Aspects of Toeplitz operators and matrices: asymptotics, norms, singular values

H Rabe

12516139

Thesis submitted for the degree *Philosophiae Doctor* in Mathematics at the Potchefstroom Campus of the North-West University

Promoter:	**Prof ACM Ran**
Co-promoter:	**Prof G Groenewald**
Assistant-promoter:	**Prof JH Fourie**

October 2015

Acknowledgements

To study full-time towards a Phd in mathematics is a great opportunity, especially in South Africa, and one that I am grateful to have had. With substantial support from the National Research Foundation (NRF), the Vrije Universiteit in Amsterdam, and the North-West University, I was able to focus all my attention on completing this project without distraction.

I would like to thank my main supervisor, André Ran, from whom I learned a lot during these four years, and for his expert guidance. I'm also grateful to my assistant supervisors, Gilbert Groenewald and Jan Fourie, for their support and willingness to assist in all matters related to my Phd.

During my visits to Amsterdam I also made a lot of good friends and met interesting people. All the evenings out at Brasserie Blazer, Frankrijk, Niew Anita, etc. with Michelangelo, Nienke, Thomas, Blaz, Simone, Pablo, Niccola, Pia, Jente, and all the others were really fun and provided the necessary escape from the research every weekend. Being in Europe for eight months in total also allowed for the occasional travel opportunities, and I managed to see quite a bit of the surrounding countries as well.

I'm also grateful to my family and local friends who I could visit regularly during my studies - this always helped me to recharge before carrying on with the work.

Summary

Keywords: Toeplitz matrices, fisection, singular values, eigenvalues, eigenvectors

The research contained in this thesis can be divided into two related, but distinct parts. The first chapter deals with block Toeplitz operators defined by rational matrix function symbols on discrete sequence spaces. Here we study sequences of operators that converge to the inverses of these Toeplitz operators via an invertibility result involving a special representation of the symbol of these block Toeplitz operators. The second part focuses on a special class of matrices generated by banded Toeplitz matrices, i.e., Toeplitz matrices with a finite amount of non-zero diagonals. The spectral theory of banded Toeplitz matrices is well developed, and applied to solve questions regarding the behaviour of the singular values of Toeplitz-generated matrices. In particular, we use the behaviour of the singular values to deduce bounds for the growth of the norm of the inverse of Toeplitz-generated matrices.

In chapter 2, we use a special state-space representation of a rational matrix function on the unit circle to define a block Toeplitz operator on a discrete sequence space. A discrete Riccati equation can be associated with this representation which can be used to prove an invertibility theorem for these Toeplitz operators. Explicit formulas for the inverse of the Toeplitz operators are also derived that we use to define a sequence of operators that converge in norm to the inverse of the Toeplitz operator. The rate of this convergence, as well as that of a related Riccati difference equation is also studied. We conclude with an algorithm for the inversion of the finite sections of block Toeplitz operators.

Chapter 3 contains the main research contribution of this thesis. Here we derive sharp growth rates for the norms of the inverses of Toeplitz-generated matrices. These results are achieved by employing powerful theory related to the Avram-Parter theorem that describes the distribution of the singular values of banded Toeplitz matrices. The investigation is then extended to include the behaviour of the extreme and general singular values of Toeplitz-generated matrices.

We conclude with Chapter 4, which sets out to answer a very specific question regarding the singular vectors of a particular subclass of Toeplitz-generated matrices. The entries of each singular vector seems to be a permutation (up to sign) of the same set of real numbers. To arrive at an explanation for this phenomenon, explicit formulas are derived for the singular values of the banded Toeplitz matrices that serve as generators for the matrices in question. Some abstract algebra is also employed together with some results from the previous chapter to describe the permutation phenomenon. Explicit formulas are also shown to exist for the inverses of these particular Toeplitz-generated matrices as well as algorithms to calculate the norms and norms of the inverses. Finally, some additional results are compiled in an appendix.

Opsomming

Die navorsing saamgevat in hierdie proefskrif kan verdeel word in twee aparte, maar tog verwante dele. Die eerste hoofstuk handel oor blok Toeplitz operatore gedefinieer deur rationale matriks funksie simbole op diskrete funksie ruimtes. Hier bestudeer ons rye wat konvergeer na die inverses van blok Toeplitz operatore via 'n omkeerbaarheids resultaat wat n spesiale voorstelling van die simbool bevat.

Die tweede gedeelte fokus op 'n spesiale klas van matrikse wat gegenereer word deur band Toeplitz matrikse, met ander woorde, Toeplitz matrikse wat 'n eindige hoeveelheid nie-nul diagonal bevat. Die spektraal teorie van band Toeplitz matrikse is hoogs ontwikkeld, en word toegepas om vrae rakend die gedrag van die singuliere waardes van Toeplitz gegenereerde matrikse op te los. In besonder gebruik ons die gedrag van die singuliere waardes om grense vir die groei van die norms van die inverses van Toeplitz gegenereerde matrikse te bepaal.

In hoofstuk 2 gebruik ons 'n spesiale voorstelling van die rationale matriks funksie op die eenheid sirkel om n blok Toeplitz operator op 'n diskrete ry ruimte te definieer. 'n Diskrete Riccati vergelyking kan met hierdie voorstelling geassosieer word wat dan gebruik kan word om 'n omkeerbaarheids stelling vir Toeplitz operatore te bewys. Eksplisiete formulas vir die inverse van die Toeplitz operatore word ook afgelei wat gebruik word om 'n ry operatore te definieer wat in norm konvergeer na die inverse van die Toeplitz operator. Die tempo van hierdie konvergensie, asook die van 'n verwante Riccati vergelyking word bestudeer. Ons eindig die hoofstuk af met n algoritme vir die berekening van die inverses van die eindige seksies van blok Toeplitz operatore.

Hoofstuk drie bevat die belangrikste navorsings bydrae van hierdie proefskrif. Hier lei ons akkurate groei tempos af van die norms van die inverses van Toeplitz gegenereerde matrikse. Hierdie resultate word verkry deur die toepassing van kragtige teorie verwant aan die Avram-Parter stelling. Hierdie stelling beskryf die verspreiding van singuliere waardes van Toeplitz matrikse. Die ondersoek word dan uitgebrei om the gedrag van die ekstreem en algemene singuliere waardes van Toeplitz gegenereerde matrikse te in te sluit.

In die finale hoofstuk 4, beantwoord ons 'n baie unieke vraag aangaande die singuliere vektore van 'n spesifieke subklas van Toeplitz gegenereerde matrikse. Die inskrywings van elke singuliere vektor wil voorkom om permutasies (uitsluitend die teken) van dieselfde versameling reele getalle te wees. Om n verduideliking vir hierdie gedrag te vind, lei ons eksplisiete formules af vir die singuliere waardes van die band Toeplitz matriks wat die genereerder is van die subklas van matrikse wat ondersoek word. Sekere abstrakte algebra teorie work ook gebruik, tesame met resultate van die vorige hoofstuk om the permutasie verskynsel te verduidelik. Eksplisiete formules vir die inverses van die spesifieke subklas van Toeplitz gegenereerde matrikese word afgelei, asook algoritmes vir die berekening van die norms en norms van die. 'n Bylaag word ook aangeheg wat additionele resultate bevat.

Contents

Chapter 1

Introduction

The study of Toeplitz operators and matrices has been an active field of research for more or less a century, starting in the early twentieth century with Otto Toeplitz, after whom these operators and matrices have been named. Research in this field has yielded thousands of research papers, ranging from application driven problems in numerical analysis, physics, probability theory, control theory and differential equations, to very deep theoretical results involving more abstract constructs such as Von Neumann and C^* algebras. The present investigation lies somewhere in between these two extremes, and will focus on providing new insights into some standard concepts related to Toeplitz operators and matrices. These include norms, convergence, singular values, singular vectors, eigenvalues and asymptotics.

Toeplitz operators can live on a variety of spaces, ranging from function spaces, to the more concrete l_p sequence spaces. In general though, they can all be characterized as a type of multiplication operator which is closely related to convolution equations and the operators they induce. In fact, the subclass of Wiener-Hopf integral operators define Toeplitz operators on certain Lebesque function spaces. The pioneering work on the equations producing these operators was done by N. Wiener and E. Hopf, and their work encouraged further study by many other mathematicians including M. G. Kreĭn. In the nineteen sixties, I.C. Gohberg and I.A. Fel'dman continued research in this area and compiled their work in the book [13]. More recently, many books have been published that make Toeplitz operators part of its main focus, e.g. [9, 10, 16, 14, 5, 20]. Currently, research into Toeplitz operators is still thriving, and the body of knowledge that has been established is immense and growing.

The contribution in this thesis involves Toeplitz operators and matrices defined on discrete sequence spaces. In this setting these operators have matrix representations with the well-known property that their diagonals consist of the same entries. The majority of our findings concern finite matrices and rely heavily on results that have been compiled in [5]. The first part of this work, contained in Chapter 2, deals mostly with the convergence (in norm) of a particular sequence of operators to block Toeplitz operators. Chapter 3 is dedicated to a class of Toeplitz-generated (T-gen) matrices. This is a class of $n \times n$ matrices of the form $X_n = T_n + f_n \cdot (T_n^{-1})^*$, where T_n is a banded Toeplitz matrix and f_n some sequence of positive numbers converging to zero. This chapter will deal with the norms, norms of inverses and singular values of T-gen matrices as their sizes grow to infinity. In Chapter 4, a special example of the T-gen class is studied and numerous

additional results are derived.

The rest of this chapter will be dedicated to establishing general background results that are applicable to the research in the following chapters.

1.1 Toeplitz operators

For our purposes, we will consider bounded Toeplitz operators defined on the sequence spaces $l_p(\mathbb{C}^m)$ and l_2. The former will be considered in Chapter 2 while the latter will apply for the rest of the chapters. On these sequence spaces the corresponding matrix representation of a Toeplitz operator is well-known - it is characterized by having constant diagonal elements. When these elements are chosen as the Fourier coefficients of an analytic (possibly matrix) function defined on an annulus that contains the unit circle, $\mathbb{T}$, it forces the Toeplitz operator to be bounded. This analytic function is called the *symbol* of the Toeplitz operator. The entries of the Toeplitz matrix are assigned as follows. Let

$$T = \begin{bmatrix} a_0 & a_{-1} & a_{-2} & \dots \\ a_1 & a_0 & a_{-1} & \dots \\ a_2 & a_1 & a_0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

be a Toeplitz operator defined on $l_p(\mathbb{C}^m)$ or l_2. The symbol of this operator has the form $a(t) = \sum_{i=-\infty}^{\infty} a_i t^i$, $t \in \mathbb{T}$. In the case of $l_p(\mathbb{C}^m)$, the a_i's denote $m \times m$ matrices with complex entries, and then T is referred to as a *block* Toeplitz operator. For l_2, the a_i's denote complex numbers.

By associating a symbol with every bounded Toeplitz operator, we will see that this allows us to investigate important properties of Toeplitz operators by just considering its symbol, and transferring questions about infinite dimensional operators to the domain of the complex plane. This is true for the norm of T, and for analytic symbols associated with Toeplitz operators, we have that

$$\|T\| = \|a(t)\|_\infty := \operatorname{ess\,sup}_t |a(t)|, \quad t \in \mathbb{T}.$$

This statement is valid for both the block and scalar case (Chapter XXIII, Corollary 3.2, [15]).

Invertibility of Toeplitz operators can also be determined by analyzing its symbol. When considering the space $l_p(\mathbb{C}^m)$, we will restrict our Toeplitz operators to having symbols that are rational matrix functions, $R(t)$. This means that the entries of $R(t)$ are quotients of two polynomials. A special *Wiener-Hopf factorisation* of the symbol is required to arrive at the desired theorem concerning invertibility, and we state a theorem regarding this factorisation - see Chapter XXIV, Theorem 3.1, [15].

Theorem 1.1.1. *Let $R(t)$ be a rational $m \times m$ matrix function with no poles on $\mathbb{T}$, and assume that $\det R(t) \neq 0$ for all $t \in \mathbb{T}$. Then there exist integers $\kappa_1 \leq \kappa_2 \leq \cdots \leq \kappa_m$ and*

rational $m \times m$ matrix functions R_- and R_+, which have no poles on $\mathbb{T}$, such that

$$R(t) = R_-(t) \begin{bmatrix} t^{\kappa_1} & & & 0 \\ & t^{\kappa_2} & & \\ & & \ddots & \\ 0 & & & t^{\kappa_m} \end{bmatrix} R_+(t), \qquad t \in \mathbb{T}, \tag{1.1}$$

and

- *R_+ has no poles on $|t| \leq 1$,*
- *$\det R_+(t) \neq 0$ for $|t| \leq 1$,*
- *R_- has no poles on $|t| \geq 1$, (∞ included)*
- *$\det R_-(t) \neq 0$ for $|t| \geq 1$, (∞ included).*

In particular, R_-^{-1} and R_+^{-1} exist, the functions R_- and R_-^{-1} are minus-functions and R_+ and R_+^{-1} are plus-functions.

By minus- and plus-functions we mean functions whose Fourier coefficients with strictly positive index, respectively negative index are zero. Note that this theorem applies also to the scalar case.

If all the indices $\kappa_1 \dots \kappa_m$ are equal to zero in the Wiener-Hopf factorisation (1.1), it is called a *right canonical factorisation*. A *left canonical factorisation* is defined similarly, except that the order of the first and last factors have switched. We can now characterize the invertibility of Toeplitz operators (Chapter XXIV, Theorem 4.1, [15]).

Theorem 1.1.2. *Let T be a block Toeplitz operator on $l_2(\mathbb{C}^m)$ defined by a rational matrix function $R(t)$. Then T is invertible, if and only,*

- *$\det R(t) \neq 0$ for each $t \in \mathbb{T}$,*
- *$R(t)$ admits a (right) canonical factorisation relative to $\mathbb{T}$.*

In this case the inverse of T is obtained in the following way. Construct a canonical factorisation $R(t) = R_-(t)R_+(t), t \in \mathbb{T}$, and write the Fourier series

$$R_-(t)^{-1} = \sum_{j=-\infty}^{0} R_j^- t^j, \qquad R_+(t)^{-1} = \sum_{j=0}^{\infty} R_j^+ t^j.$$

Then $T^{-1} = [t_{ij}]_{i,j=0}^{\infty}$, where

$$t_{ij} = \begin{cases} \sum_{r=0}^{j} R_{i-r}^+ R_{r-j}^-, & i \geq j, \\ \sum_{r=0}^{i} R_{i-r}^+ R_{r-j}^-, & i \leq j. \end{cases}$$

From the previous theorem we can see that it is possible in principle to calculate the entries of the inverse of a given block Toeplitz operator. However, the theorem does not provide explicit formulas for the factors in the factorisation of the symbol, and by implication we do not have the Fourier coefficients of these factors. Fortunately, there is a way to find explicit formulas for the Wiener-Hopf factorisation, and it relies on a *realization* of the rational matrix symbol. There exist more than one of these realizations of the symbol, and we state a well-known version here, taken from [15].

Theorem 1.1.3. *A rational $m \times m$ matrix function $R(t)$ without poles on $\mathbb{T}$ admits the following representation:*

$$R(t) = I + C(tG - A)^{-1}B, \qquad t \in \mathbb{T}. \tag{1.2}$$

Here G and A are square matrices of the same size $n \times n$, say, $\det(tG - A) \neq 0$ for each $t \in \mathbb{T}$, and B and C are matrices of sizes $n \times m$ and $m \times n$, respectively.

When this realization is used in conjuction with other results (Section XXIV.5 - XXIV.8, [15]), it is possible to arrive at a theorem which gives formulas for the entries of the inverse of a invertible block Toeplitz operator - see Chapter XXIV, Theorem 8.1, [15].

For the purposes of Chapter 2, we will use a different realization of the symbol as in [12]:

$$R(t) = R_0 + tC(I - tA)^{-1}B + \gamma(tI - \alpha)^{-1}\beta, \qquad t \in \mathbb{T}. \tag{1.3}$$

Here, A and α are square matrices of size $n \times n$ and $\nu \times \nu$ respectively, and have the property of stability, i.e., their eigenvalues are contained in the open unit disk. The remaining matrices R_0, B, C, β, γ and I (identity), all represent matrices of appropriate sizes.

In addition to characterizing invertible Toeplitz operators via their symbols and realizations, we can also analyze invertibility via certain algebraic Riccati equations associated with the realization of the symbol.

1.1.1 Algebraic Riccati equations and Toeplitz operator symbols

Algebraic Riccati equations, a special class of matrix equations, arise in many applications and occur in different forms, depending on the applications or theoretical questions considered [24]. We take our definitions from this reference work by P. Lancaster and L. Rodman. A symmetric *discrete algebraic Riccati equation*, or DARE, has the following form:

$$X = A^*XA + Q - (C + B^*XA)^*(R + B^*XB)^{-1}(C + B^*XA),$$

where A, B, C, Q and R are given matrices of sizes $n \times n$, $n \times m$, $m \times n$, $n \times n$ and $m \times m$, respectively. Assuming that R and Q are Hermitian, we want to find a Hermitian solution X to this equation.

With a symmetric symbol and a realization (1.2) thereof, it is possible to associate a symmetric DARE. The solution of the DARE can be related to the invertibility of the Toeplitz operator with associated symbol - see Section 4.7, [21] and the discussion and references given in [12]. This result was improved on in [12], where the rational matrix symbol is not assumed to be symmetric. In that case, the symbol has a different realization (1.3), and its associated algebraic Riccati equation is no longer symmetric. Indeed, it has the form

$$Q = \alpha QA + (\beta - \alpha QB)(R_0 - \gamma QB)^{-1}(C - \gamma QA),$$

and is called a *non-symmetric discrete algebraic Riccati equation*, or NDARE.

1.1.2 The Finite Section Method

The Finite Section Method (FSM) is a strategy to approximate the solution, x, to an infinite system of equations, $Ax = y$, defined by

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots \\ a_{21} & a_{22} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

where

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix}, \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix},$$

with A some block operator defined on $l_2(\mathbb{C}^m)$, i.e., whose entries are $m \times m$ matrices and $x_k, y_k \in \mathbb{C}^m$. The idea of the FSM is to approximate x by solving matrix equations of finite size. To do this, we consider the matrix equation $A_n x_n = y_n$,

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_n^{(1)} \\ \vdots \\ x_n^{(n)} \end{bmatrix} = \begin{bmatrix} y_n^{(1)} \\ \vdots \\ y_n^{(n)} \end{bmatrix}.$$

Here A_n is called the n-th finite section of A. We say that the FSM converges for A, or $A \in \Pi\{A_n\}$, if A_n is invertible for n large enough, and if for each $y = (y_1, y_2, \dots)$ in $l_2(\mathbb{C}^m)$, the vector $x(n) = (x_n^{(1)}, \dots, x_n^{(n)}, 0, 0, \dots)$, where $(x_n^{(1)}, \dots, x_n^{(n)})$ is the solution of the finite system with right hand side $(y_1, \dots, y_n)$, converges in the norm of $l_2(\mathbb{C}^m)$ to x.

Let P_n be the projection on $l_2(\mathbb{C}^m)$ defined by

$$P_n : \{x_1, x_2, \dots\} \longmapsto \{x_1, \dots, x_n, 0, 0, \dots\},$$

where $x_k \in \mathbb{C}^m$. Then

$$A_n = P_n A P_n | \operatorname{Im} P_n.$$

For bounded linear operators on $l_2(\mathbb{C}^m)$, the following holds (Section 6.2, [10]):

$$A \in \Pi\{A_n\} \iff A \text{ is invertible and the sequence } \{A_n\} \text{ is stable.}$$

A sequence is said to be *stable* if A_n is invertible for large n and $\limsup_{n\to\infty} \|A_n^{-1}\| < \infty$.

For Toeplitz operators with continuous matrix valued symbols, (of which our rational matrix function symbols are a subset), we have the following theorem (Theorem 6.9, [10]).

Theorem 1.1.4. *Let the matrix-valued symbol $a(t)$ of the Toeplitz operator, $T(a)$, be continuous on $\mathbb{T}$. Then $\{T_n(a)\}$ is stable if and only if $T(a)$ and its block transpose, $T(\tilde{a})$, are invertible.*

Observe that the matrix-valued function $\tilde{a}(t)$ is defined as the symbol of the block transpose of $T(a)$.

This leads to the fact that

$$T(a) \in \Pi\{T_n(a)\} \iff T(a) \text{ and } T(\tilde{a}) \text{ are invertible.}$$

For an in-depth discussion of the FSM, including more general classes of symbols, see for instance the book [9], chapter 7.

1.1.3 Fredholmness

Let $A : X \longmapsto Y$ be a bounded linear operator acting between two Banach spaces, X and Y. The operator A is said to be *Fredholm* if $\operatorname{Im} A$ is closed and the numbers $n(A) = \dim\ker A$ and $d(A) = \operatorname{codim}\operatorname{Im} A$ are finite. As usual, 'dim ker' denotes the dimension of the subspace of X, formed by the kernel of A, while 'codim Im' denotes the dimension of the subspace, say Y', where $Y = \operatorname{Im} A \oplus Y'$. The *index* of A is then defined as

$$\operatorname{ind}(A) = n(A) - d(A).$$

Toeplitz operators have a unique relationship with the Fredholm property, and again, the symbol of the operator is definitive in this regard. The following theorem from [15] (see also [9, 16]) formalizes this connection.

Theorem 1.1.5. *Let T be a block Toeplitz operator on $l_2(\mathbb{C}^m)$, defined by a rational matrix function $R(t)$. Assume that $\det R(t) \neq 0$ for all $t \in \mathbb{T}$, and let*

$$R(t) = R_-(t)([t^{\kappa_j}\delta_{ij}]_{i,j=1}^m)R_+(t), \quad t \in \mathbb{T}$$

be a Wiener-Hopf factorisation of $R(t)$ relative to $\mathbb{T}$. Then T is a Fredholm operator with

$$n(T) = \sum_{\kappa_j \leq 0} -\kappa_j, \qquad d(T) = \sum_{\kappa_j \geq 0} \kappa_j.$$

1.2 Toeplitz matrices

In chapters 3 and 4, we will mostly be concerned with Toeplitz matrices in the finite dimensional domain, although the study of their properties is often related to their infinite counterparts. As with Toeplitz operators, many properties of Toeplitz matrices are directly related to its corresponding symbol.

As the title of this thesis suggests, we are specifically interested in norms and singular values, and we study their evolution as the matrix sizes grow to infinity, i.e., asymptotically. A lot is known about both the norms and singular values of Toeplitz matrices, and we do not aim to add depth to the understanding as such (see for instance the books [5] and [10]). However, we will use their properties to prove interesting results regarding the new class of T-gen matrices, whose definition was inspired by a statistical problem - see chapter 2 for details. We also note that these T-gen matrices, $X_n = T_n + f_n \cdot (T_n^{-1})^*$, have entries that are dependent on their size, due to the presence of the sequence f_n. Certain finite rank perturbations will also be introduced under which our main results will remain invariant.

It is also important to keep in mind that T_n is a banded Toeplitz matrix and its associated symbol can be represented by a finite series, $b(t) = \sum_{-r}^{r} b_j t^j$, here referred to as a Laurent polynomial. This assumption unlocks certain results, not available for Toeplitz matrices whose infinite counterparts have general symbols.

1.2.1 Singular values

The singular values of any $m \times n$ matrix, say A, are defined in terms of its *singular value decomposition* (SVD). It can be shown that any matrix has a SVD, and it takes the form

$$A = UDV^*,$$

where U and V are unitary matrices of size $m \times m$ and $n \times n$ respectively. D denotes a diagonal $m \times n$ matrix whose diagonal entries are the nonnegative square roots of the eigenvalues of AA^*, and these values are called the *singular values* of A. The columns of U are the eigenvectors of AA^*, and the columns of V are the eigenvectors of A^*A. The book [18] provides a thorough development of the SVD.

In the case of Toeplitz matrices, a lot of work has been done on describing the behaviour and distribution of their singular values. Some of these results have been included in the reference works, [5, 10], of which the former restricts itself to treating banded Toeplitz matrices, which is of particular importance for our investigations here. We state a few key results here, starting with a particularly elegant result that has become known as the *splitting phenomenon* ([31, 32]). We note that we index singular values in decending order, i.e., σ_1 is the maximal singular value with σ_n the minimal singular value.

Theorem 1.2.1. *Let $b(t)$ be a Laurent polynomial and suppose $T(b)$ is Fredholm of index $k \in \mathbb{Z}$. Then the smallest $|k|$ singular values, $\sigma_n(T_n(b)) \leq \sigma_{n-1} \leq \cdots \leq \sigma_{n-k+1}(T_n(b))$, go to zero with exponential speed,*

$$\sigma_{n-j}(T_n(b)) = O(e^{-\alpha n}), \qquad 0 \leq j \leq k-1.$$

Here $\alpha > 0$ is dependent on the the symbol $b(t)$. The remaining singular values are bounded from below by a positive constant, d (dependent on $b(t)$), for sufficiently large n,

$$\sigma_{n-j}(T_n(b)) \geq d > 0, \qquad k \leq j \leq n-1.$$

We know from Theorem 1.1.5, that the previous theorem only applies to symbols that do not vanish on the unit circle. Our banded Toeplitz matrices, T_n, that generate the class of T-gen matrices are assumed to have symbols that do vanish on $\mathbb{T}$, implying that $T(b)$ is not Fredholm, and this has a significant effect on the behaviour of T_n's smallest singular values ([4, 5]).

Theorem 1.2.2. *Let $b(t)$ be a non-constant Laurent polynomial and suppose $T(b)$ is not Fredholm. Let $\alpha \in \mathbb{N}$ be the maximal order of the zeros of $|b(t)|$ on $\mathbb{T}$. Then for each natural number $k \geq 1, \sigma_{n-k} = O(1/n^\alpha)$ as $n \to \infty$.*

Here the order of the zero, say α_0, indicate the smallest natural number such that

$$\frac{d^{\alpha_0}}{dt^{\alpha_0}} b(t_0) \neq 0,$$

where $b(t_0) = 0$.

Interestingly, Fredholmness does not play a role in the behaviour of the maximal singular values of banded Toeplitz matrices ([4, 5]):

Theorem 1.2.3. *Let $b(t)$ be a non-constant Laurent polynomial. Denote by $\beta \in \mathbb{N}$ the maximal order of the zeros of $\|b\|_\infty - |b|$ on the unit circle. Then for each $k \geq 0$,*

$$\|b\|_\infty - D_k \frac{1}{n^\beta} \leq \sigma_k \leq \|b\|_\infty$$

with some constant $D_k \in (0, \infty)$ independent of n.

These theorems show the behaviour of extreme singular values in the banded case, but how do the remaining ones behave, or can we say something about their distribution? The answer to this is contained in the Avram-Parter theorem, ([25, 1]), of which we give a slightly different formulation which is based on [41].

Theorem 1.2.4. *Let $b(t)$ be a Laurent polynomial and let $f : \mathbb{R} \longmapsto \mathbb{C}$ be a function with compact support. If f is continuous or of bounded variation, then*

$$\lim_{n \longrightarrow \infty} \frac{1}{n} \sum_{k=1}^{n} f(\sigma_{n-k}(T_n(b))) = \frac{1}{2\pi} \int_0^{2\pi} f(|b(e^{i\theta})|) d\theta.$$

We note that this theorem applies to non-banded and matrix valued symbols as well, see for instance [10]. The approach followed by Zizler, Zuidwijk, Taylor and Arimoto in [41] to prove this theorem, using functions of bounded variation, leads to a very useful result. Indeed, the following lemma is the most important result used to arrive at the estimates we achieve in chapter 3.

Lemma 1.2.5. *Let $b(t)$ be a Laurent polynomial of the form $b(t) = \sum_{j=-r}^{r} b_j t^j$, $t \in \mathbb{T}$. If $E \subset \mathbb{R}$ is any segment, then*

$$|N_n(E) - n\mu(E)| \leq 14r \qquad \text{for all} \quad n \geq 1,$$

where

$$N_n(E) = \sum_{k=1}^{n} \chi_E(\sigma_k(T_n(b)))$$

is the number of singular values of $T_n(b)$ in E and

$$\mu(E) = \frac{1}{2\pi} \int_0^{2\pi} \chi_E(|b(e^{i\theta})|) d\theta = \frac{1}{2\pi} |\{t \in \mathbb{T} : |b(t)| \in E\}|,$$

with $|.|$ denoting the Lebesgue measure on the unit circle.

1.2.2 Eigenvalues

As for the singular values of Toeplitz matrices, much can be said of its eigenvalues and their distribution. In the simple case of tridiagonal Toeplitz matrices, we have explicit formulas for both the eigenvalues and eigenvectors as given by the following theorem from [5].

Theorem 1.2.6. *The eigenvalues of $T_n(b)$ $(b(t) = b_0 + b_1 t + b_{-1}t^{-1})$ are*

$$\lambda_j = b_0 + 2\sqrt{b_1 b_{-1}} \cos \frac{\pi j}{n+1} \qquad (j = 1, \ldots, n),$$

and an eigenvector for λ_j is $x_j = (x_1^{(j)}, \ldots, x_n^{(j)})^T$ with

$$x_k^{(j)} = \left(\sqrt{\frac{b_1}{b_{-1}}}\right)^k \sin \frac{k\pi j}{n+1} \qquad (k = 1, \ldots, n).$$

Assuming Hermitian banded Toeplitz matrices, the behaviour of their eigenvalues closely resembles that of its singular values. We know that $T_n(b)$ is Hermitian, if and only if $b(t)$ is real-valued. Let $m = \min_{t\in\mathbb{T}} b(t)$ and $M = \max_{t\in\mathbb{T}} b(t)$. If we put $a(t) = b(t) - m$, it turns out that the eigenvalues of $T_n(a)$ coincide with its singular values, and this leads to the similarity of the behaviour of the eigenvalues of $T_n(b)$ with its singular values. Compare the following theorem ([5], [33]) with Theorems 1.2.2 and 1.2.3.

Theorem 1.2.7. *Let $b(t)$ be a non-constant real-valued Laurent polynomial, let $\mathcal{R}(b) = [m, M]$, and denote by 2α and 2β the maximal order of the zeros of $b(t) - m$ and $M - b(t)$, respectively. Then for each fixed k,*

$$\lambda_{n-k}(T_n(b)) - m \simeq \frac{1}{n^{2\alpha}}, \qquad M - \lambda_k(T_n(b)) \simeq \frac{1}{n^{2\beta}},$$

where the notation $x_n \simeq y_n$ means that there exist constants $C_1, C_2 \in (0, \infty)$ such that $C_1 y_n \leq x_n \leq C_2 y_n$ for all sufficiently large n.

In addition to the similarity of the extreme eigenvalues and singular values of $T_n(b)$, the Avram-Parter theorem (Theorem 1.2.4) remains true for real-valued b with the singular values, $\sigma_{n-k}(T_n(b))$, replaced by eigenvalues, $\lambda_{n-k}(T_n(b))$ (Corollary 10.5, [5]).

For general non-Hermitian banded Toeplitz matrices, the distribution of the eigenvalues is more involved and requires additional background material that falls outside the scope of this thesis. We refer the reader to [5] and the references contained therein for a thorough analysis of the topic.

A substantial amount of work has also been done on the effects of perturbing a small number of entries of Toeplitz matrices, including changes in the spectrum. In chapter 4 we follow the exposition of section 14.1 in [5] to arrive at explicit formulas for the eigenvalues and eigenvectors of a particular perturbed Toeplitz matrix. This result is then used to explain a permutation phenomenon arising in the singular vectors of a subclass of T-gen matrices.

1.2.3 Invertibility

Invertibility of Toeplitz matrices is very important in our research since the study of the norms of the inverses of T-gen matrices directly require the invertibility of the Toeplitz matrices that generate them (see Theorem 3.1.1). Criteria for invertibility involves the symbol of the associated operator, and this is evident in the following theorem which is originally due to Baxter [2] and Reich [30] for the case l_1, with generalisations proven later by Gohberg and Fel'dman.

Let c_0 denote the closed subspace of l_∞ consisting of sequences converging to zero, and let the *Wiener algebra* W be the set of all functions $a : \mathbb{T} \longrightarrow \mathbb{C}$ of the form $a(t) = \sum_{n=-\infty}^{\infty} a_n t^n$ with $\sum_{n=-\infty}^{\infty} |a_n| < \infty$.

Theorem 1.2.8. *Let X be one of the spaces c_0 or l_p, ($1 \leq p \leq \infty$), and let $a \in W$. Then,*

$$\lim_{n\longrightarrow\infty} \|T_n^{-1}(a)\| < \infty \quad \text{if } T(a) \text{ is invertible,}$$
$$\lim_{n\longrightarrow\infty} \|T_n^{-1}(a)\| = \infty \quad \text{if } T(a) \text{ is not invertible.}$$

Therefore, $\{T_n(a)\}_{n=1}^{\infty}$ is stable if and only if a has no zeros on $\mathbb{T}$ and admits a right canonical factorisation.

This theorem implicitly gives a criterium for the invertiblily of finite sections of Toeplitz operators, provided they are invertible. When considering T-gen matrices, we have mentioned that the banded Toeplitz matrices that generate them have associated symbols that do vanish on the unit circle. Consequently, this theorem does not give us conditions under which finite sections of non-Fredholm Toeplitz operators will be invertible, but it does tell us that if $T_n^{-1}(a)$ exists for all n sufficiently large, $\|T_n^{-1}(a)\|$ is unbounded. Fortunately there is another useful result that determines when these matrices are invertible - see [3] or Theorem 4.27 from [5]. We will state it here, after the appropriate notation.

Let $\mathcal{R}(a) = a(\mathbb{T})$ denote the range of the symbol a, $\operatorname{conv}\mathcal{R}(a)$ the convex hull of $\mathcal{R}(a)$, and $\partial \operatorname{conv}\mathcal{R}(a)$ the boundary of $\operatorname{conv}\mathcal{R}(a)$.

Theorem 1.2.9. *Suppose $a \in W$ does not vanish identically and $\mathcal{R}(a)$ is not a line segment containing the origin in its interior. If*

$$0 \notin \operatorname{conv}\mathcal{R}(a) \quad \text{or} \quad 0 \in \partial \operatorname{conv}\mathcal{R}(a),$$

then $T_n(a)$ is invertible for all $n \geq 1$.

Since the symbol associated with our Toeplitz matrices contains at least one zero, but is not identically zero, it satisfies the assumptions of this theorem. However, because of the presence of the zero, the first condition is never satisfied and we are left with testing the second condition. It is important to note that this theorem does not give both a necessary and sufficient condition for invertibility and therefore does not give a complete characterization of invertibility. In chapter 3 we will discuss an example of a banded symbol with a zero that does not satisfy the second condition, and hence we cannot say anything about its invertibility from this theorem.

On the other hand, given that a Toeplitz matrix is invertible, the Gohberg-Semencul-Heinig formulas provide a quick way of calculating its inverse [20]. Consider again the finite Toeplitz $n \times n$ matrix $T_n(a) = \left[a_{i-j}\right]_{i,j=0}^{n-1}$. These formulas describe the inverse of the Toeplitz matrix in terms of the solutions of four matrix-vector equations. To be precise, let the vectors x, y, u, v be solutions to

$$T_n(a)x = e_1, \quad T_n(a)y = e_n, \quad T_n(a)^T u = e_n, \quad T_n(a)^T v = e_1.$$

Denote by $\operatorname{Toep}(x)$ the upper triangular Toeplitz matrix with x^T as its first row. Also denote by S the $n \times n$ forward shift matrix, that is, the matrix $\operatorname{Toep}(e_2)^T$. Then the first

coordinates of x and v are equal, so $x_1 = v_1$, and likewise, the last coordinates of y and u are equal, so $y_n = u_n$, and the inverse of $T_n(a)$ is given by

$$T_n(a)^{-1} = \text{Toep}\,(x)^T v_1^{-1}\text{Toep}\,(v) - \text{Toep}\,(Sy)^T y_n^{-1}\text{Toep}\,(Su).$$

The above formulas do not only apply to the scalar case, but also to block Toeplitz matrices. In the scalar case, the Gohberg-Semencul formulas are adequate ([20]). Also see [**?**].

1.2.4 Norms

The norm of an operator, A, on l_p sequence spaces is defined as

$$\|A\|_p = \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p},$$

where $\|x\|_p^p := \sum_{n=0}^{\infty} |x_n|^p$ for $1 \leq p < \infty$ and $\|x\|_\infty := \sup_{n\geq 0} |x_n| < \infty$. From chapter 5 of [5], or originally in [8], we know that

$$\lim_{n\longrightarrow\infty} \|T_n(b)\|_p = \|T(b)\|_p$$

and more precisely,

$$\|T(b)\|_p = \|T_n(b)\|_p + O(\frac{1}{n}),$$

for $1 < p < \infty$. As discussed in the previous section, the finite matrix $T_n(b)$ will be identified with its associated finite section of $T(b)$, defined on l_p.

For our purposes, we will only be interested in the case $p = 2$. In this case, $\|\cdot\|_2$ coincides with the *spectral* norm of a matrix, i.e., the maximum singular value ([19]). Thus, $\|A\|_2 = \max_n\{\sigma_n(A)\}$. We can then immediately see from Theorem 1.2.3, that the $O(1/n)$ estimate given here for general values of p, can be greatly improved for the case $p = 2$, since $\sigma_1(T_n(b)) = \|T_n(b)\|_2$.

We are also interested in the norms of the inverses of Toeplitz matrices. Since $\|T_n^{-1}(b)\|_2 = 1/\sigma_n T_n(b)$, Theorem 1.2.2 can directly be applied to estimate the growth of $\|T_n^{-1}(b)\|_2$.

1.2.5 Variable coefficient Toeplitz matrices

From the definition of our T-gen matrices, $X_n = T_n + f_n \cdot (T_n^{-1})^*$, it is clear that its entries depend on the sequence f_n, which changes as the size of X_n grows. One might be tempted to use the phrase *variable coefficient* to describe these matrices, but we steer clear from such a description as it has been widely used in the literature, e.g. [34, 35, 11, 36, 6, 7] to describe specific classes of Toeplitz matrices with variable coefficients that differ greatly from T-gen matrices. In [6] for example, they consider matrices defined as follows. Let $a : [0,1] \times [0,1] \times \mathbb{T} \longrightarrow \mathbb{C}$ be a continuous function with Fourier representation in its last variable,

$$a(x,y,t) = \sum_{n=-\infty}^{\infty} \hat{a}(x,y)t^n, \quad \hat{a}(x,y) = \int_{\mathbb{T}} a(x,y,t)t^{-n}\frac{|dt|}{2\pi}.$$

The $(N+1) \times (N+1)$ variable coefficient matrix generated by a is then defined as

$$A_N(a) = \left(\hat{a}_{j-k} \left(\frac{j}{N}, \frac{k}{N} \right) \right)_{j,k=0}^{\infty}.$$

Since the inverses of Toeplitz matrices are generally not Toeplitz themselves, it is easy to construct a myriad of examples of T-gen matrices that are not Toeplitz themselves. If in addition, we added particular finite rank perturbations (see chapter 3), they enlarge the class even further and we are clearly not in the realm of variable coefficient Toeplitz matrices anymore.

1.2.6 Main results

Here follows a list of the main results achieved.

- **Chapter** 2**:** Theorem 2.1.5 shows that the inverse of any invertible block Toeplitz operator, T, with rational matrix symbol can be approached in norm by the product of the inverses of related Toeplitz operators. The Toeplitz operators in the product are related to the original operator via their symbols. Indeed, their symbols converge to the inverses of the factors of the right canonical factorisation of the symbol of T.

- **Chapter** 3**:** Theorem 3.1.1 and Theorem 3.1.2 provide growth estimates for the norm of the inverse of a sequence of T-gen matrices and related finite rank perturbations - see [26]. Theorem 3.4.1 significantly improves on these results and states the growth of these norms via an order estimate [27].

- **Chapter** 4**:** Theorem 4.1.1 provides explicit formulas for the eigenvectors and eigenvalues of a special perturbed tridiagonal Toeplitz matrix. This result is then used in Section 4.3, which explains a curious permutation phenomenon regarding the singular vector entries of a special class of T-gen matrices. In addition, the norms of this special class of matrices as well as their inverses are exactly determined for any size n [28].

