Computing Error Bounds in Solving Linear Systems

By J. Schröder

1. Introduction. Let there be given a system of m linear algebraic equations for m unknowns

$$Gu = r \tag{1.1}$$

and consider an iteration procedure

$$u_{n+1} = Mu_n + s \qquad (n = 0, 1, 2, \cdots) \tag{1.2}$$

such that the system

$$u = Mu + s \tag{1.3}$$

is equivalent to (1.1). All elements of the m-dimensional vectors and the $m \times m$-matrices which occur are assumed to be real.

In the space of m-dimensional vectors $u = (u^i)$, v, $\cdots$ we define an order relation and an absolute value by writing

$$u \leqq v \quad \text{if and only if} \quad u^i \leqq v^i \qquad (i = 1, 2, \cdots, m),$$

and

$$| u | = (| u^i |).$$

Similarly, for $m \times m$-matrices $A = (a_{ij})$, B, $\cdots$ we use the notation

$$A \leqq B \quad \text{if and only if} \quad a_{ij} \leqq b_{ij} \quad (i, j = 1, 2, \cdots, m),$$

and

$$| A | = (| a_{ij} |).$$

Let B denote a matrix such that

$$| M | \leqq B$$

for M in equation (1.2) (use, for example, $B = | M |$ if M is explicitly known). If there exists a vector v_0 such that

$$| u_p - u_{p+1} | \leqq v_0 - Bv_0 \quad \text{for some index } p, \tag{1.4}$$

then the given equation has a solution u^* for which

$$| u^* - u_{p+n} | \leqq B^n v_0 \qquad (n = 0, 1, 2, \cdots) \tag{1.5}$$

holds (Theorem 1).

We present a method for computing a vector v_0 with the property (1.4). The calculation of v_0 and the vectors $B^n v_0$ in (1.5) constitutes an iterative procedure parallel to approximation procedure (1.2). This estimation procedure, described

Received August 2, 1961. This work was sponsored by the Mathematics Research Center, U.S. Army, Madison, Wisconsin.

at the end of Section 2, can be programmed for computers as easily as approximation procedure (1.2).

If the matrix B is irreducible and noncyclic and if the spectral radius $\rho(B)$ of the matrix B satisfies

$$(1.6) \qquad \rho(B) < 1,$$

then for some p the method yields a vector v_0 which satisfies (1.4) (Theorem 3).

Section 4 is concerned with the methods of simultaneous and successive displacements (with B chosen as in (4.4)). Some numerical examples (up to 15 unknowns) are given. In these examples, the estimation procedure yields a suitable vector v_0, i.e., an error bound, after at most three steps, thus the time needed for estimation is considerably smaller than the time needed for solving the given system by iteration (one need not start the estimation procedure and the approximation procedure simultaneously).

For many of the known estimation methods one has to calculate an upper bound σ of $\rho(B)$ and this bound has to satisfy the condition

$$(1.7) \qquad \sigma < 1.$$

For example, this inequality (1.7) may be the row-sum criterion for the method of successive displacements. In this paper condition (1.7) is weakened to (1.6) where $\rho(B)$ need not be known.

Of course, condition (1.6) still restricts the class of iteration procedures (1.2) to which the estimation method can be applied. Note that the procedure (1.2) converges for an arbitrary vector u_0 if and only if

$$\rho(M) < 1$$

where the spectral radius $\rho(M)$ of M satisfies

$$\rho(M) \leqq \rho(|M|) \leqq \rho(B).$$

However, every convergence condition which works with upper bounds b_{ij} of the moduli $|m_{ij}|$ instead of with elements m_{ij} cannot be weaker than condition (1.6).

For example, when solving difference equations for the Laplace equation by the method of successive displacements, one has, in general, $M \geqq O$ and $\rho(M) < 1$, thus, $\rho(B) < 1$ for $B = M$. In the case of the biharmonic equation, however, the method of successive displacements in general yields $\rho(|M|) > 1$.

2. Derivation of the Method. Let R be the set of m-dimensional real vectors $u = (u^i)$, v, $\cdots$, and let $A = (a_{ij})$, B, M, $\cdots$ denote real $m \times m$-matrices. The notation $u \leqq v$, $|u|$, $A \leqq B$, and $|A|$ shall be defined as in the introduction.

Consider an equation

$$(2.1) \qquad u = Mu + s$$

where M denotes a given matrix, s is a given vector and u is unknown. With $Tu = Mu + s$ we can write this equation as $u = Tu$.

Let B denote a fixed matrix such that

$$(2.2) \qquad |M| \leqq B,$$

then let $H[u, v]$ be the function

$$H[u, v] = \tfrac{1}{2}(B + M)\, u - \tfrac{1}{2}(B - M)\, v + s.$$

This function is increasing in u and decreasing in v, and for $u = v$ we get $H[u, u] = Tu$.

We consider the iteration procedure

$$x_{n+1} = H[x_n, y_n], \qquad y_{n+1} = H[y_n, x_n] \qquad (n = 0, 1, 2, \cdots). \tag{2.3}$$

Because of the described properties of $H[u, v]$, for this procedure the following statements hold (see [5]).

Let the conditions

$$x_0 \leqq x_1, \qquad x_0 \leqq y_0, \qquad y_1 \leqq y_0$$

be satisfied.

