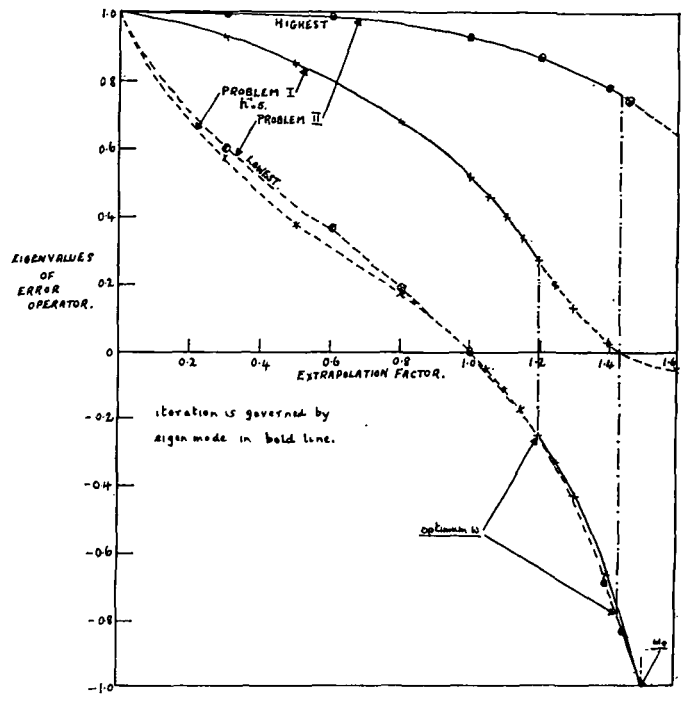

Fig. 1.—Extrapolated Aitken iteration: Problems I and II

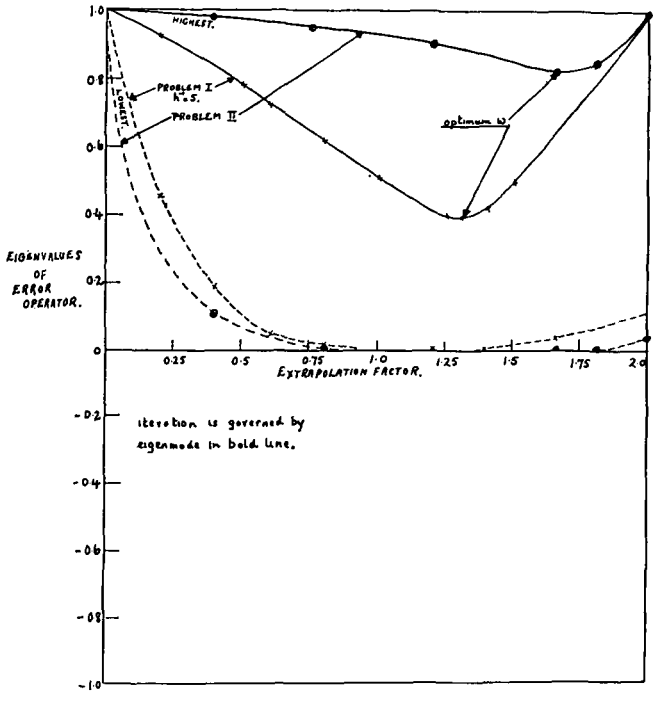

Fig. 2.—Symmetric successive over-relaxation: Problems I and II

Thus, the optimum ω is given as the solution of the quartic equation,

$$2(1 - \omega_0 + \omega_0\tau_1 + \omega_0^2 k_1)(1 - \omega_0 + \omega_0\tau_N + \omega_0^2 k_N) = \omega_0\tau_N(1 - \omega_0 + \omega_0\tau_1 + \omega_0^2 k_1) + \omega_0\tau_1(1 - \omega_0 + \omega_0\tau_N + \omega_0^2 k_N). \quad (3.9)$$

