Normal Approximants for Certain Toeplitz Matrices and Toeplitz Operators

by

Robert Norman Harrison

B Sc , University of Winnipeg, 1972

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

We accept this thesis as conforming
to the required standard

[redacted]

Dr J Phillips, Supervisor (Department of Mathematics and Statistics)

[redacted]

Dr I F Putnam, Member (Department of Mathematics and Statistics)

[redacted]

Dr A R Sourour, Member (Department of Mathematics and Statistics)

[redacted]

Dr D D Olesky, External Examiner (Department of Computer Science)

Abstract

In this thesis, we consider the normal approximation problem for certain Toeplitz matrices and Toeplitz operators, with respect to the operator norm. For upper triangular Toeplitz matrices, we exhibit a natural normal approximant, and show that it is a best normal approximant in certain fundamental cases. In other cases, we analyze its effectiveness as a distance estimate by comparing it to the standard Hermitian approximant. For Toeplitz operators induced by certain continuous functions on the unit circle, we obtain upper and lower bounds on the distance to the set of normal operators, in terms of the image of the unit circle. In particular, if the continuous function is one-to-one and the image is a circle, then the distance to the normals is the radius of the image circle.

Examiners

[redacted]

Dr J. Phillips, Supervisor (Department of Mathematics and Statistics)

[redacted]

Dr I.F. Putnam, Member (Department of Mathematics and Statistics)

[redacted]

Dr A.R. Sourour, Member (Department of Mathematics and Statistics)

[redacted]

Dr D.D. Olesky, External Examiner (Department of Computer Science)

Contents

List of Tables

Acknowledgements

I would like to thank Dr. John Phillips, who has been a dedicated and knowledgeable supervisor. In every phase of my thesis, he provided extensive encouragement, patience and direction, thus enabling me to complete my work and to enjoy the process.

To the committee members, Dr. Ian Putnam and Dr. Ahmed Sourour, I would like to extend my thanks for their discussions throughout the course of my work and especially for their time and energy in reading my thesis.

I am grateful to Dr. Dale Olesky, the external examiner, for his thorough reading in the final stages of my thesis and his subsequent recommendations.

Also, my thanks to Dr. Bob Miers, for his guidance in commencing this project, together with his encouragement, on and off the court.

Finally, my deepest appreciation goes to my wife, Judy, whose blend of encouragement, humour, patience and total commitment were invaluable to me during this process.

Dedication

I proudly dedicate this thesis
to my parents
I feel their unconditional support
in everything I do

Introduction

The intention of this thesis is to discuss the problem of approximating Toeplitz operators and Toeplitz matrices by normal operators and normal matrices. The general setting for operator approximation problems is the algebra of operators (bounded linear transformations) on a fixed Hilbert space H, which corresponds to the algebra of $n \times n$ complex matrices when H has finite dimension n.

The discussion is aimed at a reader who is familiar with the basic facts about Hilbert space and the basic facts about the algebra of operators on a Hilbert space. We do not assume the reader is familiar with Toeplitz operators or Toeplitz matrices, however, we do assume a standard first year graduate course in real and complex analysis which covers measure theory, point set topology, elementary functional analysis and basic theory of L^p spaces.

Operator Approximation Problems

The general type of operator approximation problem consists of approximating an arbitrary operator by an operator in some given class of operators—the idea being to find a nearest approximant in the given class, with respect to the operator norm. Of course, finding a nearest approximant, always involves finding the distance to the given class as a subproblem. The motivation for this type of problem is to get information about arbitrary operators via approximation by operators from well-understood classes. Not surprisingly, the classes of operators that have attracted the most interest as approximants are: Hermitians, positives, normals, unitaries, projections and compacts.

In 1955, Fan and Hoffman [FH55] obtained the first operator approximation

results by showing that a best Hermitian approximant to a matrix A is its Hermitian part $(A + A^*)/2$, while a best unitary approximant to A is a unitary matrix U that occurs in the polar decomposition $A = UP$. In 1972, van Riemsdijk [vR72] proved that the Hermitian approximation of Fan and Hoffman can be extended to infinite dimensions. In the same year, Halmos [Hal72] studied and solved the more intricate[1] problem of best approximation by positive operators—renewing interest in operator approximation problems. This included the problem of best approximation by normal operators.

The Normal Approximation Problem

In contrast to Hermitian approximation and positive approximation, the problem of normal approximation seems less tractable and little is known. For example, in the operator norm, it is not even known what a closest normal matrix to an arbitrary 3×3 matrix is. Holmes [Hol74] suggests that the reason for the difficulty is that the normal operators lack any readily apparent geometric structure. The normals are closed in the space of all operators, but they are not convex (in fact they are nowhere dense) and hence usual approximation techniques do not apply. Moreover, Rogers [Rog76] has shown that not every operator has a nearest normal approximant when H is infinite dimensional—thus motivating interest in finding special classes of operators that do have best normal approximants.

Phillips [Phi77] showed that the class of binormal operators (which has the space of all 2×2 matrices as a special case) does have best normal approximants. In fact, he developed a formula for constructing a nearest normal approximant to a given binormal operator.

Instead of trying to find a known class that has best normal approximants, Holmes [Hol74] took a different approach. He defined the class of antinormal op-

[1] A best positive approximant for a normal operator is its positive part. However, in general, the positive part of an operator is not a best positive approximant for it.

erators as all those nonzero operators for which the zero operator is a best normal approximant, and then proceeded to study properties of this special class.

Little else is known about special classes of operators that have best normal approximants. Halmos [Hal74] studied the *spectral approximation problem* of approximating arbitrary normal operators by normal operators with spectrum contained in a given non-empty subset of the complex plane. More recently, Bhatia, Horn and Kittaneh [BHK91] have shown that Phillips' result for binormal operators is actually valid with respect to any unitarily invariant norm. Very recently, Lin [Lin94] has solved an old and important problem in linear algebra and operator theory concerning when *almost normal* implies *near to normal*. More precisely, for each $\varepsilon > 0$, Lin has proven that there exists a $\delta > 0$, such that, if T is any (finite) square matrix satisfying $\|T\| \leq 1$ and $\|T^*T - TT^*\| < \delta$, then there is a normal matrix N such that $\|T - N\| < \varepsilon$.

Remark. Lin's existence theorem does not give quantitative information about δ as a function of ε. In particular, it does not give the range of $\delta = \delta(\varepsilon)$ for $\varepsilon \leq 1$, and hence, it cannot be used directly as a normal approximation result about arbitrary matrices. For example, if $\|T\| \leq 1$ and $\|T^*T - TT^*\| = d$, Lin's Theorem does not imply that there is an $\varepsilon \leq 1$ such that $\delta(\varepsilon) > d$, and hence, it does not yield useful information about the distance from T to the normals.

The Main Purpose of This Thesis

The main purpose of this thesis is to make a modest beginning at including Toeplitz operators and finite dimensional Toeplitz matrices as classes of operators for which we can find nearest normal approximants.

Why consider Toeplitz operators when looking for classes that have best normal approximants? The simplest Toeplitz operator is the unilateral (forward) shift operator—a motivating example for the Brown-Douglas-Fillmore theorem—it is es-

sentially normal, Fredholm, but not normal Moreover, it has been very thoroughly analyzed In particular, the normal approximation problem has been solved for the unilateral shift—it is at distance 1 from the normals and the zero operator is a nearest normal approximant Important subclasses of the Toeplitz operators consist of operators that are essentially normal, Fredholm, and not normal and hence seem to be promising classes to consider

Why consider Toeplitz matrices? A Toeplitz matrix can be thought of as a compression of a Toeplitz operator, and hence the class of Toeplitz matrices seem like a promising finite dimensional class to consider on its own merit Moreover, studying a finite dimensional case is usually a natural place to start understanding a problem

Toeplitz Operators

If φ is a bounded measureable function on the unit circle $\mathbf{T}$, with respect to normalized Lebesgue measure, then φ is in the Hilbert space $L^2(\mathbf{T})$, and multiplication by φ is an operator on $L^2(\mathbf{T})$ called the *Laurent operator* induced by φ, denoted M_φ These multiplication operators are the prototypes of normal operators

For each integer n, the function $e_n \quad \mathbf{T} \to \mathbf{T}$ defined by $e_n(z) = z^n$ is continuous and $\{ e_n \quad n = 0, \pm 1, \pm 2, \quad \}$ is the standard orthonormal basis for $L^2(\mathbf{T})$ The Hardy space, $H^2(\mathbf{T})$, is the closed span of $\{ e_n \quad n = 0, 1, 2, \quad \}$ If P is the orthogonal projection from $L^2(\mathbf{T})$ onto $H^2(\mathbf{T})$ then the *Toeplitz operator* induced by φ, denoted T_φ, is the compression of M_φ to $H^2(\mathbf{T})$

$$T_\varphi f = P M_\varphi f = P(\varphi \quad f)$$

for every f in $H^2(\mathbf{T})$

The unilaterally infinite matrix of T_φ with respect to the orthonormal basis $\{ e_n \quad n = 0, 1, 2, \quad \}$ has a distinctive form—it is constant on all diagonals that are

parallel to the main diagonal

$$T_\varphi = \begin{bmatrix} c_0 & c_{-1} & c_{-2} & \\ c_1 & c_0 & c_{-1} & \\ c_2 & c_1 & c_0 & \\ & & & \end{bmatrix}$$

Moreover, the constants are intimately related to φ. For each integer n, c_n is the nth Fourier coefficient of φ.

The unilateral (forward) shift operator mentioned above is the Toeplitz operator $U = T_{e_1}$. It is called the forward shift operator since $Ue_n = e_{n+1}$ for every nonegative integer n. Note that the matrix of U has 1's on the first subdiagonal and 0's everywhere else.

Important subclasses of the Toeplitz operators correspond to important subclasses of the bounded measureable functions on $\mathbf{T}$: continuous, analytic, one-to-one, etc.

Finite Dimensional Toeplitz Matrices

We will begin our study of the normal approximation problem with finite dimensional Toeplitz matrices. A Toeplitz matrix is defined to be a square finite dimensional matrix that is constant along diagonals parallel to the main diagonal. Clearly, each n-dimensional Toeplitz matrix can be thought of as the compression of a Toeplitz operator on $H^2(\mathbf{T})$ to the subspace spanned by $\{ e_0, e_1, \quad , e_{n-1} \}$. It may also be possible to think of them as Toeplitz-like operators using the compact abelian group $\mathbf{Z}/(2n-1)$ in place of the compact abelian group $\mathbf{T}$.

As a first attempt at finding nearest normals to Toeplitz matrices, we will restrict

our attention to upper triangular Toeplitz matrices

$$T = \begin{bmatrix} a_0 & a_1 & & & a_{n-1} \\ & a_0 & a_1 & & a_{n-2} \\ & & & & \\ & 0 & & & a_1 \\ & & & & a_0 \end{bmatrix}$$

The Organization of This Thesis

In Chapter 1, we establish basic definitions related to the problem of normal approximation and develop some of their immediate consequences. We include some maps which can be used to simplify the normal approximation problem. We also present a detailed survey of the known normal approximation results of Holmes [Hol74], Rogers [Rog76] and Phillips [Phi77] which have been mentioned in this introduction. In particular, we analyze the derivation of the formula obtained by Phillips (for a nearest normal to a binormal operator) as it applies to 2×2 matrices. As a bonus, we develop a characterization of 2×2 normal matrices—in a way suggested by Phillips' nearest normal approximant.

In Chapter 2, the problem of normal approximants for (finite dimensional) upper triangular Toeplitz matrices is studied. The upper triangular Toeplitz matrices with exactly one nonzero diagonal play a distinguished role in our analysis—we call such matrices *superdiagonal* Toeplitz matrices. We develop a formula for getting a nearest normal matrix to an $n \times n$ superdiagonal Toeplitz matrix. For fixed n, it is interesting that all these normal approximants are in the commutative C^*-algebra generated by a special unitary matrix, and hence their sum is a normal matrix. Since each upper triangular Toeplitz matrix, T, can be written as a sum of superdiagonal ones, N = the sum of the individual approximants is a *natural* normal approximant to consider.

In Section 2.2, we verify that N is a best normal approximant in the 2×2 case.

Enticingly, it is also a best normal approximant for Jordan blocks (which are special upper triangular Toeplitz matrices) In fact, it is a best normal approximant for any scalar translation of any superdiagonal matrix However, we show via a 3×3 counterexample that N is not a best normal approximant in the general case In the non-superdiagonal 3×3 case, we carry out a comparative analysis of N versus Holmes' approximant (the Hermitian part of T) and find certain cases where N does improve on Holmes' upper distance estimate In an attempt to find better normal aprroximants, we develop a simple characterization of 3×3 normal matrices which are constant on the main diagonal—these include 3×3 normal Toeplitz matrices Leaving the 3×3 case, we show that for every $k \geq 2$, there are special classes of $2k \times 2k$ upper triangular Toeplitz matrices which admit N as a best normal approximant

As a final finite dimensional topic, we consider direct sums of upper triangular Toeplitz matrices in Section 2 3 Our main result implies that, if A is a direct sum, and each summand is either a Toeplitz matrix for which we have found a best normal approximant in Section 2 2, or a normal matrix, or any 2×2 upper triangular matrix, then any direct sum of corresponding best normal approximants will be a best normal approximant for A In the course of developing this result, we also give examples where we find a best normal approximant for a direct sum, even though we do not know all of the individual best normal approximants

In Chapter 3, the problem of normal approximants for Toeplitz operators is studied The main results in this chapter give upper and lower bounds on the distance from certain Toeplitz operators to the set of normal operators In particular, if φ is a continuous complex-valued function on the unit circle $\mathbf{T}$, then the distance from T_φ to the normals is less than or equal to the radius of the smallest disk containing $\varphi(\mathbf{T})$ and if φ is also one-to-one (so that $\varphi(\mathbf{T})$ is a Jordan curve), then the distance from T_φ to the normals is greater than or equal to the radius of the largest disk contained *inside* $\varphi(\mathbf{T})$ It follows that, if φ is continuous on $\mathbf{T}$ and one-to-one and

$\varphi(\mathbf{T})$ is a circle of radius a, then the distance from T_φ to the normals is equal to a. When applied to the unilateral shift, this result is in agreement with the known fact, that its distance from the normals is 1. As a final result for the thesis, we present a simple example (due to A. R. Sourour), which demonstrates that, if A is a direct sum of arbitrary Toeplitz operators, then a direct sum of corresponding best normal approximants is not neccessarily a best normal approximant for A.

Chapter 1

Definitions and Preliminary Results

In this chapter, we establish basic definitions related to the problem of normal approximation and develop some of their immediate consequences. We also elaborate on the known normal approximation results mentioned in the introduction. In particular, we analyze the derivation of the formula obtained by Phillips [Phi77] for a nearest normal matrix to an arbitrary 2×2 matrix.

1.1 Basic Facts and Notation

We begin by reviewing basic facts and setting basic notation for discussing a Hilbert space and the algebra of operators on it. As in the introduction, we use the term *operator* to mean bounded linear transformation. The symbols $\mathbf{N}$, $\mathbf{Z}$, $\mathbf{R}$, $\mathbf{R}^+$ and $\mathbf{C}$ are used to denote, respectively, the sets of positive integers, integers, real numbers, nonnegative real numbers and complex numbers.

Suppose H is a complex separable Hilbert space with inner product $\langle\ ,\ \rangle$ and norm $\|\ \|$. Let $\mathcal{B}(H)$ denote the algebra of operators on H with the usual operator norm. When H has finite dimension n, we identify H with the Hilbert space, $\mathbf{C}^n$, of $n \times 1$ complex column vectors and we identify $\mathcal{B}(H)$ with the algebra, M_n, of $n \times n$ complex matrices.

For each operator A in $\mathcal{B}(H)$, denote the operator norm of A by $\|A\|$, the spectrum of A by $\sigma(A)$ and the spectral radius of A by $r(A)$, so that

$$\begin{aligned} \|A\| &= \sup\{\, \|Ax\| \ \ x \in H \text{ and } \|x\| = 1 \,\} \\ \sigma(A) &= \{\, \lambda \in \mathbf{C} \ \ \lambda I - A \text{ is not invertible in } \mathcal{B}(H) \,\} \end{aligned}$$

$$r(A) = \sup_{\lambda \in \sigma(A)} |\lambda|$$

$\mathcal{B}(H)$ equipped with the operator norm is a unital Banach algebra—i e it is a complete metric space in the induced metric, $\|I\| = 1$ where I denotes the identity operator on H and $\|AB\| \leq \|A\| \, \|B\|$ for all A, B in $\mathcal{B}(H)$ Hence, we have the usual Banach algebra formula for the spectral radius of A

$$r(A) = \lim_{n \to \infty} \|A^n\|^{1/n}$$

For each operator A in $\mathcal{B}(H)$, its adjoint A^*, is the unique operator satisfying $\langle Ax, y \rangle = \langle x, A^* y \rangle$ for all x, y in H In the case where $A = [a_{jk}]$ is an $n \times n$ matrix, A^* is just the conjugate transpose of A (i e $(A^*)_{jk} = \overline{a}_{kj}$)

The map $A \mapsto A^*$ is a conjugate-linear map on $\mathcal{B}(H)$ with $(A^*)^* = A$, $(AB)^* = B^* A^*$ and $\|A^*\| = \|A\|$ for all A, B in $\mathcal{B}(H)$, hence, $\mathcal{B}(H)$ is a Banach $*$-algebra In addition, each operator A also satisfies the important *C^*-property* $\|A^* A\| = \|A\|^2$, hence, $\mathcal{B}(H)$ is a C^*-algebra

1 1 1 Hermitian Operators

An operator A is called *Hermitian* or *self-adjoint* if $A^* = A$ Equivalently, A is Hermitian if and only if $\langle Ax, x \rangle \in \mathbf{R}$ for all x in H Moreover, if A is Hermitian then $\sigma(A) \subseteq \mathbf{R}$

Two important classes of Hermitian operators are the projections and the positive operators

An operator P is *idempotent* if $P^2 = P$ and it is a *projection* if it is both Hermitian and idempotent (i e $P = P^* = P^2$)

An operator B is *positive* if $\langle Bx, x \rangle \geq 0$ for all x in H Equivalently, B is positive if and only if B is Hermitian and $\sigma(B) \subseteq \mathbf{R}^+$ Note that for every $A \in \mathcal{B}(H)$, the operator $A^* A$ is positive since $\langle A^* Ax, x \rangle = \langle Ax, Ax \rangle = \|Ax\|^2 \geq 0$ for all x in H

One of the important properties of Hermitian operators is that the norm of a Hermitian operator is equal to its spectral radius This property is a simple

consequence of the C^*-property and the spectral radius formula—if A is Hermitian, then $\|A^2\| = \|AA\| = \|A^*A\| = \|A\|^2$ and then, by induction, $\|A^{2^n}\| = \|A\|^{2^n}$ for all positive integers n It follows that

$$r(A) = \lim_{n\to\infty} \|A^n\|^{1/n} = \lim_{n\to\infty} \|A^{2^n}\|^{1/2^n} = \lim_{n\to\infty} (\|A\|^{2^n})^{1/2^n} = \|A\|$$

This property of Hermitian operators in conjunction with the C^*-property, yields a way to express the norm of an arbitrary operator A in terms of the spectral radius of the Hermitian operator A^*A

$$\|A\| = \|A^*A\|^{1/2} = (r(A^*A))^{1/2} \tag{1 1}$$

Remark Since formula 1 1 only depends on the C^*-property and the spectral radius formula, it follows that the operator norm is the only norm on $\mathcal{B}(H)$ that makes it into a C^*-algebra

Algebraically, the adjoint operation on $\mathcal{B}(H)$ is similar to the conjugation operation on the complex numbers and the Hermitian operators play a role in $\mathcal{B}(H)$ similar to the role played by the real numbers in the complex numbers Every operator A has a unique representation, $A = B + iC$, where B and C are Hermitian operators In fact $B = \frac{1}{2}(A + A^*)$ is called the real part of the operator A and $C = \frac{1}{2i}(A - A^*)$ is called the imaginary part of the operator A According to Halmos [Hal57, page 42]

> The fact that in general the real and imaginary parts of an operator fail to commute is what makes operator theory significantly harder than the corresponding theory of complex numbers and motivates the definition of a normal operator as one for which this pathology does not occur

1.1.2 Normal Operators

By definition, an operator A is *normal* if $A^*A = AA^*$ In other words, an operator is normal if and only if it commutes with its adjoint Using this definition, it is easy

to verify Halmos's description of normal operators as those operators for which the real and imaginary parts commute—if $A = B + iC$ where B and C are Hermitian, then $A^*A - AA^* = (B - iC)(B + iC) - (B + iC)(B - iC) = 2i(BC - CB)$ and therefore $A^*A = AA^*$ if and only if $BC = CB$

Obviously, every Hermitian operator is normal Other well-known classes of normal operators are the scalar operators, the unitary operators, the skew-Hermitian operators and the diagonalizable operators (which include the scalar operators)

The *scalar operators* are $\{ \lambda I \mid \lambda \in \mathbf{C} \}$

An operator U is *unitary* if $U^*U = UU^* = I$ Clearly, every unitary operator is both normal and invertible with $U^{-1} = U^*$ Moreover, the set of unitary operators is a subgroup of the (multiplicative) group of invertible operators In fact, the set of unitary operators is exactly the set of automorphisms of H In the analogy between operators and complex numbers, the unitaries correspond to the complex numbers of modulus 1

An operator S is *skew-Hermitian* if $S^* = -S$ In this case, the real part of S is 0 and hence skew-Hermitian operators are analogous to purely imaginary complex numbers If S is skew-Hermitian, then $S^*S = -S^2 = SS^*$ and hence S is normal

An operator D is called *diagonalizable* if H has an orthonormal basis $\{ f_j \}$, where every f_j is an eigenvector for D, in this case, if $Df_j = \lambda_j f_j$ for all j, then the matrix of D with respect to the basis $\{ f_j \}$ is the diagonal matrix with $D_{jj} = \lambda_j$ for all j and the matrix of the operator D^* with respect to this basis is the diagonal matrix with $(D^*)_{jj} = \overline{\lambda}_j$ for all j With D as above, $(D^*D - DD^*)f_j = |\lambda_j|^2 f_j - |\lambda_j|^2 f_j = 0$ for all j which implies $D^*D - DD^* = 0$ and hence D is a normal operator

The following proposition gives an interesting characterization of normal operators that will be used later

Proposition 1. *If A is an operator in $\mathcal{B}(H)$, then A is normal if and only if $\|Ax\| = \|A^*x\|$ for every x in H*

Proof Since $\|Ax\|^2 = \langle Ax, Ax\rangle = \langle A^*Ax, x\rangle$ for every x in H and, similarly, $\|A^*x\|^2 = \langle A^*x, A^*x\rangle = \langle AA^*x, x\rangle$ for every x in H, the forward implication is very obvious

$$\begin{aligned} A \text{ normal} &\Leftrightarrow A^*A = AA^* \\ &\Rightarrow \langle A^*Ax, x\rangle = \langle AA^*x, x\rangle \text{ for every } x \text{ in } H \\ &\Leftrightarrow \|Ax\| = \|A^*x\| \text{ for every } x \text{ in } H \end{aligned}$$

The reverse implication follows immediately from the basic, but not so obvious fact[1] *If S and T are operators in $\mathcal{B}(H)$ and $\langle Sx, x\rangle = \langle Tx, x\rangle$ for every x in H, then $S = T$* ∎

The potential of this simple characterization of normal operators is suggested by the following remarks of Halmos [Hal57, page 43]

> One source of the importance of the concept of normality is that many facts about Hermitian operators do not depend on the identity $Ax = A^*x$ but only on the identity $\|Ax\| = \|A^*x\|$ all such facts are valid for normal operators

For example, we have pointed out the nontrivial fact that the norm of a Hermitian operator is equal to its spectral radius Proposition 1 can be used to show that the same is true for normal operators—if A is a normal operator, then for all x in H, $\|A^2x\| = \|A(Ax)\| = \|A^*(Ax)\| = \|A^*Ax\|$ By taking supremums over $\|x\| = 1$, we get $\|A^2\| = \|A^*A\| = \|A\|^2$, just as in the Hermitian operator case! The rest of the proof is the same as the proof given for Hermitian operators in Section 1.1.1

Another such fact, which is trivial for Hermitian operators, but which turns out to be important for normal operators is given by the following immediate corollary of Proposition 1

Corollary 2 *If A is a normal operator in $\mathcal{B}(H)$, then* $\ker A = \ker A^*$

[1] To prove this fact, verify the *polarization identity* $\langle Sx, y\rangle = \frac{1}{4}\sum_{k=0}^{3} i^k \langle S(x + i^k y), x + i^k y\rangle$ and then use it to show if $\langle Sx, x\rangle = \langle Tx, x\rangle$ for all x, then $\langle Sx, y\rangle = \langle Tx, y\rangle$ for all x, y It follows that $Sx = Tx$ for all x and hence $S = T$

Proof $x \in \ker A \iff 0 = \|Ax\| = \|A^*x\| \iff x \in \ker A^*$ ∎

Remark One reason for the importance of this property of normal operators has to do with the *polar decomposition* of normal operators. In the usual polar decomposition, $A = V|A|$, V is a partial isometry with $\ker V = \ker A$ and final space $(\operatorname{ran} A)^- = (\ker A^*)^\perp$. If A is normal, then we can write $A = U|A|$, where U is unitary—for example, let U equal V on $(\ker A)^\perp$ and I on $\ker A$—this works because $\ker A = \ker A^*$. The main reason this property will be important for us, is that it implies normal operators, which are Fredholm, have index 0.

1 1 3 Compact Operators and Fredholm Operators

Our main purpose in this section is to define Fredholm operators and to state the main results that we will need in Chapter 3. Since the definition of a Fredholm operator depends on the set of compact operators, we begin by defining compact operators and stating some elementary results about them. In stating these results, the topology on H is always the norm topology and the topology on $\mathcal{B}(H)$ is always the operator norm topology. Proofs of these results can be found in [Dou72, Chapter 5].

Let $\operatorname{ball} H = \{ f \in H \mid \|f\| \leq 1 \}$ denote the (closed) unit ball in H. If H is infinite dimensional, then $\operatorname{ball} H$ is not a compact subset of H. If T is any operator, then it is an interesting exercise to show that $T(\operatorname{ball} H)$ is a closed subset of H.

An operator T is a *compact operator* if $T(\operatorname{ball} H)$ is a compact subset of H. Equivalently, T is compact if and only if for each bounded sequence $\{ f_n \}$ in H, the sequence $\{ Tf_n \}$ has a convergent subsequence in H. The set of compact operators, $\mathcal{K}(H)$, is a closed (two-sided) $*$-ideal in $\mathcal{B}(H)$. Here "$*$" means that if $T \in \mathcal{K}(H)$ then $T^* \in \mathcal{K}(H)$ and "ideal" means that linear combinations of operators in $\mathcal{K}(H)$ are in $\mathcal{K}(H)$ and products of operators with at least one operator in $\mathcal{K}(H)$ are in $\mathcal{K}(H)$.

An operator T is *finite rank* if the dimension of the range of T is finite. The set of finite rank operators, $\mathcal{F}(H)$, is a (two-sided) $*$-ideal in $\mathcal{B}(H)$. Each finite rank operator T is a compact operator (since $T(\text{ball}\, H)$ is a closed and bounded subset of the finite dimensional space $T(H)$). Hence, if H is finite dimensional, then $\mathcal{K}(H) = \mathcal{B}(H)$. However, if H is infinite dimensional, then I is not in $\mathcal{K}(H)$ (since $I(\text{ball}\, H) = \text{ball}\, H$ is not compact in H) and hence $\mathcal{K}(H) \neq \mathcal{B}(H)$.

In the remainder of this discussion of compact operators and Fredholm operators, we assume that H is infinite dimensional.

The fundamental fact about compact operators on a Hilbert space is that $\mathcal{K}(H)$ is the closure of $\mathcal{F}(H)$ in $\mathcal{B}(H)$. In other words, an operator K is compact if and only if K is the norm limit of finite rank operators.

Since we are assuming that H is an infinite dimensional Hilbert space, we have that $\mathcal{K}(H)$ is a proper closed (two-sided) $*$-ideal in $\mathcal{B}(H)$. It follows that the quotient algebra $\mathcal{Q}(H) = \mathcal{B}(H)/\mathcal{K}(H)$ is a unital Banach $*$-algebra with quotient norm

$$\|T + \mathcal{K}(H)\| = \inf_{K \in \mathcal{K}(H)} \|T - K\|$$

and identity $I + \mathcal{K}(H)$. In fact, $\mathcal{Q}(H)$ is a C^*-algebra. It is called the *Calkin algebra*. We denote the natural homomorphism from $\mathcal{B}(H)$ to $\mathcal{Q}(H)$ by π.

