Electronic Transactions on Numerical Analysis.
Volume 36, pp. 99-112, 2010.

ISSN 1068-9613.

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

THE STRUCTURED DISTANCE TO NEARLY NORMAL MATRICES*

LAURA SMITHIES†

Dedicated to Richard S. Varga on the occasion of his 80th birthday

Abstract. In this note we examine the algebraic variety $\mathcal{I}_\Lambda$ of complex tridiagonal $n \times n$ matrices T, such that $T^*T - TT^* = \Lambda$, where Λ is a fixed real diagonal matrix. If $\Lambda = \mathbf{0}$ then $\mathcal{I}_\Lambda$ is $\mathcal{N}_\mathbf{T}$, the set of tridiagonal normal matrices. For $\Lambda \neq \mathbf{0}$, we identify the structure of the matrices in $\mathcal{I}_\Lambda$ and analyze the suitability for eigenvalue estimation using normal matrices for elements of $\mathcal{I}_\Lambda$. We also compute the Frobenius norm of elements of $\mathcal{I}_\Lambda$, describe the algebraic subvariety $\mathcal{M}_\Lambda$ consisting of elements of $\mathcal{I}_\Lambda$ with minimal Frobenius norm, and calculate the distance from a given complex tridiagonal matrix to $\mathcal{I}_\Lambda$.

Key words. nearness to normality, tridiagonal matrix, Kreĭn spaces, eigenvalue estimation, Geršgorin type sets

AMS subject classifications. 65F30, 65F35, 15A57, 15A18, 47A25

1. Introduction.

In this note, we establish a generalization of the matrix nearness problem which was solved by S. Noschese, L. Pasquini, and L. Reichel in [5]. The structure for the type of generalization of nearness to normality which is considered in this note was first suggested to me by Roger Horn, in connection with [2]. The homework set in [3, page 128] also discusses this type of generalization of normality.

Let $\mathcal{T}^r$ denote the set of all *real* $n \times n$ irreducible tridiagonal matrices and let $\mathcal{I}^r$ denote the algebraic variety of real normal irreducible tridiagonal $n \times n$ matrices. The paper [5] presents the following, for any fixed $T \in \mathcal{T}^r$:

(i) a formula for Frobenius distance $d_F(T, \mathcal{I}^r)$ and an easily calculated upper bound on this distance;
(ii) a formula for a real normal tridiagonal matrix $\hat{T}$, such that $\|T - \hat{T}\|_F$ is equal to $d_F(T, \mathcal{I}^r)$;
(iii) simplified versions of (i) and (ii) for Toeplitz matrices.

In order to generalize the above problem, we must fix some notation. Let $\mathbb{C}^{n\times n}$ denote the set of complex $n \times n$ matrices and let $M \in \mathbb{C}^{n\times n}$. Define the *adjoint* of M, $M^* = \bar{M}^t$, to be the conjugate transpose of M. Recall that M is defined to be *self adjoint* if $M = M^*$ and *normal* if the *commutant* of M and its adjoint, $[M, M^*] := MM^* - M^*M$, equals $\mathbf{0}$ in $\mathbb{C}^{n\times n}$. Throughout this note, fix a real diagonal matrix:

$$\Lambda = \mathrm{Diag}(\lambda_j), \text{ such that } \Lambda \neq \mathbf{0}, \;\; \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n, \quad \text{and} \quad \sum_{l=1}^{n} \lambda_l = 0. \tag{1.1}$$

Let $\mathcal{T}$ denote the set of tridiagonal complex matrices. The purpose of this paper is to investigate the set

$$\mathcal{I}_\Lambda = \{M \in \mathcal{T} \;:\; MM^* - M^*M = \Lambda\},$$

and to describe the distance from normality for the elements of $\mathcal{I}_\Lambda$. More precisely,

*Received March 13, 2009. Accepted for publication August 9, 2009. Published online on January 20, 2010. Recommended by L. Reichel.
†Department of Mathematical Sciences, Kent State University, Kent, OH 44242 (smithies@math.kent.edu).

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

(i) we provide simple formulas, in terms of Λ, for the elements of $\mathcal{I}_\Lambda$;
(ii) for $T \in \mathcal{T}$, we give a formula for the Frobenius distance $d_F^2(T, \mathcal{I}_\Lambda)$, and an easily calculated upper bound for this distance;
(iii) we provide formulas for the elements of the subvariety $\mathcal{M}_\Lambda$ of $\mathcal{I}_\Lambda$ whose Frobenius norm is minimal, and for the unique element $M_\Lambda \in \mathcal{M}_\Lambda$ with only nonnegative entries;
(iv) we combine the above results with those of [2] to describe for any $M \in \mathcal{I}_\Lambda$ the distance to normality both in the Frobenius norm and in the sense of the suitability of M for eigenvalue estimation through normal matrices.

This paper is organized as follows. Section 2 recalls some elementary results and introduces notation which will be used throughout the paper; Section 3 presents a characterization of the elements of $\mathcal{I}_\Lambda$. In Section 4, we give a formula for the distance $d_F^2(T, \mathcal{I}_\Lambda)$ from $T \in \mathcal{T}$ to $\mathcal{I}_\Lambda$, and we describe the algebraic variety $\mathcal{M}_\Lambda$ of the elements of $\mathcal{I}_\Lambda$ of minimal Frobenius norm. In Section 5, we describe the distance from normality for the elements of $\mathcal{I}_\Lambda$, in part, by applying results from [2]. The final section discusses some conclusions and possible extensions.

2. Background and notation. This section defines notation used in the sequel and recalls some elementary results. Table 2.1 collects our most important notation.

TABLE 2.1
Sets used in this paper.

$\mathcal{T}$ = the tridiagonal matrices in $\mathbb{C}^{n\times n}$	
$\mathcal{N}$ = the normal matrices in $\mathbb{C}^{n\times n}$	$\mathcal{N}_{\mathbf{T}} = \mathcal{N} \cap \mathcal{T}$
$\mathcal{S}$ = self adjoint ($A^* = A$) in $\mathbb{C}^{n\times n}$	$\mathcal{S}_{\mathbf{T}} = \mathcal{S} \cap \mathcal{T}$
$\mathcal{A}$ = anti-self adjoint ($A^* = -A$) in $\mathbb{C}^{n\times n}$	$\mathcal{A}_{\mathbf{T}} = \mathcal{A} \cap \mathcal{T}$
$\mathcal{I}^r$ = real, irreducible matrices in $\mathcal{N}_{\mathbf{T}}$	
$\Lambda = \mathrm{Diag}(\lambda_j), \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n, \quad \lambda_1 > 0, \quad \sum_{l=1}^n \lambda_l = 0.$	
$\mathcal{I}_\Lambda = \{M \in \mathcal{T} \ : \ [M, M^*] = \Lambda\}$	
$\mathcal{I}_{4\Lambda} = \{M \in \mathcal{T} \ : \ [M, M^*] = 4\Lambda\}$	
$\mathcal{P}_\Lambda = \{(A, B) \in \mathcal{S}_{\mathbf{T}} \oplus \mathcal{A}_{\mathbf{T}} \ : \ [A, B] = -2\Lambda\}$	

Let $T = \mathrm{Tri}(\sigma, \delta, \tau)$ denote the element T of $\mathcal{T}$ with lower and upper bands σ and τ in $\mathbb{C}^{n-1}$ and diagonal $\delta \in \mathbb{C}^n$. That is,

$$T = \begin{bmatrix} \delta_1 & \tau_1 & & \mathbf{0} \\ \sigma_1 & \delta_2 & \ddots & \\ & \ddots & \ddots & \tau_{n-1} \\ \mathbf{0} & & \sigma_{n-1} & \delta_n \end{bmatrix}. \tag{2.1}$$

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

Note that we use the terminology *anti-self adjoint* to refer to the constraint $A^* = -A$ on $A \in \mathbb{C}^{n\times n}$. When $A \in \mathbb{R}^{n\times n}$ this simplifies to $A^t = -A$, and the terminology anti-symmetric and skew-symmetric are also used. We use a superscript, as in $\mathcal{A}^r$, when our matrices are constrained to be real.

