SIAM J. SCI. COMPUT.
Vol. 15, No. 2, pp. 348–358, March 1994
© 1994 Society for Industrial and Applied Mathematics

Downloaded 06/17/13 to 130.236.83.211. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

MAX-MIN PROPERTIES OF MATRIX FACTOR NORMS*

A. GREENBAUM† AND L. GURVITS‡

Abstract. Given a set of real matrices $C_0, C_1, \ldots, C_k$, conditions are considered under which the equality

$$\min_{\alpha_1,\ldots,\alpha_k} \max_{\|w\|=1} \left\| \left(C_0 + \sum_{i=1}^{k} \alpha_i C_i \right) w \right\| = \max_{\|w\|=1} \min_{\alpha_1,\ldots,\alpha_k} \left\| \left(C_0 + \sum_{i=1}^{k} \alpha_i C_i \right) w \right\|$$

holds. It is shown that if the matrices C_i, $i = 0, 1, \ldots, k$ are normal and commute with one another, then the equality holds. In particular, this implies that if $C_i = A^i$ or $C_i = A^{k-i}$, where A is a normal matrix, then the equality holds. An example is given to show that the equality may fail for noncommuting matrices, when $k > 1$. It is shown that the equality holds for arbitrary matrices if $k = 1$.

Key words. GMRES, Arnoldi, matrix approximation problem, normal matrix, minmax

AMS subject classifications. 65F10, 49K35

1. Introduction. The following problem arises in the analysis of iterative methods for solving linear systems and computing eigenvalues. To solve a linear system, $Ax = b$, given an initial guess x^0 for the solution, the generalized minimal residual (GMRES) method [7] generates approximate solutions x^k, $k = 1, 2, \ldots$ of the form

$$x^k = x^0 + \sum_{i=1}^{k} \alpha_{ik} A^{i-1} r^0,$$

where $r^0 \equiv b - Ax^0$ is the initial residual. The residual vectors $r^k \equiv b - Ax^k$ are of the form

$$r^k = r^0 - \sum_{i=1}^{k} \alpha_{ik} A^i r^0,$$

and the coefficients $\alpha_{1k}, \ldots, \alpha_{kk}$ are chosen to make the 2-norm of r^k as small as possible. A bound on the 2-norm of the residual at any step k is given by

$$\|r^k\| \le \min_{\alpha_1,\ldots,\alpha_k} \left\| I - \sum_{i=1}^{k} \alpha_i A^i \right\| \cdot \|r^0\|.$$

The question arises as to whether this bound is ever attained; that is, whether there is an initial residual r^0, such that

$$\min_{\alpha_1,\ldots,\alpha_k} \left\| \left(I - \sum_{i=1}^{k} \alpha_i A^i \right) r^0 \right\| = \min_{\alpha_1,\ldots,\alpha_k} \left\| I - \sum_{i=1}^{k} \alpha_i A^i \right\| \cdot \|r^0\|.$$

In other words, we have the following max-min problem: Is the inequality

$$\max_{\|r^0\|=1} \min_{\alpha_1,\ldots,\alpha_k} \left\| \left(I - \sum_{i=1}^{k} \alpha_i A^i \right) r^0 \right\| \le \min_{\alpha_1,\ldots,\alpha_k} \max_{\|r^0\|=1} \left\| \left(I - \sum_{i=1}^{k} \alpha_i A^i \right) r^0 \right\| \tag{1}$$

actually an equality?

*Received by the editors June 17, 1992; accepted for publication (in revised form) June 9, 1993.

†Courant Institute of Mathematical Sciences, New York University, New York, New York (greenbau@cs.nyu.edu). This author's work was supported in part by the Applied Mathematical Sciences Program of the U.S. Department of Energy under contract DEFG0288ER25053, and was conducted while visiting the Institute for Mathematics and its Applications at the University of Minnesota, Minneapolis, Minnesota.

‡Siemens Research Corporation, 755 College Road East, Princeton, New Jersey 08540. This author's research was performed while visiting the Institute for Mathematics and its Applications at the University of Minnesota, Minneapolis, Minnesota (gurvits@learning.siemens.com).