An operator T in $\mathcal{B}(H)$ is defined to be a *Fredholm operator* if $\pi(T)$ is an invertible element of the Calkin algebra. We denote the set of Fredholm operators by $\text{Fred}(H)$. Clearly, every invertible operator and every compact perturbation of an invertible operator is Fredholm. What makes Fredholm operators interesting is that not all of them arise this way. It follows immediately from the definition that $\text{Fred}(H)$ is an open subset of $\mathcal{B}(H)$ which is self-adjoint, closed under multiplication and invariant under compact perturbations. Here "self-adjoint" means if $T \in \text{Fred}(H)$ then $T^* \in \text{Fred}(H)$ and "invariant under compact perturbations" means that the sum of a Fredholm operator and a compact operator is still Fredholm.

The following characterization of Fredholm operators is fundamental to the development of Fredholm operator theory

Theorem (Atkinson). *An operator T in $\mathcal{B}(H)$ is Fredholm if and only if the range of T is closed,* $\dim \ker T$ *is finite and* $\dim \ker T^*$ *is finite*

Because of Atkinson's Theorem, if T is a Fredholm operator on H, then

$$\operatorname{ind}(T) = \dim \ker T - \dim \ker T^*$$

is a well defined integer called the *index* of the Fredholm operator T For $n \in \mathbf{Z}$, let $\operatorname{Fred}_n(H) = \{ T \in \operatorname{Fred}(H) \mid \operatorname{ind}(T) = n \}$ denote the set of Fredholm operators with index n

Remarks If T is an invertible operator, then T is a Fredholm operator of index 0 (since $\ker T = \ker T^* = \{0\}$) If N is a normal operator, which is also Fredholm, then N has index 0 (since $\ker N = \ker N^*$ is finite dimensional)

The index function, ind $\operatorname{Fred}(H) \to \mathbf{Z}$ is continuous Moreover, if S, T are Fredholm and K is compact, then

(1) $\operatorname{ind}(T^*) = -\operatorname{ind}(T)$

(2) $\operatorname{ind}(ST) = \operatorname{ind}(S) + \operatorname{ind}(T)$

(3) $\operatorname{ind}(T + K) = \operatorname{ind}(T)$

Note that for each $n \in \mathbf{Z}$, property (3) implies that the set $\operatorname{Fred}_n(H)$ is invariant under compact pertubations

Remark The fact that $\operatorname{Fred}_0(H)$ is invariant under compact pertubations leads to the *Fredholm alternative* for compact operators if K is compact and λ is a nonzero complex number, then λ is an eigenvalue of K of finite multiplicity or $K - \lambda I$ is invertible

For our purposes, the most important result from this Fredholm theory of operators on a Hilbert space H, is stated in the following theorem [Dou72, Theorem 5 36]

Theorem *If H is a Hilbert space, then the components of* $\mathrm{Fred}(H)$ *are precisely the sets* $\{\mathrm{Fred}_n(H) \mid n \in \mathbf{Z}\}$. *Moreover, the index map is a continuous homomorphism from* $\mathrm{Fred}(H)$ *onto* $\mathbf{Z}$ *which is invariant under compact perturbation.*

For each operator T, the spectrum of $\pi(T)$ in the Calkin algebra is called the *essential spectrum of* T, which we denote by $\sigma_e(T)$. Since all invertible operators are Fredholm, we have that $\sigma_e(T) \subseteq \sigma(T)$. By the definition of Fredholm operator, we have that T is Fredholm if and only if $0 \notin \sigma_e(T)$. More generally, $T - \lambda I$ is Fredholm if and only if $\pi(T - \lambda I) = \pi(T) - \lambda\pi(I)$ is invertible if and only if $\lambda \notin \sigma(\pi(T)) = \sigma_e(T)$.

Remark The idea behind defining Fredholm operators in terms of invertibility in the Calkin algebra is to generalize the the concept of *invertibility* to include the idea of asymptotic invertibility and thereby to retain some of the nice properties of invertibility. A similar idea has been applied to other *essential* properties of operators (*i.e.* properties that are preserved by π). For example, an operator T in $\mathcal{B}(H)$ is defined to be *essentially normal* if $\pi(T)$ is a normal element in the Calkin algebra. An example of the success of this approach is the Brown-Douglas-Fillmore Theorem [BDF73] which says that an essentially normal operator T can be written as the sum of a normal operator and a compact operator if and only if for every $\lambda \notin \sigma_e(T)$, $\mathrm{ind}(T - \lambda I) = 0$. A motivating example for the Brown-Douglas-Fillmore Theorem is the unilateral shift (a Toeplitz operator). In Section 3 1, we will see that the unilateral shift is an essentially normal, Fredholm operator which has nonzero index.

1 1 4 More Facts about Normal Operators

Before concluding this review section, there are a few more well-known facts about normal operators that should be mentioned. The following information is drawn from Davidson's book [Dav88, Chapter 0].

Using the fact that $\mathcal{B}(H)$ is a C^*-algebra with identity, we can apply C^*-algebra theory to it

Let $\mathcal{A}$ be an abelian Banach algebra with identity, then as a Banach space, its dual space $\mathcal{A}^*$ is the Banach space of bounded linear functionals from $\mathcal{A}$ to $\mathbf{C}$ A *multiplicative linear functional* on $\mathcal{A}$ is an algebra homomorphism from $\mathcal{A}$ to $\mathbf{C}$ Let $\Phi_{\mathcal{A}}$ denote the set of *nonzero* multiplicative linear functionals on $\mathcal{A}$, then since $\mathcal{A}$ is an abelian Banach algebra with identity, $\Phi_{\mathcal{A}}$ is nonempty and for every φ in $\Phi_{\mathcal{A}}$ $\varphi(1) = 1$ and $\|\varphi\| = 1$ It follows that $\Phi_{\mathcal{A}}$ is contained in the unit ball of $\mathcal{A}^*$, which by the Banach-Alaoglu Theorem is compact in the weak* topology Moreover, $\Phi_{\mathcal{A}}$ is a weak* closed subset of the unit ball of $\mathcal{A}^*$ and hence is also weak* compact

Let $\Phi_{\mathcal{A}}$ be equipped with the relative weak* topology of $\mathcal{A}^*$, then for every a in $\mathcal{A}$, the function $\hat{a}$ $\Phi_{\mathcal{A}} \to \mathbf{C}$, defined by $\hat{a}(\varphi) = \varphi(a)$, is continuous The map $a \mapsto \hat{a}$ from $\mathcal{A}$ to $C(\Phi_{\mathcal{A}})$ is called the *Gelfand transform*

Gelfand-Naimark Theorem *If $\mathcal{A}$ is an abelian C^*-algebra with identity, then the Gelfand transform is an isometric $*$-isomorphism of $\mathcal{A}$ onto $C(\Phi_{\mathcal{A}})$*

Note In particular, for every a in $\mathcal{A}$, $\|a\| = \|\hat{a}\|$ and $\sigma(a) = \sigma(\hat{a}) = \text{ran}(\hat{a})$ where $\text{ran}(\hat{a})$ denotes the range of the function $\hat{a}$ It follows that for every a in $\mathcal{A}$, $\hat{a}$ maps $\Phi_{\mathcal{A}}$ onto $\sigma(a)$

In our case, when N is a normal operator in $\mathcal{B}(H)$, the C^*-subalgebra of $\mathcal{B}(H)$ generated by N and I, denoted $C^*(N)$, is an abelian C^*-algebra with identity In fact, $C^*(N)$ is just the norm-closure of the linear span of $\{ N^m(N^*)^n \quad m, n \in \mathbf{N} \}$ Moreover, $\hat{N}$ is a special continuous function from $\Phi_{C^*(N)}$ onto $\sigma(N)$, it is a continuous bijection, from the compact space $\Phi_{C^*(N)}$ to the Hausdorff space $\sigma(N)$, and hence it is a homeomorphism! Therefore, we can identify $\Phi_{C^*(N)}$ and $\sigma(N)$ when applying the Gelfand-Naimark theorem to $C^*(N)$

C^*-Functional Calculus *If N is a normal operator in $\mathcal{B}(H)$, then $C^*(N)$ is iso-*

metrically $*$-isomorphic to $C(\sigma(N))$. Moreover, the inverse map yields a functional calculus such that $\bar{f}(N) = f(N)^*$ and $\sigma(f(N)) = f(\sigma(N))$ for all f in $C(\sigma(N))$.

Note. Since $C^*(N)$ is an abelian $*$-algebra, every operator in $C^*(N)$ is normal. In particular, every complex polynomial in z and $\bar{z}$, $p(z) = \sum_{k=0}^{K} \sum_{m=0}^{M} c_{k,m}\, z^k\, \bar{z}^m$, is continuous on all of $\mathbf{C}$ and hence $p(N) = \sum_{k=0}^{K} \sum_{m=0}^{M} c_{k,m}\, N^k\, (N^*)^m$ is a normal operator in $C^*(N)$.

Remark. This functional calculus can be used to show that positive operators have unique positive square roots (since they are normal operators with spectrum in $\mathbf{R}^+$ and the square root function is continuous on $\mathbf{R}^+$) and then for every operator A, we can define $|A| = (A^*A)^{1/2}$. This functional calculus can also be used to characterize Hermitian, positive and projection operators as normal operators with spectrum in $\mathbf{R}$, $\mathbf{R}^+$ and $\{0, 1\}$, respectively. For example, to show the Hermitian statement, observe that in either direction, A is normal and hence $\sigma(A^* - A) = \left\{ \overline{\lambda} - \lambda \quad \lambda \in \sigma(A) \right\}$, which is $\{0\}$ if and only if $\sigma(A) \subseteq \mathbf{R}$. Clearly then, A Hermitian implies that $\sigma(A) \subseteq \mathbf{R}$. In the other direction, $A^* - A$ is normal, so that $\|A^* - A\| = r(A^* - A) = 0$ (since $\sigma(A) \subseteq \mathbf{R}$) and therefore $A^* = A$. The main idea in getting the positive and projection statements, is to look at $|A| - A$ and $A^2 - A$, respectively.

Clearly, the set of normal operators is closed under scalar multiplication. However, in general, the the set of normal operators is not closed under operator addition or operator multiplication. For example in the 2×2 matrices, let

$$A = \begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix} \text{ and } B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

then A is normal since it is diagonal and B is normal since it is Hermitian. However, neither

$$AB = \begin{bmatrix} 1 & 1 \\ i & i \end{bmatrix} \text{ nor } A + B = \begin{bmatrix} 2 & 1 \\ 1 & 1+i \end{bmatrix}$$

is normal—this statement can be verified by applying the definition of normal op-

erator to AB and $A + B$, or, wait until the next section where we show that

$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is normal if and only if $c = (e^{i\theta})^2\,\overline{b}$ where $a - d = e^{i\theta}\,|a - d|$

The following theorem due to Fuglede [Fug50] gives a special commutativity property of normal operators, which immediately yields a sufficient condition for the sum and product of two normal operators to be normal. For an elegant proof of Fuglede's Theorem, see the proof by Rosenblum [Ros58] using the exponential function.

Fuglede's Theorem. *Let H be a complex Hilbert space, and let N be a normal operator on H. Then every operator on H that commutes with N, also commutes with N^*.*

Corollary. *If A and B are normal operators in $\mathcal{B}(H)$ and A commutes with B, then both AB and $A + B$ are normal operators.*

Proof Let $N_1 = AB$ and $N_2 = A + B$. Then verify $N_j^* N_j - N_j N_j^* = 0$ using the fact that, by the theorem, all of A, A^*, B and B^* commute with one another. ∎

Note In particular, since every scalar operator is normal and commutes with all operators, $N + \lambda I$ is normal for every normal operator N and complex λ.

This completes our review of basic facts and definitions that we need for discussing a Hilbert space and the operators on it.

1 2 Preliminary Normal Approximation Results

In this section, we establish basic definitions and terminology related specifically to the problem of normal approximation in $\mathcal{B}(H)$ and survey some known results about normal approximation due to Holmes [Hol74], Rogers [Rog76] and Phillips [Phi77]

Let $\mathcal{N} = \mathcal{N}(H)$ denote the set of normal operators on H

Let $\mathcal{S}$ be a fixed nonempty subset of $\mathcal{B}(H)$ Then for every operator A, the *distance* from A to $\mathcal{S}$ is denoted by $\operatorname{dist}(A, \mathcal{S})$ where

$$\operatorname{dist}(A, \mathcal{S}) = \inf \{ \|A - X\| \quad X \in \mathcal{S} \}$$

The $\mathcal{S}$ approximation problem for A involves finding at least one X_0 in $\mathcal{S}$ such that $\|A - X_0\| = \operatorname{dist}(A, \mathcal{S})$ In this context, X_0 is called a *best* $\mathcal{S}$ approximant for A or a *nearest* $\mathcal{S}$ approximant to A

Note Since the zero operator is normal, $\operatorname{dist}(A, \mathcal{N})$ is well defined for every A in $\mathcal{B}(H)$, moreover, we have an immediate upper bound on this distance

$$\operatorname{dist}(A, \mathcal{N}) \leq \|A - 0\| = \|A\|$$

1 2 1 Invariant Maps for the Normal Approximation Problem

Before looking at the normal approximation results of Holmes, Rogers and Phillips, we establish some tools for simplifying the normal approximation problem Following an idea by Ruhe [Ruh87, p 586], we say that the normal approximation problem is *invariant* under an invertible map $F \quad \mathcal{B}(H) \to \mathcal{B}(H)$, if for every operator A, $\operatorname{dist}(F(A), \mathcal{N}) = \operatorname{dist}(A, \mathcal{N})$ *and* N is a nearest normal approximant of A if and only if $F(N)$ is a nearest normal approximant of $F(A)$

Our main result gives three invariant maps for the normal approximation problem To prove one of them, we remind the reader of the following well-known fact

Lemma 1 *The operator norm is strongly unitarily invariant*
That is, $\|UAV\| = \|A\|$ for all operators A and all unitary operators U and V

Proof For every $x \in H$, $\|Ux\|^2 = \langle Ux, Ux\rangle = \langle U^*Ux, x\rangle = \langle x, x\rangle = \|x\|^2$, so $\|Ux\| = \|x\|$ Therefore, $\|UAx\| = \|Ax\|$ for all $x \in H$, hence $\|UA\| = \|A\|$ Then, $\|UAV\| = \|AV\| = \|(AV)^*\| = \|V^*A^*\| = \|A^*\| = \|A\|$ ∎

Proposition 2. *The normal approximation problem is invariant under*

(1) *Unitary equivalence That is, under $A \mapsto U^*AU$ for every unitary U*

(2) *Scalar translation That is, under $A \mapsto A + \lambda I$ for every complex λ*

(3) *Rotation That is, under $A \mapsto e^{i\theta}A$ for every real θ*

Proof The proofs of all three statements are very similar We will prove (1) completely and leave the details of (2) and (3) to the reader

To prove (1), let U be an arbitrary unitary operator and define $F(A) = U^*AU$ for every operator A The first step is to verify that F is invertible It is well known that F is an automorphism of $\mathcal{B}(H)$, but for completeness, we prove that F is invertible directly Let A and B be arbitrary operators, then $F(UAU^*) = A$, so F is surjective, and $\|F(A) - F(B)\| = \|U^*(A - B)U\| = \|A - B\|$, so F is injective

To see that the normal approximation problem is invariant under F, let A be an arbitrary operator Then for every normal operator N
$\|A - N\| = \|U^*(A - N)U\| = \|U^*AU - U^*NU\| \geq \text{dist}(U^*AU, \mathcal{N})$ since U^*NU is normal Hence, $\text{dist}(A, \mathcal{N}) \geq \text{dist}(U^*AU, \mathcal{N})$ Similarly, for every normal operator M $\|U^*AU - M\| = \|U^*(A - UMU^*)U\| = \|A - UMU^*\| \geq \text{dist}(A, \mathcal{N})$ since UMU^* is normal Hence, $\text{dist}(U^*AU, \mathcal{N}) \geq \text{dist}(A, \mathcal{N})$

We now have $\text{dist}(A, \mathcal{N}) = \text{dist}(U^*AU, \mathcal{N})$ and $\|A - N\| = \|U^*AU - U^*NU\|$ Hence N is a nearest normal approximant of A if and only if U^*NU is a nearest normal approximant of U^*AU This completes the proof of (1)

The proofs of (2) and (3) are similar The key facts for (2) are that $\|A - N\| = \|(A + \lambda I) - (N + \lambda I)\|$ and N is normal if and only if $N + \lambda I$ is normal The key

facts for (3) are $\|A - N\| = \|e^{i\theta}A - e^{i\theta}N\|$ and N is normal if and only if $e^{i\theta}N$ is normal ∎

Although the normal approximation problem is not formally invariant under scalar multiplication, it is *effectively invariant* in the sense that it is modified in a simple, reversible way given by the following corollary

Corollary 3 *For every operator A and every complex λ,*

$$\operatorname{dist}(\lambda A, \mathcal{N}) = |\lambda| \operatorname{dist}(A, \mathcal{N})$$

and if $\lambda \neq 0$, then N is a nearest normal approximant of A if and only if λN is a nearest normal approximant of λA

Proof The proof is similar to (1) in the proposition—we leave the details to the reader The key facts are for $\lambda \neq 0$, $\|A - N\| = \frac{1}{|\lambda|} \|\lambda A - \lambda N\|$ and N is normal if and only if λN is normal ∎

In order to get some idea of how these invariant maps are tools for simplifying the normal approximation problem, consider the normal approximation problem for $n \times n$ matrices By Schur's theorem, every $n \times n$ matrix is unitarily equivalent to an upper triangular matrix—but the normal approximation problem is invariant under unitary equivalence—so the normal approximation problem for $n \times n$ matrices reduces to the normal approximation problem for upper triangular $n \times n$ matrices Our other invariant maps for the normal approximation problem are less powerful in general, but the following example shows that they can be useful in special cases

Example The normal approximation problem for nondiagonal upper triangular 2×2 Toeplitz matrices *reduces* to the normal approximation problem for the matrix $\left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ since

$$\frac{1}{a_1}\left(\begin{bmatrix} a_0 & a_1 \\ 0 & a_0 \end{bmatrix} - a_0 I\right) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

and since the normal approximation problem is invariant under scalar translation and effectively invariant under scalar multiplication

1 2 2 Distance Estimates

If every operator in $\mathcal{B}(H)$ has a best $\mathcal{S}$ approximant, Holmes [Hol74] calls the set $\mathcal{S}$ *proximinal* in $\mathcal{B}(H)$ and points out that the sets of Hermitian, positive and compact operators are all proximinal in $\mathcal{B}(H)$—these results are established in [FH55, vR72], [Hal72] and [HK71], respectively. Holmes suggests that the normal approximation problem appears to be "deeper" than these others, primarily since the normals are not convex and hence, most of the usual approximation techniques are lost.

In [Hol74], Holmes proposes to begin the study of the normal approximation problem by trying to characterize those (nonzero) operators for which the zero operator is a best normal approximant. He calls such operators *antinormal*, since their distance to the set of normal operators is maximal.

Although Holmes' main results about antinormal operators are interesting, the only result that we will use is his preliminary theorem, which establishes upper and lower bounds for the distance from an arbitrary operator to the set of normal operators. For the interested reader, we briefly summarize his main results about antinormals before taking a detailed look at his theorem on distance bounds.

Holmes' main theorem establishes that a sufficient condition for an operator T in $\mathcal{B}(H)$ to be antinormal is that its distance to the set of unitary operators be $1 + \|T\|$, the largest possible distance to the unitaries (for every unitary U, $\|U - T\| \leq \|U\| + \|T\| = 1 + \|T\|$). Holmes observes that Halmos has shown that the unilateral shift (on a separable, infinite dimensional Hilbert space) satisfies the conditions of this theorem [Hal82, p. 273], hence the unilateral shift is antinormal.

Although Holmes was unable to fully characterize antinormal operators, the content of his other theorem about antinormal operators excludes invertible operators and compact operators from consideration. It says that, if T is an invertible operator in $\mathcal{B}(H)$, then $\operatorname{dist}(T, \mathcal{N}) \leq \frac{1}{2} \sup \{ |\lambda - \mu| : \lambda, \mu \in \sigma(|T|) \} < \frac{1}{2}\|T\| < \|T\|$, which implies that T is not antinormal, and then, with a little more work, that no

compact operator can be antinormal either This concludes our brief summary of Holmes' main results about antinormal operators

Holmes' theorem on distance bounds uses the already mentioned fact, that a best Hermitian approximant for an arbitrary operator is its real part Since this result is very important in its own right, and surprisingly easy to prove, we include it here before looking at Holmes' theorem

Note If $A = B + iC$ is an operator in $\mathcal{B}(H)$ with B, C Hermitian, then $\|B\| \leq \|A\|$ and $\|C\| \leq \|A\|$ To see this fact, *copy* the complex number proof B and C are the unique real and imaginary parts of A, respectively, with $B = \frac{1}{2}(A + A^*)$ and $C = \frac{1}{2i}(A - A^*)$, hence $\|B\|, \|C\| \leq \frac{1}{2}(\|A\| + \|A^*\|) = \|A\|$

Theorem 4 *If A is an operator in $\mathcal{B}(H)$, then the real part of A is a best Hermitian approximant for A*

Proof This proof is from the preamble by Halmos in [Hal72] Write $A = B + iC$ where B, C are Hermitian, then for every Hermitian operator R

$$\|A - R\| = \|(B - R) + iC\| \geq \|C\| = \|A - B\|$$

It follows that $\|A - B\| = \inf\{\, \|A - R\| \;\; R \text{ is Hermitian} \,\}$ and hence B, the real part of A, is a best Hermitian approximant for A ∎

Here is Holmes' distance estimate and his proof

Theorem 5 (Holmes) *If T is an operator in $\mathcal{B}(H)$, then*

$$\tfrac{1}{2} \sup_{\|x\|=1} |\, \|Tx\| - \|T^*x\| \,| \leq \operatorname{dist}(T, \mathcal{N}) \leq \tfrac{1}{2} \|T - T^*\|$$

Proof For the right-hand inequality, write $T = B + iC$ with B, C the real and imaginary parts of T, respectively Since B is normal, we have that

$$\operatorname{dist}(T, \mathcal{N}) \leq \|T - B\| = \|iC\| = \tfrac{1}{2} \|T - T^*\|$$

Moreover, since B is a *best* Hermitian approximant for T, this is the smallest upper bound we can get using Hermitian operators

For the left-hand inequality, take any unit vector x and any normal operator N, then by Proposition 1 of Section 1 1, $||Nx|| = ||N^*x||$ and so

$$\begin{aligned} |\,||Tx|| - ||T^*x||\,| &= |\,||Tx|| - ||Nx|| + ||N^*x|| - ||T^*x||\,| \\ &\leq |\,||Tx|| - ||Nx||\,| + |\,||N^*x|| - ||T^*x||\,| \\ &\leq ||(T-N)x|| + ||(N^*-T^*)x|| \\ &\leq ||T-N|| + ||N^*-T^*|| \\ &= 2\,||T-N|| \end{aligned}$$

Since this is true for any unit vector x and any normal operator N,

$$\sup_{||x||=1} |\,||Tx|| - ||T^*x||\,| \leq \inf_{N\in\mathcal{N}} 2\,||T-N|| = 2\,\mathrm{dist}(T,\mathcal{N})$$

Multipling by $\frac{1}{2}$, we have the left-hand inequality ∎

1 2 3 Not Every Operator Has A Best Normal Approximant

In [Rog76], Rogers shows that if H is an infinite dimensional (separable, complex) Hilbert space, then the set of normal operators is not proximinal in $\mathcal{B}(H)$ As mentioned earlier, it is this result of Rogers which motivates interest in trying to find special classes of operators that do have best normal approximants

We do not use this result of Rogers directly, however for the interested reader, we include a statement of his main theorem along with an outline of his proof We then give Rogers' example of an operator that satisfies the hypothesis of his theorem and hence does not have a best normal approximant

Theorem 6 (Rogers). *If T is an operator with dense range such that $\mathrm{dist}(T,\mathcal{N}) \leq \frac{1}{2}\,||T||$ and such that the kernel of T contains a maximal vector for*

T^*—i e , a nonzero vector f satisfying $\|T^*f\| = \|T^*\|\ \|f\|$, then T fails to have a best normal approximant

Outline of the proof Rogers' proof of this theorem is very clever His main tool is a lemma establishing that if f is a maximal vector for A, then $\langle Af, Ag\rangle = 0$ whenever $\langle f, g\rangle = 0$ To see this fact, first show that $\|Af\| = \|A\|\ \|f\|$ if and only if $A^*Af = \|A^*A\|\, f$, and then the lemma follows immediately

Essentially, his approach in the theorem is to assume T has a best normal approximant N and then get a contradiction Without loss of generality, Rogers takes $\|T\| = 1$ and $\|f\| = 1$, then he has $Tf = 0$ and $\|T^*f\| = 1$ By Holmes' distance estimate $\operatorname{dist}(T, \mathcal{N}) \geq \frac{1}{2}\,|\,\|Tf\| - \|T^*f\|\,| = \frac{1}{2}$, and so, by the hypothesis of the theorem, he has that $\operatorname{dist}(T, \mathcal{N}) = \frac{1}{2}$ Moreover, the distance from N^*f to 0 and to T^*f is less than or equal to $\frac{1}{2}$, hence $N^*f = \frac{1}{2}T^*f$ An important consequence of this fact is that f is a maximal vector for $N - T$ and $N^* - T^*$ as well as for T^*

The final stage of Rogers' proof is to show that Nf is orthogonal to the range of T—once that is done, he has his contradiction—by hypothesis, the range of T is dense in H and so $Nf = 0$, but $\|Nf\| = \|N^*f\| = \frac{1}{2}$ Since $H = \ker T \oplus (\ker T)^{\perp}$, it suffices to show that $\langle Nf, Tg\rangle = 0$, for every $g \in (\ker T)^{\perp}$ The clever part of the proof, is *how* Rogers accomplishes this final step—he cleverly introduces and removes terms that are zero For example, his first step is

$$\langle Nf, Tg\rangle = \langle Nf, Ng\rangle - \langle Nf, (N - T)g\rangle$$

and, along the way, he uses that $Tf = 0$, $\langle T^*f, T^*g\rangle = 0$, $\langle (N - T)f, (N - T)g\rangle = 0$ and $\langle (N - T)^*f, (N - T)^*g\rangle = 0$

Example (Rogers) The hypotheses of Rogers' theorem are satisfied by T = the adjoint of the unilateral weighted shift with weight sequence $(1, \frac{1}{2}, \frac{1}{3}, \ldots)$ and hence T is an example of an operator that does not have a best normal approximant

That is, with respect to some orthonormal basis $\{e_1, e_2, \quad\}$ of H

$$T^* e_j = \tfrac{1}{j} e_{j+1} \text{ for all } j$$

The matrices of T^* and T with respect to the basis $\{e_j\}$ are

$$T^* = \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ & \frac{1}{2} & 0 & & \\ & & \frac{1}{3} & 0 & \\ & & & & \end{bmatrix} \quad \text{and} \quad T = \begin{bmatrix} 0 & 1 & & & \\ & 0 & \frac{1}{2} & & \\ & & 0 & \frac{1}{3} & \\ & & & 0 & \\ & & & & \end{bmatrix}$$

Clearly, the range of T is dense and e_1 is in the kernel of T

To verify that e_1 is actually a maximal vector for T^*, we need to verify that $\|T^* e_1\| = \|T^*\| \ \|e_1\|$ However, $\|T^* e_1\| = \|e_2\| = 1$ and $\|T^*\| \ \|e_1\| = \|T^*\|$, so, we only need to verify that $\|T^*\| = 1$ To see this, suppose $x = \sum c_j e_j$ is a unit vector, then $\|T^* x\|^2 = \|T^*(\sum c_j e_j)\|^2 = \|\sum c_j T^* e_j\|^2 = \|\sum c_j \ \frac{1}{j} e_{j+1}\|^2 = \sum |\frac{1}{j} c_j|^2 \leq \sum |c_j|^2 = \|x\|^2 = 1$ Hence, $\|T^*\| \leq 1$, but $\|T^* e_1\| = \|e_2\| = 1$, so $\|T^*\| = 1$

It only remains to show that $\text{dist}(T, \mathcal{N}) \leq \frac{1}{2}\|T\| = \frac{1}{2}$ To do that, Rogers first defines the operators N_k for $k \geq 2$ via

$$N_k e_j = \begin{cases} \frac{1}{2} e_k & \text{if } j = 1 \\ \frac{1}{2} e_{j-1} & \text{if } j = 2, \quad , k \\ \frac{1}{2} e_j & \text{if } j > k \end{cases}$$

The matrix of N_k with respect to the basis $\{e_j\}$ is

$$N_k = \left[\begin{array}{ccccc|c} 0 & \frac{1}{2} & & & & \\ & 0 & \frac{1}{2} & & & \\ & & 0 & & & \\ & & & & \frac{1}{2} & \\ \frac{1}{2} & & & & 0 & \\ \hline & & & & & \frac{1}{2} \\ & & & & & & \frac{1}{2} \\ & & & & & \end{array}\right]$$

where the upper left block is $k \times k$

Rogers then states that for all such k-values, N_k is normal (since $2\,N_k$ is a unitary operator) and $\|N_k - T\| \leq \frac{1}{2} + \frac{1}{k}$, hence $\text{dist}(T, \mathcal{N}) \leq \frac{1}{2}$. Carefully verifying these statements about N_k is an excellent opportunity for us to introduce

1. A special $k \times k$ unitary matrix that plays an important role in the search for normal approximants of $k \times k$ Toeplitz matrices.

