ON THE BANDED TOEPLITZ STRUCTURED DISTANCE TO SYMMETRIC POSITIVE SEMIDEFINITENESS

SILVIA NOSCHESE* AND LOTHAR REICHEL†

Abstract. This paper is concerned with the determination of a close real banded positive definite Toeplitz matrix in the Frobenius norm to a given square real banded matrix. While it is straightforward to determine the closest banded Toeplitz matrix to a given square matrix, the additional requirement of positive definiteness makes the problem difficult. We review available theoretical results and provide a simple approach to determine a banded positive definite Toeplitz matrix.

Key words. Matrix nearness problem, Toeplitz structure, symmetric positive definite matrix, structured distance, banded Toeplitz matrix

1. Introduction. Matrix nearness problems are the focus of much research in linear algebra; see, e.g., [9, 10, 21, 23, 38]. In particular, characterizations of the algebraic variety of normal matrices and distance measures to this variety have received considerable attention [11, 12, 13, 15, 19, 26, 28, 30, 31, 39]. Normal Toeplitz matrices are characterized in [13, 20, 26], and the distance of Toeplitz matrices to the algebraic variety of normal Toeplitz matrices, measured with the Frobenius norm and referred to as the *Toeplitz structured distance to normality*, is investigated in [36], where also an application to preconditioning is described. The present paper is concerned with the distance of banded Toeplitz matrices to the variety of similarly structured positive semidefinite matrices, and with determining the closest matrix in this variety.

We denote banded Toeplitz matrices in $\mathbb{R}^{n \times n}$ with bandwidth $2k+1$ by

$$
T_{(k)} = (n; k; \sigma, \delta, \tau) = \begin{bmatrix}
\delta & \tau_1 & \tau_2 & \dots & \tau_k & & & & \text{\Large 0} \\
\sigma_1 & \delta & \tau_1 & & & & & & \\
\sigma_2 & \sigma_1 & \ddots & & & & \ddots & & \\
\vdots & & & \ddots & & & & \tau_k & \\
\sigma_k & & \ddots & \ddots & \ddots & & & \vdots & \\
& & & & & & \tau_1 & \tau_2 & \\
& & \ddots & & & \sigma_1 & \delta & \tau_1 & \\
\text{\Large 0} & & & \sigma_k & \dots & \sigma_2 & \sigma_1 & \delta
\end{bmatrix}. \tag{1.1}
$$

Some of the scalars σ_j, δ, and τ_j may vanish. We say that the matrix (1.1) has bandwidth $2k+1$, or equivalently, is $(2k+1)$-banded, even if σ_k or τ_k vanish. All Toeplitz matrices in this paper are real. Banded Toeplitz matrices arise in many applications including signal processing, time-series analysis, and numerical methods for the solution of differential equations; see, e.g., [5, 14, 18, 27, 43]. Low-rank modifications of symmetric banded Toeplitz matrices are considered in [32].

Necessary and sufficient conditions for a banded Toeplitz matrix (1.1) to be normal are given in [35], where also the distance of $(2k+1)$-banded Toeplitz matrices of

order n, with k less than or equal to the integer part of $n/2$, to the algebraic variety of normal Toeplitz matrices of the same bandwidth, is investigated. The distance is measured with the Frobenius norm. Since the given matrix (1.1) and the closest normal Toeplitz matrix both are banded Toeplitz matrices, we refer to their distance as the *banded Toeplitz structured distance to normality*. Whether this distance measure is more meaningful than the Toeplitz structured distance depends on the application.

Let $A = [a_{ij}]_{i,j=1}^n \in \mathbb{R}^{n \times n}$ and define the Frobenius norm

$$\|A\|_F = \sqrt{\sum_{i,j=1}^n a_{ij}^2}.$$

The distance from A to the set of symmetric positive semidefinite $n \times n$ matrices in the Frobenius norm is given by

$$\delta_F^+(A) := \min\{\|E\|_F : E \in \mathbb{R}^{n \times n}, A + E \text{ symmetric positive semidefinite}\}. \quad (1.2)$$

This distance can be expressed as

$$\delta_F^+(A) = \sqrt{\sum_{\lambda_i(B) < 0} \lambda_i(B)^2 + \|C\|_F^2}, \quad (1.3)$$

where B and C are the symmetric and skew-symmetric parts of A, respectively, and $\lambda_1(B), \lambda_2(B), \ldots, \lambda_n(B)$ are the eigenvalues of B. The nearest symmetric positive semidefinite matrix in the Frobenius norm is $A_+ := (B + H)/2$, with H the symmetric polar factor of B defined as follows: Consider the spectral factorization $B = Z \operatorname{diag}(\lambda_i(B)) Z^T$ with Z an orthogonal matrix. Then $H = Z \operatorname{diag}(|\lambda_i(B)|) Z^T$; see Higham [22, Theorem 2.1]. Unfortunately, neither the matrices $A + E$ nor A_+ generally have the same structure as A. In fact, the polar factor of a symmetric banded Toeplitz matrix typically is neither Toeplitz nor banded.

We are interested in determining the *banded Toeplitz structured distance to symmetric positive semidefiniteness*, Δ_F^+, as well as the projection, $T_{(k)}^+$, of a given banded Toeplitz matrix $T_{(k)}$ in the set of the similarly structured symmetric positive semidefinite matrices. The matrix $T_{(k)}^+$ has potential application to preconditioning; see Section 5. Our next example shows that the closest Toeplitz matrix in the Frobenius norm to a symmetric positive definite matrix might not be symmetric positive definite. Therefore, it is not straightforward to determine $\delta_F^+(A)$ and Δ_F^+ even in this special situation.

EXAMPLE 1. *Let*

$$A = \begin{bmatrix} 100 & 99 & 0 \\ 99 & 100 & 1/2 \\ 0 & 1/2 & 1 \end{bmatrix}.$$

The spectrum of A is approximately $\{0.6461, 1.3532, 199.0006\}$. Thus, A is symmetric positive definite. The closest Toeplitz matrix in the Frobenius norm to A is obtained by averaging the entries of A on every diagonal,

$$T = \begin{bmatrix} 67 & 49.75 & 0 \\ 49.75 & 67 & 49.75 \\ 0 & 49.75 & 67 \end{bmatrix},$$

2

and has the spectrum $\{-3.3571, \ 67.0000, \ 137.357\}$. *Hence, T is symmetric indefinite.*

This paper is organized as follows. Section 2 discusses the structured distance of banded Toeplitz matrices to normality in the Frobenius norm and Section 3 is concerned with the structured distance to the set of positive semidefinite Toeplitz matrices. Also an approach to inexpensively determine a banded symmetric positive definite matrix that is close to a given banded Toeplitz matrix is described. The special case of tridiagonal matrices is considered in Section 4. Some remarks on the application of the symmetric positive definite banded Toeplitz matrices determined in Sections 3 and 4 to the solution of linear systems of equations are provided in Section 5. Concluding remarks can be found in Section 6.

