ON RICHARDSON'S METHOD FOR SOLVING LINEAR SYSTEMS WITH POSITIVE DEFINITE MATRICES

By David Young[1]

1. Introduction. In 1910 L. F. Richardson [6][2] presented an iterative method for solving systems of linear equations which arise in the finite difference solutions of boundary value problems associated with elliptic partial differential equations For a linear system of the form

$$\sum_{j=1}^{N} a_{i,j} u_j + d_i = 0, \qquad (i = 1, 2, \cdots, N) \tag{1}$$

where $u_1, u_2, \cdots, u_N$ are the unknowns, the iterative formula for Richardson's method is

$$u_i^{(n)} = u_i^{(n-1)} + \beta_n \left\{ \sum_{j=1}^{N} a_{i,j} u_j^{(n-1)} + d_i \right\}, \qquad (n \geqq 1), \qquad (i = 1, 2, \cdots, N) \tag{2}$$

where the trial values $u_1^{(0)}, u_2^{(0)}, \cdots, u_N^{(0)}$ are arbitrary and where the constants β_n are "relaxation factors" to be chosen.

It appears that even though Richardson's method was proposed more than forty years ago it had been used very little until quite recently. Richardson gave only rough suggestions for choosing the β_n, and if the choice of these numbers is based only on the suggestions given, then the rate of convergence is not much larger than that of the method wherein one uses a single fixed β_n for all iterations. Recently Flanders and Shortley, [2], used a method similar to Richardson's method for problems of finding by iteration the smallest eigenvalues of certain matrices. They used a theorem, proved by Markoff, [5], on Tschebyscheff polynomials to choose numbers analogous to the β_n of (2) and thereby obtained more rapid convergence.

This theorem on Tschebyscheff polynomials was applied to the solution of linear systems independently by Lanczos, [13], by Shortley, [14], and by the author, [12] The method used by Lanczos differs from Richardson's method and does not appear to be quite as simple from a computational point of view, although the possible gain in rate of convergence should be the same Shortley used Tschebyscheff polynomials in the solution of the difference equation analogue of Laplace's equation in two and three dimensions. He also obtained asymptotic estimates for the gain in the rate of convergence. His procedure, while ultimately equivalent to Richardson's method, differs in a practical computational aspect since it involves the use of linear operators which are polynomials in a single linear transformation defined on an N-dimensional vector space. In most cases the method requires the storage of many numbers and,

[1] This paper was prepared in part by research assigned to the Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland, by the Office of the Chief of Ordnance under Project No. TB3-0007K It was completed under contract DA-36-034-ORD-966, placed by the Office of Ordnance Research with the University of Maryland

[2] Numbers in brackets, [ ], refer to the bibliography at the end of the paper.

although useful for computing machines with large storage capacities, it does not appear to be as promising as Richardson's method for use on faster machines having fairly limited storage capacities.

In the present paper we shall assume that the $N \times N$ matrix A of coefficients $a_{i,j}$ in (1) is symmetric and positive definite. We shall use Tschebyscheff polynomials to select the β_n which, in a certain sense, yield the fastest possible convergence for Richardson's method. If the β_n are suitably chosen, then the rate of convergence is asymptotically proportional to the square root of the rate of convergence of the method using the best single β_n, as the latter rate of convergence tends to zero. For a linear system associated with the usual difference analogue of the Dirichlet problem, the number of iterations varies inversely as the first power of the mesh size as compared to the second power of the mesh size for the method using a fixed β_n. Hence the gain in using Richardson's method is large for problems involving a fine mesh.

The best choice of the β_n depends on having good estimates of the maximum and minimum eigenvalues of A. Upper bounds present little difficulty, but lower bounds are often difficult to estimate. In Sec. 3 the dependence of the rate of convergence on the accuracy of the estimated lower bound of the eigenvalues of A is considered. The relative decrease in the rate of convergence is shown to be nearly proportional to the relative error in the estimated lower bound. In Sec. 4 the problem of the control of roundoff errors is discussed.

It is believed that the theorem on the gain in the rate of convergence for Richardson's method is new, as are the theorems on the effects of errors in estimating the minimum eigenvalue of A.

In Sec. 5 we give a brief comparison of Richardson's method and other methods including the Successive Overrelaxation Method and gradient methods. The order-of-magnitude gain in the convergence rate is the same for Richardson's method as for the Successive Overrelaxation Method, and Richardson's method can be shown to apply under more general conditions. On the other hand, the other method is better adapted for machines since it requires less storage, is simpler, has a much smaller tendency to accumulate roundoff errors, and in many cases it can be shown to converge at least twice as fast.

2. Convergence. Let $v = (v_1, v_2, \cdots, v_N)$ denote generically a vector in the vector space V of N - tuples of real numbers. Since the matrix A is symmetric and positive definite, there exist N positive eigenvalues $\nu_1, \nu_2, \cdots, \nu_N$ and N linearly independent eigenvectors $v^{(1)}, v^{(2)}, \cdots, v^{(N)}$ such that

$$\sum_{j=1}^{N} a_{i,j} v_j^{(k)} = \nu_k v_i^{(k)}, \qquad (i = 1, 2, \cdots, N), (k = 1, 2, \cdots, N). \tag{3}$$

Moreover, the eigenvectors are orthogonal under the inner product defined by

$$(x, y) = \sum_{i=1}^{N} x_i y_i$$

and hence form an orthogonal basis for V_N.