2. A result that gives us the exact norms for a special class of operators/matrices and which can also be used to get norm estimates for more general operators.

The Basic Circulant Matrix

In the setting of $n \times n$ complex matrices, the following matrix is called the *basic circulant matrix*.

$$C = C_n = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & 0 & & \\ & & & & 1 \\ 1 & & & & 0 \end{bmatrix}$$

Let $(x_1, \ldots, x_n)$ denote a row vector in $\mathbf{C}^n$ and let $x = (x_1, \ldots, x_n)^T$ denote its *transpose*—a column vector in $\mathbf{C}^n$. Note that $Cx = (x_2, \ldots, x_n, x_1)^T$, hence C is a backward circulant on $\mathbf{C}^n$. Also note that $C^*x = (x_n, x_1, \ldots, x_{n-1})^T$, hence C^* is a foreward circulant on $\mathbf{C}^n$. It is now easy to see that $C^*C = CC^* = I$ and hence C is unitary.

If we think of C_n as the matrix of an operator on n-dimensional Hilbert space with respect to the basis $\{e_1, \ldots, e_n\}$, then the matrix C_n corresponds to the operator

$$C_n\, e_j = \begin{cases} e_n & \text{if } j = 1 \\ e_{j-1} & \text{if } j = 2, \ldots, n \end{cases}$$

In Rogers' example, let H_k denote the span of $\{e_1, \ldots, e_k\}$, then $H = H_k \oplus (H_k)^\perp$ and we can decompose $N_k = \frac{1}{2}C_k \oplus \frac{1}{2}I$, where I represents the identity operator on $(H_k)^\perp$. As Rogers said, $2\,N_k = C_k \oplus I$ is unitary, hence N_k is normal.

Weighted Partial Permutations

In Rogers' example, we showed that the weighted (forward) shift T^* with weight sequence $(1, \frac{1}{2}, \frac{1}{3}, \ldots)$ has norm 1—the largest weight. We would like to extend this result to a larger class of operators which includes T^*, T, N_k and that will help us get Rogers' norm estimate for $N_k - T$.

The most important feature of our proof that $\|T^*\| = 1$, turns out to be that T^* maps basis vectors to weighted basis vectors, *injectively*. The operators N_k satisfy this property as well, but, the weighted (backward) shift T does not quite satisfy it. T maps e_1 to 0, but then maps all the other basis vectors to weighted basis vectors injectively—that is enough for the proof to carry through.

Proposition 7 *Suppose H is a separable, complex Hilbert space (possibly finite dimensional) with orthonormal basis $\{e_j\}_{j\in J}$ and suppose $\{w_j\}_{j\in J}$ is a bounded complex weight sequence. Define $J_0 = \{\, j \in J \;:\; w_j = 0 \,\}$ and $J_1 = J \setminus J_0$.*

If $\pi : J_1 \to J$ is injective, then the operator A, defined by

$$A\, e_j = \begin{cases} 0 & \text{if } j \in J_0 \\ w_j\, e_{\pi(j)} & \text{if } j \in J_1, \end{cases}$$

is bounded with $\|A\| = \sup\limits_{j\in J} |w_j|$.

Proof Let $M = \sup_{j\in J} |w_j|$. If $M = 0$, then every $w_j = 0$ and A is the zero operator, hence $\|A\| = 0 = M$. Otherwise, $M > 0$ and J_1 is not empty.

To see that $\|A\| \le M$, suppose $x = \sum_{j\in J} c_j e_j$ is a unit vector. Then $\|Ax\|^2 = \|A(\sum_{j\in J} c_j e_j)\|^2 = \|\sum_{j\in J} c_j A e_j\|^2 = \|\sum_{j\in J_1} c_j\, w_j\, e_{\pi(j)}\|^2 = \sum_{j\in J_1} |c_j|^2\, |w_j|^2 = \sum_{j\in J} |c_j|^2\, |w_j|^2 \le M^2 \sum_{j\in J} |c_j|^2 = M^2\, \|x\|^2 = M^2$. Therefore, A is bounded and $\|A\| \le M$.

To see that $\|A\| \ge M$, observe that for every $j \in J$,

$$\|A\, e_j\| = \left\{ \begin{array}{ll} \|0\| & \text{if } w_j = 0 \\ \|w_j\, e_{\pi(j)}\| & \text{if } w_j \neq 0 \end{array} \right\} = |w_j|$$

This implies $\|A\| \geq |w_j|$ for all $j \in J$ and hence $\|A\| \geq \sup_{j \in J} |w_j| = M$ ∎

Remark 1 If H is finite dimensional in the proposition, we can always extend π to be a permutation on all of J, hence it seems appropriate to call the operator A a weighted permutation When H is infinite dimensional we may not be able to extend π to be a permutation on all of J, hence, calling A a *weighted partial permutation* seems more appropriate

Remark 2 Any finite or infinite matrix with at most 1 nonzero entry in each row and each column satisfies the hypothesis of the proposition

We can now use Proposition 7 to get Rogers' norm estimate on $N_k - T$

$$N_k - T = \left[\begin{array}{ccccc|cc} 0 & (\frac{1}{2}-1) & & & & & \\ & 0 & (\frac{1}{2}-\frac{1}{2}) & & & & \\ & & 0 & & & & \\ & & & \ddots & (\frac{1}{2}-\frac{1}{k-1}) & & \\ \frac{1}{2} & & & & 0 & \frac{-1}{k} & \\ \hline & & & & & \frac{1}{2} & \frac{-1}{k+1} \\ & & & & & & \frac{1}{2} \\ & & & & & & \end{array}\right] = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix}$$

Therefore

$$\begin{aligned} \|N_k - T\| &= \| \begin{bmatrix} A & 0 \\ 0 & \frac{1}{2}I \end{bmatrix} + \begin{bmatrix} 0 & B \\ 0 & C - \frac{1}{2}I \end{bmatrix} \| \\ &\leq \| \begin{bmatrix} A & 0 \\ 0 & \frac{1}{2}I \end{bmatrix} \| + \| \begin{bmatrix} 0 & B \\ 0 & C - \frac{1}{2}I \end{bmatrix} \| \\ &= \tfrac{1}{2} + \tfrac{1}{k} \qquad \text{(by Proposition 7)} \end{aligned}$$

Since this is true for every $k \geq 2$, it follows that $\operatorname{dist}(T, \mathcal{N}) \leq \frac{1}{2} = \frac{1}{2}\|T\|$, which completes the verification that the weighted (backward) shift T, satisfies the hypothesis of Rogers' theorem and hence does not have a best normal approximant

1 2 4 Every Binormal Operator Has A Best Normal Approximant

An operator on $H \oplus H$ is called *binormal*, if it is unitarily equivalent to a 2×2 operator matrix

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}$$

in which T_{11}, T_{12}, T_{21}, T_{22} are *commuting, normal* operators on H. Such operators are generalized 2×2 complex matrices—every 2×2 complex matrix is a binormal operator on $\mathbf{C} \oplus \mathbf{C}$. Strengthening the analogy with the 2×2 complex matrices, Radjavi and Rosenthal [RR73, Theorem 7.20] proved that each binormal operator is unitarily equivalent to a 2×2 upper triangular operator matrix

$$\begin{bmatrix} A & B \\ 0 & C \end{bmatrix}$$

in which A, B, C, are commuting normal operators on H.

Since the normal approximation problem is invariant under unitary equivalence, the normal approximation problem for binormal operators reduces to these upper triangular representations. In [Phi77], Phillips shows that every upper triangular representation of a binormal operator has a nearest normal approximant and hence, every binormal operator has a nearest normal approximant. In fact, for each upper triangular representation of a binormal operator, he actually exhibits a nearest normal approximant. More precisely:

Theorem 8 (Phillips). *Let T be a binormal operator whose upper triangular form is*

$$T = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix}$$

Then,

(1) $\operatorname{dist}(T, \mathcal{N}) = \frac{1}{2}\|B\|$,

(2) $X_0 = \begin{bmatrix} A & \frac{1}{2}B \\ \frac{1}{2}V^2B^* & C \end{bmatrix}$ *is a nearest normal to T,*

where $(A - C) = V|A - C|$ and V is unitary.

Remark Since A commutes with C, $A - C$ is a normal operator It follows that $A - C$ does have at least one *unitary* polar decomposition and hence, the theorem does exhibit at least one nearest normal approximant

Remark In [BHK91], Bhatia-Horn-Kittaneh show that the statement of Phillips' theorem remains true if the operator norm is replaced by any unitarily invariant norm

The fact that Phillips' theorem exhibits an explicit formula which gives a nearest normal approximant to an arbitrary binormal operator is very significant in the history of the normal approximation problem—no such formula is known for any other class of operators Although this state of affairs is a significant motivating factor for this thesis, we do not use the general result in the sequel We refer the interested reader to [Phi77] for a complete proof On the other hand, since we will be looking at $n \times n$ Toeplitz matrices, it is important for us to see how the proof of this theorem goes in the special case of 2×2 complex matrices—that case includes the 2×2 Toeplitz matrices As a bonus to the reader, since binormal operators are generalized 2×2 complex matrices, the proof for 2×2 complex matrices can serve as a model for the general proof

In order to help us show how Phillips' proof for binormal operators works in the special case of 2×2 complex matrices, we take the opportunity to first prove a proposition which has independent interest It characterizes 2×2 normal matrices—in a way suggested by the form of Phillips' nearest normal approximant

Proposition 9 *Let* $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ *be a* 2×2 *complex matrix Then* A *is normal if and only if* $c = v^2 \overline{b}$ *where* $a - d = v\,|a - d|$ *and* $|v| = 1$

Proof By straight computation

$$A^*A - AA^* = \begin{bmatrix} |c|^2 - |b|^2 & (\overline{a} - \overline{d})\, b - (a - d)\, \overline{c} \\ (a - d)\, \overline{b} - (\overline{a} - \overline{d})\, c & |b|^2 - |c|^2 \end{bmatrix}$$

It follows that A is normal if and only if $|c| = |b|$ and $(a-d)\,\overline{b} = (\overline{a}-\overline{d})\,c$

If $a \neq d$, then there is a unique v such that $a - d = v\,|a-d|$ and this v has modulus 1 Then, since $a - d \neq 0$ A is normal if and only if

$$|c| = |b| \quad \text{and} \quad c = \frac{a-d}{\overline{a}-\overline{d}}\,\overline{b} = \frac{v\,|a-d|}{\overline{v}\,|a-d|}\,\overline{b} = v^2\,\overline{b}$$

Since $|v| = 1$, this yields A is normal if and only if $c = v^2\,\overline{b}$ Hence, the proposition is true in the $a \neq d$ case

If $a = d$, then $(a-d)\,\overline{b} = (\overline{a}-\overline{d})\,c$ for all values of b and c and so the situation simplifies to A is normal if and only if $|c| = |b|$ However

$$\begin{aligned} |c| = |b| &\iff |c| = |\overline{b}| \\ &\iff c = \omega\,\overline{b} \qquad \text{for some } |\omega| = 1 \\ &\iff c = v^2\,\overline{b} \qquad \text{for some } |v| = 1 \end{aligned}$$

In this case, any v satisfies $a - d = v\,|a-d|$ and hence, the proposition is true in the $a = d$ case ∎

Remark It might be worthwhile to think of this proposition in its 2 cases separately If $a \neq d$, then $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ is normal if and only if c is the unique number $c = v^2\,\overline{b}$ where $a - d = v\,|a-d|$ However $\left[\begin{smallmatrix} a & b \\ c & a \end{smallmatrix}\right]$ is normal if and only if $|c| = |b|$

Note If $a \neq d$ but $a - d \in \mathbf{R}$, then A is normal if and only if $c = (\pm 1)^2\,\overline{b} = \overline{b}$ In particular, if the a and d are *both* real numbers, then A is normal if and only if A is Hermitian

Proof The following is a proof of Phillips' theorem for the *special case* of 2×2 matrices More precisely, this is how Phillips' proof would look in this simple case

By Schur's theorem, every 2×2 matrix is unitarily equivalent to an upper triangular matrix—but the normal approximation problem is invariant under unitary equivalence—so the normal approximation problem for 2×2 matrices reduces to the normal approximation problem for upper triangular 2×2 matrices

Suppose T is an upper triangular 2×2 complex matrix

$$T = \begin{bmatrix} a & b \\ 0 & c \end{bmatrix}$$

Let

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$$

be an arbitrary normal matrix. Then

$$T - X = \begin{bmatrix} a - x_{11} & b - x_{12} \\ -x_{21} & c - x_{22} \end{bmatrix}$$

and so

$$\begin{aligned} \|T - X\|^2 &\geq \max \left\{ \|\mathrm{col}_1(T - X)\|^2, \|\mathrm{col}_2(T - X)\|^2 \right\} \\ &\geq \max \{ |x_{21}|^2, |b - x_{12}|^2 \} \\ &= \max \{ |x_{12}|^2, |b - x_{12}|^2 \} \qquad \text{since } X \text{ is normal} \end{aligned}$$

Therefore

$$\begin{aligned} \|T - X\| &\geq \max \{ |x_{12}|, |b - x_{12}| \} \\ &\geq \tfrac{1}{2} (|x_{12}| + |b - x_{12}|) \\ &\geq \tfrac{1}{2} | x_{12} + b - x_{12} | \\ &= \tfrac{1}{2} |b| \end{aligned}$$

However, if $a - c = v\,|a - c|$ and $|v| = 1$, then

$$X_0 = \begin{bmatrix} a & \frac{1}{2} b \\ \frac{1}{2} v^2 \overline{b} & c \end{bmatrix}$$

is a normal matrix by Proposition 9, and

$$\|T - X_0\| = \left\| \begin{bmatrix} 0 & \frac{1}{2} b \\ -\frac{1}{2} v^2 \overline{b} & 0 \end{bmatrix} \right\| = \tfrac{1}{2} |b|$$

It follows that $\mathrm{dist}(T, \mathcal{N}) = \frac{1}{2} |b|$ and that X_0 is a nearest normal approximant to the 2×2 upper triangular matrix T. This completes the proof of Phillips' theorem for the special case of 2×2 matrices. ∎

Remark In the general theorem, Phillips has to consider $2n \times 2n$ normal matrices where "n is some ordinal number" It is not necessary for him to characterize such matrices! He can show that X_0 is normal directly, by an "easy calculation" In order to show that $\text{dist}(T, \mathcal{N}) \geq \frac{1}{2}\|B\|$, his main tool is the fact that for any $2n \times 2n$ normal matrix X, $\|\text{col}_k(X)\| = \|\text{row}_k(X)\|$ for all $k \in \{1, \quad, 2n\}$ Since this property of normal matrices will be needed in our work with finite dimensional matrices, we take the opportunity to prove it here as one of our preliminary normal approximation results

Proposition 10 *If $A = [A_{ij}]$ is a normal $n \times n$ matrix, then for all $k \in \{1, \quad, n\}$ $\|\text{col}_k(A)\| = \|\text{row}_k(A)\|$*

Proof Let $\{e_1, \quad, e_n\}$ be the standard ordered basis for $\mathbf{C}^n$ Then for every $k \in \{1, \quad, n\}$,

$$\begin{aligned}
\|\text{col}_k(A)\|^2 &= \|Ae_k\|^2 = \|A^*e_k\|^2 \qquad \text{since } A \text{ is normal} \\
&= \|\text{col}_k(A^*)\|^2 = \sum_{j=1}^{n} |(A^*)_{jk}|^2 \\
&= \sum_{j=1}^{n} |\overline{A}_{kj}|^2 = \sum_{j=1}^{n} |A_{kj}|^2 \\
&= \|\text{row}_k(A)\|^2 \quad \blacksquare
\end{aligned}$$

When $n = 2$, we get the following corollary which is all we really needed in order to mimic Phillips' proof in the 2×2 case

Corollary 11 *If $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is a normal 2×2 matrix, then $|c| = |b|$*

Proof By the Proposition, $|a|^2 + |c|^2 = \|\text{col}_1(A)\|^2 = \|\text{row}_1(A)\|^2 = |a|^2 + |b|^2$ and therefore, $|c| = |b|$ $\blacksquare$

This completes our discussion of preliminary normal approximation results As stated in the Introduction, little else is known about the normal approximation

problem. For example, in contrast to the 2×2 matrix case, no one has exhibited an explicit formula which gives a nearest normal approximant to an arbitrary 3×3 upper triangular matrix. This state of affairs motivates the search for special classes of operators for which nearest normal approximants can be found.

The remainder of this thesis is devoted to studying the normal approximation problem for two special classes of operators. In Chapter 2, we begin with the $n \times n$ upper triangular Toeplitz matrices and in Chapter 3, we consider Toeplitz operators on $H^2(\mathbf{T})$.

Chapter 2

Normal Approximants for Upper Triangular Toeplitz Matrices

In this chapter, we make a modest beginning on the study of the normal approximation problem for (finite dimensional) Toeplitz matrices by restricting our attention to upper triangular Toeplitz matrices. That is, upper triangular $n \times n$ matrices, where $n \geq 2$, of the form

$$T = \begin{bmatrix} a_0 & a_1 & & a_{n-1} \\ 0 & \ddots & \ddots & \\ & \ddots & \ddots & a_1 \\ 0 & & 0 & a_0 \end{bmatrix} \tag{2.1}$$

which are constant along the main diagonal and constant along each diagonal parallel to the main diagonal. Upper triangular Toeplitz matrices with exactly one nonzero diagonal will play a distinguished role here—we call such matrices superdiagonal Toeplitz matrices.

Our first step is to exhibit a formula for a nearest normal matrix to an arbitrary $n \times n$ superdiagonal Toeplitz matrix. For fixed n, it is interesting that all these normal approximants are in the commutative C^*-algebra generated by a special unitary matrix (the basic circulant), and hence their sum is a normal matrix. Since all upper triangular Toeplitz matrices can be written as a sum of superdiagonal ones, the sum of the individual approximants is a *natural* normal approximant to consider. We verify that this sum is a best normal approximant in the 2×2 case. Enticingly, it is also a best normal approximant for Jordan blocks (which are special upper triangular Toeplitz matrices). However, we show via a 3×3 counterexample that it is not a best normal approximant in the general case. We carry out a

comparative analysis in the 3×3 case but, unfortunately, are unable to decide upon a best normal approximant

As a final finite dimensional topic, we consider direct sums of upper triangular Toeplitz matrices We show that for direct sums of superdiagonal Toeplitz matrices and Jordan blocks, a best normal approximant is the direct sum of the individual best normal approximants However, we don't expect that a direct sum of individual best normal approximants will likely be a best normal approximant in general

2.1 Superdiagonal Toeplitz Matrices

A **nonzero** Toeplitz matrix T is *superdiagonal* if it is upper triangular, as in equation (2 1), and exactly one of $a_0, \quad , a_{n-1}$ is not zero If a_k is the one nonzero value, then T is naturally called a kth superdiagonal Toeplitz matrix[1] In addition, since we want to think of an arbitrary scalar multiple of a kth superdiagonal Toeplitz matrix, as still being a kth superdiagonal Toeplitz matrix, we also call the $n \times n$ **zero** matrix *superdiagonal* and we allow it to be arbitrarily called a kth superdiagonal Toeplitz matrix, for any $k = 0, \quad , (n-1)$ In this section, we exhibit a formula for a nearest normal matrix to an arbitrary $n \times n$ superdiagonal Toeplitz matrix

2 1 1 The Basic Superdiagonal Matrix

The particular first superdiagonal Toeplitz matrix

$$S = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}_{n \times n} \tag{2 2}$$

which has 1's on the first superdiagonal and 0's elsewhere is a fundamental matrix in the theory of upper triangular Toeplitz matrices We call S the *basic superdiagonal*

[1] In order to avoid having to always discuss the main diagonal of an upper triangular matrix separately, we find it convenient to think of the main diagonal as the 0th superdiagonal

matrix (to emphasize its similarity to the basic circulant matrix that we introduced in Section 1 2 3) or *the* first superdiagonal Toeplitz matrix (to emphasize that it is a Toeplitz matrix) This matrix S is also commonly called the *backward shift* because of its effect on vectors in $\mathbf{C}^n$ If $\{ e_1, \quad , e_n \}$ is the standard basis for $\mathbf{C}^n$, then

$$Se_j = \begin{cases} 0 & \text{if } j = 1 \\ e_{j-1} & \text{if } j = 2, \quad , n \end{cases}$$

and, if $x = (x_1, \quad , x_n)^T$ is a column vector in $\mathbf{C}^n$, then $Sx = (x_2, \quad , x_n, 0)^T$, $S^2 x = (x_3, \quad , x_n, 0, 0)^T, \quad , S^{n-1} x = (x_n, 0, \quad , 0)^T$ and $S^n x = 0$ It follows that S^2 can be called the 2-backward shift Moreover, S^2 has 1's on the 2nd superdiagonal, 0's elsewhere and hence can also be called *the* 2nd superdiagonal Toeplitz matrix Similarly, for $k \in \{ 1, \quad , n-1 \}$, S^k can be called the k-backward shift or *the* kth superdiagonal Toeplitz matrix At the extremes, $S^n = 0$ is the n-backward shift, $S^0 = I$ is the 0-backward shift and *the* 0th superdiagonal Toeplitz matrix

We can now demonstrate why S is a fundamental matrix in the theory of upper triangular Toeplitz matrices To begin with, every kth superdiagonal Toeplitz matrix can be written in the form $a_k S^k$ Moreover, every upper triangular Toeplitz matrix can obviously be written as a sum of superdiagonal Toeplitz matrices, so any upper triangular Toeplitz matrix T can be written as a polynomial matrix in S as follows

$$T = a_0 I + a_1 S + a_2 S^2 + \quad + a_{n-1} S^{n-1} \tag{2 3}$$

Conversely, any such polynomial matrix in S is a sum of superdiagonal Toeplitz matrices and hence is an upper triangular Toeplitz matrix

The following lemma exhibits a nearest normal matrix to S Although this lemma is really just a special case of our main theorem about superdiagonal Toeplitz matrices, we choose to prove it separately as an easy-to-visualize, detailed model for our general proof

Lemma 1 *For $n \geq 2$, the $n \times n$ basic superdiagonal matrix*

$$S = \begin{bmatrix} 0 & 1 & & \\ & & & \\ & & & 1 \\ & & & 0 \end{bmatrix}$$

is at distance $\frac{1}{2}$ from the set of normal matrices and a nearest normal approximant is given by

$$N = \begin{bmatrix} 0 & \frac{1}{2} & & \\ & & & \\ & & & \frac{1}{2} \\ \frac{1}{2} & & & 0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 0 & 1 & & \\ & & & \\ & & & 1 \\ 1 & & & 0 \end{bmatrix} = \frac{1}{2} C$$

where C is the (unitary) $n \times n$ basic circulant matrix

Proof Clearly N is normal (it is a multiple of a unitary) Moreover,

$$S - N = \begin{bmatrix} 0 & \frac{1}{2} & & \\ & & & \\ & & & \frac{1}{2} \\ -\frac{1}{2} & & & 0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 0 & 1 & & \\ & & & \\ & & & 1 \\ -1 & & & 0 \end{bmatrix} = \frac{1}{2} B \qquad (2\ 4)$$

By inspection, $S-N$ is a weighted permutation [Section 1 2 3] and hence $\|S - N\| = \frac{1}{2}$

Remark The matrix B in equation (2 4) is unitary, hence $S - N = \frac{1}{2} B$ is actually a normal matrix and $\|S - N\|$ could have been obtained by $\|S - N\| = \frac{1}{2} \|B\| = \frac{1}{2}$

At this point, we know that $\operatorname{dist}(S, \mathcal{N}) \leq \|S - N\| = \frac{1}{2}$ In order to show that $\operatorname{dist}(S, \mathcal{N}) \geq \frac{1}{2}$, we follow the method used by Phillips in [Phi77] Let $X = [x_{ij}]$ be an arbitrary $n \times n$ normal matrix, then

$$S - X = \begin{bmatrix} -x_{11} & (1 - x_{12}) & & -x_{1n} \\ -x_{21} & -x_{22} & & \\ & & & (1 - x_{n-1,n}) \\ -x_{n1} & -x_{2n} & & -x_{nn} \end{bmatrix}$$

and

$$\begin{aligned}
\|S-X\|^2 &\geq \max\left\{ \|\mathrm{col}_1(S-X)\|^2, \ldots, \|\mathrm{col}_n(S-X)\|^2 \right\} \\
&\geq \max\left\{ \|\mathrm{col}_1(S-X)\|^2, \|\mathrm{col}_2(S-X)\|^2 \right\} \\
&= \max\left\{ \|-\mathrm{col}_1(X)\|^2, \|\mathrm{col}_2(S-X)\|^2 \right\} \quad \text{(by inspection of } S-X\text{)} \\
&= \max\left\{ \|\mathrm{row}_1(X)\|^2, \|\mathrm{col}_2(S-X)\|^2 \right\} \quad \text{(since } X \text{ is normal)} \\
&\geq \max\{\, |x_{12}|^2, |1-x_{12}|^2 \,\}
\end{aligned}$$

Therefore

$$\begin{aligned}
\|S-X\| &\geq \max\{\, |x_{12}|, |1-x_{12}| \,\} \\
&\geq \tfrac{1}{2}\left(|x_{12}| + |1-x_{12}| \right) \\
&\geq \tfrac{1}{2}\,|\, x_{12} + (1-x_{12}) \,| \\
&= \tfrac{1}{2} \\
&= \|S-N\| \\
&\geq \mathrm{dist}(S, \mathcal{N})
\end{aligned}$$

Since this is true for any normal X, it follows that $\mathrm{dist}(S,\mathcal{N}) = \|S-N\| = \frac{1}{2}$ and that $N = \frac{1}{2}C$ is a nearest normal matrix to S ∎

Remark The preceeding proof is the one we actually used in discovering that $\frac{1}{2}C$ is a nearest normal to S It can be shortened by using Holmes' distance estimate [Section 1 2 2] to obtain that $\mathrm{dist}(S,\mathcal{N}) \geq \frac{1}{2}$ More precisely, if $\{ e_1, \ldots, e_n \}$ is the standard ordered basis for $\mathbf{C}^n$, then

$$\mathrm{dist}(S,\mathcal{N}) \geq \tfrac{1}{2}\,|\, \|Se_n\| - \|S^*e_n\| \,| = \tfrac{1}{2}\,|\, 1-0 \,| = \tfrac{1}{2}$$

The key ingredient of our main theorem about superdiagonal Toeplitz matrices is that for $k \neq 0$, the kth superdiagonal Toeplitz matrix S^k is also at distance $\frac{1}{2}$ from the set of normal matrices and a nearest normal approximant is given by $\frac{1}{2}C^k$

Moreover, $S^k - \frac{1}{2}C^k = \frac{1}{2}B^k$ where B is the unitary matrix defined in equation (2 4) and hence $S^k - \frac{1}{2}C^k$ is a normal matrix In order to prove these statements, we need to analyze powers of the basic circulant matrix C and of the unitary matrix B in the same way that we analyzed powers of the basic superdiagonal matrix

2.1 2 Two Special Unitary Circulants

We have already seen [Section 1 2 3] that the basic circulant matrix

$$C = \begin{bmatrix} 0 & 1 & & \\ & & & \\ & & & 1 \\ 1 & & & 0 \end{bmatrix}_{n \times n} \tag{2 5}$$

is a unitary matrix which operates as a backward circulant on $\mathbf{C}^n$ If $\{ e_1, \quad , e_n \}$ is the standard ordered basis for $\mathbf{C}^n$, then

$$Ce_j = \begin{cases} e_n & \text{if } j = 1 \\ e_{j-1} & \text{if } j = 2, \quad , n \end{cases}$$

and if $x = (x_1, \quad , x_n)^T$ is a column vector in $\mathbf{C}^n$, then $Cx = (x_2, \quad , x_n, x_1)^T$, $C^2 x = (x_3, \quad , x_n, x_1, x_2)^T$, $\quad , C^{n-1} x = (x_n, x_1, \quad , x_{n-1})^T$ and $C^n x = x$ It follows that for $k \in \{ 1, \quad , n \}$, C^k can be called the k-backward circulant