Then, the vectors x_n and y_n $(n = 0, 1, 2, \cdots)$ defined by (2.3) *satisfy the inequalities*

$$x_0 \leqq x_1 \leqq x_2 \leqq \cdots \leqq x_n \leqq y_n \leqq \cdots \leqq y_2 \leqq y_1 \leqq y_0. \tag{2.4}$$

Moreover, the sequences $\{x_n\}$ and $\{y_n\}$ converge to limit-vectors x^* and y^*, respectively, such that

$$x^* = H[x^*, y^*], \qquad y^* = H[y^*, x^*] \tag{2.5}$$

and

$$x_n \leqq x^* \leqq y^* \leqq y_n \qquad (n = 0, 1, 2, \cdots). \tag{2.6}$$

The inequalities (2.4) can be proved by induction, and the convergence of the sequences $\{x_n\}$ and $\{y_n\}$ then follows from the fact that these sequences are monotonic and bounded.

Adding the two equations in (2.5) and noting that (2.6) holds, we get, in addition, the following statement:

The vector

$$u^* = \tfrac{1}{2}(x^* + y^*)$$

is a solution of the given equation (2.1), *and for this solution the estimate*

$$x_n \leqq u^* \leqq y_n \qquad (n = 0, 1, 2, \cdots)$$

holds.

Now let $\{u_n\}$ and $\{v_n\}$ be sequences of vectors which satisfy the equations

$$u_{n+1} = Mu_n + s \qquad (n = 0, 1, 2, \cdots) \tag{2.7}$$

and

$$v_{n+1} = Bv_n \qquad (n = 0, 1, 2, \cdots), \tag{2.8}$$

respectively. Then, the vectors x_n and y_n defined by

$$u_n = \tfrac{1}{2}(x_n + y_n), \qquad v_n = \tfrac{1}{2}(y_n - x_n) \qquad (n = 0, 1, 2, \cdots),$$

i.e.,

$$x_n = u_n - v_n, \qquad y_n = u_n + v_n \qquad (n = 0, 1, 2, \cdots),$$

satisfy (2.3). We reformulate the main results stated above in terms of the vectors u_n and v_n instead of x_n and y_n.

THEOREM 1. *Let the conditions*

$$v_0 \geqq o \quad \text{and} \quad | u_0 - u_1 | \leqq v_0 - v_1 \tag{2.9}$$

be satisfied.

Then, the vectors u_n *and* v_n $(n = 0, 1, 2, \cdots)$ *defined by* (2.7) *and* (2.8) *satisfy the inequalities*

$$v_n \geqq o, \qquad | u_n - u_{n+1} | \leqq v_n - v_{n+1} \qquad (n = 0, 1, 2, \cdots). \tag{2.10}$$

Moreover, the sequences $\{u_n\}$ *and* $\{v_n\}$ *converge to vectors* u^* *and* v^*, *respectively, such that*

$$u^* = Mu^* + s \quad \text{and} \quad v^* = Bv^*,$$

and for u^* *the error estimate*

$$| u^* - u_n | \leqq v_n \qquad (n = 0, 1, 2, \cdots) \tag{2.11}$$

holds.

From Theorem 1 we start to develop a method of error estimation for the iteration procedure (2.7). Clearly, we can replace the vectors u_i in the theorem, which occur by u_{i+p} (p denoting a fixed non-negative integer). The assumption (2.9) then takes the form

$$v_0 \geqq o, \quad | u_p - u_{p+1} | \leqq v_0 - v_1 \tag{2.12}$$

and the estimate (2.11) becomes

$$| u^* - u_{n+p} | \leqq v_n \qquad (n = 0, 1, 2, \cdots). \tag{2.13}$$

When the vectors u_p and u_{p+1} are known one could try to construct a suitable vector v_0 using the special properties of the given problem. However, the error estimate would then, in general, be much more complicated than the calculation of the approximations u_n because one can easily program the procedure (2.7) for computers. Therefore, we will establish a method of error estimation which also can be programmed quite easily.

A Method of Error Estimation for the Iteration Procedure $u_{n+1} = Mu_n + s$. Starting with some vector

$$w_0 \geqq o \quad (\text{for example } w_0 = o)$$

and using the formula

$$w_{n+1} = Bw_n + \delta_{n+1} \quad \text{with} \quad \delta_{n+1} = | u_{n+1} - u_n | \tag{2.14}$$

calculate vectors w_n up to the first index $n = p$ for which

$$w_p \geqq w_{p+1}, \tag{2.15}$$

provided that such an index exists. Calculate in addition vectors z_n ($n = p, p + 1, \cdots$) defined by

$$z_p = w_p, \qquad z_{n+1} = Bz_n \qquad (n = p, p + 1, \cdots).$$

THEOREM 2. *If there exists an index p such that* (2.15) *holds, then the sequence* $\{u_n\}$ *converges to a solution* $u^* = Mu^* + s$ *and*

$$(2.16) \qquad | u^* - u_n | \leqq z_n \qquad (n = p, p + 1, \cdots).$$

Proof. Suppose p is an index such that (2.15) holds. Then, let $\{v_n\}$ denote the sequence defined by (2.8) for $v_0 = w_p$. Clearly, one has $v_0 \geqq o$. Moreover, also the second inequality in (2.12) is satisfied because this inequality is equivalent to the condition (2.15). Thus the sequence $\{u_n\}$ converges to a solution u^* which satisfies the inequality (2.13), and this inequality is equivalent to (2.16).