Since the determinant of A does not vanish there exists a unique solution of (1) which we denote by u. Let us define the error of the n-th iterated vector by

$$e^{(n)} = u^{(n)} - u. \tag{4}$$

Evidently, by (1), (2), and (4) we have

$$e_i^{(n)} = e_i^{(n-1)} + \beta_n \left\{ \sum_{j=1}^{N} a_{i,j} e_j^{(n-1)} \right\}, \qquad (i = 1, 2, \cdots, N) \tag{5}$$

or

$$e^{(n)} = Y(\beta_n)[e^{(n-1)}] \tag{5'}$$

where $Y(\beta_n)$ is a linear operator in V_N.

Given any $e^{(0)}$, there exist constants $c_1, c_2, \cdots, c_N$ such that

$$e^{(0)} = \sum_{k=1}^{N} c_k v^{(k)}. \tag{6}$$

Using (5) and (6) we obtain, for any integer p

$$e^{(p)} = \sum_{k=1}^{N} c_k v^{(k)} \prod_{n=1}^{p} (1 + \beta_n \nu_k), \tag{7}$$

and by the orthogonality of the eigenvectors we have

$$\| e^{(p)} \|^2 = (e^{(p)}, e^{(p)}) = \sum_{k=1}^{N} c_k^2 \| v^{(k)} \|^2 \left\{ \prod_{n=1}^{p} (1 + \beta_n \nu_k) \right\}^2, \tag{8}$$

or

$$\| e^{(p)} \|^2 \leqq \| e^{(0)} \|^2 \boldsymbol{\lambda}_p^2 \tag{9}$$

where

$$\lambda_p = \operatorname{Max} \left| \prod_{n=1}^{p} (1 + \beta_n \nu_k) \right|, \qquad (k = 1, 2, \cdots, N). \tag{10}$$

For a given integer m, in order to make the least upper bound for all $e^{(0)} \in V_N$ of the ratio $\| e^{(m)} \| / \| e^{(0)} \|$ as small as possible we would choose the $\beta_n^{(m)}$ such that λ_m is a minimum. (Here the superscript m is introduced to indicate the dependence on m). However, if we know only that for some numbers a, b

$$0 < a \leqq \nu_k \leqq b, \qquad (k = 1, 2, \cdots, N) \tag{11}$$

then it appears natural to consider as the *optimum* set of $\beta_n^{(m)}$ those which minimize λ_m', where

$$\lambda_m' = \operatorname*{Max}_{a \leqq \nu \leqq b} \left| \prod_{n=1}^{m} (1 + \beta_n^{(m)} \nu) \right|. \tag{12}$$

Brauer [1] has shown that for all k

$$\nu_k \leqq \text{Max}\left\{\sum_{j=1}^{N} | a_{i,j} |\right\}, \qquad (i = 1, 2, \cdots, N);$$

hence we have an upper bound for the eigenvalues of A. There do not appear to be in the literature any explicit formulas for lower bounds, however.

For convenience, let us introduce the new variable γ defined by

$$\gamma = -\frac{2\nu}{b - a} + \frac{b + a}{b - a}. \tag{13}$$

The interval $a \leqq \nu \leqq b$ is mapped onto the interval $-1 \leqq \gamma \leqq 1$ in such a way that $\gamma = 1$ corresponds to $\nu = a$ and $\gamma = -1$ corresponds to $\nu = b$. Next, let $P_m(\nu)$ denote the polynomial

$$P_m(\nu) = \prod_{n=1}^{m} (1 + \beta_n^{(m)} \nu) \tag{14}$$

and let $Q_m(\gamma)$ be defined by

$$Q_m(\gamma) = P_m(\nu) \tag{15}$$

Since $P_m(0) = 1$ we have $Q_m[(b + a)/(b - a)] = 1$. The problem of minimizing λ_m' is thus equivalent to the problem of finding a polynomial in γ of degree m which equals unity for $\gamma = (b + a)/(b - a)$ and which has the smallest maximum absolute value in the interval $-1 \leqq \gamma \leqq 1$. By a theorem of Markoff [5], (A proof is given by Flanders and Shortley, [2]), the desired polynomial is $S_m(\gamma)$, where

$$S_m(\gamma) = T_m(\gamma)/T_m\left(\frac{b + a}{b - a}\right). \tag{16}$$

Here $T_m(\gamma)$ is the m-th order Tschebyscheff polynomial given by

$$T_m(\gamma) = \cos (m \cos^{-1} \gamma). \tag{17}$$

We note that (17) is valid even when γ is greater than unity, for although $\cos^{-1} \gamma$ is complex, $T_m(\gamma)$ is real In fact, for $\gamma > 1$ we can replace (17) by

$$T_m(\gamma) = \tfrac{1}{2}[(\gamma + (\gamma^2 - 1)^{\frac{1}{2}})^m + (\gamma + (\gamma^2 - 1)^{\frac{1}{2}})^{-m}] = \cosh (m \cosh^{-1} \gamma) \tag{18}$$

