On the Round-Off Error in the Method of Successive Over-Relaxation

By M. Stuart Lynn

Abstract. The asymptotic behavior of the round-off error, which accumulates when the well-known iterative method of (point) successive over-relaxation is used to solve a large-scale system of linear equations, is examined by means of a statistical model. The local round-off errors are treated as independent random variables and expressions for the mean and variance of the accumulated round-off error are obtained, as the number of iterations tends to infinity.

1. Introduction. Consider the system of linear equations

$$\mathbf{C}\mathbf{x} = \mathbf{b}, \tag{1.1}$$

where $\mathbf{C}$ is an $n \times n$ real matrix and $\mathbf{b}$ an $n \times 1$ vector. In the numerical solution of (1.1), round-off errors accumulate. Wilkinson [14, 15] and Turing [11] have considered the effect of this where direct methods of solution are involved. As far as iterative methods are concerned, bounds for the round-off error occurring in a general iterative procedure have been obtained by Urabe [12] and Descloux [2].

In order to attempt to obviate the usual criticism that such bounds may be unrealistic, we shall use the technique, familiar in other fields of numerical analysis, e.g. Henrici [7], of treating the local round-off errors (see below) as independent random variables, and then obtaining expressions for the asymptotic behaviour of the mean and variance of the accumulated round-off error. In connection with iterative procedures, the statistical model employed was first used by Abramov [1] in studying the round-off error generated by the Jacobi procedure for solving (1.1). More recently, Golub [5] has used similar techniques to analyze the Richardson second-order and Chebyshev semi-iterative methods. We shall consider another method, namely the method of (point) successive over-relaxation: Young [16], Varga [13].

The statistical model which is employed is certainly open to many severe criticisms, e.g. Forsythe [3], more so, in fact, than in other fields of numerical analysis where perhaps it can be used with more confidence. Principally, the fundamental assumption that the local round-off errors are either independent or random is certainly questionable, particularly after a large number of iterations, and the procedure reaches a 'state of numerical convergence': Sibagaki [10], Urabe [12]. However, the somewhat curious nature of the results obtained from this kind of analysis (see particularly Theorem 2) are perhaps not without interest. We would again refer the reader to Golub [4, 5] for a more detailed appraisal of the use of a statistical model in connection with iterative techniques, compared with the use of bounds.

Received February 25, 1963. Revised May 20, 1963. This work was carried out mainly under the sponorship of the Office of Naval Research, and partly within the research program of the National Physical Laboratory. The paper is published by permission of the Director of the Laboratory.

Having obtained expressions for the mean and variance of the accumulated round-off error, it would not be strictly correct to apply the central limit theorem and deduce a normal distribution, since it can readily be shown that the accumulated round-off error remains uniformly bounded as the number of iterations tends to infinity. However, using an analysis precisely analogous to that used by Golub [5] we can nevertheless obtain probabilistic bounds for the round-off error. The contents of this are expressed in Theorem 3.

The method of point successive over-relaxation or the Young-Frankel method may, as usual, be defined as follows. Let $\mathbf{C}$ be decomposed by

$$\mathbf{C} = \mathbf{S} - \mathbf{L} - \mathbf{U} = \mathbf{S} - \mathbf{B} \tag{1.2}$$

where $\mathbf{S}$ is a diagonal matrix and $\mathbf{L}$, $\mathbf{U}$ are strictly lower and upper triangular matrices, respectively. For arbitrary $\mathbf{x}^{(0)}$, the sequence $\{\mathbf{x}^{(k)}\}$ of vectors is defined by

$$\mathbf{S}\mathbf{x}^{(k+1)} = \omega(\mathbf{L}\mathbf{x}^{(k+1)} + \mathbf{U}\mathbf{x}^{(k)} + \mathbf{b}) + (1 - \omega)\mathbf{S}\mathbf{x}^{(k)} \qquad k = 0, 1, 2, \cdots, \tag{1.3}$$

where ω is a *relaxation-parameter* which necessarily lies in the range

$$0 < \omega < 2. \tag{1.4}$$

We shall henceforth assume that $\mathbf{C}$ is symmetric positive-definite, so that

$$\mathbf{C} = \mathbf{C}^*, \tag{1.5}$$

and

$$\mathbf{L}^* = \mathbf{U}, \qquad \mathbf{B} = \mathbf{B}^*, \tag{1.6}$$

where, in general, $\mathbf{A}^*$ denotes the (conjugate) transpose of a matrix $\mathbf{A}$ (since we shall always be working with real matrices, the conjugation operation is not involved). Condition (1.4) is then both necessary and sufficient to ensure convergence of (1.3) to $\mathbf{x}$ as $k \to \infty$: Ostrowski [9]. Without loss of generality, we may also assume that

$$\mathbf{S} = \mathbf{I} \tag{1.7}$$

the identity matrix, since we may write (1.1) in the form,

$$\mathbf{S}^{-1/2}\mathbf{A}\mathbf{S}^{-1/2}\mathbf{S}^{1/2}\mathbf{x} = \mathbf{S}^{-1/2}\mathbf{b}, \tag{1.8}$$

or

$$\mathbf{A}_1\mathbf{y} = \mathbf{b}_1, \tag{1.9}$$

where

$$\mathbf{A}_1 = \mathbf{S}^{-1/2}\mathbf{A}\mathbf{S}^{-1/2}, \qquad \mathbf{y} = \mathbf{S}^{1/2}\mathbf{x}, \qquad \mathbf{b}_1 = \mathbf{S}^{-1/2}\mathbf{b}, \tag{1.10}$$

and $\mathbf{A}_1$ is symmetric positive-definite with unity diagonal entries.