Finally, the spectral radius of the Extrapolated Modified Aitken Method at the optimum ω is

$$\bar{\mu}(\mathcal{Q}_\omega) = \frac{\rho_1 - \tau_1}{\rho_1 + \tau_1} \text{ for } \omega = \omega_0 = \left[1 - \frac{\tau_1}{2(\omega_0^{-1} - 1) + 2\omega_0 k_1 + \tau_1}\right] \left[1 + \frac{\tau_1}{(2\omega_0^{-1} - 1) + 2\omega_0 k_1 + \tau_1}\right]^{-1} \quad (3.10)$$

Furthermore, since the eigenvalues of $\mathcal{Q}_\omega$ are real this method can also be accelerated by means of Chebyshev polynomials. In this case, we require that each contribution to $e^{(0)}$ from the eigenvectors of $\mathcal{Q}_\omega$ shall be weighted in $e^{(n)}$ by the factor

$$\frac{T_n\left(\frac{\theta}{\delta}\right)}{T_n\left(\frac{1}{\delta}\right)}$$

where $\delta < 1$ is the spectral radius of $\mathcal{Q}_\omega$.

Thus, we now describe the iteration in full by the equations

$$(\boldsymbol{I} - \omega\boldsymbol{L})(\boldsymbol{I} - \omega\boldsymbol{U})\hat{\boldsymbol{x}}^{(n+1)} = (\omega^2\boldsymbol{L}\boldsymbol{U} + (1 - \omega)\boldsymbol{I})\boldsymbol{x}^{(n)} + \omega\boldsymbol{d}$$
$$\boldsymbol{x}^{(n+1)} = \boldsymbol{x}^{(n)} + \alpha_n(\hat{\boldsymbol{x}}^{(n+1)} - \boldsymbol{x}^{(n)}) + \beta_n(\boldsymbol{x}^{(n)} - \boldsymbol{x}^{(n-1)}), \quad (3.11)$$

where the coefficients α_n and β_n are determined from the relations,

$$\alpha_n = \frac{2}{\delta}\frac{T_n\left(\frac{1}{\delta}\right)}{T_{n+1}\left(\frac{1}{\delta}\right)}$$

$$\beta_n = \frac{T_{n-1}\left(\frac{1}{\delta}\right)}{T_{n+1}\left(\frac{1}{\delta}\right)}, (n = 1, 2 \ldots) \quad (3.12)$$

and $\alpha_0 = 1$.

Obviously, the iteration process will work satisfactorily for any ω in the range $0 < \omega < \omega_f$, but once again, as in the S.S.O.R., the best convergence rate is attained when the optimum ω_0 is used. The average rate of convergence per complete iteration for the optimum acceleration of the Extrapolated Modified Aitken method is not less than

$$R_2 = -\frac{1}{n}\log\left\{\frac{1}{T_n\left(\frac{1}{\delta}\right)}\right\}, \quad (3.13)$$

where $\delta = \dfrac{1-\nu}{1+\nu}$ as given in (3.10) and

$$\nu = \frac{\tau_1}{2(\omega_0^{-1} - 1) + 2\omega_0 k_1 + \tau_1}. \quad (3.14)$$

Now, for $\tau_1 \ll 2(\omega_0^{-1} - 1) + 2\omega_0 k_1 + \tau_1$ and large n, (3.13) further simplifies to

$$R_2 \simeq 2\sqrt{\nu}. \quad (3.15)$$

We shall now compare R_2 with R_1 for the Laplace problem and deduce some general conclusions on the average convergence of the new method. Substituting $k_1 = 1$, $\tau_1 = \varepsilon$ and $\omega_0 = 1{\cdot}5$ in (3.14) we obtain as a lower bound for ν, the result

$$\nu = 0{\cdot}428\varepsilon, \quad (3.16)$$

and thus the accelerated Extrapolated Aitken iteration will have an average convergence rate at least as large as

$$R_2 \simeq 1{\cdot}31\sqrt{\varepsilon}.$$

This compares unfavourably with equation (2.11) for the S.S.O.R. We cannot provide a more direct comparison, for since the optimum ω's differ in the two methods, the quantities k_1 and τ_1 will differ also. However, the lower bound in (3.16) was obtained when we assumed $k_1 = 1$. If we now assume $\rho_1 = O(\varepsilon^{\frac{1}{2}})$ and $\tau_1 = O(\varepsilon)$ then we have $\nu = \dfrac{\tau_1}{\rho_1} = O(\varepsilon^{\frac{1}{2}})$, and it follows immediately that the Extrapolated Modified Aitken iteration with Chebyshev acceleration can converge faster than the S.O.R. method. For this particular case to occur in Problem I of the numerical experiments, we need