In order to make $Q_m(\gamma) = S_m(\gamma)$, we equate corresponding zeros. Let

$$t_1^{(m)}, t_2^{(m)}, \cdots, t_m^{(m)}$$

denote the zeros of $T_m(\gamma)$; these are given by

$$t_n^{(m)} = \cos [(2n - 1)\pi/2m], \qquad (n = 1, 2, \cdots, m). \tag{19}$$

Evidently the $t_n^{(m)}$ are the zeros of $S_m(\gamma)$ Next, the zeros of $Q_m(\gamma)$ are the values of γ corresponding to those values of ν which are the roots of the equation $P_m(\nu) = 0$. By (14) the roots of the latter equation are

$$\nu_n = -(\beta_n^{(m)})^{-1} \qquad (n = 1, 2, \cdots, m),$$

and by (13) the corresponding zeros of $S_m(\gamma)$ are given by

$$\gamma_n = \frac{2}{b-a}(\beta_n^{(m)})^{-1} + \frac{b+a}{b-a}. \tag{20}$$

Equating the zeros of $Q_m(\gamma)$ and $S_m(\gamma)$ and solving for $\beta_n^{(m)}$ we get

$$\beta_n^{(m)} = 2[(b-a)t_n^{(m)} - (b+a)]^{-1}, \quad (n = 1, 2, \cdots, m) \tag{21}$$

In the interval $-1 \leqq \gamma \leqq 1$, the maximum absolute value of $T_m(\gamma)$ equals unity, hence by (12), (14), (15), (16), and (18) we get

$$\lambda_m' = \operatorname*{Max}_{-1 \leqq \gamma \leqq 1} | S_m(\gamma) | = \left[T_m\left(\frac{b+a}{b-a}\right)\right]^{-1} < 1. \tag{22}$$

In the sense that λ_m' is minimized, the choice of the $\beta_n^{(m)}$ given by (21) is the best possible if exactly m iterations are to be performed. On the other hand, if after m iterations have been completed it is felt that more accuracy is needed, one can perform m' more iterations, where m' may equal m, using the appropriate $\beta_n^{(m')}$. However, although it would presumably have been more efficient to have used $\beta_n^{(m+m')}$ from the start, once the $\beta_n^{(m)}$ have been used, it is not efficient to do this since for $m'' > m$, very few of the $\beta_n^{(m)}$ will be included among the $\beta_n^{(m'')}$. It therefore appears best to choose a reasonable value of m and to use the $\beta_n^{(m)}$ in a cyclic order such as

$$\beta_n^{(m)} = \beta_r^{(m)} \tag{23}$$

where r is an integer such that $r \equiv n \pmod{m}$ and $1 \leqq r \leqq m$.

Since $\lambda_m \leqq \lambda_m'$ it follows from (9), (18) and (22) that with the choice of $\beta_n^{(m)}$ determined by (23), the method converges for any m. Incidentally, if $0 > \beta_n > -2/b$ for all n, then the method converges, since for $0 < \nu \leqq b$, we have

$$| 1 + \beta_n \nu | < 1.$$

However, this condition is by no means necessary for convergence.

We remark that, although once a value of m has been chosen, the $\beta_n^{(m)}$ are determined by (21), nevertheless the order in which the $\beta_n^{(m)}$ are used within a cycle of m iterations is arbitrary. This ordering does not affect the *theoretical* rate of convergence, but, as we shall see in Sec. 4, it may be important because of the growth of roundoff errors.

3. Rate of convergence. We define the *rate of convergence* of a linear transformation T on the vector space V_N by

$$R(T) = -\log \Lambda \tag{24}$$

where Λ is the *spectral norm* of T, that is, Λ is the maximum of the absolute values of the eigenvalues of T. It is easy to show that if $\Lambda < 1$, then the rate of convergence of T is approximately inversely proportional to q, the number of times T must be applied to an arbitrary vector v so that $\| T^q[v] \|$ will be less

than a specified fraction of $\| v \|$, see for instance [11]. Thus if v were the initial error vector and an iteration scheme with transformation T were used repeatedly, the number of iterations needed to achieve a desired accuracy would be approximately inversely proportional to $R(T)$.

However, in Richardson's method, using the $\beta_n^{(m)}$ in a cyclic order, one does not use the same transformation repeatedly but, rather, a product of transformations. Therefore, it seems reasonable to consider the *average* rate of convergence

$$R_m = -(1/m) \log \Gamma_m \tag{25}$$

where Γ_m is the spectral norm of the transformation $\prod_{n=1}^{m} Y(\beta_n^{(m)})$. The $Y(\beta_n^{(m)})$ are defined by (5) and (5′).