2. The Model. Now suppose that, instead of solving (1.3) exactly, we actually compute vectors $\tilde{\mathbf{x}}^{(k)}$, where (replacing $\mathbf{S}$ by $\mathbf{I}$)

$$\tilde{\mathbf{x}}^{(k+1)} = \omega(\mathbf{L}\tilde{\mathbf{x}}^{(k+1)} + \mathbf{U}\tilde{\mathbf{x}}^{(k)} + \mathbf{b}) + (1 - \omega)\tilde{\mathbf{x}}^{(k)} + \mathbf{e}^{(k+1)}, \quad k = 0, 1, 2, \cdots. \tag{2.1}$$

Here, $\{\mathbf{e}^{(k)}\}(k = 0, 1, 2, \cdots)$ are the *local round-off errors* in the sense that they are the errors committed at each iteration. Let

$$\mathbf{r}^{(k)} = \bar{\mathbf{x}}^{(k)} - \mathbf{x}^{(k)} \qquad (k = 0, 1, 2, \cdots) \tag{2.2}$$

be the *accumulated round-off error vector* at the k^{th} iteration. Then from (1.3) and (2.1),

$$\mathbf{r}^{(k+1)} = \omega(\mathbf{L}\mathbf{r}^{(k+1)} + \mathbf{U}\mathbf{r}^{(k)}) + (1 - \omega)\mathbf{r}^{(k)} + \mathbf{e}^{(k+1)}, \qquad k = 0, 1, 2, \cdots. \tag{2.3}$$

where

$$\mathbf{r}^{(0)} = \mathbf{e}^{(0)} = 0 \tag{2.4}$$

We treat the $\{\mathbf{e}^{(k)}\}$ as independent random variables (see the remarks in §1) with expected value

$$E[\mathbf{e}^{(k)}] = \boldsymbol{\varepsilon}_k ,$$

where we assume that

$$\boldsymbol{\varepsilon}_k \equiv \boldsymbol{\varepsilon} \qquad (k > 0) \tag{2.5}$$

where $\boldsymbol{\varepsilon}$ is constant; and with co-variance

$$D\mathbf{e}^{(k)} = \sigma^2\mathbf{R} \quad (k > 0); \qquad D\mathbf{e}^{(0)} = \mathbf{0} \tag{2.6}$$

where $\mathbf{R}$ is some positive definite symmetric matrix, again assumed constant, which we further necessarily assume commutes with $\mathbf{L}$ and $\mathbf{U}$, and hence with $\mathbf{A}$ and $\mathbf{B}$. By definition, we have that

$$\begin{aligned} D\mathbf{e}^{(k)} &= E[(\mathbf{e}^{(k)} - \boldsymbol{\varepsilon})(\mathbf{e}^{(k)} - \boldsymbol{\varepsilon})^*] \\ &= E[\mathbf{e}^{(k)}\mathbf{e}^{(k)*}] - \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^*, \end{aligned} \tag{2.7}$$

and so

$$E[\mathbf{e}^{(k)}\mathbf{e}^{(k)*}] = \sigma^2\mathbf{R} + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^* \tag{2.8}$$

Since we also assume that the $\{\mathbf{e}^{(k)}\}$ are independent, we also have that

$$\begin{aligned} E[\mathbf{e}^{(k)}\mathbf{e}^{(j)*}] &= E[\mathbf{e}^{(k)}]E[\mathbf{e}^{(j)}]^* \\ &= \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^* \qquad (k \neq j) \end{aligned} \tag{2.9}$$

Let

$$\mathbf{r}_k = E[\mathbf{r}^{(k)}] \tag{2.10}$$

denote the expected-value of the accumulated round-off error at the k^{th} iteration, and let

$$\mathbf{V}_k = D\mathbf{r}^{(k)}$$

denote the co-variance matrix of the same. We are interested in the behavior of $\mathbf{r}_k$ and $\mathbf{V}_k$ as $k \to \infty$.

Now from (2.3)

$$\mathbf{r}^{(k+1)} = \mathbf{T}\mathbf{r}^{(k)} + (\mathbf{I} - \omega\mathbf{L})^{-1}\mathbf{e}^{(k+1)},$$

where

$$\mathbf{T} = (\mathbf{I} - \omega\mathbf{L})^{-1}(\omega\mathbf{U} + \overline{1 - \omega}\mathbf{I}).$$

Thus

$$\mathbf{r}^{(k+1)} = \sum_{j=1}^{k+1} \mathbf{T}^{k+1-j}(\mathbf{I} - \omega\mathbf{L})^{-1}\mathbf{e}^{(j)}, \tag{2.11}$$

and so, since E is a linear operator,

$$\begin{aligned}\mathbf{r}_{k+1} &= \left\{\sum_{j=1}^{k+1} \mathbf{T}^{k+1-j}\right\}(\mathbf{I} - \omega\mathbf{L})^{-1}\boldsymbol{\varepsilon} \\ &= (\mathbf{I} - \mathbf{T}^{k+1})(\mathbf{I} - \mathbf{T})^{-1}(\mathbf{I} - \omega\mathbf{L})^{-1}\boldsymbol{\varepsilon} \\ &= (\mathbf{I} - \mathbf{T}^{k+1})\omega^{-1}\mathbf{C}^{-1}\boldsymbol{\varepsilon}.\end{aligned}$$

Hence we have:

THEOREM 1. *Under the foregoing statistical assumptions, and if* ω *satisfies* (1.4), *then*:

$$\mathbf{r}_k = (\mathbf{I} - \mathbf{T}^k)\omega^{-1}\mathbf{C}^{-1}\boldsymbol{\varepsilon},$$

and

$$\mathbf{r} = \frac{1}{\omega}\mathbf{C}^{-1}\boldsymbol{\varepsilon}, \tag{2.12}$$

where

$$\mathbf{r} = \lim_{k\to\infty} \mathbf{r}_k\,. \tag{2.13}$$

It should be noted that Theorem 1 does not rely upon the assumed independence of the $\{\mathbf{e}^{(k)}\}$.

