A CS guide to the quantum singular value transformation

Ewin Tang^{*} Kevin Tian[†]

Abstract

We present a simplified exposition of some pieces of [GSLW19], which introduced a quantum singular value transformation (QSVT) framework for applying polynomial functions to blockencoded matrices. The QSVT framework has garnered substantial recent interest from the quantum algorithms community, as it was demonstrated by [GSLW19] to encapsulate many existing algorithms naturally phrased as an application of a matrix function. First, we posit that the lifting of quantum singular processing (QSP) to QSVT is better viewed not through Jordan's lemma [Jor75; Reg06] (as was suggested by [GSLW19]) but as an application of the *cosine-sine decomposition*, which can be thought of as a more explicit and stronger version of Jordan's lemma. Second, we demonstrate that the constructions of bounded polynomial approximations given in [GSLW19], which use a variety of ad hoc approaches drawing from Fourier analysis, Chebyshev series, and Taylor series, can be unified under the framework of *truncation of Chebyshev series*, and indeed, can in large part be matched via a bounded variant of a standard meta-theorem from [Tre19]. We hope this work finds use to the community as a companion guide for understanding and applying the powerful framework of [GSLW19].

Contents

1	Introduction	2
	1.1 Our contributions	2
	1.2 Notation	3
2	QSVT and the Cosine-Sine decomposition	3
	2.1 Existence of the CS decomposition	3
	2.2 QSVT and quantum signal processing	5
	2.3 Simplified QSVT in the computational basis	7
	2.4 Simplified QSVT in general bases	10
3	QSVT and Chebyshev Series	10
	3.1 Chebyshev polynomials	11
	3.2 Chebyshev series for standard functions	12
	3.3 Bounded approximations via Chebyshev series: a user's guide	14
	3.4 Separating bounded and unbounded polynomial approximations	18
	3.5 Proof of Theorem 21	20
\mathbf{A}	Proofs of quantum signal processing	25
в	More applications of the CS decomposition	27
	B.1 Principal angles	27
	B.2 Jordan's lemma	28
С	Proof of Carlini's formula	29
D	Deferred proofs from Section 3.5	31

^{*}University of Washington, ewint@cs.washington.edu.

 $^{^{\}dagger}{\rm Microsoft\ Research,\ tiankevin@microsoft.com}$

1 Introduction

We present a "user-friendly guide" to understanding technical aspects of the quantum singular value transformation (QSVT), an elegant framework for designing quantum algorithms, particularly those that can be phrased as a linear algebraic task on a quantum state $|\psi\rangle$, viewed as a vector of amplitudes $\sum_{k=1}^{d} \psi_k |k\rangle$. This includes Hamiltonian simulation [LC17], i.e. preparing $e^{it\mathbf{H}} |\psi\rangle$ for a Hamiltonian **H**; quantum linear system solving [HHL09], i.e. preparing $\mathbf{A}^{-1} |\psi\rangle$ for a sparse matrix **A**; and quantum random walks [Sze04], i.e. approximating large powers of a Markov chain transition matrix or discriminating its singular values. QSVT was introduced by [LC19] in the Hermitian case and then generalized and subsequently popularized by a paper of Gilyén, Su, Low, and Wiebe [GSLW19] which demonstrates that these important, seemingly disparate quantum algorithms can be seen as consequences of a single unifying primitive.

Our aim is to expose the beauty of [GSLW19] and make it more accessible to an audience with a background in computer science, by simplifying or providing alternatives to its more mathematically dense proofs. This goal may be viewed as complementary to prior expositions of [GSLW19] such as [MRTC21], which focused on describing applications of QSVT to the design of quantum algorithms. In contrast, the aim of our work is to directly provide streamlined proofs or alternatives to the main technical results in [GSLW19].

1.1 Our contributions

QSVT via the CS decomposition. In Section 2, we give an alternate exposition of the *qubitization* technique given in [GSLW19, Section 3.2]. This technique lifts *quantum signal processing*, a product decomposition for computing bounded scalar polynomials, to QSVT, its matrix counterpart. In particular, QSVT implements quantum signal processing separately on each of the singular values of a "block encoded matrix" by mapping them through a polynomial transformation, while preserving the block encoding structure. Our exposition of QSVT is by way of the *Cosine-Sine* decomposition, a strengthening of Jordan's lemma [Jor75] that is more amenable to the computations in the proof of QSVT's correctness. We believe our proof strategy simplifies the exposition of QSVT in Section 3.2 of [GSLW19] (and other related expositions, e.g. Chapters 7 and 8 of [Lin22]). Specifically, the viewpoint we adopt elucidates the action of QSVT on the block structure of encoded matrices, removing much of the casework and eigenspace-by-eigenspace reasoning of prior expositions.

Bounded approximations via truncated Chebyshev series. In Section 3, we apply the technique of truncating *Chebyshev Series* to match or nearly-match all of the polynomial approximation results needed throughout [GSLW19, Section 5] through an arguably simpler framework. Our starting point is a classical theorem of Trefethen (Theorem 20) which bounds the error incurred by Chebyshev truncation for smooth functions. We derive, as a consequence of Trefethen's result, a new "bounded Chebyshev truncation" analog (Theorem 21) which applies to piecewise-smooth functions, and is compatible with the QSVT framework. Unlike its analog in the original work [GSLW19, Corollary 66], our Theorem 21 does not use Taylor series or Fourier series in its proof: the only approximation theory tools used are standard properties of Chebyshev polynomials. Our result is comparable to [GSLW19, Corollary 66]; as discussed in Remark 22, it loses a logarithmic factor in some regimes, but uses a weaker assumption on the function to be approximations to further provide a user's guide on how to apply Theorem 21 to derive bounded approximations to functions.

In Section 3.4, we give a new separation result lower bounding the degree of polynomials approximating the exponential function exp, under a boundedness requirement. Boundedness is crucial in QSVT applications (see Remark 11) due to the spectra of quantum gates. Notably, bounded approximations to exp have found use in designing Gibbs sampling techniques for quantum optimization [AG19; BGJST23]. We show that in parameter regimes of interest for these applications, a quadratically larger degree is required to achieve a bounded approximating polynomial. While lower bounds on QSVT have previously been demonstrated [GSLW19, Section 6]), our separation result is purely an approximation theory statement (independent of its use in QSVT), and its (simple) proof uses different techniques than [GSLW19]. We hope this result sheds light on when one can hope to obtain low-degree approximations for QSVT.

1.2 Notation

Matrices have bolded variable names, and \mathbf{I} is the identity matrix with dimension specified by context. For a matrix \mathbf{U} , \mathbf{U}^{\dagger} denotes its conjugate transpose. A square matrix \mathbf{U} is *unitary* if $\mathbf{U}^{\dagger}\mathbf{U} = \mathbf{U}\mathbf{U}^{\dagger} = \mathbf{I}$, and we call it "partitioned" if it has a block matrix structure with two row blocks and two column blocks (which is clear from context). When writing a matrix as blocks, an empty block denotes a zero block, and \cdot denotes that the block contains arbitrary entries. For brevity, we omit the dimensions of blocks when these sizes are not important to the computation: all block matrices occurring in products are compatible in the standard way.

The "computational basis" is the standard basis in \mathbb{C}^d . We denote $[d] \coloneqq \{1, 2, \ldots, d\}, i \coloneqq \sqrt{-1}$, and σ_z is the Pauli matrix $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. The maximum absolute value of a real function f on [a, b] is denoted $\|f\|_{[a,b]}$. We write a = b to mean there are universal constants $0 < C_1 \leq C_2$ with $C_1 a \leq b \leq C_2 a$.

2 QSVT and the Cosine-Sine decomposition

Briefly, whenever some aspect of a problem can be usefully formulated in terms of two-block by two-block partitions of unitary matrices, the CS decomposition will probably add insights and simplify the analysis. —Paige and Wei, [PW94]

2.1 Existence of the CS decomposition

We begin by proving the existence of the CS decomposition (CSD), a decomposition of a partitioned unitary matrix, following Paige and Wei [PW94]. We describe its application to the quantum singular value transformation (QSVT) in the following Sections 2.2, 2.3, and 2.4, wherein we give an alternate proof of the main QSVT result in [GSLW19] stated as Theorem 10.

The main idea of the CSD is that when a unitary matrix **U** is split into two-by-two blocks \mathbf{U}_{ij} for $i, j \in \{1, 2\}$, one can produce "simultaneous singular value decompositions (SVDs)" of the blocks, of the form $\mathbf{U}_{ij} = \mathbf{V}_i \mathbf{D}_{ij} \mathbf{W}_j^{\dagger}$.¹ If the reader cares just about the application to QSVT, they can read Theorem 1 and skip to Section 2.2.

For additional intuition on the CSD, we refer the reader to Appendix B, in which we derive principal angles and Jordan's lemma as consequences.

Theorem 1. Let $\mathbf{U} \in \mathbb{C}^{d \times d}$ be a unitary matrix, partitioned into blocks of size $\{r_1, r_2\} \times \{c_1, c_2\}$:

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{pmatrix}, \text{ where } \mathbf{U}_{ij} \in \mathbb{C}^{r_i \times c_j} \text{ for } i, j \in \{1, 2\}$$

Then, there exists unitary $\mathbf{V}_i \in \mathbb{C}^{r_i \times r_i}$ and $\mathbf{W}_j \in \mathbb{C}^{c_j \times c_j}$ for $i, j \in \{1, 2\}$ such that

$$\begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{V}_1 & \\ & \mathbf{V}_2 \end{pmatrix} \begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{W}_1 & \\ & \mathbf{W}_2 \end{pmatrix}^{\dagger},$$

where blanks represent zero matrices and $\mathbf{D}_{ij} \in \mathbb{R}^{r_i \times c_j}$ are diagonal matrices, possibly padded with zero rows or columns. Specifically, we can write

$$\mathbf{D} := \begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & | \mathbf{I} & | \\ \mathbf{C} & | \mathbf{S} & | \\ \mathbf{I} & | & \mathbf{0} \\ \mathbf{I} & | & \mathbf{0} \\ \mathbf{S} & | & -\mathbf{C} \\ \mathbf{0} & | & | & -\mathbf{I} \end{pmatrix}$$
(1)

where **I**, **C**, and **S** blocks are square diagonal matrices where **C** and **S** have entries in (0,1) on the diagonal, and **0** blocks may be rectangular.² Because **D** is unitary, we also have $\mathbf{C}^2 + \mathbf{S}^2 = \mathbf{I}$.

 $^{^{1}}$ In fact, there is some sense in which the SVD and the CSD are special cases of the same object, a *generalized Cartan*

decomposition. We recommend the survey by Edelman and Jeong for readers curious about this connection [EJ23]. ²Blocks may be non-existent. The I blocks may not necessarily be the same size, but C and S are the same size.

Remark 2. The form of **D** naturally induces decompositions $\mathbb{C}^d = \mathfrak{X}_0 \oplus \mathfrak{X}_C \oplus \mathfrak{X}_1$ and $\mathbb{C}^d = \mathfrak{Y}_0 \oplus \mathfrak{Y}_C \oplus \mathfrak{Y}_1$ into direct sums of three spaces. Hence, $\mathbf{D} : \mathbb{C}^d \to \mathbb{C}^d$ can be seen as a map $\mathbf{D} : \mathfrak{X}_0 \oplus \mathfrak{X}_C \oplus \mathfrak{X}_1 \to \mathfrak{Y}_0 \oplus \mathfrak{Y}_C \oplus \mathfrak{Y}_1$, such that **D** is a direct sum of three linear maps.

$$\begin{pmatrix} \mathbf{0} & \mathbf{I} & \mathbf{S} \\ \mathbf{I} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} \\ \mathbf{S} & -\mathbf{C} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \\ \mathbf{X}_0 \to \mathbf{y}_0 \end{pmatrix}}_{\mathcal{X}_0 \to \mathbf{y}_0} \oplus \underbrace{\begin{pmatrix} \mathbf{C} & \mathbf{S} \\ \mathbf{S} & -\mathbf{C} \\ \mathbf{X}_C \to \mathbf{y}_C \end{pmatrix}}_{\mathcal{X}_1 \to \mathbf{y}_1} \oplus \underbrace{\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix}}_{\mathcal{X}_1 \to \mathbf{y}_1}.$$

The key resulting intuition for QSVT is that, supposing everything is square, these blocks can be further decomposed into 2×2 blocks of the following rotation matrix form

$$\begin{pmatrix} \lambda_i & \sqrt{1-\lambda_i^2} \\ \sqrt{1-\lambda_i^2} & -\lambda_i \end{pmatrix}$$

from this representation, where $\{\lambda_i\}$ are the singular values of \mathbf{U}_{11} (see Lemma 14).

For completeness we now recall how to derive the CS decomposition from other common decompositions (namely, the SVD and QR decompositions), following the proof of [PW94].³

Proof of Theorem 1. We begin by considering $\mathbf{U}_{11} \in \mathbb{C}^{r_1 \times c_1}$. Let $\mathbf{V}_1 \mathbf{D}_{11} \mathbf{W}_1^{\dagger}$ be a SVD of \mathbf{U}_{11} , where $\mathbf{D}_{11} \in \mathbb{R}^{r_1 \times c_1}$ and \mathbf{V}_1 and \mathbf{W}_1 are square unitaries. Since $\|\mathbf{U}_{11}\|_{\text{op}} \leq \|\mathbf{U}\|_{\text{op}} = 1$, all its singular values are between zero and one, so we can specify that \mathbf{D}_{11} takes the form

$$\mathbf{D}_{11} = \begin{pmatrix} \mathbf{0} & & \\ & \mathbf{C} & \\ & & \mathbf{I} \end{pmatrix}$$

for a diagonal matrix \mathbf{C} with $0 < \mathbf{C}_{ii} < 1$ for all *i*. Now, take QR decompositions of $\mathbf{U}_{21}\mathbf{W}_1 \in \mathbb{C}^{r_2 \times c_1}$ and $\mathbf{U}_{12}^{\dagger}\mathbf{V}_1 \in \mathbb{C}^{c_2 \times r_1}$. These decompositions give unitaries \mathbf{V}_2 and \mathbf{W}_2 such that $\mathbf{D}_{21} \coloneqq \mathbf{V}_2^{\dagger}\mathbf{U}_{21}\mathbf{W}_1$ and $\mathbf{D}_{12}^{\dagger} \coloneqq \mathbf{W}_2^{\dagger}\mathbf{U}_{12}^{\dagger}\mathbf{V}_1$ are upper triangular with non-negative entries on the diagonal. By design, the \mathbf{V}_i and \mathbf{W}_j we have defined give us the decomposition

$$\begin{pmatrix} \mathbf{V}_1 & \\ & \mathbf{V}_2 \end{pmatrix}^{\dagger} \begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{W}_1 & \\ & \mathbf{W}_2 \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{V}_2^{\dagger} \mathbf{U}_{22} \mathbf{W}_2 \end{pmatrix}}_{\mathbf{D}}.$$

We claim that the blocks \mathbf{D}_{21} and \mathbf{D}_{12} satisfy the desired form from (1). This will (almost) be our final decomposition. We will only give the argument for \mathbf{D}_{21} ; the one for \mathbf{D}_{12} is similar. The columns of \mathbf{D} are orthonormal and \mathbf{D}_{21} is upper triangular with non-negative entries on its diagonal. So, all of the rightmost blocks of \mathbf{D}_{21} (under the I from \mathbf{D}_{11}) must be zero because of orthonormality, and further, the top-left block of \mathbf{D}_{21} must be I, using upper triangularity and non-negativity of the diagonal (inducting by row, the top-left entry is 1, forcing its row to be zeroes, and so on down the diagonal). Since the rows of \mathbf{D} are orthonormal, the I block in \mathbf{D}_{21} forces the block to its right to be the zero matrix. Finally, upper triangularity and column orthonormality forces the middle block to be non-negative diagonal with $\mathbf{S}^2 + \mathbf{C}^2 = \mathbf{I}$. The logic is displayed below:

$$\mathbf{D}_{21}=egin{pmatrix} ?&?&?\ ?&?&?\ &?&?\ &?&\end{pmatrix}
ightarrowegin{pmatrix} \mathbf{I}&?\ &?&\ &0\end{pmatrix}
ightarrowegin{pmatrix} \mathbf{I}&?\ &&0\end{pmatrix}
ightarrowegin{pmatrix} \mathbf{I}&&\ &&0\end{pmatrix}
ightarrowegin{pmatrix} \mathbf{I}&?\ &&0\end{pmatrix}
ightarrowegin{pmatrix} \mathbf{I}&?\ &&0\end{pmatrix}
ightarrowegin{pmatrix} \mathbf{I}&?\ &&0\end{pmatrix}
ightarrowegin{pmatrix} \mathbf{I}&\\ \mathbf{I}&\mathbf{I}&\mathbf{I}\\ \mathbf{I}&\mathbf{I}\end{pmatrix}
ightarrowegin{pmatrix} \mathbf{I}&&\ &&0\end{pmatrix}
ightarrowegin{pmatrix} \mathbf{I}&&\ &&0\\
ightarro$$

What we have argued so far suffices to show that (recalling \mathbf{D} is unitary)

$$\mathbf{D} = egin{pmatrix} \mathbf{0} & & \mathbf{I} & & \ \mathbf{C} & & \mathbf{S} & \ & \mathbf{I} & & \mathbf{0} \ \hline \mathbf{I} & & \mathbf{0} & & \ & \mathbf{S} & & ?_{11} & ?_{12} \ & & \mathbf{0} & ?_{21} & ?_{22} \end{pmatrix}.$$

³Our exposition will be slightly different, since Paige and Wei use a QR decomposition with a *lower* triangular matrix, where we use the more standard one with an *upper* triangular matrix.

For brevity we will only sketch the rest of the argument about the bottom-right block \mathbf{D}_{22} . First, $\mathbf{P}_{11} = -\mathbf{C}$ follows from unitarity of the following block of \mathbf{D} , which shows $\mathbf{CS} + \mathbf{SP}_{11} = 0$:

$$\begin{pmatrix} \mathbf{C} & \mathbf{S} \\ \mathbf{S} & ?_{11} \end{pmatrix}.$$

The blocks $?_{21}$ and $?_{12}$ must then be zero using unitarity, considering the fifth (block) row and column in **D**. Finally, $?_{22}$ is unitary and can be rotated to the (negative) identity by changing **W**₂:

$$\mathbf{W}_2 \leftarrow egin{pmatrix} \mathbf{I} & & \ & \mathbf{I} & \ & & -?_{22}^\dagger \end{pmatrix} \mathbf{W}_2.$$

In Appendix B, we demonstrate that for specific choices of \mathbf{U} , the form of the CS decomposition reveals interesting properties about the interactions between subspaces. In particular, the diagonals of \mathbf{C} and \mathbf{S} can naturally be seen as the cosines and sines of "principal angles" between subspaces. For those interested in applications to physics, these cosines and sines correspond to reflection and transmission probabilities in scattering theory, where this decomposition is known as the *polar decomposition* [MPK88; ML92; Bee97].

2.2 QSVT and quantum signal processing

We now apply the machinery of Section 2.1 to prove correctness of the QSVT framework of [GSLW19]. We first recall the situation treated by QSVT, requiring the notion of a block encoding.

Definition 3 (Variant of [GSLW19, Definition 43]). Given $\mathbf{A} \in \mathbb{C}^{r \times c}$, we say unitary $\mathbf{U} \in \mathbb{C}^{d \times d}$ is a block encoding of \mathbf{A} if there are $\mathbf{B}_{L,1} \in \mathbb{C}^{d \times r}$, $\mathbf{B}_{R,1} \in \mathbb{C}^{d \times c}$ with orthonormal columns such that $\mathbf{B}_{L,1}^{\dagger} \mathbf{U} \mathbf{B}_{R,1} = \mathbf{A}$. We denote $\mathbf{\Pi}_{L} = \mathbf{B}_{L,1} \mathbf{B}_{L,1}^{\dagger}$, $\mathbf{\Pi}_{R} = \mathbf{B}_{R,1} \mathbf{B}_{R,1}^{\dagger}$ to be the corresponding projections onto the spans of $\mathbf{B}_{L,1}$ and $\mathbf{B}_{R,1}$, respectively.

In other words, if **U** is a block encoding of **A**, then in the right basis, it has **A** as a submatrix.⁴ That is, if $\mathbf{B}_{L} = \begin{pmatrix} \mathbf{B}_{L,1} & \mathbf{B}_{L,2} \end{pmatrix}$ and $\mathbf{B}_{R} = \begin{pmatrix} \mathbf{B}_{R,1} & \mathbf{B}_{R,2} \end{pmatrix}$ are unitary completions of $\mathbf{B}_{L,1}$ and $\mathbf{B}_{R,2}$,

$$\mathbf{B}_{\mathrm{L}}^{\dagger}\mathbf{U}\mathbf{B}_{\mathrm{R}} = egin{pmatrix} \mathbf{A} & \cdot \ \cdot & \cdot \end{pmatrix} ext{ and } \mathbf{B}_{\mathrm{L}}^{\dagger}(\mathbf{\Pi}_{\mathrm{L}}\mathbf{U}\mathbf{\Pi}_{\mathrm{R}})\mathbf{B}_{\mathrm{R}} = egin{pmatrix} \mathbf{A} & \mathbf{0} \ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

In Section 2.3, we consider the special case when $\mathbf{B}_{L,1}$ and $\mathbf{B}_{R,1}$ are the first r and c columns of the identity, respectively. Under this restriction, the following statements about submatrices are true in the computational basis:

$$\mathbf{U} = \begin{pmatrix} \mathbf{A} & \cdot \\ \cdot & \cdot \end{pmatrix} \text{ and } \mathbf{\Pi}_{\mathrm{L}} \mathbf{U} \mathbf{\Pi}_{\mathrm{R}} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$
(2)

This simplification is for the purposes of exposition, since then \mathbf{U} is clearly a block matrix which we can apply the CS decomposition to. Indeed, it is without loss of generality: in Section 2.4, we recover general statements by unitary transformations which reduce to the special case above.

