SCUOLA
NORMALE
SUPERIORE

Classe di Scienze Matematiche, Fisiche e Naturali

PhD Thesis

Exploiting rank structures in the numerical solution of Markov chains and matrix functions

Candidate
Stefano Massei

Advisor
Prof. Dario Andrea Bini

Contents

Introduction

A recurrent theme in numerical analysis is to provide instruments capable of treating large scale problems. With "large scale" we mean that the size of the data to be processed is comparable with the total amount of our memory resources. In this situation we need to design algorithms with a linear or linear-polylogarithmic complexity in terms of time and storage space. Generally, the successful strategies rely on exploiting hidden structures in the input data —if present— to speed up the procedures designed for small scale instances.

Intuitively, one avoids numerical algorithms involving operations with big, fully populated matrices. The reason is the large number of floating point operations; e.g. the most advanced algorithm for performing the multiplication of two general $n \times n$ matrices require $\mathcal{O}(n^\gamma)$ flops, with $2 < \gamma < 3$ and 2 is an insuperable barrier. It is therefore not surprising that researchers try to take advantage of matrices with a few non zero elements. In particular they studied the class of *sparse matrices* i.e., those who have $\mathcal{O}(n)$ non zero entries. An important subset of sparse matrices are the banded matrices whose elements are located close to the main diagonal.

Techniques for efficient storage and calculation of sparse matrices have been studied widely [86]. However, if we want to apply sparse computations in an algorithm we have to attenuate as much as possible the fill-in of the intermediate results. From this point of view, sparsity turns out to be fragile. An immediate example of loss of sparsity is given by the inverse of a band matrix, see Figure 1. For this reason, it is interesting to look for more flexible structures.

$$\begin{bmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & -1 & 2 & -1 & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 & 5 & 5 \\ 1 & 2 & 3 & 4 & 5 & 6 & 6 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{bmatrix}$$

Figure 1.: Example of loss of sparsity; the inverse of a tridiagonal matrix

Recently, much attention has been payed to a generalization of banded matrices: the quasiseparable matrices. The latter are characterized in terms of the submatrices entirely contained into their strictly lower or upper triangular part, which we call off-diagonal submatrices. More precisely, a matrix is said quasiseparable if each of its off-diagonal submatrices has rank bounded by a small constant. The quasiseparability rank is defined as the maximum of the ranks of the off-diagonal submatrices. Roughly speaking, we replace the property of containing many zero entries with the one of containing large sub-blocks with low-rank. Observe that, banded matrices enjoy both the sparsity property and the quasiseparability because a generic off-diagonal block has a rank less or equal to the bandwidth. In general, quasiseparable matrices may not be sparse but can still be represented with a relatively small number of parameters.

Quasiseparable rank enjoy some nice properties, such as the invariance under inversion and the sub-additivity under addition and multiplication. In the last two decades, the problem of taking advantage of this structure has been studied by many authors such as Boito [24], Börm, Grasedyck, Hackbusch [26], Chandrasekaran[33], Eidelman [38, 40], Gemignani [39], Mastronardi, Van Barel, Vandebril et al [97, 98, 96].

A crucial target of these studies was to find a representation that enables one to perform the storage and the matrix operations cheaply. The one we take into account in this thesis belongs to the family of Hierarchical representations originally introduced by Hackbusch [53, 54] in the context of integral and partial differential equations and studied also by Börm and Grasedyck [26]. This tool consists in a class of recursive block representations with structured sub-matrices that allows the treatment of a number of data-sparse patterns. Here, we consider a particular member of this family —sometimes called *hierarchical off-diagonal low-rank* representation (HODLR)— which has a simple formulation and an effective impact in handling quasiseparable matrices.

A substantial part of this thesis is devoted to analyze the numerical preservation of the quasiseparable structure in some procedures used for solving linear and quadratic matrix equations and for computing functions of matrices. Then, by means of the HODLR representation, algorithms for large scale problems are obtained and tested. The ideas are generalized to the setting of semi-infinite matrices where —in order to deal with an infinite amount of data— we assume some further Toeplitz structure. In particular, a new arithmetic is provided for a class of semi-infinite quasi-Toeplitz matrices. Finally, this tool is used to solve equations and to compute functions of semi-infinite quasi-Toeplitz matrices.

The first two chapters are introductory. In Chapter 1 we specify our notations and we recall some classical results. Chapter 2 provides a brief description of the probabilistic model where most of the questions that inspired this work have been raised.

In Chapter 3 we introduce quasiseparable matrices and slight variations of this structure that have been considered in the literature. Then, the HODLR representation and HODLR arithmetic are described, emphasizing improvements to the computational complexity.

In Section 3.5 the applicability of the HODLR representation is related to fast decay properties of the off-diagonal singular values.

Chapter 4 provides a collection of results concerning singular values inequalities. The goal of this part is to build a framework for the study of the numerical quasiseparable rank. Section 4.6 is about the link between displacement properties and singular values. In particular, we discuss some earlier ideas of Beckermann [7] —published only recently [8]— that make a connection with rational approximation theory.

In Chapter 5 we address the problem of estimating the numerical quasiseparable rank of the matrix $B = f(A)$, where A is quasiseparable and $f(z)$ is a holomorphic function. This task has been previously studied in [46, 47] in the case $f(z) = e^z$. More precisely, the authors prove that computing e^A via a quadrature formula applied to the contour integral definition, yields an approximation of the result with a low quasiseparable rank. We introduce a different analysis which studies the interplay between the off-diagonal singular values of the matrices A and B. The numerical preservation of the structure is related to the existence of good low-degree polynomial approximation of f on a set containing the spectrum of A. In Section 5.3 the approach is generalized to meromorphic function. The key tool of this analysis is Theorem 5.3.1 which provides an extension of the Cauchy integral formula to the case in which some poles lie inside the contour of integration. We then discuss the theoretical bounds and test some strategies for computing functions of quasiseparable matrices with linear-polylogarithmic complexity. The analysis of the quasiseparable rank of matrix functions have been published as an original contribution in [76].

In Chapter 6, we describe the *cyclic reduction* (CR) [17, 23, 16] used as direct method for solving tridiagonal block Toeplitz linear systems and as iterative algorithm for solving certain quadratic matrix equations arising in the study of QBD stochastic processes. The iterative scheme of CR requires to generate some matrix sequences defined by recurrence relations that involve basic arithmetic operations. We deal with the issue of analyzing the quasiseparable rank of the members of these sequences when their starting points have a low quasiseparable rank. In order to do that, in Section 6.3 we introduce $\varphi(z)$ the so-called functional interpretation of the algorithm, that is a Laurent matrix polynomial depending on the matrices generated by the iterations of CR. In Section 6.5 we prove the presence of an exponential decay for the off-diagonal singular values of the members of the sequences. The rate of decay turns out to be linked to the domain of invertibility of $\varphi(z)$. In Section 6.6, we provide a refinement of this approach, based on the displacement rank theory [7, 8]. Using the tools of Section 4.6 we relate the preservation of the quasiseparable structure to the existence of high quality solution of particular *Zolotarev* problems [101]. The latter are rational approximation problems encountered in logarithmic potential theory [87]. In Section 6.7.1 we report numerical evidences that confirm a dramatical speed up when using CR with the HODLR representation for solving quadratic matrix equations. We conclude the chapter with Section 6.7.2 where we test the performances of the CR with HODLR representation as direct method for solving certain generalized

Sylvester equations. Our studies about the numerical preservation of the quasiseparable structure in the CR are published in [20] and in [21].

Finally, in Chapter 7 we address the issue of performing arithmetic operations with structured semi-infinite matrices. We introduce the class of *semi-infinite continuosly quasi-Toeplitz* matrices (CQT-matrices), that is matrices of the form $T(a) + E$, where $T(a)$ is the semi-infinite Toeplitz matrix associated with the symbol $a(z) = \sum_{j=-\infty}^{\infty} a_j z^j$ such that $\sum_{j=-\infty}^{\infty} |j a_j| < \infty$ and $E = (e_{i,j})_{i,j \in \mathbb{Z}^+}$ is such that $\sum_{i,j=1}^{+\infty} |e_{i,j}|$ is finite. In particular, the entries of $T(a)$ enjoy a decay moving away from the main diagonal while those of E have a decay along every directions. Therefore, even if they have an infinite number of non zero entries, they can be represented with arbitrary precision using a finite quantity of parameters. In Section 7.3, we show that the set of CQT-matrices equipped with a suited norm is a Banach algebra. In Section 7.4 we provide a practical representation for CQT-matrices and algorithms that perform the arithmetic operations in this class. The latter exploit the decomposition of CQT-matrices for splitting the computations among evaluation interpolation techniques for power series and the finite/low rank arithmetic. With a little more effort, the algorithms for CQT-matrices are extended to handling finite large scale Toeplitz matrices having small size corrections in their corners. The complexity of the resulting procedures depends on the growth of the corrections and on the bandwidth of the Toeplitz part in the intermediate results.

We then apply CQT-matrices to semi-infinite versions of the problems encountered in Chapter 5 and 6. In Section 7.6, we consider quadratic matrix equations with CQT coefficients. We show that in this framework CR generates sequences of CQT-matrices and we provide some convergence properties. The method is tested on some instances in which truncation methods fail.

In Section 7.7 we deal with the computation of functions of finite and semi-infinite CQT-matrices. Two approaches are studied: one based on the power series expansion and the other based on the contour integral definition. Both theoretical and computational aspects are analyzed. The work on the CQT arithmetic and its applications is contained in [19] and [18].

At the end we draw some conclusions about the work presented and we discuss some themes that deserve further investigations.

Chapter 1

Notation and basic tools

With the symbols $\mathbb{C}, \mathbb{R}, \mathbb{R}^+, \mathbb{Z}, \mathbb{Z}^+$ and $\mathbb{N}$ we indicate the sets of complex, real, positive real, relative integer, positive integer and non-negative integer numbers, respectively. We write $\boldsymbol{i}$ for the imaginary unit. With the notations $\mathbb{R}^m$ and $\mathbb{R}^{m\times n}$ we denote the spaces of m-vectors and of $m \times n$-matrices with coefficients in $\mathbb{R}$, respectively. Analogous definitions hold for the set of matrices and vectors with entries in the other sets of numbers.

We often make use of the following subsets of the complex plane:

- $\mathbb{T} := \{z \in \mathbb{C} : \ |z| = 1\}$,
- $B(z_0, r) := \{z \in \mathbb{C} : \ |z - z_0| < r\}$,
- $\mathbb{A}(r, R) := \{z \in \mathbb{C} : \ r < |z| < R\}$.

Moreover, given a generic $A \subset \mathbb{C}$ we denote its border with ∂A and its closure with $\overline{A}$. For example, $\partial B(0,1) = \mathbb{T}$ and $\overline{B(0,1)} = \{z \in \mathbb{C} : \ |z| \leqslant 1\}$. Concerning the matrix and the vector notation, we use the superscripts t and $*$ to indicate the transposition operator and the conjugate transposition operator, respectively. If the latter are used together with the inverse operation we write $-t$ and $-*$. The symbols I and J refer to the square identity and the square counter identity matrices. When their dimensions need to be specified we add a subscript, e.g.,

$$I_m := \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} \in \mathbb{R}^{m\times m}, \qquad J_m := \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix} \in \mathbb{R}^{m\times m}.$$

1.1 Low-rank approximation of matrices

The basis of a number of results in this thesis is the *singular value decomposition* and its properties.

Theorem 1.1.1 (SVD). *Let $A \in \mathbb{C}^{m\times n}$. Then there exist unitary matrices $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ such that*

$$A = U\Sigma V^*, \qquad \Sigma_{ij} = \begin{cases} \sigma_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

and $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_{\min(m,n)} \geqslant 0$.

The triple (U, Σ, V) is called a singular value decomposition *(SVD) of A. The columns of U and V are called left and right* singular vectors*, respectively while the numbers σ_i are called* singular values*.*

Using the notation of the previous theorem, for $1 \leqslant k < \min(m, n)$ we can define the rank-k matrix $\mathcal{T}_k(A) := U_k\Sigma_k V_k^*$ where U_k and V_k are the matrices obtained selecting the first k columns of U and V respectively and Σ_k is the diagonal $k \times k$-matrix with elements $\sigma_1, \ldots, \sigma_k$. $\mathcal{T}_k(A)$ is usually called the *truncated SVD of order k* and it has the following best approximation property.

Theorem 1.1.2 (Eckart-Young-Mirsky). *Let $A \in \mathbb{C}^{m\times n}$ and $1 \leqslant k < \min(m, n)$, then*

$$\|A - \mathcal{T}_k(A)\| = \min\{\|A - B\| : \ B \in \mathbb{C}^{m\times n} \text{ has rank at most } k\},$$

for every unitarily invariant norm $\|\cdot\|$. In particular, the property holds for the Euclidean norm and

$$\|A - \mathcal{T}_k(A)\|_2 = \sigma_{k+1}.$$

1.2 THE NULLITY THEOREM

The Nullity Theorem directly relates the rank of a sub block in a matrix and a particular one in its inverse. It was discovered by Gustafson [50] for matrices over principal ideals and it has been rephrased for matrices over fields by Fiedler and Markham [43]. Barrett and Feinsilver also provided theorems of this kind [4, 5].

Theorem 1.2.1 (Theorem 1.33 in [97]). *Let $A \in \mathbb{C}^{m\times m}$ be a nonsingular matrix partitioned as*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

with $A_{11} \in \mathbb{C}^{p\times q}$. The inverse B of A is partitioned as

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

with $B_{11} \in \mathbb{C}^{q\times p}$. Then the dimensions of $\ker(A_{11})$ and $\ker(B_{11})$ are equal.

The next corollary is particularly interesting for the study of the off-diagonal blocks. Given α and β subsets of indeces we denote with $A(\alpha, \beta)$ the submatrix of A obtained selecting the rows corresponding to α and the columns corresponding to β.

Corollary 1.2.2 (Corollary 1.36 in [97]). *Suppose $A \in \mathbb{C}^{m\times m}$ is a nonsingular matrix and α, β are nonempty subsets of $M := \{1, \dots, m\}$ with $|\alpha| < m$ and $|\beta| < m$. Then*

$$rank(A^{-1}(\alpha, \beta)) = rank(A(M \setminus \beta, M \setminus \alpha)) + |\alpha| + |\beta| - m.$$

In particular

$$rank(A^{-1}(\alpha, M \setminus \alpha)) = rank(A(\alpha, M \setminus \alpha)).$$

This corollary states in fact that the rank of all blocks of the matrix just below and just above the diagonal will be maintained under inversion.

1.3 SHERMAN-MORRISON-WOODBURY FORMULA

A useful matrix identity is the celebrated *Sherman-Morrison-Woodbury formula*. The latter claims that the inverse of a rank-k correction of some matrix can be expressed as a rank-k correction of the inverse of the original matrix.

