Singular Value Decay for Solutions of Sylvester Equations

by

Brock Klippenstein

A thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Mathematics
University of Manitoba
Winnipeg

Abstract

A Sylvester equation is an operator equation of the form $AX - XB = C$. A fact that has been proven multiple times before, see [37, 41], is that if C has low rank, then A and B satisfying certain conditions imply X has a low rank approximation. Another set of conditions was given by Beckermann and Townsend in 2019, [7], where they impose the conditions that A and B are both normal, and have disjoint and well-separated spectra. In this thesis, we explore cases where the normality condition can be relaxed. Our main tool is unitary operator dilations, whereby one can realize a given operator as the corner of a unitary operator acting on a bigger space. The basic problem becomes the lifting of the original Sylvester equation to a new one involving the unitary dilation. This is reminiscent of the so-called intertwining dilation theorem, but it requires a completely new analysis as we require additional conditions if we wish to be able to guarantee the solution has a low rank approximation.

Our main result states that if we trade in normality of A for a norm condition on A and B, then we can unitarily dilate A. This in turn allows us to conclude that X has a low rank approximation provided our condition is satisfied, B is normal, and C has low rank. Due to the similarity in the conditions for the theorem by Beckermann and Townsend and the conditions required to solve a Sylvester equation quickly using algorithms such as the alternating direction implicit (ADI) method, our dilation method also allows us to show the ADI method does not require too many iterations without requiring A to be normal.

Acknowledgments

I would like to start with thanking my supervisors Raphaël Clouâtre and Richard Mikaël Slevinsky for their guidance throughout this program. Additionally, it was not easy for them to choose a project from the intersection of both of their research interest, hence I am also very grateful for their brilliant choice of a project. Moreover, when I chose my program, I certainly had doubts, but I can now confidently say I made the right decision.

Next, I would like to thank the University of Manitoba along with the Faculty of Graduate Studies for all their immense financial support throughout this program.

List of Symbols

$\mathcal{B}(\mathcal{H})$	set of bounded linear operators on a Hilbert space $\mathcal{H}$
T^*	adjoint of an operator T
$\sigma(T)$	spectrum of an operator T
spr(T)	spectral radius of an operator T
$s_k(T)$	k^{th} singular value of an operator T
$\mathbb{C}$	set of complex numbers
$\mathbb{C}^{m\times n}$	set of all matrices of size $m \times n$
M_n	set of all complex square matrices of size $n \times n$
$\|T\|$	spectral norm of an operator T
$\|T\|_F$	Frobenius norm of a matrix T
$\|f\|_X$	norm of the continuous function f over the compact set X
$\mathcal{R}_k$	set of functions p/q for polynomials p and q with degree at most k
$Z_k(E,F)$	k^{th} Zolotarev number over sets E and F

Contents

List of Figures

1

Introduction

When solving certain differential equations numerically, such as the two-dimensional Poisson equation, see [19], one may start by discretizing the equation to obtain a matrix equation of the form

$$AX - XB = C \tag{1.1}$$

where X is a discretized approximation of the solution to the original equation and A, B and C depend on the structure of the differential equation. An important property here is that as the size of the matrices increase, we get a more accurate approximation of the solution. The equation 1.1 is called a Sylvester equation with solution X, coefficients A and B, and forcing C.

As shown in [25], Sylvester equations can also be used to find the connection coefficients between families of orthogonal polynomials. Additionally, certain classes of matrices such as Cauchy matrices solve very well-structured Sylvester equations with so-called low displacement rank, [7].

One property of matrices that can be very useful is having low rank, or more generally, having a low rank approximation. Clearly one reason for this is the desire to not have to store n^2 entries for a dense $n \times n$ matrix. Further, there are algorithms which have their computational complexity dependent on the rank of certain

matrices. For example, see [22] for how to quickly compute a low rank matrix-vector product.

Given the utility of low rank matrices, it is important to determine what conditions on $AX - XB = C$ guarantee X has a low rank approximation. See [37, 41] for two sets of conditions which guarantee this. However, the result we are most interested in here comes from Beckermann and Townsend in [7] where it was proven that if A and B are normal matrices[1], have disjoint and well-separated spectra, and the rank of C is low, then X has a low rank approximation. An important fact they use to prove their theorem is the Eckart-Young theorem, see Theorem 2.3, and how it implies that proving X has a low-rank approximation is equivalent to showing the singular values[2] of X decay quickly.

With this theorem in mind, we would like to explore if there is any way to relax the condition of normality. It is not hard to see that anything can happen to the singular values of X if A and B are both non normal. In fact, given any matrix B, any low rank matrix C, and an invertible matrix X, then assuming they each have appropriate sizes, by defining $A = (C + XB)X^{-1}$, we obtain a Sylvester equation $AX - XB = C$. Thanks to [37], we can even get a similar result except with much more structure. They proved the surprising result that given any set of singular values $\{s_k\}_{k=1}^n$, we can find some X which has these singular values and which solves $AX + XA^* = C$ where C has rank one.

See [25, Eq. 39] for a Sylvester equation $AX - XB = C$ where B is normal, and the rank of C is low. Although A is non normal, the singular values of X still have fast decay. A noteworthy property here is that $\|A\| \, \|B^{-1}\| < 1$, where $\|\cdot\|$ denotes

[1]If T is a operator on a Hilbert space, then it is normal if $T^*T = TT^*$ where T^* is the adjoint of T, defined in the usual sense.

[2]The singular values of an operator X are the square root of the elements in the spectrum of X^*X denoted by $\{s_k(X)\}_{k=1}^{\infty}$, listed in descending order. If X is an $n \times m$ matrix, then we will say $s_j(X) = 0$ if $j > \min\{n, m\}$.

the spectral norm:

$$\|T\| = \sup_{x \in \mathcal{H}} \frac{\|Tx\|}{\|x\|}$$

where T is a linear operator on some Hilbert space $\mathcal{H}$. This seems to indicate that a general result on singular value decay can be proven under these conditions. Indeed, we will prove this.

By using operator dilations, we will investigate the possibility of relaxing the conditions of the theorem by Beckermann and Townsend. Dilating an operator is the act of taking an operator and extending it to an operator that has some desired property, usually unitary, isometric, or normal. A major theme of dilation theory consists of attempting to dilate an algebraic equation in a manner that allows us to preserve the structure of the original equation. See [35, 31] for several of these results. The result that interests us the most here is known as the intertwining dilation theorem, which states that if $AX = XB$ where A and B are contractions[3], then we can unitarily dilate A and B to U_A and U_B, respectively, such that $U_A Y = Y U_B$ where Y is a dilation of X. Notice that $AX = XB$ is a Sylvester equation $AX - XB = C$ with $C = 0$. Hence, it is natural to investigate the case when C is non zero.

For further notation, we will let $\mathbb{C}$ be the set of complex numbers. Additionally, M_n will denote the set of $n \times n$ matrices with complex entries, and more generally, $\mathbb{C}^{m \times n}$ will be the set of all $m \times n$ matrices with complex entries. Further, if T is any operator on a Hilbert space $\mathcal{H}$ and $c \in \mathbb{C}$, we will denote $A + cI$ by simply $A + c$, where I is the identity on $\mathcal{H}$. Moreover, we will define the spectrum of an operator T to be $\{\lambda \in \mathbb{C} : T - \lambda \text{ is not invertible}\}$, and denote it by $\sigma(T)$.

The structure of this thesis is as follows. In chapter 2, we will state a key C*-algebra result, known as functional calculus, along with some of its useful consequences. Then, we will introduce operator dilations as well as present some explicit forms of dilations. Additionally, we will give a brief overview on compact operators

[3]An operator A is a contraction if it has spectral norm at most one.

as well as the Eckart-Young theorem.

Next, in chapter 3, we will discuss Zolotarev numbers accompanied by the necessary special functions. As for the numbers themselves, we will mention some properties of the numbers over special sets. The results we will be most concerned with will be simple upper bounds on the Zolotarev numbers.

Then in chapter 4, we will start with reviewing some basic properties of Sylvester equations. Here we will also go over the most important part of the survey aspect of this thesis, which will be the decay of the singular values of the solution, and how to solve a Sylvester equation. Due to most literature assuming normality for the coefficients, we will focus mostly on this case. However, we will also briefly review some of the approaches to the non normal case.

Finally, in chapter 5, we state most of our original contributions. We will begin with surveying previous results on dilations, which give us the motivation to use dilations here. Then, we will explore dilating a Sylvester equation in a manner that allows us to get some information on the decay of the singular values of the solution, where we pay special attention to the case where $\|A\| \, \|B^{-1}\| < 1$. Additionally, there will be a discussion on how dilating a Sylvester equation can help with solving it. Lastly, we will give an example of a very well-structured Sylvester equation that does not have good decay on the singular values of the solution.

2

Operator Theory

2.1 C*-Algebra Preliminaries

Suppose $\mathcal{A}$ is an associative algebra that is additionally a Banach space with the norm $|||\cdot|||$. If in addition $|||AB||| \leq |||A|||B|||$ for all $A, B \in \mathcal{A}$, then $\mathcal{A}$ is a Banach algebra. Let $* : \mathcal{A} \to \mathcal{A}$ be an involution such that for each $A, B \in \mathcal{A}$ and $\lambda \in \mathbb{C}$,

1. $(A^*)^* = A$
2. $(A + B)^* = A^* + B^*$
3. $(\lambda A)^* = \overline{\lambda} A^*$
4. $(AB)^* = B^* A^*$
5. $|||A^* A||| = |||A|||^2$.

Then, $\mathcal{A}$ is a C*-algebra. A very common example of a C*-algebra is $\mathcal{B}(\mathcal{H})$, the set of bounded linear operators on a Hilbert space $\mathcal{H}$ with the norm $\|\cdot\|$ and T^* defined to be the usual adjoint of an operator. Another good example is $C(X)$, the set of continuous functions on a locally compact Hausdorff space X with norm given as

$$\|f\|_X = \sup_{x \in X} |f(x)|, \qquad f \in C(X)$$

with $f^*(x) = \overline{f(x)}$. A special case of this example is when $X = \sigma(T)$, the spectrum of an operator T.

These two examples are very important in the study of C*-algebras. The first is important thanks to the Gelfand-Naimark-Segal construction which tells us that every C*-algebra is isometrically $*$-isomorphic to a C*-subalgebra of $B(\mathcal{H})$ for some Hilbert space $\mathcal{H}$, [20]. The second example is important because of this next theorem, which has a proof found in [14].

Theorem 2.1. Let $\mathcal{A}$ be a C*-algebra with an identity and let $N \in \mathcal{A}$ be normal. Denote the smallest unital C*-subalgebra containing N by $C^*(N)$. Then, there is a unital isometric $*$-isomorphism $\psi : C(\sigma(N)) \to C^*(N)$. Further, if $f(z) = z$, then $\psi(f) = N$.

This theorem is referred to as the functional calculus, and it gives us two results that we will need. First, let $r \in C(\sigma(N))$ be a rational function given by

$$r(z) = \frac{\prod_{j=0}^{m}(z - a_j)}{\prod_{k=0}^{n}(z - b_k)}.$$

Then by the usual definition of $r(N)$, we get

$$r(N) = \frac{\prod_{j=0}^{m}(N - a_j)}{\prod_{k=0}^{n}(N - b_k)}.$$

On the other hand, thanks to how ψ acts on the identity function,

$$\psi(r) = \frac{\prod_{j=0}^{m}(N - a_j)}{\prod_{k=0}^{n}(N - b_k)}.$$

Hence, we get $\psi(r) = r(N)$. Finally, since ψ is isometric,

$$\|r(N)\| = \|r\|_{\sigma(N)}.$$

Second, we can use this result to show positive elements have square roots. By positive, we mean an operator $N \in \mathcal{B}(\mathcal{H})$ such that $\langle Nx, x\rangle \geq 0$ for all $x \in \mathcal{H}$. We will be interested in taking square roots of operators of the form $a^2 - A^*A$ for some $a \geq 0$, which we can easily see are positive if $a \geq \|A\|$.

2.2 Dilation Theory

Dilating an operator is essentially the act of extending some operator up to another with some desired property. Formally, given Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2, \mathcal{K}_1$ and $\mathcal{K}_2$, and an operator $A \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$, we say $B \in \mathcal{B}(\mathcal{K}_1, \mathcal{K}_2)$ is a dilation of A if $\mathcal{H}_i \subset \mathcal{K}_i$, for $i = 1, 2$, and $A = P_2 B P_1$ for the projection[1] P_i from $\mathcal{K}_i$ onto $\mathcal{H}_i$. Additionally, we may require that B has some nice property. The most common types of dilations are normal, isometric, and unitary. Here, we focus on unitary and isometric dilations. Many of the results we show are very well-known and can be found in the literature, in particular, the very influential [35, 31]. As we shall see, any contraction has a unitary dilation. Moreover, thanks to Theorem 2.1, unitary operators can be understood through function theory on the unit circle. Thus, any contraction is related to a continuous function on a subset of the unit circle.

2.2.1 Introduction

Evidently, we first want to determine when an operator has a unitary, or at least an isometric, dilation. As we just mentioned, it has a unitary dilation if the operator is a contraction, meaning it has norm at most one. In fact, for any contraction

[1] Throughout this thesis, we will refer to an orthogonal projection as simply a projection.

$A \in \mathcal{B}(\mathcal{H})$,

$$U = \begin{pmatrix} A & \sqrt{I - AA^*} \\ \sqrt{I - A^*A} & -A^* \end{pmatrix} \in \mathcal{B}(\mathcal{H} \oplus \mathcal{H})$$

is well-defined and unitary. We can also see that this is clearly a dilation of A via the projection P given by

$$P = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \in \mathcal{B}(\mathcal{H} \oplus \mathcal{H})$$

where I is the identity on $\mathcal{H}$.

Unsurprisingly, the most natural way to show U is unitary is to observe that

$$A(I - A^*A) = (I - AA^*)A$$

and thus, by taking limits and using the Weierstrass approximation theorem, [3, Thm. 11.17], we get that

$$A\sqrt{I - A^*A} = \sqrt{I - AA^*}A. \tag{2.1}$$

Alternatively, see [39] for a very interesting approach where they first prove U is unitary using a clever trick. Then, they conclude that equation 2.1 holds.

In the literature, two nice classes of dilations are strong and minimal dilations. Suppose $A \in \mathcal{B}(\mathcal{H})$, $B \in \mathcal{B}(\mathcal{K})$ and $A = PBP$ for the projection P from $\mathcal{K}$ into $\mathcal{H}$. We will call B a strong dilation of A if $A^n = PB^nP$ for all integers $n \geq 0$. As will be shown in the following section, every contraction has a strong unitary dilation.

Before we state the definition of a minimal dilation, given a norm space X, we will define $\overline{X}$ to be the normed closure of X.

1. If B is an isometry, then B is a minimal isometric dilation of A if

$$\mathcal{K} = \overline{\operatorname{span}_{n\geq 0} B^n \mathcal{H}}$$

where the span is taken over all non-negative integers.

2. If B is unitary, then B is a minimal unitary dilation of A if

$$\mathcal{K} = \overline{\operatorname{span}_{n\in \mathbb{Z}} B^n \mathcal{H}}$$

where $\mathbb{Z}$ denotes the set of integers.

See [31] for properties such as existence and uniqueness of minimal unitary dilations. However, the short answer is that they always exist, and are unique up to a reasonably defined isomorphism.

2.2.2 Explicit Forms

Now we state some common forms of unitary and isometric dilations of contractions. Assume here that $A \in \mathcal{B}(\mathcal{H})$ is a contraction. We already saw that

$$U_2 = \begin{pmatrix} A & \sqrt{I - AA^*} \\ \sqrt{I - A^*A} & -A^* \end{pmatrix} \in \mathcal{B}(\mathcal{H} \oplus \mathcal{H})$$

is a unitary dilation of A. This can be generalized to the $n+1 \times n+1$ block matrix

$$U_{n+1} = \begin{pmatrix} A & 0 & \dots & 0 & \sqrt{I - AA^*} \\ \sqrt{I - A^*A} & 0 & \dots & 0 & -A^* \\ 0 & I & & & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & & I & 0 \end{pmatrix} \in \mathcal{B}\left(\mathcal{H}^{n+1}\right).$$

This unitary dilation is used in [28] with a finite dimensional $\mathcal{H}$ to obtain a few results. The two most notable are that U_{n+1} satisfies

$$A^k = P_{\mathcal{H}} U_{n+1}^k P_{\mathcal{H}}$$

for all $0 \leq k \leq n$, and that

$$\mathcal{H}^n = \operatorname*{span}_{1 \leq k \leq n} U_{n+1}^k \mathcal{H}.$$

These two properties are known as U_{n+1} being an n-dilation and an n-minimal dilation, respectively.

The next two explicit forms we look at were first introduced in [31]. First,

$$U_V = \begin{pmatrix} V & I - VV^* \\ 0 & V^* \end{pmatrix} \in \mathcal{B}\left(\mathcal{K} \oplus \mathcal{K}\right)$$

is a strong unitary dilation of any isometry $V \in \mathcal{B}(\mathcal{K})$. Next, set

$$V_A = \begin{pmatrix} A & 0 & \dots \\ \sqrt{I - A^*A} & 0 & \dots \\ 0 & I & \\ \vdots & & \ddots \end{pmatrix} \in \mathcal{B}\left(\ell^2(\mathcal{H})\right)$$

where for a Hilbert space $\mathcal{H}$, we set

$$\ell^2(\mathcal{H}) = \left\{ (x_1, x_2, \dots) \in \bigoplus_{n=1}^{\infty} \mathcal{H} : \sum_{n=1}^{\infty} \|x_n\|^2 < \infty \right\}.$$

Then, ${V_A}^2$ is a strong isometric dilation of $A \in \mathcal{B}(\mathcal{H})$. Combining these two dilations, we get a strong unitary dilation for any contraction.

It should also be noted that by replacing I with $\|A\| I$ in any of the dilations

[2]Notice that this is essentially an infinite-block version as the above finite dimensional unitary dilation.

given here, we may dilate any operator A to U where $\frac{U}{\|A\|}$ is unitary or isometric, provided A is non zero.

2.3 Compact Operators

If $T \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ and $B \subset \mathcal{H}_1$ denotes the closed unit ball of $\mathcal{H}_1$, then recall that T is said to be compact if $\overline{T(B)}$ is a compact subset of $\mathcal{H}_2$, in which case we write $T \in \mathcal{K}(\mathcal{H}_1, \mathcal{H}_2)$.