Downloaded 06/17/13 to 130.236.83.211. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

A similar question arises in analyzing the Arnoldi method [1] for computing eigenvalues. Given an initial vector q with $\|q\| = 1$, the Arnoldi iteration constructs a sequence of monic polynomials P_k, $k = 1, 2, \ldots$ whose coefficients are chosen to minimize $\|p_k(A)q\|$ over all monic polynomials p_k of degree k. The roots of these polynomials are taken as approximate eigenvalues of the matrix A. The question arises as to whether, for each k, there is an initial vector q such that the monic polynomial P_k constructed by the Arnoldi process also minimizes $\|p_k(A)\|$. A similar max-min statement of the problem asks if the inequality

$$\max_{\|q\|=1} \min_{\alpha_1,\ldots,\alpha_k} \left\| \left(A^k - \sum_{i=1}^{k} \alpha_i A^{k-i} \right) q \right\| \leq \min_{\alpha_1,\ldots,\alpha_k} \max_{\|q\|=1} \left\| \left(A^k - \sum_{i=1}^{k} \alpha_i A^{k-i} \right) q \right\| \tag{2}$$

is actually an equality.

In this paper, we consider a somewhat more general question: Given an arbitrary sequence of real matrices $C_0, C_1, \ldots, C_k$, under what circumstances will the equality

$$\min_{\alpha_1,\ldots,\alpha_k} \max_{\|w\|=1} \left\| \left(C_0 + \sum_{i=1}^{k} \alpha_i C_i \right) w \right\| = \max_{\|w\|=1} \min_{\alpha_1,\ldots,\alpha_k} \left\| \left(C_0 + \sum_{i=1}^{k} \alpha_i C_i \right) w \right\| \tag{3}$$

hold? It is shown that if the matrices C_i, $i = 0, 1, \ldots, k$ are normal and commute with one another, then (3) holds. This result has actually been known for some time, in a slightly different form [6, p. 292]. We include a proof of the result here, since it is not widely known in the numerical analysis community and we point out how our theorem is equivalent to an established result in approximation theory. In particular, this implies that if $C_i = A^i$ or $C_i = A^{k-i}$, where A is a normal matrix, then the equality holds. This generalizes some known results showing that equality holds in (1) and (2) when the matrix A is normal [2], [4], [5]. An example is given to show that (3) may fail for noncommuting matrices, when $k > 1$. It is shown that the equality (3) holds for arbitrary matrices if $k = 1$. The question of whether equality holds in (1) and (2) when $k > 1$ remains open, as does the question of more general conditions on the matrices $C_0, C_1, \ldots, C_k$ that would ensure that (3) holds.

Throughout this paper, we assume that the matrices and vectors appearing in our max-min statements and related theorems are real (though, of course, the eigenvalues and eigenvectors of these matrices may be complex). We will use the notation $A > 0$ to mean that the symmetric matrix A is positive definite. For a vector w, $\|w\|$ will always denote the 2-norm, and for a matrix A, $\|A\|$ will denote the corresponding matrix norm, $\max_{\|w\|=1} \|Aw\|$.

The next section gives the main theorems and examples.

2. Main theorems. The first theorem gives conditions under which some linear combination of given symmetric matrices is positive definite.

THEOREM 2.1. *Let $A_1, \ldots, A_k$ be real n-by-n symmetric matrices. There exist scalars $\alpha_1, \ldots, \alpha_k$ such that*

$$\sum_{i=1}^{k} \alpha_i A_i > 0 \tag{4}$$

if and only if for every set $\{w_1, \ldots, w_m\}$, $m \leq n$, of real nonzero orthogonal n-vectors ($w_p \cdot w_j = 0$, for $p \neq j$), there is an i such that

$$\sum_{j=1}^{m} \langle A_i w_j, w_j \rangle \neq 0. \tag{5}$$

If the symmetric matrices $A_1, \ldots, A_k$ commute: $A_i A_j = A_j A_i$, $i, j = 1, \ldots, k$, then there exist scalars $\alpha_1, \ldots, \alpha_k$ for which (4) *holds if and only if for every individual real nonzero n-vector w, there is an i such that*

$$\langle A_i w, w \rangle \neq 0. \tag{6}$$

Proof. To see the necessity of (5), note that if, for some orthogonal set $\{w_1, \ldots, w_m\}$, we have

$$\sum_{j=1}^{m} \langle A_i w_j, w_j \rangle = 0$$

for all i, then also

$$\sum_{j=1}^{m} \left\langle \sum_{i=1}^{k} \alpha_i A_i w_j, w_j \right\rangle = \operatorname{tr}\left(W^T \left(\sum_{i=1}^{k} \alpha_i A_i \right) W \right) = 0$$

for any $\alpha_1, \ldots, \alpha_k$, where W is the matrix whose columns are $w_1, \ldots, w_k$. But since every principal submatrix of a positive definite matrix is positive definite, this implies that $\sum_{i=1}^{k} \alpha_i A_i$ cannot be positive definite.

