ROUND-OFF ERROR ANALYSIS OF ITERATIONS FOR LARGE LINEAR SYSTEMS

H. Woźniakowski
Department of Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

April 1977

This work was partly done during the author's visit at Carnegie-Mellon University and it was supported in part by the Office of Naval Research under Contract N00014-76-C-0370; NR 044-422 and by the National Science Foundation under Grant MCS75-222-55.

ABSTRACT

We deal with the rounding error analysis of successive approximation iterations for the solution of large linear systems $Ax = b$. We prove that Jacobi, Richardson, Gauss-Seidel and SOR iterations are <u>numerically stable</u> whenever $A = A^* > 0$ and A has Property A. This means that the computed result x_k approximates the exact solution α with relative error of order $\zeta\|A\| \cdot \|A^{-1}\|$ where ζ is the relative computer precision. However with the exception of Gauss-Seidel iteration the residual vector $\|Ax_k - b\|$ is of order $\zeta\|A\|^2 \|A^{-1}\| \|\alpha\|$ and hence the remaining three iterations are <u>not well-behaved</u>.

1. INTRODUCTION

This paper deals with the rounding error analysis in floating point arithmetic of successive approximation iterations for the solution of large sparse linear systems $Ax = b$.

We summarize the results of this paper. Basic concepts of numerical stability and good-behavior are recalled in Section 2. We give necessary and sufficient conditions for numerical stability and good-behavior in Sections 3 and 5. In Section 4 we deal with several examples of successive approximation iterations. We prove that Jacobi, Richardson, Gauss-Seidel and SOR iterations <u>are numerically stable</u> whenever $A = A^* > 0$ and A has Property A. In Section 6 we show that with the exception of Gauss-Seidel iteration they <u>are not well-behaved</u>. In the last section we indicate that good-behavior of any numerically stable method can be achieved by the use of iterative refinement even if all computations are performed in single precision.

2. PRELIMINARIES

In this section we briefly recall what we mean by numerical stability and good-behavior of an iteration for solving a linear system $Ax = b$ where A is a $n \times n$ nonsingular complex matrix and b is a $n \times 1$ vector. We shall assume throughout this paper that $\|\cdot\|$ denotes the spectral norm.

Let $\{x_k\}$ be a computed sequence of successive approximations of the solution $\alpha = A^{-1}b$ by an iteration φ in t digit floating point arithmetic fl, see Wilkinson [63].

An iteration φ is called <u>numerically stable</u> if

$$(2.1) \quad \overline{\lim_{k}} \|x_k - \alpha\| \leq \zeta c_1 \operatorname{cond}(A) \|\alpha\| + O(\zeta^2)$$

where $\zeta = 2^{-t}$ is the relative computer precision, c_1 is a constant which depends only on the size n of the problem, and $\operatorname{cond}(A) = \|A\| \cdot \|A^{-1}\|$ is the condition number of A.

An iteration φ is called <u>well-behaved</u> (or equivalently φ has good-behavior) if

$$(2.2) \quad \overline{\lim_{k}} \|Ax_k - b\| \leq \zeta c_2 \|A\| \|\alpha\| + O(\zeta^2)$$

where $c_2 = c_2(n)$.

It is easy to verify that good-behavior implies numerical stability but not, in general, vice versa. Furthermore, φ is well-behaved iff there exist matrices E_k such that for large k

$$(2.3) \quad (A + E_k)\, x_k = b \text{ and } \|E_k\| \leq \zeta c_3 \|A\| + O(\zeta^2)$$

for $c_3 = c_3(n)$.

Thus good-behavior means that x_k is the exact solution of a slightly perturbed system or equivalently that the residual vector $r_k = Ax_k - b$ is small in the sense of (2.2).

Recall that commonly used direct methods such as Gaussian elimination with pivoting, Householder method, modified Gram-Schmidt, or Gram-Schmidt with reorthogonalization are well-behaved. Let us also mention that Chebyshev iteration is numerically stable but, in general, is <u>not</u> well-behaved; see Woźniakowski [75] where a detailed discussion of these concepts may be found.

3. NUMERICAL STABILITY OF SUCCESSIVE APPROXIMATION ITERATIONS

We consider the numerical solution of a large linear system

$$(3.1)\quad Ax = b$$

where A is a nonsingular complex $n \times n$ matrix and b is a $n \times 1$ complex vector. We assume that A is a sparse matrix of high order and $\alpha = A^{-1}b$ is the solution of (3.1).