3. The Co-Variance Matrix. We now consider the behavior of $\mathbf{V}_k$ as $k \to \infty$. From (2.8), (2.10) and the commutativity assumptions, we have immediately that†

$$\mathbf{V}_k = \sigma^2\mathbf{R}[(\mathbf{I} - \omega\mathbf{U})(\mathbf{I} - \omega\mathbf{L})]^{-1} \sum_{j=1}^{k} \mathbf{T}^{k-j}\mathbf{T}^{*k-j}. \tag{3.1}$$

Since $\mathbf{T}$ is neither symmetric nor normal, we lack any method for finding an explicit expression for the right-hand side of (3.1) (bounds do not interest us, since they defeat the object of a statistical investigation). This is unlike the case for Jacobi iteration as studied by Abramov [1], and it is in this context that the present problem possesses a separate interest. This is similarly unlike the situation for the Richardson second-order method as studied by Golub [5].

For a particular class of matrices, however, we may adopt a different approach; this is the class of matrices which have Property A and are σ_1-ordered: Young [16]. For our purposes we may define these as matrices for which $\mathbf{B}$ in (1.2) has

† The author is indebted to the referee for correcting an error in the author's original analysis, and for the more fruitful formulation and proof of Theorem 1.

the form:

$$\mathbf{B} = \overset{m}{\overbrace{\begin{matrix} 0 & F \\ F^* & 0 \end{matrix}}}^{m} \tag{3.2}$$

for some integer m, $0 < m < n$, so that F is an $m \times (n - m)$ sub-matrix. Thus we have:

Definition 2.1. If $\mathbf{B}$ in (1.2) has the form (3.2), we shall say that $\mathbf{C}$ in (1.2) has Property A and is σ_1-ordered.

In the terminology of Varga [13] $\mathbf{B}$ is weakly cyclic of index 2 and is in its normal form.

It is also worth remarking, perhaps, that if $\mathbf{C} = \mathbf{C}^*$, has Property A and is σ_1-ordered, then the assumption that $\mathbf{R}$ commutes with both $\mathbf{L}$ and $\mathbf{U}$ is satisfied if (i) $\mathbf{R}$ has the form:

$$\mathbf{R} = \overset{m}{\overbrace{\begin{matrix} R_1 & 0 \\ 0 & R_2 \end{matrix}}}^{m} \tag{3.3}$$

where $R_1 = R_1^*$ is an $m \times m$ sub-matrix, and $R_2 = R_2^*$ is an $(n - m) \times (n - m)$ sub-matrix, both of which are, of course, positive-definite; and (ii) $FR_1 = R_2F$.

Our principal result is:

THEOREM 2. *If* $\mathbf{C} = \mathbf{C}^*$ *is symmetric positive-definite, has Property A and is* σ_1*-ordered, if* $0 < \omega < 2$, *and if (2.5), (2.6) and (2.9) are satisfied, then*

$$\mathbf{V} = \sigma^2/[\omega(2 - \omega)] \cdot \mathbf{R}\mathbf{C}^{-1}, \tag{3.4}$$

where

$$\mathbf{V} = \lim_{k \to \infty} \mathbf{V}_k, \tag{3.5}$$

and

$$\mathbf{V}_k = D\mathbf{r}^{(k)}, \tag{3.6}$$

the co-variance of $\mathbf{r}^{(k)}$.

Proof. The proof proceeds by a sequence of lemmas. As an immediate consequence of (3.2), we find that:

LEMMA 1. *If* $\mathbf{C}$ *has Property A and is* σ_1*-ordered, then*

$$\mathbf{L}^2 = \mathbf{U}^2 = \mathbf{0}, \tag{3.7}$$

and hence $\mathbf{B}^2$ *commutes with both* $\mathbf{L}$ *and* $\mathbf{U}$.

LEMMA 2. *If* $\mathbf{C} = \mathbf{C}^*$ *has Property A and is* σ_1*-ordered, then the method defined by (1.3) converges if and only if the roots of*

$$\theta^2 - \lambda_i \theta + \alpha = 0 \qquad (i = 1, \cdots, n) \tag{3.8}$$

are less than one in modulus, where λ_i *are the eigenvalues of*

$$\mathbf{A} = 2(1 - \omega)\mathbf{I} + \omega^2\mathbf{B}^2, \tag{3.9}$$

and

$$\alpha = (1 - \omega)^2. \tag{3.10}$$

Proof. This is merely a re-formulation of the familiar relationship of Young [16], equation (2.4). We note that, for future reference,

$$0 \leqq \alpha < 1 \tag{3.11}$$

LEMMA 3. $\{\mathbf{r}^{(k)}\}(k = 0, 1, 2, \cdots)$ *satisfy the relationships*:

$$\mathbf{r}^{(k+2)} = \mathbf{A}\mathbf{r}^{(k+1)} - \alpha\mathbf{r}^{(k)} + \mathbf{f}^{(k+2)} \qquad (k = 0, 1, 2, \cdots) \tag{3.12}$$

$$\mathbf{r}^{(0)} = \mathbf{f}^{(0)} = 0; \qquad \mathbf{r}^{(1)} = \mathbf{f}^{(1)} = (\mathbf{I} + \omega\mathbf{L})\mathbf{e}^{(1)}, \tag{3.13}$$

where

$$\mathbf{f}^{(k+2)} = (\mathbf{I} + \omega\mathbf{L})\mathbf{e}^{(k+2)} + (\omega\mathbf{U} + \overline{1 - \omega}\mathbf{I})\mathbf{e}^{(k+1)} \qquad (k = 0, 1, 2, \cdots) \tag{3.14}$$