Evidently, if v is an eigenvector of A with eigenvalue ν, then v is also an eigenvector of $Y(\beta)$ with eigenvalue $(1 + \beta\nu)$. Therefore, by (12) and (22) we have

$$\Gamma_m \leqq \left[T_m \left(\frac{b + a}{b - a} \right) \right]^{-1}, \tag{26}$$

and by (18) and (25)

$$R_m \geqq \frac{1}{m} \log T_m \left(\frac{b + a}{b - a} \right) = -\frac{1}{m} \log \frac{2\alpha^m}{1 + \alpha^{2m}} \tag{27}$$

where, for convenience, we let

$$\alpha = \frac{\sigma}{1 + (1 - \sigma^2)^{\frac{1}{2}}} \tag{28}$$

$$\sigma = \frac{b - a}{b + a}. \tag{29}$$

But since a is an eigenvalue of A, it follows from (13)–(17) that

$$\Gamma_m \geqq T_m[(b + a)/(b - a)]^{-1};$$

hence

$$R_m = -\frac{1}{m} \log \frac{2\alpha^m}{1 + \alpha^{2m}} = \frac{1}{m} \log T_m(\sigma^{-1}). \tag{30}$$

For the method using a single relaxation factor, $m = 1$ and by (19) and (21) we have $\beta_n^{(1)} = -2/(b + a)$ for all n. The spectral norm of $Y(\beta_n^{(1)})$ is given by

$$\Gamma_1 = \sigma \tag{31}$$

and the rate of convergence is

$$R_1 = -\log \sigma. \tag{32}$$

Let us now compare R_m with R_1 as σ approaches unity.

THEOREM 1. Let r be an integer not less than 2 and let $m = m(\sigma)$ denote an integer such that $m \geqq \log 2^r/(-\log \alpha)$, where α is given by (28). If R_m and R_1 satisfy (30) and (32) respectively, then

$$\underline{\mathrm{Lim}}_{\sigma \to 1-} (R_m/R_1^{\frac{1}{2}}) \geqq (1 - r^{-1})2^{\frac{1}{2}}. \tag{33}$$

PROOF. Since $m \geqq \log 2^r/(-\log \alpha)$ and $r \geqq 2$, it follows from (30) that $R_m \geqq (1 - r^{-1})(-\log \alpha)$. Therefore by (32) we have

$$\underline{\mathrm{Lim}}_{\sigma \to 1-} (R_m/R_1^{\frac{1}{2}}) \geqq (1 - r^{-1}) \mathrm{Lim}_{\sigma \to 1-} [(-\log \alpha)/(-\log \sigma)^{\frac{1}{2}}].$$

But, using L'Hospital's rule and the identity

$$d\alpha/d\sigma = (1 - \sigma^2)^{-\frac{1}{2}}[1 + (1 - \sigma^2)^{\frac{1}{2}}]^{-1}$$

the limit in the right member of the last expression equals

$$2\mathrm{Lim}_{\sigma \to 1-} [(-\log \sigma)^{\frac{1}{2}}/(1 - \sigma^2)^{\frac{1}{2}}] = 2\{\mathrm{Lim}_{\sigma \to 1-} [(-\log \sigma)/(1 - \sigma^2)]\}^{\frac{1}{2}}$$

$$= 2\{\mathrm{Lim}_{\sigma \to 1-} (2\sigma^2)^{-1}\}^{\frac{1}{2}} = 2^{\frac{1}{2}}$$

and the theorem follows.

For sufficiently large r, the right member of (33) can be made arbitrarily close to $2^{\frac{1}{2}}$. On the other hand, since at least m iterations must be performed, it would not be efficient to require that m be much larger than $-\log \rho/R_m$ where ρ is that fraction of $\| e^{(0)} \|$ to which the norm of the error must be reduced. This is true since $(-\log \rho)/R_m$ is the approximate number of iterations which would be required for convergence if each iteration affected the convergence equally. If $\rho \leqq 2^{-q}$, for some integer q, then r may be as large as $q + 1$ without m being excessive.

As an illustration, let us consider the set of linear equations one obtains in the numerical solution of the Dirichlet problem for the unit square by finite differences, see for instance [7], [8] or [11]. Let us assume that there are N interior net points and assign to each an integer i such that $1 \leqq i \leqq N$, and denote the value of the difference equation solution at the i-th point by u_i. Then in (1) we have $a_{i,i} = 4$ for all i and if $i \neq j$ then $a_{i,j} = 0$ or -1 depending on whether the points i, j are adjacent or non-adjacent. The d_i of (1) are functions of the boundary values, and do not affect the convergence of the iteration process.

It is not difficult to show, see for instance [11], that if h is the mesh size, then exact upper and lower bounds for the eigenvalues of A are, respectively, $a = 4(1 - \cos \pi h)$ and $b = 4(1 + \cos \pi h)$. Therefore by (29) we have $\sigma = \cos \pi h$ and $R_1 = -\log \sigma = \frac{1}{2}\pi^2 h^2 + O(h^4)$. If $m \geqq \log 2^r/(-\log \alpha)$, where

$$\alpha = \cos \pi h/(1 + \sin \pi h)$$

and $r \geqq 2$, then by Theorem 1 we have, as $h \to 0$

$$R_m \sim \tfrac{1}{2}\pi h.$$

Since the required number of iterations is approximately inversely proportional to the rate of convergence, it follows that the number of iterations varies as h^{-1} with Richardson's method using $m \geqq \log 2^r/(-\log \alpha)$, (with $r \geqq 2$), as $h \to 0$, compared with h^{-2} using $m = 1$.

