Numerical Analysis Project
Manuscript NA-87-01

February 1987

The Convergence of Inexact Chebyshev and Richardson Iterative Methods for Solving Linear Systems

Gene H. Golub
Michael L. Overton

Numerical Analysis Project
Computer Science Department
Stanford University
Stanford, California 94305

THE CONVERGENCE OF INEXACT CHEBYSHEV AND RICHARDSON ITERATIVE METHODS FOR SOLVING LINEAR SYSTEMS

Gene H. Golub* and Michael L. Overton**

(Also issued on a report from the Centre for Mathematical Analysis, Canberra, Australia - CMA-R46-86)

February 1987

* Computer Science Department, Stanford University, Stanford, CA 94305 (Golub at SCORE.Stanford,edu). The work of this author was in part supported by the NSF (DCR 84-12314) and by the U.S. Army (DAAG 2983-K-0124).

**Centre for Mathematical Analysis and Mathematical Sciences Research Institute, Australian National University. (On leave from Courant Institute of Mathematical Sciences, New York University.)
(NA Overton at SCORE.Stanford,edu) The work of this author was in part supported by NSF (DCR-85-02014).

ABSTRACT

The Chebyshev and second-order Richardson methods are classical iterative schemes for solving linear systems. We consider the convergence analysis of these methods when each step of the iteration is carried out inexactly. This has many applications, since a preconditioned iteration requires, at each step, the solution of a linear system which may be solved inexactly using an “inner” iteration. We derive an error bound which applies to the general nonsymmetric inexact Chebyshev iteration. We show how this simplifies slightly in the case of a symmetric or skew-symmetric iteration, and we consider both the cases of underestimating and overestimating the spectrum. We show that in the symmetric case, it is actually advantageous to underestimate the spectrum when the spectral radius and the degree of inexactness are both large. This is not true in the case of the skew-symmetric iteration. We show how similar results apply to the Richardson iteration. Finally, we describe numerical experiments which illustrate theresults and suggest that the Chebyshev and Richardson methods, with reasonable parameter choices, may be more effective than the conjugate gradient method in the presence of inexactness.

1. INTRODUCTION

The Chebyshev method and the second-order Richardson methods are classical iterative schemes for the solution of linear systems of equations. Their convergence analysis is well known; see Golub and Varga (1961). We consider the analysis of these methods where each step of the iteration is carried out inexactly. This has important applications, since one often wants to employ a matrix splitting, introducing a linear system to be solved at each step of the main or "outer" iteration. When the exact solution of these "inner" systems is not possible or practical, an "inner" iteration may be used instead. These inner iterations are generally terminated before they have converged to full accuracy: in other words the inner systems are solved inexactly. We are concerned with the effect this has on the convergence of the outer iteration.

Other papers which study the use of inner and outer iterations include Gunn (1964). Nicolaides (1975). Pereyra (1967), Nichols (1973) and Dembo. Eisenstat and Steihaug (1982). In an earlier paper (Golub and Overton (1982)) we gave a convergence analysis for the inexact Richardson method alone, but the analysis there is somewhat different since it does not take advantage of the freedom of choice in the initial i terates. The analysis given here is also applicable to studying round-off error accumulation; see also Golub (1962).

The paper is organized as follows. In Section 2 we give a general convergence analysis for the inexact Chebyshev method for solving nonsymmetric linear systems. The exact version of this method is the one given by Manteuffel (1977). We give the analysis for the nonsymmetric case because it is not much more complicated than the symmetric version.

In Section 3 we show how the results apply to the symmetric iteration, distinguishing between the cases where the relevant eigenvalues are "underestimated" and "overestimated". We show that underestimating the spectrum can actually speed up convergence if the spectral radius is large and the iteration is quite inexact. In Section 4 we apply the results of Section 2 to the case where the preconditioned matrix defining the outer system is the sum of the identity matrix and a skew-symmetric matrix. We refer to the corresponding iteration as "skew-symmetric". This situation provides one motivation for our work, since it arises as an effective way to solve nonsymmetric systems with positive definite symmetric part, using a skew-symmetric/symmetric splitting. Again. we discuss the issue of overestimating and underestimating the spectrum.

In Section 5 we give a convergence analysis for the inexact second-order Richardson method, restricting it for simplicity to the case of a symmetric iteration where the spectrum is overestimated. The second-order Richardson method is closely related to the SOR method; see Golub and Varga (1961, p.151). In Section 6 we discuss the question of how to choose the parameters to minimize the total amount of work.

It seems to be much more difficult to provide a similar analysis for the conjugate gradient method. However, numerical results given in the final section indicate that the Chebyshev and Richardson methods, with reasonable parameter choices, are less sensitive than the conjugate gradient method to the degree of inexactness of the iteration.

We use $\|\cdot\|$ to denote the Euclidean vector and induced matrix norm. Unless otherwise specified, a statement involving k will be understood to apply for $k = 0,1,2,\ldots$.

2. THE INEXACT CHEBYSHEV METHOD

We wish to solve the real n x n system of linear equations

$$Ax = b \tag{2.1}$$

using the splitting

$$A = M - N . \tag{2.2}$$

In this section, no symmetry assumption is made on the matrix A . However, we need to make assumptions on the eigenvalues of $M^{-1}A$ in order to obtain a convergent method even in the exact case. Manteuffel (1977) showed that the appropriate assumption is that the spectrum of $M^{-1}A$ is contained in an ellipse which lies in the right half of the complex plane. Let ℓ and u be estimates of the foci of such an ellipse; these parameters will define the Chebyshev method. Since $M^{-1}A$ is real, we restrict ℓ and u to be either real (with $\ell < u$) or a complex conjugate pair. . We also make the simplifying assumption that $M^{-1}A$ is diagonalizable. If this is not the case, see the discussion given by Manteuffel (1977, p.309).

The exact Chebyshev method is defined by

$$x_1 = x_0 + z_0 \tag{2.3}$$

$$x_{k+1} = x_{k-1} + \omega_{k+1}(\alpha z_k + x_k - x_{k-1}) \text{ , } k = 1,2,\ldots \tag{2.4}$$

where

$$Mz_k = r_k \tag{2.5}$$

$$r_k = b - Ax_k \tag{2.6}$$

$$\alpha = \frac{2}{\ell+u} \tag{2.7}$$

$$I\text{-}1 = \frac{u+\ell}{u-\ell}$$

(2.9) $$\omega_{k+1} = 2\mu \frac{c_k}{c_{k+1}}$$

(2.10) $$c_{k+1} = 2\mu c_k - c_{k-1} \text{ , } k = 1.2 \text{ ,} \ldots$$

(2.11) $$c_0 = 1 \text{ , } c_1 = \mu$$

and x_0 is given. Note that c_k is the value of the k^{th} Chebyshev polynomial evaluated at μ . The inexact version of the method is obtained by replacing (2.5) by

(2.12) $$Mz_k = r_k + q_k$$

(2.13) $$\|q_k\| \leq \delta \|r_k\|$$

where $\delta \in (0.1)$. This specifies that each inner iteration, if started with the approximation $z_k = 0$, must reduce its residual norm $\|Mz_k - r_k\|$ by a factor of δ .