Similarly, if $\omega_1, \quad , \omega_n$ are nonzero complex numbers, then the matrix

$$U = \begin{bmatrix} 0 & \omega_1 & & \\ & & & \\ & & & \omega_{n-1} \\ \omega_n & & & 0 \end{bmatrix} \quad \text{with adjoint} \quad U^* = \begin{bmatrix} 0 & & & \overline{\omega}_n \\ \overline{\omega}_1 & & & \\ & & & \\ & & \overline{\omega}_{n-1} & 0 \end{bmatrix}$$

is a weighted backward circulant and for $k \in \{ 1, \quad , n \}$, U^k is a weighted k-backward circulant Moreover

$$U^*U = \begin{bmatrix} |\omega_n|^2 & & & \\ & |\omega_1|^2 & & \\ & & & \\ & & & |\omega_{n-1}|^2 \end{bmatrix} \quad \text{and} \quad UU^* = \begin{bmatrix} |\omega_1|^2 & & & \\ & |\omega_2|^2 & & \\ & & & \\ & & & |\omega_n|^2 \end{bmatrix}$$

Hence, U is unitary if and only if $|\omega_j| = 1$ for all $j = 1, \ldots, n$

Our immediate goal is to analyze powers of the special unitary circulant matrices C and B which are both of the form

$$U_\omega = \begin{bmatrix} 0 & 1 & & \\ & & \ddots & \\ & & & 1 \\ \omega & & & 0 \end{bmatrix} \tag{2 6}$$

Remark Although we will not directly appeal to a visual argument to prove our general superdiagonal result, our general proof is motivated by the fact that for $k \in \{1, \ldots, (n-1)\}$ C^k has 1's on the kth superdiagonal, 1's on the $(n-k)$th subdiagonal, and 0's elsewhere Similarly, our observation that $S^k - \frac{1}{2}C^k = \frac{1}{2}B^k$ is because B^k has 1's on the kth superdiagonal, (-1)'s on the $(n-k)$th subdiagonal, and 0's elsewhere The following nonvisual lemma is what we will formally use to prove our main theorem—the visual corollary is presented as an interesting observation Both of these results handle C and B at the same time by looking at the special weighted circulant matrix U_ω defined in equation (2 6)

Lemma 2 *Given any $n \geq 2$ and any nonzero complex number ω, let*

$$U = U_\omega = \begin{bmatrix} 0 & 1 & & \\ & & \ddots & \\ & & & 1 \\ \omega & & & 0 \end{bmatrix}$$

be the corresponding $n \times n$ weighted circulant matrix and let $\{e_1, \ldots, e_n\}$ be the standard ordered basis for $\mathbf{C}^n$ Then, for all $k = 1, \ldots, (n-1)$

$$U^k e_j = \begin{cases} \omega e_{n-k+j} & \text{if } j = 1, \ldots, k \\ e_{j-k} & \text{if } j = (k+1), \ldots, n \end{cases} \tag{2 7}$$

Proof It is true for $k = 1$ If $n = 2$, then we are done If $n > 2$, then make the induction hypothesis that it is true for a k-value satisfying $1 \leq k < (n-1)$

If $j = 1, \ldots, k$ then by the induction hypothesis $U^{k+1} e_j = U\omega e_{n-k+j}$ However, $n - k + j \geq n - k + 1 > 1$ since $j \geq 1$ and $k < n$ Hence $U^{k+1} e_j = \omega U e_{n-k+j} = \omega e_{n-k+j-1} = \omega e_{n-(k+1)+j}$

If $j = k + 1$ then by the induction hypothesis $U^{k+1} e_j = U e_{j-k} = U e_1 = \omega e_n = \omega e_{n-(k+1)+j}$

If $j = (k+2), \ldots, n$ then by the induction hypothesis $U^{k+1} e_j = U e_{j-k}$ However, $j - k > 1$, hence $U^{k+1} e_j = e_{j-k-1} = e_{j-(k+1)}$ ∎

Corollary 3 *Let ω and U be as in the Lemma Then for $k \in \{1, \ldots, (n-1)\}$, U^k has 1's on the kth superdiagonal, ω's on the $(n-k)$th subdiagonal and 0's elsewhere Furthermore, $U^n = \omega I$*

Proof Let $\{e_1, \ldots, e_n\}$ be the standard ordered basis for $\mathbf{C}^n$ Then for each $k = 1, \ldots, (n-1)$, we can think of U^k as a row vector of column vectors as follows

$$\begin{aligned} U^k &= [U^k e_1, \ldots, U^k e_k, U^k e_{k+1}, \ldots, U^k e_n] \\ &= [\omega e_{n-k+1}, \ldots, \omega e_{n-k+k}, e_{k+1-k}, \ldots, e_{n-k}] \quad \text{by the Lemma} \\ &= [\omega e_{n-k+1}, \ldots, \omega e_n, e_1, \ldots, e_{n-k}] \end{aligned}$$

By inspection, U^k has ω's in entries $(n-k+1, 1), (n-k+2, 2), \ldots, (n, k)$—this is precisely the $(n-k)$th subdiagonal of U^k, 1's in entries $(1, k+1), (2, k+2), \ldots, (n-k, n)$—this is precisely the kth superdiagonal of U^k, 0's elsewhere

Furthermore, from our knowledge of U^{n-1} and U, it follows that

$$U^n = UU^{n-1} = U[\omega e_2, \ldots, \omega e_n, e_1] = [\omega e_1, \ldots, \omega e_{n-1}, \omega e_n] = \omega I \text{ ∎}$$

2.1 3 The Main Theorem

We now have the tools to prove our main theorem about normal approximation of arbitrary superdiagonal Toeplitz matrices

Theorem 4 *Suppose T is an arbitrary $n \times n$ superdiagonal Toeplitz matrix, where $n \geq 2$. Let S denote the $n \times n$ basic superdiagonal matrix and let C denote the $n \times n$ basic circulant matrix.*

(1) *If $T = a_0 S^0 = a_0 I$ is a 0th superdiagonal Toeplitz matrix, then T is normal. It follows that $\operatorname{dist}(T, \mathcal{N}) = 0$ and that T is its own unique nearest normal approximant.*

(2) *If $T = a_k S^k$ is a kth superdiagonal Toeplitz matrix, where $0 < k < n$, then $\operatorname{dist}(T, \mathcal{N}) = \frac{1}{2}|a_k|$ and $\frac{1}{2} a_k C^k$ is a nearest normal approximant to T.*

Proof (1) is obvious—it is included in the statement of this theorem so that normal approximants of all of the superdiagonal Toeplitz matrices are recorded in one place.

(2) is trivial if $a_k = 0$. If $a_k \neq 0$, then (2) can be reduced to the case where $a_k = 1$. By Corollary 3 of Section 1 2

$$\operatorname{dist}(a_k S^k, \mathcal{N}) = |a_k| \operatorname{dist}(S^k, \mathcal{N})$$

and since $a_k \neq 0$, N is a nearest normal to S^k if and only if $a_k N$ is a nearest normal to $a_k S^k$. Therefore, it will suffice to show that $\operatorname{dist}(S^k, \mathcal{N}) = \frac{1}{2}$ and that $N = \frac{1}{2} C^k$ is a nearest normal to S^k. In other words, it will suffice to prove (2) for $a_k = 1$. To accomplish this, we follow the steps we used in Lemma 1 (the $k = 1$ case), but without showing the matrices.

Since $N = \frac{1}{2} C^k$ is a multiple of the unitary C, we have that N is normal. Let $\{e_1, \ldots, e_n\}$ denote the standard ordered basis for $\mathbf{C}^n$. Then using Lemma 2 and the fact that S^k is the k-backward shift

$$(S^k - \tfrac{1}{2} C^k) e_j = \begin{cases} (0 - \frac{1}{2} e_{n-k+j}) & \text{if } j = 1, \ldots, k \\ (e_{j-k} - \frac{1}{2} e_{j-k}) & \text{if } j = (k+1), \ldots, n \end{cases}$$

$$= \frac{1}{2} \begin{cases} -e_{n-k+j} & \text{if } j = 1, \ldots, k \\ e_{j-k} & \text{if } j = (k+1), \ldots, n \end{cases}$$

$$= \tfrac{1}{2} B^k e_j$$

where B is the special weighted circulant matrix corresponding to $\omega = -1$ in Lemma 2. Since $S^k - \frac{1}{2} C^k$ and $\frac{1}{2} B^k$ agree on a basis, we have that $S^k - \frac{1}{2} C^k = \frac{1}{2} B^k$. However, B is unitary since all of its weights have modulus 1. Hence $S^k - \frac{1}{2} C^k$ is actually normal and $\|S^k - \frac{1}{2} C^k\| = \frac{1}{2} \|B^k\| = \frac{1}{2}$.

At this point we have that $\operatorname{dist}(S^k, \mathcal{N}) \leq \|S^k - N\| = \frac{1}{2}$. To show that $\operatorname{dist}(S^k, \mathcal{N}) \geq \frac{1}{2}$, let $X = [x_{ij}]$ be an arbitrary normal matrix, then

$$\begin{aligned}
\|S^k - X\|^2 &\geq \max\left\{ \|\mathrm{col}_1(S^k - X)\|^2, \quad , \|\mathrm{col}_n(S^k - X)\|^2 \right\} \\
&\geq \max\left\{ \|\mathrm{col}_1(S^k - X)\|^2, \|\mathrm{col}_{k+1}(S^k - X)\|^2 \right\} \\
&= \max\left\{ \| - \mathrm{col}_1(X)\|^2, \|\mathrm{col}_{k+1}(S^k - X)\|^2 \right\} \quad (\text{since } \mathrm{col}_1(S^k) = 0) \\
&= \max\left\{ \|\mathrm{row}_1(X)\|^2, \|\mathrm{col}_{k+1}(S^k - X)\|^2 \right\} \quad (\text{since } X \text{ is normal}) \\
&\geq \max\{ |x_{1,k+1}|^2, |1 - x_{1,k+1}|^2 \}
\end{aligned}$$

As in Lemma 1, this implies $\|S^k - X\| \geq \frac{1}{2} = \|S^k - N\| \geq \operatorname{dist}(S^k, \mathcal{N})$. Since X was an arbitrary normal matrix, it follows that $\operatorname{dist}(S^k, \mathcal{N}) = \frac{1}{2} = \|S^k - N\|$ and hence $N = \frac{1}{2} C^k$ is a nearest normal approximant. ∎

Remark. The preceeding proof is the one we actually used in discovering that $\frac{1}{2} C^k$ is a nearest normal to S^k. As in the remark following Lemma 1, the proof here can also be shortened by using Holmes' distance estimate to obtain that $\operatorname{dist}(S^k, \mathcal{N}) \geq \frac{1}{2}$.

This completes our analysis of the normal approximation problem for superdiagonal Toeplitz matrices—given any superdiagonal Toeplitz matrix, we can calculate how far it is from the set of normal matrices and we can exhibit a nearest normal approximant.

2.2 Upper Triangular Toeplitz Matrices

Given an arbitrary $n \times n$ upper triangular Toeplitz matrix, T, where $n \geq 2$, we can write T as a sum of n superdiagonal matrices as we did in equation (2 3)

$$T = a_0 I + a_1 S + a_2 S^2 + \quad + a_{n-1} S^{n-1}$$

where S is the $n \times n$ basic superdiagonal matrix Moreover, we know a best normal approximant for each of these n superdiagonal matrices individually $a_0 I$ is its own best normal approximant and for $k = 1, \quad , (n-1)$, $\frac{1}{2} a_k C^k$ is a best normal approximant for $a_k S^k$, where C is the unitary $n \times n$ basic circulant matrix Since each of these best normal approximants is in the commutative C^*-algebra generated by the unitary C (and the identity), their sum

$$N = a_0 I + \tfrac{1}{2} a_1 C + \tfrac{1}{2} a_2 C^2 + \quad + \tfrac{1}{2} a_{n-1} C^{n-1}$$

is normal and hence is a natural normal approximant to consider

In this section, we reserve N to denote this sum of individual normal approximants We have already mentioned that, in general, N may not be a nearest normal to T However, N is always an interesting normal approximant for T since $T - N$ has some special properties Using part of the proof of Theorem 4 of Section 2 1,

$$\begin{aligned} T - N &= 0I + a_1 (S - \tfrac{1}{2} C) + \quad + a_{n-1} (S^{n-1} - \tfrac{1}{2} C^{n-1}) \\ &= \tfrac{1}{2} a_1 B + \quad + \tfrac{1}{2} a_{n-1} B^{n-1} \end{aligned}$$

where B is the unitary circulant with 1's on the first superdiagonal, $B_{n1} = -1$ and 0's elsewhere It follows that $T - N$ is normal since it is in the commutative C^*-algebra generated by the unitary B Therefore $\|T - N\| = r(T - N) = \max \{ |\lambda| \ \lambda \in \sigma(T - N) \}$ However, $T - N = p(B)$ where p is the polynomial $p(z) = \frac{1}{2} a_1 z + \quad + \frac{1}{2} a_{n-1} z^{n-1}$ Hence $\sigma(T - N) = \{ p(\lambda) \ \lambda \in \sigma(B) \}$ One way to find $\sigma(B)$ is to observe that $B^n = (-1) I$, so that $\{ -1 \} = \sigma(B^n) = \{ \lambda^n \ \lambda \in \sigma(B) \}$

and hence $\sigma(B) \subseteq \{\lambda \;\; \lambda^n = -1\}$ However, we can see that every such nth root, λ, of (-1) is an eigenvalue of B by observing that $(1, \lambda, \lambda^2, \quad , \lambda^{n-1})^T$ is an associated eigenvector Hence $\sigma(B) = \{\lambda \;\; \lambda^n = -1\}$ Putting these observations together, we have that

$$\|T - N\| = \max_{\lambda^n=-1} |\tfrac{1}{2}a_1\lambda + \quad + \tfrac{1}{2}a_{n-1}\lambda^{n-1}| \tag{2 8}$$

If nothing else, this formula for $\|T - N\|$ gives us an upper bound for $\mathrm{dist}(T, \mathcal{N})$

Recall, we also have Holmes' upper bound for $\mathrm{dist}(T, \mathcal{N})$ [Section 1 2 2] which is based on the distance from T to $H = \frac{1}{2}(T + T^*)$, the real part of T

$$\mathrm{dist}(T, \mathcal{N}) \leq \|T - H\| = \tfrac{1}{2}\|T - T^*\|$$

Moreover, we know H is a best Hermitian approximant for T and hence H is also a natural normal approximant for us to consider In this section, we reserve H for the real part of the upper triangular Toeplitz matrix T

2 2 1 The 2×2 Case

As promised earlier, we now verify that N is a nearest normal to T in the 2×2 case In this case

$$T = a_0 I + a_1 S = \begin{bmatrix} a_0 & a_1 \\ 0 & a_0 \end{bmatrix},$$

$$N = a_0 I + \tfrac{1}{2}a_1 C = \begin{bmatrix} a_0 & \frac{1}{2}a_1 \\ \frac{1}{2}a_1 & a_0 \end{bmatrix},$$

$$T - N = 0I + \tfrac{1}{2}a_1 B = \begin{bmatrix} 0 & \frac{1}{2}a_1 \\ -\frac{1}{2}a_1 & 0 \end{bmatrix},$$

and

$$\|T - N\| = \tfrac{1}{2}|a_1|$$

One way to see that N is a nearest normal to T is to recall that we know $\frac{1}{2}a_1 C$ is a nearest normal to $a_1 S$ However, by Proposition 2 of Section 1 2, the normal

approximation problem is invariant under scalar translation Hence $a_0 I + \frac{1}{2}a_1 C = N$ is a nearest normal to $a_0 I + a_1 S = T$

Another way to see that N is a nearest normal to T is to recall the 2×2 case of Phillips' Theorem [Section 1 2 4] which says $\operatorname{dist}(T, \mathcal{N}) = \frac{1}{2}|a_1|$ and a nearest normal to T is

$$X_0(v) = \begin{bmatrix} a_0 & \frac{1}{2}a_1 \\ \frac{1}{2}v^2\overline{a_1} & a_0 \end{bmatrix}$$

for any v with $|v| = 1$ The fact that N is normal and $\|T - N\| = \frac{1}{2}|a_1| = \operatorname{dist}(T, \mathcal{N})$ is enough to verify that N is a nearest normal to T Moreover, if $a_1 = |a_1|e^{i\theta}$ is a polar form of a_1, then $X_0(e^{i\theta}) = N$ That is, in the 2×2 case, N is one of the nearest normals to T given in Phillips' Theorem—as long as $a_1 \neq 0$, the 2×2 case of Phillips' Theorem gives us infinitely many nearest normals to the 2×2 upper triangular Toeplitz matrix T—one for each $|v| = 1$ For example, another nearest normal to T is

$$X_0(1) = \begin{bmatrix} a_0 & \frac{1}{2}a_1 \\ \frac{1}{2}\overline{a_1} & a_0 \end{bmatrix}$$

which is the real part of T **if** a_0 is a real number

Remark In the general 2×2 case, we cannot say that H, the real part of T, is a nearest normal to T since

$$T - H = \begin{bmatrix} a_0 & a_1 \\ 0 & a_0 \end{bmatrix} - \begin{bmatrix} \operatorname{Re}(a_0) & \frac{1}{2}a_1 \\ \frac{1}{2}\overline{a_1} & \operatorname{Re}(a_0) \end{bmatrix} = \begin{bmatrix} i\operatorname{Im}(a_0) & \frac{1}{2}a_1 \\ -\frac{1}{2}\overline{a_1} & i\operatorname{Im}(a_0) \end{bmatrix}$$

and so, if a_0 is not a real number, then

$$\|T - H\| \geq \|\operatorname{col}_2(T - H)\| > \tfrac{1}{2}|a_1| = \operatorname{dist}(T, \mathcal{N})$$

However, since we know that the normal approximation problem is invariant under scalar translation, this distinction about a_0 being real or not is somewhat superficial since we can reduce the 2×2 upper triangular Toeplitz problem to the case where $a_0 = 0$ by replacing T with $T - a_0 I$, if necessary Then, for this **reduced** 2×2 problem, we have both N and H as nearest normals to T

Remark Similarly, we can reduce the normal approximation problem for $n \times n$ upper triangular Toeplitz matrices to the case where $a_0 = 0$ As an immediate consequence of this observation and the fact that we know a best normal approximant for any superdiagonal Toeplitz matrix, $a_k S^k$, we also know a best normal approximant for any upper triangular Toeplitz matrix, $T = a_0 I + a_k S^k$, which is just a scalar translation of a superdiagonal Toeplitz matrix In fact, the sum of the 2 individual best normal approximants, $N = a_0 I + \frac{1}{2} a_k C^k$, is a best normal approximant in this case To prove this statement, we can copy our first proof for 2×2 upper triangular Toeplitz matrices almost verbatim We know $\frac{1}{2} a_k C^k$ is a nearest normal to $a_k S^k$ and we know the normal approximation problem is invariant under scalar translation Hence $a_0 I + \frac{1}{2} a_k C^k = N$ is a nearest normal to $a_0 I + a_k S^k = T$ Moreover, $\operatorname{dist}(a_0 I + a_k S^k, \mathcal{N}) = \operatorname{dist}(a_k S^k, \mathcal{N}) = \frac{1}{2} |a_k|$

In particular, we can solve the normal approximation problem for any $n \times n$ Jordan block

2 2 2 Jordan Blocks

Recall that for $n \geq 2$ and $\lambda \in \mathbf{C}$, the $n \times n$ Jordan block corresponding to λ is given by

$$J_n(\lambda) = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix} = \lambda I + S$$

where I is the $n \times n$ identity matrix and S is the $n \times n$ basic superdiagonal matrix

By the immediately preceeding remark, we know that a best normal approximant for $J_n(\lambda) = \lambda I + S$ is $N = \lambda I + \frac{1}{2} C$, where C is the $n \times n$ basic circulant matrix Moreover, $\operatorname{dist}(J_n(\lambda), \mathcal{N}) = \frac{1}{2}$ for every $n \geq 2$ and $\lambda \in \mathbf{C}$

Since 1×1 Jordan blocks, $J_1(\lambda) = [\lambda]$, play an important role in the theory of matrices, we mention for completeness that they are their own, unique, best normal

approximants

Preview As a special topic at the end of this chapter, we will show that a best normal approximant for a direct sum of Jordan blocks is a direct sum of corresponding best normal approximants This is an enticing result—if A is any square matrix, then it is known that A is *similar* to a direct sum of Jordan blocks, J (*i e* there exists an invertible matrix P such that $A = PJP^{-1}$) We do not pursue this result immediately since, unfortunately, it does not really give us information about arbitrary square matrices—in general, the normal approximation problem is not invariant under similarity For example, a nearest normal to $J_2(0) = \left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ is $N = \frac{1}{2}\left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$, but for the invertible $P = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 2 \end{smallmatrix}\right]$ we have that $PNP^{-1} = \frac{1}{4}\left[\begin{smallmatrix} 0 & 1 \\ 4 & 0 \end{smallmatrix}\right]$ is not even normal and hence cannot be a nearest normal to $PJ_2(0)P^{-1}$

2 2 3 The 3×3 Case

Let S denote the 3×3 basic superdiagonal matrix and let C denote the 3×3 basic circulant matrix The general 3×3 upper triangular Toeplitz matrix has the form

$$T = \begin{bmatrix} a_0 & a_1 & a_2 \\ 0 & a_0 & a_1 \\ 0 & 0 & a_0 \end{bmatrix} = a_0 I + a_1 S + a_2 S^2$$

However, the normal approximation problem is invariant under scalar translation so we can reduce the problem to the case where $a_0 = 0$ by replacing T with $T - a_0 I$, if necessary At this point,

$$T = \begin{bmatrix} 0 & a_1 & a_2 \\ 0 & 0 & a_1 \\ 0 & 0 & 0 \end{bmatrix} = a_1 S + a_2 S^2$$

If $a_1 = 0$, then $T = a_2 S^2$ is a 2nd superdiagonal matrix and $N = \frac{1}{2} a_2 C^2$ is a best normal approximant Similarly if $a_2 = 0$ Therefore, it only remains to solve the $a_1 a_2 \neq 0$ case

Since the normal approximation problem is effectively invariant under scalar multiplication, we can reduce this case to the case where $a_1 = 1$ by replacing T

with $\frac{1}{a_1}T$, if necessary. Hence, it only remains to solve the normal approximation problem for 3×3 upper triangular Toeplitz matrices of the form

$$T = \begin{bmatrix} 0 & 1 & a \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = S + a\,S^2 \tag{2 9}$$

In this case, the sum of normal approximants is given by

$$N = \tfrac{1}{2}\,C + \tfrac{1}{2}\,a\,C^2 = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2}a \\ \frac{1}{2}a & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2}a & 0 \end{bmatrix} \tag{2 10}$$

and the real part of T is given by

$$H = \tfrac{1}{2}\,(T + T^*) = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2}a \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2}\overline{a} & \frac{1}{2} & 0 \end{bmatrix} \tag{2 11}$$

Hence

$$T - N = \tfrac{1}{2}\begin{bmatrix} 0 & 1 & a \\ -a & 0 & 1 \\ -1 & -a & 0 \end{bmatrix} = \tfrac{1}{2}\,B + \tfrac{1}{2}\,a\,B^2 \tag{2 12}$$

and

$$T - H = \tfrac{1}{2}\begin{bmatrix} 0 & 1 & a \\ -1 & 0 & 1 \\ -\overline{a} & -1 & 0 \end{bmatrix} = \tfrac{1}{2}\,(T - T^*) \tag{2 13}$$

We will soon show that for a suitable value of a, $\|T - H\| < \|T - N\|$ and hence, the sum of best normal approximants is **not** a best normal approximant for every upper triangular Toeplitz matrix. To facilitate the proof of our example and to prepare for a comparative analysis of N and H as normal approximants for $T = S + aS^2$, we first look into computing $\|T - N\|$ and $\|T - H\|$ for arbitrary values of a.

The following proposition exhibits a formula for $\|T - N\|$ which is valid for any $a \in \mathbf{C}$. Its corollary simplifies the corresponding formula when $a \in \mathbf{R}$.

Proposition 1 *If* $T = S + aS^2$ *and* $N = \frac{1}{2}C + \frac{1}{2}aC^2$, *where* $a = b + ic$ *with* $b, c \in \mathbf{R}$, *then*

$$\|T - N\| = \tfrac{1}{2} \max\left\{ \sqrt{1 + |a|^2 - 2b}\,,\ \sqrt{1 + |a|^2 + b + \sqrt{3}\,|c|} \right\}$$

Proof Recall that since $T - N$ is a polynomial in the unitary circulant B, we have a formula for $\|T - N\|$ Using the formula given by equation (2 8) we get

$$\|T - N\| = \max_{\lambda^3 = -1} |\tfrac{1}{2}\lambda + \tfrac{1}{2}a\lambda^2| = \max_{\lambda^3 = -1} \tfrac{1}{2}\,|\lambda|\,|1 + a\lambda| = \tfrac{1}{2} \max_{\lambda^3 = -1} |1 + a\lambda|$$

First observe that for any λ

$$|1 + a\lambda|^2 = (1 + a\lambda)(1 + \overline{a\lambda}) = 1 + a\lambda + \overline{a\lambda} + |a\lambda|^2 = 1 + |a|^2 + \mathrm{Re}\,(2a\lambda)$$

In our case, we only have to check the 3 cube roots of (-1)

$$\begin{aligned} \lambda = -1 \quad &\Rightarrow \quad \mathrm{Re}\,(2a\lambda) = \mathrm{Re}\,[2(b + ic)(-1)] = -2b \\ &\Rightarrow \quad |1 + a\lambda|^2 = 1 + |a|^2 - 2b \end{aligned}$$

$$\begin{aligned} \lambda = e^{i\pi/3} = \tfrac{1}{2}(1 + i\sqrt{3}) \quad &\Rightarrow \quad \mathrm{Re}\,(2a\lambda) = \mathrm{Re}\,[2(b + ic)\tfrac{1}{2}(1 + i\sqrt{3})] = b - \sqrt{3}\,c \\ &\Rightarrow \quad |1 + a\lambda|^2 = 1 + |a|^2 + b - \sqrt{3}\,c \end{aligned}$$

$$\begin{aligned} \lambda = e^{-i\pi/3} = \tfrac{1}{2}(1 - i\sqrt{3}) \quad &\Rightarrow \quad \mathrm{Re}\,(2a\lambda) = \mathrm{Re}\,[2(b + ic)\tfrac{1}{2}(1 - i\sqrt{3})] = b + \sqrt{3}\,c \\ &\Rightarrow \quad |1 + a\lambda|^2 = 1 + |a|^2 + b + \sqrt{3}\,c \end{aligned}$$

By inspection of these 3 candidates for $|1 + a\lambda|^2$, we have our result

$$\|T - N\| = \tfrac{1}{2} \max_{\lambda^3 = -1} |1 + a\lambda| = \tfrac{1}{2} \max\left\{ \sqrt{1 + |a|^2 - 2b}\,,\ \sqrt{1 + |a|^2 + b + \sqrt{3}\,|c|} \right\} \ \blacksquare$$

Corollary 2 *If* $T = S + aS^2$ *and* $N = \frac{1}{2}C + \frac{1}{2}aC^2$, *where* $a \in \mathbf{R}$, *then*

$$\|T - N\| \ = \ \tfrac{1}{2} \max\left\{ \sqrt{1 + |a|^2 - 2a}\,,\ \sqrt{1 + |a|^2 + a} \right\}$$

$$= \begin{cases} \tfrac{1}{2}\sqrt{1 + |a|^2 + a} & \text{if } a \geq 0 \\ \\ \tfrac{1}{2}\sqrt{1 + |a|^2 + 2|a|} = \tfrac{1}{2}\,(1 + |a|) & \text{if } a < 0 \end{cases}$$

Proof The first equality comes directly from the proposition and the fact that for $a \in \mathbf{R}$, we have $c = \operatorname{Im} a = 0$ and $b = \operatorname{Re} a = a$ The second equality, is obvious from the first ∎