A successive approximation iteration is defined as follows:

(i) Transform $Ax = b$ to an equivalent system

$$(3.2)\quad x = Hx + h,\ (\alpha = H\alpha + h).$$

Sometimes $H = H(A)$ is chosen to minimize the spectral radius $\sigma(H)$ of H, $\sigma(H) < 1$, in a certain class of $\{H(A)\}$.

(ii) Solve (3.2) by the iteration

$$(3.3)\quad x_{k+1} = Hx_k + h, \qquad k = 0,1,\ldots$$

where x_0 is a given initial approximation.

Using different transformations we get different iterations; see Section 4 where Jacobi, Richardson, Gauss-Seidel and successive overrelaxation (SOR) iterations are considered.

Let $e_k = x_k - \alpha$. From (3.3) we get the theoretical error formula

$$(3.4)\quad e_k = H^k e_0.$$

Thus the theoretical iteration is convergent for any x_0 iff the spectral radius $\sigma(H)$ is less than 1. Furthermore the character of convergence mainly depends on $\sigma(H)$ since

$$\lim_{k} \frac{\| e_k \|}{(\sigma(H) + \epsilon)^k} = 0$$

for any $\epsilon > 0$.

Due to the sparseness of A in many cases we can compute the product Hx_k and x_{k+1} in time and storage proportional to n rather than n^2. However in floating point arithmetic fl we cannot compute Hx_k or x_{k+1} from (3.3) exactly.

Assume that

$$(3.5) \quad fl(Hx_k + h) = (H + \delta H_k)\, x_k + (I + \delta I_k)\, h = Hx_k + h + \xi_k$$

where $\| \delta H_k \| \le \zeta c_1 \| H \|$, $\| \delta I_k \| \le \zeta c_2$, c_1 and c_2 depend only on n and

$$(3.6) \quad \xi_k = \delta H_k\, x_k + \delta I_k (I - H)\, \alpha.$$

Note that (3.5) holds for most algorithms used in numerical practice with c_1 and c_2 of order unity.

Thus, instead of the theoretical relation (3.3) we get

$$(3.7) \quad x_{k+1} = Hx_k + h + \xi_k.$$

It follows that the error formula for the computed sequence $e_k = x_k - \alpha$ is equal to

$$(3.8) \quad e_{k+1} = H^{k+1}\, e_0 + \sum_{i=0}^{k} H^{k-i}\, \xi_i,$$

compare with (3.4).

From (3.5) and (3.6) the vectors ξ_k have a bound

$$(3.9)\quad \|\xi_k\| \le \zeta c_3(\|H\| + \|I - H\|)\|\alpha\| + \zeta c_1\|x_k - \alpha\|$$

for $c_3 = \max(c_1, c_2)$.

Let $\{\eta_i\}$ be a sequence such that $\|\eta_i\| \le 1$. Define

$$(3.10)\quad k(H) = (\|H\| + \|I - H\|) \sup_{\|\eta_i\| \le 1} \overline{\lim_k} \left\| \sum_{i=0}^{k} H^{k-i} \eta_i \right\|.$$

From (3.8), (3.9) and (3.10) it easily follows that

$$(3.11)\quad \overline{\lim_k} \|x_k - \alpha\| \le \zeta\, k(H)\, c_3\|\alpha\| + O(\zeta^2).$$

We want to determine when (3.11) is sharp. In order to do this we must assume something more about ξ_k. Recall that the vector ξ_k is the rounding error vector at the kth iterative step, see (3.6) and (3.9). In general, ξ_k can have an arbitrary direction and $\left\| \sum_{i=0}^{k} H^{k-i} \xi_i \right\|$ can be of order $\zeta k(H) c_3\|\alpha\|$. To make this point clear we shall assume throughout this paper that $\{\xi_k\}$ can be any sequence satisfying (3.9). Thus to prove the sharpness of (3.11) it is enough to define $\{\xi_k\}$ such that $\xi_k = (\|H\| + \|I - H\|)\, c_3\|\alpha\|\, \eta_k^*$ where the supremum in (3.10) is attainable for η_k^*.

Note that

$$(3.12)\quad k(H) \le (\|H\| + \|I - H\|) \sum_{i=0}^{\infty} \|H^i\|$$

and the inequality in (3.12) holds for a hermitian H, $H = H^*$.