Let

$$e_k = x - x_k$$

so that

(2.14) $$\begin{aligned} e_{k+1} &= e_{k-1} - \omega_{k+1}(\alpha z_k + e_{k-1} - e_k) \\ &= \omega_{k+1} K e_k + (1-\omega_{k+1}) e_{k-1} + \omega_{k+1} p_k \text{ , } k = 1,2, \ldots \end{aligned}$$

and

(2.15) $$e_1 = Ke_0 + p_0$$

where

(2.16) $$p_k = -\alpha M^{-1} q_k \ , \ k = 1,2,\ldots$$

and

(2.17) $$K = I - \alpha M^{-1} A \ .$$

Now let the eigenvalue decomposition of $M^{-1}A$ be given by

(2.18) $$M^{-1}A = V\Lambda V^{-1}$$

where $\{\lambda_j\}$, the entries in the diagonal matrix A , are the eigenvalues of $M^{-1}A$. Then

(2.19) $$K = V\Sigma V^{-1}$$

where $(a_j\}$, the entries in the diagonal matrix Σ , are the eigenvalues of K and are given by

(2.20) $$\sigma_j = 1 - \alpha\lambda_j \ .$$

Note that if $\{\lambda_j\}$ do indeed lie in an ellipse in the right half plane with foci ℓ and u , then $(a_j\}$ are contained in an ellipse which lies strictly inside the strip of the complex plane defined by $|\mathcal{Re}\ a| \lesssim 1$. Furthermore the quantities $\{\mu\sigma_j\}$, which will be important later, are contained in an ellipse with foci at ±1 and major axis of length $|\mu|$.

Now let

(2.21) $$fk = c_k V^{-1} e_k \ ; \ gk = c_k V^{-1} p_k \quad \text{except } g_0 = \frac{1}{2} V^{-1} p_0 \ .$$

Then we may simplify (2.14), using (2.9) - (2.11). to obtain the diagonalized system of difference equations:

(2.22) $$f_{k+1} = 2\mu\Sigma f_k - f_{k-1} + 2\mu g_k \ , \ k = 1,2,\ldots$$

with

(2.23) $$f_1 = \mu\Sigma f_0 + 2\mu g_0 .$$

We now state a useful lemma.

LEMMA 1 *Consider the inhomogeneous complex* difference *equation in* ξ_k :

(2.24) $$\xi_{k+1} = 2\gamma\beta\xi_k - \gamma^2\xi_{k-1} + \eta_k , \quad k = 1,2,\ldots .$$

The solution is

(2.25) $$\xi_k = \pi_k\xi_1 - \gamma^2\pi_{k-1}\xi_0 + \sum_{\ell=1}^{k-1} \pi_{k-\ell}\eta_\ell$$

where, if $\beta \neq \pm 1$,

(2.26) $$\pi_k = \gamma^{k-1} \frac{\sinh k\theta}{\sinh \theta} , \quad \theta = \cosh^{-1}\beta$$

and otherwise

(2.27) $$\pi_k = \pm k\gamma^{k-1} .$$

Proof The characteristic equation of (2.24) is

$$\xi^2 - 2\gamma\beta\xi + \gamma^2 = 0 .$$

The roots are

$$\gamma e^{\pm\theta}$$

for all β , where in defining θ by (2.26) we use the same branch of $\cosh^{-1}$ used by Manteuffel, so that $\mathcal{R}e\ \theta \geq 0$, $0 \leq \mathcal{I}m\ \theta < 2\pi$. The result follows from the standard theory of difference equations. □

We now apply Lemma 1 to each of the equations in (2.22). The characteristic equations are

$$\xi^2 - 2\mu\sigma_j\xi + 1 = 0 .$$

Assuming for convenience that $\mu\sigma_j \neq \pm 1$, we see that the solution to (2.22) is

$$f_k = D_k f_1 - D_{k-1} f_0 + 2\mu \sum_{\ell=1}^{k-1} D_{k-\ell} g_\ell$$

where D_k is a diagonal matrix with entries

$$\frac{\sinh k\theta_j}{\sinh \theta_j}$$

and

$$\theta_j = \cosh^{-1}(\mu\sigma_j) .$$

(If $\mu\sigma_j = \pm 1$, (2.27) must be used in place of (2.26).) Using (2.23) we therefore have

$$f_k = (\mu D_k \Sigma - D_{k-1}) f_0 + 2\mu \sum_{\ell=0}^{k-1} D_{k-\ell} g_\ell .$$

Now consider the j^{th} entry of the diagonal matrix $\mu D_k \Sigma - D_{k-1}$. Using the identity

(2.28) $$\frac{\sinh k\theta}{\sinh \theta} \cosh \theta - \frac{\sinh(k-1)\theta}{\sinh \theta} = \cosh k\theta$$

we see that it simplifies to the value

$$\cosh k\theta_j .$$

The standard convergence result for the exact Chebyshev method immediately follows. In the exact case, all the inhomogeneous terms vanish, and we see that

$$\|f_k\| \leq \max_j |\cosh k\theta_j| .$$

which gives the error bound

(2.29) $$\|V^{-1}e_k\| \le \max_j \frac{|\cosh k\theta_j|}{|\cosh k\psi|}$$

where

$$\psi = \cosh^{-1}\mu$$

using a standard property of Chebyshev polynomials.

In order to continue the analysis for the inexact case, we use the facts that

$$|\cosh k\theta_j| \le \rho^k \ , \quad \frac{\sinh k\theta_j}{|\sinh \theta_j|} \le k\rho^{k-1}$$

where

(2.30) $$\rho = \max_j |e^{\theta_j}| .$$

to obtain

(2.31) $$\|f_k\| \le \rho^k\|f_0\| + 2\ |\mu| \sum_{\ell=0}^{k-1} (k-\ell)\rho^{k-\ell-1}\|g_\ell\| .$$

To proceed further we must relate $\|g_\ell\|$ to $\|f_\ell\|$.