Lemma 1.3.1. *Let $A \in \mathbb{C}^{m\times m}$ and $C \in \mathbb{C}^{k\times k}$ be non singular matrices, $U \in \mathbb{C}^{m\times k}$ and $U \in \mathbb{C}^{k\times m}$. Then $A + UCV$ is non singular if and only if $C^{-1} + VA^{-1}U$ is non singular and*

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

1.4 LAURENT SERIES AND ANALYTIC FUNCTIONS

Denote by $\mathcal{W}$ the Wiener class formed by the functions $a(z) = \sum_{i=-\infty}^{+\infty} a_i z^i : \mathbb{T} \to \mathbb{C}$ such that $\sum_{i=-\infty}^{+\infty} |a_i| < +\infty$. It is well-known that $\mathcal{W}$ is a Banach algebra, that is, a vector space closed under multiplication, endowed with the norm $\|a\|_{\mathcal{W}} := \sum_{i\in\mathbb{Z}} |a_i|$ which makes the space complete and such that $\|ab\|_{\mathcal{W}} \leqslant \|a\|_{\mathcal{W}}\|b\|_{\mathcal{W}}$ for any $a(z), b(z) \in \mathcal{W}$. The coefficients a_i are called the Fourier coefficients of the function $a(z)$. We refer the reader to the first chapter of the book [28] for more details.

The regularity of a function defined by a Laurent series implies decay properties of its Fourier coefficients. A result which we are going to use several times is the following.

Theorem 1.4.1 (Theorem 4.4c in [58]). *Let $f(z) = \sum_{i=-\infty}^{+\infty} a_i z^i$ be analytic in the annulus $\mathbb{A}(r, R)$ with $r < 1 < R$. Then $\forall \rho_1 \in (1, R)$ and $\forall \rho_2 \in (r, 1)$ it holds*

$$|a_i| \leqslant \gamma_1 \rho_1^{-i}, \qquad |a_{-i}| \leqslant \gamma_2 \rho_2^{i}, \quad i = 0, 1, 2, \dots,$$

where $\gamma_i = \max_{|z|=\rho_i} |f(z)|$, i=1,2.

1.5 NON-NEGATIVE MATRICES

The result which constitutes the basis of the theory of non-negative matrices is the following.

Theorem 1.5.1 (Perron-Frobenius, Theorem 1.4 in [12])**.** *Let $A \in \mathbb{R}^{m\times m}$ be a non-negative matrix, then*

(*i*) *the spectral radius of A is also an eigenvalue and is called the* Perron value*,*

(*ii*) *A admits left and right non-negative eigenvectors associated with the Perron value.*

Moreover, if A is irreducible then

(*iii*) *the Perron value is a simple eigenvalue,*

(*iv*) *the left and right eigenvectors associated with the Perron value are positive. All other left and right eigenvectors have at least one strictly negative entry.*

(*v*) *If $B \in \mathbb{R}^{m\times m}$ verifies $B \geqslant A$ and $B \neq A$ then the spectral radius of B is greater then the one of A.*

Chapter 2

Motivation: Matrix Analytic Methods in Markov chains

In this chapter we briefly describe the origin of some of the questions that inspired this work. The exposure of the topic is concise and aims to emphasize the linear algebra aspects that come into play. For a complete picture which takes into account probabilistic interpretations we refer to the books [81, 71, 17, 56].

2.1 Discrete time Markov chains with discrete states

Stochastic processes are probabilistic tools used to model systems that evolve in time, e.g., queues, fluid flows, populations, prices of assets and many others. More precisely, a stochastic process is a family $\{X_t : t \in T\}$ of random variables indexed by a totally ordered set T and having values in a common set E. T and E are called time space and state space, respectively. The subclass of stochastic processes that we are going to consider is introduced in the following definition.

Definition 2.1.1. *The family of random variables $\{X_t : t \in T\}$ with state space E is said a homogeneous discrete time Markov chains with discrete states if*

- $T = \mathbb{N}$,
- *E is a denumerable set,*
- $\mathbb{P}(X_{n+1} = j \mid X_n, X_{n-1} \ldots, X_0) = \mathbb{P}(X_{n+1} = j \mid X_n)$ $\forall n \in \mathbb{N}$ *and* $\forall j \in E$,
- $\mathbb{P}(X_{n+1} = j \mid X_n = i) = \mathbb{P}(X_1 = j \mid X_0 = i)$ $\forall n \in \mathbb{N}$ *and* $\forall i, j \in E$.

The third request is the so-called Markov property and roughly says that in every moment the future state of the system depends only on the conditions of the present. The last condition ensures that the dynamic of the transitions remains unchanged over

the time. In particular, to every such process we can associate the matrix $P \in \mathbb{R}^{|E|\times|E|}$ defined by $p_{ij} := \mathbb{P}(X_{n+1} = j \mid X_n = i)$. P is called the *transition probability matrix* and it is non negative and row stochastic.

Notice that, since E is a denumerable set we can embed it into $\mathbb{Z}^m$ for a certain $m \in \mathbb{Z}^+$. This means that we are considering all the processes that can be modeled as random walks on the integer coordinates of a certain region in $\mathbb{Z}^m$.

2.2 STATIONARY DISTRIBUTION OF POSITIVE RECURRENT PROCESSES

A problem of interest in these settings is to study the behavior of the process asymptotically, i.e., try to forecast in which state the system will be as the time tends to infinity. Obviously —because of the random component— the prevision has to be characterized by a certain probability and one expects that this depends also on the starting state of the system. Mathematically, we are interested in finding

$$\lim_{n\to\infty} \mathbb{P}(X_n = j \mid X_0 = i) \quad \forall i, j \in E.$$

Assuming some additional hypotheses, the latter quantities do not depend on the starting state and form a probability measure at all.

Theorem 2.2.1 (Part of Theorem 1.17 in[17])**.** *Let $\{X_t : t \in \mathbb{N}\}$ be a homogeneous discrete time Markov chain with discrete states and suppose that*

- *the transition probability matrix P is irreducible and aperiodic,*
- *every state is positive recurrent, i.e., the probability of return to the state is* 1 *and the expected number of visits to it is finite.*

Then $\exists \pi \in \mathbb{R}^{|E|}$ such that $\pi > 0$, $\|\pi\|_1 = 1$ and

$$\lim_{n\to\infty} \mathbb{P}(X_n = j \mid X_0 = i) = \pi_j$$

for all j, independently on i. Moreover π verifies

$$\pi^t P = \pi^t.$$

The vector π is usually called the *stationary distribution* or the *steady state vector* of the process.

2.3 MATRIX GEOMETRIC PROPERTY FOR QUASI-BIRTH-DEATH PROCESSES

Here, we assume that the discrete state space E is two dimensional. As case study, consider a queue described by the number of customers X_n —called the *level*— and another feature φ_n called the *phase*. Suppose that at each time step the number of

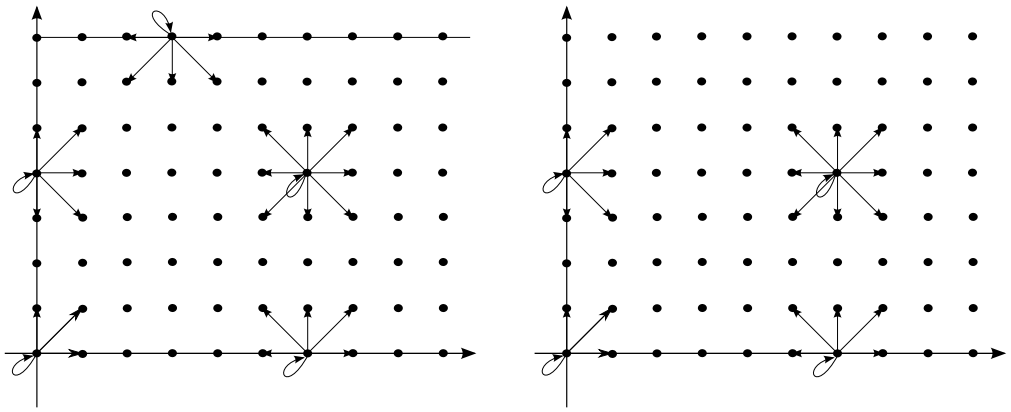

Figure 2.1.: Possible transitions of a double QBD processes on $\mathbb{N} \times S$; on the left the case of finite phase space ($m < \infty$), on the right the case of infinite phase space ($m = \infty$)

customers can either remain constant or increase/decrease by a unit, avoiding negative values. Finally, assume that the probability of these transitions are independent on the level, unless when the queue is empty where we have some boundary conditions. Instead, they depend on some external factors which vary in time, e.g., congestion of the network. These factors are modeled with the phase and we label its possible values with the set $S := \{0, \dots, m-1\}$, where m could be finite or infinite.

Under these assumptions we get a Markov chain with state space $\mathbb{N} \times S$. Using the lexicographic-order on the states we get a transition probability matrix of the form

$$P = \begin{bmatrix} \widetilde{P}_0 & \widetilde{P}_1 & & & \\ P_{-1} & P_0 & P_1 & & \\ & P_{-1} & P_0 & P_1 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}, \qquad \widetilde{P}_0, \widetilde{P}_1, P_{-1}, P_0, P_1 \in \mathbb{R}^{m \times m}, \tag{2.1}$$

where

$$\begin{aligned} (A_k)_{ij} &= P(\alpha_n = k,\ \varphi_{n+1} = j \mid X_n > 0, \varphi_n = i), \qquad k = -1, 0, 1, \\ \left(\widetilde{A}_k\right)_{ij} &= P(\alpha_n = k,\ \varphi_{n+1} = j \mid X_n = 0, \varphi_n = i), \qquad k = 0, 1 \end{aligned}$$

and α_n represents the variation of the customers at time n. The matrices $\widetilde{P}_0 + \widetilde{P}_1$ and $P_{-1} + P_0 + P_1$ are row stochastic. The processes with this block tridiagonal Toeplitz-like transitions are called *level independent Quasi-Birth-Death* (QBD) and are the most popular among the phase-type queue models.

Now, consider the stationary distribution $\pi = [\pi_0, \pi_1, \dots]$, $\pi_i \in \mathbb{R}^m$ $i \geqslant 0$, block partitioned according to (2.1). The vector π enjoy the celebrated *matrix geometric property*, stated in the following theorem.

Theorem 2.3.1 (Theorem 5.18 in [17], [71] for the case $m = +\infty$)**.** *If the QBD process described by* (2.1) *verifies the hypotheses of Theorem 2.2.1 then its stationary distribution* π *verifies*

$$\begin{aligned} &\pi_n^t = \pi_0 R^n \qquad \forall n \geqslant 0, \\ &\pi_0^t(\widetilde{P}_0 + \widetilde{P}_1 G) = \pi_0^t, \\ &\|\pi_0^t (I - R)^{-1}\|_1 = 1, \end{aligned}$$

where the matrices R *and* G *are the minimal non negative solutions of*

$$X = X^2 P_{-1} + X P_0 + P_1, \quad X \in \mathbb{R}^{m \times m}$$

and

$$X = P_{-1} + P_0 X + P_1 X^2, \quad X \in \mathbb{R}^{m \times m},$$

respectively. Moreover, it holds

$$R = P_1 (I - U)^{-1},$$

with $U = P_0 + P_1 G$*.*

Therefore, the computational strategy used to retrieve the stationary distribution is the following

(i) compute G solving $-P_{-1} + (I - P_0)X - P_1 X^2 = 0$,

(ii) compute $U = P_0 + P_1 G$ and $R = P_1 (I - U)^{-1}$,

(iii) retrieve π_0 which solves $\pi_0^t(\widetilde{P}_0 + \widetilde{P}_1 G) = \pi_0^t$ and verifies $\|\pi_0^t (I - R)^{-1}\|_1 = 1$,

(iv) compute as many blocks π_n we want by means of the relation $\pi_n = \pi_0 R^n$.

To ease the notation, we relabel the coefficients of the equation in (i) as

$$A_{-1} + A_0 X + A_1 X^2 = 0 \tag{2.2}$$

where $A_{-1} = -P_1$, $A_0 = I - P_0$ and $A_1 = -P_1$.

2.4 NUMERICAL LINEAR ALGEBRA ISSUES

The problem of practically computing the stationary distribution of a positive recurrent QBD process is well understood when the phase space is finite ($m < \infty$). The crucial step is solving the quadratic matrix equation (2.2) and this can be done successfully by means of the *cyclic reduction* algorithm that we describe in Chapter 6. This method is based on an iterative scheme which performs basic arithmetic operations on the blocks A_i. Without any further assumptions its cost is cubic in their dimension m. On the other

hand, there are several models from the applications in which the blocks A_i exhibit special structures. Very often the A_is enjoy a band structure, e.g., in the *Double QBD process* [79] they are tridiagonal, see also Figure 2.1. In spite that, we lose almost immediately the band structure when executing the CR, because of the inversion operations required. Empirically, what seems to be preserved is the quasiseparable property, that we introduce in Chapter 3. Thus, one of our goals is to design a version of CR which exploits the rank structure of the model and can be implemented at a substantially lower cost.

When the phase state is infinite ($m = +\infty$) the situation becomes more challenging. Explicit solutions are known only for very special cases. Theoretical approaches focus on estimating the tail asymptotics of π [78, 79, 65]. Moreover, due to the infinite size of the blocks none of the computational steps (i)-(iv) is obvious. Calculations can be carried on assuming some strong structures on the blocks A_i, e.g., [91]. Truncation methods can be applied, but there is no guarantee that the solutions obtained in this way approximate the original solution, e.g., [72, 6, 70]. For these reasons, another target of this work is to provide a numerical framework able to handle blocks A_i of infinite size, avoiding truncation.

Chapter 3

Quasiseparable matrices

Detecting rank structures in a mathematical model often means to dramatically reduce the memory and time consumption of the resolution procedures. In particular, the rank structures we are going to analyze are flexible with respect to matrix operation. This fact is a consequence of the Nullity Theorem for the inverse operation and it is almost trivial for matrix addition and multiplication.

In this chapter we introduce the set of quasiseparable matrices and some of its slight variations. Moreover, we address the problem of taking advantage of the structure by means of a suitable representation and a fast arithmetic.

3.1 DEFINITION AND PROPERTIES

Because of the extensive and parallel work on the rank structures, some concepts have not been uniformly introduced and so there may be ambiguities in the use of terms as *quasiseparable*, *semiseparable* and their generalizations. To overcome this drawback, we specify here our notations, which take as references [97, 98].

We start by defining the rank under or above a certain diagonal.

Definition 3.1.1. *Let $A \in \mathbb{C}^{n\times n}$, $p, q \in \mathbb{Z}$ and $r_l, r_u \in \mathbb{N}$, we say that $r_{lw}^{(p)}(A) = r_l$ if*

$$\max_{i\in\mathcal{I}_1} rank(A_{i:n,1:i+p}) = r_l$$

with

$$\mathcal{I}_1 = \{\max(1, 1-p), \ldots, \min(n-p, n)\}.$$

Analogously, we say that $r_{up}^{(q)}(A) = r_u$ if

$$\max_{i\in\mathcal{I}_2} rank(A_{1:i,i+q:n}) = r_u.$$

with

$$\mathcal{I}_2 = \{\max(1, 1-q), \ldots, \min(n-q, n)\}.$$

We can now introduce the central notion of this work.

Definition 3.1.2. *Let $A \in \mathbb{C}^{n\times n}$, we say that A has* quasiseparable rank (k_l, k_u) *if*

$$r_{lw}^{(-1)}(A) \leqslant k_l, \qquad\qquad r_{up}^{(1)}(A) \leqslant k_u.$$

We write $q_{rank}(A) = (k_l, k_u)$. In the case $k_l = k_u = k$ we just write $q_{rank}(A) = k$. In such cases we also say that the matrix is (k_l, k_u)-quasiseparable and k-quasiseparable, respectively.

