MINIMAL RESIDUAL METHOD STRONGER THAN POLYNOMIAL PRECONDITIONING *

V. FABER†, W. JOUBERT†, E. KNILL†, AND T. MANTEUFFEL‡

Abstract. This paper compares the convergence behavior of two popular iterative methods for solving systems of linear equations: the s-step restarted minimal residual method (commonly implemented by algorithms such as GMRES(s)), and $(s-1)$-degree polynomial preconditioning. It is known that for normal matrices, and in particular for symmetric positive definite matrices, the convergence bounds for the two methods are the same. In this paper we demonstrate that for matrices unitarily equivalent to an upper triangular Toeplitz matrix, a similar result holds, namely, either both methods converge or both fail to converge. However, we show this result cannot be generalized to all matrices. Specifically, we develop a method for obtaining examples of real matrices for which GMRES(s) converges for every initial vector, but every $(s-1)$-degree polynomial preconditioning stagnates or diverges for some initial vector.

Key words. linear systems, iterative methods, nonsymmetric, non-normal matrix, GMRES, polynomial preconditioning, convergence

AMS subject classifications. 65F10, 65F15

1. Introduction. A chief goal of numerical linear algebra is to solve linear systems of the form

$$Au = b \tag{1}$$

in a reliable and fast way. Here $A \in \mathbb{C}^{N \times N}$ is nonsingular and is possibly the result of a preconditioning operation such as $Q\hat{A}u = Q\hat{b}$.

The set of *polynomial methods* (sometimes loosely referred to as *Krylov subspace methods*) has proven to be extremely powerful for solving many types of linear systems. These are defined by

$$u^{(n)} = u^{(0)} + q_{n-1}(A)r^{(0)} \tag{2}$$

* This work was supported in part by the Department of Energy through grant W-7405-ENG-36, with Los Alamos National Laboratory.

† Computer Research and Applications group, Los Alamos National Laboratory, Los Alamos, NM 87545. E-mail: vxf@lanl.gov, wdj@lanl.gov, knill@lanl.gov

‡ University of Colorado at Boulder, Boulder, CO. E-mail tmanteuf@newton.colorado.edu

or

$$r^{(n)} = [I - Aq_{n-1}(A)]r^{(0)} \tag{3}$$

where $u^{(0)}$ is the initial guess, $\{u^{(i)}\}_{i\geq 0}$ denote iterates, $r^{(i)} = b - Au^{(i)}$ are the associated residuals, and each q_{n-1} is a polynomial of degree no greater than $n-1$. Examples of such methods are the conjugate gradient method, the biconjugate gradient method, the minimal residual method, and polynomial preconditioned conjugate gradient methods (see [Ashby, Manteuffel and Saylor, 1990], [Joubert and Manteuffel, 1990] for overviews of such methods).

Polynomial methods owe their strength to the fact that the properties of polynomials lend themselves to rapid convergence rates for many cases, in particular when A is Hermitian and positive definite (HPD). However, a comprehensive theory of convergence of polynomial methods for general matrices has remained elusive. The purpose of this paper is to address the issue of convergence rates of some of these methods.

A natural choice for a polynomial method is to require that q_{n-1} be "optimal", in some sense. For example, given A, b and $u^{(0)}$, let

(4) q_{n-1} be a polynomial (cf. (3)) of degree at most $n-1$ which minimizes $||r^{(n)}||$,

where $||\cdot||$ is used here and throughout to refer to the standard 2-norm. This defines the *minimal residual method*, of which the GMRES algorithm is the best-known implementation [Saad and Schultz, 1986]. To limit the average work per iteration, this method is typically restarted every s steps, leading to algorithms such as GMRES(s). The resulting method is

$$r^{(ms+s)} = [I - Aq_{s-1;m}(A)]r^{(ms)}, \quad q_{s-1;m} \text{ selected by (4) based on } r^{(ms)}. \tag{5}$$

The average work per iteration for such algorithms applied to general matrices is proportional to sN; larger values of s generally improve convergence but also increase the work per iteration.

A considerably cheaper algorithm is *polynomial preconditioning* coupled with the basic one-step iterative method, namely

$$r^{(ms)} = [I - Aq_{s-1}(A)]^m r^{(0)}, \tag{6}$$

where the polynomial q_{s-1} is chosen in some appropriate fashion. (Of course, polynomial preconditioning can also be accelerated, for example by applying GMRES to the preconditioned system $q_{s-1}(A)Au = q_{s-1}(A)b$.) Provided that a good polynomial q_{s-1} can be found, this algorithm requires only order N work per iteration, independent of s. Furthermore, the algorithm can be a big win on certain computer architectures for which inner product computations are particularly expensive, since GMRES requires inner product computations but polynomial preconditioning does not.

Good polynomials q_{s-1} are not always easy to find, so we consider here the *optimal polynomial preconditioning* of degree $s-1$ for a matrix A, defined as a polynomial q_{s-1} of degree no greater than $s-1$ which solves the minimization problem

$$\text{minimize } ||I - Aq_{s-1}(A)||. \tag{7}$$

It can be shown that such a minimizer exists and under reasonable assumptions in fact is unique [Greenbaum and Trefethen, 1992].

The performance of this preconditioner is in some sense the best possible for a polynomial preconditioner. However, it should be noted that more sophisticated optimization procedures might be considered, such as

$$\text{minimize } ||\prod_{i=1}^{m}[I - Aq_{s-1;i}(A)]r^{(0)}||, \tag{8}$$

$$\text{minimize } ||[I - Aq_{s-1}(A)]^m||, \tag{9}$$

which may in some cases yield faster convergence. In particular, (8) selects a set of polynomials to give optimality globally over all s-step cycles rather than locally for each cycle (as GMRES(s) does), and (9) selects a single polynomial preconditioner which performs well over an aggregated set of cycles without regard to its single-cycle performance. The study of the convergence behavior of these methods is beyond the scope of this paper.

The methods (5) and (6,7) are similar, but they differ in the following important respect: (6,7) uses the same polynomial repeatedly, whereas (5) selects the best polynomial for each cycle. If an adequate polynomial can be found, then (6,7) is much more economical than (5). This is true especially when s is large, which is usually desirable in order to increase the convergence rate [Joubert, 1994b]. However, it is not clear whether (6,7) converges as fast as (5). The purpose of this study is to investigate the relative rates of convergence of (6,7) compared to (5).

It can be shown that the convergence behavior of the restarted minimal residual method is bounded by

$$\frac{||r^{(ms)}||}{||r^{(0)}||} \leq \left[\max_r \min_{q_{s-1}} ||[I - Aq_{s-1}(A)]r||\right]^m, \tag{10}$$

whereas the convergence behavior of the basic iterative method applied to optimal polynomial preconditioning is bounded by

$$\frac{||r^{(ms)}||}{||r^{(0)}||} \leq \left[\min_{q_{s-1}} ||I - Aq_{s-1}(A)||\right]^m = \left[\min_{q_{s-1}} \max_r ||[I - Aq_{s-1}(A)]r||\right]^m. \tag{11}$$

This motivates the question of the relative behavior of

$$\psi_n(A) = \max_r \min_{q_{s-1}} ||[I-Aq_{s-1}(A)]r|| \qquad \text{and} \qquad \varphi_n(A) = \min_{q_{s-1}} \max_r ||[I-Aq_{s-1}(A)]r||. \tag{12}$$

The two functions $\psi_n(A)$ and $\varphi_n(A)$ will be used as measures of the convergence behavior of these two popular iterative methods.

We should say a few words about the tightness of the bounds (10), (11). It is not clear that the inequality (10) is sharp, in the sense that for every A, m and s there is an $r^{(0)}$ for which 10 is an equality. This difficulty is due to the nonlinear nature of the minimization process (5). However, it does hold that $\psi_s(A) = 1$ if and only if there is an $r^{(0)}$ such that $r^{(0)} = r^{(s)} = r^{(2s)} \ldots$, i.e. the iterative method stagnates. Similarly, bound (11) may not be sharp, and in view of (9), polynomial preconditioners may exist which have better multi-cycle convergence than the (locally) optimal polynomial preconditioner described here. However, the (locally) optimal polynomial preconditioner is in some sense based on the best information known for a single cycle, and $\varphi_s(A) = 1$ if and only if this polynomial preconditioner coupled with the basic iterative method stagnates for some $r^{(0)}$. Furthermore, this assumes the optimal preconditioner can be economically found; more standard preconditioners may give even worse performance.

The comparison of $\psi_n(A)$ and $\varphi_n(A)$ can tell us whether replacing the more strongly convergent GMRES with the faster polynomial preconditioning can be done without destroying convergence. For some classes of matrices, e.g. HPD matrices and normal matrices (i.e. matrices A for which $AA^* = A^*A$, where * denotes conjugate transpose—this includes Hermitian, skew-Hermitian, unitary, and circulant matrices, for example), it is known that $\psi_n(A) = \varphi_n(A)$, so both methods have the same convergence rate for such matrices ([Greenbaum, 1979], [Joubert, 1994a], [Greenbaum and Gurvits, 1994]). In this paper we

show further that the class of upper triangular Toeplitz matrices A satisfy $\psi_n(A) = 1$ iff $\varphi_n(A) = 1$, that is, replacing GMRES(s) with the optimal polynomial preconditioning of degree $s-1$ cannot cause stagnation. On the other hand, we do give an example over the real numbers of a matrix for which restarted GMRES(s) converges but the optimal polynomial preconditioning of degree $s-1$ can stagnate. That is, GMRES is overall a more robust iterative method than the corresponding polynomial preconditioning.

Here is an outline of the remainder of the paper. In Section 2 a general theoretical framework for ψ_n and φ_n is established, various elementary results are obtained and known results are summarized. In Section 3 the results for Toeplitz matrices are presented, and in Section 4 an example for which $\psi_n \neq \varphi_n$ is given. Implications of this result are discussed in Section 5.

2. General Results on Convergence. This section gives the basic framework of tools used to analyze the convergence behavior of these iterative methods. Furthermore, a combination of existing and new results is given on the convergence behavior of the minimal residual method and optimal polynomial preconditioning.

2.1. Convergence Bounds: Definitions and Elementary Results. The convergence bounds for the minimal residual method and for optimal polynomial preconditioning are given below. These definitions are slightly more general than the definitions given in Section 1, in that they differentiate between the solution of real and complex linear systems.

Let $\mathbb{K}$ denote either the field of real numbers $\mathbb{R}$ or the complex numbers $\mathbb{C}$. Let $\mathbb{K}_i[z]$ denote polynomials over $\mathbb{K}$ of degree no greater than i. Then for $A \in \mathbb{K}^{N\times N}$, let

$$\begin{aligned}
\varphi_{n,\mathbb{K}}(A) &= \inf_{q\in\mathbb{K}_{n-1}[z]} \sup_{v\in\mathbb{K}^N:||v||=1} ||(I - Aq(A))v|| \\
&= \inf_{q\in\mathbb{K}_{n-1}[z]} ||I - Aq(A)||, \\
\psi_{n,\mathbb{K}}(A) &= \sup_{v\in\mathbb{K}^N:||v||=1} \inf_{q\in\mathbb{K}_{n-1}[z]} ||(I - Aq(A))v||.
\end{aligned}$$

For both φ and ψ, when $\mathbb{K} = \mathbb{R}$, the infimum can be taken over either real or complex polynomials without affecting the values of φ and ψ [Joubert, 1994b].

Let us now confirm that in fact the convergence bound for the minimal residual method is at least as strong as that for optimal polynomial preconditioning. The proposition also sheds some light on what happens to the bounds when A is singular. Define the degree of a matrix $d(A)$ as $\min\{\deg(P) : P(A) = 0,\ P \text{ monic}\}$. Then

PROPOSITION 2.1. *Let $A \in \mathbb{K}^{N\times N}$. Then $0 \le \psi_{n,\mathbb{K}}(A) \le \varphi_{n,\mathbb{K}}(A) \le 1$. If $d = d(A) \le N$ is the degree of the minimal polynomial of A, then $0 < \psi_{n,\mathbb{K}}(A) \le \varphi_{n,\mathbb{K}}(A)$ for any $n < d$. For $n \ge d$ and for A also nonsingular, $\psi_{n,\mathbb{K}}(A) = \varphi_{n,\mathbb{K}}(A) = 0$. If A is singular, then $\psi_{n,\mathbb{K}}(A) = \varphi_{n,\mathbb{K}}(A) = 1$ for any n.*

Proof. The first inequality is easily shown; see [Joubert, 1994b]. The result for $n < d$ is shown as follows. If $\psi_{n,\mathbb{K}}(A) = 0$, then for all $v \in \mathbb{K}^N$, $||v|| = 1$, $\inf_q ||(I - Aq(A))v|| = 0$. It is easily seen that if v is chosen to contain nonzero components of all generalized eigenvalues of A, this leads to a contradiction. For $n \ge d$, if A is invertible, the monic minimal polynomial for A can be renormalized so that the constant term is 1. If A is not invertible, then v can be chosen from the null space of A, and $(I - Aq(A))v = v$ for any q. □

Define $f_n : \mathbb{K}^N \times \mathbb{K}^{N\times N} \to \mathbb{K}$ by

$$f_n(v, A) = \inf_{q\in\mathbb{K}_{n-1}[x]} ||(I - Aq(A))v||^2.$$

This function defines the convergence of the minimal residual method applied to a specific vector: for $A \in \mathbb{K}^{N\times N}$, $\psi_{n,\mathbb{K}}(A)^2 = \sup_{v\in\mathbb{K}^N} f_n(v, A)/||v||^2$.