2.3.1 Elementary Properties

Here we review some basic properties of compact operators that will be used in section 5. The basic properties given here can be found in nearly any book on functional analysis, for example see [10]. For someone unfamiliar with compact operators, it should be noted that the simplest example of a compact operator is one with finite rank.

Proposition 2.2. Let $\{K_n\}_{n=1}^{\infty} \subset \mathcal{K}(\mathcal{H}_1, \mathcal{H}_2)$, $A \in \mathcal{B}(\mathcal{H}_2, \mathcal{H}_3)$, $B \in \mathcal{B}(\mathcal{H}_4, \mathcal{H}_1)$ and $\lambda \in \mathbb{C}$. Then,

1. $\lambda K_1 + K_2$ is compact.
2. K_1^* is compact.
3. AK_1 and $K_1 B$ are compact.
4. If the limit
$$K = \lim_{n \to \infty} K_n$$
exists, then K is compact.

In particular, $\mathcal{K}(\mathcal{H})$ forms a closed ideal in the algebra $\mathcal{B}(\mathcal{H})$.

2.3.2 Eckart-Young Theorem

In [17], Eckart and Young proved a result stating how close we can approximate one matrix with another of lower rank in terms of the Frobenius norm[3]. This was later generalized in [30] for any unitarily invariant norm. For the spectral norm, the result was also generalized to any compact operator, [13, Lemma 1.4]. Before we state the exact result, recall that $\{s_k(A)\}_{k=1}^{\infty}$ denotes the singular values of a compact operator. If $A \in \mathbb{C}^{m\times n}$, then by convention we will say $s_k(A) = 0$ for $k > \min\{m, n\}$.

Theorem 2.3. 1. Let $A \in \mathbb{C}^{m\times n}$ with $m \geq n$ and $k \geq 0$. Denote the diagonal matrix with non zero entries as $\{s_j(A)\}_{j=k+1}^{n}$ by $\Sigma_k \in \mathbb{C}^{n\times n}$. Then if $\|\cdot\|_{\mathrm{UI}}$ is any unitarily invariant norm,

$$\min_{\text{rank } B\leq k} \|A - B\|_{\mathrm{UI}} = \|\Sigma_k\|_{\mathrm{UI}}\,.$$

2. Let A be a compact operator. Then,

$$\inf_{\text{rank } B\leq k} \|A - B\| = s_{k+1}(A).$$

The finite dimensional case has been extensively studied and even has a known form of the rank k matrix B in which the infimization

$$s_{k+1}(A) = \|A - B\|$$

is attained. See [48] for how B is essentially a truncation of the singular value decomposition of A. On the other hand, the literature is much sparser in the infinite dimensional case, but see [11] for more on the singular value expansion of an operator, including if the operator is not compact.

[3]For a matrix M, the Frobenius norm, denoted $\|M\|_F$, is the Euclidean norm of the vector obtained from stacking the columns of M on top of each other.

The Eckart-Young theorem motivates the concept of low numerical rank. We will say an operator T is of low numerical rank if

$$s_k(T) \leq \|T\| \, C\alpha^k, \qquad k \geq 1$$

for some $C > 0$ and $0 < \alpha < 1$.

In section 5, we will require a relation between the k^{th} singular value of X with the k^{th} singular value of Y given a bound on $\|X - Y\|$. Thanks to the Eckart-Young theorem, we can easily derive such a relation.

Corollary 2.4. Let $X, Y \in \mathcal{K}(\mathcal{H}_1, \mathcal{H}_2)$. Then,

1. $$s_k(X) \leq s_k(Y) + \|X - Y\| \, .$$

2. If $X = T_1 Y T_2$ for contractions T_1 and T_2, then

$$s_k(X) \leq s_k(Y).$$

3. If

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},$$

then

$$s_{2k+1}(X) \leq s_{k+1}(X_1) + s_{k+1}(X_2).$$

Further, if rank $X_1 \leq r$, then

$$s_{k+r+1}(X) \leq s_{k+1}(X_1) + s_{k+1}(X_2).$$

Proof. 1. Let $\epsilon > 0$ and by the Eckart-Young theorem, take Z with rank at most

$k-1$ such that

$$\|Y - Z\| < s_k(Y) + \epsilon.$$

Then, by the Eckart Young theorem again,

$$s_k(X) \leq \|X - Z\| \leq \|X - Y\| + \|Y - Z\| < \|X - Y\| + s_k(Y) + \epsilon.$$

Since ϵ was arbitrary, we get the result.

2. Again let $\epsilon > 0$ and take Z as in the previous part. Notice that the rank of $T_1 Z T_2$ is at most $k-1$ and thus,

$$s_k(X) \leq \|X - T_1 Z T_2\| = \|T_1(Y - Z)T_2\| \leq \|Y - Z\| \leq s_k(Y) + \epsilon$$

which gives us the result.

3. Once more fix $\epsilon > 0$ and let Z_1 and Z_2 have rank at most k such that

$$\|X_j - Z_j\| < s_{k+1}(X_j) + \epsilon$$

for $j = 1, 2$. If we set

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix},$$

then Z has rank at most $2k$, and so

$$s_{2k+1}(X) \leq \|X - Z\| \leq \|X_1 - Z_1\| + \|X_2 - Z_2\| < s_{k+1}(X_1) + s_{k+1}(X_2) + 2\epsilon$$

giving us the desired inequality.

Now if we assume further that the rank of X_1 is at most r, then we can take the Z_1 above to have rank at most r as well. This implies the rank of Z is at

most $r+k$ and thus by a similar calculation,

$$s_{r+k+1}(X) < s_{k+1}(X_1) + s_{k+1}(X_2) + 2\epsilon.$$

□

We can also combine two of the above statements to get relations between the ratios of singular values of X and Y in terms of $\|Y - X\|$. For example, if we assume $s_k(X)$ and $s_k(Y)$ are both non zero, part 1 of Corollary 2.4 states that

$$\frac{1}{s_\ell(X) + \|Y - X\|} \leq \frac{1}{s_\ell(Y)},$$

and with simple algebra,

$$\frac{1}{s_\ell(X)} \leq \left(1 + \frac{\|Y - X\|}{s_\ell(X)}\right) \frac{1}{s_\ell(Y)}.$$

Now if we assume further that $X = P_1 Y P_2$ for projections P_1 and P_2, then combining the previous equation with part 2 of Corollary 2.4, we obtain

$$\frac{s_k(X)}{s_\ell(X)} \leq \left(1 + \frac{\|Y - X\|}{s_\ell(X)}\right) \frac{s_k(Y)}{s_\ell(Y)}.$$

3

Numerical Analysis

The goal of this chapter is to discuss Zolotarev numbers. Before that can be done, we must first discuss some special functions.

3.1 Special Functions

3.1.1 Capacity of a condenser

Let $E, F \subset \mathbb{C}$ be disjoint closed sets such that their complements E^C and F^C are both connected. We start with defining the capacity of the condenser (E, F). Denote G to be the complement of $E \cup F$ in the extended complex plane. Let H solve Laplace's equation in G with boundary conditions 0 on ∂E and 1 on ∂F. Next, set Γ to be a contour consisting of a finite number of analytic Jordan curves contained in G such that Γ separates E and F. Write the derivative along the normal to Γ by $\frac{\partial}{\partial n}$ and define the capacity of (E, F) by

$$C(E, F) = \frac{1}{2\pi} \int_\Gamma \frac{\partial H}{\partial n} ds.$$

See [21], [38], and [42] for more information on the capacity of a condenser. Thanks to [4], we can write $C(E, F)$ as

$$C(E, F) = \inf_h \iint_G |\nabla h|^2 \, dxdy$$

with the infimum taken over all h which are 1 on ∂E, 0 on ∂F, and are continuously differentiable on G.

3.1.2 Elliptic Functions

As we will see, elliptic functions play a big role in determining Zolotarev numbers over certain sets. Hence, we give a brief survey here. The functions we discuss can be found in [29, Ch. 19].

We start with a function u_k defined on $[0, 1]$ given by

$$u_k(x) = \int_0^{\sin x} \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}}$$

where $0 < k < 1$. Next, we can define the complete elliptic integral as $K(k) = u_k\left(\frac{\pi}{2}\right)$. An identity stated in [29, Eq. 19.8.19] gives

$$K\left(\sqrt{1-k^2}\right) = \frac{2}{1+k} K\left(\frac{1-k}{1+k}\right). \tag{3.1}$$

Another identity from [1, Table XXI] known as Gauss' transformation states

$$K(k) = \frac{1}{1+k} K\left(\frac{2\sqrt{k}}{1+k}\right). \tag{3.2}$$

Next, we define the so-called Grötzsch ring function μ on $(0, 1)$ by

$$\mu(k) = \frac{\pi}{2} \frac{K\left(\sqrt{1-k^2}\right)}{K(k)}.$$

Two simple upper bounds are given by

$$\mu(k) \leq \ln\left(2\frac{1+\sqrt{1-k^2}}{k}\right) \leq \ln\left(\frac{4}{k}\right) \tag{3.3}$$

which can be found in [29, Eq. 19.9.5]. We can also combine equations 3.1 and 3.2 to get

$$\mu\left(\frac{2\sqrt{k}}{1+k}\right) = \frac{\mu(k)}{2}. \tag{3.4}$$

3.1.3 Möbius Transforms

A Möbius transform is a function T which maps extended complex plane onto the extended complex plane and has the form

$$T(z) = \frac{az+b}{cz+d}, \qquad ad - bc \neq 0$$

for $a, b, c, d \in \mathbb{C}$. There is a considerable amount of literature pertaining to Möbius transforms, see [32, 33].

Corresponding to four points of distinct real numbers (a, b, c, d), we define the cross ratio as

$$\gamma = \frac{|c-a||d-b|}{|d-a||c-b|}.$$ [1]

A property of Möbius transforms, see [32], states that there is a Möbius transformation mapping (a, b, c, d) to (a', b', c', d') if and only if they have the same cross ratio.

[1] Notice that the cross ratio is dependent on the order of (a, b, c, d).

3.2 Zolotarev Numbers

3.2.1 Definition and Elementary Properties

In [50], Zolotarev introduced four problems. The first two are related to polynomials, and the last two are related to rational functions. We focus only on the fourth problem here, which concerns finding a rational function minimized over one set while maximized over another. The main results we will state are relatively simple upper bounds on the Zolotarev numbers over E and F where each of E and F is either a circle or a line. This whole chapter is a survey, except for section 3.2.4 where we derive an upper bound on the Zolotarev numbers over a circle and line in terms of the Zolotarev numbers over two lines.

Define the set $\mathcal{R}_k$ to be

$$\mathcal{R}_k = \{r(z) : r(z) = \frac{p(z)}{q(z)} \text{ where } p \text{ and } q \text{ are polynomials with degree at most } k\}.$$

Then, given $E, F \subset \mathbb{C}$, Zolotarev's fourth problem consists of finding $r \in \mathcal{R}_k$ such that $|r(z)|$ is as small as possible for $z \in E$, while $|r(z)|$ is as big as possible for $z \in F$. More formally, we want r such that the extremal value

$$Z_k(E, F) = \inf_{r \in \mathcal{R}_k} \frac{\sup_{z \in E} |r(z)|}{\inf_{z \in F} |r(z)|}$$

is obtained[2][3]. The number $Z_k(E, F)$ is known as the k^{th} Zolotarev number over the

[2]If either r has a pole in E or a zero in F, by default we set

$$\frac{\sup_{z \in E} |r(z)|}{\inf_{z \in F} |r(z)|} = \infty.$$

[3]It should be noted that if E and F are not disjoint, then upon taking $x \in E \cap F$, one will notice that

$$\frac{\sup_{z \in E} |r(z)|}{\inf_{z \in F} |r(z)|} \geq \frac{|r(x)|}{|r(x)|}$$

and thus $Z_k(E, F) = 1$.

sets E and F. If the infimum is attained at r_k, then we call r_k the extremal function for $Z_k(E, F)$. For special sets such as disjoint real lines or disjoint disks, Zolotarev numbers have been extensively studied, as we will see shortly.

Given two sets E and F, we will say they are well-separated if

$$Z_n(E, F) \leq C\alpha^n, \qquad n \geq 0$$

for some $C \geq 0$ and $0 < \alpha < 1$. Two examples of well-separated sets are disjoint intervals, and disjoint disks, which will be proven in the following two sections.

We now state without proof some immediate properties of Zolotarev numbers, which can each be proven using the extremal function.

Proposition 3.1. Let $E, F \subset \mathbb{C}$ and n, m integers. Then,

1. $Z_0(E, F) = 1$.
2. $Z_n(E, F) \leq Z_m(E, F)$ if $m \leq n$.
3. $Z_n(E, F) = Z_n(F, E)$.
4. If $E_1 \subset E_2$ and $F_1 \subset F_2$, then $Z_n(E_1, F_1) \leq Z_n(E_2, F_2)$.
5. If T is a Möbius transform, then $Z_n(T(E), T(F)) = Z_n(E, F)$.[4]

This next proposition gives us a relation between Zolotarev numbers and the capacity of a condenser. The first part is thanks to [21] and the second part can be found in [42].

Proposition 3.2. Let (E, F) be a condenser. Then,

1.
$$\lim_{n\to\infty} Z_n(E, F)^{\frac{1}{n}} = \exp\left(-\frac{1}{C(E, F)}\right). \tag{3.5}$$

[4]Recall that Möbius transforms are defined on the extended complex plane, and hence are always defined on E and F.

2.

$$\exp\left(-\frac{n}{C(E,F)}\right) \leq Z_n(E,F).$$

3.2.2 Lines

Now we state some properties of $Z_n(E,F)$ where E and F are disjoint real intervals. We will denote $E=[a,b]$ and $F=[c,d]$. Historically, the case that has produced the most results is when $F=-E$, and thus we will assume this for now and denote $F=[-b,-a]$. In [27], an explicit form of $Z_n(E,F)$ is given. This was later refined in [7]. Using this formula, they derived a relatively simple bound on $Z_n(E,F)$. We state both results here.

Theorem 3.3. Let $E=[a,b]$ for $0<a<b$. Then,

$$Z_n(E,-E) = 4\rho^{-2n}\prod_{j=1}^{\infty}\frac{\left(1+\rho^{-8jn}\right)^4}{\left(1+\rho^{4n-8jn}\right)^4} \leq 4\rho^{-2n}, \qquad \rho=\exp\left(\frac{\pi^2}{2\mu\left(\frac{a}{b}\right)}\right). \tag{3.6}$$

Next, thanks to a Möbius transformation, we can extend these results to the case of general disjoint intervals. The following ideas can all be found in [7].

If $E=[a,b]$ and $F=[c,d]$ are disjoint, then there is a Möbius transform T such that $T(E)=[\alpha,\beta]$ and $T(F)=[-\beta,-\alpha]$ for $0<\alpha<\beta$ where the cross ratio of $\{a,b,c,d\}$ is the same as that of $\{-\beta,-\alpha,\alpha,\beta\}$. Thus, we can use equation 3.6 with

$$\rho=\exp\left(\frac{\pi^2}{2\mu\left(\frac{\alpha}{\beta}\right)}\right).$$

Now we seek a simpler bound not involving the Grötzsch ring function. To do this, we begin by recalling that T preserves the cross ratio here. That is,

$$\gamma=\frac{|c-a||d-b|}{|c-b||d-a|}=\frac{\left(1+\frac{\alpha}{\beta}\right)^2}{4\frac{\alpha}{\beta}},$$

which implies

$$\frac{1}{\sqrt{\gamma}} = \frac{2\sqrt{\frac{\alpha}{\beta}}}{1+\frac{\alpha}{\beta}}.$$

Therefore, using equations 3.4 and 3.3, we can derive

$$\mu\left(\frac{\alpha}{\beta}\right) \leq \ln(16\gamma).$$

Placing this into equation 3.6,

$$Z_k(E,F) \leq 4\left[\exp\left(\frac{\pi^2}{2\ln(16\gamma)}\right)\right]^{-2k}. \tag{3.7}$$

It is important to note that this bound always decays to zero for disjoint intervals. To see why, without a loss of generality, assume that $a \leq b < c \leq d$. Next, by shifting, assume $a = 0$. Thus,

$$\gamma = \frac{c(d-b)}{d(c-b)}$$

which is at least one since

$$cd - db = d(c-b) \leq c(d-b) = cd - cb$$

which holds since $c \leq d$. Therefore, the bound in equation 3.7 is

$$Z_k(E,F) \leq 4\left[e^{\alpha}\right]^{-2k}$$

where $\alpha > 0$, which clearly converges to zero.

3.2.3 Disks

We now state some results on Zolotarev numbers over two disjoint disks E and F. It was proven by [43] that if E and F are both centred on the real line, then the

extremal function for $Z_k(E,F)$ is given by

$$r_k = r_1^k, \qquad r_1(z) = \frac{z+\phi}{z+\psi} \tag{3.8}$$

where ϕ and ψ have an explicit, but complicated form. We can then use a simple Möbius transform to generalize this result to any two disjoint disks. To see this, we denote a closed disk by $D(a,r) = \{|z-a| \leq r\}$. Then, if $D(a_1,r_1)$ and $D(a_2,r_2)$ are disjoint and $\alpha_1 \neq \alpha_2$ are real numbers, we set

$$T(z) = az + b, \qquad a = \frac{\alpha_2 - \alpha_1}{a_2 - a_1}, \qquad b = \alpha_2 - a_2 a = \alpha_1 - a_1 a.$$

A very routine calculation shows

$$T\left(D\left(a_1,r_1\right)\right) = D\left(\alpha_1, |a|r_1\right), \qquad T\left(D\left(a_2,r_2\right)\right) = D\left(\alpha_2, |a|r_2\right).$$

We verify the first equality here. Let $|z-a_1| \leq r_1$, then

$$|T(z) - \alpha_1| = |az + \alpha_1 - a_1 a - \alpha_1| = |a||z-a_1| \leq |a| r_1.$$

By combining equations 3.5 and 3.8, we get a neat relation between the capacity and the Zolotarev numbers over two disks:

$$Z_k(E,F) = e^{-\frac{k}{C(E,F)}}.$$

On the other hand, it is established in [7] that if $0 < r_1 < r_2$ and $c \in \mathbb{C}$, then

$$Z_k\left(\{z : |z-c| \leq r_1\}, \{z : |z-c| \geq r_2\}\right) = \left(\frac{r_1}{r_2}\right)^k.$$

This immediately gives us an upper bound on the Zolotarev numbers over two con-

centric circles:

$$Z_k\left(\{z : |z-c| = r_1\}, \{z : |z-c| = r_2\}\right) \leq \left(\frac{r_1}{r_2}\right)^k. \tag{3.9}$$

3.2.4 Circle and Line

Here, we are concerned with the case where F is a circle centered at the origin and E is an interval on the real line outside of F[5]. By a trivial Möbius transform, we will assume that F is the unit circle, and $E = [a, b]$ for $a > 1$. Our results consist of two upper bounds. One which follows by an estimate at the extremal function, and another that uses Möbius transforms to relate this problem to that of the Zolotarev numbers over two intervals. We present this result in the following theorem.

