Electronic Transactions on Numerical Analysis.
Volume 41, pp. 159-166, 2014.
Copyright © 2014, Kent State University.
ISSN 1068-9613.

ETNA
Kent State University
http://etna.math.kent.edu

MAX-MIN AND MIN-MAX APPROXIMATION PROBLEMS FOR NORMAL MATRICES REVISITED[*]

JÖRG LIESEN[†] AND PETR TICHÝ[‡]

In memory of Bernd Fischer

Abstract. We give a new proof of an equality of certain max-min and min-max approximation problems involving normal matrices. The previously published proofs of this equality apply tools from matrix theory, (analytic) optimization theory, and constrained convex optimization. Our proof uses a classical characterization theorem from approximation theory and thus exploits the link between the two approximation problems with normal matrices on the one hand and approximation problems on compact sets in the complex plane on the other.

Key words. matrix approximation problems, min-max and max-min approximation problems, best approximation, normal matrices

AMS subject classifications. 41A10, 30E10, 49K35, 65F10

1. Introduction.

Let A be a real or complex square matrix, i.e., $A \in \mathbb{F}^{n\times n}$ with $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. Suppose that f and $\varphi_1, \dots, \varphi_k$ are given (scalar) functions so that $f(A) \in \mathbb{F}^{n\times n}$ and $\varphi_1(A), \dots, \varphi_k(A) \in \mathbb{F}^{n\times n}$ are well defined matrix functions in the sense of [9, Definition 1.2]. (In the case $\mathbb{F} = \mathbb{R}$, this requires a subtle assumption which is explicitly stated in (2.4) below.) Let $\mathcal{P}_k(\mathbb{F})$ denote the linear span of the functions $\varphi_1, \dots, \varphi_k$ with coefficients in $\mathbb{F}$ so that in particular $p(A) \in \mathbb{F}^{n\times n}$ for each linear combination $p = \alpha_1\varphi_1 + \cdots + \alpha_k\varphi_k \in \mathcal{P}_k(\mathbb{F})$.

With this notation, the optimality property of many useful methods of numerical linear algebra can be formulated as an approximation problem of the form

$$\min_{p\in\mathcal{P}_k(\mathbb{F})} \|f(A)v - p(A)v\|, \tag{1.1}$$

where $v \in \mathbb{F}^n$ is a given vector and $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{F}^n$. In (1.1) we seek a best approximation (with respect to the given norm) of the vector $f(A)v \in \mathbb{F}^n$ from the subspace of $\mathbb{F}^n$ spanned by the vectors $\varphi_1(A)v, \dots, \varphi_k(A)v$. An example of such a method is the GMRES method [15] for solving the linear algebraic problem $Ax = b$ with $A \in \mathbb{F}^{n\times n}$, $b \in \mathbb{F}^n$, and the initial guess $x_0 \in \mathbb{F}^n$. Its optimality property is of the form (1.1) with $f(z) = 1$, $\varphi_i(z) = z^i$, for $i = 1, \dots, k$, and $v = b - Ax_0$.

If the given vector v has unit norm, which usually can be assumed without loss of generality, then an upper bound on (1.1) is given by

$$\min_{p\in\mathcal{P}_k(\mathbb{F})} \|f(A) - p(A)\|, \tag{1.2}$$

where $\|\cdot\|$ denotes the matrix norm associated with the Euclidean vector norm, i.e., the matrix 2-norm or spectral norm on $\mathbb{F}^{n\times n}$. In (1.2) we seek a best approximation (with respect to the given norm) of the matrix $f(A) \in \mathbb{F}^{n\times n}$ from the subspace of $\mathbb{F}^{n\times n}$ spanned by the

[*] Received October 22, 2013. Accepted April 14, 2014. Published online on July 4, 2014. Recommended by R. Freund. The work of the second author was supported by the Grant Agency of the Czech Republic under grant No. P201/13-06684 S and by the project M100301201 of the institutional support of the Academy of Sciences of the Czech Republic.

[†] Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany (liesen@math.tu-berlin.de).

