THE CONVERGENCE OF THE METHOD OF CONJUGATE GRADIENTS AT ISOLATED EXTREME POINTS OF THE SPECTRUM

G. W. Stewart

Departments of Mathematics and
Computer Science
Carnegie-Mellon University
Pittsburgh, Pa.

August, 1973

This work was supported in part by the Office of Naval Research under Contract No. N00014-67-A-0314-0018.

ABSTRACT

Let A be a positive definite matrix with a simple eigenvalue λ_1 that lies outside an interval $[\alpha,\beta]$ containing the remaining eigenvalues. Let the method of conjugate gradients be applied to the solution of the linear system $Az = b$ producing a sequence of iterates $z_0, z_1, \ldots$ and an associated sequence of error vectors $e_i = z - z_i$. In this paper bounds are obtained for the component of the error vector lying along the eigenvector associated with λ_1. The bounds imply that, provided λ_1 is well separated from $[\alpha,\beta]$, this component will decrease rapidly, even when the matrix A is moderately ill conditioned.

1. Introduction

In this paper we shall be concerned with the method of conjugate gradients [2] for solving the system of linear equations

$$Az = b, \tag{1.1}$$

where A is a positive definite matrix of order n. The method generates a sequence of independent A-conjugate directions $d_1, d_2, \ldots, d_n$ satisfying

$$d_i^H A d_j = 0 \qquad (i \neq j), \tag{1.2}$$

from which the solution z_n of (1.1) may be obtained by means of the following algorithm:

(1.3)

1) Choose a starting vector z_0
2) For $k = 1,2,\ldots,n$
   1) $r_{k-1} = b - Az_{k-1}$
   2) $\mu_k = d_k^H r_{k-1} / d_k^H A d_k$
   3) $z_k = z_{k-1} + \mu_k d_k$

The algorithm (1.3) will work for any set of conjugate directions satisfying (1.2) (see [7] for a proof and generalizations). What distinguishes the method of conjugate gradients from other conjugate direction algorithms is that the direction d_k is taken to be a linear combination of the first k members of the Krylov sequence

$$d_1, Ad_1, A^2 d_1, \ldots, A^{n-1} d_1. \tag{1.4}$$

This permits d_{k+1} to be written as a linear combination of Ad_k, d_k, and d_{k-1}, with a resulting savings in work and storage that strongly recommends the

method to large sparse problems. The further specialization of choosing $d_1 = -r_0$ (cf. (1.3)) results in even simpler formulas and circumvents any difficulties associated with linear dependencies among the members of the Krylov sequence (1.4).

While the conjugate gradient method can be regarded as a direct method for solving the system (1.1), producing an exact answer after a finite number of steps, it can also be a good iterative method in the sense that the sequence $z_1, z_2, \ldots$ early approximates the solution of (1.1). For example if A is well conditioned, the z_i must approach the solution at a fast rate that increases with the well conditioning of A [1]. Even for the moderately ill conditioned systems associated with the numerical solution of partial differential equations, the method may produce acceptably accurate solutions surprisingly quickly [1,6], a phenomenon which is not covered by the existing theory and is not well understood. The behavior of the method in these applications apparently depends rather delicately on the spectrum of the matrix A and its relation to the solution.

The purpose of this paper is to make a start toward a more refined theory by examining the special case where A has a largest or smallest eigenvalue that is isolated from the rest of the spectrum. For the case of a largest isolated eigenvalue we shall show that the component of the error along the corresponding eigenvector must diminish rapidly at a rate that is independent of the condition number of A. Unfortunately the order constant depends on a number that is bounded by the square root of the condition number of A; however the derivation suggests that this bound will in many cases be an overestimate. For a smallest eigenvalue the results are not as nice, but they still imply reasonably fast reduction of the component of the error along the corresponding eigenvector.