$$k_1 \approx 0{\cdot}25,\ \omega_0 = 1{\cdot}2 \text{ and } \tau_1 \approx 0{\cdot}25 = O(\varepsilon)$$

whence

$$\rho_1 = 2(\omega_0^{-1} - 1) + 2\omega_0 k_1 + \tau_1 = 0{\cdot}45 \approx O(\varepsilon^{\frac{1}{2}}).$$

These results correspond to similar conditions, stated by Wachspress and Habetler in their study of the S.S.O.R. method.

4. We now discuss the computational aspects of performing the proposed iteration (3.1) with a generated rather than a stored matrix. We shall consider only matrices arising from the lowest-order finite-difference representation of second-order elliptic partial differential equations on a rectangular grid of mesh points, and as illustration we obtain the numerical solution of Problem I above.

We use the subscripts i and j to denote the column and row locations of the point (i, j) on a rectangle of $(M \times P)$ mesh points. For such points, the 5-point finite-difference equations are of the form

$$-l_{i,j}\phi_{i-1,j} - b_{i,j}\phi_{i,j-1} + \phi_{i,j} - t_{i,j}\phi_{i,j+1} - r_{i,j}\phi_{i+1,j} = d_{i,j} \quad (4.1)$$

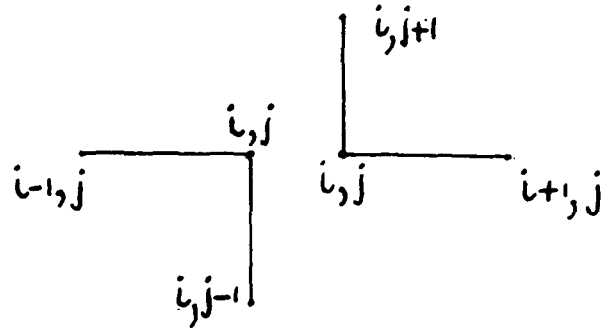

Fig. 3.—Problem I

within the rectangle specified by $0 \leqslant i \leqslant m$ and $0 \leqslant j \leqslant P$. The coefficients l, b, t, r relate mnemonically to grid points to the left, bottom, top and right of (i, j).

On the boundaries of the domain the equation (4.1) holds with the coefficients

$$\left.\begin{aligned} l_{1,j} = r_{M-1,j} &= 0 \quad \text{for } 1 \leqslant j \leqslant P-1 \\ \text{and} \quad b_{i,1} = t_{i,p} - 1 &= 0 \quad \text{for } 1 \leqslant i \leqslant M-1 \end{aligned}\right\} \quad (4.2)$$

for a columnwise ordering of points.

The equations (4.1) for $1 \leqslant j \leqslant P-1$ and $1 \leqslant i \leqslant M-1$ combine to form the matrix equation (1.1) in the following manner. The point (i, j) of the network is the $[(i-1)(P-1)+j]^{th}$ equation in the matrix array. Let q denote $[(i-1)(P-1)+j]$, so that the equation (4.1) expressed in terms of the elements $a_{l,m}$ of A is

$$-a_{q,q-p+1}x_{q-p+1} - a_{q,q-1}x_{q-1} + x_q - a_{q,q+1}x_{q+1} - a_{q,q+p-1}x_{q+p-1} = d_q \quad (4.3)$$

for $1 \leqslant q \leqslant (P-1)(M-1)$ with the boundary conditions (4.2) being expressed in the form

$$\left.\begin{aligned} &a_{q,q-1} = 0, \quad \text{when } q-1 \text{ is a multiple of } P-1, \\ &a_{q,q+1} = 0, \quad \text{when } q \text{ is a multiple of } P, \\ &a_{q,q-P+1} = 0, \text{ when } q - P + 1 \text{ is negative} \\ &\text{and} \\ &a_{q,q+P-1} = 0, \text{ when } Q + P - 1 \text{ exceeds } (P-1)(M-1). \end{aligned}\right\} \quad (4.4)$$