[‡] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 18207 Prague, Czech Republic (tichy@cs.cas.cz).

matrices $\varphi_1(A), \ldots, \varphi_k(A)$. An example of this type is the Chebyshev matrix approximation problem with $A \in \mathbb{F}^{n\times n}$, $f(z) = z^k$, and $\varphi_i(z) = z^{i-1}$, $i = 1, \ldots, k$. This problem was introduced in [8] and later studied, for example, in [3] and [17].

In order to analyse how close the upper bound (1.2) can possibly be to the quantity (1.1), one can maximize (1.1) over all unit norm vectors $v \in \mathbb{F}^n$ and investigate the sharpness of the inequality

$$\begin{aligned}(1.3) \qquad \max_{\substack{v\in\mathbb{F}^n\\ \|v\|=1}} \min_{p\in\mathcal{P}_k(\mathbb{F})} \|f(A)v - p(A)v\| &\le \min_{p\in\mathcal{P}_k(\mathbb{F})} \|f(A) - p(A)\| \\ &= \min_{p\in\mathcal{P}_k(\mathbb{F})} \max_{\substack{v\in\mathbb{F}^n\\ \|v\|=1}} \|f(A)v - p(A)v\|.\end{aligned}$$

From analyses of the GMRES method it is known that the inequality (1.3) can be strict. For example, certain nonnormal matrices $A \in \mathbb{R}^{4\times 4}$ were constructed in [2, 16] for which (1.3) is strict with $k = 3$, $f(z) = 1$, and $\varphi_i(z) = z^i$, $i = 1, 2, 3$. More recently, nonnormal matrices $A \in \mathbb{R}^{2n\times 2n}$, $n \ge 2$, were derived in [4] for which the inequality (1.3) is strict for all $k = 3, \ldots, 2n-1$, $f(z) = 1$, and $\varphi_i(z) = z^i$, $i = 1, \ldots, k$.

On the other hand, the following result is well known.

THEOREM 1.1. *Under the assumptions made in the first paragraph of the introduction, if $A \in \mathbb{F}^{n\times n}$ is normal, then equality holds in (1.3).*

At least three different proofs of this theorem or variants of it can be found in the literature. Greenbaum and Gurvits proved it for $\mathbb{F} = \mathbb{R}$ using mostly methods from matrix theory; see [7, Section 2] as well as Section 3 below for their formulation of the result. Using (analytic) methods of optimization theory, Joubert proved the equality for the case of the GMRES method with $f(z) = 1$, $\varphi_i(z) = z^i$, $i = 1, \ldots, k$, and he distinguished the cases $\mathbb{F} = \mathbb{R}$ and $\mathbb{F} = \mathbb{C}$; see [11, Theorem 4]. Finally, Bellalij, Saad, and Sadok also considered the GMRES case with $\mathbb{F} = \mathbb{C}$, and they applied methods from constrained convex optimization; see [1, Theorem 2.1].

In this paper we present yet another proof of Theorem 1.1, which is rather simple because it fully exploits the link between matrix approximation problems for normal matrices and scalar approximation problems in the complex plane. We observe that when formulating the matrix approximation problems in (1.3) in terms of scalar approximation problems, the proof of Theorem 1.1 reduces to a straightforward application of a well-known characterization theorem of polynomials of best approximation in the complex plane. While the proof of the theorem for $\mathbb{F} = \mathbb{C}$ can be accomplished in just a few lines, the case $\mathbb{F} = \mathbb{R}$ contains some technical details that require additional attention.

The characterization theorem from approximation theory we use in this paper and some of its variants have been stated and applied also in other publications in this context, in particular in [1, Theorem 5.1]. To our knowledge the theorem has, however, not been used to give a simple and direct proof of Theorem 1.1.

Personal note. We have written this paper in memory of our colleague Bernd Fischer, who passed away on July 15, 2013. Bernd's achievements in the analysis of iterative methods for linear algebraic systems using results of approximation theory, including his nowadays classical monograph [5], continue to inspire us in our own work. One of Bernd's last publications in this area (before following other scientific interests), written jointly with Franz Peherstorfer (1950–2009) and published in 2001 in ETNA [6], is also based on a variant of the characterization theorem that we apply in this paper.

2. Characterization theorem and proof of Theorem 1.1.

In order to formulate the characterization theorem of best approximation in the complex plane, we follow the treatment of Rivlin and Shapiro [14] that has been summarized in Lorentz' book [13, Chapter 2].