These results were found to hold reasonably well when a problem involving $h = 1/20$ was solved on the ORDVAC computing machine, (See [15]). The number of iterations required to satisfy a given convergence criterion with $m = 20$ was about one-ninth the number of iterations required with $m = 1$.

In general, however, we are not so fortunate as to have exact upper and lower bounds for the eigenvalues of A. As we have seen, lower bounds are the most difficult to obtain and we therefore proceed to discuss the effect on the rate of convergence of using a_1 in (21) to determine the $\beta_n^{(m)}$ instead of the true value a.

Let C_σ denote the set of symmetric positive definite matrices such that if $A \in C_\sigma$ then $\sigma = (b - a)/(b + a)$, where a, b denote respectively the minimum and maximum eigenvalues of A. Further, for any matrix $A \in C_\sigma$ let $R_m(\sigma, \theta, A)$ denote the rate of convergence of Richardson's method for a linear system with matrix A such that the m relaxation factors $\beta_n^{(m)}$ are determined from (21) using the correct value of b but replacing a by θa. Evidently by (30) we have

$$R_m(\sigma, 1, A) = (1/m) \log T_m(\sigma^{-1}). \tag{34}$$

First let us consider the case $\theta \leqq 1$. If $a_1 = \theta a$ then all eigenvalues of A lie in the range $a_1 \leqq \nu \leqq b$; hence we have

$$R_m(\sigma, \theta, A) \geqq (1/m) \log T_m(\sigma_1^{-1}), \qquad (\theta \leqq 1), \tag{35}$$

where $\sigma_1 = (b - a_1)/(b + a_1)$. We now prove

THEOREM 2. If $m = m(\sigma)$ is an integer such that $m \geqq \log 4/(-\log \alpha)$, where α is given by (28), and if $\theta \leqq 1$, then

$$\overline{\lim_{\sigma \to 1-}} \left(1 - \frac{R_m(\sigma, \theta, A)}{R_m(\sigma, 1, A)}\right) \leqq (1 - \theta)\theta^{\frac{1}{2}}. \tag{36}$$

PROOF. From (34) and (35) it follows that $R_m(\sigma, 1, A) - R_m(\sigma, \theta, A) \leqq (1/m)[\log T_m(\sigma^{-1}) - \log T_m(\sigma_1^{-1})]$. Let $c_1 = (1 - \sigma_1)/(1 + \sigma_1)$ and $c = (1 - \sigma)/(1 + \sigma)$. By the mean value theorem we have

$$\frac{1}{m} \log T_m(\sigma^{-1}) - \frac{1}{m} \log T_m(\sigma_1^{-1}) = (1 - \theta) \left\{\frac{d}{d\theta}\left[\frac{1}{m} \log T_m(\sigma_1^{-1})\right]\right\}_{\theta = \bar{\theta}}$$

where $\theta < \bar{\theta} < 1$. But

$$-\frac{d}{d\theta}\left[\frac{1}{m} \log T_m(\sigma_1^{-1})\right] = \frac{2c}{(1 - c_1)^2} \tanh [m \cosh^{-1}(\sigma_1^{-1})][\sigma_1^{-2} - 1]^{-\frac{1}{2}}$$

$$\leqq \frac{c^{\frac{1}{2}}}{1 - c_1} \left(\frac{c}{c_1}\right)^{\frac{1}{2}} = \frac{c^{\frac{1}{2}}}{1 - c_1} \theta^{-\frac{1}{2}}$$

Therefore

$$\frac{1}{m}\log T_m(\sigma^{-1}) - \frac{1}{m}\log T_m(\sigma_1^{-1}) \leqq (1-\theta)\frac{c^{\frac{1}{2}}}{1-\bar{c}_1}\theta^{-\frac{1}{2}} \tag{37}$$

where $c_1 \leqq \bar{c}_1 \leqq c$. Moreover, since $m \geqq \log 4/(-\log \alpha)$, it follows that

$$R_m(\sigma, 1, A) \geqq -\tfrac{1}{2}\log \alpha.$$

Also, by using L'Hospital's rule we get $\operatorname{Lim}_{\sigma\to 1-}[(-\log\alpha)/2c^{\frac{1}{2}}] = 1$, hence if we divide both sides of (37) by $m^{-1}\log T_m(\sigma^{-1})$ and take the limit superior, we obtain the desired result.