We remark that, in contrast to the sets $\mathcal{I}^r$ defined in [5], the matrices in our sets $\mathcal{I}_\Lambda$ and $\mathcal{P}_\Lambda$ are allowed to be reducible. However, the choice to arrange the entries of Λ in non-increasing order, plays the role of irreducibility. More precisely, cf., [8, page 11], let ϕ be a permutation of $\{1, \cdots, n\}$ and let P denote the corresponding permutation matrix, i.e., $P_{j,k} = I_{j,\phi(k)}$. We say that a matrix $A \in \mathbb{C}^{n\times n}$ is *reducible* if there exists a rearrangement of coordinates with respect to which A is block diagonal. That is, if there exists a permutation $P \in \mathbb{R}^{n\times n}$, such that

$$PAP^* = \begin{bmatrix} A_{1,1} & A_{1,2} \\ 0 & A_{2,2} \end{bmatrix}.$$

It is easy to check that $P\,\mathrm{Diag}(\lambda_j)P^* = \mathrm{Diag}(\lambda_{\phi(j)})$ and that $[M, M^*] = \Lambda$ if and only if $[PMP^*, (PMP^*)^*] = \mathrm{Diag}(\lambda_{\phi(j)})$. Our conditions on Λ imply that Λ cannot equal $\mathrm{Diag}(\lambda_{\phi(j)})$ for all permutations ϕ, and consequently, we do not limit our solution sets $\mathcal{P}_\Lambda$ and $\mathcal{I}_\Lambda$ to irreducible matrices.

Let $A, B \in \mathbb{C}^{n\times n}$. Recall that the Frobenius inner product and induced norm are defined as

$$(A, B)_F = \mathrm{Tr}(B^*A), \text{ and } \|A\|_F = \sqrt{(A, A)_F}.$$

For $x = (x_j)_{j=1}^n$ and $y = (y_j)_{j=1}^n$ in $\mathbb{C}^n$, their Euclidean inner product is

$$< x, y >= \sum_{j=1}^n x_j \bar{y}_j.$$

The corresponding vector norm is $\|x\|_2 = \sqrt{< x, x >}$ and the induced operator norm is $\|A\|_2 = \sup\{\|Ax\|_2 \ : \ \|x\|_2 = 1\}$.

Let $\Pi_{\mathcal{A}}$ and $\Pi_{\mathcal{S}}$ denote the projections of $\mathbb{C}^{n\times n}$ onto $\mathcal{S}$ and $\mathcal{A}$, respectively. That is, for $M \in \mathbb{C}^{n\times n}$ and $i = \sqrt{-1}$, let

$$\Pi_{\mathcal{S}}(M) = \frac{M + M^*}{2} = \mathrm{Re}(M) \quad \text{and} \quad \Pi_{\mathcal{A}}(M) = \frac{M - M^*}{2} = i\mathrm{Im}(M)$$

be its self adjoint and anti-self adjoint parts. It is easy to check that

$$\mathbb{C}^{n\times n} = \mathcal{S} \oplus \mathcal{A},$$

where $\oplus$ denotes the direct sum. More precisely, if $C \in \mathcal{S} \cap \mathcal{A}$ then $C^* = C$ and $C^* = -C$. So, $C = 0$. This, combined with the fact that both the set of self adjoint matrices and the set of anti-self adjoint matrices are closed under subtraction, implies that the decomposition of any $M \in \mathbb{C}^{n\times n}$ into a sum of a self adjoint and an anti-self adjoint matrix is unique. The natural structure on $\mathcal{S}_{\mathbf{T}} \oplus \mathcal{A}_{\mathbf{T}}$, from the point-of-view of functional analysis, is that of a Kreĭn space, as we discuss in Section 5.

The sets $\mathcal{S}^r$ and $\mathcal{A}^r$ are orthogonal with respect to the Frobenius inner product. However, this is not true of $\mathcal{S}$ and $\mathcal{A}$. For example, if $D = \mathrm{Diag}(\delta_j) \in \mathbb{R}^{n\times n}$, then $(D, iD) \in \mathcal{S} \oplus \mathcal{A}$ and $(D, iD)_F = \mathrm{Tr}(-iD^2) = -i\|\delta\|_2^2$. In general, if $A \in \mathcal{S}$ and $B \in \mathcal{A}$ then $(A, B)_F$ is a pure imaginary number. Indeed, $(A, B)_F = -\overline{(A, B)_F}$ because $\mathrm{Tr}(B^*A) = \overline{\mathrm{Tr}(A^*B)} = \overline{\mathrm{Tr}(-AB^*)} = -\overline{\mathrm{Tr}(B^*A)}$. On the other hand, the properties of the trace function imply that

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

$\|A-B\|_F^2 = \|A\|_F^2 - 2\mathrm{Re}((A,B)_F) + \|B\|_F^2$. Thus, sums and differences of matrices from $\mathcal{S}$ and $\mathcal{A}$ also satisfy the Pythagorean constraints:

$$\|A \pm B\|_F^2 = \|A\|_F^2 + \|B\|_F^2, \ \ \forall (A,B) \in \mathcal{S} \oplus \mathcal{A}. \tag{2.2}$$

Notice that M is tridiagonal if and only if M^* is, and this happens if and only if the self adjoint and anti-self adjoint parts of M are tridiagonal. Moreover, a routine calculation shows that for any $M \in \mathbb{C}^{n\times n}$,

$$[M, M^*] = -2[\Pi_{\mathcal{S}}(M), \Pi_{\mathcal{A}}(M)]. \tag{2.3}$$

Finally, we measure the (squared Frobenius) distance from a matrix $M \in \mathbb{C}^{n\times n}$ to a subset $\mathcal{X}$ of $\mathbb{C}^{n\times n}$ as

$$d_F^2(M, \mathcal{X}) = \inf\{\|E\|_F^2 : M + E \in \mathcal{X}\} = \inf\{\|M - E\|_F^2 : E \in \mathcal{X}\}.$$

Equivalently, for any $(A,B) \in \mathcal{S} \oplus \mathcal{A}$ and $\mathcal{X} \subseteq \mathcal{S} \oplus \mathcal{A}$ define

$$d_F^2((A,B), \mathcal{X}) = \inf\{\|A - E\|_F^2 + \|B - H\|_F^2 : (E,H) \in \mathcal{X}\}. \tag{2.4}$$

Note that equation (2.2) implies that the equivalence of $\mathbb{C}^{n\times n} = \mathcal{S} \oplus \mathcal{A}$ which is given by $M \to (\Pi_{\mathcal{A}}(M), \Pi_{\mathcal{S}}(M))$ and $(A,B) \to A + B$ is an isometry with the natural choices of topology.

3. Structure of $\mathcal{I}_\Lambda$. Recall that we have fixed the non-zero, real diagonal matrix Λ with entries which satisfy the conditions (1.1). In this section we will study the structure of tridiagonal complex matrices $M \in \mathcal{T}$, such that $[M, M^*] = \Lambda$. We shall see that the computations involved simplify with the following equivalences:

LEMMA 3.1. *Define the sets $\mathcal{I}_\Lambda$, $\mathcal{I}_{4\Lambda}$, and $\mathcal{P}_\Lambda$ as in Table 2.1. Then*

$$T \in \mathcal{I}_\Lambda \ \text{ if and only if } \ 2T \in \mathcal{I}_{4\Lambda},$$

and

$$M \in \mathcal{I}_{4\Lambda} \ \text{ if and only if } \ (\Pi_{\mathcal{S}}(M), \Pi_{\mathcal{A}}(M)) \in \mathcal{P}_\Lambda.$$

Moreover, for every $H \in \mathbb{C}^{n\times n}$,

$$4d_F^2(H, \mathcal{I}_\Lambda) = d_F^2(2H, \mathcal{I}_{4\Lambda}) = d_F^2((\Pi_{\mathcal{S}}(2H), \Pi_{\mathcal{A}}(2H)), \mathcal{P}_\Lambda).$$