Let $\underline{K}_n(v, A) \in \mathbb{C}^{N\times n}$ be defined by $\underline{K}_n(v, A)e_i = A^{i-1}v$, where e_i is the standard unit basis vector. Also define the degree of a vector $d(v, A)$ as $\min\{\deg(P) : P(A)v = 0,\ P \text{ monic}\}$. Note $A\underline{K}_n(v, A) = \underline{K}_n(Av, A)$ is full rank if and only if $d(Av, A) \ge n$, and when A is nonsingular $d(Av, A) = d(v, A)$. Then for $A\underline{K}_n(v, A)$ full rank,

$$f_n(v, A) = \hat{f}_n(v, A) \equiv v^*v - v^*A\underline{K}_n(v, A)[\underline{K}_n(v, A)^*A^*A\underline{K}_n(v, A)]^{-1}\underline{K}_n(v, A)^*A^*v. \tag{13}$$

PROPOSITION 2.2. *For fixed nonsingular A, $f_n(\cdot, A)$, considered as a function of $2N$ real variables under the identification $\mathbb{R}^{2N} \cong \mathbb{C}^N$, is C^∞ on the complement of the closed set $S = \{v : d(Av, A) < n\}$. If A is nonsingular then $f_n(\cdot, A)$ is continuous everywhere, and if furthermore $n \le d(A)$ then S has measure zero.*

Proof. The first statement follows from the above remarks and the fact that $\hat{f}_n(v, A)$ is a rational function of the real and imaginary parts of the elements of v and A. When A is nonsingular and $n \ge d(A)$, $f_n(\cdot, A)$ is uniformly zero. Otherwise, it suffices to show that for $v_i \notin S$ and $v_i \to v \in S$, $f_n(v_i, A) \to f_n(v, A)$. Let $p(A)$ be the minimal polynomial for v with respect to A; note that $\deg p < n$. Let $\tilde{p}(z) = p(z)/p(0)$. Then $0 \le f_n(v_i, A) \le ||\tilde{p}(A)v_i||^2 \to 0 = f_n(v, A)$ □

2.2. Continuity of the Bound Functions. To understand the convergence of polynomial iterative methods, it is desirable to get a better understanding of how the bound functions $\psi_{n,\mathbb{K}}$ and $\varphi_{n,\mathbb{K}}$ behave. See [Joubert, 1994b] and [Joubert, 1994a] for elementary results on these functions. In what follows we demonstrate in particular that $\psi_{n,\mathbb{K}}$ and $\varphi_{n,\mathbb{K}}$ are continuous functions on the open set of nonsingular matrices. It will be noted, however, that they are not generally continuous everywhere; for example, $\psi_{N,\mathbb{K}} = \varphi_{N,\mathbb{K}} = 1$ for A singular but $\psi_{N,\mathbb{K}} = \varphi_{N,\mathbb{K}} = 0$ for A nonsingular.

We begin with the following

LEMMA 2.3. *For A, $A_i \in \mathbb{C}^{N\times N}$ with $A_i \to A$, and for r, $r_i \in \mathbb{C}^N$ with $r_i \to r$, $\liminf_{i\to\infty} d(r_i, A_i) \geq d(r, A)$.*

Proof. Let P_i be the minimal polynomial for r_i with respect to A_i, and let P be the minimal polynomial for r with respect to A. Note first that the eigenvalues of all $\{A_i\}$ form a bounded set: if $\{\lambda, v\}$ is an eigenpair of A_i, then $|\lambda| = ||A_i v||/||v|| \leq ||A_i||$, but since $||A_i||$ is bounded near $||A||$, $|\lambda|$ must be bounded. Thus the polynomials P_i must reside within a bounded set, because the coefficients of P_i are products and sums of the eigenvalues of A_i.

Suppose there exists a subsequence i_j such that $\deg(P_{i_j}) = d(r_{i_j}, A_{i_j}) < d(r, A)$ for all j. By boundedness, this subsequence has a convergent subsequence P_{i_k} with $P_{i_k} \to \hat{P}$ for some polynomial $\hat{P}$. We note that necessarily $\deg(\hat{P}) < d(r, A)$, by the choice of P_{i_j}. Furthermore,

$$||\hat{P}(A)r|| \leq ||\hat{P}(A)r - P_{i_k}(A)r|| + ||P_{i_k}(A)r - P_{i_k}(A_{i_k})r_{i_k}||$$

since $P_{i_k}(A_{i_k})r_{i_k} = 0$. Since $P_{i_k} \to \hat{P}$, $||\hat{P}(A)r - P_{i_k}(A)r|| \to 0$. Furthermore, since the $\{P_{i_k}\}$ are bounded, $A_{i_k} \to A$ and $r_{i_k} \to r$, we have for $P_{i_k}(z) = \sum c_{i_k j} z^j$, $||P_{i_k}(A)r - P_{i_k}(A_{i_k})r_{i_k}|| \leq \sum |c_{i_k,j}| \cdot ||A^j r - A^j_{i,k} r_{i_k}|| \to 0$. Thus $\hat{P}(A)r = 0$, a contradiction. □

Note that in fact we may have $d(r_i, A_i) \to d > d(r, A)$, e.g. $A = \mathrm{diag}[1 + 1/i, 1 + 2/i]$, $r_i = [1, 1]^T$.

LEMMA 2.4. *Let $f : S_1 \times S_2 \to \mathbb{R}$ be continuous, and let $S_1 \subseteq \mathbb{C}^{n_1}$ compact and $S_2 \subseteq \mathbb{C}^{n_2}$ open. Then $F(u) = \sup_{v\in S} f(v, u)$ and $G(u) = \inf_{v\in S} f(v, u)$ are continuous.*

Proof. We prove the result for F. Note $F(u) < \infty$ for any u. Let $u_i \to u$, with u_i, $u \in S_2$. We first show $\sup_{v\in S} |f(v, u_i) - f(v, u)| \to 0$. Otherwise, there exists $\epsilon > 0$, a subsequence i_j and vectors $v_{i_j} \in S$ such that $|f(v_{i_j}, u_{i_j}) - f(v_{i_j}, u)| \geq \epsilon$ for all j. By the

compactness of S_1, there exists i_k a subsequence of i_j such that $v_{i_k} \to \hat{v} \in S$. Then

$$|f(v_{i_k}, u_{i_k}) - f(v_{i_k}, u)| \leq |f(v_{i_k}, u_{i_k}) - f(\hat{v}, u)| + |f(\hat{v}, u) - f(v_{i_k}, u)| \to 0$$

by the continuity of f, a contradiction.

Thus for any $\epsilon > 0$ there is some m such that for all $i > m$,

$$f(v, u) - \epsilon < f(v, u_i) < f(v, u) + \epsilon$$

for any $v \in S_1$. Taking suprema over S_1 yields

$$F(u) - \epsilon \leq F(u_i) \leq F(u) + \epsilon$$

implying $|F(u) - F(u_i)| \leq \epsilon$, giving the result. □

THEOREM 2.5. *The function $\psi_{n,\mathbb{K}}(\cdot)$ is continuous on the open set of nonsingular matrices in $\mathbb{K}^{N\times N}$.*

Proof. Note $\psi_{n,\mathbb{K}}(A)^2 = \sup_{v\in\mathbb{K}^N, ||v||=1} f_n(v, A)$, where f_n is as defined earlier. If $n \geq d(A)$, then $\psi_{n,\mathbb{K}}(A) = 0$ and the result follows by letting P_n be a scaling of the minimal polynomial for A: $0 \leq \psi_{n,\mathbb{K}}(A_i) \leq \varphi_{n,\mathbb{K}}(A_i) \leq ||P_n(A_i)|| \to 0$ for $A_i \to A$. Otherwise we proceed as follows.

Since the set of nonsingular matrices constitute an open set in $\mathbb{K}^{N\times N}$, the previous Lemma will give the result if it can be shown that the map $f_n(\cdot, \cdot)$ is continuous on $\{r : ||r|| = 1\} \times \{A : A \text{ nonsingular}\}$. Select A, A_i nonsingular and $||r|| = ||r_i|| = 1$, and let $A_i \to A$ and $r_i \to r$. If $d(r, A) \leq n$, then for P_n a scaling of the minimal polynomial of r with respect to A, $0 \leq f_n(r_i, A_i) \leq ||P_n(A_i)r_i|| \to 0$. Otherwise, by previous Lemma, for some m, $d(r_i, A_i) \geq d(r, A) > n$ for all $i > m$. Then $f_n(r_i, A_i) = \hat{f}_n(r_i, A_i)$ for all $i > m$, and $f_n(r, A) = \hat{f}_n(r, A)$. Since the rational function $\hat{f}_n(\cdot, \cdot)$ is continuous on the open set of elements for which it is defined, the result follows. □

THEOREM 2.6. *The function $\varphi_{n,\mathbb{K}}(\cdot)$ is continuous on the open set of nonsingular matrices in $\mathbb{K}^{N\times N}$.*

Proof. Since $\varphi_{n,\mathbb{R}}(A) = \varphi_{n,\mathbb{C}}(A)$ for $A \in \mathbb{R}^{N\times N}$, the result need only be shown for $\mathbb{K} = \mathbb{C}$. We have

$$\varphi_{n,\mathbb{C}}(A) = \inf_{P_n(0)=1} ||P_n(A)||.$$

As in the previous theorem, assume $n < d(A)$. Note that $\varphi_{n,\mathbb{C}}(A) = \inf_{P_n\in S} ||P_n(A)||$, where $S = \{P_n \in \mathbb{C}_n[z] : P_n(0) = 1, ||P_n(A)|| \leq 2\}$, and as usual $\deg P_n \leq n$. Note

that S is nonempty: $P(z) = 1$ defines a polynomial found in S. Importantly, by the linear independence of $\{A^i\}_{i=0}^n$, the set S is bounded. Also note S is closed. Thus, using the notation of Lemma 2.4, we let $f(P_n, A) = ||P_n(A)||$ to obtain the result via that Lemma. □

The previous two theorems confirm the continuity of the bound functions. This is an expected result: a small perturbation of the matrix A should cause only a small perturbation in the behavior of the iterative method.

2.3. The Generalized Field of Values. In the study of conjugate gradient methods, the field of values $F(A) = \{x^*Ax : x \in \mathbb{C}^N, ||x|| = 1\}$ [Householder, 1964] plays a prominent role in determining convergence behavior of the methods. In this study we will make use of the concept of the *generalized* field of values (see e.g. [Horn and Johnson, 1991]) to develop a more powerful set of results.

Let us first establish some notation. For any matrices $A = \{a_{i,j}\} \in \mathbb{K}^{r_a \times c_a}$ and $B = \{b_{i,j}\} \in \mathbb{K}^{r_b \times c_b}$, define the standard *tensor product of matrices* $A \otimes B \in \mathbb{K}^{r_a r_b \times c_a c_b}$ by $e^*_{j+r_b(i-1)}(A \otimes B)e_{l+c_b(k-1)} = a_{i,k}b_{j,l}$ [Lancaster and Tismenetsky, 1985]. Note that when A, B, C and D are matrices of dimension such that AC and BD are well-defined, then $(A \otimes B)(C \otimes D) = AC \otimes BD$.

For a set of matrices $\{A_i\}_{i=1}^n \subseteq \mathbb{C}^{N \times N}$, the matrix $R = \sum_{i=1}^n e_i \otimes A_i$ can be considered as a vector of quadratic forms $v \mapsto (I \otimes v^*)R(1 \otimes v) \in \mathbb{C}^n$. Thus R may may be thought of as representing a map from $\mathbb{C}^N$ to $\mathbb{C}^n$.

Let the *generalized field of values* over a field $\mathbb{K}$ for a set of matrices $\{A_i\}_{i=1}^n \subseteq \mathbb{C}^{N \times N}$ be defined by

$$F_{\mathbb{K}}(\{A_i\}_{i=1}^n) = \{(I \otimes v^*)(\textstyle\sum_{i=1}^n e_i \otimes A_i)(1 \otimes v) : \ v \in \mathbb{K}^N, \ ||v|| = 1\}$$

$$= \{\textstyle\sum_{i=1}^n e_i v^* A_i v : \ v \in \mathbb{K}^N, \ ||v|| = 1\} \subseteq \mathbb{C}^n.$$

Note that the quantity $F_{\mathbb{C}}(\{A\})$ coincides with the standard field of values of a matrix [Householder, 1964].

Also define the *conical* generalized field of values,

$$\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n) = \{\sum_{i=1}^n e_i v^* A_i v : \ v \in \mathbb{K}^N\} \subseteq \mathbb{C}^n.$$

It is clear that this object is a cone, i.e., for real $\alpha > 0$, $f \in \check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n) \Rightarrow \alpha f \in \check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n)$. Note $1 \oplus F_{\mathbb{K}}(\{A_i\}_{i=1}^n) = H_{\mathbb{K}}(e_1) \cap \check{F}_{\mathbb{K}}(\{I\} \cup \{A_i\}_{i=1}^n)$, where $H_{\mathbb{K}}(v)$ denotes

the hyperplane $\{u \in \mathbb{K}^N | v^* u = 1\}$ for a vector $v \in \mathbb{K}^N$, and more generally $H_{\mathbb{K}}(v, r_0) = \{u \in \mathbb{K}^N : v^* u = r_0\}$. Note also that the conical field of values is preserved by simultaneous congruence transformation: for $P \in \mathbb{K}^{N \times N}$ nonsingular, $\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n) = \check{F}_{\mathbb{K}}(\{P^* A_i P\}_{i=1}^n)$.