3. Theoretical Investigation of the Method. We now investigate the conditions under which the method of error estimation described above will be successful.

For this, we assume that the matrix B is irreducible (non-decomposable) and noncyclic. Then, also because $B \geqq O$, according to Frobenius [3] the following statements hold.

Matrix B has an eigenvalue $\lambda > 0$, called the maximal root of B, such that λ is greater than the modulus of each other eigenvalue of B. Corresponding to λ there exists an eigenvector $\varphi = (\varphi^i)$ with

$$(3.1) \qquad \varphi^i > 0 \qquad (i = 1, 2, \cdots, m),$$

and there are no eigenvectors or principal vectors (generalized eigenvectors) corresponding to λ which are linearly independent of φ.

Thus, each vector $u \in R$ can be written as a sum

$$(3.2) \qquad u = u^{(1)}\varphi + \psi$$

where $u^{(1)}$ is a constant and $\psi = \psi(u)$ is a linear combination of the eigenvectors and principal vectors of B belonging to eigenvalues different from λ.

Then let P_1 and P_2 denote the two projection matrices defined by

$$P_1 u = u^{(1)}\varphi \qquad (\text{for } u \in R), \qquad P_2 = I - P_1,$$

and moreover let

$$B_1 = BP_1, \qquad B_2 = BP_2.$$

Then the equation

$$B = B_1 + B_2$$

represents a spectral decomposition of matrix B with B_1 belonging to λ and B_2 belonging to the remainder of the spectrum of B. We have

$$(3.3) \qquad B^n u = B_1^n u + B_2^n u = \lambda^n u^{(1)}\varphi + B_2^n u.$$

Let $\rho(A)$ denote the spectral radius of a matrix A, i.e., the maximum of the moduli of its eigenvalues. Then, we have $\rho(B_1) = \rho(B) = \lambda$ and

$$(3.4) \qquad \rho(B_2) < \lambda.$$

If u is a vector such that $u \geqq o$, then $B^n u \geqq o$ $(n = 0, 1, 2, \cdots)$ and it follows from (3.3) that

$$(3.5) \qquad u^{(1)}\varphi + (\lambda^{-1}B_2)^n u \geqq o.$$

Because of (3.4) the inequality (3.5) yields $u^{(1)}\varphi \geqq o$ for $n \to \infty$, hence $u^{(1)} \geqq 0$. This proves the statement that

$$(3.6) \qquad u \geqq o \quad \textit{implies} \quad u^{(1)} \geqq 0.$$

If $u^i > 0$ for all $i = 1, 2, \cdots, n$, then there exists a constant $\alpha > 0$ such that $u \geqq \alpha\varphi$ and we now derive from (3.3)

$$u^{(1)}\varphi + (\lambda^{-1}B_2)^n u = \lambda^{-n}B^n u \geqq \lambda^{-n}\alpha B^n\varphi = \alpha\varphi.$$

Hence, for $n \to \infty$ we obtain the result

$$(3.7) \qquad u^i > 0 \ (\textit{for all } i = 1, 2, \cdots, n) \ \textit{implies } u^{(1)} > 0.$$

We write

$$u > o \quad \text{if and only if} \quad u \geqq o \text{ and } u \neq o,$$

and we now also prove that

$$(3.8) \qquad u > o \quad \textit{implies} \quad u^{(1)} > 0.$$

Let $\mu > \lambda$, then the equation

$$\sum_{i=0}^{\infty} \mu^{-i}B^i = (I - \mu^{-1}B)^{-1}$$

holds, and the matrix $(I - \mu^{-1}B)^{-1}$ has all elements positive because B is irreducible. (This follows from Theorem 2.2 in [7], applied to $A = \mu I$, $z = \varphi$ and B as defined above.) Therefore, if $u > o$, all components of the vector

$$v = (I - \mu^{-1}B)^{-1}u = \sum_{i=0}^{\infty} \mu^{-i}B^i u$$

are positive, hence it follows from (3.7) that $v^{(1)} > 0$. Finally, we conclude from the equation

$$v^{(1)}\varphi = P_1 v = P_1\sum_{i=0}^{\infty} \mu^{-i}B^i u = \sum_{i=0}^{\infty} \mu^{-i}\lambda^i u^{(1)}\varphi = u^{(1)}(1 - \lambda\mu^{-1})^{-1}\varphi$$

and that $u^{(1)} > 0$.

THEOREM 3. *Suppose that matrix B is irreducible and noncyclic, and assume moreover that the maximal root $\lambda = \rho(B)$ satisfies*

$$(3.9) \qquad \lambda < 1.$$

Then there exists an index p such that (2.15) *holds.*

Proof. Consider the sequence $\{\omega_n\}$ defined by

$$\omega_0 = w_0, \qquad \omega_{n+1} = B\omega_n + \delta_{n+1} \qquad (n = 0, 1, 2, \cdots).$$

We shall prove that

$$(3.10) \qquad \omega_n - \omega_{n+1} \geqq o \quad \text{for } n \text{ large enough},$$

which is sufficient for the existence of an index p with the desired property.