Theorem 3.4. Let $E = [a, b]$ for $1 < a < b$ and $F = \{z : |z| = 1\}$. Then,

$$Z_{2k}(E, F) \leq \min\left\{\left(\frac{\frac{b}{a}-1}{a+b-\frac{2}{a}}\right)^{2k}, Z_k\left(\left[0, \frac{1}{\alpha}\right], \left[\frac{1}{\alpha-d^2}, \frac{1}{\alpha-c^2}\right]\right)\right\}$$

where

$$c = \frac{a+1}{-a+1}, \qquad d = \frac{b+1}{-b+1}, \qquad \alpha > c^2.$$

Additionally,

$$Z_k\left(\left[0, \frac{1}{\alpha}\right], \left[\frac{1}{\alpha-d^2}, \frac{1}{\alpha-c^2}\right]\right) \leq 4\left[\exp\left(\frac{\pi^2}{2\ln(16\gamma)}\right)\right]^{-2k}, \qquad \gamma = \left[\frac{(a+1)(1-b)}{(1-a)(b+1)}\right]^2.$$

Proof. First, set

$$r_1(z) = \frac{z-w}{z-\frac{1}{w}}, \qquad w = \frac{a+b}{2},$$

and notice that

$$r_1'(z) = \frac{w-\frac{1}{w}}{\left(z-\frac{1}{w}\right)^2}$$

[5]By the Möbius transform $\frac{1}{z}$, we can also deal with the case when E is inside of F, assuming E only consists of either non negative or non positive numbers.

which is non zero on E and is defined everywhere on E. Hence,

$$\begin{aligned}\sup_{z\in E}|r_1(z)| &= \max\{|r_1(a)|,|r_1(b)|\}\\ &= \max\left\{\frac{(b-a)(b+a)}{2\left(a[a+b]-2\right)},\frac{(b-a)(b+a)}{2\left(b[a+b]-2\right)}\right\}\\ &= \frac{(b-a)(b+a)}{2\left(a[a+b]-2\right)}.\end{aligned}$$

On the other hand, if $z=e^{i\phi}$,

$$\begin{aligned}|r_1(z)|^2 &= \frac{(\cos\phi-w)^2+\sin^2\phi}{\left(\cos\phi-\frac{1}{w}\right)^2+\sin^2\phi}\\ &= \frac{1-2w\cos\phi+w^2}{1-2\frac{1}{w}\cos\phi+\frac{1}{w^2}}\\ &= \frac{w^2\left(1-2w\cos\phi+w^2\right)}{w^2-2w\cos\phi+1}\\ &= w^2.\end{aligned}$$

Therefore,

$$\frac{1}{\inf_{z\in F}|r_1(z)|}=\frac{1}{w},$$

and hence,

$$Z_1(E,F)\le\frac{(b-a)(b+a)}{2\left(a[a+b]-2\right)}\frac{2}{a+b}=\frac{\frac{b}{a}-1}{a+b-\frac{2}{a}}.$$

As for Z_k for $k>1$, we take $r_k=r_1^k$ to get

$$Z_k(E,F)\le\left(\frac{\frac{b}{a}-1}{a+b-\frac{2}{a}}\right)^k. \tag{3.10}$$

Alternatively, consider the Möbius transformation given by

$$T_1(z)=i\frac{z+1}{-z+1}.$$

A few immediate and useful properties about T_1 is that if $z=e^{i\phi}$, then $T_1(z)=$

$\frac{\sin\phi}{\cos\phi-1}$, and thus T_1 sends the unit circle to the real line with $T_1(1)=\infty$. Additionally, if $x\neq 1$ is real, $T_1(x)$ is on the imaginary axis. Further, if $x>1$, then $T_1(x)$ is on the negative imaginary axis. Thus, our original problem turns into finding the Zolotarev numbers where $G=\mathbb{R}$ is the set of real numbers, and $H=i[c,d]$ where $c<d<0$ for $c=\frac{a+1}{-a+1}$ and $d=\frac{b+1}{-b+1}$.

Next, notice that if we take $f(z)=z^2$, then $f(G)$ is the non-negative real line, and $f(H)$ is a negative real interval. Since this is not a Möbius transform, we do not necessarily get equality between the two corresponding Zolotarev numbers. However, if r_k is the extremal function for $Z_k(f(G),f(H))$, then $r_k(f)$ gives us a bound on $Z_{2k}(G,H)$. That is,

$$Z_{2k}(G,H)\leq Z_k(f(G),f(H))=Z_k\left([0,\infty),\left[-c^2,-d^2\right]\right). \tag{3.11}$$

Finally, we take $\alpha>c^2$ and set

$$T_2(z)=\frac{1}{z+\alpha}.$$

This shows that

$$Z_k\left([0,\infty),\left[-c^2,-d^2\right]\right)=Z_k\left(\left[0,\frac{1}{\alpha}\right],\left[\frac{1}{\alpha-d^2},\frac{1}{\alpha-c^2}\right]\right).$$

We can then use equation 3.7 to obtain a simple bound:

$$Z_{2k}(E,F)\leq 4\left[\exp\left(\frac{\pi^2}{2\ln(16\gamma)}\right)\right]^{-2k},\qquad \gamma=\left[\frac{(a+1)(1-b)}{(1-a)(b+1)}\right]^2. \tag{3.12}$$

□

Although it is not clear that the bound

$$Z_k(E,F) \leq \left(\frac{\frac{b}{a}-1}{a+b-\frac{2}{a}} \right)^k$$

decays to zero, this is in fact true. One way to see this is to notice that

$$b+2 < a^2+ab+a.$$

Then, rearrangements show the desired result.

Neither of the bounds in Theorem 3.4 is always better than the other. It seems that when a is close to one, our method of using Möbius transforms is better. However, as a increases, our method of estimating the extremal function becomes superior. This can be seen in figure 3.1.

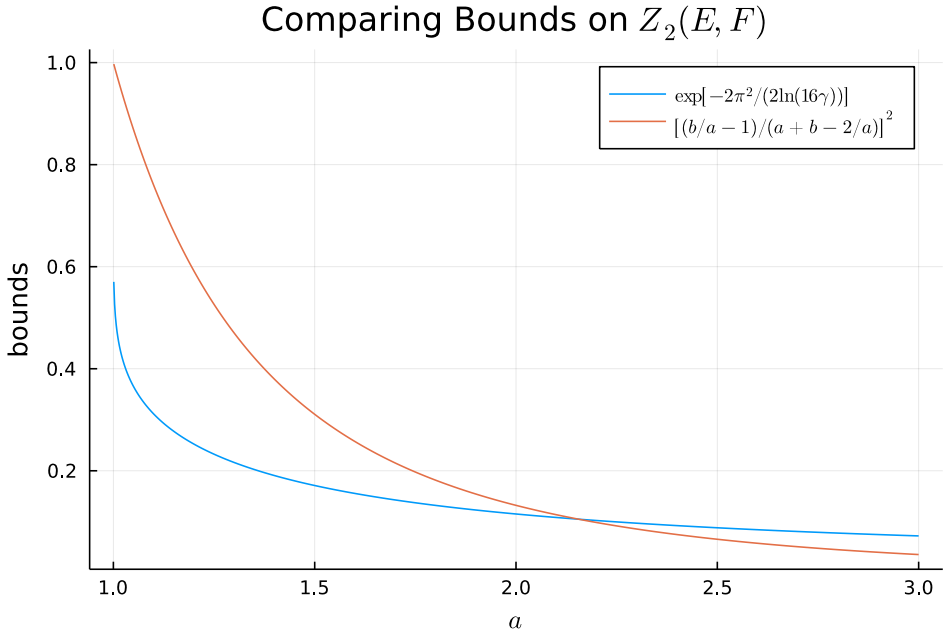

Figure 3.1: This plot compares the decaying factor of 3.12 and 3.10 for varying a and fixed $b = 10$.

It may seem quite surprising that after using the Möbius transforms and a theorem involving the exact Zolotarev number over two intervals, we are still able to beat

that bound by simply making a guess at the extremal function. One might think that the reason for this is that the bound obtained in equation 3.12 is a bound on a bound on the Zolotarev numbers, and thus perhaps one of these bounds is very weak. However, these next plots tell us this isn't the case.

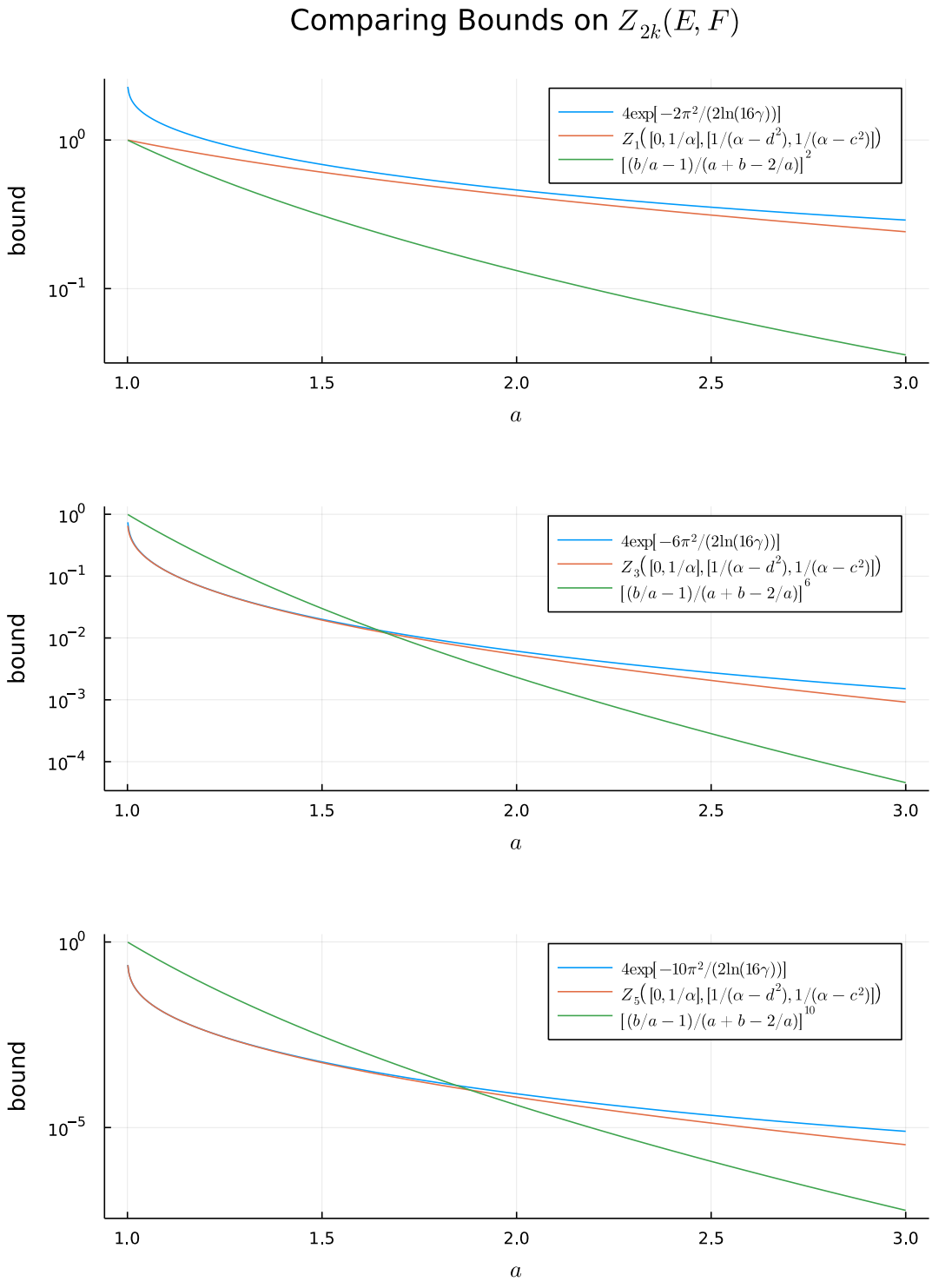

Figure 3.2: These plots compare the bounds on $Z_{2k}(E, F)$ using the exact formula for the Zolotarev numbers over two intervals given in equation 3.6 along with its simple bound also in equation 3.6, as well as the bound obtained from our guess at the extremal function in equation 3.10. The top, middle and bottom compare the bounds for $k = 1, 3, 5$, respectively. Again, we vary a and fix $b = 10$.

Thanks to figure 3.2, we can see that the bound given in equation 3.6 is quite sharp, and not the reason for the simple bound obtained from an estimate of the extremal function being better for certain values of a. Upon reviewing our process with the Möbius transforms to obtain

$$Z_{2k}(E,F) \leq Z_k\left(\left[0,\frac{1}{\alpha}\right],\left[\frac{1}{\alpha-d^2},\frac{1}{\alpha-c^2}\right]\right),$$

one will notice that we used several equalities with one inequality, which occurred in equation 3.11. Therefore, this step must be what causes this method to produce a non optimal bound. Hence, we seek a method which does not contain this step. This is something we are not able to do, and leave as an open question.

4

Sylvester Equations

A Sylvester equation is an equation of the form

$$AX - XB = C \tag{4.1}$$

where $A \in \mathcal{B}(\mathcal{H}_1), B \in \mathcal{B}(\mathcal{H}_2)$ and $X, C \in \mathcal{B}(\mathcal{H}_2, \mathcal{H}_1)$ for Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$. The general idea behind Sylvester equations is that A, B and C are all given, and X is to be determined.

4.1 Basic Properties

4.1.1 Existence and Uniqueness

The most well-known fact about Sylvester equations gives necessary and sufficient conditions on A and B for there to exist a unique solution. This result was proven first by Sylvester in [46] for the case when $\mathcal{H}_1$ and $\mathcal{H}_2$ are finite dimensional, and later independently extended to the case of general operators by both [12] and [40]. Throughout this thesis, we will always assume the linear transformations are bounded, but see [26] for Sylvester equations with unbounded operators.

Theorem 4.1. The equation $AX - XB = C$ has a unique solution X for every C if and only if $\sigma(A) \cap \sigma(B) = \emptyset$.

With this theorem in mind, one might wonder if we can guarantee just one of existence or uniqueness. To determine this, it is important to understand the idea behind the proof of Theorem 4.1, which is to define the linear operator $\mathcal{F} : \mathcal{B}(\mathcal{H}_2, \mathcal{H}_1) \to \mathcal{B}(\mathcal{H}_2, \mathcal{H}_1)$ by $\mathcal{F}(Z) = AZ - ZB$. Then, existence of a solution for every C is equivalent to $\mathcal{F}$ being surjective, and uniqueness for every C is equivalent to $\mathcal{F}$ being injective. If $\mathcal{H}_1$ and $\mathcal{H}_2$ are finite dimensional, then $\mathcal{F}$ is injective if and only if it is surjective. Thus, we can see that uniqueness is equivalent to existence. However, if at least one of $\mathcal{H}_1$ or $\mathcal{H}_2$ is infinite dimensional, then the answer is unclear, and to answer it, some variations of the spectrum are first required.

Let $T \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$. Then define the approximate defect spectrum, denoted as $\sigma_\delta(T)$, by

$$\sigma_\delta(T) = \{\lambda \in \mathbb{C} : T - \lambda \text{ is not onto}\}.$$

We also define the approximate point spectrum, denoted by $\sigma_\pi(T)$, to be

$$\sigma_\pi(T) = \{\lambda \in \mathbb{C} : \|(T - \lambda)x_n\| \to 0 \text{ for some } \{x_n\}_{n=1}^\infty \subset \mathcal{H}_1 \text{ with } \|x_n\| = 1\}.$$

It is easy to see that $\sigma_\pi(T), \sigma_\delta(T) \subset \sigma(T)$ for any $T \in \mathcal{B}(\mathcal{H})$. See [15] for additional properties.

Now as mentioned above, thanks to [15], these two sets allow us to get two slightly different results from Theorem 4.1.

Theorem 4.2. The equation $AX - XB = C$ has a solution X for every C if and only if $\sigma_\delta(A) \cap \sigma_\pi(B) = \emptyset$.

Theorem 4.3. There is an $m > 0$ such that $AX - XB = C$ implies $\|C\| \geq m \|X\|$ if and only if $\sigma_\pi(A) \cap \sigma_\delta(B) = \emptyset$.

4.1.2 Sensitivity

Given $AX - XB = C$, we want to find an upper bound on the norm of X. The most standard approach is to use the separation of A and B:

$$\operatorname{sep}_{|||\cdot|||}(A, B) = \inf_{|||Z|||=1} |||AZ - ZB|||$$

where $|||\cdot|||$ is a norm on the space of operators. Let us assume until further notice that A and B are both acting on finite dimensional spaces, as in this case, the infimum becomes a minimum. Thus, if $\sigma(A) \cap \sigma(B) = \emptyset$, by Theorem 4.1, $\operatorname{sep}_{|||\cdot|||}(A, B) > 0$. It then easily follows that

$$|||X||| \leq \frac{|||C|||}{\operatorname{sep}_{|||\cdot|||}(A, B)}.$$

Now we must ask what a lower bound on the separation is. The following ideas can be found in [44, 45]. First, if U and V are both unitary and conformable, meaning they have appropriate sizes, then

$$\operatorname{sep}(A, B)_{|||\cdot|||_{\mathrm{UI}}} = \operatorname{sep}_{|||\cdot|||_{\mathrm{UI}}}(UAU^*, VBV^*).$$

Additionally, if D and E are two diagonal matrices, then

$$\operatorname{sep}_{\|\cdot\|_F}(D, E) = \min_{\substack{\lambda \in \sigma(D) \\ \mu \in \sigma(E)}} |\lambda - \mu|.$$

Combining these two facts, we immediately get this following theorem.