2. Structured distance to normality. This section reviews notation and results from [34, 35] that will be useful in the sequel.

THEOREM 2.1. *([35]) The real $(2k+1)$-banded Toeplitz matrix $T_{(k)} = (n; k; \sigma, \delta, \tau)$ of order n with $k \leq \lfloor n/2 \rfloor$ is normal if and only if it is either symmetric or shifted skew-symmetric (i.e., obtained by adding to a skew-symmetric matrix a multiple of the identity). Consider the sum*

$$\sum_{h=1}^{k} (n-h)\sigma_h \tau_h \qquad (2.1)$$

associated with the matrix $T_{(k)} = (n; k; \sigma, \delta, \tau)$. If this sum is positive, then the projection of $T_{(k)}$ onto the algebraic variety of similarly structured normal matrices is the real symmetric $(2k+1)$-banded Toeplitz matrix

$$T_{1,(k)}^* = (n; k; \frac{\sigma + \tau}{2}, \delta, \frac{\sigma + \tau}{2}).$$

If, instead, the sum (2.1) is negative, then the projection of $T_{(k)}$ onto the algebraic variety of similarly structured normal matrices is the real shifted skew-symmetric $(2k+1)$-banded Toeplitz matrix

$$T_{2,(k)}^* = (n; k; \frac{\sigma - \tau}{2}, \delta, \frac{\tau - \sigma}{2}).$$

Finally, if the sum (2.1) vanishes, then both matrices $T_{1,(k)}^*$ and $T_{2,(k)}^*$ are closest matrices to $T_{(k)}$ in the algebraic variety of similarly structured normal matrices. Moreover, the squared banded Toeplitz structured distance to normality from $T_{(k)}$ is

$$\Delta_F(T_{(k)})^2 = \frac{1}{2} \min \left\{ \sum_{j=1}^{k} (n-j)(\sigma_j - \tau_j)^2, \sum_{j=1}^{k} (n-j)(\sigma_j + \tau_j)^2 \right\}.$$

Note that the matrices $T_{(k)}$ and $T_{(k)} - \delta I_n = (n; k; \sigma, 0, \tau) \in \mathbb{R}^{n \times n}$ have the same banded Toeplitz structured distance to normality; however, they have different projections onto the algebraic variety of similarly structured normal matrices (at distance $\sqrt{n}|\delta|$).

Theorem 2.1 greatly simplifies in the tridiagonal case and the following result holds.

3

COROLLARY 2.2. ([34, Theorem 3.3]) The squared 3-banded Toeplitz structured distance to normality from $T_{(1)}$ is

$$\Delta_F(T_{(1)})^2 = \frac{n-1}{2} \min\left\{(\sigma_1 - \tau_1)^2, (\sigma_1 + \tau_1)^2\right\} = \frac{n-1}{2} \left\||\sigma_1| - |\tau_1|\right\|^2.$$

Moreover, the normal tridiagonal Toeplitz matrix closest to $T_{(1)}$ is the symmetric matrix $T^*_{1,(1)} = (n; 1; \frac{\sigma_1 + \tau_1}{2}, \delta, \frac{\sigma_1 + \tau_1}{2})$ if $\sigma_1 \tau_1 \geq 0$, and the shifted skew-symmetric matrix $T^*_{2,(1)} = (n; 1; \frac{\sigma_1 - \tau_1}{2}, \delta, \frac{\tau_1 - \sigma_1}{2})$ if $\sigma_1 \tau_1 \leq 0$.

3. Structured distance to symmetric positive semidefiniteness. It follows from (1.3) that the squared distance of a real $(2k+1)$-banded (possibly nonsymmetric) Toeplitz matrix $T_{(k)} = (n; k; \sigma, \delta, \tau)$ to the set of symmetric positive semidefinite matrices is

$$\delta_F^+(T_{(k)})^2 = \sum_{\lambda_i^{(k)} < 0} (\lambda_i^{(k)})^2 + \sum_{i=1}^{k} \frac{n-i}{2}(\sigma_i - \tau_i)^2, \tag{3.1}$$

where the $\lambda_i^{(k)}$ denote the eigenvalues of the closest symmetric matrix (in the Frobenius norm) to $T_{(k)}$. We note that the closest symmetric matrix to $T_{(k)}$ in the Frobenius norm is $T^*_{1,(k)} = (n; k; \frac{\sigma + \tau}{2}, \delta, \frac{\sigma + \tau}{2})$, but the latter matrix is not guaranteed to be positive definite. Moreover, $T^*_{2,(k)} - \delta I_n = (n; k; \frac{\sigma - \tau}{2}, 0, \frac{\tau - \sigma}{2})$ is the closest skew-symmetric matrix to $T_{(k)}$ (in the Frobenius norm) and

$$\|T^*_{2,(k)} - \delta I_n\|_F^2 = \sum_{i=1}^{k} \frac{n-i}{2}(\sigma_i - \tau_i)^2.$$

We are interested in determining the closest symmetric positive semidefinite $(2k+1)$-banded Toeplitz matrix $T^+_{(k)}$ to $T_{(k)}$, as well as the distance

$$\Delta_F^+(T_{(k)}) := \|T_{(k)} - T^+_{(k)}\|_F.$$

There is no simple expression available for $T^+_{(k)}$ even when $T_{(k)}$ is symmetric. We therefore seek to determine an approximation $\widetilde{T}^+_{(k)}$ of $T^+_{(k)}$, as well as the upper bound

$$\widetilde{\Delta}_F^+(T_{(k)}) := \|T_{(k)} - \widetilde{T}^+_{(k)}\|_F$$

for $\Delta_F^+(T_{(k)})$. A natural approach to determine a suitable approximation $\widetilde{T}^+_{(k)}$ of $T^+_{(k)}$ is to shift the matrix $T^*_{1,(k)}$ so that all eigenvalues of the shifted matrix are nonnegative and one of them is zero. The following well-known result (see, e.g., [18, 5]) provides an approach to choose the shift, which we denote by γ.