To see the sufficiency of (5), suppose that no linear combination $\sum_{i=1}^{k} \alpha_i A_i$ is positive definite. Then the linear subspace spanned by $A_1, \ldots, A_k$ and the convex cone of positive definite matrices can be separated. That is, there is a symmetric matrix B such that

$$\operatorname{tr}\left(B \sum_{i=1}^{k} \alpha_i A_i \right) = 0 \tag{7}$$

for all $\alpha_1, \ldots, \alpha_k$ and

$$\operatorname{tr}(BP) > 0 \tag{8}$$

for all positive definite matrices P. Write B in the form $B = QDQ^T$, where D is the diagonal matrix of eigenvalues of B and Q is the orthogonal matrix of eigenvectors. Then (8) implies that

$$\operatorname{tr}(QDQ^T P) = \operatorname{tr}(D\, Q^T P Q) > 0. \tag{9}$$

Since the diagonal elements of $Q^T P Q$ can be any positive numbers, (9) implies that the diagonal elements of D are nonnegative, with at least one of these being positive. But from (7) it follows that

$$\operatorname{tr}(BA_i) = \operatorname{tr}(QDQ^T A_i) = \operatorname{tr}((QD^{1/2})^T A_i (QD^{1/2})) = 0$$

for all i. Taking $w_1, \ldots, w_m$ to be the nonzero columns of $QD^{1/2}$, this says

$$\sum_{j=1}^{m} w_j^T A_i w_j = 0,$$

which contradicts (5).

Downloaded 06/17/13 to 130.236.83.211. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Downloaded 06/17/13 to 130.236.83.211. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

The second part of the theorem can be proved similarly, using the fact that if the matrices $A_1, \ldots, A_k$ commute, then they can be simultaneously diagonalized. That is, there exists an orthogonal matrix Q such that

$$A_i = Q\Lambda_i Q^T, \quad QQ^T = Q^T Q = I, \quad \Lambda_i = \operatorname{diag}(\lambda_{i1}, \ldots, \lambda_{in}).$$

Again, the necessity of (6) is clear, and to see that it is sufficient, we note that if no linear combination of $A_1, \ldots, A_k$ and hence of $\Lambda_1, \ldots, \Lambda_k$ is positive definite, then there is a diagonal matrix $D = \operatorname{diag}(d_1, \ldots, d_n)$ such that

$$\operatorname{tr}\left(D \sum_{i=1}^{k} \alpha_i \Lambda_i\right) = 0, \tag{10}$$

for all $\alpha_1, \ldots, \alpha_k$ and

$$\operatorname{tr}(DP) > 0 \tag{11}$$

for all positive definite diagonal matrices P. From (11) it follows that the diagonal elements of D are nonnegative, with at least one being positive. Define w to be the vector $(\sqrt{d_1}, \ldots, \sqrt{d_n})^T$. From (10) we have for each i,

$$\operatorname{tr}(D\Lambda_i) = \sum_{j=1}^{n} d_j \lambda_{ij} = \langle \Lambda_i w, w \rangle = 0,$$

which contradicts (6). □

The second part of Theorem 2.1 is really a result about vectors rather than matrices: Given a set of n-vectors, $\phi^{(1)}, \ldots, \phi^{(k)}$ (corresponding to the diagonal matrices $\Lambda_1, \ldots, \Lambda_k$ in the theorem), there is a linear combination of these vectors that has all positive elements if and only if for every set of nonnegative numbers $w_1^2, \ldots, w_n^2$, not all zero, there is an i such that

$$\sum_{j=1}^{n} \phi_j^{(i)} w_j^2 \neq 0.$$

In this form, the second part of Theorem 2.1, as well as Theorem 2.3, which is given later, are known. See, for example, [6]. We include proofs of these theorems here for completeness.

We now use Theorem 2.1 to establish conditions under which the optimal coefficients $\alpha_1, \ldots, \alpha_k$ on both sides of equality (3) are zero.