We first suppose that $\omega_0 = w_0 = o$. Then, if $u_0 = u_1$ the inequality (2.15) is satisfied for $p = 0$. Therefore, we will assume that $\delta_1 = |u_1 - u_0| > o$.

The difference $\omega_n - \omega_{n+1}$ can be written as

$$\omega_n - \omega_{n+1} = (I - B)(B^{n-1}\delta_1 + B^{n-2}\delta_2 + \cdots + \delta_n) - \delta_{n+1} \quad (n = 1, 2, \cdots)$$

with

$$\delta_j = | M^{j-1} \epsilon_1 | \quad (j = 1, 2, \cdots) \tag{3.11}$$

and

$$\epsilon_1 = u_1 - u_0 .$$

In the following we consider two cases described by $\rho(M) < \lambda$ and $\rho(M) = \lambda$, respectively.

Case I. Let

$$\rho(M) < \lambda. \tag{3.12}$$

Using (3.3), we decompose $\omega_n - \omega_{n+1}$ into two summands

$$\omega_n - \omega_{n+1} = S_n^1 + S_n^2 \quad (n = 1, 2, \cdots) \tag{3.13}$$

with

$$S_n^1 = (1 - \lambda)(\lambda^{n-1}\delta_1^{(1)} + \lambda^{n-2}\delta_2^{(1)} + \cdots + \delta_n^{(1)})\varphi,$$

$$S_n^2 = (P_2 - B_2)\eta_n - \delta_{n+1}$$

and

$$\eta_n = B_2^{n-1}\delta_1 + B_2^{n-2}\delta_2 + \cdots + \delta_n .$$

Because of (3.6), all coefficients $\delta_j^{(1)}$ are non-negative. Therefore, we get

$$\lambda^{-n+1}S_n^1 \geqq (1 - \lambda)\delta_1^{(1)}\varphi \quad (n = 1, 2, \cdots), \tag{3.14}$$

and from (3.1), (3.8), and (3.9) we show that the vector on the right side of (3.14) has all components positive.

Because the spectral radii of B_2 and M are smaller than λ and because δ_j is of the form (3.11), the series $\sum_{n=0}^{\infty} \lambda^{-n}B_2^n$ and

$$\sum_{n=1}^{\infty} \lambda^{-n+1}\delta_n \tag{3.15}$$

converge. Therefore, the product series

$$\sum_{n=1}^{\infty} \lambda^{-n+1}\eta_n = \sum_{i=0}^{\infty} \lambda^{-i}B_2^i \cdot \sum_{j=1}^{\infty} \lambda^{-j+1}\delta_j \tag{3.16}$$

also converges. In particular, the summands $\lambda^{-n+1}\delta_n$ and $\lambda^{-n+1}\eta_n$ of the series (3.15) and (3.16) converge to the null vector. Thus, we have

$$\lim_{n\to\infty} \lambda^{-n+1}S_n^2 = o, \tag{3.17}$$

and this relation, together with the inequality (3.14), indicates that (3.10) holds in Case I.

Case II. Suppose now that $\rho(M) = \lambda$. Then, according to a result of Wielandt [8], M can be written as

$$M = e^{i\alpha}D^{-1}BD \tag{3.18}$$

where α denotes a real number and D is a diagonal matrix with diagonal elements of modulus 1. Because M is supposed to be real the relation (3.18) holds for $\alpha = 0$ or $\alpha = \pi$ and a diagonal matrix D with diagonal elements ± 1.

In this case the vectors δ_j take the form

$$\delta_j = | M^{j-1} \epsilon_1 | = | D^{-1} B^{j-1} D \, \epsilon_1 | = | B^{j-1} \zeta | = | \lambda^{j-1} \zeta^{(1)} \varphi + B_2^{j-1} \zeta |$$

with $\zeta = D \, \epsilon_1$ $(j = 1, 2, \cdots)$.

If $\zeta^{(1)} = 0$, then we start again with the decomposition (3.13), for which (3.14) holds. As in Case I, we can prove again that (3.17) holds. For this, we now use $\delta_j = | B_2^{j-1} \zeta |$ and (3.4) instead of (3.11) and (3.12). As in Case I, (3.14) and (3.17) together yield (3.10).

Now let $\zeta^{(1)} \neq 0$. Then we have

$$\delta_j = | \zeta^{(1)} | \lambda^{j-1} | \psi_j | \quad \text{with} \quad \psi_j = \varphi + [\zeta^{(1)}]^{-1} (\lambda^{-1} B_2)^{j-1} \zeta \quad (j = 1, 2, \cdots).$$

Because of (3.1) and since the second summands of the ψ_j converge to o for $j \to \infty$ there exists a number j_0 such that $\psi_j \geqq o$ for $j \geqq j_0$; hence

$$\delta_j = | \zeta^{(1)} | \lambda^{j-1} \varphi + (sgn \; \zeta^{(1)}) B_2^{j-1} \zeta \quad \text{for} \quad j \geqq j_0 . \tag{3.19}$$

Suppose now that $n > j_0 + 1$. Then we write

$$\omega_n - \omega_{n+1} \quad \text{as a sum} \quad \omega_n - \omega_{n+1} = S_n^{\,3} + S_n^{\,4} \qquad (n > j_0 + 1)$$

where

$$S_n^{\,3} = (I - B)(B^{n-1} \delta_1 + B^{n-2} \delta_2 + \cdots + B^{n-j_0} \delta_{j_0})$$

and

$$S_n^{\,4} = (I - B)(B^{n-j_0-1} \delta_{j_0+1} + \cdots + \delta_n) - \delta_{n+1} .$$