Proof. The map $T \to 2T$ obviously defines a bijection between $\mathcal{I}_\Lambda$ and $\mathcal{I}_{4\Lambda}$. Equation (2.3) yields the equivalence between $\mathcal{I}_{4\Lambda}$ and $\mathcal{P}_\Lambda$. Now, let $H \in \mathbb{C}^{n\times n}$. Then,

$$d_F^2(2H, \mathcal{I}_{4\Lambda}) = \inf_{2T\in\mathcal{I}_{4\Lambda}} \|2H - 2T\|_F^2 = 4 \inf_{T\in\mathcal{I}_\Lambda} \|H - T\|_F^2 = 4d_F^2(H, \mathcal{I}_\Lambda),$$

and the Pythagorean property (2.2), implies that

$$d_F^2(2H, \mathcal{I}_{4\Lambda}) = d_F^2((\Pi_{\mathcal{S}}(2H), \Pi_{\mathcal{A}}(2H)), \mathcal{P}_\Lambda). \qquad \square$$

DEFINITION 3.1. *Let $\Lambda = \mathrm{Diag}(\lambda_j)$ be as in (1.1). For each $j = 1, \cdots, n$, define*

$$S_j = \sum_{l=1}^{j} \lambda_l.$$

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

REMARK 3.1. *We remark that for any $M \in \mathbb{C}^{n\times n}$, $[M, M^*]$ is self adjoint and has trace zero. Thus, our fixed right hand side in the equation $[M, M^*] = \Lambda$ is, up to choice of ordering for the real numbers λ_j, the general non-zero, diagonal right hand side.*

The motivations for choosing a diagonal matrix for the right hand side Λ will be discussed in the final section. The essential point is that we are interested in describing the extent to which a matrix M fails to be normal in terms of the rank of the commutant $[M, M^*]$. Our motivation for assuming that the λ_j are in non-increasing order is indicated by the following lemma.

LEMMA 3.2. *Let $\Lambda = \mathrm{Diag}(\lambda_l)$ satisfy condition (1.1) and let S_j denote the j-th partial sum. Then*

$$S_j > 0, \quad \textit{for all } \; j = 1, \cdots, n-1.$$

Proof. Since $\Lambda \neq \mathbf{0}$, we must have $\lambda_1 > 0$. Now let j be minimal, such that $S_j < 0$. Then $\lambda_j < 0$. Since the λ_l are non-increasing, this means $\lambda_l < 0$ for all $l = j, \cdots, n$. Therefore, $S_l < S_j < 0$ for all $l > j$. However, by assumption $S_n = 0$. □

We can now simplify our study of the structure of the matrices in $\mathcal{I}_\Lambda$. Let $\mathcal{S}_\mathbf{T}$ (respectively $\mathcal{A}_\mathbf{T}$) denote the self adjoint (respectively anti-self adjoint) matrices in $\mathcal{T}$, the complex tridiagonal matrices. Recall that we have defined

$$\mathcal{P}_\Lambda = \{(A, B) \in \mathcal{S}_\mathbf{T} \oplus \mathcal{A}_\mathbf{T} \;:\; [A, B] = -2\Lambda\}.$$

Because of Lemma 3.1, each $(A, B) \in \mathcal{P}_\Lambda$ corresponds to a unique $T = \frac{A+B}{2}$ in $\mathcal{I}_\Lambda$. Thus, the following lemma yields a characterization of the elements of $\mathcal{I}_\Lambda$.

LEMMA 3.3. *Let $\Lambda = \mathrm{Diag}(\lambda_l)$ satisfy condition (1.1). Let $(A, B) \in \mathcal{S}_\mathbf{T} \oplus \mathcal{A}_\mathbf{T}$ and denote the entries by*

$$A = \mathrm{Tri}(\bar{z}_j, a_j, z_j) \textit{ and } B = \mathrm{Tri}(-\bar{w}_j, ib_j, w_j).$$

Then $(A, B) \in \mathcal{P}_\Lambda$ if and only if the entries satisfy

$$\begin{array}{llll} (i) & \mathrm{Re}(z_j\bar{w}_j) & = S_j, & \forall j = 1, \cdots, n-1; \\ (ii) & z_j w_{j+1} & = w_j z_{j+1}, & \forall j = 1, \cdots, n-2; \\ (iii) & a_j & = a_{j+1}, & \forall j = 1, \cdots, n-1; \\ (iv) & b_j & = b_{j+1}, & \forall j = 1, \cdots, n-1. \end{array} \tag{3.1}$$

Proof. Note that since $A = A^*$, its diagonal (a_j) is real and since $B = -B^*$, the diagonal of B is imaginary. Let $C = [A, B]$. Since A and B are tridiagonal, C is pentadiagonal. Moreover, $C = C^*$. A routine calculation shows

$$\begin{array}{lcl} C_{j,j} & = & -2[\mathrm{Re}(z_j\bar{w}_j) - \mathrm{Re}(z_{j-1}\bar{w}_{j-1})], \\ C_{j,j+1} & = & (a_j - a_{j+1})w_j + i(b_{j+1} - b_j)z_j, \\ C_{j,j+2} & = & z_j w_{j+1} - w_j z_{j+1}. \end{array}$$

Equating C to -2Λ, it is easy to see that conditions (i) and (ii) of (3.1) hold. Recall that Λ is assumed to be non-zero and hence $\mathrm{Re}(z_j\bar{w}_j) = S_j \neq 0$, for all $j = 1, \cdots, n-1$. Thus, $z_j \neq 0$ and $w_j \neq 0$ for all j. The remaining equations say

$$(a_{j+1} - a_j) = i(b_{j+1} - b_j)\frac{z_j}{w_j} = i(b_{j+1} - b_j)\frac{z_j\overline{w_j}}{|w_j|^2}.$$

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

The left hand side is real, the right hand side has non-zero imaginary part, unless b_j is constant. Therefore, a_j and b_j are constant.

Conversely, let $(A, B) \in \mathcal{S}_{\mathbf{T}} \oplus \mathcal{A}_{\mathbf{T}}$ satisfy conditions (3.1), and let $C = [A, B]$. Clearly, $C_{j,j+1} = (a_j - a_{j+1})w_j + (b_{j+1} - b_j)z_j = 0$ and $C_{j,j+2} = z_j w_{j+1} - w_j z_{j+1} = 0$. Moreover,

$$C_{j,j} = -2[\mathrm{Re}(z_j \bar{w}_j) - \mathrm{Re}(z_{j-1}\bar{w}_{j-1})] = -2[S_j - S_{j-1}] = -2\lambda_j.$$

Thus, $(A, B) \in \mathcal{P}_\Lambda$. □

In view of the previous lemma, we denote the entries of an element (A, B) of $\mathcal{P}_\Lambda$ by $(A, B) = (Z + aI, W + ibI)$, where $a, b \in \mathbb{R}$, $Z = \mathrm{Tri}(\bar{z}_j, 0, z_j)$ and $W = \mathrm{Tri}(-\bar{w}_j, 0, w_j)$. Note that

$$[A, B] = [Z + aI, W + ibI] = [Z, W].$$

With this in mind, we can now describe the elements of $\mathcal{P}_\Lambda$ for $\Lambda \neq \mathbf{0}$. Let $\mathbb{T} = [0, 2\pi)/\sim$ denote the torus, and let $\mathbb{R}^+$ denote the positive real numbers. We will use the both of the notations θ and $e^{i\theta}$ to identify elements of $\mathbb{T}$.