We now describe the QSVT framework, a lifting of a 2×2 matrix polynomial construction defined via "phase factors" (Definition 4), to higher dimensions. The construction in the 2×2 case is referred to in the literature as quantum signal processing (QSP). We recall the basics of QSP here.

Definition 4 (Quantum signal processing). A sequence of phase factors $\Phi = \{\phi_j\}_{0 \le j \le n} \in \mathbb{R}^{n+1}$ defines a quantum signal processing *circuit*⁵

$$\mathbf{QSP}(\Phi, x) \coloneqq \begin{pmatrix} e^{\imath\phi_0} & 0\\ 0 & e^{-\imath\phi_0} \end{pmatrix} \prod_{j=1}^n \underbrace{\begin{pmatrix} x & \sqrt{1-x^2} \\ \sqrt{1-x^2} & -x \end{pmatrix}}_{=:\mathbf{R}(x)} \underbrace{\begin{pmatrix} e^{\imath\phi_j} & 0\\ 0 & e^{-\imath\phi_j} \end{pmatrix}}_{e^{\imath\phi_j\sigma_z}}.$$
 (3)

⁴Methods for preparing block encodings often produce block encodings of $\frac{1}{\alpha}\mathbf{A}$ for some scaling factor α . This factor α appears in gate complexities of applications. Some authors parametrize this notion, e.g. by saying **U** is a (α, ε) -block encoding if **U** is a block encoding of some $\tilde{\mathbf{A}}$ such that $\|\tilde{\mathbf{A}} - \frac{1}{\alpha}\mathbf{A}\| \leq \frac{\varepsilon}{\alpha}$ in operator norm. ⁵We define QSP with the reflection operation $\mathbf{R}(x)$; a different convention is to use the rotation $e^{i \arccos(x)\sigma_x} = \frac{1}{\alpha}e^{-\frac{1}{\alpha}\mathbf{A}}$

 $[\]left(\frac{x}{\sqrt{1-x^2}}, \frac{\sqrt{1-x^2}}{x}\right)$, denoted W(x) in [GSLW19]. These two types of circuits are equivalent up to a shift in phase factors [MRTC21, Appendix A.2]. Using W(x) is perhaps more natural, since then this corresponds to alternating rotations in the σ_X and σ_Z basis.

Here and elsewhere, the product goes from 1 on the left-hand side to n on the right-hand side (by this convention, rotations are applied from ϕ_n to ϕ_0).

The idea of QSP is that we can perform a *known* polynomial, satisfying an achievability condition defined below, on an *unknown* (parameterized) operator via interleaved rotations.

Definition 5 (QSP-achievable polynomial). We say that a polynomial $p(x) \in \mathbb{C}[x]$ is QSP-achievable if there is a sequence of phase factors $\Phi = \{\phi_j\}_{0 \le j \le n} \in \mathbb{R}^{n+1}$ such that

$$\mathbf{QSP}(\Phi, x) = \begin{pmatrix} p(x) & \cdot \\ \cdot & \cdot \end{pmatrix}.$$
 (4)

Through elementary calculations, [GSLW19] gives a characterization of QSP-achievability; in particular, they show that every bounded, real polynomial p_{\Re} is achievable in the sense that there is a QSP-achievable polynomial whose real part is p_{\Re} , which we summarize here. For self-containedness, we give a proof in Appendix A, with a formal statement in Theorem 38.

Theorem 6 ([GSLW19, Corollary 10]). Let $p_{\Re}(x) \in \mathbb{R}[x]$ be a real polynomial of degree n. Then there exists a $p \in \mathbb{C}[x]$ such that $p_{\Re} = \Re(p)$ and p is QSP-achievable with some $\Phi = \{\phi_j\}_{0 \leq j \leq n} \in \mathbb{R}^{n+1}$ if and only if the following conditions hold:

- (a) p_{\Re} is even or odd;
- (b) $|p_{\Re}(x)| \leq 1$ for $x \in [-1, 1]$.

If p is QSP-achievable, then we can achieve $\Re(p)$ via a linear combination of unitary circuits on top of QSP; see Remark 11. We next introduce a generalization of Definition 4 to higher dimensions.

Definition 7 ([GSLW19, Definition 15]). The phased alternating sequence associated with a partitioned unitary **U** (following notation of Definition 12) and $\Phi = \{\phi_i\}_{0 \le j \le n} \in \mathbb{R}^{n+1}$ is

$$\mathbf{U}_{\Phi} \coloneqq \begin{cases} e^{\imath\phi_{0}(2\mathbf{\Pi}_{\mathrm{L}}-\mathbf{I})} \mathbf{U} e^{\imath\phi_{1}(2\mathbf{\Pi}_{\mathrm{R}}-\mathbf{I})} \prod_{j=1}^{\frac{n-1}{2}} \mathbf{U}^{\dagger} e^{\imath\phi_{2j}(2\mathbf{\Pi}_{\mathrm{L}}-\mathbf{I})} \mathbf{U} e^{\imath\phi_{2j+1}(2\mathbf{\Pi}_{\mathrm{R}}-\mathbf{I})} & \text{if } n \text{ is odd, and} \\ e^{\imath\phi_{0}(2\mathbf{\Pi}_{\mathrm{L}}-\mathbf{I})} \prod_{j=1}^{\frac{n}{2}} \mathbf{U}^{\dagger} e^{\imath\phi_{2j-1}(2\mathbf{\Pi}_{\mathrm{L}}-\mathbf{I})} \mathbf{U} e^{\imath\phi_{2j}(2\mathbf{\Pi}_{\mathrm{R}}-\mathbf{I})} & \text{if } n \text{ is even.} \end{cases}$$

Remark 8. The phased alternating sequence \mathbf{U}_{Φ} can be seen as a generalization of the quantum signal processing circuit $\mathbf{QSP}(\Phi, x)$. When d = 2 and r = c = 1, $2\mathbf{\Pi}_{\mathrm{L}} - \mathbf{I} = 2\mathbf{\Pi}_{\mathrm{R}} - \mathbf{I} = \boldsymbol{\sigma}_{z}$, so

$$\mathbf{QSP}(\Phi, x) = [\mathbf{R}(x)]_{\Phi} \text{ where } \mathbf{R}(x) = \begin{pmatrix} x & \sqrt{1-x^2} \\ \sqrt{1-x^2} & -x \end{pmatrix}.$$

Finally, we define the matrix polynomial we wish to target via the QSVT framework as follows.

Definition 9 ([GSLW19, Definition 16]). Let $f : \mathbb{R} \to \mathbb{C}$ be even or odd, and let $\mathbf{A} \in \mathbb{C}^{r \times c}$ have SVD $\mathbf{A} = \sum_{i \in [\min(r,c)]} \sigma_i u_i v_i^{\dagger}$. Then we define

$$f^{(\mathrm{SV})}(\mathbf{A}) = \begin{cases} \sum_{i \in [\min(r,c)]} f(\sigma_i) u_i v_i^{\dagger} & f \text{ is odd} \\ \sum_{i \in [c]} f(\sigma_i) v_i v_i^{\dagger} & f \text{ is even} \end{cases}$$

where σ_i is defined to be zero for $i > \min(r, c)$.

When f(x) = p(x) is an even or odd polynomial, $p^{(SV)}(\mathbf{A})$ can be written as a polynomial in the expected way, e.g. if $p(x) = x^2 + 1$, $p^{(SV)}(\mathbf{A}) = \mathbf{A}^{\dagger}\mathbf{A} + \mathbf{I}$ and if $p(x) = x^3 + x$, $p^{(SV)}(\mathbf{A}) = \mathbf{A}\mathbf{A}^{\dagger}\mathbf{A} + \mathbf{A}$. With this definition in hand, we are now ready to state the main result of the QSVT framework.

Theorem 10 ([GSLW19, Theorem 17]). Let partitioned unitary $\mathbf{U} \in \mathbb{C}^{d \times d}$ be a block encoding of **A**. Let $\Phi = \{\phi_j\}_{0 \le j \le n} \in \mathbb{R}^{n+1}$ be the sequence of phase factors such that $\mathbf{QSP}(\Phi, x)$ computes the

degree-n polynomial $p(x) \in \mathbb{C}[x]$, as in Definition 5. Then \mathbf{U}_{Φ} is a block encoding of $p^{(SV)}(\mathbf{A})$:

if p is odd,
$$\Pi_{\mathrm{L}} \mathbf{U}_{\Phi} \Pi_{\mathrm{R}} = \begin{pmatrix} p^{(\mathrm{SV})}(\mathbf{A}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = p^{(\mathrm{SV})}(\Pi_{\mathrm{L}} \mathbf{U} \Pi_{\mathrm{R}}),$$

and if p is even, $\Pi_{\mathrm{R}} \mathbf{U}_{\Phi} \Pi_{\mathrm{R}} = \begin{pmatrix} p^{(\mathrm{SV})}(\mathbf{A}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \Pi_{\mathrm{R}} p^{(\mathrm{SV})}(\Pi_{\mathrm{L}} \mathbf{U} \Pi_{\mathrm{R}}) \Pi_{\mathrm{R}}$

As we will see, when $\mathbf{B}_{L,1}$ and $\mathbf{B}_{R,1}$ are in the computational basis, the CS decomposition (Theorem 1) readily reduces the proof of Theorem 10 to substantially simpler subproblems (see Lemma 14).

Remark 11 (QSVT to quantum algorithms). Theorem 10 typically admits quantum algorithms in the following way. Suppose our goal is to apply \mathbf{A}^{-1} to a quantum state $|\psi\rangle$, where \mathbf{A} is a matrix with singular values in $[\frac{1}{\kappa}, 1]$ that we have in the block encoding \mathbf{U} . First, we take an odd, bounded polynomial $p(x) \in \mathbb{R}[x]$ such that $|p(x) - \frac{\kappa}{x}| \leq \varepsilon$ for $x \in [\frac{1}{\kappa}, 1]$. Then by Theorem 6, there is a phase sequence Φ which implements a $q \in \mathbb{C}[x]$ such that $\Re(q) = p$. By Theorem 10, \mathbf{U}_{Φ} is a block encoding of $q^{(SV)}(\mathbf{A})$ and a calculation shows that $\mathbf{U}_{-\Phi}$ is a block encoding of the same polynomial but with coefficients conjugated, $[q^*]^{(SV)}(\mathbf{A})$.

The circuit $(\mathbf{H} \otimes \mathbf{I})(|0\rangle\langle 0| \otimes \mathbf{U}_{\Phi} + |1\rangle\langle 1| \otimes \mathbf{U}_{-\Phi})(\mathbf{H} \otimes \mathbf{I})$ is a block encoding of $\frac{1}{2}(q^{(SV)}(\mathbf{A}) + [q^*]^{(SV)}(\mathbf{A})) = p^{(SV)}(\mathbf{A})$, and Corollary 19 of [GSLW19] shows that one can implement this circuit with controlled \mathbf{U} and \mathbf{U}^{\dagger} 's, along with other gates based on $\mathbf{\Pi}_{\mathrm{L}}$ and $\mathbf{\Pi}_{\mathrm{R}}$.

Equipped with a block encoding of $p^{(SV)}(\mathbf{A})$, we can accomplish our desired algorithmic task. To apply $p^{(SV)}(\mathbf{A})$ to an input state $|\psi\rangle \in \mathbb{C}^c$, we rotate $|\psi\rangle$ in d-dimensional state space to be aligned with the block in the block encoding. After applying \mathbf{U} and post-selecting, the output state is $p^{(SV)}(\mathbf{A}) |\psi\rangle$ normalized to have unit norm, and $\|p^{(SV)}(\mathbf{A}) - \mathbf{A}^{-1}\| \leq \varepsilon$ due to our choice of polynomial.⁶

2.3 Simplified QSVT in the computational basis

In this section, we provide a proof of Theorem 10 in the computational basis. We begin with some helpful notation in this special case, following the partitioning given by Theorem 1.

Definition 12 (Variant of [GSLW19, Definition 12]). Let $\mathbf{U} \in \mathbb{C}^{d \times d}$ be a block encoding of $\mathbf{A} \in \mathbb{C}^{r \times c}$ where $\mathbf{B}_{L,1}$ and $\mathbf{B}_{R,1}$ are the first r and c columns of the identity, respectively (see (2)). By Theorem 1, there is a CS decomposition compatible with the partitioning of \mathbf{U} :

$$\mathbf{U} = \begin{pmatrix} \mathbf{A} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{pmatrix}}_{\mathbf{V}} \underbrace{\begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{pmatrix}}_{\mathbf{D}} \underbrace{\begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix}^{\dagger}}_{\mathbf{W}^{\dagger}}.$$

In Definition 12, we applied Theorem 1 to obtain an SVD of $\mathbf{A} = \mathbf{V}_1 \mathbf{D}_{11} \mathbf{W}_1$ that we have extended to the *d*-dimensional U. Throughout the remainder of this section, $\mathbf{B}_L, \mathbf{B}_R, \mathbf{\Pi}_L$, and $\mathbf{\Pi}_R$ are defined consistently with the choice of $\mathbf{B}_{L,1}$ and $\mathbf{B}_{R,1}$ in Definition 12: $\mathbf{B}_L = \mathbf{B}_R = \mathbf{I}$, and $\mathbf{\Pi}_L$ and $\mathbf{\Pi}_R$ are the identity but with all but the first *r* and *c* 1's set to 0, respectively. We next observe that this SVD commutes appropriately with exponentiated projections respecting the partition.

Lemma 13 (Variant of [GSLW19, Lemma 14]). Let $\phi \in \mathbb{R}$. Following notation of Definition 12,

$$e^{\imath\phi(2\Pi_{\mathrm{L}}-\mathbf{I})} = \begin{pmatrix} e^{\imath\phi}\mathbf{I} \\ e^{-\imath\phi}\mathbf{I} \end{pmatrix}, \ e^{\imath\phi(2\Pi_{\mathrm{R}}-\mathbf{I})} = \begin{pmatrix} e^{\imath\phi}\mathbf{I} \\ e^{-\imath\phi}\mathbf{I} \end{pmatrix},$$

with appropriate block sizes,⁷ and

$$\begin{pmatrix} e^{i\phi}\mathbf{I} & \\ & e^{-i\phi}\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{V}_1 & \\ & \mathbf{V}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{V}_1 & \\ & \mathbf{V}_2 \end{pmatrix} \begin{pmatrix} e^{i\phi}\mathbf{I} & \\ & e^{-i\phi}\mathbf{I} \end{pmatrix},$$
$$\begin{pmatrix} \mathbf{W}_1 & \\ & \mathbf{W}_2 \end{pmatrix} \begin{pmatrix} e^{i\phi}\mathbf{I} & \\ & e^{-i\phi}\mathbf{I} \end{pmatrix} = \begin{pmatrix} e^{i\phi}\mathbf{I} & \\ & e^{-i\phi}\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{W}_1 & \\ & \mathbf{W}_2 \end{pmatrix}.$$

⁶A warning: when applying QSVT on an approximate block encoding of $\tilde{\mathbf{A}}$ with $\|\tilde{\mathbf{A}} - \mathbf{A}\| \le \varepsilon$, the error may not propagate as expected. This is because, if f satisfies $|f'| \le L$, $\|f^{(SV)}(\mathbf{A}) - f^{(SV)}(\tilde{\mathbf{A}})\| \le L \|\mathbf{A} - \tilde{\mathbf{A}}\|$ is not true in general, even up to constants. As Section 3.3 of [GSLW19] discusses, sometimes one must lose logarithmic factors here.

⁷These block sizes are such that the blocks of the product $e^{i\phi(2\Pi_{\rm L}-{\bf I})}\mathbf{U}e^{i\varphi(2\Pi_{\rm R}-{\bf I})}$ match. For example, if the top-left block of \mathbf{U} is $r \times c$, then the top-left block of $e^{i\phi(2\Pi_{\rm L}-{\bf I})}$ is $r \times r$ and the top-left block of $e^{i\phi(2\Pi_{\rm L}-{\bf I})}$ is $c \times c$.

We next state our main technical claim, whose proof is deferred to the end of the section.

Lemma 14. Let $\Phi \in \mathbb{R}^{n+1}$ be the sequence of phase factors implementing the degree-*n* polynomial $p(x) \in \mathbb{C}[x]$ via quantum signal processing (Definition 5). Then we can compute the corresponding QSVT circuit (Definition 7) applied to the following partitioned unitaries. First, the unitary associated with zero singular values.

$$\begin{bmatrix} \begin{pmatrix} \mathbf{0}^{r \times c} & \mathbf{I}_r \\ \mathbf{I}_c & \mathbf{0}^{c \times r} \end{bmatrix}_{\Phi} = \begin{pmatrix} p(0)\mathbf{I}_c & \cdot \\ \cdot & \cdot \end{pmatrix} \text{ for } n \text{ even and } \begin{pmatrix} \mathbf{0}^{r \times c} & \cdot \\ \cdot & \cdot \end{pmatrix} \text{ for } n \text{ odd.}$$
(5)

Next, the unitary associated with one singular values.

$$\begin{bmatrix} \begin{pmatrix} \mathbf{I}_r & \mathbf{0}^{r \times c} \\ \mathbf{0}^{c \times r} & -\mathbf{I}_c \end{bmatrix}_{\Phi} = \begin{pmatrix} p(1)\mathbf{I}_r & \cdot \\ \cdot & \cdot \end{pmatrix}.$$
 (6)

Finally, the unitary for intermediate singular values: let $\mathbf{C}, \mathbf{S} \in \mathbb{C}^{r \times r}$ be diagonal with $\mathbf{C}^2 + \mathbf{S}^2 = \mathbf{I}$.

$$\begin{bmatrix} \begin{pmatrix} \mathbf{C} & \mathbf{S} \\ \mathbf{S} & -\mathbf{C} \end{bmatrix}_{\Phi} = \begin{pmatrix} p^{(\mathrm{SV})}(\mathbf{C}) & \cdot \\ \cdot & \cdot \end{pmatrix}.$$
 (7)

Using this lemma, our main QSVT result (Theorem 10) in the setting of Definition 12 follows.

Proof of Theorem 10, special case. We recall the definition of \mathbf{U}_{Φ} :

$$\mathbf{U}_{\Phi} = \begin{cases} e^{i\phi_0(2\mathbf{\Pi}_{\mathrm{L}}-\mathbf{I})} \mathbf{U} e^{i\phi_1(2\mathbf{\Pi}_{\mathrm{R}}-\mathbf{I})} \prod_{\substack{j=1\\j=1}}^{\frac{n-1}{2}} \mathbf{U}^{\dagger} e^{i\phi_{2j}(2\mathbf{\Pi}_{\mathrm{L}}-\mathbf{I})} \mathbf{U} e^{i\phi_{2j+1}(2\mathbf{\Pi}_{\mathrm{R}}-\mathbf{I})} & \text{if } n \text{ is odd, and} \\ e^{i\phi_0(2\mathbf{\Pi}_{\mathrm{L}}-\mathbf{I})} \prod_{\substack{j=1\\j=1}}^{\frac{n}{2}} \mathbf{U}^{\dagger} e^{i\phi_{2j-1}(2\mathbf{\Pi}_{\mathrm{L}}-\mathbf{I})} \mathbf{U} e^{i\phi_{2j}(2\mathbf{\Pi}_{\mathrm{R}}-\mathbf{I})} & \text{if } n \text{ is even.} \end{cases}$$

Using that **V** and \mathbf{W}^{\dagger} from the CS decomposition $\mathbf{U} = \mathbf{V}\mathbf{D}\mathbf{W}^{\dagger}$ commute with their adjacent exponentiated reflections (Lemma 13), we continue:

$$= \begin{cases} \mathbf{V}e^{i\phi_0(2\mathbf{\Pi}_{\mathrm{L}}-\mathbf{I})}\mathbf{D}e^{i\phi_1(2\mathbf{\Pi}_{\mathrm{R}}-\mathbf{I})} \left(\prod_{j=1}^{\frac{n-1}{2}} \mathbf{D}^{\dagger}e^{i\phi_{2j}(2\mathbf{\Pi}_{\mathrm{L}}-\mathbf{I})}\mathbf{D}e^{i\phi_{2j+1}(2\mathbf{\Pi}_{\mathrm{R}}-\mathbf{I})}\right) \mathbf{W}^{\dagger} & \text{if } n \text{ is odd, and} \\ \mathbf{W}e^{i\phi_0(2\mathbf{\Pi}_{\mathrm{L}}-\mathbf{I})} \left(\prod_{j=1}^{\frac{n}{2}} \mathbf{D}^{\dagger}e^{i\phi_{2j-1}(2\mathbf{\Pi}_{\mathrm{L}}-\mathbf{I})}\mathbf{D}e^{i\phi_{2j}(2\mathbf{\Pi}_{\mathrm{R}}-\mathbf{I})}\right) \mathbf{W}^{\dagger} & \text{if } n \text{ is even.} \end{cases}$$
$$= \begin{cases} \mathbf{V}\mathbf{D}_{\Phi}\mathbf{W}^{\dagger} & \text{if } n \text{ is odd, and} \\ \mathbf{W}\mathbf{D}_{\Phi}\mathbf{W}^{\dagger} & \text{if } n \text{ is even.} \end{cases}$$
(8)