Chapter 2

Block Toeplitz Operators and the NDARE

2.1 Introduction

In [12] a connection was made between rational matrix functions on the unit circle and a related NDARE (2.3). Here the rational matrix function serves as the symbol of a block Toeplitz operator. Necessary and sufficient conditions were found regarding the right canonical factorization of the symbol, and a unique stable solution of the corresponding NDARE. Hence, the invertibility of a Toeplitz operator with rational matrix symbol can be related to the existence of a unique stabilizing solution of (2.3). Also, the FSM can be applied to provide a constructible sequence that converges to the unique stable solution of the NDARE. This sequence, however, can be constructed independently of the FSM as well.

Our main goal here will be the uniform approximation of the inverses of these block Toeplitz operators. Let

$$T = \begin{bmatrix} R_0 & R_{-1} & R_{-2} & \dots \\ R_1 & R_0 & R_{-1} & \dots \\ R_2 & R_1 & R_0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{2.1}$$

be a block Toeplitz operator whose entries are $m \times m$ matrices, which are the Fourier coefficients of a rational matrix function $R(t) = \sum_{i=-\infty}^{\infty} R_i t^i$. This matrix function is called the corresponding symbol. The operator T is defined on $l_p(\mathbb{C}^m)$ $(1 \leq p < \infty)$, i.e., the vector-valued l_p spaces. Theorem 1.1.2 tells us that T is invertible if and only if $R(t)$ has a right canonical factorisation and $\det R(t) \neq 0, t \in \mathbb{T}$. Such a factorisation can be expressed as $R(t) = \psi(t)\theta(t)$ on the unit circle $\mathbb{T}$, where $\theta(t)$ and $\theta(t)^{-1}$ have no poles on the unit disk $\{t \in \mathbb{C} : |t| \leq 1\}$, while $\psi(t)$ and $\psi(t)^{-1}$ have no poles in $\{t \in \mathbb{C} : |t| \geq 1\}$. In addition, $T^{-1} = T_{\theta^{-1}} T_{\psi^{-1}}$, where $T_{\theta^{-1}}$ and $T_{\psi^{-1}}$ are both block Toeplitz operators with corresponding symbols. As in the introduction, Toeplitz operator will always imply block Toeplitz operator in this chapter.

As mentioned in the introduction, we will employ a special representation, or realiza-

tion of the symbol of our Toeplitz operator. We state it again for convenience:

$$R(t) = R_0 + tC(I - tA)^{-1}B + \gamma(tI - \alpha)^{-1}\beta. \tag{2.2}$$

Here, A and α are square matrices of size $n \times n$ and $\nu \times \nu$ respectively, and have the property of stability, i.e., their eigenvalues are contained in the open unit disk. The remaining matrices R_0, B, C, β, γ and I (identity), all represent matrices of appropriate sizes. We shall refer to (2.2) as a **stable representation** of $R(t)$. With (2.2) we associate the NDARE

$$Q = \alpha QA + (\beta - \alpha QB)(R_0 - \gamma QB)^{-1}(C - \gamma QA). \tag{2.3}$$

Q is said to be a **stabilizing solution** to the above equation if the matrix $R_0 - \gamma QB$ is invertible, Q is a solution to (2.3) and both

$$A_\circ = A - B(R_0 - \gamma QB)^{-1}(C - \gamma QA) \quad \text{and} \tag{2.4}$$
$$\alpha_\circ = \alpha - (\beta - \alpha QB)(R_0 - \gamma QB)^{-1} \tag{2.5}$$

are stable.

We now state a few results from [12] which will be important for our study.

Theorem 2.1.1. *Let $R(t)$ be a $m \times m$ rational matrix function with no poles on $\mathbb{T}$ with (2.2) as a stable representation. Then $R(t)$ admits a right canonical factorization with respect to $\mathbb{T}$ if and only if the NDARE (2.3) has a stabilizing solution Q, and in that case a canonical factorization $R(t) = \psi(t)\theta(t)$ is obtained by taking*

$$\theta(t) = D + tC_\circ(I - tA)^{-1}B, \quad \psi(t) = \delta + \gamma(tI - \alpha)^{-1}\beta_\circ, \tag{2.6}$$

where

$$C_\circ = \delta^{-1}(C - \gamma QA), \quad \beta_\circ = (\beta - \alpha QB)D^{-1},$$

and δ and D are any invertible matrices satisfying $\delta D = R_0 - \gamma QB$. Moreover, the inverses of the factors are given by

$$\theta^{-1}(t) = D^{-1} - tD^{-1}C_\circ(I - tA_\circ)^{-1}BD^{-1}$$
$$\psi^{-1}(t) = \delta^{-1} - t\delta^{-1}\gamma(It - \alpha_\circ)^{-1}\beta_\circ\delta^{-1},$$

where $A_\circ$ and $\alpha_\circ$ are given by (2.4) and (2.5) respectively. Finally, if the NDARE (2.3) has a stabilizing solution, then this solution is unique and given by

$$Q = [\beta\ \alpha\beta\ \alpha^2\beta\ \ldots]T^{-1}\begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \end{bmatrix}, \tag{2.7}$$

where T is the Toeplitz operator (2.1) with symbol $R(t)$.

In what follows, we will write $Q = \omega T^{-1} W$ with

$$\omega = [\beta \ \alpha\beta \ \alpha^2\beta \cdots], \qquad W = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \end{bmatrix}.$$

We also define

$$\omega_n = [\beta \ \alpha\beta \ \cdots \alpha^{n-1}\beta], \qquad W_n = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}.$$

Lemma 2.1.2. *Assume the n-th finite section T_n of T is invertible, and put $Q_n = \omega_n T_n^{-1} W_n$. Then T_{n+1} is invertible if and only if $R_0 - \gamma Q_n B$ is invertible, and in that case the matrix $Q_{n+1} = \omega_{n+1} T_{n+1}^{-1} W_{n+1}$ is given by*

$$Q_{n+1} = \alpha Q_n A + (\beta - \alpha Q_n B)(R_0 - \gamma Q_n B)^{-1}(C - \gamma Q_n A). \tag{2.8}$$

Proposition 2.1.3. *Let $R(t)$ be given by the stable representation* (2.2)*, and consider the Ricatti difference equation* (2.8)*. Assume the Finite Section Method converges for the Toeplitz operator with symbol $R(t)$. Then there exists a positive integer k such that the following holds*

(i) *T_n, the n-th finite section of T, is invertible for all $n \geq k$;*

(ii) *$R_0 - \gamma Q_n B$ is invertible where*

$$Q_n = \omega_n T_n^{-1} W_n$$

is the solution to equation (2.8) *for all $n \geq k$. (Here the subscript n denotes the n-th section or truncation of the vectors and operator as specified in Theorem 2.1.1).*

(iii) *Q_n converges to Q and $R_0 - \gamma QB$ is invertible;*

(iv) *the matrices $\alpha_\circ$ and $A_\circ$ are stable.*

In this case, Q is the stabilizing solution to the Ricatti equation (2.2)*.*

Remark 2.1.4.

In [12] the results were stated for the case where the Toeplitz operator was defined on $l_2(\mathbb{C}^m)$. However, these results go through trivially on $l_p(\mathbb{C}^m)$ $(1 \leq p < \infty)$ since no special Hilbert space properties were employed and most of the proofs play out in the finite-dimensional domain. See [13], chapter VIII, sections 3-5 for relevant results. We

now give the main results of this chapter:

Let Q_n be any sequence of matrices converging to the stabilizing solution Q of the NDARE. Put

$$\begin{aligned}
A_{\circ n} &= A - B(R_0 - \gamma Q_n B)^{-1}(C - \gamma Q_n A),\\
\alpha_{\circ n} &= \alpha - (\beta - \alpha Q_n B)(R_0 - \gamma Q_n B)^{-1}\gamma,\\
\beta_{\circ n} &= (\beta - \alpha Q_n B)D_n^{-1},\\
C_{\circ n} &= \delta_n^{-1}(C - \gamma Q_n A).
\end{aligned}$$

Here, δ_n and D_n are chosen as the identity and $R_0 - \gamma Q_n B$, respectively. Define rational matrix functions θ_n and ψ_n via their inverses, θ_n^{-1} and ψ_n^{-1}, analogous to θ^{-1} and ψ^{-1}:

$$\begin{aligned}
\theta_n^{-1}(t) &= D_n^{-1} - tD_n^{-1}C_{\circ n}(I - tA_{\circ n})^{-1}BD_n^{-1}\\
\psi_n^{-1}(t) &= I - t\gamma(t - \alpha_{\circ n})^{-1}\beta_{\circ n},
\end{aligned}$$

and introduce also

$$S_n = T_{\theta_n^{-1}}T_{\psi_n^{-1}}, \tag{2.9}$$

where $T_{\theta_n^{-1}}$ and $T_{\psi_n^{-1}}$ are the Toeplitz operators with symbol θ_n^{-1} and ψ_n^{-1}, respectively.

Theorem 2.1.5. *Given an invertible Toeplitz operator T with a rational matrix function symbol. Let Q be the stabilizing solution of NDARE, and let Q_n be any sequence of matrices converging to Q. Then, with S_n given by* (2.9) *we have*

$$\lim_{n\to\infty} \|S_n - T^{-1}\| = 0.$$

The following proposition gives a relation between the convergence rate of S_n to T^{-1} and the convergence rate of Q_n to Q.

Proposition 2.1.6. *The inequality $\|S_n - T^{-1}\| \le c\|Q_n - Q\|$ holds for n large enough, where c is a positive constant.*

Returning to the Q_n given in Lemma 2.1.2, for which we know that $Q_n \to Q$, the convergence rate is known to be quadratic in the symmetric case (see Theorem 13.2.1, [24]). In contrast to the general case of a non-symmetric Riccati difference equation (NRDE) as in (2.8), we can only show linear convergence.

Proposition 2.1.7. *For the non-symmetric Riccati difference equation* (2.8) *we have*

$$\|Q_{n+1} - Q_n\| \le k\|Q_n - Q_{n-1}\|,$$

where k is some positive constant.

We will also show that a quadratic convergence rate does not apply in general to equation (2.8). This is in contrast to symmetric discrete algebraic Riccati equations where quadratic convergence holds ([24]).

2.2 Approximation of the inverse Toeplitz operator

From Theorem 2.1.1 we know that $T^{-1} = T_{\theta^{-1}} T_{\psi^{-1}}$ when Q in (2.7) is a stabilizing solution to the NDARE (2.3). Taking a closer look at Theorem 2.1.1, one sees that we have explicit formulas for the symbols θ and ψ as well as their inverses. Also notice that Q is present in them.

Assume Q_n is any sequence of operators (with appropriate size) which converge to the stabilizing solution Q. We can define functions, θ_n^{-1} and ψ_n^{-1}, analogous to θ^{-1} and ψ^{-1}, and hope that they can be used as symbols for Toeplitz operators, $T_{\theta_n^{-1}}$ and $T_{\psi_n^{-1}}$, respectively. We then investigate the convergence of the sequence

$$\|T_{\theta_n^{-1}} T_{\psi_n^{-1}} - T^{-1}\|.$$

Now, define

$$S_n := T_{\theta_n^{-1}} T_{\psi_n^{-1}},$$

with

$$\begin{aligned} \theta_n^{-1}(t) &= D_n^{-1} - tD_n^{-1} C_{\circ n} (I - tA_{\circ n})^{-1} B D_n^{-1} \quad \text{and} \\ \psi_n^{-1}(t) &= I_n - t\gamma(t - \alpha_{\circ n})^{-1} \beta_{\circ n}, \end{aligned}$$

where

$$\begin{aligned} A_{\circ n} &= A - B(R_\circ - \gamma Q_n B)^{-1}(C - \gamma Q_n A), \\ \alpha_{\circ n} &= \alpha - (\beta - \alpha Q_n B)(R_\circ - \gamma Q_n B)^{-1}\gamma, \\ \beta_{\circ n} &= (\beta - \alpha Q_n B) D_n^{-1}, \\ C_{\circ n} &= (C - \gamma Q_n A). \end{aligned}$$

Here, D_n is chosen as $R_0 - \gamma Q_n B$. With these definitions in hand, we are ready to prove our main theorem.

Proof of Theorem 2.1.5. We need to check that S_n is well-defined, i.e., that θ_n^{-1} and ψ_n^{-1} are themselves well-defined and have no poles on the unit circle $\mathbb{T}$. Since we know that $R_0 - \gamma QB$ is invertible, there exists a k such that $R_\circ - \gamma Q_n B$ is invertible for all $n \geq k$. To see this, note that $\det(R_\circ - \gamma QB) \neq 0$ and that the determinant is a continuous function. It follows that $A_{\circ n}, \alpha_{\circ n}$ and $\beta_{\circ n}$ are well-defined for n large enough. The choice of δ_n also guarantees that $C_{\circ n}$ is well-defined.

What remains is to show that θ_n^{-1} and ψ_n^{-1} have no poles on $\mathbb{T}$ and we prove this by showing that $A_{\circ n}$ and $\alpha_{\circ n}$ are stable for n large enough.

It's easy to see that $A_{\circ n} \to A_\circ$ pointwise, and hence in norm, since they are bounded linear operators of finite rank. Suppose that for all $m \in \mathbb{N}$ there exists a $n \geq m$ such that $A_{\circ n}$ is not stable. This implies the existence of a sequence of eigenvalues λ_n, with $|\lambda_n| \geq 1$ such that $A_{\circ n} x_n = \lambda_n x_n$. Without loss of generality, we can choose the x_n's such that

$\|x_n\| = 1$ for all n. Since the unit ball is compact in a finite dimensional space, x_n has a convergent subsequence x_{n_k} with limit denoted by $x_\circ$. Also, the eigenvalues λ_n exist in a compact set. Indeed, $|\lambda_n| \leq \|A_{\circ n}\|$ and $\|A_{\circ n}\|$ is bounded. Now consider the sequence of eigenvalues λ_{n_k}, corresponding to the sequence x_{n_k}. Again we can find a convergent subsequence $\lambda_{n_{k_l}}$ with limit, say $\lambda_\circ$, with its corresponding sequence of eigenvectors $x_{n_{k_l}}$ converging to $x_\circ$. Therefore we have $\lambda_{n_{k_l}} x_{n_{k_l}} \to \lambda_\circ x_\circ$ and $A_{\circ n_{k_l}} x_{n_{k_l}} \to A_\circ x_\circ$ and together they give us $A_\circ x_\circ = \lambda_\circ x_\circ$. However, $A_\circ$ is stable, and $|\lambda_\circ| \geq 1$ since $|\lambda_n| \geq 1$ which is a contradiction, and consequently our assumption is false. This proves the stability of $A_{\circ n}$ for all n sufficiently large. A completely analogous proof provides us with the stability of $\alpha_{\circ n}$, leading to a well-defined S_n.

As for $A_{\circ n}$, its equally straightforward to see that $\alpha_{\circ n} \to \alpha_\circ$, $\beta_{\circ n} \to \beta_\circ$ and $C_{\circ n} \to C_\circ$. The stage is now set to estimate $\|S_n - T^{-1}\|$, but first, a few more definitions.

Define $\tilde{\theta}_n^{-1}(t) := I - tC_{\circ n}(I - tA_{\circ n})^{-1}BD_n^{-1}$ and $\tilde{\theta}^{-1}(t) := I - tC_\circ(I - tA_\circ)^{-1}BD^{-1}$. Then $\theta_n^{-1} = D_n^{-1}\tilde{\theta}_n^{-1}$ and $\theta^{-1} = D^{-1}\tilde{\theta}^{-1}$.

We now test the convergence of S_n:

$$
\begin{aligned}
&\|S_n - T^{-1}\| \\
=&\|T_{\theta_n^{-1}} T_{\psi_n^{-1}} - T_{\theta^{-1}} T_{\psi^{-1}}\| \\
=&\|(T_{\theta_n^{-1}} - T_{\theta^{-1}}) T_{\psi_n^{-1}} + T_{\theta^{-1}}(T_{\psi_n^{-1}} - T_{\psi^{-1}})\| \\
=&\|(T_{D_n^{-1}} T_{\tilde{\theta}_n^{-1}} - T_{D^{-1}} T_{\tilde{\theta}^{-1}}) T_{\psi_n^{-1}} + T_{\theta^{-1}}(T_{\psi_n^{-1}} - T_{\psi^{-1}})\| \\
=&\|[(T_{D_n^{-1}} - T_{D^{-1}}) T_{\tilde{\theta}_n^{-1}} + T_{D^{-1}}(T_{\tilde{\theta}_n^{-1}} - T_{\tilde{\theta}^{-1}})] T_{\psi_n^{-1}} + T_{\theta^{-1}}(T_{\psi_n^{-1}} - T_{\psi^{-1}})\| \\
\leq&[\|D_n^{-1} - D^{-1}\| \|\tilde{\theta}_n^{-1}\|_\infty + \|D^{-1}\| \|\tilde{\theta}_n^{-1} - \tilde{\theta}^{-1}\|_\infty] \|\psi_n^{-1}\|_\infty + \|\theta^{-1}\|_\infty \|\psi_n^{-1} - \psi^{-1}\|_\infty. \quad (2.10)
\end{aligned}
$$

It remains to show that $\|\psi_n^{-1} - \psi^{-1}\|_\infty$ and $\|\tilde{\theta}_n^{-1} - \tilde{\theta}^{-1}\|_\infty$ go to zero as n goes to infinity since their convergence implies that $\|\psi_n^{-1}\|_\infty$ and $\|\tilde{\theta}_n^{-1}\|_\infty$ are bounded respectively. Also, $\|D_n^{-1} - D^{-1}\|$ converges to zero by construction. First we check the convergence of ψ_n^{-1}:

$$
\begin{aligned}
&\|\psi_n^{-1} - \psi^{-1}\|_\infty \\
=&\sup_{|t|=1} \|[I - \gamma(t - \alpha_{\circ n})^{-1}\beta_{\circ n}] - [I - \gamma(t - \alpha_\circ)^{-1}\beta_\circ]\| \\
\leq&\|\gamma\| \sup_{|t|=1} \|[(t - \alpha_{\circ n})^{-1} - (t - \alpha_\circ)^{-1}]\beta_{\circ n} + (t - \alpha_\circ)^{-1}(\beta_{\circ n} - \beta_\circ)\| \\
=&\|\gamma\| \sup_{|t|=1} \|(t - \alpha_{\circ n})^{-1}[(t - \alpha_\circ) - (t - \alpha_{\circ n})](t - \alpha_\circ)^{-1}\beta_{\circ n} + (t - \alpha_\circ)^{-1}(\beta_{\circ n} - \beta_\circ)\| \\
=&\|\gamma\| \sup_{|t|=1} \|(t - \alpha_{\circ n})^{-1}(\alpha_{\circ n} - \alpha_\circ)(t - \alpha_\circ)^{-1}\beta_{\circ n} + (t - \alpha_\circ)^{-1}(\beta_{\circ n} - \beta_\circ)\| \\
\leq&\|\gamma\|[\sup_{|t|=1} \|(t - \alpha_{\circ n})^{-1}\| \|\alpha_{\circ n} - \alpha_\circ\| \sup_{|t|=1} \|(t - \alpha_\circ)^{-1}\| \|\beta_{\circ n}\| \\
&+ \sup_{|t|=1} \|(t - \alpha_\circ)^{-1}\| \|\beta_{\circ n} - \beta_\circ\|].
\end{aligned}
$$

Considering the last inequality, we claim that the right hand side will go to zero as $n \to \infty$ if we can show that $\|(t - \alpha_{\circ n})^{-1}\|_\infty$ is bounded. Note that all other terms involving n converge to zero or are bounded.

Let $\sup_{|t|=1}\|(t-\alpha_\circ)^{-1}\| = K$ and assume that for every $k \in \mathbb{N}$ there exits a $n \geq k$ such that $\sup_{|t|=1}\|(t-\alpha_{\circ n})^{-1}\| > 2K$. Since $\|(t-\alpha_{\circ n})^{-1}\|$ is a continuous function for all n large enough, and $\mathbb{T}$ is a compact set, this function will assume its supremum for some $t_n \in \mathbb{T}$. Thus,
$\|(t_n-\alpha_{\circ n})^{-1}\| = \sup_{|t|=1}\|(t-\alpha_{\circ n})^{-1}\| > 2K$. Now form a convergent subsequence $t_{n_l} \to t_\circ$ and hence $\|(t_{n_l}-\alpha_{\circ n_l})^{-1}\| > 2K$. On the other hand, taking the limit as $l \to \infty$ we get

$$\lim_{l\to\infty}\|(t_{n_l}-\alpha_{\circ n_l})^{-1}\| = \|(t_\circ-\alpha_\circ)^{-1}\| \leq K.$$

Therefore there exits a l (and hence a n), such that for all $p \geq n$ we have
$\|(t_p-\alpha_{\circ p})^{-1}\| \leq 2K$. This contradicts $\|(t_{n_l}-\alpha_{\circ n_l})^{-1}\| > 2K$ and therefore our assumption is false, proving the boundedness of $\|(t-\alpha_{\circ n})^{-1}\|_\infty$.

We conclude with the convergence of $\tilde{\theta}_n^{-1}$:

$$\begin{aligned}
\|\tilde{\theta}_n^{-1}-\tilde{\theta}^{-1}\|_\infty &= \sup_{|t|=1}\|[I-tC_{\circ n}(I-tA_{\circ n})^{-1}BD_n^{-1}]-[I-tC_\circ(I-tA_\circ)^{-1}BD^{-1}]\| \\
&= \sup_{|t|=1}\|-t[C_{\circ n}[t(t^{-1}-A_{\circ n})]^{-1}BD_n^{-1}-C_\circ[t(t^{-1}-A_\circ)]^{-1}BD^{-1}]\| \\
&= \sup_{|t|=1}\|C_{\circ n}(t^{-1}-A_{\circ n})^{-1}BD_n^{-1}-C_\circ(t^{-1}-A_\circ)^{-1}BD^{-1}\|
\end{aligned}$$

For readability sake, we denote $(t^{-1}-A_{\circ n})^{-1}$ by $F_{\circ n}^{-1}$ and $(t^{-1}-A_\circ)^{-1}$ by $F_\circ^{-1}$. We then write

$$\begin{aligned}
&\sup_{|t|=1}\|C_{\circ n}(t^{-1}-A_{\circ n})^{-1}BD_n^{-1}-C_\circ(t^{-1}-A_\circ)^{-1}BD^{-1}\| \\
&= \|C_{\circ n}F_{\circ n}^{-1}BD_n^{-1}-C_\circ F_\circ^{-1}BD^{-1}\|_\infty \\
&= \|(C_{\circ n}F_{\circ n}^{-1}-C_\circ F_\circ^{-1})BD_n^{-1}+C_\circ F_\circ^{-1}B(D_n^{-1}-D^{-1})\|_\infty \\
&= \|[(C_{\circ n}-C_\circ)F_{\circ n}^{-1}+C_\circ(F_{\circ n}^{-1}-F_\circ^{-1})]BD_n^{-1}+C_\circ F_\circ^{-1}B(D_n^{-1}-D^{-1})\|_\infty \\
&= \|[(C_{\circ n}-C_\circ)F_{\circ n}^{-1}+C_\circ(F_{\circ n}^{-1}(F_\circ-F_{\circ n})F_\circ^{-1})]BD_n^{-1}+C_\circ F_\circ^{-1}B(D_n^{-1}-D^{-1})\|_\infty \\
&\leq [\|C_{\circ n}-C_\circ\|\|F_{\circ n}^{-1}\|_\infty+\|C_\circ\|\|F_{\circ n}^{-1}\|_\infty\|A_{\circ n}-A_\circ\|\|F_\circ^{-1}\|_\infty]\|B\|\|D_n^{-1}\| \\
&\quad+\|C_\circ\|\|B\|\|F_{\circ n}^{-1}\|_\infty\|D_n^{-1}-D^{-1}\|.
\end{aligned}$$

Notice again that all factors involving n are either bounded, or convergent with limit zero, including $F_{\circ n}^{-1}$ by the same argument as before. It then follows directly that $\lim_{n\to\infty}\|\tilde{\theta}_n^{-1}-\tilde{\theta}^{-1}\|_\infty = 0$ and thus we have shown that

$$\lim_{n\to\infty}\|S_n-T^{-1}\| = 0.$$

□

Example 2.2.1.

Let us now consider an example, showing how T^{-1} is approximated. To simplify calculations, we choose a scalar valued symbol,

$$R(t) = \frac{1}{2}+\frac{3}{4}t-\frac{2}{3}t^{-1}.$$

Clearly this defines a tri-diagonal Toeplitz operator and the state space representation of this symbol can be found by setting A and α to zero, and choosing the rest of the constants as follows,

$$R_0 = \frac{1}{2}, \quad B = \frac{3}{2}, \quad C = \frac{1}{2}, \quad \beta = 1 \quad \text{and} \quad \gamma = -\frac{2}{3}.$$

First we apply Theorem 2.1.1 to find a right canonical factorization of the symbol and then Proposition 2.1.3 to construct a converging sequence S_n.
In this case, (2.3) reduces to

$$Q = \beta(R_0 - \gamma QB)^{-1}C = \left(\frac{1}{2} + Q\right)^{-1}\frac{1}{2}.$$

Solving this quadratic equation, we find that there are two possible values, $Q = -1$ and $Q = \frac{1}{2}$. One of these is the desired stabilizing solution and to find which one, we test $A_\circ$ and $\alpha_\circ$. Take $Q = \frac{1}{2}$. Then,

$$A_\circ = -B(R_0 - \gamma QB)^{-1}C = -\frac{3}{2}\left(Q + \frac{1}{2}\right)^{-1}\left(\frac{1}{2}\right) = -\frac{3}{4}.$$

Also,

$$\alpha_\circ = -\beta(R_0 - \gamma QB)^{-1}\gamma = -1\left(Q + \frac{1}{2}\right)^{-1}\left(-\frac{2}{3}\right) = \frac{2}{3}.$$

Thus, $Q = \frac{1}{2}$ is the stabilizing solution, and we can calculate the factors in the canonical factorization of $R(t)$. Using the equations in (2.6) directly, we see that

$$\theta(t) = 1 + \frac{3}{4}t \quad \text{and} \quad \psi(t) = 1 - \frac{2}{3}t^{-1}.$$

Since our symbol $R(t)$ is scalar and analytic on $\mathbb{T}$, and defines an invertible Toeplitz operator T, it is well known that the FSM converges for T. Therefore we can apply Proposition 2.1.3 in conjunction with Theorem 2.1.5 to show how S_n approximates T^{-1}. Recall that $S_n = T_{\theta_n^{-1}}T_{\psi_n^{-1}}$. We need to construct θ_n^{-1} and ψ_n^{-1} and verify for which n they are well-defined. Notice that

$$\theta_n(t) = \frac{1}{2} + Q_n + \frac{3}{4}t; \quad \theta_n^{-1}(t) = \frac{1}{\frac{1}{2} + Q_n + \frac{3}{4}t} \quad \text{and}$$
$$\psi_n(t) = 1 - \frac{2t^{-1}}{\frac{3}{2} + 3Q_n}; \quad \psi_n^{-1}(t) = \frac{1}{1 - \frac{2t^{-1}}{\frac{3}{2}+3Q_n}}.$$

Remember that by definition, $Q_n = \omega_n T_n^{-1} W_n$. Thus, $Q_1 = \beta T_1^{-1} C = 1$ since $T_1^{-1} = 2$. According to Lemma 2.1.2, T_2^{-1} will now be invertible since $R_0 - \gamma Q_1 B = \frac{3}{2}$ is invertible. Luckily we need not check every single factor.

Notice that in general, $R_0 - \gamma Q_n B = \frac{1}{2} + Q_n$. Hence, for all positive Q_n, this factor will be invertible, and consequently, T_{n+1} also. But, we also see from Lemma 2.1.2 that $Q_{n+1} = \frac{1}{2}\left(\frac{1}{2} + Q_n\right)^{-1}$, which also produces positive Q_{n+1}'s from positive Q_n's. Therefore,

starting with $Q_1 = 1$, we see that T_n will be invertible for all $n \geq 2$ and we can now determine for which n the functions θ_n^{-1} and ψ_n^{-1} are well-defined. Consider

$$\theta_n^{-1}(t) = \frac{1}{\frac{1}{2} + Q_n + \frac{3}{4}t}.$$

θ_n^{-1} can only be ill-defined if $t = -1$, forcing $Q_n = \frac{1}{4}$. Will this ever happen? It turns out that it will not. We show this via a recursive argument. From our NDAR difference equation, we have $Q_1 = 1$ and $Q_2 = \frac{1}{3}$ and we claim that

$$\frac{1}{4} < Q_n < \frac{1}{2} \Longrightarrow \frac{1}{2} < Q_{n+1} < \frac{2}{3}. \tag{2.11}$$

This can be verified by using $Q_{n+1} = \frac{1}{2}\big(Q_n + \frac{1}{2}\big)^{-1}$. In the same manner, it can be shown that

$$\frac{1}{2} < Q_{n+1} < \frac{2}{3} \Longrightarrow \frac{3}{7} < Q_{n+2} < \frac{1}{2} \Longrightarrow \frac{1}{2} < Q_{n+3} < \frac{7}{13}.$$

From this inequality, we see that Q_{n+3} satisfies the second group of inequalities from (2.11). Thus we can write

$$\frac{3}{7} < Q_n < \frac{2}{3}, \quad \text{for all} \quad n \geq 4,$$

and a well-defined θ_n^{-1} is guaranteed. Now consider

$$\psi_n^{-1}(t) = \frac{1}{1 - \frac{2t^{-1}}{\frac{3}{2} + 3Q_n}}.$$

Here we see that the only value for which ψ_n^{-1} can be ill-defined is $t = 1$, forcing $Q_n = \frac{1}{6}$. From the previous arguments, we already know that $Q_n \neq \frac{1}{6}$ and we can conclude that S_n is well-defined for all n. From the formulas for θ_n^{-1} and ψ_n^{-1}, it is clear that they converge to θ^{-1} and ψ^{-1} respectively, and consequently S_n will converge to T^{-1}.

A natural question that arises concerns the speed of convergence for S_n. If we know something about the rate of convergence for Q_n, we may be able to express the convergence of S_n in terms of Q_n. This is dealt with in the next section.

2.3 Convergence rate of S_n

Proof of Proposition 2.1.6. First, a few recurring factors are shown to converge at the same rate as Q_n. This will simplify the many calculations necessary to arrive at the desired result. Here they are:

- $$\|C_{\circ n} - C_\circ\| = \|(C - \gamma Q_n A) - (C - \gamma Q A)\| \leq \|\gamma\|\|A\|\|Q_n - Q\| = K_1\|Q_n - Q\|.$$

- $$\begin{aligned}\|D_n^{-1} - D^{-1}\| &= \|(R_\circ - \gamma Q_n B)^{-1} - (R_\circ - \gamma QB)^{-1}\| \\ &= \|(R_\circ - \gamma Q_n B)^{-1}(R_\circ - \gamma Q_n B - R_\circ + \gamma QB)(R_\circ - \gamma QB)^{-1}\| \\ &\leq 2\|D_n^{-1}\|\|\gamma\|\|Q_n - Q\|\|B\|\|D^{-1}\| \\ &= K_2\|Q_n - Q\|,\end{aligned}$$

where $\|D_n^{-1}\| = \|(R_\circ - \gamma Q_n B)^{-1}\| \leq 2\|D^{-1}\|$ for n large enough, since $Q_n \to Q$.

- $$\begin{aligned}\|\beta_{\circ n} - \beta_\circ\| &= \|(\beta - \alpha Q_n B)D_n^{-1} - (\beta - \alpha QB)D^{-1}\| \\ &= \|[(\beta - \alpha Q_n B) - (\beta - \alpha QB)]D_n^{-1} + (\beta - \alpha QB)(D_n^{-1} - D^{-1})\| \\ &= \|(-\alpha Q_n B + \alpha QB)D_n^{-1} + (\beta - \alpha QB)(D_n^{-1} - D^{-1})\| \\ &\leq \|\alpha\|\|Q_n - Q\|\|B\|\|D_n^{-1}\| + \|\beta - \alpha QB\|K_2\|Q_n - Q\| \\ &\leq \|Q_n - Q\|(\|\alpha\|\|B\|2\|D^{-1}\| + \|\beta - \alpha QB\|K_2) \\ &= K_3\|Q_n - Q\|\end{aligned}$$

- $$\begin{aligned}\|A_{\circ n} - A_\circ\| &= \|(A - BD_n^{-1}C_{\circ n}) - (A - BD^{-1}C_\circ)\| \\ &= \| - BD_n^{-1}C_{\circ n} + BD^{-1}C_\circ\| \\ &\leq \|B\|[\|D_n^{-1} - D^{-1}\|\|C_{\circ n}\| + \|D^{-1}\|\|C_{\circ n} - C_\circ\|] \\ &\leq \|B\|[K_2\|Q_n - Q\|\|C_{\circ n}\| + \|D^{-1}\|K_1\|Q_n - Q\|] \\ &\leq \|Q_n - Q\|[\|B\|2\|C_\circ\|K_2 + \|D^{-1}\|K_1] \\ &= K_4\|Q_n - Q\|,\end{aligned}$$

  where $\|C_{\circ n}\| \leq 2\|C_\circ\|$ for n large enough.

- $$\begin{aligned}&\|\alpha_{\circ n} - \alpha_\circ\| \\ =&\|[\alpha - (\beta - \alpha Q_n B)(R_\circ - \gamma Q_n B)^{-1}\gamma] - [\alpha - (\beta - \alpha QB)(R_\circ - \gamma QB)^{-1}\gamma]\| \\ =&\|[\alpha - (\beta - \alpha Q_n B)D_n^{-1}\gamma] - [\alpha - (\beta - \alpha QB)D^{-1}\gamma]\| \\ =&\| - (\beta - \alpha Q_n B)D_n^{-1}\gamma + (\beta - \alpha QB)D^{-1}\gamma\| \\ \leq&\|\gamma\|\|[(\beta - \alpha Q_n B) - (\beta - \alpha QB)]D_n^{-1} + (\beta - \alpha QB)(D_n^{-1} - D^{-1})\| \\ =&\|\gamma\|\| - \alpha(Q_n - Q)BD_n^{-1} + (\beta - \alpha QB)(D_n^{-1} - D^{-1})\| \\ \leq&\|\gamma\|[\|\alpha\|\|Q_n - Q\|\|B\|\|D_n^{-1}\| + \|\beta - \alpha QB\|\|D_n^{-1} - D^{-1}\|] \\ \leq&\|\gamma\|[\|\alpha\|\|Q_n - Q\|\|B\|\|D_n^{-1}\| + \|\beta - \alpha QB\|K_2\|Q_n - Q\|] \\ \leq&\|Q_n - Q\|(\|\gamma\|\|\alpha\|\|B\|2\|D^{-1}\| + \|\beta - \alpha QB\|K_2) \\ =&K_5\|Q_n - Q\|.\end{aligned}$$

  Using the above estimates, we will show that $\|\tilde{\theta}_n^{-1} - \tilde{\theta}^{-1}\|_\infty$ and $\|\psi_n^{-1} - \psi^{-1}\|_\infty$ converge at the same rate as $\|Q_n - Q\|$, and consequently, $\|S_n - T^{-1}\|$ as well.

- $$\begin{aligned}\|\psi_n^{-1}-\psi^{-1}\|_\infty &\le \|\gamma\|[\sup_{|t|=1}\|(t-\alpha_{\circ n})^{-1}\|\|\alpha_{\circ n}-\alpha_\circ\|\sup_{|t|=1}\|(t-\alpha_\circ)^{-1}\|\|\beta_{\circ n}\| \\ &\quad+\sup_{|t|=1}\|(t-\alpha_\circ)^{-1}\|\|\beta_{\circ n}-\beta_\circ\|] \\ &\le \|\gamma\|[2KK_5\|Q_n-Q\|K_2\|\beta_\circ\|+KK_3\|Q_n-Q\|] \\ &= \|Q_n-Q\|(\|\gamma\|[2KK_5K_2\|\beta_\circ\|+\|\gamma\|KK_3) \\ &= K_6\|Q_n-Q\|\end{aligned}$$

  with $\|\beta_{\circ n}\| \le 2\|\beta_\circ\|$ for n large enough. (Recall that $K=\sup_{|t|=1}\|(t-\alpha_\circ)^{-1}\|$.)

- $$\begin{aligned}&\|\tilde\theta_n^{-1}-\tilde\theta^{-1}\|_\infty \\ \le&[\|C_{\circ n}-C_\circ\|\|F_{\circ n}^{-1}\|_\infty+\|C_\circ\|\|F_{\circ n}^{-1}\|_\infty\|A_{\circ n}-A_\circ\|\|F_\circ^{-1}\|_\infty]\|B\|\|D_n^{-1}\| \\ &+\|C_\circ\|\|B\|\|F_\circ^{-1}\|_\infty\|D_n^{-1}-D^{-1}\| \\ \le&[K_1\|Q_n-Q\|2\|F_\circ^{-1}\|_\infty+\|C_\circ\|2\|F_\circ^{-1}\|_\infty K_4\|Q_n-Q\|\|F_\circ^{-1}\|_\infty]\|B\|2\|D^{-1}\| \\ &+\|C_\circ\|\|B\|2\|F_\circ^{-1}\|_\infty K_2\|Q_n-Q\| \\ =&\|Q_n-Q\|[\|B\|2\|D^{-1}\|(K_12\|F_\circ^{-1}\|_\infty+\|C_\circ\|2\|F_\circ^{-1}\|_\infty K_4\|F_\circ^{-1}\|_\infty) \\ &+\|C_\circ\|\|B\|2\|F_\circ^{-1}\|_\infty K_2] \\ =&K_7\|Q_n-Q\|,\end{aligned}$$

with $\|F_{\circ n}^{-1}\|_\infty \le 2\|F_\circ^{-1}\|_\infty$ for n large enough. Finally, we are ready to put it all together. It follows from inequality 2.10 that

$$\begin{aligned}&\|S_n-T^{-1}\| \\ \le&[\|D_n^{-1}-D^{-1}\|\|\tilde\theta_n^{-1}\|_\infty+\|D^{-1}\|\|\tilde\theta_n^{-1}-\tilde\theta^{-1}\|_\infty]\|\psi_n^{-1}\|_\infty+\|\theta^{-1}\|_\infty\|\psi_n^{-1}-\psi^{-1}\|_\infty \\ \le&[K_2\|Q_n-Q\|2\|\tilde\theta^{-1}\|_\infty+\|D^{-1}\|K_7\|Q_n-Q\|]2\|\psi^{-1}\|_\infty+\|\theta^{-1}\|_\infty K_6\|Q_n-Q\| \\ =&\|Q_n-Q\|[2\|\psi^{-1}\|_\infty(K_22\|\tilde\theta^{-1}\|_\infty+\|D^{-1}\|K_7)+\|\theta^{-1}\|_\infty K_6] \\ =&c\|Q_n-Q\|,\end{aligned}$$

where $\|\psi_n^{-1}\|_\infty \le 2\|\psi^{-1}\|_\infty$ and $\|\tilde\theta_n^{-1}\|_\infty \le 2\|\tilde\theta^{-1}\|_\infty$ for n large enough. □

Example 2.3.1.