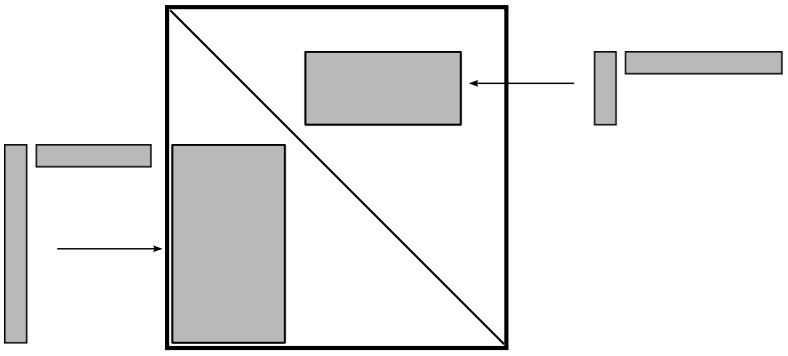

Figure 3.1.: Graphic description of the quasiseparable structure

The quasiseparable rank turns out to be invariant under inversion and sub-additive with respect to the matrix sum and product.

Theorem 3.1.3. *Let A be a (k_l, k_u)-quasiseparable matrix and B a (j_l, j_u)-quasiseparable matrix.*

(i) If A is invertible then also A^{-1} is a (k_l, k_u)-quasiseparable matrix.

(ii) $A + B$ is a $(k_l + j_l, k_u + j_u)$-quasiseparable matrix.

(iii) $A \cdot B$ is a $(k_l + j_l, k_u + j_u)$-quasiseparable matrix.

Proof. Propery (i) is a consequence of the Nullity Theorem (in particular Corollary 1.2.2). Properties (ii) and (iii) follow from the direct computation of the generic off-diagonal block and the sub-additivity of the rank. □

Definition 3.1.4. *A square matrix $A = (a_{ij}) \in \mathbb{C}^{n\times n}$ is called* (p, q)-band matrix *if*

$$i - j > p \quad \Rightarrow \quad a_{ij} = 0 \quad \textit{and} \quad j - i > q \quad \Rightarrow \quad a_{ij} = 0.$$

We indicate with $\mathbb{B}_p^q$ the set of all (p, q)-banded matrices.

We say that A is a strict band matrix *if all the elements on the extreme diagonals are different from zero.*

It is easy to verify that a (p, q)-band matrix is also a (p, q)-quasiseparable matrix. In fact, every submatrix that we can select into its strctly lower (upper) triangular part has at most p (q) rows and columns different from zero.

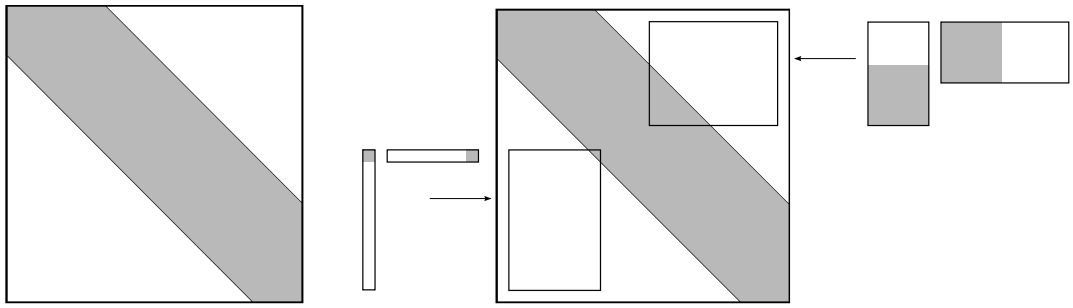

Figure 3.2.: Graphic description of the band structure; in grey, the non zero entries

3.2 SOME SUBSETS OF QUASISEPARABLE MATRICES

Here, we introduce some special subclasses —of the set of quasiseparable matrices— which we are going to use in Section 6.4. Before that and according to MATLAB notation, we define the *tril* and *triu* operators.

Definition 3.2.1. *Let $A \in \mathbb{C}^{n\times n}$, we indicate with $tril(A,p)$ and $triu(A,p)$ the matrices*

$$tril(A,p)_{ij} = \begin{cases} a_{ij} & \text{if } i \geqslant j-p \\ 0 & \text{elsewhere} \end{cases}, \qquad triu(A,p)_{ij} = \begin{cases} a_{ij} & \text{if } i \leqslant j-p \\ 0 & \text{elsewhere} \end{cases}.$$

When $p = 0$ we just write $tril(A)$ and $triu(A)$.

The first generalization —that we consider— is the inclusion of the main diagonal into the rank structure.

Definition 3.2.2. *Let $A \in \mathbb{C}^{n\times n}$, we say that A has* semiseparable rank (k_l, k_u) *if*

$$r_{lw}^{(0)}(A) \leqslant k_l, \qquad r_{up}^{(0)}(A) \leqslant k_u.$$

We write $s_{rank}(A) = (k_l, k_u)$. In the case $k_l = k_u = k$ we just write $s_{rank}(A) = k$. In such cases we also say that the matrix is (k_l, k_u)-semiseparable and k-semiseparable, respectively.

Two possible extensions of (k_l, k_u)-semiseparable matrix —with additional representability properties— are the following.

Definition 3.2.3. *Let $r_l, r_u \in \mathbb{N}$, a matrix $A \in \mathbb{C}^{n\times n}$ is called* (r_l, r_u)-generator representable semiseparable *if $\exists U, V \in \mathbb{C}^{n\times r_l}$ and $W, Z \in \mathbb{C}^{n\times r_u}$ such that*

$$tril(A, r_l - 1) = tril(UV^*) \quad \text{and} \quad triu(A, 1 - r_u) = triu(WZ^*).$$

We call the quadruple (U, V, W, Z) the generator.

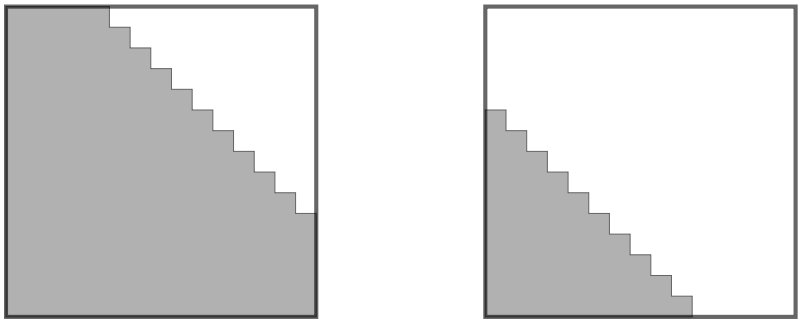

Figure 3.3.: In grey, the parts selected by $\mathrm{tril}(A,k)$ and $\mathrm{tril}(A,-k)$ for $k>0$

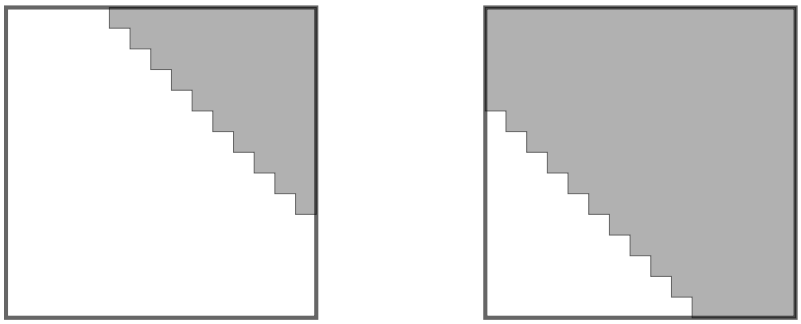

Figure 3.4.: In grey, the parts selected by $\mathrm{triu}(A,k)$ and $\mathrm{triu}(A,-k)$ for $k>0$

Definition 3.2.4. *Let $r_l, r_u \in \mathbb{N}$, a matrix $A \in \mathbb{C}^{n\times n}$ is called* extended (r_l,r_u)-generator representable semiseparable *if $\exists U,V \in \mathbb{C}^{n\times r_l}$ and $W,Z \in \mathbb{C}^{n\times r_u}$ such that*

$$tril(A) = tril(UV^*) \quad and \quad triu(A) = triu(WZ^*).$$

We call the quadruple (U,V,W,Z) the generator.

We indicate with $\mathbb{G}_{r_l}^{r_u}$ the set of all extended (r_l,r_u)-generator representable semiseparable matrices.

The first class is the set of matrices having $r_{lw}^{(r_l-1)} \leqslant r_l$, $r_{up}^{(1-r_u)} \leqslant r_u$ representable through generators. The second one requires the same properties limited to the lower and upper triangular part of the matrix. Obviously, for equal parameters, the second class contains the first one and they coincide for $r_l = r_u = 1$. As a further consequence of the Nullity Theorem, the semiseparable matrices are strictly related to the band matrices. More precisely, it holds the following.

Theorem 3.2.5 ([97] Section 8.3)**.** *The inverse of an invertible strict (p,q)-band matrix is an invertible (p,q)-generator representable semiseparable and vice versa.*

If we do not assume strictness then we lose the representability property, but the rank structure still holds.

3.3 REPRESENTING A QUASISEPARABLE MATRIX

In order to take advantage of the quasiseparable structure we need a representation that enables us to perform the storage and the matrix operations cheaply. The typical request is a linear or linear-polylogarithmic complexity with respect to the size of the matrix. There are various representations that meet these demands. The one we consider belongs to the family of hierarchical representations ($\mathcal{H}$-matrices) pioneered by Hackbusch [53, 54] for handling matrices coming from the discretization of integral and partial differential equations. This is a class of recursive block representation with structured sub-matrices that allows the treatment of a number of data-sparse patterns. We focus on a particular member of this family —sometimes called hierarchical off-diagonal low-rank representation (HODLR)— which has a simple formulation and guarantees a significant speed up of the algorithms where is employed.

We also mention the works —on other representations for quasiseparable matrices— done by Eidelman, Gohberg and Haimovici [38, 40] and by Van Barel, Vandebril and Mastronardi [97, 98, 96].

3.3.1 *HODLR representation*

We introduce the HODLR representation in an informal and constructive way. For a detailed treatment of hierarchical formats ($\mathcal{H}$-matrices) see [26] and [54].

Let $A \in \mathbb{C}^{m\times m}$ be a k-quasiseparable matrix and consider the partitioning

$$A = \begin{bmatrix} A_{11} & A_{22} \\ A_{21} & A_{22} \end{bmatrix},$$

where $A_{11} \in \mathbb{C}^{m_1\times m_1}$, $A_{22} \in \mathbb{C}^{m_2\times m_2}$, with $m_1 := \lfloor \frac{m}{2} \rfloor$ and $m_2 := \lceil \frac{m}{2} \rceil$. Observe that the antidiagonal blocks A_{12} and A_{21} do not involve any element of the main diagonal of A, hence we can represent them with outer products of length k. Moreover, the diagonal blocks A_{11} and A_{22} are square matrices which are again k-quasisepable. Therefore it is possible to re-apply these procedure recursively. We stop when the diagonal blocks reach a minimal dimension m_{min}, and we store them as full matrices. The process is described graphically in Figure 3.5. Informally, we call a matrix which admits such partitioning, a HODLR matrix of rank k.

If m_{min} and k are negligible with respect to m then the storage cost of each sub-matrix is $O(m)$. Since the levels of the recursion are $O(\log(m))$, this yields a linear-polylogarithmic memory consumption with respect to the size of the matrix.

3.4 HODLR-MATRIX ARITHMETIC

HODLR representation acts on a matrix by compressing many of its sub-blocks. Therefore, it is natural to expect that the arithmetic operations are performed in a block-recursive

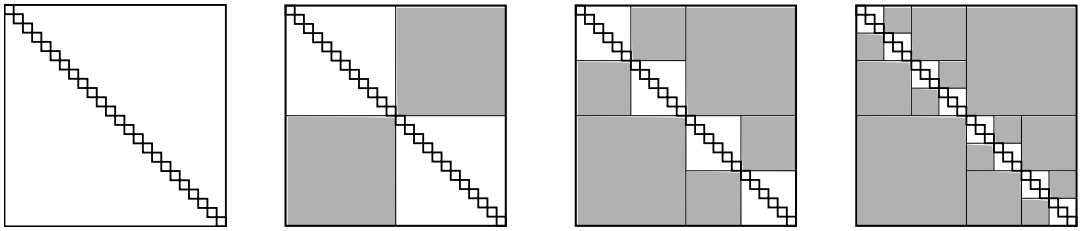

Figure 3.5.: The behavior of the block partitioning in the HODLR-matrix representation. The blocks filled with grey are represented as low-rank outer products, the diagonal blocks in the last step are stored as full matrices.

fashion. The basic steps of these procedures require arithmetic operations between low-rank matrices and/or $m_{min} \times m_{min}$-matrices.

In order to simplify the analysis of the procedures we assume $m = 2^p \cdot m_{min}$ and the rank of the off-diagonal blocks uniformly bounded by k. The first condition ensures to deal —at each level of the recursion— with square matrices of equal dimension. The computational cost of an operation involving matrices of dimesion m is indicated with $\mathcal{C}_*(m)$ where in place of $*$ a description of the operation is provided. In such description the symbols H, R and v are used for indicating HODLR matrices, low-rank matrices and vectors, respectively.

3.4.1 *Low-rank matrix arithmetic*

Here, we recall the well-known strategies used to perform basic operations with low-rank matrices at linear cost. For $A, B \in \mathbb{C}^{m \times m}$ of rank k we assume the outer product factorizations $A = U_A V_A^*$ and $B = U_B V_B^*$, $U_A, U_B, V_A, V_B \in \mathbb{C}^{m \times k}$.

- **Matrix-vector multiplication:** $Ax = U_A \cdot (V_A^* x)$, we perform two matrix vector multiplications where the matrix has either few rows or few columns.
- **Matrix product:** $AB = U_A \cdot (V_B U_B^* V_A)^* = (U_A V_A^* U_B) \cdot V_B^*$, according to the minimum between the rank of A and B, we have to compute only one of the two factor of the representation of the result. This can be done by performing the multiplication of three matrices each of them having either few columns or rows.
- **Matrix sum:** $A + B = [U_A U_B][V_A V_B]^*$, without performing operations we get a representation of length $k_A + k_B$ where these two quantities indicate the rank of A and B respectively.

It may happen that the length of the representation overestimates the real rank of the matrix. For example when we compute $A + B$, the procedure returns a representation with length $k_A + k_B$ that could be greater than the exact rank of $A + B$. Furthermore,

considering low-rank approximations —up to a certain threshold ϵ— can provide significant savings in the length of the representations and so in the computational cost. For these reasons we describe an efficient way to compress a low-rank representation.

- **Matrix compression:** $A = U_A V_A^* \longrightarrow_\epsilon \widetilde{U}_A \widetilde{V}_A^*$. Compute $U_A = Q_U R_U$ and $V_A = Q_V R_V$ the QR decompositions of U_A and V_A, respectively. Then, retrieve the SVD decomposition of $R_U R_V^* = W \Sigma Z^*$. Truncate (W, Σ, Z) according to ϵ, getting $(\widetilde{W}, \widetilde{\Sigma}, \widetilde{Z})$. Finally, compute $\widetilde{U}_A = Q_U \cdot \widetilde{W} \cdot \widetilde{\Sigma}$ and $\widetilde{V}_A = Q_V \cdot Z$.

