SEVERAL OBSERVATIONS ON THE USE OF CONJUGATE GRADIENT METHODS

by

A. K. Cline

March, 1978

CNA-133
TR-76

This report is being released jointly with the Computer Sciences Department at The University of Texas at Austin under the number TR-76.

CENTER FOR NUMERICAL ANALYSIS
THE UNIVERSITY OF TEXAS AT AUSTIN

SEVERAL OBSERVATIONS ON THE USE OF CONJUGATE GRADIENT METHODS

by

A. K. Cline*

Abstract

The conjugate gradient algorithm for symmetric, positive definite systems should be viewed as an iterative algorithm (independent of its theoretical finite convergence). Its effectiveness for a particular problem depends upon the number of iterations required to achieve a certain accuracy. In this paper we discuss the theoretical bounds on convergence as well as experimental results for particular problem classes. We also compare the standard conjugate gradient algorithm with the minimum residual variant and consider the use of such algorithms for dense and banded systems.

Introduction

The conjugate gradient algorithm, commonly credited to Stiefel and Hestenes [3] and [6], can be an important tool for the solution of large, sparse, positive definite systems of linear equations. Let us denote such a system by

$$Ax = b,$$

and make several important observations which distinguish this method from some of the others used for the problem. First, the matrix A need only by provided as a procedure, i.e., given a vector z we must have the capability of computing Az, but how this is done does not affect the method. (This procedure for computing

---

*This work was performed under NASA Grant 47-102-001 while the author was in residence at the Institute for Computer Applications in Science and Engineering at NASA Langley Research Center in Hampton, Virginia.

Az may be very closely related to a physical basis for the problem, and a matrix A in sparse form or otherwise need not be stored.) Second, no additional parameters other than the right-hand side b and the procedure for A need be involved in the method, i.e., no relaxation factors or similar constants need be estimated. Third, the method is optimal in the sense that after k steps of the iteration the approximate solution x_k is "best" from among linear combinations of b, Ab, ..., $A^{k-1}b$. (This sense of "best" will be made more precise later.)

In this paper we seek to discuss several questions related to the use of the conjugate gradient algorithm. In section 1, we define the standard algorithm as well as the minimum residual variant, then discuss their convergence properties. We give bounds on the convergence and consider the likelihood of the bounds being attained. In section 2, we compare the two algorithms theoretically and experimentally. In section 3, the question of the affect on convergence of the eigenspectrum of A having isolated extreme values is considered. Finally, in section 4, we examine the use of the conjugate gradient method to solve dense or banded systems.

1. The Algorithms and Rates of Convergence

The standard conjugate gradient algorithm for $Ax = b$ can be expressed as an iteration:

$$\Delta x_0 = 0 \ ;$$

$$x_0 = 0 \ ;$$

$$\Delta r_0 = 0 \ ;$$

$$r_0 = b \ .$$

For $k = 1,2,\ldots$

$$e_k = \begin{cases} 0 & k = 1 \\ q_{k-1}\,\dfrac{r_{k-1}^T r_{k-1}}{r_{k-2}^T r_{k-2}} & ; \end{cases}$$

$$q_k = \frac{r_{k-1}^T A r_{k-1}}{r_{k-1}^T r_{k-1}} - e_k \; ;$$

$$\Delta x_k = \frac{1}{q_k}\,(e_k \Delta x_{k-1} + r_{k-1}) \quad ;$$

$$x_k = x_{k-1} + \Delta x_k \quad ;$$

$$\Delta r_k = \frac{1}{q_k}\,(e_k \Delta r_{k-1} - A r_{k-1}) \quad ;$$

$$r_k = r_{k-1} + \Delta r_k \; .$$

The iteration is terminated when the residual norm, $\|r_k\| = (r_k^T r_k)^{1/2}$, is sufficiently small. If an initial approximation $\hat{x}$ other than zero is known then it can be used by substituting "$x_0 = \hat{x}$" and "$r_0 = b - A\hat{x}$" for "$x_0 = 0$" and "$r_0 = 0$", respectively. Henceforth, we shall assume the algorithm as given, however.

Notice the arithmetic and storage necessary. At each step one application of A to r_{k-1} is required as well as two inner products, four vector additions, four products of scalars and vectors, plus a minor amount of scalar arithmetic. Storage is required for six vectors (b, x, Δx, r, Δr, and Ar) although five are sufficient if it is permissible to overwrite r on b and thus destroy b.

The preceding algorithm (referred to as the cg-method in Stieffel and Hestenes [3]) is almost identical to another algorithm (referred to here as the "minimum residual variant" but in [3] as the cg1-method) in which all equations

are the same but all inner products are replaced by "A-inner products" (i.e., $r_{k-2}^T r_{k-2}$, $r_{k-1}^T r_{k-1}$, and $r_{k-1}^T A r_{k-1}$ are replaced by $r_{k-2}^T A r_{k-2}$, $r_{k-1}^T A r_{k-1}$, and $r_{k-1}^T A^2 r_{k-1} = (Ar_{k-1})^T (Ar_{k-1})$, respectively). The preceding comments on termination, initialization, arithmetic, and storage all apply to this method as well, except that the quantity $\|r_k\|^2$, which is required for checking the termination condition, is not a by-product of this variant and thus its computation requires one additional inner-product per iteration over the standard method.

Using a simple induction argument, the following properties may be verified for either algorithm:

1. The approximate solution x_k is a linear combination of b, Ab, ..., $A^{k-1}b$.
2. The quantity r_k is the residual $b-Ax_k$ and is a linear combination of b, Ab, ..., $A^k b$.
3. The residuals are orthogonal for the standard algorithm (i.e., $r_k^T r_j = 0$ if $k \neq j$) and A-orthogonal for the minimum residual variant (i.e., $r_k^T A r_j = 0$ if $k \neq j$).