Proof. Since $\mathbf{L}^2 = \mathbf{0}$, one can immediately verify that

$$(\mathbf{I} - \omega\mathbf{L})^{-1} = (\mathbf{I} + \omega\mathbf{L}).$$

Hence, from (2.3), for $k = 0, 1, 2, 3, \cdots$,

$$\begin{aligned}
\mathbf{r}^{(k+2)} &= (\mathbf{I} - \omega\mathbf{L})^{-1}[(\omega\mathbf{U} + \overline{1 - \omega}\mathbf{I})\mathbf{r}^{(k+1)} + \mathbf{e}^{(k+2)}] \\
&= (\mathbf{I} + \omega\mathbf{L})[(\omega\mathbf{U} + \overline{1 - \omega}\mathbf{I})\mathbf{r}^{(k+1)} + \mathbf{e}^{(k+2)}] \\
&= (\overline{1 - \omega}\mathbf{I} + \omega^2\mathbf{B}^2)\mathbf{r}^{(k+1)} + \omega\mathbf{U}(\mathbf{I} - \omega\mathbf{L})\mathbf{r}^{(k+1)} + \omega(1 - \omega)\mathbf{L}\mathbf{r}^{(k+1)} \\
&\qquad + (\mathbf{I} + \omega\mathbf{L})\mathbf{e}^{(k+2)} \\
&= (\overline{1 - \omega}\mathbf{I} + \omega^2\mathbf{B}^2)\mathbf{r}^{(k+1)} + \omega\mathbf{U}[\omega\mathbf{U}\mathbf{r}^{(k)} + (1 - \omega)\mathbf{r}^{(k)} + \mathbf{e}^{(k+1)}] \\
&\qquad + \omega(1 - \omega)\mathbf{L}\mathbf{r}^{(k+1)} + (\mathbf{I} + \omega\mathbf{L})\mathbf{e}^{(k+2)} \\
&= (\overline{1 - \omega}\mathbf{I} + \omega^2\mathbf{B}^2)\mathbf{r}^{(k+1)} + \omega(1 - \omega)[\mathbf{L}\mathbf{r}^{(k+1)} + \mathbf{U}\mathbf{r}^{(k)}] \\
&\qquad + (\mathbf{I} + \omega\mathbf{L})\mathbf{e}^{(k+2)} + \omega\mathbf{U}\mathbf{e}^{(k+1)} \\
&= (2\overline{1 - \omega}\mathbf{I} + \omega^2\mathbf{B}^2)\mathbf{r}^{(k+1)} - (1 - \omega)^2\mathbf{r}^{(k)} + \mathbf{f}^{(k+2)} \\
&= \mathbf{A}\mathbf{r}^{(k+1)} - \alpha\mathbf{r}^{(k)} + \mathbf{f}^{(k+2)}
\end{aligned}$$

as required. The proof of (3.13) is entirely similar.

Now from (3.12) and (3.13), it follows that for certain polynomials $p_{kj}(\mathbf{A})$ in $\mathbf{A}$,

$$\mathbf{r}^{(k)} = \sum_{j=1}^{k} p_{kj}(\mathbf{A})\mathbf{f}^{(j)} \qquad (k = 1, 2, 3, \cdots) \tag{3.15}$$

Hence

$$\mathbf{r}_k = E[\mathbf{r}^{(k)}] = \sum_{j=1}^{k} p_{kj}(\mathbf{A})E[\mathbf{f}^{(j)}] \qquad (k = 1, 2, 3, \cdots)$$

$$\mathbf{r}_0 = 0; \qquad \mathbf{r}_1 = E[\mathbf{f}^{(1)}] = (\mathbf{I} + \omega\mathbf{L})\boldsymbol{\varepsilon}$$

Thus, since $\mathbf{A} = \mathbf{A}^*$,

$$\mathbf{V}_k = E[\mathbf{r}^{(k)}\mathbf{r}^{(k)*}] - \mathbf{r}_k\mathbf{r}_k^* = \sum_{i,j=1}^{k} p_{ki}(\mathbf{A})\,\mathbf{E}_{ij}p_{kj}(\mathbf{A}) \tag{3.17}$$

where

$$\mathbf{E}_{ij} = E[\mathbf{f}^{(i)}\mathbf{f}^{(j)*}] - E[\mathbf{f}^{(i)}]E[\mathbf{f}^{(j)*}] \tag{3.18}$$

We now wish to compute the $\mathbf{E}_{ij}(i, j = 1, 2, \cdots)$. Various cases arise. From (2.5), (2.8), (2.9) and the fact that $\mathbf{R}$ commutes with $\mathbf{B}$, $\mathbf{L}$ and $\mathbf{U}$, we have, omitting details:

$$\begin{aligned}
&\text{(i) } i = j \neq 1: && \mathbf{E}_{i,i} = \sigma^2\mathbf{R}[2(1-\omega)\mathbf{I} + \omega^2(\mathbf{I} + \mathbf{B} + \mathbf{B}^2)]\\
&\text{(ii) } i = j = 1: && \mathbf{E}_{1,1} = \sigma^2\mathbf{R}[\mathbf{I} + \omega\mathbf{B} + \omega^2\mathbf{L}\mathbf{U}]\\
&\text{(iii) } 2 \leqq j = i+1 \leqq k: && \mathbf{E}_{i,i+1} = \sigma^2\mathbf{R}[\omega^2\mathbf{L} - (1-\omega)\mathbf{I}]\\
&\text{(iv) } 2 \leqq i = j+1 \leqq k: && \mathbf{E}_{i+1,i} = \sigma^2\mathbf{R}[\omega^2\mathbf{U} - (1-\omega)\mathbf{I}]\\
&\text{(v) } |i-j| > 2: && \mathbf{E}_{i,j} = \mathbf{0}
\end{aligned} \tag{3.19}$$