This reduces the problem to computing \mathbf{D}_{Φ} . Recall from Remark 2 that the structure of \mathbf{D} is

$$\begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{I} & \mathbf{S} \\ \mathbf{I} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} \\ \mathbf{S} & -\mathbf{C} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \\ \mathbf{X}_0 \to \mathbf{y}_0 \end{pmatrix} \oplus \underbrace{\begin{pmatrix} \mathbf{C} & \mathbf{S} \\ \mathbf{S} & -\mathbf{C} \\ \mathbf{X}_C \to \mathbf{y}_C \end{pmatrix} \oplus \underbrace{\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix}}_{\mathcal{X}_1 \to \mathbf{y}_1}.$$

Similarly, for $\phi \in \mathbb{R}$,

$$e^{i\phi(2\Pi_{\mathbf{L}}-\mathbf{I})} = \begin{pmatrix} e^{i\phi}\mathbf{I} & \\ \hline & e^{-i\phi}\mathbf{I} \end{pmatrix} = \underbrace{\begin{pmatrix} e^{i\phi}\mathbf{I} & \\ & e^{-i\phi}\mathbf{I} \end{pmatrix}}_{y_0 \to y_0} \oplus \underbrace{\begin{pmatrix} e^{i\phi}\mathbf{I} & \\ & e^{-i\phi}\mathbf{I} \end{pmatrix}}_{y_C \to y_C} \oplus \underbrace{\begin{pmatrix} e^{i\phi}\mathbf{I} & \\ & y_{1} \to y_{1} \end{pmatrix}}_{y_1 \to y_1},$$
$$e^{i\phi(2\Pi_{\mathbf{R}}-\mathbf{I})} = \begin{pmatrix} e^{i\phi}\mathbf{I} & \\ & e^{-i\phi}\mathbf{I} \end{pmatrix} = \underbrace{\begin{pmatrix} e^{i\phi}\mathbf{I} & \\ & e^{-i\phi}\mathbf{I} \end{pmatrix}}_{\chi_0 \to \chi_0} \oplus \underbrace{\begin{pmatrix} e^{i\phi}\mathbf{I} & \\ & e^{-i\phi}\mathbf{I} \end{pmatrix}}_{\chi_C \to \chi_C} \oplus \underbrace{\begin{pmatrix} e^{i\phi}\mathbf{I} & \\ & e^{-i\phi}\mathbf{I} \end{pmatrix}}_{\chi_1 \to \chi_1},$$

Leveraging this direct sum decomposition of **D**, applying Lemma 14 to each block yields

$$\mathbf{D}_{\Phi} = \begin{bmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \end{bmatrix}_{\Phi} \oplus \begin{bmatrix} \begin{pmatrix} \mathbf{C} & \mathbf{S} \\ \mathbf{S} & -\mathbf{C} \end{pmatrix} \end{bmatrix}_{\Phi} \oplus \begin{bmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix} \end{bmatrix}_{\Phi}$$
$$= \begin{cases} \begin{pmatrix} \mathbf{0} & \cdot \\ \cdot & \cdot \end{pmatrix} \oplus \begin{pmatrix} p^{(SV)}(\mathbf{C}) & \cdot \\ \cdot & \cdot \end{pmatrix} \oplus \begin{pmatrix} p(1)\mathbf{I} & \cdot \\ \cdot & \cdot \end{pmatrix} & \text{if } n \text{ is odd, and} \\ \begin{pmatrix} p(0)\mathbf{I} & \cdot \\ \cdot & \cdot \end{pmatrix} \oplus \begin{pmatrix} p^{(SV)}(\mathbf{C}) & \cdot \\ \cdot & \cdot \end{pmatrix} \oplus \begin{pmatrix} p(1)\mathbf{I} & \cdot \\ \cdot & \cdot \end{pmatrix} & \text{if } n \text{ is even.} \end{cases}$$

So, for n odd, recalling (8) and p(0) = 0, we have

$$\begin{split} \mathbf{\Pi}_{\mathrm{L}}\mathbf{U}_{\Phi}\mathbf{\Pi}_{\mathrm{R}} &= \mathbf{\Pi}_{\mathrm{L}}\mathbf{V}\mathbf{D}_{\Phi}\mathbf{W}^{\dagger}\mathbf{\Pi}_{\mathrm{R}} \\ &= \begin{pmatrix} \mathbf{I} & \\ & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{V}_{1} & \\ & \mathbf{V}_{2} \end{pmatrix} \mathbf{D}_{\Phi} \begin{pmatrix} \mathbf{W}_{1}^{\dagger} & \\ & \mathbf{W}_{2}^{\dagger} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \\ & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{V}_{1} & \\ & \mathbf{0} \end{pmatrix} \mathbf{D}_{\Phi} \begin{pmatrix} \mathbf{W}_{1}^{\dagger} & \\ & \mathbf{0} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\mathbf{V}_{1} \begin{pmatrix} \mathbf{0} & p^{(\mathrm{SV})}(\mathbf{C}) & \\ & p^{(1)}\mathbf{I} \end{pmatrix} \mathbf{W}_{1}^{\dagger} & \\ & & \end{pmatrix} = \begin{pmatrix} \frac{p^{(\mathrm{SV})}(\mathbf{A}) \mid \mathbf{0}}{\mathbf{0} \mid \mathbf{0}} \end{pmatrix}. \end{split}$$

Similarly, for n even, we have

$$\begin{split} \mathbf{\Pi}_{\mathrm{R}} \mathbf{U}_{\Phi} \mathbf{\Pi}_{\mathrm{R}} &= \mathbf{\Pi}_{\mathrm{R}} \mathbf{W} \mathbf{D}_{\Phi} \mathbf{W}^{\dagger} \mathbf{\Pi}_{\mathrm{R}} \\ &= \begin{pmatrix} \mathbf{W}_{1} & \\ & \mathbf{0} \end{pmatrix} \mathbf{D}_{\Phi} \begin{pmatrix} \mathbf{W}_{1}^{\dagger} & \\ & \mathbf{0} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\mathbf{W}_{1} \begin{pmatrix} p(\mathbf{0})\mathbf{I} & \\ & p^{(\mathrm{SV})}(\mathbf{C}) & \\ & & p(\mathbf{1})\mathbf{I} \end{pmatrix} \mathbf{W}_{1}^{\dagger} & \\ & & \end{pmatrix} = \begin{pmatrix} \frac{p^{(\mathrm{SV})}(\mathbf{A}) \mid \mathbf{0}}{\mathbf{0} \mid \mathbf{0}} \end{pmatrix}. \end{split}$$

We conclude the section by proving Lemma 14.

Proof of Lemma 14. The basic intuition behind this argument is that, by assumption and (3),

$$\begin{pmatrix} e^{i\phi_0} & 0\\ 0 & e^{-i\phi_0} \end{pmatrix} \prod_{j=1}^n \begin{pmatrix} x & \sqrt{1-x^2}\\ \sqrt{1-x^2} & -x \end{pmatrix} \begin{pmatrix} e^{i\phi_j} & 0\\ 0 & e^{-i\phi_j} \end{pmatrix} = \begin{pmatrix} p(x) & \cdot\\ \cdot & \cdot \end{pmatrix}$$

So, supposing we could evaluate the polynomial at a matrix $x \leftarrow \mathbf{C}$, we get that

$$\binom{e^{i\phi_{0}}\mathbf{I}}{e^{-i\phi_{0}}\mathbf{I}} \prod_{j=1}^{n} \begin{pmatrix} \mathbf{C} & \sqrt{\mathbf{I}-\mathbf{C}^{2}} \\ \sqrt{\mathbf{I}-\mathbf{C}^{2}} & -\mathbf{C} \end{pmatrix} \begin{pmatrix} e^{i\phi_{j}}\mathbf{I} \\ e^{-i\phi_{j}}\mathbf{I} \end{pmatrix} = \begin{pmatrix} p(\mathbf{C}) & \cdot \\ \cdot & \cdot \end{pmatrix} .$$

This should hold because block matrix multiplication operates by the same rules as scalar matrix multiplication, but requires care to handle the non-square case. We formalize this argument in an elementary manner. First, we consider (5). Let $\mathbf{U} = \begin{pmatrix} \mathbf{0} & \mathbf{I}_r \\ \mathbf{I}_c & \mathbf{0} \end{pmatrix}$. When *n* is even,

$$\begin{aligned} \mathbf{U}_{\Phi} &= \begin{pmatrix} e^{i\phi_{0}}\mathbf{I} \\ e^{-i\phi_{0}}\mathbf{I} \end{pmatrix} \prod_{j=1}^{\frac{n}{2}} \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix}^{\dagger} \begin{pmatrix} e^{i\phi_{2j-1}}\mathbf{I} \\ e^{-i\phi_{2j-1}}\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \begin{pmatrix} e^{i\phi_{2j}}\mathbf{I} \\ e^{-i\phi_{2j}}\mathbf{I} \end{pmatrix} \\ &= \begin{pmatrix} e^{i\phi_{0}}\mathbf{I} \\ e^{-i\phi_{0}}\mathbf{I} \end{pmatrix} \prod_{j=1}^{\frac{n}{2}} \begin{pmatrix} e^{i(\phi_{2j}-\phi_{2j-1})}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & e^{-i(\phi_{2j}-\phi_{2j-1})}\mathbf{I} \end{pmatrix} \\ &= \begin{pmatrix} e^{i\sum_{k=0}^{n}(-1)^{k}\phi_{k}}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & e^{-i\sum_{k=0}^{n}(-1)^{k}\phi_{k}}\mathbf{I} \end{pmatrix}. \end{aligned}$$

Taking **I** and **0** to be 1-dimensional scalars 1 and 0, this computation and Definition 4 also show that $p(0) = e^{i \sum_{k=0}^{n} (-1)^k \phi_k}$ yielding the desired conclusion. Similarly, when *n* is odd,

$$\begin{aligned} \mathbf{U}_{\Phi} &= \begin{pmatrix} e^{i\phi_{0}}\mathbf{I} \\ e^{-i\phi_{0}}\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \begin{pmatrix} e^{i\phi_{1}}\mathbf{I} \\ e^{-i\phi_{1}}\mathbf{I} \end{pmatrix} \\ & \prod_{j=1}^{\frac{n-1}{2}} \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix}^{\dagger} \begin{pmatrix} e^{i\phi_{2j}}\mathbf{I} \\ e^{-i\phi_{2j}}\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \begin{pmatrix} e^{i\phi_{2j+1}}\mathbf{I} \\ e^{-i\phi_{2j+1}}\mathbf{I} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{0} & e^{i\sum_{k=0}^{n}(-1)^{k}\phi_{k}}\mathbf{I} \\ e^{-i\sum_{k=0}^{n}(-1)^{k}\phi_{k}}\mathbf{I} & \mathbf{0} \end{pmatrix}. \end{aligned}$$

Next, we prove (6). Let $\mathbf{U} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_c \end{pmatrix}$. Notice that \mathbf{U} commutes with the other matrices in the expression \mathbf{U}_{Φ} . As an immediate consequence,

$$\mathbf{U}_{\Phi} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & (-1)^{n} \mathbf{I} \end{pmatrix} \prod_{k=0}^{n} \begin{pmatrix} e^{i\phi_{k}} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & e^{-i\phi_{k}} \mathbf{I} \end{pmatrix} = \begin{pmatrix} e^{i\sum_{k=0}^{n}\phi_{k}} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & (-1)^{n} e^{-i\sum_{k=0}^{n}\phi_{k}} \mathbf{I} \end{pmatrix}.$$

As before, the same computation specialized to a 2-dimensional $\mathbf{U} = \boldsymbol{\sigma}_z$ shows that $p(1) = e^{i \sum_{k=0}^{n} \phi_k}$, giving the desired claim. Finally to prove (7), let the diagonal entries of \mathbf{C} be $\{c_i\}_{i \in [r]}$. Then \mathbf{U} is the direct sum of r matrices of the form $\mathbf{R}(c_i)$, where we recall we defined \mathbf{R} in Definition 4. Applying Definition 4 to each 2×2 block, and comparing to Definition 9, yields the conclusion. \Box

2.4 Simplified QSVT in general bases

We now finish the proof of Theorem 10 in the case of general bases \mathbf{B}_{L} , \mathbf{B}_{R} . For disambiguation we let \mathbf{B}_{L} , \mathbf{B}_{R} , $\mathbf{B}_{L,1}$, $\mathbf{B}_{R,1}$, $\mathbf{\Pi}_{L}$, $\mathbf{\Pi}_{R}$ refer to an arbitrary basis and associated subspace in Definition 3, and (for this section only) we let $\overline{\mathbf{B}}_{L}$, $\overline{\mathbf{B}}_{R}$, $\overline{\mathbf{B}}_{L,1}$, $\overline{\mathbf{H}}_{R,1}$, $\overline{\mathbf{\Pi}}_{L}$, $\overline{\mathbf{\Pi}}_{R}$ refer to the computational basis where $\overline{\mathbf{B}}_{L,1}$ and $\overline{\mathbf{B}}_{R,1}$ have the same number of columns as $\mathbf{B}_{L,1}$ and $\mathbf{B}_{R,1}$. These are related via

$$\mathbf{B}_{\mathrm{L},1} = \mathbf{B}_{\mathrm{L}}\overline{\mathbf{B}}_{\mathrm{L},1}, \ \mathbf{\Pi}_{\mathrm{L}} = \mathbf{B}_{\mathrm{L}}\overline{\mathbf{\Pi}}_{\mathrm{L}}\mathbf{B}_{\mathrm{L}}^{\dagger}, \ \mathbf{B}_{\mathrm{R},1} = \mathbf{B}_{\mathrm{R}}\overline{\mathbf{B}}_{\mathrm{R},1}, \ \mathbf{\Pi}_{\mathrm{R}} = \mathbf{B}_{\mathrm{R}}\overline{\mathbf{\Pi}}_{\mathrm{R}}\mathbf{B}_{\mathrm{R}}^{\dagger}.$$
(9)

Finally, we define

$$\overline{\mathbf{U}} \coloneqq \mathbf{B}_{\mathrm{L}}^{\dagger} \mathbf{U} \mathbf{B}_{\mathrm{R}} \iff \mathbf{U} = \mathbf{B}_{\mathrm{L}} \overline{\mathbf{U}} \mathbf{B}_{\mathrm{R}}^{\dagger}.$$
(10)

We prove the general case by reducing to the special case of Theorem 10 we proved in the prior Section 2.3, as suggested earlier. The following observation will be useful:

$$\overline{\Pi}_{L}\overline{U}\overline{\Pi}_{R} = \mathbf{B}_{L}^{\dagger}\Pi_{L}U\Pi_{R}\mathbf{B}_{R}.$$
(11)

Proof of Theorem 10, general case. For simplicity we will only prove the odd case, as the reduction in the even case is essentially identical. Recall that when n is odd,

$$\begin{split} \mathbf{U}_{\Phi} &= e^{i\phi_{1}(2\mathbf{\Pi}_{\mathrm{L}}-\mathbf{I})}\mathbf{U}\prod_{j\in[\frac{n-1}{2}]}e^{i\phi_{2j}(2\mathbf{\Pi}_{\mathrm{R}}-\mathbf{I})}\mathbf{U}^{\dagger}e^{i\phi_{2j+1}(2\mathbf{\Pi}_{\mathrm{L}}-\mathbf{I})}\mathbf{U} \\ &= \mathbf{B}_{\mathrm{L}} \begin{pmatrix} e^{i\phi_{1}(2\overline{\mathbf{\Pi}}_{\mathrm{L}}-\mathbf{I})}\overline{\mathbf{U}}\prod_{j\in[\frac{n-1}{2}]}e^{i\phi_{2j}(2\overline{\mathbf{\Pi}}_{\mathrm{R}}-\mathbf{I})}\overline{\mathbf{U}}^{\dagger}e^{i\phi_{2j+1}(2\overline{\mathbf{\Pi}}_{\mathrm{L}}-\mathbf{I})}\overline{\mathbf{U}} \end{pmatrix} \mathbf{B}_{\mathrm{R}}^{\dagger} \\ &= \mathbf{B}_{\mathrm{L}} \begin{pmatrix} p^{(\mathrm{SV})}(\overline{\mathbf{A}}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{B}_{\mathrm{R}}^{\dagger}, \text{ where } \overline{\mathbf{A}} \coloneqq \overline{\mathbf{\Pi}}_{\mathrm{L}}\overline{\mathbf{U}}\overline{\mathbf{\Pi}}_{\mathrm{R}}. \end{split}$$

In the second line, we used the definitions (9), (10) and the fact that $\overline{\mathbf{B}}_{\mathrm{L}}$, $\overline{\mathbf{B}}_{\mathrm{R}}$ are unitary. Finally, in the third line we used the special case of Theorem 10 we proved earlier, applied to $\overline{\mathbf{U}}$. The conclusion follows by the claim

$$\mathbf{B}_{\mathrm{L}} \begin{pmatrix} p^{(\mathrm{SV})}(\overline{\mathbf{A}}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{B}_{\mathrm{R}}^{\dagger} = p^{(\mathrm{SV})}(\mathbf{A}), \text{ where } \mathbf{A} = \mathbf{\Pi}_{\mathrm{L}} \mathbf{U} \mathbf{\Pi}_{\mathrm{R}}.$$
(12)

Indeed, as a consequence of (11), **A** and $\overline{\mathbf{A}}$ have the same singular values, their left singular vectors are related via rotation by \mathbf{B}_{L} , and their right singular vectors are related via rotation by \mathbf{B}_{R} . Comparing with the definition of $p^{(\mathrm{SV})}$ from Definition 9 yields the claim (12).

3 QSVT and Chebyshev Series

So it becomes tempting to look at approximation methods that go beyond interpolation, and to warn people that interpolation is dangerous... the trouble with this is that for almost all the functions encountered in practice, Chebyshev interpolation works beautifully! —Trefethen, [Tre19]

3.1 Chebyshev polynomials

The QSVT framework gives a generic way of applying bounded polynomials to matrices. In applications of interest, the main goal is actually to apply a non-polynomial function; to capture these applications, it is important to develop tools for approximating the relevant functions with bounded polynomials. In this section, we introduce Chebyshev polynomials, our main tool for constructing approximations. We present only properties which are needed to achieve our results.

Definition 15 (Chebyshev polynomial). The degree-*n* Chebyshev polynomial (of the first kind), denoted $T_n(x)$, is the function that satisfies, for all $z \in \mathbb{C}$,

$$T_n(\frac{1}{2}(z+z^{-1})) = \frac{1}{2}(z^n+z^{-n}).$$

For $z = \exp(i\theta)$ for $\theta \in [-\pi, \pi]$ we may identify $x \coloneqq \frac{1}{2}(z + z^{-1})$ for $x = \cos \theta$. This identification yields another familiar definition of the Chebyshev polynomials,

$$T_n(\cos(\theta)) = \cos(k\theta).$$

From these definitions we have that $||T_n(x)||_{[-1,1]} \leq 1$, and that T_n has the same parity as n, i.e. $T_n(-x) = (-1)^n T_n(x)$. We can invert $x = \frac{1}{2}(z+z^{-1})$ to obtain $z = x \pm \sqrt{x^2 - 1}$. Consequently,

$$T_n(x) = \frac{1}{2} \left(\left(x + \sqrt{x^2 - 1} \right)^n + \left(x - \sqrt{x^2 - 1} \right)^n \right).$$

Definition 16 (Chebyshev coefficients). Let $f : [-1,1] \to \mathbb{C}$ be Lipschitz (i.e. $|f(x)-f(y)| \le C|x-y|$ for finite C). Then f has a unique decomposition into Chebyshev polynomials

$$f(x) = \sum_{k=0}^{\infty} a_k T_k(x),$$

where the Chebyshev coefficients a_k absolutely converge, and can be computed via the following integral (counterclockwise) around the complex unit circle, where πi is replaced by $2\pi i$ when k = 0:

$$a_k = \frac{1}{\pi i} \int_{|z|=1} z^{k-1} f(\frac{1}{2}(z+z^{-1})) \mathrm{d}z.$$
(13)

This comes from the Cauchy integral formula: $f(\frac{1}{2}(z+z^{-1}))$ maps [-1,1] (twice) onto the unit circle, so that when we write f as a Laurent series $\sum_{k\in\mathbb{Z}} b_k(z^k+z^{-k})$, the coefficients b_k match up with the coefficients a_k (up to a factor of two). For more details, see Theorem 3.1 of [Tre19]. We will typically construct polynomial approximations via *Chebyshev truncation*, defined as follows.

Definition 17 (Chebyshev truncation). For a function $f : [-1,1] \to \mathbb{C}$ written as a Chebyshev series $f(x) = \sum_{k=0}^{\infty} a_k T_k(x)$, we denote the degree-*n* Chebyshev truncation of *f* as

$$f_n(x) = \sum_{k=0}^n a_k T_k(x).$$

To construct polynomial approximations time-efficiently, we should avoid computing the integral (13). So, instead of the Chebyshev truncation, we would actually compute the *Chebyshev interpolant* [Tre19, Chapter 3], the degree-*n* polynomial that agrees with *f* at the n + 1 Chebyshev points $\cos(\frac{j\pi}{n})$ for $0 \le j \le n$. Computing this interpolant only requires n + 1 evaluations of *f* to specify, and $O(n \log(n))$ additional time to compute its Chebyshev coefficients, via a fast cosine transform. Chebyshev truncation and Chebyshev interpolation are closely related through standard bounds (see Eq. 4.9, [Tre19]); we focus on the latter as it is conceptually cleaner, but all bounds we prove extend to interpolation up to a factor of two.⁸

 $^{^{8}}$ This justifies our use of the quote from [Tre19] at the beginning of Section 3: though it is about Chebyshev interpolation rather than truncation, the spirit is the same.