Theorem 4.4. Let A and B be normal matrices. Then,

$$\operatorname{sep}_{\|\cdot\|_F}(A, B) = \min_{\substack{\lambda \in \sigma(A) \\ \mu \in \sigma(B)}} |\lambda - \mu|.$$

We can then get a lower bound on $\operatorname{sep}_{\|\cdot\|}$ by using the relation given in [24] which

states

$$\|T\| \leq \|T\|_F \leq \sqrt{\operatorname{rank}(T)}\,\|T\| \leq \sqrt{\min\{m,n\}}\,\|T\|$$

where $T \in \mathbb{C}^{n\times m}$. Indeed, this relation immediately implies

$$\frac{1}{\sqrt{\min\{m,n\}}}\operatorname{sep}_{\|\cdot\|_F}(A,B) \leq \operatorname{sep}_{\|\cdot\|}(A,B) \leq \operatorname{sep}_{\|\cdot\|_F}(A,B)$$

where $A \in M_m$ and $B \in M_n$.

Let us now return to the case where A and B are potentially infinite dimensional. The main issue here is that we cannot instantly guarantee the separation is non zero. Hence, we must use the structure of the Sylvester equation to find a bound on $|||X|||$. We give one example of this approach here.

Example 4.5. Suppose $AX - XB = C$ where $|||A|||\,|||B^{-1}||| < 1$. Then, simple algebra gives

$$|||X||| \leq |||A|||\,\Big|\Big|\Big|B^{-1}\Big|\Big|\Big|\,|||X||| + |||C|||\,\Big|\Big|\Big|B^{-1}\Big|\Big|\Big|.$$

Thus,

$$|||X||| \leq \frac{|||C|||\,|||B^{-1}|||}{1 - |||A|||\,|||B^{-1}|||}.$$

In particular, the operator $\mathcal{F}(Z) = AZ - ZB$ is injective

Notice that in the previous example, $\operatorname{sep}_{|||\cdot|||}(A,B)$ is non zero. This follows from

$$\frac{|||AX - XB|||}{|||X|||} \geq \frac{1 - |||A|||\,|||B^{-1}|||}{|||B^{-1}|||},$$

and thus

$$\operatorname{sep}_{|||\cdot|||}(A,B) \geq \frac{1 - |||A|||\,|||B^{-1}|||}{|||B^{-1}|||}.$$

4.2 Singular Values of the Solution

We want to find conditions which guarantee X has a low rank approximation, meaning there is X' such that $\|X - X'\|$ is sufficiently small and X' has low rank, if $AX - XB = C$ where C has low rank. Recall the Eckart-Young theorem which tells us it is sufficient to prove the singular values of X decay quickly. The first main result here was proven by [7] in the case when every operator was finite dimensional. Without changing any details, their proof extends to the case when A and B are any operators on Hilbert spaces and X is any compact operator, as we show below.

Recall that $\mathcal{R}_k$ denotes the set of all rational functions $r = p/q$ where p and q have degree at most k. Additionally, $\{s_k(X)\}_{k=1}^{\infty}$ denotes the set of singular value of X, listed in descending order. By convention, if $X \in \mathbb{C}^{m\times n}$ and $k > \min\{m, n\}$, we will say that $s_k(X) = 0$.

Theorem 4.6. Let $AX - XB = C$ and suppose rank $C = v$, and X is compact. Then for all $j, k \geq 1$,

$$s_{j+vk}(X) \leq \|r_k(A)\| \left\|r_k(B)^{-1}\right\| s_j(X)$$

for any $r_k \in \mathcal{R}_k$ such that $r_k(A)$ and $r_k(B)^{-1}$ are defined.

Proof. Let $p(z)$ and $q(z)$ be polynomials with degree at most k. Define

$$\Gamma = p(A)Xq(B) - q(A)Xp(B),$$

then the claim is that

$$\text{rank } \Gamma \leq vk. \tag{4.2}$$

First, let $t \leq s \leq k$ and notice that

$$\begin{aligned} A^s X B^t - A^t X B^s &= A^t \left(A^{s-t} X - X B^{s-t} \right) B^t \\ &= \sum_{j=0}^{s-t-1} A^{t+j} \left(AX - XB \right) B^{s-1-j} \\ &= \sum_{j=0}^{s-t-1} A^{t+j} C B^{s-1-j} \end{aligned} \tag{4.3}$$

where the second equality comes from the sum telescoping. Next, we require a similar formula for the case when $s \leq t \leq k$. To obtain this, notice that $B^*(-X^*) - (-X^*)A^* = C^*$, and thus

$$\begin{aligned} A^s X B^t - A^t X B^s &= \left[\left(A^s X B^t - A^t X B^s \right)^* \right]^* \\ &= - \left[B^{*^t} (-X^*) A^{*^s} - B^{*^s} (-X^*) A^{*^t} \right]^* \\ &= - \left[\sum_{j=0}^{t-s-1} B^{*^{s+j}} C^* A^{*^{t-1-j}} \right]^* \\ &= - \sum_{j=0}^{t-s-1} A^{t-1-j} C B^{s+j} \end{aligned}$$

where the third equality follows from equation 4.3. Therefore, there are $c_\alpha, c_\beta \in \mathbb{C}$ such that

$$\Gamma = \sum_{\alpha=0}^{k-1} \left[c_\alpha A^\alpha C \sum_{\beta=0}^{k-1} c_\beta B^\beta \right].$$

This is a sum of k terms with each having rank at most the rank of C, which is v. Hence, equation 4.2 is established.

Now let $r_k(z) = p(z)/q(z) \in \mathcal{R}_k$ be such that $r_k(A)$ and $r_k(B)^{-1}$ are defined. This implies that $p(B)$ and $q(A)$ are invertible and thus the operator Y defined by

$$Y = -q(A)^{-1} \Gamma p(B)^{-1} = X - r_k(A) X r_k(B)^{-1}$$

has rank at most vk. Fix $\epsilon > 0$ and by the Eckart-Young theorem, let X_{j-1} be such

that

$$\|X - X_{j-1}\| \leq s_j(X) + \epsilon, \qquad \text{rank } X_{j-1} \leq j - 1.$$

Also define $Y_{j-1} = r_k(A)X_{j-1}r_k(B)^{-1}$ which has rank at most $j-1$, and thus $Y+Y_{j-1}$ has rank at most $j + vk - 1$. Therefore,

$$\begin{aligned} s_{j+vk}(X) &\leq \|X - (Y + Y_{j-1})\| \\ &= \left\|r_k(A)(X - X_{j-1})r_k(B)^{-1}\right\| \\ &\leq \|r_k(A)\| \left\|r_k(B)^{-1}\right\| \|X - X_{j-1}\| \\ &\leq \|r_k(A)\| \left\|r_k(B)^{-1}\right\| (s_j(X) + \epsilon). \end{aligned}$$

Since $\epsilon > 0$ was arbitrary, we get the desired result. □

Now we discuss two special cases where Theorem 4.6 shows the singular values of X decay quickly.

Corollary 4.7. Suppose $AX - XB = C$ where A and B are normal, rank $C = v$, and X is compact. Then, if $j, k \geq 1$,

$$s_{j+vk}(X) \leq Z_k\left(\sigma(A), \sigma(B)\right) s_j(X).$$

Proof. By Theorem 4.6, we have

$$s_{j+vk}(X) \leq \|r_k(A)\| \left\|r_k(B)^{-1}\right\| s_j(X)$$

where $r_k = p/q$ is any rational function in R_k is such that $r_k(A)$ and $r_k(B)^{-1}$ are defined. Notice that is equivalent to q not having any roots in $\sigma(A)$, and p not having any roots in $\sigma(B)$.

Now let $\epsilon > 0$ and $\tilde{r}_k$ be such that

$$\frac{\sup_{z\in\sigma(A)} |\tilde{r}_k(z)|}{\inf_{z\in\sigma(B)} |\tilde{r}_k(z)|} \leq Z_k\left(\sigma(A), \sigma(B)\right) + \epsilon.$$

By part 1 of Proposition 3.1, $\tilde{r}_k(A)$ and $\tilde{r}_k(B)^{-1}$ are both defined. Therefore, since A and B are normal, invoking Theorem 2.1 gives

$$\begin{aligned} s_{j+vk}(X) &\leq \|r_k(A)\| \left\|r_k(B)^{-1}\right\| s_j(X) \\ &= \frac{\sup_{z\in\sigma(A)} |\tilde{r}_k(z)|}{\inf_{z\in\sigma(B)} |\tilde{r}_k(z)|} s_j(X) \\ &\leq \left[Z_k\left(\sigma(A), \sigma(B)\right) + \epsilon\right] s_j(X). \end{aligned}$$

Since $\epsilon > 0$ was arbitrary, we get the announced result. □

Alternatively, we can notice that if $\|A\| \, \|B^{-1}\| < 1$, we need only a clever choice of $r_k(z) = z^k$.

Corollary 4.8. Let $AX - XB = C$ where B is invertible and $\|A\| \, \|B^{-1}\| < 1$. If rank $C = v$ and X is compact,

$$s_{j+vk}(X) \leq \left[\|A\| \left\|B^{-1}\right\|\right]^k s_j(X).$$

Although Theorem 4.6 holds regardless, meaning for a compact X,

$$s_{j+vk}(X) \leq \|r_k(A)\| \left\|r_k(B)^{-1}\right\| s_j(X)$$

regardless of what conditions A, B, and C satisfy, for it to give meaningful information we can see that we desire the case where C has low rank, A and B are normal, and $\sigma(A)$ is disjoint and well-separated from $\sigma(B)$. For generalizing the condition of C having low rank, see [47] for how they impose the condition that C has low numerical rank. In this thesis, section 5 in particular, we explore different ways of

generalizing the normality requirement. Additionally, below we will briefly mention how to deal with the case where the spectrum of A is disjoint but not well-separated from that of B.

The most basic relaxation of the normality condition on A would be to assume A is a diagonalizable matrix. If we assume this and write its diagonalization as $A = PDP^{-1}$, then we obtain

$$\|r_k(A)\| = \left\|Pr_k(D)P^{-1}\right\| \leq \kappa_2(P) \|r_k\|_{\sigma(A)} \tag{4.4}$$

where $\kappa_2(P) = \|P\| \|P^{-1}\|$. With a proper choice of r_k and assuming B is normal,

$$\|r_k(A)\| \left\|r_k(B)^{-1}\right\| \leq \kappa_2(P) Z_k \left(\sigma(A), \sigma(B)\right)$$

will decay quickly as k increases, provided the spectra of A and B are well-separated. Although $\kappa_2(P)$ is independent of k, as we saw in the introduction, in most applications of Sylvester equations, the accuracy of the approximation increases as the size of A and B increases. This in turn can cause $\kappa_2(P)$ to blow up with the size of P. Additionally, it can be quite difficult to get a bound on $\kappa_2(P)$. Hence, there are disadvantages with this method.

In [5], they discuss other methods of relaxing the normality condition, which consist of bounding $\|r(A)\|$ in terms of $\|r\|_{\sigma(A)}$ and a constant. Additionally, they provide a very interesting example. Intuitively, one may think that if at least one of A or B was not normal, then another approach would be to assume they were close to a normal operator. However, their example concludes that this is not correct.

Now we briefly mention how to deal with the spectra of A and B being disjoint but not well-separated, assuming A is normal. By unitarily diagonalizing A, we can

block divide it as

$$A = \begin{pmatrix} A_{11} & \\ & A_{22} \end{pmatrix}$$

where $A_{11} \in M_m$. Then block divide X and C accordingly to obtain

$$\begin{pmatrix} A_{11} & \\ & A_{22} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} - \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} B = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}$$

which results in the two Sylvester equations:

$$A_{11}X_1 - X_1B = C_1, \qquad A_{22}X_2 - X_2B = C_2.$$

Now suppose the spectrum of A_{11} consists of the part of the spectrum of A which is not well-separated from the spectrum of B. Thus, by part 2 and 3 of Corollary 2.4, we get

$$\frac{s_{k+m+1}(X)}{s_\ell(X)} \leq \frac{s_{k+1}(X_1)}{s_\ell(X_1)} + \frac{s_{k+1}(X_2)}{s_\ell(X_2)}.$$

Hence, if $k \geq m$ and m is small, there is no need to be concerned about the first equation as we can simply conclude that

$$\frac{s_{k+m+1}(X)}{s_\ell(X)} \leq \frac{s_{k+1}(X_2)}{s_\ell(X_2)}. \tag{4.5}$$

Finally, notice that the spectra of A_{22} and B are well separated, and thus the singular values of X_2 decay quickly if B is normal.

4.3 Solving Sylvester Equations

Assume here that $AX - XB = C$ where $A \in M_n$ and $B \in M_m$, and thus $X, C \in \mathbb{C}^{n\times m}$. In this section, we explore different ways to solve a Sylvester equation.

The simplest way to solve a Sylvester equation begins with converting the equa-

tion to something of the form $Mx = b$ where M is a matrix and x and b are vectors. To do this, transform the Sylvester equation into

$$M\text{vec } X = \text{vec } C \tag{4.6}$$

where vec T is the vector obtained from taking the columns of a matrix T and stacking them on top of each other. It is shown in [23, Ch. 4] that the matrix M is given by

$$M = (I_m \otimes A) - (B^T \otimes I_n)$$

where $\otimes$ is the Kronecker product. Solving for X is then just a matter of solving the equation 4.6 and then converting vec X back to its matrix form. It is quite clear that an advantage of this method is the simplicity. However, there is one big disadvantage, which is the time complexity to solve the equation using this method. Converting $AX - XB = C$ into equation 4.6 is $\mathcal{O}(m^2n^2)$, while solving equation 4.6 using Gauss-Jordan is $\mathcal{O}(m^3n^3)$. Thus, in total, the time complexity to solve $AX - XB = C$ using this method is $\mathcal{O}(m^3n^3)$.

Next, we look at the most universally popular method of solving Sylvester equations, which was first introduced in [6] by Bartels and Stewart. The first step consists of taking the Schur decomposition, see [24], of A and B and writing them as $A = UA'U^*$ and $B = QB'Q^*$ where A' is lower triangular, and B' is upper triangular. Then, replace A and B in $AX - XB = C$ with their Schur decomposition and multiply the equation on the left by U^* and on the right by Q to get

$$A'X' - X'B' = C'$$

where $T' = U^*TQ$ for $T = X, C$. Next, inductively solve for the entries of X' in the triangular system. For example, if we consider the $(1, 1)$ entry of $A'X' - X'B' = C'$,

we see that

$$A'_{11}X'_{11} - X'_{11}B'_{11} = C'_{11}$$

which easily gives us X'_{11} if $\sigma(A) \cap \sigma(B) = \emptyset$. Next, if we look at the $(2,1)$ entry, we get

$$A'_{11}X'_{12} - X'_{11}B'_{12} - X'_{12}B'_{22} = C'_{12}$$

in which X'_{12} is the only unknown. We can then obtain every entry of X' by continuing in this fashion. After X' has been found, set $X = UX'Q^*$. The cost to run the Bartels-Stewart algorithm is $\mathcal{O}(n^3 + m^3)$ to convert the original Sylvester equation to the triangular form, and $\mathcal{O}(n^2m + nm^2)$ to solve the triangular system, then another $\mathcal{O}(n^2m + nm^2)$ to compute X from X'. Hence, in total, this method costs $\mathcal{O}(n^3 + m^3)$. Notice that while this method is a significant improvement from the first, it is still expensive in time. Additionally, if the size of A and B is big, it can also be very expensive to store the dense matrices U and Q, even if A and B are sparse. Hence, next we discuss some algorithms that can be faster and can also avoid the need to store large dense matrices.

4.3.1 Alternating Direction Implicit Method

The alternating direction implicit (ADI) is an iterative algorithm which gives an approximation of X if $AX - XB = C$. It was first introduced in 1955 in [36], and has a lot of literature concerning it, see for example, [49] for a survey. If k iterations are to be performed, then shifts $\{\alpha_i\}_{i=1}^k, \{\beta_j\}_{j=1}^k \subset \mathbb{C}$ must first be chosen. Next, for an initial guess X_0, perform the following steps.

1. Solve for $X_{j+\frac{1}{2}}$ in

$$(A - \beta_{j+1})X_{j+\frac{1}{2}} = X_j(B - \beta_{j+1}) + C.$$

2. Solve for X_{j+1} in

$$X_{j+1}(B - \alpha_{j+1}) = (A - \alpha_{j+1})X_{j+\frac{1}{2}} - C.$$

We do not address how to quickly solve linear systems here, but see [18] for one method known as Thomas' algorithm to solve the linear system $Ax = b$ in $\mathcal{O}(n)$ where $A \in M_n$ is tridiagonal, and x and b are vectors.

Next, we need to determine how many iterations we need if we want $\|X_k - X\|$ to be a certain precision. See [8] for several properties of this method after k iterations. The most applicable here is

$$X_k - X = \prod_{j=1}^{k} \frac{A - \alpha_j}{A - \beta_j}(X_0 - X) \prod_{j=1}^{k} \frac{B - \beta_j}{B - \alpha_j}$$

where division by a matrix means multiplying by its inverse. Therefore,

$$\frac{\|X_k - X\|}{\|X_0 - X\|} \leq \|r_k(A)\| \left\|r_k(B)^{-1}\right\|, \qquad r_k(z) = \prod_{j=1}^{k} \frac{z - \alpha_j}{z - \beta_j}.$$

Two things are now very evident. First, ideally A and B are normal, as in this case,

$$\frac{\|X_k - X\|}{\|X_0 - X\|} \leq \|r_k\|_{\sigma(A)} \left\|r_k^{-1}\right\|_{\sigma(B)} = \frac{\sup_{z \in \sigma(A)} |r_k(z)|}{\inf_{z \in \sigma(B)} |r_k(z)|}.$$

Second, with a proper choice of shifts,

$$\frac{\|X_k - X\|}{\|X_0 - X\|} \leq Z_k\left(\sigma(A), \sigma(B)\right) + \epsilon$$

for some $\epsilon > 0$ sufficiently small.

See [19] for how to choose the shifts when A and B are both normal and both of their spectra are disjoint real intervals. Additionally, the following example gives us one option for choosing the shifts when A is unitary and B is Hermitian.