Next, let us consider the case where $\theta \geqq 1$. The rate of convergence is equal to $-\log \Gamma_m$, where $\Gamma_m = \operatorname{Max}\left|\prod_{n=1}^{m}(1+\bar{\beta}_n^{(m)}\nu)\right|$ the maximum being taken over all eigenvalues ν of A, and where $\bar{\beta}_n^{(m)} = 2[(b-a_1)t_n^{(m)} - (b+a_1)]^{-1}$ with $a_1 = \theta a$. Letting $\gamma_1 = (-2\nu + b + a_1)(b-a_1)^{-1}$ we obtain as before

$$\prod_{n=1}^{m}(1+\bar{\beta}_n^{(m)}\nu) = T_m(\gamma_1)/T_m(\sigma_1^{-1}).$$

But since $\gamma_1(a) = 1 + 2(a_1 - a)/(b-a_1)$, and since $T_m(x)$ is an increasing function for $x \geqq 1$, it follows that $\Gamma_m = T_m(\gamma_1(a))/T_m(\sigma_1^{-1})$ and that

$$R_m(\sigma,\theta,A) = \frac{1}{m}\left[\log T_m(\sigma_1^{-1}) - \log T_m\left(1+\frac{2(c_1-c)}{1-c_1}\right)\right], \qquad (\theta \geqq 1), \tag{38}$$

where $c = a/b = (1-\sigma)/(1+\sigma)$ and $c_1 = a_1/b = (1-\sigma_1)/(1+\sigma_1)$. We now prove

THEOREM 3. If $m = m(\sigma)$ is an integer such that $\log 4 \leqq m/(-\log\alpha) \leqq K$, where K is a constant, and if $\theta \geqq 1$, then

$$\overline{\operatorname{Lim}_{\sigma\to 1-}}\left(1 - \frac{R_m(\sigma,\theta,A)}{R_m(\sigma,1,A)}\right) \leqq (\theta-1)\left(K - \frac{15}{17}\theta^{-\frac{1}{2}}\right). \tag{39}$$

PROOF. For fixed m, we have by (18) and (38)

$$-\frac{d}{d\theta}[R_m(\sigma,\theta,A)] = \tanh\left[m\cosh^{-1}\left(1+\frac{2(c_1-c)}{1-c_1}\right)\right]\frac{2c(1-c)}{(1-c_1)^2}$$
$$\cdot\left[\left(1+\frac{2(c_1-c)}{1-c_1}\right)^2 - 1\right]^{-\frac{1}{2}} - \tanh\left[m\cosh^{-1}(\sigma_1^{-1})\right]\frac{2c\sigma_1^{-1}}{(1+c_1)^2}(1-\sigma_1^2)^{-\frac{1}{2}}.$$

Since $\tanh x \leqq x$ and $\log(1+x) \leqq x$ for $x \geqq 0$, we have

$$-\frac{d}{d\theta}[R_m(\sigma,\theta,A)] \leqq m\left[1+\left(\frac{c_1-c}{1-c_1}\right)^{\frac{1}{2}}\right]\frac{2c(1-c)}{(1-c_1)^2}$$
$$- \tanh\left[m\cosh^{-1}(\sigma_1^{-1})\right]\frac{c}{(1+c_1)c_1^{\frac{1}{2}}\sigma_1}.$$

But $\overline{\lim}_{\sigma\to 1-} mc^{\frac{1}{2}} \leqq \lim_{\sigma\to 1-} [Kc^{\frac{1}{2}}/(-\log \alpha)] = K/2$, as in the proof of Theorem 2. Therefore, using the mean value theorem we get

$$\overline{\lim_{\sigma\to 1-}}\left(1 - \frac{R_m(\sigma, \theta, A)}{R_m(\sigma, 1, A)}\right) \leqq \overline{\lim_{\sigma\to 1-}}\left[(\theta - 1)\frac{c^{\frac{1}{2}}}{R_m(\sigma, 1, A)}\left\{\frac{2mc^{\frac{1}{2}}(1 - c)}{(1 - \bar{c}_1)^2}\right.\right.$$

$$\left.\left.\cdot\left[1 + \left(\frac{\bar{c}_1 - c}{1 - \bar{c}_1}\right)^{\frac{1}{2}}\right] - \tanh\,[m\cosh^{-1}(\bar{\sigma}_1^{-1})](c/\bar{c}_1)^{\frac{1}{2}}(1 + \bar{c}_1)^{-1}\bar{\sigma}_1^{-1}\right\}\right]$$

where $c \leqq \bar{c}_1 \leqq c_1$. Since $\cosh^{-1}(\bar{\sigma}_1^{-1}) \geqq \cosh^{-1}(\sigma^{-1}) = -\log \alpha$, we have $\tanh\,[m\cosh^{-1}(\bar{\sigma}_1^{-1})] \geqq \tanh(\log 4) = 15/17$. Passing to the limit and recalling that $\underline{\lim}_{\sigma\to 1-} [R_m(\sigma, 1, A)/c^{\frac{1}{2}}] \geqq 1$, and $\underline{\lim}_{\sigma\to 1-} (c/\bar{c}_1)^{\frac{1}{2}} \geqq \theta^{-\frac{1}{2}}$, the theorem follows.

From Theorems 2 and 3 we may conclude that the relative decrease in the rate of convergence, due to using a_1 instead of a, is approximately proportional to the relative error in estimating a, and that the constant of proportionality is not large. This is true whether the error is the result of overestimation or of underestimation.