LEMMA 2 *We have*

$$\|g_\ell\| \le \epsilon\|f_\ell\|$$

where

(2.32) $$\epsilon = \delta\alpha\|V^{-1}M^{-1}\| \|AV\| .$$

Proof The proof is a straightforward consequence of (2.21). (2.16). (2.13) and (2.6). ❑

Continuing the error analysis, we can substitute $\epsilon\|f_\ell\|$ for $\|g_\ell\|$ in (2.31) and rewrite it to obtain

$$\|f_k\| - 2\epsilon|\mu| \sum_{\ell=1}^{k-1} (k-\ell)\rho^{k-\ell-1}\|f_\ell\| \le (\rho^k+2\epsilon|\mu|k\rho^{k-1})\|f_0\| \ , \ k = 1,2,\ldots,m$$

where m is introduced to indicate the last computed iterate x_m . This system of m linear inequalities can be written using matrix notation as

$$(I-EL)\begin{bmatrix}\|f_1\| \\ \vdots \\ \|f_n\|\end{bmatrix} \le s$$

where the non-negative matrix L and the vector s are defined accordingly. Now $L^m = 0$ so

$$(I-\epsilon L)^{-1} = I + \epsilon L + \epsilon^2L^2 + \ldots + \epsilon^{m-1}L^{m-1} \ .$$

Thus $(I-\epsilon L)^{-1}$ is a non-negative matrix and

$$\begin{bmatrix}\|f_1\| \\ \vdots \\ \|f_n\|\end{bmatrix} \le (I-\epsilon L))^{-1}s \ .$$

Let us define $t = [\tau_1,\ldots,\tau_m]^T = (I-\epsilon L)^{-1}s$. Let $\tau_0 = \|f_0\|$. By definition, τ_k satisfies

'(2.33) $$\tau_k = \rho^k\tau_0 + 2\epsilon|\mu|\sum_{\ell=0}^{k-1} (k-\ell)\rho^{k-\ell-1}\tau_\ell \ .$$

We therefore have

$$\tau_{k+1} - \rho\tau_k = 2\epsilon|\mu| \sum_{\ell=0}^{k} \rho^{k-\ell}\tau_\ell$$

$$\rho\tau_k - \rho^2\tau_{k-1} = 2\epsilon|\mu| \sum_{\ell=0}^{k-1} \rho^{k-\ell}\tau_\ell$$

and so

$$(2.34) \qquad \tau_{k+1} = 2(\rho+\epsilon|\mu|)\tau_k - \rho^2\tau_{k-1}\,, \quad k = 1,2,\ldots .$$

Given $\tau_0 = \|f_0\|$, and using

$$\tau_1 = (\rho+2\epsilon|\mu|)\tau_0$$

(which follows from (2.33)). equation (2.34) completely defines τ_k , the error bound on $\|f_k\|$. This homogeneous difference equation may now be solved using Lemma 1 again. We obtain

$$(2.35) \qquad \tau_k = \left[\rho^{k-1}\frac{\sinh k\varphi}{\sinh\varphi}(\rho+2\epsilon|\mu|) - \rho^k\frac{\sinh(k-1)\varphi}{\sinh\varphi}\right]\tau_0$$

where

$$(2.36) \qquad Q = \cosh^{-1}\left(1 + \frac{\epsilon|\mu|}{\rho}\right).$$

Using (2.28) again. (2.35) simplifies to become

$$\tau_k = \rho^k\left[\cosh k\varphi + \frac{k\varphi\mu|}{p}\frac{\sinh}{\sinh Q}\right|\tau_0 \ .$$

Finally, we divide through by c_k to obtain:

THEOREM 1 Assume that $M^{-1}A$ is *diagonalizable with spectrum* $\{\lambda_j\}$. *The error of the inexact Chebyshev* method is bounded by

$$(2.37) \qquad \|V^{-1}e_k\| \leq \frac{\rho^k\left[\cosh k\varphi + \frac{\epsilon|\mu|}{\rho}\frac{\sinh k\varphi}{\sinh\varphi}\right]}{|\cosh k\psi|}\|V^{-1}e_0\|\,,$$

where $\rho = \max_{j}|\exp(\cosh^{-1}\mu\sigma_j)|$, $\psi = \cosh^{-1}\mu$, $\varphi = \cosh^{-1}(1 + \frac{\epsilon\mu}{\rho})$, *and* ϵ , μ , σ_j *and* V *are* given by (2.32). (2.8). (2.20) *and* (2.18). □

Note that the final error bound has three factors multiplying onto $\|V^{-1}e_0\|$, two in the numerator and one in the denominator. One of them depends on ϵ , which measures the degree of inexactness of the inner iterations. The other two are essentially the same as those in the standard error bound (2.29) for the exact method. since ρ^k is an upper bound for $|\cosh k\theta_j|$. We could simplify the formula for ρ to

$$\rho = \max_j \left|\mu\sigma_j + \sqrt{\mu^2\sigma_j^2-1}\right|$$

but note, as Manteuffel (1977, p.312) points out, that care must be taken in choosing the branch of the square root to conform with the definition of $\cosh^{-1}$ used in Lemma 1.

In the case of the exact method, the standard error bound (2.29) converges to zero if and only if the spectrum $\{\lambda_j\}$ lies in an ellipse with foci ℓ and u which is strictly in the right half plane. Fur thermore, the method is asymptotically optimal over any such ellipse (Manteuffel (1977, p.315)).

3 . THE SYMMETRIC VERSION OF THE INEXACT CHEBYSHEV METHOD

Suppose now that A and M are symmetric and $M^{-1}A$ is positive definite, so that the spectrum $\{\lambda_j\}$ lies on the positive part of the real line. The parameters ℓ and u now become estimates of the end points of the line segment containing $\{\lambda_j\}$. Let us first review the well known results for the exact method.. The error bound (2.29) shows that convergence takes place if and only if $|\sigma_j| < 1$, i.e. $0 < \lambda_j < \ell + u$. This is consistent with the discussion for ellipses at the end of the previous section, since such a spectrum always lies inside an ellipse with the required properties. The method is optimal over the line segment $[\ell,u]$. Now assume Al $\leq . . . \leq \lambda_n$ and, for convenience. that $\sigma_1 \geq |\sigma_j|$, $j = 1,...,n$. If theparameters ℓ and

U overestimate the spectrum, i.e. $\ell < \lambda_1$, $\lambda_n < u$, then the numerator of the standard error bound (2.29) is less that one and may be written

$$\cos k \cos^{-1}(\mu\sigma_1) \tag{3.1}$$

since $\mu\sigma_1 < 1$ and θ_1 is imaginary. The optimal choice of the parameters is well known to be $\ell = \lambda_1$, $u = \lambda_n$, which gives a numerator of one in (2.29).

Now consider the inexact method. The analysis of the previous section simplifies slightly, since the matrix V of (2.18) may be written

$$V = M^{-1/2}\tilde{V} \tag{3.2}$$

where $\tilde{V}$ is orthogonal. The quantity ϵ in (2.32) simplifies accordingly. Otherwise there are no changes, and we obtain:

THEOREM 2 Assume M *and* A are symmetric and $M^{-1}A$ is posi*tive definite*. *The* error of the *inexact Chebyshev* method *is* bounded by

$$\|M^{1/2}e_k\| \le B_E B_I \|M^{1/2}e_0\| \tag{3.3}$$

where

$$\begin{aligned}
B_E &= \frac{\rho^k}{\cosh k\psi} \\
B_I &= \cosh k\varphi + \frac{\epsilon\mu}{\rho}\,\frac{\sinh k\varphi}{\sinh \varphi} \\
\rho &= |\mu\sigma_1 + \sqrt{\mu^2\sigma_1^2-1}\ | \\
\psi &= \cosh^{-1}\mu \\
\varphi &= \cosh^{-1}\left(1 + \frac{\epsilon\mu}{\rho}\right) \\
\epsilon &= \delta\alpha\|M^{-1/2}\|\ \|AM^{-1/2}\|
\end{aligned} \tag{3.4}$$

and μ , α , σ_1 , are given by (2.8). (2.7) *and* (2.20). *with* $\sigma_1 \geq |\sigma_j|$. □

(Note that we have introduced notation B_E and BI for the exact and inexact factors in the error bound, and we have simplified the formulas for ρ and ϵ .)