It is possible to find c in Proposition 2.1.6 above, but it requires a lot of tedious calculations and only an outline will be presented here. The first step involves the calculation of the constants (K_1,...,K_5) in Proposition 2.1.6 which will enable us to establish K_6, K_7 and then c. It is important to note that for this example we do not use the same estimates as in the inequalities of Proposition 2.1.6. For example,

$$\|D_n^{-1}-D^{-1}\| \le \|(R_0-\gamma Q_nB)^{-1}\|\|\gamma\|\|Q_n-Q\|\|B\|\|(R_0-\gamma QB)^{-1})\|.$$

The first and last factors on the right-hand side of the inequality is just D_n^{-1} and D^{-1} respectively. From Example 2.2.1 we know that $D_n^{-1}=\frac{1}{\frac{1}{2}+Q_n}$ and $D^{-1}=1$. Now, we

do not estimate $\|D_n^{-1}\| \leq 2\|D^{-1}\|$. We know that $\frac{3}{7} < Q_n < \frac{2}{3}$ for $n \geq 4$. To simplify our calculations, we just use $\frac{1}{3}$ as a lower bound for Q_n. Applying this bound, we can directly calculate and verify that $\|D_n^{-1}\| \leq \frac{6}{5}$, and consequently find that $K_2 = \frac{6}{5}$. The same arguments are applied to find the following constants:

$$K_1 = 0; \quad K_2 = \frac{6}{5}; \quad K_3 = \frac{6}{5}; \quad K_4 = \frac{9}{10}; \quad K_5 = \frac{4}{5}$$

With these in hand we first consider

$$\begin{aligned}\|\psi_n^{-1} - \psi^{-1}\|_\infty \leq \|\gamma\|[&\sup_{|t|=1}\|(t-\alpha_{\circ n})^{-1}\|\|\alpha_{\circ n} - \alpha_\circ\| \sup_{|t|=1}\|(t-\alpha_\circ)^{-1}\|\|\beta_{\circ n}\| \\ &+ \sup_{|t|=1}\|(t-\alpha_\circ)^{-1}\|\|\beta_{\circ n} - \beta_\circ\|]. \end{aligned} \tag{2.12}$$

As before, we calculate each of the factors in the above inequality directly. It is easy to see that

$$\sup_{|t|=1}\|(t-\alpha_\circ)^{-1}\| \;=\; \sup_{|t|=1}\frac{1}{\left|t-\frac{2}{3}\right|} \;\leq\; \sup_{|t|=1}\frac{1}{\left||t|-\frac{2}{3}\right|} \;=\; 3$$

using the inequality $|x-y| \geq ||x|-|y||$. Using the same reasoning, we also have

$$\sup_{|t|=1}\|(t-\alpha_{\circ n})^{-1}\| \;=\; \sup_{|t|=1}\frac{1}{|t-\frac{2}{3Q_n+\frac{3}{2}}|} \;\leq\; \sup_{|t|=1}\frac{1}{|1-\frac{2}{3Q_n+\frac{3}{2}}|}.$$

But, we know that $Q_n > \frac{1}{3}$ for $n \geq 4$ and inserting $Q_n = \frac{1}{3}$ we find that

$$\sup_{|t|=1}\frac{1}{|1-\frac{2}{3Q_n+\frac{3}{2}}|} < 5.$$

Inserting all these estimates into equation (2.12), it follows that

$$\begin{aligned}\|\psi_n^{-1} - \psi^{-1}\| &< \frac{2}{3}\left[\frac{72}{5}|Q_n - Q| + \frac{18}{5}|Q_n - Q|\right] \\ &= 12|Q_n - Q|,\end{aligned}$$

and $K_6 = 12$. We now determine K_7 by considering

$$\begin{aligned}\|\tilde{\theta}_n^{-1} - \tilde{\theta}^{-1}\|_\infty \leq [&\|C_{\circ n} - C_\circ\|\|F_{\circ n}^{-1}\|_\infty + \|C_\circ\|\|F_{\circ n}^{-1}\|_\infty\|A_{\circ n} - A_\circ\|\|F_\circ^{-1}\|_\infty]\|B\|\|D_n^{-1}\| \\ &+ \|C_\circ\|\|B\|\|F_{\circ n}^{-1}\|_\infty\|D_n^{-1} - D^{-1}\|.\end{aligned}$$

Recall from the proof of Theorem 2.1.1 that $\|F_\circ^{-1}\|_\infty = \sup_{|t|=1}|(t^{-1}-A_\circ)^{-1}|$ and $\|F_{\circ n}^{-1}\|_\infty = \sup_{|t|=1}|(t^{-1}-A_{\circ n})^{-1}|$. Applying the same arguments mentioned above, we find that

$$\sup_{|t|=1}\|(t^{-1}-A_\circ)^{-1}\| \;=\; \sup_{|t|=1}\frac{1}{\left|t^{-1}-\frac{-3}{4}\right|} \;\leq\; \sup_{|t|=1}\frac{1}{\left||t^{-1}|-\frac{3}{4}\right|} \;=\; 4,$$

and

$$\sup_{|t|=1}\|(t^{-1}-A_{\circ n})^{-1}\| \;=\; \sup_{|t|=1}\frac{1}{\left|t^{-1}-\frac{-3}{4}\frac{1}{\frac{1}{2}+Q_n}\right|} \;\leq\; \sup_{|t|=1}\frac{1}{\left||t^{-1}|-\frac{3}{4}\frac{1}{\frac{1}{2}+Q_n}\right|} \;<\; 10.$$

The last inequality follows from the fact that $\frac{1}{3} < Q_n < \frac{2}{3}$ for $n \geq 4$. Hence we can write

$$\begin{aligned}\|\tilde{\theta}_n^{-1} - \tilde{\theta}^{-1}\|_\infty &< \left[0 + 5.\frac{9}{10}|Q_n - Q|4\right]\frac{9}{5} + 9|Q_n - Q| \\ &= \frac{207}{5}|Q_n - Q| < 42|Q_n - Q|,\end{aligned}$$

and we choose $K_7 = 42$. We are almost ready to find c, but from inequality 2.10

$$\|S_n - T^{-1}\| \leq [\|D_n^{-1} - D^{-1}\|\|\tilde{\theta}_n^{-1}\|_\infty + \|D^{-1}\|\|\tilde{\theta}_n^{-1} - \tilde{\theta}^{-1}\|_\infty]\|\psi_n^{-1}\|_\infty + \|\theta^{-1}\|_\infty\|\psi_n^{-1} - \psi^{-1}\|_\infty$$

we notice that we still have to find estimates for $\|\tilde{\theta}_n^{-1}\|_\infty$, $\|\psi_n^{-1}\|_\infty$ and $\|\theta^{-1}\|_\infty$. Here we apply aforementioned reasoning to see that

$$\|\theta^{-1}\|_\infty = \sup_{|t|=1}|\theta(t)^{-1}| \leq 4 \quad \text{and} \quad \|\psi_n^{-1}\|_\infty = \sup_{|t|=1}|\psi(t)_n^{-1}| < 5.$$

Also,

$$\begin{aligned}\|\tilde{\theta}_n^{-1}\|_\infty &= \sup_{|t|=1}|\tilde{\theta}_n^{-1}(t)| \\ &= \sup_{|t|=1}|1 - tC_{\circ n}(1 - tA_{\circ n})^{-1}BD_n^{-1}| \\ &= \sup_{|t|=1}|1 - C_{\circ n}(t^{-1} - A_{\circ n})^{-1}BD_n^{-1}| \\ &= \sup_{|t|=1}\left|1 - \frac{3}{4}F_{\circ n}^{-1}D_n^{-1}\right| \\ &\leq 1 + \frac{3}{4}\sup_{|t|=1}|F_{\circ n}^{-1}||D_n^{-1}| \\ &< 10.\end{aligned}$$

All the necessary estimates have been gathered and we compute c:

$$\begin{aligned}\|S_n - T^{-1}\| &\leq [\|D_n^{-1} - D^{-1}\|\|\tilde{\theta}_n^{-1}\|_\infty + \|D^{-1}\|\|\tilde{\theta}_n^{-1} - \tilde{\theta}^{-1}\|_\infty]\|\psi_n^{-1}\|_\infty \\ &\quad + \|\theta^{-1}\|_\infty\|\psi_n^{-1} - \psi^{-1}\|_\infty \\ &< [12|Q_n - Q| + 42|Q_n - Q|]5 + 48|Q_n - Q| \\ &= 318|Q_n - Q|.\end{aligned}$$

2.4 Convergence rate of the NDAR difference equation

We begin by showing the linear convergence of the NDAR difference equation as stated in Proposition 2.1.7.

Proof. Recall the NDAR difference equation(s):

$$Q_{n+1} = \alpha Q_n A + (\beta - \alpha Q_n B)(R_0 - \gamma Q_n B)^{-1}(C - \gamma Q_n A)$$

and

$$Q_n = \alpha Q_{n-1}A + (\beta - \alpha Q_{n-1}B)(R_0 - \gamma Q_{n-1}B)^{-1}(C - \gamma Q_{n-1}A).$$

With a few substitutions we can write the following difference equation,

$$\begin{aligned} Q_{n+1} - Q_n &= \alpha(Q_n - Q_{n-1})A + \beta_{\circ n}C_{\circ n} - \beta_{\circ(n-1)}C_{\circ(n-1)} \\ &= \alpha(Q_n - Q_{n-1})A + (\beta_{\circ n} - \beta_{\circ(n-1)})C_{\circ n} \\ &\quad + \beta_{\circ(n-1)}(C_{\circ n} - C_{\circ(n-1)}). \end{aligned}$$

Taking norms and using the estimates for the different factors as in the proof of Proposition 2.1.6, we have

$$\|Q_{n+1} - Q_n\| \leq k\|Q_n - Q_{n-1}\|.$$

□

In other words, we have linear convergence for Q_n. We have to wonder whether this is the best we can do since we know that the symmetric discrete algebraic Riccati equation converges quadratically (see [24]). It turns out that in general, quadratic convergence does not hold under a weak assumption, and we will show that this is always true for the scalar case. For the non-scalar case we will provide a counter example.

Considering the NDAR difference equation (2.8) for the scalar case, one can perform a few rearrangements to arrive at

$$\begin{aligned} Q_{n+1} - Q_n = \alpha(Q_n - Q_{n-1})A + (X_n - X_{n-1})Y_nZ_n \\ + X_{n-1}[(Y_n - Y_{n-1})Z_{n-1} + Y_{n-1}(Z_n - Z_{n-1})], \end{aligned}$$

where

$$X_n = \beta - \alpha Q_n B, \qquad Y_n = (R_0 - \gamma Q_n B)^{-1} \qquad \text{and} \qquad Z_n = C - \gamma Q_n A.$$

Taking norms on both sides and factoring out $(Q_n - Q_{n-1})$ in the previous equation, we can write

$$\begin{aligned} &|Q_{n+1} - Q_n| \\ =&|\alpha(Q_n - Q_{n-1})A - \alpha(Q_n - Q_{n+1})BY_nZ_n \\ &+ X_{n-1}[Y_n\gamma(Q_n - Q_{n-1})BY_{n-1}Z_n - Y_{n-1}\gamma(Q_n - Q_{n-1})A]| \\ =&|Q_n - Q_{n-1}||\alpha A - \alpha BY_nZ_n + X_{n-1}Y_n\gamma BY_{n-1}Z_n - X_{n-1}Y_{n-1}\gamma A| \\ =&|Q_n - Q_{n-1}||G_n|, \end{aligned}$$

where $G_n = |\alpha A - \alpha BY_nZ_n + X_{n-1}Y_n\gamma BY_{n-1}Z_n - X_{n-1}Y_{n-1}\gamma A|$. Now, assuming quadratic convergence holds, it is true that

$$|Q_{n+1} - Q_n| \leq c|Q_n - Q_{n-1}|^2,$$

where c is some positive real constant. Inserting the previous equality for $|Q_{n+1} - Q_n|$ we get

$$\begin{aligned} |Q_n - Q_{n-1}|\|G_n\| &\leq c|Q_n - Q_{n-1}|^2 \\ \implies \quad \|G_n\| &\leq c|Q_n - Q_{n-1}|. \end{aligned}$$

We know that G_n is bounded since all its factors are bounded, and if $\lim_{n\longrightarrow\infty} G_n \neq 0$, the above inequality is false since the right-hand side can be made arbitrarily small.

This is exactly the case in Example 2.2.1. Here we have a significantly simplified difference equation. Recall that we had $\frac{3}{7} < Q_n < \frac{2}{3}$ for all $n \geq 4$. Then

$$\begin{aligned} |Q_{n+1} - Q_n| &= \left| \frac{\frac{1}{2}}{Q_n + \frac{1}{2}} - \frac{\frac{1}{2}}{Q_{n-1} + \frac{1}{2}} \right| \\ &= \frac{1}{2} \left| \frac{Q_n - Q_{n-1}}{(Q_n + \frac{1}{2})(Q_{n-1} + \frac{1}{2})} \right|. \end{aligned}$$

Combining with the bounds for Q_n we obtain

$$\frac{18}{49}|Q_n - Q_{n-1}| \leq |Q_{n+1} - Q_n| \leq \frac{98}{169}|Q_n - Q_{n-1}|.$$

Therefore Q_n does not converge quadratically.

For the non-scalar case we have the following example:

Example 2.4.1.

Let the Toeplitz operator T have the symbol

$$\begin{aligned} R(t) &= \begin{bmatrix} 2 + \frac{t}{2} - \frac{t^{-1}}{2} & 0 \\ 0 & 2 + \frac{t}{2} - \frac{t^{-1}}{2} \end{bmatrix} \\ &= \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} + \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} t + \begin{bmatrix} -\frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{bmatrix} t^{-1}. \end{aligned}$$

Also remember that $R(t) = R_0 + tC(I - tA)^{-1}B + \gamma(tI - \alpha)^{-1}\beta$. Setting $A = \alpha = 0$, we see that $R(t) = R_0 + CBt + \gamma\beta t^{-1}$. Choosing our matrices as follows,

$$B = \gamma = I, \qquad C = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}, \qquad \beta = \begin{bmatrix} -\frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{bmatrix}, \qquad R_0 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

we arrive at the symbol chosen above. Our first step is to calculate the solution Q to the NDARE and then show that $A_\circ$ and $\alpha_\circ$ are both stable. It is easy to see that

$$R_0 - \gamma QB = \begin{bmatrix} 2 - Q^{(1)} & 0 \\ 0 & 2 - Q^{(4)} \end{bmatrix} \qquad \Longrightarrow \qquad (R_0 - \gamma QB)^{-1} = \begin{bmatrix} \frac{1}{2-Q^{(1)}} & 0 \\ 0 & \frac{1}{2-Q^{(4)}} \end{bmatrix},$$

where we assumed $Q = \begin{bmatrix} Q^{(1)} & 0 \\ 0 & Q^{(4)} \end{bmatrix}$ and this will turn out to be a good choice for Q. In this case the NDARE reduces to

$$\begin{aligned} Q &= \beta(R_0 - \gamma QB)^{-1}C \\ &= \begin{bmatrix} -\frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{2-Q^{(1)}} & 0 \\ 0 & \frac{1}{2-Q^{(4)}} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \\ &= \begin{bmatrix} -\frac{1}{4}\left(\frac{1}{2-Q^{(1)}}\right) & 0 \\ 0 & -\frac{1}{4}\left(\frac{1}{2-Q^{(4)}}\right) \end{bmatrix}. \end{aligned}$$

From this matrix equation we can solve the following polynomial equation to determine Q (Notice that Q^1 and Q^4 will have the same solution set):

$$(Q^{(1)})^2 - 2Q^{(1)} - \frac{1}{4} = 0 \qquad \Longrightarrow \qquad Q^{(1)} = \frac{2 \pm \sqrt{5}}{2}.$$

Moving on to the question of stability, we have

$$\begin{aligned} A_\circ &= \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{2-Q^{(1)}} & 0 \\ 0 & \frac{1}{2-Q^{(4)}} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \\ &= \begin{bmatrix} -\frac{1}{2}\left(\frac{1}{2-Q^{(1)}}\right) & 0 \\ 0 & -\frac{1}{2}\left(\frac{1}{2-Q^{(4)}}\right) \end{bmatrix} \end{aligned}$$

and since $Q^{(1)}$ and $Q^{(4)}$ have the same solution set there are only two possible eigenvalues for $A_\circ$:

$$\begin{aligned} \lambda &= -\frac{1}{2}\left(\frac{1}{2 - \frac{2+\sqrt{5}}{2}}\right) \\ &= -\frac{1}{2-\sqrt{5}} \\ &> 1 \end{aligned}$$

and therefore this solution does not provide a stable $A_\circ$. On the other hand,

$$\begin{aligned} |\lambda| &= \frac{1}{2}\left(\frac{1}{2 - \frac{2-\sqrt{5}}{2}}\right) \\ &= \frac{1}{2+\sqrt{5}} \\ &< 1 \end{aligned}$$

which gives a stable $A_\circ$. We need to check this solution for $\alpha_\circ$ to guarantee invertibility of T.

$$\begin{aligned} \alpha_\circ &= \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{2-Q^{(1)}} & 0 \\ 0 & \frac{1}{2-Q^{(4)}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2}\left(\frac{1}{2-Q^{(1)}}\right) & 0 \\ 0 & \frac{1}{2}\left(\frac{1}{2-Q^{(4)}}\right) \end{bmatrix} \end{aligned}$$

We see that this equation will produce the same eigenvalues (except for a sign change) as for $A_\circ$. Therefore,

$$Q = \begin{bmatrix} \frac{2-\sqrt{5}}{2} & 0 \\ 0 & \frac{2-\sqrt{5}}{2} \end{bmatrix}$$

is a stable solution to the NDARE and T is invertible. We still want to investigate the convergence of $Q_n \longrightarrow Q$, but for this we need to show that the FSM converges for our

Toeplitz operator. We just need to check that the block transpose $T^{\#}$ of our original operator is invertible ([13]). Luckily, because of the almost symmetric structure of T - only a sign change in the upper and lower diagonal entries distinguish T from $T^{\#}$ - $T^{\#}$ will also be invertible via equations analogous to the above. This guarantees the convergence of Q_n and we proceed to analyze the speed of convergence. Consider the NDAR difference equation,

$$\begin{bmatrix} Q_{n+1}^{(1)} & 0 \\ 0 & Q_{n+1}^{(4)} \end{bmatrix} = \begin{bmatrix} -\frac{1}{4}\left(\frac{1}{2-Q_n^{(1)}}\right) & 0 \\ 0 & -\frac{1}{4}\left(\frac{1}{2-Q_n^{(4)}}\right) \end{bmatrix}.$$

In what follows we want $Q_n^{(1)} = Q_n^{(4)}$ for all $n \in \mathbb{N}$ large enough. Will this be the case? Indeed it will be, and actually $k = 1$ in Proposition 2.1.3. To see this, we apply Lemma 2.1.2. Notice that $T_1 = R_0$ and therefore T_1 is invertible. Let

$$Q_1 = \omega_1 T_1^{-1} W_1 = \begin{bmatrix} -\frac{1}{8} & 0 \\ 0 & -\frac{1}{8} \end{bmatrix}.$$

If we consider the NDAR difference equation it becomes clear that $Q_n^{(1)} = Q_n^{(4)} < 0$ for all $n \in \mathbb{N}$. This will also guarantee the invertibility of $R_0 - \gamma Q_n B$ for all $n \in \mathbb{N}$, leading to the invertibily of all T_n^{-1}. Now,

$$\begin{aligned}
&Q_{n+1} - Q_n \\
&= \begin{bmatrix} -\frac{1}{4}\left(\frac{1}{2-Q_n^{(1)}} - \frac{1}{2-Q_{n-1}^{(1)}}\right) & 0 \\ 0 & -\frac{1}{4}\left(\frac{1}{2-Q_n^{(4)}} - \frac{1}{2-Q_n^{(4)}}\right) \end{bmatrix} \\
&= \begin{bmatrix} -\frac{1}{4} & 0 \\ 0 & -\frac{1}{4} \end{bmatrix} \begin{bmatrix} \frac{Q_n^{(1)} - Q_{n-1}^{(1)}}{(2-Q_n^{(1)})(2-Q_{n-1}^{(1)})} & 0 \\ 0 & \frac{Q_n^{(4)} - Q_{n-1}^{(4)}}{(2-Q_n^{(4)})(2-Q_{n-1}^{(4)})} \end{bmatrix} \\
&= \begin{bmatrix} -\frac{1}{4}\frac{1}{(2-Q_n^{(1)})(2-Q_{n-1}^{(1)})} & 0 \\ 0 & -\frac{1}{4}\frac{1}{(2-Q_n^{(4)})(2-Q_{n-1}^{(4)})} \end{bmatrix} \begin{bmatrix} Q_n^{(1)} - Q_{n-1}^{(1)} & 0 \\ 0 & Q_n^{(4)} - Q_{n-1}^{(4)} \end{bmatrix} \\
&= G_n(Q_n - Q_{n-1}),
\end{aligned}$$

where G_n is the left matrix factor in the penultimate equality above. Taking norms, we have

$$\begin{aligned}
\|Q_{n+1} - Q_n\| &= \|G_n(Q_n - Q_{n-1})\| \\
&= \|G_n\|\|(Q_n - Q_{n-1})\|.
\end{aligned}$$

The last equality follows from the fact that all the factors are constant diagonal matrices and when we consider the spectral norm, the splitting of the norm does not result in an inequality. Assuming quadratic convergence, we arrive at

$$\begin{aligned}
\|Q_{n+1} - Q_n\| = \|G_n\|\|(Q_n - Q_{n-1})\| \leq c\|Q_n - Q_{n-1}\|^2 \\
\implies \|G_n\| \leq c\|Q_n - Q_{n-1}\|.
\end{aligned}$$

However, it is clear that $\lim_{n\to\infty}\|G_n\| \neq 0$, where $\lim_{n\to\infty}\|Q_n - Q_{n-1}\| = 0$ and consequently quadratic convergence does not hold.

Although this is a single counter-example, it must be noted that we could have chosen any size for the matrix symbol and used the same entries on the diagonals. Moreover, this example is part of a class of invertible Toeplitz operators for which the FSM converges. In fact, let T be a tridiagonal Toeplitz operator with $A = \alpha = 0$ and let all other factors in its state space representation be constant diagonal matrices. Thus,

$$\begin{aligned} Q_{n+1} - Q_n &= \beta[(R_0 - \gamma Q_n B)^{-1} - (R_0 - \gamma Q_{n-1}B)^{-1}]C \\ &= \beta(R_0 - \gamma Q_n B)^{-1}[R_0 - \gamma Q_n B - (R_0 - \gamma Q_{n-1}B)](R_0 - \gamma Q_{n-1}B)^{-1}C \\ &= \beta(R_0 - \gamma Q_n B)^{-1}(Q_n - Q_{n-1})(R_0 - \gamma Q_{n-1}B)^{-1}C \\ &= \beta(R_0 - \gamma Q_n B)^{-1}(R_0 - \gamma Q_{n-1}B)^{-1}C(Q_n - Q_{n-1}) \\ &= G_n(Q_n - Q_{n-1}), \end{aligned}$$

and we can use the same arguments as for the specific example to show that quadratic convergence does not hold when $\lim_{n\to\infty} G_n \neq 0$. Clearly,
$G_n = \beta(R_0 - \gamma Q_n B)^{-1}(R_0 - \gamma Q_{n-1}B)^{-1}C$ is a diagonal matrix.

It is also necessary to note that this counter example holds only for this iterative method of calculating Q via the NDAR difference equation. There might still be other ways to construct a sequence of Q_n's that converge quadratically.

2.5 Algorithm for calculating T_n^{-1}

Proposition 2.1.3 is also useful to calculate the inverses of the sections of our Toeplitz operator T via a recursive formula. Consider the $(n+1)$-th finite section of a Toeplitz operator T. Its special structure permits the following representation:

$$T_{n+1} = \begin{bmatrix} R_0 & \Gamma_n \\ \Xi_n & T_n \end{bmatrix} \text{on} \begin{bmatrix} \mathbb{C}^m \\ l_+^2(\mathbb{C}^m) \end{bmatrix}$$

where

$$\Gamma_n = [R_{-1} \quad R_{-2} \quad \cdots \quad R_{-n} \quad 0 \quad \cdots] : l_+^2(\mathbb{C}^m) \to \mathbb{C}^m,$$

$$\Xi_n = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_n \\ 0 \\ \vdots \end{bmatrix} : \mathbb{C}^m \to l_+^2(\mathbb{C}^m).$$

For the case when $n = 0$, we assume the vectors and operator above to be zero. If we assume T_{n+1} is an invertible operator for all n on $P_{n+1}[l_+^2(\mathbb{C}^m)]$, i.e., the projection onto the first $n+1$ coordinates of $l_+^2(\mathbb{C}^m)$, we have the following formula for T_{n+1}^{-1}:

$$T_{n+1}^{-1} = \begin{bmatrix} \Delta_n^{-1} & -\Delta_n^{-1}\Gamma_n T_n^{-1} \\ -T_n^{-1}\Xi_n\Delta_n^{-1} & T_n^{-1} + T_n^{-1}\Xi_n\Delta_n^{-1}\Gamma_n T_n^{-1} \end{bmatrix} \text{on} \begin{bmatrix} \mathbb{C}^m \\ l_+^2(\mathbb{C}^m) \end{bmatrix}$$

with the Schur complement, $\Delta_n = R_0 - \Gamma_n T_n^{-1} \Xi_n$, a well-defined invertible operator on $\mathbb{C}^m$. Noticing that $\Gamma_n = \gamma\omega_n$ and $\Xi_n = W_n B$, we find that

$$T_{n+1}^{-1} = \begin{bmatrix} (R_0 - \gamma\omega_n T_n^{-1} W_n B)^{-1} & -(R_0 - \gamma\omega_n T_n^{-1} W_n B)^{-1} \gamma\omega_n T_n^{-1} \\ -T_n^{-1} W_n B (R_0 - \gamma\omega_n T_n^{-1} W_n B)^{-1} & T_n^{-1} + T_n^{-1} W_n B (R_0 - \gamma\omega_n T_n^{-1} W_n B)^{-1} \gamma\omega_n T_n^{-1} \end{bmatrix}. \tag{2.13}$$

With this recursive relation we can calculate any T_{n+1}^{-1}. Indeed, starting with $n = 0$, we see that

$$T_1^{-1} = \begin{bmatrix} R_0^{-1} \end{bmatrix}$$

as expected. Inserting T_1^{-1} back into the right hand side of the equation above, we see that all the remaining variables are known, or calculable, and we iterate to find all T_{n+1}^{-1}.

From equation (2.13) we notice the presence of Q_n, and assuming that Proposition 2.1.3 holds, we do not have to calculate the factors $\omega_n T_n^{-1} W_n$ directly. Instead we find Q_n recursively from equation (2.8) which is in general a much quicker calculation. This allows us to define to following algorithm:

(i) $(n = 0)$: $Q_0 = 0$ since all vectors and matrices with index $n = 0$ were defined to be zero vectors and matrices respectively. This produces $T_1^{-1} = [R_0^{-1}]$.

(ii) $(n = 1, 2, 3, ...)$: Find Q_n via equation (2.8). With T_n^{-1} from the previous step and the other known vectors we can calculate T_{n+1}^{-1}.

Notice that in the calculation of T_{n+1}^{-1}, we only invert the Schur complement which is a matrix of fixed rank. The remaining computations consist of basic matrix multiplications which is restricted to two recurring factors, $\gamma\omega_n T_n^{-1}$ and $T_n^{-1} W_n B$.

Chapter 3

Norm asymptotics for a special class of Toeplitz-generated matrices

3.1 Introduction

Motivated by a question posed by statisticians [37], we consider the following $n \times n$ matrix

$$K_n = \begin{bmatrix} 2+\frac{1}{m} & -1 & 0 & \cdots & \cdots & 0 \\ 1+\frac{1}{m} & 1+\frac{1}{m} & -1 & 0 & & \vdots \\ \vdots & \frac{1}{m} & & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & -1 & 0 \\ \vdots & \vdots & \ddots & \ddots & & -1 \\ 1+\frac{1}{m} & \frac{1}{m} & \cdots & \cdots & \frac{1}{m} & 1+\frac{1}{m} \end{bmatrix}, \tag{3.1}$$

where $m = [cn]$, for some constant c (we will omit the rounding of cn from here on as it will not impact on the analysis). The problem is then to compute the norm of the inverse, or at least to estimate the norm of the inverse asymptotically as n goes to infinity.

We observe that the matrix K_n can be written as a Toeplitz matrix plus a rank-one perturbation,

$$K_n = T_n(\frac{1}{cn}) + ee_1^*,$$

where e is the all-ones vector of size n and e_1 is the first standard basis vector. Here, $T_n(\frac{1}{cn})$ is the $n \times n$ Toeplitz matrix

$$T_n(\frac{1}{cn}) = \begin{bmatrix} 1+\frac{1}{cn} & -1 & 0 & \cdots & \cdots & 0 \\ \frac{1}{cn} & 1+\frac{1}{cn} & -1 & 0 & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -1 & 0 \\ \vdots & \ddots & \ddots & \ddots & & -1 \\ \frac{1}{cn} & \cdots & \cdots & \cdots & \frac{1}{cn} & 1+\frac{1}{cn} \end{bmatrix}, \tag{3.2}$$

depending on the positive real-valued sequence $\frac{1}{cn}$ (which depends on the size). From now on we will assume that matrices with subscript n indicate $n \times n$ matrices.

To estimate the norm of the inverse of $T_n(\frac{1}{cn})$, we write it as

$$T_n(\frac{1}{cn}) = T_{0,n} + \frac{1}{cn}L,$$

where

$$T_{0,n} = \begin{bmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & \ddots & -1 \\ & & & 1 \end{bmatrix}, L = \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & & \ddots & \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Now we observe that $L = (T_{0,n}^{-1})^*$. Thus,

$$T_n(\frac{1}{cn}) = T_{0,n} + \frac{1}{cn}(T_{0,n}^{-1})^*.$$

In other words, we decomposed $T_n(\frac{1}{cn})$ into the sum of some Toeplitz matrix plus a constant times the adjoint of its inverse. It turns out that this decomposition is a special example of a much larger class of matrices for which we will prove our main result.

Define $X_n := T_n + f_n(T_n^{-1})^*$, where T_n is an invertible banded Toeplitz matrix - viewed as a finite section of a fixed and banded infinite Toeplitz matrix, and f_n a sequence of positive real numbers. Observe that X_n does not have to be Toeplitz, but it is 'generated' by the banded Toeplitz matrix T_n. Our main result is as follows.

Theorem 3.1.1. *Let the sequence f_n satisfy the conditions*

- $\lim_{n\to\infty} f_n = 0$,
- $\exists\, c > 0$ *such that* $n\sqrt{f_n^{\epsilon+1}} > c$ *for n large enough and some* $\epsilon > 0$. *In other words,* $f_n > O(\frac{1}{n^2})$.

Assume that the symbol $b(e^{i\theta}) = \sum_{j=-r}^{r} b_j e^{ij\theta}$ associated with the infinite Toeplitz matrix $T = (b_{j-k})_{j,k=1}^{\infty}$ is such that it has at least one zero on $\mathbb{T}$.

Let $X_n = T_n + f_n(T_n^{-1})^*$, whenever T_n is invertible. Then X_n is invertible and,

$$\lim_{n\to\infty} 2\sqrt{f_n}\|X_n^{-1}\| = 1.$$

The first property of the symbol b guarantees that the range of $|b|$ contains some interval $[0, c)$, which ensures that the singular values of T_n eventually surround f_n. And clearly, the second property is needed for X_n to make sense. It is not easy to characterize invertible Toeplitz matrices with associated symbols that have zeros on $\mathbb{T}$ in general. We do have Theorem 1.2.9, and as discussed in the introduction, there are symbols with zeros that this theorem does not apply to. For example, consider the symbol

$$b(t) = \frac{1}{2}t^3 + \frac{1}{2}t^{-1} = \frac{1}{2}e^{3i\theta} + \frac{1}{2}e^{-i\theta}, \quad 0 \le \theta \le 2\pi.$$

It has a zero at $\theta = \pi/4$, but does not satisfy the second condition of Theorem 1.2.9. This can clearly be seen in Figure 3.1.

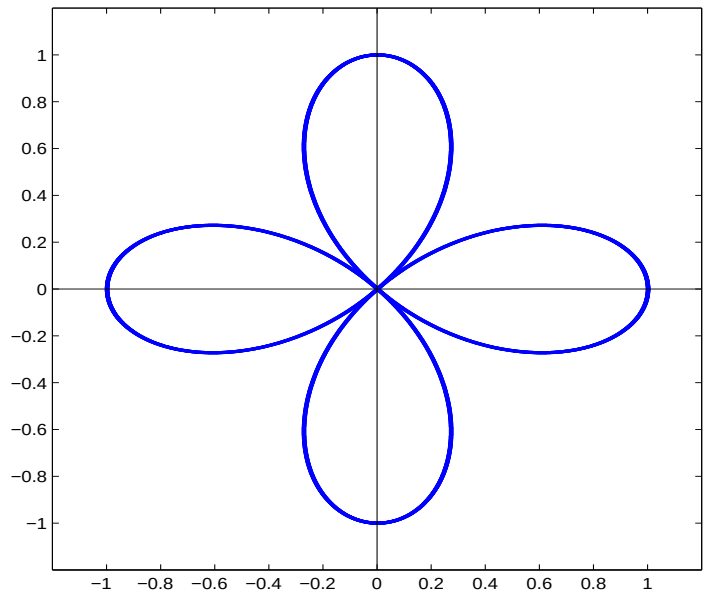

Figure 3.1: The range of $b(t) = \frac{1}{2}t^3 + \frac{1}{2}t^{-1}, t \in \mathbb{T}$

On the other hand, many classes of symbols are easily seen to be invertible. For instance, if T is triangular with nonzero diagonal entries b_0, or if Re $b(t) \geq 0$ for $t \in \mathbb{T}$ but Re b is not identically zero on $\mathbb{T}$. In the concrete case considered here, we have $b(t) = 1 - t^{-1}$, and hence all the conditions mentioned are satisfied.

Under the same conditions, we can prove this result for a perturbation of X_n as well. Define

$$Z_n := V_n T_n + f_n (T_n^{-1})^* + W_n,$$

where $W_n = (T_n^{-1})^* \sum_{j=1}^n w_j e_j e_j^*$, $V_n = \sum_{j=1}^n v_j e_j e_j^*$, and where v_j and w_j are positive real numbers. As usual, e_j is the j-th standard basis vector. Then we also have the following theorem.

Theorem 3.1.2. *Under the conditions of Theorem 3.1.1, let*

$$Z_n = V_n T_n + f_n (T_n^{-1})^* + W_n,$$

with V_n and W_n as defined above. In addition, assume the following conditions on the entries of V_n and W_n:

- $v_j \geq 1$ *for* $1 \leq j \leq q_v$ *where q_v (independent of n) is fixed.*
- $v_j = 1$ *for* $n \geq j > q_v$.
- $w_j \geq 0$ *for* $1 \leq j \leq q_w$ *where q_w (independent of n) is fixed.*
- $w_j = 0$ *for* $n \geq j > q_w$.

Then,

$$\lim_{n\to\infty} 2\sqrt{f_n}\|Z_n^{-1}\| = 1.$$

Given the matrices V_n and W_n, it is reasonable to ask which entries of X_n will be perturbed? First we rewrite Z_n as $Z_n = V_n T_n + (T_n^{-1})^* \sum_{j=1}^n (f_n + w_j) e_j e_j^*$. Clearly V_n and $\sum_{j=1}^n (f_n + w_j) e_j e_j^*$ are diagonal matrices that act on T_n from the left and $(T_n^{-1})^*$ from the right respectively, and therefore the first q_v rows and q_w columns of X_n will be perturbed.

In the current literature on asymptotics of singular values or eigenvalues (see [5] for a recent survey), the starting point has been to analyze the behaviour of a Toeplitz matrix or operator with a fixed symbol. In our case, the presence of the sequence f_n causes the entries of X_n and Z_n to change as their sizes grow.

To achieve the above-mentioned results, we will establish upper bounds and lower bounds for the norms of X_n^{-1} and Z_n^{-1}, and then apply the well-known squeeze theorem to arrive at the desired result. Sections 3.2 and 3.3 will be devoted to this task. In Section 3.4 we continue to establish a more precise description of the behaviour of $\|X_n^{-1}\|$ and $\|Z_n^{-1}\|$. This is followed in Section 3.5 by considering the case where T_n is Fredholm. The rest of the chapter (3.6 - 3.8) will address questions regarding the behaviour of $\|X_n\|$, $\|Z_n\|$ and the singular values of of T-gen matrices in general.

Sections 3.1 - 3.5 are largely based on [26, 27] with some refinements and additions.

3.2 Upper bounds for $\|X_n^{-1}\|$ and $\|Z_n^{-1}\|$

The following general lemma will provide the upper bounds for $\|X_n^{-1}\|$ and be critical in calculating these bounds for $\|Z_n^{-1}\|$.

Lemma 3.2.1. *Let Y be a square matrix of the form $c_1 A + c_2 (A^{-1})^*$ where A is invertible, and c_1 and c_2 are positive real numbers. Then, $\|Y^{-1}\| \leq \frac{1}{2\sqrt{c_1 c_2}}$.*

Proof. We prove that $\sigma_n(Y) \geq 2\sqrt{c_1 c_2}$, which gives the result of the lemma. Compute

$$YY^* = (c_1 A + c_2 (A^{-1})^*)(c_1 A^* + c_2 A^{-1}) = 2c_1 c_2 I + c_1^2 AA^* + c_2^2 (A^{-1})^* A^{-1},$$

and consider the matrix $c_1^2 AA^* + c_2^2 (A^{-1})^* A^{-1}$. This is a matrix of the form $c_1^2 H + c_2^2 H^{-1}$, where H is positive definite. Let $f : \mathbb{R}^+ \to \mathbb{R}^+$ be the function $f(x) = c_1^2 x + c_2^2 \frac{1}{x}$, then $c_1^2 H + c_2^2 H^{-1}$ is $f(H)$. By the spectral mapping theorem, the eigenvalues of $c_1^2 H + c_2^2 H^{-1}$ are the values of f acting on the eigenvalues of H. Hence $c_1^2 H + c_2^2 H^{-1} \geq (\min_{x>0} f(x)) I$. Since the minimum of f is attained for $x = \frac{c_2}{c_1}$ and $f(\frac{c_2}{c_1}) = 2c_1 c_2$, we see that $c_1^2 AA^* + c_2^2 (A^{-1})^* A^{-1} \geq 2c_1 c_2 I$. This gives $YY^* \geq 4c_1 c_2 I$, proving the lemma. □

An alternative proof runs as follows: Let $A = UDV^*$ be a SVD for A. It follows that $(A^{-1})^* = UD^{-1}V^*$ and therefore $Y^{-1} = V(c_1 D + c_2 D^{-1})^{-1} U^*$. Its singular values are the entries on the diagonal of the matrix $(c_1 D + c_2 D^{-1})^{-1}$. Consider the function $f(x) = c_1 x + c_2 \frac{1}{x}$. By the inequality between the arithmetic and geometric means, $f(x) \geq 2\sqrt{c_1 c_2}$, and the minimum $2\sqrt{c_1 c_2}$ is attained at $\sqrt{\frac{c_2}{c_1}}$. This implies that the largest entry of $(c_1 D + c_2 D^{-1})^{-1}$ is less than or equal to $\frac{1}{2\sqrt{c_1 c_2}}$, and hence $\|Y^{-1}\| \leq \frac{1}{2\sqrt{c_1 c_2}}$.

As a concrete example, take $Y_n = T_n(\frac{1}{n})$, and observe that in this case the matrix (3.2) is intimately connected to the matrix (3.1). In fact, $Y_n = X_n$ for the sequence $f_n = \frac{1}{n}$. Here $T_{0,n}$ plays the role of A with $c_1 = 1$ and $c_2 = \frac{1}{n}$. Therefore, Lemma 3.2.1 implies

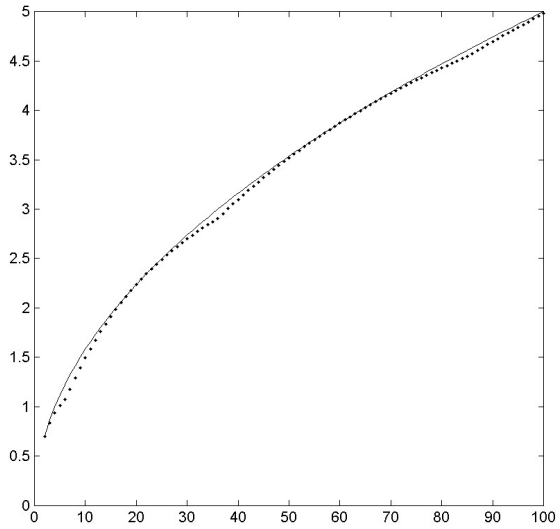

Figure 3.2: The norm of $T_n^{-1}(\frac{1}{n})$ (dots) and $\frac{\sqrt{n}}{2}$ (line) plotted as functions of n.

that $\|T_n^{-1}(\frac{1}{n})\| \leq \frac{\sqrt{n}}{2}$. From Figure 3.2 it seems at least visually then that asymptotically, $\|T_n^{-1}(\frac{1}{n})\|$ and $\frac{\sqrt{n}}{2}$ should behave the same.

We now prove a proposition that gives the upper bounds for $\|X_n^{-1}\|$ and $\|Z_n^{-1}\|$.

Proposition 3.2.2. *Let T_n be an invertible banded Toeplitz matrix and f_n a sequence such that $f_n > 0$ for all $n \in \mathbb{N}$.*

(i) Put $X_n = T_n + f_n(T_n^{-1})^*$. Then,

$$\|X_n^{-1}\| \leq \frac{1}{2\sqrt{f_n}}.$$

(ii) Put $Z_n = V_nT_n + f_n(T_n^{-1})^* + W_n$ with $W_n = (T_n^{-1})^* \sum_{j=1}^n w_j e_j e_j^*$ and $V_n = \sum_{j=1}^n v_j e_j e_j^*$. Also, let $w_j \geq 0$ and $v_j > 0$ for $1 \leq j \leq n$. Then,

$$\|Z_n^{-1}\| \leq \frac{1}{2} \max_{1\leq j\leq n} \left\{\frac{1}{\sqrt{v_j}}\right\} \max_{1\leq j\leq n} \left\{\frac{1}{\sqrt{f_n + w_j}}\right\}.$$

Proof. (i): A direct consequence of Lemma 3.2.1.