Using these properties, the following optimality condition can be proved:

4. From the subspace spanned by b, Ab, ..., $A^{k-1}b$, x_k is the unique vector which minimizes

   $\|A^{-1/2}(b-Ax)\|$ for the standard algorithm

   or

   $\|b-Ax\|$ for the minimum residual variant.

Notice first that the quantity

$$\|A^{-1/2}(b-Ax)\|^2 = (b-Ax)^T A^{-1}(b-Ax)$$
$$= (A^{-1}b-x)^T A(A^{-1}b-x),$$

which is the "A-norm" of the error $A^{-1}b-x$. Furthermore, this quantity differs only by a constant (equal to $b^T A^{-1}b$) from the quadratic expression $x^TAx - 2b^Tx$, and thus this is minimized by x_k over the subspace spanned by $b, \ldots, A^{k-1}b$ with the standard algorithm. Notice also the justification for the name "minimum residual variant" for the second algorithm.

These properties guarantee that the n-th residual (where n is the order of the system) must be orthogonal (or A-orthogonal) to n independent vectors $r_0, \ldots, r_{n-1}$ which implies it is zero. Thus we obtain an exact solution in at most n iterations. This observation, although true, is not relevant to the computational aspects of the algorithm for two reasons. First, when finite precision arithmitic is used, exact orthogonality may not hold and the n-th residual may be far from negligible. Second, for very large problems, n iterations of the algorithms may be excessively expensive, and in such cases the algorithm must produce acceptable solutions in far less than n iterations to be of value.

In section 2, these two algorithms will be compared. Except in that section, all other comments about a conjugate gradient algorithm pertain to the minimum residual variant.

The theories of orthogonal polynomials (see Stiefel [7]) and approximation (see Kaniel [4] and Belford and Kaufman [2]) have been applied to produce estimates of the convergence behavior of the algorithm. Their results have been applied in the remainder of this section.

Since the approximate solution x_k after k iterations is a linear combination of $b, Ab, \ldots, A^{k-1}b$, we may represent it as $\bar{P}_{k-1}(A)b$, where $\bar{P}_{k-1}$ is some polynomial of degree $k-1$. Notice that any linear combination of $b, Ab, \ldots, A^{k-1}b$ can be written as a polynomial in A of degree $k-1$ applied to b, but $\bar{P}_{k-1}$ has the property that over all such polynomials P_{k-1} the corresponding residual $r = b - Ax = b - AP_{k-1}(A)b$ is smallest in Euclidean norm. Thus, if we let $A = UDU^T$ be an orthogonal eigenvalue decomposition for A, where $D = \text{diag}(d_i)$ and $U^TU = UU^T = I$, we have

$$\|r_k\|^2 = \|b-Ax_k\|^2 = \|b-A\bar{P}_{k-1}(A)\cdot b\|^2 = \|(I-A\bar{P}_{k-1}(A))b\|^2$$

$$= \|U(I-D(\bar{P}_{k-1}(D)))U^Tb\|^2.$$

From the invariance of the Euclidean norm under orthogonal transformation, we obtain

$$\|r_k\|^2 = \sum_{i=1}^{n} (1 - d_i\bar{P}_{k-1}(d_k))^2 b_i'^2 ,$$

where $b' = U^Tb$. From the minimizing property of the polynomial $\bar{P}_{k-1}$, we also have that for any other polynomial P_{k-1} of degree $k-1$,

$$\|r_k\|^2 \leq \sum_{i=1}^{n} (1 - d_iP_{k-1}(d_i))^2 b_i'^2$$

$$\leq \max_i (1 - d_iP_{k-1}(d_i))^2 \cdot \sum_{i=1}^{n} b_i'^2$$

$$\leq \max_i (1 - d_iP_{k-1}(d_i))^2 \|b'\|^2$$

$$\leq \max_i (1 - d_iP_{k-1}(d_i))^2 \|b\|^2.$$

The term $1 - d_i P_{k-1}(d_i)$ is simply a polynomial Q_k of degree k evaluated at d_i. Q_k has the property, however, that $Q_k(0) = 1$. It should be recognized that for any such Q_k of degree k and with $Q_k(0) = 1$, there is a unique P_{k-1} of degree $k-1$, satisfying $Q_k(\lambda) = 1 - \lambda P_{k-1}(\lambda)$. We summarize the preceding results as a theorem which shall be referred to as the polynomial bound, henceforth.

Theorem 1. Let $r_k = b - Ax_k$ be the residual obtained by k steps of the minimum residual variant conjugate gradient algorithm; then for any polynomial P_{k-1} of degree $k-1$, we have

$$\frac{\|r_k\|}{\|b\|} \leq \max_i \left|1 - d_i P_{k-1}(d_i)\right|.$$

Alternatively, for any polynomial Q_k of degree k satisfying $Q_k(0) = 1$, we have

$$\frac{\|r_k\|}{\|b\|} \leq \max_i \left|Q_k(d_i)\right|.$$

All of the results to follow which bound $\|r_k\|/\|b\|$ are obtained by using particular selections of such polynomials P_{k-1} and Q_k.