PROPOSITION 3.1. The set $\{\lambda_h^{(n)}\}_{h=1}^n$ of eigenvalues of the symmetric banded Toeplitz matrix $T_{(k)} = (n; k; \sigma, \delta, \sigma)$, ordered so that $\lambda_1^{(n)} \geq \lambda_2^{(n)} \geq \ldots \geq \lambda_n^{(n)}$, are distributed as $\{g(\frac{h\pi}{n+2})\}_{h=1}^n$, where g is the symbol for the matrix $T_{(k)}$, i.e.,

$$g(\theta) = \delta + 2\sum_{j=1}^{k} \sigma_j \cos(j\theta), \quad \theta \in (-\pi, \pi).$$

4

Moreover, if $g(\theta) \geq 0$, $\forall \theta \in (-\pi, \pi)$, then $T_{(k)}$ is positive semidefinite or positive definite.

REMARK 1. *Notice that the matrix $T_{(k)} = (n; k; \sigma, \delta, \sigma)$ is positive definite if its symbol $g(\theta)$ has only isolated zeros; see, e.g., [42].*

Let $T_{(k)}$ be a general banded Toeplitz matrix. One obtains a symmetric positive semidefinite $(2k+1)$-banded Toeplitz matrix by shifting the symmetric matrix $T^*_{1,(k)}$ by γI_n, where

$$\gamma = \max\{0, \sum_{j=1}^{k} |\sigma_j + \tau_j| - \delta\}. \tag{3.2}$$

We remark that an application of Gershgorin disks suggests the same shift.

The following result can be used to bound the distance between the spectra of $T^*_{1,(k)}$ and $\widetilde{T}^+_{(k)} := T^*_{1,(k)} + \gamma I_n$, with γ defined by (3.2).

PROPOSITION 3.2. *([3]) Let the matrices $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times n}$ be symmetric and measure the distance between the matrices A and B in the Frobenius norm,*

$$d_{A,B} := \|A - B\|_F.$$

Let the eigenvalues of the matrices A and B be ordered to be nonincreasing as a function of their index. Introduce the vectors

$$\lambda(A) = [\lambda_1(A), \lambda_2(A), \ldots, \lambda_n(A)]^T, \quad \lambda(B) = [\lambda_1(B), \lambda_2(B), \ldots, \lambda_n(B)]^T,$$

containing all eigenvalues $\lambda_j(A)$ of A and $\lambda_j(B)$ of B, respectively, and define the distance between these vectors by means of the Euclidean norm $\|\cdot\|$,

$$d_{\lambda(A,B)} := \|\lambda(A) - \lambda(B)\|.$$

Then

$$d_{A,B} \geq d_{\lambda(A,B)}. \tag{3.3}$$

In our context, we obtain equality in (3.3),

$$d_{T^*_{1,(k)}, \widetilde{T}^+_{(k)}} = \|T^*_{1,(k)} - \widetilde{T}^+_{(k)}\|_F = \sqrt{n}\gamma = \|\lambda(T^*_{1,(k)}) - \lambda(\widetilde{T}^+_{(k)})\| = d_{\lambda(T^*_{1,(k)}, \widetilde{T}^+_{(k)})}.$$

We define for future reference

$$\widetilde{\Delta}^+_F(T^*_{1,(k)}) = \|T^*_{1,(k)} - \widetilde{T}^+_{(k)}\|_F.$$

REMARK 2. *Notice that the matrix δI_n may be considered a $(2k+1)$-banded Toeplitz matrix for any k less than or equal to the integer part of $n/2$. It is symmetric positive semidefinite if and only if $\delta \geq 0$.*

Using (1.3), the (unstructured) nearness to symmetric positive semidefiniteness of $T^*_{1,(k)}$ can be bounded by

$$\delta^+_F(T^*_{1,(k)})^2 = \sum_{\lambda_i(T^*_{1,(k)}) < 0} \lambda_i(T^*_{1,(k)})^2$$

$$\leq \sum_{i=1}^{n} \lambda_i(T^*_{1,(k)})^2 = \|T^*_{1,(k)}\|_F^2 = \sum_{i=1}^{k} \frac{n-i}{2}(\sigma_i + \tau_i)^2 + n\delta^2,$$

5

where equality is attained when the spectrum of $T_{1,(k)}^*$ is confined to the negative real axis. In this case, the closest symmetric positive semidefinite matrix to $T_{1,(k)}^*$ is the zero matrix O_n, because

$$\max\{0,\delta\} = \max\left\{0,\sum_{i=1}^{n}\lambda_i(T_{1,(k)}^*)\right\} = 0.$$

Thus, for $\Delta_F^+(T_{1,(k)}^*)$, i.e., for the banded Toeplitz structured nearness to symmetric positive semidefiniteness of $T_{1,(k)}^*$, one obtains the lower and upper bounds

$$\delta_F^+(T_{1,(k)}^*) \leq \Delta_F^+(T_{1,(k)}^*) \leq \min\{\|T_{1,(k)}^* - \max\{0,\delta\}I_n\|_F, \widetilde{\Delta}_F^+(T_{1,(k)}^*)\}.$$

We conclude with the observation that, since $\|B+C\|_F^2 = \|B\|_F^2 + \|C\|_F^2$ if $B = B^T$ and $C = -C^T$, the above inequalities also hold for the banded Toeplitz structured distance of $T_{(k)}$ to the set of symmetric positive semidefinite matrices. We have

$$
\begin{aligned}
\delta_F^+(T_{(k)})^2 &\leq \Delta_F^+(T_{(k)})^2 \\
&\leq \min\{\|T_{1,(k)}^* - \max\{0,\delta\}I_n\|_F, \widetilde{\Delta}_F^+(T_{1,(k)}^*)\}^2 + \|T_{2,(k)}^* - \delta I_n\|^2,
\end{aligned}
$$
(3.4)

where equality is achieved if the spectrum of $T_{1,(k)}^*$ is confined to the negative real axis, so that $\delta_F^+(T_{(k)}) = \Delta_F^+(T_{(k)}) = \|T_{(k)}\|_F$.