Now let us assume that $\lambda_1 + \lambda_n$ is known and that $\ell + u = \lambda_1 + \lambda_n$. Since the centre of the spectrum is fixed by this assumption, *a* in (2.7) has the correct value and so al = $-\sigma_n$. With a fixed, the iteration depends only on the parameters μ and δ . Note that in the limit, as $\mu \rightarrow \infty$, the Chebyshev method becomes the first-order Richardson method, which is closely related to the Gauss-Seidel method. We have already mentioned that in the exact case, i.e. when $\delta = 0$, the optimal value is given by $\mu\sigma_1 = 1$, i.e. $\mu = 1/\sigma_1$. It is now quite instructive to examine the error bound (3.3) to see how to choose the optimal value of μ if $\delta > 0$.

As a first step, let us review how the standard optimal choice $\mu = 1/\sigma_1$ may be explained when $\delta = 0$. Recall that

$$\cosh^{-1}\mu = \ell n\left[\mu + \sqrt{\mu^2-1}\right]$$

and so asymptotically

$$\cosh k\psi \approx \frac{1}{2}\left[\mu + \sqrt{\mu^2-1}\right]^k$$

as long as $\mu > 1$. (Certainly we may assume $\mu > 1$ since convergence is not obtained when $a_1 \geq 1$). The factor which determines convergence of the exact error bound B_E is therefore

$$(3.5) \qquad \frac{\left|\mu\sigma_1+\sqrt{\mu^2\sigma_1^2-1}\right|}{\mu+\sqrt{\mu^2-1}} .$$

Now this quantity has the value

$$\left|\frac{\sigma_1}{1+\sqrt{1-\sigma_1^2}}\right] \qquad (3.6)$$

if $\mu\sigma_1 = 1$ and a larger value if $\mu\sigma_1 > 1$, with limiting value al as $\mu \to \infty$. Therefore $\mu = 1/\sigma_1$ is preferable to a larger value for μ . Unfortunately we cannot draw any conclusion about smaller values of μ , which are obtained when ℓ and u overestimate the interval $[\lambda_1,\lambda_n]$, from simply looking at B_E . This is because $\rho = 1$ even when $\mu <$ $1/\sigma_1$. and so B_E does not reflect any possible advantage gained by overestimating. In order to conclude that overestimating is not advantageous , one needs to look at the factor (3.1) which we had to bound above by ρ^k in the inexact analysis.

Now let us suppose that $\delta > 0$. The factor B_E does not change, but now we need to consider BI . Let us neglect the second term in B_I which has a second order effect. We see that the factor which controls convergence of the main term in BI is

$$1 + \frac{\epsilon\mu}{\rho} + \sqrt{\left(1+\frac{\epsilon\mu}{\rho}\right)^2-1} \ . \qquad (3.7)$$

Now μ/ρ is given by

$$\frac{\mu}{\rho} = \frac{\mu}{|\mu\sigma_1+\sqrt{\mu^2\sigma_1^2-1}\,|} \ . \qquad (3.8)$$

This quantity has the value $1/\sigma_1$ when $\mu\sigma_1 = 1$, and a smaller value for larger μ , with limiting value $1/(2\sigma_1)$ as $\mu \to \infty$. We must now combine the effects of (3.5) and (3.7) to draw our conclusions. There are essentially four cases.

Case 1: $\sigma_1 << 1$, $\epsilon << 1$. This is the standard exact case. The bound (3.5) increases by up to a factor of two as $\mu \rightarrow \infty$.

Case 2: $\sigma_1 \approx 1$, $\epsilon << 1$. In this case there is no great advantage in having $\mu = 1/\sigma_1$. There is neither a decrease nor a significant increase in the bound as $\mu \rightarrow \infty$. Of course the convergence is very slow.

Case 3: $\sigma_1 << 1$, $\epsilon \approx 1$. In this case the advantage of having $\mu = 1/\sigma_1$ in (3.5) is cancelled by the advantage of a larger value of μ in (3.7). (3.8). In fact, in the limiting case both factors have the value two and cancel. Thus there is no disadvantage to underestimating the eigenvalues if the iteration is quite inexact.

Case 4: $\sigma_1 \approx 1$, $\epsilon \approx 1$. In this case the advantage of $\mu = 1/\sigma_1$ in the exact case disappears and we are left with an overall advantageas $\mu \rightarrow \infty$. In other words, although an iteration of this type will certainly be slow (if it converges at all), it can be speeded up by underestimating the eigenvalues!

The predictions made in these four cases are substantiated by numerical experiments which we now report. Let M and N be symmetric matrices of order 20, with $M = \mathrm{Diag}(\gamma)$, where γ is a real number to be specified. Let N and the right-hand side b be randomly generated as specified by Appendix A. Table I reports the results of using the inexact Chebyshev method to solve (2.1) with δ = (0.0, 0.9) and μ^{-1} = (1.0, σ_1 and 0.001) , where al depends on γ . The table reports results for $\gamma = 10$ (σ_1=0.4017, $(\lambda_1+\lambda_n)/2$ = 0.85222) , $\gamma = 6$ (σ_1=0.757, $(\lambda_1+\lambda_n)/2$ = 0.75370) and $\gamma = 5$ (σ_1=0.972, $(\lambda_1+\lambda_n)/2$ = 0.70444), respectively. In all cases a was set to the exact value $2/(\lambda_1+\lambda_n)$. The inner iterations were simulated by perturbing the residuals r_k of the outer iteration,' as specified in Appendix A, and then solving (2.12) exactly. The initial value x_0 was set to zero. The quantity reported is $\|r_{20}\|$, the norm of the residual of the outer iteration after 20 steps.

TABLE1

(SYMMETRIC INEXACT CHEBYSHEVITERATION)

$\|r_{20}\|$

δ	γ=10 (σ_1=.4017)			γ=6 (σ_1=.757)			γ=5 (σ_1=.972)		
	μ-1.			μ^{-1}:			μ^{-1}:		
	1.0	.4017	.001	1.0	.757	.001	1.0	.972	.001
0.0	3.2E-2	1.5E-12	1.2E-8(1)	8.8E-2	6.5E-6	5.8E-3	3.1E-1	1.2E-1	8.5E-1(2)
0.9	7.6E+1	3.3E-3	1.7E-3(3)	7.9E+3	1.1E+1	7.3E-1	1.7E+5	2.4E+4	2.0E+1(4)

Note: (1), (2), (3), (4) indicate Cases 1. 2, 3, 4 respectively.

The results clearly support the predictions made for the behaviour of the error bound as $\mu \to \infty$ in the four cases. In particular, for $\gamma = 5$ and $\delta = 0.9$, the choice $\mu^{-1} = 0.001$ instead of $\mu^{-1} = \sigma_1$ changes a divergent iteration to one which is (barely) convergent. The table include results for $\mu = 1$ for completeness. We have already noted that no conclusion about the value of $\mu^{-1} > \sigma_1$, i.e. overestimating the eigenvalues, can be made by analysis of the bound (3.3).