Comparing (3.11) with the definition of numerical stability (2.1) we see that to get numerical stability of the successive approximation iteration, k(H) has to be of order cond(A). Thus we have proven

<u>Theorem 3.1</u>

If (3.5) holds then the successive approximation iteration given by (3.12) and (3.3) is <u>numerically stable</u> iff

(3.13) $k(H) \le c_5 \operatorname{cond}(A)$

where $c_5 = c_5(n)$ and $k(H)$ is given by (3.10). ■

In the next section we determine for which transformations $k(H)$ is comparable with $\operatorname{cond}(A)$. We want to end this section by showing that for $\|H\|$ not too close to unity we get numerical stability. More precisely let $q \in [0,1)$ be a number not too close to unity ($q \le .9$, say). If $\|H\| \le q$ then due to (3.12) $k(H) \le (2q+1)/(1-q)$ and (3.13) holds with $c_5 = (2q+1)/\{(1-q) \operatorname{cond}(A)\} \le (2q+1)/(1-q)$. This means that the successive approximation iteration is <u>always</u> numerically stable for a class of problems for which $\|H\| \le q$. However, usually for ill-conditioned problems (for large $\operatorname{cond}(A)$), some eigenvalues of H have moduli close to 1 and $k(H)$ is large. Furthermore we shall see that even for well-conditioned problems it can happen that $k(H)$ is large which indicates an unstable case of the successive approximation iteration.

4. EXAMPLES OF NUMERICAL STABILITY

In this section we consider some examples of transformations from $Ax = b$ to $x = Hx + g$ and we find conditions assuring numerical stability.

For the sake of simplicity we assume throughout this section that A is a hermitian, positive definite matrix and A has a form

$$(4.1) \quad A = I - B$$

where B is hermitian and has zero diagonal elements. Furthermore we assume that $\|A\| < 2$. Let λ_{min} and λ_{max} be the smallest and the largest eigenvalue of A. Thus $0 < \lambda_{min} \le 1$ and $1 \le \lambda_{max} \le 2$. Note that $\mathrm{cond}(A) = \lambda_{max}/\lambda_{min}$.

<u>Example 4.1</u> Jacobi Iteration

In this case $H = B$ and $h = b$. Thus assumption (2.5) holds for any reasonable algorithm for computing $Hx_k + h$. Since $H = I - A$ is hermitian then

$$\|H\| = \sigma(H) = \max(1 - \lambda_{min}, \lambda_{max} - 1) < 1.$$

Note that $\sigma(H)$ is close to 1 if λ_{min} is close to zero (which means that the problem is ill-conditioned) or λ_{max} is close to two which can happen even for well-conditioned problems.

From (3.12) we get

$$(4.2) \quad k(H) = \frac{\sigma(H) + \lambda_{max}}{1 - \sigma(H)} .$$

In general $k(H)$ can considerably exceed the condition number $\mathrm{cond}(A)$ even for very small n. For instance let $n = 3$ and

$$(4.3)\quad A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix},\ 0 < a < 1/2,$$

whose eigenvalues are $1 - a$, $1 - a$ and $1 + 2a$, see Young [71, p. 111]. We have $\sigma(H) = 2a$ and

$$k(H) = \frac{1 + 4a}{1 - 2a},\ \mathrm{cond}(A) = \frac{1 + 2a}{1 - a}\ .$$

Thus

$$\lim_{a \to 1/2^-} k(H) = +\infty \quad \text{and} \quad \lim_{a \to 1/2} \mathrm{cond}(A) = 4$$

which means that (3.13) does not hold for values of a close to 1/2. We performed some numerical tests on the PDP-10 computer where

$$\zeta \doteq 3\times 10^{-9} \quad \text{with} \quad \alpha = [1,1,1]^T \quad \text{for} \quad a = \frac{1}{2} - 10^{-i},\ \ i = 2,3,4 \text{ and } 5.$$

The best computed results had relative error of order 10^{-9+i} which confirms theoretical considerations. Thus Jacobi iteration for very well-conditioned system (4.3) with the value of a close to 1/2 is <u>numerically unstable</u>.

To assure that $k(H)$ is of order $\mathrm{cond}(A)$ we have to assume something more concerning the eigenvalues of A.

<u>Theorem 4.1</u>

Jacobi iteration is numerically stable for $A = A^* > 0$ and A is of the form 4.1 iff

$$(4.4)\quad \frac{\lambda_{\min}}{2 - \lambda_{\max}} \le c_6$$

$c_6 = c_6(n)$. ∎

Proof

Assume that (4.4) holds. Consider two cases.

Case I. Let $1 - \lambda_{min} \geq \lambda_{max} - 1$. Then $k(H) = (1 - \lambda_{min} + \lambda_{max})/\lambda_{min} \leq 2\ \mathrm{cond}(A)$ and (3.13) holds with $c_5 = 2$.