We now turn our attention to trying to compute $\|T - H\|$ By thinking of T in terms of its Hermitian decomposition, $T = H + iK$ where H and K are Hermitian, we see that $T - H = iK$ is skew-Hermitian and hence normal It follows that $\|T - H\| = r(T-H)$ To find $\sigma(T-H)$ we can look for the roots of its characteristic polynomial, $\det[\lambda I - (T - H)]$ To avoid some fractions, we choose to first find $\sigma[2(T-H)]$ by looking at its characteristic polynomial, $p(\lambda) = \det[\lambda I - 2(T - H)]$ Referring to $T - H$ in equation (2 13), we have

$$\begin{aligned} p(\lambda) &= \det\left(\begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} - \begin{bmatrix} 0 & 1 & a \\ -1 & 0 & 1 \\ -\overline{a} & -1 & 0 \end{bmatrix}\right) \\ &= \det\begin{bmatrix} \lambda & -1 & -a \\ 1 & \lambda & -1 \\ \overline{a} & 1 & \lambda \end{bmatrix} \\ &= (\lambda^3 + \overline{a} - a) - (-|a|^2\lambda - \lambda - \lambda) \\ &= \lambda^3 + (2 + |a|^2)\lambda + (\overline{a} - a) \end{aligned} \tag{2 14}$$

Finding the roots of $p(\lambda)$ becomes easy when $a \in \mathbf{R}$ In that case, $p(\lambda) = \lambda^3 + (2 + |a|^2)\lambda = \lambda[\lambda^2 + (2 + |a|^2)]$ which has roots $0, \pm i\sqrt{2 + |a|^2}$ Hence, when $a \in \mathbf{R}$, we have $\|2(T - H)\| = r[2(T - H)] = \sqrt{2 + |a|^2}$ and therefore $\|T - H\| = \frac{1}{2}\sqrt{2 + |a|^2}$

The following proposition restates this fact for future reference

Proposition 3 *If $T = S + aS^2$ where $a \in \mathbf{R}$ and $H = \frac{1}{2}(T + T^*)$ is the real part of T, then*

$$\|T - H\| = \tfrac{1}{2}\sqrt{2 + |a|^2}$$

Proof Has already been done completely in the discussion leading up to the statement of the proposition ∎

Example When $a = 2$, we have

$$T = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

By Corollary 2

$$||T - N|| = \tfrac{1}{2}\sqrt{1 + |2|^2 + 2} = \tfrac{1}{2}\sqrt{7}$$

By Proposition 3

$$||T - H|| = \tfrac{1}{2}\sqrt{2 + |2|^2} = \tfrac{1}{2}\sqrt{6}$$

Therefore, $||T - H|| < ||T - N||$ and hence we have an example of a 3×3 upper triangular Toeplitz matrix for which N is **not** a best normal approximant

Remark Similarly, for any $a \in \mathbf{R}$, Corollary 2 and Proposition 3 make it is easy to compare N and H as normal approximants for $T = S + aS^2$ The results of this comparative analysis are contained in the following proposition

Proposition 4 *If* $a \in \mathbf{R}$ *and*

$$T = \begin{bmatrix} 0 & 1 & a \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

then the following summarizes how

$$N = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2}a \\ \frac{1}{2}a & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2}a & 0 \end{bmatrix} \quad \textit{and} \quad H = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2}a \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2}a & \frac{1}{2} & 0 \end{bmatrix}$$

compare as normal approximants for T

(1) *If* $a \in (-\frac{1}{2}, 1)$ *then* $||T - N|| < ||T - H||$
(2) *If* $a \notin [-\frac{1}{2}, 1]$ *then* $||T - N|| > ||T - H||$
(3) *If* $a = 1$ *then* $N = H$ *and hence* $||T - N|| = ||T - H||$
(4) *If* $a = -\frac{1}{2}$ *then* $N \neq H$ *but* $||T - N|| = ||T - H||$

Proof By Corollary 2

$$\|T - N\| = \begin{cases} \frac{1}{2}\sqrt{1 + |a|^2 + a} & \text{if } a \geq 0 \\ \\ \frac{1}{2}\sqrt{1 + |a|^2 + 2|a|} & \text{if } a < 0 \end{cases}$$

and by Proposition 3

$$\|T - H\| = \tfrac{1}{2}\sqrt{2 + |a|^2} = \tfrac{1}{2}\sqrt{1 + |a|^2 + 1} \quad \text{for any } a \in \mathbf{R}$$

In our case, we know that $a \in \mathbf{R}$, so by inspection of these 2 formulas

(a) $\|T - N\| < \|T - H\| \iff (a \geq 0 \text{ and } a < 1) \text{ or } (a < 0 \text{ and } 2|a| < 1)$
$\iff a \in [0, 1) \text{ or } a \in (-\frac{1}{2}, 0)$
$\iff a \in (-\frac{1}{2}, 1)$ Hence (1) is true

(b) $\|T - N\| > \|T - H\| \iff (a \geq 0 \text{ and } a > 1) \text{ or } (a < 0 \text{ and } 2|a| > 1)$
$\iff a > 1 \text{ or } a < -\frac{1}{2}$
$\iff a \notin [-\frac{1}{2}, 1]$ Hence (2) is true

(c) $\|T - N\| = \|T - H\| \iff (a \geq 0 \text{ and } a = 1) \text{ or } (a < 0 \text{ and } 2|a| = 1)$
$\iff a = 1 \text{ or } a = -\frac{1}{2}$

By simply substituting into N and H, it is easy to verify that $N = H$ when $a = 1$ and $N \neq H$ when $a = -\frac{1}{2}$ Hence (3) and (4) are true ∎

Remark Proposition 4 does not exhibit best normal approximants for the 3×3 matrix $T = S + aS^2$ when $a \in \mathbf{R}$ It only compares N and H as normal approximants for T in this case At this point, it is only when $a = 0$, and hence T is the 3×3 basic superdiagonal matrix, that we know for certain N is a best normal approximant for T, and then since $0 \in (-\frac{1}{2}, 1)$, we know by Proposition 4, that H is not a best normal approximant for T

So far, the superdiagonal cases in which we have successfully exhibited a best normal approximant have involved two tasks. For example, to show that the $n \times n$ normal matrix $\frac{1}{2}C$ is a nearest normal to S, we showed

1. $\|S - \frac{1}{2}C\| = \frac{1}{2}$.
2. For any normal matrix X, $\|S - X\| \geq \frac{1}{2}$ and hence $\frac{1}{2}$ is a lower bound on $\text{dist}(S, \mathcal{N})$.

Since the lower bound obtained for $\text{dist}(S, \mathcal{N})$ equals $\|S - \frac{1}{2}C\|$ we can conclude that $\frac{1}{2}C$ is a nearest normal to S. Similarly, when Phillips exhibited a nearest normal operator to a binormal operator [Phi77], his lower bound on the distance to the normal operators was equal to the distance to his particular normal operator.

In this 3×3 case, we conjecture that N is a nearest normal to $T = S + aS^2$ for some values of a other than 0. In particular, for values of a near 0. The fact that we have been able to obtain a formula for $\|T - N\|$ is encouraging. Unfortunately, the best lower bound that we have been able to obtain for $\text{dist}(T, \mathcal{N})$ is given by the following proposition.

Proposition 5 *If* $a \in \mathbf{C}$ *and* $T = \begin{bmatrix} 0 & 1 & a \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$, *then* $\text{dist}(T, \mathcal{N}) \geq \frac{1}{2}\sqrt{1 + |a|^2}$.

Proof Let $\{ e_1, e_2, e_3 \}$ denote the standard ordered basis for $\mathbf{C}^3$. Using Holmes' distance estimate from Section 1 2 2,

$$\begin{aligned} \text{dist}(T, \mathcal{N}) &\geq \tfrac{1}{2} \sup_{\|x\|=1} | \, \|Tx\| - \|T^*x\| \, | \\ &\geq \tfrac{1}{2} | \, \|Te_3\| - \|T^*e_3\| \, | \\ &= \tfrac{1}{2} \left| \sqrt{|a|^2 + 1 + 0} - \sqrt{0 + 0 + 0} \right| \\ &= \tfrac{1}{2}\sqrt{1 + |a|^2} \quad \blacksquare \end{aligned}$$

However, by Proposition 1, if $a = b + ic$ where $b, c \in \mathbf{R}$, then the distance from T to N is given by

$$\|T - N\| = \tfrac{1}{2} \max\left\{ \sqrt{1 + |a|^2 - 2b}, \sqrt{1 + |a|^2 + b + \sqrt{3}\,|c|} \right\}$$

Therefore, whenever $a \neq 0$, $\frac{1}{2}\sqrt{1 + |a|^2}$ is strictly less than $\|T - N\|$ and hence we are unable to use this lower bound to conclude that N is a nearest normal to T for any values of a other than 0

In fact, finding a nearest normal to the 3×3 Toeplitz matrix $T = S + aS^2$ has eluded us for all values of $a \neq 0$ Although our efforts have not produced a nearest normal when $a \neq 0$, they have produced a simple characterization of 3×3 normal matrices which are constant on the main diagonal Besides being interesting in its own right, this characterization immediately yields an easy way to generate normal matrices to act as candidates for nearest normals To finish off our discussion in this 3×3 case, we shall develop this characterization, and then give some examples to indicate how it may be useful in future research

In attempting to either raise the lower bound on $\operatorname{dist}(T, \mathcal{N})$ or to find new candidates for a nearest normal to T, the problem of characterizing 3×3 normal matrices came up It turned out that if we only used the fact that an arbitrary 3×3 normal matrix X has $\|\mathrm{col}_j(X)\| = \|\mathrm{row}_j(X)\|$ for $j = 1, 2, 3$, as we had done for superdiagonal matrices and Phillips had done for binormal matrices, then we could only show that $\|T - X\| \geq \frac{1}{2}\max\{1, |a|\}$ Since this is not as good as the lower bound obtained in Proposition 5 using Holmes' distance estimate, we were led to the problem of learning more about 3×3 normal matrices in the hope that we would be able to refine our estimates of $\|T - X\|$ or find new candidates for a nearest normal

We have been unable to obtain a useful characterization of arbitrary 3×3 normal matrices However, in a search for nearest normals to 3×3 Toeplitz matrices, it is compelling to consider 3×3 normal matrices which agree with the Toeplitz matrix

on the main diagonal. Our main characterization result about normal 3×3 matrices is a simple characterization of 3×3 normal matrices which are constant on the main diagonal. Motivated by the fact that our original candidates N and H, are actually 3×3 Toeplitz matrices, we also specify the corresponding characterization of normal 3×3 Toeplitz matrices. In order to assist in the proof of these characterizations, we begin with the following lemma.

Lemma 6 *Suppose a, b, α, β are complex numbers, then*

$$(*) \quad |a|^2 + |b|^2 = |\alpha|^2 + |\beta|^2 \quad \textit{and} \quad \overline{a}b = \overline{\alpha}\beta$$

if and only if

$$\begin{array}{ll} & (1) \quad a = \alpha = 0 \textit{ and } |b| = |\beta| \\ \textit{or} & \\ & (2) \quad b = \lambda\overline{\alpha} \textit{ and } \beta = \lambda\overline{a} \textit{ where } |a| = |\alpha| \textit{ or } |\lambda| = 1 \end{array}$$

Proof We first show that (1) $\Rightarrow$ $(*)$. By (1), $a = \alpha = 0$, so $\overline{a}b - \overline{\alpha}\beta = \overline{0}b - \overline{0}\beta = 0$ and $|a|^2 + |b|^2 - |\alpha|^2 - |\beta|^2 = |0|^2 + |b|^2 - |0|^2 - |\beta|^2 = |b|^2 - |\beta|^2$. However, (1) also gives that $|b| = |\beta|$, so $|b|^2 - |\beta|^2 = 0$. Hence (1) $\Rightarrow$ $(*)$.

Secondly, we show that (2) $\Rightarrow$ $(*)$. By (2), $b = \lambda\overline{\alpha}$ and $\beta = \lambda\overline{a}$, so $\overline{a}b - \overline{\alpha}\beta = \overline{a}\lambda\overline{\alpha} - \overline{\alpha}\lambda\overline{a} = 0$ and $|a|^2 + |b|^2 - |\alpha|^2 - |\beta|^2 = |a|^2 + |\lambda|^2|\alpha|^2 - |\alpha|^2 - |\lambda|^2|a|^2 = (|a|^2 - |\alpha|^2)(1 - |\lambda|^2)$. However, (2) also gives that $|a| = |\alpha|$ or $|\lambda| = 1$, so $(|a|^2 - |\alpha|^2)(1 - |\lambda|^2) = 0$. Hence (2) $\Rightarrow$ $(*)$.

Finally, we show that $(*)$ $\Rightarrow$ (1) or (2) by cases.

If $a = \alpha = 0$, then $(*)$ implies $|0|^2 + |b|^2 = |0|^2 + |\beta|^2$ and $\overline{0}b = \overline{0}\beta$ which simply says $|b| = |\beta|$. Hence, in this $a = \alpha = 0$ case, we have $(*)$ $\Rightarrow$ (1).

In case $a \neq 0$, we get that $\lambda = \frac{\beta}{\overline{a}}$ is well defined and then $\beta = \lambda\overline{a}$. By $(*)$, we know $\overline{a}b = \overline{\alpha}\beta$ and so $b = \frac{\beta}{\overline{a}}\overline{\alpha} = \lambda\overline{\alpha}$. But then $(*)$ also gives that $0 = |a|^2 + |b|^2 - |\alpha|^2 - |\beta|^2 = |a|^2 + |\lambda|^2|\alpha|^2 - |\alpha|^2 - |\lambda|^2|a|^2 = (|a|^2 - |\alpha|^2)(1 - |\lambda|^2)$ and so $|a| = |\alpha|$ or $|\lambda| = 1$. Hence, in this $a \neq 0$ case, we have $(*)$ $\Rightarrow$ (2).

The only remaining case is the $\alpha \neq 0$ case, however, by the symmetry of a and α in $(*)$ and in (2), if $\alpha \neq 0$, then we also have $(*) \Rightarrow$ (2) ∎

Proposition 7. *Suppose*

$$M = \begin{bmatrix} z & a_{12} & b_{13} \\ b_{21} & z & a_{23} \\ a_{31} & b_{32} & z \end{bmatrix} = zI + \begin{bmatrix} 0 & a_{12} & b_{13} \\ b_{21} & 0 & a_{23} \\ a_{31} & b_{32} & 0 \end{bmatrix} = zI + X$$

is a 3×3 *matrix that is constant on the main diagonal, then* M *is normal if and only if* $X = M - zI$ *has one of the following 2 forms*

$$(1) \quad X = \begin{bmatrix} 0 & 0 & b_{13} \\ b_{21} & 0 & 0 \\ 0 & b_{32} & 0 \end{bmatrix} \qquad \text{where } |b_{13}| = |b_{21}| = |b_{32}|$$

or

$$(2) \quad X = \begin{bmatrix} 0 & a_{12} & \lambda\overline{a}_{31} \\ \lambda\overline{a}_{12} & 0 & a_{23} \\ a_{31} & \lambda\overline{a}_{23} & 0 \end{bmatrix} \qquad \text{where } |a_{12}| = |a_{23}| = |a_{31}| \text{ or } |\lambda| = 1$$

Proof Since the scalar matrix zI commutes with all matrices, we have that M is normal if and only if $X = M - zI$ is normal. By definition, X is normal if and only if $X^*X = XX^*$, so we compute

$$X^*X = \begin{bmatrix} 0 & \overline{b}_{21} & \overline{a}_{31} \\ \overline{a}_{12} & 0 & \overline{b}_{32} \\ \overline{b}_{13} & \overline{a}_{23} & 0 \end{bmatrix} \begin{bmatrix} 0 & a_{12} & b_{13} \\ b_{21} & 0 & a_{23} \\ a_{31} & b_{32} & 0 \end{bmatrix}$$

$$= \begin{bmatrix} |a_{31}|^2 + |b_{21}|^2 & \overline{a}_{31}b_{32} & a_{23}\overline{b}_{21} \\ a_{31}\overline{b}_{32} & |a_{12}|^2 + |b_{32}|^2 & \overline{a}_{12}b_{13} \\ \overline{a}_{23}b_{21} & a_{12}\overline{b}_{13} & |a_{23}|^2 + |b_{13}|^2 \end{bmatrix}$$

$$XX^* = \begin{bmatrix} 0 & a_{12} & b_{13} \\ b_{21} & 0 & a_{23} \\ a_{31} & b_{32} & 0 \end{bmatrix} \begin{bmatrix} 0 & \overline{b}_{21} & \overline{a}_{31} \\ \overline{a}_{12} & 0 & \overline{b}_{32} \\ \overline{b}_{13} & \overline{a}_{23} & 0 \end{bmatrix}$$

$$= \begin{bmatrix} |a_{12}|^2 + |b_{13}|^2 & \overline{a}_{23}b_{13} & a_{12}\overline{b}_{32} \\ a_{23}\overline{b}_{13} & |a_{23}|^2 + |b_{21}|^2 & \overline{a}_{31}b_{21} \\ \overline{a}_{12}b_{32} & a_{31}\overline{b}_{21} & |a_{31}|^2 + |b_{32}|^2 \end{bmatrix}$$

Since X^*X and XX^* are both self-adjoint, we get that X is normal if and only if X^*X and XX^* agree in the 6 entries (1,1), (2,2), (3,3), (2,3), (3,1) and (1,2) We arrange these 6 conditions into 3 groups of 2 conditions as follows

$$(G_1) \qquad |a_{31}|^2 + |b_{21}|^2 = |a_{12}|^2 + |b_{13}|^2 \quad \text{and} \quad \overline{a}_{31}b_{21} = \overline{a}_{12}b_{13}$$

$$(G_2) \qquad |a_{12}|^2 + |b_{32}|^2 = |a_{23}|^2 + |b_{21}|^2 \quad \text{and} \quad \overline{a}_{12}b_{32} = \overline{a}_{23}b_{21}$$

$$(G_3) \qquad |a_{23}|^2 + |b_{13}|^2 = |a_{31}|^2 + |b_{32}|^2 \quad \text{and} \quad \overline{a}_{23}b_{13} = \overline{a}_{31}b_{32}$$

Remark These 3 groups of conditions can be thought of as conditions on corresponding columns and rows of X More precisely, G_j says that $\|\mathrm{col}_j(X)\|^2 = \|\mathrm{row}_j(X)\|^2$ and a certain product in $\mathrm{col}_j(X)$ equals a certain product in $\mathrm{row}_j(X)$

By Lemma 6, these 3 groups are true if and only if the following 3 groups are true

$$(H_1) \qquad \begin{bmatrix} a_{31} = a_{12} = 0 \\ \text{and } |b_{21}| = |b_{13}| \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} b_{21} = \lambda_1\overline{a}_{12} \text{ and } b_{13} = \lambda_1\overline{a}_{31} \\ \text{where } |a_{31}| = |a_{12}| \text{ or } |\lambda_1| = 1 \end{bmatrix}$$

$$(H_2) \qquad \begin{bmatrix} a_{12} = a_{23} = 0 \\ \text{and } |b_{32}| = |b_{21}| \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} b_{32} = \lambda_2\overline{a}_{23} \text{ and } b_{21} = \lambda_2\overline{a}_{12} \\ \text{where } |a_{12}| = |a_{23}| \text{ or } |\lambda_2| = 1 \end{bmatrix}$$

$$(H_3) \qquad \begin{bmatrix} a_{23} = a_{31} = 0 \\ \text{and } |b_{13}| = |b_{32}| \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} b_{13} = \lambda_3\overline{a}_{31} \text{ and } b_{32} = \lambda_3\overline{a}_{23} \\ \text{where } |a_{23}| = |a_{31}| \text{ or } |\lambda_3| = 1 \end{bmatrix}$$

We are now set to show that M is normal if and only if X has form (1) or (2)

We first show that if X has form (1) then M is normal If X has form (1), then $a_{12} = a_{23} = a_{31} = 0$ and $|b_{13}| = |b_{21}| = |b_{32}|$ So H_1, H_2 and H_3 are all true and hence M is normal

Next we show that if X has form (2) then M is normal. If X has form (2), then $b_{13} = \lambda\overline{a}_{31}$, $b_{21} = \lambda\overline{a}_{12}$ and $b_{32} = \lambda\overline{a}_{23}$ where $|a_{12}| = |a_{23}| = |a_{31}|$ or $|\lambda| = 1$. Again H_1, H_2 and H_3 are all true and hence M is normal.

Finally, we show that M being normal implies X has form (1) or (2), by cases.

In case $a_{12} = a_{23} = a_{31} = 0$, we use that M normal implies G_1, G_2 and G_3 are all true. By simply substituting $a_{12} = a_{23} = a_{31} = 0$ into G_1, G_2 and G_3 we get that $|b_{13}| = |b_{21}| = |b_{32}|$. Hence, in this $a_{12} = a_{23} = a_{31} = 0$ case, we have that M normal implies that X has form (1).

Otherwise, at least one of a_{12}, a_{23}, a_{31} is not 0. For example, if $a_{12} \neq 0$, then we use that M normal implies H_1, H_2 and H_3 are all true. Since $a_{12} \neq 0$ H_1 gives that $b_{21} = \lambda_1\overline{a}_{12}$ and $b_{13} = \lambda_1\overline{a}_{31}$ where $|a_{31}| = |a_{12}|$ or $|\lambda_1| = 1$, and H_2 gives that $b_{21} = \lambda_2\overline{a}_{12}$ and $b_{32} = \lambda_2\overline{a}_{23}$ where $|a_{12}| = |a_{23}|$ or $|\lambda_2| = 1$. Then, since $0 = b_{21} - b_{21} = (\lambda_1 - \lambda_2)\overline{a}_{12}$ and $a_{12} \neq 0$, we have that $\lambda_1 = \lambda_2$. Letting λ denote this common value, H_1 and H_2 now say $b_{13} = \lambda\overline{a}_{31}$, $b_{21} = \lambda\overline{a}_{12}$ and $b_{32} = \lambda\overline{a}_{23}$ where $|a_{31}| = |a_{12}| = |a_{23}|$ or $|\lambda| = 1$. Therefore, in this $a_{12} \neq 0$ case, we get that M normal implies X has form (2). In the cases where $a_{23} \neq 0$ or $a_{31} \neq 0$, similar proofs show that M normal implies X has form (2). ∎

Remark. Consider the following decomposition of M.

$$M = zI + \begin{bmatrix} 0 & a_{12} & 0 \\ 0 & 0 & a_{23} \\ a_{31} & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & b_{13} \\ b_{21} & 0 & 0 \\ 0 & b_{32} & 0 \end{bmatrix} = zI + A + B$$

By applying the proposition, observe that A is normal if and only if $|a_{31}| = |a_{12}| = |a_{23}|$ and B is normal if and only if $|b_{13}| = |b_{21}| = |b_{32}|$. If one knows these two facts, then the proposition can be remembered as follows: M is normal if and only if (1) $M - zI = B$ and B is normal, or (2) $M - zI = A + \lambda A^*$ where A is normal or $|\lambda| = 1$.

Since 3×3 Toeplitz matrices are constant on the main diagonal, the proposition applies to them as a special case.

Corollary 8.

$$\textit{The } 3 \times 3 \textit{ Toeplitz matrix } M = \begin{bmatrix} z & x & y \\ \eta & z & x \\ \xi & \eta & z \end{bmatrix} \textit{ is normal}$$

if and only if

$$(G_1) \quad |x|^2 + |y|^2 = |\xi|^2 + |\eta|^2 \textit{ and } \overline{x}y = \overline{\xi}\eta$$

if and only if

$$(1) \quad M - zI = \begin{bmatrix} 0 & 0 & y \\ \eta & 0 & 0 \\ 0 & \eta & 0 \end{bmatrix} \quad \textit{where } |y| = |\eta|$$

or

$$(2) \quad M - zI = \begin{bmatrix} 0 & x & \lambda\overline{\xi} \\ \lambda\overline{x} & 0 & x \\ \xi & \lambda\overline{x} & 0 \end{bmatrix} \quad \textit{where } |x| = |\xi| \textit{ or } |\lambda| = 1$$

Proof The fact that, M is normal if and only if $M - zI$ has form (1) or (2), is just the statement of the proposition applied to this Toeplitz M

The fact that, M is normal if and only if G_1 is true, can be obtained from the proof of the proposition Referring to the proof of the proposition, we get that M is normal if and only if the conditions G_1, G_2 and G_3 are all true However, because of the special form of this Toeplitz M, G_3 is equivalent to G_1 and G_2 is always satisfied ∎

Remark The significance of including the condition G_1 in the statement of this corollary is that it is a simple condition for deciding if a given 3×3 Toeplitz matrix is normal By contrast, given an arbitrary 3×3 matrix $M = zI + X$, which is constant on the main diagonal, checking all the conditions G_1, G_2 and G_3 is hardly better than computing $X^*X - XX^*$

To finish off our discussion of the 3×3 case, we give some examples to clarify where we stand and to indicate how these characterization results may be useful in future research

Examples For The 3×3 Case

The first two examples summarize the 3×3 upper triangular Toeplitz matrices

$$T = \begin{bmatrix} a_0 & a_1 & a_2 \\ 0 & a_0 & a_1 \\ 0 & 0 & a_0 \end{bmatrix} = a_0 I + a_1 S + a_2 S^2$$

for which we can exhibit a nearest normal approximant. In all examples, we consider our special "sum" of normal approximants

$$N = a_0 I + \tfrac{1}{2} a_1 C + \tfrac{1}{2} a_2 C^2 = \begin{bmatrix} a_0 & \frac{1}{2}a_1 & \frac{1}{2}a_2 \\ \frac{1}{2}a_2 & a_0 & \frac{1}{2}a_1 \\ \frac{1}{2}a_1 & \frac{1}{2}a_2 & a_0 \end{bmatrix}$$

and we also consider the real part of T

$$H = \tfrac{1}{2}(T + T^*) = \begin{bmatrix} a_0 & \frac{1}{2}a_1 & \frac{1}{2}a_2 \\ \frac{1}{2}\overline{a}_1 & a_0 & \frac{1}{2}a_1 \\ \frac{1}{2}\overline{a}_2 & \frac{1}{2}\overline{a}_1 & a_0 \end{bmatrix}$$

which we know is a nearest Hermitian matrix to T.

Example 1. If $a_2 = 0$ so that

$$T = \begin{bmatrix} a_0 & a_1 & 0 \\ 0 & a_0 & a_1 \\ 0 & 0 & a_0 \end{bmatrix}$$

then N is a nearest normal approximant to T with

$$\operatorname{dist}(T, \mathcal{N}) = \|T - N\| = \left\| \begin{bmatrix} 0 & \frac{1}{2}a_1 & 0 \\ 0 & 0 & \frac{1}{2}a_1 \\ -\frac{1}{2}a_1 & 0 & 0 \end{bmatrix} \right\| = \tfrac{1}{2}|a_1|$$

In this case

$$\|T - H\| = \left\| \begin{bmatrix} 0 & \frac{1}{2}a_1 & 0 \\ -\frac{1}{2}\overline{a}_1 & 0 & \frac{1}{2}a_1 \\ 0 & -\frac{1}{2}\overline{a}_1 & 0 \end{bmatrix} \right\| \geq \|\operatorname{col}_2(T - H)\| = \tfrac{1}{2}\sqrt{2}\,|a_1|$$

and hence H is not a nearest normal to T (except in the trivial $a_1 = 0$ case).