So far our attention has been restricted to the optimal choice of parameters with respect to convergence of the outer iteration. However, our real interest should be to choose optimal values for δ and μ which minimize a measure of the total amount of work, perhaps the total number of inner iterates. We defer this topic to Section 6.

4. THE SKEW-SYMMETRIC VERSION OF THE INEXACT CHEBYSHEV METHOD

A reasonable method for solving $Ax = b$ when A is a nonsymmetric matrix with positive definite symmetric part is to split A into its symmetric and skew-symmetric parts, i.e. let

$$(4.1) \qquad M = \tfrac{1}{2}(A^T+A) \ , \ N = \tfrac{1}{2}(A^T-A) \ .$$

Then the eigenvalues of $M^{-1}N$ are imaginary and $\{\lambda_j\}$, the eigenvalues of $M^{-1}A$, occur in complex conjugate pairs with real part equal to one. We therefore set ℓ and u to a complex conjugate pair with real part equal to one. Then $a = 1$ in (2.7), and the quantities $\{\sigma_j\}$ and μ are imaginary. Assume $|\sigma_1| \geq |\sigma_j|$. $j = 1,\dots,n$.

Unlike the symmetric case, convergence of the exact method may be obtained for arbitrarily large $|\sigma_1|$, if $|\mu|$ is small enough. This reflects the fact that a is fixed at $a = 1$, while in the symmetric case a is used to scale the spectrum. The optimal value of μ in the exact case is well known to be $\mu = 1/\sigma_1$. By optimal, we mean asymptotically optimal, in the same sense used in the previous section. Although the exact symmetric Chebyshev method has an optimality property which may be stated for each iteration k , this is not true in the skew-symmetric case (see Manteuffel (1977, p.314)). Freund and Ruscheweyh (1986) give a discussion of optimal methods for a spectrum contained in a line segment parallel to the imaginary axis.

Now consider the inexact skew-symmetric method. We have (3.2). where $\tilde{V}$ is unitary. Theorem 2 therefore holds unchanged for this case. However, we must be careful with the notation. Since $\mu\sigma_1$ is negative, the correct branch to use for the square root in the definition of ρ in (3.4) is the negative branch. Similarly, asymptotically

$$(4.2)\qquad |\cosh k\psi| \approx \frac{1}{2}\left|\mu + \sqrt{\mu^2-1}\right|^k$$

where the branch of the square root is chosen to have the same imaginary sign as μ . Consequently the right-hand side of (4.2) is

$$\frac{1}{2}\left[|\mu| + \sqrt{|\mu|^2+1}\,\right]$$

using the normal positive square root. Thus the factor which determines convergence of the exact error bound B_E is

$$(4.3)\qquad \frac{|\mu\sigma_1| + \sqrt{|\mu\sigma_1|^2-1}}{|\mu| + \sqrt{|\mu|^2+1}} \;.$$

This quantity has the value

$$\frac{|\sigma_1|}{1 + \sqrt{1+.|\sigma_I|^2}}$$

if $|\mu\sigma_1| = 1$ and a larger value if $|\mu\sigma_1| > 1$, with limiting value $|\sigma_1|$ as $|\mu| \to \infty$. Therefore, as in the symmetric case, $|\mu\sigma_1| = 1$ is preferable to a larger value for $|\mu|$. As before, we cannot draw any, conclusion about smaller values of $|\mu|$.

Now let us suppose that $\delta > 0$ and consider the factor B_I . As before, the factor which controls convergence of the main term in B_I is (3.7). As in the symmetric case, the quantity $|\mu|/\rho$ has the value 1/ la, I for $|\mu\sigma_1| = 1$ and a limiting value of $1/(2|\sigma_1|)$ as $|\mu| \to \infty$. We now combine the effects of B_E and B_I . Again there are four cases, but the conclusions are quire different from those drawn in the symmetric case.

Case 1: la, I $<< 1$. $\epsilon << 1$. This is the standard exact case. The bound (4.3) increases by up to a factor of two as $|\mu| \to \infty$.

Case 2: $|a, | >> 1$, $\epsilon << 1$. In this case letting $|\mu| \to \infty$ is a disaster. The bound (4.3) increases by up to a factor of $|\sigma_1|$.

Case 3: $|a, | << 1$, $\epsilon \approx 1$. As in Case 3 for the symmetric iteration, factors of two cancel in (4.3) and (3.7). (3.8). There is neither an advantage nor a disadvantage in letting $|\mu| \to \infty$.

Case 4: $|a, | >> 1$, $\epsilon \approx 1$. This is completely different from Case 4 for the symmetric iteration. As in Case 2 (above), letting $|\mu| \to \infty$ is a disaster. Note that the degree of inexactness of the iteration is almost irrelevant.

Again, these predictions are verified by numerical experimentation. Consider the same randomly generated problem as before, where we now replace the lower triangle of N by its negative, so that N is skew-symmetric. Table II reports the results of using the inexact Chebyshev method to solve (2.1) with δ = (0.0, 0.9) and $|\mu^{-1}|$ = (10/γ, $|\sigma_1|$ and .001) , where $|\sigma_1$ I depends on γ . The table reports results for $\gamma = 10$ (σ_1=0.3676i) , $\gamma = 1$ (a_1=3.676i) and $\gamma = 0.01$ (σ_1=367.6i) . As already noted, $a = 1$ in all cases. It is straightforward to implement the method using only real arithmetic. The inner iterations were simulated by perturbing r_k as before. Again, x_0 was set to zero, and the quantity reported is $\|r_{20}\|$.

Again, the results clearly support the predictions made. As in the symmetric case, letting $|\mu| \to \infty$ is harmless if δ is large and $|\sigma_1|$ is small. Unlike in the symmetric case, letting $|\mu| \to \infty$ is detrimental to convergence if $|\sigma_1|$ is moderate or large. We have included results for small $|\mu|$ for completeness. Since underestimating the eigenvalues

($|\mu| \rightarrow \infty$) has such negative consequences, overestimating may be a better choice in general. However, see the experimental results in Section 7.

TABLE II

(SKEW-SYMMETRIC INEXACT CHEBYSHEV ITERATION)

$\|r_{20}\|$

δ	$\gamma=10$ ($\lvert\sigma_1\rvert=0.367$)			$\gamma=1$ ($\lvert\sigma_1\rvert=3.67$)			$\gamma=.01$ ($\lvert\sigma_1\rvert=367.6$)		
	$\lvert\mu^{-1}\rvert$:			$\lvert\mu^{-1}\rvert$:			$\lvert\mu^{-1}\rvert$:		
	1.0	.367	.001	10.0	3.67	.001	1000	367.6	,001
0.0	5.2E-8	4.1E-15	2.1E-9(1)	3.2E-1	1.9E-2	2.1E+11	1.8E+0	1.5E+0	OF(2)
0.9	4.8E-6	4.6E-6	1.1E-5(3)	5.2E-1	1.5E+0	2.7E+11	1.9E+0	1.0E+2	OF(4)

Notes: OF indicates overflow. (1), (2), (3), (4) indicate Cases 1, 2. 3, 4 respectively.