4. Effect of roundoff. On the majority of large automatic computing machines all quantities must have modulus less than unity. Unless a floating decimal, or binary, point routine is used, at the expense of a large part of the storage capacity of the machine, numbers with modulus greater than unity must be premultiplied by a suitable scaling factor. However if in a certain storage register, the sizes of the numbers vary over wide ranges, in order to insure that the largest number has modulus less than unity, few significant digits can be available for the smaller numbers. For this reason in our problem it is desirable that the $u_i^{(m)}$ should not vary too much in magnitude.

This situation is not easy to achieve, however, if the quantities $|\beta_n^{(m)}|$ are large, as in the case where a, the lower bound of the eigenvalues of A, is small. In fact, by (21) we find that $\beta_1^{(m)} \to -a^{-1}$ as $m \to \infty$. Thus in the example of the preceding section, if $h^{-1} = 20$, then $a = .0492$, and if $m = 20$ then $\beta_1^{(m)} = -16.28$. For larger m, $|\beta_1^{(m)}|$ increases but never exceeds 20.32.

Let us now consider the error $e^{(p)}$ given by (7). Although the product

$$\prod_{n=1}^{p}(1 + \beta_n^{(m)}\nu_k)c_k$$

which is theoretically equal to the coefficient of $v^{(k)}$ in the expansion

$$e^{(p)} = \textstyle\sum_{k=1}^{N} c_k^{(p)} v^{(k)}$$

tends to zero as $p \to \infty$, the actual coefficient $c_k^{(p)}$ may not tend to zero. For, because of roundoff there exists some positive δ such that, in general, the actual coefficient will exceed δ in magnitude; thus one is led to consider the product

$$\prod_{n=p_1}^{p_1+p_2} |1 + \beta_n^{(m)}\nu_k| \,\delta \tag{40}$$

for all k, p_1, p_2 This product gives a more realistic estimate for the upper bound of $c_k^{(p_1+p_2)}$, and may become so large that the convergence of the method will be slowed or even prevented, and in severe cases the capacity of the machine will be exceeded.

The growth of roundoff errors may be inhibited by using the $\beta_n^{(m)}$ in a certain order within each cycle. It is clear that to use the $\beta_n^{(m)}$ in descending order of magnitude, as in (21), is not suitable since for large ν_k the largest of the factors $| 1 + \beta_n^{(m)}\nu_k |$ in (40) are used first. It is somewhat better to use the order of ascending magnitudes In this case the product $\prod_{n=1}^{p} | (1 + \beta_n^{(m)}\nu_k) |$ is reduced somewhat before the large factors are used. However the product cannot become smaller than δ because of roundoff. Then as the large $\beta_n^{(m)}$ are used the coefficient of $v^{(k)}$ may become as large as

$$\prod_n | 1 + \beta_n^{(m)}\nu_k | \delta$$

where the product is taken over all factors such that $| 1 + \beta_n^{(m)}\nu_k | \geqq 1$.

A better method, which has been more successful, is to use the ordering

$$\beta_{(m/2+1)}^{(m)}, \beta_{m/2}^{(m)}, \beta_{(m/2+2)}^{(m)}, \cdots, \beta_m^{(m)}, \beta_1^{(m)}$$

when m is even, and

$$\beta_{(m+1)/2}^{(m)}, \beta_{[(m+1)/2+1]}^{(m)}, \beta_{[(m+1)/2-1]}^{(m)}, \cdots, \beta_m^{(m)}, \beta_1^{(m)}$$

when m is odd. Here we assume that the $\beta_n^{(m)}$ are labeled with descending magnitude, i.e. $| \beta_1^{(m)} | \geqq | \beta_2^{(m)} | \geqq \cdots \geqq | \beta_m^{(m)} |$. This procedure has the advantage of using alternately large and small $\beta_n^{(m)}$, thus preventing uninterrupted growth of roundoff errors. Of course, still further improvements are possible.

We remark that the growth of roundoff errors can be considerably reduced if one uses polynomial operators, as was done by Shortley, [14] However, as already noted, this requires the storage of many additional numbers

5. Comparison with other methods. We now discuss briefly the relation of Richardson's method with other iterative methods such as the Gauss-Seidel method, the Successive Overrelaxation Method [11], and gradient methods including the method of steepest descent [10] and [4] and the method of conjugate gradients, [9] and [4].

With gradient methods one modifies the tentative solution at each stage in such a way as to reduce the value of a certain positive definite quadratic form. These methods appear to be very promising, especially the method of conjugate gradients where theoretically one obtains the exact solution after N iterations. Nevertheless, a considerable amount of further study and testing on computing machines remains before these methods can be accurately evaluated as tools for solving linear systems. Of course the same is true of Richardson's method, but it appears that the gradient methods require more computations per iteration. Furthermore, although estimates of the maximum and minimum eigenvalues are not needed to insure theoretical convergence for the gradient methods, never-

theless, these estimates may be necessary in actual practice in order to control the growth of roundoff errors.