(ii): We start by rewriting Z_n as

$$\begin{aligned} Z_n &= V_nT_n + f_n(T_n^{-1})^* + W_n \\ &= V_nT_n + (T_n^{-1})^* \sum_{j=1}^n (f_n + w_j) e_j e_j^*. \end{aligned}$$

Denote $E_n = \sum_{j=1}^{n}(f_n + w_j)e_j e_j^*$ and notice that it is a diagonal matrix with positive entries. We write $E_n^{1/2}$ for its square root and now we can express Z_n as

$$\begin{aligned} Z_n &= (V_n T_n E_n^{-1/2} + (T_n^{-1})^* E_n^{1/2})E_n^{1/2} \\ &= V_n^{1/2}(V_n^{1/2} T_n E_n^{-1/2} + V_n^{-1/2}(T_n^{-1})^* E_n^{1/2})E_n^{1/2} \end{aligned}$$

If we set $A_n = V_n^{1/2} T_n E_n^{-1/2}$, we see that $(A_n^{-1})^* = V_n^{-1/2}(T_n^{-1})^* E_n^{1/2}$ and

$$Z_n = V_n^{1/2}(A_n + (A_n^{-1})^*)E_n^{1/2}, \qquad Z_n^{-1} = E_n^{-1/2}(A_n + (A_n^{-1})^*)^{-1} V_n^{-1/2}.$$

We can now find an upper bound for the norm of Z_n^{-1} since Lemma 3.2.1 applies directly to the middle factor in the expression above. Indeed,

$$\begin{aligned} E_n^{-1/2} &= \mathrm{Diag}\left(\sqrt{\frac{1}{f_n + w_1}}, \ldots, \sqrt{\frac{1}{f_n + w_n}}\right), \\ V_n^{-1/2} &= \mathrm{Diag}\left(\sqrt{\frac{1}{v_1}}, \ldots, \sqrt{\frac{1}{v_n}}\right), \end{aligned}$$

giving

$$\begin{aligned} \|Z_n^{-1}\| &\le \|V_n^{-1/2}\| \|(A_n + (A_n^{-1})^*)^{-1}\| \|E_n^{-1/2}\| \\ &\le \frac{1}{2} \max_{1 \le j \le n} \left\{ \frac{1}{\sqrt{v_j}} \right\} \max_{1 \le j \le n} \left\{ \frac{1}{\sqrt{f_n + w_j}} \right\}, \end{aligned}$$

as desired. □

If we choose $f_n = \frac{1}{n}$, $V_n = I_n$, $T_n = T_{0,n}$ and $w_1 = 1$ with all other $w_j = 0$, then $Z_n = K_n$ (compare (3.1)). Proposition 3.2.2 then gives $\|K_n^{-1}\| \le \frac{\sqrt{n}}{2}$ for $n \ge 2$. Figure 3.3 shows a plot of the functions $\frac{\sqrt{n}}{2}$ and $\|K_n^{-1}\|$. Again the close relation between these two quantities warrants the suspicion that their asymptotic behaviour should be the same.

3.3 Norm asymptotics of X_n^{-1} and Z_n^{-1}

In this section we prove our main result for X_n^{-1} and Z_n^{-1} as defined before. The norm of the inverse of $T_n(\frac{1}{n})$ and K_n will then immediately follow as special cases.

To arrive at our main result, we need the following two lemmas. The first was stated in the introduction but given here again for ease of reference. The second is a technical lemma that we shall prove.

Lemma 3.3.1. *Let b be a Laurent polynomial of the form $b(t) = \sum_{j=-r}^{r} b_j t^j$, $t \in \mathbb{T}$. If $E \subset \mathbb{R}$ is any segment, then*

$$|N_n(E) - n\mu(E)| \le 14r \qquad \text{for all} \quad n \ge 1,$$

where $N_n(E)$ is the number of singular values of $T_n(b)$ in E and

$$\mu(E) = \frac{1}{2\pi} |\{t \in \mathbb{T} : |b(t)| \in E\}|,$$

with $|.|$ denoting the Lebesgue measure on the unit circle.

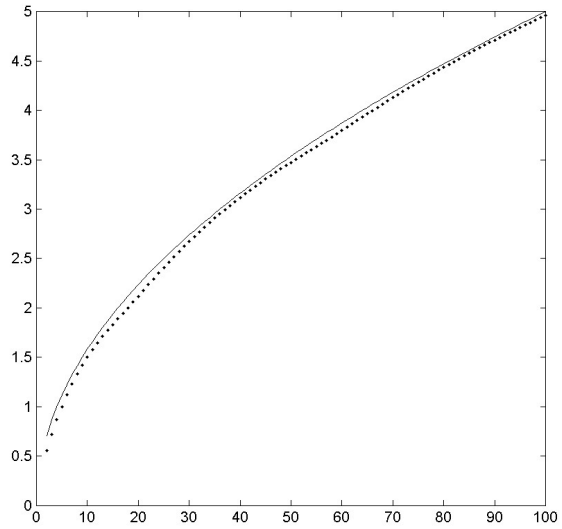

Figure 3.3: The norm of K_n^{-1} (dots) and $\frac{\sqrt{n}}{2}$ (line) plotted as functions of n

Lemma 3.3.2. *Let $b(t)$ be a Laurent polynomial. We will show that the following items hold true:*

- *If $b(t)$ is not identically zero, then $|b(t)|$ has a finite number of zeros on the unit circle.*
- *$\frac{d}{d\theta}|b(e^{i\theta})|$ exists and is bounded on the interval $[0, 2\pi]$ except where $|b(e^{i\theta})| = 0$.*
- *If $|b(t)|$ is not constant on $\mathbb{T}$, then $\frac{d}{d\theta}|b(e^{i\theta})|$ has only a finite number of zeros in the interval $[0, 2\pi]$.*

Proof. Note that $|b(t)| = |t^r b(t)|$ on $\mathbb{T}$ and $p(t) := t^r b(t)$ is a polynomial of degree $2r$. Therefore p, and consequently also $|b(t)|$, has at most $2r$ zeros on $\mathbb{T}$, provided $p \neq 0$.

The function $|b(e^{i\theta})|^2$ is a trigonometric polynomial and hence its derivative is bounded. Outside its zeros on $[0, 2\pi]$, we have

$$\frac{d}{d\theta}|b(e^{i\theta})|^2 = 2|b(e^{i\theta})|\frac{d}{d\theta}|b(e^{i\theta})|.$$

It remains to show that for each zero $t_0 = e^{i\theta_0}$ of $b(t)$, there is a neighbourhood U of θ_0 such that $\frac{d}{d\theta}|b(e^{i\theta_0})|$ is bounded on $U \setminus \{\theta_0\}$. For a sufficiently small neighbourhood U we have $|b(t)| = |t - t_0|^\alpha \psi(t)$ with a positive integer α and a positive C^∞ function $\psi(t)$. Thus,

$$|b(e^{i\theta})| = 2^\alpha \left| \sin \frac{\theta - \theta_0}{2} \right|^\alpha \psi(e^{i\theta}),$$

and this obviously has a bounded derivative in $U \setminus \{\theta_0\}$, proving the second item.

Finally, to prove the third item we argue as follows: By assumption, $\frac{d}{d\theta}|b(e^{i\theta})|^2 = 2|b(e^{i\theta})|\frac{d}{d\theta}|b(e^{i\theta})|$ does not vanish identically on $[0, 2\pi]$ minus the zero set of $|b(e^{i\theta})|$. By

applying the first item to $\frac{d}{d\theta}|b(e^{i\theta})|^2$, the right-hand side has only finitely many zeros on $[0, 2\pi]$, proving the final item. □

Incidentally, by Proposition 5.6 of [5], the Laurent polynomials with constant modulus on $\mathbb{T}$ are just the monomials γt^m, with $\gamma \in \mathbb{C}$ and $m \in \mathbb{Z}$.

Proof of Theorem 3.1.1. Let $T_n = U_n D_n V_n^*$ be a SVD. Hence,

$$X_n = T_n + f_n(T_n^{-1})^* = U_n D_n V_n^* + f_n U_n D_n^{-1} V_n^* = U_n(D_n + f_n D^{-1})V_n^*$$

and

$$X_n^{-1} = V_n(D_n + f_n D_n^{-1})^{-1} U_n^*.$$

The singular values of X_n^{-1} are the entries of the diagonal matrix $(D_n + f_n D_n^{-1})^{-1}$. If we set $g_n(x) := x + f_n x^{-1}$, then $1/g_n(x)$ produces the singular values of X_n^{-1} when we evaluate the function in the singular values of T_n. We observe by elementary calculus that $1/g_n(x)$ has its maximum at the point $x = \sqrt{f_n}$, and if we can find a singular value of T_n close to this point, we are in business. To show that there exist singular values arbitrarily close to $\sqrt{f_n}$, Lemma 3.3.1 will be pivotal in our argument.

Indeed, we apply this lemma with

$$E_n = [\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}}, \sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}]$$

for $\epsilon > 0$ and some k to be determined later. It is clear that E_n is an interval around the point $\sqrt{f_n}$ with decreasing length as n grows, and we will prove that this interval contains a singular value of $T_n(b)$ (in our case T_n) for all n large enough.

The first item of Lemma 3.3.2 allows us to choose an interval (θ_0, θ_f), where $|b(e^{i\theta_0})| = 0$ and θ_f is the first point to the right (left hand version follows similarly) of θ_0 such that $\frac{d}{d\theta}|b(e^{i\theta_f})| = 0$. This is possible since there are only a finite number of points where $\frac{d}{d\theta}|b(e^{i\theta_f})| = 0$ (third item, Lemma 3.3.2). (Note that it could happen that there is no point to the right (or left) of θ_0 where the derivative is zero. In this case $|b(e^{i\theta})|$ would be monotonically increasing on $(\theta_0, 2\pi]$ (or monotonically decreasing on $[0, \theta_0)$) since a negative (or positive) derivative would imply a point in the interval where the derivative must be zero, contradicting our assumption).

Now, $\frac{d}{d\theta}|b(e^{i\theta})|$ is positive in (θ_0, θ_f) (if the derivative is negative in the interval it would imply that there must be a point closer to θ_0 where the derivative is zero) and bounded (second item, Lemma 3.3.2). This implies that $|b(e^{i\theta})|$ is monotonically increasing, and hence for n large enough there exist θ_f^b and θ_i^b in the interval such that $|b(e^{i\theta_f^b})| = \sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}$ and $|b(e^{i\theta_f^i})| = \sqrt{f_n} - \sqrt{f_n^{\epsilon+1}}$. In other words, $|b(e^{i\theta})|$ maps the interval $[\theta_i^b, \theta_f^b]$ onto E_n.

Now let $F(\theta) = M\theta$ denote a function of a straight line through the origin, with $M \geq \frac{d}{d\theta}|b(e^{i\theta})|$ for $\theta \in [\theta_i^b, \theta_f^b]$, and large enough such that

$$\theta_i^F = \frac{\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}}}{M} \qquad \text{and} \qquad \theta_f^F = \frac{\sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}}{M}$$

are the two points in $[0, 2\pi]$ where F maps the interval $[\theta_i^F, \theta_f^F]$ exactly onto our interval E_n. Since F is a straight line, we have that

$$M = \frac{F(\theta_f^F) - F(\theta_i^F)}{\theta_f^F - \theta_i^F}.$$

By the mean value theorem we also have

$$\frac{|b(e^{i\theta_f^b})| - |b(e^{i\theta_i^b})|}{\theta_f^b - \theta_i^b} = \frac{d}{d\theta}|b(e^{i\theta_c^b})|$$

for some θ_c^b in $(0, 2\pi)$. Now,

$$\left|\frac{|b(e^{i\theta_f^b})| - |b(e^{i\theta_i^b})|}{\theta_f^b - \theta_i^b}\right| = \left|\frac{d}{d\theta}|b(e^{i\theta_c^b})|\right| \leq M = \frac{F(\theta_f^F) - F(\theta_i^F)}{\theta_f^F - \theta_i^F}.$$

Since $F(\theta_f^F) = |b(e^{i\theta_f^b})|$ and $F(\theta_i^F) = |b(e^{i\theta_i^b})|$, the previous inequality implies that $\theta_f^F - \theta_i^F \leq |\theta_f^b - \theta_i^b|$, i.e.,

$$|\{\theta \in [0, 2\pi] \mid F(\theta) \in E_n\}| \leq |\{\theta \in [0, 2\pi] \mid |b(e^{i\theta})\}| \in E_n|.$$

This can be directly related to $\mu(E_n)$ via Lemma 3.3.1. Since $\mu(E_n)$ remains unchanged if we replace $e^{i\theta} \in \mathbb{T}$ with $\theta \in [0, 2\pi]$, we find that

$$\mu(E_n) = \frac{1}{2\pi}\left|\{\theta \in [0, 2\pi] \mid |\sum_{j=-r}^{r} b_j e^{ij\theta}| \in E_n\}\right|.$$

Therefore,

$$\frac{n}{2\pi}\frac{1}{M}|E_n| = \frac{n}{2\pi}\frac{1}{M}(k+1)\sqrt{f_n^{\epsilon+1}} \leq n\mu(E_n)$$

for n large enough. Now taking k and n large enough, and keeping in mind the restriction on the decay of f_n, we see that

$$n\mu(E_n) \geq \frac{n\sqrt{f_n^{\epsilon+1}}(k+1)}{2\pi M} > 15r, \tag{3.3}$$

so that for this value of k we must have $N_n(E_n) \geq r$ from the inequality

$$|N_n(E_n) - n\mu(E_n)| \leq 14r.$$

That means that there is at least one singular value of T_n in the interval

$$E_n = [\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}}, \sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}]$$

for n large enough. Denote this sequence of singular values as

$$\sigma_n = \sqrt{f_n} + s_n\sqrt{f_n^{\epsilon+1}},$$

where $|s_n| \leq k$ for all n large enough. If we insert σ_n into the function $1/g_n(x)$, we get a singular value of X_n^{-1}, which is immediately a lower bound for its norm, i.e.,

$$\|X_n^{-1}\| \geq \frac{1}{\sqrt{f_n} + s_n\sqrt{f_n^{\epsilon+1}} + \frac{f_n}{\sqrt{f_n}+s_n\sqrt{f_n^{\epsilon+1}}}}.$$

Recall that we also have an upper bound, and combining this with the previous lower bound, we have

$$\frac{1}{2\sqrt{f_n}} \geq \|X_n^{-1}\| \geq \frac{\sqrt{f_n} + s_n\sqrt{f_n^{\epsilon+1}}}{\left(\sqrt{f_n} + s_n\sqrt{f_n^{\epsilon+1}}\right)^2 + f_n}$$

or multiplying through with a constant,

$$\begin{aligned} 1 \geq 2\sqrt{f_n}\|X_n^{-1}\| \geq & \frac{2\sqrt{f_n}(\sqrt{f_n} + s_n\sqrt{f_n^{\epsilon+1}})}{(\sqrt{f_n} + s_n\sqrt{f_n^{\epsilon+1}})^2 + f_n} \\ = & \frac{2f_n + 2s_n\sqrt{f_n^{\epsilon+2}}}{2f_n + 2s_n\sqrt{f_n^{\epsilon+2}} + s_n^2 f_n^{\epsilon+1}} \\ = & \frac{2 + 2s_n f_n^{\frac{\epsilon}{2}}}{2 + 2s_n f_n^{\frac{\epsilon}{2}} + s_n^2 f_n^{\epsilon}}. \end{aligned}$$

Taking limits of the first, second and last term in the inequality above, we see that

$$1 \geq \lim_{n\to\infty} 2\sqrt{f_n}\|X_n^{-1}\| \geq 1,$$

proving the theorem. □

Corollary 3.3.3. *Choosing $f_n = \frac{1}{n}$ and $T_n = T_{0,n}$, we have the desired result for $T_n^{-1}(\frac{1}{n})$, i.e., $\lim_{n\to\infty} \frac{2}{\sqrt{n}}\|T_n^{-1}(\frac{1}{n})\| = 1$.*

We will apply the same strategy for $\lim_{n\to\infty} \|Z_n^{-1}\|$ as described above, but with a slight modification involving the interlacing of singular values.

Consider a general $m \times n$ matrix A. Form the decreasing sequence $\alpha_1 \geq \cdots \geq \alpha_{\min(m,n)}$ of the singular values of A. We set $\alpha_t = 0$ for $\max(m,n) \geq t > \min(m,n)$. Let $Q_{mp}[Q_{nq}]$ be the set of all $\binom{m}{p}[\binom{n}{q}]$ sequences $\{i_1 \dots i_p\}[\{j_1 \dots j_q\}]$ of integers

$$1 \leq i_1 \leq \cdots \leq i_p \leq m[1 \leq j_1 \leq j_q \leq n].$$

By $A[\omega|\tau]$ we denote the submatrix of A, consisting of the elements in the intersection of the $i_1 \dots i_p$ ($\omega := \{i_1 \dots i_p\}$) rows and $j_1 \dots j_q$ ($\tau := \{j_1 \dots j_q\}$) columns of A. Form the decreasing sequence $\beta_{\omega\tau,1} \geq \cdots \geq \beta_{\omega\tau,\min(p,q)}$ of singular values of $A[\omega|\tau]$. Let $\beta_{\omega\tau,t} = 0$ for $t > \min(p,q)$. Then from [39] we have the following theorem.

Theorem 3.3.4. *If $\omega \in Q_{mp}$, $\tau \in Q_{nq}$:*

- $\alpha_i \geq \beta_{\omega\tau,i} \qquad i = 1, 2, \dots, \min(p,q)$
- $\beta_{\omega\tau,i} \geq \alpha_{i+(m-p)+(n-q)} \qquad i \leq \min(p+q-m, p+q-n)$

Proof of Theorem 3.1.2. Recall that $Z_n = V_n T_n + f_n(T_n^{-1})^* + W_n$. Let W_n have non-zero entries only from the $(1,1)$ to (q_w, q_w) entry. Let V_n have entries greater or equal to one only from the $(1,1)$ to (q_v, q_v) entry, and all other entries equal to one. These restrictions on V_n and W_n are necessary, since then Proposition 3.2.2 still gives the upper bound of $\frac{1}{2\sqrt{f_n}}$ for n large enough. Choose $p = \max\{q_w, q_v\}$. Now, let Z_n play the role of A and

U_{n-p} be the submatrix $A[\omega|\tau]$, with $\omega = \tau = \{1, \ldots, n\}$, as in the discussion preceding Theorem 3.3.4. Notice that we obtain U_{n-p} by deleting the first p rows and columns of Z_n. In this case, the second item in Theorem 3.3.4 can be rewritten as

$$\beta_i \geq \alpha_{i+2p} \qquad \text{and} \qquad \beta_{n-2p} \geq \alpha_n, \qquad (i \leq n-2p),$$

where we omitted the subscripts ω and τ for convenience. Combining this inequality with the fact that $\|Z_n^{-1}\| = \frac{1}{\alpha_n}$, we have $\|Z_n^{-1}\| \geq \frac{1}{\beta_{n-2p}}$.

Now we do not know enough about U_{n-p} to say something meaningful about its singular values to relate it to our original matrix Z_n. However, we will show that we can relate its singular values to those of X_n and use it to calculate a lower bound for $\|Z_n^{-1}\|$ in a manner analogous to the proof of $\|X_n^{-1}\|$.

Define

$$Q = \begin{bmatrix} 0 & I_{n-p} \end{bmatrix} \tag{3.4}$$

as the block matrix where 0 is a $p \times (n-p)$ block of zeros and I_{n-p} denotes the identity matrix of size $n-p$. Here $p = \max\{q_w, q_v\}$ - see the definition of Z_n above. Doing a straightforward multiplication, we see that

$$QZ_nQ^* = U_{n-p}.$$

Now we can calculate what U_{n-p} is since we have an expression for Z_n:

$$\begin{aligned} U_{n-p} &= Q(V_nT_n + f_n(T_n^{-1})^* + W_n)Q^* \\ &= QV_nT_nQ^* + f_nQ(T_n^{-1})^*Q^* + QW_nQ^* \\ &= QT_nQ^* + f_nQ(T_n^{-1})^*Q^* \\ &= Q(T_n + f_n(T_n^{-1})^*)Q^* \\ &= QX_nQ^*. \end{aligned}$$

The third equality follows from the easily checked fact that $QW_nQ^* = 0$ and $QV_n = Q$. This equality can now be used via the following result from [19], Lemma 3.3.1, to relate the singular values of Z_n, X_n and U_{n-p}.

Lemma 3.3.5. *Let $C \in \mathbb{C}^{m\times n}$, $V_k \in \mathbb{C}^{m\times k}$ and $W_k \in \mathbb{C}^{n\times k}$ be given, where $k \leq \min\{m,n\}$ and V_k, W_k have orthonormal columns. Then,*

$$\sigma_i(V_k^*CW_k) \leq \sigma_i(C), \quad i = 1, \ldots, k.$$

In our case, let $C = X_n$, $V_k = Q^*$ and $W_k = Q^*$. Then Lemma 3.3.5 tells us that

$$\sigma_i(U_{n-p}) \leq \sigma_i(X_n), \quad i = 1, \ldots, n-p$$

since $U_{n-p} = QX_nQ^*$. Also, from the previous lemma it follows that $\beta_{n-2p} \leq \sigma_{n-2p}(X_n)$. Therefore,

$$\alpha_n \leq \beta_{n-2p} \leq \sigma_{n-2p}(X_n).$$

Now $\sigma_{n-2p}(X_n)$ is the $(2p+1)$'th smallest singular value of X_n. Define

$$E_{n-2p} = [\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}}, \sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}].$$

Analogous to the proof of Theorem 3.1.1, we arrive at the inequality

$$(n-2p)\mu(E_{n-2p}) \geq \frac{(n-2p)\sqrt{f_n^{\epsilon+1}}(k+1)}{2\pi}.$$

Observe that this inequality is similar to (3.3) and plays the same role in the current proof. Now, taking k and n large as before, we have

$$(n-2p)\mu(E_{n-2p}) \geq \frac{(n-2p)\sqrt{f_n^{\epsilon+1}}(k+1)}{2\pi} > 2p+14r+1$$

which implies that $N_{n-2p}(E_{n-2p}) \geq 2p+1$. We now obtain that for large enough n there exist $2p+1$ distinct singular values of T_n, taking multplicity into account, inside E_{n-2p}. Note that these singular values have the form

$$\sigma_n^{(j)} = \sqrt{f_n} + s_n^{(j)}\sqrt{f_n^{\epsilon+1}}, \qquad 1 \leq j \leq 2p+1$$

where $|s_n^{(j)}| \leq k$. The associated sequences of singular values of X_n can be written as

$$\sigma_n^{(j)}(X_n) = g_n(\sqrt{f_n} + s_n^{(j)}\sqrt{f_n^{\epsilon+1}}) \qquad 1 \leq j \leq 2p+1.$$

Now form the sequence $\sigma_n^{\max}(X_n) = \max_{1\leq j\leq 2p+1}\{\sigma_n^{(j)}\}$. Clearly, $\sigma_n^{\max}(X_n) \geq \sigma_{n-p}(X_n) \geq \beta_{n-2p}$ and therefore $\|Z_n^{-1}\| \geq \frac{1}{\beta_{n-2p}} \geq \frac{1}{\sigma_n^{\max}(X_n)}$ where

$$\frac{1}{\sigma_n^{\max}(X_n)} = \frac{1}{g_n(\sqrt{f_n} + s_n^{(j_n)}\sqrt{f_n^{\epsilon+1}})},$$

with $j_n \in \{1,\ldots,2p+1\}$. We also have $\frac{1}{2\sqrt{f_n}} \geq \|Z_n^{-1}\|$ for n large enough from Proposition 3.2.2, and again, following the same line of reasoning as before, we arrive at

$$\lim_{n\to\infty} 2\sqrt{f_n}\|Z_n^{-1}\| = 1. \qquad \square$$

Corollary 3.3.6. *Choosing $T_n = T_{0,n}$, $f_n = 1/n$, $V_n = I_n$ and $w_1 = 1$ in W_n with all other $w_j = 0$, we arrive at the special case where $Z_n = K_n$ - compare* (3.1)*. Then,* $\lim_{n\to\infty} \frac{2}{\sqrt{n}}\|K_n^{-1}\| = 1$.

We have shown that, asymptotically, $\frac{1}{2\sqrt{f_n}}$, $\|X_n\|^{-1}$ and $\|Z_n\|^{-1}$ behave the same. However, the main results (Theorem 3.1.1, Theorem 3.1.2) still allow for $\frac{1}{2\sqrt{f_n}} - \|X_n\|^{-1}$ and $\frac{1}{2\sqrt{f_n}} - \|Z_n\|^{-1}$ to be greater than any fixed positive constant for all n. From the graphs provided in this section, it seems that these quantities have a stronger relation, and might even converge in some sense. The next section is devoted to this question and provides stronger versions of Theorem 3.1.1 and Theorem 3.1.2.

Sections 3.4 - 3.5 contain the results of [27], including some improvements and a few additional results.

3.4 Convergence of $\frac{1}{2\sqrt{f_n}} - \|X_n^{-1}\|$

We shall prove the following stronger versions of Theorems 3.1.1 and 3.1.2:

Theorem 3.4.1. *Assuming the same conditions as in Theorem 3.1.1, except that $\epsilon > 1/2$, we have*

$$\left(\frac{1}{2\sqrt{f_n}} - \|X_n^{-1}\|\right) = O\left(f_n^{\epsilon-\frac{1}{2}}\right). \tag{3.5}$$

Note that $\epsilon > 1/2$ implies $f_n > O\left(\frac{1}{n^{\frac{4}{3}}}\right)$.

This theorem applies to Z_n^{-1} in place of X_n^{-1} as well when the conditions of Theorem 3.1.2 are assumed with the same exception as above.

Recall that

$$\frac{1}{2\sqrt{f_n}} \geq \|X_n^{-1}\| \geq \frac{1}{g_n(\sigma_n)} = \frac{\sqrt{f_n} + s_n\sqrt{f_n^{\epsilon+1}}}{2f_n + 2s_n\sqrt{f_n^{\epsilon+2}} + s_n^2 f_n^{\epsilon+1}} := R_n,$$

from the proof of Theorem 3.1.1, with $\sigma_n = \sqrt{f_n} + s_n\sqrt{f_n^{\epsilon+1}}$. We can rewrite this inequality as

$$\frac{1}{2\sqrt{f_n}} - R_n \geq \frac{1}{2\sqrt{f_n}} - \|X_n^{-1}\| \geq 0. \tag{3.6}$$

Ideally we would like

$$\lim_{n\to\infty} \frac{1}{2\sqrt{f_n}} - R_n = 0,$$

but this is not possible since R_n depends on s_n which is not necessarily convergent and we cannot take this limit as n tends to infinity.

Thus we need to define a similar sequence to R_n by replacing s_n with some other quantity of which we can take the limit, or that does not depend on n. Also, this associated expression of R_n, say $\tilde{R}_n$, has to be a lower bound for R_n so that we can replace R_n with $\tilde{R}_n$ in (3.6).

Define

$$\tilde{R}_n = \frac{\sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}}{2f_n + 2k\sqrt{f_n^{\epsilon+2}} + k^2 f_n^{\epsilon+1}}$$

where we just replaced all the s_n with k. Therefore,

$$\tilde{R}_n = \frac{1}{g_n(\sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}})}.$$

In the following Lemma we prove that $\tilde{R}_n \leq R_n$.

Lemma 3.4.2. *Let $\tilde{R}_n$ and R_n be as defined above. Then for n and k large enough, $R_n \geq \tilde{R}_n$.*

Proof. As mentioned before, $\frac{1}{g_n(x)}$ has its maximum at $\sqrt{f_n}$ and notice that it is monotonically increasing for $0 \le x \le \sqrt{f_n}$, and monotonically decreasing for $x \ge \sqrt{f_n}$. Therefore, if s_n is negative or positive respectively, we have

$$\frac{1}{g_n(\sigma_n)} \ge \frac{1}{g_n(\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}})} \qquad \text{or} \qquad \frac{1}{g_n(\sigma_n)} \ge \frac{1}{g_n(\sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}})}.$$

We don't know whether s_n will be positive or negative, but we will show that

$$\frac{1}{g_n(\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}})} \ge \frac{1}{g_n(\sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}})}$$

for n and k large enough. This covers both possibilities and then we are done.

Consider the difference

$$\begin{aligned}
&\frac{1}{g_n(\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}})} - \frac{1}{g_n(\sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}})} \\
=&\frac{1}{\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}} + f_n\frac{1}{\sqrt{f_n}-\sqrt{f_n^{\epsilon+1}}}} - \frac{1}{\sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}} + f_n\frac{1}{\sqrt{f_n}+k\sqrt{f_n^{\epsilon+1}}}} \\
=&\frac{\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}}}{(\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}})^2 + f_n} - \frac{\sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}}{(\sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}})^2 + f_n} \\
=&\frac{\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}}}{2f_n - 2\sqrt{f_n^{\epsilon+2}} + f_n^{\epsilon+1}} - \frac{\sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}}{2f_n + 2k\sqrt{f_n^{\epsilon+2}} + k^2 f_n^{\epsilon+1}} \\
=&\frac{\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}}}{u_n} - \frac{\sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}}{v_n} \\
=&\frac{(\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}})(2f_n + 2k\sqrt{f_n^{\epsilon+2}} + k^2 f_n^{\epsilon+1}) - (\sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}})(2f_n - 2\sqrt{f_n^{\epsilon+2}} + f_n^{\epsilon+1})}{w_n} \\
=&\frac{f_n^{\frac{2\epsilon+3}{2}}(k^2-1) + f_n^{\frac{3\epsilon+3}{2}}(-k^2-k)}{w_n},
\end{aligned}$$

where

$$\begin{aligned}
u_n &= 2f_n - 2\sqrt{f_n^{\epsilon+2}} + f_n^{\epsilon+1}, \\
v_n &= 2f_n + 2k\sqrt{f_n^{\epsilon+2}} + k^2 f_n^{\epsilon+1}, \\
w_n &= u_n v_n.
\end{aligned}$$

Notice that u_n, v_n and w_n are all positive for n large enough since the limit of f_n is zero as n goes to infinity. Then, for n large enough and $k > 1$,

$$f_n^{\frac{2\epsilon+3}{2}}(k^2-1) + f_n^{\frac{3\epsilon+3}{2}}(-k^2-k) > 0.$$

□

Proof of Theorem 3.4.1. Applying Lemma 3.4.2, we can rewrite (3.6) with R_n replaced by $\tilde{R}_n$:

$$\frac{1}{2\sqrt{f_n}} - \frac{\sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}}{2f_n + 2k\sqrt{f_n^{\epsilon+2}} + k^2 f_n^{\epsilon+1}} \ge \frac{1}{2\sqrt{f_n}} - \|X_n^{-1}\| \ge 0.$$

Clearly we want the left side of the inequality to converge to zero, so that is what we will show. Now,

$$\begin{aligned}
&\frac{1}{2\sqrt{f_n}} - \frac{\sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}}{2f_n + 2k\sqrt{f_n^{\epsilon+2}} + k^2 f_n^{\epsilon+1}} \\
=&\frac{1}{2\sqrt{f_n}} \left(1 - \frac{2f_n + 2k\sqrt{f_n^{\epsilon+2}}}{2f_n + 2k\sqrt{f_n^{\epsilon+2}} + k^2 f_n^{\epsilon+1}} \right) \\
=&\frac{1}{2\sqrt{f_n}} \left(\frac{k^2 f_n^{\epsilon+1}}{2f_n + 2k\sqrt{f_n^{\epsilon+2}} + k^2 f_n^{\epsilon+1}} \right) \\
=&\frac{k^2 f_n^{\epsilon+1}}{4f_n^{\frac{3}{2}} + 4kf_n^{\frac{\epsilon+3}{2}} + 2k^2 f_n^{\frac{2\epsilon+3}{2}}}.
\end{aligned}$$

The last expression will only converge to zero if $\epsilon > 1/2$. Assuming that is the case, let $\epsilon = 1/2 + \delta$, where $\delta > 0$, and then

$$\frac{k^2 f_n^{\epsilon+1}}{4f_n^{\frac{3}{2}} + 4kf_n^{\frac{\epsilon+3}{2}} + 2k^2 f_n^{\frac{2\epsilon+3}{2}}} = \frac{k^2 f_n^{\frac{3}{2}+\delta}}{4f_n^{\frac{3}{2}} + 4kf_n^{\frac{\frac{1}{2}+\delta+3}{2}} + 2k^2 f_n^{\frac{4+2\delta}{2}}} = O(f_n^{\delta}) = O\left(f_n^{\epsilon-\frac{1}{2}}\right).$$

□

Recall that ϵ is a free but bounded parameter, since by the conditions of Theorem 3.1.1 $n\sqrt{f_n^{\epsilon+1}}$ must remain greater than a positive constant c for n large enough. It was used in defining the interval $E_n = [\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}}, \sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}]$ that contained some singular values of T_n. Notice that higher values of ϵ imply that this interval's length decrease more rapidly as n grows. With this in mind, we can find the largest $\epsilon > 1/2$ for a given f_n such that the conditions of Theorem 3.1.1 are satisfied. This ϵ would then give the fastest order of convergence in the equality above.

We note that the proof of Theorem 3.4.1 for the case Z_n^{-1} is analogous. The values of s_n and k might be different, but that does not change the arguments used above.

Corollary 3.4.3. *Let f_n be of the form $f_n = 1/n^x$, and satisfy the conditions of Theorem 3.4.1. Then,*

$$\frac{1}{2\sqrt{f_n}} - \|X_n^{-1}\| = O\left(\frac{1}{n^{\frac{4-3x}{2}}}\right).$$

This means that there is an inverse relationship between the order of f_n and the order of $\frac{1}{2\sqrt{f_n}} - \|X_n^{-1}\|$.

Proof. For f_n to satisfy the conditions of Theorem 3.1.1, x must be such that $0 < x < 2$. Now, we want to find the maximal ϵ for the particular choice of f_n such that $n\sqrt{f_n^{\epsilon+1}}$ remains greater than some positive constant c for n large enough:

$$n\sqrt{f_n^{\epsilon+1}} = n(\frac{1}{n^x})^{\frac{\epsilon+1}{2}} = \frac{n}{n^{\frac{x(\epsilon+1)}{2}}} > c.$$

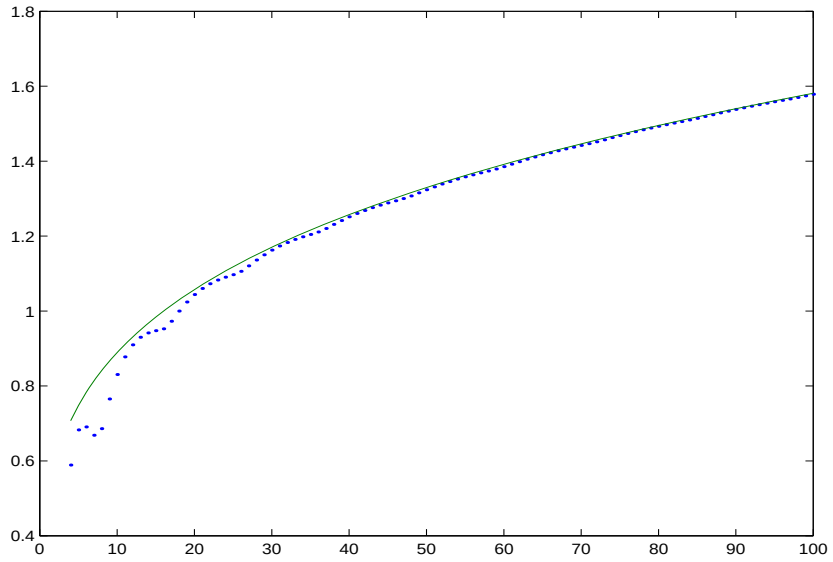

Figure 3.4: The norm of Z_n^{-1} (dots) and $\frac{n^{\frac{1}{4}}}{2}$ (line) plotted as functions of n.

Recall that $\epsilon > 1/2$, and this implies that the maximum value x can attain is $x = 4/3$. Also, for the last inequality to be true, we must have

$$\frac{x(\epsilon+1)}{2} \leq 1 \qquad \Longrightarrow \qquad \epsilon \leq \frac{2}{x} - 1.$$

Since we are interested in the maximal value of ϵ, we choose $\epsilon = \frac{2}{x} - 1$. Inserting these values into the convergence estimate, we have

$$\frac{1}{2\sqrt{f_n}} - \|X_n^{-1}\| = O\left(f_n^{\epsilon-\frac{1}{2}}\right) = O\left((\frac{1}{n^x})^{\frac{2}{x}-\frac{3}{2}}\right) = O\left(\frac{1}{n^{\frac{4-3x}{2}}}\right).$$

□

Example 3.4.4.

Let T_n be the banded Toeplitz matrix with symbol $b(t) = 3 - t - 2t^{-1}$, and $f_n = \frac{1}{\sqrt{n}}$. Choose $v_1 = 1, v_2 = 2, v_3 = 3$ with all other $v_j = 1$ in matrix V. Also, let $w_1 = 0.1, w_2 = 0.1$ with all other $w_j = 0$ in matrix W. Now form the matrix Z as defined earlier.

Notice that the matrix Z as defined above satisfies all the assumptions of Theorem 3.1.2. Graph 3.4 shows the behaviour of $\|Z_n^{-1}\|$.

Remark 3.4.5.

Our class of Toeplitz-generated matrices, $X_n = T_n + f_n(T_n^{-1})^*$, can be viewed as a class of matrices formed by a simple 'function' acting on the generator matrix T_n. We could then ask what would happen if we extended this class to more complicated 'functions'? It turns out that we can generalize Theorem 3.1.1 in the sense that we consider a bigger class of Toeplitz-generated matrices, but we pay a price in that we have to assume that T_n is positive definite.

3.4.1 The positive definite case

By restricting T_n to be positive definite, we gain the advantage of it having a SVD of the form $T_n = U_n D_n U_n^*$. This allows us to to enlarge the class, $X_n = T_n + f_n(T_n^{-1})^*$, to the more general case where we can take arbitrary powers (positive and negative) of the generator:

$$X_n = T_n^a + f_n T_n^{-b},$$

where a and b are strictly positive integers. In this section we denote the symbol of our generator matrix by B, to not confuse with the power b in the equation above.

Our modified theorem reads as follows.

Theorem 3.4.6. *Let f_n satisfy the conditions*

- $\lim_{n\to\infty} f_n = 0$,
- $\exists\, c > 0$ *such that* $n f_n^{\frac{\epsilon+1}{a+b}} > c$ *for n large enough and some* $\epsilon > 0$.

Assume that the symbol $B(e^{i\theta}) = \sum_{j=-r}^{r} B_j e^{ij\theta}$ ($B_j = 0$ for $|j| > r$) associated with the infinite Toeplitz matrix $T = (\beta_{j-k})_{j,k=1}^{\infty}$ has the following three properties:

- *B has at least one zero on $\mathbb{T}$ but is not identically zero,*
- *for n sufficiently large the principal $n \times n$ truncations T_n of T are invertible.*
- *T_n is positive definite for all n.*

Let $X_n = T_n^a + f_n T_n^{-b}$, where $a, b > 0$. Then,

$$\lim_{n\to\infty} \left[\left(\frac{b f_n}{a}\right)^{\frac{a}{a+b}} + f_n \left(\frac{b f_n}{a}\right)^{\frac{-b}{a+b}} \right] \|X_n^{-1}\| = 1.$$

Before proving this theorem, we first establish a new upper bound for $\|X_n^{-1}\|$.