For example, let all the eigenvalues of A be contained in the interval $[\alpha,\beta]$, where $\alpha > 0$, and then let T_k denote the k-th degree Chebychef polynomial on the interval $[\alpha,\beta]$ normalized so $T_k(\beta) = 1$, i.e.,

$$T_k(\lambda) = \frac{1}{2}\left[\left(\theta + \sqrt{\theta^2 - 1}\right)^k + \left(\theta - \sqrt{\theta^2 - 1}\right)^k\right]$$

where

$$\theta = \theta(\lambda) = \frac{2}{\beta - \alpha} \cdot \left[\lambda - \frac{\alpha + \beta}{2}\right].$$

The polynomial $T_k(\lambda)/T_k(0)$ assumes the value 1 at zero; hence we define this as our polynomial Q_k, and since $\max_i |T_k(d_i)| \leq \max_{\lambda\epsilon[\alpha,\beta]} |T_k(\lambda)| = 1$, from the properties of Chebychef polynomials, we have:

Theorem 2. The approximate solution x_k of the minimum residual variant satisfies

$$\frac{\|b-Ax_k\|}{\|b\|} \leq |T_k(0)|^{-1}$$

where T_k is the Chebychef polynomial of degree k on an interval $[\alpha,\beta]$ containing the eigenvalues of A and normalized with $T_k(\beta) = 1$.

If one were to use the Chebychef iteration method (see Varga [8]) on this problem, the same estimate would be obtained (in fact, with this method $r_k = (T_k(0))^{-1} T_k(A)\, b$). Its application would require knowledge of suitable constants α and β, however, and, again from the optimal nature of the minimal residual variant, the conjugate gradient residual at any step does not exceed in norm the Chebychef iteration residual at that step.

Returning to the bound of Theorem 2, we now seek to explore the nature of the sequence $\left\{|T_k(0)|^{-1}\right\}_{k=1}$. Since

$$\theta(0) = -\frac{\beta+\alpha}{\beta-\alpha} = -\frac{\beta/\alpha+1}{\beta/\alpha-1},$$

$$|T_k(0)| = \frac{1}{2}\left[\left(\frac{\beta/\alpha+1}{\beta/\alpha-1} + \sqrt{\left(\frac{\beta/\alpha+1}{\beta/\alpha-1}\right)^2 - 1}\right)^k + \left(\frac{\beta/\alpha+1}{\beta/\alpha-1} - \sqrt{\left(\frac{\beta/\alpha+1}{\beta/\alpha-1}\right)^2 - 1}\right)^k\right].$$

$$= \frac{1}{2}\left[\left(\frac{\beta/\alpha+1+2\sqrt{\beta/\alpha}}{\beta/\alpha-1}\right)^k + \left(\frac{\beta/\alpha+1-2\sqrt{\beta/\alpha}}{\beta/\alpha-1}\right)^k\right].$$

$$= \frac{1}{2}\left[\left(\frac{\sqrt{\beta/\alpha}+1}{\sqrt{\beta/\alpha}-1}\right)^k + \left(\frac{\sqrt{\beta/\alpha}-1}{\sqrt{\beta/\alpha}+1}\right)^k\right].$$

Using simple properties of the hyperbolic cosine, it can be shown that also

$$|T_k(0)| = \cosh k \cosh^{-1}\left(\frac{\beta/\alpha + 1}{\beta/\alpha - 1}\right),$$

and we notice that for any fixed k $|T_k(0)|$ increases (hence the bound $|T_k(0)|^{-1}$ decreases) as β/α decreases. Thus the tightest bound of this type is obtained when $\beta = \max d_i$ (the largest eigenvalue) and $\alpha = \min d_i$ (the smallest eigenvalue), in which case the ratio $\beta/\alpha = \kappa$, the condition number of the matrix A. We state this as a corollary.

Corollary. The approximate solution x_k, of the minimum residual variant, satisfies

$$\frac{\|b-Ax_k\|}{\|b\|} \leq \left(\cosh k \cosh^{-1}\left(\frac{\kappa+1}{\kappa-1}\right)\right)^{-1} = 2\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\right]^{-1}$$

where κ is the condition number of A.

It immediately follows that if we want $\|b-Ax_k\|/\|b\|$ to be less than some given, positive tolerance ϵ, then we should have

$$\cosh k \cosh^{-1}\left(\frac{\kappa+1}{\kappa-1}\right) \geq \frac{1}{\epsilon},$$

thus

$$k \geq \cosh^{-1}\left(\frac{1}{\epsilon}\right) / \cosh^{-1}\left(\frac{\kappa+1}{\kappa-1}\right).$$

Figure 1 provides these values of k corresponding to condition numbers $\kappa = 10^1, 10^2, \ldots, 10^6$, and residual tolerances $\epsilon = 10^{-1}, 10^{-2}, \ldots, 10^{-8}$.

Recall that these numbers answer the question "how many steps of conjugate gradient may be necessary?"

To further explore the nature of the convergence of the algorithm, notice that since $\kappa \geq 1$, $0 \leq \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} < 1 < \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}$. Thus the term $\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k$ in the second bound of the corollary becomes negligible with increasing k. Furthermore, since it is non-negative, we simply overestimate if we ignore it. This results in the following:

Corollary. The approximate solution x_k of the minimum residual variant satisfies

$$\frac{\|b-Ax_k\|}{\|b\|} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k = 2\left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^k,$$

and if $k \geq \log\left(\frac{1}{2\epsilon}\right) \Big/ \log\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)$ for $\epsilon > 0$, then $\|b-Ax_k\| / \|b\| \leq \epsilon$.

This approximation is very good; in fact, if the iteration bounds in Figure 1 based on the first corollary were replaced by the values given by this corollary, then only several entries would change, and they by one step.