The above procedure has a cost linear in m. In fact —apart from products of low-rank matrices— it only requires the QR factorization of two $m \times k$ matrices, that costs $O(mk^2)$ and the SVD of a $k \times k$ matrix which needs $O(k^3)$ operations.

In the procedures regarding the HODLR arithmetic, whenever we speak about products and sums of low-rank matrices it is always meant that —together with the operation— a final compression step is performed up to a certain threshold ϵ. This means that a certain off-diagonal block of the result is a low-rank approximation of the corresponding block of the exact result.

Moreover, given a matrix A and an accuracy parameter ϵ we indicate with $\mathcal{H}_\epsilon(A)$ its HODLR approximation, where the off-diagonal blocks are compressed with respect to the treshold ϵ.

3.4.2 *Matrix-vector multiplication*

With a little abuse of notation, we symbolised the block structure of the data —that come into play— in this manner

$$\begin{bmatrix} \mathcal{H}_{\frac{m}{2}} & \mathcal{R}_{\frac{m}{2}} \\ \mathcal{R}_{\frac{m}{2}} & \mathcal{H}_{\frac{m}{2}} \end{bmatrix} \begin{bmatrix} v_{\frac{m}{2}} \\ v_{\frac{m}{2}} \end{bmatrix} = \begin{bmatrix} \mathcal{H}_{\frac{m}{2}} \cdot v_{\frac{m}{2}} + \mathcal{R}_{\frac{m}{2}} \cdot v_{\frac{m}{2}} \\ \mathcal{R}_{\frac{m}{2}} \cdot v_{\frac{m}{2}} + \mathcal{H}_{\frac{m}{2}} \cdot v_{\frac{m}{2}} \end{bmatrix}.$$

Intuitively, the symbols $\mathcal{H}_{\frac{m}{2}}, \mathcal{R}_{\frac{m}{2}}$ and $v_{\frac{m}{2}}$ stand for $\frac{m}{2} \times \frac{m}{2}$-HODLR matrix, $\frac{m}{2} \times \frac{m}{2}$-low-rank matrix and vector of dimension $\frac{m}{2}$, respectively.

The products $\mathcal{R}_{\frac{m}{2}} \cdot v_{\frac{m}{2}}$ can be computed with the low-rank matrix arithmetic, while recursion is applied for retrieving $\mathcal{H}_{\frac{m}{2}} \cdot v_{\frac{m}{2}}$. Finally, two sums of vectors have to be computed. This means

$$\mathcal{C}_{Hv}(m) = 2\mathcal{C}_{H \cdot v}(\frac{m}{2}) + 2\mathcal{C}_{R \cdot v}(\frac{m}{2}) + 2\mathcal{C}_{v+v}(\frac{m}{2}).$$

3.4.3 *Matrix addition*

Using the same notation as before, we represent the sum of two HODLR-matrices as follows:

$$\begin{bmatrix} \mathcal{H}_{\frac{m}{2}} & \mathcal{R}_{\frac{m}{2}} \\ \mathcal{R}_{\frac{m}{2}} & \mathcal{H}_{\frac{m}{2}} \end{bmatrix} + \begin{bmatrix} \mathcal{H}_{\frac{m}{2}} & \mathcal{R}_{\frac{m}{2}} \\ \mathcal{R}_{\frac{m}{2}} & \mathcal{H}_{\frac{m}{2}} \end{bmatrix} = \begin{bmatrix} \mathcal{H}_{\frac{m}{2}} + \mathcal{H}_{\frac{m}{2}} & \mathcal{R}_{\frac{m}{2}} + \mathcal{R}_{\frac{m}{2}} \\ \mathcal{R}_{\frac{m}{2}} + \mathcal{R}_{\frac{m}{2}} & \mathcal{H}_{\frac{m}{2}} + \mathcal{H}_{\frac{m}{2}} \end{bmatrix}.$$

Again, the antidiagonal blocks can be computed with the low-rank arithmetic while recursion is applied for computing the diagonal blocks. This gives

$$\mathcal{C}_{H+H}(m) = 2\mathcal{C}_{H+H}(\frac{m}{2}) + 2\mathcal{C}_{R+R}(\frac{m}{2}).$$

Analogously, one can compute the sum of a HODLR matrix and a low-rank matrix, representing the outcome as an HODLR matrix:

$$\begin{bmatrix} \mathcal{H}_{\frac{m}{2}} & \mathcal{R}_{\frac{m}{2}} \\ \mathcal{R}_{\frac{m}{2}} & \mathcal{H}_{\frac{m}{2}} \end{bmatrix} + \begin{bmatrix} \mathcal{R}_{\frac{m}{2}} & \mathcal{R}_{\frac{m}{2}} \\ \mathcal{R}_{\frac{m}{2}} & \mathcal{R}_{\frac{m}{2}} \end{bmatrix} = \begin{bmatrix} \mathcal{H}_{\frac{m}{2}} + \mathcal{R}_{\frac{m}{2}} & \mathcal{R}_{\frac{m}{2}} + \mathcal{R}_{\frac{m}{2}} \\ \mathcal{R}_{\frac{m}{2}} + \mathcal{R}_{\frac{m}{2}} & \mathcal{H}_{\frac{m}{2}} + \mathcal{R}_{\frac{m}{2}} \end{bmatrix}.$$

It holds

$$\mathcal{C}_{H+R}(m) = 2\mathcal{C}_{H+R}(\frac{m}{2}) + 2\mathcal{C}_{R+R}(\frac{m}{2}).$$

3.4.4 *Matrix multiplication*

Before looking at the product between two HODLR matrices we point out that the product of a HODLR matrix with a low-rank matrix can be carried out with k matrix-vector multiplications involving a HODLR matrix. That is

$$\mathcal{C}_{H\cdot R}(m) = k\mathcal{C}_{H\cdot v}(m)$$

and it is important to underline that the result is represented with the same low-rank format of the right factor. The same relation holds for $\mathcal{C}_{R\cdot H}(m)$.

With these tools we can deal with the multiplication of two HODLR matrices

$$\begin{bmatrix} \mathcal{H}_{\frac{m}{2}} & \mathcal{R}_{\frac{m}{2}} \\ \mathcal{R}_{\frac{m}{2}} & \mathcal{H}_{\frac{m}{2}} \end{bmatrix} \cdot \begin{bmatrix} \mathcal{H}_{\frac{m}{2}} & \mathcal{R}_{\frac{m}{2}} \\ \mathcal{R}_{\frac{m}{2}} & \mathcal{H}_{\frac{m}{2}} \end{bmatrix} = \begin{bmatrix} \mathcal{H}_{\frac{m}{2}}\mathcal{H}_{\frac{m}{2}} + \mathcal{R}_{\frac{m}{2}}\mathcal{R}_{\frac{m}{2}} & \mathcal{H}_{\frac{m}{2}}\mathcal{R}_{\frac{m}{2}} + \mathcal{R}_{\frac{m}{2}}\mathcal{H}_{\frac{m}{2}} \\ \mathcal{R}_{\frac{m}{2}}\mathcal{H}_{\frac{m}{2}} + \mathcal{H}_{\frac{m}{2}}\mathcal{R}_{\frac{m}{2}} & \mathcal{R}_{\frac{m}{2}}\mathcal{R}_{\frac{m}{2}} + \mathcal{H}_{\frac{m}{2}}\mathcal{H}_{\frac{m}{2}} \end{bmatrix}.$$

Remembering that the operations $\mathcal{H}_{\frac{m}{2}}\mathcal{R}_{\frac{m}{2}}$ and $\mathcal{R}_{\frac{m}{2}}\mathcal{H}_{\frac{m}{2}}$ return low-rank matrices we note that also 4 sums —between HODLR and low-rank matrices— are involved. This yields

$$\mathcal{C}_{H\cdot H}(m) = 2\mathcal{C}_{H\cdot H}(\frac{m}{2}) + 2\mathcal{C}_{H\cdot R}(\frac{m}{2}) + 2\mathcal{C}_{R\cdot H}(\frac{m}{2}) + 2\mathcal{C}_{R\cdot R}(\frac{m}{2}) + 4\mathcal{C}_{H+R}.$$

3.4.5 *Matrix inversion*

We start again to label the sub-blocks of the HODLR partitioning because the blocks of the outcome are not independently calculated.

In order to carry on the (approximate) computation of the inverse we need to assume the invertibility of one of the two diagonal blocks in addition to the invertibility of the whole matrix. For example if the upper left submatrix is invertible we use the block inversion formula

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{bmatrix},$$

where $S := A_{22} - A_{21}A_{11}^{-1}A_{12}$. Computing the quantities in this order

(1) A_{11}^{-1}
(2) $A_{21}A_{11}^{-1}$
(3) $A_{11}^{-1}A_{12}$
(4) $A_{21}A_{11}^{-1}A_{12}$
(5) $A_{22} - A_{21}A_{11}^{-1}A_{12}$
(6) S^{-1}
(7) $-S^{-1}A_{21}A_{11}^{-1}$
(8) $-A_{11}^{-1}A_{12}S^{-1}$
(9) $A_{11}^{-1}A_{12}S^{-1}A_{21}$
(10) $A_{11}^{-1} + A_{11}^{-1}A_{12}S^{-1}A_{21}$

we get

$$\mathcal{C}_{inv(H)}(m) = 2\mathcal{C}_{inv(H)}(\frac{m}{2}) + 2\mathcal{C}_{H\cdot R}(\frac{m}{2}) + 2\mathcal{C}_{R\cdot H}(\frac{m}{2}) + 2\mathcal{C}_{H+R}(\frac{m}{2}) + 2\mathcal{C}_{R\cdot R}(\frac{m}{2}).$$

3.4.6 *Triangular systems*

Here, we address the problem of implementing the forward and backward substitution for HODLR matrices. That is, we consider linear systems whose coefficient matrix is both HODLR and triangular. In the case of forward substitution we want to solve

$$\begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

Performing these operations:

- solve recursively $A_{11}x_1 = b_1$,
- set $z = b_2 - A_{21}x_1$,
- solve recursively $A_{22}x_2 = z$,

we retrieve

$$\mathcal{C}_{fw(H)}(m) = 2\mathcal{C}_{fw(H)}(\frac{m}{2}) + \mathcal{C}_{H\cdot v}(\frac{m}{2}) + \mathcal{C}_{v+v}(\frac{m}{2}).$$

Analogous computations for the backward substitution provide

$$\mathcal{C}_{bw(H)}(m) = 2\mathcal{C}_{bw(H)}(\frac{m}{2}) + \mathcal{C}_{H\cdot v}(\frac{m}{2}) + \mathcal{C}_{v+v}(\frac{m}{2}).$$

3.4.7 *LU decomposition*

The tools developed in the previous section pave the way to compute the LU decomposition of a HODLR matrix:

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}.$$

If the conditions for the existence are satisfied we can proceed as follows

- compute recursively the decomposition $A_{11} = L_{11}U_{11}$,

Operation	Computational complexity
Storage	$O(km\log(m))$
Matrix-vector multiplication	$O(km\log(m))$
Matrix-matrix addition	$O(k^2m\log(m))$
Matrix-matrix multiplication	$O(k^2m\log(m)^2)$
Matrix-inversion	$O(k^2m\log(m)^2)$
Solve linear system	$O(k^2m\log(m)^2)$

Table 3.1.: Computational complexity of the HODLR-matrix arithmetic

- compute $U_{12} = L_{11}^{-1}A_{12}$ and $U_{21} = A_{21}U_{11}^{-1}$,
- compute recursively the decomposition $A_{22} - L_{21}U_{12} = L_{22}U_{22}$.

Observe that for computing $U_{12} = L_{11}^{-1}A_{12}$ we exploit the fact that both A_{12} and U_{12} have rank k. To be precise, we just need to compute the left factor of U_{12} solving k triangular systems with the columns of the left factor of A_{12} as right-hand sides. Analogously we compute $U_{21} = A_{21}U_{11}^{-1}$ using backward substitution. This yields

$$\mathcal{C}_{LU(H)}(m) = 2\mathcal{C}_{LU(H)}(\frac{m}{2}) + k\mathcal{C}_{fw(H)}(\frac{m}{2}) + k\mathcal{C}_{bw(H)}(\frac{m}{2}) + C_{R\cdot R}(\frac{m}{2}) + C_{H+R}(\frac{m}{2}).$$

3.4.8 *Complexity estimates*

The recursive relations —we derived in the previous sections— can be used together with the Master theorem [35][Section 4.3] for estimating the computational complexity of the operations in the HODLR-matrix arithmetic. The results of this analysis are resumed in Table 3.1 and show the linear-polylogarithmic complexity of the matrix operations, with respect to the size. The operation "Solve linear system" comprises to compute the LU factorization of the coefficient matrix and to solve the two triangular linear systems. For a more detailed treatment of these estimates see [54].

Remark 3.4.1. *We want to highlight that the matrix-vector product is the only operation which is computed exactly. The others are approximations affected by the accuracy at which we filter the singular values in the compression step. In fact, an important feature of HODLR representation is the ability to perform matrix operations and compute the representation of the results without knowing the quasiseparable rank a priori.*

3.5 FAST DECAY OF THE OFF-DIAGONAL SINGULAR VALUES

The use of HODLR-matrix arithmetic raises the question of whether a representation with small rank in the off-diagonal blocks is feasible.

The following result states that if the off-diagonal singular values of A decay *fast* then there exists a HODLR matrix —with low-rank— close to A. That is, for a relatively small k there is a perturbation δA of *small* norm such that $A + \delta A$ is a HODLR matrix with rank of the off-diagonal blocks at most k.

Theorem 3.5.1. *Let $f(l)$ be a function over the positive integers, and let $A \in \mathbb{C}^{m\times m}$ be a matrix such that $\sigma_l(B) \leqslant f(l)$ for every off-diagonal block B in A. Then, for any l there exists a perturbation matrix δA such that $A + \delta A$ is a HODLR matrix of rank l and $\|\delta A\|_2 \leqslant f(l) \cdot \log_2 m$.*

Proof. First, recall that if the nonzero singular values of a matrix B are $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_k$ then, for any $j < k$ we may write B as a matrix of rank j plus a perturbation δB such that $\|\delta B\|_2 = \sigma_{j+1}$. Now consider an HODLR partitioning of A with minimal blocks of dimension m_{min}. Notice that the depth of this recursive partition is $\sigma = \lceil \log_2(\frac{m}{m_{min}}) \rceil$. This way, for each off-diagonal block B of this partitioning and for any integer j, there exists a perturbation matrix that makes this block of rank j. The 2-norm of this perturbation is equal to $\sigma_{j+1}(B) \leqslant f(j+1)$. We may form the matrix δA which collects all these perturbations of each off-diagonal block of the above partitioning. This way, if $j = l - 1$, the off-diagonal blocks of $A + \delta A$ have rank at most l. We can now show that $\|\delta A\|_2 \leqslant f(l) \cdot \log_2 m$. We have

$$\delta A = \sum_{i=0}^{\sigma} \delta A_i, \qquad \sigma \leqslant \log_2 m,$$

where δA_i is the correction obtained by putting together all the blocks at level i of subdivision, that is,

$$\delta A_0 = \begin{bmatrix} 0 & \delta X_1^{(0)} \\ \delta X_2^{(0)} & 0 \end{bmatrix}, \qquad \delta A_1 = \left[\begin{array}{cc|cc} 0 & \delta X_1^{(1)} & & \\ \delta X_2^{(1)} & 0 & & \\ \hline & & 0 & \delta X_3^{(1)} \\ & & \delta X_4^{(1)} & 0 \end{array}\right], \qquad \ldots$$

Since the summands are just permutations of block diagonal matrices their 2-norm is the maximum of the 2-norms of the (block) diagonal entries, and this gives the desired bound. □

We want to highlight that the feasibility of the HODLR representation in an algorithm is directly connected with the off-diagonal singular values of the matrices involved. In particular, we do not need a preservation of the exact quasiseparable rank but is sufficient that only a few singular values stay above a reasonable threshold. That is why in the next chapters we focus on finding fast decaying bounds for the off-diagonal singular values.