We now discuss the generation of the two triangular matrices $(I - \omega L)$ and $(I - \omega U)$ on the network of grid lines. In the matrix array, they are of the form,

$$\left.\begin{aligned} &-\omega a_{q,q-p+1}x_{q-p+1} - \omega a_{q,q-1}x_{q-1} + x_q \\ \text{and} \quad &x_q - \omega a_{q,q+1}x_{q+1} - \omega a_{q,q+p-1}x_{q+p-1} \end{aligned}\right\} \quad (4.5)$$

subject to the boundary conditions (4.4), and by reference to Fig. 3 we see that these are equivalent to the expressions

$$-\omega l_{i,j}\phi_{i-1,j} - \omega b_{i,j}\phi_{i,j-1} + \phi_{i,j} \text{ and}$$
$$\phi_{i,j} - \omega t_{i,j}\phi_{i,j+1} - \omega r_{i,j}\phi_{i+1,j} \quad (4.6)$$

subject to the boundary conditions (4.2).

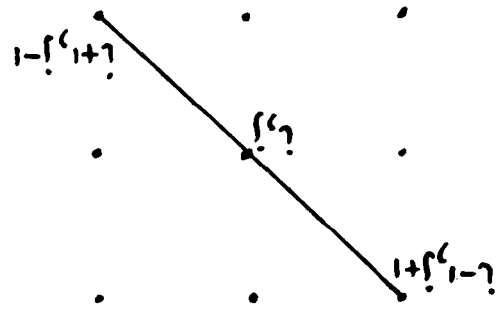

Fig. 4

Similarly, the matrix $[\omega^2 \boldsymbol{LU} + (1 - \omega)I]$ has the form,

$$\omega^2 a_{q,q-p} a_{q-p,q-p+1} x_{q-p+2} + [(\omega - 1) + \omega^2(a_{q,q-p} a_{q-p,q} + a_{q,q-1} a_{q-1,q})] x_q + \omega^2 a_{q,q-1} a_{q-1,q+p-1} x_{q+p-2} \quad (4.7)$$

for $1 \leqslant q \leqslant (P - 1)(M - 1)$, subject to the conditions (4.4). This matrix is directly related to the simple stencil, illustrated in Fig. 4, at the point (i, j) on the mesh, and it can be generated on the mesh by the simple equation at the point (i, j),

$$\omega^2 l_{i,j} t_{i-1,j} \phi_{i-1,j+1} + [(\omega - 1) + \omega^2(l_{i,j} r_{i,j} + b_{i,j} t_{i,j})] \phi_{i,j} + \omega^2 b_{i,j} r_{i,j-1} \phi_{i+1,j-1} \quad (4.8)$$

subject to the boundary conditions (4.2).

Finally, we compare the amount of work done in each iteration. For each of the two sweeps of the S.S.O.R. method we need to do five multiplications and five additions per equation whilst in the Extrapolated Modified Aitken method a complete iteration involves seven multiplications and seven additions per equation. This is true only when quantities such as $l_{i,j} t_{i-1,j}$, etc. in (4.8) are either readily available in an auxiliary store or are simple fractions which can be immediately deduced, i.e., such as in the Laplace Equation. Hence, the latter method can be more efficient by a factor of approximately 1·4.

5. Experimental programs were written for the Manchester University Mercury computer to perform the iteration procedures discussed in this paper for the numerical solution of the Problem I above for the mesh sizes, $h^{-1} = 5$, 10 and 20, and Problem II, for the mesh size, $h^{-1} = 16$.