We may now use the concept of generalized field of values to find characterizations of when $\psi_{n,\mathbb{K}}(A)$ or $\varphi_{n,\mathbb{K}}(A)$ is equal to 1, i.e. when the corresponding iterative methods can stagnate. This allows the performance of these methods to be studied in terms of the geometric properties of these objects, in particular, their convexity properties, as in the case of the standard field of values.

The following two theorems characterize stagnation of the methods in terms of properties of the generalized field of values.

THEOREM 2.7. *For nonsingular matrices $A \in \mathbb{K}^{N \times N}$, $\psi_{n,\mathbb{K}}(A) = 1$ if and only if $0 \in F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$.*

Proof. Suppose that $\psi_{n,\mathbb{K}}(A) = 1$. By continuity of $f_n(\cdot, A)$ and by compactness, there exists a vector $v \in \mathbb{K}^N$, $||v|| = 1$, such that $f_n(v, A) = \inf_{q \in \mathbb{K}_{n-1}[z]} ||(I - Aq(A))v|| = 1$. Note that for such v, $d(v, A) > n$. In view of the definition of $\hat{f}_n$ (13), this can hold only if v is perpendicular to the space generated by $Av, A^2 v, \ldots, A^n v$, which implies that $v^* A^i v = 0$ for $1 \leq i \leq n$. Thus $0 \in F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$.

Conversely, if $0 \in F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$, then for some nonzero $v \in \mathbb{K}^N$ with $||v|| = 1$, $v^* A^i v = 0$ for $1 \leq i \leq n$. This implies that v is perpendicular to each $A^i v$ for $1 \leq i \leq n$ and for any q,

$$||(I - Aq(A))v||^2 = ||v - q_0 Av - q_1 A^2 v - \ldots - q_{n-1} A^2 v||^2 = ||v||^2 + ||Aq(A)v||^2 \geq 1.$$

The result follows. □

THEOREM 2.8. *For nonsingular matrices $A \in \mathbb{K}^{N \times N}$, $\varphi_{n,\mathbb{K}}(A) = 1$ if and only if $0 \in \mathrm{cvx}(F_{\mathbb{K}}(\{A^i\}_{i=1}^n))$, the convex hull of the generalized field of values.*

Proof. By the Hahn-Banach theorem, $0 \notin \mathrm{cvx}(F_{\mathbb{K}}(\{A^i\}_{i=1}^n))$ iff there exists a hyperplane separating 0 from $F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$, i.e. for some $c \in \mathbb{K}^n$, $\mathrm{Re}\, c^* w > 0$ for all $w \in F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$. (In the complex case, let $c = c_r + ic_i$ and $w = w_r + iw_i \in F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$, so that $\mathrm{Re}\, c^* w = c_r^* w_r + c_i^* w_i$. Let $\hat{c} = c_r \oplus c_i$ and $\hat{w} = w_r \oplus w_i$. Then $\mathrm{Re}\, c^* w = \hat{c}^* \hat{w}$, and the result follows from the Hahn-Banach theorem in $\mathbb{R}^{2n}$.)

Let $C(c, z) = \sum_{i=1}^n (c^* e_i) z^{i-1}$. Then $\mathrm{Re}(c^* F_{\mathbb{K}}(\{A^i\}_{i=1}^n) = \mathrm{Re}(F_{\mathbb{K}}(AC(c, A)))$.

Suppose $\varphi_{n,\mathbb{K}}(A) < 1$. Then there exist $P(z) = 1 - zq(z)$ and ρ such that $||P(A)v||^2 = 1 - 2\,\mathrm{Re}\, v^* Aq(A)v + ||Aq(A)v||^2 \leq \rho < 1$ for every $v \in \mathbb{K}^N$ with $||v|| = 1$. Thus $1 -$

$2\,\mathrm{Re}\,v^*Aq(A)v \le \rho$ for such v, or $\mathrm{Re}\,v^*Aq(A)v \ge (1-\rho)/2$. Defining c by $q(z) = C(c,z)$, we obtain $\mathrm{Re}\,c^*F_{\mathbb{K}}(\{A^i\}_{i=1}^n) \ge (1-\rho)/2 > 0$, giving $0 \notin \mathrm{cvx}(F_{\mathbb{K}}(\{A^i\}_{i=1}^n))$.

Suppose there exists $C(c,z) = \sum(c^*e_i)z^{i-1}$ with $\mathrm{Re}(F_{\mathbb{K}}(AC(c,A))) > 0$. We have $||[I-\epsilon AC(c,A)]v||^2 = ||v||^2 - 2\epsilon\,\mathrm{Re}\,v^*AC(c,A)v + \epsilon^2||AC(c,A)v||^2$ for $\epsilon > 0$. By compactness, there exists $\delta > 0$ such that $\mathrm{Re}(F_{\mathbb{K}}(AC(c,A))) > \delta$. Hence, for a sufficiently small ϵ, $||[I-\epsilon AC(c,A)]v||^2 < 1-\epsilon\delta$ for all v with $||v|| = 1$. Thus $\varphi_{n,\mathbb{K}}(A) < 1$. □

The principles behind these two theorems will be used heavily in Section 4 to construct the counterexample. In particular, note that if the generalized field of values associated with the powers of A is convex, then either both methods converge or both diverge. On the other hand, if the generalized field of values is nonconvex at the origin, in the sense of Theorem 2.8, then restarted GMRES will necessarily converge but the associated polynomial preconditioning may diverge.

2.4. Bounds Based on the Generalized Field of Values. Theorems 2.7 and 2.8 can be quantified in terms of the distances between 0 and either of the sets $F_{\mathbf{K}}(\{A^i\}_{i=1}^n)$ and $\mathrm{cvx}[F_{\mathbf{K}}(\{A^i\}_{i=1}^n)]$. In particular, the convergence rates of the two methods can be bounded in terms of these distances.

First consider the map $\Gamma_A : \mathbb{K}^n \to \mathbb{K}^{N\times N}$ defined by $\Gamma_A(c) = AQ(A)$, where $Q(z) = \sum_{i=1}^n (e_i^*c)z^{i-1}$. Then, Γ_A is a bounded map under the induced norm

$$||\Gamma_A|| = \sup_{c\in\mathbb{K}^n,||c||=1} ||AQ(A)|| = \sup_{c\in\mathbb{K}^n,||c||=1}\ \sup_{v\in\mathbb{K}^N,||v||=1} ||AQ(A)v||.$$

Observe that $||\Gamma_A|| \le \sum_{i=1}^n ||A^i||$.

The following two theorems give bounds on $\psi_{n,\mathbb{K}}(A)$ and $\varphi_{n,\mathbb{K}}(A)$. Here, let $A \in \mathbb{K}^{N\times N}$, and let $S_{n,\mathbb{K}}$ denote the sphere in $\mathbb{K}^n$.

THEOREM 2.9. *Let $\eta = \sup\{\rho \ge 0 : (\rho\cdot S_{n,\mathbb{K}}) \cap F_{\mathbb{K}}(\{A^i\}_{i=1}^n) = \{\}\}$, the distance from $F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$ to the origin. Then $\psi_{n,\mathbb{K}}(A) \le (1-(\eta/||\Gamma_A||)^2)^{1/2}$.*

Proof. Given $v \in \mathbb{K}^N$, $||v|| = 1$, let $w = v^*A\underline{K}_n(v,A)$ and let c satisfy $\mathrm{Re}\,c^*w \ge \eta$, $||c|| = 1$, for example by $c = w/||w||$. Let $Q(z) = \sum_{i=1}^n (e_i^*c)z^{i-1}$. Then let $\hat{Q} = \epsilon Q$ for $\epsilon > 0$.

$$||(I - A\hat{Q}(A))v||^2 = v^*v - 2\epsilon\,\mathrm{Re}\,v^*AQ(A)v + \epsilon^2||AQ(A)v||^2 \le v^*v - 2\epsilon\eta + \epsilon^2||\Gamma_A||^2.$$

Setting $\epsilon = \eta/||\Gamma_A||^2$ gives the result. □

THEOREM 2.10. *Let* $\eta = \sup\{\rho \geq 0 : \rho \cdot S_{n,\mathbb{K}} \cap \mathrm{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n] = \{\}\}$, *the distance from* $\mathrm{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n]$ *to the origin. Then* $\varphi_{n,\mathbb{K}}(A) \leq (1 - (\eta/||\Gamma_A||)^2)^{1/2}$.

Proof. The result is trivial for $\eta = 0$. Otherwise, note that

$$\eta \cdot S_{n,\mathbf{K}} \cap \mathrm{cvx}[F_{\mathbf{K}}(\{A^i\}_{i=1}^n]$$

must contain exactly one point. It has at least one point since $f(v) = ||v^* A \underline{K}_n(v, A)||$ must attain its infimum on $\{v : ||v|| = 1\}$. On the other hand, if the intersection contains two distinct points γ_1 and γ_2, then $(\gamma_1 + \gamma_2)/2$ is in the interior of $\eta \cdot S_{n,\mathbf{K}}$ and also in $\mathrm{cvx}[F_{\mathbf{K}}(\{A^i\}_{i=1}^n]$ due to convexity, a contradiction due to closedness. Now, let γ be the point in this intersection.

We claim that for $c = \gamma/||\gamma||$, $c^* w \geq \eta$ for every $v \in \mathbb{K}^N$, $||v|| = 1$, with $w = v^* A \underline{K}_n(v, A)$. This follows if we can show the result for all $w \in \mathrm{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)]$. Otherwise there exists $w \in \mathrm{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)]$ with $\mathrm{Re}\, \gamma^* w < ||\gamma||^2$, where $||\gamma|| = \eta$. Let $w_\epsilon = \epsilon w + (1 - \epsilon)\gamma$, $0 \leq \epsilon \leq 1$. Note $w_\epsilon \in \mathrm{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)]$ for any such ϵ. Furthermore,

$$||w_\epsilon||^2 = ||\gamma||^2 - 2\epsilon||\gamma||^2 + 2\epsilon\, \mathrm{Re}\, \gamma^* w + \mathcal{O}(\epsilon^2) < ||\gamma||^2$$

for sufficiently small ϵ. For this ϵ, w_ϵ is in the interior of $\eta \cdot S_{n,\mathbb{K}}$ and also in $\mathrm{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n]$, a contradiction.

Defining Q and $\hat{Q}$ as in the proof of the previous theorem we see that

$$||(I - A\hat{Q}(A))v||^2 = v^* v - 2\epsilon\, \mathrm{Re}\, v^* AQ(A)v + \epsilon^2 ||AQ(A)v||^2 \leq v^* v - 2\epsilon\eta + \epsilon^2 ||\Gamma_A||^2,$$

for every $v \in \mathbb{K}^N$. Setting $\epsilon = \eta/||\Gamma_A||^2$ gives the result. □

These bounds will become useful in the numerical example given later.

2.5. Deriving Results for Other Matrices. This subsection gives a collection of results which add further insight on the behavior of the bound functions and will be useful later for extending results to wider classes of matrices.

The first result shows that the generalized field of values is convex and the bound functions are equal for normal matrices.

THEOREM 2.11. *For* $A \in \mathbb{K}^{N \times N}$ *nonsingular and normal (e.g. Hermitian, real symmetric, skew-Hermitian, real skew-symmetric, unitary, or circulant),* $\check{F}_{\mathbf{K}}(\{A^i\}_{i=0}^n)$ *is convex, and in fact* $\psi_{n,\mathbb{K}}(A) = \varphi_{n,\mathbf{K}}(A)$.

Proof. See [Joubert, 1994a], [Greenbaum and Gurvits, 1994], [Joubert, 1990]. □

The following result shows that for a block diagonal matrix, the convergence rate of either of the methods is no better than the convergence rate of that method for any of the diagonal submatrices of the block diagonal matrix.

THEOREM 2.12. *For* $A_i \in \mathbb{K}^{N_i \times N_i}$, $i = 1, 2$, $\psi_{n,\mathbb{K}}(A_1) \leq \psi_{n,\mathbb{K}}(\mathrm{diag}[A_1, A_2])$ *and* $\varphi_{n,\mathbb{K}}(A_1) \leq \varphi_{n,\mathbb{K}}(\mathrm{diag}[A_1, A_2])$. *Furthermore,* $\check{F}_{\mathbb{K}}(\{A_1^i\}_{i=0}^n) \subseteq \check{F}_{\mathbb{K}}(\{\mathrm{diag}[A_1, A_2]^i\}_{i=0}^n)$.

Proof. The first result follows easily from

$$\psi_{n,\mathbb{K}}(A_1) = \sup_{v \in \mathbb{K}^{N_1} : ||v||=1} \inf_{P \in \mathbb{C}_n[x]:P(0)=1} ||P(A)v||$$

$$= \sup_{v \in \mathbb{K}^{N_1} : ||v||=1} \inf_{P \in \mathbb{C}_n[x]:P(0)=1} ||P(\begin{bmatrix} A_1 & \\ & A_2 \end{bmatrix}) \begin{bmatrix} v \\ 0 \end{bmatrix}||$$

$$\leq \sup_{v \in \mathbb{K}^{N_1+N_2} : ||v||=1} \inf_{P \in \mathbb{C}_n[x]:P(0)=1} ||P(\mathrm{diag}[A_1, A_2])v|| = \psi_{n,\mathbb{K}}(\mathrm{diag}[A_1, A_2]),$$

$$\varphi_{n,\mathbb{K}}(A_1) = \inf_{P \in \mathbb{C}_n[x]:P(0)=1} \sup_{v \in \mathbb{K}^{N_1} : ||v||=1} ||P(A)v||$$

$$= \inf_{P \in \mathbb{C}_n[x]:P(0)=1} \sup_{v \in \mathbb{K}^{N_1} : ||v||=1} ||P(\begin{bmatrix} A_1 & \\ & A_2 \end{bmatrix}) \begin{bmatrix} v \\ 0 \end{bmatrix}||$$

$$\leq \inf_{P \in \mathbb{C}_n[x]:P(0)=1} \sup_{v \in \mathbb{K}^{N_1+N_2} : ||v||=1} ||P(\mathrm{diag}[A_1, A_2])v|| = \varphi_{n,\mathbb{K}}(\mathrm{diag}[A_1, A_2]).$$

The subset inclusion result follows by similar line of argument. □

The next result shows that for the special case of a submatrix replicated down the main diagonal of a block diagonal matrix, the convergence rate for polynomial preconditioning is unchanged by the replication.