Since the details of the analysis are quite fussy, we shall sketch the underlying ideas here. For definiteness suppose that A has an eigenvalue λ_1 of unity corresponding to the eigenvector x_1 and that the rest of the spectrum of A is confined to the interval $[0,1/2]$. Then if $x_1^H d_1$ is not too small, the vectors $A^{k-1}d_1$, suitably scaled, will approach x_1 (linearly with ratio at least $1/2$), and of course a linear combination of the first k members of the Krylov sequence can be contrived to give an even better approximation to x_1. Otherwise put, the columns space of the matrix $D_k = (d_1, d_2, \ldots, d_k)$ (denoted by $\mathcal{R}(D_k)$) will contain a good approximation to x_1.

Let $e_k = z_n - z_k$ denote the error in the k-th iterate z_k. In [7] the author has shown that

$$e_k = (I - P_k)\, e_0,$$

where

$$P_k = D_k (D_k^H A D_k)^{-1} D_k^H A \tag{1.5}$$

is the projector onto $\mathcal{R}(D_k)$ along the orthogonal complement of $\mathcal{R}(AD_k)$. Since x_1 nearly belongs to $\mathcal{R}(D_k)$, it might be expected that the complementary projector $I - P_k$ would nearly annnihilate the component of the error along x_1. If P_k were an orthogonal projector, this expectation could be easily verified. However P_k is an oblique projector, potentially quite oblique since A may be ill conditioned, and the proper treatment of this obliqueness accounts for most of the detail in the sequel.

In the next section we shall use standard techniques to estimate how accurate an approximation to x_1 we can expect to find in $\mathcal{R}(D_k)$. In Section 3 we shall examine the structure of the projector P_k. These results will be applied in Section 4 to prove a general theorem bounding the component of the

error along x_1. We shall follow Householder's notational conventions [3,8], and we shall use the Euclidean vector norm defined by

$$\|x\|^2 = x^H x$$

as well as the subordinate spectral matrix norm defined by

$$\|A\| = \sup_{\|x\|=1} \|Ax\|.$$

Part of this work was done while I was visiting the IBM Thomas J. Watson Research Center, where I was particularly encouraged by a series of stimulating discussions with Dr. Philip Wolfe.

2. Obtaining an Accurate Eigenvector

Let the matrix A have eigenvalues $\lambda_1,\lambda_2,\ldots,\lambda_n$ corresponding to the orthonormal system of eigenvectors $x_1,x_2,\ldots,x_n$. Since we shall be concerned with the behavior of the method of conjugate gradients at either the largest or the smallest eigenvalue, we shall denote that eigenvalue by λ_1 and let $[\alpha,\beta]$ be the smallest interval containing the remaining eigenvalues. Set

$$\Lambda = \operatorname{diag}(\lambda_1,\lambda_2,\ldots,\lambda_n),\ \Lambda_2 = \operatorname{diag}(\lambda_2,\lambda_3,\ldots,\lambda_n)$$

and

$$X = (x_1,x_2,\ldots,x_n),\ X_2 = (x_2,x_3,\ldots,x_n).$$

It follows that

$$AX = X\Lambda, \quad AX_2 = X_2\Lambda_2,$$

and hence for any polynomial π

$$\pi(A)X = X\pi(\Lambda),\ \pi(A)X_2 = \pi(\Lambda_2)X_2.$$

In this section we shall attempt to determine how accurate an approximation to x_1 can be found in $\mathcal{R}(D_k)$, where D_k was defined in the last section. This is equivalent to finding a linear combination

$$y = \gamma_0 d_1 + \gamma_1 A d_1 + \ldots + \gamma_{k-1} A^{k-1} d_1$$

of the members of the Krylov sequence (1.4) that is a good approximation to x. If we introduce the polynomial

$$\pi(\lambda) = \gamma_0 + \gamma_1 \lambda + \ldots + \gamma_{k-1} \lambda^{k-1}$$

then we must determine π so that

$$y = \pi(A) d_1$$

is a good approximation.