3.2 Chebyshev series for standard functions

If the function f one wishes to approximate is standard, closed forms of the Chebyshev coefficients may be known, so one can take a Chebyshev truncation and explicitly bound the error:

$$\|f - f_n\|_{[-1,1]} = \left\|\sum_{k=n+1}^{\infty} a_k T_k(x)\right\|_{[-1,1]} \le \sum_{k=n+1}^{\infty} |a_k| \|T_k(x)\|_{[-1,1]} = \sum_{k=n+1}^{\infty} |a_k|.$$

In other words, by choosing n such that the coefficient tail sum is bounded by ε , we obtain an ε -uniform approximation on [-1, 1]. We list some standard Chebyshev coefficient series here, which can help for converting a Taylor or Fourier series to a Chebyshev series. The notation \sum' means that if a term exists for T_0 in the summation, it is halved:

$$x^{m} = 2^{1-m} \sum_{n=0}^{\lfloor m/2 \rfloor} {}' \binom{m}{n} T_{m-2n}(x).$$
 [MH02, Eq. 2.14]

$$e^{tx} = 2\sum_{n=0}^{\infty} {}^{\prime}I_n(t)T_n(x).$$
 [MH02, Eq. 5.18]

$$\sinh(tx) = 2\sum_{n=0}^{\infty} I_{2n+1}(t)T_{2n+1}(x).$$
 [MH02, Eq. 5.19]

$$\cosh(tx) = 2\sum_{n=0}^{\infty} {}^{\prime}I_{2n}(t)T_{2n}(x).$$
 [MH02, Eq. 5.20]

In the above display, I_n denotes the modified Bessel function of the first kind. This function is typically defined as the solution to a differential equation, but for us it suffices to define $I_n(t)$ as "the n^{th} Chebyshev coefficient of e^{tx} ," so by (13) (and pairing Laurent coefficients),

$$I_n(t) \coloneqq \frac{1}{2\pi i} \oint_{|z|=1} z^{n-1} e^{\frac{t}{2}(z+z^{-1})} dz = \frac{1}{2\pi i} \oint_{|z|=1} z^{-n-1} e^{\frac{t}{2}(z+z^{-1})} dz.$$

In the remainder of the section, to demonstrate the "direct coefficient bound" style of approximation error analysis, we analyze the use of Chebyshev truncation to give a uniform approximation to $f(x) = e^{tx}$ for $t \in \mathbb{R}$ on the interval [-1, 1]. The main result of this section is the following.

Theorem 18 ([SV14, Theorem 4.1], [GSLW19, Lemmas 57, 59]). Let $\varepsilon > 0$ and let p(x) be the degree-*n* Chebyshev truncation of e^{tx} for $t \in \mathbb{R}$. Then $\|p(x) - e^{tx}\|_{[-1,1]} \leq \varepsilon$ for

$$n \approx \begin{cases} |t| + \frac{\log(1/\varepsilon)}{\log(e + \log(1/\varepsilon)/|t|)} & \varepsilon \le 1\\ \sqrt{|t|\log(e^{|t|}/\varepsilon)} & \varepsilon > 1 \end{cases}.$$

To elaborate on what this bound states, there are four regimes of ε .

- 1. When $\varepsilon \ge e^{|t|}$, the zero polynomial $p(x) \equiv 0$ suffices.
- 2. When $1 \le \varepsilon < e^{|t|}$, we have $n = \sqrt{|t|\log(e^{|t|}/\varepsilon)}$, or $n = \sqrt{|t|\log(1/\delta)}$, rewriting $\varepsilon = \delta e^{|t|}$ to scale with the maximum value of $e^{|t|x}$ on [-1, 1]. This is the bound shown in [SV14].
- 3. When $e^{-C|t|} \leq \varepsilon < 1$ for some universal constant C, the scaling is n = |t|.
- 4. When $\varepsilon < e^{-C|t|}$, the scaling is $\frac{\log(1/\varepsilon)}{\log(e+\log(1/\varepsilon)/|t|)}$. This is the bound shown in [GSLW19].

Theorem 18 was first proven by combining results from [SV14] and [GSLW19]. A recent work [AA22] obtained the same upper bound, as well as a lower bound showing Theorem 18 is tight. The bounds from [SV14; GSLW19] are each loose in certain regimes ([SV14]'s bound, $\sqrt{|t|\log(e^{|t|}/\varepsilon)} + \log(e^{|t|}/\varepsilon)$, is loose in regime 4, whereas [GSLW19] assumes $\varepsilon < 1$), potentially due to the different proof techniques employed. Specifically, while [GSLW19] proceeds by a standard Bessel function inequality to bound the tail terms of the Chebyshev truncation, [SV14] proceeds by approximating monomials in the Taylor expansion of e^{tx} with Chebyshev truncation. As noted by [LC17], this strategy bounds the tail terms of the Chebyshev truncation by an easier-to-understand series that dominates it.

We give another (arguably more straightforward) proof of Theorem 18, by bounding the error of Chebyshev truncation (a strategy also employed by [AA22]). To achieve the right bound in the $\varepsilon > 1$ regime, we require a sharper bound on I_n , the Chebyshev coefficients of e^{tx} .

Lemma 19 (Carlini's formula). For $t \in \mathbb{R}$,

$$|I_n(t)| < \frac{\exp(\sqrt{t^2 + n^2})(\sqrt{(n/t)^2 + 1} - n/|t|)^n}{2(n^2 + t^2)^{1/4}}.$$

We contribute an independent proof of Lemma 19 in Appendix C, which was proven without using Bessel-style techniques in [AA22]. While we would guess that this bound is well-known, we could not find this statement in the literature.⁹ The equivalent bound on the (unmodified) Bessel function of the first kind J_k is due to Carlini [Wat95, Chapter 1.4], and can be viewed as a "real-valued" analog of Lemma 19 following the equivalence $I_k(t) = i^{-k}J_k(it)$.¹⁰ Our proof follows [Wat17] (who handled the real-valued version), and begins with a representation of a Bessel function as a contour integral. We bound this integral via the *method of steepest descent*, where the contour is changed to a real-valued one, using that the integrand is analytic. Using Lemma 19, we now prove Theorem 18.

Proof of Theorem 18. By symmetry it suffices to take $t \ge 0$. We split into cases based on ε . Case 1: $\varepsilon \le 1$. Define $r(t, \varepsilon)$ to be the value r such that $\varepsilon = (t/r)^r$. We choose

$$n = \lceil r(3t, \varepsilon) \rceil.$$

Note that the function $(t/r)^r$ is decreasing for $r \ge t$ and hence in the regime $\varepsilon \le 1$, we have $n \ge 3t$. Then, recalling that p is the degree-n Chebyshev truncation, we have by Lemma 19 that

$$\begin{split} \|p(x) - e^{tx}\|_{[-1,1]} &\leq 2\sum_{k=n+1}^{\infty} |I_k(t)| \leq 2\sum_{k=n}^{\infty} |I_k(t)| \\ &\leq 2\sum_{k=n}^{\infty} \frac{\exp(\sqrt{t^2 + k^2})(\sqrt{(k/t)^2 + 1} - k/t)^k}{2(k^2 + t^2)^{1/4}} \\ &\leq \sum_{k=n}^{\infty} \exp(\sqrt{t^2 + k^2})(\sqrt{(k/t)^2 + 1} - k/t)^k \\ &\leq \sum_{k=n}^{\infty} \exp(k\sqrt{(t/n)^2 + 1})(\sqrt{(n/t)^2 + 1} - n/t)^k \end{split}$$

by $k \ge n$ and since $\sqrt{x^2 + 1} - x$ decreases in x

$$\leq \sum_{k=n}^{\infty} \left(\exp(\sqrt{(t/n)^2 + 1}) \cdot \frac{t}{2n} \right)^k \qquad \qquad \text{by } \sqrt{1 - x^2} - x \leq \frac{1}{2x}$$

$$\leq \sum_{k=n} \left(\exp(\sqrt{10/9}) \cdot \frac{t}{2n} \right)^k \qquad \text{by } n \geq 3t$$

$$\leq \sum_{k=n}^{\infty} \left(\frac{3t}{2n}\right)^k \leq \left(\frac{3t}{n}\right)^n \sum_{k=1}^{\infty} \frac{1}{2^k} \leq \varepsilon \qquad \qquad \text{by } n \geq r(3t,\varepsilon).$$

The desired bound on n in this regime then follows from [GSLW19, Lemma 59] which shows that, for $\varepsilon \in (0, 1)$ and t > 0, $r(t, \varepsilon) = \Theta(t + \frac{\log(1/\varepsilon)}{\log(e + \log(1/\varepsilon)/t)})$.

Case 2: $\varepsilon > 1$. For $\varepsilon \in (1,2]$ the conclusion follows from our proof when $\varepsilon = 1$, so assume $\varepsilon > 2$,¹¹

⁹Off-the-shelf bounds like Kapteyn's inequality [DLMF, Eq. 10.14.8] are not quite tight enough for our purposes, since a very fine degree of control is necessary in the challenging regime where none of n, t, or t/n remain constant.

¹⁰Note that the real version of this statement is perhaps more non-trivial, since then the terms in the power series for the Bessel function are no longer nonnegative. Qualitatively similar statements may have also been made by Laplace, but for this claim Watson cites a book of the Mécanique Céleste without an English translation [Wat95, page 7]. This felt like a good place to stop our investigation of Bessel function bounds. ¹¹If t is a sufficiently small constant, then applying the $\varepsilon = 1$ case gives a constant-degree polynomial. Otherwise,

¹¹If t is a sufficiently small constant, then applying the $\varepsilon = 1$ case gives a constant-degree polynomial. Otherwise, t is sufficiently large to outweigh constant-factor changes in ε (and hence additive changes in $\log \frac{1}{\varepsilon}$).

for $\varepsilon \ge e^t$ the zero polynomial suffices so assume $\varepsilon < e^t$. Let $\delta = \frac{\varepsilon - 1}{5}$ and $m = \lceil 3t \rceil$. We choose

$$n = \left\lceil \sqrt{100t \left(t + \log \frac{1}{\delta}\right)} \right\rceil \ge 10\sqrt{t},$$

where we use $t + \log \frac{1}{\delta} \ge 1$ for our range of ε . The claim then follows from $\delta = \Theta(\varepsilon)$ and combining:

$$2\sum_{k=n+1}^{m-1} |I_k(t)| \le 5\delta, \ 2\sum_{k=m}^{\infty} |I_k(t)| \le 1.$$
(14)

The second claim in (14) was already shown by our earlier derivation setting $\varepsilon = 1$, since $m \ge r(3t, 1) = 3t$. For bounding the first sum, the following estimate will be helpful: for $k \le 3t$,

$$\exp\left(\sqrt{t^2 + k^2}\right) \left(\sqrt{\left(\frac{k}{t}\right)^2 + 1} - \frac{k}{t}\right)^k \leq \exp\left(\sqrt{t^2 + k^2} + k\left(\sqrt{\left(\frac{k}{t}\right)^2 + 1} - \frac{k}{t} - 1\right)\right)$$
$$= \exp\left(t\left(1 + \frac{k}{t}\right)\left(\sqrt{\left(\frac{k}{t}\right)^2 + 1} - \frac{k}{t}\right)\right)$$
$$\leq \exp\left(t\left(1 - 0.01\left(\frac{k}{t}\right)^2\right)\right).$$
(15)

The last equation used $(1+x)(\sqrt{1+x^2}-x) \le 1-0.01x^2$ for $0 \le x \le 3$. Since $m-1 \le 3t$,

$$\begin{split} \sum_{k=n+1}^{m-1} |I_k(t)| &\leq \frac{1}{2(n^2+t^2)^{1/4}} \sum_{k=n+1}^{m-1} \exp\left(t\left(1-0.01\left(\frac{k}{t}\right)^2\right)\right) \\ &= \frac{e^t}{2(n^2+t^2)^{1/4}} \sum_{k=n+1}^{m-1} \exp\left(-\frac{k^2}{100t}\right) \\ &\leq \frac{e^t}{2(n^2+t^2)^{1/4}} \int_n^\infty \exp\left(-\frac{x^2}{100t}\right) \mathrm{d}x \\ &= \frac{e^t\sqrt{25t}}{(n^2+t^2)^{1/4}} \int_{\frac{n}{\sqrt{100t}}}^\infty \exp\left(-x^2\right) \mathrm{d}x \\ &\leq \frac{e^t\sqrt{25t}}{(n^2+t^2)^{1/4}} \cdot \left(\frac{1}{2}\exp\left(-\frac{n^2}{100t}\right)\right) \leq \frac{5}{2}\exp\left(t-\frac{n^2}{100t}\right) \leq \frac{5}{2}\delta. \end{split}$$

The first line used Lemma 19 and (15), and the second-to-last used the Gaussian tail bound (20):

$$\int_{\frac{n}{\sqrt{100t}}}^{\infty} \exp(-x^2) dx < \exp\left(-\frac{n^2}{100t}\right) \cdot \frac{1}{1 + \frac{n}{\sqrt{100t}}} \le \frac{1}{2} \exp\left(-\frac{n^2}{100t}\right),$$

where we used $n \ge 10\sqrt{t}$ and $\exp(t - \frac{n^2}{100t}) \le \delta$ by construction.

3.3 Bounded approximations via Chebyshev series: a user's guide

Two issues often arise when using polynomial approximations for QSVT. First, we may not know explicitly what the Chebyshev coefficients of our desired function are. Second, even when we do, Chebyshev truncation may be bad for our purposes, since our criteria is *different* from uniform approximation on [-1, 1]. For example, quantum linear systems requires a polynomial approximation close to x^{-1} on $[-1, -\frac{1}{\kappa}] \cup [\frac{1}{\kappa}, 1]$, but it merely needs to be bounded on $[-\frac{1}{\kappa}, \frac{1}{\kappa}]$. This bounded requirement is necessary to use the machinery of Section 2.2 (see Remark 11).

Since quantum computing researchers are resourceful, we often see good polynomial approximations derived through ad hoc techniques to tame the function at poorly-behaved points. For example, [CKS17] performs Chebyshev truncation on the polynomial $(1 - (1 - x^2)^b) \cdot \frac{1}{x}$ instead of $\frac{1}{x}$, and this has the desired properties. However, as [GSLW19] points out, there are generic ways to find approximations to piecewise smooth functions which satisfy the " ε -close on smooth pieces, but bounded near points of discontinuity" requirement of QSVT, with log $\frac{1}{\varepsilon}$ scaling in the degree. Their proof of this claim assumes that the Taylor series coefficients of the smooth functions are bounded.

We give a variant of this result that only assumes that the smooth functions are bounded on ellipses in complex space, and has a self-contained proof based on Chebyshev series. This alternate version matches [GSLW19] for sufficiently small ε , and otherwise loses a logarithmic factor (though under a weaker assumption on the function to be approximated, see Remark 22). To do so, we combine a powerful meta-technique for bounding the Chebyshev coefficients of analytic functions with applications of explicit thresholding functions. This meta-technique is stated as Theorem 20; in many prominent settings, a direct application already yields near-optimal polynomial approximations.

Theorem 20 ([Tre19, Theorems 8.1 and 8.2]). Let f be an analytic function in [-1,1] and analytically continuable to the interior of the Bernstein ellipse $E_{\rho} = \{\frac{1}{2}(z+z^{-1}): |z|=\rho\}$, where it satisfies $|f(x)| \leq M$. Then its Chebyshev coefficients satisfy $|a_0| \leq M$ and $|a_k| \leq 2M\rho^{-k}$ for $k \geq 1$. Consequently, for each $n \geq 0$, its Chebyshev projections satisfy

$$||f - f_n||_{[-1,1]} \le \frac{2M\rho^{-n}}{\rho - 1},$$

and choosing $n = \lceil \frac{1}{\log(\rho)} \log \frac{2M}{(\rho-1)\varepsilon} \rceil$, we have $||f - f_n||_{[-1,1]} \le \varepsilon$.

Proof. Recall from (13) (and since inverting z does not change the contour integral) that for $k \ge 1$,

$$a_k = \frac{1}{\pi i} \int_{|z|=1} z^{-(k+1)} f(\frac{1}{2}(z+z^{-1})) \mathrm{d}z.$$

The boundary of E_{ρ} is given by $\frac{1}{2}(z+z^{-1})$ for $|z|=\rho$, and f is analytic in E_{ρ} , so we may choose a different contour without affecting the value of the integral:

$$a_k = \frac{1}{\pi i} \int_{|z|=\rho} z^{-(k+1)} f(\frac{1}{2}(z+z^{-1})) \mathrm{d}z.$$

The conclusion follows from the facts that the circumference of $|z| = \rho$ is $2\pi\rho$ and the function is bounded by M. A similar argument gives the case k = 0, where (13) has $2\pi i$ in the denominator. \Box

Theorem 20 shows that if one can analytically continue f to a Bernstein ellipse with $\rho = 1 + \alpha$ for small α , then a degree $\approx \frac{1}{\alpha}$ polynomial obtains good approximation error on [-1, 1]. Unfortunately, since the approximation in Theorem 20 is based on Chebyshev truncation, the approximation rapidly blows up outside the range [-1, 1] (in Lemma 32, we give estimates on the growth of Chebyshev polynomials, i.e. that the n^{th} polynomial grows as $O(|x|^n)$ for x sufficiently outside [-1, 1]). In interesting applications of the QSVT framework, this is an obstacle. For example, to use QSVT for matrix inversion, we need a polynomial approximation to x^{-1} on $[\delta, 1]$ that is bounded on [-1, 1]. Upon linearly remapping $[\delta, 1]$ to [-1, 1], this corresponds to a bounded approximation on [-b, 1] for some b > 1, so Chebyshev truncations give us a very poor degree of control.

To this end, we provide the following "bounded approximation" variant of Theorem 20, as a user-friendly way of extending it to applications of the QSVT framework.

Theorem 21. Let f be an analytic function in [-1, 1] and analytically continuable to the interior of E_{ρ} where $\rho = 1 + \alpha$, where it is bounded by M. For $\delta \in (0, \frac{1}{C} \min(1, \alpha^2))$ where C is a sufficiently large constant, $\varepsilon \in (0, 1)$, and b > 1, there is a polynomial q of degree $O(\frac{b}{\delta} \log \frac{b}{\delta \varepsilon})$ such that

$$\begin{split} \|f - q\|_{[-1,1]} &\leq M\varepsilon, \\ \|q\|_{[-(1+\delta),1+\delta]} &\leq M, \\ \|q\|_{[-b,-(1+\delta)]\cup[1+\delta,b]} &\leq M\varepsilon. \end{split}$$

Proof sketch. We give a formal proof in Section 3.5, but briefly summarize our proof strategy here.

- 1. Applying Theorem 20 gives f_n of degree $n \approx \frac{1}{\alpha}$ approximating f in the interval [-1, 1], but f_n does not satisfy the other required conclusions due to its growth outside [-1, 1].
- 2. We multiply f_n by a "threshold" r based on the Gaussian error function erf, whose tails decay much faster than the Chebyshev polynomials grow outside [-1, 1]. Our function r has the property that inside [-1, 1], it is close to 1, and outside $[-(1 + \delta), 1 + \delta]$, it is close to 0.

3. Using bounds on the growth of erf, we show $r \cdot f_n$ is bounded on a Bernstein ellipse of radius $1 + \frac{\delta}{h}$ appropriately rescaled, and applying Theorem 20 once more gives the conclusion.

The final proof requires some care to obtain the claimed scalings on the windows of approximation, but we include this tedium to make the theorem statement as simple to use as possible. \Box

Remark 22. Theorem 21 is an alternative to Corollary 66 of [GSLW19]. Translating the statement there to our setting, the polynomial approximation it would achieve has degree $O(\frac{b}{\delta} \log \frac{M}{\varepsilon})$. Our approximation of degree $O(\frac{b}{\delta} \log \frac{b}{\delta\varepsilon})$ is comparable, matching when ε and $\frac{\delta}{b}$ are polynomially related.

We note that our Theorem 21 has a $\log \frac{b}{\delta \varepsilon}$ dependence (instead of $\log \frac{1}{\varepsilon}$) because we use a slightly weaker type of assumption: not only does [GSLW19, Corollary 66] assume that its function $f(x) = \sum_{k=0}^{\infty} a_k x^k$ is analytic and bounded by M on a disk of radius $1 + \delta$, it also assumes that the Taylor series coefficients $|a_k|$ satisfy $\sum_{k=0}^{\infty} |a_k| (1 + \delta)^k \leq M$. Without this final condition, boundedness merely implies that $|a_k| = O((1+\delta)^{-k})$, and this slight weakening leads to an additional logarithmic factor when ε is large. In most applications, the difference is negligible; both strategies have an additional polynomial overhead on $\frac{b}{\delta}$, which typically dominates a $\log \frac{b}{\delta}$ dependence.

Finally, the precondition of Theorem 21 is weaker than the requirement of Corollary 66 of [GSLW19] in another sense. Specifically, [GSLW19] assumes a locally bounded Taylor series in a scaled unit circle in the complex plane, whereas we only require a bound on a (potentially much smaller) Bernstein ellipse, which could enable more applications.

We use the rest of the section to provide a user's guide on applying Theorem 21 to boundedly approximate various piecewise smooth functions. All of our applications proceed as follows.

- 1. We linearly rescale the "region of interest," i.e. the part of \mathbb{R} where we wish to approximate a function via bounded polynomials, to the interval [-1, 1].
- 2. We apply Theorem 21 to the rescaled function for appropriate choices of b and δ , so the region where the approximation must be bounded is captured upon undoing the rescaling.
- 3. If additional properties of the bounded approximation are desired, e.g. a parity requirement, we use the additional implications of Theorem 21 to obtain these properties.