Let Γ be a compact subset of $\mathbb{F}$, where either $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, and let $C(\Gamma)$ denote the set of continuous functions on Γ. If Γ consists of finitely many single points (which is the case of interest in this paper), then $g \in C(\Gamma)$ means that the function g has a well defined (finite) value at each point of Γ. For $g \in C(\Gamma)$ we denote the maximum norm on Γ by

$$\|g\|_\Gamma \equiv \max_{z \in \Gamma} |g(z)|.$$

Now let $f \in C(\Gamma)$ and $\varphi_1, \dots, \varphi_k \in C(\Gamma)$ be given functions with values in $\mathbb{F}$. As above, let $\mathcal{P}_k(\mathbb{F})$ denote the linear span of the functions $\varphi_1, \dots, \varphi_k$ with coefficients in $\mathbb{F}$. For $p \in \mathcal{P}_k(\mathbb{F})$, define

$$\Gamma(p) \equiv \{z \in \Gamma : \ |f(z) - p(z)| = \|f - p\|_\Gamma\}.$$

A function $p_* = \alpha_1\varphi_1 + \cdots + \alpha_k\varphi_k \in \mathcal{P}_k(\mathbb{F})$ is called a *polynomial of best approximation* for f on Γ when

$$\|f - p_*\|_\Gamma \ = \min_{p \in \mathcal{P}_k(\mathbb{F})} \|f - p\|_\Gamma. \tag{2.1}$$

Under the given assumptions, such a polynomial of best approximation exists; see, e.g., [13, Theorem 1, p. 17]. The following well known result (see, e.g., [13, Theorem 3, p. 22] or [14, pp. 672-674]) characterizes the polynomials of best approximation.

THEOREM 2.1. *In the notation established above, the following two statements are equivalent:*

1. *The function $p_* \in \mathcal{P}_k(\mathbb{F})$ is a polynomial of best approximation for f on Γ.*
2. *For the function $p_* \in \mathcal{P}_k(\mathbb{F})$ there exist ℓ pairwise distinct points $\mu_i \in \Gamma(p_*)$, $i = 1, \dots, \ell$, where $1 \le \ell \le k+1$ for $\mathbb{F} = \mathbb{R}$ and $1 \le \ell \le 2k+1$ for $\mathbb{F} = \mathbb{C}$, and ℓ real numbers $\omega_1, \dots, \omega_\ell > 0$ with $\omega_1 + \cdots + \omega_\ell = 1$, such that*

$$\sum_{j=1}^{\ell} \omega_j \left[f(\mu_j) - p_*(\mu_j)\right]\overline{p(\mu_j)} = 0, \quad \text{for all } p \in \mathcal{P}_k(\mathbb{F}). \tag{2.2}$$

A well known geometric interpretation of the condition (2.2) is that the origin is contained in the convex hull of the points

$$\left\{ \left([f(\mu) - p_*(\mu)]\overline{\varphi_1(\mu)}, \dots, [f(\mu) - p_*(\mu)]\overline{\varphi_k(\mu)} \right) \in \mathbb{F}^k \ : \ \mu \in \Gamma(p_*) \right\};$$

see, e.g., [13, Equation (5), p. 21]. Here we will not use this interpretation but rewrite (2.2) in terms of an algebraic orthogonality condition involving vectors and matrices. Using that condition we will be able to prove Theorem 1.1 in a straightforward way. We will distinguish the cases of complex and real normal matrices because the real case contains some subtleties.

2.1. Proof of Theorem 1.1 for $\mathbb{F} = \mathbb{C}$. Let $A \in \mathbb{C}^{n \times n}$ be normal. Then A is unitarily diagonalizable, $A = Q\Lambda Q^H$ with $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ and $QQ^H = Q^HQ = I_n$. In the notation established above, let $\Gamma = \{\lambda_1, \dots, \lambda_n\}$ and suppose that $p_* \in \mathcal{P}_k(\mathbb{C})$ is a polynomial of best approximation for f on Γ so that statement 2 from Theorem 2.1 applies to p_*. With this setting, the matrix approximation problem (1.2) can be seen as the scalar best approximation problem (2.1), i.e.,

$$\min_{p \in \mathcal{P}_k(\mathbb{C})} \|f(A) - p(A)\| = \min_{p \in \mathcal{P}_k(\mathbb{C})} \|f(\Lambda) - p(\Lambda)\| = \min_{p \in \mathcal{P}_k(\mathbb{C})} \|f - p\|_\Gamma \,.$$