Example 4.9. Let A be unitary and B be normal with $\sigma(B) \subset [a, b]$ with $a > 1$. As we established in Theorem 3.4, an estimate for the extremal function here is

$$r_k(z) = \frac{\left(z - \frac{2}{a+b}\right)^k}{\left(z - \frac{a+b}{2}\right)^k}.$$

Hence, our shifts should be $\alpha_j = \frac{2}{a+b}$ and $\beta_j = \frac{a+b}{2}$. After J iterations and an initial guess of $X_0 = 0$, this would give us

$$\|X - X_J\| \leq \left(\frac{\frac{b}{a} - 1}{a + b - \frac{2}{a}}\right)^J \|X\|.$$

It is critical to ask what the time complexity of the ADI method is, as we would like it to beat $\mathcal{O}(n^3 + m^3)$. To see what it is, set $N = \max\{n, m\}$, and notice that it is clear that without additional conditions on A or B, each iteration of ADI costs $\mathcal{O}(N^3)$, and thus k iterations cost $\mathcal{O}(kN^3)$. Hence, we clearly need conditions on A and B that make each iteration quicker. In [19], they impose the condition that $(A - \alpha)x = c$ and $(B - \beta)x = c$ must be solvable in $\mathcal{O}(n)$ and $\mathcal{O}(m)$[1], respectively, which is what we assume here. This implies that each iteration costs only $\mathcal{O}(N^2)$. Hence, ADI costs $\mathcal{O}(kN^2)$ and beats Bartels-Stewart provided $k \ll N$. Now it only remains to establish when not too many iterations are required. However, as we can see above, this just requires that $Z_k(\sigma(A), \sigma(B))$ decays sufficiently fast.

4.3.2 Factored Alternating Direction Implicit Method

Next, we look at the factored ADI (fADI) algorithm. In essence, fADI is a variation of ADI where, as the name suggests, the output has a factored form.

Similar to ADI, if k iterations are to be ran, the first step of fADI consists of choosing shifts $\{\alpha_i\}_{i=1}^k, \{\beta_j\}_{j=1}^k \subset \mathbb{C}$ so that $A - \beta_j$ and $B - \alpha_j$ are both invertible

[1]This appears to be a very demanding condition, however in practice it is common for A and B to be sparse, and thus this condition is satisfied.

for each j. Next, factor C as $C = FG^*$ for $F \in \mathbb{C}^{n\times r}$ and $G \in \mathbb{C}^{m\times r}$. Now make initial guesses Z_0, Y_0, and compute Y_{j+1} and Z_{j+1} in

$$Y_{j+1} = \begin{pmatrix} (A-\beta_{j+1})^{-1}F & (A-\alpha_{j+1})(A-\beta_{j+1})^{-1}Y_j \end{pmatrix}$$
$$Z_{j+1} = \begin{pmatrix} (\beta_{j+1}-\alpha_{j+1})G^*(B-\alpha_{j+1})^{-1} \\ Z_j(B-\beta_{j+1})(B-\alpha_{j+1})^{-1} \end{pmatrix}.$$

After each iteration, define $X_{j+1} = Y_{j+1}Z_{j+1}$. Then, X_{j+1} approximately solves $AX - XB = FG^*$.

At first glance, the relation between ADI and fADI is not clear. However, direct computations with both algorithms show that for $k \geq 0$,

$$X_{k+1} = \frac{A-\alpha_{k+1}}{A-\beta_{k+1}} X_k \frac{B-\beta_{k+1}}{B-\alpha_{k+1}} + (\beta_{k+1}-\alpha_{k+1})(A-\beta_{k+1})^{-1}C(B-\alpha_{k+1})^{-1}.$$

This implies the error analysis of fADI is identical to that of ADI. Additionally, we can see that $\operatorname{rank}(X_{k+1}) \leq \operatorname{rank}(X_k) + \operatorname{rank}(C)$, which implies that

$$\operatorname{rank}(X_k) \leq k\operatorname{rank}(C) + \operatorname{rank}(X_0).$$

5

Dilating a Sylvester Equation

We saw thanks to Corollary 4.7 that if $AX - XB = C$ where A and B are normal, have disjoint and well-separated spectra, and C has low rank, then X has a low rank approximation. In this chapter, we will try to relax the normality condition using operator dilations. On top of the numerical implications of unitarily dilating the coefficients of a Sylvester equation, this next section will show us that whether a Sylvester equation can be dilated is a natural question from an operator theory perspective as well.

5.1 Motivation

Here, we are concerned with algebraic equations involving contractions and determining whether each of the contractions can be unitarily dilated in a manner that preserves this structure. We will look at commuting contractions, as well as intertwining contractions. These results can be found in [35, 31].

If $\mathcal{K}$ is a Hilbert space and $\mathcal{H}$ is a subspace of $\mathcal{K}$, we will denote the projection from $\mathcal{K}$ to $\mathcal{H}$ by $P_{\mathcal{H}}$. This first result tells us that we get a very strong result when each of our contractions are isometries.

Lemma 5.1. Let $\{V_k\}_{k=1}^n \subset \mathcal{B}(\mathcal{H})$ be a set of commuting isometries. Then, there is a set of commuting unitaries $\{U_k\}_{k=1}^n \subset \mathcal{B}(\mathcal{K})$ such that for each $1 \leq k \leq n$, U_k is a unitary dilation of V_k. Further,

$$\prod_{k=1}^{n} V_k^{m_k} = P_{\mathcal{H}} \left[\prod_{k=1}^{n} U_k^{m_k} \right] P_{\mathcal{H}}$$

for any set of nonnegative integers $\{m_k\}_{k=1}^n$.

If our set of commuting contractions consisted of only two contractions, then thanks to [2], we can dilate them both to commuting isometries. Combining this fact with Lemma 5.1, we get Ando's dilation theorem.

Theorem 5.2. Let $\{A_1, A_2\} \subset \mathcal{B}(\mathcal{H})$ be commuting contractions. Then, there are commuting unitaries $\{U_1, U_2\} \subset \mathcal{B}(\mathcal{K})$ such that for $k = 1, 2$, U_k is a unitary dilation of A_k. Further,

$$A_1^n A_2^m = P_{\mathcal{H}} U_1^n U_2^m P_{\mathcal{H}}$$

for any nonnegative integers n and m.

Obviously, one would wonder if we could obtain a similar result if we had an arbitrary number of commuting contractions. However, thanks to Parrott in [34], we cannot even necessarily get a similar result for three commuting contractions.

One downside of the previous results is that we do not know if any of the dilations are minimal. The commutant lifting theorem shows us that if we sacrifice the unitary dilation of one of the contractions in Theorem 5.2, then we can get minimality of the other dilation.

Theorem 5.3. Let $A \in \mathcal{B}(\mathcal{H})$ be a contraction, and $U \in \mathcal{B}(\mathcal{K})$ its minimal unitary dilation. If $B \in \mathcal{B}(\mathcal{H})$ commutes with A, then there is $C \in \mathcal{B}(\mathcal{K})$ commuting with U such that $\|C\| = \|B\|$. Further,

$$BA^n = P_{\mathcal{H}} C U^n P_{\mathcal{H}}$$

for any nonnegative integer n.

Recall that X is said to intertwine A and B if $AX = XB$. We can also dilate an intertwining equation while preserving the structure.

Theorem 5.4. Let $A \in \mathcal{B}(\mathcal{H}_1)$ and $B \in \mathcal{B}(\mathcal{H}_2)$ be contractions with minimal unitary dilations $U_A \in \mathcal{B}(\mathcal{K}_1)$ and $U_B \in \mathcal{B}(\mathcal{K}_2)$, respectively. If X intertwines A and B, then there is a Y intertwining U_A and U_B such that $\|Y\| = \|X\|$. Further,

$$A^n X = XB^n = P_{\mathcal{H}_2} Y U_A^n P_{\mathcal{H}_1} = P_{\mathcal{H}_1} U_B^n Y P_{\mathcal{H}_2}$$

for any nonnegative integer n.

Two very different methods of proving Theorem 5.4 are given by [35] and [31]. A very nice construction is given by [31], whereas [35] first observes that X intertwines A and B if and only if the two operators

$$\begin{pmatrix} 0 & 0 \\ X & 0 \end{pmatrix}, \qquad \begin{pmatrix} B & 0 \\ 0 & A \end{pmatrix}$$

commute. Then, the result follows from Theorem 5.3.

Notice that $AX = XB$ is a Sylvester equation $AX - XB = C$ with $C = 0$. Hence, it is very natural to investigate what can be said when C is non zero. In this chapter, we will explore this idea with intentions of using these results to conclude the singular values of X decay quickly.

5.2 Preliminaries for Dilating a Sylvester Equation

Before we attempt to perform any dilations, it is important to first look at the conditions that must be met if we were to dilate A and B to normal operators.

For example, we saw in Theorem 5.4 that we were able to preserve the norm of X. However, if our intention is to conclude that the singular values of X decay quickly, preserving the norm is not required. To get an idea of what is required, for now suppose we have dilated A to U_1 and B to U_2 where U_1 and U_2 are normal. Let Z solve

$$U_1 Z - Z U_2 = C'$$

which we know exists by Theorem 4.1 if we assume $\sigma(U_1) \cap \sigma(U_2) = \emptyset$. Next, if we assume Z is compact, we may invoke Corollary 4.7 to conclude that the singular values of Z decay quickly provided C' has low numerical rank. Now, what we need is some way to relate the singular values of X to that of Z. Recall part 1 of Corollary 2.4 which tells us if X and Z are both compact, and X' denotes the trivial extension[1] of X to the same space Z acts on,

$$s_k(X') \leq s_k(Z) + \|X' - Z\|.$$

However, $s_k(X') = s_k(X)$ for each $k \geq 1$, and thus we get a relation between the singular values of X and Z. Here on out, we will denote X' by X.

Next, assuming $s_\ell(X)$ and $s_\ell(Z)$ are both non zero, we can again use part 1 of Corollary 2.4 to get

$$\frac{s_k(X)}{s_\ell(X)} \leq K_\ell \frac{s_k(Z) + \|X - Z\|}{s_\ell(Z)}, \qquad K_\ell = 1 + \frac{\|X - Z\|}{s_\ell(X)}^2. \tag{5.1}$$

Here we can see that regardless of how quickly $\frac{s_k(Z)}{s_\ell(Z)}$ decays, the right-hand side of equation 5.1 is bounded below by $\frac{K_\ell \|X-Z\|}{s_\ell(Z)}$. Thus, we must be able to make $\|X - Z\|$ arbitrarily small. This can be done by imposing the condition that Z is a dilation of

[1]Suppose $\mathcal{H}_i$ and $\mathcal{K}_i$ are Hilbert spaces with $\mathcal{H}_i \subset \mathcal{K}_i$ for $i = 1, 2$, and $R \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ and $T \in \mathcal{B}(\mathcal{K}_1, \mathcal{K}_2)$. If $\mathcal{K}_1 = \mathcal{H}_1 \oplus \mathcal{H}_1^\perp$ and $Tx = Rx$ for each $x \in \mathcal{H}_1$ and $Ty = 0$ for each $y \in \mathcal{H}_1^\perp$, then we will call T a trivial extension of R.

[2]Refer to the discussion following Corollary 2.4 on how to derive these relations.

X, in which case we may use part 2 of Corollary 2.4 to say that

$$\frac{s_k(X)}{s_\ell(X)} \leq K_\ell \frac{s_k(Z)}{s_\ell(Z)}.^{3}$$

Alternatively, we may relax this condition and instead require that Z be such that $\epsilon = \|Y - Z\|$ can be made arbitrarily small for some compact dilation Y of X. As if this is true, then

$$\frac{s_k(X)}{s_\ell(X)} \leq K'_\ell \frac{s_k(Y)}{s_\ell(Y)} \leq K''_\ell \frac{s_k(Z) + \epsilon}{s_\ell(Z)}, \qquad K'_\ell = 1 + \frac{\|X - Y\|}{s_\ell(X)}, \qquad K''_\ell = K'_\ell \left(1 + \frac{\epsilon}{s_\ell(Z)}\right).$$

We may now summarize the conditions we must impose if we are to use this method to show the singular values of X decay quickly. Assume we start with $AX - XB = C$, where the coefficients are assumed to be non zero, and after a dilation procedure, we obtain Z along with normal operators U_1 and U_2. If the resulting Sylvester equation is $U_1 Z - Z U_2 = C'$, then we require the following conditions.

1. The spectrum of U_1 is disjoint and well-separated from the spectrum of U_2.
2. The rank of C' must be numerically low.
3. Z must be compact.
4. Either Z is an dilation of X, or for each $\epsilon > 0$, there is a compact dilation Y_ϵ of X such that $\|Y_\epsilon - Z\| < \epsilon$.

In the next two sections, we give conditions on the norm of A and B so that given $AX - XB = C$, we may dilate at least one of A and B to a unitary operator while satisfying the four given conditions above, with special emphasis on the case when $\|A\| \, \|B^{-1}\| < 1$. First, we look at the scenario where the dilations are finite

[3] It is important to note that in the following sections we will have ℓ fixed and k varying.

dimensional, but there are approximations involved. Then, we will remove these approximations, however, we will require infinite dimensional dilations to compensate. Third, we show how dilations can help prove that ADI and fADI converge quickly. Finally, we will provide an example where either the infinite or finite dilations can be used to find a Sylvester equation with A a unitary operator, $\sigma(A)$ and $\sigma(B)$ are well separated, rank $C = 1$, but the singular values of X have slow decay.

5.3 Finite Dimensional Dilations

Recall from section 2 that given an operator A, if we put $U_{n,A}$ to be the $n \times n$ block matrix given by

$$U_{n,A} = \begin{pmatrix} A & 0 & \dots & 0 & D_{A^*} \\ D_A & 0 & \dots & 0 & -A^* \\ 0 & \|A\| & & & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & & \|A\| & 0 \end{pmatrix}, \qquad D_A = \sqrt{\|A\|^2 - A^*A},$$

then $\frac{U_{n,A}}{\|A\|}$ is unitary. Throughout this section, $U_{n,A}$ will denote this dilation.

Before we give the first result, we observe two quick facts.

Proposition 5.5. Suppose $\|A^n\| \, \|B^n\|$ converges to zero. Then, the series

$$S = \sum_{n=1}^{\infty} A^n C B^n$$

is convergent.

Proof. Take $m \geq 1$ be such that $\|A^m\| \, \|B^m\| < 1$. Next, rewrite the series as

$$S = \sum_{n=1}^{\infty} A^{m(n-1)+1} \left[\sum_{k=0}^{m-1} A^k C B^k \right] B^{m(n-1)+1}.$$

Thus,

$$\|S\| \leq \|C\| \sum_{k=0}^{m-1} \|A\|^{k+1} \|B\|^{k+1} \sum_{n=1}^{\infty} \left[\|A^m\| \|B^m\|\right]^{n-1}$$

which is a convergent geometric series. □

Lemma 5.6. Suppose A and B are finite dimensional matrices, B is invertible and $\|A\| = 1$. If $\|B^{-n}\|$ converges to zero, then there is an $m \geq 1$ such that $n \geq m$ implies $\mathcal{F}(Z) = A^*Z + ZB^{n-1}$ is bijective.

Proof. Since $\|B^{-n}\|$ converges to zero, there is $m \geq 1$ such that $n \geq m$ implies $\left\|B^{-(n-1)}\right\| < 1$. The result then follows from example 4.5. □

Theorem 5.7. Suppose $AX - XB = C$ where B is invertible and $\|A\|^m \|B^{-m}\|$ converges to zero. Additionally, assume there is some m such that $\mathcal{F}(Z) = A^*Z + ZB^{n-1}$ is surjective for all $n \geq m$. Then, for all $\epsilon > 0$ there is an $n > 1$ and a dilation Y of X such that

$$U_{n,A}Z - ZB = C'$$

for C' trivially extending C and $\|Z - Y\| < \epsilon$.

Proof. Without loss of generality, assume $\|A\| = 1$. Let n be such that $\left\|B^{-(n-1)}\right\| < 1$ and there is X_n solving $A^*X_n + X_nB^{n-1} = D_AX$. Define $\{X_k\}_{k=2}^n$ to be a sequence where $X_k = X_{k+1}B$ for $2 \leq k \leq n-1$. Now observe that

$$U_{n,A} \underbrace{\begin{pmatrix} X \\ X_2 \\ X_3 \\ \vdots \\ X_n \end{pmatrix}}_{Y} - \begin{pmatrix} X \\ X_2 \\ X_3 \\ \vdots \\ X_n \end{pmatrix} B = \begin{pmatrix} C + D_{A^*}X_n \\ D_AX - A^*X_n - X_2B \\ X_2 - X_3B \\ \vdots \\ X_{n-1} - X_nB \end{pmatrix}. \tag{5.2}$$

By construction, clearly the third entry down to the last entry on the right-hand side are all zero. To see that the second entry is zero, first notice that by construction,

$X_2 = X_n B^{n-2}$. Then,

$$D_A X - A^* X_n - X_2 B = D_A X - A^* X_n - X_n B^{n-1} = 0.$$

Next, write the right-hand side of equation 5.2 as

$$\begin{pmatrix} C + D_{A^*} X_n \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \underbrace{\begin{pmatrix} C \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{C'} + \underbrace{\begin{pmatrix} D_{A^*} X_n \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{R}.$$

The fact that $\|D_{A^*} X_n\|$ can be made arbitrarily small follows from the fact that $\left\|B^{-(n-1)}\right\| < 1$ and example 4.5 applied to $A^* X_n + X_n B^{n-1} = D_A X$ which implies

$$\|X_n\| \leq \frac{\left\|B^{-(n-1)}\right\| \|D_A X\|}{1 - \|B^{-(n-1)}\|}.$$

Now we claim that $\sigma(B) \cap \{|z| = 1\} = \emptyset$. To see this, notice that by the spectral mapping Theorem, [14, Lemma I.2.2],

$$\sigma\left(B^{-n}\right) = \sigma\left(B^{-1}\right)^n,$$

we have

$$\operatorname{spr}\left(B^{-n}\right) = \operatorname{spr}\left(B^{-1}\right)^n.$$

Hence, using the fact that the spectral radius is at most the norm of an operator, [14, Thm. I.2.1],

$$\lim_{n\to\infty} \operatorname{spr}\left(B^{-1}\right)^n \leq \lim_{n\to\infty} \left\|B^{-n}\right\| = 0$$

which implies $\operatorname{spr}(B^{-1}) < 1$. Next, notice that

$$\sigma(B) = \left\{ \frac{1}{\lambda} : \lambda \in \sigma\left(B^{-1}\right) \right\} \tag{5.3}$$

which is true since if λ is non zero,

$$B^{-1} - \lambda = \lambda B^{-1}\left(\frac{1}{\lambda} - B\right)$$

and thus, the complements of the sets in equation 5.3 are equal. Therefore, if the claim was false, there would be some $\lambda \in \sigma(B^{-1})$ such that $|\lambda| = 1$, contradicting $\operatorname{spr}(B^{-1}) < 1$.