Chapter 4

Studying the singular values

The singular values constitute an informative feature of a matrix and are the basis of a number of applications in statistics, functional analysis and linear algebra. In particular the l-th singular value can be interpreted as a measure of the maximum linear independence we can reach by choosing l columns (or rows) in the matrix. This can indicate the presence of good low-rank approximations. More formally, the Eckart-Young-Mirsky theorem ensures that if the l-th singular value is small then the matrix can be well-approximated by another one with rank $l-1$. This property makes the singular values a more flexible instrument than the exact rank in detecting data sparsity patterns. This could be crucial when we deal with perturbed data or other numerical effects. In particular, the leitmotif of this work is based on Theorem 3.5.1 and consists in approximating matrices having few significant off-diagonal singular values with HODLR matrices.

In this chapter we provide a framework for the analysis of the off-diagonal singular values after that a matrix computation is performed. The final aim is to improve the understanding of the quasiseparable preservation in an algorithm.

4.1 singular values of products

The modern theory of singular values inequalities has been developed starting from the efforts of Chang [34], Weyl [99], Horn [61], Ky Fan [42] and Polya [84] in the mid-twentieth century and it has been taken up more recently by Audenaert [2, 3], Tao [94], Drury [37], Zhan [100] and others. This topic is wide and beyond the purpose of this work, we refer to [62] for a complete overview and for the proof of the results we are going to report.

The singular values are highly connected with the eigenvalues through the following theorem due to Weyl

Theorem 4.1.1 (Theorem 3.3.2 in [62]). *Let $A \in \mathbf{C}^{m\times m}$ with singular values $\sigma_1 \geqslant \ldots \geqslant \sigma_m \geqslant 0$ and eigenvalues $|\lambda_1| \geqslant \ldots \geqslant |\lambda_m|$. Then $\forall l = 1, \ldots, m$*

$$\prod_{h=1}^{l} |\lambda_h| \leqslant \prod_{h=1}^{l} \sigma_h$$

and equality holds when $l = m$.

This result is crucial because it implies that several algebraic inequalities have their analogous singular value form. The following is what we can say about the singular values of products.

Theorem 4.1.2 (Horn, Theorem 3.3.4 in [62]). *Consider the product $A_1 \cdot \ldots \cdot A_p$, $p \geqslant 2$, where $A_1, \ldots, A_p$ are complex matrices with compatible dimensions. Then $\forall l = 1, \ldots, p$*

$$\prod_{h=1}^{l} \sigma_h(A_1 \cdot \ldots \cdot A_p) \leqslant \prod_{h=1}^{l} \sigma_h(A_1) \cdot \ldots \cdot \sigma_h(A_p).$$

If $A_h \in \mathbf{C}^{m\times m} \quad \forall h = 1, \ldots, p$ then it also holds

$$\prod_{h=1}^{m} \sigma_h(A_1 \cdot \ldots \cdot A_p) = \prod_{h=1}^{m} \sigma_h(A_1) \cdot \ldots \cdot \sigma_h(A_p).$$

Observing that the singular values are a non increasing sequence of non negative real numbers it follows that the l-th singular value is less than the geometric mean of the first l. Therefore, an obvious consequence of the Horn theorem is the following bound.

Corollary 4.1.3. *Consider the product $A_1 \cdot \ldots \cdot A_p$, $p \geqslant 2$, where $A_1, \ldots, A_p$ are complex matrices with compatible dimensions. Then $\forall l \in \mathbb{Z}^+$*

$$\sigma_l(A_1 \cdot \ldots \cdot A_p) \leqslant \left(\prod_{h=1}^{l} \sigma_h(A_1) \cdot \ldots \cdot \sigma_h(A_p) \right)^{\frac{1}{l}}.$$

We conclude this section with a technical lemma which can be used to relate individually a singular value of the product with the correspondent in one of the factors.

Lemma 4.1.4. *Consider two matrices $A \in \mathbf{C}^{m\times n}$, $B \in \mathbf{C}^{n\times n}$, such that B is invertible. Then it holds that*

$$\frac{\sigma_l(A)}{\|B^{-1}\|_2} \leqslant \sigma_l(AB) \leqslant \|B\|_2 \cdot \sigma_l(A), \qquad \frac{\sigma_l(A)}{\|B^{-1}\|_2} \leqslant \sigma_l(BA^*) \leqslant \|B\|_2 \cdot \sigma_l(A),$$
$$\frac{1}{\kappa(B)} \frac{\sigma_l(A)}{\sigma_1(A)} \leqslant \frac{\sigma_l(AB)}{\sigma_1(AB)} \leqslant \kappa(B) \frac{\sigma_l(A)}{\sigma_1(A)}, \qquad \frac{1}{\kappa(B)} \frac{\sigma_l(A)}{\sigma_1(A)} \leqslant \frac{\sigma_l(BA^*)}{\sigma_1(BA^*)} \leqslant \kappa(B) \frac{\sigma_l(A)}{\sigma_1(A)}.$$

Proof. We prove only the first statement since the other statements follow directly from it. Consider the reduced SVD decompositions of the two matrices given by

$$A = U_A \Sigma_A V_A^*, \quad B = U_B \Sigma_B V_B^*.$$

Recall that the singular values of a generic matrix $M \in \mathbb{C}^{m\times n}$ are the square roots of the eigenvalues of M^*M. In particular, by exploiting the Rayleigh quotient, we can write

$$\sigma_l(M)^2 = \max_{\substack{\dim(V)=l \\ V\subseteq\mathbb{R}^n}} \min_{x\in V} \frac{x^*M^*Mx}{x^*x}.$$

Now note that $AB = U_A\Sigma_A V_A^* U_B \Sigma_B V^*$ and since unitary matrices do not change the singular values we have that $\sigma_l(AB) = \sigma_l(\Sigma_A Q \Sigma_B)$ where $Q = V_A^* U_A$. Then, we can express $\sigma_l(AB)^2$ as

$$\sigma_l(AB)^2 = \max_{\substack{\dim(V)=l \\ V\subseteq\mathbb{R}^n}} \min_{x\in V} \frac{x^*\Sigma_B^* Q^* \Sigma_A^* \Sigma_A Q \Sigma_B x}{x^*x} = \max_{\substack{\dim(V)=l \\ V\subseteq\mathbb{R}^n}} \min_{x\in V} \frac{(\Sigma_B x)^* Q^* \Sigma_A^2 Q (\Sigma_B x)}{x^*x}.$$

By setting $y = \Sigma_B x$ and recalling that Σ_B must be invertible by hypothesis we have that

$$\sigma_l(AB)^2 = \max_{\substack{\dim(V)=l \\ V\subseteq\mathbb{R}^n}} \min_{y\in V} \frac{y^*Q^*\Sigma_A^2 Qy}{y^*y} \cdot \frac{y^*y}{x^*x}$$

so that by using the fact that Q is unitary and that $\frac{x^*x}{\|B^{-1}\|_2^2} \leqslant y^*y \leqslant \|B\|_2^2 x^*x$ we obtain

$$\frac{\sigma_l(A)^2}{\|B^{-1}\|_2^2} = \frac{\sigma_l(\Sigma_A)^2}{\|B^{-1}\|_2^2} \leqslant \sigma_l(AB)^2 \leqslant \|B\|_2^2 \cdot \sigma_l(\Sigma_A)^2 = \|B\|_2^2 \cdot \sigma_l(A)^2.$$

□

4.2 OFF-DIAGONAL SINGULAR VALUES OF THE INVERSE

The quasiseparable rank is maintained under inversion, but what can we say about the numerical rank of the off-diagonal blocks? It turns out that we are able to relate the off-diagonal singular values in the inverse with those of the starting matrix given some mild hypotheses. Notice that, even if the following result speaks about a maximal subdiagonal block —by means of transposition or restriction— analogous statements hold for any submatrix under or over the principal diagonal.

Lemma 4.2.1. *Let $\mathcal{A} \in \mathbb{C}^{m\times m}$ be an invertible matrix and let $1 \leqslant i \leqslant m-1$. Consider the block decomposition*

$$\mathcal{A} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \quad \mathcal{A}^{-1} = \begin{pmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{pmatrix}$$

where A and $\tilde{A}$ are $i \times i$ matrices. We have the following properties

1. *If D is invertible then*

$$\frac{1}{\|D\|_2\|S_D\|_2}\sigma_l(C) \leqslant \sigma_l(\tilde{C}) \leqslant \|D^{-1}\|_2 \cdot \|S_D^{-1}\|_2 \cdot \sigma_l(C),$$
$$\frac{1}{\kappa(D)\kappa(S_D)}\frac{\sigma_l(C)}{\sigma_1(C)} \leqslant \frac{\sigma_l(\tilde{C})}{\sigma_1(\tilde{C})} \leqslant \kappa(D)\cdot\kappa(S_D)\frac{\sigma_l(C)}{\sigma_1(\tilde{C})}$$

where $S_D = A - BD^{-1}C$ is the Schur complement of D.

2. *If A is invertible then*

$$\frac{1}{\|A\|_2\|S_A\|_2}\sigma_l(C) \leqslant \sigma_l(\tilde{C}) \leqslant \|A^{-1}\|_2 \cdot \|S_A^{-1}\|_2 \cdot \sigma_l(C),$$
$$\frac{1}{\kappa(A)\kappa(S_A)}\frac{\sigma_l(C)}{\sigma_1(C)} \leqslant \frac{\sigma_l(\tilde{C})}{\sigma_1(\tilde{C})} \leqslant \kappa(A)\cdot\kappa(S_A)\frac{\sigma_l(C)}{\sigma_1(\tilde{C})}$$

where $S_A = D - CA^{-1}B$ is the Schur complement of A.

Proof. Let us consider part 1. We prove the first equation, the second easily follows from the first one. If D is invertible we can consider the analytic inversion formula

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} S_D^{-1} & S_D^{-1}BD^{-1} \\ -D^{-1}CS_D^{-1} & D^{-1} + D^{-1}CS_D^{-1}BD^{-1} \end{pmatrix}$$

and in particular we have $\tilde{C} = -D^{-1}CS_D^{-1}$. Repeatedly applying Lemma 4.1.4 to $\tilde{C}$ gives us that

$$\frac{1}{\|D\|_2}\sigma_l(CS_D^{-1}) \leqslant \sigma_l(\tilde{C}) \leqslant \|D^{-1}\|_2 \cdot \sigma_l(CS_D^{-1})$$

and eventually that

$$\frac{1}{\|S_D\|_2}\sigma_l(C) \leqslant \sigma_l(CS_D^{-1}) \leqslant \|S_D^{-1}\|_2 \cdot \sigma_l(C).$$

The combination of these inequalities gives us the thesis.

For proving part 2 we can proceed in the same manner, relying on the inversion formula

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}BS_A^{-1}CA^{-1} & -A^{-1}BS_A^{-1} \\ -S_A^{-1}CA^{-1} & S_A^{-1} \end{pmatrix}.$$

□

4.3 SINGULAR VALUES OF SUMS AND SERIES

In this section we start to look for class of matrices whose singular values decay fast. More precisely, we try to retrieve upper bounds of the form

$$\sigma_l(A) \leqslant \gamma e^{-\alpha l}, \tag{4.1}$$

where α and γ are non negative constants. Given α and γ, we can bound uniformly the numerical rank of the matrices for which (4.1) holds.

The next result provides estimates of the kind (4.1) for infinite series of low-rank matrices with decaying Euclidean norms.

Lemma 4.3.1. *Let $A = \sum_{j=-\infty}^{+\infty} A_j$ and $A^+ = \sum_{j=0}^{+\infty} A_j$ with $A_j \in \mathbb{C}^{m\times n}$ having rank k and such that $\|A_j\|_2 \leqslant \gamma e^{-\alpha|j|}$. Then*

$$\sigma_l(A) \leqslant \frac{2\gamma}{1-e^{-\alpha}} \cdot e^{-\alpha\frac{l-k}{2k}}, \qquad \sigma_l(A^+) \leqslant \frac{\gamma}{1-e^{-\alpha}} \cdot e^{-\alpha\frac{l-k}{k}}.$$

Proof. Note that $\sum_{|j|<\lceil\frac{l-k}{2k}\rceil} A_j$ is at most a rank-$(l-1)$ approximation of A. This implies that

$$\sigma_l(A) \leqslant \left\| A - \sum_{|j|<\lceil\frac{l-k}{2k}\rceil} A_j \right\|_2 = \left\| \sum_{|j|\geqslant\lceil\frac{l-k}{2k}\rceil} A_j \right\|_2 \leqslant \sum_{|j|\geqslant\lceil\frac{l-k}{2k}\rceil} \gamma e^{-\alpha|j|} =$$
$$= 2\gamma e^{-\alpha\lceil\frac{l-k}{2k}\rceil} \sum_{j\geqslant 0} e^{-\alpha j} = \frac{2\gamma}{1-e^{-\alpha}} \cdot e^{-\alpha\lceil\frac{l-k}{2k}\rceil}.$$

The same arguments can be applied to obtain the bound on $\sigma_l(A^+)$. □

Remark 4.3.2. *In the particular case $k = 1$ the above result yields*

$$\sigma_l(A) \leqslant \frac{2\gamma}{1-e^{-\alpha}} \cdot e^{-\alpha\frac{l-1}{2}} \qquad \sigma_l(A^+) \leqslant \frac{\gamma}{1-e^{-\alpha}} \cdot e^{-\alpha(l-1)}.$$

Using the previous result we can see what happens to the rate of the power law when we perform an arithmetic mean of matrices with the same exponential decay in their singular values.