Essentially this bound implies linear convergence with a factor of $\left(1 - \frac{2}{\sqrt{\kappa}+1}\right)$, and these numbers have been included in Figure 1. Notice that with well-conditioned problems, convergence is very rapid: about a bit per iteration with $\kappa = 10$. However, for poorly conditioned problems we see that each extra bit may require a large number of steps (e.g., 346 steps for $\kappa = 10^6$). One approach to improving the behavior of the conjugate gradient algorithm is the transformation of the original problem $Ax = b$ into perhaps $(S^TAS)(S^{-1}x) = S^Tb$,

NUMBER OF ITERATIONS NECESSARY TO YIELD RESIDUALS OF A GIVEN MAGNITUDE

Size of Residual	CONDITION NUMBER					
	10^1	10^2	10^3	10^4	10^5	10^6
10^{-1}	5	15	48	150	474	1497
10^{-2}	9	27	84	265	838	2650
10^{-3}	12	38	121	381	1202	3801
10^{-4}	16	50	157	496	1566	4952
10^{-5}	19	61	193	611	1930	6104
10^{-6}	23	73	230	726	3394	7255
10^{-7}	26	84	266	841	2659	8406
10^{-8}	30	96	303	956	3023	9557
Linear Convergence Factor	.519	.818	.938	.980	.994	.998

Figure 1. Theoretical Bounds--Minimum Residual Variant

where S^TAS is better conditioned than A. This is discussed in [1].

It should be emphasized, though, that these are bounds. Furthermore, they are based upon theory including the assumption of A-orthogonality of the residual vectors. It has been our experience, though, that while in practice the theoretical assumption of A-orthogonality may be grossly violated, these bounds have never been. A related question is, then, can these bounds be met (at least theoretically)? In [4], Kaniel asserted the existence of systems for which the bounds of the first corollary would be met; the following theorem provides a construction of such a system.

Theorem 3. Let A be a (k+1) x (k+1) matrix with eigenvalue decomposition UDU^T where $D = \text{diag}\,(d_i)$ and

$$d_i = \frac{\alpha+\beta}{2} + \frac{\alpha-\beta}{2} \cos \frac{i-1}{k} \pi , \quad i = 1,\ldots,k+1,$$

for some $\beta > \alpha > 0$. Further, suppose $b' = U^Tb$ satisfies

$$b'_i = \begin{cases} \frac{1}{2} \frac{1}{\sqrt{d_i}} , & i = 1,\ k+1 \\ \frac{1}{\sqrt{d_i}} , & i = 2,\ldots,\ k. \end{cases}$$

The residual after k steps of the minimum residual conjugate gradient algorithm satisfies

$$\frac{\|r_k\|}{\|b\|} = |T_k(0)|^{-1}$$

where T_k is the Chebychef polynomial on the interval $[\alpha,\beta]$ normalized so $T_k(\beta) = 1$. This is the maximal value for $\|r_k\|/\|b\|$ for the minimum residual conjugate gradient algorithm applied to systems with matrices with eigenvalues in $[\alpha,\beta]$.

Proof: We intend to show that the solution after k steps is

$$x_k = A^{-1}\left(I - \frac{1}{T_k(0)}\, T_k(A)\right) b.$$

First, it is clear that $A^{-1}\left(I - \frac{1}{T_k(0)}\, T_k(A)\right)$ is a polynomial of degree $k-1$ in A and thus x_k is contained in the linear span of $b, Ab, \ldots, A^{k-1}b$. The associated residual is

$$r_k = \frac{1}{T_k(0)}\, T_k(A)\, b$$

and, from the orthogonality condition, we may assert that this is the k-th residual if it is A-orthogonal to the residuals $r_0, r_1, \ldots, r_{k-1}$. Since the solutions $x_0, x_1, \ldots, x_{k-1}$ have the form $x_j = \bar{P}_{j-1}(A)\, b$ where $\bar{P}_{j-1}$ is a polynomial of degree $j-1$, the residuals must also be polynomials: $I - A\bar{P}_{j-1}(A)\, b$. These are of degree up to $k-1$, and it suffices to show that $r_k^T A r = 0$ for any r which is a polynomial of degree up to $k-1$ in A applied to b. Equivalently, we may show $r_k^T A r = 0$ for a set of independent r's which span the same space. Thus we choose r of the form $T_j(A)\, b$ for $j = 0, \ldots, k-1$ and show

$$0 = r_k^T A r = \frac{1}{T_k(0)}\, b^T T_k(A)\, A T_j(A)\, b\,.$$

(Notice we do not claim that the j-th residual r_j is $T_j(A)\, b$ or $\frac{1}{T_j(0)}\, T_j(A)\, b$; only that by using the lower degree Chebychef polynomials we may span the same space as the proper t_j's.)

Using the eigenvalue decomposition, we have

$$b^T T_k(A) A T_j(A)\, b = b'^T T_k(D)\, D\, T_j(D)\, b',$$

$$= \sum_{i=1}^{k+1} d_i T_k(d_i) T_j(d_i) b_i'^2.$$

Then from the definition of b' we have

$$b^T T_k(A) A T_j(A) b = \sum_{i=1}^{k+1"} T_k(d_i) T_j(d_i),$$

where " indicates summation with only half the first and last terms. Now using the definition of d_i and the fact that

$$T_j(d) = \cos\, j \cos^{-1}\left(\frac{2}{\beta-\alpha}\left(d - \frac{\alpha+\beta}{2}\right)\right), \quad \text{for} \quad j = 0,1,\ldots,k,$$

we have

$$b^T T_k(A) A T_j(A) b = \sum_{i=1}^{k+1"} \cos\,(i-1)\pi \cos\, j\,\frac{(i-1)}{k}\,\pi\,,$$

$$= \frac{1}{2}\sum_{i=1}^{k+1"} \cos\,(i-1)\,\frac{k+j}{k}\,\pi + \frac{1}{2}\sum_{i=1}^{k+1"} \cos(i-1)\,\frac{k-j}{k}\,\pi,$$

which is easily shown to be zero using trigonometric identities. Thus r_k is the k-th residual and