Let $\{\mu_i\}(i = 1, \cdots, n)$ be the real eigenvalues of $\mathbf{B}$, so that

$$\lambda_i = 2(1-\omega) + \omega^2\mu_i^2 \qquad (i = 1, \cdots, n)$$

where $\lambda_i(i = 1, \cdots, n)$ are defined in the statement of Lemma 2. Now, if $p(\mathbf{B})$ is any polynomial in $\mathbf{B}$, then the eigenvalues of $p(\mathbf{B})$ are $p(\mu_i)(i = 1, \cdots, n)$ and the corresponding eigenvectors are precisely the n linearly independent, orthonormal (say) eigenvectors $\mathbf{x}_i(i = 1, \cdots, n)$ of $\mathbf{B}$. Then from (3.17) and (3.19), and using the spectral decomposition of $\mathbf{B}$ and $\mathbf{A}$,

$$\begin{aligned}
\mathbf{V}_k = \sigma^2\mathbf{R}\Big\{&[2(1-\omega)\mathbf{I} + \omega^2(\mathbf{I} + \mathbf{B} + \mathbf{B}^2)]\cdot \sum_{i=2}^{k} p_{k,i}^2(\mathbf{A})\\
&+ [\mathbf{I} + \omega\mathbf{B} + \omega^2\mathbf{L}\mathbf{U}]p_{k,1}^2(\mathbf{A})\\
&+ [\omega^2\mathbf{B} - 2(1-\omega)\mathbf{I}]\cdot \sum_{i=1}^{k-1} p_{k,i}(\mathbf{A})\, p_{k,i-1}(\mathbf{A})\Big\}\\
= \sigma^2\mathbf{R}\cdot \sum_{j=1}^{n}\Big\{&[2(1-\omega) + \omega^2(1 + \mu_j + \mu_j^2)]\cdot \sum_{i=2}^{k} p_{k,i}^2(\lambda_j)\\
&+ [\mathbf{I} + \omega\mathbf{B} + \omega^2\mathbf{L}\mathbf{U}]p_{k,1}^2(\lambda_j)\\
&+ [\omega^2\mu_j - 2(1-\omega)]\cdot \sum_{i=1}^{k-1} p_{k,i}(\lambda_j)p_{k,i-1}(\lambda_j)\Big\}\mathbf{x}_j\,\mathbf{x}_j^*
\end{aligned} \tag{3.20}$$

where $p_{kj}(\varphi)$, for any scalar, φ, denotes that polynomial in φ which has the same coefficients as $p_{k,j}(\mathbf{A})$.

In order to examine the behaviour of (3.20) as $k \to \infty$, we must consider the $p_{k,j}$ in more detail. Substituting (3.15) into (3.12), (3.13) and (3.14), and 'comparing coefficients' of the $\{\mathbf{f}^{(k)}\}$ (which is permissible since the resulting expression is to be true for *all* $\{\mathbf{e}^{(k)}\}$ and hence *all* $\{\mathbf{f}^{(k)}\}$), we have the recurrence relationships:

$$p_{k,k}(\mathbf{A}) = \mathbf{I}$$

$$p_{k,k-1}(\mathbf{A}) = \mathbf{A} \tag{3.21}$$

$$p_{k,i}(\mathbf{A}) - \mathbf{A}p_{k-1,i}(\mathbf{A}) + \propto p_{k-2,i}(\mathbf{A}) = \mathbf{0} \qquad (i = 1, 2, \cdots, k-2)$$

for $k = 1, 2, \cdots$. Similarly, $p_{k,j}(\varphi)$ satisfies the recurrence relationships

$$p_{k,k}(\varphi) = 1$$

$$p_{k,k-1}(\varphi) = \varphi \tag{3.22}$$

$$p_{k,i}(\varphi) - \varphi p_{k-1,i}(\varphi) + \alpha\, p_{k-2,i}(\varphi) = 0 \qquad (i = 1, 2, \cdots, k-2)$$

for $k = 1, 2, \cdots$. Now consider the difference equation for $p_i(\varphi)$

$$p_{i+2}(\varphi) - \varphi p_{i+1}(\varphi) + \alpha\, p_i(\varphi) = 0 \qquad (i = 0, 1, 2, \cdots) \tag{3.23}$$

$$p_0(\varphi) = 1; \qquad p_1(\varphi) = \varphi$$

Then from the familiar theory of difference equations with constant coefficients, we know that

$$p_i(\varphi) = \begin{cases} (t_1^{i+1} - t_2^{i+1})/(t_1 - t_2), & \text{if } t_1 \neq t_2 \\ (i+1)t_1^i, & \text{it } t_1 = t_2 \end{cases} \tag{3.24}$$

where t_1, t_2 are the roots of

$$t^2 - \varphi t + \alpha = 0 \tag{3.25}$$

From (3.23) we see that $p_i(\varphi)$ is a polynomial in φ. It is easy to verify from (3.22), and noting that both the $p_i(\varphi)$ and the $p_{k,j}(\varphi)$ are uniquely defined, that

$$p_{k,j}(\varphi) = p_{k-j}(\varphi) \qquad (j = 1, \cdots, k) \tag{3.26}$$

Hence

$$p_{k,j}(\mathbf{A}) = p_{k-j}(\mathbf{A}) \tag{3.27}$$

where $p_i(\mathbf{A})$ has the obvious meaning. It follows that

$$\sum_{i=2}^{k} p_{k,i}^2(\varphi) = \sum_{i=2}^{k} p_{k-i}^2(\varphi) = \sum_{i=0}^{k-2} p_i^{\,2}(\varphi) \tag{3.28}$$

$$\sum_{i=2}^{k-1} p_{k,i}(\varphi)p_{k,i+1}(\varphi) = \sum_{i=2}^{k-1} p_{k-i}(\varphi)p_{k-i-1}(\varphi) = \sum_{i=1}^{k-2} p_i(\varphi)p_{i-1}(\varphi) \tag{3.29}$$

Furthermore, from (3.24), if $|t_s| < 1 (s = 1, 2)$, then $p_i(\varphi) \to 0$ as $i \to \infty$, and hence from Lemma 2,

$$p_{k,1}(\lambda_j) = p_{k-1}(\lambda_j) \to 0 \quad \text{as} \quad k \to \infty \qquad j = 1, \cdots, n. \tag{3.30}$$

We now need:

LEMMA 4. *Suppose, in* (3.25), *that* $|t_s| < 1 (s = 1, 2)$, *and let*

$$M_k(\varphi) = \sum_{i=0}^{k} p_i^{\,2}(\varphi); \qquad N_k(\varphi) = \sum_{i=1}^{k} p_i(\varphi)p_{i-1}(\varphi). \tag{3.31}$$

Then

$$M(\varphi) = \lim_{k\to\infty} M_k(\varphi), \quad \text{and} \quad N(\varphi) = \lim_{k\to\infty} N_k(\varphi) \tag{3.32}$$

both exist, and in fact,

$$M(\varphi) = [(1+\alpha)/(1-\alpha)]/[(1+\alpha)^2 - \varphi^2]$$
$$N(\varphi) = [\varphi/(1-\alpha)]/[(1+\alpha)^2 - \varphi^2] \tag{3.33}$$

Proof. This is a simple computation using (3.24), the fact $t_1 + t_2 = \varphi$, $t_1 t_2 = \alpha$ and that $|t_s| < 1 (s = 1, 2)$.