Example 2 If $a_1 = 0$ so that

$$T = \begin{bmatrix} a_0 & 0 & a_2 \\ 0 & a_0 & 0 \\ 0 & 0 & a_0 \end{bmatrix}$$

then N is a nearest normal approximant to T with

$$\text{dist}(T, \mathcal{N}) = \|T - N\| = \| \begin{bmatrix} 0 & 0 & \frac{1}{2}a_2 \\ -\frac{1}{2}a_2 & 0 & 0 \\ 0 & -\frac{1}{2}a_2 & 0 \end{bmatrix} \| = \tfrac{1}{2}|a_2|$$

In this case

$$\|T - H\| = \| \begin{bmatrix} 0 & 0 & \frac{1}{2}a_2 \\ 0 & 0 & 0 \\ -\frac{1}{2}\overline{a}_2 & 0 & 0 \end{bmatrix} \| = \tfrac{1}{2}|a_2|$$

and hence H is also a nearest normal to T

The remaining examples all involve the general case where $a_1 \neq 0$ and $a_2 \neq 0$ Replacing T with $\frac{1}{a_1}(T - a_0 I)$ if necessary, we can assume T has the deceptively simple looking form

$$T = \begin{bmatrix} 0 & 1 & a \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{where } a = \tfrac{a_2}{a_1} \neq 0$$

The goal of these remaining examples is to clarify what is stopping us from exhibiting a nearest normal in such cases and to indicate how our characterization of 3×3 normal matrices which are constant on the main diagonal may be useful in future research

Example 3 If $a = 1$ so that

$$T = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

then, in this particular case

$$N = H = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

Applying Corollary 2 or Proposition 3, we have

$$\|T - N\| = \|T - H\| = \tfrac{1}{2}\sqrt{1 + |1|^2 + 1} = \tfrac{1}{2}\sqrt{3}$$

Applying Proposition 5, we have

$$\operatorname{dist}(T, \mathcal{N}) \geq \tfrac{1}{2}\sqrt{1 + |1|^2} = \tfrac{1}{2}\sqrt{2}$$

and hence all we know for sure is that

$$\tfrac{1}{2}\sqrt{2} \leq \operatorname{dist}(T, \mathcal{N}) \leq \tfrac{1}{2}\sqrt{3}$$

In looking for nearer normals to this particular T, it is compelling to look for normals which agree with it on the main diagonal By Proposition 7, we only have to consider 2 different forms

If X is a normal matrix which has the form

$$X = \begin{bmatrix} 0 & 0 & b_1 \\ b_2 & 0 & 0 \\ 0 & b_3 & 0 \end{bmatrix} \quad \text{where } |b_1| = |b_2| = |b_3|$$

then

$$\|T - X\| = \left\| \begin{bmatrix} 0 & 1 & 1 - b_1 \\ -b_2 & 0 & 1 \\ 0 & -b_3 & 0 \end{bmatrix} \right\| \geq 1 > \tfrac{1}{2}\sqrt{3} = \|T - N\|$$

and hence there is no normal X of this form which is nearer to T than $N = H$

Is there a normal matrix of the constant main diagonal form

$$X = \begin{bmatrix} 0 & a_{12} & \lambda\overline{a}_{31} \\ \lambda\overline{a}_{12} & 0 & a_{23} \\ a_{31} & \lambda\overline{a}_{23} & 0 \end{bmatrix} \quad \text{where } |a_{12}| = |a_{23}| = |a_{31}| \text{ or } |\lambda| = 1$$

such that $\|T - X\| < \|T - N\|$? Since X is a function of 4 variables, it is not surprising that we have not been able to obtain an algebraic expression for $\|T - X\|$, never mind our ultimate problem of finding how to minimize $\|T - X\|$

Of course, we can also use this form of X to help us generate normal matrices to act as candidates for nearest normals In an attempt to generate "reasonable" candidates for this Toeplitz T, it is tempting to try normal Toeplitz matrices of the form

$$X = \begin{bmatrix} 0 & x & \lambda\overline{\xi} \\ \lambda\overline{x} & 0 & x \\ \xi & \lambda\overline{x} & 0 \end{bmatrix} \quad \text{where } |x| = |\xi| \text{ or } |\lambda| = 1$$

Motivated by the fact that $N = H$ is of this form, with $x = \frac{1}{2}$ and $\lambda\overline{\xi} = \frac{1}{2}$, we have tried various numeric experiments with $x = \frac{1}{2}$ and $\lambda\overline{\xi} = \frac{1}{2}$, but they have not yielded any normal Toeplitz matrices nearer to T than $N = H$ Although these normal Toeplitz matrices are functions of 3 variables, and hence difficult to work with analytically, some analytic experimentation (*i e* arbitrarily adding constraints) has yielded a normal Toeplitz matrix nearer to T than $N = H$ We begin experimenting with $\xi = x$, then

$$X = \begin{bmatrix} 0 & x & \lambda\overline{x} \\ \lambda\overline{x} & 0 & x \\ x & \lambda\overline{x} & 0 \end{bmatrix} = \begin{bmatrix} 0 & x & y \\ y & 0 & x \\ x & y & 0 \end{bmatrix} = xC + yC^2 \in C^*(C)$$

and

$$T - X = \begin{bmatrix} 0 & 1-x & 1-y \\ -y & 0 & 1-x \\ -x & -y & 0 \end{bmatrix}$$

This difference of two Toeplitz matrices, $T - X$, is also a Toeplitz matrix If it is not normal, then we are forced to find $\|T - X\|$ using $\|T - X\|^2 = \|(T - X)^*(T - X)\|$ On the other hand, by Corollary 8,

$$\begin{aligned}
& T - X \text{ is normal} \\
\iff\ & \overline{x}y = (1-\overline{x})(1-y) \ \text{ and } \ |x|^2 + |y|^2 = |1-x|^2 + |1-y|^2 \\
\iff\ & \overline{x}y = 1 - \overline{x} - y + \overline{x}y \text{ and } |x|^2 + |y|^2 = 1 - 2\,\mathrm{Re}\ x + |x|^2 + 1 - 2\,\mathrm{Re}\ y + |y|^2 \\
\iff\ & \overline{x} + y = 1 \ \text{ and } \ \mathrm{Re}\ x + \mathrm{Re}\ y = 1 \\
\iff\ & \overline{x} + y = 1
\end{aligned}$$

By arbitrarily adding this constraint

$$T - X = \begin{bmatrix} 0 & 1-x & \overline{x} \\ -(1-\overline{x}) & 0 & 1-x \\ -x & -(1-\overline{x}) & 0 \end{bmatrix}$$

is a normal matrix and so $\|T - X\| = r(T - X)$. Let $p(\lambda)$ denote the characteristic polynomial of $T - X$, then

$$\begin{aligned} p(\lambda) &= \det[\lambda I - (T - X)] = \det \begin{bmatrix} \lambda & -(1-x) & -\overline{x} \\ (1-\overline{x}) & \lambda & -(1-x) \\ x & (1-\overline{x}) & \lambda \end{bmatrix} \\ &= (\lambda^3 + (1-x)^2 x - (1-\overline{x})^2 \overline{x}) - (-|x|^2 \lambda - |1-x|^2 \lambda - |1-x|^2 \lambda) \\ &= \lambda^3 + (|x|^2 + 2|1-x|^2)\lambda + [(1-x)^2 x - (1-\overline{x})^2 \overline{x}] \end{aligned}$$

For $x \in \mathbf{C}$, finding the roots of this cubic polynomial, $p(\lambda)$, is a computational problem. However, by arbitrarily adding the constraint that $x \in \mathbf{R}$, $\|T - X\| = \sqrt{x^2 + 2(1-x)^2} = \sqrt{3x^2 - 4x + 2}$. The quadratic function $f(x) = 3x^2 - 4x + 2$ attains its minimum value when $0 = f'(x) = 6x - 4$. Hence, the minimum value of $\|T - X\|$ over $\{ X = xC + yC^2 : \overline{x} + y = 1 \text{ and } x \in \mathbf{R} \}$, occurs when $x = \frac{2}{3}$ and is

$$\|T - X\| = \sqrt{f(\tfrac{2}{3})} = \sqrt{\tfrac{2}{3}} \approx 0.8165$$

Recall that, $\|T - N\| = \frac{1}{2}\sqrt{3} = \sqrt{\frac{3}{4}} \approx 0.8660$, and hence when $x = \frac{2}{3}$ and $y = 1 - \overline{x} = \frac{1}{3}$ we have that the normal Toeplitz matrix

$$X_0 = \tfrac{2}{3}C + \tfrac{1}{3}C^2 = \begin{bmatrix} 0 & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} & 0 \end{bmatrix}$$

is nearer to T than $N = H$. However, $\|T - X_0\|$ is still greater than $\frac{1}{2}\sqrt{2} = \sqrt{\frac{1}{2}} \approx 0.7071$, our best known lower bound on $\operatorname{dist}(T, \mathcal{N})$, and hence we cannot declare X_0 as a nearest normal to T.

Remark 1 In general, when we try to analytically broaden our "reasonable" search for new candidates, we run into computational difficulties beyond our abilities Even in this particular example, if we dropped our last constraint that $x \in \mathbf{R}$, then we are faced with finding the roots of the nontrivial cubic characteristic polynomial $p(\lambda)$ in terms of the single complex variable x, followed by minimizing their maximum modulus Computations get even worse if we drop the constraint that $\overline{x} + y = 1$ since then $T - X$ is not necessarily normal and we are faced with finding the roots of the cubic characteristic polynomial of $(T-X)^*(T-X)$ in terms of the 2 complex variables x and y

Remark 2 It is interesting that, with respect to the ℓ_2 norm, defined for $n \times n$ matrices $A = [a_{ij}]$ by

$$\|A\|_{\ell_2} = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2},$$

we have that $X_0 = \frac{2}{3}C + \frac{1}{3}C^2$ is the ℓ_2-nearest matrix to $T = S + S^2$, of the form $X = xC + yC^2$ with $x, y \in \mathbf{R}$ This is true since

$$\begin{aligned} \|T - X\|_{\ell_2}^2 &= \| \begin{bmatrix} 0 & 1-x & 1-y \\ -y & 0 & 1-x \\ -x & -y & 0 \end{bmatrix} \|_{\ell_2}^2 \\ &= [2(1-x)^2 + x^2] + [(1-y)^2 + 2y^2] \\ &= [3x^2 - 4x + 2] + [3y^2 - 2y + 1] = f(x) + g(y) \end{aligned}$$

is minimized when $0 = f'(x) = 6x - 4$ and $0 = g'(y) = 6y - 2$ That is, when $x = \frac{2}{3}$ and $y = \frac{1}{3}$

Remark 3 In fact, this X_0 is actually the ℓ_2-nearest matrix to $T = S + S^2$ in $C^*(C)$ In general, if U is unitary in M_n with $U^n = \omega I$, then the map $E \quad M_n \mapsto M_n$ defined for $A \in M_n$ by

$$E(A) = \frac{1}{n} \sum_{k=1}^{n} U^k A U^{-k}$$

maps A into the commutant of the commutative $C^*(U)$ However, if $\sigma(U)$ has n distinct points, then $C^*(U)$ is unitarily equivalent to D_n, the subalgebra of diagonal

matrices in M_n. Since D_n is its own commutant, we have that $C^*(U)$ is its own commutant and that E maps M_n into $C^*(U)$. However, for every $A \in C^*(U)$, $E(A) = A$ and so E maps M_n onto $C^*(U)$. Thinking of M_n simply as a Hilbert space, with inner product

$$\langle A, B \rangle = \text{trace}\,(B^*A) \quad \text{for all } A, B \text{ in } M_n,$$

we have that the Hilbert space norm is the ℓ_2 norm, and that $E^* = E$ since $\langle E(A), B\rangle = \langle A, E(B)\rangle$ for all A, B in M_n. We also have, $E^2 = E$ and so E is the othogonal projection onto $C^*(U)$. It follows that, in the Hilbert space norm of M_n, $E(A)$ is the nearest matrix to A in $C^*(U)$. In other words, $E(A)$ is the ℓ_2-nearest matrix to A in $C^*(U)$.

In our case, $T = S + S^2$ is in M_3 and C is a unitary in M_3 satisfying $C^3 = I$ and having 3 distinct points in its spectrum. Therefore, $E(T) = \frac{1}{3}\sum_{k=1}^{3} C^k T C^{-k}$ is the ℓ_2-nearest matrix to T in $C^*(C)$. By computing $E(T)$, we obtain $E(T) = \frac{2}{3}C + \frac{1}{3}C^2 = X_0$. Alternatively, since C has real-valued entries and this T has only real-valued entries, we know $E(T)$ will be of the form $xC + yC^2$ where $x, y \in \mathbf{R}$ and hence $E(T)$ must be X_0.

The next example shows that for $T = S + aS^2$, the ℓ_2-nearest matrix to T in $C^*(C)$ is not always better than N and H.

Example 4 If $a \in [-1, 2]$ and

$$T = \begin{bmatrix} 0 & 1 & a \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

then, by Proposition 4, we know how N and H compare to one another as normal approximants for T.

How does $E(T)$, the ℓ_2-nearest matrix to T in $C^*(C)$, compare to N and H over this interval? Computing $E(T) = \frac{1}{3}\sum_{k=1}^{3} C^k T C^{-k}$ we obtain $E(T) = \frac{2}{3}C + \frac{1}{3}aC^2$ as our "ℓ_2-candidate". An analytic comparison is computationally difficult since it

turns out that $T - E(T)$ is normal if and only if $|a| = 1$. However, a numerical comparison is enough to verify that $E(T)$ is not always a better normal approximant than N and H.

Table 2.1 on page 73 was obtained using *Matlab*. It numerically analyzes $E(T)$, N and H as normal approximants for $T = S + aS^2$ by tabulating $\|T - E(T)\|$, $\|T - N\|$, $\|T - H\|$ and $\frac{1}{2}\sqrt{1 + |a|^2}$ (our greatest known lower bound on $\operatorname{dist}(T, \mathcal{N})$) for some values of $a \in [-1, 2]$.

For the values of a tabulated, observe that $E(T)$ is a better normal approximant than N and H only for $a = -0.6, -0.5, 0.8, 0.9, 1.0, 1.1$.

This concludes our examples for the 3×3 case.

Remark. These 4 examples are representative of our results for the 3×3 case to this point in time (July 1995).

Future research into the 3×3 case may be able to use our characterization of 3×3 normal matrices which are constant on the main diagonal to generate better normal approximants.

Future research may also be able to generate better normal approximants by replacing the unitary C, with some other unitary U (satisfying $U^3 = \omega I$ and having 3 distinct points in its spectrum), and then looking at the Hilbert space projection from M_3 onto $C^*(U)$. An obvious place to start would be with other unitary circulants. In particular, the special unitary circulants which have 1's on the first superdiagonal, ω in the bottom left entry and 0's elsewhere are natural candidates.

This second approach is the more appealing to try first since the theory is already in place to extend the method to the $n \times n$ case. In contrast, getting a characterization of normal $n \times n$ matrices which are constant on the main diagonal does not seem likely.

a	$\|T - E(T)\|$	$\|T - N\|$	$\|T - H\|$	$\frac{1}{2}\sqrt{1 + \|a\|^2}$
-1 0	0 9107	1 0000	**0 8660**	0 7071
-0 9	0 8667	0 9500	**0 8382**	0 6727
-0 8	0 8260	0 9000	**0 8124**	0 6403
-0 7	0 7891	0 8500	**0 7890**	0 6103
-0 6	**0 7566**	0 8000	0 7681	0 5831
-0 5	**0 7287**	0 7500	0 7500	0 5590
-0 4	0 7059	**0.7000**	0 7348	0 5385
-0 3	0 6884	**0 6500**	0 7228	0 5220
-0 2	0 6761	**0 6000**	0 7141	0 5099
-0 1	0 6690	**0 5500**	0 7089	0 5025
0 0	0 6667	**0 5000**	0 7071	0 5000
0 1	0 6688	**0.5268**	0 7089	0 5025
0 2	0 6749	**0 5568**	0 7141	0 5099
0 3	0 6846	**0 5895**	0 7228	0 5220
0 4	0 6972	**0 6245**	0 7348	0 5385
0 5	0 7125	**0 6614**	0 7500	0 5590
0 6	0 7300	**0 7000**	0 7681	0 5831
0 7	0 7494	**0 7399**	0 7890	0 6103
0 8	**0 7704**	0 7810	0 8124	0 6403
0 9	**0 7929**	0 8231	0 8382	0 6727
1 0	**0 8165**	0 8660	0 8660	0 7071
1 1	**0 8745**	0 9097	0 8958	0 7433
1 2	0 9333	0 9539	**0 9274**	0 7810
1 3	0 9930	0 9987	**0 9605**	0 8201
1 4	1 0532	1 0440	**0 9950**	0 8602
1 5	1 1141	1 0897	**1 0308**	0 9014
1 6	1 1755	1 1358	**1 0677**	0 9434
1 7	1 2373	1 1822	**1 1057**	0 9862
1 8	1 2995	1 2288	**1 1446**	1 0296
1 9	1 3621	1 2757	**1 1843**	1 0735
2 0	1 4250	1 3229	**1.2247**	1 1180

Table 2 1 A numerical comparison of 3 normal approximants for $T = S + aS^2$ in M_3 For each value of a, the smallest norm is printed in boldface

2 2 4 The $n \times n$ Case For $n \geq 4$

What can we say about the normal approximation problem for arbitrary $n \times n$ upper triangular Toeplitz matrices when $n \geq 4$?

Let S denote the $n \times n$ basic superdiagonal matrix. Then the general $n \times n$ upper triangular Toeplitz matrix is given by $T = a_0 I + a_1 S + \cdots + a_{n-1} S^{n-1}$.

Let C denote the $n \times n$ basic circulant matrix. Then the normal matrix $N = a_0 I + \frac{1}{2} a_1 C + \cdots + \frac{1}{2} a_{n-1} C^{n-1}$ is a nearest normal to T if at most one of $a_1, \ldots, a_{n-1}$ is nonzero. In that case, if at most a_k is nonzero, then $\|T - N\| = \operatorname{dist}(T, \mathcal{N}) = \frac{1}{2}|a_k| = \frac{1}{2}\sqrt{|a_1|^2 + \cdots + |a_{n-1}|^2}$. In any case, by formula (2 8) we always have

$$\|T - N\| = \max_{\lambda^n = -1} |\tfrac{1}{2} a_1 \lambda + \cdots + \tfrac{1}{2} a_{n-1} \lambda^{n-1}| \tag{2 15}$$

The normal matrix $H = \frac{1}{2}(T + T^*)$, the real part of T, is always a nearest Hermitian matrix to T. Moreover, $T - H$ is always normal (skew-Hermitian) and hence $\|T - H\| = r(T - H)$.

It follows that, we always have $\operatorname{dist}(T, \mathcal{N}) \leq \min\{\, \|T - N\|, \|T - H\| \,\}$. Moreover, we know that there are $n \times n$ upper triangular Toeplitz matrices for which N is a better normal approximant than H. For example, when $T = a_1 S$, we have $\|T - N\| = \frac{1}{2}|a_1|$ while $\|T - H\| \geq \|\mathrm{col}_2(T - H)\| = \frac{1}{2}\sqrt{2}\,|a_1|$.

Here is an easy-to-calculate lower bound on $\operatorname{dist}(T, \mathcal{N})$ which can be obtained from Holmes' lower bound [Section 1 2 2] using the fact that the normal approximation problem is invariant under scalar translation. Let $\{e_1, \ldots, e_n\}$ denote the standard ordered basis for $\mathbf{C}^n$. Then

$$\begin{aligned}
\operatorname{dist}(T, \mathcal{N}) &= \operatorname{dist}(T - a_0 I, \mathcal{N}) \\
&\geq \tfrac{1}{2} \sup_{\|x\|=1} |\, \|(T - a_0 I)x\| - \|(T - a_0 I)^* x\| \,| \\
&\geq \tfrac{1}{2} |\, \|(T - a_0 I)e_n\| - \|(T - a_0 I)^* e_n\| \,| \\
&= \tfrac{1}{2}\sqrt{|a_1|^2 + \cdots + |a_{n-1}|^2}
\end{aligned}$$

When at most one of $a_1, \quad , a_{n-1}$ is nonzero, we know that this lower bound equals $\text{dist}(T, \mathcal{N})$ Unfortunately, we do not know that this lower bound equals $\text{dist}(T, \mathcal{N})$ for every T If it does not equal $\text{dist}(T, \mathcal{N})$ for a particular T, then we will have to find $\text{dist}(T, \mathcal{N})$ in some other way if we ever hope to identify a nearest normal to that T At this point in time, the only way we can recognize that a normal matrix X is a nearest normal to T is if $\|T - X\|$ equals this lower bound Otherwise, all we can say is

$$\tfrac{1}{2}\sqrt{|a_1|^2 + \quad + |a_{n-1}|^2} \leq \text{dist}(T, \mathcal{N}) \leq \|T - X\| \tag{2 16}$$

In addition to N and H, another natural normal approximant to consider is $E(T) = E_C(T) = \frac{1}{n}\sum_{k=1}^{n} C^k T C^{-k}$ By Remark 3 on page 70, it is the ℓ_2-nearest matrix to T in $C^*(C)$ By the same remark, if U is any $n \times n$ unitary with $U^n = \omega I$ and n distinct points in its spectrum, then $E_U(T)$ is the ℓ_2-nearest matrix to T in $C^*(U)$ and hence we have a way to generate more normal approximants which are at least ℓ_2-related to T One hope for future research is that we may be able to obtain better normal approximants using unitaries other than C

Remark Based on our experience with the 3×3 case, it would seem that this is essentially all we can say about the normal approximation problem for arbitrary $n \times n$ upper triangular Toeplitz matrices when $n \geq 4$ However, the following 4×4 example suggests that whenever n is an even number, we will be able to describe additional, special classes of $n \times n$ upper triangular Toeplitz matrices, which allow certain pairs of $a_1, \quad , a_{n-1}$ to be nonzero and which have N as a best normal approximant

Special Classes within the 4×4 Case

Surprisingly, if $T = \begin{bmatrix} a_0 & a_1 & 0 & a_3 \\ 0 & a_0 & a_1 & 0 \\ 0 & 0 & a_0 & a_1 \\ 0 & 0 & 0 & a_0 \end{bmatrix}$ satisfies a simple condition, satisfied by

every such T in $M_4(\mathbf{R})$, then we can show that N is a best normal approximant

The general 4×4 upper triangular Toeplitz matrix is given by $T = a_0 I + a_1 S + a_2 S^2 + a_3 S^3$ Based on our experience with the 3×3 case, we did not expect N, H, or $E(T)$ to be identifiable as a best normal approximant[2] if more than one of a_1, a_2, a_3 are nonzero However, in the course of some real-valued numerical experiments to compare these approximants, we were surprised to discover that there was one case where we could numerically identify a best normal approximant To illustrate, Table 2 2 numerically analyzes $E(T)$, N and H as normal approximants for

$$T = \begin{bmatrix} 0 & 1 & 0 & a \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = S + aS^3$$

by tabulating their distances to T along with $\frac{1}{2}\sqrt{1+|a|^2}$ (our greatest known lower bound on $\operatorname{dist}(T, \mathcal{N})$) for some values of $a \in [-2, 2]$

a	$\|T - E(T)\|$	$\|T - N\|$	$\|T - H\|$	$\frac{1}{2}\sqrt{1+\|a\|^2}$
-2	1 5308	**1 1180**	1 3090	**1 1180**
-1	0 8090	**0 7071**	1 0000	**0 7071**
0	0 7500	**0 5000**	0 8090	**0 5000**
1	0 8090	**0 7071**	**0 7071**	**0 7071**
2	1 5308	**1 1180**	1 1514	**1 1180**

Table 2 2 A numerical analysis of 3 normal approximants for $T = S + aS^3$ in M_4 For each value of a, the smallest norm and the lower bound are printed in boldface

For the values of a tabulated in Table 2 2, observe that $\|T - N\|$ always numerically equals $\frac{1}{2}\sqrt{1+|a|^2}$ and hence N appears to be a nearest normal to T for those values of a The following lemma verifies that N actually is a nearest normal to T for every $a \in \mathbf{R}$

[2]In this 4×4 case, $E(T) = a_0 I + \frac{3}{4}a_1 C + \frac{1}{2}a_2 C^2 + \frac{1}{4}a_3 C^3$

Lemma 9 *In M_4, let $T = S + aS^3$ and $N = \frac{1}{2}C + \frac{1}{2}aC^3$ If $a \in \mathbf{R}$, then $\|T - N\| = \frac{1}{2}\sqrt{1 + |a|^2}$ and hence N is a nearest normal to T when $a \in \mathbf{R}$ In general, if $a = b + ic$ where $b, c \in \mathbf{R}$, then $\|T - N\| = \frac{1}{2}\sqrt{1 + |a|^2 + 2|c|}$*

Proof By equation 2 16, $\operatorname{dist}(T, \mathcal{N})$ is between $\frac{1}{2}\sqrt{1 + |a|^2}$ and $\|T - N\|$ By equation 2 15

$$\|T - N\| = \max_{\lambda^4 = -1} |\tfrac{1}{2}\lambda + \tfrac{1}{2}a\lambda^3| = \max_{\lambda^4 = -1} \tfrac{1}{2}|\lambda||1 + a\lambda^2| = \tfrac{1}{2} \max |1 \pm ai|$$

At this point it is clear that if $a \in \mathbf{R}$, then $\|T - N\| = \frac{1}{2}\sqrt{1 + |a|^2}$ and hence N is a nearest normal to T

In case $a = b + ic$ where $b, c \in \mathbf{R}$, $|1 \pm ai| = |1 \pm ib \mp c| = \sqrt{1 \pm 2c + c^2 + b^2}$ and hence, $\|T - N\| = \frac{1}{2} \max |1 \pm ai| = \frac{1}{2}\sqrt{1 + |a|^2 + 2|c|}$ ∎

Remark This proof suggests that whenever $n \geq 4$ is an even number, we will be able to describe special classes of $n \times n$ upper triangular Toeplitz matrices, which allow certain pairs of $a_1, \ldots, a_{n-1}$ to be nonzero and which have N as a best normal approximant For example in M_{2k} for $k \geq 2$, we can describe $(k - 1)$ such classes as follows For every $m \in \{1, \ldots, (k-1)\}$, If $T = S^m + aS^{m+k}$ and $a \in \mathbf{R}$, then N is a nearest normal approximant to T, since in all cases

$$\|T - N\| = \max_{\lambda^{2k} = -1} \tfrac{1}{2}|\lambda^m||1 + a\lambda^k| = \tfrac{1}{2} \max |1 \pm ai| = \tfrac{1}{2}\sqrt{1 + |a|^2}$$

The following proposition uses the lemma to describe a larger class of 4×4 upper triangular Toeplitz matrices for which N is a best normal approximant

Proposition 10 *If* $T = \begin{bmatrix} a_0 & a_1 & 0 & a_3 \\ 0 & a_0 & a_1 & 0 \\ 0 & 0 & a_0 & a_1 \\ 0 & 0 & 0 & a_0 \end{bmatrix} = a_0 I + a_1 S + a_3 S^3$ *and* $\overline{a}_1 a_3 \in \mathbf{R}$

(in particular, if T is in $M_4(\mathbf{R})$), then $N = a_0 I + \frac{1}{2}a_1 C + \frac{1}{2}a_3 C^3$ is a nearest normal to T with $\|T - N\| = \operatorname{dist}(T, \mathcal{N}) = \frac{1}{2}\sqrt{|a_1|^2 + |a_3|^2}$

Proof If $\overline{a}_1 a_3 = 0$, then at most one of a_1, a_3 is nonzero and so we know that N is a nearest normal to T and that $\operatorname{dist}(T, \mathcal{N}) = \frac{1}{2}\sqrt{|a_1|^2 + |a_3|^2}$

Otherwise, $a_1 \neq 0$ and so $a = \frac{a_3}{a_1} = \frac{\overline{a}_1 a_3}{|a_1|^2} \in \mathbf{R}$ Then

$$\begin{aligned}
\operatorname{dist}(T, \mathcal{N}) &= |a_1| \operatorname{dist}(\tfrac{1}{a_1}(T - a_0 I), \mathcal{N}) \\
&= |a_1| \operatorname{dist}(S + aS^3, \mathcal{N}) \\
&= |a_1| \, \|\tfrac{1}{a_1}(T - a_0 I) - (\tfrac{1}{2}C + \tfrac{1}{2}aC^3)\| \quad \text{(by the Lemma)} \\
&= |a_1| \, \|\tfrac{1}{a_1}[T - a_0 I - \tfrac{1}{2}a_1 C - \tfrac{1}{2}a_3 C^3]\| \\
&= \|T - N\|
\end{aligned}$$

It follows that N is a nearest normal to T and $\operatorname{dist}(T, \mathcal{N}) = |a_1| \operatorname{dist}(S + aS^3, \mathcal{N}) = |a_1| \, \frac{1}{2}\sqrt{1 + |a|^2} = \frac{1}{2}\sqrt{|a_1|^2 + |a_3|^2}$ ∎

Remark Similarly, whenever $k \geq 2$, we will be able to describe $(k-1)$ special classes of $2k \times 2k$ upper triangular Toeplitz matrices, which allow certain pairs of $a_1, \quad , a_{2k-1}$ to be nonzero and which have N as a best normal approximant More precisely, for every $m \in \{1, \quad , (k-1)\}$, if $T = a_0 I + aS^m + bS^{m+k}$ and $\overline{a}b \in \mathbf{R}$, then $N = a_0 I + \frac{1}{2}aC^m + \frac{1}{2}bC^{m+k}$ is a nearest normal approximant to T with $\|T - N\| = \operatorname{dist}(T, \mathcal{N}) = \frac{1}{2}\sqrt{|a|^2 + |b|^2}$

This completes our discussion of the $n \times n$ case for $n \geq 4$ and brings us to the end of our discussion of the normal approximation problem for upper triangular Toeplitz matrices

We conclude this section with a brief summary of how our results can be used to attack the normal approximation problem for an arbitrary upper triangular Toeplitz matrix

2 2 5 Brief Summary of Results

Suppose $n \geq 2$ and $T = a_0 I + a_1 S + \cdots + a_{n-1} S^{n-1}$ is an arbitrary $n \times n$ upper triangular Toeplitz matrix

In this section, we have shown that if at most one of $a_1, \ldots, a_{n-1}$ are nonzero (this includes all 2×2 cases, all superdiagonal matrices, and all scalar translated superdiagonal matrices including Jordan blocks) or if $n = 2k \geq 4$ and exactly one of $(a_1, a_{1+k}), \ldots, (a_{k-1}, a_{k-1+k})$ is not $(0,0)$ and that (a_m, a_{m+k}) satisfies $\overline{a}_m a_{m+k} \in \mathbf{R}$, then

$$N = a_0 I + \tfrac{1}{2} a_1 C + \cdots + \tfrac{1}{2} a_{n-1} C^{n-1}$$

is a nearest normal to T with

$$\|T - N\| = \operatorname{dist}(T, \mathcal{N}) = \tfrac{1}{2}\sqrt{|a_1|^2 + \cdots + |a_{n-1}|^2}$$

For arbitrary $a_1, \ldots, a_{n-1}$, we have that

$$\tfrac{1}{2}\sqrt{|a_1|^2 + \cdots + |a_{n-1}|^2} \leq \operatorname{dist}(T, \mathcal{N}) \leq \|T - N\|$$

Moreover, in cases where $\|T - N\| \neq \tfrac{1}{2}\sqrt{|a_1|^2 + \cdots + |a_{n-1}|^2}$, we have begun developing a list of reasonable normal approximants to help close in on $\operatorname{dist}(T, \mathcal{N})$ from above. Using $H = \frac{1}{2}(T + T^*)$, we get Holmes' upper bound on $\operatorname{dist}(T, \mathcal{N})$ and we have seen cases where H is nearer to T than N. We also include $E(T) = E_C(T) = \frac{1}{n}\sum_{k=1}^{n} C^k T C^{-k}$, the ℓ_2-nearest matrix to T in $C^*(C)$, since we have found some cases where $E(T)$ is nearer to T than either N or H. One hope for future research is to replace C with some other unitary U, satisfying $U^n = \omega I$ and having n distinct points in its spectrum, and for which there are cases where $E_U(T) \in C^*(U)$ is nearer to T than N, H or $E(T)$.