THEOREM 2.2. *Let $C_1, \ldots, C_k$ be real square matrices such that each pair $(C_i + C_i^T)$ and $(C_j + C_j^T)$ commute. Suppose*

$$\min_{\alpha_1, \ldots, \alpha_k} \max_{\|w\|=1} \left\| \left(I + \sum_{i=1}^{k} \alpha_i C_i\right) w \right\| = 1. \tag{12}$$

Then

$$\max_{\|w\|=1} \min_{\alpha_1, \ldots, \alpha_k} \left\| \left(I + \sum_{i=1}^{k} \alpha_i C_i\right) w \right\| = 1. \tag{13}$$

Downloaded 06/17/13 to 130.236.83.211. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Proof. For a given vector w, we have

$$\min_{\alpha_1,\ldots,\alpha_k} \left\| \left(I + \sum_{i=1}^{k} \alpha_i C_i \right) w \right\| = 1$$

if and only if $\langle C_i w, w\rangle = 0$ for all i; i.e., if and only if $\langle (C_i + C_i^T)w, w\rangle = 0$ for all i. Suppose (13) does not hold. Then for any vector $w \neq 0$ there is an i such that

$$\langle (C_i + C_i^T)w, w\rangle \neq 0.$$

From Theorem 2.1, there is a linear combination

$$\sum_{i=1}^{k} \alpha_i (C_i + C_i^T)$$

that is positive definite. For ϵ sufficiently small, then,

$$\left\| I - \epsilon \sum_{i=1}^{k} \alpha_i C_i \right\|^2 = \left\| I - \epsilon \sum_{i=1}^{k} \alpha_i (C_i + C_i^T) \right\| + \mathcal{O}(\epsilon^2) < 1,$$

which contradicts the assumption (12). □

Theorem 2.2 is now used to establish certain conditions under which equality (3) holds.

THEOREM 2.3. *Let $C_0, C_1, \ldots, C_k$ be nonsingular normal matrices that commute. Then*

$$\min_{\alpha_1,\ldots,\alpha_k} \max_{\|w\|=1} \left\| \left(C_0 + \sum_{i=1}^{k} \alpha_i C_i \right) w \right\| = \max_{\|w\|=1} \min_{\alpha_1,\ldots,\alpha_k} \left\| \left(C_0 + \sum_{i=1}^{k} \alpha_i C_i \right) w \right\|. \tag{14}$$

Proof. Suppose $\hat{\alpha}_1, \ldots, \hat{\alpha}_k$ minimize $\|C_0 + \sum_{i=1}^{k} \alpha_i C_i\|$. We can assume without loss of generality that this minimal norm is 1. We will consider two cases.

1. First, suppose all singular values of $U \equiv C_0 + \sum_{i=1}^{k} \hat{\alpha}_i C_i$ are equal. Then U is a real orthogonal matrix and it commutes with each matrix C_i. The same holds for the inverse matrix U^T. We can write

$$\begin{aligned} \min_{\alpha_1,\ldots,\alpha_k} \left\| C_0 + \sum_{i=1}^{k} \alpha_i C_i \right\| &= \min_{\beta_1,\ldots,\beta_k} \left\| U^T \left(C_0 + \sum_{i=1}^{k} (_i a \hat{\alpha}_i + \beta_i) C_i \right) \right\| \\ &= \min_{\beta_1,\ldots,\beta_k} \left\| I + \sum_{i=1}^{k} \beta_i U^T C_i \right\| = 1. \end{aligned}$$

Because the matrices C_i are normal and commute with each other and with U and U^T, each pair $(U^T C_i + C_i^T U)$ and $(U^T C_j + C_j^T U)$ commute. Therefore, from Theorem 2.2, we have

$$\max_{\|w\|=1} \min_{\beta_1,\ldots,\beta_k} \left\| \left(I + \sum_{i=1}^{k} \beta_i U^T C_i \right) w \right\| = 1,$$

and from this the desired result follows.