Case II. Let $1 - \lambda_{min} < \lambda_{max} - 1$. Then $k(H) = (2\lambda_{max} - 1)/(2 - \lambda_{max})$. But from (4.4) we have $1/(2 - \lambda_{max}) \leq c_6/\lambda_{min}$ and $k(H) \leq 2 c_6\ \mathrm{cond}(A)$ and once more (3.13) holds with $c_5 = 2 c_6$.

The necessity of (4.4) easily follows from the above example (4.3) with $a \to 1/2^-$. Since $\lambda_{min} = 1 - a \doteq 1/2$ and $\lambda_{max} = 1 + 2a \doteq 2$, the lefthand side of (4.4) tends to infinity as a tends to $1/2^-$ which causes instability of Jacobi iteration. ■

Note that if A has Property A or equivalently B has the form

$$(4.5)\quad B = \begin{pmatrix} 0_1 & F \\ F & 0_2 \end{pmatrix}$$

where 0_1 and 0_2 are square mull matrices (see Young [71, p. 42]) then $\lambda_{min} = 2 - \lambda_{max}$ and (4.4) holds with $c_6 = 1$. Thus we get

Corollary 4.1

If $A = A^* > 0$ and A has the form 4.1 and Property A then Jacobi iteration is numerically stable. ■

Example 4.2 Richardson Iteration

In this case

$$(4.6)\quad H = I - c\,A \text{ and } h = -c\,b$$

where $c = 2/(\lambda_{min} + \lambda_{max})$. Then $\|H\| = \sigma(H) = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}$ and $k(H) = \frac{3\lambda_{max} - \lambda_{min}}{2\lambda_{min}} \le \frac{3}{2}$ cond(A) which due to (3.13) proves

<u>Theorem 4.2</u>

If $A = A^* > 0$ then Richardson iteration is numerically stable. ∎

<u>Example 4.3</u> Gauss-Seidel Iteration

Assume that $A = I - B$ has Property A. Thus

$$B = L + U = \begin{pmatrix} 0_1 & F \\ F & 0_2 \end{pmatrix}$$

where L and U are strictly lower and strictly upper triangular matrices. Gauss-Seidel Iteration is defined by

$$H = (I - L^{-1})\, U = \begin{pmatrix} 0_1 & F \\ 0 & F^* F \end{pmatrix},$$
$$h = (I - L^{-1})\, b.$$

It is easy to verify that

$$\text{(4.6)} \quad H^k = \begin{pmatrix} 0_1 & F\,(F^* F)^{k-1} \\ 0 & (F^* F)^k \end{pmatrix}, \quad \|H^k\| = \sigma(B)^{2k-1}\sqrt{1 + \sigma^2(B)}, \quad \forall k \ge 1$$

From (3.12) we get

$$k(H) \le (1 + 2\,\sigma(B)\sqrt{1 + \sigma^2(B)})(1 + \sqrt{1 + \sigma^2(B)} \sum_{k=1}^{\infty} \sigma(B)^{2k-1}) \le$$
$$\le (1 + 2\sqrt{2})(1 + \sqrt{2} / 2 * (1 - \sigma(B))^{-1}).$$

Since $\sigma(B) = 1 - \lambda_{min}$ we have $(1 - \sigma(B))^{-1} = \text{cond}(A) / \lambda_{max} \le \text{cond}(A)$. This proves that

$$k(H) \le c_5 \text{ cond}(A) \quad \text{with} \quad c_5 \le (1 + 2\sqrt{2})(1 + \sqrt{2}/2) \doteq 6.5.$$

Hence we have proven

Theorem 4.3

If $A = A^* > 0$ and A has the form 4.1 and Property A then Gauss-Seidel iteration is numerically stable. ∎

Example 4.4 Successive Overrelaxation Iteration (SOR)

Assume that $A = I - B$ has Property A. SOR iteration is defined by

$$(4.9) \qquad \begin{aligned} H &= (I - w L)^{-1}(w U + (1-w)I), \\ h &= w(I - w L)^{-1} b \end{aligned}$$

where the optimal w is given by

$$w = \frac{2}{1 + \sqrt{1 - \sigma^2(B)}} .$$

It is easy to verify that

$$\sigma(H) = w - 1 = \left(\frac{\sqrt{\text{cond}(A)} - 1}{\sqrt{\text{cond}(A)} + 1}\right)^2 .$$