This completes our discussion of the normal approximation problem for upper triangular Toeplitz matrices.

As a final topic for this chapter, we include a section on the normal approximation problem for direct sums of such matrices.

2 3 Direct Sums of Upper Triangular Toeplitz Matrices

In this section, we consider the normal approximation problem for direct sums of upper triangular Toeplitz matrices (even though such direct sums are not Toeplitz matrices) Here is the fundamental question

> *If A is a direct sum of upper triangular Toeplitz matrices, will a direct sum of corresponding best normal approximants be a best normal approximant for A?*

As mentioned when we discussed Jordan blocks in Section 2 2 2, if we restrict ourselves to a direct sum of Jordan blocks, then the answer is *yes* However, when we look at a direct sum of arbitrary upper triangular Toeplitz matrices, it seems like there is absolutely no logical reason why the answer should always be *yes* On the other hand, we have not been able to come up with a counterexample In fact, our search for a counterexample has led us to our main results for this section If we restrict ourselves to only taking direct sums of upper triangular Toeplitz matrices for which we have found best normal approximants in Section 2 2, then a direct sum of corresponding best normal approximants is a best normal approximant

We actually prove this result for a more general class of matrices Recall that, if $n \geq 2$ and $T = a_0 I + a_1 S + \cdots + a_{n-1} S^{n-1}$ is any of the Toeplitz matrices for which we have found a best normal approximant in Section 2 2, then $N = a_0 I + \frac{1}{2} a_1 C + \cdots + \frac{1}{2} a_{n-1} C^{n-1}$ is a best normal approximant, and

$$\operatorname{dist}(T, \mathcal{N}) = \tfrac{1}{2} \sqrt{|a_1|^2 + \cdots + |a_{n-1}|^2} \tag{2 17}$$

In generalizing this distance condition so that it will be less entry-dependent, we actually obtain a distance condition that is less Toeplitz-dependent, but which definitely includes all the matrices for which we have found best normal approximants in Section 2 2 Our main results are formulated in terms of these more general ma-

trices and we use examples to indicate their implications for direct sums of upper triangular Toeplitz matrices

We begin by stating two well-known results which combine to give the norm of a finite direct sum of two or more matrices

Lemma 1 *If* $A \in M_n$ *and* $B \in M_k$, *then* $A \oplus B = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \in M_{n+k}$ *and* $\|A \oplus B\| = \max\{\|A\|, \|B\|\}$

Corollary 2 *For* $k \geq 2$, *if* $A_j \in M_{n_j}$ *for* $j = 1, \ldots, k$, *then*

$$\|A_1 \oplus \cdots \oplus A_k\| = \max\{\|A_j\|\}_{j=1}^{k}$$

Before going to our main results, we can get some insight into the problem by considering a direct sum of arbitrary square matrices and seeing what goes wrong if we try to show that a direct sum of corresponding best normal approximants is a best normal approximant For $j = 1, 2$, suppose B_j is a nearest normal to A_j in M_{n_j} Let $\mathcal{N}_{n_j}$ denote the set of normal matrices in M_{n_j} and let $\mathcal{N} = \mathcal{N}_{n_1+n_2}$

For $A = A_1 \oplus A_2$ and $B = B_1 \oplus B_2$, we have that B is normal and

$$\|A - B\| = \left\| \begin{bmatrix} A_1 - B_1 & 0 \\ 0 & A_2 - B_2 \end{bmatrix} \right\| = \max_{j=1,2} \|A_j - B_j\| = \max_{j=1,2} \operatorname{dist}(A_j, \mathcal{N}_{n_j})$$

At this point, we know that $\operatorname{dist}(A, \mathcal{N}) \leq \|A - B\|$ In order to show that B is a nearest normal to A, we must show that $\operatorname{dist}(A, \mathcal{N}) \geq \|A - B\|$ That turns out to be a problem

An obvious first approach is to take an arbitrary X in $\mathcal{N}$ and partition it to match our direct sum partition of A Using unit vectors $\begin{bmatrix} x \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ y \end{bmatrix}$ in $\mathbf{C}^{n_1} \oplus \mathbf{C}^{n_2}$, we can show that

$$\|A - X\| = \left\| \begin{bmatrix} A_1 - X_{11} & -X_{12} \\ -X_{21} & A_2 - X_{22} \end{bmatrix} \right\| \geq \max_{j=1,2} \|A_j - X_{jj}\|$$

If we knew the X_{jj} were normal, then we would have $\|A_j - X_{jj}\| \geq \operatorname{dist}(A_j, \mathcal{N}_{n_j})$, and that would be sufficient to yield $\|A - X\| \geq \|A - B\|$ However, that is too

much to ask X normal does not imply X_{11} and X_{22} are normal For a counterexample, take X to be the unitary 4×4 basic circulant matrix, and partition it into 2×2 blocks, then $X_{11} = X_{22} = \left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ is not normal In fact, it seems unlikely that in the general case, B would always be a nearest normal to A In an attempt to find a normal X nearer to A than B, we restricted A to be a direct sum of upper triangular Toeplitz matrices for which *we* know best normal approximants (so that we would know $\|A - B\|$) However, in our chosen attempts we were always able to show that $\|A - X\| \geq \|A - B\|$ (by using $\|\mathrm{col}_j(X)\| = \|\mathrm{row}_j(X)\|$) and hence we always had that our direct sum of best normal approximants was a best normal approximant for A

Another possible approach for the general case is to try using Holmes' lower bound on $\mathrm{dist}(A, \mathcal{N})$ [Section 1 2 2] We begin by introducing some new notation related to Holmes' distance estimate

For every $T \in M_n$, we define

$$\begin{aligned} \mathrm{Holmes}(T, x) &= \tfrac{1}{2} \left| \|Tx\| - \|T^*x\| \right| \text{ for every } x \in \mathbf{C}^n \\ \mathrm{Holmes}(T) &= \sup_{\|x\|=1} \mathrm{Holmes}(T, x) \end{aligned}$$

With this new notation, $\mathrm{Holmes}(T)$ is Holmes' lower bound on $\mathrm{dist}(T, \mathcal{N})$ and for every unit vector x, $\mathrm{Holmes}(T, x)$ is a lower bound on $\mathrm{dist}(T, \mathcal{N})$

If we apply this to our $A = A_1 \oplus A_2$, we can use unit vectors $\left[\begin{smallmatrix} x \\ 0 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 0 \\ y \end{smallmatrix}\right]$ in $\mathbf{C}^{n_1} \oplus \mathbf{C}^{n_2}$ to show that

$$\mathrm{dist}(A, \mathcal{N}) \geq \mathrm{Holmes}(A) \geq \max_{j=1,2} \mathrm{Holmes}(A_j)$$

If we knew that $\mathrm{dist}(A_j, \mathcal{N}_{n_j}) = \mathrm{Holmes}(A_j)$ for $j = 1, 2$, that would be sufficient to yield $\mathrm{dist}(A, \mathcal{N}) \geq \max_{j=1,2} \mathrm{dist}(A_j, \mathcal{N}_{n_j}) = \|A - B\|$

This is too much to ask from an arbitrary square matrix In fact, we can take a Toeplitz matrix as a counterexample Let $T = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$, then $\mathrm{Holmes}(T) < 0\,23 < 0\,5 = \mathrm{dist}(T, \mathcal{N})$

In our first attempts to get a counterexample, we considered some of the simplest upper triangular Toeplitz matrices for which we know best normal approximants—quite naturally they had 0's on the main diagonal But every such sample T satisfied $\text{Holmes}(T) = \text{dist}(T, \mathcal{N})$ In fact, every such T will Every $T = a_0 I + a_1 S +$ $+ a_{n-1} S^{n-1}$ for which we know a best normal approximant satisfies

$$\text{dist}(T, \mathcal{N}) = \tfrac{1}{2}\sqrt{|a_1|^2 + \quad + |a_{n-1}|^2} = \text{Holmes}(T - a_0 I, e_n) \qquad (2\ 18)$$

where e_n is the nth standard basis vector Hence, when $a_0 = 0$, we have

$$\text{Holmes}(T) \leq \text{dist}(T, \mathcal{N}) = \text{Holmes}(T, e_n) \leq \text{Holmes}(T)$$

Hence, if A is a direct sum of such matrices, then B will be a best normal approximant for A Moreover, we can customize this second approach so that it at least includes *all* the matrices for which we have found best normal approximants in Section 2 2

We begin by formulating a more general distance condition satisfied by all Toeplitz matrices for which we know best normal approximants By equation (2 18), they all satisfy $\text{dist}(T, \mathcal{N}) = \text{Holmes}(T - a_0 I, e_n) \leq \text{Holmes}(T - a_0 I)$ However, we always have $\text{dist}(T, \mathcal{N}) = \text{dist}(T - a_0 I, \mathcal{N}) \geq \text{Holmes}(T - a_0 I)$ and hence, all Toeplitz matrices for which we know best normal approximants, satisfy $\text{dist}(T, \mathcal{N}) = \text{Holmes}(T - a_0 I)$ Since we want to include 1×1 matrices (Jordan blocks) in our applications, we observe that all 1×1 matrices trivially satisfy this condition (in fact, all normal matrices do)

The following three results do not assume matrices are Toeplitz, however, we will mainly apply them to Toeplitz matrices In fact all three results are formulated in terms of matrices satisfying the artificially constructed distance condition

$$\text{dist}(T, \mathcal{N}) = \text{Holmes}(T - \lambda I) \text{ for some } \lambda \in \mathbf{C} \qquad (2\ 19)$$

This condition is satisfied by all normal matrices and by all the Toeplitz matrices for which we know best normal approximants It is even satisfied by arbitrary 2×2

upper triangular matrices $T = \left[\begin{smallmatrix} a & b \\ 0 & c \end{smallmatrix}\right]$ since

$$\text{Holmes}(T - cI) = \text{Holmes}(T - cI, e_2) = \tfrac{1}{2}|b| = \text{dist}(T, \mathcal{N})$$

We make no further attempt to find matrices which satisfy it To help the reader get better acquainted with this artificial condition, we introduce each result with a Toeplitz example to which it applies As usual, in M_n, we use S to denote the basic superdiagonal matrix and C to denote the basic circulant matrix

Example Lemma 3 will allow us to say that a nearest normal to $A = (I + 2S) \oplus (S + S^2) \in M_2 \oplus M_3$ is $B = (I + C) \oplus (\frac{1}{2}C + \frac{1}{2}C^2)$, even though B_2 is not a nearest normal to A_2 *Proof* $\|A_1 - B_1\| = 1 = \text{Holmes}(A_1 - I)$ and $\|A_2 - B_2\| = \frac{1}{2}\sqrt{3} < 1$

Lemma 3 *For $j = 1, 2$, suppose A_j is a matrix in M_{n_j} and that B_j is a normal matrix in $\mathcal{N}_{n_j}$ Let $A = A_1 \oplus A_2$ and $B = B_1 \oplus B_2$ be their corresponding direct sums in M_n where $n = n_1 + n_2$*

If $\|A - B\| = \|A_1 - B_1\|$ and $\|A_1 - B_1\| = \text{Holmes}(A_1 - \lambda I_{n_1})$ for some $\lambda \in \mathbf{C}$, then B is a nearest normal to A

Remark Observe that the hypotheses require B_1 to be a nearest normal to A_1 and A_1 to satisfy a special distance condition, since

$$\begin{aligned} \text{Holmes}(A_1 - \lambda I_{n_1}) &\leq \text{dist}(A_1 - \lambda I_{n_1}, \mathcal{N}_{n_1}) = \text{dist}(A_1, \mathcal{N}_{n_1}) \\ &\leq \|A_1 - B_1\| = \text{Holmes}(A_1 - \lambda I_{n_1}) \end{aligned}$$

and hence $\text{dist}(A_1, \mathcal{N}_{n_1}) = \|A_1 - B_1\| = \text{Holmes}(A_1 - \lambda I_{n_1})$ Moreover, A_1 has a distinguished role in the direct sum since $\text{dist}(A_1, \mathcal{N}_{n_1}) \geq \|A_2 - B_2\| \geq \text{dist}(A_2, \mathcal{N}_{n_2})$ By the same equation, it is also clear that B_2 need not neccessarily be a nearest normal to A_2 but it cannot be arbitrarily far away either

Proof To see that B is normal, compute

$$B^*B - BB^* = (B_1^*B_1 \oplus B_2^*B_2) - (B_1B_1^* \oplus B_2B_2^*) = 0 \oplus 0 = 0$$

Therefore $\operatorname{dist}(A, \mathcal{N}_n) \leq \|A - B\| = \operatorname{Holmes}(A_1 - \lambda I_{n_1})$ by the hypotheses of the lemma

It remains to show that $\operatorname{dist}(A, \mathcal{N}_n) \geq \|A - B\| = \operatorname{Holmes}(A_1 - \lambda I_{n_1})$

However, $\operatorname{dist}(A, \mathcal{N}_n) = \operatorname{dist}(A - \lambda I_n, \mathcal{N}_n) \geq \operatorname{Holmes}(A - \lambda I_n)$ and hence, it will suffice to show that $\operatorname{Holmes}(A - \lambda I_n) \geq \operatorname{Holmes}(A_1 - \lambda I_{n_1})$

Now

$$A - \lambda I_n = \begin{bmatrix} A_1 - \lambda I_{n_1} & 0 \\ 0 & A_2 - \lambda I_{n_2} \end{bmatrix}$$

and so, for every $x \in \mathbf{C}^{n_1}$,

$$\|(A - \lambda I_n) \begin{bmatrix} x \\ 0 \end{bmatrix}\| = \| \begin{bmatrix} (A_1 - \lambda I_{n_1})x \\ 0 \end{bmatrix} \| = \|(A_1 - \lambda I_{n_1})x\|$$

and

$$\|(A - \lambda I_n)^* \begin{bmatrix} x \\ 0 \end{bmatrix}\| = \| \begin{bmatrix} (A_1 - \lambda I_{n_1})^* x \\ 0 \end{bmatrix} \| = \|(A_1 - \lambda I_{n_1})^* x\|$$

It follows that

$$\operatorname{Holmes}(A_1 - \lambda I_{n_1}) = \sup_{\|x\|=1} \operatorname{Holmes}(A - \lambda I_n, \begin{bmatrix} x \\ 0 \end{bmatrix}) \leq \operatorname{Holmes}(A - \lambda I_n)$$

This completes the proof that $\operatorname{dist}(A, \mathcal{N}_n) \geq \operatorname{Holmes}(A_1 - \lambda I_{n_1}) = \|A - B\|$

We now have $\operatorname{dist}(A, \mathcal{N}_n) = \|A - B\|$ and hence, B is a nearest normal to A ∎

The next proposition is a generalization of the lemma In preparation, we observe that since we know the direct sum of two normals is normal, we have, by induction, that any direct sum of two or more normals, $B = (B_1 \oplus \quad \oplus B_{k-1}) \oplus B_k$, is normal

Example Proposition 4 will allow us to say that a nearest normal to $A = (S + S^2) \oplus J_1(1) \oplus \sqrt{3}S \in M_3 \oplus M_1 \oplus M_3$ is $B = (\frac{1}{2}C + \frac{1}{2}C^2) \oplus J_1(1) \oplus \frac{1}{2}\sqrt{3}C$, even though B_1 is not a nearest normal to A_1 *Proof* $\|A_3 - B_3\| = \frac{1}{2}\sqrt{3} = \operatorname{Holmes}(A_3)$, $\|A_1 - B_1\| = \frac{1}{2}\sqrt{3}$ and $\|A_2 - B_2\| = 0$

Proposition 4 *For $j = 1, \ldots, k$ where $k \geq 2$, suppose A_j is a matrix in M_{n_j} and that B_j is a normal matrix in $\mathcal{N}_{n_j}$. Let $A = A_1 \oplus \cdots \oplus A_k$ and $B = B_1 \oplus \cdots \oplus B_k$ be their corresponding direct sums in M_n where $n = n_1 + \cdots + n_k$.*

If there exists an $i \in \{1, \ldots, k\}$ such that
$\|A - B\| = \|A_i - B_i\|$ *and* $\|A_i - B_i\| = \text{Holmes}(A_i - \lambda I_{n_i})$ *for some $\lambda \in \mathbf{C}$, then B is a nearest normal to A.*

Proof In case $i = 1$, $B = B_1 \oplus (B_2 \oplus \cdots \oplus B_k)$ is a direct sum of 2 normals which corresponds to $A = A_1 \oplus (A_2 \oplus \cdots \oplus A_k)$. Moreover, $\|A - B\| = \|A_1 - B_1\|$ and $\|A_1 - B_1\| = \text{Holmes}(A_1 - \lambda I_{n_1})$. By Lemma 3, B is a nearest normal to A.

In case $i \neq 1$, there exists a unitary U in M_n such that
$UAU^* = A_i \oplus \cdots \oplus A_k \oplus A_1 \oplus \cdots \oplus A_{i-1}$ and then UBU^* is a direct sum of normals which corresponds to UAU^*. Moreover $\|UAU^* - UBU^*\| = \|A - B\| = \|A_i - B_i\|$ and $\|A_i - B_i\| = \text{Holmes}(A_i - \lambda I_{n_i})$. However, A_i and B_i are the first matrices in the direct sums UAU^* and UBU^*, respectively. Therefore, by the $i = 1$ case, UBU^* is a nearest normal to UAU^*. However, the normal approximation problem is invariant under unitary equivalence and hence B is a nearest normal to A.

Therefore, in all cases, B is a nearest normal to A. ∎

Example Proposition 5 will allow us to say that if A is a direct sum and each summand is either a Toeplitz matrix for which we have found a best normal approximant in Section 2 2 or a normal matrix or a 2×2 upper triangular matrix, then any direct sum of corresponding best normal approximants will be a best normal approximant for A. *Proof* We have already observed that every such matrix satisfies the distance condition (2 19).

Proposition 5 *For $j = 1, \ldots, k$ where $k \geq 2$, suppose B_j is a nearest normal to A_j in M_{n_j}. Let $A = A_1 \oplus \cdots \oplus A_k$ and $B = B_1 \oplus \cdots \oplus B_k$ be their corresponding direct sums in M_n where $n = n_1 + \cdots + n_k$.*

If, for every $j \in \{1, \ldots, k\}$, A_j *satisfies* $\operatorname{dist}(A_j, \mathcal{N}_{n_j}) = \operatorname{Holmes}(A_j - \lambda_j I_{n_j})$, *for some* $\lambda_j \in \mathbf{C}$, *then* B *is a nearest normal to* A.

Proof By Lemma 1, we know
$\|A - B\| = \max\{\|A_j - B_j\|\}_{j=1}^{k} = \|A_i - B_i\|$ for some $i \in \{1, \ldots, k\}$. Since we are given that $\operatorname{dist}(A_i, \mathcal{N}_{n_i}) = \operatorname{Holmes}(A_i - \lambda_i I_{n_i})$, Proposition 4 yields that B is a nearest normal to A. ∎

Remark Although it is a little anticlimactic, we confirm our previous statements claiming that if A is a direct sum of Jordan blocks, then any direct sum of corresponding best normal approximants will be a best normal approximant for A. It is easy to see that as long as all the Jordan blocks in the direct sum are not 1×1, then $\operatorname{dist}(A, \mathcal{N}) = \frac{1}{2}$.

This concludes our consideration of the normal approximation problem for direct sums of upper triangular Toeplitz matrices.

It also concludes this chapter on the normal approximation problem for upper triangular Toeplitz matrices. In the next chapter we turn our attention to the normal approximation problem for Toeplitz operators.

Chapter 3

Normal Approximants for Toeplitz Operators

In this chapter, the problem of normal approximants for Toeplitz operators is studied. The main results in this chapter give upper and lower bounds on the distance from certain Toeplitz operators to the set of normal operators. In particular, if φ is a continuous complex-valued function on the unit circle $\mathbf{T}$, then the distance from T_φ to the normals is less than or equal to the radius of the smallest disk containing $\varphi(\mathbf{T})$ and if φ is also one-to-one (so that $\varphi(\mathbf{T})$ is a Jordan curve) then the distance from T_φ to the normals is greater than or equal to the radius of the largest disk contained *inside* $\varphi(\mathbf{T})$. In order to prove these results we need some known facts about Toeplitz operators and Fredholm operators. The facts we need about Fredholm operators can be found in Section 1.1.3. We compile the facts we need about Toeplitz operators in Section 3.1. Our distance estimates are presented in Section 3.2. As a final topic, we present a simple example (due to A. R. Sourour), which demonstrates that, if A is a direct sum of arbitrary Toeplitz operators, then a direct sum of corresponding best normal approximants is not neccessarily a best normal approximant for A.

3.1 Toeplitz Operators

Our main objective in this section is to compile some standard results about Toeplitz operators with continuous symbol (for use in the next section).

3 1.1 Preliminaries

We begin by briefly reminding the reader of the set up we gave in the Introduction The unit circle $\mathbf{T}$ is considered as a measure space with respect to normalized Lebesgue measure The set $\{ e_n \quad n \in \mathbf{Z} \}$ is the standard orthonormal basis for $L^2(\mathbf{T})$ ($i\ e$ for $n \in \mathbf{Z}$ and $z \in \mathbf{T}$, $e_n(z) = z^n$) The Hardy space $H^2(\mathbf{T})$ is the closed span of $\{ e_n \quad n \geq 0 \}$ and P is the projection of $L^2(\mathbf{T})$ onto $H^2(\mathbf{T})$

If $\varphi \in L^\infty(\mathbf{T})$, then the *Laurent operator* M_φ on $L^2(\mathbf{T})$ is just the multiplication operator defined by $M_\varphi f = \varphi f$ for f in $L^2(\mathbf{T})$ and the *Toeplitz operator* T_φ on $H^2(\mathbf{T})$ is defined by $T_\varphi f = P(M_\varphi f) = P(\varphi f)$ for f in $H^2(\mathbf{T})$ and hence is the compression of M_φ to $H^2(\mathbf{T})$

The bilaterally infinite matrix of M_φ with respect to the basis $\{ e_n \quad n \in \mathbf{Z} \}$ has a distinctive form—it is constant on all diagonals parallel to the main diagonal To illustrate, if we partition the basis via $\{ \quad , e_{-2}, e_{-1} \mid e_0, e_1, \quad \}$, then the corresponding partitioned matrix of M_φ is given by

$$M_\varphi = \left[\begin{array}{cc|cc} & & & \\ & c_0 & c_{-1} & c_{-2} & c_{-3} \\ & c_1 & c_0 & c_{-1} & c_{-2} \\ & & & & \\ \hline & & & & \\ & c_2 & c_1 & c_0 & c_{-1} \\ & c_3 & c_2 & c_1 & c_0 \\ & & & & \end{array}\right] \tag{3 1}$$

Moreover, the constants $\{ c_n \quad n \in \mathbf{Z} \}$ are intimately related to φ For $n \in \mathbf{Z}$, $c_n = \hat{\varphi}(n)$, where $\hat{\varphi}(n)$ denotes the nth Fourier coefficient of φ More precisely, for $n \in \mathbf{Z}$

$$c_n = \hat{\varphi}(n) = \int_{\mathbf{T}} \varphi\, \overline{e}_n = \tfrac{1}{2\pi} \int_0^{2\pi} \varphi(e^{it})\, e^{-int}\, dt$$

The unilaterally infinite matrix of T_φ with respect to the basis $\{ e_n \quad n \geq 0 \}$ is

just the compression of M_φ onto $H^2(\mathbf{T})$ and hence

$$T_\varphi = \begin{bmatrix} c_0 & c_{-1} & c_{-2} & \\ c_1 & c_0 & c_{-1} & \\ c_2 & c_1 & c_0 & \\ & & & \end{bmatrix} \tag{3 2}$$

corresponds to the bottom right block of our partitioned representation of M_φ in (3 1) above

The theory of Laurent operators is straightforward The mapping $\varphi \mapsto M_\varphi$ from $L^\infty(\mathbf{T})$ into $\mathcal{B}(L^2(\mathbf{T}))$ is an isometric unital $*$-algebra homomorphism In other words, for $\varphi, \psi \in L^\infty(\mathbf{T})$ and $\lambda, \mu \in \mathbf{C}$ we have $\|M_\varphi\| = \|\varphi\|_\infty$, $M_1 = I$, $M_{\overline{\varphi}} = (M_\varphi)^*$, $M_{\lambda\varphi+\mu\psi} = \lambda M_\varphi + \mu M_\psi$ and $M_{\varphi\psi} = M_\varphi M_\psi$ It follows that, $\varphi \mapsto M_\varphi$ is an isometric $*$-isomorphism of $L^\infty(\mathbf{T})$ onto a C^*-subalgebra of $\mathcal{B}(L^2(\mathbf{T}))$ and hence $\sigma(M_\varphi) = \sigma(\varphi)$ Moreover, since $L^\infty(\mathbf{T})$ is commutative, we have that Laurent operators commute with one another and in particular, every Laurent operator is normal

The Bilateral Shift The Laurent operator $W = M_{e_1}$ is called the *bilateral (forward) shift* since for all $n \in \mathbf{Z}$, $We_n = e_{n+1}$ Its adjoint, $W^* = M_{\overline{e}_1} = M_{e_{-1}}$ is called the bilateral backward shift and satisfies $W^*e_n = e_{n-1}$ for all $n \in \mathbf{Z}$ Observe that for all $n \in \mathbf{Z}$, $W^*We_n = e_n = WW^*e_n$ Therefore, $W^*W = WW^* = I$ and hence, W is unitary

In contrast to the theory of Laurent operators, the theory of Toeplitz operators is much more complicated However, part of the theory is easy Since each Toeplitz operator is the compression of a Laurent operator to $H^2(\mathbf{T})$, we have that the mapping $\varphi \mapsto T_\varphi$ maps $L^\infty(\mathbf{T})$ into $\mathcal{B}(H^2(\mathbf{T}))$ and that $\|T_\varphi\| \leq \|M_\varphi\| = \|\varphi\|_\infty$ In addition, $\varphi \mapsto T_\varphi$, is unital, linear and preserves adjoints In particular, since we will be looking at scalar translations of Toeplitz operators in the next section, we observe that if $\varphi \in L^\infty(\mathbf{T})$ and $\lambda \in \mathbf{C}$, then $T_\varphi - \lambda I = T_\varphi - \lambda T_1 = T_{\varphi-\lambda}$, a Toeplitz operator

The complications in the theory of Toeplitz operators are caused by their multiplicative properties. In contrast to the Laurent operator case, if φ, ψ are arbitrary functions in $L^\infty(\mathbf{T})$, then it is seldom true that $T_{\varphi\psi} = T_\varphi T_\psi$ or that $T_\varphi T_\psi = T_\psi T_\varphi$. In fact, it is seldom true that $T_\varphi T_\psi$ is a Toeplitz operator. The standard example for all these statements is furnished by the following analysis of the unilateral shift.