All iterations were initiated from the same approximation $\phi^{(0)}$ (i.e., an arbitrary vector whose components were obtained by crudely interpolating between the boundary conditions) and continued until the convergence criterion

$$\max_{i,j} |\phi_{i,j}^{(n+1)} - \phi_{i,j}^{(n)}| < 5 \times 10^{-5}$$

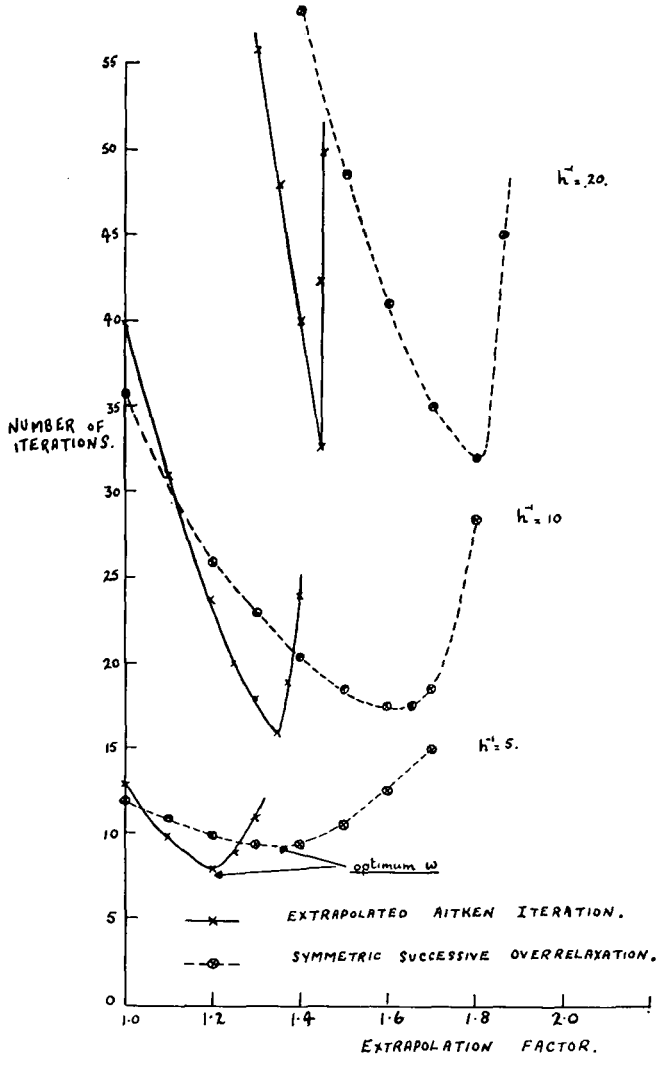

Fig. 5

was satisfied, for $1 \leqslant i \leqslant M - 1$ and $1 \leqslant j \leqslant M - 1$, where M is the number of mesh points in the i and j directions.

Since the iteration method proposed by (3.12) differs from the S.S.O.R. method (2.1) prior to the acceleration by Chebyshev polynomials, we first compare the results from numerical experiments for the two iteration processes without Chebyshev acceleration.

The S.S.O.R. method for Problem I, $h^{-1} = 5$, 10, 20, and Problem II, $h^{-1} = 16$, are shown in Figs. 5 and 6, respectively, the number of iterations being plotted against the over-relaxation factor ω. The mechanism by which the optimum factor was obtained has already been discussed, with reference to Fig. 2. In detail, the fundamental eigenfunction of the S.S.O.R. method was determined by iteration, and the quantities k_1 and τ_1 evaluated from equations defined in (2.6). Then, from the analysis shown in Section 2 we have the final results given in Table 1.

Table 1

	k_1	τ_1	ω EQN. (2.6)	ω (EXP)	$\bar{\mu}$ EQN. (2.7)	$\bar{\mu}$ (EXP)
Problem I, $h^{-1} = 5$	0·2056	0·2302	1·3	1·3 (Fig. 5)	0·39	0·39 (Fig. 2)
Problem II, $h^{-1} = 16$	0·2496	0·0174	1·7	1·7 (Fig. 6)	0·825	0·825 (Fig. 2)

Table 2

	k_1	τ_1	k_N	τ_N	ω EQN. (3.9)	ω (EXP)	$\bar{\mu}$ EQN. (3.10)	$\bar{\mu}$ (EXP)
Problem I, $h^{-1} = 5$	0·2010	0·2051	0·0278	0·6669	1·2	1·2 (Fig. 5)	0·26	0·26 (Fig. 1)
Problem II, $h^{-1} = 16$	0·2493	0·0171	0·0904	0·3986	1·43	1·43 (Fig. 6)	0·77	0·77 (Fig. 1)