A simple application of Theorem 21 is obtaining degree- $O(\frac{1}{\delta} \log \frac{1}{\delta \varepsilon})$ polynomial approximations to the sign and rectangle functions (where our guarantee is ε -closeness outside of a δ interval around the points of discontinuity, as described in [GSLW19, Lemmas 25 and 29]).¹² We leave this as an exercise. We begin with a bounded approximation to the rescaled exponential function in Corollary 23. Such bounds have previously seen use in quantum applications of the multiplicative weights framework via QSVT, to design faster approximate solvers for linear programs [AG19; BGJST23].

Corollary 23. Let $\varepsilon \in (0, 1)$, and let $f(x) = \exp(\beta x)$ for $\beta \ge 1$. There exists a polynomial p of degree $O(\beta \log \frac{\beta}{\varepsilon})$ such that $\|p\|_{[-1,1]} = O(1)$ and $\|p - f\|_{[-1,0]} \le \varepsilon$.

Proof. First, we rescale so the region of interest is [-1,1]: let $g(y) = f(\beta(\frac{1}{2}(y-1)))$. Note that g(y) is analytic everywhere and bounded by a constant on E_{ρ} for $\rho = 1 + \beta^{-\frac{1}{2}}$. To see this, the magnitude of $\frac{1}{2}(z-1)$ for $z \in E_{\rho}$ is maximized when z is furthest from 1, and Fact 30 shows this magnitude is $O(\frac{1}{\beta})$. Hence, applying Theorem 21 with b = 3 and $\delta = \Theta(\frac{1}{\beta})$ for a sufficiently small constant yields the claim upon shifting the region of interest back, since $\frac{1}{2}(y-1) = 1$ for y = 3. \Box

Next, in Corollary 24 we provide an analog of Lemma 70 in [GSLW19], regarding the bounded approximation of arcsin, using the framework of Theorem 21.

Corollary 24. Let $\delta, \varepsilon \in (0, 1)$, and let $f(x) = \frac{2}{\pi} \arcsin(x)$. There exists an odd polynomial p(x) of degree $O(\frac{1}{\sqrt{\delta}} \log \frac{1}{\delta \varepsilon})$ such that $\|p\|_{[-1,1]} \leq 1$ and $\|p - f\|_{[-(1-\delta),1-\delta]} \leq \varepsilon$.

¹²This result shows that direct Chebyshev truncation is sometimes not enough for these slightly different approximation guarantees: the sign function has Chebyshev series $\sum_{k\geq 0} \frac{4}{\pi} \frac{(-1)^k}{2k+1} T_{2k+1}(x)$ [Tre19, Exercise 3.6], which cannot be truncated without paying $\Omega(1)$ error.

Proof. First, we rescale so the region of interest is [-1, 1]: let $\overline{\operatorname{arcsin}}(x) = \operatorname{arcsin}((1 - \delta)x)$. The arcsin function is analytic on $\mathbb{C} \setminus ((-\infty, 1] \cup [1, \infty))$, so we choose $\rho = 1 + \sqrt{2\delta}$ so that $\overline{\operatorname{arcsin}}(x)$ is analytic on the interior of E_{ρ} by the first bound in Fact 30. By the maximum modulus principle, the maximum of $\overline{\operatorname{arcsin}}$ is achieved on the boundary of the ellipse. We can bound this using that, for $|z| \leq 1$ (so the Taylor series [DLMF, Eq. 4.24.1] converges),

$$|\arcsin z| = \left|\sum_{n=0}^{\infty} \frac{(2n)!}{2^{2n}(n!)^2} \frac{z^{2n+1}}{2n+1}\right| \le \sum_{n=0}^{\infty} \frac{(2n)!}{2^{2n}(n!)^2} \frac{|z|^{2n+1}}{2n+1} \le \arcsin|z| \le \frac{\pi}{2}.$$
 (16)

We can further verify by Fact 30 that $|z| \leq 1 + \delta$ for $z \in E_{\rho}$, so the above display yields

$$\left|\overline{\operatorname{arcsin}}(z)\right| = \left|\operatorname{arcsin}((1-\delta)z)\right| \le \frac{\pi}{2}.$$

So, by Theorem 21 with $b \leftarrow \frac{1}{1-\delta}$ and $\delta \leftarrow \frac{\delta}{C}$ for a sufficiently large C, there is a polynomial q with $\|q - \overline{\arcsin}\|_{[-1,1]} \leq \frac{\pi}{2}\varepsilon$ and $\|q\|_{[-(1-\delta)^{-1},(1-\delta)^{-1}]} \leq \frac{\pi}{2}$. Letting $p((1-\delta)x) = \frac{2}{\pi}q(x)$, we have the desired bounds. The degree of p is $O(\frac{1}{\delta}\log\frac{1}{\delta\varepsilon})$, and it is odd by Corollary 35 as $\overline{\arcsin}$ is odd. \Box

In Corollary 25, we further apply Theorem 21 to the "fractional query" setting of Corollary 72 in [GSLW19], which requires approximations to $\exp(it \arcsin(x))$ for small t. As in [GSLW19], we provide bounded approximations to $\cos(t \arcsin(x))$ and $\sin(t \arcsin(x))$ through our framework.

Corollary 25. Let $\varepsilon \in (0,1)$ and $t \in [-1,1]$. There exists an even polynomial p and an odd polynomial q of degree $O(\log \frac{1}{\varepsilon})$ such that $\|p\|_{[-1,1]} \leq 1$, $\|q\|_{[-1,1]} \leq 1$, and

$$\|p(x) - \cos(t \arcsin(x))\|_{[-\frac{1}{2}, \frac{1}{2}]} \le \varepsilon, \ \|q(x) - \sin(t \arcsin(x))\|_{[-\frac{1}{2}, \frac{1}{2}]} \le \varepsilon.$$

Proof. First, we rescale so the region of interest is [-1,1]: let $f(x) = \cos(t \operatorname{arcsin}(\frac{x}{2}))$ and $g(x) = \sin(t \operatorname{arcsin}(\frac{x}{2}))$. These are analytic on $\mathbb{C} \setminus ((-\infty, -2] \cup [2, \infty))$, since that is where $\operatorname{arcsin}(\frac{x}{2})$ is analytic. Let $\rho = 2$, so f and g are analytic on the interior of E_{ρ} . We observe that for all $z \in \mathbb{C}$,

$$|\cos(z)| = \frac{1}{2} \left| e^{\imath z} + e^{-\imath z} \right| \le \frac{1}{2} |e^{\imath z}| + \frac{1}{2} |e^{-\imath z}| \le \cosh(|z|),$$

as cosh is increasing and the imaginary part of z is at most |z|. A similar argument shows $|\sin(z)| \leq \cosh(|z|)$. By Fact 30 we observe that every point in the interior of $\frac{1}{2}E_{\rho}$ has modulus $\leq \frac{3}{4}$, and $|\arcsin|$ is bounded in this region by $\frac{\pi}{2}$ (see (16)), so for $z \in E_{\rho}$, $|f(z)| = |\cos(t \arcsin(\frac{z}{2}))| \leq \cosh(\frac{\pi}{2})$, and we may analogously bound g on E_{ρ} . Taking b = 2 and δ to be a sufficiently small constant in Theorem 21, and rescaling the region of interest, gives the conclusion. The parities of p and q follow from Corollary 35 and the parities of $\cos(t \arcsin(x))$ and $\sin(t \arcsin(x))$.

Finally, Corollary 26 gives a variant of Corollaries 67 and 69 in [GSLW19], regarding the bounded approximation of negative power functions. Our bound has a slightly worse logarithmic factor in some regimes (as discussed in Remark 22), but otherwise agrees with the bounds in [GSLW19] up to a constant factor, using arguably a more standard approach.

Corollary 26. Let $\delta, \varepsilon \in (0, 1)$, and let $f(x) = |\frac{\delta}{x}|^c$ for c > 0. There exist both even and odd polynomials p(x) of degree $O(\frac{\max(1,c)}{\delta} \log \frac{1}{\delta \varepsilon})$ such that $\|p\|_{[-1,1]} \leq 3$ and $\|p - f\|_{[\delta,1]} \leq \varepsilon$.

Proof. Assume δ is sufficiently small, else taking a smaller δ only affects the bound by a constant. We rescale the region of interest: $x = \frac{1-\delta}{2}y + \frac{1+\delta}{2}$ is in $[\delta, 1]$ for $y \in [-1, 1]$, so let

$$g(y) \coloneqq \delta^c \Big(\frac{1-\delta}{2} y + \frac{1+\delta}{2} \Big)^{-c}.$$

We require a bound of g on E_{ρ} for $\rho = 1 + \sqrt{\delta/4 \max(1, c)}$. Since f is largest closest to the origin,

g is largest at the point closest to $-\frac{1+\delta}{1-\delta}$, i.e. $-\frac{1}{2}(\rho+\rho^{-1}) > -(1+\frac{\delta}{8\max(1,c)})$ by Fact 30. Further,

$$g\left(-\frac{1}{2}(\rho+\rho^{-1})\right) \leq g\left(-\left(1+\frac{\delta}{8\max(1,c)}\right)\right)$$
$$\leq \delta^{c}\left(-\frac{1-\delta}{2}\left(1+\frac{\delta}{8\max(1,c)}\right) + \frac{1+\delta}{2}\right)^{-c}$$
$$= \left(1-\frac{1-\delta}{16\max(1,c)}\right)^{-c} \leq \frac{3}{2}.$$

Let $\widetilde{\delta} = \frac{\delta}{4C \max(1,c)}$ for sufficiently large C, and b = 4. Theorem 21 yields q(y) satisfying:

$$\|q(y) - g(y)\|_{[-1,1]} \le \varepsilon, \ \|q(y)\|_{[-(1+\widetilde{\delta}),1+\widetilde{\delta}]} \le 2^c, \ \|q(y)\|_{[-4,-(1+\widetilde{\delta})]\cup[1+\widetilde{\delta},4]} \le \varepsilon.$$

Shifting back $y = \frac{2}{1-\delta}(x - \frac{1+\delta}{2})$, it is clear for sufficiently large C that $y = -\frac{1+3\delta}{1-\delta}$ (which corresponds to $x = -\delta$) has $y < -(1 + \tilde{\delta})$, and $y = -\frac{3+\delta}{1-\delta}$ (which corresponds to x = -1) has y > -4. So,

$$\left\|q\left(\frac{2}{1-\delta}\left(x-\frac{1+\delta}{2}\right)\right) - f(x)\right\|_{[\delta,1]} \le \varepsilon,$$

$$\left\|q\left(\frac{2}{1-\delta}\left(x-\frac{1+\delta}{2}\right)\right)\right\|_{[-\delta,\delta]} \le 2^{c},$$

$$\left\|q\left(\frac{2}{1-\delta}\left(x-\frac{1+\delta}{2}\right)\right)\right\|_{[-1,-\delta]} \le \varepsilon.$$
(17)

Depending on whether we wish the final function to be even or odd, we take

$$p(x) = q\left(\frac{2}{1-\delta}\left(x-\frac{1+\delta}{2}\right)\right) \pm q\left(\frac{2}{1-\delta}\left(-x-\frac{1+\delta}{2}\right)\right).$$

Then the guarantees of (17) give $\|p(x) - f(x)\|_{[\delta,1]} \leq 2\varepsilon$ and $\|p(x)\|_{[-1,1]} \leq 3$, and we rescale ε to conclude. The final degree of the polynomial is the degree of q(y): $O(\frac{\max(1,c)}{\delta\varepsilon}\log\frac{1}{\delta\varepsilon})$.

3.4 Separating bounded and unbounded polynomial approximations

In this section, we show that QSVT's requirement that the polynomials it implements be bounded can worsen the quality of approximations. Specifically, we prove a simple separation result which shows that polynomial approximations may necessarily require larger degree under an additional boundedness constraint. This follows from the observation that bounded degree-d polynomials can have derivative as large as d^2 near the boundary of [-1, 1], yet are bounded by O(d) on the interior. This is formalized by the following classical inequality due to Bernstein [Ber12].

Proposition 27 (Theorem 2, [Sch41]). Let p be a degree-d polynomial with rational coefficients satisfying $|p(x)| \leq 1$ for all $x \in [-1, 1]$. Then

$$d \ge |p'(x)|\sqrt{1-x^2}$$
 for all $x \in [-1,1]$.

Proposition 27 is troublesome for obtaining the type of bound we want since it depends on derivatives of p, the approximation, rather than f, the function to be approximated. We next give a simple extension of Proposition 27, with degree lower bounds depending on a quantity $\sup_{x,y} \frac{|f(x)-f(y)|-2\varepsilon}{|x-y|}$ which can be viewed as a "robust" Lipschitz constant of f. For example, if f is a differentiable function with derivative $\geq L$ on an interval of length at least $\frac{4\varepsilon}{L}$, then this quantity is $\geq \frac{L}{2}$, and taking $\varepsilon \to 0$ recovers the maximum derivative of f.

Proposition 28. For $\Delta \in (0,1]$ and $S \subset [-\Delta, \Delta]$, let $f(x) : S \to [-1,1]$ be a function with polynomial approximation p(x) such that, for some approximation error $\varepsilon > 0$,

$$|p(x) - f(x)| \le \varepsilon \text{ for all } x \in S, \text{ and}$$
(18)

$$|p(x)| \le 1 \text{ for all } x \in [-1, 1].$$
 (19)

Then

$$\deg(p) \ge \sqrt{1 - \Delta^2} \sup_{\substack{x, y \in S \\ x \neq y}} \frac{|f(x) - f(y)| - 2\varepsilon}{|x - y|}.$$

Proof. Consider some distinct $x, y \in S$, and let

$$L = \frac{|f(x) - f(y)| - 2\varepsilon}{|x - y|}.$$

Then by (18),

$$|p(x) - p(y)| \ge |f(x) - f(y)| - 2\varepsilon = L|x - y|,$$

so by the intermediate value theorem, for some ξ between x and y, $|p'(\xi)| \ge L$. Since p(x) is bounded by 1 in [-1, 1], we can apply Proposition 27 to get that

$$\deg(p) \ge |p'(\xi)|\sqrt{1-\xi^2} \ge L\sqrt{1-\Delta^2}$$

We take the supremum over all x, y to get the desired bound.

We now discuss some implications of Proposition 28 for quantum algorithm design. In recent works on quantum optimization [AG19; BGJST23], approximations to $\exp(\beta y)$ on $y \in [-1, \delta]$ for constant δ are used to speed up Gibbs sampler subroutines for solving zero-sum games via QSVT. It is a well-known result that a degree $O(\sqrt{\beta})$ polynomial approximates $\exp(-\beta y)$ up to additive error 0.1 on [-1,0] [SV14, Theorem 4.1]. However, only a $\approx \beta$ degree polynomial approximation was known when the polynomial is further required to be bounded in $[0, \delta]$ (a small interval outside the region of approximation). Because the boundedness requirement comes from the use of QSVT, and is not needed in the classical setting, the state-of-the-art quantum runtime for zero-sum games [BGJST23] incurs an overhead of $\sqrt{\beta} = \sqrt{1/\varepsilon}$ compared to classical counterparts (while saving on dimension-dependent factors).

Corollary 29 applies Proposition 28 to show that for $\delta = \omega(\beta^{-1})$, adding the boundedness requirement necessitates an approximation of larger degree, up to quadratically worse when $\delta = \Omega(1)$. This implies that the degree achieved by Corollary 23 is nearly-tight. This negatively resolves the open question posed by [BGJST23], which was whether Corollary 29 could be modified to remove the last remaining overhead in $\frac{1}{\varepsilon}$ (when $\delta = \Omega(1)$). We rule out this approach, suggesting it is necessary to fundamentally change the application of QSVT to obtain this conjectured speedup.

Corollary 29. Let $\beta \ge 1$, $\delta \in (0,1]$, and let q(x) be a degree-d polynomial which satisfies

$$|q(x) - \exp(\beta x)| \le 0.1 \text{ for } x \in [-1, 0] \text{ and}$$

 $|q(x)| \le 1 \text{ for } x \in [0, \delta].$

Then $d = \Omega(\beta \sqrt{\delta}).$

Proof. Consider the change of variable $x = \frac{1+\delta}{2}t - \frac{1-\delta}{2}$ which maps [-1,1] to $[-1,\delta]$. Then, for $f(t) = \exp(\beta x(t))$ and p(t) = q(x(t)), we know that $|f(t) - p(t)| \le 0.1$ for $t \in [-1, \frac{1-\delta}{1+\delta}]$ and $|p(t)| \le 1$ for $t \in [-1, 1]$. Further,

$$\frac{|f(\frac{1-\delta}{1+\delta}) - f(\frac{2}{1+\delta}(-\frac{1}{\beta} + \frac{1-\delta}{2}))| - 0.2}{|\frac{1-\delta}{1+\delta} - \frac{2}{1+\delta}(-\frac{1}{\beta} + \frac{1-\delta}{2})|} = \frac{1 - 1/e - 0.2}{\frac{2}{\beta(1+\delta)}} = \Omega(\beta).$$

So, applying Proposition 28 with $S = \{\frac{2}{1+\delta}(-\frac{1}{\beta} + \frac{1-\delta}{2}), \frac{1-\delta}{1+\delta}\}$, we get that

$$d = \Omega\left(\beta\sqrt{1 - \left(\frac{1-\delta}{1+\delta}\right)^2}\right) = \Omega(\beta\sqrt{\delta}).$$

A similar quadratic gap occurs for quantum algorithms for solving linear systems $\mathbf{A}x = b$ when \mathbf{A} is positive definite [OD21]. Classical methods for this problem like conjugate gradient have a $\sqrt{\kappa}$ condition number dependence, which arises because $\frac{1}{x}$ has a good polynomial approximation on $[\frac{1}{\kappa}, 1]$ with degree $\approx \sqrt{\kappa}$.¹³ However, QSVT requires approximations to be bounded on [-1, 1]; by applying a similar argument as in Corollary 29, this implies that a degree of $\Omega(\kappa)$ is necessary to achieve same approximation quality with the boundedness constraint. Orsucci and Dunjko work around this issue by observing that if we have a block-encoding of $\mathbf{I} - \mathbf{A}$, then the function to be approximated is now $\frac{1}{1-x}$, which can be done with degree $\approx \sqrt{\kappa}$, since the ill-conditioned part of the function is on the boundary of [-1, 1], rather than the interior.

3.5 Proof of Theorem 21

We conclude with a proof of Theorem 21. Our proof builds upon several elementary bounds on Bernstein ellipses and the growth of Chebyshev polynomials, as well as the construction of explicit thresholding functions. For ease of exposition, we state all the helper bounds we use in this section, but defer their proofs to Appendix D. We begin with our bounds on the sizes of Bernstein ellipses.

Fact 30. The Bernstein ellipse E_{ρ} for $\rho \geq 1$ satisfies

interior
$$(E_{\rho}) \subset \left\{ x + iy \mid x, y \in \mathbb{R}, \, |x| \leq \frac{1}{2}(\rho + \rho^{-1}) \text{ and } |y| \leq \frac{1}{2}(\rho - \rho^{-1}) \right\}.$$

Further, for $\rho = 1 + \delta \leq 2$,

$$1 + \frac{\delta^2}{4} \le \frac{1}{2}(\rho + \rho^{-1}) = 1 + \frac{\delta^2}{2(1+\delta)} \le 1 + \frac{\delta^2}{2},$$
$$\frac{3}{4}\delta \le \frac{1}{2}(\rho - \rho^{-1}) = \delta - \frac{\delta^2}{2(1+\delta)} \le \delta.$$

This yields the following containment fact, whose proof is deferred to Appendix D.

Lemma 31. For $\delta \in (0, 1)$, $(1+\delta)E_{1+\alpha}$ is contained in the interior of E_{σ} , where $\sigma = 1+3(\alpha+\sqrt{\delta})$. We also use the following bounds on Chebyshev polynomials, deferring a proof to Appendix D. **Lemma 32.** There are universal constants C, c > 0 such that, for $n \ge 0$ and $x, y \in \mathbb{R}$, $|y| \le c$,

$$|T_n(x+\imath y)| \le \begin{cases} (1+C\sqrt{|y|})^n & |x| \le 1\\ (x+\sqrt{x^2-1}+C\sqrt{|xy|})^n & |x| > 1. \end{cases}$$

To ameliorate the polynomial growth of Chebyshev polynomials from Lemma 32, we apply a threshold function with tails which decay superexponentially. Our thresholding is based on the Gaussian error function erf; we define erf and recall some standard bounds on it in the following.

Fact 33 (Eqs. 7.8.3, 7.8.7, [DLMF]). For $z \in \mathbb{C}$, $\operatorname{erf} : \mathbb{C} \to \mathbb{C}$ by $\operatorname{erf}(z) \coloneqq \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$. Then,

$$1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-t^{2}} dt < \frac{2e^{-x^{2}}}{\sqrt{\pi}(1+x)} < 2e^{-x^{2}},$$
(20)

$$|\operatorname{erf}(ix)| = \frac{2}{\sqrt{\pi}} \int_0^x e^{t^2} dt < \frac{2(e^{x^2} - 1)}{\sqrt{\pi}|x|} < 2e^{x^2} \quad (when \ x \ge 1).$$
(21)

For $z \in \mathbb{R}$, we note that $\frac{1}{2} + \frac{1}{2} \operatorname{erf}(z)$ is the cumulative distribution function for a Gaussian with mean 0 and variance $\frac{1}{2}$, which interpolates between 0 and 1; consequently, one may view erf (appropriately rescaled as necessary) as a "smoothed" variant of the sign function

$$\operatorname{sgn}(x) \coloneqq \begin{cases} -1 & x < 0\\ 0 & x = 0\\ 1 & x > 0 \end{cases}$$

¹³We can see this explicitly. Approximating x^{-1} on $[\kappa^{-1}, 1]$ is equivalent to approximating $\frac{1}{x-a}$ on [-1, 1] for $a = 1 + \Theta(\kappa^{-1})$. There is an explicit expression for the Chebyshev coefficients of $\frac{1}{x-a} = \sum_{k=0}^{\infty} a_k T_k(x)$ [MH02, Eq. (5.14)]: $|a_k| \sim \frac{1}{\sqrt{a^2-1}} (a - \sqrt{a^2-1})^k$. This is $\varepsilon \| \frac{1}{x-a} \|_{[-1,1]}$ when taking $k = \Theta(\sqrt{\kappa})$.