The Unilateral Shift. The Toeplitz operator $U = T_{e_1}$ is called the *unilateral (forward) shift* since for all $n \in \mathbf{N}$, $Ue_n = e_{n+1}$. Its adjoint, $U^* = T_{\overline{e_1}} = T_{e_{-1}}$ is called the unilateral backward shift. For $n \geq 1$ it satisfies $U^*e_n = P(e_{-1}e_n) = Pe_{n-1} = e_{n-1}$ and hence it is a backward shift on $\{ e_1, e_2, \quad \}$. However, $U^*e_0 = P(e_{-1}e_0) = 0$ and hence we can think of U^* as shifting e_0 backwards, out of the basis. Observe that for all $n \in \mathbf{N}$, $U^*Ue_n = e_n$ and hence $U^*U = I$ and U is an isometry. On the other hand, $UU^*e_n = e_n$ only for $n \geq 1$. For $n = 0$, $UU^*e_0 = 0$. By letting P_0 denote the projection of $H^2(\mathbf{T})$ onto span$\{ e_0 \}$, we can write $UU^* = I - P_0$. It follows immediately that U is not a normal operator.

Moreover, we have an example of two Toeplitz operators, U and U^*, such that $U^*U \neq UU^*$ and hence Toeplitz operators need not commute. In addition, the matrix of $UU^* = I - P_0$ with respect to the basis $\{ e_0, e_1, \quad \}$, is not constant on the main diagonal and hence, we also have that the product of two Toeplitz operators need not be a Toeplitz operator. To preview an upcoming result, we observe that although $T_{e_1e_{-1}} \neq T_{e_1}T_{e_{-1}}$ we do have that $T_{e_1e_{-1}} - T_{e_1}T_{e_{-1}} = P_0$ is a compact operator (finite rank) in $\mathcal{B}(H^2(\mathbf{T}))$.

Remark. It is possible to put restrictions on φ, ψ so as to insure that $T_{\varphi\psi} = T_\varphi T_\psi$. For example, if $\psi \in H^\infty(\mathbf{T}) = \left\{ \phi \in L^\infty(\mathbf{T}) \quad \hat{\phi}(n) = 0 \text{ for } n < 0 \right\}$, then $P\psi = \psi$ and the result follows. However, there is more to be gained by restricting our attention to continuous functions on the (compact) unit circle.

Remark about the unilateral shift. Let π denote the natural map from $\mathcal{B}(H^2(\mathbf{T}))$ onto the Calkin algebra, then by inspection of the above analysis of U, we have

that $\pi(U)$ is invertible in the Calkin algebra and hence U is Fredholm. In fact, $\pi(U)$ is a unitary in the Calkin algebra. We also have that $\ker U = \{0\}$ and $\ker U^* = \text{span}\{e_0\}$. Therefore, $\text{ind}(U) = -1$ and hence the unilateral shift is an example of a Fredholm operator that is not a compact perturbation of an invertible. Moreover, since it is an essentially unitary (Fredholm) operator with nonzero index, Brown-Douglas-Fillmore [BDF73] use it as an example of an essentially normal operator which cannot be written as a compact perturbation of a normal operator.

3.1.2 Standard Results About Toeplitz Operators

We now survey some standard results about Toeplitz operators. The main results are presented as theorems. Proofs can be found in [Mur90, Section 3.5].

Before restricting our attention to Toeplitz operators induced by continuous functions on $\mathbf{T}$, we state a theorem about the spectrum and norm of arbitrary Toeplitz operators.

Hartman-Wintner Theorem *If $\varphi \in L^\infty(\mathbf{T})$ and $\sigma(\varphi)$ denotes the spectrum of φ in $L^\infty(\mathbf{T})$, then $\sigma(\varphi) \subseteq \sigma(T_\varphi)$ and $\|T_\varphi\| = r(T_\varphi) = \|\varphi\|_\infty$.*

The remaining results that we need are for Toeplitz operators induced by continuous functions on $\mathbf{T}$. Such operators are called *Toeplitz operators with continuous symbol*. If $\varphi, \psi \in C(\mathbf{T})$, then, in general, we still do not have $T_{\varphi\psi} = T_\varphi T_\psi$—our unilateral shift example only involves the functions e_1 and e_{-1}, which are continuous on $\mathbf{T}$. However, for $\varphi, \psi \in C(\mathbf{T})$, we always have that $T_\varphi T_\psi - T_{\varphi\psi}$ is a compact operator and hence the images of $T_\varphi T_\psi$ and $T_{\varphi\psi}$ are equal in the Calkin algebra.

For the remainder of this section, let H^2 denote $H^2(\mathbf{T})$ and let π denote the natural map from $\mathcal{B}(H^2)$ onto the Calkin algebra, $\mathcal{B}(H^2)/\mathcal{K}(H^2)$. As mentioned above, the fact that $T_\varphi T_\psi - T_{\varphi\psi}$ is a compact operator for every $\varphi, \psi \in C(\mathbf{T})$ implies that $\pi(T_{\varphi\psi}) = \pi(T_\varphi T_\psi)$ for every $\varphi, \psi \in C(\mathbf{T})$. We know that the mapping $\varphi \mapsto T_\varphi$ from $C(\mathbf{T})$ into $\mathcal{B}(H^2)$ is unital, linear and preserves adjoints. It follows

that the mapping $\varphi \mapsto \pi(T_\varphi)$ from $\mathbf{C}(\mathbf{T})$ into $\mathcal{B}(H^2)/\mathcal{K}(H^2)$ is a unital $*$-algebra homomorphism, and hence $\{\, \pi(T_\varphi) \;\; \varphi \in C(\mathbf{T}) \,\}$ is a commutative $*$-subalgebra of $\mathcal{B}(H^2)/\mathcal{K}(H^2)$. It follows that every Toeplitz operator with continuous symbol is essentially normal.

Let $\mathcal{T}$ denote the C^*-subalgebra of $\mathcal{B}(H^2)$ generated by $\{\, T_\varphi \;\; \varphi \in C(\mathbf{T}) \,\}$. $\mathcal{T}$ is called the *Toeplitz algebra*. An important result is that the Toeplitz algebra contains the ideal of compact operators on H^2 and hence $\mathcal{T}/\mathcal{K}(H^2)$ is a C^*-subalgebra of $\mathcal{B}(H^2)/\mathcal{K}(H^2)$. Moreover, $\mathcal{T}/\mathcal{K}(H^2)$ contains $\pi(T_1) = \pi(I)$, the identity of $\mathcal{B}(H^2)/\mathcal{K}(H^2)$. It follows that elements of $\mathcal{T}/\mathcal{K}(H^2)$ are invertible in $\mathcal{T}/\mathcal{K}(H^2)$ if and only if they are invertible in $\mathcal{B}(H^2)/\mathcal{K}(H^2)$ and we have the following theorem.

Theorem 1 *The map*

$$\varphi \mapsto \pi(T_\varphi) \;\; C(\mathbf{T}) \to \mathcal{T}/\mathcal{K}(H^2) \tag{3.3}$$

is a $*$-isomorphism.

As an immediate consequence of this isomorphism, we have that if $\varphi \in C(\mathbf{T})$, then $\sigma(\pi(T_\varphi)) = \sigma(\varphi)$ and $\pi(T_\varphi)$ is invertible in $\mathcal{T}/\mathcal{K}(H^2)$ if and only if φ is invertible in $C(\mathbf{T})$. In other words, we have the following theorem.

Theorem 2 *If $\varphi \in C(\mathbf{T})$, then $\sigma_e(T_\varphi) = \varphi(\mathbf{T})$ and T_φ is Fredholm if and only if φ never vanishes.*

When T_φ is Fredholm, its index is intimately related to φ.

Theorem 3 *If $\varphi \in C(\mathbf{T})$ and φ never vanishes, then*

$$\operatorname{ind}(T_\varphi) = -\operatorname{wn}(\varphi, 0)$$

where $\operatorname{wn}(\varphi, 0)$ denotes the winding number of φ about 0.

Remark In addition, it can be shown that, if $\varphi \in C(\mathbf{T})$, then T_φ is invertible if and only if T_φ is Fredholm of index 0. Therefore, if $\varphi \in C(\mathbf{T})$, then we can write the spectrum of T_φ as follows:

$$\sigma(T_\varphi) = \varphi(\mathbf{T}) \cup \{ \lambda \in \mathbf{C} \backslash \varphi(\mathbf{T}) \quad T_\varphi - \lambda I = T_{\varphi - \lambda} \text{ is (Fredholm) of non-zero index} \}$$

Example We apply some of these results to the unilateral shift, $U = T_{e_1}$. Since $e_1 \in C(\mathbf{T})$ and $e_1(\mathbf{T}) = \mathbf{T}$, we have that e_1 does not vanish on $\mathbf{T}$ and hence, U is Fredholm with $\operatorname{ind}(U) = -\operatorname{wn}(e_1, 0) = -1$. This agrees with our earlier analysis. Moreover, if λ is not on $e_1(\mathbf{T}) = \mathbf{T}$, then $T_{e_1 - \lambda}$ is Fredholm with

$$\operatorname{ind}(T_{e_1 - \lambda}) = -\operatorname{wn}(e_1 - \lambda, 0) = -\operatorname{wn}(e_1, \lambda) = \begin{cases} 0 & \text{if } |\lambda| > 1 \\ -1 & \text{if } |\lambda| < 1 \end{cases}$$

Therefore, by the immediately preceeding Remark, $\sigma(U) = \mathbf{T} \cup \{ \lambda \quad |\lambda| < 1 \}$. That is, the spectrum of the unilateral shift is precisely the unit disk.

This completes our survey of standard results. In the next section, we apply these results to get estimates on the distance from a Toeplitz operator with continuous symbol to the set of normal operators in $\mathcal{B}(H^2)$.

3 2 Distance Estimates for Toeplitz Operators

In this section, we use the theory of Toeplitz operators and the theory of Fredholm operators to obtain upper and lower bounds on the distance from certain Toeplitz operators with continuous symbol to the set, $\mathcal{N}$, of normal operators in $\mathcal{B}(H^2)$.

Proposition 1 *If $\varphi \in C(\mathbf{T})$, so that $T_\varphi \in \mathcal{B}(H^2)$, then*
$\operatorname{dist}(T_\varphi, \mathcal{N}) \leq \inf \{ r \quad \text{there is a disk of radius } r \text{ containing } \varphi(\mathbf{T}) \}$

Proof For every $\lambda \in \mathbf{C}$ and $r > 0$, let $D_r(\lambda) = \{ \mu \in \mathbf{C} \quad |\mu - \lambda| < r \}$ denote the open disk of radius r about λ.

Since φ is continuous and $\mathbf{T}$ is compact, we have that $\varphi(\mathbf{T})$ is a compact subset of $\mathbf{C}$ and hence there are disks of finite radius containing $\varphi(\mathbf{T})$.

Let $D_r(\lambda)$ be an arbitrary but fixed disk such that $\varphi(\mathbf{T}) \subseteq D_r(\lambda)$ By letting $\psi = \varphi - \lambda$, we have that ψ is continuous and $\psi(\mathbf{T}) = \varphi(\mathbf{T}) - \lambda \subseteq D_r(\lambda) - \lambda = D_r(0)$ It follows that $\|\psi\|_\infty \leq r$

Since the normal approximation problem is invariant under scalar translation $\operatorname{dist}(T_\varphi, \mathcal{N}) = \operatorname{dist}(T_\varphi - \lambda I, \mathcal{N})$ However, $T_\varphi - \lambda I = T_{\varphi - \lambda} = T_\psi$ Therefore

$$\begin{aligned} \operatorname{dist}(T_\varphi, \mathcal{N}) &= \operatorname{dist}(T_\psi, \mathcal{N}) \\ &\leq \|T_\psi\| \quad \text{since } 0 \text{ is normal} \\ &= \|\psi\|_\infty \quad \text{by the Hartman-Wintner Theorem} \\ &\leq r \end{aligned}$$

However, $D_r(\lambda)$ was an arbitrary disk of radius r containing $\varphi(\mathbf{T})$ and hence $\operatorname{dist}(T_\varphi, \mathcal{N}) \leq \inf \{ r \text{ there is a disk of radius } r \text{ containing } \varphi(\mathbf{T}) \}$ ∎

Proposition 2 *If $\varphi \in C(\mathbf{T})$ and there exists a component of $\mathbf{C} \setminus \varphi(\mathbf{T})$ in which the winding number of φ is nonzero, then*

$$\operatorname{dist}(T_\varphi, \mathcal{N}) \geq \sup \left\{ r \;\middle|\; \begin{array}{l} \textit{there is a disk of radius } r \textit{ in } \mathbf{C} \setminus \varphi(\mathbf{T}) \textit{ in} \\ \textit{which the winding number of } \varphi \textit{ is nonzero} \end{array} \right\}$$

Proof Let $D_r(\lambda)$ be an arbitrary but fixed disk contained inside a component of $\mathbf{C} \setminus \varphi(\mathbf{T})$ in which the winding number of φ is nonzero It follows that the winding number of φ is constant on $D_r(\lambda)$ Let $n = \operatorname{wn}(\varphi, \lambda)$ denote this nonzero winding number

By letting $\psi = \varphi - \lambda$, we have that ψ is continuous and $\psi(\mathbf{T}) = \varphi(\mathbf{T}) - \lambda$ It follows that $D_r(0) = D_r(\lambda) - \lambda$ is in a component of $\mathbf{C} \setminus \psi(\mathbf{T})$ where the winding number of ψ is also constantly equal to n Since $\psi(\mathbf{T})$ does not intersect $D_r(0)$, we have

$$r \leq \inf_{z \in \mathbf{T}} |\psi(z)| = |\psi|_{\min}$$

Moreover, since 0 is not in $\psi(\mathbf{T})$, ψ never vanishes on $\mathbf{T}$ and hence T_ψ is Fredholm with $\operatorname{ind}(T_\psi) = -\operatorname{wn}(\psi, 0) = -n$

For every $T \in \mathcal{B}(H^2)$, let $B_r(T) = \left\{ A \in \mathcal{B}(H^2) \;\middle|\; \|T - A\| < r \right\}$ denote the open ball of radius r about T in $\mathcal{B}(H^2)$

Claim $B_r(T_\psi) \subseteq \mathrm{Fred}(H^2)$

Note Assuming this claim is true, $B_r(T_\psi)$ is a connected set of Fredholm operators and hence they all have index equal to $\mathrm{ind}(T_\psi) = -n \neq 0$ It follows that there are no normal operators in $B_r(T_\psi)$, (since every normal Fredholm operator has index 0) and therefore $\mathrm{dist}(T_\psi, \mathcal{N}) \geq r$ As in Proposition 1, $\mathrm{dist}(T_\varphi, \mathcal{N}) = \mathrm{dist}(T_\varphi - \lambda I, \mathcal{N}) = \mathrm{dist}(T_\psi, \mathcal{N})$, and so we have $\mathrm{dist}(T_\varphi, \mathcal{N}) \geq r$ However, $D_r(\lambda)$ was an arbitrary disk of radius r contained inside a component of $\mathbf{C} \setminus \varphi(\mathbf{T})$ in which the winding number of φ is nonzero and hence

$$\mathrm{dist}(T_\varphi, \mathcal{N}) \geq \sup \left\{ r \;\middle|\; \begin{array}{l} \text{there is a disk of radius } r \text{ in } \mathbf{C} \setminus \varphi(\mathbf{T}) \text{ in} \\ \text{which the winding number of } \varphi \text{ is nonzero} \end{array} \right\}$$

as we set out to prove It only remains to prove that $B_r(T_\psi) \subseteq \mathrm{Fred}(H^2)$

Proof Take any $A \in B_r(T_\psi)$ To show that A is Fredholm, it will suffice to show that $\pi(A)$ is invertible However, we know $\pi(T_\psi)$ is invertible, and so it will suffice to show that $\|\pi(T_\psi) - \pi(A)\| < \|\pi(T_\psi)^{-1}\|^{-1}$

Using the fact that $C(\mathbf{T}) \cong \mathcal{T}/\mathcal{K}(H^2)$ via $\phi \mapsto \pi(T_\phi)$ we have

$$\|\pi(T_\psi)^{-1}\| = \|\pi(T_{\frac{1}{\psi}})\| = \left\| \frac{1}{\psi} \right\|_\infty = \sup_{z \in \mathbf{T}} \frac{1}{|\psi(z)|} = \frac{1}{|\psi|_{\min}}$$

and then

$$\|\pi(T_\psi) - \pi(A)\| = \|\pi(T_\psi - A)\| \leq \|T_\psi - A\| < r \leq |\psi|_{\min} = \|\pi(T_\psi)^{-1}\|^{-1}$$

We now have that $\pi(A)$ is invertible and hence A is Fredholm However, A was arbitrary in $B_r(T_\psi)$ and therefore, $B_r(T_\psi) \subseteq \mathrm{Fred}(H^2)$ ∎

Corollary 3. *If $\varphi \in C(\mathbf{T})$ and φ is 1 1, so that $\varphi(\mathbf{T})$ is a Jordan curve, then* $\mathrm{dist}(T_\varphi, \mathcal{N}) \geq \sup \{ r$ there is a disk of radius r contained *inside* $\varphi(\mathbf{T}) \}$

Proof Since $\varphi(\mathbf{T})$ is a Jordan curve, it determines exactly two components—an *inside* and an *outside*. The winding number of φ in the unbounded outside is 0. However, since φ is 1 – 1, $\varphi(\mathbf{T})$ goes around the inside exactly once, and hence the winding number of φ in the inside is ± 1.

Since $\varphi \in C(\mathbf{T})$ and the inside is the only component determined by φ in which the winding number is nonzero, Proposition 2 reduces to

$\operatorname{dist}(T_\varphi, \mathcal{N}) \geq \sup\{r$ – there is a disk of radius r contained *inside* $\varphi(\mathbf{T})\}$ ∎

Corollary 4 *Suppose $\varphi \in C(\mathbf{T})$ and $\varphi(\mathbf{T})$ is an ellipse with axis lengths a, b where $a \leq b$. If the winding number of φ inside the ellipse is nonzero, then* $\operatorname{dist}(T_\varphi, \mathcal{N}) \in [a, b]$

Proof Since $\varphi \in C(\mathbf{T})$, Proposition 1 applies and hence

$\operatorname{dist}(T_\varphi, \mathcal{N}) \leq \inf\{r$ – there is a disk of radius r containing the ellipse$\} = b$

Since $\varphi \in C(\mathbf{T})$ and $\varphi(\mathbf{T})$ is an ellipse, its inside is the only possible component in which the winding number of φ is nonzero and, by hypothesis, it is nonzero. Therefore Proposition 2 applies and hence

$$\begin{aligned} \operatorname{dist}(T_\varphi, \mathcal{N}) &\geq \sup\{r \text{ – there is a disk of radius } r \text{ contained inside the ellipse}\} \\ &= a \quad \blacksquare \end{aligned}$$

Corollary 5 *Suppose $\varphi \in C(\mathbf{T})$ and $\varphi(\mathbf{T})$ is a circle of radius a. If the winding number of φ inside the circle is nonzero, then* $\operatorname{dist}(T_\varphi, \mathcal{N}) = a$

Proof By Corollary 4, $\operatorname{dist}(T_\varphi, \mathcal{N}) \in [a, a]$ ∎

Example We apply this last corollary to the unilateral shift $U = T_{e_1}$. Since e_1 is continuous on $\mathbf{T}$, $e_1(\mathbf{T}) = \mathbf{T}$ is the unit circle and the winding number of e_1 inside the circle is $1 \neq 0$, we have that $\operatorname{dist}(U, \mathcal{N}) = 1$. Moreover, since U is an isometry, we have that $\|U\| = 1$ and hence 0 is a best normal approximant for U.

Remark As we mentioned in the Introduction, the fact that the unilateral shift is at distance 1 from the normals has been known for a long time. For a different

proof see [Hal82, Problem 144] or see Holmes' proof [Hol74] that the unilateral shift is antinormal (*i e* a nearest normal to U is 0)

Example If n is any nonzero integer, then e_n is continuous on $\mathbf{T}$, $e_n(\mathbf{T}) = \mathbf{T}$ and the winding number of e_n inside the circle is $n \neq 0$ It follows that for each nonzero integer n, $\operatorname{dist}(T_{e_n}, \mathcal{N}) = 1$ Moreover, $\|T_{e_n}\| = \|e_n\|_\infty = 1$, and hence 0 is a best normal approximant for T_{e_n} Of course, for the $n = 0$ case, $T_{e_0} = I$ is normal

Remark If we compress the unilateral backward shift $U^* = T_{e_{-1}}$ to the span of $\{ e_0, \quad , e_{n-1} \}$ and look at its matrix representation, we have the $n \times n$ basic superdiagonal matrix It seems curious that U^* is at distance 1 from the normals while every such $n \times n$ basic superdiagonal matrix is at distance $\frac{1}{2}$ from the normals

3 3 Direct Sums of Toeplitz Operators

When considering the normal approximation problem for direct sums of Toeplitz operators, the fundamental question is as follows

> *If A is a direct sum of Toeplitz operators, will a direct sum of corresponding best normal approximants be a best normal approximant for A?*

The purpose of this final section is to present a simple example (devised by A R Sourour in a private discussion) which demonstrates that a direct sum of corresponding best normal approximants is not neccessarily a best normal approximant for an arbitrary A The example involves our most well known Toeplitz operators, the unilateral shift and its adjoint

Example (Sourour). Let $U = T_{e_1}$ denote the unilateral shift, then 0 is a best normal approximant for U and for U^* However, $0 = 0 \oplus 0$ is not a best normal approximant for $A = U^* \oplus U$ since a normal operator can be exhibited which is a better normal approximant than 0

Proof The distance from A to 0 is $\|A - 0\| = \|A\| = \max\{\|U^*\|, \|U\|\} = 1$

The main part of this example is to exhibit a normal operator N, such that $\|A - N\| < 1$ and hence 0 is not a nearest normal to A

As a first step we write the matrix of A as a bilaterally infinite partitioned matrix with respect to the partitioned basis

$$\mathcal{E} = \left\{ \quad e_2 \oplus 0 \quad e_1 \oplus 0 \quad e_0 \oplus 0 \,\middle|\, 0 \oplus e_0 \quad 0 \oplus e_1 \quad 0 \oplus e_2 \quad \right\}$$

Since $U^*(e_0) = 0$, $U^*(e_n) = e_{n-1}$ for $n > 1$, and $U(e_n) = e_{n+1}$ for all $n \geq 0$, we get

$$A = \left[\begin{array}{c|c} U^* & 0 \\ \hline 0 & U \end{array}\right] = \left[\begin{array}{ccc|ccc} & & & & & \\ 0 & & & & & \\ 1 & 0 & & & & \\ & 1 & 0 & & & \\ \hline & & & 0 & 0 & \\ & & & & 1 & 0 \\ & & & & & 1 \\ & & & & & \end{array}\right]$$

Now, let W denote the bilateral shift operator on $\mathcal{H} = H^2(\mathbf{T}) \oplus H^2(\mathbf{T})$ with respect to the basis $\mathcal{E}$ W is unitary, and hence the operator $N = \frac{1}{2}W$ is a normal operator on $\mathcal{H}$ The matrix of $A - N$ with respect to $\mathcal{E}$ is given by

$$A - N = \left[\begin{array}{ccc|ccc} & & & & & \\ 0 & & & & & \\ \frac{1}{2} & 0 & & & & \\ & \frac{1}{2} & 0 & & & \\ \hline & & & -\frac{1}{2} & 0 & \\ & & & & \frac{1}{2} & 0 \\ & & & & & \frac{1}{2} \\ & & & & & \end{array}\right]$$

By inspection of this matrix, $A-N$ is a weighted shift, our prototype for weighted partial permutations in Section 1 2 3 Therefore $\|A - N\| = \frac{1}{2}$, the supremum of all the weights

It follows that $0 = 0 \oplus 0$ is not a best normal approximant for $A = U^* \oplus U$ ∎

Remark The operator N constructed in Sourour's example is actually a best normal approximant for $A = U^* \oplus U$

Proof By Holmes' distance estimate [Section 1 2 2]

$$\begin{aligned} \operatorname{dist}(A, \mathcal{N}) &\geq \tfrac{1}{2} \sup_{\|x\|=1} \big| \|Ax\| - \|A^*x\| \big| \\ &\geq \tfrac{1}{2} \big| \|A(0 \oplus e_0)\| - \|A^*(0 \oplus e_0)\| \big| \\ &= \tfrac{1}{2} \big| \|0 \oplus e_1)\| - \|0 \oplus 0\| \big| \\ &= \tfrac{1}{2} \end{aligned}$$

We now have $\frac{1}{2} \leq \operatorname{dist}(A, \mathcal{N}) \leq \|A - N\| = \frac{1}{2}$ Therefore $\operatorname{dist}(A, \mathcal{N}) = \frac{1}{2}$ and N is a best normal approximant for A ∎

Bibliography

[BDF73] L Brown, R G Douglas, and P A Fillmore *Unitary Equivalence Modulo the Compact Operators and Extensions of C^*-Algebras*, volume 345 of *Lecture Notes in Mathematics*, pages 58–128 Springer-Verlag, 1973

[BHK91] Rajendra Bhatia, Roger A Horn, and Fuad Kittaneh Normal approximants to binormal operators *Linear Algebra and Its Applications*, **147** 169–179, (1991)

[Dav88] Kenneth R Davidson *Nest algebras*, volume 191 of *Pitman Research Notes in Mathematics Series* Longman Scientific & Technical, Harlow, Essex, England, 1988

[Dou72] Ronald G Douglas *Banach Algebra Techniques in Operator Theory*, volume 49 of *Pure and Applied Mathematics* Academic Press, Inc , New York, 1972

[FH55] Ky Fan and A J Hoffman Some metric inequalities in the space of matrices *Proceedings of the American Mathematical Society*, **6** 111–116, (1955)

[Fug50] B Fuglede A commutativity theorem for normal operators *Proceedings of the National Academy of Sciences U S A* , **36** 35–40, (1950)

[Hal57] Paul R Halmos *Introduction to Hilbert Space and the Theory of Spectral Multiplicity* Chelsea Publishing Company, New York, second edition, 1957

[Hal72] Paul R Halmos Positive approximants of operators *Indiana University Mathematics Journal*, **21** 951–960, (1972)

[Hal74] Paul R Halmos Spectral approximants of normal operators *Proceedings of the Edinburgh Mathematical Society*, **19** 51–58, (1974)

[Hal82] Paul R Halmos *A Hilbert Space Problem Book*, volume 19 of *Graduate Texts in Mathematics* Springer-Verlag, New York, second edition, 1982

[HK71] Richard B Holmes and Bernard R Kripke Best approximation by compact operators *Indiana University Mathematics Journal*, **21** 255–263, (1971)

[Hol74] Richard B Holmes Best approximation by normal operators *Journal of Approximation Theory*, **12** 412–417, (1974)

[Lin94] Huaxin Lin Almost commuting selfadjoint matrices and their applications Preprint, 1994

[Mur90] Gerard J Murphy *C^*-Algebras and Operator Theory* Academic Press, Inc , San Diego, 1990

[Phi77] John Phillips Nearest normal approximation for certain operators *Proceedings of the American Mathematical Society*, **67** 236–240, (1977)

[Rog76] Donald D Rogers On proximinal sets of normal operators *Proceedings of the American Mathematical Society*, **61** 44–48, (1976)

[Ros58] M Rosenblum On a theorem of Fuglede and Putnam *J London Math Soc* , **33** 376–377, (1958)

[RR73] Heydar Radjavi and Peter Rosenthal *Invariant Subspaces* Springer-Verlag, New York, 1973

[Ruh87] Axel Ruhe Closest normal matrix finally found! *BIT*, **27** 585–598, (1987)

[vR72] D J van Riemsdijk Some metric inequalities in the space of bounded linear operators on a separable Hilbert space *Nieuw Archief voor Wiskunde*, **20** 216–230, (1972)

Vita

Surname Harrison

Given Names Robert Norman

Place of Birth Winnipeg, Manitoba, Canada

Educational Institutions Attended

University of Victoria	1991 to 1995
University of Manitoba	1972 to 1974
University of Winnipeg	1969 to 1972
University of Manitoba	1967 to 1969

Degrees Awarded

B Sc	University of Winnipeg	1972

Partial Copyright License

I hereby grant the right to lend my thesis to users of the University of Victoria Library, and to make single copies only for such users or in response to a request from the library of any other university, or similar institution, on its behalf or for one of its users. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by me or a member of the university designated by me. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Title of Thesis:

Normal Approximants for Certain Toeplitz Matrices and Toeplitz Operators

Author: [redacted]

Robert Norman Harrison

September 29, 1995