Using (3.3) we decompose $S_n^{\,3}$ further into a sum

$$S_n^{\,3} = S_n^{\,31} + S_n^{\,32}$$

with

$$S_n^{\,31} = (1 - \lambda)(\lambda^{n-1} \delta_1^{\,(1)} + \lambda^{n-2} \delta_2^{\,(1)} + \cdots + \lambda^{n-j_0} \delta_{j_0}^{(1)}) \varphi,$$

$$S_n^{\,32} = (I - B_2) B_2^{\,n-j_0} (B_2^{\,j_0-1} \delta_1 + \cdots + \delta_{j_0}).$$

Since the coefficients $\delta_j^{\,(1)}$ are non-negative, we get

$$\lambda^{-n+1} S_n^{\,31} \geqq (1 - \lambda) \delta_1^{\,(1)} \varphi \qquad (n > j_0 + 1), \tag{3.20}$$

where the vector on the right side of the inequality has all components positive.

The second summands $S_n^{\,32}$ satisfy

$$\lim_{n \to \infty} \lambda^{-n+1} S_n^{\,32} = o. \tag{3.21}$$

This follows from (3.4) because j_0 is a fixed number.

Using (3.19), we also split up $S_n^{\,4}$ into the following sum:

$$S_n^{\,4} = S_n^{\,41} + S_n^{\,42}$$

with

$$S_n^{41} = |\zeta^{(1)}|[(I - B)(\lambda^{j_0}B^{n-j_0-1} + \cdots + \lambda^{n-1}I)\varphi - \lambda^n\varphi]$$
$$= |\zeta^{(1)}|\lambda^{n-1}[(n - j_0)(1 - \lambda) - \lambda]\varphi$$

and

$$S_n^{42} = (sgn\ \zeta^{(1)})[(I - B)(B^{n-j_0-1}B_2^{j_0} + \cdots + B_2^{n-1})\zeta - B_2^n\zeta]$$
$$= (sgn\ \zeta^{(1)})B_2^{n-1}[(n - j_0)(I - B_2) - B_2]\zeta.$$

We have

$$\lambda^{-n+1}S_n^{41} \geqq o \text{ for } n \text{ large enough,} \tag{3.22}$$

and from (3.4) we deduce that

$$\lim_{n\to\infty} \lambda^{-n+1}S_n^{42} = o. \tag{3.23}$$

Altogether, from the relations (3.20), (3.21), (3.22), and (3.23) it follows that (3.10) holds in Case II also.

Finally, let $\omega_0 = w_0 > o$. Then, the difference $\omega_n - \omega_{n+1}$ gets an additional summand

$$\tilde{S}_n = (I - B)B^n w_0 = (1 - \lambda)\lambda^n w_0^{(1)}\varphi + (I - B_2)B_2^n w_0 \quad (n = 1, 2, \cdots).$$

These summands $\tilde{S}_n$ satisfy

$$\lim \lambda^{-n+1}\tilde{S}_n = (1 - \lambda)\lambda w_0^{(1)}\varphi$$

where the vector on the right side has all components positive. Thus, (3.10) also holds in this case $w_0 > o$ (even if $u_1 - u_0 = o$, i.e., $\delta_1^{(1)} = 0$ in (3.14) and (3.20)).

Corollary. *Suppose that matrix B is irreducible and let $u_1 \neq u_0$. Then, condition (3.9) is necessary for the existence of a vector v_0 which satisfies*

$$v_0 \geqq o \quad and \quad |u_0 - u_1| \leqq v_0 - Bv_0. \tag{3.24}$$

Remark. This corollary, together with Theorem 3, says, roughly speaking, that if B is irreducible and noncyclic, the method of estimation in Section 2 is always successful if one can get an estimate with Theorem 2.

Proof of the corollary. Since $u_1 \neq u_0$, it follows from (3.24) that

$$v_0 > o \quad \text{and} \quad (I - B)v_0 > o,$$

hence, in view of (3.8)

$$v_0^{(1)} > o \quad \text{and} \quad [(I - B)v_0]^{(1)} = (1 - \lambda)v_0^{(1)} > o.$$

These two inequalities can hold only if $\lambda < 1$.

4. Applications to Numerical Examples. A given system of m linear equations

$$Gu = r \quad \text{with} \quad g_{ii} > 0 \qquad (i = 1, 2, \cdots, m) \tag{4.1}$$

can be written in the form (2.1) as follows. Let $G = D - C_1 - C_2$ where D is the diagonal matrix with diagonal elements g_{ii} and C_1 is some lower triangular matrix.

Then equation (4.1) is equivalent to (2.1) with $s = (D - C_1)^{-1}r$ and

$$M = (D - C_1)^{-1}C_2 . \tag{4.2}$$

In case $C_1 = O$ the iteration procedure

$$u_{n+1} = Mu_n + s \qquad (n = 0, 1, 2, \cdots) \tag{4.3}$$

is the method of simultaneous displacements of the system (4.1). On the other hand, if C_2 is an upper triangular matrix (4.3) represents the method of successive displacements for equation (4.1).