THEOREM 3.3. *We have the following upper bounds for the squared* $(2k+1)$-*banded Toeplitz structured distance to symmetric positive semidefiniteness of* $T_{(k)} = (n;k;\sigma,\delta,\tau)$:

$$\min\left\{\sum_{i=1}^{k}(n-i)(\sigma_i^2+\tau_i^2), n\max\left\{0,\sum_{i=1}^{k}|\sigma_i+\tau_i|-\delta\right\}^2 + \sum_{i=1}^{k}\frac{n-i}{2}(\sigma_i-\tau_i)^2\right\}$$

if $\delta > 0$, *and*

$$\min\left\{\sum_{i=1}^{k}(n-i)(\sigma_i^2+\tau_i^2)+n\delta^2, n\left(\sum_{i=1}^{k}|\sigma_i+\tau_i|-\delta\right)^2 + \sum_{i=1}^{k}\frac{n-i}{2}(\sigma_i-\tau_i)^2\right\}$$

if $\delta \leq 0$. *These bounds can be computed in* $\mathcal{O}(k)$ *arithmetic floating point operations (flops).*

Proof. The bounds follow from the upper bound in (3.4), by replacing $\widetilde{\Delta}_F^+(T_{1,(k)}^*)$ by $\sqrt{n}\gamma$, with γ given by (3.2), and by observing that

$$\sum_{i=1}^{k}\frac{n-i}{2}(\sigma_i+\tau_i)^2 + \sum_{i=1}^{k}\frac{n-i}{2}(\sigma_i-\tau_i)^2 = \sum_{i=1}^{k}(n-i)(\sigma_i^2+\tau_i^2).$$

Moreover, it is straightforward to observe that the cost of computing these bounds increases linearly with the bandwidth of the $(2k+1)$-banded Toeplitz matrix $T_{(k)}$. This concludes the proof. $\square$

EXAMPLE 2. *Consider the symmetric pentadiagonal Toeplitz matrices* $T_{(2)}(p) = (100;2;\sigma,\delta,\tau)$, *with entries* $\sigma_1 = \tau_1 = 0.05$, $\sigma_2 = \tau_2 = p$, *and* $\delta = 0.1$, *where* p *ranges from* 0.04 *to* 0.06 *with step* 0.0001. *Figure 3.1 shows for each* p *the squared distances* $d_1 = \sum_{i=1}^{2}\frac{n-i}{2}(\sigma_i+\tau_i)^2$ *and* $d_2 = n\max\{0,\sum_{i=1}^{2}|\sigma_i+\tau_i|-\delta\}^2$ *in red and green,*
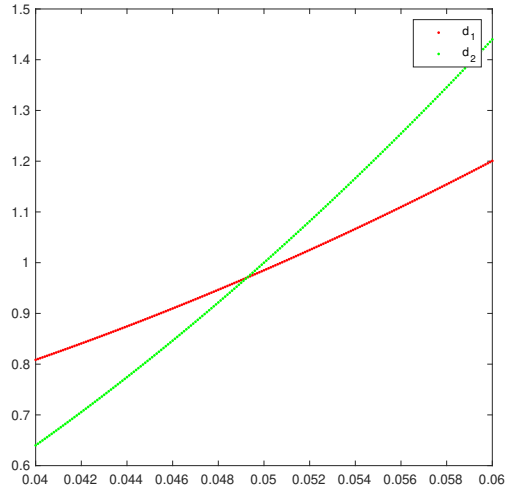
FIG. 3.1. *Example 2. Upper bound for the 5-banded Toeplitz structured nearness to symmetric positive semidefiniteness of $T_{(2)}(p) = (100; 2; [0.05, p], 0.1, [0.05, p])$, with $p = 0.04 : 0.0001 : 0.06$.*

respectively. The squared distance $d_3 = \sum_{i=1}^{2} \frac{n-i}{2}(\sigma_i - \tau_i)^2$ always vanishes, since the matrices considered are symmetric.

It is easy to verify that the upper bounds in (3.4) for $\Delta_F^+(T_{(2)}(p))^2$ are attained by $d_2 = \min\{d_1, d_2\} + d_3$ for p ranging from 0.04 to 0.0492 and by $d_1 = \min\{d_1, d_2\} + d_3$ for p ranging from 0.0493 to 0.06.

4. Structured distance to symmetric positive semidefiniteness in the tridiagonal case. This section considers real tridiagonal Toeplitz matrices. The eigenvalues of such matrices are known to be

$$\lambda_i(T_{(1)}) = \delta + 2\sqrt{\sigma_1 \tau_1} \, \cos \frac{i\pi}{n+1}, \quad i = 1, 2, \ldots, n.$$

It is therefore possible to construct the symmetric tridiagonal semidefinite Toeplitz matrix $\widetilde{T}_{(1)}^+$ by adding the matrix $\gamma_n I_n$, where $\gamma_n \geq 0$ depends on the order n, to the symmetric part of $T_{(1)}$, i.e., to $T_{1,(1)}^*$, to obtain $\widetilde{T}_{(1)}^+$.

4.1. The case $\delta = 0$. This subsection considers real tridiagonal Toeplitz matrices with vanishing diagonal entries. We may assume that the matrix $T_{(1)} = (n; 1; \sigma_1, 0, \tau_1)$ is of odd order n. The eigenvalues of $T_{1,(1)}^*$ then are given by

$$\lambda_i^{(1)} = |\sigma_1 + \tau_1| \, \cos \frac{i\pi}{n+1}, \quad i = 1, 2, \ldots, n.$$

Thus, they are real and allocated symmetrically with respect to the origin, one of them vanishes; see, e.g., [34]. If, instead, $T_{(1)}$ is of even order, then no eigenvalue of $T_{1,(1)}^*$ vanishes.

The closest symmetric positive semidefinite (not necessarily tridiagonal Toeplitz) matrix computed by the algorithm in [22] has the same $(n-1)/2$ [or $n/2$, if n is even]

7

positive eigenvalues as $T_{1,(1)}^*$, whereas the remaining eigenvalues vanish. A matrix with such a spectrum cannot be a tridiagonal Toeplitz matrix; see, e.g., [34].