2. Now suppose some singular values are less than 1. We can write the real Schur decomposition of each matrix C_i in the form

$$C_i = Q D_i Q^T, \qquad Q Q^T = Q^T Q = I,$$

Downloaded 06/17/13 to 130.236.83.211. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

where each D_i is a block diagonal matrix with 1-by-1 or 2-by-2 blocks on the main diagonal. It suffices to consider the block diagonal matrices D_i. Define $\hat{D} \equiv D_0 + \sum_{i=1}^{k} \hat{\alpha}_i D_i$, and order the elements so that $\hat{D}$ is of the form

$$\hat{D} = \begin{pmatrix} U & 0 \\ 0 & K \end{pmatrix},$$

where U is, say, a t-by-t matrix whose eigenvalues are all equal to 1 in magnitude and K is an $(n-t)$-by-$(n-t)$ matrix whose eigenvalues are all less than 1 in magnitude. Note that each matrix D_i has this same block structure

$$D_i = \begin{pmatrix} D_{i1} & 0 \\ 0 & D_{i2} \end{pmatrix},$$

since any off-diagonal block X in D_i would have to satisfy the homogeneous Sylvester equation: $UX - XK = 0$ in order that D_i and $\hat{D}$ commute. Since the spectrum of U does not intersect the spectrum of K, this equation has only the trivial solution $X = 0$. We can write

$$\begin{aligned} \min_{\alpha_1,\ldots,\alpha_k} \left\| D_0 + \sum_{i=1}^{k} \alpha_i D_i \right\| &= \min_{\beta_1,\ldots,\beta_k} \left\| D_0 + \sum_{i=1}^{k} (\hat{\alpha}_i + \beta_i) D_i \right\| \\ &= \min_{\beta_1,\ldots,\beta_k} \left\| \begin{pmatrix} U & 0 \\ 0 & K \end{pmatrix} + \sum_{i=1}^{k} \beta_i \begin{pmatrix} D_{i1} & 0 \\ 0 & D_{i2} \end{pmatrix} \right\|. \end{aligned} \tag{15}$$

Since $\|U\| > \|K\|$, the same coefficients $\beta_1, \ldots, \beta_k$ that minimize the norm of the matrix in (15) (namely, $\beta_i = 0$, $i = 1, \ldots, k$) also minimize the norm of the upper left t-by-t block, and we have

$$\min_{\beta_1,\ldots,\beta_k} \left\| U - \sum_{i=1}^{k} \beta_i D_{i1} \right\| = 1.$$

Since the singular values of the minimal norm matrix of this form are all equal to 1 and since the matrices D_{i1} are normal and commute with each other, it now follows from part 1 that there is a t-vector $\hat{w}$ with $\|\hat{w}\| = 1$ such that

$$\min_{\beta_1,\ldots,\beta_k} \left\| U - \sum_{i=1}^{k} \beta_i D_{i1} \right\| = \min_{\beta_1,\ldots,\beta_k} \left\| \left(U - \sum_{i=1}^{k} \beta_i D_{i1} \right) \hat{w} \right\|.$$

Defining $\tilde{w}$ to be the n-vector whose first t elements are equal to those of $\hat{w}$ and whose remaining elements are zero, we find

$$\min_{\alpha_1,\ldots,\alpha_k} \left\| D_0 + \sum_{i=1}^{k} \alpha_i D_i \right\| = \min_{\alpha_1,\ldots,\alpha_k} \left\| \left(D_0 + \sum_{i=1}^{k} \alpha_i D_i \right) \tilde{w} \right\|,$$

from which the desired result follows. □

Stated in terms of vectors, Theorem 2.3 states the following. Let f denote the vector consisting of the eigenvalues of C_0, and let $g^{(1)}, \ldots, g^{(k)}$ denote the vectors consisting of the eigenvalues of $C_1, \ldots, C_k$. The left-hand side of (14) is the distance between f and the closest vector to f in the infinity norm from the space G spanned by $g^{(1)}, \ldots, g^{(k)}$. The right-hand side of (14) is the difference between f and the closest vector to f in some weighted L_2 norm

Downloaded 06/17/13 to 130.236.83.211. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

from the space G. The weights are the squared components of the vector w in the direction of each eigenvector of C_i. With this notation, Theorem 2.3 can be stated as follows.

There exists a vector w in R^n with $\|w\|_2 = 1$ such that

$$\min_{g \in G} \|f - g\|_\infty = \|f - \bar{g}\|_\infty = \|f - \bar{g}\|_w = \min_{g \in G} \|f - g\|_w,$$

where $\|v\|_w^2 \equiv \sum_{i=1}^n v_i^2 w_i^2$.