Lemma 4.3.3. *Let $A = \frac{1}{k}\sum_{i=1}^{k} A_i \in \mathbb{C}^{m\times m}$ be the mean of k matrices that satisfy the uniform relation*

$$\sigma_l(A_i) \leqslant \gamma e^{-\alpha l}, \quad \forall l = 1, \dots, m.$$

Then it holds that

$$\sigma_l(A) \leqslant \tilde{\gamma} e^{-\alpha\frac{l-k}{k}}, \quad \tilde{\gamma} = \frac{\gamma}{1-e^{-\alpha}}$$

Proof. Note that every matrix A_i can be expanded as the sum of its singular vectors (here taken to infinity for convenience by setting $\sigma_j(A) = 0$ for every $j > m$):

$$A_i = \sum_{j=1}^{\infty} \sigma_j(A) u_{i,j} v_{i,j}^*, \quad \|u_{i,j}\|_2 = \|v_{i,j}\|_2 = 1$$

This allows to write

$$A = \frac{1}{k}\sum_{i=1}^{k} A_i = \sum_{j=1}^{\infty} \left(\frac{1}{k} \sum_{i=1}^{k} \sigma_j(A_i) u_{i,j} v_{i,j}^* \right) = \sum_{j=1}^{\infty} \tilde{A}_j.$$

It is very easy to check that $\tilde{A}_j$ are rank k matrices such that $\|A_j\|_2 \leqslant \gamma e^{-\alpha j}$. This implies that we can apply Lemma 4.3.1 and obtain the thesis. □

To conclude, we specify that if we can bound the singular values of a certain matrix, then adding a low rank correction only shifts the bound.

Lemma 4.3.4. *Let $A, B \in \mathbb{C}^{m\times m}$ and suppose that B has rank k. Then*

$$\sigma_{l+k}(A+B) \leqslant \sigma_l(A).$$

Proof. For the Eckart-Young-Mirsky theorem $\forall l = 1, \dots, m$ $\exists \widetilde{A}$ of rank $l-1$ such that $\|A - \widetilde{A}\|_2 = \sigma_l(A)$. Therefore, since $\widetilde{A} + B$ has rank less or equal than $l+k-1$ we have

$$\sigma_{l+k}(A+B) \leqslant \|(A+B) - (\widetilde{A}+B)\|_2 = \sigma_l(A).$$

□

4.4 SINGULAR VALUES OF OUTER PRODUCTS AND QR FACTORIZATION

Lemma 4.3.1 can be used for estimating the singular values of a sum of dyads $X = \sum_{i=1}^{s} u_i v_i^*$, $u_i \in \mathbb{C}^m$, $v_i \in \mathbb{C}^n$, $s \in \mathbb{Z}^+ \cup \{\infty\}$, relating the decay rate in the singular values with those of the quantity $\|u_i v_i^*\|_2$. This bound can be rude in certain settings. For example, the singular values of X decay fast also when the vectors u_i and/or v_i tend to become dependent even if $\|u_i v_i^*\|_2$ has a slow decay rate.

For describing this phenomenon, we try to rephrase the expression $X = \sum_{i=1}^{s} u_i v_i^*$ as $X = \sum_{i=1}^{s} \tilde{u}_i \tilde{v}_i^*$ where $\tilde{u}_i$ and $\tilde{v}_i$ are chosen as "orthogonal as possible". To this aim, we study the QR decomposition of the matrices

$$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_s \end{bmatrix}, \qquad V = \begin{bmatrix} v_1 & v_2 & \cdots & v_s \end{bmatrix}, \qquad X = UV^t.$$

Indicating their QR decompositions with (Q_u, R_u) and (Q_v, R_v) we get

$$X = Q_u R_u R_v^* Q_v^*, \qquad Q_u \in \mathbb{C}^{m\times m}, \quad R_u \in \mathbb{C}^{m\times s}, \quad Q_v \in \mathbb{C}^{n\times n}, \quad R_v \in \mathbb{C}^{n\times s},$$

therefore the singular values of X coincide with those of the matrix $R_u R_v^*$. A property of "progressive dependence" in the vectors u_i and v_i can be translated as an entry-wise decay in the entries of R_u and R_v.

The rest of the section is devoted to show how entry-wise decay properties in the factors R_u and R_v imply the decay in the singular values of X.

Theorem 4.4.1. *Let $U = Q_u R_u$ and $V = Q_v R_v$ be QR factorizations of $U \in \mathbb{C}^{m\times s}$ and $V \in \mathbb{C}^{n\times s}$. Let $\alpha_u, \beta_u, \gamma_u, \alpha_v, \beta_v$ and γ_v be positive constants such that*

$$|R_{u,ij}| \leqslant \gamma_u e^{-\alpha_u i - \beta_u j}, \qquad |R_{v,ij}| \leqslant \gamma_v e^{-\alpha_v i - \beta_v j}$$

for any i, j.

Then, the singular values of the matrix $X = UV^*$ can be bounded by

$$\sigma_l(X) \leqslant \gamma e^{-\alpha(l+1)}.$$

where $\gamma := \frac{\gamma_u \gamma_v e^{\alpha - \tilde{\alpha}}}{(1-e^{-(\beta_u+\beta_v)})(1-e^{-2\tilde{\alpha}})}$, $\alpha := \max\{\alpha_u, \alpha_v\}$ and $\tilde{\alpha} := \min\{\alpha_u, \alpha_v\}$.

Proof. As pointed out previously, we can focus on the singular values of $S = R_u R_v^*$. Notice that the element in position (i, j) of S can be bounded in absolute value by

$$|S_{ij}| \leqslant \left(\sum_{l=1}^{s} e^{-(\beta_u+\beta_v)l} \right) \gamma_u \gamma_v e^{-\alpha_u i - \alpha_v j} \leqslant \tilde{\gamma} e^{-\alpha_u i - \alpha_v j}, \qquad \tilde{\gamma} = \frac{\gamma_u \gamma_v}{1 - e^{-(\beta_u+\beta_v)}}.$$

We can estimate the l-th singular value by setting the first $l-1$ rows or columns of S to zero according to the maximum between α_u and α_v. Notice that this is equivalent to apply a rank $l-1$ correction and use the Eckart-Young-Mirsky theorem. For example, consider the case $\alpha = \alpha_u$ and let S_l be the matrix composed by the last $m - l + 1$ rows of S. In particular, we have $\sigma_l(S) \leqslant \|S_l\|_2$. Observe that the entries of S_l satisfy the relation $(S_l)_{ij} \leqslant \tilde{\gamma} e^{-\alpha l} e^{-\tilde{\alpha}(i+j-1)}$. We have

$$\left\| \frac{e^{\alpha l}}{\tilde{\gamma}} S_l \right\|_F^2 = \sum_{i=1}^{m-l} \sum_{j=1}^{n} \left| \frac{e^{\alpha l}}{\tilde{\gamma}} (S_k)_{i,j} \right|^2 \leqslant \sum_{t=1}^{\infty} t e^{-2\tilde{\alpha} t} = -\frac{1}{2} \frac{d}{d\tilde{\alpha}} \left(\sum_{t=1}^{\infty} e^{-2\tilde{\alpha} t} \right) = \frac{e^{-2\tilde{\alpha}}}{(1 - e^{-2\tilde{\alpha}})^2}.$$

Since $\|S_l\|_2 \leqslant \|S_l\|_F$ we have $\sigma_l(S) \leqslant \gamma e^{-\alpha(l+1)}$. □

Remark 4.4.2. *Theorem 4.4.1 claims that the decay rate in the singular values of X is at least the maximum between the rates along the columns of the two R factors, α_u and α_v respectively. It is worth to point out that the typical behavior observed experimentally for $\sigma_l(X)$, involves a decay rate of $\alpha_u + \alpha_v$. To shed some lights on the issue, we consider the case $\alpha = \alpha_u = \alpha_v$ and we represent a matrix with a "Hankel-like" exponential decay in its entries (the matrix S in the previous theorem) with a product DAD where*

$$A \in \mathbb{C}^{m \times m}, \qquad D = \begin{bmatrix} \rho & & \\ & \ddots & \\ & & \rho^m \end{bmatrix}, \quad \rho = e^{-\alpha}.$$

Then, as a consequence of Horn's theorem 4.1.2 we can write

$$\prod_{i=1}^{l} \sigma_i(DAD) \leqslant \|A\|_2^l \cdot \rho^{l(l+1)}, \qquad \prod_{i=1}^{m} \sigma_i(DAD) = \prod_{i=1}^{m} \sigma_i(A) \cdot \rho^{m(m+1)}.$$

Observe that using Corollary 4.1.3 we re-obtain the coarse bound $\sigma_l(DAD) \leqslant \|A\|_2 \rho^{l+1}$. Assuming $\sigma_1(DAD) \approx \|A\|_2 \cdot \rho^2$ and that the asymptotic gap between two consecutive singular values is constant, i.e.

$$\frac{\sigma_{i+1}(DAD)}{\sigma_i(DAD)} \approx \rho^c \quad \forall i = 1, \dots m-1,$$

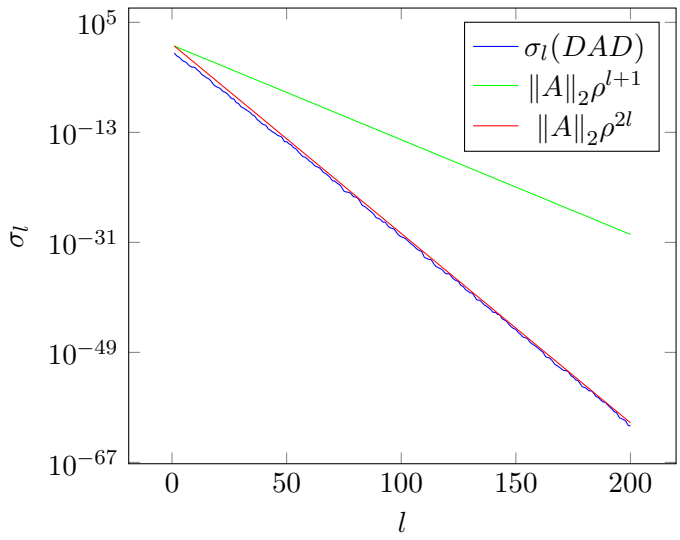

Figure 4.1.: Typical singular values distribution of a matrix with "Hankel-like" decay in its entries. With the notation of Remark 4.4.2 we pick the entries of $A \in \mathbb{R}^{200\times 200}$ from a distribution $\mathcal{N}(0,1)$ and we consider $\rho = e^{-\alpha} = 0.7$. The green line and the red line represent the functions $\|A\|_2 e^{-\alpha(l+1)}$ and $\|A\|_2 e^{-2\alpha l}$ respectively

we have that necessarily $c = 2$. *In other words the decay rate is equal to* 2α.

Moreover, it is important to remark that the "Hankel-like" decay alone is not sufficient to ensure the double decay rate. Consider for example

$$A = \begin{bmatrix} 1 & & & \\ & & & 1 \\ & & \iddots & \\ & \iddots & & \\ 1 & & & \end{bmatrix} \in \mathbb{R}^{m\times m}, \qquad DAD = \begin{bmatrix} \rho^2 & & & \\ & & & \rho^{m+2} \\ & & \iddots & \\ & \iddots & & \\ \rho^{m+2} & & & \end{bmatrix}.$$

In this case $\sigma_1(DAD) = \rho^2$ *and* $\sigma_i(DAD) = \rho^{m+2}\ \forall i > 1$, *so there is no double decay rate.*

Due to technical reasons linked with the structured outer products we are going to analyze in the next sections, we consider the presence of the counter identity matrix

$$J_s = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix} \in \mathbb{R}^{s\times s}$$

between the multiplication of the two R factors.

In this case it is possible to adapt the argument used to prove Theorem 4.4.1 for retrieving bounds on the singular values. In addition, the outcome turns out to be negligible as the number of columns in the outer product increases. We suppose for simplicity the same decay along the rows and columns of R_u and R_v respectively.

Theorem 4.4.3. *Let $U = Q_u R_u$ and $V = Q_v R_v$ be QR factorizations of $U \in \mathbb{C}^{m\times s}$ and $V \in \mathbb{C}^{m\times s}$. Let α, β and γ be positive constants such that $|R_{u,ij}|, |R_{v,ij}| \leqslant \gamma e^{-\alpha i - \beta j}$ for any i, j. Then the matrix $X = UJ_sV^*$ has singular values bounded by*

$$\sigma_l(X) \leqslant \gamma e^{-\alpha(l+1)}, \qquad \gamma := \frac{\gamma^2 s e^{-(s+1)\beta}}{(1 - e^{-2\alpha})}.$$

Proof. We can write $X = UJ_sV^* = Q_u R_u J_s R_v^* Q_v^*$, so its singular values coincide with the ones of $S = R_u J_s R_v^*$. The element in position (i, j) of S is obtained as the a sum

$$S_{ij} = \sum_{l=1}^{s} R_{u,il} \cdot R_{v,j(s-l)}, \qquad |R_{U,il} \cdot R_{v,j(s+1-l)}| \leqslant \gamma^2 e^{-\alpha(i+j)-\beta(s+1)}$$

according to our hypotheses. Since the bound on the elements in the above summation is independent of l we can write $|S_{ij}| \leqslant \gamma^2 s e^{-\beta(s+1)} e^{-\alpha(i+j)}$. The thesis can then be obtained by following the same procedure as in Theorem 4.4.1. □

Remark 4.4.4. *Observe that the larger s the closer the quantity $se^{-\beta s}$ is to 0. Therefore for sufficiently big s the resulting matrix X is negligible.*

4.5 SINGULAR VALUES OF STRUCTURED OUTER PRODUCTS

In this section we analyze certain outer products which enjoy the two-way decay property in the R factor we analyzed in Section 4.4.

We consider products UV^* in which the columns of U and V are of the form $p_i(A)b$, for generic $A \in \mathbb{C}^{m\times s}$ $b \in \mathbb{C}^s$ and some sequence of polynomials $\{p_i\}$. Formally, we introduce the following two classes of matrices.

Definition 4.5.1. *Given $A \in \mathbb{C}^{m\times s}$ and $b \in \mathbb{C}^s$ we define*

$$\mathcal{KM}_n(A, b) := \left[\, b \,\middle|\, Ab \,\middle|\, \ldots \,\middle|\, A^{n-1}b \,\right] \in \mathbb{C}^{m\times n}. \tag{4.2}$$

The span of the first i columns is indicated with $\mathcal{K}_i(A, b)$ and it is called the Krylov subspace *of dimension i generated by A and b.*

Definition 4.5.2. *Given $A \in \mathbb{C}^{m\times s}$, $b \in \mathbb{C}^s$ and $p(x) = \sum_{i=0}^{n-1} a_i x^i \in \mathbb{C}[x]$ we define*

$$\mathcal{HM}_p(A, b) := \left[a_{n-1}b \,\middle|\, (a_{n-1}A + a_{n-2}I)b \,\middle|\, \ldots \,\middle|\, \sum_{i=0}^{n-1} a_i A^i b \right] \in \mathbb{C}^{m\times n}. \tag{4.3}$$

It is evident that the polynomials used for generating (4.2) correspond to the monomial basis. Instead, the columns in (4.3) correspond to the so called Horner shifts (which are the intermediate results obtained in the evaluation of a polynomial by means of the Horner rule [58]) of $p(A)b$. In the following we refer to the patterns of (4.2) and (4.3) as *Krylov* and *Horner* matrices, respectively.

The rest of the section is dedicated to proving the element-wise decay in the QR factorization of Krylov and Horner matrices. Then, we draw the conclusions on the singular values of outer products between matrices which have these structures.

4.5.1 *Polynomial interpolation tools*

The key argument —for proving the decay property of Krylov and Horner matrices— is a connection between the entries of their R factors and the residual of a minimax polynomial approximation problem. The rate of decay in the R factor is related to the rate of convergence of this approximation problem, with respect to the degree of the approximant. The latter can be estimated with a classical result of Bernstein and depends on some geometric notions that we need to introduce.