PROPOSITION 4.1. *The distance of the symmetric part $T_{1,(1)}^*$ of $T_{(1)}$ to symmetric positive semidefiniteness is*

$$\delta_F^+(T_{1,(1)}^*) = \frac{\sqrt{n-1}}{2}|\sigma_1 + \tau_1|.$$

Proof. According to (3.1), one has

$$\delta_F^+(T_{1,(1)}^*)^2 = \sum_{\lambda_i^{(1)} < 0} (\lambda_i^{(1)})^2.$$

The eigenvalues of $T_{1,(1)}^*$ are allocated symmetrically with respect to the origin. Therefore,

$$\delta_F^+(T_{1,(1)}^*)^2 = \frac{1}{2}\|T_{1,(1)}^*\|_F^2 = \frac{n-1}{4}(\sigma_1 + \tau_1)^2.$$

This concludes the proof. $\square$

We are in a position to determine upper and lower bounds for $\Delta_F^+(T_{1,(1)}^*)$.

COROLLARY 4.2.

$$\frac{\sqrt{n-1}}{2}|\sigma_1 + \tau_1| \leq \Delta_F^+(T_{1,(1)}^*) \leq \sqrt{\frac{n-1}{2}}|\sigma_1 + \tau_1| \tag{4.1}$$

for all $n = 1, 2, \ldots$.

Proof. Theorem 3.3 applied to $T_{1,(1)}^* = (n; 1; \frac{\sigma_1 + \tau_1}{2}, 0, \frac{\sigma_1 + \tau_1}{2})$ and Proposition 4.1 yield the lower and upper bounds

$$\frac{\sqrt{n-1}}{2}|\sigma_1 + \tau_1| \leq \Delta_F^+(T_{1,(1)}^*) \leq \min\left\{\sqrt{\frac{n-1}{2}}|\sigma_1 + \tau_1|, \sqrt{n}|\sigma_1 + \tau_1|\right\},$$

from which (4.1) straightforwardly follows. $\square$

Alternatively, one might consider shifting $T_{1,(1)}^*$ by a multiple of the identity so that all eigenvalues become nonnegative. Since the eigenvalues of $T_{1,(1)}^*$ are allocated symmetrically with respect to the origin, this means that we could add a multiple of the identity, $\gamma_n I_n$, to $T_{1,(1)}^*$, with γ_n equal to the spectral radius

$$\rho(T_{1,(1)}^*) = |\sigma_1 + \tau_1|\cos\frac{\pi}{n+1}.$$

Notice that $\gamma_n = 0$ if and only if $\sigma_1 = -\tau_1$, i.e., $T_{1,(1)}^* = O_n$ and $T_{2,(1)}^* = T_{(1)}$. This way, one would get

$$\widetilde{\Delta}_F^+(T_{1,(1)}^*) = \sqrt{n}|\sigma_1 + \tau_1|\cos\frac{\pi}{n+1}.$$

However, it is easy to show that

$$\sqrt{\frac{n-1}{2}}|\sigma_1 + \tau_1| \leq \sqrt{n}|\sigma_1 + \tau_1|\cos\frac{\pi}{n+1}, \tag{4.2}$$

8

for all $n = 1, 2, \dots$. Hence, the upper bound in (4.1) is sharper. Indeed, direct computations show equality in (4.2) for $n \in \{1, 2\}$ and, for $n \geq 3$, one has

$$\frac{1}{2} \leq \cos^2\left(\frac{\pi}{n+1}\right).$$

We next show lower and upper bounds for the 3-banded Toeplitz structured distance of $T_{(1)} = (n; 1; \sigma_1, 0, \tau_1)$ to symmetric positive semidefiniteness in the Frobenius norm.

THEOREM 4.3. *For the squared 3-banded Toeplitz structured distance to symmetric positive semidefiniteness of the matrix $T_{(1)} = (n; 1; \sigma_1, 0, \tau_1)$, we have the lower and upper bounds*

$$\frac{n-1}{4}(3\,\sigma_1^2 + 3\,\tau_1^2 - 2\,\sigma_1\tau_1) \leq \Delta_F^+(T_{(1)})^2 \leq (n-1)(\sigma_1^2 + \tau_1^2). \tag{4.3}$$

Proof. The upper bound follows from

$$\Delta_F^+(T_{(1)})^2 \leq \|T_{(1)}\|_F^2 = (n-1)(\sigma_1^2 + \tau_1^2).$$

The lower bound is obtained from

$$\Delta_F^+(T_{(1)})^2 \geq \delta_F^+(T_{(1)})^2 = \delta_F^+(T_{1,(1)}^*)^2 + \|T_{2,(1)}^*\|_F^2,$$

where the equality is a consequence of (1.3). The first term on the right-hand side is given by Proposition 4.1 and the second term is evaluated in a straightforward manner to give the lower bound (4.3). $\square$

EXAMPLE 3. *Consider the downshift matrix*

$$T_{(1)} = (n; 1; 1, 0, 0) = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & 0 & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 0 & \cdots & \cdots & 0 & 1 & 0 \end{bmatrix}. \tag{4.4}$$

Its squared distances in the Frobenius norm to the sets of the symmetric and skew-symmetric (banded Toeplitz) matrices are both $\frac{n-1}{2}$. Thus, the squared banded Toeplitz structured distance to normality is $\frac{n-1}{2}$. It is shown in [33, Section 9] that the squared (non-banded) Toeplitz structured distance to normality is $\frac{n-1}{n}$, and a circulant being at this distance is described. Moreover, it is shown in [16, Proposition 2.1] that the squared distance to the set of the symmetric positive semidefinite matrices in the Frobenius norm is $\delta_F^+(T_{(1)})^2 = \frac{3(n-1)}{4}$. These results are consistent with Theorem 4.3.