$$\|r_k\| = \frac{1}{|T_k(0)|}\,\|T_k(A)\,b\| = \frac{1}{|T_k(0)|}\cdot\|T_k(D)\,b'\|$$

$$= \frac{1}{|T_k(0)|}\cdot\left(\sum_{i=1}^{k+1}(\cos(i-1)\pi\cdot b_i')^2\right)^{1/2}$$

$$= \frac{1}{|T_k(0)|}\left(\sum_{i=1}^{k+1} b_i'^2\right)^{1/2}$$

$$= \frac{1}{|(T_k(0)|}\cdot\|b\|.$$

The claim of maximality was proved in the discussion preceding the theorem. ∎

Even though we now know that the bounds given in Figure 1 are attainable theoretically (and numerical experiments with the system from the theorem support the theory), it may be the case that for particular distributions of eigenvalues, these bounds may be severe overestimates. In Figure 2, we display the actual experimental iteration counts to solve a system of order 1000, where the eigenvalues were equally spaced on the interval $[\alpha,\beta]$ and the components of the right-hand side in the directions of the eigenvectors (i.e., the quantities b_i') were all equal. This is referred to as Test Problem 1. We see that for the ill-conditioned cases there is relatively slow convergence between 10^{-1} and 10^{-2}, but then the rate increases and is essentially independent of condition number.

The case of equally spaced eigenvalues was not selected as an extreme for showing rapid convergence. We shall see in section 3 that for rapid convergence, it is better for the eigenvalues to be sparse at the extremes and dense in the center. The equal spacing is perhaps middle ground between the case where eigenvalues are packed in the extremes (slow convergence) and eigenvalues are packed in the center (rapid convergence). On a particular problem (or class of problems) often a common spectral behavior is predictable, and it may be possible to make better estimates about iteration counts than by using Figure 1. Two important classes in which the distributions are known but disappointing are the tridiagonal matrices with constant diagonals (as found in two-point boundary value differential equations) and the block-tridiagonal matrices associated with the 5-point difference operator approach to Poisson's equation. Both of these classes have spectral distributions that are essentially the Chebychef points of the extreme example of the theorem, and thus we should not expect convergence behavior too much better than that of Figure 1. However, in these cases, it is possible to transform the problem, essentially replacing the condition number with its square root. (For details, consult [1].)

NUMBER OF ITERATIONS NECESSARY TO YIELD RESIDUALS OF A GIVEN MAGNITUDE

Size of Residual	CONDITION NUMBER					
	10^1	10^2	10^3	10^4	10^5	10^6
10^{-1}	4	7	9	10	10	10
10^{-2}	7	19	48	101	128	148
10^{-3}	11	31	83	127	148	164
10^{-4}	14	43	107	145	164	178
10^{-5}	18	54	126	162	178	191
10^{-6}	21	65	143	176	191	203
10^{-7}	25	77	158	189	203	215
10^{-8}	29	88	172	201	215	226

Figure 2. Test Problem 1: Minimum Residual Variant

2. Standard Conjugate Gradient vs. Minimum Residual Variant

In section 1, two different conjugate gradient algorithms were defined. The first minimizes the norm of $A^{-1/2}$ times the residual (which is $A^{1/2}$ times the error, $x - x_k$), the second minimizes the norm of the residual itself (which is A times the error). As shown in [1], it is possible to define an algorithm that minimizes $r^T(CA)^{-1}r$ (which is $(x-x_k)^T(A^{-1}C)^{-1}(x-x_k)$) for any symmetric positive definite matrix C such that $C^{-1}r$ is computable for residuals r.

No method has come to our attention which minimizes simply the error over the space $b, Ab, \ldots, A^{k-1}b$; the standard algorithm comes closest to this in minimizing $A^{1/2}$ times the error. As mentioned in section 1, this algorithm also minimizes a quadratic expression $x^TAx - 2x^Tb$, which may be a discretization of an energy integral. If one is concerned about small errors or minimum energy, then perhaps this is the algorithm to be used, but it is our opinion that more often, especially when the vector b may contain experimental error, a small residual is desired. In the conjugate gradient software implementing the standard algorithm, termination is not based upon a sufficiently small error or even $A^{1/2}$ times errors (since these things are unattainable without the true solution x), but instead upon small residual. Thus one motivation for adopting the minimum residual variant could be,"If we must terminate based upon small residual we should select the algorithm that produces small residuals as fast as possible."

The remark has been made, however, that the residuals produced by the standard algorithm are close to those of the minimum residual variant. The question is then raised, "By how much can the two residuals differ?" A bound can be obtained as follows: Let $\bar{r}_k$ denote the residual from the standard algorithm after k steps (recall it minimizes $\|A^{-1/2}(b-Ax_k)\|$ over x_k selected from the span of $b, Ab, \ldots, A^{k-1}b$) and r_k denote the minimum residual variant

residual after k steps (it minimizes $\|b-Ax_k\|$ over the same set). We have

$$\|\bar{r}_k\| = \|A^{1/2}A^{-1/2}\bar{r}_k\| \leq \|A^{1/2}\| \cdot \|A^{-1/2}\bar{r}_k\|.$$

From the above property of $\bar{r}_k$ that $\|A^{-1/2}\bar{r}_k\| \leq \|A^{-1/2}r_k\|$,

$$\|\bar{r}_k\| \leq \|A^{1/2}\| \cdot \|A^{-1/2}r_k\| \leq \|A^{1/2}\| \cdot \|A^{-1/2}\| \cdot \|r_k\| = \kappa^{1/2}\|r_k\|.$$

It thus is the case that the standard residual does not exceed the square root of the condition number times the minimum residual. Can this bound be met? The answer is not known to this author but certainly cannot be for the trivial case of 2 x 2 systems. In this particular case where

$$D = \begin{bmatrix} \kappa & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad b' = \begin{bmatrix} 1/\sqrt{\kappa} \\ 1 \end{bmatrix}$$

(and this can easily be shown to be the extreme case for 2 x 2 systems), we have

$$\|\bar{r}_1\| = \frac{(\kappa-1)\sqrt{\kappa+1}}{2\kappa}$$

$$\|r_1\| = \frac{(\kappa-1)}{\sqrt{\kappa}\sqrt{\kappa+1}} \quad .$$

Thus $\|\bar{r}_1\| = \frac{1}{2}\left(\sqrt{\kappa} + 1/\sqrt{\kappa}\right) \cdot \|r_1\|$. This is about half the value predicted by the bound.