Furthermore from Young [71, p. 248] it follows that

$$\begin{aligned} \|H\| &= \sigma^k(H)\{k(\sigma(H)^{1/2} + \sigma(H)^{-1/2}) + \sqrt{k^2(\sigma(H)^{1/2} + \sigma(H)^{-1/2})^2 + 1}\} \\ &\le 2.3\ k\ \sigma^k(H)(\sigma(H)^{1/2} + \sigma(H)^{-1/2}) \end{aligned}$$

which yields

$$k(H) \le (1 + 2\ \|H\|)[1 + 2.3(\sigma(H)^{1/2} + \sigma(H)^{-1/2}) \sum_{k=1}^{\infty} k\ \sigma(H)^k] \le$$
$$10.2(1 + 4.6(1 - \sigma(H))^{-2}).$$

Since

$$(1 - \sigma(H))^{-2} = \text{cond}(A)(1 + \text{cond}(A)^{-1/2})^4 \;/\; 16$$

we have $k(H) \le c_5$ cond(A) with $c_5 \le 10.2 * 5.6 \doteq 57$. However if cond(A) is large then c_5 is less than 4. Hence we have proven

<u>Theorem 4.4</u>

If $A = A^* > 0$ and A has the form 4.1 and Property A then SOR iteration is numerically stable. ∎

5. GOOD-BEHAVIOR OF SUCCESSIVE APPROXIMATION ITERATIONS

Recall that we transform the linear system $Ax + g = 0$ to an equivalent system $(I - H)\, x = h$ which is solved by constructing $\{x_k\}$ such that

$$(5.1) \quad x_{k+1} = H\, x_k + h.$$

We define two different sequences of residuals vectors, $A(x_k - \alpha)$ for the original system and $(I - H)(x_k - \alpha)$ for the transformed one. Let

$$(5.2) \quad r_k = M(x_k - \alpha)$$

where $M = A$ or $M = I - H$. We want to verify good-behavior of the successive approximation iteration with respect to A or $I - H$. Due to (2.2) we need to prove that

$$(5.3) \quad \overline{\lim_k}\, \| r_k \| \le \zeta\, c_2\, \|M\|\, \|\alpha\| + O(\zeta^2)$$

for a constant $c_2 = c_2(n)$. From (3.8) we get

$$(5.4) \quad r_{k+1} = M\, H^{k+1}\, e_0 + \sum_{i=0}^{k} M\, H^{k-i}\, \xi_i$$

where ξ_i is given by (3.6) and (3.10).

Let $\{\eta_i\}$ be a sequence such that $\|\eta_i\| \le 1$. Define

$$(5.5) \quad k(M,H) = (\|H\| + \|I - H\|) \sup_{\|\eta_i\| \le 1} \overline{\lim_k}\, \Big\| \sum_{i=0}^{k} M\, H^{k-i}\, \eta_i \Big\| .$$

Note that $k(I,H) = k(H)$.

From 5.4 it easily follows

$$(5.6) \quad \overline{\lim_k}\, \| r_k \| \le \zeta\, k(M,H)\, c_3 \|\alpha\| + O(\zeta^2).$$

Since (5.6) is sharp, (5.3) yields

Theorem 5.1

If (3.5) holds then the successive approximation iteration is well-behaved with respect to M iff

(5.7) $k(M,H) \leq c_6 \ \|M\|$.

where $c_6 = c_6(n)$.

Remark 5.1

We showed in Section 3 that $\|H\| \leq q$ where q is not too close to unity implies numerical stability of the successive approximation iteration. It is also obvious that $\|H\| \leq q$ yields good-behavior since

$$k(M,H) \leq \frac{2q+1}{1-q} \ \|M\|$$

and (5.7) holds with $c_6 = (2q+1) / (1-q)$. ∎

In general, it is rather hard to evaluate k(M,H). However for many cases it is enough to know some bounds on k(M,H).

Lemma 5.1

Let $\lambda \neq 0$ be an eigenvalue of H, $H\xi = \lambda\xi$ with $\|\xi\| = 1$. Then

(5.8) $k(M,H) \geq \frac{1}{1-|\lambda|} \ \|M\xi\|$.