Our approach is inspired by the one of Benzi and Boito in [10, 9], where the authors proved the numerical preservation of sparsity patterns in matrix functions. For a classic reference of the complex analysis behind the next definitions and theorems we refer to [74, 41].

Definition 4.5.3 (Logarithmic capacity)**.** *Let $F \subseteq \mathbb{C}$ be a nonempty, compact and connected set, and denote with G_∞ the connected component of the complement containing the point at the infinity. Since G_∞ is simply connected, in view of the Riemann Mapping Theorem there exists a conformal map $\Phi(z)$ which maps G_∞ to the complement of a disc. If we impose the normalization conditions*

$$\Phi(\infty) = \infty, \qquad \lim_{z\to\infty} \frac{\Phi(z)}{z} = 1$$

then this disc is uniquely determined. We say that its radius ρ is the logarithmic capacity *of F and we write $\mathrm{lc}(F) = \rho$. Let $\Psi = \Phi^{-1}$, for every $R > \rho$ we indicate with C_R the image under Ψ of the circle $\{|z| = R\}$.*

In our settings it is not restrictive to consider a Jordan region F with a rectifiable boundary ∂F. Moreover, for almost every point $z \in \partial F$ it is defined a tangent vector which makes an angle $\theta(z)$ with the positive real axis. We call the quantity

$$\mathcal{V} := \int_{\partial F} |d\theta(z)|$$

the *total rotation* of F. In general $\mathcal{V} \geqslant 2\pi$ and if F is convex then $\mathcal{V} = 2\pi$, see [41].

The logarithmic capacity and the total variation are strictly related to the following well-known result of Bernstein about the polynomial approximation in the complex plane.

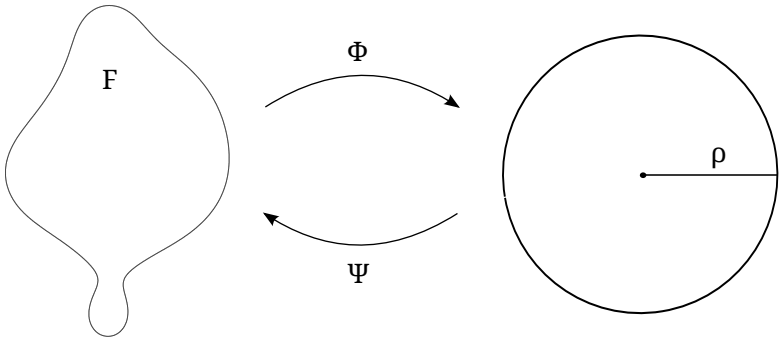

Figure 4.2.: Conformal transformations through the Riemann map and its inverse

Lemma 4.5.4 (Corollary 2.2 in [41])**.** *Let F be a Jordan region whose boundary is of finite total rotation $\mathcal{V}$ and of logarithmic capacity ρ. If $f(z)$ is an analytic function on $\mathbb{C}$ then $\forall r > \rho$ and any integer $i \geqslant 0$ there exists a polynomial $p_i(z)$ of degree at most i such that*

$$\|f(z) - p_i(z)\|_{\infty,F} := \min_{z \in F} |f(z) - p_i(z)| \leqslant \frac{M(r)\mathcal{V}}{\pi(1 - \frac{\rho}{r})} \left(\frac{\rho}{r}\right)^{i+1}.$$

with $M(r) := \max_{C_r} |f(z)|$.

In order to exploit Lemma 4.5.4 in our framework, we need to introduce some quantitative notions concerning the geometry of the set F.

Definition 4.5.5. *Given $F \subseteq \mathbb{C}$ compact, connected with $\mathrm{lc}(F) = \rho \in (0, 1)$, we indicate with R_F the quantity*

$$R_F := \sup\{R > \rho : \ C_R \text{ is strictly contained in the unit circle}\}.$$

Definition 4.5.6. *We say that $F \subset \mathbb{C}$ is* enclosed *by the triple of parameters $(\rho, R_F, \mathcal{V}_F)$ if $\exists F'$ Jordan region whose boundary has finite total rotation $\mathcal{V}_F$, $\mathrm{lc}(F') = \rho$, $R_{F'} = R_F$ and $F \subseteq F'$.*

In the next section we see that the set where we perform the polynomial approximation is the spectrum of the matrix which generates a Krylov matrix or a Horner matrix. That is why we also introduce the following definitions.

Definition 4.5.7. *We say that $A \in \mathbb{C}^{m \times m}$ is* enclosed *by $(\rho, R_A, \mathcal{V}_A)$ if the set of its eigenvalues is enclosed by $(\rho, R_A, \mathcal{V}_A)$.*

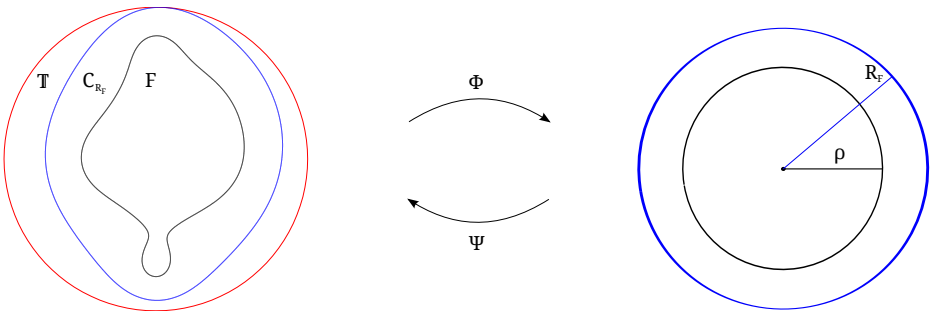

Figure 4.3.: Red line: the unit circle; blue line: C_{R_F}; black line: the region F

Definition 4.5.8. *Let J be the Jordan canonical form of $A \in \mathbb{C}^{m\times m}$. Let $\mathbb{V} := \{V \in \mathbb{C}^{m\times m} : V^{-1}AV = J\}$. We define the* spectral condition number *as the quantity*

$$\kappa_s(A) := \inf_{V\in\mathbb{V}} \| V \|_2 \| V^{-1} \|_2 .$$

4.5.2 *Decay in the entries of the R factor for Krylov matrices*

Theorem 4.5.9. *Let $A \in \mathbb{C}^{m\times m}$ be a diagonalizable matrix enclosed by $(\rho, R_A, \mathcal{V}_A)$, $\rho \in (0,1)$, $b \in \mathbb{C}^m$ and $U = \mathcal{KM}_n(A,b)$.*

Then $\forall r \in (\rho, R_A)$ the entries of the R factor in the QR decomposition of U satisfy

$$|R_{ij}| \leqslant c(r)\cdot\kappa_s(A)\cdot\left(\frac{\rho}{r}\right)^i \delta^j$$

where $\delta = \max_{z\in C_r}|z|$ and $c(r) = \frac{\mathcal{V}_A}{\delta\pi(1-\frac{\rho}{r})}\cdot\|b\|_2$.

Proof. Let $QR = U$ be the QR factorization of U and $V^{-1}AV = D$ the spectral decomposition of A. Notice that the quantity $\|R_{i+1:j,j}\|_2$ is equal to the norm of the projection of u_j on the orthogonal to the space spanned by the first i columns of U, that is $\mathcal{K}_i(A,b)^\perp$. It is well-known that the Krylov subspace $\mathcal{K}_i(A,b)$ contains all the vectors of the form $p(A)b$ where p has degree at most $i-1$. In particular, we have:

$$\begin{aligned}
|R_{i+1,j}| \leqslant \|R_{i+1:j,j}\|_2 &\leqslant \min_{deg(p)=i-1}\|p(A)b - u_j\|_2 \\
&= \min_{deg(p)=i-1}\|p(A)b - A^{j-1}b\|_2 \\
&\leqslant \min_{deg(p)=i-1}\|p(D) - D^{j-1}\|_2\|V^{-1}\|_2\|V\|_2\|b\|_2 \\
&\leqslant \frac{M(r)\mathcal{V}_A}{\pi(1-\frac{\rho}{r})}\left(\frac{\rho}{r}\right)^i \kappa_s(A)\|b\|_2,
\end{aligned}$$

where $M(r) = \max_{C_r}|z|^{j-1} = \delta^{j-1}$. □

Non diagonalizable case

The diagonalizability hypothesis can be relaxed using different strategies. We first propose to rely on a well-known result by Crouzeix [36] based on the numerical range. Then, we discuss another approach consisting in estimating the minimax approximation error on the Jordan canonical form.

Numerical range

In the spirit of the results found in [9], we can give an alternative formulation that avoids the requirement of diagonalizability. The price to pay consists in having to estimate the minimax error bound on a set larger than the spectrum. To be precise, we need to consider the numerical range of the matrix A.

Definition 4.5.10. *Let A be a matrix in $\mathbb{C}^{m\times m}$. We define its numerical range $\mathcal{W}(A)$ as the set*

$$\mathcal{W}(A) = \{x^*Ax \mid x \in \mathbb{C}^m, \quad \|x\|_2 = 1\} \subseteq \mathbb{C}.$$

The numerical range is a compact convex subset of $\mathbb{C}$ which contains the eigenvalues of A. When A is normal $\mathcal{W}(A)$ is exactly the convex hull of the eigenvalues of A. Moreover, it has a strict connection with the evaluation of matrix functions, which is described by the following result.

Theorem 4.5.11 (Crouzeix [36])**.** *There is a universal constant $2 \leqslant \mathcal{C} \leqslant 11.08$ such that, given $A \in \mathbb{C}^{m\times m}$, and a continuous function $g(z)$ on $\mathcal{W}(A)$, analytic in its interior, the following inequality holds:*

$$\|g(A)\|_2 \leqslant \mathcal{C} \cdot \|g(z)\|_{\infty,\mathcal{W}(A)}.$$

Whenever the numerical range $\mathcal{W}(A)$ has a logarithmic capacity smaller than 1 it is possible to extend Theorem 4.5.9.

Corollary 4.5.12. *Let $A \in \mathbb{C}^{m\times m}$ be such that its numerical range $\mathcal{W}(A)$ is enclosed by $(\rho, R_{\mathcal{W}(\mathcal{A})}, \mathcal{V}_{\mathcal{W}(A)})$, $\rho \in (0,1)$ and $b \in \mathbb{C}^m$. Moreover, let $b \in \mathbb{C}^m$ and $U = \mathcal{KM}_n(A,b)$.*

Then $\forall r \in (\rho, R_{W(A)})$ the entries of the R factor in the QR decomposition of U satisfy

$$|R_{ij}| \leqslant c(r) \cdot \left(\frac{\rho}{r}\right)^i \delta^j$$

where $\delta = \max_{z\in C_r} |z|$ and $c(r) = \frac{\mathcal{C}\cdot\mathcal{V}_{\mathcal{W}(A)}}{\delta\pi(1-\frac{\rho}{r})} \cdot \|b\|_2$.

Proof. It is sufficient to follow the same steps of the proof of Theorem 4.5.9 employing Theorem 4.5.11 to bound R_{ij}. □

Jordan canonical form

An alternative to the above approach is to rely on the Jordan canonical form in place of the eigendecomposition. More precisely, we can always write any matrix A as $A = V^{-1}JV$ with J being block diagonal with bidiagonal blocks (the so-called Jordan blocks). This implies that the matrix $f(J)$ is block diagonal with blocks $f(J_t)$ where $f(J_t)$ have the following form:

$$J_t = \begin{bmatrix} \lambda_t & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_t \end{bmatrix} \in \mathbb{C}^{m_t \times m_t}, \quad f(J_t) = \begin{bmatrix} f(\lambda_t) & f'(\lambda_t) & \dots & \frac{f^{(m_t-1)}(\lambda_t)}{(m_t-1)!} \\ & \ddots & \ddots & \vdots \\ & & \ddots & f'(\lambda_t) \\ & & & f(\lambda_t) \end{bmatrix}.$$

We can evaluate the matrix function $f(A)$ by $f(A) = V^{-1}f(J)V$.
Estimating the norm $\|R_{i+1:j,j}\|_2$ as in the proof of Theorem 4.5.9 we get

$$|R_{i+1,j}| \leqslant \min_{deg(p)=i-1} \|p(A)b - u_j\|_2 \leqslant \min_{deg(p)=i-1} \|p(J) - J^{j-1}\|_2 \cdot \kappa_s(A) \cdot \|b\|_2 \tag{4.4}$$

where $p(J) = \mathrm{diag}(p(J_t))$, $J^j = \mathrm{diag}(J_t^j)$ and

$$p(J_t) - J_t^j = \begin{bmatrix} p(\lambda_t) - \lambda_t^j & p'(\lambda_t) - j\lambda_t^{j-1} & \dots & \frac{p^{(m_t-1)}(\lambda_t)}{(m_t-1)!} - \frac{j!}{(j-m_t)!(m_t-1)!}\lambda_h^{j-m_t} \\ & \ddots & \ddots & \vdots \\ & & \ddots & p'(\lambda_t) - j\lambda_t^{j-1} \\ & & & p(\lambda_t) - \lambda_t^j \end{bmatrix}. \tag{4.5}$$

We can rephrase (4.4) as a problem of simultaneous approximation of a function and its derivatives.

Lemma 4.5.13. *Let $\mathcal{S}$ be a simply connected subset of the complex plane and suppose that $\exists z_0 \in \mathcal{S}$ such that each element of $\mathcal{S}$ can be connected to z_0 with a path of length less than 1. Let $p(z)$ be a degree i polynomial approximating the holomorphic function $f'(z)$ in $\mathcal{S}$, such that $|f'(z) - p(z)| \leqslant \epsilon \ \forall z \in \mathcal{S}$. Then there exists a polynomial $q(z)$ of degree $i+1$ with $q'(z) = p(z)$ such that*

$$|q(z) - f(z)| \leqslant \epsilon \qquad \forall z \in \mathcal{S}.$$

Proof. Define $q(z)$ as follows:

$$q(z) = f(z_0) + \int_\gamma p(z), \qquad \gamma \text{ any path connecting } z_0 \text{ and } z.$$

The above definition uniquely determines $q(z)$, and we know that it is a polynomial of degree $i+1$. Given $z \in \mathcal{S}$ choose γ a path connecting z_0 to z with length less than 1, we have:

$$|f(z) - q(z)| = |f(z_0) + \int_\gamma f'(z) - f(z_0) - \int_\gamma p(z)| \leqslant \int_\gamma |f'(z) - p(z)| \leqslant \epsilon.$$

□

If $m_{t'}$ is the maximum size among all the Jordan blocks we can find a minimax approximating polynomial for the $m_{t'}$ derivative of z^j. The above Lemma guarantees that, with the latter choice, the matrix (4.5) has the (i,j)-th entry bounded in modulus by $\frac{\epsilon}{(j-i)!}$ when $j \geqslant i$. An easy computation shows that both the 1 and ∞ norms of

$$T = \epsilon \begin{bmatrix} 1 & 1 & \frac{1}{2!} & \cdots & \frac{1}{(m_{t'}-1)!} \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \frac{1}{2!} \\ & & & \ddots & 1 \\ & & & & 1 \end{bmatrix}$$

are bounded by ϵe, where e is the Napier constant. Then, we have $\|p(J) - J^k\|_2 \leqslant \|T\|_2 \leqslant \sqrt{\|T\|_1 \|T\|_\infty} \leqslant \epsilon e$. Using this relation one can prove the next result by following the same steps as in the proof of Theorem 4.5.9.