Building upon erf, we state our family of thresholding functions, deferring proofs to Appendix D.

Lemma 34 (Thresholding function). For $\mu, s > 0$, let $r(z) \coloneqq \frac{1}{2}(\operatorname{erf}(s(\mu+z)) + \operatorname{erf}(s(\mu-z)))$. When $z \in \mathbb{R}, 0 \le r(z) \le 1$. When $x, y \in \mathbb{R}$ and z = x + iy, $|r(z) - r(x)| \le \exp(-s^2(\mu - |x|)^2)|\operatorname{erf}(isy)|$.

Evidently for $z \in \mathbb{R}$, the function r behaves as a threshold: sufficiently inside $[-\mu, \mu]$, it is close to 1, and sufficiently outside it is close to 0. The size of the "growth window" near μ is roughly $\frac{1}{s}$, and Lemma 34 shows $r(x + iy) \approx r(x)$ for small y. Leveraging these tools, we now prove Theorem 21.

Proof of Theorem 21. Without loss of generality, we rescale so that M = 1. To obtain the theorem statement, it suffices to prove that there exists a polynomial q of degree $O(\frac{b}{\delta} \log \frac{1}{\alpha \varepsilon})$ such that

$$\|f - q\|_{[-(1-\delta),1-\delta]} \leq \varepsilon,$$

$$\|q\|_{[-1,1]} \leq 1 + \varepsilon,$$

$$\|q\|_{[-b,-1]\cup[1,b]} \leq \varepsilon.$$
(22)

To see this, consider f as in the theorem statement. Let $\delta' = \frac{\delta}{1+\delta} = \Theta(\delta)$, so that $\frac{1}{1-\delta'} = 1 + \delta$. Then $f(\frac{y}{1-\delta'})$ is analytic and bounded by M for y in the interior of $(1-\delta')E_{\rho}$, which contains $E_{1+\frac{\alpha}{4}}$ by Lemma 31 and $\sqrt{\delta'} < \sqrt{\delta} < \frac{\alpha}{C}$. Applying (22), we get a function q(y) such that $q((1-\delta')x)$ satisfies the guarantees described above, with the intervals scaled up by a factor of $\frac{1}{1-\delta'}$:

$$\begin{aligned} |f(x) - q((1 - \delta')x)| &\leq \varepsilon & \text{for } x \in [-1, 1], \\ |q((1 - \delta')x)| &\leq 1 + \varepsilon & \text{for } x \in [-\frac{1}{1 - \delta'}, \frac{1}{1 - \delta'}] = [-(1 + \delta), 1 + \delta], \\ |q((1 - \delta')x)| &\leq \varepsilon & \text{for } x \in [-\frac{b}{1 - \delta'}, -(1 + \delta)] \cup [1 + \delta, \frac{b}{1 - \delta'}] \end{aligned}$$

To conclude, consider $\frac{1}{1+\varepsilon}q((1-\delta')x)$. We make q slightly smaller so that it is bounded by 1 in $[-(1+\delta), (1+\delta)]$. This only affects the closeness to f by a constant factor: for $x \in [-1, 1]$,

$$|f(x) - \frac{1}{1+\varepsilon}q((1-\delta')x)| \le |f(x) - q((1-\delta')x)| + (1-\frac{1}{1+\varepsilon})|q((1-\delta')x)| \le 2\varepsilon.$$

The degree of $\frac{1}{1+\varepsilon}q((1-\delta')x)$ is degree of q(x), as desired. We now proceed to prove (22). By Theorem 20, there is a polynomial with degree $n = \lceil \frac{1}{\alpha} \log \frac{6}{\alpha \varepsilon} \rceil$ with $||f - f_n||_{[-1,1]} \le \frac{\varepsilon}{3}$, and the Chebyshev coefficients of $f_n = \sum_{k=0}^n a_k T_k(x)$, satisfy $|a_k| \le 2\rho^{-k}$. Next, let $\tilde{p}(z) \coloneqq r(z)f_n(z)$ be the truncation f_n multiplied by the function r(z) from Lemma 34 with

$$\mu \coloneqq 1 - \frac{\delta}{2}, \ s \coloneqq \frac{C_s}{\delta} \sqrt{\log \frac{1}{\alpha \varepsilon}},$$

and C_s is a constant to be chosen later. Let $\tilde{\rho} := 1 + \frac{\delta}{b}$; we will show \tilde{p} is bounded on $bE_{\tilde{\rho}}$, and then our final approximation q will be an application of Theorem 20 to approximate \tilde{p} on [-b, b]. To this end, it suffices to bound $\tilde{p}(z)$ for all $z \in S$, where the strip S is defined as

$$S \coloneqq \{z = x + \imath y \mid |y| \le \delta\}$$

because Fact 30 implies $S \supseteq bE_{\tilde{\rho}}$. We begin by bounding r(z) for $z \in S$:

$$|r(x + iy)| \le r(x) + e^{-s^{2}(\mu - |x|)^{2}} |\operatorname{erf}(isy)| \le r(x) + e^{-s^{2}(\mu - |x|)^{2}} \left| \operatorname{erf}\left(iC_{s}\sqrt{\log\frac{1}{\alpha\varepsilon}}\right) \right| \le r(x) + 2e^{-s^{2}(\mu - |x|)^{2}} (\alpha\varepsilon)^{-C_{s}^{2}}.$$
(23)

The inequalities above respectively used Lemma 34, the definition of S, and (21). We now combine (23) with Lemma 32 to bound \tilde{p} on S. First, consider when $z = x + iy \in S$ and $x \in [-1, 1]$. This is the bottleneck of the argument, where \tilde{p} is largest. We bound

$$\begin{split} |\tilde{p}(z)| &= |r(z)| |f_n(z)| \le \left(1 + \exp(-s^2(\mu - |x|)^2) \operatorname{poly}\left(\frac{1}{\alpha\varepsilon}\right)\right) \left| \sum_{k=0}^n a_k T_k(z) \right| \\ &\le \operatorname{poly}\left(\frac{1}{\alpha\varepsilon}\right) \left(2\sum_{k=0}^n \rho^{-k} |T_k(z)|\right) \\ &\le \operatorname{poly}\left(\frac{1}{\alpha\varepsilon}\right) \left(\sum_{k=0}^n \left(\frac{1 + K\sqrt{\delta}}{\rho}\right)^k\right) = \operatorname{poly}\left(\frac{1}{\alpha\varepsilon}\right). \end{split}$$

The first inequality was (23), the second used the guarantees of Theorem 20, the third used Lemma 32, and the last used $\delta \ll \alpha^2$, $n = \text{poly}(\frac{1}{\alpha \varepsilon})$. Next, for $z = x + iy \in S$ with $|x| \ge 1$,

$$r(x) = \frac{1}{2} (\operatorname{erf}(s(\mu + |x|)) - \operatorname{erf}(s(|x| - \mu))) \le \frac{1}{2} (1 - \operatorname{erf}(s(|x| - \mu))) < e^{-s^2(|x| - \mu)^2}, \quad (24)$$

where the first inequality was $\operatorname{erf}(z) \leq 1$ for $z \in \mathbb{R}$, and the last was (20). Further, Lemma 32 yields

$$|f_n(z)| \le \sum_{k=0}^n 2\rho^{-k} \left(|x| + \sqrt{x^2 - 1} + K\sqrt{|xy|} \right)^k \le 2n \left(|x| + \sqrt{x^2 - 1} + K\sqrt{|xy|} \right)^n \le 2n \exp\left(n \left(|x| - 1 + \sqrt{x^2 - 1} + K\sqrt{|xy|} \right) \right).$$
(25)

Continuing, we combine (23), (24), and (25) to conclude

$$\begin{aligned} |\widetilde{p}(z)| &\leq \exp\left(-s^{2}(|x|-\mu)^{2}\right) \operatorname{poly}\left(\frac{1}{\alpha\varepsilon}\right) |f_{n}(z)| \\ &\leq \exp\left(-s^{2}(|x|-\mu)^{2} + n\left(|x|-1+\sqrt{x^{2}-1}+K\sqrt{|xy|}\right)\right) \operatorname{poly}\left(\frac{1}{\alpha\varepsilon}\right) \\ &\leq \exp\left(\left(-\frac{C_{s}^{2}}{\delta^{2}}(|x|-\mu)^{2} + \frac{1}{\sqrt{\delta}}\left(|x|-1+\sqrt{x^{2}-1}+K\sqrt{|x|\delta}\right)\right) \log\frac{1}{\alpha\varepsilon}\right) \operatorname{poly}\left(\frac{1}{\alpha\varepsilon}\right). \end{aligned}$$
(26)

Here we used that $n \leq \frac{1}{\sqrt{\delta}} \log \frac{1}{\alpha \varepsilon}$ under the assumed relationship between δ and α , and the definition of s. For sufficiently large C_s , it is straightforward to see that for all $|x| \geq 1$, since the left-hand side asymptotically grows faster than each term in the right-hand side,

$$\frac{C_s^2}{2\delta^2} \left(|x| - \left(1 - \frac{\delta}{2}\right) \right)^2 \ge \frac{1}{\sqrt{\delta}} \left(|x| - 1 + \sqrt{x^2 - 1} + K\sqrt{|x|\delta} \right),\tag{27}$$

and hence for this choice of C_s , plugging this into the previous bound gives that $\widetilde{p}(x+iy) \leq \text{poly}(\frac{1}{\alpha\varepsilon})$ for $|x| \geq 1$.¹⁴ Later, we will need a tighter bound when y = 0 and $|x| \geq 1$: in this setting, the second additive term in (23) vanishes, and hence taking C_s such that (27) holds, repeating the arguments in (26) without the $\text{poly}(\frac{1}{\alpha\varepsilon})$ overhead gives for sufficiently large C_s ,

$$\widetilde{p}(x) \le \exp\left(-\frac{C_s^2}{2\delta^2}(|x|-\mu)^2\log\frac{1}{\alpha\varepsilon}\right) \le \exp\left(-\frac{C_s^2}{8}\log\frac{1}{\alpha\varepsilon}\right) \le \frac{\varepsilon}{3} \text{ for all } x \in \mathbb{R} \text{ with } |x| \ge 1.$$
(28)

Thus, we have shown that for all $z \in S$, $|\tilde{p}(z)| \leq \text{poly}(\frac{1}{\alpha\varepsilon})$, and as the product of analytic functions, \tilde{p} is analytic. Next, for all $z \in \mathbb{C}$ let $\hat{p}(z) \coloneqq \tilde{p}(bz)$. We have shown \hat{p} is bounded on $E_{\tilde{\rho}}$, and hence Theorem 20 gives a Chebyshev truncation \hat{p}_m such that $\|\hat{p} - \hat{p}_m\|_{[-1,1]} \leq \frac{\varepsilon}{3}$, for

$$m = O\bigg(\frac{b}{\delta}\log\frac{b}{\delta\varepsilon}\bigg).$$

Our final approximation is $q(z) \coloneqq \widehat{p}_m(\frac{z}{b})$. By the definitions of q and \widehat{p} , the relationship between \widehat{p} and \widetilde{p} implies $||q - \widetilde{p}||_{[-b,b]} \le \frac{\varepsilon}{3}$. Combined with (28), this implies the third bound in (22),

$$\|q\|_{[-b,-1]\cup[1,b]} \le \varepsilon.$$

The first bound $||f - q||_{[-(1-\delta),1-\delta]} \leq \varepsilon$ in (22) follows from $||q - \widetilde{p}||_{[-(1-\delta),1-\delta]} \leq \frac{\varepsilon}{3}$, $||f_n - f||_{[-(1-\delta),1-\delta]} \leq \frac{\varepsilon}{3}$, and $||f_n - \widetilde{p}||_{[-(1-\delta),1-\delta]} \leq (1+\frac{\varepsilon}{3})||1 - r||_{[-(1-\delta),1-\delta]} \leq \frac{\varepsilon}{3}$ choosing C_s sufficiently large. Finally, the second bound in (22) follows from

$$\begin{aligned} \|q\|_{[1-\delta,1]} &\leq \|q-\widetilde{p}\|_{[1-\delta,1]} + \|\widetilde{p}\|_{[1-\delta,1]} \leq \frac{\varepsilon}{3} + \|\widetilde{p}\|_{[1-\delta,1]} \\ &\leq \frac{\varepsilon}{3} + \|f_n\|_{[1-\delta,1]} \leq \varepsilon + \|f\|_{[1-\delta,1]} \leq 1+\varepsilon, \end{aligned}$$

where we used the closeness bounds between (\tilde{p}, q) and (f_n, f) , as well as the assumed bound on f over E_{ρ} (which contains $[1 - \delta, 1]$). The bound $||q||_{[-1, -(1-\delta)]} \leq 1 + \varepsilon$ follows symmetrically. \Box

¹⁴We note that the poly in (26) hides a C_s -dependent exponent, which grows faster than (27).

In some of our applications in Section 3.3, we used the following property of our approximations constructed via Theorem 21, which we record here for convenience.

Corollary 35. In the setting of Theorem 21, if f is even or odd, so is q.

Proof. It is straightforward to check that all operations we perform on f (Chebyshev truncation, multiplication by an even function r, and another Chebyshev truncation) preserve parity.

Acknowledgements

ET thanks t.f. for providing useful references. ET also thanks David Gosset, Beni Yoshida, and Richard Cleve for the invitation to Waterloo and the hospitality; ET first read about the CS decomposition during a quiet night at the Perimeter Institute. ET is supported by the NSF GRFP (DGE-1762114). KT thanks Jonathan Kelner for encouraging him to learn about the CS decomposition, Christopher Musco for encouraging him to read [Tre19], and Yang P. Liu for a helpful conversation about Jordan's lemma many moons ago. We thank Carlo Beenakker for drawing our attention to the application of the CS decomposition to scattering theory.

References

- [DLMF] NIST Digital Library of Mathematical Functions. https://dlmf.nist.gov/, Release
 1.10 of 2023-06-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider,
 R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds. URL: https://dlmf.nist.gov/ (pages 13, 17, 20).
- [AA22] Amol Aggarwal and Josh Alman. "Optimal-degree polynomial approximations for exponentials and gaussian kernel density estimation". In: 37th Computational Complexity Conference, CCC 2022. Vol. 234. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022, 22:1–22:23. DOI: 10.4230/LIPIcs.CCC.2022.22 (pages 12, 13).
- [AG19] Joran van Apeldoorn and András Gilyén. Quantum algorithms for zero-sum games. 2019. DOI: 10.48550/ARXIV.1904.03180 (pages 2, 16, 19).
- [Bee97] C. W. J. Beenakker. "Random-matrix theory of quantum transport". In: Reviews of Modern Physics 69.3 (July 1997), pp. 731-808. DOI: 10.1103/revmodphys.69.731. arXiv: cond-mat/9612179 [cond-mat.mes-hall] (page 5).
- [Ber12] S. N. Bernstein. "Sur l'ordre de la meilleure approximation des fonctions continues par les polynômes de degré donné". In: *Mémoires de l'Académie Royale de Belgique* 4 (1912), pp. 1–103 (page 18).
- [BGJST23] Adam Bouland, Yosheb Getachew, Yujia Jin, Aaron Sidford, and Kevin Tian. Quantum speedups for zero-sum games via improved dynamic gibbs sampling. 2023. DOI: 10. 48550/ARXIV.2301.03763 (pages 2, 16, 19).
- [CKS17] Andrew M. Childs, Robin Kothari, and Rolando D. Somma. "Quantum algorithm for systems of linear equations with exponentially improved dependence on precision". In: *SIAM Journal on Computing* 46.6 (2017), pp. 1920–1950. DOI: 10.1137/16M1087072 (page 14).
- [EJ23] Alan Edelman and Sungwoo Jeong. "Fifty three matrix factorizations: a systematic approach". In: *SIAM Journal on Matrix Analysis and Applications* 44.2 (Apr. 2023), pp. 415–480. DOI: 10.1137/21m1416035. arXiv: 2104.08669 [math.NA] (page 3).
- [GSLW19] András Gilyén, Yuan Su, Guang Hao Low, and Nathan Wiebe. "Quantum singular value transformation and beyond: Exponential improvements for quantum matrix arithmetics". In: Proceedings of the 51st ACM Symposium on the Theory of Computing (STOC). ACM, June 2019, pp. 193–204. DOI: 10.1145/3313276.3316366. arXiv: 1806.01838 (pages 1–3, 5–7, 12–17, 25, 26, 28).
- [HHL09] Aram W. Harrow, Avinatan Hassidim, and Seth Lloyd. "Quantum algorithm for linear systems of equations". In: *Physical Review Letters* 103 (15 Oct. 2009), p. 150502. DOI: 10.1103/PhysRevLett.103.150502 (page 2).

- [Jor75] Camille Jordan. "Essai sur la géométrie à n dimensions". In: Bulletin de la Société Mathématique de France 3 (1875), pp. 103-174. ISSN: 0037-9484. DOI: 10.24033/bsmf.
 90. URL: http://www.numdam.org/item?id=BSMF_1875_3_103_2 (pages 1, 2, 28).
- [LC17] Guang Hao Low and Isaac L. Chuang. "Optimal Hamiltonian simulation by quantum signal processing". In: *Physical Review Letters* 118.1 (Jan. 2017), p. 010501. DOI: 10.1103/PhysRevLett.118.010501. arXiv: 1606.02685 [quant-ph] (pages 2, 12).
- [LC19] Guang Hao Low and Isaac L. Chuang. "Hamiltonian simulation by qubitization". In: *Quantum* 3 (July 2019), p. 163. DOI: 10.22331/q-2019-07-12-163 (page 2).
- [Lin22] Lin Lin. Lecture notes on quantum algorithms for scientific computation. Jan. 20, 2022. arXiv: 2201.08309 [quant-ph]. URL: https://math.berkeley.edu/~linlin/qasc/ (page 2).
- [MH02] John C Mason and David C Handscomb. *Chebyshev polynomials*. Chapman and Hall/CRC, 2002 (pages 12, 20).
- [ML92] Th. Martin and R. Landauer. "Wave-packet approach to noise in multichannel mesoscopic systems". In: *Physical Review B* 45.4 (Jan. 1992), pp. 1742–1755. DOI: 10.1103/ physrevb.45.1742 (page 5).
- [MPK88] P.A Mello, P Pereyra, and N Kumar. "Macroscopic approach to multichannel disordered conductors". In: Annals of Physics 181.2 (Feb. 1988), pp. 290–317. DOI: 10.1016/0003-4916(88)90169-8 (page 5).
- [MRTC21] John M. Martyn, Zane M. Rossi, Andrew K. Tan, and Isaac L. Chuang. "Grand unification of quantum algorithms". In: *PRX Quantum* 2 (4 Dec. 2021), p. 040203. DOI: 10.1103/PRXQuantum.2.040203. arXiv: 2105.02859 [quant-ph] (pages 2, 5).
- [OD21] Davide Orsucci and Vedran Dunjko. "On solving classes of positive-definite quantum linear systems with quadratically improved runtime in the condition number". In: *Quantum* 5 (Nov. 2021), p. 573. DOI: 10.22331/q-2021-11-08-573. arXiv: 2101.11868 [quant-ph] (page 20).
- [PW94] C. C. Paige and M. Wei. "History and generality of the CS decomposition". In: Linear Algebra and Its Applications 208/209 (1994), pp. 303–326. ISSN: 0024-3795. DOI: 10.1016/0024-3795(94)90446-4 (pages 3, 4).
- [Reg06] Oded Regev. Quantum computation lecture 2: Witness-preserving amplification of QMA. Accessed: 2022/12/20. 2006. URL: https://cims.nyu.edu/~regev/teaching/ quantum_fall_2005/ln/qma.pdf (pages 1, 28).
- [Sch41] A. C. Schaeffer. "Inequalities of A. Markoff and S. Bernstein for polynomials and related functions". In: *Bull. Amer. Math. Soc.* 47 (1941), pp. 565–579. ISSN: 0002-9904. DOI: 10.1090/S0002-9904-1941-07510-5 (page 18).
- [SV14] Sushant Sachdeva and Nisheeth K. Vishnoi. "Faster algorithms via approximation theory". In: Foundations and Trends in Theoretical Computer Science 9.2 (2014), pp. 125–210. ISSN: 1551-305X. DOI: 10.1561/040000065 (pages 12, 19).
- [Sze04] Mario Szegedy. "Quantum speed-up of markov chain based algorithms". In: 45th Symposium on Foundations of Computer Science (FOCS 2004). IEEE Computer Society, 2004, pp. 32–41. DOI: 10.1109/FOCS.2004.53 (page 2).
- [Tre19] Lloyd N. Trefethen. Approximation theory and approximation practice, extended edition. Extended edition [of 3012510]. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2019, pp. xi+363. ISBN: 978-1-611975-93-2. DOI: 10.1137/1. 9781611975949 (pages 1, 10, 11, 15, 16, 23).
- [Wat17] G. N. Watson. "Bessel functions and Kapteyn series". In: Proceedings of the London Mathematical Society s2-16.1 (1917), pp. 150–174. DOI: 10.1112/plms/s2-16.1.150 (page 13).
- [Wat95] G. N. Watson. A treatise on the theory of Bessel functions. Cambridge Mathematical Library. Reprint of the second (1944) edition. Cambridge University Press, Cambridge, 1995, pp. viii+804. ISBN: 0-521-48391-3 (page 13).