Applying our approach to constructing an approximate nearest symmetric tridiagonal positive semidefinite matrix $\widetilde{T}_{(1)}^+$ to $T_{(1)}$, we obtain

$$\widetilde{T}_{(1)}^+ = (n; 1; \frac{1}{2}, \cos\frac{\pi}{n+1}, \frac{1}{2}) = \begin{bmatrix} \cos\frac{\pi}{n+1} & \frac{1}{2} & \cdots & \cdots & 0 \\ \frac{1}{2} & \cos\frac{\pi}{n+1} & \frac{1}{2} & \cdots & 0 \\ 0 & \frac{1}{2} & \cdots & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \frac{1}{2} \\ 0 & \cdots & \cdots & \frac{1}{2} & \cos\frac{\pi}{n+1} \end{bmatrix}. \tag{4.5}$$

9

and

$$\widetilde{\Delta}_F^+(T_{(1)})^2 = \|T_{(1)} - \widetilde{T}_{(1)}^+\|^2 = \|T_{1,(1)}^* - \widetilde{T}_{(1)}^+\|_F^2 + \|T_{2,(1)}^*\|_F^2 = n\cos^2\frac{\pi}{n+1} + \frac{n-1}{2}.$$

Moreover, the squared distance between the spectrum of $T_{1,(1)}^*$ and the spectrum of $\widetilde{T}_{(1)}^+$ is $\|\lambda^{(1)} - \widetilde{\lambda}^+\|^2 = n\cos^2\frac{\pi}{n+1}$, whereas the squared distance between the spectrum of $T_{1,(1)}^*$ and the spectrum of the (not necessarily tridiagonal Toeplitz) closest symmetric positive semidefinite matrix is $\|\lambda^{(1)} - \lambda^+\|^2 = \frac{n-1}{4}$. Here $\lambda^{(1)}$ is the vector of all eigenvalues of $T_{1,(1)}^*$ ordered nonincreasingly, $\widetilde{\lambda}^+$ denotes the vector of eigenvalues of $\widetilde{T}_{(1)}^+$ ordered in the same manner, and λ^+ is a vector of all eigenvalues of the closest symmetric positive semidefinite matrix ordered similarly; $\|\cdot\|$ denotes the Euclidean norm.

Finally, regard the approximate nearest symmetric positive semidefinite tridiagonal Toeplitz matrix $O_n \in \mathbb{R}^{n\times n}$. The squared distance from $T_{(1)}$ to O_n is $\|T_{(1)}\|_F^2 = n-1$, whereas the squared distance between the spectrum of $T_{1,(1)}^*$ and the spectrum of O_n is $\|\lambda^{(1)}\|^2 = \frac{n-1}{2}$. We obtain

$$\frac{3(n-1)}{4} = \delta_F^+(T_{(1)})^2 \le \Delta_F^+(T_{(1)})^2 \le \|T_{(1)}\|_F^2 = n-1.$$

4.2. The case $\delta \neq 0$. We consider general tridiagonal Toeplitz matrices $T_{(1)} = (n; 1; \sigma_1, \delta, \tau_1)$. The distance δ_F^+ to symmetric positive semidefiniteness depends on the eigenvalues of $T_{1,(1)}^*$, which are given by

$$\lambda_i^{(1)} = \delta + |\sigma_1 + \tau_1| \cos\frac{i\pi}{n+1}, \qquad i = 1, 2, \ldots, n.$$

When $\delta < |\sigma_1 + \tau_1| \cos\frac{\pi}{n+1}$, the closest symmetric positive semidefinite matrix computed as in [22] has the same nonnegative eigenvalues as $T_{1,(1)}^*$ and the remaining eigenvalues are zero. If, instead, $\delta \ge |\sigma_1 + \tau_1| \cos\frac{\pi}{n+1}$, then $T_{1,(1)}^*$ is symmetric positive semidefinite and the following argument is not needed.

Assume that the matrix $T_{1,(1)}^*$ is not symmetric positive semidefinite. Then we can add a multiple of the identity, $\gamma_n I_n$, to $T_{1,(1)}^*$ with

$$\gamma_n = |\delta - |\sigma_1 + \tau_1| \cos\frac{\pi}{n+1}| \tag{4.6}$$

so that the matrix $\widetilde{T}_{(1)}^+ = T_{1,(1)}^* + \gamma_n I_n$ is positive semidefinite. We have

$$\widetilde{T}_{(1)}^+ = (n; 1; (\sigma_1 + \tau_1)/2, |\sigma_1 + \tau_1| \cos\frac{\pi}{n+1}, (\sigma_1 + \tau_1)/2).$$

Alternatively, one may consider $\max\{0, \delta\}I_n$ as an approximate nearest symmetric positive semidefinite tridiagonal Toeplitz matrix.

THEOREM 4.4. *For the squared 3-banded Toeplitz structured distance to symmetric positive semidefiniteness of the matrix $T_{(1)} = (n; 1; \sigma_1, \delta, \tau_1)$, we have the upper bounds*

$$\min\left\{(n-1)(\sigma_1^2 + \tau_1^2), n\max\left\{0, |\sigma_1 + \tau_1|\cos\frac{\pi}{n+1} - \delta\right\}^2 + \frac{n-1}{2}(\sigma_1 - \tau_1)^2\right\}$$
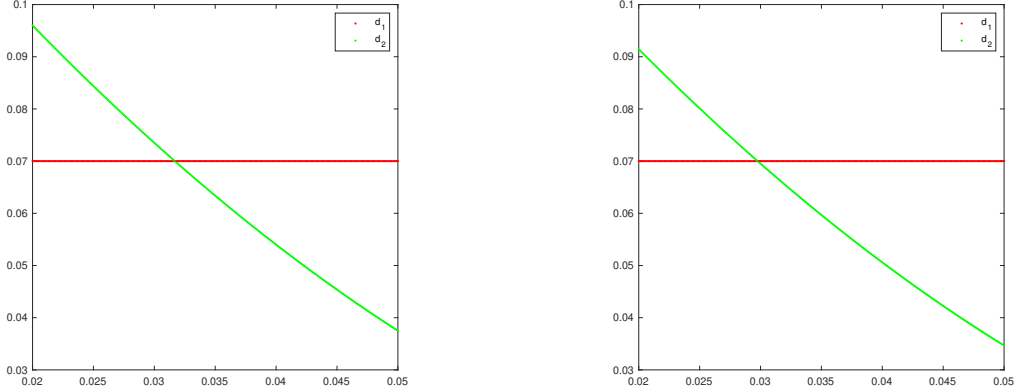
FIG. 4.1. *Example 4. Upper bounds for the tridiagonal Toeplitz structured distance to symmetric positive semidefiniteness of $T_{(1)}(p) = (15; 1; 0.05, p, 0.05)$ for $p = 0.02 : 0.0001 : 0.05$. In the left graph the upper bounds are determined by Theorem 3.3, whereas in the right graph they are determined by Theorem 4.4.*

if $\delta > 0$, and

$$\min\left\{ (n-1)(\sigma_1^2 + \tau_1^2) + n\delta^2, n\left(|\sigma_1 + \tau_1|\cos\frac{\pi}{n+1} - \delta \right)^2 + \frac{n-1}{2}(\sigma_1 - \tau_1)^2 \right\}$$

if $\delta \leq 0$. The cost of computing these upper bounds is $\mathcal{O}(1)$ flops.