Thus, by Theorem 4.1, let Z solve $U_{n,A}Z - ZB = C'$. Then, $U_{n,A}(Y - Z) - (Y - Z)B = R$. Next, by Theorem 4.1, the solution is unique, and by [9, Thm. 9.1], it has the explicit form

$$Y - Z = -\sum_{j=0}^{\infty} U_{n,A}^{j} R B^{-(j+1)}$$

where convergence follows from Proposition 5.5. Further,

$$\|Y - Z\| \leq \|R\| \sum_{k=1}^{m} \left\|B^{-1}\right\|^{k+1} \sum_{j=1}^{\infty} \left\|B^{-m}\right\|^{j-1},$$

where the only term dependent on n is $\|R\|$, which decreases as n increases. Hence, choosing n sufficiently large makes $\|Y - Z\|$ arbitrarily small. □

There are two useful consequences of Theorem 5.7. First, we will show that if in addition B is normal, then we use Theorem 5.7 to show the singular values of X decay quickly. Second, under additional conditions we can use Theorem 5.7 to dilate B as well.

We also point out that the operator B can be dilated instead of A, simply by manipulating the original Sylvester equation. For example, by taking the adjoint,

$AX - XB = C$ turns into

$$-B^*X - X\left(-A^*\right) = C^*.$$

We may then invoke Theorem 5.7 if we assume $\|B\|^n \left\|A^{-n}\right\|$ converges to zero. As we will see, taking this trick one step further and assuming $\|A\| \left\|B^{-1}\right\| < 1$, we can dilate both A and B.

Now we show how Theorem 5.7 can help to find a bound on the decay of the singular values of X.

Corollary 5.8. Let $AX - XB = C$ where A and B are finite dimensional matrices. Assume that $\|A\| = 1$ and B is Hermitian with $\sigma(B) \subset [a, b]$ for $a > 1$, and the rank of C is v. Then if $\epsilon > 0$,

$$s_{\ell+2vk}(X) \leq 4\left[\exp\left(\frac{\pi^2}{2\ln(16\gamma)}\right)\right]^{-2k}\left[s_\ell(X) + \frac{\sqrt{2}\,\|X\|}{a-1} + \epsilon\right] + \epsilon$$

for

$$\gamma = \left[\frac{(a+1)(1-b)}{(1-a)(b+1)}\right]^2.$$

Proof. By Lemma 5.6, the conditions of Theorem 5.7 are satisfied. Hence, let $\epsilon > 0$ and let $n \geq 1$ from Theorem 5.7 be such that $U_{n,A}Z - ZB = C'$ where $\|Z - Y\| < \epsilon$ for the dilation Y of X defined in the proof of Theorem 5.7. By Corollary 4.7, we have

$$s_{\ell+2vk}(Z) \leq Z_{2k} s_\ell(Z), \qquad Z_{2k} = Z_{2k}\left(\{|z| = 1\}, [a, b]\right).$$

Next, by part 1 of Corollary 2.4,

$$s_{\ell+2vk}(Y) \leq s_{\ell+2vk}(Z) + \epsilon \leq Z_{2k} s_\ell(Z) + \epsilon,$$

and by part 2 of Corollary 2.4,

$$\begin{aligned} s_{\ell+2vk}(X) \leq s_{\ell+2vk}(Y) &\leq s_{\ell+2vk}(Z) + \epsilon \\ &\leq Z_{2k} s_\ell(Z) + \epsilon \\ &\leq Z_{2k} \left[s_\ell(Y) + \epsilon \right] + \epsilon \\ &\leq Z_{2k} \left[s_\ell(X) + \|X - Y\| + \epsilon \right] + \epsilon \end{aligned}$$

where the last two inequalities follow from part 1 of Corollary 2.4.

Now to bound $\|Y - X\|$, we start with

$$\|Y - X\| \leq \sum_{k=2}^{n} \|X_k\| . \tag{5.4}$$

Next, we can write each X_k as $X_k = X_n B^{n-k}$. Now using $A^* X_n + X_n B^{n-1} = D_A X$ and [9, Thm. 9.1], we can write X_n as

$$X_n = -\sum_{j=0}^{\infty} A^{*^j} D_A X \left[B^{n-1} \right]^{-(j+1)} .$$

Therefore,

$$X_k = -\left[\sum_{j=0}^{\infty} A^{*^j} D_A X \left[B^{-(n-1)} \right]^j \right] B^{-(k-1)},$$

which implies

$$\|X_k\| \leq \|D_A\| \, \|X\| \, \frac{\|B^{-1}\|^{k-1}}{1 - \|B^{-1}\|^{n-1}}.$$

Placing this into equation 5.4 gives

$$\begin{aligned} \|Y - X\| &\leq \|D_A\| \, \|X\| \, \frac{1}{1 - \|B^{-1}\|^{n-1}} \sum_{k=1}^{n-1} \left\| B^{-1} \right\|^k \\ &\leq \|D_A\| \, \|X\| \, \frac{1}{1 - \|B^{-1}\|^{n-1}} \frac{\|B^{-1}\| \left(1 - \|B^{-1}\|^{n-1} \right)}{1 - \|B^{-1}\|} . \end{aligned}$$

Next, notice that $\|D_A\|^2 = \|I - A^* A\| \leq 1 + \|A\|^2 = 2$, and normality of B implies

$\|B^{-1}\| \leq \frac{1}{a}$. Placing it all together,

$$\|Y - X\| \leq \frac{\sqrt{2}\,\|X\|}{a-1}.$$

Finally, by Corollary 4.7 and Theorem 3.4,

$$Z_{2k} \leq 4\left[\exp\left(\frac{\pi^2}{2\ln(16\gamma)}\right)\right]^{-2k}, \qquad \gamma = \left[\frac{(a+1)(1-b)}{(1-a)(b+1)}\right]^2.$$

□

Alternatively, we could use equation 3.10 to bound the Zolotarev numbers. This would give us

$$s_{\ell+vk}(X) \leq \left[\frac{\frac{b}{a}-1}{a+b-\frac{2}{a}}\right]^k \left[s_\ell(X) + \frac{\sqrt{2}\,\|X\|}{a-1} + \epsilon\right] + \epsilon.$$

Next, we can compare this to the method where we diagonalized A in equation 4.4. Recall that if we diagonalize A,

$$s_{\ell+vk}(X) \leq \alpha Z_k\left(\sigma(A), \sigma(B)\right) s_\ell(X)$$

where $\alpha \geq 1$. Whereas the previous example tells us that if $\|A\| = 1$,

$$s_{\ell+vk}(X) \leq \alpha\left[Z_k\left(\mathbb{T}, \sigma(B)\right) + \epsilon\right] s_\ell(X)^4$$

where $\mathbb{T}$ is the unit circle and again $\alpha \geq 1$. We can see that one advantage with the dilations approach is that provided the spectra of $U_{n,A}$ and B are well-separated, α will not be too large, whereas diagonalizing A could still lead to α being very large.

Now as we announced earlier, we dilate both A and B provided $\|A\|\,\|B^{-1}\| < 1$.

[4]The ϵ can be completely ignored as well. This will be shown using Theorem 5.11.

Corollary 5.9. Let $AX - XB = C$ where A and B are both finite dimensional matrices. Suppose B is invertible and $\|A\| \, \|B^{-1}\| < 1$. Then, for each $\epsilon > 0$, there is some Z dilating X^*, and there are unitary dilations of A and $\frac{B^{-1^*}}{\|B^{-1}\|}$, $U_{n,A}$ and $U_{m,B^{-1^*}/\|B^{-1}\|}$ respectively, such that

$$U_{m,B^{-1^*}/\|B^{-1}\|} Z' - Z' \frac{U_{n,A}}{\|B^{-1}\|} = C'$$

where rank C' = rank C, and $\|Z - Z'\| < \epsilon$.

Proof. Without loss of generality, assume $\|A\| = 1$. Next, notice that since A and B are finite dimensional operators, Lemma 5.6 implies the conditions of Theorem 5.7 hold. Hence, let $\epsilon_1 > 0$ and note that in the proof of Theorem 5.7, we constructed Y dilating X which solves $U_{n,A}Y - YB = C_1 + R_1$ where $U_{n,A}$ is the finite dimensional dilation of A given at the beginning of this section, C_1 extends C trivially, and $\|R_1\| < \epsilon_1$. Now take $U_{n,A}Y - YB = C_1 + R_1$ and multiply on the left by $U_{n,A}^*$, on the right by B^{-1}, and take the adjoint of both sides, and divide by $\|B^{-1}\|$ to get

$$\underbrace{\frac{B^{-1^*}}{\|B^{-1}\|}}_{F} Y^* - Y^* \underbrace{\frac{U_{n,A}}{\|B^{-1}\|}}_{G} = \underbrace{\frac{B^{-1^*}}{\|B^{-1}\|} C_1^* U_{n,A}}_{C_2} + \underbrace{\frac{B^{-1^*}}{\|B^{-1}\|} R_1^* U_{n,A}}_{R_2}.$$

Observe that $G^{-1} = \left(\frac{U_{n,A}}{\|B^{-1}\|}\right)^{-1} = \|B^{-1}\| \, U_{n,A}^*$ and thus

$$\|F\| \left\|G^{-1}\right\| = \left\|B^{-1}\right\| \left\|U_{n,A}^*\right\| = \left\|B^{-1}\right\| < 1.$$

Next, since the dilation is finite dimensional and $\|F\| \, \|G^{-1}\| < 1$, Lemma 5.6 states that the conditions of Theorem 5.7 are satisfied. Thus, we may invoke Theorem 5.7 once more. That is, if $\epsilon_2 > 0$, there is a finite dimensional unitary dilation of F,

$U_{m,B^{-1^*}/\|B^{-1}\|}$, such that

$$U_{m,B^{-1^*}/\|B^{-1}\|}Z - ZG = C_3 + R_3 + R_4$$

where Z is a dilation of Y^*, C_3 and R_3 are trivial extensions of C_2 and R_2, respectively, and $\|R_4\| < \epsilon_2$. Now observe that

$$\|R_3 + R_4\| \leq \epsilon_1 + \epsilon_2.$$

Finally, if Z' solves $U_{m,B^{-1^*}/\|B^{-1}\|}Z' - Z'G = C_3$, then $Z - Z'$ solves

$$U_{m,B^{-1^*}/\|B^{-1}\|}(Z - Z') - (Z - Z')G = R_3 + R_4.$$

The fact that $Z - Z'$ can be made arbitrarily small follows from example 4.5. Additionally, Z is a dilation of Y^* which is a dilation of X^*, and thus Z is a dilation of X^*. □

We could use this result to get a bound on the decay on the singular values of X. However, by equation 3.9, this bound would be

$$s_{j+vk}(X) \leq K_j \left[\|A\| \left\|B^{-1}\right\| + \epsilon\right]^k s_j(X)$$

where $K_j \geq 1$ and v is the rank of C, which is what Corollary 4.8 tells us with $K_j = 1$.

5.4 Infinite Dimensional Dilations

In this section, we will prove similar results as the last section. However, here we trade in the approximations and finite dimensional dilations for exact solutions and infinite dimensional dilations. We begin with recalling the explicit form for dilating

an operator A to V_A where $\frac{V_A}{\|A\|}$ is an isometry, which is given by

$$V_A = \begin{pmatrix} A & 0 & \dots \\ D_A & 0 & \dots \\ 0 & \|A\| & \\ \vdots & & \ddots \end{pmatrix}.$$

Next, if

$$U_A = \begin{pmatrix} V_A & \|V_A\|^2 - V_A V_A^* \\ 0 & -V_A^* \end{pmatrix},$$

then $\frac{U_A}{\|V_A\|}$ is unitary. Throughout this section, U_A and V_A will denote these operators.

These forms give us two immediate dilation results. Before we state them, if we say we can trivially dilate A in $AX - XB = C$, we mean we can dilate A to A' and obtain a new Sylvester equation $A'X' - X'B = C'$ where X' and C' are both trivial extensions of X and C, respectively. Similarly, we say we can trivially dilate B if we can dilate B to B' to obtain $AX' - X'B' = C'$ where X' and C' are trivial extensions of X and C, respectively.

Proposition 5.10. Let $AX - XB = C$.

1. If B is a contraction, then we can trivially isometrically dilate B.

2. If A is an isometry, then we can trivially unitarily dilate A.

Proof. 1. A direct calculation shows

$$A\begin{pmatrix} X & 0 & \dots \end{pmatrix} - \begin{pmatrix} X & 0 & \dots \end{pmatrix}\begin{pmatrix} B & 0 & \dots \\ \sqrt{I - B^*B} & 0 & \dots \\ 0 & I & \\ \vdots & & \ddots \end{pmatrix} = \begin{pmatrix} C & 0 & \dots \end{pmatrix}.$$

Part 2 can be proven similarly using the explicit form of a unitary dilation given above.

□

We now present one method of dilating one of the coefficients of a Sylvester equation to a scalar multiple of a unitary. Thanks to the previous proposition, if we desire to unitarily dilate A, the result is easily achieved upon isometrically dilating A.

Theorem 5.11. Suppose $AX - XB = C$ where B is invertible and $\|A\|^n \|B^{-n}\|$ converges to zero. Then, there a dilation Z of X such that

$$U_A Z - ZB = C'$$

where C' is a trivial extension of C. Further, if X is compact, then so is Z.

Proof. Without loss of generality, assume $\|A\| = 1$. Define $\{X_n\}_{n=1}^{\infty}$ as $X_1 B = \sqrt{I - A^*A}X$ and $X_n B = X_{n-1}$ for $n \geq 2$. Then,

$$\underbrace{\begin{pmatrix} A & 0 & \dots & & \\ D_A & 0 & \dots & & \\ 0 & I & & & \\ \vdots & & \ddots & & \\ & & & I & \\ & & & & \ddots \end{pmatrix}}_{V_A} \underbrace{\begin{pmatrix} X \\ X_1 \\ X_2 \\ \vdots \\ X_n \\ \vdots \end{pmatrix}}_{Y} - \begin{pmatrix} X \\ X_1 \\ X_2 \\ \vdots \\ X_n \\ \vdots \end{pmatrix} B = \begin{pmatrix} AX - XB \\ \sqrt{I - A^*A}X - X_1 B \\ X_1 - X_2 B \\ \vdots \\ X_{n-1} - X_n B \\ \vdots \end{pmatrix} = \underbrace{\begin{pmatrix} C \\ 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \end{pmatrix}}_{C'}$$

where V_A is an isometry. To see that Y has finite norm, notice that

$$Y^*Y - X^*X = \sum_{n=1}^{\infty} \left(B^{-n}\right)^* X^* \left(I - A^*A\right) X B^{-n}$$

which converges by Proposition 5.5 since $\|A\| = 1$.

Next, we can trivially dilate V_A to a unitary to get

$$\begin{pmatrix} V_A & I - V_A V_A^* \\ 0 & V_A^* \end{pmatrix} \begin{pmatrix} Y \\ 0 \end{pmatrix} - \begin{pmatrix} Y \\ 0 \end{pmatrix} B = \begin{pmatrix} C' \\ 0 \end{pmatrix}.$$

Finally, to see that Y is compact assuming X is compact, first notice that each X_n is compact, and since

$$\lim_{n\to\infty} \left\| Y - X^{(n)} \right\| = 0, \qquad X^{(n)} = \begin{pmatrix} X \\ X_1 \\ \vdots \\ X_n \\ 0 \\ \vdots \end{pmatrix}$$

Y is compact by Proposition 2.2. □

Now we could refine Corollary 5.8 using this infinite dimensional dilation. The only differences would be that here, we would have $\epsilon = 0$. Moreover, we would require a different approach to bound $\|Y - X\|$. If we impose the same conditions as Corollary 5.8, this could be done by noticing that

$$\begin{aligned} \|Y - X\| &\leq \sum_{n=1}^{\infty} \|A\|^{n-1} \left\| D_A X B^{-n} \right\| \\ &\leq \sqrt{2}\, \|X\| \sum_{n=1}^{\infty} \left[\|A\| \left\| B^{-1} \right\| \right]^n \\ &= \sqrt{2}\, \|X\| \, \frac{\|A\| \, \|B^{-1}\|}{1 - \|A\| \, \|B^{-1}\|}. \end{aligned}$$

Therefore, if we scale so that $\|A\| = 1$ and assume $\|B^{-1}\| < \frac{1}{a}$, we obtain the same

bound:

$$\|Y - X\| \leq \frac{\sqrt{2}\,\|X\|}{a-1}.$$

Next, we want to be able to say something similar about B. To do this, we use the same idea in Corollary 5.9

Corollary 5.12. Let $AX - XB = C$ with B invertible and $\|A\|\,\|B^{-1}\| < 1$. Then, there is a dilation, Z, of X^* such that

$$\frac{U_A^*}{\|A\|^2}(-Z^*) - (-Z^*)U_{B^{-1^*}}^* = S \tag{5.5}$$

where S has the same rank as C. Further, Z is compact if X is.

Proof. Invoke Theorem 5.11 to get

$$U_A Y - YB = C'$$

where Y is a dilation of X and C' is a trivial extension of C. Now multiply on the left by U_A^{-1} and on the right by B^{-1} to get

$$YB^{-1} - U_A^{-1}Y = U_A^{-1}C'B^{-1}.$$

Next, take the adjoint of both sides

$$\underbrace{B^{-1^*}}_{F}Y^* - Y^*\underbrace{U_A^{-1^*}}_{G} = B^{-1^*}C'^*U_A^{-1^*}.$$

We want to use 5.11 once more. To do this, we require that $\|F\|\,\|G^{-1}\| < 1$. Notice that $\|F\| = \|B^{-1}\|$, and since $\frac{U_A}{\|A\|}$ is unitary, $U_A^{-1} = \frac{U_A^*}{\|A\|^2}$, which implies $U_A^{-1^*} = \frac{U_A}{\|A\|^2}$. Hence,

$$\left\|G^{-1}\right\| = \left\|\left(\frac{U_A}{\|A\|^2}\right)^{-1}\right\| = \|U_A^*\| = \|U_A\| = \|A\|.$$

Therefore, $\|F\|\,\|G^{-1}\| = \|B^{-1}\|\,\|A\| < 1$, and so there is Z which is a dilation of Y^*, an S trivially extending $B^{-1^*}C'^*U_A^{-1^*}$, and a U_F such that $\frac{U_F}{\|B^{-1}\|}$ is unitary, and

$$U_F Z - Z\frac{U_A}{\|A\|^2} = S.$$

Finally, if X is compact, then by Theorem 5.11, Y is compact. Hence, by Theorem 5.11 once more, Z is compact. □

Next, we show that any Sylvester equation can have one of the coefficients unitarily dilated, but this comes at a cost of affecting the other coefficient. Before we give the exact statement, we require a generalization of the isometric dilation given at the start of this section. For any operator $T \in \mathcal{B}(\mathcal{H})$, take $\{T_n\}_{n=1}^{\infty}$ such that

$$W_T = \begin{pmatrix} T & 0 & \dots \\ T_1 & 0 & \dots \\ 0 & T_2 & \\ \vdots & & \ddots \end{pmatrix} \in \mathcal{B}\left(\ell^2(\mathcal{H})\right).$$

That is, the isometric dilation V_T of a contraction and W_T have the same zero entries.