A Proofs of quantum signal processing

Our goal in this section is to characterize which polynomials are QSP-achievable. Looking at the form of QSP, we can express its entries via polynomials satisfying a recurrence.

Lemma 36 (QSP as a recurrence). For some phase factors $\Phi = \{\phi_j\}_{0 \le j \le n} \in \mathbb{R}^{n+1}$,

$$\mathbf{QSP}(\{\phi_j\}_{k \le j \le n}, x) = \begin{pmatrix} p_k(x) & q_k^*(-x)\sqrt{1-x^2} \\ q_k(x)\sqrt{1-x^2} & p_k^*(-x) \end{pmatrix},$$
(29)

where $p_n(x) = e^{i\phi_n}$ and $q_n(x) = 0$, and $p_k(x)$ and $q_k(x)$ satisfy the following recurrence relation:

$$p_k(x) = e^{i\phi_k} (xp_{k+1}(x) + (1 - x^2)q_{k+1}(x)),$$
(30)

$$q_k(x) = e^{-i\phi_k} (p_{k+1}(x) - xq_{k+1}(x)).$$
(31)

Proof. The base case is because

$$\mathbf{QSP}(\{\phi_n\}, x) = \begin{pmatrix} e^{i\phi_n} & 0\\ 0 & e^{-i\phi_n} \end{pmatrix}.$$

The inductive case is a computation:

$$\begin{aligned} \mathbf{QSP}(\{\phi_j\}_{k \le j \le n}, x) \\ &= e^{i\phi_k \sigma_z} \mathbf{R}(x) \cdot \mathbf{QSP}(\{\phi_j\}_{k+1 \le j \le n}, x) \\ &= \begin{pmatrix} e^{i\phi_k \sigma_z} \mathbf{R}(x) \cdot \mathbf{QSP}(\{\phi_j\}_{k+1 \le j \le n}, x) \\ e^{-i\phi_k \sqrt{1-x^2}} & -e^{-i\phi_k x} \end{pmatrix} \begin{pmatrix} p_{k+1}(x) & q_{k+1}^*(-x)\sqrt{1-x^2} \\ q_{k+1}(x)\sqrt{1-x^2} & p_{k+1}^*(-x) \end{pmatrix} \\ &= \begin{pmatrix} e^{i\phi_k}(xp_{k+1}(x) + (1-x^2)q_{k+1}(x)) & e^{i\phi_k}(p_{k+1}^*(-x) + xq_{k+1}^*(-x))\sqrt{1-x^2} \\ e^{-i\phi_k}(p_{k+1}(x) - xq_{k+1}(x))\sqrt{1-x^2} & e^{-i\phi_k}(-xp_{k+1}^*(-x) + (1-x^2)q_{k+1}^*(-x)) \end{pmatrix} \\ &= \begin{pmatrix} p_k(x) & q_k^*(-x)\sqrt{1-x^2} \\ q_k(x)\sqrt{1-x^2} & p_k^*(-x) \end{pmatrix}. \end{aligned}$$

Theorem 37 (Variant of [GSLW19, Theorem 3]). A degree-n polynomial $p(x) \in \mathbb{C}[x]$ is QSP-achievable if and only if there is a polynomial q(x) such that:

- (a) q has degree $\leq n 1$;
- (b) (p,q) are (even, odd) or (odd, even);
- (c) $|p(x)|^2 + (1 x^2)|q(x)|^2 \equiv 1.$

Proof. First, we consider the "only if" direction. Suppose p(x) is QSP-achievable with the phase factors $\Phi \in \mathbb{R}^{n+1}$. Then, by Lemma 36, there is some q(x) such that

$$\mathbf{QSP}(\Phi, x) = \begin{pmatrix} p(x) & q^*(-x)\sqrt{1-x^2} \\ q(x)\sqrt{1-x^2} & p^*(-x) \end{pmatrix},$$

derived from the recurrence described in that lemma. From this recurrence, we can verify that at all times, conditions (a) and (b) are satisfied. Finally, condition (c) is always satisfied because $\mathbf{QSP}(\Phi, x)$ is a product of unitary matrices, and so is unitary: the first column having norm one is equivalent to $|p(x)|^2 + (1-x^2)|q(x)|^2 = p(x)p^*(x) + (1-x^2)q(x)q^*(x) = 1$, and this argument works for every $x \in [-1, 1]$. Because it holds for infinitely many x, the equality holds as polynomials.

Second, we consider the "if" direction. Suppose we have some p(x) of degree n and q(x) satisfying (a), (b), and (c). We want to construct phase factors that implement p(x). We proceed by induction: when n = 0, this means that p(x) is scalar and q(x) has degree ≤ -1 (meaning it must be zero). Thus, $p(x) \equiv e^{i\phi}$ for some ϕ ; we can implement this with $\Phi = \{\phi\}$. For the inductive step, consider p(x) of degree n + 1. If we could show that there exists some φ such that

$$(e^{i\varphi\sigma_{z}}\mathbf{R}(x))^{\dagger} \begin{pmatrix} p(x) & q^{*}(-x)\sqrt{1-x^{2}} \\ q(x)\sqrt{1-x^{2}} & p^{*}(-x) \end{pmatrix} = \begin{pmatrix} p_{\downarrow}(x) & q_{\downarrow}^{*}(-x)\sqrt{1-x^{2}} \\ q_{\downarrow}(x)\sqrt{1-x^{2}} & p_{\downarrow}^{*}(-x) \end{pmatrix}$$
(32)

for $p_{\downarrow}, q_{\downarrow}$ some even/odd polynomials of one degree lower than p and q, then we would be done. By assumption, the matrices on the left-hand side of (32) are unitary, so the right-hand side matrix is also unitary. Thus, p_{\downarrow} and q_{\downarrow} satisfy all the properties of the induction hypothesis, and there are phase factors $\{\phi_0, \ldots, \phi_n\} \in \mathbb{R}^{n+1}$ giving the equality

$$(e^{i\varphi\sigma_{z}}\mathbf{R}(x))^{\dagger} \begin{pmatrix} p(x) & q^{*}(-x)\sqrt{1-x^{2}} \\ q(x)\sqrt{1-x^{2}} & p^{*}(-x) \end{pmatrix} = \mathbf{QSP}(\{\phi_{0},\dots,\phi_{n}\},x).$$
(33)

$$\begin{pmatrix} p(x) & q^*(-x)\sqrt{1-x^2} \\ q(x)\sqrt{1-x^2} & p^*(-x) \end{pmatrix} = \mathbf{QSP}(\{\varphi, \phi_0, \dots, \phi_n\}, x)$$
(34)

So it comes down to finding the right value of φ that could remove a degree from p and q in (32). By properties (a) and (b), we can write

$$p(x) = a_{n+1}x^{n+1} + a_{n-1}x^{n-1} + \dots$$
(35)

$$q(x) = b_n x^n + a_{n-2} x^{n-2} + \dots ag{36}$$

The condition (c) implies that $|a_{n+1}| = |b_n|$. Now we perform the matrix calculation. Since $\mathbf{R}(x)$ is its own inverse, $(e^{i\varphi\sigma_z}\mathbf{R}(x))^{\dagger} = \mathbf{R}(x)e^{-i\varphi\sigma_z}$, so

$$(e^{i\varphi\boldsymbol{\sigma}_{z}}\mathbf{R}(x))^{\dagger} \begin{pmatrix} p(x) & q^{*}(-x)\sqrt{1-x^{2}} \\ q(x)\sqrt{1-x^{2}} & p^{*}(-x) \end{pmatrix}$$
(37)

$$= \begin{pmatrix} e^{-i\varphi}x & e^{i\varphi}\sqrt{1-x^2} \\ e^{-i\varphi}\sqrt{1-x^2} & -e^{i\varphi}x \end{pmatrix} \begin{pmatrix} p(x) & q^*(-x)\sqrt{1-x^2} \\ q(x)\sqrt{1-x^2} & p^*(-x) \end{pmatrix}$$
(38)

$$= \begin{pmatrix} e^{-i\varphi}p(x) + e^{i\varphi}(1-x^2)q(x) & (e^{i\varphi}p^*(-x) + e^{-i\varphi}xq^*(-x))\sqrt{1-x^2} \\ (e^{-i\varphi}p(x) - e^{i\varphi}xq(x))\sqrt{1-x^2} & -e^{i\varphi}xp^*(-x) + e^{-i\varphi}(1-x^2)q^*(-x) \end{pmatrix}$$
(39)

So, we need the following polynomials to have lower degree:

$$p_{\downarrow}(x) = e^{-i\varphi} p(x) + e^{i\varphi} (1 - x^2) q(x)$$
(40)

$$q_{\downarrow}(x) = e^{-i\varphi}p(x) - e^{i\varphi}xq(x) \tag{41}$$

The "leading" coefficient of x^{n+1} for p_{\downarrow} and x^n for q_{\downarrow} are the same: $e^{-i\varphi}a_{n+1} - e^{i\varphi}b_n$. If we choose φ such that $e^{i\varphi} = \sqrt{a_{n+1}/b_n}$, then this coefficient is 0, and so the degrees of p_{\downarrow} and q_{\downarrow} are $\leq n-1$ and $\leq n-2$, as desired.

The characterization of when a polynomial p(x) is QSP-achievable is still somewhat difficult to understand. With more work, we can give a clearer understanding of QSP-achievable polynomials, if we give up the imaginary degree of freedom in our polynomials. Generalizing the notion of pbeing QSP-achievable, we say that the pair of polynomials (p,q) is QSP-achievable if there are phase factors such that p and q are the two polynomials in the characterization of Lemma 36.

Theorem 38 ([GSLW19, Theorem 5, Lemma 6]). Let $p_{\Re}(x), q_{\Re}(x) \in \mathbb{R}[x]$ be real-valued polynomials with p of degree n. Then there exist $p, q \in \mathbb{C}[x]$ such that (p,q) is QSP-achievable and $p_{\Re} = \Re(p), q_{\Re} = \Re(q)$ if and only if

- (a) q_{\Re} has degree $\leq n 1$;
- (b) (p_{\Re}, q_{\Re}) are (even, odd) or (odd, even);
- (c') $(p_{\Re}(x))^2 + (1-x^2)(q_{\Re}(x))^2 \le 1$ for $x \in [-1,1]$.

To interpret this claim, it implies that if we have real polynomials where the "unit norm" constraint is merely an inequality (c'), then we can add imaginary components to make it an equality, so that by Theorem 37 these supplemented polynomials are achievable. Theorem 6 follows as a corollary of this theorem, taking q = 0.

Proof. The "only if" direction is more straightforward: if (p,q) is QSP-achievable, then the real parts of p and q satisfy (a), (b), and (c') by Theorem 37.

The "if" direction requires some work: given p_{\Re} and q_{\Re} , we need to find some $p_{\Im} \in \mathbb{R}[x]$ and $q_{\Im} \in \mathbb{R}[x]$ of the right degree and parity such that $p \coloneqq p_{\Re} + ip_{\Im}$ and $q \coloneqq q_{\Re} + iq_{\Im}$ satisfy

$$|p(x)|^{2} + (1 - x^{2})|q(x)|^{2} = p_{\Re}^{2} + p_{\Im}^{2} + (1 - x^{2})(q_{\Re}^{2} + q_{\Im}^{2}) \equiv 1.$$

This would imply the conclusion via Theorem 37. Consider $P = 1 - p_{\Re}^2 - (1 - x^2)q_{\Re}^2$, which is an even polynomial with real coefficients. By assumption (c'), we know P is non-negative in $x \in [-1, 1]$. If we can write $P = A^2 + (1 - x^2)B^2$ where $\deg(A) \leq \deg(P)$, $\deg(B) \leq \deg(P) - 1$, and (A, B) are (odd, even) or (even, odd), then we are done. If we have two polynomials P and Q that can be expressed in the above way, then their product can:

$$PQ = (A_P^2 + (1 - x^2)B_P^2)(A_Q^2 + (1 - x^2)B_Q^2)$$

= $|A_P + i\sqrt{1 - x^2}B_P|^2|A_Q + i\sqrt{1 - x^2}B_Q|^2$
= $|(A_PA_Q - (1 - x^2)B_PB_Q) + i\sqrt{1 - x^2}(A_PB_Q + A_QB_P)|^2$
= $(A_PA_Q - (1 - x^2)B_PB_Q)^2 + (1 - x^2)(A_PB_Q + A_QB_P)^2.$

So, it suffices to prove that this representation is possible for all "irreducible" polynomials, i.e. even polynomials with real coefficients that are non-negative in [-1, 1], that cannot be decomposed into the product of two polynomials satisfying the same criteria. Using the fundamental theorem of algebra, we can give a complete list of such polynomials up to scaling by a positive number.

- 1. (Polynomials with roots (0,0)) $R(x) = x^2$; here, A = x and B = 0.
- 2. (Polynomials with roots (-s, -s, s, s) for $s \in (0, 1)$) $R(x) = (x^2 s^2)^2$; here, $A = (x^2 s^2)$ and B = 0.
- 3. (Polynomials with roots (-s, s) for $s \ge 1$) $R(x) = s^2 x^2$; here, $A = \sqrt{s^2 1}x$ and B = s.
- 4. (Polynomials with roots (-is, is) for s > 0) $R(x) = x^2 + s^2$ for C > 0; here, $A = \sqrt{s^2 + 1}x$ and B = s.
- 5. (Polynomials with roots (s+it, s-it, -s+it, -s-it) for s, t > 0) $R(x) = x^4 + 2x^2(t^2 s^2) + (s^2 + t^2)^2$; here, $A = cx^2 (s^2 + t^2)$ and $B = \sqrt{c^2 1}x$ for $c = s^2 + t^2 + \sqrt{2(s^2 + 1)t^2 + (s^2 1)^2 + t^4}$.

Because all of these polynomials can be written in the desired representation, all polynomials satisfying the criteria can. $\hfill \Box$

B More applications of the CS decomposition

To shed light on the CS decomposition as capturing interactions between subspaces, in this section we derive two further applications beyond QSVT. We note that we do not claim the intuition provided in this section is very helpful for understanding the particular application of QSVT. However, we hope the reader is sufficiently convinced of the virtue of the CS decomposition as a technical tool, and this section serves to provide additional background on this tool.

B.1 Principal angles

In this section we consider two rank-*a* subspaces $\mathfrak{X} = \text{Image}(\mathbf{\Pi}_x) \subset \mathbb{C}^d$ and $\mathfrak{Y} = \text{Image}(\mathbf{\Pi}_y) \subset \mathbb{C}^d$, for some $a \in [d]$. For $k \in [a]$, we define the k^{th} principal angle between \mathfrak{X} and \mathfrak{Y} recursively via

$$\cos(\theta_k) \coloneqq \max_{\substack{x \in \mathcal{X} \\ \|x\|_2 = 1 \ \|y\|_2 = 1}} \max_{\substack{y \in \mathcal{Y} \\ \|y\|_2 = 1}} \langle x, y \rangle \text{ subject to } x \perp x_i, y \perp y_i \text{ for all } i < k,$$
(42)

where x_k , y_k are the *principal vectors* realizing the maximum above. In other words, the first principal angle θ_1 is the largest angle between a vector in \mathcal{X} and a vector in \mathcal{Y} ; θ_2 is the largest angle between vectors in the subspaces of \mathcal{X} and \mathcal{Y} orthogonal to the vectors achieving θ_1 , and so on. This definition only depends on the subspaces, and so is agnostic to the choice of basis for \mathcal{X} and \mathcal{Y} .

Lemma 39. Let $\mathbf{X}, \mathbf{Y} \in \mathbb{C}^{d \times a}$ be such that their columns are orthonormal bases for \mathfrak{X} and \mathfrak{Y} , respectively. Then, the values $\{\cos(\theta_k)\}_{k \in [a]}$ are the singular values of $\mathbf{X}^{\dagger}\mathbf{Y}$. Further, letting \mathbf{VCW}^{\dagger} be the SVD of $\mathbf{X}^{\dagger}\mathbf{Y}$, the principal vectors between $\mathfrak{X}, \mathfrak{Y}$ are columns of \mathbf{XV}, \mathbf{YW} .

Proof. This result follows from the variational characterization of singular values and vectors. Let $\mathbf{C} = \mathbf{diag}(c)$. Fix $k \in [a]$ and suppose inductively the conclusion holds for all i < k. Recall that the SVD is recursively defined by

$$c_k = \max_{\|v\|_2 = \|w\|_2 = 1} v^{\dagger} \mathbf{X}^{\dagger} \mathbf{Y} w, \text{ subject to } v \perp v_i, \ w \perp w_i \text{ for all } i < k.$$
(43)

Inductively assume $x_i = \mathbf{X}v_i$ for all i < k. Notice that every unit vector in \mathfrak{X} can be written as $\mathbf{X}v$ for some unit $v \in \mathbb{C}^a$, and since $\mathbf{X}^{\dagger}\mathbf{X} = \mathbf{I}$, we have $\mathbf{X}v \perp \mathbf{X}v_i$ iff $v \perp v_i$. By reasoning similarly for \mathcal{Y} , we conclude the optimization problems (42) and (43) are the same under the transformation $x \leftarrow \mathbf{X}v$ and $y \leftarrow \mathbf{Y}w$. Hence setting $x_k \leftarrow \mathbf{X}v_k$ and $y_k \leftarrow \mathbf{Y}w_k$ we may continue inducting. \Box

Now we explain the connection between the above digression and the CS decomposition. Lemma 39 shows that we can find the principal angles between \mathcal{X} and \mathcal{Y} by taking orthogonal bases, \mathbf{X} and \mathbf{Y} , and computing the SVD of $\mathbf{X}^{\dagger}\mathbf{Y}$. We could further ask: what are the principal angles of the orthogonal subspaces, $\mathcal{X}_{\perp} = \{u \mid \langle u, x \rangle = 0 \text{ for all } x \in \mathcal{X}\}$ and $\mathcal{Y}_{\perp} = \{u \mid \langle u, y \rangle = 0 \text{ for all } y \in \mathcal{Y}\}$? We can apply the same lemma on bases for the subspaces, \mathbf{X}^{\perp} and \mathbf{Y}^{\perp} , to compute them; however, we can say more. First, let the following matrices be unitary completions of \mathbf{X} , \mathbf{Y} :

$$\begin{pmatrix} \mathbf{X} & \mathbf{X}_{\perp} \end{pmatrix}, \begin{pmatrix} \mathbf{Y} & \mathbf{Y}_{\perp} \end{pmatrix}.$$
(44)

We next take the CS decomposition (Theorem 1) of the product (which is also unitary),

$$\mathbf{U} = \begin{pmatrix} \mathbf{X} & \mathbf{X}_{\perp} \end{pmatrix}^{\dagger} \begin{pmatrix} \mathbf{Y} & \mathbf{Y}_{\perp} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^{\dagger} \mathbf{Y} & \mathbf{X}^{\dagger} \mathbf{Y}_{\perp} \\ \mathbf{X}_{\perp}^{\dagger} \mathbf{Y} & \mathbf{X}_{\perp}^{\dagger} \mathbf{Y}_{\perp} \end{pmatrix}.$$

This gives us $\mathbf{V}_1, \mathbf{V}_2, \mathbf{W}_1, \mathbf{W}_2$ such that

$$\begin{pmatrix} \mathbf{V}_1^{\dagger} \mathbf{X}^{\dagger} \mathbf{Y} \mathbf{W}_1 & \mathbf{V}_1^{\dagger} \mathbf{X}^{\dagger} \mathbf{Y}_{\perp} \mathbf{W}_2 \\ \mathbf{V}_2^{\dagger} \mathbf{X}_{\perp}^{\dagger} \mathbf{Y} \mathbf{W}_1 & \mathbf{V}_2^{\dagger} \mathbf{X}_{\perp}^{\dagger} \mathbf{Y}_{\perp} \mathbf{W}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{pmatrix}$$

of the form in Theorem 1. This gives simultaneous SVDs for each block, and thus gives the principal angles and vectors for all combinations of $\mathcal{X}, \mathcal{X}_{\perp}, \mathcal{Y}$, and \mathcal{Y}_{\perp} . In particular, we can see that the principal angles of $(\mathcal{X}, \mathcal{Y})$ and $(\mathcal{X}_{\perp}, \mathcal{Y}_{\perp})$ are related: up to padding by 0's and 1's, they are identical!

Further, we can take $\mathbf{X} \leftarrow \mathbf{X}\mathbf{V}_1$, $\mathbf{Y} \leftarrow \mathbf{Y}\mathbf{W}_1$, etc. without affecting the induced subspaces (since e.g. $\mathbf{X}\mathbf{V}_1\mathbf{V}_1^{\dagger}\mathbf{X}^{\dagger} = \mathbf{X}\mathbf{X}^{\dagger}$), but such that after this transformation we simply have

$$\begin{pmatrix} \mathbf{X}^{\dagger}\mathbf{Y} & \mathbf{X}^{\dagger}\mathbf{Y}_{\perp} \\ \mathbf{X}_{\perp}^{\dagger}\mathbf{Y} & \mathbf{X}_{\perp}^{\dagger}\mathbf{Y}_{\perp} \end{pmatrix} = \begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{pmatrix}.$$
 (45)

That is, we can choose canonical basis representations of \mathfrak{X} , \mathfrak{Y} , and their complement subspaces consistently, such that every pairing directly induces the "principal angles and vectors" defined above. We remark that this argument extends just fine to \mathfrak{X} , \mathfrak{Y} of different dimensions.