Finally, then, applying Lemma 2 in Lemma 4, and using this in (3.28), (3.29) and (3.30), we have, on letting $k \to \infty$ in (3.20):

$$\begin{aligned}\mathbf{V} &= \lim_{k\to\infty} \mathbf{V}_k \\ &= \sigma^2\mathbf{R}\sum_{j=1}^{n} \{[2(1-\omega) + \omega^2(1+\mu_j+\mu_j^2)]\cdot[1+(1-\omega)^2]/[1-(1-\omega)^2] \\ &\quad + \lambda_j[\omega^2\mu_j - 2(1-\omega)]/[1-(1-\omega)^2]\} \frac{\mathbf{x}_j\mathbf{x}_j^*}{[1+(1-\omega)^2] - \lambda_j^2} \\ &= \sigma^2\mathbf{R}\sum_{j=1}^{n} Q_j(\omega)\mathbf{x}_j\mathbf{x}_j^*, \end{aligned} \tag{3.34}$$

where, since $\lambda_j = 2(1-\omega) + \omega^2\mu_j^2$, and simplifying:

$$\begin{aligned} Q_j(\omega) &= \frac{[\omega^4\mu_j^3 + \omega^4\mu_j^2 + \omega^2(2-\omega)^2\mu_j + \omega^2(2-\omega)^2]}{\omega^3[(2-\omega)^2 + \omega^2\mu_j^2](1-\mu_j^2)} \\ &= 1/[\omega(2-\omega)(1-\mu_j)]. \end{aligned} \tag{3.35}$$

Thus

$$\mathbf{V} = \sigma^2\mathbf{R}/[\omega(2-\omega)]\cdot\sum_{j=1}^{n}\mathbf{x}_j\mathbf{x}_j^*/(1-\mu_j) = \sigma^2/[\omega(2-\omega)]\cdot\mathbf{R}\mathbf{C}^{-1},$$

which completes the proof of Theorem 2.

Apart from the many objections that could be lodged from a purely statistical standpoint, it would appear from Theorem 2 that it could be dangerous to over-relax excessively, that is, to employ values of ω close to 2. It is well-known, however Young [16], that, for certain problems, the optimal value of $\omega = \omega_b$, considered purely from the standpoint of the asymptotic rate of convergence, tends to 2 as $n \to \infty$. Furthermore, for ill-conditioned problems in general, where $\mu = \max_i |\mu_i|$ is very close to unity, then ω_b is very close to 2. The Young theory of successive over-relaxation also indicates that it is preferable to over-estimate ω_b rather than to underestimate it. It is in the context of these results that perhaps Theorem 2 may have some significance. Much depends upon the size of $\sigma^2\mathbf{R}$, however, about which it is difficult to make a priori assumptions, and which will, in general, be a function of ω. The presence of $\mathbf{C}^{-1}$ in (2.13) and (3.4) is merely confirmation of the well-known fact that the round-off error depends very definitely upon the conditioning of $\mathbf{C}$ [12, 15, 2] even in an iterative method.

Furthermore, as was stated in the introduction, the central limit theorem is not applicable to the distribution of $\mathbf{r}^{(k)}$ as $k \to \infty$, since it may easily be shown that $\mathbf{r}^{(k)}$ remains uniformly bounded for all k. However, using Lemma 3 of Golub [5] in

precisely the same manner as is applied there to the Richardson second-order method, we may use Theorems 1 and 2 to obtain probabilistic bounds:

THEOREM 3. *Under the same hypotheses as in Theorem 2, then for all* $0 \leqq \beta \leqq 1$,

$$P\{\| \mathbf{r}^{(\infty)} - \mathbf{r} \|^2 \leqq [\sigma^2 \operatorname{tr}(\mathbf{R}\mathbf{C}^{-1})]/[\omega(2-\omega)\beta]\} \geqq 1 - \beta, \tag{3.36}$$

where P *denotes the probability function,* $\operatorname{tr}(\cdots)$ *the trace of a matrix,* $\| \cdots \|$ *the Euclidean norm, and*

$$\mathbf{r}^{(\infty)} = \lim_{k\to\infty} \mathbf{r}^{(k)}, \qquad \mathbf{r} = \lim_{k\to\infty} \mathbf{r}_k = \frac{1}{\omega}\mathbf{C}^{-1}\boldsymbol{\varepsilon}. \tag{3.37}$$

We remark that, if $\mathbf{R} = \mathbf{I}$, then we may bound $\operatorname{tr}(\mathbf{C}^{-1})$ in (3.36) by $n/(1-\mu)$, where $\mu = \max_i |\mu_i|$.