Lemma 3.4.7. *If the matrix $X_n = T_n^a + f_n T_n^{-b}$ is defined as above. Then,*

$$\|X_n^{-1}\| \leq \frac{1}{(\frac{b f_n}{a})^{\frac{a}{a+b}} + f_n(\frac{b f_n}{a})^{\frac{-b}{a+b}}}.$$

Proof. Let $T_n = U_n D_n U_n^*$ be a SVD. Then $T_n^a = U D^a U^*$, $T_n^{-b} = U D^{-b} U^*$ and

$$X_n^{-1} = U(D^a + f_n D^{-b})^{-1} U^*.$$

Define $g_n(x) := x^a + f_n x^{-b}$ for positive values of x. Then if $1/g_n(x)$ has an absolute maximum for all positive x, it implies that this maximum will be an upper bound for the norm of $\|X_n^{-1}\|$ - compare Lemma 3.2.1. By taking the derivative of $1/g_n(x)$ and setting it to zero, we find its maximum to be at $x = (b f_n/a)^{1/(a+b)}$ with a value of

$$\frac{1}{(\frac{b f_n}{a})^{\frac{a}{a+b}} + f_n(\frac{b f_n}{a})^{\frac{-b}{a+b}}}.$$

□

Since most of the proof of our theorem here follows directly from the proof of Theorem 3.1.1, we will only highlight where adjustments need to be made to the previous proof.

Proof of Theorem 3.4.6. From the calculation of the upper bound in the lemma here, we know that a singular value of T_n that is 'closest' to $(bf_n/a)^{1/(a+b)}$ will determine the norm of X_n^{-1}. As before, we will show that there is a singular value arbitrarily close to $(bf_n/a)^{1/(a+b)}$ if n is large enough. Again, we will apply Lemma 3.3.1, but this time define our interval

$$E_n = \left[\left(\frac{bf_n}{a}\right)^{\frac{1}{a+b}} - \left(\frac{bf_n}{a}\right)^{\frac{\epsilon+1}{a+b}}, \left(\frac{bf_n}{a}\right)^{1/(a+b)} + k\left(\frac{bf_n}{a}\right)^{\frac{\epsilon+1}{a+b}} \right].$$

From here on everything follows exactly as the proof of Theorem 3.1.1 up to the inequality,

$$\frac{n}{2\pi}\frac{1}{m}|E_n| = \frac{n}{2\pi}\frac{1}{m}(k+1)\left(\frac{bf_n}{a}\right)^{\frac{\epsilon+1}{a+b}} \leq n\mu(E_n).$$

Taking k and n large enough we see that

$$n\mu(E_n) \geq \frac{n(k+1)(\frac{bf_n}{a})^{\frac{\epsilon+1}{a+b}}}{2\pi M} > 15r,$$

and this together with Lemma 3.3.1 implies that there is at least one singular value in the interval E_n. We denote this sequence of singular values as

$$\sigma_n = \left(\frac{bf_n}{a}\right)^{\frac{1}{a+b}} + s_n\left(\frac{bf_n}{a}\right)^{\frac{\epsilon+1}{a+b}}.$$

We can now insert this formula into $1/g(x)$ which will produce a lower bound for the norm of X_n^{-1}:

$$\frac{1}{(\frac{bf_n}{a})^{\frac{a}{a+b}} + f_n(\frac{bf_n}{a})^{\frac{-b}{a+b}}} \geq \|X_n^{-1}\| \geq \frac{1}{\left[\left(\frac{bf_n}{a}\right)^{\frac{1}{a+b}} + s_n\left(\frac{bf_n}{a}\right)^{\frac{\epsilon+1}{a+b}}\right]^a + f_n\left[\left(\frac{bf_n}{a}\right)^{\frac{1}{a+b}} + s_n\left(\frac{bf_n}{a}\right)^{\frac{\epsilon+1}{a+b}}\right]^{-b}}.$$

Also,

$$\begin{aligned}
1 \geq & \left[\left(\frac{bf_n}{a} \right)^{\frac{a}{a+b}} + f_n \left(\frac{bf_n}{a} \right)^{\frac{-b}{a+b}} \right] \|X_n^{-1}\| \\
\geq & \frac{(\frac{bf_n}{a})^{\frac{a}{a+b}} + f_n(\frac{bf_n}{a})^{\frac{-b}{a+b}}}{\left[\left(\frac{bf_n}{a}\right)^{\frac{1}{a+b}} + s_n \left(\frac{bf_n}{a}\right)^{\frac{\epsilon+1}{a+b}} \right]^a + f_n \left[\left(\frac{bf_n}{a}\right)^{\frac{1}{a+b}} + s_n \left(\frac{bf_n}{a}\right)^{\frac{\epsilon+1}{a+b}} \right]^{-b}} \\
= & \frac{\left[(\frac{bf_n}{a})^{\frac{a}{a+b}} + f_n(\frac{bf_n}{a})^{\frac{-b}{a+b}} \right] \left[\left(\frac{bf_n}{a}\right)^{\frac{1}{a+b}} + s_n \left(\frac{bf_n}{a}\right)^{\frac{\epsilon+1}{a+b}} \right]^b}{\left[\left(\frac{bf_n}{a}\right)^{\frac{1}{a+b}} + s_n \left(\frac{bf_n}{a}\right)^{\frac{\epsilon+1}{a+b}} \right]^{a+b} + f_n} \\
= & \frac{\left[(\frac{bf_n}{a})^{\frac{a}{a+b}} + f_n(\frac{bf_n}{a})^{\frac{-b}{a+b}} \right] \left[(\frac{bf_n}{a})^{\frac{b}{a+b}} + \cdots + s_n^b (\frac{bf_n}{a})^{\frac{b(\epsilon+1)}{a+b}} \right]}{\left[(\frac{bf_n}{a}) + \cdots + s_n^{a+b} (\frac{bf_n}{a})^{\epsilon+1} \right] + f_n} \\
= & \frac{\frac{bf_n}{a} + f_n + F_1(f_n)}{\frac{bf_n}{a} + f_n + F_2(f_n)} \\
= & \frac{\frac{b}{a} + 1 + \frac{F_1(f_n)}{f_n}}{\frac{b}{a} + 1 + \frac{F_2(f_n)}{f_n}},
\end{aligned}$$

where $F_1(f_n)$ and $F_2(f_n)$ are both functions of powers of f_n higher than one. If we then take the limit left and right of this inequality as n approaches infinity,

$$1 \geq \lim_{n\to\infty} \left[\left(\frac{bf_n}{a} \right)^{\frac{a}{a+b}} + f_n \left(\frac{bf_n}{a} \right)^{\frac{-b}{a+b}} \right] \|X_n^{-1}\| \geq 1,$$

implying that

$$\lim_{n\to\infty} \left[\left(\frac{bf_n}{a} \right)^{\frac{a}{a+b}} + f_n \left(\frac{bf_n}{a} \right)^{\frac{-b}{a+b}} \right] \|X_n^{-1}\| = 1.$$

□

We now state the positive definite version of Theorem 3.4.1 for the case $b = a$.

Theorem 3.4.8. *Assuming the same conditions as in Theorem 3.4.6, except that $\epsilon > a$, we have*

$$\frac{1}{2\sqrt{f_n}} - \|X_n^{-1}\| = O\left(f_n^{\frac{\epsilon-a}{2a}} \right), \tag{3.7}$$

where $X_n = T_n^a + f_n T_n^{-a}$, with $a > 0$ and $\epsilon > a$ implies $f_n > O\left(\frac{1}{n^{\frac{2a}{a+1}}}\right)$.

Note that in this case we cannot have different positive and negative powers.

As for the analysis preceding the proof of Theorem 3.4.1, we will again define a R_n and $\tilde{R}_n$. From the proof of the previous theorem (assuming $a = b$), we have

$$\frac{1}{f_n^{\frac{1}{2}} + f_n \cdot f_n^{\frac{-1}{2}}} \geq \|X_n^{-1}\| \geq \frac{1}{\left(f_n^{\frac{1}{2a}} + s_n f_n^{\frac{\epsilon+1}{2a}} \right)^a + f_n \left(f_n^{\frac{1}{2a}} + s_n f_n^{\frac{\epsilon+1}{2a}} \right)^{-a}} := R_n,$$

and we also define

$$\tilde{R}_n := \frac{1}{\left(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}}\right)^a + f_n\left(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}}\right)^{-a}} = \frac{1}{g_n\left(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}}\right)}.$$

To prove the current theorem, we will need a new version of Lemma 3.4.2, as our function $1/g_n(x)$ is different from the standard case, as well as the interval E_n which plays a role in determining which values we need to evaluate $1/g_n(x)$ for.

Recall from Chapter 2, that to prove Lemma 3.4.2, we had to show that

$$\frac{1}{g_n(\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}})} \geq \frac{1}{g_n(\alpha_n)},$$

where

$$\alpha_n = \sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}.$$

This time, all arguments of Lemma 3.4.2 remain valid, except that now, $\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}}$ and α_n is replaced by

$$f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}} \qquad \text{and} \qquad \alpha_n := f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}}$$

respectively. Using this notation, $\tilde{R}_n = 1/g_n(\alpha_n)$, and hence we calculate,

$$\begin{aligned}
&\frac{1}{g_n(f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})} - \frac{1}{g_n(\alpha_n)} \\
&= \frac{1}{(f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^a + f_n(f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^{-a}} - \frac{1}{(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^a + f_n(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^{-a}} \\
&= \frac{(f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^a}{(f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^{2a} + f_n} - \frac{(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^a}{(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^{2a} + f_n} \\
&= \frac{[(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^{2a} + f_n](f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^a - [(f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^{2a} + f_n](f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^a}{[(f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^{2a} + f_n][(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^{2a} + f_n]}.
\end{aligned}$$

Notice that the denominator of the last expression is positive and we only have to check that the numerator is positive. Thus,

$$\begin{aligned}
&[(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^{2a} + f_n](f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^a - [(f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^{2a} + f_n](f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^a \\
&= (f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^a(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^{2a} + f_n(f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^a \\
&- (f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^{2a}(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^a - f_n(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^a \\
&= (f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^a(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^a[(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^a - (f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^a] \\
&- f_n[(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^a - (f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^a].
\end{aligned}$$

The term in the square brackets are clearly positive, and therefore we only need to check that the product in the round brackets are greater than f_n:

$$\begin{aligned}
&(f_n^{\frac{1}{2a}} - f_n^{\frac{\epsilon+1}{2a}})^a(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^a \\
&= (f_n^{\frac{1}{a}} + kf_n^{\frac{\epsilon+2}{2a}} - f_n^{\frac{\epsilon+2}{2a}} - kf_n^{\frac{2\epsilon+2}{2a}})^a.
\end{aligned}$$

It is clear that

$$kf_n^{\frac{\epsilon+2}{2a}} - f_n^{\frac{\epsilon+2}{2a}} - kf_n^{\frac{2\epsilon+2}{2a}} > 0$$

when k and n is large enough, and this implies that

$$(f_n^{\frac{1}{a}} + kf_n^{\frac{\epsilon+2}{2a}} - f_n^{\frac{\epsilon+2}{2a}} - kf_n^{\frac{2\epsilon+2}{2a}})^a > f_n.$$

We have thus shown that $R_n \geq \tilde{R}_n$, and consequently,

$$\frac{1}{f_n^{\frac{1}{2}} + f_n \cdot f_n^{\frac{-1}{2}}} \geq \|X_n^{-1}\| \geq \tilde{R}_n,$$

which if rearranged, leads to the inequality

$$\frac{1}{2\sqrt{f_n}} - \frac{1}{\left(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}}\right)^a + f_n\left(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}}\right)^{-a}} \geq \frac{1}{2\sqrt{f_n}} - \|X_n^{-1}\| \geq 0.$$

We compute

$$\begin{aligned}
&\frac{1}{2\sqrt{f_n}} - \frac{1}{\left(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}}\right)^a + f_n\left(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}}\right)^{-a}} \\
&= \frac{1}{2\sqrt{f_n}} - \frac{\left(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}}\right)^a}{\left(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}}\right)^{2a} + f_n} \\
&= \frac{\left(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{a}}\right)^{2a} + f_n - 2\sqrt{f_n}\left(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}}\right)^a}{2\sqrt{f_n}[(f_n^{\frac{1}{2a}} + kf_n^{\frac{\epsilon+1}{2a}})^{2a} + f_n]} \\
&= \frac{\left(f_n + kf_n^{\frac{2a+\epsilon}{2a}} + \cdots + k^{2a}f_n^{2\epsilon+2}\right) + f_n - 2\sqrt{f_n}\left(\sqrt{f_n} + kf_n^{\frac{a+\epsilon}{2a}} + \cdots + k^a f_n^{\frac{\epsilon+1}{2}}\right)}{2\sqrt{f_n}(f_n + \cdots + k^{2a}f_n^{\epsilon+1}) + 2\sqrt{f_n}f_n} \\
&= \frac{\left(kf_n^{\frac{2a+\epsilon}{2a}} + \cdots + k^{2a}f_n^{2\epsilon+2}\right) - 2\left(kf_n^{\frac{2a+\epsilon}{2a}} + \cdots + k^a f_n^{\frac{\epsilon+2}{2}}\right)}{2\sqrt{f_n}(f_n + \cdots + k^{2a}f_n^{\epsilon+1}) + 2\sqrt{f_n}f_n}.
\end{aligned}$$

We want this expression to converge to zero as n tends to infinity. For this to happen we need the manimal power of f_n in the numerator to be strictly greater than the minimal power of f_n in the denominator, i.e.,

$$\frac{2a+\epsilon}{2a} > \frac{3}{2},$$

leading to $\epsilon > a$. If we denote $\epsilon = a + \delta$, the expression above would be

$$O\left(f_n^{\frac{\delta}{2a}}\right) = O\left(f_n^{\frac{\epsilon-a}{2a}}\right).$$

The order of convergence here can be determined by finding the maximum value of ϵ which satisfies the conditions of this theorem with respect to the sequence f_n that is considered.

Example 3.4.9.

Choose X_n such that T_n is the tridiagonal matrix with associated symbol $B(t) = 2 - t - t^{-1}$ and constants $a = 2$, $b = 3$. Then Figures 3.5, 3.6 and 3.7 corresponds to the sequences $f_n = 1/n, 1/n^2, 1/n^3$, respectively, and shows the growth of $\|X_n^{-1}\|$.

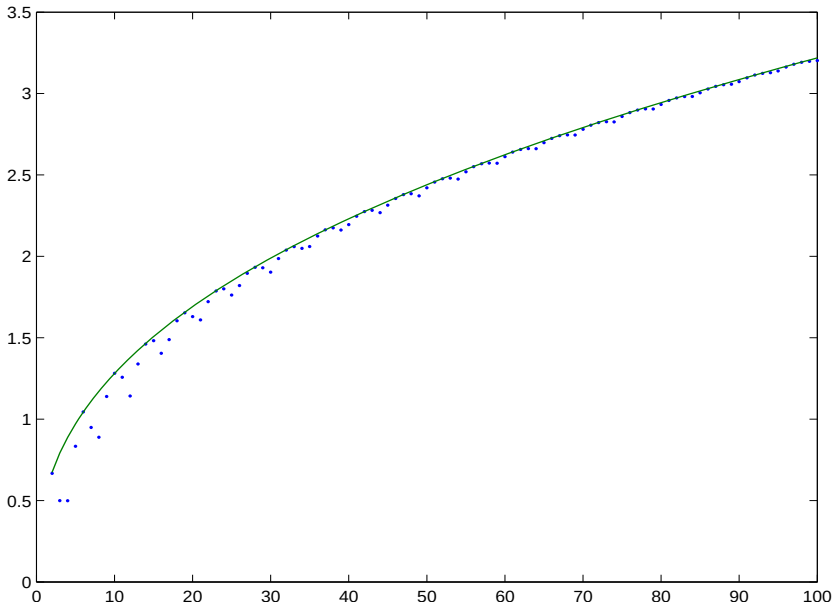

Figure 3.5: The norm of X_n^{-1} (dots) and upper bound (line), plotted as functions of n - $f_n = 1/n$.

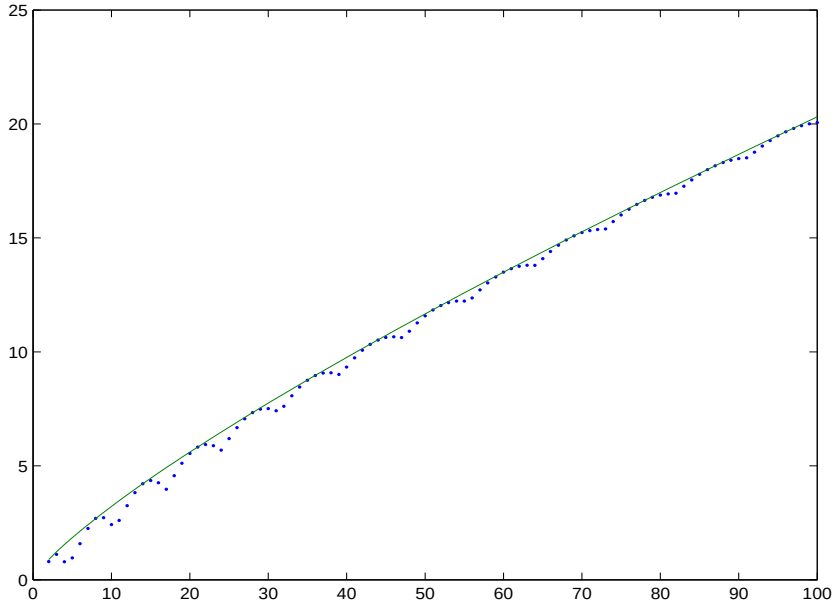

Figure 3.6: The norm of X_n^{-1} (dots) and upper bound (line), plotted as functions of n - $f_n = 1/n^2$.

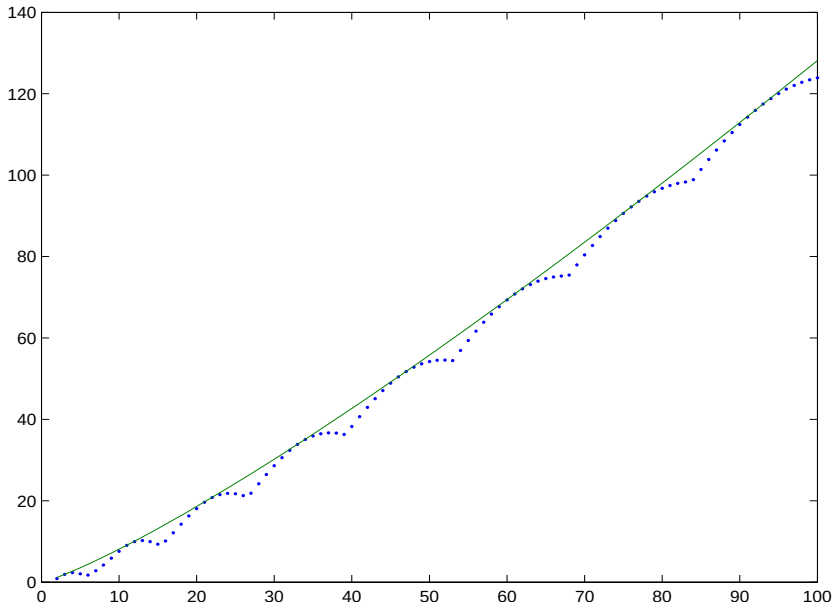

Figure 3.7: The norm of X_n^{-1} (dots) and upper bound (line), plotted as functions of n - $f_n = 1/n^3$.

3.5 Extension to Fredholm case

The assumptions on the symbol b of the Toeplitz operator T (Theorem 3.1.1), implies that T is not Fredholm (see e.g., [5]). We will show, via an example, that the possibility to carry over our results in a meaningful way to the case where T is Fredholm, but not invertible, is very limited. Indeed, the class of sequences $\{f_n\}$ to which Theorem 3.1.1 applies is a fairly wide class. If we want to extend the result to Toeplitz operators which are Fredholm, then this will be possible only for sequences $\{f_n\}$ which have a very specific decay rate. The reason for this is that we need $\sqrt{f_n}$ to be close to a singular value of T_n (for large n). Because of the splitting phenomenon (Theorem 1.2.1), when T is Fredholm but not invertible, there are only a finite number of singular values of T_n that go to zero (precisely as many as the absolute value of the Fredholm index), and they do so with a specific decay rate.

Example 3.5.1.

Consider the Toeplitz operator T with symbol $b(t) = 1 + \frac{\gamma}{t}$, where γ is a complex number with $|\gamma| > 1$. It is easy to see that $b(t)$ does not vanish on the unit circle and has a winding number of -1. It follows that T is Fredholm, but not invertible (see [16]).

Let us denote by J_n the $n \times n$ upper triangular Toeplitz matrix with a one on the second diagonal, and zeros elsewhere. Then the $n \times n$ finite section of T, T_n, is given by $T_n = I_n + \gamma J_n$. As $J_n^n = 0$, it follows that

$$T_n^{-1} = (I_n + \gamma J_n)^{-1} = \sum_{k=0}^{n-1} (-1)^k \gamma^k J_n^k,$$

and since $\|J_n\| = 1$, we conclude that

$$\|T_n^{-1}\| \leq \sum_{k=0}^{n-1} |\gamma|^k = \frac{|\gamma|^n - 1}{|\gamma| - 1} \leq \frac{|\gamma|^n}{|\gamma| - 1}.$$

Next, we observe that the last column of T_n^{-1} is equal to the vector

$$y := \left[(-1)^j \gamma^j\right]_{j=0}^{n-1}.$$

Let x be the vector consisting of zeros except for 1 as its last entry. Since the spectral norm is induced by the Euclidean vector norm and $\|T_n x\|_2 = \|y\|_2$, we have

$$\|T_n^{-1}\| \geq \| \left[(-1)^j \gamma^j\right]_{j=0}^{n-1} \|_2 = \sqrt{1 + |\gamma|^2 + \cdots + |\gamma|^{2n-2}} = \sqrt{\frac{|\gamma|^{2n} - 1}{|\gamma|^2 - 1}}.$$

Since $|\gamma| > 1$, for every $\varepsilon < 1$ there is a number N (depending on γ) such that for $n \geq N$

$$\sqrt{\frac{|\gamma|^{2n} - 1}{|\gamma|^2 - 1}} \geq \varepsilon \frac{|\gamma|^n}{\sqrt{|\gamma|^2 - 1}}.$$

It follows that, for $n \geq N$

$$\varepsilon \frac{|\gamma|^n}{\sqrt{|\gamma|^2 - 1}} \leq \|T_n^{-1}\| \leq \frac{|\gamma|^n}{|\gamma| - 1}.$$

Since the smallest singular value of T_n is one over the norm of its inverse, we obtain

$$(|\gamma| - 1)\frac{1}{|\gamma|^n} \leq \sigma_n(T_n) \leq \frac{1}{\varepsilon}\sqrt{|\lambda|^2 - 1}\frac{1}{|\lambda|^n}.$$

The splitting phenomenon applies to our T_n and shows that only one singular value goes to zero as n approaches infinity. Now consider $X_n = T_n + f_n(T_n^{-1})^*$. Recall that $\|X_n^{-1}\| \leq \frac{2}{\sqrt{f_n}}$ (Proposition 3.2.2), and that for this estimate to be an asymptotic equivalence, one needs a singular value of T_n near $\sqrt{f_n}$ for large enough n. In order for an analogue of Theorem 3.1.1 to hold in the particular case we are now considering, we need $f_n \to 0$, but also that $\sqrt{f_n}$ should be close to $\sigma_n(T_n)$. This forces $\sqrt{f_n}$ to go to zero at the same rate as $\sigma_n(T_n)$, that is, it forces f_n to be $O(\frac{1}{|\gamma|^{2n}})$. This is a far more restrictive condition on the sequence (f_n) than what is required in Theorem 3.1.1.

To be precise, let us assume that f_n is not of the same order as $\sigma_n(T_n)$. That is, either $\sigma_n(T_n) \leq \alpha_n\sqrt{f_n}$, where α_n is any sequence of positive numbers converging to zero, or $\sigma_n(T_n) \geq \alpha_n\sqrt{f_n}$, where α_n is any sequence of positive numbers converging to infinity.

To know the behaviour of $\|X_n^{-1}\|$, we must know which of the singular values of T_n will produce $\|X_n^{-1}\|$ through the function $1/g_n(x)$. First let us consider all the singular values of T_n except $\sigma_n(T_n)$. We know from the splitting phenomenon, that they are all bigger than some positive constant and smaller than $\|b\|_\infty := \max_{t\in\mathbb{T}} |b(t)|$. Now, the corresponding singular values of X_n^{-1} which we find by evaluating $1/g_n(x)$ in these singular values of T_n are all bounded for all n. This can be seen by calculating

$$\lim_{n\to\infty} \frac{1}{g_n(x)} = \lim_{n\to\infty} \frac{x}{x^2 + f_n} = \frac{1}{x}.$$

Since these singular values of T_n are bounded for all n, $1/x$ will also be bounded. This implies that $\|X_n^{-1}\|$ is bounded if it happens to be produced by a singular value of T_n other than $\sigma_n(T_n)$. In this case it is clear then that

$$\lim_{n\to\infty} 2\sqrt{f_n}\|X_n^{-1}\| = 0,$$

Let us see then what happens if $\sigma_n(T_n)$ determines $\|X_n^{-1}\|$. Recall that $1/g_n(x)$ is monotonically decreasing and increasing for $x \geq \sqrt{f_n}$ and $x \leq \sqrt{f_n}$ respectively. Therefore, assuming that $\sigma_n(T_n) \leq \alpha_n\sqrt{f_n}$ (α_n converges to zero), or $\sigma_n(T_n) \geq \alpha_n\sqrt{f_n}$ (α_n converges to infinity)

$$2\sqrt{f_n}\|X_n^{-1}\| = \frac{2\sqrt{f_n}}{g_n(\sigma_n(T_n))} \leq \frac{2\sqrt{f_n}}{g_n(\alpha_n\sqrt{f_n})} = \frac{2\alpha_n f_n}{\alpha_n^2 f_n + f_n} = \frac{2}{\alpha_n + \frac{1}{\alpha_n}}$$

and

$$\lim_{n\to\infty} 2\sqrt{f_n}\|X_n^{-1}\| \leq \lim_{n\to\infty} \frac{2}{\alpha_n + \frac{1}{\alpha_n}} = 0.$$

This shows that in this case the conclusion of Theorem 3.1.1 does not hold, and that the asymptotic behaviour of $\|X_n^{-1}\|$ and its upper bound is not the same.

This leaves us with the case when f_n has the same decay rate as $\sigma_n(T_n)^2$. Assume that for the sequence f_n there are positive constants β and η such that the inequalities

$$\frac{\beta}{|\gamma|^2} < \sqrt{f_n} < \frac{\eta}{|\gamma|^n}, \quad \frac{\beta}{|\gamma|^2} < \sigma_n(T_n) < \frac{\eta}{|\gamma|^n}, \tag{3.8}$$

are satisfied simultaneously. Then we compute

$$\frac{1}{g_n(\frac{\beta}{|\gamma|^n})} = \frac{\beta}{\frac{\beta^2}{|\gamma|^n} + f_n|\gamma|^n}.$$

Since

$$\frac{2\beta^2}{|\gamma|^n} < \frac{\beta^2}{|\gamma|^n + f_n\gamma|^n} < \frac{\eta^2 + \beta^2}{|\gamma|^n},$$

we have that

$$|\gamma|^n\frac{\beta}{\eta^2 + \beta^2} < g_n(\frac{\beta}{|\gamma|^n}) < |\gamma|^n\frac{1}{2\beta}.$$

Similarly, we compute that

$$\frac{1}{g_n(\frac{\eta}{|\gamma|^n})} = \frac{\eta}{\frac{\eta^2}{|\gamma|^n} + f_n|\gamma|^n}.$$

Since

$$\frac{\beta^2 + \eta^2}{|\gamma|^n} < \frac{\eta^2}{|\gamma|^n} + f_n|\gamma|^n < \frac{2\eta^2}{|\gamma|^n},$$

we have that

$$|\gamma|^n\frac{1}{2\eta} < \frac{1}{g_n(\frac{\eta}{|\gamma|^n})} < |\gamma|^n\frac{\eta}{\eta^2 + \beta^2}.$$

By our initial assumption, $\frac{\beta}{|\gamma|^n} < \sigma_n(T_n) < \frac{\eta}{|\gamma|^n}$, and the fact that $\sqrt{f_n}$ (where $1/g_n(x)$ takes its maximum) is bounded by the same two values, it follows that

$$\frac{1}{g_n(\sigma_n(T_n))} > |\gamma|^n \min\left(\frac{1}{2\eta}, \frac{\beta}{\eta^2+\beta^2}\right) = |\gamma|^n \frac{\beta}{\eta^2+\beta^2}.$$

We explain the last equality as follows. We want $1/(\eta^2/\beta + \beta) < 1/2\eta$. Consider

$$\frac{\eta^2}{\beta} + \beta - 2\eta = \frac{1}{\beta}\left(\eta^2 + \beta^2 - 2\eta\beta\right) = \frac{1}{\beta}(\eta-\beta)^2.$$

Since $\eta > \beta$ by assumption, the expression in the equality above is positive.

Then

$$\|X_n^{-1}\| = \frac{1}{g_n(\sigma_n(T_n))} > |\gamma|^n \frac{\beta}{\eta^2+\beta^2}.$$

Since $\sqrt{f_n} > \frac{\sqrt{\beta}}{|\gamma|^n}$, we conclude that

$$2\sqrt{f_n}\|X_n^{-1}\| > \frac{2\beta^2}{\eta^2+\beta^2}.$$

In addition, we already know from Theorem 3.1.1 that $2\sqrt{f_n}\|X_n^{-1}\| < 1$. Thus, if (3.8) holds, then at least the sequence $\{2\sqrt{f_n}\|X_n\|^{-1}\}$ is bounded from below by a positive constant. Note that the condition does not seem to imply that the limit of this sequence necessarily exists, neither that the limit, if it exists, must be one.

However, if we replace the positive constants β and η by two positive sequences, β_n and η_n, with the property that $\lim \frac{\eta_n}{\beta_n} = 1$, we will have the same result as in Theorem 3.1.1. Clearly this is an extremely forced condition, and shows how sensitive the Fredholm case is to the choice of f_n.

It is also relevant to consider the perturbation of X_n (Z_n) in the Fredholm case, and to see if a similar analysis as for the example above can be performed to obtain an analogue of Theorem 3.1.2. In general, it seems that this generalization is limited by an extra condition and in our particular example we cannot perturb X_n at all. The reason for this is related to the interlacing of singular values. We know from Theorem 3.1.2 that Z_n differs from X_n by a finite number of rows and/or columns, say p. We can then form the submatrix U_{n-p} obtained from X_n by deleting this finite number of rows and columns. The basic idea in the proof of Theorem 3.1.2 is to use the interlacing of the singular values of Z_n and U_{n-p} to find a lower bound for Z_n^{-1}. In fact, we need $2p+14r+1$ singular values in the interval $E_n = [\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}}, \sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}]$ to achieve the result. Now in our case, there is only one singular value tending to zero, and therefore we cannot find $2p+14r+1$ singular values close to $\sqrt{f_n}$, and therefore cannot apply the interlacing technique. This does not mean that Z_n^{-1} does not have the desired behaviour; it just means that one can not apply the techniques used previously to analyze $\|Z_n^{-1}\|$.

The general Fredholm case however, allows for a finite number of singular values (equal to the absolute value of its index) to approach zero. Therefore, if the Fredholm index equals i, any perturbation of X_n up to size $i-1$ is permitted and, potentially, the interlacing of singular values can be combined with similar techniques as shown here in our example.

3.6 The norms of X_n and Z_n

The main results of this chapter deal with the asymptotic behaviour of the norms of the inverses of X_n and Z_n. It is natural then to ask what happens to the norm of the matrices themselves? We shall shed some light on this question and show that their behaviour is not as predictable as that of their inverses.

We consider three scenarios regarding the norms of X_n and Z_n: Convergence to zero, general boundedness, or unboundedness. First we consider X_n, and show that convergence to zero is impossible.

Let $T_n = U_n D_n V_n^*$ be a SVD. As before, we calculate $X_n = U_n(D_n + f_n D_n^{-1})V_n^*$. The entries of D_n are the singular values of T_n, and hence its maximum value never approaches zero ([5], Chapter 5). Also, the entries of $D_n + f_n D_n^{-1}$ contain the singular values of X_n, and since $\|X_n\|$ is determined by the largest singular value it cannot tend to zero since the values of D_n does not converge to zero.

To determine to largest singular value of X_n, we need to evaluate the function $g_n(x) = x + f_n \frac{1}{x}$ for positive values of x. Its derivative is $1 - f_n \frac{1}{x^2}$ and is monotone decreasing on the interval $(0, \sqrt{f_n}]$, and monotone increasing on $[\sqrt{f_n}, \infty)$. It follows that $\|X_n\|$ attains its maximum value at either the minimum, or maximum singular value of T_n. We know that $\|T_n(b)\|$ is a bounded sequence of positive numbers since $\|T_n(b)\|_2 = \|T(b)\|_2 + O(1/n)$ ([5], Chapter 5).

Let σ_n, σ_1 and σ_l be the smallest, largest and l'th singular values of T_n respectively. We omit the dependance on T_n here from the singular values since they will always be the singular values of T_n for the rest of this chapter. Now,

$$\lim_{n\to\infty} g_n(\sigma_1) = \lim_{n\to\infty} (\sigma_1 + f_n \frac{1}{\sigma_1}) = \lim_{n\to\infty} \sigma_1 = \|T(b)\|.$$

Notice that if $g_n(\sigma_n) \leq g_n(\sigma_1)$ for all n large enough, $\|X_n\| = g_n(\sigma_1)$, and

$$\lim_{n\to\infty} \|X_n\| = \|T(b)\|,$$

forcing $\|X_n\|$ to be bounded.

On the other hand, if $g_n(\sigma_1) \leq g_n(\sigma_n)$ for all large n, the situation changes since $\lim_{n\to\infty} \sigma_n = 0$, and this could possibly lead to the case where $\lim_{n\to\infty} g_n(\sigma_n) = \infty$. To investigate this, we consider

$$\lim_{n\to\infty} g_n(\sigma_n) = \lim_{n\to\infty} (\sigma_n + f_n \frac{1}{\sigma_n}) = \lim_{n\to\infty} f_n \frac{1}{\sigma_n}.$$

Recall the assumption $n\sqrt{f_n^{\epsilon+1}} > c$ on the sequence f_n from Theorem 3.1.1. This implies that f_n must converge to zero slower than $O(1/n^2)$. Thus, if σ_n is $O(1/n^2)$, or just converge to zero faster than f_n, $g_n(\sigma_n)$ $(= \|X_n\|)$ will be an unbounded sequence approaching infinity.

To obtain the speed of convergence of σ_n for our finite sections, $T_n(b)$, we need to divide our class of symbols ($T(b)$ non-Fredholm) into two subclasses, i.e., non-constant and constant Laurent polynomials.

In the non-constant case, we state the following theorem from section 1.2.1 again.

Theorem 3.6.1. *Let b be a non-constant Laurent polynomial and suppose $T(b)$ is not Fredholm. Let $\alpha \in \mathbb{N}$ be the maximal order of the zeros of $|b|$ on $\mathbb{T}$. Then for each natural number $k \geq 1, \sigma_{n-k} = O(1/n^{\alpha})$ as $n \to \infty$.*

Here the order of the zero, say α_0, indicate the smallest natural number such that

$$\frac{d}{dt}^{\alpha_0} b(t_0) \neq 0,$$

where $b(t_0) = 0$.

A constant Laurent polynomial is such that it has constant modulus, $|b(t)| = b_0, \forall t \in \mathbb{T}$, where $b_0 \in \mathbb{C}$. It can be shown to be of the form $b(t) = \gamma t^m$, with constants $\gamma, m \in \mathbb{Z}$. See [5], Section 5.2 for details. Recall that in the non-Fredholm case, the symbol must vanish on the unit circle in at least a single point. We can therefore ignore the constant Laurent polynomials since they cannot vanish on the unit circle, and hence, are Fredholm. The theorem above is then generally applicable in our case.

This theorem in conjunction with the assumption on f_n shows that α has to equal one for $\|X_n\|$ to have the possibility of being bounded:

$$\lim_{n\to\infty} g_n(\sigma_n) = \lim_{n\to\infty} f_n \frac{1}{\sigma_n}.$$

If $\|X_n\|$ is not bounded, the following order of growth applies:

$$\|X_n\| = O(f_n n^{\alpha}).$$

In the bounded case, assume $O(1/n^2) < f_n \leq O(1/n)$ and $d/n \leq \sigma_n \leq e/n$, where d and e are constants. Then,

$$\begin{aligned} \|X_n\| = g_n(\sigma_n) &= \sigma_n + f_n \frac{1}{\sigma_n} \\ &\leq \frac{e}{n} + f_n \frac{n}{d}. \end{aligned}$$

Now, since $\|X_n\|$ cannot converge to zero, the inequality above is only valid if $f_n = O(1/n)$. What this means is that if $\alpha = 1$ and $f_n < O(1/n)$, $\|X_n\|$ will be determined by σ_1, not σ_n.

We can also state the order of convergence for the first case where $\|X_n\| = g_n(\sigma_1)$. The following result, also stated in section 1.2.1, gives the convergence rate for all the upper singular values of $T_n(b)$.

Theorem 3.6.2. *Let b be a non-constant Laurent polynomial. Denote by $\beta \in \mathbb{N}$ the maximal order of the zeros of $\|b\|_\infty - |b|$ on the unit circle. Then for each $k \geq 0$,*

$$\|b\|_\infty - D_k \frac{1}{n^{\beta}} \leq \sigma_k \leq \|b\|_\infty$$

with some constant $D_k \in (0, \infty)$ independant of n.

Applying this theorem, we have

$$\begin{aligned}\|X_n\| &= g_n(\sigma_1)\\ &= \sigma_1 + f_n\frac{1}{\sigma_1}\\ &= \|T(b)\| + O(1/n^\beta) + O(f_n).\end{aligned}$$

Example 3.6.3.

We return to our regular example $X_n = T_n(\frac{1}{n})$. Recall that $T_n = T_{0,n}$ and $f_n = 1/n$. We will show that $g_n(\sigma_n) \le g_n(\sigma_1)$ for n large enough, and consequently, $\lim_{n\to\infty}\|X_n\| = \|T(b)\| = \|b\|_\infty = 2$, where $b(t) = 1 - t^{-1}$.

In the next chapter, we prove that the singular values of $T_{0,n}$ are given by $\sqrt{2-2\cos(\theta_m)}$, where $\theta_m = \frac{2m+1}{2n+1}\pi$ and $0 \le m \le n-1$ (Theorem 4.1.1). We can rewrite these singular values as $2\sin(\theta_m/2)$ using a trigonometric identity. Inserting the θ_m's into $2\sin(\theta_m/2)$, it is clear that $\sigma_n = 2\sin(\frac{1}{4n+2}\pi)$ and $\sigma_1 = 2\sin(\frac{2n-1}{4n+2}\pi)$. Thus,

$$\begin{aligned}&g_n(\sigma_1) - g_n(\sigma_n)\\ =&2\Big[\sin(\frac{2n-1}{4n+2}\pi) - \sin(\frac{1}{4n+2}\pi)\Big] + \frac{1}{2n}\Big[\frac{1}{\sin(\frac{2n-1}{4n+2}\pi)} - \frac{1}{\sin(\frac{1}{4n+2}\pi)}\Big].\end{aligned}$$

Taking limits,

$$\begin{aligned}&\lim_{n\to\infty} g_n(\sigma_1) - g_n(\sigma_n)\\ =&\lim_{n\to\infty} 2\Big[\sin(\frac{2n-1}{4n+2}\pi) - \sin(\frac{1}{4n+2}\pi)\Big] + \lim_{n\to\infty}\frac{1}{2n}\Big[\frac{1}{\sin(\frac{2n-1}{4n+2}\pi)} - \frac{1}{\sin(\frac{1}{4n+2}\pi)}\Big]\\ =&\,2 - \frac{1}{2}\lim_{n\to\infty}\frac{1}{n}\frac{1}{\sin(\frac{1}{4n+2}\pi)}\\ =&\,2 - \frac{1}{2}\lim_{n\to\infty}\frac{\frac{1}{n}}{\sin(\frac{1}{4n+2}\pi)}\\ =&\,2 + \frac{1}{2}\lim_{n\to\infty}\frac{\frac{1}{n^2}}{\cos(\frac{1}{4n+2}\pi)\frac{d}{dn}\frac{\pi}{4n+2}}\\ =&\,2 - \frac{1}{2}\lim_{n\to\infty}\frac{\frac{1}{n^2}}{\cos(\frac{1}{4n+2}\pi)}\frac{(4n+2)^2}{4\pi}\\ =&\,2 - \frac{1}{2}\lim_{n\to\infty}\frac{16 + \frac{16}{n} + \frac{4}{n^2}}{\cos(\frac{1}{4n+2}\pi)4\pi}\\ =&\,2 - \frac{2}{\pi} > 0.\end{aligned}$$

Figure 3.8 shows the behaviour of $\|T_n(\frac{1}{n})\|$.