THEOREM 3.4. *Let $\Lambda = \mathrm{Diag}(\lambda_l)$ satisfy condition (1.1), and let S_j to be the j-th partial sum of the λ_l. The ordered pairs from $\mathcal{S}_{\mathbf{T}} \oplus \mathcal{A}_{\mathbf{T}}$ which lie in $\mathcal{P}_\Lambda$ are parametrized bijectively by $\mathbb{R}^3 \times \mathbb{R}^+ \times \mathbb{T}^{n-1}$. Specifically, each $(n+3)$-tuple $(a, b, c, |w_1|, \theta_1, \cdots, \theta_{n-1})$ defines the complex, tridiagonal matrices $A = Z + aI$ and $B = W + ibI$, where the entries of Z and W satisfy*

$$\begin{array}{llll} & (i) & w_j = |w_j| e^{i\theta_j}, & \text{where } |w_j|^2 = \frac{S_j |w_1|^2}{\lambda_1}, \\ \text{and} & (ii) & z_j = r w_j, & \text{where } r = \frac{\lambda_1 + ic}{|w_1|^2}, \end{array} \tag{3.2}$$

for all $j = 1, \cdots, n-1$.

Proof. First, let $(a, b, c, |w_1|, \theta_1, \cdots, \theta_{n-1})$, be a fixed element in $\mathbb{R}^3 \times \mathbb{R}^+ \times \mathbb{T}^{n-1}$, and let $(Z + aI, W + ibI)$ be defined by plugging this $(n+3)$-tuple into the given formulas. It is easy to check that $z_1 \bar{w}_1 = \lambda_1 + ic$, and that conditions (i)-(iv) of Lemma 3.3 hold. Thus, $(A, B) \in \mathcal{P}_\Lambda$.

Now, let $(A, B) \in \mathcal{P}_\Lambda$. By the previous lemma $(A, B) = (Z + aI, W + ibI)$ for some $a, b \in \mathbb{R}$, and for all $j = 1, \cdots, n-1$, $\mathrm{Re}(z_j \bar{w}_j) = S_j$ and $S_j \neq 0$, by Lemma 3.2. So, $w_j \neq 0$ for all $j = 1, \cdots, n-1$. Define $\theta_j = \mathrm{Arg}(w_j)$ and let c denote $\mathrm{Im}(z_1 \bar{w}_1)$. We will show that (Z, W) is given by plugging the $(n+1)$-tuple $(c, |w_1|, \theta_1, \cdots, \theta_{n-1})$ into the above formulas. Since $(Z, W) \in \mathcal{P}_\Lambda$,

$$\mathrm{Re}(z_j \bar{w}_j) = S_j, \quad \text{and} \quad \frac{z_j}{w_j} = \frac{z_{j+1}}{w_{j+1}}.$$

Now,

$$\mathrm{Re}(\frac{z_j}{w_j}) = \frac{\mathrm{Re}(z_j \bar{w}_j)}{|w_j|^2} = \frac{S_j}{|w_j|^2}, \text{ and } \mathrm{Re}(\frac{z_j}{w_j}) = \mathrm{Re}(\frac{z_1}{w_1}) = \frac{\mathrm{Re}(z_1 \bar{w}_1)}{|w_1|^2} = \frac{\lambda_1}{|w_1|^2}.$$

Thus,

$$|w_j|^2 = \frac{S_j |w_1|^2}{\lambda_1}.$$

Similarly,

$$\mathrm{Im}(\frac{z_j}{w_j}) = \mathrm{Im}(\frac{z_1}{w_1}) = \frac{\mathrm{Im}(z_1 \bar{w}_1)}{|w_1|^2} = \frac{c}{|w_1|^2}.$$

Combined with our second formula for $\mathrm{Re}(\frac{z_j}{w_j})$, this tells us that

$$z_j = \frac{\lambda_1 + ic}{|w_1|^2} w_j \text{ for all } j = 1, \cdots, n-1.$$

Finally, suppose $(a, b, c, |w_1|, (\theta_j)_{j=1}^{n-1})$ and $(\hat{a}, \hat{b}, \hat{c}, |\hat{w}_1|, (\hat{\theta}_j)_{j=1}^{n-1})$ in $\mathbb{R}^3 \times \mathbb{R}^+ \times \mathbb{T}^{n-1}$ define (A, B) and $(\hat{A}, \hat{B})$ in $\mathcal{S}_{\mathbf{T}} \oplus \mathcal{A}_{\mathbf{T}}$ which are equal. Then, $a = \hat{a}$, $b = \hat{b}$, $z_j = \hat{z}_j$ and $w_j = \hat{w}_j$ for all $j = 1, \cdots, n-1$. Thus, $\theta_j = \mathrm{Arg}(w_j) = \mathrm{Arg}(\hat{w}_j) = \hat{\theta}_j$, and $c = \mathrm{Im}(z_1 \overline{w_1}) = \mathrm{Im}(\hat{z}_1 \overline{\hat{w}}_1) = \hat{c}$. □

COROLLARY 3.5. *Let $\Lambda = \mathrm{Diag}(\lambda_l)$ satisfy condition (1.1). Each $M \in \mathcal{I}_\Lambda$ is uniquely determined by an $(n+3)$-tuple*

$$(a, b, c, |w_1|, \theta_1, \cdots, \theta_{n-1}) \in \mathbb{R}^3 \times \mathbb{R}^+ \times \mathbb{T}^{n-1}.$$

Specifically, each such $(n+3)$-tuple defines the complex, tridiagonal matrix of the form

$$M = \frac{1}{2} \,\mathrm{Tri}((\bar{r} - 1)\, \overline{\omega}, a + ib, (r+1)\omega),$$

where $r = \frac{\lambda_1 + ic}{|w_1|^2}$, and $\omega = (\frac{\sqrt{S_j}|w_1|e^{i\theta_j}}{\sqrt{\lambda_1}}) \in \mathbb{C}^{n-1}$.

Proof. By the previous theorem, we know how $\mathbb{R}^3 \times \mathbb{R}^+ \times \mathbb{T}^{n-1}$ parametrizes $\mathcal{P}_\Lambda$ and we know from Lemma 3.1 that each $M \in \mathcal{I}_\Lambda$ is uniquely defined by $M = \frac{A+B}{2}$ for some $(A, B) \in \mathcal{P}_\Lambda$. □

4. Distance formulas. In this section we establish a formula for the elements M_Λ^θ in $\mathcal{I}_\Lambda$ of minimal Frobenius norm. Specifically, we show the minimal elements form an algebraic subvariety $\mathcal{M}_\Lambda$ which is isomorphic to $\mathbb{T}^{n-1}$. We also define M_Λ the unique matrix in $\mathcal{M}_\Lambda$ with *nonnegative* entries. Equivalently, $-M_\Lambda$ is the unique Z-matrix, cf., [8], in $\mathcal{M}_\Lambda$. We also give a formula for the distance from a fixed $T \in \mathcal{T}$ to $\mathcal{I}_\Lambda$, and an easily computed upper bound on this distance.

First we need some preliminary calculations on the behavior of the Frobenius norm on tridiagonal matrices. Let $T \in \mathcal{T}$ and write

$$T = \mathrm{Tri}(\sigma, \delta, \tau), \quad D = \mathrm{Diag}(\delta_j), \quad \text{and } M = \mathrm{Tri}(\sigma, 0, \tau).$$

Then

$$\|T\|_F^2 = \|D\|_F^2 + \|M\|_F^2 = \|\delta\|_2^2 + \|\sigma\|_2^2 + \|\tau\|_2^2. \tag{4.1}$$

In particular,

$$\text{if } A = \mathrm{Tri}(\pm\bar{\sigma}, \delta, \sigma), \text{ then } \|A\|_F^2 = \|\delta\|_2^2 + 2\|\sigma\|_2^2. \tag{4.2}$$

Formula (4.1) follows from the straightforward calculations that $\mathrm{Tr}(D^*M) = 0$, $\|D\|_F^2 = \|\delta\|_2^2$, and

$$\mathrm{Tr}(M^*M) = \sum_{j=1}^{n} M_{j,j-1}^* M_{j-1,j} + M_{j,j+1}^* M_{j+1,j} = \sum_{j=2}^{n} \overline{\tau}_{j-1}\tau_{j-1} + \sum_{j=1}^{n-1} \overline{\sigma}_j \sigma_j.$$