4. Numerical Experiment. As a simple numerical example, we consider the one-dimensional Dirichlet problem

$$\begin{aligned} y'' &= f(x) \\ y(0) = \beta_0\,; \qquad y(1) &= \beta_1 \end{aligned} \tag{4.1}$$

The system of linear equations which arises when this problem is discretised has the form:

$$\mathbf{C}\mathbf{y} = -h^2\mathbf{f} + \mathbf{g}, \tag{4.2}$$

where $\mathbf{C}$ is the $n \times n$ matrix

$$\mathbf{C} = \boldsymbol{\pi}\begin{bmatrix} 2 & -1 & 0 & \cdot & \cdot & 0 \\ -1 & 2 & -1 & \cdot & \cdot & \cdot \\ 0 & -1 & 2 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & -1 & 2 \end{bmatrix}\boldsymbol{\pi}^* \tag{4.3}$$

for some permutation matrix, $\boldsymbol{\pi}$, and

$$\begin{aligned} \mathbf{f} &= \boldsymbol{\pi}[f_1, \cdots, f_n]^*, \\ \mathbf{g} &= \boldsymbol{\pi}[\beta_0, 0, \cdots, 0, \beta_1]^*, \end{aligned} \tag{4.4}$$

where $f_i = f(x_i)$ and $x_i = ih$, $(n+1)h = 1$. In order to apply our theory, $\boldsymbol{\pi}$ is such that $\mathbf{C}$ has the σ_1-ordering. Now it can be shown (Marcus [8]) that if $\mathbf{C}^{-1} = (\gamma_{ij})$, then

$$\gamma_{ij} = \begin{cases} i(n+1-j)/(n+1), & \text{if } j \geqq i \\ j(n+1-i)/(n+1), & \text{if } j \leqq i \end{cases} \tag{4.5}$$

and hence, letting $\gamma_i = \sum_{j=1}^{n} \gamma_{ij}$,

$$\gamma_i = i(n+1-i)/2. \tag{4.6}$$

Suppose in Theorem 1 that

$$\boldsymbol{\varepsilon} = \boldsymbol{\pi}[\epsilon_1, \epsilon_2, \cdots, \epsilon_n]^*, \qquad \mathbf{r} = \boldsymbol{\pi}[\rho_1, \rho_2, \cdots, \rho_n] \tag{4.7}$$

Then, for some permutation $\pi = \pi(i)$ of $\{1, \cdots, n\}$,

$$\rho_{\pi(i)} = \left\{(n+1-i)\sum_{j<i} j\epsilon_j + i\sum_{j\geqq i}(n+1-j)\epsilon_j\right\} \Big/ [\omega(n+1)] \tag{4.8}$$

If we assume that all the ϵ_j's are constant, $\epsilon_j \equiv \epsilon$ (say), then

$$\rho_{\pi(i)} = \gamma_i\epsilon/\omega = i(n+1-i)\epsilon/2\omega \tag{4.9}$$

Similarly for the variance, denoting, in Theorem 2, the diagonal elements of $\mathbf{V}$ by v_i and assuming that $\mathbf{R} = \mathbf{I}$,

$$v_{\pi(i)} = \sigma^2 i(n+1-i)/[\omega(2-\omega)(n+1)] \tag{4.10}$$

In actual numerical computations, it is obviously difficult to make *a priori* assumptions about the size of ϵ and σ^2, apart from justifying all the sundry assumptions that led to (5.9) and (5.10). In an attempt to facilitate these problems, we computed the solution to

$$\begin{aligned} y_m''(x) &= k_1(e^x + e^{1-x}) \\ y_m(0) &= y_m(1) = k_2 + m\Delta \end{aligned} \tag{4.11}$$

for certain constants k_1, k_2 and where Δ is a small increment. k_1 and k_2 were chosen ($k_1 = \frac{3}{16}$, $k_2 = \frac{5}{8}$) such that the analytical solution of (4.11) lay between $\frac{1}{2}$ and 1 for $m \leqq 100$ and $\Delta = 10^{-4}$, and so that $y_m''(x) - y_m(x)$ was of small order of magnitude in comparison to $\frac{1}{2}$. The first of these requirements was to meet the demand that the ϵ_j's be constant, and so that our floating-point arithmetic became effectively fixed-point, and the second of these requirements attempted to ensure that (2.5) was fulfilled, since then our initial guess, namely of setting y_i equal to f_i, was close to the exact solution and would not change by much during the computation.

The successive over-relaxation procedure was carried out in both single and double-precision arithmetic, the difference between the two being considered as the accumulated round-off error. The mean and variance of the latter were calculated, using $m = 1, \cdots, 100$ as a sample (see Henrici [7] for a full discussion of this kind of experiment). One-hundred iterations were performed; it was surprisingly found that both the means and the variances converged, and that $k = 100$

TABLE 1
Dependence of Expected Values upon n

ω	1.2			1.5			1.8		
n	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
10	61	71	67	64	71	56	65	71	57
20	247	257	212	253	257	183	251	257	246
30	564	560	517	576	560	432	572	560	478
40	1011	980	917	1064	980	820	1079	980	1053
50	1497	1517	1694	1504	1517	1596	1513	1517	1581

(a) Experimental values of $10^{10}\ \rho_{\max}$, fixed-point.
(b) Theoretical values of $10^{10}\ \rho_{\max}$, fixed-point.
(c) Experimental values of $10^{9}\ \rho_{\max}$, floating-point.

TABLE 2
Dependence of Variances upon n

ω	1.2			1.5			1.8		
n	(d)	(e)	(f)	(d)	(e)	(f)	(d)	(e)	(f)
10	131	76	117	137	121	144	462	320	381
20	247	144	206	342	230	315	1038	611	735
30	372	213	317	398	340	498	1316	903	1076
40	516	282	346	621	449	881	1521	1194	1298
50	724	350	590	740	559	1113	1976	1486	1727

(d) Experimental values of $10^{20}\ V_{\max}$, fixed-point.
(e) Theoretical values of $10^{20}\ V_{\max}$, fixed-point.
(f) Experimental values of $10^{18}\ V_{\max}$, floating-point.