Without loss of generality, we may assume that the eigenvalues of A are ordered so that $\lambda_j = \mu_j$ for $j = 1, \dots, \ell$. We denote

$$\delta \equiv \|f - p_*\|_\Gamma = |f(\lambda_j) - p_*(\lambda_j)|, \quad j = 1, \dots, \ell.$$

Next, we define the vector

$$v_* \equiv Q\xi, \quad \text{where} \quad \xi \equiv [\xi_1, \dots, \xi_\ell, 0, \dots, 0]^T \in \mathbb{C}^n, \quad \xi_j \equiv \sqrt{\omega_j}, \; j = 1, \dots, \ell. \tag{2.3}$$

Since Q is unitary and $\omega_1 + \cdots + \omega_\ell = 1$, we have $\|v_*\| = 1$.

The condition (2.2) can be written as

$$\begin{aligned} 0 &= \sum_{j=1}^{\ell} |\xi_j|^2 \overline{p(\lambda_j)} \, [f(\lambda_j) - p_*(\lambda_j)] = \xi^H p(\Lambda)^H \, [f(\Lambda) - p_*(\Lambda)] \, \xi \\ &= v_*^H p(A)^H \, [f(A) - p_*(A)] \, v_* \, , \quad \text{for all } p \in \mathcal{P}_k(\mathbb{C}), \end{aligned}$$

or, equivalently,

$$f(A)v_* - p_*(A)v_* \perp p(A)v_* \, , \quad \text{for all } p \in \mathcal{P}_k(\mathbb{C}).$$

It is well known that this algebraic orthogonality condition with respect to the Euclidean inner product is equivalent to the optimality condition

$$\|f(A)v_* - p_*(A)v_*\| = \min_{p \in \mathcal{P}_k(\mathbb{C})} \|f(A)v_* - p(A)v_*\|;$$

see, e.g., [12, Theorem 2.3.2].

Using the previous relations we now obtain

$$\begin{aligned} \min_{p \in \mathcal{P}_k(\mathbb{C})} \|f(A) - p(A)\| = \delta &= \left(\sum_{j=1}^{\ell} |\xi_j|^2 \delta^2 \right)^{1/2} \\ &= \left(\sum_{j=1}^{\ell} |\xi_j|^2 \, |f(\lambda_j) - p_*(\lambda_j)|^2 \right)^{1/2} \\ &= \| \, [f(\Lambda) - p_*(\Lambda)] \, \xi \| \\ &= \|Q \, [f(\Lambda) - p_*(\Lambda)] \, Q^H Q\xi\| \\ &= \|f(A)v_* - p_*(A)v_*\| \\ &= \min_{p \in \mathcal{P}_k(\mathbb{C})} \|f(A)v_* - p(A)v_*\| \\ &\leq \max_{\substack{v \in \mathbb{C}^n \\ \|v\|=1}} \min_{p \in \mathcal{P}_k(\mathbb{C})} \|f(A)v - p(A)v\|. \end{aligned}$$

This is just the reverse of the inequality (1.3) for $\mathbb{F} = \mathbb{C}$, and hence the proof of Theorem 1.1 for $\mathbb{F} = \mathbb{C}$ is complete.

2.2. Proof of Theorem 1.1 for $\mathbb{F} = \mathbb{R}$. If $A \in \mathbb{R}^{n \times n}$ is symmetric, then we can write $A = Q\Lambda Q^T$ with a real diagonal matrix Λ and a real orthogonal matrix Q. The proof presented in the previous section also works in this case. In particular, for a real matrix Q, the vector $v_* = Q\xi$ constructed in (2.3) is real, and for a real matrix A, the maximization in (1.3) is performed over $v \in \mathbb{R}^n$.