A more general version of this theorem can be deduced from [6, p. 292]. Corollary 1 in that reference states that for any p, q with $1 < p < q \leq \infty$, the vector $\bar{g}$ minimizes some weighted L_q norm of $f - g$ if and only if it minimizes some weighted L_p norm of $f - g$. To obtain the additional result that $\|f - \bar{g}\|_\infty = \|f - \bar{g}\|_w$, we must use the fact that $|f_i - \bar{g}_i|$ attains its maximum value at each of the points i for which w_i is nonzero [6, p. 302]. These facts form the basis of Lawson's algorithm for computing L_∞ approximations by solving a sequence of weighted L_2 (or L_k) approximation problems [3].

Note that the assumption of commutativity in the second part of Theorem 2.1, and hence in Theorems 2.2 and 2.3, is necessary. Consider, for example, the symmetric matrices

$$A_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \tag{16}$$

For any vector w, we have

$$\langle A_1 w, w \rangle = w_1^2 - w_2^2, \qquad \langle A_2 w, w \rangle = w_1^2 + 2w_1 w_2 - w_2^2,$$

and if the first inner product is zero and w_1 and w_2 are not both zero, then the second inner product cannot be zero. Yet there is no linear combination of A_1 and A_2 that is positive definite. We have

$$\min_{\alpha_1, \alpha_2} \|I + \alpha_1 A_1 + \alpha_2 A_2\| = 1,$$

but for any vector w,

$$\min_{\alpha_1, \alpha_2} \|(I + \alpha_1 A_1 + \alpha_2 A_2) w\| = 0.$$

In the next lemma we consider general real m-by-n matrices $C_0, C_1, \ldots, C_k$. Suppose $\hat{\alpha}_1, \ldots, \hat{\alpha}_k$ minimize

$$\left\| C_0 + \sum_{i=1}^k \alpha_i C_i \right\|.$$

Define $\hat{C} \equiv C_0 + \sum_{i=1}^k \hat{\alpha}_i C_i$ and write the singular value decomposition of $\hat{C}$ as

$$\hat{C} = U \Sigma V^T,$$

where U is an m-by-m real orthogonal matrix, V is an n-by-n real orthogonal matrix, and Σ is an m-by-n matrix of the form

$$\Sigma = \begin{pmatrix} \sigma_1 & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \sigma_n \\ 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & 0 \end{pmatrix} \quad \text{or} \quad \Sigma = \begin{pmatrix} \sigma_1 & & & 0 & \cdot & 0 \\ & \cdot & & \cdot & \cdot & \cdot \\ & & \cdot & \cdot & \cdot & \cdot \\ & & & \sigma_m & 0 & \cdot & 0 \end{pmatrix}$$

Downloaded 06/17/13 to 130.236.83.211. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

accordingly as $m \geq n$ or $m \leq n$. Assume that the singular values σ_i, $i = 1, \ldots, \min\{m, n\}$ satisfy

$$\sigma_1 = \cdots = \sigma_t > \sigma_{t+1} \geq \cdots,$$

and let V_t be the n-by-t matrix consisting of the first t columns of V, while $V_{t+1:n}$ is the n-by-$(n-t)$ matrix consisting of columns $t+1$ through n of V.

The following lemma is used to prove Theorem 2.5, but it is also of significant interest in itself.

LEMMA 2.4. *Using the above notation, the coefficients* $\hat{\alpha}_1, \ldots, \hat{\alpha}_k$ *that minimize*

$$\left\| C_0 + \sum_{i=1}^{k} \alpha_i C_i \right\| \tag{17}$$

also minimize

$$\left\| \left(C_0 + \sum_{i=1}^{k} \alpha_i C_i \right) V_t \right\|. \tag{18}$$

Proof. Suppose $\hat{\alpha}_1, \ldots, \hat{\alpha}_k$ do not minimize (18). Then there are coefficients $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_k$ such that

$$\left\| \left(C_0 + \sum_{i=1}^{k} \tilde{\alpha}_i C_i \right) V_t \right\| < \left\| \left(C_0 + \sum_{i=1}^{k} \hat{\alpha}_i C_i \right) V_t \right\|.$$

We will show that for sufficiently small values of ϵ, the coefficients $(1-\epsilon)\hat{\alpha}_i + \epsilon\tilde{\alpha}_i$ satisfy

$$\left\| C_0 + \sum_{i=1}^{k} ((1-\epsilon)\hat{\alpha}_i + \epsilon\tilde{\alpha}_i) C_i \right\| < \left\| C_0 + \sum_{i=1}^{k} \hat{\alpha}_i C_i \right\|,$$

which contradicts the assumption that $\hat{\alpha}_1, \ldots, \hat{\alpha}_k$ minimize (17).