TABLE 3
Dependence of Expected Values and Variances upon ω

ω	(a)	(b)	(c)	(d)	(e)	(f)
1.1	239		161	267	129	167
1.2	247		212	247	144	206
1.3	252		229	258	172	218
1.4	252		252	300	191	229
1.5	253		183	342	230	315
1.6	253		263	491	291	358
1.7	254		263	535	383	505
1.8	251		246	1038	611	735
1.82	274	257	217	1188	666	921
1.84	247		225	1201	752	1032
1.86	254		294	1441	844	1409
1.88	258		197	1539	988	1906
1.90	255		219	2005	1216	2124
1.92	254		240	2392	1525	2706
1.94	254		261	3335	2041	3120
1.96	274		270	4621	3073	4914
1.98	261		221	8742	6158	7091

was sufficient to ensure this convergence, which is required by the asymptotic nature of theorems 1 and 2. The procedure was carried out twice, once in fixed-point† and once in floating-point arithmetic, in both cases symmetric rounding being employed. An analysis of the arithmetic operations involved, using in particular Theorem 1.9 and equations (1-107) and (1-108) of Henrici [7], leads to the assumption that for the fixed-point computation

$$\epsilon \cong \frac{\omega u}{2}\ ; \qquad \sigma^2 \cong \frac{9\omega^2 + 16}{8}\frac{u^2}{12} \tag{4.12}$$

for $1 < \omega < 2$, where u is the basic machine unit. The fact that ϵ is not zero in

† The author would like to acknowledge the assistance of Mr. Peter Golitzen in this connection.

spite of symmetric rounding being employed stems from the particular equation solved, where we have to form products of the form $\frac{1}{2}y_i$ where $\frac{1}{2} < y_i \leqq 1$, and the only possible values of the round-off error in forming this product is either 0 or $u/2$. This is also responsible for the peculiar expression for σ^2.

Since the floating-point computations performed were in fact pseudo fixed-point, we also assume that (4.12) holds in this case, too. In the fixed-point calculation, $u = 2^{-30}$, whilst for the floating-point calculation, $u = 2^{-27}$. Thus from (4.9), (4.10) and (4.12), we expect that

$$\begin{aligned} \rho_{\pi(i)} &\cong \tfrac{1}{2}i(n + 1 - i)u/2 \\ v_{\pi(i)} &\cong [i(n + 1 - i)(9\omega^2 + 16)u^2]/[96(n + 1)\omega(2 - \omega)] \end{aligned} \tag{4.13}$$

so that if we let $\rho_{\max} = \max_i |\rho_i|$, $V_{\max} = \max_i |V_i|$,

$$\begin{aligned} \rho_{\max} &\cong (n + 1)^2u/16 \\ V_{\max} &\cong [(n + 1)(9\omega^2 + 16)u^2]/[384\omega(2 - \omega)] \end{aligned} \tag{4.14}$$

Or computational results are divided into two parts, one to show dependence upon n, and the other to show dependence upon ω. In Tables 1 and 2 we show dependence upon n for $\omega = 1.2, 1.5; 1.8$. In Table 3 the dependence upon ω for $n = 20$ is shown; in this case ω_b, the optimal value of ω, is approximately 1.74. We give theoretical values only for the fixed-point computations. It will be seen that the experimentally calculated expected values correlate quite closely with the predicted theory, whereas the experimental values of the variances seem to be slightly higher than the theoretical values. This latter situation is most probably due to the non-independence of the local round-off errors after a large number of iterations. Similar experiences were obtained in the numerical experiments of Golub and Moore [6].

The author would like to extend his appreciation to Professor P. K. Henrici for his guidance in the preparation of the dissertation upon which this paper is principally based. He would also like to acknowledge the constructive criticisms of Dr. J. H. Wilkinson.

National Physical Laboratory
Teddington, Middlesex, England

1. A. A. Abramov, "On the influence of round-off errors in the solution of Laplace's equation," *Vycis. Matem. i Vycis. Teh.*, v. 1, 1953, p. 37–41.
2. J. Descloux, "Note on the round-off error in iterative processes," *Math. Comp.*, v. 17, 1963, p. 18–27.
3. G. E. Forsythe, "Note on rounding-off errors," *SIAM Rev.*, v. 1, 1959, p. 66–67.
4. G. H. Golub, "The use of Chebyshev matrix polynomials in the iterative solution of linear equations compared to the method of successive over-relaxation," Doctoral Thesis, University of Illinois, 1959.
5. G. H. Golub, *Bounds for the Round-Off Errors in the Richardson Second-Order Method*, Nordisk Tidskrift for Informations-Behandlung, v. 2, 1962, p. 212–223.
6. G. H. Golub, & J. K. Moore, *ibid* (appendix).
7. P. K. Henrici, *Discrete-variable Methods in Ordinary Differential Equations*, John Wiley & Sons, Inc., New York, 1961.
8. M. Marcus, "Basic theorems in matrix theory," *N.B.S. Appl. Math. Ser.* No. 57, 1960.
9. A. Ostrowski, "On the linear iteration procedures for symmetric matrices," *Rendic. di Mat. e.d.s. Applicaz*, v. 13, 1954, p. 1–24.
10. W. Sibagaki, "On the idea of 'numerical convergence' and its applications," *Mem. Fac. Sci. Kyushu Univ. Ser. A*, v. 5, 1950, p. 89–97.

11. A. M. TURING, "Rounding errors in algebraic processes," *Quart. J. Appl. Math.*, v. 1, 1948, p. 287–307.
12. M. URABE, "Convergence of numerical iteration in solution of equations," *J. Sci. Hiroshima Univ. Ser. A*, v. 19, 1956, p. 479–489.
13. R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, New Jersey, 1962.
14. J. H. WILKINSON, "Rounding errors in algebraic processes," *Proc. Int. Conference on Information Processing*, UNESCO, 1959, p. 44.
15. J. H. WILKINSON, Error analysis of direct methods of matrix inversion," *J. Assoc. Comp. Mach.*, v. 8, 1961, p. 281–330.
16. D. M. YOUNG, "Iterative methods for solving-partial differential equations of the elliptic type," *Trans. Amer. Math. Soc.*, v. 76, 1954, p. 92–111.