B.2 Jordan's lemma

Next we derive Jordan's lemma [Jor75], a useful way of decomposing \mathbb{C}^d into subspaces (induced by a unitary matrix) which are jointly compatible with two projection matrices in a certain sense. We note that Jordan's lemma has seen varied implicit or explicit uses in the quantum computing literature (including a suggestion by [GSLW19]), and refer to [Reg06] for an account of this.

Lemma 40. Let $\Pi_x, \Pi_y \in \mathbb{C}^{d \times d}$ be projection matrices. There exists a unitary matrix $\mathbf{U} \in \mathbb{C}^{d \times d}$ with columns $\{u_i\}_{i \in [d]}$, and a partition of [d] into $S := \{S_j\}_{j \in [k]}$ such that $|S_j| \in \{1, 2\}$ for all $j \in [k]$, and $\mathbf{U}^{\dagger} \Pi_x \mathbf{U}$, $\mathbf{U}^{\dagger} \Pi_y \mathbf{U}$ are block-diagonal with blocks indexed by $\{S_j\}_{j \in [k]}$; each block is trace-1. Moreover for each $S_j = \{i, i'\}$ where $|S_j| = 2$, we have $\Pi_x u_i \parallel \Pi_x u_{i'}$ and $\Pi_y u_i \parallel \Pi_y u_{i'}$.

In other words, there is a choice of subspaces given by **U** (whose columns are partitioned by S) such that if $i, i' \in [n]$ where $i \in S_j$ and $i' \in S_{j'}$, $u_i^{\dagger} \Pi_x u_{i'} \neq 0$, $u_i^{\dagger} \Pi_y u_{i'} \neq 0$ iff j = j'. Moreover, the second part states each of the 2×2 blocks in $\mathbf{U}^{\dagger} \Pi_x \mathbf{U}$ and $\mathbf{U}^{\dagger} \Pi_y \mathbf{U}$ are in fact rank-1 and trace-1.

Proof of Lemma 40. We first prove this in the special case when Π_x and Π_y are dimension- $\frac{d}{2}$ projectors with "no intersection," and briefly discuss how to extend this to the general case.

Special case. Suppose Π_x and Π_y are dimension- $\frac{d}{2}$ projectors, and let us further make one following restriction. Let $\Pi_x = \mathbf{X}\mathbf{X}^{\dagger}$ and $\Pi_y = \mathbf{Y}\mathbf{Y}^{\dagger}$ where \mathbf{X} , \mathbf{Y} and their "completions" \mathbf{X}_{\perp} , \mathbf{Y}_{\perp} (in the sense that the matrices (44) are unitary) are chosen such that (45) holds, as guaranteed by Theorem 1; we make the restriction that for $\mathbf{C} = \mathbf{X}^{\dagger}\mathbf{Y}$, all of the diagonal entries of \mathbf{C} are in (0, 1).

The previous section shows this means $\operatorname{Span}(\mathbf{X}) \cap \operatorname{Span}(\mathbf{Y}) = \emptyset$. We choose the basis inducing **U** as follows. Let the columns of **X** be $\{x_j\}_{j \in [\frac{d}{2}]} \subset \mathbb{C}^d$ and the columns of **Y** be $\{y_j\}_{j \in [\frac{d}{2}]} \subset \mathbb{C}^d$. For $j \in [\frac{d}{2}]$, we let $\{u_{2j-1}, u_{2j}\} \subset \mathbb{C}^d$ be an arbitrary basis of $\operatorname{Span}\{x_j, y_j\}$. We claim such **U** meets the requirements, where each $S_j = \{2j-1, 2j\}$ in our partition. For all $j \in [\frac{d}{2}]$, since $\mathbf{X}^{\dagger}\mathbf{Y} = \mathbf{C}$,

$$\mathbf{\Pi}_x y_j = \mathbf{X} \mathbf{X}^{\dagger} y_j = \mathbf{X} (\mathbf{C} e_j) = c_j x_j.$$

So, Π_x maps $\operatorname{Span}\{x_j, y_j\}$ to $\operatorname{Span}\{x_j\}$, and similarly Π_y maps $\operatorname{Span}\{x_j, y_j\}$ to $\operatorname{Span}\{y_j\}$. This proves the second part of the lemma, namely that Π_x and Π_y act as rank-1 projectors on each block of the partition. For the first part (the block-diagonal structure), it suffices to show that for all $j \neq j' \in [\frac{d}{2}]$, $x_j \perp \operatorname{Span}\{x_{j'}, y_{j'}\}$, since we already argued Π_x maps $\operatorname{Span}\{x_j, y_j\}$ to $\operatorname{Span}\{x_j\}$. By orthonormality of \mathbf{X} , $x_j \perp x_{j'}$, and since we used the canonical choice where $\mathbf{X}^{\dagger}\mathbf{Y} = \mathbf{C}$, indeed $x_j \perp y_{j'}$ as well. To see that each block has trace 1, write $x_j = \alpha_j u_{2j-1} + \beta_j u_{2j}$. We have that $\mathbf{X}\mathbf{X}^{\dagger}u_{2j-1} = \alpha_j x_j$ and $\mathbf{X}\mathbf{X}^{\dagger}u_{2j} = \beta_j x_j$; to see this, we already argued that u_{2j-1} , u_{2j} are orthogonal to all x_i for $i \neq j$, since $x_j \perp x_i$ and $y_j \perp x_i$. Hence, the 2 × 2 block of Π_x indexed by S_j is the outer product of $(\alpha_j \quad \beta_j)$ which clearly has trace 1; a similar argument applies to Π_y .

General case. More generally, we can "pull out" vectors corresponding to the I and 0 blocks of the decomposition in Theorem 1, when the dimensions are unequal. Concretely, again let $\Pi_x = \mathbf{X}\mathbf{X}^{\dagger}$ and $\Pi_y = \mathbf{Y}\mathbf{Y}^{\dagger}$, such that $\mathbf{X}^{\dagger}\mathbf{Y} = \mathbf{C}$ and \mathbf{C} has the form guaranteed by Theorem 1. Further, assume $d_1 = \dim(\operatorname{Span}(\mathbf{X})) \geq \dim(\operatorname{Span}(\mathbf{Y})) = d_2$. Whenever there is a 1 entry in \mathbf{C} , this corresponds to a subset of size 1 in the partition with the column of \mathbf{U} set to the corresponding vector in $\operatorname{Span}(\mathbf{X}) \cap \operatorname{Span}(\mathbf{Y})$. Whenever there is a 0 entry we simply pull out the corresponding vector in $\operatorname{Span}(\mathbf{X}) \setminus \operatorname{Span}(\mathbf{Y})$ into its own block in the partition. Finally, when $\operatorname{Span}(\mathbf{X}) \oplus \operatorname{Span}(\mathbf{Y}) \neq \mathbb{C}^d$ we find any orthonormal basis of $(\operatorname{Span}(\mathbf{X}) \oplus \operatorname{Span}(\mathbf{Y}))^c$ and add them it as columns of \mathbf{U} . It is an exercise to check the overall dimension of \mathbf{U} after this process is d.

C Proof of Carlini's formula

Proof of Lemma 19. Recall $2I_n(t)$ is the n^{th} Chebyshev coefficient for $\exp(tx)$. It will be slightly more convenient for us to reparameterize and bound $I_n(nt)$. Without loss of generality $t \ge 0$, since $\exp(tx)$ and $\exp(-tx)$ have the same Chebyshev coefficients up to sign. By (13), for $n \ge 1$,

$$I_n(nt) = \frac{1}{2\pi i} \oint_{|z|=1} z^{-n-1} \exp(\frac{nt}{2}(z+z^{-1})) dz$$
$$= \frac{1}{2\pi i} \oint_{|z|=1} \exp\left(-n\left(\log(z) - \frac{t}{2}(z+\frac{1}{z})\right)\right) \frac{dz}{z}.$$

We choose a contour circling around the origin once; by Cauchy's theorem, this results in the same integral as the contour does not cross the origin. We parameterize it via $z = re^{i\theta}$ and construct r as a function of θ . Consider the (rescaled) imaginary part of the expression in the exponential:

$$\log(z) - \frac{t}{2}(z + \frac{1}{z}) = \log(r) + i\theta - \frac{t}{2n}(re^{i\theta} + \frac{1}{r}e^{-i\theta})$$
$$\implies \Im\left(\log(z) - \frac{t}{2}(z + \frac{1}{z})\right) = \theta - \frac{t}{2}(r\sin(\theta) - \frac{1}{r}\sin(\theta)).$$

We wish to make the imaginary part constant; we set it equal to ψ , and solve for r:

$$\psi = \theta - \frac{t}{2} (r \sin(\theta) - \frac{1}{r} \sin(\theta))$$
$$\implies \frac{1}{2} (r - \frac{1}{r}) = \frac{\theta - \psi}{t \sin(\theta)}$$
$$\implies r = \frac{\theta - \psi}{t \sin(\theta)} + \sqrt{\left(\frac{\theta - \psi}{t \sin(\theta)}\right)^2 + 1}.$$

Above, we use that $r = x + \sqrt{x^2 + 1}$ and $\frac{1}{r} = -x + \sqrt{x^2 + 1}$ is a solution to $\frac{1}{2}(r - \frac{1}{r}) = x$. Now, we choose to take our contour for θ from $-\pi + \psi$ to $\pi + \psi$, and we take $\psi = 0$. So the contour is

$$z = re^{i\theta} = \left(\frac{\theta}{t\sin(\theta)} + \sqrt{\left(\frac{\theta}{t\sin(\theta)}\right)^2 + 1}\right)e^{i\theta} \text{ for } \theta \in [-\pi, \pi], \text{ where } \frac{\sin \theta}{\theta} \coloneqq 1.$$

Since $r \ge 1$ always, the contour winds once counter-clockwise around zero and is valid for evaluating the integral. By design, on this contour $\Im(\log(z) - \frac{t}{2}(z - \frac{1}{z}))$ vanishes, and the real part is

$$F(\theta, t) \coloneqq \Re\left(\log(z) - \frac{t}{2}(z + \frac{1}{z})\right) = \log(r) - \frac{t}{2}(r + \frac{1}{r})\cos(\theta)$$
$$= \log\left(\frac{\theta}{t\sin(\theta)} + \sqrt{\left(\frac{\theta}{t\sin(\theta)}\right)^2 + 1}\right) - t\cos(\theta)\sqrt{\left(\frac{\theta}{t\sin(\theta)}\right)^2 + 1}.$$

Now, we consider the original integral along this contour:

$$I_n(nt) = \frac{1}{2\pi i} \oint \exp\left(-n\left(\log(z) - \frac{t}{2}(z + \frac{1}{z})\right)\right) \frac{\mathrm{d}z}{z}$$

$$= \frac{1}{2\pi i} \int_{-\pi}^{\pi} \exp\left(-n \cdot F(\theta, t)\right) \left(\frac{\mathrm{d}r(\theta)}{\mathrm{d}\theta} \cdot \frac{1}{r(\theta)} + i\right) \mathrm{d}\theta$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp\left(-n \cdot F(\theta, t)\right) \mathrm{d}\theta$$

$$= \frac{1}{\pi} \int_{0}^{\pi} \exp\left(-n \cdot F(\theta, t)\right) \mathrm{d}\theta.$$

The last two lines use that as $F(\theta, t)$ and $r(\theta)$ are even functions in θ , the piece of the integral corresponding to $\frac{\mathrm{d}r(\theta)}{r(\theta)\mathrm{d}\theta}$ vanishes. From here, it becomes a matter of bounding $F(\theta, t)$. We compute

$$\frac{\partial}{\partial \theta} F(\theta, t) = \sqrt{(t\sin(\theta))^2 + \theta^2} + \frac{(1 - \theta\cot(\theta))^2}{\sqrt{(t\sin(\theta))^2 + \theta^2}}$$

First, we notice that $\frac{\partial}{\partial \theta} F \ge 0$, so F is increasing. Second, we notice that, for $\theta \in [0, \pi/2]$,

$$\frac{\partial}{\partial \theta} F(\theta, t) \ge \theta \sqrt{(t \sin(\theta)/\theta)^2 + 1} \ge \theta \sqrt{t^2(4/\pi^2) + 1}$$

Integrating, we get that, for $\theta \in [0, \frac{\pi}{2}]$, $F(\theta, t) - F(0, t) \ge \frac{1}{2}\theta^2\sqrt{1 + 4t/\pi^2}$, and using that $F(0, t) = \log(t^{-1} + \sqrt{t^{-2} + 1}) - \sqrt{1 + t^2}$, we have the desired

$$\begin{split} I_n(nt) &= \frac{1}{\pi} \int_0^{\pi} \exp(-nF(\theta,t)) \mathrm{d}\theta \\ &\leq \frac{2}{\pi} \int_0^{\pi/2} \exp(-nF(\theta,t)) \mathrm{d}\theta \\ &\leq \frac{2}{\pi} \int_0^{\pi/2} \exp(-\frac{n}{2}\theta^2 \sqrt{1+4t^2/\pi^2}) \exp(-nF(0,t)) \mathrm{d}\theta \\ &= \frac{\exp(-nF(0,t))}{\pi} \int_0^{\pi} \exp(-\frac{1}{2}\theta^2 \sqrt{n^2+4n^2t^2/\pi^2}) \mathrm{d}\theta \\ &< \frac{\exp(-nF(0,t))}{\pi} \int_0^{\infty} \exp(-\frac{1}{2}\theta^2 \sqrt{n^2+4n^2t^2/\pi^2}) \mathrm{d}\theta \\ &= \frac{\exp(-nF(0,t))}{\sqrt{2\pi\sqrt{n^2+4n^2t^2/\pi^2}}} \\ &= \frac{\exp(\sqrt{n^2+n^2t^2})}{(4\pi^2n^2+16n^2t^2)^{1/4}(t^{-1}+\sqrt{t^{-2}+1})^n} \\ &\leq \frac{\exp(\sqrt{n^2+n^2t^2})(\sqrt{t^{-2}+1}-t^{-1})^n}{2(n^2+n^2t^2)^{1/4}}. \end{split}$$

D Deferred proofs from Section 3.5

Lemma 31. For $\delta \in (0, 1)$, $(1+\delta)E_{1+\alpha}$ is contained in the interior of E_{σ} , where $\sigma = 1 + 3(\alpha + \sqrt{\delta})$.

Proof. Recall from Theorem 20 that we can parameterize the boundary of E_{ρ} as $\frac{1}{2}(\rho + \rho^{-1})\cos\theta + \frac{1}{2}(\rho - \rho^{-1})\sin\theta$, for $\theta \in [0, 2\pi]$. Hence, to prove the stated inclusion it suffices to prove that $\frac{1}{2}(\sigma + \sigma^{-1}) \ge (1 + \delta)\frac{1}{2}(\rho + \rho^{-1})$ and $\frac{1}{2}(\sigma - \sigma^{-1}) \ge (1 + \delta)\frac{1}{2}(\rho - \rho^{-1})$. By Fact 30,

$$\begin{aligned} \frac{1}{2}(\sigma + \sigma^{-1}) - (1+\delta)\frac{1}{2}(\rho + \rho^{-1}) &= 1 + \frac{9(\alpha^2 + 2\alpha\sqrt{\delta} + \delta)}{2(1+3\alpha+3\sqrt{\delta})} - (1+\delta)\left(1 + \frac{\alpha^2}{2(1+\alpha)}\right) \\ &\geq 1 + \frac{2(\alpha^2 + 2\alpha\sqrt{\delta} + \delta)}{2(1+\alpha)} - (1+\delta)\left(1 + \frac{\alpha^2}{2(1+\alpha)}\right) \\ &= \frac{2(\alpha^2 + 2\alpha\sqrt{\delta} + \delta) - (1+\delta)\alpha^2 - 2\delta(1+\alpha)}{2(1+\alpha)} \geq 0. \end{aligned}$$

Further, since $\sigma \ge (1+\delta)(1+\alpha)$ and $\sigma - \sigma^{-1}$ increases in σ ,

$$\frac{1}{2}(\sigma - \sigma^{-1}) - (1 + \delta)\frac{1}{2}(\rho - \rho^{-1}) \ge \frac{1}{2}((1 + \delta)\rho - ((1 + \delta)\rho)^{-1}) - (1 + \delta)\frac{1}{2}(\rho - \rho^{-1}) \ge 0.$$

Lemma 32. There are universal constants C, c > 0 such that, for $n \ge 0$ and $x, y \in \mathbb{R}$, $|y| \le c$,

$$|T_n(x+\imath y)| \le \begin{cases} (1+C\sqrt{|y|})^n & |x| \le 1\\ (x+\sqrt{x^2-1}+C\sqrt{|xy|})^n & |x| > 1. \end{cases}$$

Proof. The only points not lying on the boundary of a Bernstein ellipse are the interval [-1, 1], where the Chebyshev polynomials are at most 1. Otherwise, suppose z := x + iy lies on the boundary of the Bernstein ellipse E_{ρ} , for some $\rho > 1$. This implies that for some $\theta \in [-\pi, \pi]$,

$$\frac{1}{2}(\rho + \rho^{-1})\cos\theta = x, \ \frac{1}{2}(\rho - \rho^{-1})\sin\theta = y,$$

which follows from the parameterization of the Bernstein ellipse in Theorem 20. Let $s = \cos \theta$ and $t = \frac{1}{2}(\rho + \rho^{-1})$. Noting that $\sqrt{t^2 - 1} = \frac{1}{2}(\rho - \rho^{-1})$, we then have the system of equations

$$ts = x, \ \sqrt{1 - s^2}\sqrt{t^2 - 1} = y \implies t^4 - (1 + x^2 + y^2)t^2 + x^2 = 0$$

Hence, solving a quadratic equation in $t^2=\frac{1}{4}(\rho^2+\rho^{-2})+\frac{1}{2}$ yields

$$\frac{1}{2}(\rho^2 + \rho^{-2}) = x^2 + y^2 + \sqrt{(1 + x^2 + y^2)^2 - 4x^2} =: D,$$

and then $\rho = (D + \sqrt{D^2 - 1})^{\frac{1}{2}}$. By definition, the Chebyshev polynomial satisfies

$$T_n(z) \le \frac{1}{2}(\rho^n + \rho^{-n}) \le \rho^n,$$
(46)

so it suffices to establish bounds on ρ . Next, assume without loss of generality that $y \ge 0$, as Chebyshev polynomials are either odd or even and the stated conclusions are unsigned. We bound

$$D = x^{2} + y^{2} + \sqrt{1 + x^{4} + y^{4} - 2x^{2} + 2y^{2} + 2x^{2}y^{2}}$$

= $x^{2} + y^{2} + \sqrt{(1 - x^{2})^{2} + 2y^{2}(1 + x^{2}) + y^{4}} \le x^{2} + |1 - x^{2}| + O(y\sqrt{1 + x^{2}}).$

When $|x| \leq 1$, then D = 1 + O(y) and $\rho = 1 + O(\sqrt{y})$, establishing the conclusion via (46). Otherwise, when |x| > 1, we have $D = 2x^2 - 1 + O(|x|y)$, and then

$$\begin{split} \rho &= \sqrt{2x^2 - 1 + \sqrt{4x^4 - 4x^2 + O(|x^3|y)}} \\ &= \sqrt{\left(|x| + \sqrt{x^2 - 1}\right)^2 + O\left(\sqrt{|x^3|y}\right)} = |x| + \sqrt{x^2 - 1} + O(\sqrt{|xy|}), \end{split}$$

again proving the desired claim via (46).

31

Lemma 34 (Thresholding function). For $\mu, s > 0$, let $r(z) := \frac{1}{2}(\operatorname{erf}(s(\mu+z)) + \operatorname{erf}(s(\mu-z)))$. When $z \in \mathbb{R}, 0 \le r(z) \le 1$. When $x, y \in \mathbb{R}$ and z = x + iy, $|r(z) - r(x)| \le \exp(-s^2(\mu - |x|)^2)|\operatorname{erf}(isy)|$.

Proof. To see the first claim, let $z \in \mathbb{R}$. As previously discussed we have $\operatorname{erf}(s(\mu+z)), \operatorname{erf}(s(\mu-z)) \leq 1$, giving the upper bound. For the lower bound, since erf is odd and increasing, $z \geq 0$ implies

$$\operatorname{erf}(s(\mu+z)) + \operatorname{erf}(s(\mu-z)) = \operatorname{erf}(s(\mu+z)) - \operatorname{erf}(s(-\mu+z)) \ge 0,$$

and a similar argument handles the case $z \leq 0$. Next, for the second claim we first observe

$$|\operatorname{erf}(z) - \operatorname{erf}(x)| = \frac{2}{\sqrt{\pi}} \left| \int_{t=x}^{t=x+iy} e^{-t^{2}} dt \right|$$

$$= \frac{2e^{-x^{2}}}{\sqrt{\pi}} \left| \int_{0}^{y} e^{-2ixt+t^{2}} dt \right|$$

$$\leq \frac{2e^{-x^{2}}}{\sqrt{\pi}} \int_{0}^{|y|} e^{t^{2}} dt = e^{-x^{2}} |\operatorname{erf}(iy)|.$$
 (47)

The last line used the triangle inequality. Finally,

$$\begin{aligned} |\operatorname{erf}(z) - \operatorname{erf}(x)| &\leq \frac{1}{2} |\operatorname{erf}(s(\mu + z)) - \operatorname{erf}(s(\mu + x))| + \frac{1}{2} |\operatorname{erf}(s(\mu - z)) - \operatorname{erf}(s(\mu - x))| \\ &\leq \frac{1}{2} \Big(e^{-s^2(\mu + x)^2} |\operatorname{erf}(\imath sy)| + e^{-s^2(\mu - x)^2} |\operatorname{erf}(-\imath sy)| \Big), \end{aligned}$$

where we used (47), and we conclude by noting $\mu + x, \mu - x \ge \mu - |x|$.