For matrix M given in equation (4.2) inequality (2.2) is satisfied for

$$B = (D - | C_1 |)^{-1} | C_2 | . \tag{4.4}$$

Clearly, one need not start the estimation procedure described at the end of Section 2, and the approximation procedure (4.3) simultaneously. One can define $w_0 = w_1 = \cdots = w_q$ (q denoting some non-negative integer), then start to calculate further vectors by equation (2.14). In this case, the estimation method is described by the following formulas for matrix B in equation (4.4):

$$w_q \geqq o \text{ (with } q \text{ a given non-negative integer);}$$

$$(D - | C_1 |)\tau_{n+1} = | C_2 | w_n , \; w_{n+1} = \tau_{n+1} + \delta_{n+1} \quad (n = q, q + 1, \cdots , p)$$

where $\delta_{n+1} = | u_{n+1} - u_n |$ and p is the smallest index such that $w_p \geqq w_{p+1}$;

$$z_p = w_p , \qquad (D - | C_1 |)z_{n+1} = | C_2 | z_n \qquad (n = p, p + 1, \cdots).$$

This procedure has been programmed for the CDC 1604 Computer. Of course the index p and the bounds z_n depend on the value of the chosen q. To indicate this we write $p = p(q)$ and $z_n = z_n(q)$. According to Theorem 2 we have

$$| u^* - u_n | \leqq z_n(q) \quad \text{for} \quad n \geqq p(q).$$

The following examples have been calculated using the program mentioned above with $w_q = o$.

Example 1. The system $Gu = r$ with G and r as given in Table 1 consists of difference equations which approximate a certain boundary value problem for the Laplace equation [4]. We solve this system by the method of successive displacements. In the present case, we have $C_1 = | C_1 |$ and $C_2 = | C_2 |$, thus

$$M = B. \tag{4.5}$$

The starting vector u_0 as given in Table 2 is the (rounded) solution of difference equations for a smaller mesh width. In the same table are listed the exact errors $\zeta_n = u_n - u^*$ for some indices n and the corresponding bounds $z_n(q)$ for $q = 0, 10,$ and 15. The indices $p(q)$ belonging to these values of q are

$$p(0) = 3, \qquad p(10) = 11, \qquad p(15) = 16.$$

This indicates that the estimation method was successful for $q = 0$ after three steps and for $q = 10$ and $q = 15$ after one step. Table 2 shows that the bounds for $q = 10$ and $q = 15$ are equal (up to nine decimals) but sharper than the bounds for $q = 0$. We learn from this that the estimation procedure should not be started immediately ($q = 0$), but rather after the iteration process (4.3) has become "steady." That the bounds for $q = 10$ and $q = 15$ are almost equal is a consequence of (4.5) (in this connection, see also the next example). It would have been sufficient to start the estimation procedure at $q = 15$.

Example 2. Matrix G and vector r of the second example are given in Table 3. The corresponding equation $Gu = r$ consists of difference equations approximating a boundary value problem for the differential equation $\Delta\Delta u = \varphi(x, y)$ [1]. In general, condition $\rho(B) < 1$ is not satisfied for such problems. However, the estimation method works in this case with a few unknowns. We calculate this example in order to test how the bounds $z_n(q)$ behave if $\rho(M) < \rho(B) < 1$

Numerical results for $q = 0, 10$, and 25 are given in Table 4. The corresponding indices p are

$$p(0) = 2, \qquad p(10) = 12, \qquad p(25) = 27.$$

In this example, the largest index q gives the sharpest bounds. This can always be expected if $\rho(M) < \rho(B)$, because in this case the vectors $|\zeta_n|$ decrease faster than the bounds z_n. For example, if the equation det $(M - \kappa I) = 0$ has a simple root κ_1 such that $|\kappa_1| = \rho(M)$ and all other eigenvalues of M have moduli smaller than $|\kappa_1|$, then we have in general

$$(4.6) \qquad |\zeta_{n+1}| \approx \rho(M)\,|\zeta_n| \text{ for all sufficiently large } n$$

while

$$z_{n+1} \approx \rho(B) z_n \text{ for all sufficiently large } n.$$

The number of steps $p(q) - q$ from the beginning of the estimation procedure until success is achieved is the same for the two larger values of q. This phenomenon also occurred in Example 1. It can be explained by the fact that, in general, $\delta_{n+1} \approx \rho(M)\delta_n$ for n large enough.

Further Examples. In Example 3 we have solved a system of 15 difference equations which approximate the same boundary value problem as the equations in Example 1 and which are of the same type as those equations; however, in Example 3 the row-sum criterion is not satisfied. Compared with Example 1 there has been no essential difference in the behavior of the estimation procedure. Therefore, we give only a few numerical results in Table 5. Furthermore, we applied our method to a system $Gu = r$ with a 2-cyclic matrix G (Example 4) using the

TABLE 1
Coefficients of Example 1

Matrix G								Vector r
12	−1	−1	0	−2	−2	−2	−2	1
−2	12	0	0	0	−4	−4	0	1
−2	0	12	−2	−4	0	0	−2	0
0	0	−2	14	−2	0	0	0	0
−2	0	−2	−1	13	−1	0	−1	0
−2	−2	0	0	−1	12	−1	0	0
−2	−2	0	0	0	−1	12	−1	6
−4	0	−2	0	−2	0	−2	12	2