Now we are ready to find the minimal elements of $\mathcal{I}_\Lambda$. Let $M \in \mathcal{I}_\Lambda$ and let $(A, B) = (\Pi_{\mathcal{S}}(2M), \Pi_{\mathcal{A}}(2M))$ denote the corresponding element in $\mathcal{P}_\Lambda$. Write $A = \mathrm{Tri}(\bar{z}_j, a, z_j)$ and

$B = \mathrm{Tri}(-\bar{w}_j, ib, w_j)$, where the entries are defined as in condition (3.2). Then by equations (2.2) and (4.2),

$$\|M\|_F^2 = \|A\|_F^2 + \|B\|_F^2 = a^2 n + 2\|\zeta\|_2^2 + b^2 n + 2\|\omega\|_2^2. \tag{4.3}$$

Clearly the norm of M is minimized by the choice of zero for the diagonals of A and B. Moreover, the relations (3.2) imply that $z_j = \frac{\lambda_1 + ic}{|w_1|^2} w_j$ and $|w_j|^2 = \frac{S_j}{\lambda_1}|w_1|^2$. Define

$$L = \sum_{j=1}^{n-1} S_j = \sum_{j=1}^{n-1}\sum_{l=1}^{j} \lambda_l.$$

To minimize $\|M\|_F^2$, we need to minimize

$$D_0(|w_1|^2) = 2\Big[\sum_{j=1}^{n-1}\Big(\frac{\lambda_1^2 + c^2}{|w_1|^4}\Big)\frac{S_j}{\lambda_1}|w_1|^2 + \frac{S_j}{\lambda_1}|w_1|^2\Big] = \frac{2L}{\lambda_1}\Big(\frac{\lambda_1^2 + c^2}{|w_1|^2} + |w_1|^2\Big). \tag{4.4}$$

This is minimized at $|w_1|^2 = \sqrt{\lambda_1^2 + c^2}$, and the minimal value is

$$D_0\big(\sqrt{\lambda_1^2 + c^2}\big) = \frac{4L\sqrt{\lambda_1^2 + c^2}}{\lambda_1}.$$

To summarize, let $M \in \mathcal{I}_\Lambda$ be defined by evaluating the formulas in Corollary 3.5 at $a = 0$, $b = 0$, $c \in \mathbb{R}$, $|w_1|^2 = \sqrt{\lambda_1^2 + c^2}$, and $(\theta_1, \cdots, \theta_{n-1}) \in \mathbb{T}^{n-1}$. Then

$$\|M\|_F^2 = \frac{L\sqrt{\lambda_1^2 + c^2}}{\lambda_1}.$$

This is, of course, minimized when $c = 0$. The choice $c = 0$ and, hence, $|w_1|^2 = \lambda_1$ imply that the factor $r = \frac{\lambda_1 + ic}{|w_1|^2}$ equals 1. Combined with Corollary 3.5, the above observations establish the following theorem.

THEOREM 4.1. *Let $\Lambda = \mathrm{Diag}(\lambda_l)$ satisfy condition (1.1), and define $S_j = \sum_{l=1}^{j} \lambda_l$ and $L = \sum_{j=1}^{n-1} S_j$. The minimal Frobenius norm of the elements of $\mathcal{I}_\Lambda$ is L. The subvariety of $\mathbb{C}^{n\times n}$ consisting of elements of $\mathcal{I}_\Lambda$ which have norm L is*

$$\mathcal{M}_\Lambda = \{M_\Lambda^\theta = \mathrm{Tri}(0, 0, e^{i\theta_j}\sqrt{S_j}) \;:\; (\theta_1, \cdots, \theta_{n-1}) \in \mathbb{T}^{n-1}\}.$$

In particular,

$$M_\Lambda = \mathrm{Tri}(0, 0, \sqrt{S_j})$$

is the unique element of $\mathcal{I}_\Lambda$ with $\|M_\Lambda\|_F^2 = L$, and nonnegative entries.

Recall that we measure the (squared Frobenius) distance from a matrix M to a subset $\mathcal{X}$ of $\mathbb{C}^{n\times n}$ by

$$d_F^2(M, \mathcal{X}) = \inf\{\|M - E\|_F^2 : E \in \mathcal{X}\}.$$

Given any fixed $T \in \mathcal{T}$, we want to find the distance from T to $\mathcal{I}_\Lambda$. Equivalently, (up to a factor of 4) we want to find the distance from an element $(P, Q) = (\Pi_\mathcal{S}(2T), \Pi_\mathcal{A}(2T))$ in $\mathcal{S}_\mathbf{T} \oplus \mathcal{A}_\mathbf{T}$ to the set $\mathcal{P}_\Lambda$. Write

$$P = \mathrm{Tri}(\bar{\mu}, \mathrm{Re}(\delta), \mu) \text{ and } Q = \mathrm{Tri}(-\bar{\nu}, i\mathrm{Im}(\delta), \nu).$$

Let $(A,B) \in \mathcal{P}_\Lambda$ with

$$A = \mathrm{Tri}(\bar{\zeta}, a, \zeta) \text{ and } B = \mathrm{Tri}(-\bar{\omega}, ib, \omega).$$

The distance between (P,Q) and (A,B) is

$$\|\,\mathrm{Tri}(\overline{\mu-\zeta}, \mathrm{Re}(\delta)-a, \mu-\zeta)\|_F^2 + \|\,\mathrm{Tri}(-(\overline{\nu-\omega}), \mathrm{Im}(\delta)-b, \nu-\omega)\|_F^2,$$

which, by equations (4.1) and (4.2), is

$$\|\mathrm{Re}(\delta)-a\|_2^2 + 2\|\mu-\zeta\|_2^2 + \|\mathrm{Im}(\delta)-b\|_2^2 + 2\|\nu-\omega\|_2^2.$$

The first and third term are minimized by the choices of constant n-tuples $a = \frac{\sum_{j=1}^n \mathrm{Re}(\delta_j)}{n}$ and $b = \frac{\sum_{j=1}^n \mathrm{Im}(\delta_j)}{n}$, cf., [5]. Thus, finding a closest element in $\mathcal{P}_\Lambda$ to (P,Q) reduces to minimizing

$$2\|\mu-\zeta\|_2^2 + 2\|\nu-\omega\|_2^2, \quad \text{where } \zeta,\omega \in \mathbb{C}^{n-1} \text{ satisfy (3.2).} \tag{4.5}$$

Let $\zeta,\omega \in \mathbb{C}^{n-1}$ satisfy (3.2) and let $c = \mathrm{Im}(z_1\bar{w}_1)$ and $r = \frac{\lambda_1 + ic}{|w_1|^2}$. Recall that $\zeta = r\omega$. Thus equation (4.5) is

$$2\|\mu\|_2^2 - 4\mathrm{Re}(<\mu, r\omega>) + 2|r|^2\|\omega\|_2^2 + 2\|\nu\|_2^2 - 4\mathrm{Re}(<\nu,\omega>) + 2\|\omega\|_2^2.$$

Now, $\|\omega\|_2^2 = \sum_{j=1}^{n-1} \frac{S_j}{\lambda_1}|w_1|^2 = \frac{L|w_1|^2}{\lambda_1}$, and $|r|^2 = \frac{\lambda_1^2 + c^2}{|w_1|^4}$. Therefore, equation (4.5) is

$$2\|\mu\|_2^2 + 2\|\nu\|_2^2 + (1+|r|^2)\frac{2L|w_1|^2}{\lambda_1} - 4\mathrm{Re}(<\bar{r}\mu + \nu, \omega>). \tag{4.6}$$

Let $\epsilon = \bar{r}\mu + \nu = (\epsilon_j)$ and recall that ω has the form $(w_j) = (\frac{|w_1|\sqrt{S_j}}{\sqrt{\lambda_1}} e^{i\theta_j})$, where we can choose $\theta_j = \mathrm{Arg}(\epsilon_j)$. That is, we fix θ_j so that $\mathrm{Re}(\epsilon_j e^{-i\theta_j})$ is maximal. Thus,

$$\mathrm{Re}(<\epsilon,\omega>) = \sum_{j=1}^{n-1} |\epsilon_j| \frac{|w_1|\sqrt{S_j}}{\sqrt{\lambda_1}} = \frac{|w_1|}{\sqrt{\lambda_1}} \sum_{j=1}^{n-1} |\bar{r}\mu_j + \nu_j|\sqrt{S_j}.$$

The above calculations lead us to define the function $D(c,|w_1|)$ from $\mathbb{R}\times\mathbb{R}^+$ to $\mathbb{R}^+$ by

$$D(c,|w_1|) = \frac{2L(|w_1|^4 + \lambda_1^2 + c^2)}{\lambda_1|w_1|^2} + \frac{-4}{|w_1|\sqrt{\lambda_1}} \sum_{j=1}^{n-1} \sqrt{S_j}|(\lambda_1 - ic)\mu_j + |w_1|^2\nu_j|. \tag{4.7}$$

As the following lemma indicates, finding the distance between a given pair (P,Q) in $\mathcal{S}_{\mathbf{T}} \oplus \mathcal{A}_{\mathbf{T}}$ and $\mathcal{P}_\Lambda$ is equivalent to minimizing the above function.