From now on we consider a general normal matrix $A \in \mathbb{R}^{n\times n}$. In the spectral decomposition $A = Q\Lambda Q^H$, the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ and the unitary matrix Q are in general complex. Since this would lead to a complex vector $v_* = Q\xi$ in (2.3), the previous proof requires some modifications.

As above, let $\Gamma = \{\lambda_1, \dots, \lambda_n\}$. Since A is real, the set Γ may contain non-real points (appearing in complex conjugate pairs), and thus we must allow complex-valued functions $f \in C(\Gamma)$ and $\varphi_1, \dots, \varphi_k \in C(\Gamma)$. This means that we must work with Theorem 2.1 for $\mathbb{F} = \mathbb{C}$, although A is real. However, we will assume that for each eigenvalue λ_j of A the given functions f and $\varphi_1, \dots, \varphi_k$ satisfy

$$\overline{f(\lambda_j)} = f(\overline{\lambda}_j) \quad \text{and} \quad \overline{\varphi_i(\lambda_j)} = \varphi_i(\overline{\lambda}_j), \quad i = 1, \dots, k. \tag{2.4}$$

This is a natural assumption for real matrices A since it guarantees that the matrices $f(A)$ and $\varphi_1(A), \dots, \varphi_k(A)$ are real as well; see [9, Remark 1.9] (for analytic functions it is actually a necessary and sufficient condition; see [9, Theorem 1.18]).

Now let $q_* = \sum_{i=1}^k \alpha_i \varphi_i \in \mathcal{P}_k(\mathbb{C})$ be a polynomial of best approximation for f on Γ. Then, for any eigenvalue λ_j of A,

$$\Big|f(\lambda_j) - \sum_{i=1}^k \alpha_i \varphi_i(\lambda_j)\Big| = \Big|\overline{f(\lambda_j)} - \sum_{i=1}^k \overline{\alpha}_i \overline{\varphi_i(\lambda_j)}\Big| = \Big|f(\overline{\lambda}_j) - \sum_{i=1}^k \overline{\alpha}_i \varphi_i(\overline{\lambda}_j)\Big|.$$

Since both λ_j and $\overline{\lambda}_j$ are elements of Γ, we see that also $\overline{q}_* \equiv \sum_{i=1}^k \overline{\alpha}_i \varphi_i$ is a polynomial of best approximation for f on Γ. Denote

$$\delta \equiv \|f - q_*\|_\Gamma = \|f - \overline{q}_*\|_\Gamma\,,$$

then for any $0 \le \alpha \le 1$ we obtain

$$\begin{aligned} \delta &\le \|f - \alpha q_* - (1-\alpha)\overline{q}_*\|_\Gamma = \|\alpha(f - q_*) + (1-\alpha)(f - \overline{q}_*)\|_\Gamma \\ &\le \alpha \, \|f - q_*\|_\Gamma + (1-\alpha)\, \|f - \overline{q}_*\|_\Gamma = \delta, \end{aligned}$$

which shows that any polynomial of the form $\alpha q_* + (1-\alpha)\overline{q}_*$, $0 \le \alpha \le 1$, is also a polynomial of best approximation for f on Γ. In particular, for $\alpha = \frac{1}{2}$ we obtain the *real* polynomial of best approximation

$$p_* \equiv \frac{1}{2}(q_* + \overline{q}_*) \ \in \ \mathcal{P}_k(\mathbb{R}).$$

Using $p_* \in \mathcal{P}_k(\mathbb{R})$ and (2.4) we get

$$|f(z) - p_*(z)| = |\overline{f(z) - p_*(z)}| = |f(\overline{z}) - p_*(\overline{z})|, \quad \text{for all } z \in \Gamma.$$

Therefore, the set $\Gamma(p_*)$ of all points z which satisfy $|f(z) - p_*(z)| = \|f - p_*\|_\Gamma$ is symmetric with respect to the real axis, i.e., $z \in \Gamma(p_*)$ if and only if $\overline{z} \in \Gamma(p_*)$.

For simplicity of notation we denote

$$\zeta_p(z) \equiv [f(z) - p_*(z)]\overline{p(z)}.$$

In the definition of $\zeta_p(z)$ we indicate only its dependence on p and z since f is a given function and p_* is fixed. If $p \in \mathcal{P}_k(\mathbb{R})$, then the corresponding function $\zeta_p(z)$ satisfies $\overline{\zeta_p(z)} = \zeta_p(\overline{z})$ for all $z \in \Gamma$.