Theorem 5.13. Let $AX - XB = C$ and define W_B as above where

1. B_n is invertible for each $n \geq 1$,
2. The series
$$\sum_{n=1}^{\infty} \|A\|^n \left\| \prod_{k=1}^{n} B_k^{-1} \right\|$$
is finite.

Additionally, let U_A be the dilation of A such that $\frac{U_A}{\|A\|}$ is unitary given at the beginning of the section. Then, there is Z dilating X such that

$$U_A Z - Z W_B = C'$$

where C' trivially extends C. Further, if X is compact, then so is Z.

Proof. First, by the same argument in part 1 of Proposition 5.10, we can trivially dilate B to obtain $AX' - X'W_B = C'$ where X' and C' are trivial extensions of X and C, respectively. For notational purposes, suppose $X \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$. Then for now, define $X_n^{(m)} \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ to be an arbitrary operator for each $n, m \geq 1$, and set

$$X_n = \begin{pmatrix} X_n^{(1)} & \dots & X_n^{(m)} & \dots \end{pmatrix}.$$

Additionally, define

$$Y = \begin{pmatrix} X' \\ X_1 \\ \vdots \end{pmatrix}$$

Now observe that

$$V_A Y - Y W_B = \begin{pmatrix} AX' - X'W_B \\ \sqrt{\|A\|^2 - A^*A}X' - X_1 W_B \\ \|A\| X_1 - X_2 W_B \\ \vdots \\ \|A\| X_{n-1} - X_n W_B \\ \vdots \end{pmatrix}. \tag{5.6}$$

Notice that the first entry is just C'. Next, set the second entry of equation 5.6 to be zero, and observe that

$$0 = \begin{pmatrix} \sqrt{\|A\|^2 - A^*A}X - X_1^{(1)}B - X_1^{(2)}B_1, & -X_1^{(3)}B_2 & , \dots \end{pmatrix}.$$

Let $X_1^{(2)} = \sqrt{\|A\|^2 - A^*A}XB_1^{-1}$, and $X_1^{(m)} = 0$ if $m \neq 2$.

Continue this inductively by setting the m^{th} entry of equation 5.6 to be zero and

assume $X_{n-1}^{(m)} = 0$ for all $m \neq n$. Thus,

$$\begin{aligned} 0 &= \|A\| X_{n-1} - X_n W_B \\ &= \left(-X_n^1 B - X_n^{(2)} B_1, -X_2^{(3)} B_2, \ldots, \|A\| X_{n-1}^{(n)} - X_n^{(n+1)} B_n, -X_n^{(n+2)} B_{n+1}, \ldots \right) \end{aligned}$$

implies $X_n^{(n+1)} = \|A\| X_{n-1}^{(n)} B_n^{-1}$ and $X_n^{(m)} = 0$ for $m \neq n+1$. Therefore,

$$X_n^{(n+1)} = \|A\|^{n-1} D_A X \prod_{k=1}^{n} B_k^{-1}, \qquad n \geq 1.$$

Since $X_n^{(m)} = 0$ for $m \neq n+1$, $\|X_n\| = \|X_n^{n+1}\|$, and thus

$$\|Y\| \leq \|X\| + \sum_{n=1}^{\infty} \|X_n\| \leq \|X\| + \|D_A\| \, \|X\| \sum_{n=1}^{\infty} \|A\|^{n-1} \left\| \prod_{k=1}^{n} B_k^{-1} \right\|$$

which is finite by assumption. The result is then obtained by trivially dilating A to a scalar multiple of a unitary by part 2 of Proposition 5.10 to obtain $U_A Z - Z W_B = C''$ where Z and C'' are trivial extensions of Y and C', respectively.

As for compactness, notice that if X is compact, then each X_n is compact. Additionally, if

$$X^{(n)} = \begin{pmatrix} X' \\ X_1 \\ \vdots \\ X_n \\ 0 \\ \vdots \end{pmatrix},$$

then

$$\left\| Y - X^{(n)} \right\| \leq \sum_{k=n+1}^{\infty} \|X_k\|$$

which converges to zero. Therefore, by Proposition 2.2, Y is compact, and thus so is Z. $\square$

At first glance, it may seem unclear if such a W_B exists for all sets of A and B. However, if we take B_1 to be any invertible operator, and for each $n \geq 2$, set $B_n = b$ to be a scalar multiple of the identity with $\|A\| < b$, then the two conditions of Theorem 5.13 hold. Indeed, clearly each B_n is invertible, and the second condition turns into requiring the series

$$\sum_{n=1}^{\infty} \left(\frac{\|A\|}{b} \right)^n$$

to converge, which is clearly true since $\|A\| < b$. Further, under the assumption that $\|A\| < \|B\| < 1$, we can take W_B to be the isometric dilation of B given at the beginning of the section. To see this, the first assumption of Theorem 5.13 clearly holds since $B_1 = \sqrt{I - B^*B}$ which is invertible since $\text{spr}(B^*B) \leq \|B\|^2 < 1$. Additionally, the series in the second condition is

$$\sum_{n=1}^{\infty} \|A\|^n \left\| \sqrt{I - B^*B}^{-1} \right\|$$

which is a convergent geometric series.

5.5 Using Dilations to Solve a Sylvester Equation

Recall from sections 4.3.1 and 4.3.2 that the conditions for solving a Sylvester equation quickly are very similar to the conditions required to conclude X has a low rank approximation. Hence, we will now show that the dilations presented can also be applied to show that ADI and fADI converge quickly under certain conditions.

Recall that the k iterations of the ADI method consists of first choosing the shifts $\{\alpha_i\}_{i=1}^k, \{\beta_j\}_{j=1}^k \subset \mathbb{C}$. Next, for an initial guess X_0, perform the following steps.

1. Solve for $X_{j+\frac{1}{2}}$ in

$$(A - \beta_{j+1})X_{j+\frac{1}{2}} = X_j(B - \beta_{j+1}) + C.$$

2. Solve for X_{j+1} in

$$X_{j+1}(B - \alpha_{j+1}) = (A - \alpha_{j+1})X_{j+\frac{1}{2}} - C.$$

Theorem 5.14. Suppose $AX - XB = C$ and let U be a dilation of A such that $UY - YB = C'$ where Y and C' are dilations of X and C, respectively. Let P be the projection such that $A = PUP$ and $C = PC'$. Run ADI on both $AX - XB = C$ and $UY - YB = C'$ with shifts $\{\alpha_i\}_{i=1}^k$ and $\{\beta_j\}_{j=1}^k$ such that $\{\alpha_i\}_{i=1}^k \cap \sigma(B) = \{\beta_j\}_{j=1}^k \cap \sigma(A) = \{\beta_j\}_{j=1}^k \cap \sigma(U) = \emptyset$. Assume U and $Y_0 = PX_0$ are chosen such that

$$P(U - \gamma)Y_j = (A - \gamma)PY_j \tag{5.7}$$

for each $\gamma = \alpha_j, \beta_j$ and $1 \leq j \leq k$. Then, $X_j = PY_j$.

Proof. Proceed with induction, noting that the base case is by assumption. Next, assume that $X_j = PY_j$. Beginning with applying ADI to $AX - XB = C$, observe:

$$\begin{aligned}
(A - \beta_{j+1})X_{j+\frac{1}{2}} &= X_j(B - \beta_{j+1}) + C \\
&= PY_j(B - \beta_{j+1}) + PC' \\
&= P\left[Y_j(B - \beta_{j+1}) + C'\right] \\
&= P(U - \beta_{j+1})Y_{j+\frac{1}{2}} \\
&= (A - \beta_{j+1})PY_{j+\frac{1}{2}}.
\end{aligned} \tag{5.8}$$

Since β_{j+1} is chosen so that $A - \beta_{j+1}$ is invertible, $X_{j+\frac{1}{2}} = PY_{j+\frac{1}{2}}$. Showing $X_{j+1} =$

PY_{j+1} follows from a similar calculation:

$$\begin{aligned} X_{j+1}(B-\alpha_{j+1}) &= (A-\alpha_{j+1})X_{j+\frac{1}{2}} - C \\ &= (A-\alpha_{j+1})PY_{j+\frac{1}{2}} - PC' \\ &= P(U-\alpha_{j+1})Y_{j+\frac{1}{2}} - PC' \\ &= P\left[(U-\alpha_{j+1})Y_{j+\frac{1}{2}} - C'\right] \\ &= PY_{j+1}(B-\alpha_{j+1}). \end{aligned} \tag{5.9}$$

As α_{j+1} was chosen so that $B-\alpha_{j+1}$ is invertible, $X_{j+1} = PY_{j+1}$. □

We now use this result to reproduce the result that ADI converges if $\|A\|\,\|B^{-1}\| < 1$ and B is Hermitian.

Corollary 5.15. Let $AX - XB = C$ where $\|A\| = 1$, and B is Hermitian with $\sigma(B) \subset [a,b]$ where $a > 1$. Run k iterations of ADI with an initial guess $X_0 = 0$, along with shifts $\alpha_i = \frac{2}{a+b}$ and $\beta_j = \frac{a+b}{2}$. Then,

$$\frac{\|X_k - X\|}{\|X\|} \leq \left(1 + \frac{\sqrt{2}}{a-1}\right)\left(\frac{\frac{b}{a}-1}{a+b-\frac{2}{a}}\right)^k.$$

Proof. Notice that $\|A\|\,\|B^{-1}\| < 1$, and thus Let $U_A Y - YB = C'$ be the resulting equation given in Theorem 5.11 and P the projection such that $A = PU_AP$. Run ADI on $U_AY - YB = C'$ with the same shifts and initial guess $Y_0 = 0$. We must first show equation 5.7 holds for each $j \geq 1$ and $\gamma = \alpha_j, \beta_j$. For notational purposes, let $A \in \mathcal{B}(\mathcal{H}_1)$. Thus,

$$V_A \in \mathcal{B}(\ell^2(\mathcal{H}_1)), \qquad U_A \in \mathcal{B}\left(\ell^2(\mathcal{H}_1) \oplus \ell^2(\mathcal{H}_1)\right).$$

Further, suppose $B \in \mathcal{B}(\mathcal{H}_2)$. Then,

$$Y_j = \begin{pmatrix} Y_j^1 \\ Y_j^2 \end{pmatrix}$$

where $Y_j^1, Y_j^2 \in \mathcal{B}(\mathcal{H}_2, \ell^2(\mathcal{H}_1))$. The claim is that $Y_j^2 = 0$ for each j, which we prove using induction. Notice that the base case holds since $Y_0 = 0$. Next, assume $Y_k^2 = 0$. Then, by the first step of ADI,

$$\begin{aligned} Y_{k+\frac{1}{2}} &= (U_A - \beta_{k+1})^{-1} \left[Y_k (B - \beta_{k+1}) + C' \right] \\ &= \begin{pmatrix} (V_A - \beta_{k+1})^{-1} & U' \\ 0 & (V_A^* - \beta_{k+1})^{-1} \end{pmatrix} \left[\begin{pmatrix} Y_k^1 \\ 0 \end{pmatrix} (B - \beta_{k+1}) + \begin{pmatrix} C'' \\ 0 \end{pmatrix} \right]. \end{aligned}$$

where $U' = -(V_A - \beta_{k+1})^{-1}(I - V_A V_A^*)(V_A^* - \beta_{k+1})^{-1}$ and C'' is a trivial extension of C. It then follows from direct computation that $Y_{k+\frac{1}{2}}^2 = 0$. Similarly,

$$\begin{aligned} Y_{k+1} &= \left[(A - \alpha_{k+1}) Y_{k+\frac{1}{2}} - C' \right] (B - \alpha_{k+1})^{-1} \\ &= \left[\begin{pmatrix} V_A - \alpha_{k+1} & I - V_A V_A^* \\ 0 & V_A^* - \alpha_{k+1} \end{pmatrix} \begin{pmatrix} Y_{k+\frac{1}{2}}^1 \\ 0 \end{pmatrix} - \begin{pmatrix} C'' \\ 0 \end{pmatrix} \right] (B - \alpha_{k+1})^{-1} \end{aligned}$$

which shows $Y_{k+1}^2 = 0$ by another direct computation.

For additional notation, let

$$Y_j^1 = \begin{pmatrix} X_j^1 \\ X_j^2 \\ \vdots \end{pmatrix}, \qquad X_j^m \in \mathcal{B}(\mathcal{H}_2, \mathcal{H}_1),$$

and let $P_{\mathcal{H}}$ denote the projection from $\ell^2(\mathcal{H})$ onto $\mathcal{H}$. Then, the left hand side of

equation 5.7 is

$$P_{\mathcal{H}}(U_A-\gamma)Y_j=\begin{pmatrix}P_{\mathcal{H}} & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}V_A-\gamma & I-V_AV_A^*\\ 0 & V_A^*\end{pmatrix}\begin{pmatrix}Y_j^1\\ 0\end{pmatrix}=\begin{pmatrix}P_{\mathcal{H}}(V_A-\gamma)Y_j^1\\ 0\end{pmatrix}$$

where

$$P_{\mathcal{H}}(V_A-\gamma)Y_j^1=\begin{pmatrix}(A-\gamma) & 0 & \dots\\ 0 & 0 & \\ \vdots & & \ddots\end{pmatrix}\begin{pmatrix}X_j^1\\ X_j^2\\ \vdots\end{pmatrix}=\begin{pmatrix}(A-\gamma)X_j^1\\ 0\\ \vdots\end{pmatrix}$$

Whereas the right-hand side of equation 5.7 is given by

$$(A-\gamma)P_{\mathcal{H}}Y_j=\begin{pmatrix}(A-\gamma)P_{\mathcal{H}}Y_j^1\\ 0\end{pmatrix}$$

for

$$(A-\gamma)P_{\mathcal{H}}Y_j^1=(A-\gamma)\begin{pmatrix}X_j^1\\ 0\\ \vdots\end{pmatrix}=\begin{pmatrix}(A-\gamma)X_j^1\\ 0\\ \vdots\end{pmatrix}.$$

Therefore, the conditions of Theorem 5.14 are satisfied, and thus we get $X_j = PY_j$ for $j \geq 1$. Hence, by example 4.9

$$\begin{aligned}\frac{\|X_k-X\|}{\|X\|} &\leq \left(1+\frac{\|Y-X\|}{\|X\|}\right)\frac{\|PY_k-PY\|}{\|Y\|}\\ &\leq \left(1+\frac{\|Y-X\|}{\|X\|}\right)\frac{\|Y_k-Y\|}{\|Y\|}\\ &\leq \left(1+\frac{\sqrt{2}}{a-1}\right)\left(\frac{\frac{b}{a}-1}{a+b-\frac{2}{a}}\right)^k\end{aligned}$$

where we bounded $\|Y - X\|$ using the discussion following Theorem 5.11. □

Thanks to the discussion at the end of section 4.3.2, we get that with the exact same conditions, Theorem 5.14 gives an identical result for fADI. Further, we can

achieve an identical result for Corollary 5.15 with fADI.

5.6 Constructing a Counter Example

Suppose $AX - XB = C$ where C has rank one. As we already saw, this does not imply anything about the singular values of X. On the other hand, we saw that the best-case scenario is when A and B are normal, and they have disjoint and well-separated spectra. With this in mind, one might wonder whether we can conclude something similar if we had nearly the best-case scenario. That is, if A was normal, A and B had disjoint and well-separated spectra, and C had rank one. In this section, we will show that this does not necessarily tell us much about the singular values of X either.

We begin with a fact about Jordan matrices.

Example 5.16. Let $0 < b < 1$ and define $T_b \in M_n$ as

$$(T_b)_{i,j} = \begin{cases} b, & j = i, \\ 1, & j = i+1, \\ 0, & \text{otherwise.} \end{cases}$$

Next, set $B_b = T_b^{-1}$. Now we will derive an upper bound on $s_k(B_b)$ for $2 \leq k \leq n$. We begin with noting that

$$T_b^* T_b = \begin{pmatrix} b^2 & b & & & & \\ b & b^2+1 & \ddots & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & b & \\ & & & b & b^2+1 \end{pmatrix} = \underbrace{\begin{pmatrix} b^2+1 & b & & & & \\ b & b^2+1 & \ddots & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & b & \\ & & & b & b^2+1 \end{pmatrix}}_{T'} + \underbrace{\begin{pmatrix} -1 & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix}}_{T''}.$$

For a matrix $M \in M_n$ with real eigenvalues, let $\{\lambda_i(M)\}_{i=1}^n$ denote the eigenvalues in increasing order. Then by Weyl's Theorem, [24, Thm. 4.3.1],

$$\lambda_i\left(T_b^* T_b\right) \geq \lambda_{i-j+1}(T') + \lambda_j(T''), \qquad 1 \leq j \leq i.$$

Set $i = j = 2$ to obtain

$$\lambda_2\left(T_b^* T_b\right) \geq \lambda_1(T') + \lambda_2(T'').$$

It is clear that $\lambda_2(T'') = 0$, whereas [24, Ex. 1.4.P16] states that since T' is a Toeplitz, symmetric and tridiagonal matrix, $\lambda_1(T') \geq (1-b)^2$. Therefore,

$$\lambda_2\left(T_b^* T_b\right) \geq (1-b)^2.$$

Then, since $\{\lambda_i\left(T_b^* T_b\right)\}_{i=1}^n$ is in increasing order, $\lambda_j\left(T_b^* T_b\right) \geq (1-b)^2$ for all $2 \leq j \leq n$. Therefore, $s_k\left(T_b\right) \geq 1-b$ for $1 \leq k \leq n-1$, which gives us the desired bound:

$$s_k\left(B_b\right) \leq \frac{1}{1-b}, \qquad 2 \leq k \leq n.$$

Next, we claim that given $\alpha \geq 1$, there is some $0 < b < 1$ such that

$$\lim_{n\to\infty} \alpha^n s_k(B_b)^n \left\|B_b^{-n}\right\| = 0.$$

To see this, first notice that by the spectral radius formula, [16, Lemma IX.1.8],

$$\lim_{n\to\infty} \left\|B_b^{-n}\right\|^{\frac{1}{n}} = \operatorname{spr}\left(B_b^{-1}\right),$$

for $\epsilon = \frac{1}{2\alpha s_k(B_b)}$, there is an $N \geq 1$ such that $n \geq N$ implies

$$\left\|B_b^{-n}\right\|^{\frac{1}{n}} < \text{spr}\left(B_b^{-1}\right) + \epsilon.$$

Therefore,

$$\begin{aligned}\lim_{n\to\infty} \alpha^n s_k(B_b)^n \left\|B_b^{-n}\right\| &< \lim_{n\to\infty} \left[\alpha s_k(B_b)\text{spr}\left(B_b^{-1}\right) + \frac{1}{2}\right]^n \\ &\leq \lim_{n\to\infty} \left[\alpha \frac{b}{1-b} + \frac{1}{2}\right]^n.\end{aligned}$$

The limit is zero if

$$\alpha \frac{b}{1-b} + \frac{1}{2} < 1,$$

which is true if b is chosen such that

$$b < \frac{1}{2\alpha + 1}.$$

This example allows us to write any matrix X with a left inverse as the solution to a Sylvester equation $AX - XB = C$ with right-hand side rank one and $\|A\|^n \|B^{-n}\|$ converging to zero.