Example 3.6.4.

By making a small adjustment to the choice of f_n in the previous example, $\|X_n\|$ will be determined by σ_n instead of σ_1. For this purpose, let $f_n = 4/n$, and then $X_n = T_n(\frac{4}{n})$. If we follow the limit calculation exactly as in the previous example, we see that

$$\lim_{n\to\infty} g_n(\sigma_1) - g_n(\sigma_n) = 2 - \frac{8}{\pi} < 0.$$

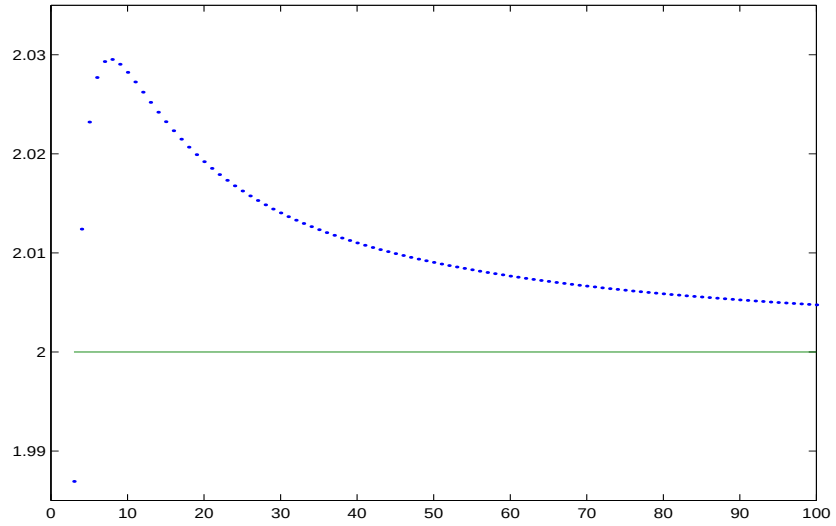

Figure 3.8: The norm of $T_n(\frac{1}{n})$ (dots) converging to 2 (line), plotted as functions of n.

Now we have to determine if $\|X_n\|$ is a bounded or unbounded sequence. Since $f_n = O(1/n)$, we need to check the convergence rate of σ_n. Applying Theorem 1.2.2, we see that $\alpha = 1$ since $b(1) = 0$ and $b'(1) \neq 0$. The symbol b vanishes only at the point 1 on the unit circle. It follows that $\|X_n\|$ is a bounded sequence. Figure 3.9 shows the behaviour of $\|T_n(\frac{4}{n})\|$.

We have covered the two possibilities where $\|X_n\|$ is a bounded sequence. The next example illustrates the unbounded case.

Example 3.6.5.

Again, we only need to choose a different f_n to cause the sequence $\|X_n\|$ to be unbounded. We already know that $\alpha = 1$ in Theorem 1.2.2, so we just need to choose f_n such that it converges at a slower rate to zero. Let $f_n = \frac{1}{n^{1/4}}$. Figure 3.10 shows the unbounded growth of $\|X_n\|$.

We now consider the sequence $\|Z_n\|$, and we will cover the same three possibilities, i.e., convergence to zero, boundedness and unboundedness. The choice of f_n and the particular perturbation of X_n will affect this behaviour.

We state the definition of Z_n again for convenience:

$$Z_n := V_n T_n + f_n (T_n^{-1})^* + W_n,$$

where $W_n = (T_n^{-1})^* \sum_{j=1}^{n} w_j e_j e_j^*$, $V_n = \sum_{j=1}^{n} v_j e_j e_j^*$, and where v_j and w_j are positive real numbers. In addition, we assumed (see Theorem 3.1.2) that

- $v_j \geq 1$ for $1 \leq j \leq q$ where q is fixed.
- $v_j = 1$ for $n \geq j > q$.

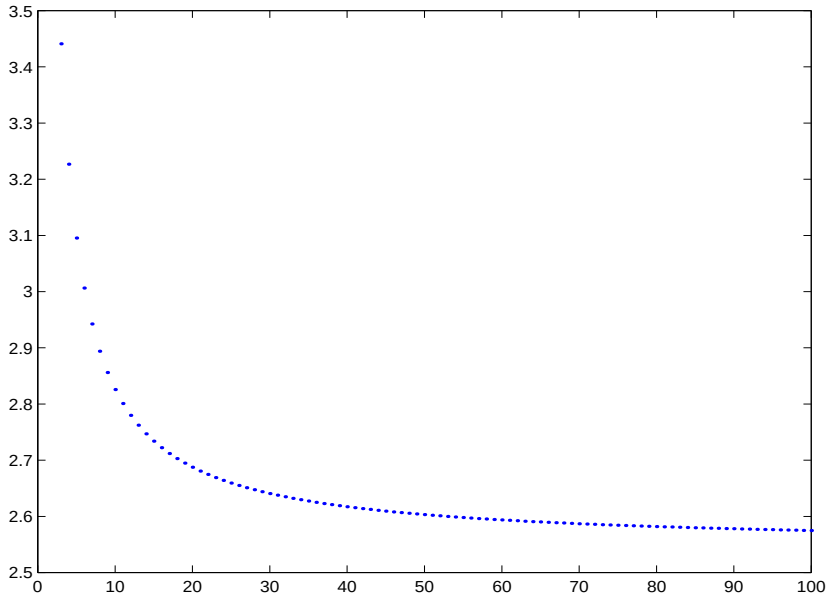

Figure 3.9: The norm of $T_n(\frac{4}{n})$ (dots), plotted as a function of n.

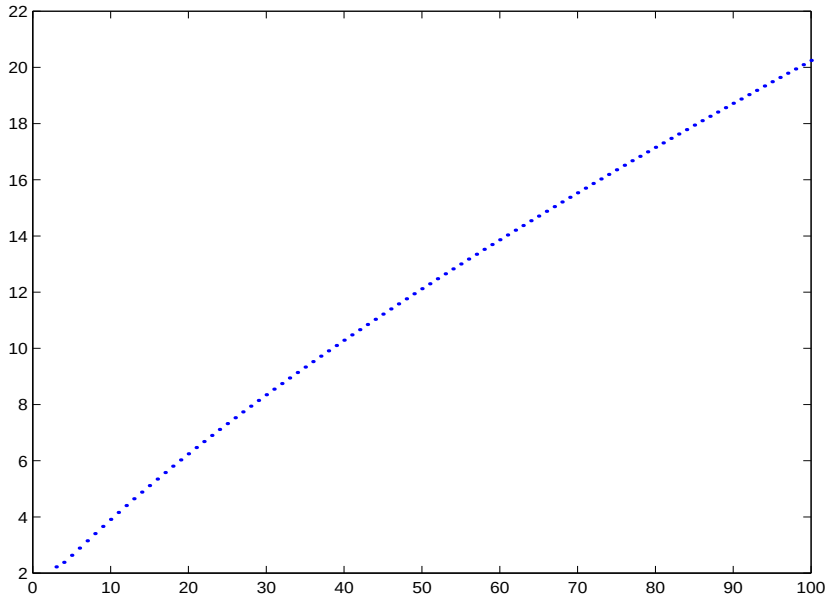

Figure 3.10: The norm of $T_n(\frac{1}{n^{1/4}})$ (dots).

- $w_j \geq 0$ for $1 \leq j \leq r$ where r is fixed.
- $w_j = 0$ for $n \geq j > r$.

From the proof of Proposition 3.2.2, we know that

$$Z_n = V_n^{1/2}(A_n + (A_n^{-1})^*)E_n^{1/2},$$

where $A_n = V_n^{1/2}T_nE_n^{-1/2}$ and $E_n = \sum_{j=1}^{n}(f_n + w_j)e_je_j^*$.

As is the case for $\|X_n\|$, $\|Z_n\|$ cannot converge to zero as n goes to infinity. We state Theorem 4.4 from [16] which we will apply to prove this assertion.

Theorem 3.6.6. *Let $\mathcal{H}$ be a Hilbert space and let $A \in \mathcal{L}(\mathcal{H})$ be compact. Then for $n = 1, 2, \ldots,$*

$$\sigma_n(A) = \min\{\|A - F\| \mid F \in \mathcal{L}(\mathcal{H}), \operatorname{rank} F \leq n - 1\}.$$

where $\sigma_n(A)$ denotes the decreasing sequence of singular values of A.

As discussed in the previous sections, Z_n differs from X_n in its first q rows and r columns. Then, $P_n := X_n - Z_n$ will be a matrix of zeros, except in its first q rows and r columns. From the definition of the rank of a matrix, P_n will be of finite rank for all n and thus compact as well. Indeed, $\operatorname{rank} P_n \leq q + r$, for all n. Now choose $F = P_n$ and $A = X_n$ in the theorem above. Note that for each n, X_n is also a compact operator. Thus, in this case, Theorem 3.6.6 gives

$$\sigma_m(X_n) = \min\{\|X_n - P_n\| \mid P_n \in \mathbb{C}^{n\times n}, \operatorname{rank} P_n \leq m - 1\}, \qquad m = 1, 2, \ldots, n.$$

Since we know that $\operatorname{rank} P_n \leq q + r$ and $X_n - P_n = Z_n$, this equality becomes

$$\sigma_m(X_n) \leq \min\{\|X_n - P_n\| \mid P_n \in \mathbb{C}^{n\times n}, \operatorname{rank} P_n \leq q+r\}, \qquad m = q+r+1, q+r+2, \ldots, n.$$

If we assume that $\lim_{n\to\infty} \|Z_n\| = 0$, the last inequality implies that the $q + r + 1$'th singular value of X_n goes to zero as n approaches infinity. However, Theorem 1.2.3 with $k = q+r$ implies that σ_{q+r} does not go to zero as n tends to infinity. The singular values of X_n are given by the entries of the matrix $D_n + f_nD^{-1}$. Therefore, for every singular value of T_n, there is a singular value of X_n that is larger and consequently, $\sigma_{q+r}(X_n) \geq \sigma_{q+r}$ and $\lim_{n\to\infty} \sigma_{q+r}(X_n) \neq 0$, contradicting our assumption that $\lim_{n\to\infty} \|Z_n\| = 0$.

To determine the asymptotic behaviour of the norm of Z_n, we evaluate the inequality

$$\begin{aligned}\|Z_n\| &\leq \|V_n^{1/2}\|\|A_n + (A_n^{-1})^*\|\|E_n^{1/2}\| \\ &= \max_j \sqrt{v_j}\|A_n + (A_n^{-1})^*\|\sqrt{\max_j w_j + f_n}.\end{aligned}$$

Again we use $\|A_n + (A_n^{-1})^*\|$ and its corresponding function, $g(x) = x + 1/x$, to bound its value. This function has a minimum at 1, and g is monotonically decreasing and increasing on the intervals $(0, 1]$ and $[1, \infty)$ respectively. The value of $\|A_n + (A_n^{-1})^*\|$ will then be determined by either the minimum, $\sigma_n(A_n)$, or maximum $(\sigma_1(A_n))$ singular values of A_n. Starting with $\sigma_1(A_n)$,

$$\sigma_1(A_n) = \|A_n\| \leq \max_j \sqrt{v_j}\frac{1}{\sqrt{f_n}}\|T_n\| \leq \frac{c}{\sqrt{f_n}},$$

where c is some positive constant. The last inequality follows from the fact that $\lim_{n\to\infty}\|T_n\| = \|T(b)\|$. If $\sigma_1(A_n)$ determines the norm of $A_n + (A_n^{-1})^*$, we can find its maximal order of growth:

$$g(\sigma_1(A_n)) \leq g(\frac{c}{\sqrt{f_n}}) = \frac{c}{\sqrt{f_n}} = O(\frac{1}{\sqrt{f_n}}).$$

To bound $\sigma_n(A_n)$, we apply Theorem 3.3.4 once again and it tells us that $\sigma_n(A_n) \leq \frac{1}{\sqrt{f_n}}\sigma_n$. As we already know, $\sigma_n = O(1/n^\alpha)$ and therefore $\sigma_n(A_n)$ is at least $O(1/\sqrt{f_n}n^\alpha)$. We can also bound $\sigma_n(A_n)$ from below by noticing that

$$\|A_n^{-1}\| \leq \|E_n^{1/2}\|\|T_n^{-1}\|\|V_n^{-1/2}\| \leq (\max_j(w_j) + \sqrt{f_n})cn^\alpha.$$

If $W_n \neq 0$, it is clear that $\|A_n^{-1}\| = O(n^\alpha)$, otherwise, $\|A_n^{-1}\| = O(\sqrt{f_n}n^\alpha)$. Then, $\sigma_n(A_n)$ is bounded from below by an $O(n^{-\alpha})$ sequence in general, and $O(1/\sqrt{f_n}n^\alpha)$ when $W_n = 0$. However, we just proved that $\sigma_n(A_n)$ is at least $O(1/\sqrt{f_n}n^\alpha)$ and therefore $\|A_n\| = O(1/\sqrt{f_n}n^\alpha)$ exactly when $W_n = 0$. Hence, for this class of Z_n (when $\|A_n + (A_n^{-1})^*\|$ is determined by $\sigma_n(A_n)$), the order of growth for $\|A_n + (A_n^{-1})^*\|$ is as follows:

$$\lim_{n\to\infty} g(\sigma_n(A_n)) \leq \lim_{n\to\infty} g(\frac{c}{\sqrt{f_n}n^\alpha}) = c\lim_{n\to\infty}\sqrt{f_n}n^\alpha = O(\sqrt{f_n}n^\alpha).$$

In the general case, $\|A_n + (A_n^{-1})^*\|$ will be $O(n^\alpha)$ by an analogous computation.

Since f_n cannot decay faster than $O(1/n)$, and $\alpha \geq 1$, we can assume that the growth of $\|A_n + (A_n^{-1})^*\|$ will be determined by $\sigma_n(A_n)$. Therefore,

$$\|Z_n\| \leq \max_j\sqrt{v_j}\|A_n + (A_n^{-1})^*\|\sqrt{\max_j w_j + f_n} \leq \max_j\sqrt{v_j}cn^\alpha\sqrt{\max_j w_j + f_n},$$

implying that in general, $\|Z_n\|$ is at most $O(n^\alpha)$. If we consider the subclass where $W_n = 0$ again, it follows from the above inequality (with cn^α replaced by $c\sqrt{f_n}n^\alpha$) that $\|Z_n\|$ is at most $O(f_n n^\alpha)$. If we assume further that $f_n = O(1/n)$ and $\alpha = 1$, $\|Z_n\|$ will be bounded. The following is an example of this special case.

Example 3.6.7.

Let $f_n = 1/n$, $T_n = T_{0,n}$ and V_n such that $v_1 = 2, v_2 = 2$ with all other $v_j = 1$ and $W_n = 0$. Figure 3.11 shows the boundedness of Z_n.

We now show that the norm of Z_n is unbounded in general when the associated X_n is Toeplitz. As we have done previously, apply Theorem 3.3.4 to Z_n and let X_{n-p} be the submatrix which is Toeplitz for the smallest possible p. Then, $\|Z_n\| = \sigma_1(Z_n) \geq \sigma_1(X_{n-p}) = \|X_{n-p}\|$.

Even if $W_n = 0$, $\|Z_n\|$ remains unbounded if f_n decays slow enough. This can be seen in the following example.

Example 3.6.8.

Let Z_n be as in the previous example, except that $f_n = 1/\sqrt{n}$. In this case, Z_n is unbounded - Figure 3.12

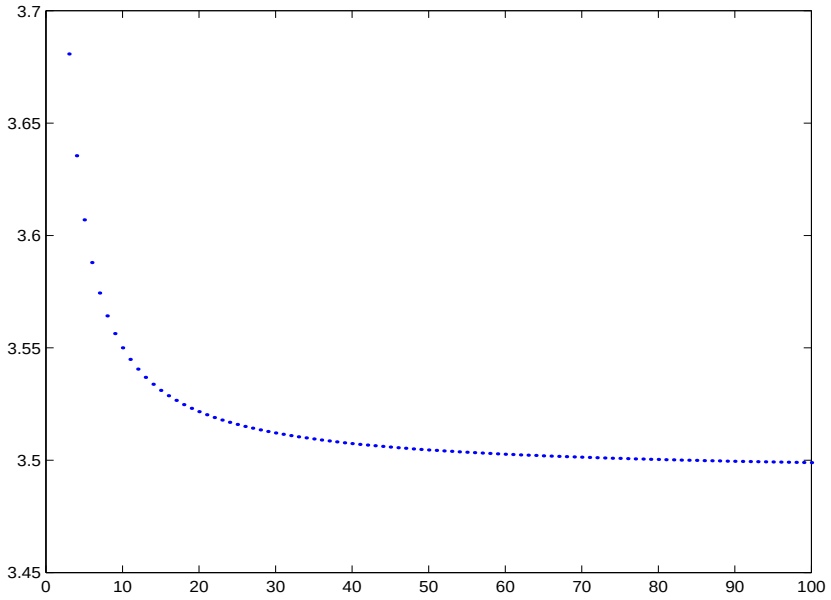

Figure 3.11: The norm of Z_n (dots), plotted as a function of n.

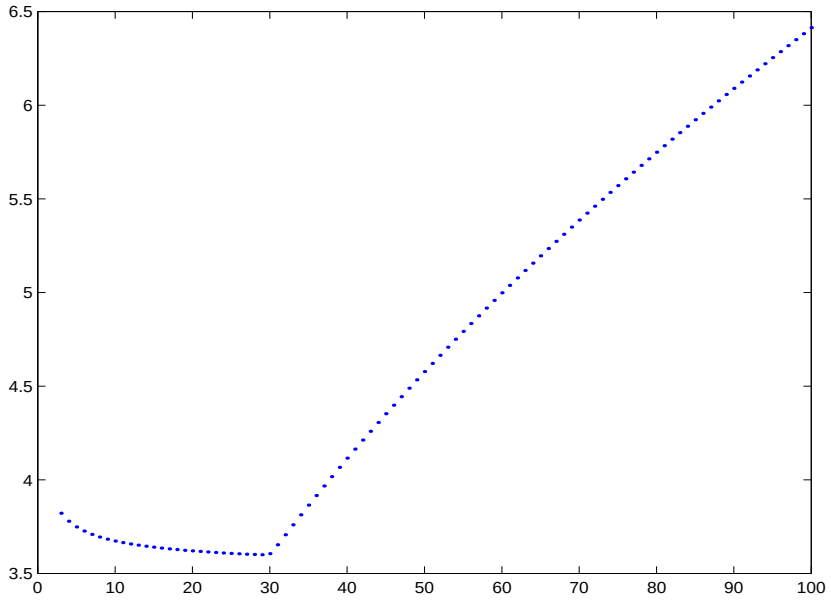

Figure 3.12: The norm of Z_n (dots), plotted as a function of n.

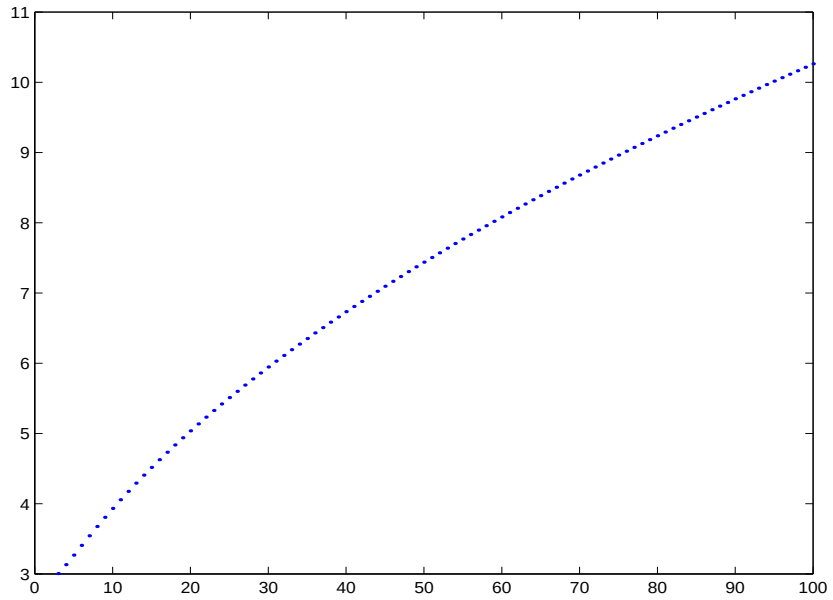

Figure 3.13: The norm of K_n (dots), plotted as a function of n.

Example 3.6.9.

Choose $Z_n = K_n$. The presence of W_n forces $\|K_n\|$ to be unbounded as can be seen in Figure 3.13.

In the last paragraph we showed that $\|Z_n\| \geq O(f_n n^\alpha)$. In the case of K_n, this does not tell us if it is necessarily unbounded as can be seen in figure 3.13. To prove unboundedness, recall that we can express

$$K_n = T_{0,n} + \frac{1}{n}(T_{0,n}^{-1})^* + W_n = T_n(\frac{1}{n}) + W_n,$$

where W_n is the matrix of ones in the first column and zeros elsewhere. We can apply the general norm inequality

$$\|X - Y\| \geq |\|X\| - \|Y\||$$

with X replaced by $T_n(\frac{1}{n})$ and Y by $-W_n$. Then,

$$\|K_n\| = \|T_n(\frac{1}{n}) + W_n\| \geq |\|T_n(\frac{1}{n})\| - \|W_n\||.$$

Taking limits left and right, we notice that we need to show that $\lim_{n\to\infty} \|W_n\|$ is unbounded, since $\|T_n(\frac{1}{n})\|$ is a bounded sequence. We can find the singular values of W_n explicitly by considering the eigenvalues of $W_n W_n^*$. This matrix is conveniently the all-ones matrix, and via direct calculation, this matrix has the eigenvalue n with any constant entry vector as corresponding eigenvector. Since $W_n W_n^*$ maps any vector to a constant entry vector, it follows that it cannot have a non-zero eigenvalue associated with a non-constant entry eigenvector. Therefore, the eigenvalues of $W_n W_n^*$ are n and zero (the latter with multiplicity $n-1$) and $\|W_n\| = \sqrt{n}$, proving that K_n is an unbounded sequence from

the last inequality above. The boundedness of $\|T_n(\frac{1}{n})\|$ is also convenient since it allows us to find an upper bound of the same order as the lower bound. Indeed,

$$\lim_{n\to\infty} \|K_n\| = \lim_{n\to\infty} \|T_n(\frac{1}{n}) + W_n\| \leq \lim_{n\to\infty} \|T_n(\frac{1}{n})\| + \lim_{n\to\infty} \|W_n\| = O(\sqrt{n}),$$

and this forces $\|K_n\| = O(\sqrt{n})$.

3.6.1 Summary

Since the norms of X_n, and particularly Z_n, behave differently under various circumstances, we provide a listed summary of the possible combinations of their growth as n tends to infinity. For X_n, we have the following:

Let $g_n(\sigma_1) \geq g_n(\sigma_n)$ for all n large enough. Then,

$$\|X_n\| = \|T(b)\| + O(f_n). \qquad \text{(Example 3.6.3)}$$

Let $g_n(\sigma_n) \geq g_n(\sigma_1)$ for all n large enough. Then,

- If $f_n = O(1/n)$ and $\alpha = 1$,

  $$\|X_n\| = C + O(\frac{1}{n}), \qquad \text{(Example 3.6.4)}$$

  where C is some positive constant.

- For any other order of f_n or value of α,

  $$\|X_n\| = O(f_n n^\alpha), \qquad \text{(Example 3.6.5)}$$

  and thus unbounded.

For Z_n, we have the following:

- If $W_n \neq 0$ for all n large enough. Then,

  $$\|Z_n\| \leq O(n^\alpha).$$

  – If the matrix X_n associated with Z_n is Toeplitz,

  $$O(n^\alpha) \geq \|Z_n\| \geq O(\|X_{n-2p}\|). \qquad \text{(Example 3.6.9)}$$

- If $W_n = 0$ for all n large enough. Then,

  $$\|Z_n\| \leq O(f_n n^\alpha).$$

  – If the matrix X_n associated with Z_n is Toeplitz,

  $$O(f_n n^\alpha) \geq \|Z_n\| \geq O(\|X_{n-2p}\|). \qquad \text{(Example 3.6.8)}$$

3.6.2 Fredholmness

Up until now we assumed non-Fredholmness for the banded Toeplitz operator, $T(b)$, with corresponding finite sections, $T_n(b)$ or simply T_n. We will revisit the results and analysis of this section and show where Fredholmness of $T(b)$ changes these findings.

For the unperturbed X_n, we had a few results which were central to the results established. Firstly, it is known that $\|T_n\| = T(b) + O(1/n)$ for any banded symbol b. If $g(\sigma_n) \leq g(\sigma_1)$ for n large enough, we showed that $\|X_n\| = \|T(b)\| + O(f_n)$. In the Fredholm case however, this scenario is not possible and it follows from the exponential decay rate of σ_n as established in Theorem 1.2.1. More precisely,

$$\begin{aligned}\lim_{n\to\infty} g_n(\sigma_n) &= \lim_{n\to\infty} \sigma_n + \frac{f_n}{\sigma_n} \\ &= 0 + \lim_{n\to\infty} \frac{f_n}{\sigma_n} \\ &= O(f_n e^{\alpha n}),\end{aligned}$$

and clearly this growth rate is higher than what could be achieved via σ_1, proving that $\|X_n\|$ can only be determined by σ_n in the Fredholm case. Also, since $\alpha > 0$ and f_n cannot converge to zero faster than $O(1/n)$, $\|X_n\|$ will always be unbounded.

For Z_n, the decay rate of σ_n also dominates σ_1 and the effect of f_n. The order of growth for $g(\sigma_1)$ is $O(\frac{1}{\sqrt{f_n}})$ and for $g(\sigma_n)$ it is $O(\sqrt{f_n}e^{\alpha n})$, and therefore $\|Z_n\|$ will also only be determined by the smallest singular values of T_n. Formally,

$$\|Z_n\| \geq \|A_{n-p} + (A_{n-p}^{-1})^*\| \sqrt{\max_j w_j + f_n} \geq \sqrt{\max_j w_j + f_n}\sqrt{f_{n-p}} e^{\alpha(n-p)} = O(\sqrt{f_n} e^{\alpha n}).$$

Note that with W_n set to zero, unboundedness is not affected here.

3.6.3 The condition numbers of X_n and Z_n

The condition number of a $n \times n$ matrix, say A_n, is defined as

$$\kappa(A_n)_p := \|A_n\|_p \|A_n^{-1}\|_p,$$

where A_n is viewed as an operator on $\mathbb{C}_n$ equipped with the l_p norm. Since we have been working with the spectral norm in this chapter, we shall omit the subscript p and write

$$\kappa(A_n) := \|A_n\| \|A_n^{-1}\|.$$

As we already know, both $\|X_n^{-1}\|$ and $\|Z_n^{-1}\|$ form unbounded sequences. The hope that (at least) $\sup_{n\in\mathbb{N}} \kappa(X_n) < \infty$ fails, since that would require $\lim_{n\to\infty} \|X_n\| = 0$ or $\lim_{n\to\infty} \|Z_n\| = 0$ which we disproved in the previous section.

Estimates of the growth of $\kappa(X_n)$ and $\kappa(Z_n)$ can obviously be obtained from employing the results from earlier sections.

3.7 The singular values of X_n and Z_n

The previous sections of this chapter all dealt with either the minimum or maximum singular values of X_n or Z_n. This directly related to the results obtained regarding the asymptotic behaviour of the norms of their inverses, and the norms of the matrices themselves.

In this section, we attempt to shed some light on the other singular values as n tends to infinity. Starting with the easier questions first, we will look at the extreme singular values, i.e., the first (largest) or last (smallest) l singular values where l is any fixed natural number. Then we will consider the 'inner' singular values of some special classes and cases. These singular values include all those that can't be defined as extreme.

3.7.1 The smallest singular values, $\sigma_{n-l}(X_n)$ and $\sigma_{n-l}(Z_n)$

As one might have guessed, the behaviour of $\sigma_{n-l}(X_n)$ and $\sigma_{n-l}(Z_n)$ is very closely related to that of $\|X_n^{-1}\|$ and $\|Z_n^{-1}\|$. In fact, they behave exactly the same. To see this, we recall the general argument used in the proof of Theorem 3.1.1 and note that it is analogous to the case of Theorem 3.1.2. The whole idea revolved around the fact that we needed a singular value of T_n close to $\sqrt{f_n}$. For this purpose we defined the interval $E_n = [\sqrt{f_n} - \sqrt{f_n^{\epsilon+1}}, \sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}]$ including the point $\sqrt{f_n}$. Eventually we arrived at the inequality

$$n\mu(E_n) \geq \frac{n\sqrt{f_n^{\epsilon+1}}(k+1)}{2\pi M} > 15r, \tag{3.9}$$

which, if used with

$$|N_n(E_n) - n\mu(E_n)| \leq 14r,$$

by taking k and n large enough, implies that $N_n(E_n) \geq r \geq 1$ (number of singular values inside E_n) is greater or equal to one. Now, the proof works for any choice of k and therefore, by choosing k even larger, we can use these inequalities to show that an arbitrary fixed number of singular values exist inside E_n. By continuing the same arguments of the proof of Theorem 3.1.1, it follows that

$$\lim_{n\to\infty} 2\sqrt{f_n}\sigma_{n-l}(X_n) = 1,$$

for any fixed k. Actually we can say more, since we will show that Theorem 3.4.1 also applies to $\sigma_{n-l}(X_n)$:

$$\frac{1}{2\sqrt{f_n}} - \sigma_{n-l}(X_n) = O(f_n^{\epsilon-\frac{1}{2}}).$$

Recall the inequality

$$\frac{1}{2\sqrt{f_n}} \geq \sigma_{n-l}(X_n) \geq \frac{\sqrt{f_n} + s_n\sqrt{f_n^{\epsilon+1}}}{2f_n + 2s_n\sqrt{f_n^{\epsilon+2}} + s_n^2 f_n^{\epsilon+1}} := R_n.$$

from the proof of Theorem 3.1.1 where we replaced $\|X_n^{-1}\|$ with $\sigma_{n-l}(X_n)$ - this follows from the previous discussion proving Theorem 3.1.1 for the minimal extreme singular

values. We can rewrite this inequality as

$$\frac{1}{2\sqrt{f_n}} - R_n \geq \frac{1}{2\sqrt{f_n}} - \sigma_{n-l}(X_n) \geq 0. \tag{3.10}$$

Now as before, we can not take the limit as n approaches infinity of R_n. For this purpose, we defined $\tilde{R}_n$ and proved that $R_n \geq \tilde{R}_n$ in Lemma 3.4.2. What is important is that $R_n \geq \tilde{R}_n$ holds for all k and n large enough. This means that the same E_n which we chose large enough to include any $\sigma_{n-l}(X_n)$, can be used in Lemma 3.4.2. This immediately implies that

$$\frac{1}{2\sqrt{f_n}} - \frac{\sqrt{f_n} + k\sqrt{f_n^{\epsilon+1}}}{2f_n + 2k\sqrt{f_n^{\epsilon+2}} + k^2 f_n^{\epsilon+1}} \geq \frac{1}{2\sqrt{f_n}} - \sigma_{n-l}(X_n) \geq 0.$$

and we can follow the proof of Theorem 3.4.1 to show that it holds for $\sigma_{n-l}(X_n)$ with l fixed.

3.7.2 The largest singular values, $\sigma_l(X_n)$ and $\sigma_l(Z_n)$

This subsection is largely based on the results obtained in section 3.6. This is not surprising as Theorem 1.2.3 tells us that all the largest singular values converge to $\|b\|_\infty = \|T(b)\|_2$ (Theorem 1.15, [5]).

For the largest singular values of X_n and Z_n, we split the investigation between the two cases since their behaviour is different. We start with X_n and denote its upper singular values by $\sigma_l(X_n)$.

We mirror the arguments made for $\|X_n\|$ here. Since $\lim_{n\to\infty} \|X_n\| \neq 0$, we know that $\lim_{n\to\infty} \sigma_l(X_n) \neq 0$, since $\lim_{n\to\infty} g_n(\sigma_l(X_n)) \neq 0$.

Now consider the situation when $g_n(\sigma_n) \leq g_n(\sigma_1)$. In this case we proved that $\|X_n\| = \|T(b)\|_2 + O(f_n)$ by calculating $\lim_{n\to\infty} g_n(\sigma_1)$. However, the convergence of $\sigma_l(T_n(b))$ to $\|T(b)\|_2$ is guaranteed by Theorem 1.2.3 and hence there is and arbitrary number of X_n's singular values that converge to $\|T(b)\|_2$ and hence $\sigma_{l,n}(X_n) = \|T(b)\| + O(f_n)$ by performing the same calculation as for $\|X_n\|$.

The other possibility occurs when $g_n(\sigma_n) \geq g_n(\sigma_1)$ and here $\|X_n\| = O(f_n n^\alpha)$ was proved. This time we apply Theorem 1.2.2 to σ_l and then $g_n(\sigma_l) = O(f_n n^\alpha)$ ensures that $\sigma_l(X_n) = O(f_n n^\alpha)$. We note again that these singular values will form bounded sequences if $f_n = O(1/n)$ and $\alpha = 1$.

As could be expected, the analysis for Z_n is more involved, and the shortcomings of section 3.6 make their appearance again. The fact that we don't have a suitable decomposition for Z_n that allows us to find its singular values, is the reason for the difficulties. Even when considering $\|Z_n\| = \sigma_1(Z_n)$, we had to rely on norm inequalities to find bounds for its value.

To uncomplicate this situation slightly, we assume that the X_n corresponding to Z_n is Toeplitz itself. This allows us to apply Theorem 3.3.4 with X_{n-p} as a submatrix of Z_n, and then

$$\sigma_1(X_{n-p}) \geq \sigma_{1+2p}(Z_n).$$

This tells us that all the largest singular values of Z_n, except the upper $2p$ ones, are interlaced with those of X_n and their behaviour is therefore the same as the largest singular values of X_n.

3.7.3 The inner singular values of X_n and Z_n

By inner singular values we mean those singular values that do not necessarily belong to the group of largest or smallest singular values. Here we take a more global approach and consider their distribution over an interval, rather than track their behaviour as n approaches infinity.

The fact that we assume a banded Toeplitz generator matrix T_n in the definition of X_n and Z_n is critical, and without it, we wouldn't be able to prove theorems like Theorem 3.1.1 or Theorem 3.4.1. The reason is Lemma 3.3.1, and here it will be central to our analysis as well. Recall that for a banded symbol $b(t) = \sum_{j=-r}^{r} b_j t^j$, $t \in \mathbb{T}$, the following is true for the Toeplitz matrix $T_n(b)$:

$$|N_n(E) - n\mu(E)| \leq 14r \qquad \text{for all} \quad n \geq 1,$$

where $N_n(E)$ is the number of singular values of $T_n(b)$ in E and

$$\mu(E) = \frac{1}{2\pi}|\{t \in \mathbb{T} : |b(t)| \in E\}|,$$

with $|.|$ denoting the Lebesgue measure on the unit circle.

We assumed in the beginning of this chapter that $b(t) = 0$ for some $t \in \mathbb{T}$ and that it is not identically zero. Therefore the range of $|b(t)|$ is the closed interval $[0, \|b\|_\infty]$. If E is chosen as any interval whose intersection with $[0, \|b\|_\infty]$ is not empty, then Lemma 3.3.1 tells us that for n large enough, E will contain at least one singular value of $T_n(b)$, and as n grows, $[0, \|b\|_\infty]$ will be populated with the singular values of $T_n(b)$.

By now we are very familiar with the diagonal matrix $D + f_n D^{-1}$ which contains the singular values of X_n. Corresponding to this matrix we also defined the function $g_n(x) = x + f_n 1/x$, which has a minimum at $x = \sqrt{f_n}$ with a positive derivative on the interval $[\sqrt{f_n}, \infty]$. Note that, as $n \to \infty$, $g_n(x)$ will behave like the straight line function $f(x) = x$ on the interval $[\sqrt{f_n}, \infty]$. Therefore, as n tends to infinity, the singular values of $T_n(b)$ in $[\sqrt{f_n}, \infty]$, mapped by the function $g_n(x)$, will produce singular values of X_n that are very close to those of $T_n(b)$, mapped by $g_n(x)$. Also, since $\sqrt{f_n}$ goes to zero, X_n will also populate the interval $[0, \|b\|_\infty]$ with its singular values, similar to what $T_n(b)$ does. However, depending on the function f_n, the function $g_n(x)$ will also map the singular values of $T_n(b)$ that are smaller than $\sqrt{f_n}$ to corresponding singular values of X_n. The negative gradient of $g_n(x)$ in this region grows quickly as x tends to zero and accounts for some of the larger, and indeed the largest singular values of X_n, especially if f_n decays slowly and if the smallest singular values of $T_n(b)$ goes to zero quickly (see Theorem 1.2.2).

In the case of Z_n, we assume again that its corresponding matrix X_n is Toeplitz. From the last paragraph of the previous subsection we have

$$\sigma_1(X_{n-p}) \geq \sigma_{1+2p}(Z_n).$$

The singular values of Z_n except the largest $2p$ ones are interlaced with those of X_{n-p}. Therefore, if $\|X_n\|$ is bounded and Z_n is not, we will see that most of the singular values of Z_n will cluster in the interval $[0, \|b\|_\infty]$ with at least the largest one approaching infinity as n grows large. Figure 3.14 shows this behaviour for a few values of n when we choose $Z_n = K_n$.

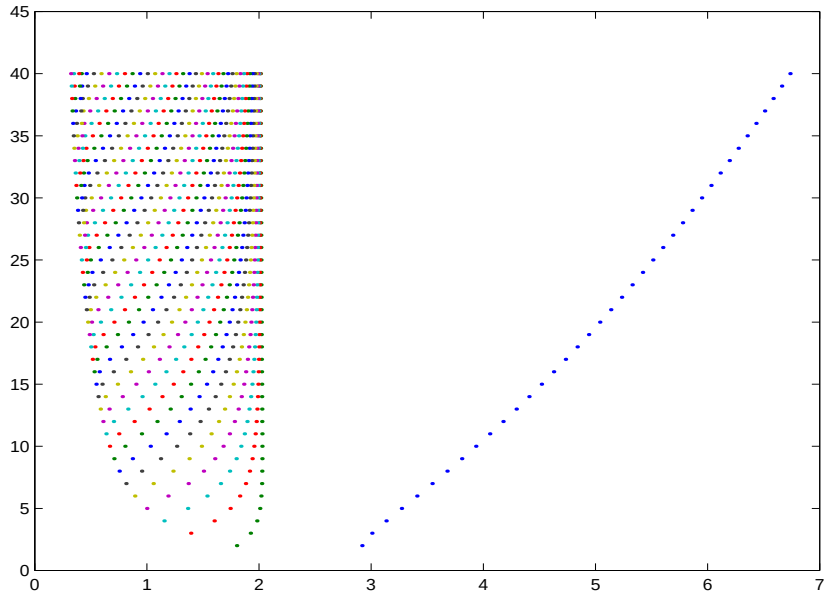

Figure 3.14: The singular values of K_n, plotted as a function of n.

From the figure it is clear that $\sigma_1(K_n)$ goes to infinity ($\mathrm{O}(\sqrt{n})$ as shown before), but $\sigma_2(K_n)$ seems to stay close to the value two. The other singular values of K_n are interlaced with those of $T_n(1/n)$, which we know lie in the interval $[0, 2]$ (Actually we will calculate the singular values of $T_n(1/n)$ explicitly in the next section). Hence, we only need to know the behaviour of $\sigma_2(K_n)$ to have a fairly complete picture of what happens to the singular values of K_n as n tends to infinity.

Consider Theorem 3.6.6, and let K_n play the role of A. Then this theorem states that

$$\sigma_2(K_n) = \min\{\|K_n - F\| \mid F \in \mathcal{L}(\mathcal{H}), \operatorname{rank} F \leq 1\}.$$

Let F_n be the $n \times n$ matrix with ones in the first column and zeros elsewhere. It has rank one, and then

$$\sigma_2(K_n) \leq \|K_n - F_n\| = \|T_n\left(\frac{1}{n}\right)\|.$$

We know that $\sigma_2(K_n) \geq \sigma_2(T_{n-1}(1/n))$ from Theorem 3.3.4, and therefore $\sigma_1(T_n(1/n)) \geq \sigma_2(K_n) \geq \sigma_2(T_{n-1}(1/n))$. Since all the upper singular values of $T_n(1/n)$ converge to 2 as n tends to infinity, the last inequality forces $\sigma_2(K_n)$ to do the same.