Proof

Define $\eta_i = \frac{\lambda^i}{|\lambda|^i}\,\xi$. Then

$$\sum_{i=0}^{k} M H^{k-i} \eta_i = (\lambda^k \sum_{i=0}^{k} \frac{1}{|\lambda|^i}) M\xi = \frac{\lambda^k}{|\lambda|^k} \frac{1 - |\lambda|^{k+1}}{1 - |\lambda|} M\xi$$

which proves (5.8). ■

Lemma 5.2

Let $M = I - H$. If an iteration is well-behaved then

$$(5.9) \qquad \max_{\lambda \in \mathrm{spect}(H)} \frac{|1 - \lambda|}{1 - |\lambda|} \le c_6 \|I - H\| .$$

Proof

From Lemma 5.1 and 5.7 we get

$$\frac{1}{1 - |\lambda|} \|M\xi\| = \frac{|1 - \lambda|}{1 - |\lambda|} \le k(M,H) \le c_6 \|I - H\|$$

for any eigenvalue of H which proves (5.9). ■

Lemma 5.2 states a necessary condition for good-behavior with $M = I - H$ which means that $|\lambda| \cong 1$ implies $\lambda \cong 1$ for any eigenvalue of H.

Lemma 5.3

Let $M = I - H$ and $H = H^*$. Then an iteration is well-behaved iff

$$(5.10) \qquad \max_{\lambda \in \mathrm{spect}(H)} \frac{1 - \lambda}{1 - |\lambda|} \le c_6 \| I - H\| .$$

Proof

Let $H = U D U^*$ where $U^* U = I$ and $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Let $z_i = [z_1^{(i)}, \ldots, z_n^{(i)}]^T = U^* \eta_i$. Then

$$\sum_{i=0}^{k} M H^{k-i} \eta_i = U(I - D) \sum_{i=0}^{k} D^{k-i} U^* \eta_i = U[(1 - \lambda_1) \sum_{i=0}^{k} \lambda_1^{k-i} z_1^{(i)} ..$$

$$\ldots, (1 - \lambda_n) \sum_{i=0}^{k} \lambda_n^{k-i} z_n^{(i)}]^T$$

and

$$(5.11) \quad k(I - H,H) \le 3 \max_j \frac{1 - \lambda_j}{1 - |\lambda_j|} \overline{\lim_k} \| z_k \| = 3 \max_j \frac{1 - \lambda_j}{1 - |\lambda_j|} .$$

Since (5.11) is sharp, (5.10) is proven. ∎

Note that (5.10) means that H does not have eigenvalues close to -1.

We end this section by showing that for $H = H^*$ it is often possible to redefine the transformed system such that (5.10) holds and yields good-behavior. Multiply $(I - H)x = h$ by $I + H$. Then $x = H^2x + (I + H)\ h$ and we can iterate

$$(5.12) \quad x_{k+1} = H^2 x_k + (I + H)\ h.$$

We shall call the iteration (5.12) as the modified successive approximation iteration. Note that $H^2 = [H^2]^* \ge 0$ and the lefthand side of (5.10) is equal to unity. Thus, if $\| I - H^2 \|$ is not too small, $\| I - H^2 \| \ge c_7$ for $c_7 \ge .1$, say, then we get good-behavior. Hence we have proven

Lemma 5.4

If $H = H^*$ and $\| I - H^2 \| \ge c_7 > 0$ then the modifed successive approximation iteration (5.12) is well-behaved for $M = I - H^2$ and $c_6 = \frac{2}{2/c_7}$. ∎

6. EXAMPLES OF GOOD-BEHAVIOR

As in Section 4 we assume that $A = A^* > 0$. Except Example 6.2 we additionally assume that A has Property A, see (4.1) and (4.5).

Example 6.1 Jacobi Iteration

In this case $H = I - A$ is hermitian and $\lambda_{min} = 2 - \lambda_{max}$. Apply Lemma 5.1 with $\lambda = 1 - \lambda_{max}$ for $M = A$ and next $M = I - H$. In both cases we get

$$k(M,H) \geq \frac{\lambda_{max}}{2 - \lambda_{max}} = \text{cond}(A)$$

which shows that Jacobi iteration is not well-behaved.

For the modified Jacobi iteration (5.12) let $\lambda = (1 - \lambda_{max})^2$. Then

$$k(A,H^2) \geq \frac{\lambda_{max}}{1 - (1 - \lambda_{max})^2} = \frac{1}{2 - \lambda_{max}} \geq \frac{\text{cond}(A)}{2}$$

which contradicts good-behavior. Finally notice that

$$\|I - H^2\| = \max_{\lambda \in \text{spect}(A)} \lambda(2 - \lambda) = c_7.$$

If one of eigenvalues of A is close to unity then $c_7 \cong 1$ which yields good-behavior of the modified Jacobi iteration for $M = I$. Thus we get

Theorem 6.1

Jacobi iteration is not well-behaved for $M = A$ or $M = I - H$. The modified Jacobi iteration is not well-behaved for $M = A$ and it is well behaved for $M = I - H$ whenever A has an eigenvalue close to unity. ∎

Example 6.2 Richardson Iteration

The matrix $H = I - cA$ with $c = 2/(\lambda_{min} + \lambda_{max})$ is also hermitian. Apply Lemma 5.1 with $\lambda = (1 - c\,\lambda_{max})^i$ and $M = A$ for $i = 1,2$. Then

$$k(A,H^i) = \left(\frac{\lambda_{max} + \lambda_{min}}{2}\right)^i \frac{\lambda_{max}^{2-i}}{\lambda_{min}} \geq \frac{\text{cond}(A)}{4}$$

which proves that Richardson and the modified Richardson iterations are not well-behaved for $M = A$.