THEOREM 2.13. *For* $A \in \mathbb{K}^{N \times N}$ *and* I *an identity matrix of any size,* $\varphi_{n,\mathbb{K}}(A \otimes I) = \varphi_{n,\mathbb{K}}(A)$. *Furthermore, for* $\{A_i\}_{i=1}^n \subseteq \mathbb{K}^{N \times N}$, $\mathrm{cvx}[\check{F}_{\mathbb{K}}(\{A_i \otimes I\}_{i=1}^n)] = \mathrm{cvx}[\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n)]$.

Proof. Using the results on tensor products from [Lancaster and Tismenetsky, 1985],

$$\varphi_{n,\mathbb{K}}(A \otimes I) = \inf_P ||P(A \otimes I)|| = \inf_P \lambda_{max}[[P(A)^* P(A)] \otimes I]^{1/2}$$

$$= \inf_P \lambda_{max}[P(A)^* P(A)]^{1/2} = \varphi_{n,\mathbb{K}}(A).$$

To show the set equality, it is sufficient to note

$$\sum_i e_i \left[\sum_j v_j \otimes e_j\right]^* [A_i \otimes I] \left[\sum_j v_j \otimes e_j\right] = \sum_i e_i \sum_j [v_j \otimes e_j]^* [A_i \otimes I][v_j \otimes e_j]$$

$$= \sum_j \sum_i e_i v_j^* A_i v_j = \sum_j (1/n) \sum_i e_i (\sqrt{n} v_j)^* A_i (\sqrt{n} v_j) \in \mathrm{cvx}[\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n)].$$

□

The following weaker result holds for the minimal residual method.

THEOREM 2.14. *For $A_i \in \mathbb{K}^{N\times N}$ and I_N an identity matrix of size N, the set $\check{F}_{\mathbb{K}}(\{A_i \otimes I_N\}_{i=1}^n)$ is convex.*

Proof. We view $\check{F}_{\mathbf{K}}(\{A_i \otimes I_N\}_{i=1}^n)$ as the image of the composition of two maps where the first one has a convex range and the second one is linear.

First let $v = \sum_i v_i \otimes e_i$ for vectors $\{v_i\}_{i=1}^N \subseteq \mathbb{K}^N$. The first map ρ takes v to $\rho(v) = \sum_{i=1}^N v_i v_i^* \in \mathbb{K}^{N\times N}$. Note that the range of ρ is exactly the convex cone of Hermitian nonnegative definite matrices in $\mathbb{K}^{N\times N}$.

The second map σ takes a matrix $P \in \mathbb{K}^{N\times N}$ to $\sigma(P) = \sum_{i=1}^n e_i \mathrm{trace}(P A_i) \in \mathbb{K}^n$. This map is linear.

Let R be the map associated with $\check{F}_{\mathbb{K}}(\{A_i \otimes I_N\}_{i=1}^n)$. To complete the proof, we show that $\sigma(\rho(v)) = R(v)$:

$$R(v) = \sum_{i=1}^n e_i v^* (A_i \otimes I_N) v = \sum_{i=1}^n e_i \sum_{j=1}^N v_j^* A_i v_j = \sum_{i=1}^n e_i \mathrm{trace}\left(\left[\sum_{j=1}^N v_j v_j^*\right] A_i\right) = \sigma(\rho(v)).$$

□

We may also characterize situations in which $\psi_{n,\mathbf{K}}(A \otimes I_k) = \varphi_{n,\mathbb{K}}(A \otimes I_k)$. For this result, let us make the following definitions. Consider $\mathbf{A} = \sum_{i=1}^n e_i \otimes A_i$, $A_i \in \mathbb{K}^{N\times N}$. For our purposes, $A_i = A^i$. For $v \in \mathbb{K}^n$, define $v \cdot \mathbf{A} = [v^T \otimes I]\mathbf{A} = \sum_i [e_i^* v] A_i$. Define $\varphi_{\mathbb{K}}(\mathbf{A}) = \inf_v \sup_x ||(I + v \cdot \mathbf{A})x||$, $\psi_{\mathbb{K}}(\mathbf{A}) = \sup_x \inf_v ||(I + v \cdot \mathbf{A})x||$, where x and v are vectors over $\mathbb{K}$ and the suprema are over $||x|| = 1$.

Let us define the *restricted* generalized field of values: for such $\mathbf{A}$ and for M a subspace of $\mathbb{C}^N$, $F_M(\mathbf{A}) = \{\sum_i e_i [x^* A_i x] : ||x|| = 1, x \in M\}$. Also, let $\Sigma_{\mathbb{K}}(B) \subseteq \mathbb{K}^N$ be the subspace over $\mathbb{K}$ spanned by the (right) singular vectors associated with the maximal singular value of $B \in \mathbb{K}^{N\times N}$.

The following result generalizes Theorem 2.8.

LEMMA 2.15. *Given* $\mathbf{A} = \sum_i e_i \otimes A_i$, $A_i \in \mathbb{K}^{N\times N}$, v *is a minimizer of* $\varphi_{\mathbb{K}}(\mathbf{A})$ *if and only if for* $B = v \cdot \mathbf{A}$, $0 \in \operatorname{cvx} F_{\Sigma_{\mathbb{K}}(I+B)}([I \otimes [I+B]]^* \mathbf{A})$.

Proof. Following the proof of Theorem 2.8, suppose first that $0 \in \operatorname{cvx} F_{\Sigma_{\mathbb{K}}(I+B)}([I \otimes [I+B]]^* \mathbf{A})$. Then for any w there exists $x_w \in \Sigma_{\mathbb{K}}(I+B)$, $||x_w|| = 1$, such that $\operatorname{Re} x_w^*(I+B)^*[w \cdot \mathbf{A}]x_w = 0$. Thus for any w and for $\epsilon > 0$,

$$||(I + (v+\epsilon w)\cdot \mathbf{A})||^2 \geq ||(I + (v+\epsilon w)\cdot \mathbf{A})x_w||^2$$

$$= ||(I + v\cdot \mathbf{A})x_w||^2 + \epsilon^2 ||(w\cdot \mathbf{A})x_w||^2 \geq ||(I+v\cdot \mathbf{A})x_w||^2 = ||I + v\cdot \mathbf{A}||^2,$$

since the cross-term is zero. Thus v is a local minimizer. By a convexity argument it can be shown that v is thus a global minimizer.

If $0 \notin \operatorname{cvx} F_{\Sigma_{\mathbb{K}}(I+B)}([I \otimes [I+B]]^* \mathbf{A})$, then as before there is w such that for all $x \in \Sigma_{\mathbb{K}}(I+B)$ such that $||x|| = 1$, $\operatorname{Re} x^*(I+B)^*[w\cdot \mathbf{A}]x < 0$. Then

$$||(I+(v+\epsilon w)\cdot \mathbf{A})x||^2 = ||(I+B)x||^2 + 2\epsilon \operatorname{Re} x^*(I+B)^*[w\cdot \mathbf{A}]x + \epsilon^2 ||(w\cdot \mathbf{A})x||^2 < ||(I+B)x||^2$$

uniformly for ϵ sufficiently small. For more general x such that $||x|| = 1$, since x has components of right singular vectors associated with smaller singular values of $I+B$, it can also be shown that $||(I+(v+\epsilon w)\cdot \mathbf{A})x||^2 < ||I+B||^2$ uniformly for sufficiently small ϵ. By this line of argument we arrive at a contradiction to v being a minimizer. □

LEMMA 2.16. *Given* $\mathbf{A} = \sum_i e_i \otimes A_i$, $A_i \in \mathbb{K}^{N\times N}$, $\psi_{\mathbb{K}}(\mathbf{A}) = \varphi_{\mathbb{K}}(\mathbf{A})$ *if and only if for some (equivalently, every) minimizer* v *of* $\varphi_{\mathbb{K}}(\mathbf{A})$ *and for* $B = v\cdot \mathbf{A}$, $0 \in F_{\Sigma_{\mathbb{K}}(I+B)}([I \otimes [I+B]]^* \mathbf{A})$.

Proof. If $0 \in F_{\Sigma_{\mathbb{K}}(I+B)}([I \otimes [I+B]]^* \mathbf{A})$, then there exists $x \in \Sigma_{\mathbb{K}}(I+B)$, $||x|| = 1$, such that $(I+B)x \perp A_i x$, for all i. Then such B solves the least squares problem for x, i.e. $||(I+B)x|| = \inf_w ||(I+w\cdot \mathbf{A})x||$, so in fact $\varphi_{\mathbb{K}}(\mathbf{A}) \leq ||I+B|| = ||(I+B)x|| \leq \psi_{\mathbb{K}}(\mathbf{A}) \leq \varphi_{\mathbb{K}}(\mathbf{A})$.

Now suppose $\psi_{\mathbb{K}}(\mathbf{A}) = \varphi_{\mathbb{K}}(\mathbf{A})$. Let v be a minimizer for φ, and let $B = v\cdot \mathbf{A}$. Let x be a maximizer for ψ, $||x|| = 1$. Then by definition of ψ, letting B' solve the least squares problem for x, $\varphi_{\mathbb{K}}(\mathbf{A}) = \psi_{\mathbb{K}}(\mathbf{A}) = ||(I+B')x|| \leq ||(I+v\cdot \mathbf{A})x|| = ||(I+B)x|| \leq \varphi_{\mathbb{K}}(\mathbf{A})$. But then x must be a maximal right singular vector of $I+B$, and in fact this inequality is an equality. Furthermore, v solves the minimization problem $\inf_w ||(I+w\cdot \mathbf{A})x||$, so $(I+B)x \perp A_i x$, giving the result. □

COROLLARY 2.17. *Given* $\mathbf{A} = \sum_i e_i \otimes A_i$, $A_i \in \mathbb{K}^{N\times N}$, *if for some minimizer* v *of* $\varphi_{\mathbb{K}}(\mathbf{A})$ *and for* $B = v \cdot \mathbf{A}$, $F_{\Sigma_{\mathbf{K}}(I+B)}([I \otimes [I+B]]^* \mathbf{A})$ *is convex, then* $\psi_{\mathbb{K}}(\mathbf{A}) = \varphi_{\mathbb{K}}(\mathbf{A})$.

These results lead to the following theorem, which shows that for k sufficiently large, ψ and φ are equal for $\mathbf{A} \otimes I_k$.

THEOREM 2.18. *Given* $\mathbf{A} = \sum_{i=1}^n e_i \otimes A_i$, $A_i \in \mathbb{K}^{N\times N}$, *if* $k \leq N$ *is the smallest dimension for the maximal singular vector space of* $I + B = I + v \cdot \mathbf{A}$ *for* v *any minimizer for* $\varphi_{\mathbf{K}}(\mathbf{A})$, *then for* I_k *the identity matrix of dimension* $k \times k$, $\psi_{\mathbf{K}}(\mathbf{A} \otimes I_k) = \varphi_{\mathbb{K}}(\mathbf{A} \otimes I_k)$.

Proof. Let B' be a minimizer for $\varphi_{\mathbb{K}}(\mathbf{A} \otimes I_k)$. Then $B' = v \cdot \mathbf{A} \otimes I_k = (v \cdot \mathbf{A}) \otimes I_k = B \otimes I_k$, where $B = v \cdot \mathbf{A}$. It is clear from Theorem 2.13 that B' is a minimizer for $\varphi(\mathbf{A} \otimes I_k)$ if and only if B is a minimizer for $\varphi(\mathbf{A})$. Note also for $I+B = U\Sigma V^*$ a singular value decomposition, $[I+B] \otimes I_k = (U \otimes I_k)(\Sigma \otimes I_k)(V \otimes I_k)^*$ is also a singular value decomposition. Thus $\Sigma_{\mathbb{K}}(I+B') = \bigoplus_{i=1}^k \Sigma_{\mathbb{K}}(I+B)$. Also, $[I \otimes [I+B']]^* [\mathbf{A} \otimes I_k] = ([I \otimes [I+B]]^* \mathbf{A}) \otimes I_k = \mathbf{C} \otimes I_k$, for $\mathbf{C} = \sum_i e_i \otimes C_i$, $C_i = [I+B]^* A_i$. We conclude that $F_{\Sigma_{\mathbf{K}}(I+B')}([I \otimes [I+B']]^* \mathbf{A}') = \{\sum_i \sum_j e_i v_j^* C_i v_j : \sum_j ||v_j||^2 = 1,\ v_j \in \Sigma_{\mathbf{K}}(I+B)\}$. It is enough to show this is convex.