Proof. The bounds follow from (3.4), where $\widetilde{\Delta}_F^+(T_{1,(k)}^*)$ is given by the shift γ_n in (4.6) (instead of the shift γ in (3.2), as in Theorem 3.3). □

EXAMPLE 4. *Consider the symmetric tridiagonal Toeplitz matrices $T_{(1)}(p) = (15; 1; \sigma_1, \delta, \tau_1)$ with $\sigma_1 = \tau_1 = 0.05$ and $\delta = p$, where p ranges from 0.02 to 0.05 with step 0.0001. The left-hand side graphs of Figure 4.1 show, for each p, the squared distances*

$$d_1 = \frac{n-1}{2}(\sigma_1 + \tau_1)^2, \quad d_2 = n\max\{0, |\sigma_1 + \tau_1| - \delta\}^2$$

that come from Theorem 3.3, in red and green, respectively. Since the squared distance $d_3 = \frac{n-1}{2}(\sigma_1 - \tau_1)^2$ vanishes, both d_1 and d_2 provide upper bounds for the squared 3-banded Toeplitz structured distance to symmetric positive semidefiniteness. The graphs on the right-hand side of Figure 4.1 show, for each p, the squared distances

$$d_1 = \frac{n-1}{2}(\sigma_1 + \tau_1)^2, \quad d_2 = n\max\{0, |\sigma_1 + \tau_1|\cos\frac{\pi}{n+1} - \delta\}^2$$

that come from Theorem 4.4, in red and green, respectively. These distances provide upper bounds for the squared 3-banded Toeplitz structured distance to symmetric positive semidefiniteness. It is easy to verify that they are sharper. For instance, for $p = 0.03$, one has $d_1 = 0.0700$ and, in the former case, $d_2 = 0.0735$ (left graph), whereas in the latter case $d_2 = 0.0695$ (right graph).

5. Remarks on applications to the solution of linear systems of equations. Consider the solution of a linear system of equations

$$Ax = b, \qquad A \in \mathbb{R}^{n \times n}, \quad x, b \in \mathbb{R}^n, \tag{5.1}$$

11

with a large matrix A. This matrix is not required to have any particular structure, but we assume that A is close to a banded symmetric positive definite Toeplitz matrix T^+. It is natural to use the matrix T^+ as a preconditioner, because linear systems of equations with a banded symmetric positive definite Toeplitz matrix can be solved rapidly and stably by exploiting the Toeplitz structure by Schur or generalized Schur algorithms described in [1, 2, 8, 27, 29][1], as well as by the method by Bini and Meini [4].

When the matrix A is symmetric positive definite, we can solve (5.1) by the preconditioned conjugate gradient method using T^+ as a preconditioner; see, e.g., [17, Algorithm 11.5.1]. If A is symmetric indefinite, then the conjugate gradient method should be replaced by the SYMMLQ algorithm; see [37] for a description of the latter. Finally, when A is nonsymmetric, a preconditioned iterative method designed for the solution of systems of equations with a nonsymmetric matrix should be used, such as preconditioned GMRES; see [17, 40]. In all these situations, the preconditioned iterative methods are simpler when using a symmetric positive definite preconditioner, because Schur and generalized Schur algorithms can be applied to rapidly solve linear systems of equations with such a preconditioner matrix. Though, the application of an indefinite or nonsymmetric preconditioner for the solution of Toeplitz systems also has been described in the literature; see [7, 25, 41].

Large banded matrices that are close to the set of banded symmetric positive definite Toeplitz matrices arise when discretizing second order differential equations in one space dimension on the interval $0 < t < 1$ at equidistant grid points using the standard symmetric second order 3-point finite difference approximation of the second derivative $-d^2/dt^2$ with some boundary conditions. For instance, Neumann boundary conditions give rise to a symmetric tridiagonal matrix of the form

$$\frac{1}{h^2}(T_{(1)} - e_1 e_1^T - e_n e_n^T), \tag{5.2}$$

where $e_j = [0, \ldots, 0, 1, 0, \ldots, 0]^T \in \mathbb{R}^n$ denotes the jth canonical basis vector, $T_{(1)} = (n, 1; -1, 2, -1)$, and $h = 1/(n+1)$. The matrix $T_{(1)}$ is symmetric positive definite, while the matrix (5.2) is singular. Consistent linear systems of equations with the latter matrix have a unique solution $x = [x_1, x_2, \ldots, x_n]^T$ such that $\sum_{j=1}^n x_j = 0$. The matrix $T_{(1)}$ can be used as a preconditioner. Analogous formulas arise in higher space dimensions.

Other techniques for determining positive definite banded Toeplitz preconditioners with applications to the solution of partial differential equations are described by Chan [6] and Hon et al. [24], who apply the Remez algorithm to compute a nonnegative low-degree trigonometric polynomial that approximates the symbol associated with a given symmetric indefinite Toeplitz matrix. This can be fairly expensive. The approach described in the present paper is much cheaper, but may give a banded Toeplitz matrix of lower quality. A careful comparison of these approaches to construct preconditioners is a topic of future work.

6. Conclusion. This paper discusses the determination of a positive definite banded Toeplitz matrix that is close to a Toeplitz matrix with the same band structure. A simple fast method is described.

---

[1]The superfast generalized Schur algorithms described in [1, 2] do not exploit bandedness.

REFERENCES

[1] G. S. Ammar and W. B. Gragg, *The generalized Schur algorithm for the superfast solution of Toeplitz systems*, in Rational Approximation and its Applications in Mathematics and Physics, J. Gilewicz, M. Pindor, and W. Siemaszko, eds., Lecture Notes in Mathematics # 1237, Springer, Berlin, 1987, pp. 315–330.

[2] G. S. Ammar and W. B. Gragg, *Superfast solution of real positive definite Toeplitz systems*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 61–76.