A ratio of $\|\bar{r}_k\|$ to $\|r_k\|$ even mildly approaching $\sqrt{\kappa}$ for large κ would be disastrous for the standard algorithm, since not only is $\sqrt{\kappa}$ large but the rate of convergence for these problems is so slow that the number of iterations required to get $\|\bar{r}_k\|$ sufficiently small could be prohibitive.

In Figure 3, we display the results of applying the standard algorithm to Test Problem 1. We see for small condition number (i.e., 10^1), the required steps are nearly identical to those for the minimal residual variant (compare with Figure 2). However, for the 10^{-1} level of residual, notice how rapidly the number of iterations increases. One can observe the residual norms increasing to more than ten times $\|b\|$ for many steps (and hence ratios of $\|\bar{r}_k\|$ to $\|r_k\|$ of more than 10 since $\|r_k\| \leq \|b\|$). But then $\|\bar{r}_k\|$ begins a very sudden descent, such that the 10^{-2} line of Figure 3 lags the same line of Figure 2 by at most 5 steps. Thereafter, the two tables agree to an even greater extent.

The conclusions one may draw from this are not firm. It appears the standard algorithm's behavior may "asymptotically" agree with the minimal residual algorithm for some problems. Certainly the initial behavior can be very different especially for poorly conditioned problems. This may be important if a good initial estimate is available and only slight reduction in residual norm is required. Such is the case in time-dependent problems when slightly different systems are being solved at each time step, and the final solution at one time step provides a starting estimate for the solution at the next time step; or with non-linear systems where the conjugate gradient algorithm is being used to solve a sequence of linearized approximations. In general, since the minimum residual variant requires only one additional inner product per step (and it is possible to dispense with this except on occasional steps) and guarantees smaller residual, it is preferred by this author.

NUMBER OF ITERATIONS NECESSARY TO YIELD RESIDUALS OF A GIVEN MAGNITUDE

Size of Residual	CONDITION NUMBER					
	10^1	10^2	10^3	10^4	10^5	10^6
10^{-1}	4	10	26	85	115	136
10^{-2}	8	22	66	113	136	153
10^{-3}	11	34	93	133	153	168
10^{-4}	15	45	114	151	168	182
10^{-5}	18	57	132	166	182	195
10^{-6}	22	68	148	180	194	206
10^{-7}	25	79	162	192	206	217
10^{-8}	29	90	176	204	217	228

Figure 3. Test Problem 1: Standard Conjugate Gradient

3. The Effect of Isolated Extreme Eigenvalues on Rate of Convergence

It was mentioned in the previous section that poor convergence (i.e., close to the theoretical bounds) is observed when eigenvalues are packed at the extremes. The case when the values are packed in the center and the extreme values are well separated is very different. Remarks such as "The method takes several steps cleaning out the components of the solution associated with the extreme values and then moves in on the packed region" and "The rate of convergence essentially depends upon the dense section of the spectrum" have been made. In this section such questions will be explored theoretically and experimentally.

First we may consider the case where the largest eigenvalue is well separated from the remainder. We assume β is exactly this largest eigenvalue and all other eigenvalues are in an interval $[\alpha,\beta']$, with $0 < \alpha \leq \beta' < \beta$. As shown in [1], we may consider the polynomial $Q_k(\lambda) = (T'_{k-1}(0))^{-1}(1-\beta^{-1}\lambda)T'_{k-1}(\lambda)$ where T'_{k-1} is the Chebychef polynomial of degree k-1 on the interval $[\alpha,\beta']$ and normalized so $T'_{k-1}(\beta') = 1$. Certainly, Q_k is of degree k and satisfies $Q_k(0) = 1$. Furthermore, since $|T'_{k-1}(\lambda)| \leq 1$ for $\lambda \in [\alpha,\beta']$, we have

$$|Q_k(\lambda)| \leq |T'_{k-1}(0)|^{-1}|1-\beta^{-1}\lambda|$$
$$\leq |T'_{k-1}(0)|^{-1}$$

and $Q_k(\beta) = 0$. Since all eigenvalues are contained in $[\alpha,\beta'] \cup \{\beta\}$, we may apply the polynomial bound and conclude that after k steps of the minimum residual variant, $\|r_k\|/\|b\|_{k-1}$ is bounded by

$$|T'_{k-1}(0)|^{-1} \sim \left(1 - \frac{2}{\sqrt{\beta'/\alpha}+1}\right)^{k-1}.$$

We may conclude then that the rate of convergence depends upon $k' = \beta'/\alpha$ and that the effect of the eigenvalue at β is only the addition of one iterative

step. This can immediately be extended to several large eigenvalues (say $\beta_1,\ldots,\beta_\ell > \beta'$) by using $Q_k(\lambda) = (T'_{k-\ell}(0))^{-1} \prod_{i=1}^{\ell} (1-\beta^{-1}\lambda)\, T'_{k-\ell}(\lambda)$. In this case the addition of the ℓ eigenvalues larger than β' incurs a lag of at most ℓ steps in the iteration from the case where all eigenvalues are in $[\alpha,\beta']$.