LEMMA 4.2. *Let $\Lambda = \mathrm{Diag}(\lambda_l)$ satisfy condition (1.1), and define $S_j = \sum_{l=1}^j \lambda_l$ and $L = \sum_{j=1}^{n-1} S_j$. Let $(P,Q) \in \mathcal{S}_{\mathbf{T}} \oplus \mathcal{A}_{\mathbf{T}}$ with*

$$P = \mathrm{Tri}(\bar{\mu}, \alpha, \mu) \text{ and } Q = \mathrm{Tri}(-\bar{\nu}, i\beta, \nu).$$

Define $D(c,|w_1|)$ as in (4.7) and let

$$D = \inf\{D(c,|w_1|) \ : \ (c,|w_1|) \in \mathbb{R}\times\mathbb{R}^+\}.$$

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

Let $\hat{a}$ and $\hat{b}$ be the constant n-tuples with entries $\sum_{j=1}^{n} \frac{\alpha_j}{n}$, and $\sum_{j=1}^{n} \frac{\beta_j}{n}$, respectively. Then, the distance from (P,Q) to $\mathcal{P}_\Lambda$ is

$$d_F^2((P,Q),\mathcal{P}_\Lambda) = \|\alpha-\hat{a}\|_2^2 + \|\beta-\hat{b}\|_2^2 + 2\|\mu\|_2^2 + 2\|\nu\|_2^2 + D. \tag{4.8}$$

Moreover, this distance is bounded above by

$$d_F^2((P,Q),\mathcal{P}_\Lambda) \leq \|\alpha-\hat{a}\|_2^2 + \|\beta-\hat{b}\|_2^2 + 2\|\mu\|_2^2 + 2\|\nu\|_2^2 + 4(L - \sum_{j=1}^{n-1} \sqrt{S_j}|\mu_j+\nu_j|).$$

Proof. Fix $(c,|w_1|)$ and define the variables $r = \frac{\lambda_1+ic}{|w_1|^2}$, $\epsilon_j = \bar{r}\mu_j+\nu_j$, and $\theta_j = \mathrm{Arg}(\epsilon_j)$, for all $j = 1,\cdots,n-1$. Let $(A,B) \in \mathcal{P}_\Lambda$ be given by the $(n+3)$-tuple $(a,b,c,|w_1|,(\theta_j))$. We saw in the discussion above, these choices of a b, and (θ_j) are optimal for this $(c,|w_1|)$, and $d_F^2((P,Q),(A,B))$ is given by (4.8). The upper bound is given by noting that the last term, $D = \inf\{D(c,|w_1|) \ : \ (c,|w_1|) \in \mathbb{R}\times\mathbb{R}^+\}$, in the distance formula is less than or equal to

$$D(0,\sqrt{\lambda_1}) = 4L - 4\sum_{j=1}^{n-1} \sqrt{S_j}|\mu_j+\nu_j|. \qquad \square$$

The difficulty of minimizing $D(c,|w_1|)$ over $\mathbb{R}\times\mathbb{R}^+$ depends, of course, on the specific matrices (P,Q). For example, if $(P,Q) = (\mathbf{0},\mathbf{0})$,

$$D(c,|w_1|) = D_0(|w_1|) = \frac{2L(|w_1|^4+\lambda_1^2+c^2)}{\lambda_1|w_1|^2}.$$

We have seen that this is minimized at $c=0$, $|w_1|^2 = \lambda_1$ and the minimum is $4L$.

The above results translate directly into distance formulas for $\mathcal{T}$.

COROLLARY 4.3. *Let $\Lambda = \mathrm{Diag}(\lambda_l)$ satisfy condition (1.1), and define $S_j = \sum_{l=1}^{j}\lambda_l$ and $L = \sum_{j=1}^{n-1} S_j$. Let $T = \mathrm{Tri}(\overline{\sigma},\gamma,\tau)$. Define $\hat{\gamma}$ to be the constant n-tuple with entry $\frac{\sum_{j=1}^{n}\gamma_j}{n} = \frac{\mathrm{Tr}(T)}{n}$, and define*

$$\delta(c,|w_1|) = \frac{L(|w_1|^4+\lambda_1^2+c^2)}{2\lambda_1|w_1|^2} - \sum_{j=1}^{n-1} \frac{|(\lambda_1-|w_1|^2-ic)\sigma_j + (\lambda_1+|w_1|^2-ic)\tau_j|\sqrt{S_j}}{|w_1|\sqrt{\lambda_1}}.$$

Let

$$\Delta = \inf\{\delta((c,|w_1|)) \ : \ (c,|w_1|) \in \mathbb{R}\times\mathbb{R}^+\}.$$

Then

$$d_F^2(T,\mathcal{I}_\Lambda) = \|\gamma-\hat{\gamma}\|_2^2 + \|\sigma\|_2^2 + \|\tau\|_2^2 + \Delta.$$

In particular, for any $T \in \mathcal{T}$ its distance from $\mathcal{I}_\Lambda$ is bounded by

$$d_F^2(T,\mathcal{I}_\Lambda) \leq \delta((0,\sqrt{\lambda_1})) = \|\gamma-\hat{\gamma}\|_2^2 + \|\sigma\|_2^2 + \|\tau\|_2^2 + L - \sum_{j=1}^{n-1} 2|\tau_j|\sqrt{S_j}.$$

Proof. $(\Pi_{\mathcal{S}}(2T), \Pi_{\mathcal{A}}(2T)) = (\mathrm{Tri}(\overline{\sigma+\tau}, 2\mathrm{Re}(\gamma), \sigma+\tau), \mathrm{Tri}(\overline{\sigma-\tau}, 2i\mathrm{Im}(\gamma), \tau-\sigma))$. If $(P, Q) = (\Pi_{\mathcal{S}}(2T), \Pi_{\mathcal{A}}(2T))$ in the previous lemma, then equation (4.8) is

$$\|2\gamma - 2\hat{\gamma}\|_2^2 + 2\|\sigma+\tau\|_2^2 + 2\|\tau-\sigma\|_2^2 + D = 4\|\gamma-\hat{\gamma}\|_2^2 + 4\|\sigma\|_2^2 + 4\|\sigma\|_2^2 + D,$$

and the function $D(c, |w_1|)$ reduces to $4\delta(c, |w_1|)$ for $\mu = \sigma + \tau$ and $\nu = \tau - \sigma$. This, combined with the fact that $d_F^2(T, \mathcal{I}_\Lambda) = \frac{1}{4} d_F^2((\Pi_{\mathcal{S}}(2T), \Pi_{\mathcal{A}}(2T)), \mathcal{P}_\Lambda))$, establishes the corollary. □

Let $T \in \mathcal{T}$. Define $T_0 = T - \frac{\mathrm{Tr}(T)}{n} I$. Then T_0 is the translation of T, by a multiple of the identity matrix, which has minimal Frobenius norm among all such translations, cf., [5]. The above bound for the squared Frobenius distance from T to $\mathcal{I}_\Lambda$ is less than or equal to (with equality holding if and only if $\tau = 0$) the sum of the squared Frobenius distances from T_0 to $\mathbf{0}$ and from $\mathbf{0}$ to $\mathcal{I}_\Lambda$. We will see the importance of the translate, T_0, in the next section.