[3] R. Bhatia, *The distance between the eigenvalues of Hermitian matrices*, Proc. Amer. Math. Soc., 96 (1986), pp. 41–42.

[4] D. A. Bini and B. Meini, *Effective methods for solving banded Toeplitz systems*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 700–719.

[5] A. Böttcher and S. Grudsky, *Spectral Properties of Banded Toeplitz Matrices*, SIAM, Philadelphia, 2005.

[6] R. H. Chan, *Toeplitz preconditioners for Toeplitz systems with nonnegative generating functions*, IMA J. Numer. Anal., 11 (1991), pp. 333–345.

[7] R. Chan, D. Potts, and G. Steidl, *Preconditioners for nondefinite Hermitian Toeplitz systems*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 647–665.

[8] Ph. Delsarte and Y. Genin, *On the splitting of classical algorithms in linear prediction theory*, IEEE Trans. Acoust., Speech, Signal Processing, ASSP-35 (1987), pp. 645–653.

[9] J. W. Demmel, *On condition numbers and the distance to the nearest ill-posed problem*, Numer. Math., 51 (1987), pp. 251–289.

[10] J. W. Demmel, *Nearest defective matrices and the geometry of ill-conditioning*, in Reliable Numerical Computation, M. G. Cox and S. Hammarling, eds., Clarendon Press, Oxford, 1990, pp. 35–55.

[11] L. Elsner and Kh. D. Ikramov, *Normal matrices: an update*, Linear Algebra Appl., 285 (1998), pp. 291–303.

[12] L. Elsner and M. H. C. Paardekooper, *On measures of nonnormality of matrices*, Linear Algebra Appl., 92 (1987), pp. 107–124.

[13] D. R. Farenick, M. Krupnik, N. Krupnik, and W. Y. Lee, *Normal Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 1037–1043.

[14] D. Fischer, G. Golub, O. Hald, C. Leiva, and O. Widlund. *On Fourier-Toeplitz methods for separable elliptic problems*, Math. Comp., 28 (1974), pp. 349–368.

[15] S. Friedland, *Normal matrices and the completion problem*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 896–902.

[16] S. Gazzola, S. Noschese, P. Novati, and L. Reichel, *Arnoldi decomposition, GMRES, and preconditioning for linear discrete ill-posed problems*, Appl. Numer. Math., 142 (2019), pp. 102–121.

[17] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed, Johns Hopkins University Press, Baltimore, 2013.

[18] U. Grenander and G. Szegő, *Toeplitz Forms and Their Applications*, Chelsea, New York, 1984.

[19] R. Grone, C. R. Johnson, E. M. Sa, and H. Wolkowicz, *Normal matrices*, Linear Algebra Appl., 87 (1987), pp. 213–225.

[20] C. Gu and L. Patton, *Commutation relations for Toeplitz and Hankel matrices*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 728–746.

[21] P. Henrici, *Bounds for iterates, inverses, spectral variation and field of values of non-normal matrices*, Numer. Math., 4 (1962), pp. 24–40.

[22] N. J. Higham, *Computing a nearest symmetric positive semidefinite matrix*, Linear Algebra Appl., 103 (1988), pp. 103–118.

[23] N. J. Higham, *Matrix nearness problems and applications*, in Applications of Matrix Theory, M. J. C. Gover and S. Barnett, eds., Oxford University Press, Oxford, 1989, pp. 1–27.

[24] S. Hon, S. Serra-Capizzano, and A. Wathen, *Band-Toeplitz preconditioners for ill-conditioned Toeplitz systems*, BIT Numer. Math., 62 (2022), pp. 465–491.

[25] T. Huckle, S. Serra-Capizzano, and C. Tablino-Possio, *Preconditioning strategies for non-Hermitian Toeplitz linear systems*, Numer. Linear Algebra Appl., 12 (2005), pp. 211–220.

[26] T. Ito, *Every normal Toeplitz matrix is either of type I or of type II*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 998–1006.

[27] T. Kailath (ed.), *Modern Signal Processing*, Hemisphere Publishing, Washington, 1985.

[28] L. László, *An attainable lower bound for the best normal approximation*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1035–1043.

[29] T. Laudadio, N. Mastronardi, and P. Van Dooren, *The generalized Schur algorithm and*

some applications, Axiom, 2018, 7, 81.

[30] S. L. LEE, *A practical upper bound for departure from normality*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 462–468.

[31] S. L. LEE, *Best available bounds for departure from normality*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 984–991.

[32] A. LUATI AND T. PROIETTI, *On the spectral properties of matrices associated with trend filters*, 2008, MPRA Paper No. 11502, `http://mpra.ub.uni-muenchen.de/11502`

[33] S. NOSCHESE, L. PASQUINI, AND L. REICHEL, *The structured distance to normality of an irreducible real tridiagonal matrix*, Electron. Trans. Numer. Anal., 28 (2007), pp. 65–77.

[34] S. NOSCHESE, L. PASQUINI, AND L. REICHEL, *Tridiagonal Toeplitz matrices: properties and novel applications*, Numer. Linear Algebra Appl., 20 (2013), pp. 302–326.

[35] S. NOSCHESE AND L. REICHEL, *The structured distance to normality of banded Toeplitz matrices*, BIT Numer. Math., 49 (2009), pp. 629–640.

[36] S. NOSCHESE AND L. REICHEL, *The structured distance to normality of Toeplitz matrices with application to preconditioning*, Numer. Linear Algebra Appl., 18 (2011), pp. 429–447.

[37] C. C. PAIGE AND M. A. SAUNDERS, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.

[38] A. RUHE, *On the closeness of eigenvalues and singular values for almost normal matrices*, Linear Algebra Appl., 11 (1975), pp. 87–94.

[39] A. RUHE, *Closest normal matrix finally found!*, BIT Numer. Math., 27 (1987), pp. 585–598.

[40] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.

[41] S. SERRA-CAPIZZANO, *Preconditioning strategies for Hermitian Toeplitz systems with nondefinite generating functions*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 1007–1019.

[42] S. SERRA-CAPIZZANO, *On the extreme eigenvalues of Hermitian (block) Toeplitz matrices*, Linear Algebra Appl., 270 (1998), pp. 109–129.

[43] L. N. TREFETHEN AND M. EMBREE, *Spectra and Pseudospectra*, Princeton University Press, Princeton, 2005.