TABLE 2

Results of Example 1 (u_0 = *starting vector*, $\zeta_n = u_n - u^*$ = *exact error*, $z_n(q)$ = *bounds for* ζ_n)

i	1	2	3	4	5	6	7	8
u_0^i	0.399 6	0.465 5	0.215 5	0.053 9	0.153 0	0.215 5	0.715 5	0.502 2
ζ_3^i	0.002 736 691	0.002 203 276	0.002 229 846	0.000 710 506	0.001 296 122	0.001 153 026	0.001 221 478	0.001 713 471
$z_3^i(0)$	0.009 759 418	0.008 644 262	0.007 523 341	0.002 570 770	0.004 525 089	0.004 433 653	0.004 477 596	0.006 008 977
ζ_{11}^i	0.000 005 720	0.000 004 516	0.000 004 751	0.000 001 519	0.000 002 721	0.000 002 393	0.000 002 550	0.000 003 577
$z_{11}^i(10)$	0.000 006 650	0.000 005 249	0.000 005 523	0.000 001 766	0.000 003 163	0.000 002 781	0.000 002 965	0.000 004 159
$z_{11}^i(0)$	0.000 020 738	0.000 016 370	0.000 017 223	0.000 005 508	0.000 009 863	0.000 008 673	0.000 009 244	0.000 012 968
ζ_{16}^i	0.000 000 121	0.000 000 096	0.000 000 101	0.000 000 032	0.000 000 058	0.000 000 051	0.000 000 054	0.000 000 075
$z_{16}^i(15)$	0.000 000 141	0.000 000 112	0.000 000 117	0.000 000 038	0.000 000 067	0.000 000 059	0.000 000 063	0.000 000 088
$z_{16}^i(10)$	0.000 000 141	0.000 000 112	0.000 000 117	0.000 000 038	0.000 000 067	0.000 000 059	0.000 000 063	0.000 000 088
$z_{16}^i(0)$	0.000 000 439	0.000 000 347	0.000 000 365	0.000 000 116	0.000 000 209	0.000 000 184	0.000 000 196	0.000 000 275

TABLE 3
Coefficients of Example 2

Matrix G				Vector r
12	−3	−3	1	1
−3	10	−2	−3	1
−3	−2	10	−3	1
2	−6	−6	11	1

TABLE 4
Results of Example 2 (u_o = starting vector, $\zeta_n = u_n - u^*$ = exact error, $z_n(q)$ = bounds for ζ_n)

i	1	2	3	4
u_0^i	1	1	1	1
ζ_2^i	0.079 260 342	0.097 059 129	0.175 309 129	0.117 789 897
$z_2^i(0)$	0.275	0.327 954 545	0.273 545 455	0.334 426 997
ζ_{12}^i	0.000 474 178	0.000 677 978	0.000 601 444	0.000 611 652
$z_{12}^i(10)$	0.000 861 331	0.001 076 452	0.000 967 820	0.001 140 014
$z_{12}^i(0)$	0.009 386 534	0.011 627 371	0.010 937 345	0.014 014 669
ζ_{27}^i	0.000 000 096	0.000 000 137	0.000 000 122	0.000 000 123
$z_{27}^i(25)$	0.000 000 174	0.000 000 217	0.000 000 196	0.000 000 230
$z_{27}^i(10)$	0.000 006 455	0.000 007 996	0.000 007 522	0.000 009 638
$z_{27}^i(0)$	0.000 076 075	0.000 094 237	0.000 088 644	0.000 113 585
ζ_{30}^i	0.000 000 018	0.000 000 025	0.000 000 022	0.000 000 023
$z_{30}^i(25)$	0.000 000 062	0.000 000 076	0.000 000 072	0.000 000 092
$z_{30}^i(10)$	0.000 002 464	0.000 003 053	0.000 002 871	0.000 003 679
$z_{30}^i(0)$	0.000 029 040	0.000 035 973	0.000 033 838	0.000 043 358

procedure of simultaneous displacements. In this case matrix B in (4.4) was 2-cyclic also, and the estimation procedure was unsuccessful.

General Remarks Concerning Practical Application. The sharpness of the bounds $z_n(q)$ depends on the chosen index q. In general, for larger q one gets sharper bounds after the estimation procedure has been successful, i.e., for $n \geqq p(q)$. Only if $\rho(M) = \rho(B)$ can one expect that the bounds are almost equal for different q, provided q is so large that the iteration process has become "steady." Moreover, the time needed for the estimation becomes smaller for larger q. Therefore, the best way might be to start the estimation procedure with an index q such that δ_q is smaller than a suitably chosen bound. Then, the approximations are improved still more for $n > q$.

There are several further possible ways to verify the program. For example, one may use the fact that in general the differences $p(q) - q$ become equal for q large enough. One may start at some index q_1 in order to find $p(q_1) - q_1$, then stop the estimation procedure and start it again with a suitable index q_2 such that $p(q_2) \approx q_2 + (p(q_1) - q_1)$.