Theorem 5.17. Let $X \in \mathbb{C}^{m\times n}$ with $m \geq n$. Given $\epsilon > 0$ and $k \geq 1$, there is X' and $b > 0$ such that if B_b is the matrix defined in example 5.16, then there exist C and A, so that

$$AX' - X'B_b = C$$

where $\|X - X'\| \leq \epsilon$, rank $C \leq k$, and $\|A\|^n \left\|B_b^{-n}\right\|$ converges to zero. Further if X is left invertible, one can take $X' = X$ with $\epsilon = 0$.

Proof. First assume X has a left inverse, X_{li}. Let B_b be in example 5.16 such that

$$\lim_{n\to\infty} \left(\|X\| \|X_{li}\|\right)^n s_{k+1}(B_b)^n \left\|B_b^{-n}\right\| = 0.$$

Let B_k have rank k such that $s_{k+1}(B_b) = \|B_b - B_k\|$. Set $C = -XB_k$ and define A as $A = X(B_b - B_k)X_{li}$. Then,

$$AX - XB_b = C$$

and,

$$\|A\|^n \leq (\|X\| \, \|X_{li}\|)^n \, s_{k+1}(B)^n$$

which shows the desired result.

Next, if X is not left invertible, then denote the singular value decomposition of X as $X = U\Sigma V^*$. Then, define $X' = U\Sigma' V^*$ where Σ' is diagonal with

$$\Sigma'_{k,k} = \max\{\Sigma_{k,k}, \epsilon\}.$$

Therefore, $\|X - X'\| \leq \epsilon$, and proceed as above with X'. □

Before we give an example of the announced Sylvester equation given at the beginning of this section, we require one more fact about Jordan matrices.

Example 5.18. Let $0 < b < \frac{1}{2}$ and again define $T_b \in M_n$ for $n \geq 2$ as

$$(T_b)_{i,j} = \begin{cases} b, & j = i, \\ 1, & j = i + 1, \\ 0, & \text{otherwise.} \end{cases}$$

Our claim is that

$$\sum_{k=1}^{\infty} \frac{\left\|T_b^k\right\|}{(1-b)^k} \leq n - 1 + f(b), \qquad \lim_{b \to 0^+} f(b) = 0. \tag{5.10}$$

To see this, we will bound both sums up to $k = n-1$, and after $k = n-1$. We begin

with noting the binomial expansion theorem, which tells us that

$$T_b^k = \sum_{j=0}^{\min\{k,n-1\}} \binom{k}{j} b^{k-j} N^j, \qquad N = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}.$$

Therefore, if $1 \leq k \leq n-1$,

$$\left\| T_b^k \right\| \leq \sum_{j=0}^{k} \binom{k}{j} b^{k-j} = 1 + b \sum_{j=0}^{k-1} \binom{k}{j} b^{k-j-1}.$$

Since $k \leq n-1$, there is $M_1 > 0$ such that

$$\left\| T_b^k \right\| \leq 1 + bM_1$$

for each $1 \leq k \leq n-1$. Therefore,

$$\begin{aligned} \sum_{k=1}^{n-1} \frac{\left\| T_b^k \right\|}{(1-b)^k} &= \sum_{k=1}^{n-1} \frac{1+bM_1}{(1-b)^k} \\ &\leq \frac{(1+bM_1)(n-1)}{(1-b)^{n-1}} \\ &= n - 1 + g(b) + b\frac{M_1(n-1)}{(1-b)^{n-1}}, \qquad \lim_{b \to 0^+} g(b) = 0. \end{aligned}$$

Next, to bound the remainder of the sum in equation 5.10, first notice that if $k \geq n$, $1 < n - b \leq k - b$, which implies

$$k^n < k^{n+1} - k^n b + b^n,$$

which is equivalent to

$$k^n - b^n < k^n(k-b).$$

Hence,

$$\frac{k^n - b^n}{k - b} < k^n. \tag{5.11}$$

Additionally,

$$\left\|T_b^k\right\| \leq \sum_{j=0}^{n-1} \binom{k}{j} b^{k-j}.$$

Therefore,

$$\begin{aligned}
\sum_{k=n}^{\infty} \frac{\left\|T_b^k\right\|}{(1-b)^k} &\leq \sum_{k=n}^{\infty} \sum_{j=0}^{n-1} \frac{\binom{k}{j} b^{k-j}}{(1-b)^k} \\
&\leq \sum_{k=n}^{\infty} \left(\frac{b}{1-b}\right)^k \sum_{j=0}^{n-1} \left(\frac{k}{b}\right)^j \qquad \text{(proof found in [24, P. 181])} \\
&= \sum_{k=n}^{\infty} \left(\frac{b}{1-b}\right)^k \left(\frac{\left[\frac{k}{b}\right]^n - 1}{\frac{k}{b} - 1}\right) \\
&= \frac{1}{b^{n-1}} \sum_{k=n}^{\infty} \left(\frac{b}{1-b}\right)^k \left(\frac{k^n - b^n}{k - b}\right) \\
&< \frac{1}{b^{n-1}} \sum_{k=n}^{\infty} \left(\frac{b}{1-b}\right)^k k^n \qquad \text{(equation 5.11)} \\
&= b \sum_{k=n}^{\infty} \left(\frac{b}{1-b}\right)^k \left(\frac{k}{b}\right)^n.
\end{aligned}$$

The last series converges thanks to the root test:

$$\begin{aligned}
\lim_{k\to\infty} \left[\left(\frac{b}{1-b}\right)^k \left(\frac{k}{b}\right)^n\right]^{\frac{1}{k}} &= \lim_{k\to\infty} \left(\frac{b}{1-b}\right) \left(\frac{k}{b}\right)^{\frac{n}{k}} \\
&= \left(\frac{b}{1-b}\right) \left[\frac{\lim_{k\to\infty} k^{\frac{1}{k}}}{\lim_{k\to\infty} b^{\frac{1}{k}}}\right]^n \\
&= \frac{b}{1-b} \\
&< 1.
\end{aligned}$$

Hence, we conclude that

$$\sum_{k=n}^{\infty} \frac{\left\|T_b^k\right\|}{(1-b)^k} \leq bM_2$$

where $M_2 > 0$. Therefore, we obtain

$$\sum_{k=1}^{\infty} \frac{\left\|T_b^k\right\|}{(1-b)^k} \leq \sum_{k=1}^{n-1} \frac{\left\|T_b^k\right\|}{(1-b)^k} + \sum_{k=n}^{\infty} \frac{\left\|T_b^k\right\|}{(1-b)^k} \leq n - 1 + g(b) + b\left[\frac{M_1(n-1)}{(1-b)^{n-1}} + M_2\right]$$

which proves the claim.

Example 5.19. Let $n \geq 2$, $m \geq n$ and $X \in \mathbb{C}^{m\times n}$ any matrix where each singular value is $s_k(X) = 1$. Let $AX - XB_b = C$ be the corresponding Sylvester equation in Theorem 5.17 where rank $C = 1$ and $b < \frac{1}{2}$. Using Theorem 5.11, we get $UY_b - Y_bB_b = C'$ where rank $C' = 1$, $\frac{U}{\|A\|}$ is the unitary dilation of $\frac{A}{\|A\|}$, and Y_b is the dilation of X, all defined in Theorem 5.11. Then, by part 1 and 2 of Corollary 2.4,

$$1 = \frac{s_k(X)}{s_1(X)} \leq (1 + \|Y_b - X\|)\,\frac{s_k(Y_b)}{s_1(Y_b)}.$$

To bound $\|Y_b - X\|$, we begin with bounding $\|A\|$. First, recall from Theorem 5.17 that if B_k has rank at most k and $s_{k+1}(B_b) = \|B_b - B_k\|$, then

$$A = X\,(B_b - B_k)\,X_{li}$$

where if the singular value decomposition of X is $X = U\Sigma V^*$, then we can write $X_{li} = V\Sigma^*U^*$. Hence, X and X_{li} both have norm one since each singular value of X is one. Therefore,

$$\|A\| = \|X(B_b - B_k)X_{li}\| \leq s_k(B_b) \leq \frac{1}{1-b}.$$

Thus, using the discussion following Theorem 5.11,

$$\begin{aligned}\|Y_b - X\| &\leq \sum_{k=1}^{\infty} \sqrt{2}\,\|X\|\,\|A\|^k \left\|B_b^{-k}\right\| \\ &\leq \sqrt{2}\sum_{j=1}^{\infty} \frac{\left\|B_b^{-k}\right\|}{(1-b)^k}\end{aligned}$$

Therefore, by example 5.18, given $\epsilon > 0$, there is some $b > 0$ such that

$$\|Y_b - X\| \leq \sqrt{2}(n-1) + \epsilon.$$

Thus, for any b sufficiently small, Y_b is the solution of a Sylvester equation with right-hand side having rank one, one coefficient is a scalar multiple of a unitary, and Y_b has slow decay of its singular values. Finally, since $\|A\| \leq s_k(B)$, for small b, the spectra of U and B are very well-separated.

6

Conclusion

In this thesis, we saw conditions on $AX - XB = C$ which implied X has a low rank approximation. As Corollary 4.7 tells us, this occurs when A and B are normal, have disjoint and well-separated spectra, and C has small rank. The goal of this thesis was to investigate if the normality condition could be relaxed in any way. Our approach was operator dilations, and thanks to Theorems 5.7 and 5.11, we showed that indeed we could if $\|A\| \, \|B^{-1}\| < 1$. Additionally, if B is normal at the start, and C has low rank, we can perform these dilations in a manner that allows us to show the decay of the singular values of X is bounded above by the Zolotarev numbers over a circle and a line. Although we derived bounds on these numbers in Theorem 3.4, these bounds are far from ideal, and thus sharpening them requires future work.

Bibliography

[1] Naum Il'ich Akhiezer. *Elements of the Theory of Elliptic Functions*, volume 79. American Mathematical Soc., 1990.

[2] Tsuyoshi Andô. On a Pair of Commutative Contractions. *Acta Sci. Math.(Szeged)*, 24(1-2):88–90, 1963.

[3] Tom M. Apostol and C.M. Ablow. Mathematical Analysis. *Physics Today*, 11(7):32, 1958.

[4] Thomas Bagby. The Modulus of a Plane Condenser. *Journal of Mathematics and Mechanics*, 17(4):315–329, 1967.

[5] Jonathan Baker, Mark Embree, and John Sabino. Fast Singular Value Decay for Lyapunov Solutions with Nonnormal Coefficients. *SIAM Journal on Matrix Analysis and Applications*, 36(2):656–668, 2015.

[6] Richard H. Bartels and George W. Stewart. Solution of the Matrix Equation AX+ XB= C [f4]. *Communications of the ACM*, 15(9):820–826, 1972.

[7] Bernhard Beckermann and Alex Townsend. Bounds on the Singular Values of Matrices with Displacement Structure. *SIAM Review*, 61(2):319–344, 2019.

[8] Peter Benner and Patrick Kürschner. Computing Real Low-Rank Solutions of Sylvester Equations by the Factored ADI Method. *Computers & Mathematics with Applications*, 67(9):1656–1672, 2014.

[9] Rajendra Bhatia and Peter Rosenthal. How and why to Solve the Operator Equation AX- XB= Y. *Bulletin of the London Mathematical Society*, 29(1):1–21, 1997.

[10] John B. Conway. *A Course in Functional Analysis*, volume 96. Springer, 2019.

[11] Daniel K. Crane and Mark S. Gockenbach. The Singular Value Expansion for Arbitrary Bounded Linear Operators. *Mathematics*, 8(8):1346, 2020.

[12] Yu L. Daleckiı. On the Asymptotic Solution of a Vector Differential Equation. In *Dokl. Akad. Nauk SSSR*, volume 92, pages 881–884, 1953.

[13] Kenneth R. Davidson. *Nest Algebras*. 1987.

[14] Kenneth R. Davidson. *C*-Algebras by Example*, volume 6. American Mathematical Soc., 1996.

[15] Chandler Davis and Peter Rosenthal. Solving Linear Operator Equations. *Canadian Journal of Mathematics*, 26(6):1384–1389, 1974.

[16] N. Dunford and J.T. Schwartz. Linear Operators. Part 2: Spectral Theory. Self Adjoint Operators in Hilbert Space, 1963. *Wiey-Interscience, New York*.

[17] Carl Eckart and Gale Young. The Approximation of one Matrix by Another of Lower Rank. *Psychometrika*, 1(3):211–218, 1936.

[18] William Ford. *Numerical Linear Algebra with Applications: Using MATLAB*. Academic Press, 2014.

[19] Daniel Fortunato and Alex Townsend. Fast Poisson Solvers for Spectral Methods. *IMA Journal of Numerical Analysis*, 40(3):1994–2018, 2020.

[20] Israel Gelfand and Mark Naimark. On the Imbedding of Normed Rings into the Ring of Operators in Hilbert Space. *Mat. Sbornik.*, 12(2):197–217, 1943.

[21] A.A. Gončar. Zolotarev Problems Connected with Rational Functions. *Mathematics of the USSR-Sbornik*, 7(4):623, 1969.

[22] Leslie Greengard and Vladimir Rokhlin. A Fast Algorithm for Particle simulations. *Journal of Computational Physics*, 73(2):325–348, 1987.

[23] Roger A Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1994.

[24] Roger A. Horn and Charles R. Johnson. *Matrix Analysis,* 2nd *Ed.* Cambridge university press, 2013.

[25] Brock Klippenstein and Richard Mikael Slevinsky. Fast Associated Classical Orthogonal Polynomial Transforms. *Journal of Computational and Applied Mathematics*, 403:113831, 2022.

[26] Nguyen Thanh Lan. On the Operator Equation A X- X B= C with Unbounded Operators A, B, and C. In *Abstract and Applied Analysis*, volume 6, pages 317–328. Hindawi, 2001.

[27] V.I. Lebedev. On a Zolotarev Problem in the Method of Alternating Directions. *USSR Computational Mathematics and Mathematical Physics*, 17(2):58–76, 1977.

[28] Eliahu Levy and Orr Moshe Shalit. Dilation Theory in Finite Dimensions: the Possible, the Impossible and the Unknown. *Rocky Mountain Journal of Mathematics*, 44(1):203–221, 2014.

[29] Daniel W. Lozier. NIST Digital Library of Mathematical Functions. *Annals of Mathematics and Artificial Intelligence*, 38(1):105–119, 2003.

[30] Leon Mirsky. Symmetric Gauge Functions and Unitarily Invariant Norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960.

[31] Béla Sz. Nagy, Ciprian Foias, Hari Bercovici, and László Kérchy. *Harmonic Analysis of Operators on Hilbert Space*. Springer Science & Business Media, 2010.

[32] Tristan Needham. *Visual Complex Analysis*. Oxford University Press, 1998.

[33] John Olsen. The Geometry of Möbius Transformations. *Rochester: University of Rochester*, 2010.

[34] Stephen Parrott. Unitary Dilations for Commuting Contractions. *Pacific Journal of Mathematics*, 34(2):481–490, 1970.

[35] Vern Paulsen. *Completely Bounded Maps and Operator Algebras*. Number 78. Cambridge University Press, 2002.

[36] Donald W. Peaceman and Henry H. Rachford, Jr. The Numerical Solution of Parabolic and Elliptic Differential Equations. *Journal of the Society for Industrial and Applied Mathematics*, 3(1):28–41, 1955.

[37] Thilo Penzl. Eigenvalue Decay Bounds for Solutions of Lyapunov Equations: the Symmetric Case. *Systems & Control Letters*, 40(2):139–144, 2000.

[38] George Pólya and Gábor Szegö. *Isoperimetric Inequalities in Mathematical Physics.(AM-27), Volume 27*. Princeton University Press, 2016.

[39] P.L. Robinson. Julia Operators and Halmos Dilations. *arXiv preprint arXiv:1803.09329*, 2018.

[40] Marvin Rosenblum. On the Operator Equation BX- XA= Q. *Duke Mathematical Journal*, 23(2):263–269, 1956.

[41] John Sabino. Solution of Large-Scale Lyapunov Equations via the Block Modified Smith Methods. Technical report, 2006.

[42] Edward B. Saff and Vilmos Totik. *Logarithmic Potentials with External Fields*, volume 316. Springer Science & Business Media, 2013.

[43] Gerhard Starke. Near-Circularity for the Rational Zolotarev Problem in the Complex Plane. *Journal of Approximation Theory*, 70(1):115–130, 1992.

[44] Gilbert Stewart. Error Bounds for Approximate Invariant Subspaces of Closed Linear Operators. *SIAM Journal on Numerical Analysis*, 8(4):796–808, 1971.

[45] Gilbert Stewart. Error and Perturbation Bounds for Subspaces Associated with Certain Eigenvalue Problems. *SIAM review*, 15(4):727–764, 1973.

[46] James J. Sylvester. Sur l'Équation en Matrices px= xq. *CR Acad. Sci. Paris*, 99(2):67–71, 1884.

[47] Alex Townsend and Heather Wilber. On the Singular Values of Matrices with High Displacement Rank. *Linear Algebra and its Applications*, 548:19–41, 2018.

[48] Charles F. Van Loan and G. Golub. Matrix Computations (Johns Hopkins Studies in Mathematical Sciences). *Matrix Computations*, 1996.

[49] Eugene Wachspress. *The ADI Model Problem*. Springer, 2013.

[50] EI Zolotarev. Application of Elliptic Functions to Questions of Functions Deviating Least and Most from Zero. *Zap. Imp. Akad. Nauk. St. Petersburg*, 30(5):1–59, 1877.