5. THE SYMMETRIC INEXACT SECOND-ORDER RICHARDSON METHOD

The second-order Richardson method is obtained from the Chebyshev method by a single change: replace (2.9) by

$$\omega_{k+1} = \omega = \frac{2}{1+\sqrt{1-\mu^{-2}}} \tag{5.1}$$

so that ω_{k+1} is a fixed quantity determined by μ . It is well known that the two methods are identical asymptotically (see Golub and Varga(1961')). Formula (5.1) is well known from the theory of SOR (Young(1971)). see Golub and Varga (1961, p.151) for an explanation of the close connection between the Richardson and SOR methods.

Because the Richardson analysisis somewhat more complicated than the Chebyshev analysis, we restrict ourselves to the symmetric positive definite case where ℓ and u overestimate (or exactly estimate) the spectrum $\{\lambda_j\}$. In the symmetric case, convergence of the exact Richardson method is obtained if al < 1 , assuming al $\geq |\sigma_j'|$, $j = 1,\ldots,n$. Since μ^{-1} is an estimate of al , assume $\mu > 1$. Then $1 \leq \omega < 2$. The assumption that ℓ and u overestimate the spectrum means that $\mu^{-1} > \sigma_1$, i.e. that ω in (5.1) is greater than its optimal value.

We now show how the analysis given in Section 2 changes for the inexact Richardson method. Equations (2.14) to (2.20) are unchanged. Replace (2.21) by

$$f_k = V^{-1}e_k \ , \ g_k = V^{-1}p_k \ , \ \text{except} \ g_0 = \frac{1}{\omega}V^{-1}p_0 \ ,$$

recalling that (3.2) holds. Instead of (2.22) we get

$$f_{k+1} = \omega\Sigma f_k + (1-\omega)f_{k-1} + \omega g_k \ , \ k = 1,2,\ldots \tag{5.2}$$

and

$$f_1 = \Sigma f_0 + \omega g_0 \ . \tag{5.3}$$

Applying Lemma 1, we see that the solution to this diagonalized system of difference equations is

$$f_k = D_k e_1 - \rho^2 D_{k-1} e_0 + \omega \sum_{\ell=1}^{k-1} D_{k-\ell} g_\ell \tag{5.4}$$

where

$$\rho = \sqrt{\omega-1} \tag{5.5}$$

and D_k is a diagonal matrix with entries

$$(5.6) \qquad \rho^{k-1}\frac{\sin k\theta_j}{\sin\theta_j}$$

where

$$(5.7) \qquad \theta_j = \cos^{-1}\left[\frac{\omega\sigma_j}{2\rho}\right] .$$

Note that the sines, rather than hyperbolic sines, appear because $|\omega\sigma_j/2\rho| < 1$, a fact which follows from the assumption that the spectrum is overestimated. If the spectrum is estimated exactly, there is a double root in the characteristic equation of (5.2). and (5.6) must be replaced using (2.27) in Lemma 1.

using (5.3). we see that (5.4) becomes

$$(5.8) \qquad f_k = (D_k\Sigma-\rho^2 D_{k-1})f_0 + \omega\sum_{\ell=0}^{k-1} D_{k-\ell}g_\ell \cdot$$

Now consider the j th entry of the diagonal matrix $D_k\Sigma - \rho^2 D_{k-1}$. It is given by

$$(5.9) \quad \rho^{k-1}\frac{\sin k\theta_j}{\sin\theta_j}\sigma_j - \rho^k\frac{\sin(k-1)\theta_j}{\sin\theta_j} = \rho^k\left[\frac{\sin k\theta_j}{\sin\theta_j}\frac{2}{\omega}\cos\theta_j - \frac{\sin(k-1)\theta_j}{\sin\theta_j}\right].$$

Using the identity (2.29) with cosh, sinh replaced by cos, sin, together with the inequality

$$\frac{\sin k\theta}{\sin\theta} \le k$$

(5.9) reduces to

$$\rho^k\left[\cos k\theta_j + (\frac{2}{\omega} - 1)\frac{\sin k\theta_j}{\sin\theta_j}\cos\theta_j\right] \le \rho^k\left[1 + k\frac{1-\rho^2}{1+\rho^2}\right]$$

(since $\omega = 1+\rho^2$ by (5.5)). The convergence result for the exact Richardson method now follows. In the exact case. all the inhomogeneous terms in (5.8) vanish, and we see that

$$\|M^{1/2}e_k\| \leq \rho^k\left[1 + k\,\frac{1-\rho^2}{1+\rho^2}\right]\|M^{1/2}e_0\|$$

which is the bound given by Golub (1959, p.23).

For the analysis of the inexact case we need now to relate $\|g_\ell\|$ to $\|f_\ell\|$. Lemma 2 applies, showing that

$$\|g_\ell\| \leq \epsilon\|f_\ell\|$$

where

$$\epsilon = \delta\alpha\|M^{-1/2}\|\ \|AM^{-1/2}\| \ . \tag{5.10}$$

We can therefore substitute $\epsilon\|f_\ell\|$ for $\|g_\ell\|$ in (5.8) to obtain

$$\|f_k\| \leq \rho^k\left[1 + k\,\frac{1-\rho^2}{1+\rho^2}\right]\|f_0\| + \omega\epsilon\sum_{\ell=0}^{k-1}(k-\ell)\rho^{k-\ell-1}\|f_\ell\| \ .$$

The same argument used in the Chebyshev analysis shows that, consequently,

$$\|f_k\| \leq \tau_k$$

where

$$\tau_k = \rho^k\left[1 + k\,\frac{1-\rho^2}{1+\rho^2}\right]\tau_0 + \omega\epsilon\sum_{\ell=0}^{k-1}(k-\ell)\rho^{k-\ell-1}\tau_\ell \ . \tag{5.11}$$

Using the same technique as before, we see that τ_k satisfies

$$\tau_{k+1} = (2\rho+\omega\epsilon)\tau_k - \rho^2\tau_{k-1}\ ,\ k = 1,2,\ldots\ . \tag{5.12}$$

We also have

$$\tau_1 = \left[\rho\left[1 + \frac{1-\rho^2}{1+\rho^2}\right] + \omega\epsilon\right]\tau_0 \tag{5.13}$$

which follows from (5.11). The homogeneous difference equation (5.12) may now be solved using Lemma 1 again. We obtain

$$\tau_k = \rho^{k-1}\,\frac{\sinh k\varphi}{\sinh\ \varphi}\,\tau_1 - \rho^k\,\frac{\sinh(k-1)\varphi}{\sinh\ \varphi}\,\tau_0 \tag{5.14}$$

where

$$\varphi = \cosh^{-1}\left(1 + \frac{\epsilon\omega}{2\rho}\right) . \tag{5.15}$$