Assume that $x_1^H d_1$ is nonzero. Then the vector

$$\frac{y}{\pi(\lambda_1) x_1^H d}$$

has its component along x_1 equal to unity. The remaining components are given by

$$v = \frac{X_2^H y}{\pi(\lambda_1) x_1^H d_1} = \frac{\pi(\Lambda_2) X_2^H d_1}{\pi(\lambda_1) x_1^H d_1}, \tag{2.1}$$

and minimizing the norm of this vector should give an accurate eigenvector. Of course we must have $\pi(\lambda_1) \neq 0$ and since π occurs in both the numerator and the denominator of (2.1) we may assume that $\pi(\lambda_1) = 1$. We then have the following theorem.

Theorem 2.1. With the above definitions,

$$\min_{\substack{\pi(\lambda_1)=1 \\ \deg\pi\le k-1}} \frac{\|X_2^H\pi(A)d_1\|}{|x_1^H\pi(A)d_1|} \le \frac{1}{T_{k-1}\left[\dfrac{\lambda_1-\frac{\beta+\alpha}{2}}{\frac{\beta-\alpha}{2}}\right]} \frac{\|X_2^Hd_1\|}{|x_1^Hd_1|},$$

where T_{k-1} is the Chebychev polynomial defined by

$$T_{k-1}(x) = \frac{(x+\sqrt{x^2-1})^{k-1} + (x-\sqrt{x^2-1})^{k-1}}{2}.$$

Proof. We have

$$\min_{\substack{\pi(\lambda_1)=1 \\ \deg(\pi)\le k-1}} \frac{\|X_2^H\pi(A)d_1\|}{|x_1^H\pi(A)d_1|} = \min_{\substack{\pi(\lambda_1)=1 \\ \deg\pi\le k-1}} \frac{\|\pi(\Lambda_2)X_2^Hd_1\|}{|x_1^Hd_1|}$$

$$\le \min_{\substack{\pi(\lambda_1)=1 \\ \deg\pi\le k-1}} \|\pi(\Lambda_2)\| \frac{\|X_2^Hd_1\|}{|x_1^Hd_1|}$$

$$\le \left[\min_{\substack{\pi(\lambda_1)=1 \\ \deg\pi\le k-1}} \max\{|\pi(\lambda)| : \lambda=\lambda_2,\ldots,\lambda_n\}\right] \frac{\|X_2^Hd_1\|}{\|x_1^Hd_1\|}$$

$$\le \left[\min_{\substack{\pi(\lambda_1)=1 \\ \deg\pi\le k-1}} \max_{\lambda\in[\alpha,\beta]} |\pi(\lambda)|\right] \frac{\|X_2^Hd_1\|}{\|x_1^Hd_2\|}.$$

The quantity in braces is well known to be

$$\frac{1}{T_{k-1}\left[\dfrac{\lambda_1-\frac{\beta+\alpha}{2}}{\frac{\beta-\alpha}{2}}\right]}.$$

Theorem 2.1 has been given implicitly in [5] and explicitly in [4]. It implies that there is a vector in $\mathcal{R}(D_k)$ whose 2-norm (as opposed to its individual components) along X_2 decreases to zero in proportion to the k-th power of

$$\frac{1}{\left|\frac{\lambda_1 - \frac{\beta+\alpha}{2}}{\frac{\beta-\alpha}{2}}\right| + \left|\frac{\lambda_1^2-\alpha-\beta+\alpha\beta}{\left(\frac{\beta-\alpha}{2}\right)^2}\right|^{1/2}}, \tag{2.2}$$

a decrease which will be quite rapid when λ_1 is reasonably well separated from the interval $[\alpha,\beta]$. It should be observed that the bound results from treating all the X_2-components of d_1 as being equally significant, which may make it a considerable overestimate in some cases. For example if α is near zero then, other things being equal, the vector $d_1 = -Ar_0$ will have a small component along the eigenvector corresponding to α.

3. The Projector P

In this section we shall be concerned with the structure of the projector P_k defined by (1.5). For convenience we shall drop the subscript k.

Let S be nonsingular and let U = DS. It follows from (1.5) that

$$P = U(U^HAU)^{-1}U^HA; \tag{3.1}$$

in other words the projector P may be defined by any matrix of full rank whose column space is the same as that of D. We shall find it convenient to take the columns of U to form an orthonormal basis for $\mathcal{R}(D)$. From the results of the last section, we know that some vector can be expected to be a good approximation to x_1, and without loss of generality we may assume that this vector is u_1, the first column of U.