For $v_i \in \Sigma_{\mathbf{K}}(I+B)$ let $v = \sum v_i \otimes e_i$ and $\rho(v) = \sum_i v_i v_i^*$. Furthermore let $\sigma(P) = \sum_{i=0}^n e_i \mathrm{trace}(PC_i)$, letting $C_0 = I$. Note this sum begins at $i = 0$. Since, without loss of generality, $k \geq \dim \Sigma_{\mathbf{K}}(I+B)$, it is easily seen that the range of ρ is convex. Since σ is linear, the range $\check{F}$ of $\sigma \circ \rho$ is convex. But then $H_{\mathbb{K}}(e_1) \cap \check{F} = F_{\Sigma_{\mathbb{K}}(I+B')}([I \otimes [I+B']]^* \mathbf{A}')$ is convex. □

It should be added that when $A_i = A^i$, under appropriate conditions the minimizer for φ is unique (see [Greenbaum and Trefethen, 1992]). We finally conclude that, given A and n, for some k no greater than N, $\psi_{n,\mathbf{K}}(A \otimes I_k) = \varphi_{n,\mathbf{K}}(A \otimes I_k)$. Note also that $\psi_{n,\mathbf{K}}(A \otimes I_k)$ is nondecreasing in k (Theorem 2.12), whereas $\varphi_{n,\mathbb{K}}(A \otimes I_k)$ is constant as a function of k (Theorem 2.13).

The result which follows allows us to characterize convergence of the two methods solely in terms of the *conical* generalized field of values rather than the standard generalized field of values. These results will be useful later in the paper.

PROPOSITION 2.19. *For* $A \in \mathbb{K}^{N\times N}$, $0 \in F_{\mathbb{K}}(\{A^i\}_{i=1}^n)$ *holds if and only if* $e_1 \in \check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n)$. *Also,* $0 \in \mathrm{cvx}[F_{\mathbf{K}}(\{A^i\}_{i=1}^n)]$ *if and only if* $e_1 \in \mathrm{cvx}[\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n)]$.

Proof. Recalling that $H_{\mathbb{K}}(e_1) = \{u \in \mathbb{K}^N : e_1^* u = 1\}$, note

$$0 \in F_{\mathbb{K}}(\{A^i\}_{i=1}^n) \iff e_1 \in 1 \oplus F_{\mathbb{K}}(\{A^i\}_{i=1}^n) = H_{\mathbf{K}}(e_1) \cap \check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n) \subseteq \check{F}_{\mathbf{K}}(\{A^i\}_{i=0}^n),$$

$$0 \in \mathrm{cvx}\, F_{\mathbb{K}}(\{A^i\}_{i=1}^n) \Longleftrightarrow e_1 \in 1 \bigoplus \mathrm{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)] = \mathrm{cvx}[1 \bigoplus F_{\mathbb{K}}(\{A^i\}_{i=1}^n)]$$

$$= \mathrm{cvx}[H_{\mathbb{K}}(e_1) \cap \check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n)] \subseteq \mathrm{cvx}[\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n)].$$

Note also that if $e_1 \in \check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n)$ then since $e_1 \in H_{\mathbf{K}}(e_1)$, $e_1 \in H_{\mathbf{K}}(e_1) \cap \check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n$. Now suppose $e_1 \in \mathrm{cvx}(\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n))$, i.e. there exist $t_i \geq 0, \sum_i t_i = 1, e_1 = \sum_i t_i [\sum_{j=0}^n e_j v_i^* A^j v_i]$ for some $v_i \in \mathbb{K}^N$. That is, $1 = \sum_i t_i v_i^* v_i$, and for $j \geq 1$, $0 = \sum_i t_i v_i^* A^j v_i$. Now, let $t_i' = t_i v_i^* v_i$, and let $v_i' = v_i / ||v_i||$ when $v_i \neq 0$ and otherwise let v_i' be any vector of norm 1. Then, for $S = \{i : \ v_i \neq 0\}$, $e_1 = \sum_{i \in S} t_i' [\sum_j e_j v_i'^* A^j v_i'] = \sum_i t_i' [\sum_j e_j v_i'^* A^j v_i']$ since $t_i' = t_i v_i^* v_i = 0$ for $i \notin S$, $\sum_i t_i' = \sum_i t_i v_i^* v_i = 1$, and $||v_i'|| = 1$ for all i, so $0 \in \mathrm{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)]$. □

The simple result below states that if the conical field of values of a set of matrices is convex, then it is also convex for a subset of the matrices. This can be useful for transferring a result on equivalence of convergence of iterative methods to a lower iteration number.

PROPOSITION 2.20. *For $A_i \in \mathbb{C}^{N \times N}$ if $\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n)$ is convex, then for $m \leq n$, $\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^m)$ is convex. More generally, if for $A_i \in \mathbb{C}^{N \times N}$, $\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n)$ is convex, and $V \in \mathbb{C}^{m \times n}$, then $V \check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n) = \check{F}_{\mathbf{K}}(\{\sum v_{i,j} A_j\}_{i=1}^n)$ is convex.*

Proof. A linear operator applied to a convex set preserves convexity. □

The following are two further results on combining multiple matrices into a block diagonal matrix.

PROPOSITION 2.21. *Let $A_i \in \mathbb{C}^{N_1 \times N_1}$ and $B_i \in \mathbb{C}^{N_2 \times N_2}$, and also suppose that $\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n)$ and $\check{F}_{\mathbb{K}}(\{B_i\}_{i=1}^n)$ are convex. Then $\check{F}_{\mathbb{K}}(\{\mathrm{diag}[A_i, B_i]\}_{i=1}^n)$ is convex.*

Proof. It is enough to show that for $t_i \geq 0$, $\sum_i t_i = 1$, $v_{i,1} \in \mathbb{K}^{N_1}$, $v_{i,2} \in \mathbb{K}^{N_2}$, there exist $v_1 \in \mathbb{K}^{N_1}$, $v_2 \in \mathbb{K}^{N_2}$ such that for $1 \leq k \leq n$, $\sum_i t_i [\sum_k e_k [v_{i,1}^* A_k v_{i,1} + v_{i,2}^* B_k v_{i,2}]] = \sum_k [e_k v_1^* A_k v_1 + v_2^* B_k v_2]]$. This may be done simply by letting

$$\sum_k e_k [v_1^* A_k v_1] = \sum_i t_i [\sum_k e_k [v_{i,1}^* A_k v_{i,1}]], \qquad \sum_k e_k [v_2^* B_k v_2] = \sum_i t_i [\sum_k e_k [v_{i,2}^* B_k v_{i,2}]].$$

□

COROLLARY 2.22. *Let $A_i \in \mathbb{C}^{N \times N}$, and suppose $\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n)$ is convex. Then $\check{F}_{\mathbb{K}}(\{A_i \otimes D_m\}_{i=1}^n)$ is convex, for $D_m \in \mathbb{C}^{m \times m}$ any diagonal matrix.*

The next two results further extend the above convexity results to the tensor product of a matrix with an appropriate normal matrix.

PROPOSITION 2.23. *Let $A_i \in \mathbb{C}^{N\times N}$, and suppose $\check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n)$ is convex. Also let $B \in \mathbb{K}^{M\times M}$ be any normal matrix with eigenvalues in $\mathbb{K}$, and let $p_i \in \mathbb{C}[x]$. Then $\check{F}_{\mathbb{K}}(\{A_i \otimes p_i(B)\}_{i=1}^n)$ is convex.*

Proof. Let $B = U\Lambda U^*$, $U \in \mathbb{K}^{N\times N}$ unitary, $\Lambda = \{\lambda_i\} \in \mathbb{K}^{N\times N}$ diagonal. Note

$$\check{F}_{\mathbb{K}}(\{A_i \otimes p_i(B)\}_{i=1}^n) = \check{F}_{\mathbb{K}}(\{A_i \otimes p_i(\Lambda)\}_{i=1}^n) = \{\sum_i e_i \sum_j (v_j \otimes e_j)^* [A_i \otimes p_i(\Lambda)](v_j \otimes e_j)\}$$

$$= \{\sum_i e_i \sum_j p_i(\lambda_j)(v_j \otimes e_j)^* [A_i \otimes e_j e_j^*](v_j \otimes e_j)\}$$

$$= \left[\sum_{i,j} p_i(\lambda_j) e_i [e_i \otimes e_j]^*\right] \{\sum_{i,j} e_i \otimes e_j [(v_j \otimes e_j)^* [A_i \otimes e_j e_j^*](v_j \otimes e_j)]\}$$

$$= \left[\sum_{i,j} p_i(\lambda_j) e_i [e_i \otimes e_j]^*\right] \check{F}_{\mathbb{K}}(\{A_i \otimes e_j e_j^*\}_{i=1\,j=1}^{n\quad M})$$

$$= \left[\sum_{i,j} p_i(\lambda_j) e_i [e_i \otimes e_j]^*\right] \oplus_{j=1}^M \check{F}_{\mathbb{K}}(\{A_i\}_{i=1}^n).$$

The result follows from noting that a direct sum of convex sets is convex, and the linear transformation of a convex set is convex. □

COROLLARY 2.24. *For A, $B \in \mathbb{K}^{N\times N}$, for B normal with eigenvalues in $\mathbb{K}$, if $\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n)$ is convex, then so is $\check{F}_{\mathbb{K}}(\{(A\otimes B)^i\}_{i=0}^n)$.*

The next proposition shows that it is possible to replace matrices with their transpose or conjugate transpose.

PROPOSITION 2.25. *For $A_i \in \mathbb{C}^{N\times N}$, suppose $\check{F}_{\mathbb{K}}(\{A_i\})$ is convex, and for each i let $\tilde{A}_i$ be either A_i, A_i^* or A_i^T. Then $\check{F}_{\mathbb{K}}(\{\tilde{A}_i\})$ is convex.*

Proof. Follows from $x^* A_i^* x = \overline{x^* A_i x}$ and $x^* A_i^T x = \overline{x^* \overline{A_i} x} = \overline{x}^* A_i \overline{x}$. □

Similarly, each matrix can be replaced with its Hermitian and skew-Hermitian parts. Here, $(M)_H$ denotes the Hermitian part of a matrix, $(M)_H = (M + M^*)/2$.

PROPOSITION 2.26. *For $A_j \in \mathbb{R}^{N\times N}$, $\check{F}_{\mathbb{R}}(\{A_j\}_{j=1}^n) = \check{F}_{\mathbb{R}}(\{(A_j)_H\}_{j=1}^n)$, and for $A_j \in \mathbb{C}^{N\times N}$, $\check{F}_{\mathbb{C}}(\{A_j\}_{j=1}^n)$ is isomorphic to $\check{F}_{\mathbb{C}}(\{(A_j)_H, (iA_j)_H/i\}_{j=1}^n)$, under the identification of $\mathbb{C}^n$ with $\mathbb{R}^{2n}$.*

Proof. For the real case, note $v^*Av = v^*(A)_H v$; for the complex case, the real and imaginary parts of v^*Av are given by $v^*(A)_H v$ and $-v^*(iA)_H v$, respectively. □

The following result indicates that shifting a matrix by a constant multiple of the identity does not change convexity of the conical field of values.

PROPOSITION 2.27. *For $A \in \mathbb{C}^{N\times N}$ and $c \in \mathbb{C}$, if $\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n)$ is convex then $\check{F}_{\mathbb{K}}(\{(A+cI)^i\}_{i=0}^n)$ is convex.*

Proof. The result follows from the fact that a linear transformation T over $\mathbb{C}$ exists such that $\check{F}_{\mathbb{K}}(\{(A+cI)^i\}_{i=0}^n) = T\check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n)$. □

3. Results for Toeplitz Matrices. In this section we demonstrate that for the special class of upper triangular Toeplitz matrices, the generalized field of values of the powers of that matrix is convex. The implication of this result, based on the results of the previous section, is that for such matrices, if GMRES(s) converges for given s, then the optimal polynomial preconditioning of corresponding degree must also converge.

Let us begin with notation. Let $D_k \in \mathbb{R}^{N\times N}$ be defined by $e_i^* D_k e_j = \delta_{i,j-k}$. These matrices form a basis for the Toeplitz matrices. Note $D_0 = I$, $D_{-k} = D_k^*$, and for $kl \geq 0$, $D_k D_l = D_{kl}$. A Toeplitz matrix over $\mathbb{K}$ is defined to be $\sum_{i=1-N}^{N-1} t_i D_i$, for $t_i \in \mathbb{K}$.

We begin by proving the convexity of $\check{F}_{\mathbb{K}}(\{D_i\}_{i=1-N}^{N-1})$. Note each $r = [r_{1-N}, \ldots r_{N-1}] \in \check{F}_{\mathbb{K}}(\{D_i\})$ can be mapped by a linear injective map to the space of rational functions of the form $p(z) = \sum_{i=1-N}^{N-1} r_i z^i$, the z-transform. Letting $r_i = x^* D_i x$, with $x \in \mathbb{K}^N$, $r_i = 0$ for $i \notin [1-N, N-1]$ and $x_i = 0$ for $i \notin [1, N]$, we can write

$$p(z) = \sum_{i=1-N}^{N-1} \left(\sum_{j=1}^{N} \bar{x}_j x_{j+i} \right) z^i = \sum_{i=0}^{N-1} x_{i+1} z^i \sum_{i=0}^{N-1} \bar{x}_{i+1} z^{-i}.$$

Thus $p(z) = q(z)\bar{q}(1/z)$, where $q(z) = \sum_{i=0}^{N-1} x_{i+1} z^i$. Here the convention $\bar{q}(z) = \sum_{i=0}^{N-1} \bar{q}_i z^i$ is assumed, and similarly for rational functions. Note then that $r \in \check{F}_{\mathbb{K}}(\{D_i\})$ if and only if the corresponding p may be factored as $q(z)\bar{q}(1/z)$ for $q \in \mathbb{K}_{n-1}[x]$.