Although theoretically correct, the actual computational results are not quite **as** pleasing. In Figure 4, we display the result of placing 999 eigenvalues equally spaced in $[.1,1]$ (thus $\kappa' = 10$) plus 1 eigenvalue at 1, 10, ..., 10^5. This is Test Problem 2. As in Test Problem 1, b' contains equal components. Theory predicts that all columns should differ by at most 1 step from the first column of Figure **2.** Instead, we see slight increases as the condition number increases. It appears that the affect of finite precision arithmetic is equivalent to perturbing the large eigenvalue β in such a way that the factor $(1-\beta^{-1}\lambda)$ from the theory is not quite zero at this eigenvalue.

We now consider the affect of a well-separated eigenvalue at the lower end of the spectrum. We assume we have $\alpha < \alpha' < \beta$, where α is the lowest eigenvalue and all other values are contained in $[\alpha',\beta]$. As in the preceding analysis, we could consider the polynomial

$$Q_k(\lambda) = (T''_{k-1}(0))^{-1}(1-\alpha^{-1}\lambda)\, T''_{k-1}(\lambda)$$

where T''_{k-1} is the Chebychef polynomial of degree k-1 on $[\alpha',\beta]$. Obviously, $Q_k(0) = 1$ and $Q_k(\alpha) = 0$. Since $|T''_{k-1}(\lambda)| \leq 1$ for $\lambda \in [\alpha',\beta]$, we have

$$|Q_k(\lambda) \leq |T''_{k-1}(0)|^{-1} \cdot |1-\alpha^{-1}\lambda|$$

$$\leq |T''_{k-1}(0)|^{-1} \cdot |1-\beta/\alpha|.$$

NUMBER OF ITERATIONS NECESSARY TO YIELD RESIDUALS OF A GIVEN MAGNITUDE

Size of Residual	CONDITION NUMBER					
	10^1	10^2	10^3	10^4	10^5	10^6
10^{-1}	5	5	5	5	6	6
10^{-2}	7	8	9	9	10	10
10^{-3}	11	13	14	14	15	16
10^{-4}	14	16	17	18	19	20
10^{-5}	18	21	22	23	25	26
10^{-6}	21	24	26	27	29	30
10^{-7}	25	28	30	32	34	36
10^{-8}	29	32	35	38	40	42

Figure 4. Test Problem 2

Since all eigenvalues are contained in the interval $\{\alpha\} \cup [\alpha', \beta]$, we may again apply the polynomial bound theorem and conclude that after k steps of the minimum residual variant $\|r_k\|/\|b\|$ is bounded by $|T''_{k-1}(0)|^{-1} \cdot (\beta/\alpha - 1) \sim (\kappa - 1) \cdot \left(1 - \frac{2}{\sqrt{\beta/\alpha'} + 1}\right)^{k-1}$. So although the rate of convergence is governed by the factor $\left(1 - \frac{2}{\sqrt{\beta/\alpha'} + 1}\right)$ which depends on the "essential condition number" $\kappa' = \beta/\alpha'$, a constant $(\kappa - 1)$ also enters and this can be immense. We see that perhaps $\log (\kappa - 1) \Big/ \left(\frac{\sqrt{k'} + 1}{\sqrt{k'} - 1}\right)$ additional steps are required simply to reduce $|T''_{k-1}(0)|^{(\kappa-1)}$ to about 1.

In Figure 5, we present the results for Test Problem 3 in which 999 eigenvalues were equally spaced in the interval $[1, 10]$, the last eigenvalue being successively at 1, 10^{-1}, 10^{-2}, ..., 10^{-5}. Again, the vector b had equal components in the direction of all eigenvalues. Thus the real condition number, β/α, is 10, 10^2, ..., 10^6, while the essential condition remains very low, 10. By examining the 10^{-2} residual level and those below, we see that the rate of convergence is about constant over all the condition numbers as the theory suggests. The theory also suggests a lag (i.e., $\log (\kappa - 1) \Big/ \left(\frac{\sqrt{\kappa'} + 1}{\sqrt{\kappa'} - 1}\right)$) of about 7, 11, 14, 18, and 21 for condition numbers 10^2, 10^3, 10^4, 10^5, and 10^6, respectively. From observation we see the actual lags are about 5, 9, 12, 16, and 19. These lags may seem negligible for this particular problem; more commonly, however, the reduction of condition number is less dramatic than from 10^6 to 10 and in such cases the assymptotic rate (i.e., the rate associated with the smaller condition number) may not be exhibited until a significant number of iterations has been performed.

NUMBER OF ITERATIONS NECESSARY TO YIELD RESIDUALS OF A GIVEN MAGNITUDE

Size of Residual	CONDITION NUMBER					
	10^1	10^2	10^3	10^4	10^5	10^6
10^{-1}	4	4	4	4	4	4
10^{-2}	7	13	17	20	24	27
10^{-3}	11	17	20	24	27	31
10^{-4}	14	20	24	27	31	34
10^{-5}	18	24	27	31	34	38
10^{-6}	22	27	31	34	38	41
10^{-7}	25	31	34	38	41	45
10^{-8}	29	34	38	41	45	48

Figure 5. Test Problem 3

In conclusion, then, it can be said that the two remarks about convergence made at the beginning of this section (i.e., that the rate depends essentially on the region of clustered eigenvalues) are to some extent true, however, much more so with the isolated large eigenvalues than with the small ones.