Using (5.13), (5.14) becomes

$$\tau_k = \rho^k\left[\frac{\sinh k\varphi}{\sinh \varphi}\left(1 + \frac{1-\rho^2}{1+\rho^2} + \frac{\omega\epsilon}{\rho}\right) - \frac{\sinh(k-1)\varphi}{\sinh \varphi}\right]\tau_0 .$$

Using (2.28) we then get

$$\tau_k = \rho^k\left[\cosh k\varphi + \left(\frac{1-\rho^2}{1+\rho^2} + \frac{\omega\epsilon}{2\rho}\right)\frac{\sinh k\varphi}{\sinh \varphi}\right]\tau_0 . \tag{5.16}$$

This bound takes account of the cancellation that takes place because of the particular choice of xl . However, a cruder but simpler bound is obtained by directly studying the roots of the characteristic equation of (5.12). namely

$$\xi^2 - (2\rho+\omega\epsilon)\xi + \rho^2 = 0 .$$

Let $\tilde{\rho}$ be the larger root of this equation, i.e.

$$\tilde{\rho} = \rho e^{\varphi} = \rho + \frac{\omega\epsilon}{2} + \left[\rho\omega\epsilon + \frac{\omega^2\epsilon^2}{4}\right]^{1/2} . \tag{5.17}$$

The solution to (5.12) is then bounded by

$$\tau_k \leq k\tilde{\rho}^{k-1}\tau_1 + (k-1)\tilde{\rho}^k\tau_0 . \tag{5.18}$$

Summarizing these results we have

THEOREM 3 *The* symmetric inexact second-order Richardson method, assuming $\lambda_j \in [\ell,u]$, $j = 1,\ldots,m$, *converges if* $\tilde{\rho} < 1$, *where* $\tilde{\rho}$. a function of δ , a , ω , is defined by (5.17). (5.10). and (5.5). *More* specifically, $\|M^{1/2}e_k\| \le \tau_k$. *where* τ_k satisfies (5.18) for *any* choice of x_1 and (5.16) for the choice of xl *defined* by (2.3). □

The bound (5.16) consists of two factors. The first, ρ^k , corresponds to $1/\cosh k\psi$ in the denominator of the exact Chebyshev bound B_E . The second factor corresponds to B_I , the inexact Chebyshev factor. The numerator of B_E does not appear here, since we assumed the spectrum is overestimated (see (3.1)). There is no difficulty in extending the analysis to the case of underestimating the spectrum, or, indeed, the nonsymmetric version of the Richardson method as given by Niethammer and Varga (1883). Essentially the factor ρ^k appearing here is replaced by $\rho_1^k\rho_2^k$, where pl = $\rho < 1$ and $\rho_2 > 1$ is the quantity ρ appearing in the Chebyshev analysis, i.e. (3.4). It follows that the conclusions in Sections 3 and 4 regarding the benefits (or otherwise) of underestimating the spectrum in the symmetric and skew-symmetric cases apply also to the inexact Richardson method.

6 . ON REDUCING THE TOTAL NUMBER OF INNER ITERATES

Let m_k be the number of inner iterates required in the k^{th} inner iteration to achieve (2.12). (2.13). A reasonable measure of the total amount of work required to generate $x_1,\ldots,x_m$ is

$$w = \sum_{k=1}^{m} m_k .$$

This measure is particularly appropriate when each step of each inner iteration involves solving another system of equations by a direct method. Let τ be a measure of relative accuracy required for the outer iteration, e.g. specifying $\|M^{1/2}e_m\| \le \tau\|M^{1/2}e_0\|$ in the symmetric or skew-symmetric case. Theorem 2 and the subsequent remarks show that the error $\|M^{1/2}e_k\|$ is. reduced at each step of the outer iteration by approximately $\bar{\rho}$, where $\bar{\rho}$, a function of μ , a , δ and σ_1 , is the product of (3.5) and (3.7). Therefore the number of iterates required for the outer iteration is approximately

$$m \approx \frac{-\log \tau}{-\log \bar{\rho}(\mu,\alpha,\delta,\sigma_1)} .$$

Now let us suppose that the iterative method used for the inner iteration is such that the associated residual is reduced by about a factor of ζ at each step. Assuming a starting value $z = 0$ for each inner iteration we have

$$m_k \approx \frac{-\log \delta}{-\log \zeta} .$$

With these assumptions the total amount of work W is approximately

$$W \approx \frac{-\log \tau}{-\log \zeta} \cdot \frac{-\log \delta}{-\log \bar{\rho}(\mu,\alpha,\delta,\sigma_1)} .$$

Now τ and ζ are fixed so a reasonable goal in choosing the parameters μ , a and δ is to minimize

$$\bar{W} = \frac{-\log \delta}{-\log \bar{\rho}(\mu,\alpha,\delta,\sigma_1)} .$$

Naturally this is difficult to do, but the formula gives some insight. As $\delta \to 0$, $\bar{W} \to \infty$, indicating that solving the inner iterations too accurately is very expensive. On the other hand, if δ is too large, $\bar{\rho}$ may be nearly one or greater than one, indicating that the outer iteration is divergent. The optimal value of δ , given μ and a , is somewhere between these extremes. We have already.explained that μ plays an important role, which varies with δ , in the size of $\bar{\rho}$. Choosing optimal parameters is clearly complicated. However, substantial insight can be obtained from numerical experiments. Fortunately, once an adequate choice of parameters has been determined for a particular problem, whole classes of related problems can usually be solved efficiently with the same parameter values.

7. NUMERICAL EXPERIMENTS AND APPLICATIONS

Consider the differential equation

$$-Au + (au)_x + au_x + (bu)_Y + bu_y + cu = f \tag{7.1}$$

where u , a , b and f are functions on the unit square: $0 \lesssim x \lesssim 1$, $0 \lesssim y \lesssim 1$. Approximating (3.1) by finite differences on a grid with s interior points in each direction results in a linear system of order s^2 ,

$$(M-N)v = d . \tag{7.2}$$

Here v is the solution to the discretized problem, N is a skew-symmetric matrix corresponding to the first-order terms in (7.1), $M = M_1 - M_2$ is a symmetric matrix, with M_1 the negative discrete Laplacian operator and M_2 a diagonal matrix corresponding to the term cu , and d corresponds to f , modified to incorporate the boundary conditions. We consider the particular differential equation whose solution is $u(x,y) = \exp(x^2+y^2)$, with coefficients $a(x,y) = b(x,y) = 10 \exp(x^2+y^2)$, $c(x,y) = 20 \exp(3x^2+3y^2)$, and with nonhomogeneous Dirichlet conditions imposed on the boundary of the square. We used the mesh size $s = 15$. We have also performed experiments for various other problems. The one reported here was chosen so that both the outer and the inner systems are moderately difficult to solve. The results seem fairly typical.

Table III gives the results of solving (7.2) using the skew-symmetric version of the inexact Chebyshev method with various choices of μ and δ . (Recall that $a = 1$.) The initial value x_0 was set to zero, and the termination condition for the outer iteration was $\|r_m\| \lesssim 10^{-4}$. Each inner iteration was carried out using a (preconditioned) conjugate gradient method, starting with the zero vector, so that (2.12). (2.13) hold. The symmetric splitting $M = M_1 - M_2$ was exploited so that each step of each inner iteration used a fast direct method to solve a system of the form $M_1\bar{z} = \bar{r}$, i.e. the Poisson equation. The entries in the table are of the form "m/W" , i.e. they report the number of outer iterates and the total number of inner iterates.