5. Applications. In this section, we bound the Frobenius distance from normality for the elements of $\mathcal{I}_\Lambda$. We also apply results from [2] to indicate how well the set of eigenvalues of an element of $\mathcal{I}_\Lambda$ can be approximated by using normal matrices and Geršgorin-type sets.

First, let us consider what the above results tell us about the Frobenius distance from normality for the elements of $\mathcal{I}_\Lambda$. Let $T \in \mathbb{C}^{n\times n}$. The direct sum structure $\mathbb{C}^{n\times n} = \mathcal{S} \oplus \mathcal{A}$ and the Pythagorean relationship (2.2) imply that

$$d_F^2(T, \mathcal{N}) \le \min\{d_F^2(T, \mathcal{S}), d_F^2(T, \mathcal{A})\} \le \min\{\|\Pi_{\mathcal{A}}(T)\|_F^2, \|\Pi_{\mathcal{S}}(T)\|_F^2\}.$$

Recall that we defined

$$T_0 = T - \frac{\mathrm{Tr}(T)}{n} I.$$

Because the matrix $\frac{\mathrm{Tr}(T)}{n} I$ is scalar, $d_F^2(T, \mathcal{N}) = d_F^2(T_0, \mathcal{N})$, cf., [5, Theorem 3.2]. Therefore,

$$\forall T \in \mathbb{C}^{n\times n} \qquad d_F^2(T, \mathcal{N}) \le \min\{\|\Pi_{\mathcal{A}}(T_0)\|_F^2, \|\Pi_{\mathcal{S}}(T_0)\|_F^2\}.$$

Now let $M \in \mathcal{I}_\Lambda$. By Corollary 3.5,

$$M = \frac{1}{2}\,\mathrm{Tri}((\bar{r}-1)\,\overline{\omega}, a+ib, (r+1)\omega), \quad \text{and} \quad M_0 = \frac{1}{2}\,\mathrm{Tri}((\bar{r}-1)\,\overline{\omega}, 0, (r+1)\omega),$$

where $r = \frac{\lambda_1 + ic}{|w_1|^2}$ and $w_j = \frac{\sqrt{S_j}\,|w_1| e^{i\theta_j}}{\sqrt{\lambda_1}}$. Then $\Pi_{\mathcal{S}}(M_0) = \frac{1}{2}\,\mathrm{Tri}(\bar{r}\bar{\omega}, 0, r\omega)$ and $\Pi_{\mathcal{A}}(M_0) = \frac{1}{2}\,\mathrm{Tri}(-\bar{\omega}, 0, \omega)$. By equation (4.2), $\|\Pi_{\mathcal{S}}(M_0)\|_F^2 = 2\|\frac{r\omega}{2}\|_2^2 = \frac{(\lambda_1^2+c^2)L}{2\lambda_1|w_1|^2}$, and $\|\Pi_{\mathcal{A}}(M_0)\|_F^2 = 2\|\frac{\omega}{2}\|_2^2 = \frac{L|w_1|^2}{2\lambda_1}$. Thus, for $M \in \mathcal{I}_\Lambda$ defined by the $(n+3)$-tuple, $(a, b, c, |w_1|, (\theta_j)_{j=1}^{n-1})$, the distance from M to $\mathcal{N}$ satisfies

$$d_F^2(M, \mathcal{N}) \le \min\{\frac{(\lambda_1^2+c^2)L}{2\lambda_1|w_1|^2}, \frac{L|w_1|^2}{2\lambda_1}\}.$$

Graphically, this bound can be described as follows. Let $M \in \mathcal{I}_\Lambda$ be given by $(a, b, c, |w_1|, (\theta_j)_{j=1}^{n-1})$. The Frobenius distance from M to the set of normal matrices is determined by where the ordered pair $(c, |w_1|^4)$ lies in relation to the parabola $y = x^2 + \lambda_1^2$. Specifically,

$$d_F^2(M, \mathcal{N}) \le \begin{cases} \frac{(\lambda_1^2+c^2)L}{2\lambda_1|w_1|^2} & \text{if} \quad |w_1|^4 > c^2 + \lambda_1^2; \\[2ex] \frac{|w_1|^2 L}{2\lambda_1} = \frac{(\lambda_1^2+c^2)L}{2\lambda_1|w_1|^2} & \text{if} \quad |w_1|^4 = c^2 + \lambda_1^2; \\[2ex] \frac{|w_1|^2 L}{2\lambda_1} & \text{if} \quad |w_1|^4 < c^2 + \lambda_1^2. \end{cases}$$

Note that $\mathcal{M}_\Lambda$, the set of elements from $\mathcal{I}_\Lambda$ with minimal Frobenius norm, corresponds to the vertex of $y = x^2 + \lambda_1^2$. For every $M_\Lambda^\theta \in \mathcal{M}_\Lambda$, the above bound says $d_F^2(M_\Lambda^\theta, \mathcal{N}) \leq L$. This is, of course, consistent with Theorem 4.1 which tells us that the normal matrix **0** has distance L from M_Λ^θ.

We now use the results of [2] to describe the distance from normality, in the sense of the numerical stability of eigenvalue estimation through normal matrices, for the elements of $\mathcal{I}_\Lambda$. Recall that a singular value decomposition (SVD), of a non-zero complex $n \times n$ matrix B, is an expression of B as a product

$$B = V\Sigma W^* = \begin{bmatrix} | & \cdots & | \\ \phi_1 & \cdots & \phi_n \\ | & \cdots & | \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n \end{bmatrix} \begin{bmatrix} - & \bar{\psi}_1^T & - \\ \vdots & \vdots & \vdots \\ - & \bar{\psi}_n^T & - \end{bmatrix}, \tag{5.1}$$

where V and W are unitary matrices in $\mathbb{C}^{n\times n}$, and Σ is a nonnegative diagonal matrix. The entries of Σ, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$, are the eigenvalues of $|B|$ arranged in non-increasing order. They are called the *singular values* of B.

Fix any non-zero $B \in \mathbb{C}^{n\times n}$ and a SVD, $B = V\Sigma W^*$, as in (5.1). In [2], we defined the *SV-normally estimated Geršgorin set*, $\Gamma^{\mathrm{NSV}}(V\Sigma W^*)$. Like the Geršgorin set for B, the set $\Gamma^{\mathrm{NSV}}(V\Sigma W^*)$ is a union of closed discs and it contains the eigenvalues of B. We also defined the *SV-normal estimator* $\epsilon_{V\Sigma W^*}$ corresponding to this SVD of B. Specifically, define for each $l = 1, \cdots, n$,

$$\epsilon_l = \sqrt{1 - |<\phi_l, \psi_l>|^2} \quad \text{and let} \quad \epsilon_{V\Sigma W^*} = \max_{1\leq l\leq n} \{\epsilon_l\}.$$

The parameter $\epsilon_{V\Sigma W^*}$ lies between 0 and 1, inclusively. It is used as a type of condition number which indicates how well the set $\Gamma^{\mathrm{NSV}}(V\Sigma W^*)$ estimates the eigenvalues of B. When $\epsilon_{V\Sigma W^*}$ is zero, $\Gamma^{\mathrm{NSV}}(V\Sigma W^*)$ is exactly the set of eigenvalues of B; when $\epsilon_{V\Sigma W^*}$ is small, the centers of the discs which comprise $\Gamma^{\mathrm{NSV}}(V\Sigma W^*)$ provides a good estimate of the spectrum of B. This is because the radii of the discs which comprise $\Gamma^{\mathrm{NSV}}(V\Sigma W^*)$ are all $R = \sqrt{2\sum_l \sigma_l^2 \epsilon_l^2}$. Roughly speaking, up to a scaling factor of σ_1, this common radius will be small when $\epsilon_{V\Sigma W^*}$ is.