Partition U in the form (u_1, U_2) and let

$$R = X^H U = \begin{pmatrix} x_1^H u_1 & x_1^H U_2 \\ X_2^H u_1 & X_2^H U_2 \end{pmatrix} \equiv \begin{pmatrix} \rho_{11} & r_{12}^H \\ r_{21} & R_{22} \end{pmatrix}.$$

Because u_1 is a good approximation to x_1, the number

$$\epsilon = \|r_{21}\|$$

must be small. Because P has orthonormal columns,

$$\|r_{12}\| \leq \epsilon$$

(this may be proved by using the singular value decomposition theorem [8, p. 328] to reduce R_{22} to diagonal form). We also have

$$|\rho_{11}| = \sqrt{1 - \epsilon^2} \leq 1$$

and

$$\|R_{22}\| \leq 1.$$

Now let

$$B = U^H A U = \begin{pmatrix} u_1^H A u_1 & u_1^H A U_2 \\ U_2^H A u_1 & U_2^H A U_2 \end{pmatrix} \equiv \begin{pmatrix} \beta_{11} & b_{21}^H \\ b_{21} & B_{22} \end{pmatrix}.$$

We have

$$\beta_{11} = \rho_{11}^2 \lambda_1 + r_{21}^H \Lambda_2 r_{21},$$

whence

$$\beta_{11} \geq \lambda_1 - 2\epsilon^2 \|A\|. \tag{3.2}$$

Also

$$b_{21} = \rho_{11} \lambda_1 r_{12} + R_{22}^H \Lambda_2 r_{21} \tag{3.3}$$

whence

$$\|b_{21}\| \leq 2\epsilon\|A\|. \tag{3.4}$$

Note that since we have assumed nothing about the dimension of $\mathcal{R}(U_2)$, this last inequality implies that for any vector v that is orthogonal to u_1 we have

$$|u_1^H A v| \leq 2\epsilon\|A\| \, \|v\| . \tag{3.5}$$

It is of course the inverse of B that is required in the definition (3.1) of P. This inverse is given explicitly by

$$B^{-1} = \begin{pmatrix} \gamma & \gamma b_{21}^H B_{22}^{-1} \\ \gamma B_{22}^{-1} b_{21} & (B_{22} - \beta_{11} b_{21} b_{21}^H)^{-1} \end{pmatrix}, \tag{3.6}$$

where

$$\gamma^{-1} = \beta_{11} - b_{21}^H B_{22}^{-1} b_{21}.$$

The inverses in the above expressions must exist, since B is positive definite. If we define

$$\kappa = \|A\| \, \|B_{22}^{-1}\| ,$$

then from (3.2) and (3.4) we have the following bound on γ:

$$\gamma^{-1} \geq \lambda_1 - 2\epsilon^2\|A\| \, (1 + 2\kappa). \tag{3.7}$$

Finally we shall need a realistic bound on the norm of P. This may be obtained by observing that

$$V = A^{1/2} U B^{-1/2}$$

has orthonormal columns and hence norm unity. But

$$P = UB^{-1/2}V^{H}A^{1/2};$$

hence

$$\|P\| \le \|U\| \, \|B^{-1/2}\| \, \|V^{H}\| \, \|A^{1/2}\| = \kappa^{1/2}. \tag{3.8}$$

4. The Main Result

In this section we shall derive bounds on the component of the error associated with the isolated eigenvalue λ_1. The general result is contained in the following theorem, in which we recapitulate a bit.