Now, Theorem 2.1 (with $\mathbb{F} = \mathbb{C}$) implies the existence of a set

$$G_* \equiv \{\mu_1, \dots, \mu_\ell\} \subseteq \Gamma(p_*) \subseteq \Gamma,$$

and the existence of positive real numbers $\omega_1, \dots, \omega_\ell$ with $\sum_{j=1}^{\ell} \omega_j = 1$ such that

$$\sum_{j=1}^{\ell} \omega_j \, \zeta_p(\mu_j) = 0, \quad \text{for all } p \in \mathcal{P}_k(\mathbb{R}), \tag{2.5}$$

where we have used that $\mathcal{P}_k(\mathbb{R}) \subset \mathcal{P}_k(\mathbb{C})$. To define a convenient real vector v_* similar to the construction leading to (2.3), we will "symmetrize" the condition (2.5) with respect to the real axis.

Taking complex conjugates in (2.5) and using that $\zeta_p(z) = \zeta_p(\overline{z})$ for any $z \in \Gamma$, we obtain another relation of the form

$$\sum_{j=1}^{\ell} \omega_j \, \zeta_p(\overline{\mu}_j) = 0, \quad \text{for all } p \in \mathcal{P}_k(\mathbb{R}),$$

and therefore

$$\frac{1}{2} \sum_{j=1}^{\ell} \omega_j \, \zeta_p(\mu_j) + \frac{1}{2} \sum_{j=1}^{\ell} \omega_j \, \zeta_p(\overline{\mu}_j) = 0, \quad \text{for all } p \in \mathcal{P}_k(\mathbb{R}). \tag{2.6}$$

Here (2.6) is the desired "symmetrized" condition. We now define the set

$$G_*^{\mathrm{sym}} \equiv \{\theta_1, \dots, \theta_m\} \equiv G_* \cup \overline{G}_*,$$

where each $\theta_i \in G_*^{\mathrm{sym}}$ corresponds to some μ_j or $\overline{\mu}_j$, and clearly $\ell \leq m \leq 2\ell$. (The exact value of m is unimportant for our construction.) Writing the condition (2.6) as a single sum over all points from G_*^{sym}, we get

$$\sum_{i=1}^{m} \widetilde{\omega}_i \, \zeta_p(\theta_i) = 0, \quad \text{for all } p \in \mathcal{P}_k(\mathbb{R}), \tag{2.7}$$

where the coefficients $\widetilde{\omega}_i$ are defined as follows.

If $\mu_j \in \mathbb{R}$, then $\zeta_p(\mu_j)$ appears in both sums in (2.6) with the same coefficient $\omega_j/2$. Since $\theta_i = \mu_j \in \mathbb{R}$, the term $\zeta_p(\theta_i)$ appears in (2.7) with the coefficient $\widetilde{\omega}_i = \omega_j$.

If $\mu_j \notin \mathbb{R}$ and $\overline{\mu}_j \notin G_*$, then $\zeta_p(\mu_j)$ appears only in the left sum in (2.6) with the coefficient $\omega_j/2$. Therefore, the term $\zeta_p(\mu_j)$ corresponds to a single term $\zeta_p(\theta_i)$ in (2.7) with the coefficient $\widetilde{\omega}_i = \omega_j/2$. Similarly, $\zeta_p(\overline{\mu}_j)$ appears only in the right sum in (2.6) with the coefficient $\omega_j/2$, and it corresponds to a single term, say $\zeta_p(\theta_s)$, in (2.7) with the coefficient $\widetilde{\omega}_s = \omega_j/2$.