The Extrapolated Modified Aitken iteration method was investigated for the same problems. Again, the number of iterations versus the extrapolation factor is plotted and shown in Figs. 5 and 6, while Fig. 1 shows the highest and lowest eigenvalues of the error operator plotted as functions of ω. As with S.S.O.R., the eigenfunctions corresponding to the highest and lowest eigenvalues at the optimum ω were determined by iteration, by using equations (3.6), and the quantities k_1, τ_1, k_N and τ_N determined. Then, by using the analysis developed in Section 3, we have the final results given in Table 2.

Furthermore, values of k_N and τ_N were determined by iteration, and equation (3.7) solved to determine ω_f and hence the range of ω for the convergence of the Extrapolated Modified Aitken method. The results obtained are given in Table 3.

We see that the Extrapolated Modified Aitken method converges for $0 < \omega < 1{\cdot}5$ for the two problems under investigation, and that excellent agreement between the theoretical and experimental results was obtained.

Finally, the iterations are compared after their acceleration by Chebyshev polynomials. The theoretical results are given in Table 4, whilst the experimental results for Problem I are shown in Fig. 7, together with those for the Simultaneous iteration (1.2), Gauss–Seidel iteration (1.4), Aitken iteration (1.4 and 1.5), the S.O.R. method (2.2) (at the optimum over-relaxation factor), and the unaccelerated Extrapolated Aitken and S.S.O.R. methods.

Table 3

	k_N	τ_N	ω_f EQN. (3.7)	ω_f (EXP)
Problem I, $h^{-1} = 5$	0·100	0·367	1·5	1·5 (Fig. 1)
Problem II, $h^{-1} = 16$	0·111	0·334	1·5	1·5 (Fig. 1)

Conclusions

For the problems discussed in this paper, the Extrapolated Modified Aitken method has been shown to have results comparable with the S.S.O.R. method. Further numerical experiments on general diffusion equations are necessary before any general conclusions can be made regarding the new iteration process, for Wachspress and Habetler have recently shown that Problem I cannot be regarded as a true model problem for the S.S.O.R. method.

Acknowledgement

The author is indebted to Dr. D. W. Martin of the National Physical Laboratory for discussion and assistance with the presentation of the material in this paper.

(*See p.* 201 *for references*)