Theorem 4.1. Let the columns of $X = (x_1, X_2)$ form an orthonormal set of eigenvectors for the positive definite matrix A. Let the eigenvalues corresponding to the columns of X_2 lie in the interval $[\alpha,\beta]$ while the eigenvalue λ_1 corresponding to x_1 lies outside $[\alpha,\beta]$. Let $U = (u_1, U_2)$ have orthonormal columns and set

$$P = U(U^{H}AU)^{-1}U^{H}A.$$

Set

$$\epsilon = \|X_2^{H}u_1\|$$

and

$$\kappa = \|A\| \, \|(U_2^{H}AU_2)^{-1}\| .$$

For any vector e_0 set

$$\sigma = |x_1^{H}e_0|, \qquad \tau = \|X_2^{H}e_0\| .$$

Then if

$$\lambda_1 - 2\epsilon^2\|A\|(1+2\kappa) > 0 \tag{4.1}$$

we have

$$|x_1^{H}(I - P)e_0| \le 2\epsilon(\epsilon\sigma+\tau)(1+\kappa^{1/2})\left[1 + \frac{\|A\|}{\lambda_1 - 2\epsilon^2\|A\|(1+2\kappa)}\right]. \tag{4.2}$$

Proof. By multiplying x_1 by a suitable constant of absolute value unity, we may express e_0 in the form

$$e_0 = \sigma x_1 + \tau x_2,$$

where $x_2 \in \mathcal{R}(X_2)$ and $\|x_2\| = 1$. We may also write e_0 in the form

$$e_0 = \sigma' u_1 + \tau' u_2,$$

where $u_1^H u_2 = 0$ and $\|u_2\| = 1$. Moreover

$$\begin{aligned} |\tau'| &= |\sigma u_2^H x_1 + \tau u_2^H x_2| \\ &\leq \epsilon\sigma + \tau. \end{aligned}$$

Since $Pu_1 = u_1$, we have

$$(I - P)e_0 = \tau'(I - P)u_2.$$

Set

$$\tau'(I - P)u_2 = \sigma'' u_1 + \tau'' u_2',$$

where again $u_1^H u_2' = 0$ and $\|u_2'\| = 1$. We shall now obtain bounds for $|\sigma''|$ and $|\tau''|$.

A bound for $|\tau''|$ is easily obtained from (3.8); for

$$|\tau''| \leq |\tau'| \|I-P\| \leq |\tau'|(1+\varkappa^{1/2}).$$

The bound for $|\sigma''|$ depends on the fine structure of P. Specifically

$$\sigma'' = \tau' u_1^H (I - P)u_2 = \tau' u_1^H P u_2 \tag{4.3}$$

Now from the definition of P and from (3.6)

$$u_1^H P u_2 = e_1^H B^{-1} U^H A u_2$$

$$= (\gamma,\ \gamma b_{21}^H B_{22}^{-1}) \begin{pmatrix} U_1^H \\ U_2^H \end{pmatrix} A u_2 \tag{4.4}$$

$$= \gamma u_1^H A u_2 + \gamma b_{21}^H B_{22}^{-1} U_2^H A u_2 .$$

We shall treat each of terms $u_1^H A u_2$ and $b_{21}^H B_{22}^{-1} U_2^H A u_2$ separately. From (3.5),

$$|u_1^H A u_2| \le 2\varepsilon \|A\| . \tag{4.5}$$

From the definition of B

$$b_{21}^H B_{22}^{-1} U_2^H A u_2 = u_2^H A U_2 (U_2^H A U_2)^{-1} U_2^H A u_2,$$

$$= u_1^T A^{1/2} Q A^{1/2} u_2,$$

where $Q = A^{1/2} U_2 (U_2^H A U_2)^{-1} U_2^H A^{1/2}$ is the orthogonal projector onto $\mathcal{R}(A^{1/2} U_2)$. It follows that

$$b_{21}^H B_{22}^{-1} U_2^H A u_2 = u_1^H A^{1/2} v,$$

where $v \in \mathcal{R}(A^{1/2} U_2)$ and $\|v\| \le \|A\|^{1/2}$. But since $B_{22} = (A^{1/2} U_2)^H (A^{1/2} U_2)$, the vector v can be written in the form

$$v = A^{1/2} w,$$

where $w \in \mathcal{R}(U_2)$ and

$$\|w\| \le \|B_{22}^{-1}\|^{1/2} \|v\| .$$

Thus

$$|b_{21}^H B_{22}^{-1} U_2^H A u_1| \le |u_1^H A w| \le 2\varepsilon \|A\| \kappa^{1/2}. \tag{4.6}$$