Let $\mathcal{P}'_{\mathbb{K}}$ be the set of symmetric rational functions over $\mathbb{K}$, i.e. $p(z) = \bar{p}(1/z)$, such that $z^{N-1}p(z) = \tilde{p}(z) \in \mathbb{K}[z]$. Note $\mathcal{P}'_{\mathbb{K}}$ is a linear space over the reals, in the sense that $p_i \in \mathcal{P}'_{\mathbb{K}}$, $a_i \in \mathbb{R}$ imply $\sum_i a_i p_i \in \mathcal{P}'_{\mathbb{K}}$. Furthermore let $\mathcal{P}_{\mathbb{K}}$ denote the set of rational functions over $\mathbb{K}$ factorable as $q(z)\bar{q}(1/z)$, $q \in \mathbb{K}_{N-1}[x]$. Clearly $\mathcal{P}_{\mathbb{K}} \subseteq \mathcal{P}'_{\mathbb{K}}$. The convexity of $\check{F}_{\mathbb{K}}(\{D_i\})$ will follow if $\mathcal{P}_{\mathbb{K}}$ can be shown to be a convex subset of the linear space $\mathcal{P}'_{\mathbb{K}}$.

Let us consider the properties of polynomials in $z^{N-1}\mathcal{P}'_{\mathbb{K}}$. Note if α is a root of $s(z) \in$

$\mathcal{P}'_{\mathbb{K}}$, then so is $1/\bar{\alpha}$, and furthermore both of these are roots of the polynomial $z^{N-1}s(z) \in z^{N-1}\mathcal{P}'_{\mathbb{K}}$. Also, for nonzero α, $|\alpha| \neq 1$ if and only if α and $1/\bar{\alpha}$ are distinct, so for $\alpha \in \mathbb{K}$, $|\alpha| \neq 1$ implies $s(z)/[(z-\alpha)(1/z-\bar{\alpha})] \in \mathcal{P}'_{\mathbb{K}}$. Similarly for $\mathbb{K} = \mathbb{R}$, $\alpha \in \mathbb{C} \setminus \mathbb{R}$ and $|\alpha| \neq 1$, the roots α, $\bar{\alpha}$, $1/\bar{\alpha}$ and $1/\alpha$ are all distinct, so $s(z)/[(z-\alpha)(1/z-\bar{\alpha})(z-\bar{\alpha})(1/z-\alpha)] \in \mathcal{P}'_{\mathbb{K}}$. This demonstrates that for $|\alpha| \neq 1$, the roots α and $1/\bar{\alpha}$ must have the same multiplicity in $z^{N-1}s(z) \in z^{N-1}\mathcal{P}'_{\mathbb{K}}$.

Note also

LEMMA 3.1. *For z on the unit circle, $s(z) \in \mathcal{P}'_{\mathbb{K}}$ takes on real values.*

Proof.

$$s(z) = \bar{s}(1/z) = \bar{s}(\bar{z}) = \overline{s(z)}.$$

□

The following result characterizes polynomials in $z^{N-1}\mathcal{P}_{\mathbb{K}}$.

LEMMA 3.2. *For nonzero rational functions $s(z)$, the polynomial $z^{N-1}s(z) \in z^{N-1}\mathcal{P}_{\mathbb{K}}$ if and only if the roots of $z^{N-1}s(z) \in \mathbb{K}[z]$, counted with multiplicity, consist of k zero roots and $(N-k-1)$ pairs of roots of the form $\{\alpha_i, 1/\bar{\alpha}_i\}$ and furthermore $s(z) \geq 0$ for all z on the unit circle.*

Proof. For $z^{N-1}s(z) \in z^{N-1}\mathcal{P}_{\mathbb{K}}$ we have

$$z^{N-1}s(z) = z^{N-1}q(z)\bar{q}(1/z) = (\alpha\bar{\alpha})z^{N-1}\prod_{i=1}^{N-k-1}(z-\alpha_i)\prod_{i=1}^{N-k-1}(1/z-\bar{\alpha}_i)$$

$$= (\alpha\bar{\alpha})z^k\left[\prod_{i=1}^{N-k-1}(-\bar{\alpha}_i)\right]\prod_{i=1}^{N-k-1}(z-\alpha_i)\prod_{i=1}^{N-k-1}(z-1/\bar{\alpha}_i)$$

for some $k \geq 0$ and for $q(z) = \alpha\prod(z-\alpha_i)$. Note that $q(z)$ has strict degree $(N-k-1)$, and $z^{N-1}s(z)$ has strict degree $2(N-1)-k$. Also note that if $s(z) \in \mathcal{P}_{\mathbb{K}}$, then for $|z| = 1$, $s(z) = q(z)\bar{q}(1/z) = q(z)\bar{q}(\bar{z}) = |q(z)|^2 \geq 0$. To show the converse, let

$$z^{N-1}s(z) = cz^k\left[\prod_{i=1}^{N-k-1}(-\bar{\alpha}_i)\right]\prod_{i=1}^{N-k-1}(z-\alpha_i)\prod_{i=1}^{N-k-1}(z-1/\bar{\alpha}_i)$$

$$= cz^{N-1}\prod_{i=1}^{N-k-1}(z-\alpha_i)\prod_{i=1}^{N-k-1}(1/z-\bar{\alpha}_i).$$

By substituting values of z for which $1/z = \bar{z}$, we obtain necessarily that $c \geq 0$, so c may be written $c = \alpha\bar{\alpha}$, giving the result. □

The previous result shows that roots α of the polynomial $z^{N-1}s(z) \in z^{N-1}\mathcal{P}_{\mathbb{K}}$ with $|\alpha| = 1$ must necessarily have even multiplicity. Thus,

COROLLARY 3.3. *For nonzero $s(z) \in \mathcal{P}'_{\mathbb{K}}$, $s(z) \notin \mathcal{P}_{\mathbb{K}}$ if and only if $s(z)$ is negative for some z on the unit circle or $s(z)$ has a root on the unit circle of odd multiplicity.*

We conclude then that

LEMMA 3.4. *For $s(z) \in \mathcal{P}'_{\mathbb{K}}$, $s(z) \in \mathcal{P}_{\mathbb{K}}$ if and only if $s(z)$ is nonnegative on the unit circle.*

Proof. The case when $s(z)$ is zero is readily dispensed with. If $s(z) \in \mathcal{P}_{\mathbf{K}}$, then as shown earlier, $s(z)$ is nonnegative on the unit circle. For the converse, it suffices to show that if $s(z) \in \mathcal{P}'_{\mathbf{K}}$ has a root $e^{i\theta}$ of odd multiplicity on the unit circle, then $s(z)$ is negative somewhere on the unit circle. In this case $s(z) = (z - e^{i\theta})^{2m+1}t(z)$ where t is nonzero at $e^{i\theta}$. Let $z = e^{i(\theta+\delta)}$.

$$s(e^{i(\theta+\delta)}) = (e^{i\theta}(e^{i\delta} - 1))^{2m+1}t(z) = e^{i(2m+1)\theta}[(-1)^m i\delta^{2m+1}]t(e^{i\theta}) + \mathcal{O}(\delta^{2m+2}).$$

Thus for sufficiently small δ, $s(e^{i(\theta+\delta)})$ and $s(e^{i(\theta-\delta)})$ must have different sign. □

Thus,

THEOREM 3.5. *$\check{F}_{\mathbb{K}}(\{D_i\}_{i=1-N}^{N-1})$ is convex.*

Proof. Both $\mathcal{P}_{\mathbb{K}}$ and the set of rational functions which are real and nonnegative on the unit circle are convex sets, so their intersection is convex, giving the result. □

COROLLARY 3.6. *For $T_i \in \mathbb{K}^{N\times N}$ Toeplitz, $\check{F}_{\mathbb{K}}(\{T_i\})$ is convex.*

COROLLARY 3.7. *For $T \in \mathbb{K}^{N\times N}$ upper triangular and Toeplitz, $\check{F}_{\mathbb{K}}(\{T^i\}_{i=0}^n)$ is convex.*

The results of Section 2 can be used to extend these results further:

COROLLARY 3.8. *For $T \in \mathbb{K}^{N\times N}$ upper triangular and Toeplitz and $B \in \mathbb{K}^{N\times N}$ normal with eigenvalues in $\mathbb{K}$, $\check{F}_{\mathbb{K}}(\{(T\otimes B)^i\}_{i=0}^n)$ is convex.*

COROLLARY 3.9. *For $T_i \in \mathbb{K}^{N_i\times N_i}$ upper triangular Toeplitz, $\check{F}_{\mathbb{K}}(\{(\mathrm{diag}\{T_i\})^j\}_{j=0}^n)$ is convex.*

COROLLARY 3.10. *For $T_i \in \mathbb{K}^{N_i\times N_i}$ upper triangular Toeplitz, and B and B_i normal matrices over $\mathbb{K}$ with eigenvalues in $\mathbb{K}$, $\check{F}_{\mathbb{K}}(\{((\mathrm{diag}\{T_i\otimes B_i\})\otimes B)^j\}_{j=0}^n)$ is convex.*

Thus, for a fairly large number of matrices, including normal matrices and direct sums of upper triangular Toeplitz matrices, if GMRES(s) converges, then the optimal polynomial preconditioner of corresponding degree must also converge—though in the latter case it is not clear that the convergence rate is necessarily the same.

One might be led to believe that the same result holds for all matrices. However, the next section shows that this is not the case.

4. Counterexamples for General Matrices. In this section a method is given for generating matrices $A \in \mathbb{K}^{N\times N}$ for which $\psi_{n,\mathbb{K}}(A) < 1$ but $\varphi_{n,\mathbb{K}}(A) = 1$, for certain values of n and N.

In particular, we will construct a real nonsingular matrix A of dimension $N = 4$ such that $\psi_{N-1,\mathbb{R}}(A) < \varphi_{N-1,\mathbb{R}}(A) = 1$.

The following step-by-step process is used to construct the counterexample. It should be noted that the method given here may be used to generate other counterexamples $A \in \mathbb{R}^{N\times N}$, possibly for larger N, for which $\psi_{N-1,\mathbb{R}}(A) < \varphi_{N-1,\mathbb{R}}(A) = 1$; however, the construction is not necessarily always guaranteed to work, and each potential counterexample must be checked to confirm its validity.

Step 1: Construct HPD $M \in \mathbb{R}^{N\times N}$ and $w \in \mathbb{R}^N$ such that $H \equiv (D(w)M)_H$ is nonnegative definite with kernel of dimension 2. Here we define $D(w) = \sum_i (e_i^* w) e_i e_i^*$.

Let us note that for Hermitian M, $H = (D(w)M)_H = \Delta \circ M$, where $\Delta = \mathbf{1}^* w + w^* \mathbf{1}$, where $\mathbf{1}$ denotes the vector of all 1's, and $B \circ C$ denotes the Hadamard product of matrices, $\{b_{ij} c_{ij}\}$. We thus seek $H = \Delta \circ M$ with kernel of dimension 2. If w has no zero entries, then this may be written $M = \Delta^{\circ -1} \circ H$, where $B^{\circ -1}$ denotes the Hadamard inverse, $\{1/b_{ij}\}$.

We define the map $\mu : \mathbb{R}^N \times \prod_{i=1}^{N-2} \mathbb{R}^N \to \mathbb{R}^{N\times N}$ by

$$\mu(w, \{x_i\}_{i=1}^{N-2}) = (\mathbf{1}^* w + w^* \mathbf{1})^{\circ -1} \circ \sum_{i=1}^{N-2} x_i x_i^*.$$

Since $\Delta^{\circ -1}$ is nonnegative definite when $w_i > 0$ for all i [Horn and Johnson, 1991], p. 348, and nonnegative definiteness is preserved by Hadamard products [Horn and Johnson, 1991], p. 309, matrices in the image of μ for such w are nonnegative definite.

To obtain a matrix M in the image of μ that is not only nonnegative definite but also positive definite, we seek w and $\{x_i\}$ such that all symmetric matrices in a neighborhood of $\mu(w, \{x_i\})$ are in the image of μ. That is, we seek a point $(w, \{x_i\})$ where the Jacobian $J(\mu)$

is of rank $N(N+1)/2$, the dimension of the space of symmetric matrices. This condition also assures a set of matrices M of positive measure which satisfy the desired condition.

It can be shown directly that for $N = 3$, $J(\mu)$, a matrix-valued function of size $N(N-1) = 6$, never has rank $N(N+1)/2 = 6$. However, when $N = 4$, the following vectors w, x_1, x_2 yield $J(\mu)$ of size $N(N-1) = 12$ and of rank $N(N+1)/2 = 10$ and corresponding $M = \mu(w, x_1, x_2)$ of rank $(N-2) = 2$:

$$w = (\,1 \quad 2 \quad 2 \quad 3\,)^T, \quad x_1 = (\,8 \quad 8 \quad 3 \quad 9\,)^T, \quad x_2 = (\,5 \quad 8 \quad 2 \quad 8\,)^T.$$

This yields

$$M = \begin{pmatrix} 89/2 & 104/3 & 34/3 & 28 \\ 104/3 & 32 & 10 & 136/5 \\ 34/3 & 10 & 13/4 & 43/5 \\ 28 & 136/5 & 43/5 & 145/6 \end{pmatrix}.$$

The nullspace V of $H = \Delta \circ M = \sum_{i=1}^{N-2} x_i x_i^*$ consists of vectors perpendicular to both x_1 and x_2. A basis of V for our example is given by

$$y_1 = (\,0 \quad -3 \quad -4 \quad 4\,)^T, \quad y_2 = (\,-8 \quad -1 \quad 24 \quad 0\,)^T.$$

Step 2: Examine the image of $\{v \in \mathbb{K}^N : v^* M v = 1\} \cap \mathrm{Ker}_{\mathbb{K}}[(D(w)M)_H]$ by the vector of quadratic maps $\mathcal{M} = \sum_{i=1}^{N} e_i \otimes (e_i e_i^* M)$. For $\mathbb{K} = \mathbb{R}$ and $n = N - 1 = 3$, this is an ellipsoidal curve, being the image of an ellipse, which under appropriate conditions is nondegenerate.