Finally, we cite the following lemma which bounds the SV-normal estimators from below. Notice that this lower bound on $\epsilon_{V\Sigma W^*}$ is *independent* of the choice of SVD for B.

LEMMA 5.1. *(See [2]) Let $B \in \mathbb{C}^{N\times N}$ and let $\epsilon_{V\Sigma W^*}$ denote the SV-normal estimator corresponding to a SVD, $B = V\Sigma W^*$. Then,*

$$\|B^*B - BB^*\|_2 \leq \|B\|_F^2 \, \epsilon_{V\Sigma W^*}.$$

The above lemma allows us to describe how well the spectrum of an element of $\mathcal{I}_\Lambda$ can be approximated with the SV-normally estimated Geršgorin set, $\Gamma^{\mathrm{NSV}}(V\Sigma W^*)$.

THEOREM 5.2. *Let $M \in \mathcal{I}_\Lambda$ be defined by the $(n+3)$-tuple $(a, b, c, |w_1|, (\theta_j)_{j=1}^{n-1})$, and recall that $L = \sum_{j=1}^{n-1}\sum_{l=1}^{j}\lambda_l$. Define*

$$H = \frac{n(a^2+b^2)}{4} + \frac{L}{2\lambda_1}\left(\frac{\lambda_1^2 + c^2}{|w_1|^2} + |w_1|^2\right).$$

Let $M = V\Sigma W^*$ be a SVD of M and denote the corresponding SV-normal estimator by $\epsilon_{V\Sigma W^*}$. Then

$$\frac{\|\Lambda\|_2}{H} \leq \epsilon_{V\Sigma W^*}.$$

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

In particular, if $M \in \mathcal{M}_\Lambda$ then

$$\frac{\|\Lambda\|_2}{L} \leq \epsilon_{V\Sigma W^*}.$$

Proof. Let

$$M = \frac{1}{2}\,\mathrm{Tri}((\bar{r}-1)\,\overline{\omega}, a+ib, (r+1)\omega),$$

be defined by $(a, b, c, |w_1|, (\theta_j)_{j=1}^{n-1})$. Since $M \in \mathcal{I}_\Lambda$, $M^*M - MM^* = \Lambda$ and we have $\|M^*M - MM^*\|_2 = \|\Lambda\|_2$. We showed in Section 4 that

$$\|M\|_F^2 = \frac{n(a^2+b^2)}{4} + \|\frac{(\bar{r}-1)\bar{\omega}}{2}\|_2^2 + \|\frac{(r+1)\omega}{2}\|_2^2.$$

Thus,

$$\|M\|_F^2 = \frac{n(a^2+b^2)}{4} + (\frac{|\bar{r}-1|^2+|r+1|^2}{4})\|\omega\|_2^2 = \frac{n(a^2+b^2)}{4} + \frac{(|r|^2+1)}{2}\|\omega\|_2^2.$$

Since $|r|^2+1 = \frac{\lambda_1^2+c^2}{|w_1|^4} + 1$ and $\|\omega\|_2^2 = \frac{L|w_1|^2}{\lambda_1}$,

$$\|M\|_F^2 = \frac{n(a^2+b^2)}{4} + \frac{L}{2\lambda_1}(\frac{\lambda_1^2+c^2}{|w_1|^2} + |w_1|^2) = H.$$

Finally, if $M \in \mathcal{M}_\Lambda$, by Theorem 4.1, $\|M\|_F^2 = L$. Thus,

$$\frac{\|M^*M - MM^*\|_2}{\|M\|_F^2} = \frac{\|\Lambda\|_2}{L} \leq \epsilon_{V\Sigma W^*},$$

in this case. □

The previous theorem has an interesting interpretation. The elements of $\mathcal{M}_\Lambda$ have minimal Frobenius norm. However, the square of the reciprocal of this Frobenius norm is a factor of our lower bound for $\epsilon_{V\Sigma W^*}$. Consequently, the condition number of the SV-normally estimated Geršgorin sets for elements of $\mathcal{M}_\Lambda$ is maximally bounded above 0. This seems to suggest the counter-intuitive idea that, regardless of which SVD is used, the radii of the set $\Gamma^{\mathrm{NSV}}(V\Sigma W^*)$ should tend to be largest for smallest elements of $\mathcal{I}_\Lambda$. However, this suggestion fails to consider how weighting factors σ_j^2 of the radius R increase with $\|M\|_F^2 = \sum_{j=1}^n \sigma_j^2$.

6. Conclusions and extensions. We conclude this note with a few remarks and some indications of further lines of inquiry for the sets $\mathcal{I}_\Lambda$. As we mentioned in Section 3, the right hand side of the matrix equation $M^*M - MM^* = \Lambda$ has to be self adjoint with trace zero, since the left hand side is. The choice to make Λ diagonal arose from a desire to simplify the calculations for $\mathcal{P}_\Lambda$ and $\mathcal{I}_\Lambda$ and to make the rank of Λ easy to identify, since this rank is a type of measure of the extent to which the matrix M fails to be normal. It would be interesting to consider how the above development changes for the general right hand side, Λ.

The intermediate set $\mathcal{P}_\Lambda$ was used to simplify the calculations for $\mathcal{I}_\Lambda$ and to help clarify how the upper and lower bands of the elements of $\mathcal{I}_\Lambda$ are related to each other. However, the set $\mathcal{P}_\Lambda$ has an interesting intrinsic functional analytic structure. Specifically, the decomposition $\mathbb{C}^{n\times n} = \mathcal{S} \oplus \mathcal{A}$ expresses $\mathbb{C}^{n\times n}$ as a *Kreĭn space*. This is an indefinite inner product space which has the structure of the direct sum of a Hilbert space and a negative Hilbert space. The structure of the operators on such spaces have been studied in detail, cf., [1, 4], and an interesting line of inquiry would be to examine the properties of the elements of $\mathcal{P}_\Lambda$ as Kreĭn space operators.

Finally, other SVD-based Geršgorin-type sets were developed in [6] and [7]. The difficulty in applying such methods generally arises from the nonuniqueness of SVDs. We see from the above development that elements of the algebraic varieties $\mathcal{M}_\Lambda$ and $\mathcal{I}_\Lambda$ have sufficiently strong structure constraints to overcome the difficulties created by the nonuniqueness of the SVD for normally estimated Geršgorin sets of [2]. An interesting question is whether this happens with other SVD Geršgorin-type sets.

REFERENCES

[1] J. BOGNAR, *Indefinite Inner Product Spaces*, Springer, Berlin, 1974.

[2] N. FONTES, J. KOVER, L. SMITHIES, AND R. S. VARGA, *Singular value decomposition normally estimated Geršgorin sets*, Electron. Trans. Numer. Anal., 26, (2007), pp. 320–329.
http://etna.math.kent.edu/vol.26.2007/pp320-329.pdf

[3] R. HORN AND C. JOHNSON, *Topics in Matrix Analysis*, 2nd ed., Cambridge University Press, Cambridge, England, 1994.

[4] H. LANGER, *Spectral functions of definitizable operators in Kreĭn spaces*, in Lecture Notes in Mathematics, vol. 948, Springer, Berlin, 1982, pp. 1–46.

[5] S. NOSCHESE, L. PASQUINI, AND L. REICHEL, *The structured distance to normality of an irreducible real tridiagonal matrix*, Electron. Trans. Numer. Anal., 28 (2007-08), pp. 65–77.
http://etna.math.kent.edu/vol.28.2007-2008/pp65-77.dir/pp65-77.pdf

[6] L. SMITHIES, *A note on polar decomposition based Geršgorin-type sets*, Linear Algebra Appl., 429 (2008), pp. 2623–2627.

[7] L. SMITHIES AND R. S. VARGA, *Singular value decomposition Geršgorin sets*, Linear Algebra Appl., 417 (2005), pp. 370–380.

[8] R. S. VARGA, Geršgorin and His Circles, Springer, Heidelberg, 2004.