Combining (3.7), (4.3), (4.4), (4.5) and (4.6) we get

$$|\sigma''| \leq |\tau'| \frac{2\epsilon\|A\| + 2\epsilon\kappa^{1/2}\|A\|}{\lambda_1 - 2\epsilon^2\|A\|\,(1+2\kappa)}$$

If we now write $\tau'(I - P)u_2$ in the form

$$\tau'(I - P)u_2 = \sigma'''x_1 + \tau'''x_2,$$

where $x_2 \in \mathcal{R}(X_2)$ and $\|x_2\| = 1$, we have

$$|x_1^H(I - P)e_0| = |\sigma'''| \leq |\sigma'''| + \epsilon|\tau''|$$

$$\leq 2\epsilon(\epsilon\sigma+\tau)(1+\kappa^{1/2}) \left[1 + \frac{\|A\|}{\lambda_1 - 2\epsilon^2\|A\|(1+2\kappa)}\right],$$

which is the desired inequality.

When $\lambda_1 > \beta$, that is when A has an isolated largest eigenvalue, the numbers λ_1 and $\|A\|$ are equal. The condition (4.1) assumes the simpler form

$$1 - 2\epsilon^2(1+2\kappa) > 0,$$

and the final bound takes the form

$$|x_1^H(I - P)e_0| \leq 2\epsilon(\epsilon\sigma+\tau)(1+\kappa^{1/2}) \left[1 + \frac{1}{1-2\epsilon^2(1+2\kappa)}\right] \tag{4.7}$$

It follows from the results of Section 2 that the method of conjugate gradients must ultimately reduce the component of the error along x_1 at a rate at least as great as the approach of the k-th power of (2.2) to zero. Unfortunately, the factor $(1+\kappa^{1/2})$ appears in the bound. The number κ is bounded by the condition number $\kappa(A) = \|A\|\,\|A^{-1}\|$, which would suggest that rapid convergence cannot be expected when A is ill conditioned. Two circumstances mitigate this hard conclusion. First, it is the square root of κ, not κ itself, that appears in the bound. Second, the number κ is the product of $\|A\|$ and the reciprocal of the

smallest eigenvalue of $B_{22} = U_2^H A U_2$. This eigenvalue is always greater than or equal to the smallest eigenvalue of A, and, especially in the earlier stages of the iteration, it may be quire a bit greater. It follows that $\varkappa$ may often be significantly smaller than $\varkappa(A)$.

In the case where $\lambda_1 < \beta$. the factor $1+\varkappa$ becomes one of order unity. However when ϵ is small, the term in braces in (4.2) becomes approximately equal to $\varkappa(A)$. Of course the quantity ϵ continues to decrease at a rate that is independent of $\varkappa(A)$.

References

1. Engeli, M., Ginsburg, Th., Rutishauser, H., and Stiefel, E., Refined Iterative Methods for Computations of the Solution and Eigenvalues of Self Ajoint Boundary Value Problems. Birkhäuser Verlag, Basel-Stuttgart (1959).

2. Hestenes, M. R. and Stiefel, E., Methods of conjugate gradients for solving linear systems," J. Res. Nat. Bur. Standards 49 (1952) 409-436.

3. Householder, A. S., The Theory of Matrices in Numerical Analysis, Blaisdel, New York 1964.

4. Kaniel, Shmuel, Estimates for some computational techniques in linear Algebra, Math. Comp. 20 (1966) 369-378.

5. Meinardus, Gunther, Über eine Verallgemeinerung einer Ungleichung von L. V. Kantorowitsch, Numer. Math. (1963) 14-23.

6. Reid, J. K., The use of conjugate gradients for systems of linear equations possessing "property A", SIAM J. Numer. Anal. 9 (1972) 325-332.

7. Stewart, G. W., Conjugate direction methods for solving systems of linear equations, Carnegie-Mellon University, Department of Computer Science Report (1972). To appear Numer. Math.

8. _____, Introduction to Matrix Computations, Academic Press, New York (1973).