Remark 3.7.1.

As we have seen throughout this section, the analysis of Z_n is more complicated than that of X_n. This can mostly be ascribed to the fact that we do not have a SVD for Z_n in terms of its generator matrix T_n. One might be tempted to try to analyze the behaviour of Z_n by considering the behaviour of the perturbation apart from that of X_n and then 'adding' their behaviour in some way to describe what happens with Z_n itself. This approach is fundamentally flawed when $W_n \neq 0$, since W_n depends on the generator matrix itself. We could expect then that two similar X_n's, with exactly the same choice of

w_j's might show different behaviour. This is indeed the case, and we will give an example of two X_n's with exactly the same set of singular values, and by choosing the same set of w_j's, they will produce an unbounded Z_n corresponding to the one X_n, and a bounded Z_n corresponding to the other.

We return to our usual example, $T_n(1/n)$ with generator $T_{0,n}$. We choose W_n to be such that $w_1 = 1$ with all other $w_j = 0$. This produces our favourite $Z_n = K_n$. For the other X_n, we choose as generator $T^*_{0,n}$. From the definition of X_n it is easy to see that in this case, $X_n = (T_n(1/n))^*$. Clearly the singular values of $T_n(1/n)$ and $(T_n(1/n))^*$ are the same and we know their behaviour. In the latter case however, the perturbation that produces Z_n is different, and $[(T^*_{0,n})^{-1}]^* W_n$ is the matrix of all zeros, except in the one-one position. Therefore,

$$\|Z_n\| \leq \|T^*_{0,n} + \frac{1}{n}[(T^*_{0,n})^{-1}]^*\| + \|W_n\| \leq C,$$

where C is some constant. This follows from the fact that $\|W_n\| = 1$ for all n and $\|T^*_{0,n} + \frac{1}{n}[(T^*_{0,n})^{-1}]^*\| = \|T_n(1/n)\|$ which is bounded.

3.8 Future work and open problems

T-gen matrices with their associated finite rank perturbations constitute a large class of matrices, reaching far beyond Toeplitz matrices themselves, although the intersection of these two sets does not represent a significant part of either class. As mentioned before, the presence of T_n^{-1} in the construction of T-gen matrices is mostly responsible for this, and generalizations to widen the T-gen class will probably not contain significantly more Toeplitz matrices. Considering the strong results on the asymptotic behaviour of the norms of the inverses, it is reasonable to want to find Toeplitz or variable coefficient Toeplitz matrices which fall within this class, or extend this class to contain more of these matrices.

Some obvious avenues of generalization do exist that are worth discussing and include non-banded Toeplitz matrices. All the work done in this chapter is with regard to Toeplitz matrices with finite non-zero diagonals. The Avram-Parter Theorem (1.2.4), which describes the distribution of singular values is applicable to the non-banded case, as well as the block case for matrix valued symbols. We can then assume that the main results established for the norms of the inverses of T-gen matrices might remain true if we enlarged our class of generators to include non-banded and block Toeplitz matrices. The problem here is that we do not have an analogous result for Lemma 3.3.1 in the non-banded or block case. However, we know that Lemma 3.2.1 applies for any matrix which at least ensures that the upper bound for the growth applies for these more general classes. The following example shows visually similar behaviour to the banded case.

Example 3.8.1.

Consider the block Toeplitz matrix

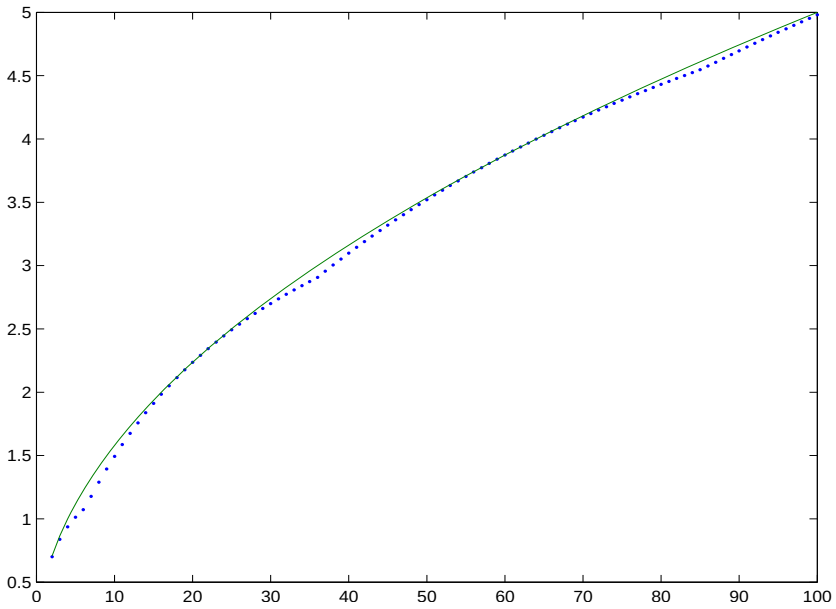

Figure 3.15: The norm of X_n (dots) and $\frac{\sqrt{n}}{2}$ (line) plotted as functions of n.

$$T_n = \begin{bmatrix} \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix} & \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \cdots \\ \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix} & \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} & \cdots \\ \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} & \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix} & \\ \vdots & \vdots & & \ddots \end{bmatrix}$$

with symbol

$$R(t) = \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} t^{-1} = \begin{bmatrix} 0 & 2 - t^{-1} \\ 1 - t^{-1} & 0 \end{bmatrix}.$$

It is clearly a banded symbol such that $\det R(t)$ vanishes on $\mathbb{T}$. Now form the block T-gen matrices, $X_n = T_n + f_n(T_n^{-1})^*$, with $f_n = 1/n$. X_n is now the $2n \times 2n$ matrix with entries 2×2 blocks. Figure 3.15 shows similar behaviour to what was seen in the first scalar example, $X_n = T_n(1/n)$ - Figure 3.2.

There also remain other less fundamental gaps in the theory developed in this chapter. Recall that we proved our main Theorem 3.4.1 for the positive definite case - Theorem 3.4.8. Here we had to assume that the arbitrary powers a and b in $X_n = T_n^a + f_n(T_n^{-b})^*$ have to be equal. Numerical investigations (e.g. figure 3.5) show that this is probably a relic of the techniques used in the proof, and not because the theorem is untrue if assumed otherwise.

When considering the norms of Z_n asymptotically, a lot can still be done. For instance, the bound given for the growth derived was not sharp. For the choice $Z_n = K_n$, we showed that $\|K_n\| = O(\sqrt{n})$, but the general bounds given were $O(n) \geq \|K_n\| \geq \|X_{n-2p}\|$ ($\alpha = 1$). In other words, this statement does not even guarantee that K_n is unbounded since $X_{n-2p} = T_{n-2p}(1/n)$ is a bounded sequence in this case. The upper bound is also off by $O(\sqrt{n})$. Clearly the presence of W_n in Z_n has a strong influence on its behaviour, but it does not seem to be trivially quantifiable.

Chapter 4

The eigenvalues and eigenvectors of a special perturbed tridiagonal Toeplitz matrix

4.1 Introduction

The main contribution of this section is an explanation of a peculiar phenomenon regarding the singular vectors of the special Toeplitz matrix $T_n(\frac{1}{cn})$ of the previous chapter. Recall that

$$T_n(\frac{1}{cn}) = \begin{bmatrix} 1+\frac{1}{cn} & -1 & 0 & \cdots & \cdots & 0 \\ \frac{1}{cn} & 1+\frac{1}{cn} & -1 & 0 & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -1 & 0 \\ \vdots & \ddots & \ddots & \ddots & & -1 \\ \frac{1}{cn} & \cdots & \cdots & \cdots & \frac{1}{cn} & 1+\frac{1}{cn} \end{bmatrix}, \tag{4.1}$$

where c is some positive constant. Recall that the matrix is of size $n \times n$. Notice that the entries of $T_n(\frac{1}{cn})$ change as its size increases. Fortunately, this does not impact on the underlying structure of this sequence of matrices. In fact, it is this underlying structure from which the phenomenon discussed in section 4.3 is born.

As shown before, we can express

$$T_n\left(\frac{1}{cn}\right) = T_{0,n} + \frac{1}{cn}(T_{0,n}^{-1})^*,$$

where

$$T_{0,n} = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & -1 \\ & & & & 1 \end{bmatrix}, \tag{4.2}$$

and $T_n(\frac{1}{cn})$ is said to be generated by the banded Toeplitz matrix $T_{0,n}$. Notice that $T_{0,n}$is independent of the constant c or n.

In section 4.3 we will show that the singular vectors of $T_n(\frac{1}{cn})$ are exactly the singular vectors of $T_{0,n}$, and hence the phenomenon in question can be explained by looking at the much simpler matrix $T_{0,n}$. Actually, since we are interested in singular values, we will be considering the matrix

$$P_n := T_{0,n}^* T_{0,n} = \begin{bmatrix} 1 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}. \tag{4.3}$$

We will show that explicit formulas exist for the eigenvalues and eigenvectors of this matrix. This result has been achieved previously by G.Strang ([38]), but our approach here seems to be new.

Theorem 4.1.1. *The eigenvalues, λ_m, and corresponding eigenvectors, x_m, of the matrix P_n are given by*

$$\lambda_m = 2 - 2\cos(\theta_m), \qquad x_m = [\sin n\theta_m, \sin(n-1)\theta_m, \ldots, \sin\theta_m]^T, \tag{4.4}$$

where

$$\theta_m = \frac{2m+1}{2n+1}\pi,$$

and $0 \le m \le n-1$.

The second section of this chapter will be dedicated to establishing Theorem 4.1.1. In Section 4.3 we will apply this result to explain the singular vector phenomenon of the matrix $T_n(\frac{1}{cn})$. The third section will be devoted to computing $T_n(\frac{1}{cn})^{-1}$, $K_n^{\pm 1}$, $\det K_n^{\pm 1}$ and $\|T_n(\frac{1}{cn})^{\pm 1}\|$.

Most of the contents of this chapter can also be found in [28].

4.2 The eigenvalues and eigenvectors of P_n

The first part of this section will follow the exposition of [5], Section 14.1. We shall modify a few of the results to be applicable in our case, and this will lead to formulas for the eigenvectors of a special class of Toeplitz matrices with an 'impurity'. P_n will be a special case of this type of matrices.

Define [1]

$$T_n(\sigma) = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix},$$

[1] $T_n(\sigma)$ arises naturally in differential equation applications - see for instance [40], Chapter 6.

where $\sigma(t) = 2 - t - t^{-1}$, $t \in \mathbb{T}$, and let E_{jj} denote the matrix whose jj entry is one and all other zero. These definitions allow us to write

$$P_n = T_n(\sigma) - E_{11}.$$

Now, denote $D_n(a) := \det T_n(a)$ where $a(t) = a_0 + a_1 t + a_{-1} t^{-1}$ $(t \in \mathbb{T})$ is the symbol of a tridiagonal Toeplitz matrix. We state the following two theorems from [5] for convenience.

Theorem 4.2.1. *The eigenvalues of $T_n(a)$ are*

$$\lambda_j = a_0 + 2\sqrt{a_1 a_{-1}} \cos \frac{\pi j}{n+1} \qquad (j = 1, \ldots, n),$$

and an eigenvector for λ_j is $x_j = (x_1^{(j)}, \ldots, x_n^{(j)})^T$ with

$$x_k^{(j)} = \left(\sqrt{\frac{a_1}{a_{-1}}} \right)^k \sin \frac{k\pi j}{n+1} \qquad (k = 1, \ldots, n).$$

Theorem 4.2.2. *Let q_1 and q_2 be the zeros of the polynomial $q^2 - a_0 q + a_1 a_{-1}$. Then*

$$D_n(a) = \frac{q_2^{n+1} - q_1^{n+1}}{q_2 - q_1} \qquad \text{if} \quad q_1 \neq q_2,$$
$$D_n(a) = (n+1) q^n \qquad \text{if} \quad q_1 = q_2 = q.$$

From now on we are only concerned with our particular symbol $\sigma(t)$ and we define $D_n(\lambda) := D_n(\sigma - \lambda)$. Let q_1 and q_2 be the zeros of the polynomial $q^2 - (2-\lambda)q + 1$, that is,

$$q_1 = \frac{(2-\lambda) + \sqrt{(2-\lambda)^2 - 4}}{2},$$
$$q_2 = \frac{(2-\lambda) - \sqrt{(2-\lambda)^2 - 4}}{2}.$$

From the previous theorem we deduce that if $q_1 \neq q_2$, then

$$D_n(\lambda) = \frac{q_2^{n+1} - q_1^{n+1}}{q_2 - q_1}.$$

We now prove a lemma which will provide necessary and sufficient conditions for a real number λ to be an eigenvalue of $T_n(\sigma)$ plus an impurity.

Lemma 4.2.3. *Let $v \in \mathbb{R}$. A number $\lambda \in \mathbb{R}$ is an eigenvalue of $T_n(\sigma) + vE_{jj}$ if and only if*

$$(2 + v - \lambda) D_{j-1}(\lambda) D_{n-j}(\lambda) - D_{j-2}(\lambda) D_{n-j}(\lambda) - D_{j-1}(\lambda) D_{n-j-1}(\lambda) = 0.$$

Proof.

$$
\begin{aligned}
&\det[T_n(\sigma-\lambda)+vE_{jj}]\\
&=\det\begin{bmatrix}
2-\lambda & -1 & & & & & & & \\
-1 & 2-\lambda & \ddots & & & & & & \\
 & \ddots & \ddots & \ddots & & & & & \\
 & & -1 & 2+v-\lambda & -1 & & & & \\
 & & & -1 & 2-\lambda & -1 & & & \\
 & & & & \ddots & \ddots & \ddots & & \\
 & & & & & \ddots & \ddots & -1 & \\
 & & & & & & -1 & 2-\lambda &
\end{bmatrix},\\
&=(2+v-\lambda)\det T_{j-1}(\sigma-\lambda)\det T_{n-j}(\sigma-\lambda)\\
&\quad-\det T_{j-2}(\sigma-\lambda)\det T_{n-j}(\sigma-\lambda)-\det T_{j-1}(\sigma-\lambda)\det T_{n-j-1}(\sigma-\lambda)\\
&=(2+v-\lambda)D_{j-1}(\lambda)D_{n-j}(\lambda)-D_{j-2}(\lambda)D_{n-j}(\lambda)-D_{j-1}(\lambda)D_{n-j-1}(\lambda).
\end{aligned}
$$

□

The following proposition will give the general formulas for the singular vectors of $T_n(\sigma)$ plus an impurity. In the special case of P_n, we will be able to simplify the formulas further, and find the eigenvectors in terms of its associated eigenvalues.

Proposition 4.2.4. *Let $v \in \mathbb{R}$ and let $\lambda \in \mathbb{R}\backslash\{0,4\}$ be an eigenvalue of the matrix $T_n(\sigma)+vE_{jj}$. Then an eigenvector for λ is given by*

$$x=(\psi_1,\ldots,\psi_j,\phi_{n-j},\ldots,\phi_1)^T,$$

where

$$
\begin{aligned}
\psi_k &= D_{n-j}(\lambda)D_{k-1}(\lambda)=\frac{q_2^{n-j+1}-q_1^{n-j+1}}{q_2-q_1}\frac{q_2^k-q_1^k}{q_2-q_1} \quad (1\le k\le j),\\
\varphi_k &= D_{j-1}(\lambda)D_{k-1}(\lambda)=\frac{q_2^j-q_1^j}{q_2-q_1}\frac{q_2^k-q_1^k}{q_2-q_1} \quad (1\le k\le n-j),\\
q_{1/2} &= \frac{(2-\lambda)\pm\sqrt{(2-\lambda)^2-4}}{2}.
\end{aligned}
$$

Proof. We must verify that

$$[T_n(\sigma-\lambda)x+vE_{jj}x]_k=0 \tag{4.5}$$

for $1 \leq k \leq n$. If $k \in \{2, \ldots, n\}$, then the left-hand side of equation (4.5) is

$$\begin{aligned}
&-\psi_{k-1} + (2-\lambda)\psi_k - \psi_{k+1} \\
=&- D_{n-j}(\lambda)[D_{k-2}(\lambda) - (2-\lambda)D_{k-1}(\lambda) + D_k(\lambda)] \\
=&- D_{n-j}(\lambda)\Big[\frac{q_2^{k-1} - q_1^{k-1}}{q_2 - q_1} - (2-\lambda)\frac{q_2^k - q_1^k}{q_2 - q_1} + \frac{q_2^{k+1} - q_1^{k+1}}{q_2 - q_1}\Big] \\
=&- \frac{D_{n-j}(\lambda)}{q_2 - q_1}[q_2^{k-1} - q_1^{k-1} - (2-\lambda)(q_2^k - q_1^k) + q_2^{k+1} - q_1^{k+1}] \\
=&- \frac{D_{n-j}(\lambda)}{q_2 - q_1}[q_2^{k-1}(q_2^2 - (2-\lambda)q_2 + 1) - q_1^{k-1}(q_1^2 - (2-\lambda)q_1 + 1)] \\
=&- \frac{D_{n-j}(\lambda)}{q_2 - q_1}[q_2^{k-1}(0) - q_1^{k-1}(0)] \\
=&0.
\end{aligned}$$

The zeros in the second last equation follow from the fact that q_1 and q_2 are the roots of the polynomial $q^2 - (2-\lambda)q + 1$. The case $k = 1$ is proved with a similar calculation. If $k = j + l$, $l \in \{1, \ldots, n-j-1\}$ then the left-hand side of (4.5) is

$$\begin{aligned}
&-x_{j+l-1} + (2-\lambda)x_{j+l} - x_{j+l+1} \\
=&- \varphi_{n-j-l+2} + (2-\lambda)\varphi_{n-j-l+1} - \varphi_{n-j-l} \\
=&- D_{j-1}(\lambda)D_{n-j-l+1}(\lambda) - (2-\lambda)D_{j-1}(\lambda)D_{n-j-l}(\lambda) + D_{j-1}(\lambda)D_{n-j-l-1}(\lambda) \\
=&- \frac{D_{j-1}}{q_2 - q_1}[(q_2^{n-j-l+2} - q_1^{n-j-l+2}) - (2-\lambda)(q_2^{n-j-l+1} - q_1^{n-j-l+1}) + (q_2^{n-j-l} - q_1^{n-j-l})] \\
=&- \frac{D_{j-1}}{q_2 - q_1}[q_2^{n-j-l}(q_2^2 - (2-\lambda)q_2 + 1) - q_1^{n-j-l}(q_1^2 - (2-\lambda)q_1 + 1)] \\
=&\ 0.
\end{aligned}$$

For $k = n$, a similar argument is used. We are left with $k = j$, and again, the left-hand side of (4.5) is

$$\begin{aligned}
&-\psi_{j-1} + (2-\lambda)\psi_j - \psi_{j+1} \\
=&- D_{n-j}(\lambda)D_{j-2}(\lambda) + (2+v-\lambda)D_{n-j}(\lambda)D_{j-1}(\lambda) - D_{n-j-1}(\lambda)D_{j-1}(\lambda) \\
=&\ 0,
\end{aligned}$$

by Lemma 4.2.3. □

Proof of Theorem 4.1.1 Applying the Gerš̃gorin disc theorem (see [18], p.344) to P_n, we see that $0 \leq \lambda \leq 4$ and this implies that q_1 and q_2 are complex numbers where only the sign of the complex part differs. Therefore, $|q_1| = |q_2| = 1$ via a direct computation, and

$$q_1 = \cos\theta + i\sin\theta \quad q_2 = \cos\theta - i\sin\theta,$$

with $\lambda = 2 - 2\cos\theta$. Clearly,

$$q_1^k = \cos k\theta + i\sin k\theta, \quad q_2^k = \cos k\theta - i\sin k\theta,$$

which implies that $q_2^k - q_1^k = -2i \sin k\theta$. In this case, since $j = 1$,

$$\psi_1 = \frac{q_2^n - q_1^n}{q_2 - q_1} = \frac{\sin n\theta}{\sin \theta}$$
$$\varphi_k = \frac{q_2^k - q_1^k}{q_2 - q_1} = \frac{\sin k\theta}{\sin \theta} \quad (1 \leq k \leq n-1)$$

and hence we have the eigenvector

$$x = [\sin n\theta, \sin(n-1)\theta, \ldots, \sin \theta]^T.$$

We now have the eigenvalue, λ, and the eigenvector, x, in terms of the angle θ. We can then solve for θ by writing

$$P_n \begin{bmatrix} \sin n\theta \\ \sin(n-1)\theta \\ \vdots \\ \sin \theta \end{bmatrix} = (2 - 2\cos\theta) \begin{bmatrix} \sin n\theta \\ \sin(n-1)\theta \\ \vdots \\ \sin \theta \end{bmatrix}.$$

The first coordinate gives $\sin n\theta - \sin(n-1)\theta = (2 - 2\cos\theta)\sin n\theta$. Using the fact that $\sin(n-1)\theta = -\cos n\theta \sin\theta + \sin n\theta \cos\theta$, we see that

$$\cos n\theta \sin\theta = \sin n\theta(1 - \cos\theta) = \sin n\theta \left(1 - \cos\left(2.\frac{\theta}{2}\right)\right) = \sin n\theta \left(2\sin^2 \frac{\theta}{2}\right).$$

Division by $\sin n\theta$ gives

$$\tan n\theta = \frac{\sin\theta}{2\sin^2\frac{\theta}{2}} = \frac{2\sin\frac{\theta}{2}\cos\frac{\theta}{2}}{2\sin^2\frac{\theta}{2}} = \frac{\cos\frac{\theta}{2}}{\sin\frac{\theta}{2}} = \frac{\cos(\frac{\pi}{2} - \frac{\theta}{2})}{\sin(\frac{\pi}{2} - \frac{\theta}{2})} = \tan\left(\frac{\pi}{2} - \frac{\theta}{2}\right).$$

Therefore,

$$n\theta = \frac{\pi}{2} - \frac{\theta}{2} + m\pi, \qquad (m \in \{0, 1, \ldots, n-1\})$$

and hence

$$\theta_m = \frac{2m+1}{2n+1}\pi, \qquad \text{and} \qquad \lambda_m = 2 - 2\cos\theta_m,$$

where $m \in \{0, 1, \ldots, n-1\}$.

□

Remark 4.2.5.

By definition, $P_n = T_{0,n}^* T_{0,n}$. An analogous proof as for Theorem 4.1.1 will also yield the 'right' singular vectors of $T_{0,n}$, i.e., the eigenvectors of $T_{0,n}T_{0,n}^*$. In fact, the entries of the right singular vectors will be the same as for the left singular vectors, except that they will appear in reverse order. This follows from applying Proposition 4.2.4 with $j = n$, instead of $j = 1$ as in the proof of Theorem 4.1.1.

4.3 A peculiar permutation phenomenon

When computing the singular value decomposition of $T_n(\frac{1}{cn})$, one sees that for certain values of n, the absolute values of the entries of the singular vectors are all permutations of the same n numbers. For example, if we take $c = 1$, and $T_n(\frac{1}{cn}) = UDV^*$, then for $n = 3, 4, 5, 6$ respectively,

$$U_3 = \begin{bmatrix} -0.737 & 0.591 & 0.328 \\ -0.591 & -0.328 & -0.737 \\ -0.328 & -0.737 & 0.591 \end{bmatrix},$$

,

$$U_4 = \begin{bmatrix} -0.657 & -0.577 & -0.429 & -0.228 \\ -0.577 & 0 & 0.577 & 0.577 \\ -0.429 & 0.577 & 0.228 & -0.657 \\ -0.228 & 0.577 & -0.657 & 0.429 \end{bmatrix},$$

$$U_5 = \begin{bmatrix} 0.597 & 0.549 & 0.456 & 0.326 & 0.170 \\ 0.549 & 0.170 & -0.326 & -0.597 & -0.456 \\ 0.456 & -0.326 & -0.549 & 0.170 & 0.597 \\ 0.326 & -0.597 & 0.170 & 0.456 & -0.549 \\ 0.170 & -0.456 & 0.597 & -0.549 & 0.326 \end{bmatrix},$$

and

$$U_6 = \begin{bmatrix} 0.551 & -0.519 & -0.457 & -0.368 & -0.258 & -0.133 \\ 0.519 & -0.258 & 0.133 & 0.457 & 0.551 & 0.368 \\ 0.457 & 0.133 & 0.551 & 0.258 & -0.368 & -0.519 \\ 0.368 & 0.457 & 0.258 & -0.519 & -0.133 & 0.551 \\ 0.258 & 0.551 & -0.368 & -0.133 & 0.519 & -0.457 \\ 0.133 & 0.368 & -0.519 & 0.551 & -0.457 & 0.258 \end{bmatrix}.$$

For $n = 3, 5, 6$, we see that the mentioned phenomenon occurs, but not for $n = 4$. Notice that for the case $n = 4$, a zero entry can be found in the third column vector.

As mentioned in the introduction, we will not need to study the singular vectors of $T_n(\frac{1}{cn})$ directly. We will show here that $T_{0,n}$ has exactly the same singular vectors.

To see this, let $T_{0,n} = UDV^*$ be a singular value decomposition (SVD). Then it follows that $(T_{0,n}^{-1})^* = UD^{-1}V^*$ and consequently,

$$T_n(\frac{1}{cn}) = U(D + f(n)D^{-1})V^*$$

and

$$T_n(\frac{1}{cn})[T_n(\frac{1}{cn})]^* = U(D + f(n)D^{-1})^2U^*.$$

A direct calculation shows that the column vectors of U are eigenvectors of $T_n(\frac{1}{cn})[T_n(\frac{1}{cn})]^*$ with associated eigenvalues the entries of $(D + f(n)D^{-1})^2$. From the previous section we

have an exact description of the entries of these singular vectors. To understand their behaviour, we must take a detour to basic abstract algebra.

Consider the set $S := \{e^{i\theta_p} : \theta_p = \frac{p}{2n+1}\pi,\ 0 \leq p \leq 4n+1\}$ of complex numbers. This set equipped with regular multiplication forms a group of $4n+2$ elements. If we define the function $\phi(e^{i\theta_p}) = [p]$ between the group S and $\mathbb{Z}_{4n+2}$, we see that S is isomorphic to $\mathbb{Z}_{4n+2}$.

Since the sines of θ and $\pi + \theta$ differ only by sign, and since we are interested in the absolute values of the sines, let us identify the angles θ and $\pi + \theta$. In the group $\mathbb{Z}_{4n+2}$, this amounts to taking the quotient group with respect to the subgroup $\{0, 2n+1\}$. If we define the function $\phi(\{[a], [a+2n+1]\}) = [a]$ between $\mathbb{Z}_{4n+2}/\{0, 2n+1\}$ and $\mathbb{Z}_{2n+1}$, we see that $\mathbb{Z}_{4n+2}/\{0, 2n+1\}$ is isomorphic to $\mathbb{Z}_{2n+1}$.

In the previous section, we derived the formulas for the entries of the singular vectors of $T_n(\frac{1}{n})$, i.e., $x_m = [\sin n\theta_m, \ldots, \sin\theta_m]^T$. We see from experiments in Matlab that the permutation phenomenon fails when one of the singular vectors contains a zero entry, i.e., when the sine of $l\theta_k$, for some $0 \leq k \leq n-1$, is a multiple of π. That happens if and only if $(2k+1)l$ is a multiple of $2n+1$. In other words,

Proposition 4.3.1. *Let n be given, and let $\theta_k = \frac{2k+1}{2n+1}\pi$. Then there is an l such that $\sin(l\theta_k) = 0$ if and only if $2k+1$ is a divisor of zero in $\mathbb{Z}_{2n+1}$.*

It is now clear why for $n = 3$ there are no zero entries in U_3 ($\mathbb{Z}_7$ has no divisors of zero), but for U_4 we do ($\mathbb{Z}_9$ has zero divisor 3).

From now on we assume that $2k+1$ is not a divisor of zero in $\mathbb{Z}_{2n+1}$. Consider the set $A_k = \{(2k+1)l \mid l = 1, 2, \cdots, n\}$. Suppose that for some $l_1 > l_2$ we have

$$(2k+1)l_1 = (2k+1)l_2 \mod (2n+1).$$

Then $(2k+1)(l_1 - l_2) = 0 \mod (2n+1)$, and this is a contradiction with $2k+1$ not being a divisor of zero in $\mathbb{Z}_{2n+1}$. So, modulo $2n+1$, the set A_k contains n different elements. This means that the corresponding set of angles, $\frac{2k+1}{2n+1}l\pi$, are also distinct. However, we have to show that taking the sine of these angles will still result in n distinct real values. Therefore, we have to check when the sine function maps different arguments to the same value up to a sign change. We have already taken care of the case $\{\theta, \theta + \pi\}$, via the isomorphism with $\mathbb{Z}_{2n+1}$, and what remains are the cases $\{\theta, -\theta\}$ and $\{\pi - \theta, \pi + \theta\}$.

With the number $(2k+1)l$ in $\mathbb{Z}_{2n+1}$ corresponds the angle

$$\theta_{k,l} = \frac{(2k+1)l}{2n+1}\pi.$$

Then $-\theta_{k,l}$ (which gives the same absolute value for the sine as $\theta_{k,l}$) corresponds to $-(2k+1)l \mod 2n+1$. Suppose that

$$(2k+1)l_1 = -(2k+1)l_2 \mod (2n+1).$$

for some l_1 and l_2. Then $(2k+1)(l_1 + l_2) = 0 \mod 2n+1$, and $l_1 = -l_2 \mod 2n+1$, as $2k+1$ is not a divisor of zero by assumption. Since $l_1, l_2 \in \{1, \ldots, n\}$, we have a contradiction.

Turning to the case $\{\pi-\theta, \pi+\theta\}$, we have that $[2n+1-(2k+1)l] \in \mathbb{Z}_{2n+1}$ corresponds to

$$\pi - \theta_{k,l} = \frac{[2n+1-(2k+1)l]}{2n+1}\pi,$$

and similarly, $\pi + \theta_{k,l}$ is related to $[2n+1+(2k+1)l] \mod 2n+1$. Assuming that

$$2n+1-(2k+1)l_1 = 2n+1+(2k+1)l_2 \mod (2n+1)$$

for some l_1 and l_2, we apply the same argument as before to arrive at a contradiction.

What we have shown up to now is that for each fixed k, the n real numbers $|\sin \frac{2k+1}{2n+1} l\pi|$ $(1 \leq l \leq n)$ are distinct. However, we want to show that these n real numbers are the same set for every k. The following proposition will aid in establishing this fact.

Proposition 4.3.2. *Suppose $2k+1$ is not a divisor of zero in $\mathbb{Z}_{2n+1}$. Then the set $A_k = \{(2k+1)l \mid l = 1, 2, \ldots, n\}$ contains exactly one element of every pair of the form $\{m, (2n+1)-m\}$, where $1 \leq m \leq n$.*

Proof. Suppose that for some m and some $l_1, l_2 \in \{1, 2, \ldots, n\}$, we have

$$\begin{aligned}(2k+1)l_1 &= m \mod (2n+1),\\ (2k+1)l_2 &= ((2n+1)-m) \mod (2n+1).\end{aligned}$$

Then

$$(2k+1)(l_1+l_2) = (2n+1) \mod (2n+1) = 0 \mod (2n+1),$$

which again contradicts the fact that $2k+1$ is not a divisor of zero, unless $l_1 + l_2 = 0 \mod (2n+1)$. However, since $2 \leq l_1 + l_2 \leq 2n$, this is impossible. □This proposition tells us that the angles $\frac{2k+1}{2n+1} l\pi$ (for any k), corresponding to $(2k+1)l \in \mathbb{Z}_{2n+1}$, can only take on one of $2n$ values since that is the number of elements in $\{m, (2n+1)-m\}$, where $1 \leq m \leq n$. If we consider the absolute value of the sine with the angle corresponding to the elements $m, (2n+1)-m \in \mathbb{Z}_{2n+1}$, we have

$$\left|\sin \frac{2n+1-m}{2n+1}\pi\right| = \left|\sin(\pi - \frac{m}{2n+1}\pi)\right| = \left|\sin \frac{m}{2n+1}\pi\right|,$$

and hence we see that there is in fact only n values that $|\sin \frac{2k+1}{2n+1} l\pi|$ with $0 \leq k \leq n-1$ and $1 \leq l \leq n$ can achieve.

Corollary 4.3.3. *If $2k+1$ is not a divisor of zero in $\mathbb{Z}_{2n+1}$, then the entries in the k-th column of U are a permutation of the numbers*

$$\frac{\sin \frac{l}{2n+1}\pi}{\sqrt{\sum_{j=1}^{n} \sin \frac{j}{2n+1}\pi}}, \qquad l = 1, 2, \cdots, n,$$

up to some signs.

In the case that $2n+1$ is prime, there are no divisors of zero in $\mathbb{Z}_{2n+1}$, and all columns of U display this phenomenon.

If $2n+1$ is not prime, there will always be odd integer zero divisors since $2n+1$ is odd. Now, $\{2k+1 \mid 0 \leq k \leq n-1\}$ accounts for all the odd elements of $\mathbb{Z}_{2n+1}$ and thus, when $2n+1$ is not prime, there will always be a zero entry in at least one of the singular vectors of $T_n(f_n)$.

4.4 Computing $T_n^{-1}(\frac{1}{cn})$, K_n^{-1} and $\det(K_n^{\pm 1})$

In this section we apply our main result to calculate the inverses of $T_n(\frac{1}{cn})$ and K_n and the determinants of $K_n^{\pm 1}$. We note that it is possible to explicitly calculate these inverses via the Gohberg-Semencul formulas as well, and these results are given in the Appendix of this chapter.

Remember that

$$K_n = \begin{bmatrix} 2+\frac{1}{m} & -1 & 0 & \cdots & \cdots & 0 \\ 1+\frac{1}{m} & 1+\frac{1}{m} & -1 & 0 & & \vdots \\ \vdots & \frac{1}{m} & & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & -1 & 0 \\ \vdots & \vdots & \ddots & \ddots & & -1 \\ 1+\frac{1}{m} & \frac{1}{m} & \cdots & \cdots & \frac{1}{m} & 1+\frac{1}{m} \end{bmatrix}, \tag{4.6}$$

where $m = [cn]$, for some constant c as defined in the previous chapter.

Let $T_{0,n} = UDV^*$ be a SVD. As calculated before, $T_n(\frac{1}{cn}) = U(D + f_n D^{-1})V^*$, and hence, $T_n^{-1}(\frac{1}{cn}) = U(D + \frac{1}{cn}D^{-1})^{-1}V^*$. Now we know U, D and V from Theorem 4.1.1 and Remark 4.2.5, and therefore

$$\begin{aligned} & T_n^{-1}(\frac{1}{cn}) \\ &= \begin{bmatrix} u_1 \sin n\theta_1 & \dots & u_n \sin n\theta_n \\ \vdots & & \vdots \\ u_1 \sin \theta_1 & \dots & u_n \sin \theta_n \end{bmatrix} \begin{bmatrix} \frac{cn\lambda_1+1}{cn\sqrt{\lambda_1}} & & \\ & \ddots & \\ & & \frac{cn\lambda_n+1}{cn\sqrt{\lambda_n}} \end{bmatrix}^{-1} \begin{bmatrix} u_1 \sin \theta_1 & \dots & u_1 \sin n\theta_1 \\ \vdots & & \vdots \\ u_n \sin \theta_n & \dots & u_n \sin n\theta_n \end{bmatrix} \\ &= \begin{bmatrix} u_1 \sin n\theta_1 & \dots & u_n \sin n\theta_n \\ \vdots & & \vdots \\ u_1 \sin \theta_1 & \dots & u_n \sin \theta_n \end{bmatrix} \begin{bmatrix} \frac{cn\sqrt{\lambda_1}}{cn\lambda_1+1} & & \\ & \ddots & \\ & & \frac{cn\sqrt{\lambda_n}}{cn\lambda_n+1} \end{bmatrix} \begin{bmatrix} u_1 \sin \theta_1 & \dots & u_1 \sin n\theta_1 \\ \vdots & & \vdots \\ u_n \sin \theta_n & \dots & u_n \sin n\theta_n \end{bmatrix}, \end{aligned}$$

where $u_i = \frac{1}{\sqrt{\sum_j^n \sin^2 j\theta_i}}$, $(i = 1, \dots, n)$ are the normalization constants and $\sqrt{\lambda_i}$ are the singular values of $T_{0,n}$.

To calculate K_n^{-1}, we need a special expression for it. Proposition 3.2.2 and the paragraph following it shows that

$$K_n^{-1} = E_n^{-1/2}(A_n + (A_n^{-1})^*)^{-1}, \tag{4.7}$$

where $E_n = \sum_{j=1}^n (\frac{1}{cn} + w_j) e_j e_j^*$ ($w_1 = 1$, all other $w_j = 0$) and $A_n = T_{0,n}E_n^{-1/2}$.

Since $E_n^{-1/2}$ is diagonal and easily computible, we consider $A_n + (A_n^{-1})^*$ and decompose it as follows:

$$\begin{aligned} A_n + (A_n^{-1})^* &= T_{0,n}E_n^{-1/2} + (T_{0,n}^{-1})^* E_n^{1/2} \\ &= (T_{0,n}^{-1})^*(T_{0,n}^* T_{0,n} E_n^{-1/2} + E_n^{1/2}) \\ &= (T_{0,n}^{-1})^*(T_{0,n}^* T_{0,n} + E_n) E_n^{-1/2}. \end{aligned}$$

Notice that

$$T_{0,n}^* T_{0,n} + E_n = \begin{bmatrix} 2+\frac{1}{cn} & -1 & & & \\ -1 & 2+\frac{1}{cn} & -1 & & \\ & & \ddots & & \\ & & & \ddots & -1 \\ & & & -1 & 2+\frac{1}{cn} \end{bmatrix}. \tag{4.8}$$

We know this matrix is positive definite and since we know all its eigenvalues and eigenvectors (Theorem 4.2.1), we can decompose it via the SVD, say $T_{0,n}^* T_{0,n} + E_n = W\Sigma W^*$. In this case, the columns of W are the eigenvectors of $A_n + (A_n^{-1})^*$ with corresponding eigenvalues the entries of Σ. Consequently,

$$\begin{aligned}
&K_n^{-1} \\
=&E_n^{-1/2}(W\Sigma W^*)^{-1} \\
=&E_n^{-1/2} W \Sigma^{-1} W^* \\
=&\begin{bmatrix} \sqrt{\frac{cn}{cn+1}} & & & \\ & \sqrt{cn} & & \\ & & \ddots & \\ & & & \sqrt{cn} \end{bmatrix} \begin{bmatrix} w_1 \sin\frac{\pi}{n+1} & \dots & \dots & w_n \sin\frac{\pi n}{n+1} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ w_1 \sin\frac{n\pi}{n+1} & \dots & \dots & w_n \sin\frac{n\pi n}{n+1} \end{bmatrix} \\
&\begin{bmatrix} 2+\frac{1}{cn}+2\cos\frac{\pi}{n+1} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 2+\frac{1}{cn}+2\cos\frac{\pi n}{n+1} \end{bmatrix}^{-1} \begin{bmatrix} w_1 \sin\frac{\pi}{n+1} & \dots & \dots & w_1 \sin\frac{n\pi}{n+1} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ w_n \sin\frac{\pi n}{n+1} & \dots & \dots & w_n \sin\frac{n\pi n}{n+1} \end{bmatrix} \\
=&\begin{bmatrix} \sqrt{\frac{cn}{cn+1}} & & & \\ & \sqrt{cn} & & \\ & & \ddots & \\ & & & \sqrt{cn} \end{bmatrix} \begin{bmatrix} w_1 \sin\frac{\pi}{n+1} & \dots & \dots & w_n \sin\frac{\pi n}{n+1} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ w_1 \sin\frac{n\pi}{n+1} & \dots & \dots & w_n \sin\frac{n\pi n}{n+1} \end{bmatrix} \\
&\begin{bmatrix} \frac{2cn(1+\cos\frac{\pi}{n+1})+1}{cn} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \frac{2cn(1+\cos\frac{\pi n}{n+1})+1}{cn} \end{bmatrix} \begin{bmatrix} w_1 \sin\frac{\pi}{n+1} & \dots & \dots & w_1 \sin\frac{n\pi}{n+1} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ w_n \sin\frac{\pi n}{n+1} & \dots & \dots & w_n \sin\frac{n\pi n}{n+1} \end{bmatrix},
\end{aligned}$$

where $w_i = \frac{1}{\sqrt{\sum_j^n \sin^2 \frac{j\pi i}{n+1}}}$, $(i = 1, \dots, n)$ are the normalization constants.