4. Use of the Conjugate Gradient Algorithm for Dense and Banded Systems

The rates of convergence and related number of steps for acceptable solutions have been discussed largely independent of the actual dimension of the system. The important consideration has been the distribution of the eigenvalues. Figure 1, it may be recalled, provides guaranteed bounds on the number of steps depending only on condition number and is totally independent of the system's dimension.

While we have suggested that the procedure for applying the matrix A to a given z be related to the nature of A as an operator, it is certainly possible to represent A as an $n \times n$ matrix of coefficients (in dense or sparse form) and simply take inner products of its rows with z (or linear combinations of A's columns if that is preferred) to determine Az. In such a case we would expect the amount of work per application to be approximately 1 multiplication and 1 addition for every non-zero in the matrix. Let us denote the number of non-zeros by N. The algorithm itself requires about $7n$ multiplications and additions per step, and if k iterations are necessary for sufficient convergence, this is $k(N+7n)$ multiplications (and the same number of additions).

We have seen that in the case of well-conditioned problems, a small number of iterations may result in acceptable accuracy, and hence the question arises: Could conjugate gradients be more efficient than a decomposition method even on dense systems? If we compare with Cholesky factorization, followed by solving a pair of triangular systems which requires about $n^3/6 + n + O(n)$ multiplications,

we may conclude that conjugate gradients will be more efficient whenever

$$k(N + 7n) < n^3/6 + n^2.$$

Since N is certainly bounded by n^2 (the situation where A is dense), we may say that this is the case if n exceeds $3\left(k - 1 + \sqrt{(k-1)^2 + 14/3k}\right)$. In Figure 6, we have taken the values of k given in the first 3 columns of Figure 1 and replaced them by sufficient values of n. Recall that these figures are based upon slowest possible convergence and dense systems; if either property were not to hold, the sufficient n would be lower.

For banded systems of bandwidth m (i.e., at most $2m + 1$ non-zeros per row), an application of A requires about $n(2m+1)$ multiplications (and the same number of additions). Taking k conjugate gradient steps here requires, then, about $kn(2m+8)$ multiplications. Comparing this with banded Cholesky factorization and banded triangular system-solving (which together require about $n/2 \cdot (m+1)(m+6)$ multiplications, see [5]), we seek to have

$$kn(2m+8) < \frac{n}{2}(m+1)(m+6)$$

i.e.,

$$k < \frac{(m+1)(m+6)}{4(m+4)}$$

or

$$m < \frac{1}{2} \cdot (4k-7)\ 1 + \sqrt{1+8(8k-3)/(4k-7)^2} \quad .$$

In Figure 7, we display the bandwidths which would result in conjugate gradient exceeding the efficiency of the banded Cholesky approach. Since the operation counts for Cholesky assume $m \ll n$, this should be considered. What also should be considered is that we have actually only assumed that no more than 2m+1 non-zeros occurred in any row. We did not assume anything about the location of the non-zeros: the matrix need not be symmetrically permutable into banded form for the validity

Size of Residual	Condition Number		
	10^1	10^2	10^3
10^{-1}	31	91	289
10^{-2}	55	163	505
10^{-3}	73	229	727
10^{-4}	97	301	943
10^{-5}	115	367	1159
10^{-6}	139	439	1381
10^{-7}	157	505	1597
10^{-8}	181	577	1817

Figure 6. Sufficient Dimension of Dense System for Conjugate Gradient Algorithm to Be More Efficient Than Cholesky Factorization

Size of Residual	Condition Number		
	10^1	10^2	10^3
10^{-1}	14	54	186
10^{-2}	30	102	330
10^{-3}	42	146	478
10^{-4}	58	194	622
10^{-5}	70	238	766
10^{-6}	86	286	914
10^{-7}	98	330	1058
10^{-8}	114	378	1206

Figure 7. Sufficient Bandwidth of Banded System for Conjugate Gradient to Be More Efficient Than Cholesky Factorization

of the conjugate gradient operation counts. Alternatively, the Cholesky approach may begin with a time-consuming bandwidth reduction stage and result in a permuted matrix with a certain bandwidth but a great deal of sparsity within the bands. Most factorizations do not exploit these internal zeros (it is as costly to determine them as it is to ignore them), and thus if the bands were only about p% full, the necessary bandwidth for conjugate gradient to be more efficient would be about p% of the bounds given in Figure 7, and this ignores any time for bandwidth reductions.

While we do not wish to suggest that the conjugate gradient approach is even competitive for all problems, there certainly exist practical problems which are full or banded and for which conjugate gradient is much more efficient.

REFERENCES

1. Axelsson, O.: On Preconditioning and Convergence Acceleration in Sparse Matrix Problems. CERN 74-10 (1974), 21 pp.

2. Belford, G. G., and Kaufman, E. H., Jr.: An Application of Approximation Theory to an Error Estimate in Linear Algebra. Math. Comp. 28 (1974), 711-713.

3. Hestenes, M. R., and Stiefel, E.: Methods of Conjugate Gradients for Solving Linear Systems. Nat. Bur. Standards J. of Res. 49 (1952), 409-436.

4. Kaniel, S.: Estimates for Some Computational Techniques in Linear Algebra, Math. Comp. 20 (1966), 369-378.

5. Martin, R. S., and Wilkinson, J. H.: Symmetric Decomposition of Positive Definite Band Matrices. Numer. Math. 7 (1965), 355-361.

6. Stiefel, E.: Einige Methoden der Relaxations - Rechnung. Z. Angew. Math. Phys. 3 (1952), 1-33.

7. Stiefel, E.: KernalPolynomials in Linear Algebra and Their Numerical Applications. Nat. Bur. Standards Appl. Math. Ser. 49 (1958), 1-22.

8. Varga, R. S.: Matrix Iterative Analysis. Prentice-Hall, Englewood Cliffs, N. J., 1962, 322 pp.