For any ϵ in (0, 1) we have

$$\begin{aligned} \left\| \left(C_0 + \sum_{i=1}^{k} ((1-\epsilon)\hat{\alpha}_i + \epsilon\tilde{\alpha}_i) C_i \right) V_t \right\| &\leq (1-\epsilon) \left\| \left(C_0 + \sum_{i=1}^{k} \hat{\alpha}_i C_i \right) V_t \right\| \\ &+ \epsilon \left\| \left(C_0 + \sum_{i=1}^{k} \tilde{\alpha}_i C_i \right) V_t \right\| < \sigma_1 - \mathcal{O}(\epsilon). \end{aligned} \tag{19}$$

For sufficiently small ϵ, we also have

$$\begin{aligned} \left\| \left(C_0 + \sum_{i=1}^{k} ((1-\epsilon)\hat{\alpha}_i + \epsilon\tilde{\alpha}_i) C_i \right) V_{t+1:n} \right\| &\leq (1-\epsilon) \left\| \left(C_0 + \sum_{i=1}^{k} \hat{\alpha}_i C_i \right) V_{t+1:n} \right\| \\ &+ \epsilon \left\| \left(C_0 + \sum_{i=1}^{k} \tilde{\alpha}_i C_i \right) V_{t+1:n} \right\| < \sigma_{t+1} + \frac{1}{2}(\sigma_1 - \sigma_{t+1}). \end{aligned} \tag{20}$$

Downloaded 06/17/13 to 130.236.83.211. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Define the matrix $K \equiv (K_1, K_2)$ by

$$K_1 = U^T \left(C_0 + \sum_{i=1}^{k} ((1-\epsilon)\hat{\alpha}_i + \epsilon\tilde{\alpha}_i) C_i \right) V_t$$

$$= (1-\epsilon)\sigma_1 \begin{pmatrix} I \\ 0 \end{pmatrix} + \epsilon U^T \left(C_0 + \sum_{i=1}^{k} \tilde{\alpha}_i C_i \right) V_t,$$

$$K_2 = U^T \left(C_0 + \sum_{i=1}^{k} ((1-\epsilon)\hat{\alpha}_i + \epsilon\tilde{\alpha}_i) C_i \right) V_{t+1:n}$$

$$= (1-\epsilon) \begin{pmatrix} 0 \\ \Sigma_{t+1:n} \end{pmatrix} + \epsilon U^T \left(C_0 + \sum_{i=1}^{k} \tilde{\alpha}_i C_i \right) V_{t+1:n},$$

where $\Sigma_{t+1:n} \equiv \operatorname{diag}(\sigma_{t+1}, \ldots, \sigma_n)$. We would like to show that $\|K\| < \sigma_1$ or, equivalently, that the matrix

$$\sigma_1^2 I - K^T K = \begin{pmatrix} \sigma_1^2 I - K_1^T K_1 & -K_1^T K_2 \\ -K_2^T K_1 & \sigma_1^2 I - K_2^T K_2 \end{pmatrix} \tag{21}$$

is positive definite. From (19) and (20) it follows that the diagonal blocks are positive definite, so it suffices to show that

$$(\sigma_1^2 I - K_1^T K_1) - K_1^T K_2 (\sigma_1^2 I - K_2^T K_2)^{-1} K_2^T K_1 > 0.$$

It is easy to check that

$$\|K_1^T K_2 (\sigma_1^2 I - K_2^T K_2)^{-1} K_2^T K_1\| = \mathcal{O}(\epsilon^2),$$

while the eigenvalues of $\sigma_1^2 I - K_1^T K_1$ are of order ϵ. For sufficiently small ϵ, then, the matrix (21) is positive definite, and this gives the desired contradiction. □

Using this lemma and Theorem 2.2 we can now prove equality (3) for general matrices, when $k = 1$.