If $\mu_j \notin \mathbb{R}$ and $\overline{\mu}_j \in G_*$, then $\overline{\mu}_j = \mu_s$ for some index $s \neq j$, $1 \leq s \leq \ell$. Therefore, the term $\zeta_p(\mu_j)$ appears in both sums in (2.6), in the left sum with the coefficient $\omega_j/2$ and in the right sum with the coefficient $\omega_s/2$. Hence, $\zeta_p(\mu_j)$ corresponds to a single term $\zeta_p(\theta_i)$ in (2.7) with the coefficient $\widetilde{\omega}_i = \omega_j/2 + \omega_s/2$. Similarly, $\zeta_p(\overline{\mu}_j)$ corresponds to the term $\zeta_p(\overline{\theta}_i)$ in (2.7) with the coefficient equal to $\omega_j/2 + \omega_s/2$.

One can easily check that $\widetilde{\omega}_i > 0$, for $i = 1, \dots, m$, and that

$$\sum_{i=1}^{m} \widetilde{\omega}_i = 1.$$

Moreover, if $\theta_j = \overline{\theta}_i$ for $j \neq i$, then $\widetilde{\omega}_j = \widetilde{\omega}_i$.

Based on the relation (2.7) we set

$$v_* \equiv Q\xi, \quad \xi \equiv [\xi_1, \ldots, \xi_n]^T \in \mathbb{R}^n,$$

where the ξ_j, $j = 1, \ldots, n$, are defined as follows: if $\lambda_j \in G_*^{\text{sym}}$, then there exits an index i such that $\lambda_j = \theta_i$, and we define $\xi_j \equiv \sqrt{\widetilde{\omega}_i}$. If $\lambda_j \notin G_*^{\text{sym}}$, we set $\xi_j = 0$.

It remains to justify that the resulting vector v_* is real. If $\lambda_j \in \mathbb{R}$, then the corresponding eigenvector q_j (i.e., the jth column of the matrix Q) is real, and $\xi_j q_j$ is real. If $\lambda_j \notin \mathbb{R}$ and $\lambda_j \in G_*^{\text{sym}}$, then also $\overline{\lambda}_j \in G_*^{\text{sym}}$, and $\overline{\lambda}_j = \lambda_i$ for some $i \neq j$. The corresponding eigenvector is $q_i = \overline{q}_j$, and since $\xi_i = \xi_j$, the linear combination $\xi_j q_j + \xi_i q_i = \xi_j(q_j + \overline{q}_j)$ is a real vector. Therefore, the resulting vector $v_* = Q\xi$ is real.

Using (2.7), analogously to the previous section, we get

$$0 = v_*^T p(A)^T \left[f(A) - p_*(A)\right] v_* \,, \quad \text{for all } p \in \mathcal{P}_k(\mathbb{R}),$$

or, equivalently,

$$\|f(A)v_* - p_*(A)v_*\| = \min_{p \in \mathcal{P}_k(\mathbb{R})} \|f(A)v_* - p(A)v_*\|$$

so that

$$\begin{aligned}\min_{p \in \mathcal{P}_k(\mathbb{R})} \|f(A) - p(A)\| = \delta &= \|f(A)v_* - p_*(A)v_*\| \\ &= \min_{p \in \mathcal{P}_k(\mathbb{R})} \|f(A)v_* - p(A)v_*\| \\ &\leq \max_{\substack{v \in \mathbb{R}^n \\ \|v\|=1}} \min_{p \in \mathcal{P}_k(\mathbb{R})} \|f(A)v - p(A)v\|.\end{aligned}$$

This is just the reverse of the inequality (1.3) for $\mathbb{F} = \mathbb{R}$, and hence the proof of Theorem 1.1 for $\mathbb{F} = \mathbb{R}$ is complete.

3. A different formulation.

Theorem 1.1 can be easily rewritten as a statement about pairwise commuting normal matrices. In the following we only discuss the complex case. The real case requires an analogous treatment as in Section 2.2.

Let $A_0, A_1, \ldots, A_k \in \mathbb{C}^{n\times n}$ be pairwise commuting normal matrices. Then these matrices can be simultaneously unitarily diagonalized, i.e., there exists a unitary matrix $U \in \mathbb{C}^{n\times n}$ so that

$$U^H A_i U = \Lambda_i = \operatorname{diag}(\lambda_1^{(i)}, \ldots, \lambda_n^{(i)}), \quad i = 0, 1, \ldots, k;$$

see, e.g., [10, Theorem 2.5.5]. Let $\Gamma \equiv \{\lambda_1, \ldots, \lambda_n\}$ be an arbitrary set containing n pairwise distinct complex numbers, and let $A \equiv U\operatorname{diag}(\lambda_1, \ldots, \lambda_n)U^H \in \mathbb{C}^{n\times n}$. We now *define* the functions $f \in C(\Gamma)$ and $\varphi_1, \ldots, \varphi_k \in C(\Gamma)$ to be any functions satisfying

$$f(\lambda_j) \equiv \lambda_j^{(0)}, \quad \varphi_i(\lambda_j) \equiv \lambda_j^{(i)}, \quad j = 1, \ldots, n, \;\; i = 1, \ldots, k.$$