In addition to the Chebyshev results, Table III gives results for solving (7.2) using a preconditioned skew-symmetric conjugate gradient method (see Concus and Golub (1976) and Widlund (1978)). The inner iterations were carried out exactly as described above, being terminated by (2.12). (2.13). The results are reported in the column headed CG. The conjugate gradient method was also run with the inner iterations carried out to full precision, in order to obtain $\sigma 1$ from the resulting tridiagonal Lanczos matrix. This run, made solely to compute la, I . is not included in the table'. It determined that $|\sigma_1| = 2.19$.

TABLE III

δ	CG	CHEBYSHEV						
					$\|\mu^{-1}\|$:			
		10.0	3.0	2.1*	2.0	1.5	1.0	.001
0.01	39/274	t	60/401	45/294	76/559	D	D	D
0.1	46/204	1991804	61/249	46/203	46/202	D	D	D
0.5	%	276/603	85/194	65/149	63/141	188/422	D	D
0.9	%	%	2991299	2191219	206/206	169/169	132/132	D

Reports # outer iterations/# inner iterations.

Notes: * The value of σ_1 is 2.101 .

† Number of inner iterates exceeded 1000.

% Number of outer iterates exceeded 500.

D Outer iteration' diverges.

A careful inspection of Table III reveals a number of interesting features :

1. Consider the column $|\mu^{-1}| = |\sigma_1| = 2.1$. The number of outer interates increases as δ increases, as expected. The total number of inner iterates, however, reaches a minimum at about $\delta = 0.5$. Clearly, a small value of δ is inefficient: also, if δ is close to one, only one step is taken in each inner iteration, which is overall less effective than using $\delta = 0.5$.

2. Consider the Chebyshev results as μ varies. For $\delta = 0.01$, there is a rapid deterioration as $|\mu|$ grows greater than $|\sigma_1^{-1}$ I . For larger δ , the deterioration is not as rapid and for large

enough δ there is a noticeable *decrease* in the number of outer iterates before a minimum is reached and a sharp increase occurs as $|\mu| \to \infty$. It may be possible to partially explain this from the results of Section 4 since the factor (3.7) may, as $|\mu|$ increases past $|\sigma_1^{-1}|$, decrease faster than (4.3) increases initially. As soon as (4.3) grows enough. divergence occurs. However, see the next remark.

3. There is one feature of Table III which cannot be explained from the results of Section 4. When $|\mu|$ is fixed greater than $|\sigma_1^{-1}|$ there is a distinct drop in the number of *outer* iterates as δ increases larger than 0.01 . (Indeed, as a consequence the lowest value for the total number of inner iterates in the entire table is for $|\mu| = 1$. $\delta = 0.9$.) Since B_E is fixed and BI increases as δ increases, this result is not reflected by the bound (3.3). This does not imply that the bound is not sharp, however. The reason for this behaviour seems to depend on the choice of method used for the inner iterations. When the inner iterations are simulated as described in Appendix A , the phenomenon disappears.

4. For any fixed δ , as $|\mu|$ decreases below $|\sigma_1^{-1}|$, the number of outer iterates increases but it does not do so very fast.

5. The conjugate gradient method is not as robust, with respect to δ , as the Chebyshev method with a reasonable value for μ .

We also tested the Richardson method on the same problem. The results were very similar. All results were obtained using double precision on a VAX/780 at Australian National University.

Using the symmetric/skew-symmetric splitting to solve nonsymmetric systems is only one possible application of the inexact method. Another possible application involves using the Arrow-Hurwicz-Uzawa method (see Glowinski, Lions & Tremolières (1976, p.96)) to solve systems of the form

$$A = \begin{bmatrix} C & B \\ -B^T & 0 \end{bmatrix} , \quad b = \begin{bmatrix} b^{(1)} \\ b^{(2)} \end{bmatrix}$$

where C is positive definite. Such systems arise frequently in the solution of Stokes equations or variational problems with constraints. The main idea of this method is to use the splitting

$$M = \begin{bmatrix} C & 0 \\ -B^T & \kappa I \end{bmatrix} , \quad N = \begin{bmatrix} 0 & -B \\ 0 & \kappa I \end{bmatrix}$$

for some $\kappa \neq 0$. An exact version therefore requires the solution of a system of equations $Cz^{(1)} = r^{(1)}$; an inexact implementation allows an approximate solution of this equation. Note that

$$M^{-1}N = \begin{bmatrix} 0 & -C^{-1}B \\ 0 & I-\kappa^{-1}B^TC^{-1}B \end{bmatrix} .$$

It follows that the eigenvalues of $M^{-1}N$ are real. The choice of κ depends on the spectrum of $B^TC^{-1}B$. Let β_1 and β_2 be respectively the largest and smallest eigenvalues of $B^TC^{-1}B$. Then κ should be chosen by

$$\kappa = \frac{\beta_1+\beta_2}{2}$$

so that the spectral radius of $M^{-1}N$ is

$$\frac{\beta_1-\beta_2}{\beta_1+\beta_2}$$

and $a = 1$ in (2.7). Other possible applications of the inexact iteration include domain decomposition.

8. CONCLUSIONS.

We have given a convergence analysis for the inexact Chebyshev and Richardson methods. We showed that in the case of the symmetric iteration, underestimating the eigenvalues speeds up a very inexact iteration if the spectral radius is large. This is not generally true in the case of the-skew-symmetric iteration. We have presented experimental results which indicate that the Chebyshev & Richardson methods, with reasonable parameter choices, are less sensitive than the conjugate gradient method to the degree of inexactness of the iteration. Finally we note that as parallel computing increases in importance. the Chebyshev and Richardson methods become more attractive than the conjugate gradient method since they do not require inner products and are therefore fully parallelizable.

APPENDIX A

So that others may reproduce the experiments, we give details of the randomly generated problem discussed in Section 3. We define N to be a symmetric matrix with a zero diagonal. We used the NAG library random number generator, initializing it by calling G05CBF with seed value 3.0. Subsequently we called G05CAF to obtain, in order, $N_{1,2}$, $N_{1,3}, \ldots, N_{1,20}$, $N_{2,3}, \ldots N_{19,20}$, $b_1, \ldots, b_{20}$, $\delta_{1,1}$, $\delta_{1,2}, \ldots, \delta_{1,20}$.

$\delta_{2,1},\ldots$. The values $\delta_{k,i}$ were used to obtain q_k satisfying (2.12). (2.13) by

$$(q_k)_i = \frac{2\|r_k\|}{\sqrt{20}}\left(\delta_{k,i} - \frac{1}{2}\right) .$$

ACKNOWLEDGEMENTS

The authors would like to thank Max Dryja and Olof Widlund for stimulating them to complete this, work. The second author would also like to thank his hosts at Australian National University for their warm hospitality. .