We now demonstrate that for $\mathbb{K} = \mathbb{R}$ this image is the intersection of the supporting hyperplane $H_{\mathbb{K}}(w, 0)$ with $\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}_{i=1}^N) \cap H_{\mathbb{K}}(\mathbf{1})$; the object $\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}_{i=1}^N) \cap H_{\mathbb{K}}(\mathbf{1})$ is chosen here in anticipation of being transformed into $F_{\mathbb{K}}(\{A^i\}_{i=1}^{N-1})$ in a later stage of this construction. This result follows easily from noting $\mathrm{Ker}_{\mathbb{K}}(D(w)M)_H = \{v \in \mathbb{K} : \mathrm{Re}[v^* D(w) M v] = 0\} = \{v \in \mathbb{K} : v^* D(w) M v = 0\}$ and

$$H_{\mathbb{K}}(w, 0) \cap \check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbb{K}}(\mathbf{1})$$

$$= \{\textstyle\sum_i e_i v^* (e_i e_i^* M) v : \sum_i w_i v^* (e_i e_i^* M) v = 0,\ \sum_i v^* (e_i e_i^* M) v = 1,\ v \in \mathbb{K}^N\}$$

$$= \{\textstyle\sum_i e_i v^* (e_i e_i^* M) v : v^* D(w) M v = 0,\ v^* M v = 1,\ v \in \mathbb{K}^N\}.$$

Note also $\mathrm{Re}[w^*[\check{F}_{\mathbf{K}}(\{e_i e_i^* M\}_{i=1}^N) \cap H_{\mathbf{K}}(\mathbf{1})]] \geq 0$, since $(D(w)M)_H$ is nonnegative definite. Furthermore, this verifies that the image of the ellipse through the quadratic maps is contained in the two-dimensional plane $H_{\mathbb{R}}(w,0) \cap H_{\mathbf{R}}(\mathbf{1})$.

For our example, let $y = ay_1 + by_2$ represent an arbitrary element of $\mathrm{Ker}_{\mathbf{K}}[(D(w)M)_H]$. Then

$$y^* \mathcal{M} y = \begin{bmatrix} a & b \end{bmatrix} \left(\begin{bmatrix} 0 & 448/3 \\ 448/3 & 2848/3 \end{bmatrix}, \begin{bmatrix} 408/5 & 588/5 \\ 588/5 & 208/3 \end{bmatrix}, \begin{bmatrix} 172/5 & -868/15 \\ -868/15 & -544 \end{bmatrix}, \right.$$

$$\left. \begin{bmatrix} -232/3 & -448/5 \\ -448/5 & 0 \end{bmatrix} \right) \begin{bmatrix} a \\ b \end{bmatrix} \equiv \begin{bmatrix} a & b \end{bmatrix} \mathcal{Q} \begin{bmatrix} a \\ b \end{bmatrix}.$$

We wish to verify that the curve generated by $(a,b)\mathcal{Q}(a,b)^T$, under the condition that (a,b) satisfies $(a,b)([y_1\ y_2]^* M [y_1\ y_2])(a,b)^T = 1$, is nondegenerate, i.e. is not contained in a single line. This is true if three values of (a,b) can be given for which the three points $(a,b)\mathcal{Q}(a,b)^T$ are not collinear. In particular,

$$(1,0)\mathcal{Q}(1,0)^T = (\,0 \quad 408/5 \quad 172/5 \quad -232/3\,)^T,$$

$$(0,1)\mathcal{Q}(0,1)^T = (\,2848/3 \quad 208/3 \quad -544 \quad 0\,)^T,$$

$$(1,1)\mathcal{Q}(1,1)^T = (\,1248 \quad 5792/15 \quad -1876/3 \quad -3848/15\,)^T$$

which can be shown not to be collinear.

What we have found then is a generalized field of values of matrices which has a supporting hyperplane whose intersection with the body is an ellipsoidal curve and thus not convex.

Step 3: Select $x \in \mathrm{cvx}[\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbf{K}}(\mathbf{1}) \cap H_{\mathbf{K}}(w,0)] \setminus [\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbf{K}}(\mathbf{1}) \cap H_{\mathbb{K}}(w,0)]$, that is, x is in the convex hull of the set just determined, but not in the set itself. Let us now confirm that such x also satisfies $x \in \mathrm{cvx}[\check{F}_{\mathbf{K}}(\{e_i e_i^* M\})] \setminus [\check{F}_{\mathbb{K}}(\{e_i e_i^* M\})]$. Note that such x satisfies $x \in \mathrm{cvx}[\check{F}_{\mathbf{K}}(\{e_i e_i^* M\}) \cap H_{\mathbb{K}}(\mathbf{1})] \setminus [\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbf{K}}(\mathbf{1})]$: specifically, $x \in \mathrm{cvx}[\check{F}_{\mathbf{K}}(\{e_i e_i^* M\}) \cap H_{\mathbf{K}}(\mathbf{1}) \cap H_{\mathbb{K}}(w,0)] \subseteq \mathrm{cvx}[\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbf{K}}(\mathbf{1})]$, and $x \notin \check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbf{K}}(\mathbf{1})$ since $x \in \mathrm{cvx}[\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbf{K}}(\mathbf{1}) \cap H_{\mathbf{K}}(w,0)] \subseteq \mathrm{cvx}[H_{\mathbf{K}}(w,0)] = H_{\mathbb{K}}(w,0)$. Note also that $x \in \mathrm{cvx}[\check{F}_{\mathbb{K}}(\{e_i e_i^* M\}) \cap H_{\mathbb{K}}(\mathbf{1})] \subseteq \mathrm{cvx}[H_{\mathbb{K}}(\mathbf{1})] = H_{\mathbb{K}}(\mathbf{1})$. Thus $x \notin \check{F}_{\mathbf{K}}(\{e_i e_i^* M\})$.

For our example, let us select f_1, f_2 as values of $(a,b)\mathcal{Q}(a,b)^T$ and let $x = \alpha f_1 + \beta f_2$ for appropriate positive values of α, β. In particular, let $f_1 = (1,0)\mathcal{Q}(1,0)^T = (0, 408/5, 172/5, -232/3)$, $f_2 = (3,-1)\mathcal{Q}(3,-1)^T = (160/3, 1472/15, 564/5, -792/5)$, $\alpha = 15/8$ and $\beta = 1/8$, yielding $x = (100, 337, 276, -442) \in \mathrm{cvx}[\check{F}_{\mathbb{K}}(\{e_i e_i^* M\})] \setminus [\check{F}_{\mathbf{K}}(\{e_i e_i^* M\})]$. Note that since $\check{F}_{\mathbf{K}}(\{e_i e_i^* M\})$ is a cone, such x may be easily scaled to insure that $x \in \mathrm{cvx}[\check{F}_{\mathbf{K}}(\{e_i e_i^* M\}) \cap H_{\mathbb{K}}(\mathbf{1}) \cap H_{\mathbb{K}}(w,0)]$. However, this scaling is not important to the argument which follows.

Step 4: For such $x \in \mathbb{K}^N$ select $d \in \mathbb{K}^N$ with distinct positive entries such that $d^{\circ i} \perp x$, for $1 \leq i \leq N-1$, where $d^{\circ i}$ denote the Hadamard powers d, $d \circ d$, $d \circ d \circ d$, ...

Let V be the Vandermonde matrix determined by d, $(V)_{i,j} = (d_j)^{i-1}$. Note that for such d, $e_i^* V x = 0$ for $i > 1$, and furthermore since $x \in H_{\mathbb{K}}(\mathbf{1})$, $Vx = e_1$.

For our example, the existence of a real solution for the entries of d can be proven formally by reducing the constraints through elimination of variables to a one variable equation. The resulting one variable equation can be seen to have a solution by the intermediate value theorem.

Writing $d = [d_1, d_2, d_3, 1]$, we have three equations:

$$100d_1 + 337d_2 + 276d_3 - 442 = 0 \tag{14}$$

$$100d_1^2 + 337d_2^2 + 276d_3^2 - 442 = 0 \tag{15}$$

$$100d_1^3 + 337d_2^3 + 276d_3^3 - 442 = 0 \tag{16}$$

We proceed by first solving eqn. (14) for d_3 and substituting in the eqns. (15) and (16) to eliminate d_3. Next we formally solve eqn. (15) for d_2 (using a computer algebra system) and substitute one of the two results in eqn. (16) to eliminate d_2 (we found that which result was used did not affect the roots of the new eqn. (16)). We then proceed to solve eqn. (16). Although both computer algebra systems we tried returned several incorrect solutions, the one corresponding to $d_1 = -0.26769$ could be verified formally, simply by checking that the lefthand side $L(d_1)$ of eqn. (16) satisfies

$$L(-1/2) = -\frac{63362781}{751538} - \frac{183\sqrt{4519}^3}{751538\sqrt{7751}}$$

$$L(0) = \frac{20482722}{375769} - \frac{61\sqrt{359810550031 14}}{\sqrt{7751}\ 375769}$$

It can be checked that the signs of these two expressions are negative and positive, respectively.

We finally obtain

$$d = (-0.267698 \ldots \quad 1.084117 \ldots \quad 0.374718 \ldots \quad 1) .$$

Step 5: For such $M \in \mathbb{K}^{N\times N}$ let $M = P^{-1}P^{-*}$, $P \in \mathbb{K}^{N\times N}$, by, for example, singular value decomposition.

For our example, $M = UDU^*$, where $D = \mathrm{diag}[(98.1079, 5.44092, 0.339205, 0.0286123)]$ and

$$U = \begin{bmatrix} -0.650955 & 0.732478 & -0.152607 & 0.12824 \\ -0.564055 & -0.330403 & 0.752453 & -0.0805637 \\ -0.180298 & -0.047945 & -0.257713 & -0.948039 \\ -0.474966 & -0.593306 & -0.586608 & 0.279797 \end{bmatrix} .$$

Then let $P^{-1} = U\sqrt{D}$.

Step 6: We now verify that for $A = PD(d)P^{-1}$, $\varphi_{N-1,\mathbb{R}}(A) = 1$ but $\psi_{N-1,\mathbb{R}}(A) < 1$. This will be done by using previous results to transform the general field of values of $\{e_i e_i M\}$ to that of $\{A^i\}$. This transformation will map the point x, located in the "hole" in the body, to the origin.

First note that for $A \in \mathbb{K}^{N\times N}$, $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, we may transfer from the standard to the conical generalized field of values:

$$0 \in F_{\mathbb{K}}(\{A^i\}_{i=1}^n) \Longleftrightarrow e_1 \in 1 \bigoplus F_{\mathbb{K}}(\{A^i\}_{i=1}^n) = H_{\mathbb{K}}(e_1) \cap \check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n),$$

$$0 \in \mathrm{cvx}\, F_{\mathbb{K}}(\{A^i\}_{i=1}^n) \Longleftrightarrow e_1 \in 1 \bigoplus \mathrm{cvx}[F_{\mathbb{K}}(\{A^i\}_{i=1}^n)]$$

$$= \mathrm{cvx}[1 \bigoplus F_{\mathbb{K}}(\{A^i\}_{i=1}^n)] = \mathrm{cvx}[H_{\mathbb{K}}(e_1) \cap \check{F}_{\mathbb{K}}(\{A^i\}_{i=0}^n)].$$

Note that for $n = N-1$ and V the Vandermonde matrix defined above, since $V(I \bigotimes v^*) = (V \bigotimes 1)(I \bigotimes v^*) = V \bigotimes v^* = (I \bigotimes v^*)(V \bigotimes I)$, we have

$$V[(I \bigotimes v^*)(\textstyle\sum_{i=1}^N e_i \bigotimes (P e_i e_i^* P^{-1}))(1 \bigotimes v)]$$

$$= (I \bigotimes v^*)(V \bigotimes I)(\textstyle\sum_{i=1}^N e_i \bigotimes (P e_i e_i^* P^{-1}))(1 \bigotimes v)$$

$$= (I \bigotimes v^*)(\textstyle\sum_{i,j} d_i^{j-1} e_j \bigotimes (P e_i e_i^* P^{-1}))(1 \bigotimes v) = (I \bigotimes v^*)(\textstyle\sum_{i=j}^N e_j \bigotimes A^{j-1})(1 \bigotimes v),$$

and so for V, $P \in \mathbb{K}^{N\times N}$

$$\check{F}_{\mathbb{K}}(\{A^{i-1}\}_{i=1}^N) = V\check{F}_{\mathbb{K}}(\{Pe_ie_i^*P^{-1}\}_{i=1}^N)$$

$$= V\check{F}_{\mathbb{K}}(\{e_ie_i^*P^{-1}P^{-*}\}_{i=1}^N) = V\check{F}_{\mathbf{K}}(\{e_ie_i^*M\}_{i=1}^N),$$

using the invariance of conical field of values over simultaneous congruence transformation. Note also that $VH_{\mathbb{K}}(\mathbf{1}) = H_{\mathbf{K}}(V^{-*}\mathbf{1}) = H_{\mathbf{K}}(e_1)$. We conclude that $H_{\mathbb{K}}(e_1) \cap \check{F}_{\mathbb{K}}(\{A^{i-1}\}_{i=1}^N) = V[H_{\mathbb{K}}(\mathbf{1}) \cap \check{F}_{\mathbf{K}}(\{e_ie_i^*M\}_{i=1}^N)]$. As a result, since

$$x \in \mathrm{cvx}[\check{F}_{\mathbf{K}}(\{e_ie_i^*M\}_{i=1}^N) \cap H_{\mathbf{K}}(\mathbf{1})] \setminus [\check{F}_{\mathbb{K}}(\{e_ie_i^*M\}_{i=1}^N) \cap H_{\mathbf{K}}(\mathbf{1})],$$

we have

$$Vx = e_1 \in \mathrm{cvx}[V\check{F}_{\mathbf{K}}(\{e_ie_i^*M\}_{i=1}^N) \cap H_{\mathbb{K}}(\mathbf{1})] \setminus [V\check{F}_{\mathbf{K}}(\{e_ie_i^*M\}_{i=1}^N) \cap H_{\mathbf{K}}(\mathbf{1})]$$

$$= \mathrm{cvx}[\check{F}_{\mathbf{K}}(\{A^{i-1}\}_{i=1}^N) \cap H_{\mathbf{K}}(e_1)] \setminus [\check{F}_{\mathbf{K}}(\{A^{i-1}\}_{i=1}^N) \cap H_{\mathbf{K}}(e_1)],$$

establishing the desired result.