In [11] it is shown that under certain assumptions on A, the Successive Over-relaxation Method (S.O.R. Method) defined by

$$(41)\qquad \begin{cases} u_i^{(n+1)} = -(\omega/a_{i,i})\left(\sum_{j=1}^{i-1} a_{i,j}\, u_j^{(n+1)} + \sum_{j=i+1}^{N} a_{i,j}\, u_j^{(n)} + d_i\right) \\ \qquad\qquad -(\omega - 1)u_i^{(n)}, \qquad (i = 1, 2, \cdots, N) \\ u_i^{(0)} \text{ arbitrary}, \qquad (i = 1, 2, \cdot\ \cdot, N) \end{cases}$$

gives the same order-of-magnitude gain in convergence rate as Richardson's method, provided ω is suitably chosen. In fact, it can be shown that if $a_{i,i}$ is independent of i, then the rate of convergence of the method is given by

$$(42)\qquad R = -2 \log \alpha$$

provided

$$(43)\qquad \omega = 1 + \alpha^2.$$

If in (41) we let $\omega = 1$ we obtain the Gauss-Seidel method. Under the same assumptions on A it is shown in [11] that the Gauss-Seidel method converges twice as fast as Richardson's method with $\beta_n = -2/(b + a)$ for all n, provided $a_{i,i}$ is independent of i. Thus Richardson's method with larger m is certainly superior to the Gauss-Seidel method.

The assumptions on A which were made to prove the preceding results are that A be symmetric and positive definite and also have Property (A) defined as follows:

There exist two disjoint subsets P and Q of W, the set of the first N positive integers, such that if $a_{i,j} \neq 0$, then $i = j$ or $i \in P$ and $j \in Q$ or $i \in Q$ and $j \in P$.

Evidently Richardson's method applies under more general conditions. Nevertheless, as shown in [11], for linear systems associated with many elliptic partial difference equations, the matrix does have Property (A). In these cases the S.O.R. Method has the following advantages for use on large automatic computing machines:

1. The storage problem is simpler since new iterated values are used as soon as obtained; hence the old values need not be retained until after a whole iteration has been completed, as is necessary with Richardson's method.

2. If the diagonal elements of A are equal, then the S.O.R. Method converges at least twice as fast as Richardson's method. This follows from (42) and (30). It appears likely that the S.O.R. Method is faster in all cases but this remains to be proved

3 The S.O.R. Method is simpler since only one value of ω is needed while for Richardson's method many different β_n are used. Moreover, roundoff errors do not build up appreciably with the S.O R. Method since $\omega \leqq 2$. On the other hand roundoff is a very serious problem with Richardson's method

BIBLIOGRAPHY

1 A. Brauer, *Limits for the characteristic roots of a matrix*, Duke Math. Jour , **13**, pp 387–395 (1946).

2. D Flanders and G. Shortley, *Numerical determination of fundamental modes*, J. App Phys , **21**, pp 1326–1332 (1950).

3. M Hestenes and E Stiefel, *Methods of conjugate gradients for solving linear systems*, Journal of Research, **49**, pp. 409–436 (1952).

4 L V Kantorovich, *On an effective method of solving extremal problems for quadratic functionals*, Comptes Rendus (Doklady) de l'Académie des Sciences de L'URSS, 1945, **XLVIII**, Nov 7.

5 W Markoff, *Über Polynome, die in einem gegeben Intervalle möglichst wenig von Null abweichen*, Math. Ann , **77**, (1916), pp 213–258 (Translation and condensation by J Grossman of Russian article published in 1892)

6 L F Richardson, *The approximate solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam*, Roy Soc Phil Trans **210A**, pp 307–357 (1910)

7 G Shortley and R Weller, *The numerical solution of Laplace's equation*, J. App. Phys , **9**, pp 334–344 (1938).

8. R Southwell, *Relaxation methods in theoretical physics*, University Press, Oxford, 1946.

9 E Stiefel, *Uber einige Methoden der Relaxationsrechnung*, Zeits f angew Math. u Phys , **3**, pp 1–33 (1952)

10 G Temple, *The general theory of relaxation methods applied to linear systems*, Proc Roy. Soc , **A169**, pp 476–500 (1938–39)

11 D Young, *Iterative methods for solving partial difference equations of elliptic type*, doctoral dissertation, Harvard Univ 1950 A revised version has been accepted for publication

12 D. Young, *On Richardson's method for solving linear systems with positive definite matrices*, abstract of a paper presented at the October 25, 1952 meeting of the American Mathematical Society, New Haven, Conn Abstract 26t, Bull. Amer Math Soc , **59**, pp 47–48, 1953

13 C Lanczos, *Solution of systems of linear equations by minimized iterations*, Journal of Research, **49**, pp 33–53 (1952).

14 G Shortley, *Use of Tschebyscheff-polynomial operators in the numerical solution of boundary-value problems*, J App Phys , **24**, pp. 392–396 (1953).

15. D Young and C Warlick, *On the use of Richardson's method for the numerical solution of Laplace's equation on the ORDVAC*, Ballistic Research Laboratories Memorandum Report No. 707, Aberdeen Proving Ground, Md., July 1953.

University of Maryland

(Received April 1, 1953)