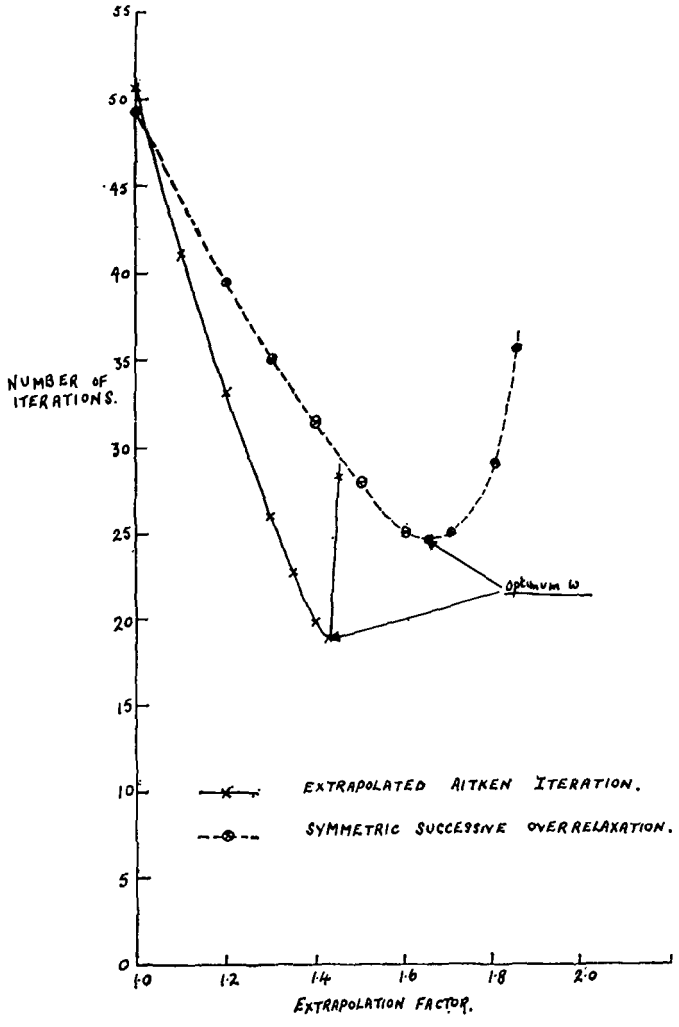

Fig. 6.—Problem II

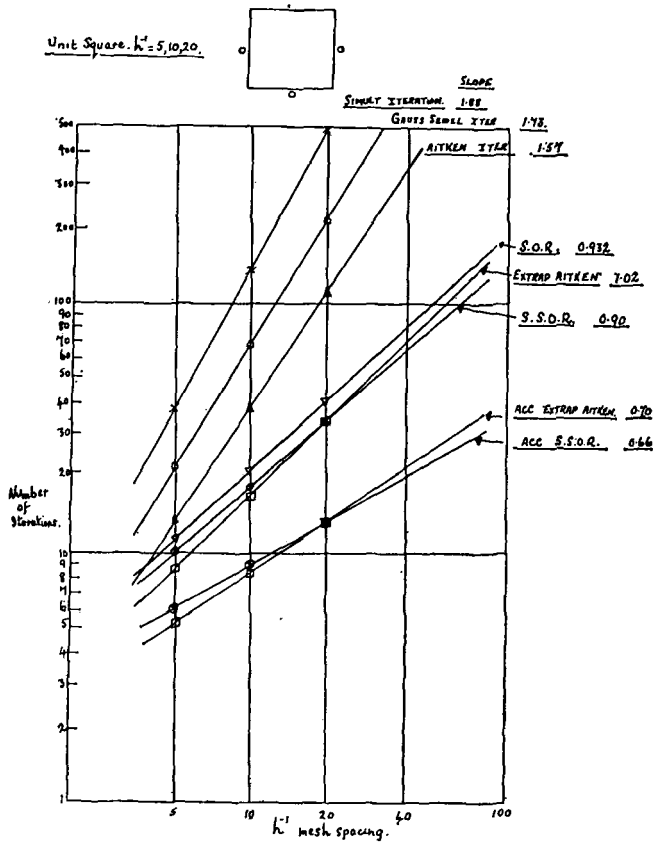

Fig. 7.—Problem I

Table 4

	CHEBYSHEV ACCELERATION			
	EXTRAPOLATED MODIFIED AITKEN METHOD		S.S.O.R. METHOD	
	ASYMPTOTIC CONVERGENCE RATE	NUMBER OF ITERATIONS REQUIRED TO REDUCE ERROR VECTOR NORM BY A FACTOR OF 5×10^{-5}	ASYMPTOTIC CONVERGENCE RATE	NUMBER OF ITERATIONS REQUIRED TO REDUCE ERROR VECTOR NORM BY A FACTOR OF 5×10^{-5}
Problem I, $h^{-1} = 5$	0·9	5	0·9	5
$h^{-1} = 10$	0·64	8	0·57	9
$h^{-1} = 20$	0·39	13	0·41	13
Problem II, $h^{-1} = 16$	0·34	12	0·38	11

References

AITKEN, A. C. (1950). "Studies in Practical Mathematics. On the Iterative Solution of a System of Linear Equations," *Proc. Roy. Soc. Edinburgh*, A.63, pp. 652–60.

FORSYTHE, G. E., and WASOW, W. R. (1960). *Finite Difference Methods for Solving Partial Differential Equations*, pp. 220–35, London: Wiley.

HABETLER, G. J., and WACHSPRESS, E. L. (1961). "Symmetric Successive Over-Relaxation in Solving Diffusion Difference Equations," *Math. Comput.*, Vol. 15, pp. 356–63.