Next note that H has eigenvalues close to -1 for ill-conditioned problems. Lemma 5.3 shows that Richardson iteration cannot have good-behavior for $M = I - H$. Finally

$$\|I - H^2\| = c^2 \max_{\lambda \in \text{spect}(A)} \lambda(\lambda_{min} + \lambda_{max} - \lambda) = c_7.$$

If one of eigenvalues of A is close to $1/c$ then $c_7 \cong 1$ which implies good-behavior of the modified Richardson iteration for $M = I - H$. Thus we have proven

Theorem 6.3

Richardson iteration is not well-behaved for $M = A$ or $M = I - H$. The modified Richardson iteration is not well-behaved for $M = A$ and it is well-behaved for $M = I - H$ whenever A has an eigenvalue close to $(\lambda_{min} + \lambda_{max}) / 2$. ∎

Example 6.3 Gauss-Seidel Iteration

The matrix H is now defined by

$$H = \begin{pmatrix} 0_1 & F \\ 0 & F^* F \end{pmatrix}.$$

From (4.6) we have

$$(I - H)H^k = \begin{pmatrix} 0_1 & F(I - F^* F)(F^* F)^{k-1} \\ 0 & (I - F^* F)(F^* F)^k \end{pmatrix},$$

$$AH^k = \begin{pmatrix} 0_1 & F(I - F^* F)(F^* F)^{k-1} \\ 0 & 0 \end{pmatrix}, \quad \forall k \geq 1.$$

We estimate k(M,H) from (5.5). Let $\eta_i = [\eta_i^{(1)^T}, \eta_i^{(2)^T}]^T$. Then

$$\sum_{i=0}^{k} (I - H)H^{k-i} \eta_i = [w_k^{(1)^T}, w_k^{(2)^T}]^T$$

where

$$w_k^{(1)} = F(I - F^* F) \sum_{i=0}^{k-1} (F^* F)^{k-1-i} \eta_i^{(2)} + \eta_k^{(1)} - F \eta_k^{(2)},$$

$$w_k^{(2)} = (I - F^* F) \sum_{i=0}^{k} (F^* F)^{k-i} \eta_i^{(2)} .$$

Since $F^* F$ is nonnegative definite then repeating the proof of Lemma 5.3 it is easy to verify that

$$\overline{\lim_k} \| w_k^{(1)} \| \leq (2 \| F \| + 1) \overline{\lim_k} \| \eta_k \| \leq 3,$$

$$\overline{\lim_k} \| w_k^2 \| \leq 1$$

which yields

$$k(I - H, H) \leq (\| H \| + \| I - H \|) \sqrt{9+1} \leq 2\sqrt{10} \doteq 6.3.$$

Due to the form of $A H^k$ it can be verified that

$$k(A, H) \leq 6.$$

Since $\| A \|$ and $\| I - H \|$ are both not less than unity we finally get

$$k(I - H, H) \le 2\sqrt{10}\ \| I - H \|, \quad k(A,H) \le 6\ \| A \|$$

which due to (5.7) proves good-behavior. Hence we have shown

Theorem 6.3

Gauss-Seidel iteration is well-behaved for $M = A$ and $M = I - H$. ∎

Example 6.4 SOR Iteration

In this case

$$H = (I - wL)^{-1}(wU + (1 - w)I)$$

where $w = 2 \,/\, (1 + \sqrt{1 - \sigma^2(B)})$ and $A = I - B$.