In our examples we used the starting vector $w_q = o$. In other cases however,

TABLE 5
Some Results of Example 3

$u_0{}^1$	$= 0.39256$
u_{18}^1	$= 0.384\ 711\ 925$
$z_{18}^1(15)$	$= 0.000\ 007\ 026$
u_{28}^1	$= 0.384\ 707\ 090$
$z_{28}^1(25)$	$= 0.000\ 000\ 145$
$z_{28}^1(15)$	$= 0.000\ 000\ 145$
u_{30}^1	$= 0.384\ 707\ 035$
$z_{30}^1(25)$	$= 0.000\ 000\ 067$
$z_{30}^1(15)$	$= 0.000\ 000\ 067$

for a vector

$$w_q = \beta \delta_q$$

with $\beta > 0$ the corresponding number p might be much smaller than for $w_q = o$.

For example, let

$$M \geqq O \quad \text{and} \quad B = M.$$

Then, for q large enough the difference $u_q - u_{q+1}$ in general is approximately proportional to φ. Suppose that $u_q - u_{q+1} = \varphi$ and choose $w_q = \beta\delta_q = \beta\varphi$. Then the first index $n = p$ for which $w_p \geqq w_{p+1}$ holds is the smallest integer such that

$$p \geqq \lambda[(1-\lambda)^{-1} - \beta] + q \quad \text{and} \quad p \geqq q.$$

For $\beta = 0$ and λ very close to 1 this is a large number, for $\beta > (1-\lambda)^{-1}$ however, one has $p = q$. Of course, for β much larger than $(1-\lambda)^{-1}$ the bound $w_p = w_q$ is not sharp.

For example, solving large systems of difference equations for the Laplace equation by the method of successive displacements one may choose

$$w_q = (1-\kappa)^{-1}\delta_q$$

where κ approximates the corresponding eigenvalue λ. Such an approximation κ is known in many cases.

Round-off errors have not been considered in our program. However, we believe that it is not difficult to do this if suitable subroutines are available. It certainly is not necessary to take into consideration all round-off errors which occur in the entire approximation and estimation procedure. One has to do this only for the last step.

Let n_0 be the index up to which the approximations and bounds have been calculated. Then, consider the vectors u_{n_0-1} and z_{n_0-1} as they are computed with a certain number of digits. If one can show that

$$z_{n_0-1} \geqq o \quad \text{and} \quad |u_{n_0-1} - u_{n_0}| \leqq z_{n_0-1} - z_{n_0} \tag{4.7}$$

hold, then as a consequence of Theorem 1, applied to u_{n_0-1} and z_{n_0-1} instead of u_0

and v_0, there exists a solution u^* such that $| u^* - u_{n_0} | \leqq z_{n_0}$ and $| u^* - u_{n_0-1} | \leqq z_{n_0-1}$.

If the entire approximation and estimation process could be done without round-off errors, then, certainly, inequalities (4.7) would be satisfied. This follows from the statement (2.10) in Theorem 1. Therefore, in general, one can expect that (4.7) can be proved also for the vectors u_{n_0-1} and z_{n_0-1} which are actually computed.

In (4.7), u_{n_0} and z_{n_0} do not denote the vectors numerically calculated, but the exact vectors defined by

$$u_{n_0} = Mu_{n_0-1} + s, \qquad z_{n_0} = Bz_{n_0-1}. \tag{4.8}$$

Thus, one has to estimate the round-off errors which occur in computing u_{n_0} and z_{n_0} by (4.8).

In order to do this, one may, for example, reckon with pairs of numbers (instead of numbers) in (4.8). For example, a subroutine for calculation with pairs of numbers has been written for the IBM 650 Computer and this subroutine has been successfully applied for error estimation for certain differential equations [6]. A similar subroutine has been developed by G. E. Collins for the IBM 704 Computer (1959). The method of calculating with pairs of numbers is called "interval arithmetic" by Collins. This method has been used for desk computers by Dwyer [2] under the name "range arithmetic."

Institut für Angewandte Mathematik
Universität Hamburg

Mathematics Research Center
University of Wisconsin
Madison, Wisconsin

1. L. Collatz, *The Numerical Treatment of Differential Equations*, third edition, Springer-Verlag, Berlin, 1961, p. 391.
2. P. S. Dwyer, *Linear Computations*, John Wiley & Sons, Inc., New York, 1951.
3. G. Frobenius, "Über Matrizen aus nichtnegativen Elementen," *S.-B. Preuss. Akad. Wiss. Berlin*, 1912, p. 456–477.
4. J. Schröder, "Zur Lösung von Potentialaufgaben mit dem Differenzenverfahren," *Z. Angew. Math. Mech.*, v. 34, 1954, p. 241–253.
5. J. Schröder, "Anwendung von Fixpunktsätzen bei der numerischen Behandlung nichtlinearer Gleichungen in halbgeordneten Räumen," *Arch. Rational Mech. Anal.*, v. 4, 1959, p. 177–192.
6. J. Schröder, "Fehlerabschätzung mit Rechenanlagen bei gewöhnlichen Differentialgleichungen erster Ordnung," *Numer. Math.*, v. 3, 1961, p. 39–61, 125–130.
7. J. Schröder, "Lineare Operatoren mit positiver Inversen," *Arch. Rational Mech. Anal.*, v. 8, 1961, p. 408–434.
8. H. Wielandt, "Unzerlegbare nichtnegative Matrizen," *Math. Z.*, v. 52, 1950, p. 642–648.