MARTIN, D. W., and TEE, G. J. (1961). "Iterative Methods for Linear Equations with Symmetric Positive Definite Matrix," *The Computer Journal*, Vol. 4, No. 3, pp. 242–53.

SHELDON, J. W. (1955). "On the Numerical Solution of Elliptic Difference Equations," *Math. Tables. Aids. Comput.*, Vol. 9, p. 101.

Book review: Dynamic programming

Applied Dynamic Programming, by RICHARD E. BELLMAN and STUART E. DREYFUS, 1963; 363 pages. (London: *Oxford University Press*, £2 15s. 6d.)

The type of problem giving rise to the technique known as Dynamic Programming can be formulated in a few words. Supposing that I want to invest £1000 in 10 projects ("activities") $A_1 \ldots A_{10}$ and that the profit of A_i in terms of the invested capital x_i is given by a function $g_i(x_i)$. How should I subdivide my £1000 into 10 portions $x_1 \ldots x_{10}$ so as to obtain the maximum, called $f_{10}(1000)$, of the total profit $g_1(x_1) + \ldots + g_{10}(x_{10})$?

The answer given by Dynamic Programming is this. Assume that we already know the solution to the problem for the case of 9 activities. Allocating £x_{10} to activity A_{10} leaves £$(1000 - x_{10})$ to be invested in $A_1 \ldots A_9$. Naturally, $x_1 \ldots x_9$ will be chosen so as to maximize the return for $A_1 \ldots A_9$, and this maximum has the known value $f_9(1000 - x_{10})$.

The return for all 10 projects is then

$$g_{10}(x_{10}) + f_9(1000 - x_{10}),$$

and by varying x_{10} the maximum value $f_{10}(1000)$ of this expression can be computed.

In general, the problem of finding the maximum $f_N(x)$ of the sum $\sum_{i=1}^{N} g_i(x_i)$ subject to $\sum_{i=1}^{N} x_i = x$ is solved by the recurrence relation

$$f_n(x) = \max_{x_n} [g_n(x_n) + f_{n-1}(x - x_n)].$$

This method, and its modifications for solving generalized versions of the basic problem, has already been extensively discussed in Richard Bellman's first book published about six years ago. In spite of this, reports of its application up to now have been few and far between and, in contrast to Linear Programming which, in some industries, is used almost as widely as PAYE calculations, its "image" is still that of an ingenious mathematical technique of doubtful practical value.

The new book should do much to alter this situation. While the theoretical connections between Dynamic Programming and classical methods of analysis (such as Lagrange Multipliers and the calculus of variations) are by no means neglected, its emphasis is on practical applications and on the achievement of actual numerical results. The remarkable flexibility of the method is shown by the wide range of its possible uses; among those mentioned in the book are the loading of a ship of given capacity so as to maximize the total value of the cargo; transportation problems with non-linear cost functions; the problem of finding a defective coin by the least number of weighings; studies of the reliability of multi-component devices; minimizing the cost due to shortages of replacement parts; optimum advertising campaigns; smoothing and scheduling problems; computation of optimal trajectories; and the minimization of the number of steps when searching for the zero of a function.*

The solution of these and other problems is illustrated by detailed flow diagrams of the programs developed for the various modifications of the method; and the tabulation of data and results, together with statements of actual running times on existing computers, leave the reader in no doubt that problems of practical importance can be and have been solved by this method. The book will be required reading for anyone concerned with problems of optimization that cannot be solved by more familiar algorithms.

Paper, printing and proof-reading are good, but one slip (on page 19), though not affecting the argument, should be corrected in future editions: the statement that 10^7 seconds (actually about four months) "is something of the order of 10^5 hours, and thus of the order of magnitude of ten years."

D. G. PRINZ

* Another interesting application, using Dynamic Programming as a subroutine in Linear Programming for minimizing the scrap in cutting stock, was published too recently to be included (P. C. Gilmore and R. E. Gomory, "A linear-programming approach to the cutting-stock problem," *Jour. Op. Res. Soc. Am.*, Vol. 9, p. 849, Nov.–Dec. 1961)