Let μ be an eigenvalue of B. Then the eigenvalues of H are equal to

$$\lambda = \frac{1}{2}(w^2\mu^2 - 2(w - 1)) \pm i\sqrt{(4(w - 1) - w^2\mu^2)\, w^2\mu^2}$$

where $i = \sqrt{-1}$, see Young [71, p. 203]. From this

$$|\lambda| = w - 1 = \sigma(H) \text{ and } |1 - \lambda| = w\sqrt{1 - \mu^2}.$$

We apply Lemma 5.1 with $M = I - H$ and next $M = A$. Then

$$k(I - H,H) \ge \frac{|1 - \lambda|}{1 - |\lambda|} = \frac{w\sqrt{1 - \mu^2}}{1 - \sigma(H)} \ge \frac{1}{2}\sqrt{\mathrm{cond}(A)}\ \sqrt{1 - \mu^2}$$

It is known that $\mu = 0$ is an eigenvalue of B whenever the size of the problem n is odd which yields

$$k(I - H,H) \ge \frac{1}{2}\sqrt{\mathrm{cond}(A)}.$$

Hence SOR is not well-behaved for $M = I - H$.

Now let $M = A$, $\mu = -\sigma(B)$ and let ξ be an eigenvector associated with $\lambda = w - 1$, $\xi = [\xi_1^T, \xi_2^T]^T$, $\|\xi\| = 1$. From Young [71, p. 237] it follows

$$A[\xi_1^T, \lambda^{-1/2}\ \xi_2^T]^T = (1 + \sigma(B))[\xi_1^T, \lambda^{-1/2}\ \xi_2^T]^T.$$

Thus

$$k(A,H) \geq \frac{\|A\xi\|}{1 - |\lambda|} = (1 - \sigma(H))^{-1}\|A[\xi_1^T, \lambda^{-1/2}\ \xi_2^T]^T -$$

$$- A[0^T, (\lambda^{-1/2} - 1)\ \xi_2^T]^T\| \geq$$

$$\frac{1}{4}\sqrt{\mathrm{cond}(A)}\ [1 + \sigma(B) - 2\ (\sigma(H)^{-1/2} - 1)]$$

which tends to infinity as cond(A) does. Hence SOR is also <u>not</u> well-behaved for $M = A$. Hence we have

<u>Theorem 6.4</u>

SOR iteration is <u>not</u> well-behaved for $M = I - H$ or $M = A$. ∎

7. FINAL REMARKS

We have shown that certain well-known iterations are numerically stable and except Gauss-Seidel they are not well-behaved. However it is possible to get good-behavior for $M = A$ using iterative refinement with single or double precision for the computation of the residual vectors.

It is shown in Jankowski and Woźniakowski [77] that if $\zeta \operatorname{cond}^2(A)$ is of order of unity then any numerically stable method (direct or iterative) with iterative refinement using only single precision is well-behaved for $M = A$. Since $\zeta \operatorname{cond}^2(A)$ is much less than unity in most practical cases, Jacobi, Richardson and SOR iterations with iterative refinement in single precision <u>are</u> well-behaved.

ACKNOWLEDGMENT

I am indebted to A. Kielbasiński and J. F. Traub for their valuable comments on this paper.

Jankowski and Woźniakowski [77] — **Jankowski**, M., **Woźniakowski**, H, "Iterative Refinement Implies Numerical Stability," to appear in BIT.

Wilkinson [63] — Wilkinson, J.H., Rounding Errors in Algebraic Processes, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

Woźniakowski [75] — Woźniakowski, H., "Numerical Stability of the Chebyshev Method for the Solution of Large Linear Systems," Dept. of Computer Science Report, Carnegie-Mellon University, 1975, to appear in Numerische Mathematik.

Young [71] — Young, D. M., Iterative Solution of Large Linear Systems, Academic Press, New York, 1971.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)

ROUND-OFF ERROR ANALYSIS OF ITERATIONS FOR LARGE LINEAR SYSTEMS

5. TYPE OF REPORT & PERIOD COVERED

Interim

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

H. Woźniakowski

8. CONTRACT OR GRANT NUMBER(s)

N00014-76-C-0370;
NR 044-422

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Carnegie-Mellon University
Computer Science Dept.
Pittsburgh, PA 15213

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS

Office of Naval Research
Arlington, VA 22217

12. REPORT DATE

April 1977

13. NUMBER OF PAGES

27

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)

UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We deal with the rounding error analysis of successive approximation iterations for the solution of large linear systems $Ax = b$. We prove that Jacobi, Richardson, Gauss-Seidel and SOR iterations are numerically stable whenever $A = A^* > 0$ and A has Property A. This means that the computed result x_k approximates the exact solution α with relative error of order $\zeta\|A\| \cdot \|A^{-1}\|$ where ζ is the relative computer precision. However with the exception of Gauss-Seidel iteration the residual vector $\|Ax_k - b\|$ is of order $\zeta\|A\|^2\|A^{-1}\| \, \|\alpha\|$ and hence the remaining three iterations are not well-behaved.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)