The final result of this construction is the matrix

$$A = \begin{pmatrix} .469258949671 & .144764925686 & -.011212551044 & .000047280410 \\ 2.610326570493 & .327595085371 & .028231187826 & -.010533891160 \\ -3.242997268120 & .452834814744 & .976573801662 & -.038644914105 \\ .162118329163 & -2.003125765058 & -.458143025300 & .417709385890 \end{pmatrix}.$$

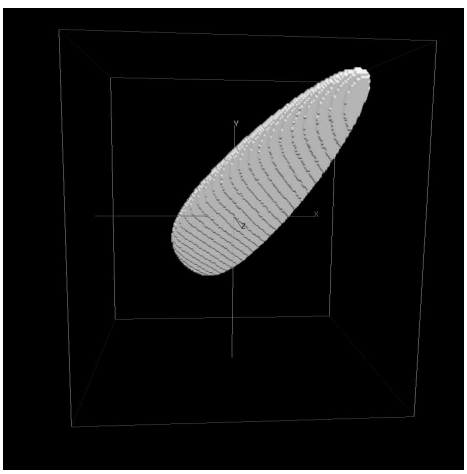

Figure 1. Generalized Field of Values

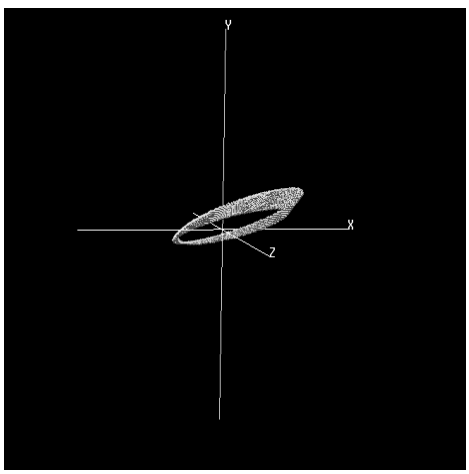

Figure 2. Slice of Generalized F.O.V.

In Figure 1 the generalized field of values $F_{\mathbf{R}}(\{A^i\}_{i=1}^3)$ is rendered graphically. The generalized field of values is plotted in a box of extents $[-4,4]$ along each axis, with the rightward x-axis corresponding to A, the upward y-axis A^2 and the frontward z-axis A^3.

The calculated distance of the generalized field of values to the origin is .001274. The calculated value of $\psi_{3,\mathbb{R}}(A)$ is approximately 0.99988.

The nonconvexity of this generalized field of values is supported by the illustration in Figure 2. This plot shows points in the generalized field of values that are on one side of the plane defined to be normal to the vector $(-.0691426, -.741006, .667929)$ and .00025 from the origin. The boundary of the convex hull of the body passes exactly through the origin. Though the field of values in Figure 1 appears to be flat near the origin, in fact a small concavity exists near the origin, indicated by the "hole" in the graph of Figure 2, and furthermore $0 \in \mathrm{cvx}[F_{\mathbf{R}}(\{A^i\}_{i=1}^3)] \setminus [F_{\mathbb{R}}(\{A^i\}_{i=1}^3)]$, so $\psi_{3,\mathbf{R}}(A) < \varphi_{3,\mathbb{R}}(A) = 1$.

5. Observations and Open Questions. Polynomial preconditioning is a popular and useful technique, insofar as it increases solution speed by reducing the requirements for inner product calculations, which is useful in its own right but has yet more advantage on certain advanced computer architectures for which inner products are particularly expensive (for a study of this issue, see for example [Joubert and Carey, 1992]).

For Hermitian positive definite problems, polynomial preconditioning is robust. As shown in [Greenbaum, 1979], not only do convergent preconditioners exist, but furthermore preconditioners with the same convergence rate as the conjugate gradient method exist. Thus, the main goal is to calculate preconditioners which give these good convergence rates.

On the other hand, the results of this paper indicate that for nonsymmetric problems, using polynomial preconditioning for the sake of increased speed may mean sacrificing robustness, the ability of the method to converge reliably to the solution of a given problem. Furthermore, this is a limitation in principle of the applicability of polynomial precoditioning as a technique. This problem is particularly critical for highly indefinite matrices which commonly arise in practice and may require very many GMRES iterations before restarting in order to converge.

Let us summarize some particular facts we know regarding this issue:

1. Due to the counterexample given above, it is at least known that for $N = 4$ and $n = 3$, there exists $A \in \mathbb{R}^{N\times N}$ such that $\psi_{n,\mathbf{R}}(A) < \varphi_{n,\mathbf{K}} = 1$.
2. Since the counterexample matrix is nonsingular, it is known by the continuity theo-

rems of Section 2 that there is in fact a set of matrices of positive measure for which $\psi_{n,\mathbb{R}}(A) < \varphi_{n,\mathbb{K}}$ ($N = 4$, $n = 3$). In other words, there is a nonzero probability of an arbitrary matrix being such that restarted GMRES converges but the associated polynomial preconditioning does not. Exactly how large the set is or how to characterize the matrices is not known.

3. As shown in the theorem below, a set of matrices of positive measure exists for which both methods stagnate. Thus, $\psi_{n,\mathbb{R}}(A) = \varphi_{n,\mathbb{R}}(A)$ on a set of matrices of positive measure. This affirms the experience that a significant number of matrices result in slow convergence or stagnation for restarted GMRES (and thus slow convergence for other iterative methods such as biconjugate gradient or QMR as well). How to ascertain whether this will happen for a given matrix easily is not known.

   THEOREM 5.1. *For every $N \geq 1$ and for every $n < N$, there exists a set of matrices $A \in \mathbb{R}^{N\times N}$ of positive measure in $\mathbb{R}^{N\times N}$ satisfying $1 = \psi_{n,\mathbb{R}}(A) = \varphi_{n,\mathbb{R}}(A)$.*

   Proof. Without loss of generality, let $n = N - 1$. Let $\hat{A} = e_1 e_N^*/2 + \sum_{i=2}^{N} e_i e_{i-1}^*$. Note for $\hat{r} = e_1$ and $g_A(r) \equiv r^* \underline{K}_N(r, A)$, $g_{\hat{A}}(\hat{r}) = e_1$. It is enough to show that for every sufficiently small real perturbation A of $\hat{A}$, the corresponding function g_A has a real solution r to the equation $g_A(r) = e_1$.

   The Jacobian function $J(g_{\hat{A}})(r)$ for $g_{\hat{A}}$ with respect to r is the matrix-valued function $2\,[r \quad (\hat{A})_H r \quad \ldots \quad (\hat{A}^{N-1})_H r]$. Then

$$J(g_{\hat{A}})(\hat{r}) = [2e_1 \quad e_2 + e_N/2 \quad e_3 + e_{N-1}/2 \quad \ldots \quad e_N + e_2/2]\,.$$

   This matrix is of full rank, by Gershgorin's theorem. Thus by the inverse function theorem there exist open sets $V \ni \hat{r}$ and $W \ni g_{\hat{A}}(\hat{r})$ such that $g_{\hat{A}} : V \to W$ is bijective.

   There exists an open disk $\hat{W}_\epsilon \subseteq W$ of radius ϵ centered at $g_{\hat{A}}(\hat{r})$. By continuity, there exists $\delta > 0$ such that for every $A = \hat{A} + \delta E$, $||E|| = 1$, the image $g_A(V)$ contains $\hat{W}_{\epsilon/2}$, the disk centered at $g_{\hat{A}}(\hat{r})$ of radius $\epsilon/2$. Thus, for each such A, $g_{\hat{A}}(\hat{r})$ is in the image of g_A, and thus $g_A(r) = g_{\hat{A}}(\hat{r}) = e_1$ has a solution r. □

4. It is known that $\psi_{n,\mathbb{R}}(A) = \varphi_{n,\mathbb{R}}(A)$ on some important measure-zero sets of matrices such as Hermitian matrices. Furthermore, for the measure-zero set of upper triangular Toeplitz matrices, at least $\psi_{n,\mathbb{R}}(A) < \varphi_{n,\mathbb{R}}(A) = 1$ cannot occur. It is not clear whether a positive measure set exists for which $\psi_{n,\mathbb{R}}(A) = \varphi_{n,\mathbb{R}}(A) < 1$. However, it is known that positive measure sets exist for which $\psi_{n,\mathbb{R}}(A) \leq \varphi_{n,\mathbb{R}}(A) < 1$,

due to continuity of the bound functions (small perturbations of an HPD matrix, for example).

5. One might ask how large the gap $\varphi_{n,\mathbb{R}}(A) - \psi_{n,\mathbb{R}}(A)$ can be. In the example given above, the gap is calculated to be approximately .00012. However, in a recent paper [Toh, 1995], a class of matrices is given for which $\varphi_{n,\mathbb{R}}(A) - \psi_{n,\mathbb{R}}(A)$ can be arbitrarily close to 1. Note that the gap cannot equal 1, since $0 = \psi_{n,\mathbb{R}}(A) < \varphi_{n,\mathbb{R}}(A)$ cannot occur. It is not known how to calculate this gap for a matrix in a simple and reliable way.

6. Conclusions. This paper has demonstrated several new results on the convergence rate of GMRES and polynomial preconditionings, including the fact that matrices exist for which restarted GMRES converges but every polynomial preconditioning of corresponding degree does not. Further research is required in order to devise practical tests for determining the convergence rates for these methods for matrices encountered in practice.

Acknowledgments. The authors would like to thank Anne Greenbaum for her comments and suggestions.

REFERENCES

[Ashby, Manteuffel and Saylor, 1990] S. F. Ashby, T. A. Manteuffel and P. E. Saylor, *A Taxonomy for Conjugate Gradient Methods,* SIAM J. Numer. Anal. **27** (1990), 1542–1568.

[Greenbaum, 1979] A. Greenbaum, *Comparison of splittings used with the conjugate gradient algorithm*, Numerische Mathematik, 33 (1979), pp. 181–194.

[Greenbaum and Gurvits, 1994] A. Greenbaum and L. Gurvits, *Max-min properties of matrix factor norms*, SIAM Journal on Scientific Computing, vol. 15, no. 2, pp. 427–439, March 1994.

[Greenbaum and Strakos, 1992] A. Greenbaum and Z. Strakos, *Matrices that Generate the Same Krylov Residual Spaces.*

[Greenbaum and Trefethen, 1992] A. Greenbaum and L. N. Trefethen, *GMRES/CR and Arnoldi/Lanczos as matrix approximation problems*, SIAM Journal on Scientific Computing, vol. 15, no. 2, pp. 427–439, March 1994.

[Horn and Johnson, 1991] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991.

[Householder, 1964] A. S. Householder, *The Theory of Matrices in Numerical Analysis.* New York: Dover, 1964.

[Joubert, 1990] W. Joubert, *Iterative Methods for the Solution of Nonsymmetric Systems of Linear Equations*, the University of Texas at Austin, Center for Numerical Analysis, Report CNA-242, February 1990.

[Joubert and Carey, 1992] Wayne D. Joubert and Graham F. Carey, *Parallelizable Restarted Iterative Methods for Nonsymmetric Linear Systems. Part I: Theory,* International Journal of Computer Mathematics, vol. 44, pp. 243–267, 1992.

[Joubert, 1994a] W. Joubert, *A Robust GMRES-Based Adaptive Polynomial Preconditioning Algorithm for Nonsymmetric Linear Systems*, SIAM Journal on Scientific Computing, vol. 15, no. 2, pp. 427–439, March 1994.

[Joubert, 1994b] Wayne Joubert, *On the Convergence Behavior of the Restarted GMRES Algorithm for Solving Nonsymmetric Linear Systems*, Numerical Linear Algebra with Applications, Volume 1, Issue 4, 1994.

[Joubert and Manteuffel, 1990] W. D. Joubert and T. A. Manteuffel, *Iterative Methods for Nonsymmetric Linear Systems,* Iterative Methods for Large Linear Systems, eds. D. R. Kincaid and L. J. Hayes, Academic Press, Boston (1990), 149–171.

[Lancaster and Tismenetsky, 1985] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, second edition, Boston, Academic Press, 1985.

[Saad and Schultz, 1986] Y. Saad and M. H. Schultz, *GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems,* SIAM J. Sci. Stat. Comp. 7 (1986), pp. 856–869.

[Toh, 1995] Kim-Chuan Toh, *GMRES vs. Ideal GMRES,* SIAM Journal on Matrix Analysis and Applications, submitted.