Then $f(A) = A_0$ and $\varphi_i(A) = A_i$ for $i = 1, \ldots, k$, so that Theorem 1.1 implies

$$\begin{aligned}\max_{\substack{v \in \mathbb{C}^n \\ \|v\|=1}} \min_{\alpha_1,\ldots,\alpha_k \in \mathbb{C}} \|A_0 v - \sum_{i=1}^k \alpha_i A_i v\| &= \max_{\substack{v \in \mathbb{C}^n \\ \|v\|=1}} \min_{\alpha_1,\ldots,\alpha_k \in \mathbb{C}} \|f(A)v - \sum_{i=1}^k \alpha_i \varphi_i(A) v\| \\ &= \min_{\alpha_1,\ldots,\alpha_k \in \mathbb{C}} \|f(A) - \sum_{i=1}^k \alpha_i \varphi_i(A)\| \\ &= \min_{\alpha_1,\ldots,\alpha_k \in \mathbb{C}} \|A_0 - \sum_{i=1}^k \alpha_i A_i\|.\end{aligned}$$

This equality is in fact the version of Theorem 1.1 proven by Greenbaum and Gurvits in [7, Theorem 2.3] for the case $\mathbb{F} = \mathbb{R}$.

Acknowledgments. We thank the anonymous referees for comments that helped to improve the presentation.

REFERENCES

[1] M. BELLALIJ, Y. SAAD, AND H. SADOK, *Analysis of some Krylov subspace methods for normal matrices via approximation theory and convex optimization*, Electron. Trans. Numer. Anal., 33 (2008/09), pp. 17–30. http://etna.mcs.kent.edu/vol.33.2008-2009/pp17-30.dir

[2] V. FABER, W. JOUBERT, E. KNILL, AND T. MANTEUFFEL, *Minimal residual method stronger than polynomial preconditioning*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 707–729.

[3] V. FABER, J. LIESEN, AND P. TICHÝ, *On Chebyshev polynomials of matrices*, SIAM J. Matrix Anal. Appl., 31 (2009/10), pp. 2205–2221.

[4] ———, *Properties of worst-case GMRES*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 1500–1519.

[5] B. FISCHER, *Polynomial Based Iteration Methods for Symmetric Linear Systems*, Wiley, Chichester, 1996.

[6] B. FISCHER AND F. PEHERSTORFER, *Chebyshev approximation via polynomial mappings and the convergence behaviour of Krylov subspace methods*, Electron. Trans. Numer. Anal., 12 (2001), pp. 205–215. http://etna.mcs.kent.edu/vol.12.2001/pp205-215.dir

[7] A. GREENBAUM AND L. GURVITS, *Max-min properties of matrix factor norms*, SIAM J. Sci. Comput., 15 (1994), pp. 348–358.

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

[9] N. J. HIGHAM, *Functions of Matrices. Theory and Computation*, SIAM, Philadelphia, 2008.

[10] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, 1990.

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

[12] J. LIESEN AND Z. STRAKOŠ, *Krylov Subspace Methods. Principles and Analysis*, Oxford University Press, Oxford, 2013.

[13] G. G. LORENTZ, *Approximation of Functions*, 2nd ed., Chelsea, New York, 1986.

[14] T. J. RIVLIN AND H. S. SHAPIRO, *A unified approach to certain problems of approximation and minimization*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 670–699.

[15] 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.

[16] K.-C. TOH, *GMRES vs. ideal GMRES*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 30–36.

[17] K.-C. TOH AND L. N. TREFETHEN, *The Chebyshev polynomials of a matrix*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 400–419.