THEOREM 2.5. *Let C_0 and C_1 be arbitrary real m-by-n matrices. Then*

$$\min_{\alpha} \max_{\|w\|=1} \|(C_0 + \alpha C_1) w\| = \max_{\|w\|=1} \min_{\alpha} \|(C_0 + \alpha C_1) w\|. \tag{22}$$

Proof. We will use induction on the number of columns n. If $n = 1$, the result is clearly true. Assume it is true for matrices with $n-1$ columns, and now consider matrices with n columns. Suppose $\hat{\alpha}$ minimizes $\|C_0 + \alpha C_1\|$. Define $\hat{C} \equiv C_0 + \hat{\alpha} C_1$ and write the singular value decomposition of $\hat{C}$ as

$$\hat{C} = U \Sigma V^T,$$

where U, V, and Σ are as defined earlier. Assume that the singular values $\sigma_i, i = 1, \ldots, \min\{m, n\}$ satisfy

$$\sigma_1 = \cdots = \sigma_t > \sigma_{t+1} \geq \cdots,$$

and let V_t be the n-by-t matrix consisting of the first t columns of V, while $V_{t+1:n}$ is the n-by-$(n-t)$ matrix consisting of columns $t+1$ through n of V.

We will consider two cases. In the first case, assume that $t < n$. According to the lemma, $\hat{\alpha}$ minimizes

$$\|(C_0 + \alpha C_1)V_t\|,$$

and so, by the induction hypothesis,

$$\|(C_0 + \hat{\alpha}C_1)V_t\| = \max_{\|w\|=1} \min_{\alpha} \|(C_0 + \alpha C_1)V_t w\|.$$

If $\hat{w}$ is the t-vector for which this maximum is attained and if we define the n-vector $\tilde{w}$ to be $V_t\hat{w}$, then we have the desired result

$$\|C_0 + \hat{\alpha}C_1\| = \min_{\alpha} \|(C_0 + \alpha C_1)\tilde{w}\|.$$

Now suppose $t = n$. We can assume without loss of generality that $\sigma_1 = 1$, and then we have

$$\min_{\alpha} \|C_0 + \alpha C_1\| = \min_{\beta} \|U^T(C_0 + (\beta + \hat{\alpha})C_1)V\|$$

$$= \min_{\beta} \left\| \begin{pmatrix} I \\ 0 \end{pmatrix} + \beta \begin{pmatrix} U_n^T C_1 V \\ \\ U_{n+1:m}^T C_1 V \end{pmatrix} \right\| = 1,$$

where U_n consists of the first n columns of U and $U_{n+1:m}$ consists of columns $n+1$ through m of U. The same coefficient β that minimizes the norm of the entire matrix also minimizes the norm of the top n-by-n block of this matrix and so we have

$$\min_{\beta} \|I + \beta U_n^T C_1 V\| = 1.$$

From Theorem 2.2 it now follows that

$$\max_{\|w\|=1} \min_{\beta} \|(I + \beta U_n^T C_1 V)w\| = 1$$

and hence that (22) holds. □

Special cases of this theorem, as well as Theorem 2.3, were derived independently and proved in a different way by Joubert [5].

3. Further discussion. Extensive numerical testing of the inequalities in (1) and (2) for a variety of matrices suggests that they are, indeed, equalities. Theorem 2.5 proves this is so for $k = 1$, but we have been unable to prove (or disprove) this result for $k > 1$. The example (16) shows that the proof must rely on special properties of polynomials, since the result is not true for arbitrary noncommuting matrices, even if they are normal.

Downloaded 06/17/13 to 130.236.83.211. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Acknowledgment. The authors thank Nick Trefethen for pointing out that the results of Theorem 2.3 are equivalent to known results about vector approximation and for indicating where these could be found in the literature.

REFERENCES

[1] W. E. ARNOLDI, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[2] A. GREENBAUM, *Comparison of splittings used with the conjugate gradient algorithm*, Numer. Math., 33 (1979), pp. 181–194.

[3] C. L. LAWSON, *Contributions to The Theory of Linear Least Maximum Approximations*, Ph.D. thesis, University of California at Los Angeles, 1961.

[4] A. GREENBAUM AND L. N. TREFETHEN, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM J. Sci. Comput., 15 (1994) pp. 359–368.

[5] W. JOUBERT, *A robust GMRES-based adaptive polynomial preconditioning algorithm for nonsymmetric linear systems*, SIAM J. Sci. Comput., 15 (1994) pp. 427–439.

[6] J. R. RICE, *The Approximation of Functions, Vol. 2: Nonlinear and Multivariate Theory*, Addison-Wesley, Reading, MA, 1969.

[7] Y. SAAD AND M. H. SCHULTZ, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

Downloaded 06/17/13 to 130.236.83.211. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php