From the previous paragraph's expression for K_n^{-1}, and using Theorem 4.2.2, it follows

that

$$\begin{aligned}
\det K_n &= \det(T_{0,n}^{-1})^* \det(T_{0,n}^* T_{0,n} + E_n) \\
&= \det \begin{bmatrix} 1 & & & & \\ 1 & 1 & & & \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ 1 & \dots & \dots & \dots & 1 \end{bmatrix} \det \begin{bmatrix} 2+\frac{1}{cn} & -1 & & & & \\ -1 & 2+\frac{1}{cn} & -1 & & & \\ & & \ddots & & & \\ & & & \ddots & -1 & \\ & & & -1 & 2+\frac{1}{cn} \end{bmatrix} \\
&= \det \begin{bmatrix} 2+\frac{1}{cn} & -1 & & & \\ -1 & 2+\frac{1}{cn} & -1 & & \\ & & \ddots & & \\ & & & \ddots & -1 \\ & & & -1 & 2+\frac{1}{cn} \end{bmatrix} \\
&= \frac{q_2^{n+1} - q_1^{n+1}}{q_2 - q_1}
\end{aligned}$$

for n large enough and where q_1 and q_2 are the roots of the equation $q^2 - (2 + \frac{1}{cn})q + 1$. Obviously, $\det K_n^{-1} = \frac{1}{\det K_n}$.

4.4.1 Computing $\|T_n(\frac{1}{n})^{-1}\|$

Another consequence of having explicit formulas for the singular values of $T_{0,n}$, is that it enables us to exactly determine $\|T_n(\frac{1}{n})^{-1}\|$. We will present a simple algorithm that will produce $\|T_n(\frac{1}{n})^{-1}\|$ for any n.

Similar to the previous chapter, we have the following equation relating the eigenvalues of $(X_n^{-1})(X_n^{-1})^*$ (or singular values of X_n^{-1}) to the eigenvalues of $T_n^* T_n$ (or singular values of $T_{0,n}$):

$$\lambda[(X_n^{-1})(X_n^{-1})^*] = \frac{1}{h_n[\lambda(T_n^* T_n)]}, \tag{4.9}$$

where $\frac{1}{h_n(x)} = \frac{x}{(f_n+x)^2}$ which has a maximum at $x = f_n$. In this particular case, we have $T_n = T_{0,n}$, $f_n = 1/n$ and $X_n = T_n(\frac{1}{n})$. Thus we see that $\|T_n(\frac{1}{n})\|$ will be realized for the singular value of $T_{0,n}$ that is the 'closest' to $1/\sqrt{n}$. Since $\frac{1}{h_n(x)}$ is not symmetric around its maximum, the closest singular values of $T_{0,n}$ to the left and right of the maximum has to be tested. To be precise,

$$\|T_n(\frac{1}{n})^{-1}\| = \frac{1}{\sqrt{h_n[(\sigma_{0,n}^{\max})^2]}},$$

where $\sigma_{0,n}^{\max}$ is the singular value that maximizes the function $\frac{1}{h_n(x)}$.

Recall from Theorem 4.1.1 that the singular values of $T_{0,n}$ are given by

$$\sigma_m = \sqrt{2 - 2\cos\theta_m}$$

where

$$\theta_m = (\frac{2m+1}{2n+1})\pi, \quad 0 \leq m \leq n-1.$$

We want to find j such that σ_j is close to $1/\sqrt{n}$. To do this, we rewrite σ_j:

$$\sigma_j^2 = 2 - 2\cos\theta_j = 4\sin^2\frac{\theta_j}{2},$$

via a trigonometric identity. This implies that we want to find j such that

$$\sin\frac{\theta_j}{2} \approx \frac{1}{2\sqrt{n}}.$$

Since we are looking at small values of sin, we can attempt to approximate $\sin\frac{\theta_j}{2}$ with $\frac{\theta_j}{2}$. In other words, we are trying to find j such that

$$\frac{\theta_j}{2} \approx \frac{1}{2\sqrt{n}}$$

is satisfied. We have an explicit expression for θ_j and hence

$$\theta_j = \left(\frac{2j+1}{2n+1}\right)\pi \approx \frac{1}{\sqrt{n}} \Longrightarrow j \approx \frac{2n+1}{2\pi\sqrt{n}} - \frac{1}{2}.$$

We now have a very simple way to determine which singular value will produce the norm of $T_n(\frac{1}{n})^{-1}$, but we still have to show that the approximation of $\sin\frac{\theta_j}{2}$ by $\frac{\theta_j}{2}$ that we used are good enough so as not to provide us with the wrong integer j. To do this we will bound their difference.

We know

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots.$$

Since we are looking at values of $\sin\frac{\theta_j}{2}$ close to $1/2\sqrt{n}$, we can assume that $\frac{\theta_j}{2} < 1$ which implies that

$$\sin\frac{\theta_j}{2} > \frac{\theta_j}{2} - \frac{(\frac{\theta_j}{2})^3}{3!}$$

and thus our error,

$$\frac{\theta_j}{2} - \sin\frac{\theta_j}{2} < \frac{(\frac{\theta_j}{2})^3}{3!} = \frac{\theta_j^3}{48}. \qquad \text{(similar for } j+1\text{)}$$

Remember that we have an expression for the choice of j, $j \approx \frac{2n+1}{2\pi\sqrt{n}} - \frac{1}{2}$. We can insert this expression into the upper bound for our error. Actually we will use $\frac{2n+1}{2\pi\sqrt{n}}$ instead, since j is an integer that can be rounded up by at most $1/2$. Now, for $n \geq 2$, $j+1 \leq \frac{2n+1}{2\pi\sqrt{n}} + 1$,

implying that

$$\begin{aligned}
\frac{\theta_{j+1}^3}{48} &\leq \left(\frac{2(\frac{2n+1}{2\pi\sqrt{n}}+1)+1}{2n+1}\right)^3 \frac{\pi^3}{48} \\
&= \left[\frac{2n+1+3\pi\sqrt{n}}{\sqrt{n}(2n+1)}\right]^3 \frac{1}{48} \\
&= \left[\frac{2+\frac{1}{n}+3\frac{\pi}{\sqrt{n}}}{2\sqrt{n}+\frac{1}{\sqrt{n}}}\right]^3 \frac{1}{48} \qquad \left[\left(2+\frac{1}{n}+3\frac{\pi}{\sqrt{n}}\right)^3 < 770\right] \\
&< \frac{770}{48}\frac{1}{(2\sqrt{n}+\frac{1}{\sqrt{n}})^3} \\
&= \frac{770}{48}\frac{1}{[\frac{1}{\sqrt{n}}(2n+1)]^3} \\
&= \frac{770}{48}\frac{1}{\frac{(2n+1)(2n+1)^3}{n\sqrt{n}}} \qquad \left[\frac{(2n+1)^2}{n\sqrt{n}} > 8\right] \\
&< \frac{770}{48\cdot 8}\frac{1}{2n+1} < \frac{\pi}{2n+1}.
\end{aligned}$$

This error bound is significant since $\frac{\theta_{j+1}}{2} - \frac{\theta_j}{2} = \frac{\pi}{2n+1}\ \forall j$, and we have just shown that

$$\frac{\theta_{j+1}}{2} - \sin\frac{\theta_{j+1}}{2} < \frac{\pi}{2n+1}.$$

This leads to the inequality

$$\frac{\theta_j}{2} < \sin\frac{\theta_{j+1}}{2} < \frac{\theta_{j+1}}{2}.$$

Let us assume that we calculated j such that

$$\frac{\theta_{j-1}}{2} < \frac{1}{2\sqrt{n}} < \frac{\theta_j}{2}.$$

Now there are two possibilities. Either

$$\sin\frac{\theta_{j-1}}{2} < \frac{1}{2\sqrt{n}} < \sin\frac{\theta_j}{2}$$

or

$$\sin\frac{\theta_j}{2} < \frac{1}{2\sqrt{n}} < \sin\frac{\theta_{j+1}}{2}$$

since $\frac{\theta_j}{2} < \sin\frac{\theta_{j+1}}{2} < \frac{\theta_{j+1}}{2}$. Therefore, we have to test the singular values of $T_{0,n}$ corresponding to $j-1$, j and $j+1$ in equation 4.9 to find $\|T_n(\frac{1}{n})^{-1}\|$. We note that it might happen that $\frac{\theta_j}{2} = \frac{1}{2\sqrt{n}}$, but this situation does not produce more singular values to test.

4.4.2 Computing $\|T_n(\frac{1}{n})\|$

In the previous chapter, section 3.6, we saw that for n large enough, $\|T_n(\frac{1}{n})\|$ is determined by the maximum singular value σ_1 (for which we have explicit formulas) of the generator matrix $T_{0,n}$. We can calculate $\|T_n(\frac{1}{n})\|$ by inserting σ_1 into the equation $g_n(x) = x+1/nx$, since if $T_{0,n} = UDV^*$ is a SVD, it follows that $T_n(\frac{1}{n}) = U(D + \frac{1}{n}D^{-1})V^*$ where D is the matrix consisting of the singular values of $T_{0,n}$. Then, to compute $\|T_n(\frac{1}{n})\|$ practically, we just need to find the n from which $\|T_n(\frac{1}{n})\|$ is determined by σ_1.

In the previous subsection, we showed that the singular values of $T_{0,n}$ can be written as $\sigma_j = 2\sin(\frac{\theta_j}{2})$, where $\theta_j = \frac{2j+1}{2n+1}\pi$ and $0 \le j \le n-1$. Recall from section 3.6, that the derivative of g_n is $1 - f_n\frac{1}{x^2}$, and is monotone decreasing on the interval $(0, \sqrt{f_n}]$, and monotone increasing on $[\sqrt{f_n}, \infty)$. It follows that $\|X_n\|$ attains its maximum value at either the minimum, or maximum singular value of T_n. Therefore, we have to show that $g_n(\sigma_n) \le g_n(\sigma_1)$ for all n:

$$\begin{aligned}
&g_n(\sigma_1) - g_n(\sigma_n) \\
&= 2\sin\left(\frac{\sigma_1}{2}\right) + \frac{1}{2n\sin(\frac{\sigma_1}{2})} - 2\sin\left(\frac{\sigma_n}{2}\right) + \frac{1}{2n\sin(\frac{\sigma_n}{2})} \\
&= \frac{4n\sin^2(\frac{\sigma_1}{2})\sin(\frac{\sigma_n}{2}) + \sin(\frac{\sigma_n}{2}) - 4n\sin(\frac{\sigma_1}{2})\sin^2(\frac{\sigma_n}{2}) - \sin(\frac{\sigma_1}{2})}{2n\sin(\frac{\sigma_1}{2})\sin(\frac{\sigma_n}{2})} \\
&= \frac{4n\sin(\frac{\sigma_1}{2})\sin(\frac{\sigma_n}{2})[\sin(\frac{\sigma_1}{2}) - \sin(\frac{\sigma_n}{2})] - [\sin(\frac{\sigma_1}{2}) - \sin(\frac{\sigma_n}{2})]}{2n\sin(\frac{\sigma_1}{2})\sin(\frac{\sigma_n}{2})}.
\end{aligned}$$

Certainly $\sin(\frac{\sigma_1}{2}) - \sin(\frac{\sigma_n}{2}) > 0$, which implies that we only need to prove that

$$4n\sin(\frac{\sigma_1}{2})\sin(\frac{\sigma_n}{2}) > 1$$

or

$$\sin(\frac{\sigma_1}{2})\sin(\frac{\sigma_n}{2}) > \frac{1}{4n}.$$

Notice that

$$\sin(\frac{\sigma_1}{2})\sin(\frac{\sigma_n}{2}) \ge \frac{1}{2}\sin(\frac{\sigma_n}{2})$$

for all all $n \ge 2$ and then we check if

$$\sin(\frac{\sigma_n}{2}) > \frac{1}{2n}.$$

We know that $\sin x \ge x - \frac{x^3}{3!}$ and thus,

$$\begin{aligned}
\sin(\frac{\sigma_n}{2}) - \frac{1}{2n} &\ge \frac{\sigma_n}{2} - \frac{\frac{\sigma_n^3}{2}}{6} - \frac{1}{2n} \\
&= \frac{\pi}{4n+2} - \frac{\pi^3}{6(4n+2)^3} - \frac{1}{2n} \\
&= \frac{n\pi - (2n+1)}{n(4n+2)} - \frac{\pi^3}{6(4n+2)^3} \\
&> 0
\end{aligned}$$

for all $n \geq 2$. Hence,

$$\left\| T_n\left(\frac{1}{n}\right) \right\| = 2\sin\left(\frac{2n-1}{4n+2}\pi\right) + \frac{1}{2n\sin\left(\frac{2n-1}{4n+2}\pi\right)}, \quad n \geq 2.$$

4.5 Appendix

4.5.1 Explicit formula for $T_n(\frac{1}{n})^{-1}$

In general explicit formulas for the inverse of a Toeplitz matrix may be found using the Gohberg-Semencul formulas or variants of it, see [20]. Usually there are a few related matrix-vector equations that have to be solved to find some vectors whose entries appear in the inverted matrix, be it either the first and last column and the first and last row of the inverse, or some other columns or rows. Fortunately, for the matrix $T_n(1/n)^{-1}$ it is possible to achieve an explicit expression for all n, although slightly complicated. We shall rely on a particular variant of the Gohberg-Semencul formulas for this purpose.

Our main result for the inverse of $T_n(1/n)$ is the following.

Theorem. *Let the numbers D_j be defined recursively by $D_1 = 1, D_2 = 2 + \frac{1}{n}$, and*

$$D_j = (2 + \frac{1}{n})D_{j-1} - D_{j-2}. \tag{4.10}$$

With these numbers, introduce the vector

$$x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{bmatrix} = \frac{1}{(1+\frac{1}{n})D_n - D_{n-1}} \begin{bmatrix} D_n - D_{n-1} \\ -\frac{1}{n}D_{n-1} \\ \vdots \\ -\frac{1}{n}D_1 \end{bmatrix}, \tag{4.11}$$

and the vector

$$y = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_{n-1} \end{bmatrix} = \frac{1}{(1+\frac{1}{n})D_n - D_{n-1}} \begin{bmatrix} D_{n-1} - D_{n-2} \\ (1+\frac{1}{n})(D_{n-1} - D_{n-2}) \\ -\frac{1}{n}(2+\frac{1}{n})D_{n-2} \\ \vdots \\ -\frac{1}{n}(2+\frac{1}{n})D_1 \end{bmatrix}. \tag{4.12}$$

Then

$$\begin{aligned} T_n(\frac{1}{n})^{-1} = \frac{1}{x_{n-1}}\{[x_i x_{n-j-1}]_{i,j=0}^{n-1} &+ \begin{bmatrix} y_0 & & & \\ y_1 & y_0 & & \\ \vdots & & \ddots & \\ y_{n-1} & \cdots & \cdots & y_0 \end{bmatrix} \begin{bmatrix} 0 & x_{n-1} & \cdots & x_1 \\ & \ddots & \ddots & \vdots \\ & & \ddots & x_{n-1} \\ & & & 0 \end{bmatrix} \\ &- \begin{bmatrix} x_0 & & & \\ x_1 & x_0 & & \\ \vdots & & \ddots & \\ x_{n-1} & \cdots & \cdots & x_0 \end{bmatrix} \begin{bmatrix} 0 & y_{n-1} & \cdots & y_1 \\ & \ddots & \ddots & \vdots \\ & & \ddots & y_{n-1} \\ & & & 0 \end{bmatrix}\}. \end{aligned} \tag{4.13}$$

Proof. Suppose that for a general square Toeplitz matrix T_n of size $n \times n$, the following equations are solvable:

$$T_n x = e_0 \qquad T_n y = e_1,$$

where e_i for $0 \leq i \leq n-1$ is the unit vector with one in its i-th position. Then T_n is invertible and its inverse is given by (4.13) (see [20]).

From the above equation it is clear that one needs to compute the vectors x and y to arrive at an explicit formula for $T_n(1/n)^{-1}$. We will first rewrite $x = T_n(1/n)^{-1}e_0$ and $y = T_n(1/n)^{-1}e_1$ before solving them. Starting with x we have

$$\begin{aligned} T_n(\frac{1}{n}) &= T_{0,n} + \frac{1}{n}(T_{0,n}^{-1})^* \\ &= (T_{0,n}^*)^{-1}(T_{0,n}^* T_{0,n} + \frac{1}{n}I) \\ \implies x = T_n(\frac{1}{n})^{-1}e_0 &= (T_{0,n}^* T_{0,n} + \frac{1}{n}I)^{-1} T_{0,n}^* e_0. \end{aligned}$$

Define

$$\tilde{T}_n := T_{0,n}^* T_{0,n} + \frac{1}{n}I = \begin{bmatrix} 1+\frac{1}{n} & -1 & & \\ -1 & 2+\frac{1}{n} & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2+\frac{1}{n} \end{bmatrix}$$

and notice that $T_{0,n}^* e_0 = e_0 - e_1$.

Putting these together we arrive at

$$\tilde{T}_n x = e_0 - e_1 \qquad \implies \qquad x = \tilde{T}_n^{-1} e_0 - \tilde{T}_n^{-1} e_1.$$

We now solve the equations $\tilde{T}_n u = e_0$ and $\tilde{T}_n v = e_1$ individually via Cramer's rule. First, a few more definitions. Let

$$D = \det \tilde{T}_n, \qquad D_1 = 1, \qquad D_2 = 2 + \frac{1}{n}, \qquad D_j = \det \begin{bmatrix} 2+\frac{1}{n} & -1 & & \\ -1 & 2+\frac{1}{n} & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2+\frac{1}{n} \end{bmatrix},$$

where D_j is a square matrix of size $j-1 \times j-1$. Applying Cramer's rule to u_0 we see that

$$u_0 = \frac{1}{D} \det \begin{bmatrix} 1 & -1 & & & \\ 0 & 2+\frac{1}{n} & \ddots & & \\ \vdots & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2+\frac{1}{n} \end{bmatrix} = \det \begin{bmatrix} 2+\frac{1}{n} & -1 & & \\ -1 & 2+\frac{1}{n} & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2+\frac{1}{n} \end{bmatrix} = \frac{D_n}{D}.$$

Similarly, we find that $u_1 = \frac{D_{n-1}}{D}$ and in general $u_j = \frac{D_{n-j}}{D}$.

Now for v we have the following:

$$v_0 = \frac{1}{D}\det\begin{bmatrix} 0 & -1 & & & \\ 1 & 2+\frac{1}{n} & \ddots & & \\ 0 & -1 & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & -1 \\ & & & -1 & 2+\frac{1}{n} \end{bmatrix} = -1.\det\begin{bmatrix} -1 & & & & \\ -1 & 2+\frac{1}{n} & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2+\frac{1}{n} \end{bmatrix}$$
$$= \frac{D_{n-1}}{D},$$

$$v_1 = \frac{1}{D}\det\begin{bmatrix} 1+\frac{1}{n} & 0 & & & & \\ -1 & 1 & -1 & & & \\ & 0 & 2+\frac{1}{n} & \ddots & & \\ & \vdots & -1 & \ddots & \ddots & \\ & & & \ddots & \ddots & -1 \\ & & & & -1 & 2+\frac{1}{n} \end{bmatrix} = (1+\frac{1}{n})\frac{D_{n-1}}{D}$$

and in general we get that $v_j = (1+\frac{1}{n})\frac{D_{n-j}}{D}$ for $1 < j \leq n-1$. We can also determine D in terms of the D_j's and it turns out that $D = (1+\frac{1}{n})D_n - D_{n-1}$.

Therefore we can write

$$x = T_n(\frac{1}{n})^{-1}e_0 = \tilde{T}_n^{-1}e_0 - \tilde{T}_n^{-1}e_1$$
$$= \frac{1}{(1+\frac{1}{n})D_n - D_{n-1}}\begin{bmatrix} D_n \\ D_{n-1} \\ \vdots \\ D_1 \end{bmatrix} - \frac{1}{(1+\frac{1}{n})D_n - D_{n-1}}\begin{bmatrix} D_{n-1} \\ (1+\frac{1}{n})D_{n-1} \\ \vdots \\ (1+\frac{1}{n})D_1 \end{bmatrix}$$
$$= \frac{1}{(1+\frac{1}{n})D_n - D_{n-1}}\begin{bmatrix} D_n - D_{n-1} \\ -\frac{1}{n}D_{n-1} \\ \vdots \\ -\frac{1}{n}D_1 \end{bmatrix}.$$

As for y, we have the following equation

$$y = T_n(\frac{1}{n})^{-1}e_1 = (T_{0,n}^*T_{0,n} + \frac{1}{n}I)^{-1}T_{0,n}^*e_1 = (T_{0,n}^*T_{0,n} + \frac{1}{n}I)^{-1}(e_1 - e_2).$$

Notice that we only have to solve the equation $\tilde{T}_n s = e_2$ since the other has already been solved during the calculation of x. Following the same procedure as before,

$$
\begin{aligned}
s_0 &= \frac{D_{n-2}}{D} \\
s_1 &= (1+\frac{1}{n})\frac{D_{n-2}}{D} \\
s_2 &= \frac{1}{D}[(1+\frac{1}{n})(2+\frac{1}{n})D_{n-2} - D_{n-2}] \\
s_j &= \frac{1}{D}[(1+\frac{1}{n})(2+\frac{1}{n})D_{n-j} - D_{n-j}], \quad 2 < j \leq n-1.
\end{aligned}
$$

and

$$
\begin{aligned}
y &= T_n(\frac{1}{n})^{-1}e_1 \\
&= \tilde{T}_n^{-1}e_1 - \tilde{T}_n^{-1}e_2 \\
&= \frac{1}{(1+\frac{1}{n})D_n - D_{n-1}}\left\{ \begin{bmatrix} D_{n-1} \\ (1+\frac{1}{n})D_{n-1} \\ \vdots \\ (1+\frac{1}{n})D_1 \end{bmatrix} - \begin{bmatrix} D_{n-2} \\ (1+\frac{1}{n})D_{n-2} \\ (1+\frac{1}{n})(2+\frac{1}{n})D_{n-2} - D_{n-2} \\ \vdots \end{bmatrix} \right\} \\
&= \frac{1}{(1+\frac{1}{n})D_n - D_{n-1}} \begin{bmatrix} D_{n-1} - D_{n-2} \\ (1+\frac{1}{n})(D_{n-1} - D_{n-2}) \\ -\frac{1}{n}(2+\frac{1}{n})D_{n-2} \\ \vdots \\ -\frac{1}{n}(2+\frac{1}{n})D_1 \end{bmatrix}.
\end{aligned}
$$

□

We now turn our attention to the D_j's since $T_n(1/n)^{-1}$ is now completely defined in terms of them.

Proposition. *Introduce*

$$
\begin{aligned}
r_1 &= 1 + \tfrac{1}{2n} + \tfrac{1}{2}\sqrt{\tfrac{4}{n} + \tfrac{1}{n^2}}, & r_2 &= 1 + \tfrac{1}{2n} - \tfrac{1}{2}\sqrt{\tfrac{4}{n} + \tfrac{1}{n^2}}, \\
k_1 &= \tfrac{r_2 - 2 - \frac{1}{n}}{1 - r_1^2}, & k_2 &= \tfrac{2 + \frac{1}{n} - r_1}{r_2^2 - 1}.
\end{aligned}
$$

Then

$$
D_j = k_1 r_1^j + k_2 r_2^j. \tag{4.14}
$$

Proof. Recall that $D_j = (2+\frac{1}{n})D_{j-1} - D_{j-2}$. The corresponding generating function for this relation is $f(t) = t^2 - (2+\frac{1}{n})t + 1$, and consequently,

$$
D_j = k_1 r_1^j + k_2 r_2^j,
$$

where r_1 and r_2 are the roots of the generating polynomial and k_1 and k_2 constants to be determined by using D_1 and D_2 in the expression for D_j. The proposition now follows in a straightforward way. □

4.5.2 Explicit formula for K_n^{-1}

Moving on to K_n^{-1}, we have the following theorem.

Theorem. *For $K_n = T_n(\frac{1}{n}) + \hat{e}_0 e_0$ we have*

$$K_n^{-1} = T_n(\frac{1}{n})^{-1} - \frac{1}{\left((1+\frac{1}{n})D_n - D_{n-1}\right)\left((2+\frac{1}{n})D_n - D_{n-1}\right)}\hat{D}, \tag{4.15}$$

where

$$\hat{D} = \begin{bmatrix} D_n \\ D_{n-1} \\ \vdots \\ D_1 \end{bmatrix} \begin{bmatrix} D_n - D_{n-1} & \cdots & D_2 - D_1 & D_1 \end{bmatrix}$$

Proof. We have from $K_n = T_n(\frac{1}{n}) + \hat{e}_0 e_0$, using the Morrison-Woodbury formula that

$$K_n^{-1} = T_n(\frac{1}{n})^{-1} - \frac{T_n(\frac{1}{n})^{-1}\hat{e}_0 e_0^* T_n(\frac{1}{n})^{-1}}{1 + e_0^* T_n(\frac{1}{n})^{-1}\hat{e}_0}.$$

For the second term on the right, we just need to calculate $T_n(\frac{1}{n})^{-1}\hat{e}_0$, $e_0^* T_n(\frac{1}{n})^{-1}$ and the denominator $1 + e_0^* T_n(\frac{1}{n})^{-1}\hat{e}_0$, and then K_n^{-1} is also explicitly known in terms of the above D_j's.

We start by computing $T_n(\frac{1}{n})^{-1}\hat{e}_0$. Observe that $T_n(\frac{1}{n})e_0 = e_0 + \frac{1}{n}\hat{e}_0$. So, $T_n(\frac{1}{n})^{-1}\hat{e}_0 = n(e_0 - T_n(\frac{1}{n})^{-1}e_0) = n(e_0 - x)$. Now

$$e_0 - x = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \frac{1}{(1+\frac{1}{n})D_n - D_{n-1}} \begin{bmatrix} D_n - D_{n-1} \\ -\frac{1}{n}D_{n-1} \\ \vdots \\ -\frac{1}{n}D_1 \end{bmatrix} = \frac{1}{(1+\frac{1}{n})D_n - D_{n-1}} \frac{1}{n} \begin{bmatrix} D_n \\ D_{n-1} \\ \vdots \\ D_1 \end{bmatrix}.$$

Next, we compute

$$\begin{aligned} 1 + e_0^* T_n(\frac{1}{n})^{-1}\hat{e}_0 &= 1 + n(1 - x_0) = \\ &= 1 + n\left(1 - \frac{D_n - D_{n-1}}{(1+\frac{1}{n})D_n - D_{n-1}}\right) = \\ &= 1 + n\left(\frac{\frac{1}{n}D_n}{(1+\frac{1}{n})D_n - D_{n-1}}\right) = \\ &= 1 + \frac{D_n}{(1+\frac{1}{n})D_n - D_{n-1}} = \\ &= \frac{(2+\frac{1}{n})D_n - D_{n-1}}{(1+\frac{1}{n})D_n - D_{n-1}}. \end{aligned}$$

Finally, we compute $e_0^* T_n(\frac{1}{n})^{-1}$. A little rearrangement shows that

$$e_0^* T_n(\frac{1}{n})^{-1} = e_0^* T_{0,n}(T_{0,n}T_{0,n}^* + \frac{1}{n}I)^{-1} = e_0^*(T_{0,n}T_{0,n}^* + \frac{1}{n}I)^{-1} := e_0^*\hat{T}_n^{-1},$$

where

$$\hat{T}_n := T_{0,n}T_{0,n}^* + \frac{1}{n}I = \begin{bmatrix} 2+\frac{1}{n} & -1 & & \\ -1 & 2+\frac{1}{n} & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 1+\frac{1}{n} \end{bmatrix}.$$

If we let $w^* = e_0^* T_n(1/n)^{-1}$, then $w^*\hat{T}_n = e_0^* \quad \Rightarrow \quad \hat{T}_n w = e_0$. Applying Cramer's rule once again, we notice that we do not have our determinants in terms of the D_j's. Rather, we have them in terms of

$$\hat{D}_j := \det \begin{bmatrix} 2+\frac{1}{n} & -1 & & \\ -1 & 2+\frac{1}{n} & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 1+\frac{1}{n} \end{bmatrix}$$

having size $j-1 \times j-1$. However, if we compute $\hat{D}_j$ by expanding the determinant along the last row, we find that

$$\hat{D}_j = (1+\frac{1}{n})D_{j-1} - D_{j-2} = D_j - D_{j-1},$$

and finally

$$e_0^* T_n(\frac{1}{n})^{-1} = w^* = \frac{1}{(1+\frac{1}{n})D_n - D_{n-1}} \begin{bmatrix} D_n - D_{n-1} & \dots & D_2 - D_1 & D_1 \end{bmatrix}.$$

Taking everything together, we obtain

$$\begin{aligned} K_n^{-1} &= T_n(\frac{1}{n})^{-1} - \frac{(1+\frac{1}{n})D_n - D_{n-1}}{(2+\frac{1}{n})D_n - D_{n-1}} n(e_0 - x)w^* = \\ &= T_n(\frac{1}{n})^{-1} - \frac{n}{(2+\frac{1}{n})D_n - D_{n-1}}(e_0 - x)\begin{bmatrix} D_n - D_{n-1} & \cdots & D_2 - D_1 & D_1 \end{bmatrix} \\ &= T_n(\frac{1}{n})^{-1} - \frac{n}{(2+\frac{1}{n})D_n - D_{n-1}} \frac{1}{(1+\frac{1}{n})D_n - D_{n-1}} \frac{1}{n}\hat{D}, \end{aligned}$$

which finally is equal to (4.15). □

4.5.3 The characteristic polynomial of P_n

In our investigation into the permutation phenomenon of the previous chapter, we analysed the characteristic polynomial of P_n, since this might have lead to a better understanding of its eigenvalues, and hence, the singular values of $T_{0,n}$. This approach did not help much, except that this polynomial has an elegant formula which we found for any n.

Since $P_n = T_{0,n}(T_{0,n})^*$ is a tridiagonal matrix, we can find a recurrence relation in n for the determinant of $P_n - \lambda I$. Indeed, denote the characteristic polynomial of $P_n - \lambda I$ by $C_n(\lambda)$ and let $C_0(\lambda) = 1$. Then,

$$C_j(\lambda) = (2 - \lambda)C_{j-1}(\lambda) - C_{j-2}(\lambda), \tag{4.16}$$

We can also denote

$$C_n(\lambda) = \sum_{k=0}^{n} a_{nk}\lambda^k.$$

Then from (4.16) we have the following recursion relation between the coefficients a_{nk}:

$$a_{nk} = 2a_{n-1\,k} - a_{n-2\,k} - a_{n-1\,k-1}. \tag{4.17}$$

Listing the first few polynomials we have

$$\begin{aligned} C_1(\lambda) &= 1 - \lambda \\ C_2(\lambda) &= 1 - 3\lambda + \lambda 2 \\ C_3(\lambda) &= 1 - 6\lambda + 5\lambda 2 - \lambda 3 \\ C_4(\lambda) &= 1 - 10\lambda + 15\lambda^2 - 7\lambda^3 + \lambda^4. \end{aligned}$$

Comparing the coefficients with the numbers in Pascal's triangle, we are tempted to conjecture that the following holds:

Proposition. *For all n,*

$$a_{nk} = (-1)^k \binom{n+k}{2k}, \qquad k = 0, 1, 2, \ldots, n, \tag{4.18}$$

and hence

$$C_n(\lambda) = \sum_{k=0}^{n} (-1)^k \binom{n+k}{2k} \lambda^k. \tag{4.19}$$

Proof. The proof is by induction on n. We shall use several times the well-known rule for the construction of Pascal's triangle: for all n and $k = 0, 1, \ldots, n$ we have

$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}, \tag{4.20}$$

where n represents the number of the row (counting from zero), and k the k-th entrie from the left or right, counting from zero. The basis for the induction is checking that (4.18) is true for $n = 0$ and $n = 1$. For $n = 0$ the only option is $k = 0$ and $a_{00} = 1$ indeed. For $n = 1$ we have to check $k = 0$ and $k = 1$: for $k = 0$ we obtain $a_{1\,0} = 1$, for $k = 1$ we obtain $a_{1\,1} = -1$ for (4.18), which fits with $C_1(\lambda) = 1 - \lambda$.

For the induction step, fix k. The right hand side of (4.18) becomes, assuming that the result is correct for $j-1$ and $j-2$:

$$\begin{aligned}(-1)^k &\left\{2\binom{n+k-1}{2k} - \binom{n+k-2}{2k} + \binom{n+k-2}{2k-2}\right\} = \\ =(-1)^k &\left\{\binom{n+k-1}{2k} + \binom{n+k-1}{2k} - \binom{n+k-2}{2k} + \binom{n+k-2}{2k-2}\right\}.\end{aligned}$$

Applying (4.20) to the second and third term we see that this is equal to

$$(-1)^k \left\{\binom{n+k-1}{2k} + \binom{n+k-2}{2k-1} + \binom{n+k-2}{2k-2}\right\}.$$

Again applying (4.20), now to the last two terms, we see that this is equal to

$$(-1)^k \left\{\binom{n+k-1}{2k} + \binom{n+k-1}{2k-1}\right\}.$$

Once more applying (4.20) we see that this is equal to

$$(-1)^k \binom{n+k}{2k}$$

as desired. □

Now we use that $\lambda = 2 - 2\cos\theta = 4\sin^2(\frac{1}{2}\theta)$, to translate the polynomials $C_n(\lambda)$ into polynomials in $\sin(\frac{1}{2}\theta)$. We shall denote these trigonometric polynomials by $\widehat{C}_j(\theta)$, thus

$$\widehat{C}_n(\theta) = \sum_{k=0}^{n} (-1)^k \binom{n+k}{2k} 4^k \sin^{2k}(\frac{1}{2}\theta). \tag{4.21}$$

Bibliography

[1] F. Avram. On bilinear forms in Gaussian random variables and Toeplitz matrices. *Probab. Theory Related Fields* 79 (1988), 37–45.

[2] G. Baxter. A norm inequality for a finite-section Wiener-Hopf equation. *Illinois J. Math.* 7 (1963), 97–103.

[3] A. Böttcher, S. Grudsky. On the condition numbers of large semidefinite Toeplitz matrices. *Linear Algebra Appl.* 279 (1998) 285–301.

[4] A. Böttcher, S. Grudsky. *Toeplitz Matrices, Asymptotic Linear Algebra, and Functional Analysis*, Hindustan Book Agency, New Delhi, 2000, and Birkhäuser Verlag, Basel, 2000.

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

[6] A. Böttcher, S. Grudsky. Uniform boundedness of Toeplitz matrices with variable coefficients. *Integral Equations Operator Theory* 60 (2008), 313-328.

[7] A. Böttcher, S. Grudsky. *Variable-coefficient Toeplitz matrices with symbols beyond the Wiener algebra*. Numerical methods for structured matrices and applications, 191202, Oper. Theory Adv. Appl., 199, Birkhuser Verlag, Basel, 2010.

[8] A. Böttcher, S. Grudsky, A. Kozak, B. Silbermann. Norms of large Toeplitz band matrices, *SIAM J. Matrix Anal. Appl.* 21 (1999), 547–561.

[9] A. Böttcher, B. Silbermann. *Analysis of Toeplitz Operators*. Springer-Verlag, Berlin, 1990.

[10] A. Böttcher, B. Silbermann. *Introduction to Large Truncated Toeplitz Matrices*. Springer-Verlag, New York, 1999.

[11] T. Ehrhardt, B. Shao. Asymptotic behavior of variable-coefficient Toeplitz determinants. *J. Fourier Anal. Appl.* 7 (2001), no. 1, 71-92.

[12] A.E Frazho, M.A Kaashoek, A.C.M. Ran. The non-symmetric discrete algebraic Riccati equation and canonical factorization of rational matrix functions on the unit circle. *Integral Equations Operator Theory*, 66 (2010), 215-229.

[13] I.C. Gohberg, I.A. Feldman. *Convolution equations and projection methods for their solution*, Transl. Math. Monographs, 41, Amer. Math. Soc., Providence, R.I., 1974.

[14] I. Gohberg, S. Goldberg, M.A. Kaashoek. *Classes of Linear Operators Vol. I.* Operator Theory: Advances and Applications, Vol. 49. Birkhäuser Verlag, Basel, 1990.

[15] I. Gohberg, S. Goldberg, M.A. Kaashoek. *Classes of Linear Operators Vol. II.* Operator Theory: Advances and Applications, Vol. 63. Birkhäuser Verlag, Basel, 1993.

[16] I. Gohberg, S. Goldberg, M.A. Kaashoek. *Basic Classes of Linear Operators.* Birkhäuser Verlag, Basel, 2003.

[17] U. Grenander, G. Szegö. *Toeplitz forms and their applications.* University of California Press, Berkeley and Los Angeles, 1958.

[18] R.A. Horn, C.R. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.

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

[20] G. Heinig, K. Rost. *Algebraic methods for Toeplitz-like matrices and operators.* Operator Theory: Advances and Applications, 13. Birkhäuser Verlag, Basel, 1984.

[21] V. Ionescu, C. Oara, M. Weiss. *Generalized Riccati Theory and robust control. A Popov function approach.* John Wiley, Chichester, 1999.

[22] X. Jin. *Developments and applications of block Toeplitz iterative solvers.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.

[23] S. Kimitei. *Algorithms for Toeplitz matrices with applications to image deblurring.* Lap Lambert Academic Publishing GmbH KG, 2011.

[24] P. Lancaster and L. Rodman, *Algebraic Riccati equations*, ClarendonPress, Oxford, 1995.

[25] S. V. Parter. On the distribution of the singular values of Toeplitz matrices. *Linear Algebra Appl.* 80 (1986), 115–130.

[26] H. Rabe, A.C.M. Ran. Asymptotics of the smallest singular value of a class of Toeplitz-generated matrices and related finite rank perturbations. *Integral Equations Operator Theory* 77 (2013), 385-396.

[27] H. Rabe, A.C.M. Ran. Asymptotics of the smallest singular value of a class of Toeplitz-generated matrices II. *Integral Equations Operator Theory* 79 (2014), 243–253.

[28] H. Rabe, A.C.M. Ran. A peculiar permutation phenomenon arising from the singular vector entries of a special class of Toeplitz matrices. *Linear Algebra Appl.* 459 (2014), 368–383.

[29] V. Rabinovich, S. Roch, B. Silbermann. *Limit Operators and Their Applications in Operator Theory.* Operator Theory: Advances and Applications, Vol.150. Birkhäuser Verlag, Basel, 2004.

[30] E. Reich. On non-Hermitian Toeplitz matrices. *Math. Scand.* 10 (1962), 145–152.

[31] S. Roch, B. Silbermann. A note on singular values of Cauchy-Toeplitz matrices. *Linear Algebra Appl.* 275/276 (1998), 531–536.

[32] S. Roch, B. Silbermann. Index calculus for approximation methods and singular value decomposition. *J. Math. Anal. Appl.* 225 (1998), 401 – 426.

[33] S. Serra Capizzano. On the extreme spectral properties of Toeplitz matrices generated by L^1 functions with several minima/maxima. *BIT*, 36 (1996), 135–142.

[34] B. Shao. A trace formula for variable-coefficient Toeplitz matrices with symbols of bounded variation. *J. Math. Anal. Appl.* 222 (1998), no. 2, 505-546.

[35] B. Shao. A trace formula for a class of variable-coefficient block Toeplitz matrices. *Integral Equations Operator Theory* 45 (2003), 359-374.

[36] E. Shargorodsky. Toeplitz matrices with variable coefficients, pseudodifferential operators, and Strichartz's method. *Math. Nachr.* 283 (2010), no. 1, 126-138.

[37] S. Sniekers, A.W. van der Vaart. Private communication.

[38] G. Strang. The Discrete Cosine Transform. *SIAM Rev.* 41 (1999), 135–147.

[39] R.C. Thompson. Principal submatrices IX: Interlacing inequalities for singular values of submatrices. *Linear Algebra Appl.* 5 (1972) 1–12.

[40] C.A. Tracy. *Lectures on Differential Equations*, Davis, California, 2015, https://www.math.ucdavis.edu/ tracy/courses/math22B/22BBook.pdf

[41] P. Zizler, R.A. Zuidwijk, F.A Taylor, S. Arimoto. A finer aspect of eigenvalue distribution of selfadjoint banded Toeplitz matrices. *SIAM J. Matrix Anal. Appl.* 24 (2002), 59–67.