Theorem 4.5.14. *Let $A \in \mathbb{C}^{m\times m}$, $b \in \mathbb{C}^m$, $U = \mathcal{KM}_n(A,b)$ and F be the convex hull of the spectrum of A. Suppose that $F \subseteq B(0,1)$ is enclosed by $(\rho, R_F, \mathcal{V}_F)$, $\rho \in (0,1)$ and indicate with $m_{t'}$ the size of the largest Jordan block of A.*

Then $\forall r \in (\rho, R_F)$ the entries of the R factor in the QR decomposition of U satisfy

$$|R_{ij}| \leqslant c(r) \cdot \kappa_s(A) \cdot \left(\frac{\rho}{r}\right)^{i-(m_{t'}-1)} \delta^j,$$

where $\delta = \max_{z\in C_r} |z|$ and $c(r) = \frac{e\cdot\mathcal{V}_F}{\delta\pi(1-\frac{\rho}{r})} \cdot \|b\|_2$.

4.5.3 *Decay in the entries of the R factor for Horner matrices*

Theorem 4.5.15. *Let $A \in \mathbb{C}^{m\times m}$ be a diagonalizable matrix enclosed by $(\rho, R_A, \mathcal{V}_A)$, $\rho \in (0,1)$ and $b \in \mathbb{C}^m$. Moreover let $p(x) = \sum_{j=0}^{s-1} a_j x^j$ and $U = \mathcal{HM}_p(A,b)$ where the finite sequence $\{a_j\}_{j=0,\dots,s-1}$ verifies*

$$|a_j| \leqslant \hat{\gamma} \cdot \hat{\rho}^{j+1}, \quad \hat{\gamma} > 0, \quad \hat{\rho} \in (0,1), \qquad j = 0, \dots, s-1.$$

Then the R factor in the QR decomposition of U is entry-wise bounded by

$$|R_{ij}| \leqslant c \cdot \kappa_s(A) \cdot \left(\frac{\rho}{R_A}\right)^i \hat{\rho}^{i+(s-j)}$$

where $c = \frac{\hat{\rho}\hat{\gamma}\mathcal{V}_A}{\pi(1-\hat{\rho})(1-\frac{\rho}{R_A})}\|b\|_2$.

Proof. Here we assume that $a_{s-1} \neq 0$. This is not restrictive because if $j < s-1$ is the largest j such that $a_{j'} = 0$ for any $j' > j$ the first $s-1-j$ columns of U are zero, and can be ignored. Observe that the j-th column of U is of the form $q(A)b$ where q is the polynomial defined by the coefficients a_j in reversed order, i.e.,

$$q(x) := \sum_{n=0}^{j-1} a_{s-j+n} x^n.$$

The subspace spanned by the first i columns of U contains all the vectors of the form $p(A)b$ where p is a polynomial of degree at most $i-1$. With the same argument used for proving Theorem 4.5.9 we can bound the entries of R in this way

$$|R_{ij}| \leqslant \min_{deg(p)=i-1} \left\| p(D) - \sum_{n=0}^{j-1} a_{s-j+n} D^n \right\|_2 \cdot \kappa_s(A) \cdot \|b\|_2.$$

Moreover

$$\begin{aligned}
\min_{deg(p)=i-1} \left\| p(D) - \sum_{n=0}^{j-1} a_{s-j+n} D^n \right\|_2 &= \min_{deg(p)=i-1} \left\| p(D) - \sum_{n=i}^{j-1} a_{s-j+n} D^n \right\|_2 \\
&\leqslant \sum_{n=i}^{j-1} |a_{s-j+n}| \min_{deg(p)=i-1} \|p(D) - D^n\|_2 \\
&\leqslant \sum_{n=i}^{j-1} \hat{\gamma}\hat{\rho}^{s-j+1+n} \min_{deg(p)=i-1} \|p(D) - D^n\|_2 \\
&\leqslant \sum_{n=i}^{j-1} \hat{\gamma}\hat{\rho}^{s-j+1+n} \frac{\mathcal{V}_A}{\pi(1-\frac{\rho}{R_A})} \left(\frac{\rho}{R_A}\right)^i \\
&\leqslant \frac{\hat{\rho}\hat{\gamma}\mathcal{V}_A}{\pi(1-\hat{\rho})(1-\frac{\rho}{R_A})} \hat{\rho}^{s-j+i} \left(\frac{\rho}{R_A}\right)^i,
\end{aligned}$$

where we used Lemma 4.5.4 with $r = R_A$. □

Remark 4.5.16. *In view of the above arguments we can rephrase Theorem 4.5.9 for non diagonalizable matrices. We obtain similar statements involving* $\mathrm{lc}(\mathcal{W}(A))$ *in place of* $\mathrm{lc}(A)$ *or with a shifted column decay. The same technique can be used to generalize the results of the next sections. The proofs and statements are analogous to the diagonalizable case. Therefore, they are not reported.*

4.5.4 *Decay in the singular values of Krylov/Horner outer products*

Before starting —due to technical reasons— we need to introduce the following quantity.

Definition 4.5.17. *Given $A \in \mathbb{C}^{m\times m}$ enclosed by $(\rho, R_A, \mathcal{V}_A)$ and a parameter $R \in \mathbb{R}^+$ we define*

$$\Lambda(\rho, R_A, \mathcal{V}_A, R) := \frac{\mathcal{V}_A^2}{\pi^2(R-1)(1-\frac{\rho}{R_A})\sqrt{1-(\frac{\rho}{RR_A})^2}} \cdot \min_{\rho<r<R_A} \frac{1}{\delta(r)(1-\delta(r)^2)(\frac{r}{\rho}-1)\sqrt{(1-\frac{\rho^2}{r^2})}},$$

where $\delta(r) := \max\{\frac{1}{R}, \max_{C_r}|z|\}$.

Now, we have all the ingredients for studying the singular values of outer products between Krylov and Horner matrices. This can be done combining the decay properties of Section 4.5 with the results of Section 4.4. This strategy provides the following theorem which will be useful in Chapter 5.

Theorem 4.5.18. *Let $A_1 \in \mathbb{C}^{m\times m}$ and $A_2 \in \mathbb{C}^{n\times n}$ be two diagonalizable matrices enclosed by $(\rho, R_A, \mathcal{V}_A)$ with $\rho \in (0,1)$. Moreover, let $b_1 \in \mathbb{C}^m$ and $b_2 \in \mathbb{C}^n$. Then for any polynomial $p(x) = \sum_{j=0}^{s-1} a_j x^j$ which verifies*

$$|a_j| \leqslant \hat{\gamma} \cdot R^{-(j+1)}, \quad R > 1, \quad j \in \{0, \dots, s-1\},$$

the singular values of

$$X = \mathcal{KM}_s(A_1, b_1) \cdot J_s \cdot \mathcal{HM}_p(A_2, b_2)^* \tag{4.6}$$

can be bounded by

$$\sigma_l(X) \leqslant \gamma \cdot e^{-(\alpha+\alpha')(l+1)}, \qquad \alpha = \log\left(\frac{R_A}{\rho}\right), \qquad \alpha' = \log(R),$$

where γ is defined as

$$\gamma := \hat{\gamma} \cdot \kappa_s(A_1)\kappa_s(A_2)\|b_1\|_2\|b_2\|_2 \cdot \Lambda(\rho, R_A, \mathcal{V}_A, R).$$

Proof. Consider the matrices U and V defined as follows:

$$U = \left[\begin{array}{c|c|c|c} b_1 & A_1 b_1 & \dots & A_1^{s-1} b_1 \end{array}\right], \quad V = \left[\begin{array}{c|c|c|c} a_{s-1}b_2 & (a_{s-1}A_2 + a_{s-2}I)b_2 & \dots & \displaystyle\sum_{j=0}^{s-1} a_j A_2^j b_2 \end{array}\right],$$

so that we have $X = UJV^*$ as in Equation (4.6). Moreover, let (Q_u, R_u) and (Q_v, R_v) be the QR factorizations of U and V respectively. Applying Theorem 4.5.9 and Theorem 4.5.15 we get that $\forall r \in (\rho, R_A)$

$$|R_{u,ij}| \leqslant c_1(r) \cdot e^{-\eta i - \beta j} \qquad \text{and} \qquad |R_{v,ij}| \leqslant c_2 \cdot e^{-(\alpha+\alpha')i - \beta(s-j)},$$

with $\eta = \log\left(\frac{r}{\rho}\right)$, $\beta = |\log(\delta)|$ and

$$c_1(r) = \frac{\mathcal{V}_{A_1}}{\delta\pi(1-\frac{\rho}{r})} \cdot \kappa_s(A_1) \cdot \|b_1\|_2, \qquad c_2 = \frac{\hat{\rho}\hat{\gamma}\mathcal{V}_{A_2}}{\pi(1-\hat{\rho})(1-\frac{\rho}{R_A})}\kappa_s(A_2)\|b_2\|_2.$$

In order to bound the singular values of X we look at those of $S = R_u J_s R_v^*$. The entry (i,j) of S is obtained as the sum:

$$S_{ij} = \sum_{h=1}^{s} R_{u,ih} \cdot R_{v,j(s-h)}, \qquad |R_{u,ih} \cdot R_{v,j(s-h)}| \leqslant c \cdot e^{-\eta i-(\alpha+\alpha')j-2\beta h},$$

where $c = c_1(r) \cdot c_2$. Summing all the bounds on the addends we obtain

$$|S_{ij}| \leqslant \frac{c}{1-e^{-2\beta}} e^{-\eta i-(\alpha+\alpha')j}.$$

Again, we can estimate the l-th singular value by setting the first $l-1$ columns of S to zero. Let S_l be the matrix composed by the last $m-l+1$ rows of S. In particular, the entries of S_l satisfy the relation $(S_l)_{ij} \leqslant \tilde{\gamma} e^{-(\alpha+\alpha')l} e^{-\eta i-(\alpha+\alpha')(j-1))}$ where $\tilde{\gamma} = \frac{c}{1-e^{-2\beta}}$. Therefore:

$$\left\| \frac{e^{(\alpha+\alpha')l}}{\tilde{\gamma}} S_l \right\|_F^2 = \sum_{i=1}^{m-l}\sum_{j=1}^{n} |\frac{e^{(\alpha+\alpha')l}}{\tilde{\gamma}}(S_k)_{i,j}|^2 \leqslant \frac{e^{-2\eta}}{(1-e^{-2\eta})(1-e^{(-\alpha+\alpha')})}.$$

Since $\|S_l\|_2 \leqslant \|S_l\|_F$ we have $\sigma_l(S) \leqslant \frac{\tilde{\gamma}e^{-\eta}}{\sqrt{(1-e^{-2\eta})(1-e^{-2(\alpha+\alpha')})}} e^{-(\alpha+\alpha')l} = \gamma e^{-(\alpha+\alpha')l}$. □

For simplicity we assumed the matrices to be diagonalizable, but we highlight that it is easy to recover analogous estimates for the general framework employing the techniques of Section 4.5.2.

4.6 SINGULAR VALUES AND DISPLACEMENT RANK

In this section, we explore the possibility of using rational interpolation techniques in order to estimate the singular values of matrices with particular algebraic properties. We rely on the concept of displacement rank and on some result by B. Beckermann [7], of which we report the proof.

Definition 4.6.1. *Given matrices $A, B, X \in \mathbb{C}^{m\times m}$ the* displacement rank *of X with respect to the pair (A,B) is defined as*

$$\rho_{A,B}(X) = \mathit{rank}(AX - XB).$$

Example 4.6.2. *Consider the Krylov matrix $X = \mathcal{KM}_m(A, b)$ with $A \in \mathbb{C}^{m\times m}$ and the shift operator*

$$\Pi = \begin{bmatrix} 0 & & & 1 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}.$$

With a direct computation we can verify that

$$AX - X\Pi = (A^m b - b) \cdot \begin{bmatrix} 0 & \dots & 0 & 1 \end{bmatrix} \quad \Longrightarrow \quad \rho_{A,\Pi}(X) = 1.$$

Even the outer product between two Krylov matrices has a small displacement rank. Consider for example $X = \mathcal{KM}_m(A, a) \cdot \mathcal{KM}_m(B, b)^*$ with A and B square matrices and B invertible. Then, the following relation holds

$$AX - XB^{-*} = A^m a(B^{m-1}b)^* - a(B^{-1}b)^* \quad \Longrightarrow \quad \rho_{A,B^{-*}}(X) = 2.$$

Observe that we do not really need the matrix B to be invertible. It is sufficient to write the Moore-Penrose pseudoinverse $B^\dagger$ in place of B^{-1} in the displacement equation.

We need also to introduce the set $\mathcal{R}_{n,d}$ of rational functions over $\mathbb{C}$ where n and d are the degree of the numerator and of the denominator, respectively.

For a matrix X with a small displacement rank it is possible to provide bounds on its singular values in terms of the optimal values of some *Zolotarev* problems [101] according to the following result of B. Beckermann [7].

Theorem 4.6.3. *Let $X \in \mathbb{C}^{m\times m}$ and suppose that there exist two normal matrices $A, B \in \mathbb{C}^{m\times m}$ such that $\rho_{A,B}(X) = d$. Then, indicating with E and F the spectrum of A and B respectively, for the singular values of X it holds:*

$$\frac{\sigma_{1+l\cdot d}(X)}{\|X\|_2} \leqslant Z_l(E, F) := \inf_{r(x)\in\mathcal{R}_{l,l}} \frac{\max_{x\in E} |r(x)|}{\min_{x\in F} |r(x)|}, \quad l = 1, 2, \dots .$$

Proof. Consider $p(x) := \sum_{i=0}^{l} p_i x^i$ and $q(x) := \sum_{i=0}^{l} q_i x^i$ polynomials of degree l and define $r(x) := \frac{p(x)}{q(x)}$. We prove that the matrix

$$\Delta := q(A)Xp(B) - p(A)Xq(B)$$

has rank at most $l \cdot d$. Without loss of generality we consider the case $d = 1$ and suppose $AX - XB = uv^*$. We can prove by induction that $A^kX - XB^k = \sum_{h=0}^{k-1} A^h uv^* B^{k-1-h}$. For $k = 1$ the property trivially holds. For $k > 1$ one has:

$$\begin{aligned} A^kX = A^{k-1}XB + A^{k-1}uv^* &= XB^k + \left(\sum_{h=0}^{k-2} A^h uv^* B^{k-1-h}\right) B + A^{k-1}uv^* \\ &= XB^k + \sum_{h=0}^{k-1} A^h uv^* B^{k-1-h}. \end{aligned}$$

Now, observe that

$$\Delta = q(A)Xp(B) - p(A)Xq(B) = \sum_{i \neq j}^{d} (q_i p_j - q_j p_i)(A^i X B^j - A^j X B^i),$$

and if $i > j$ (the other case is analogous)

$$A^i X B^j - A^j X B^i = A^j (A^{i-j} X - X B^{i-j}) B^j = A^j \left(\sum_{h=0}^{i-j-1} A^h u v^* B^{k-1-h} \right) B^j.$$

In particular all the addends involved in the expansion of Δ can be expressed as sum of dyads whose left vectors belong to the Krylov space $\mathcal{K}_l(A, u)$ and so l is an upper bound for the rank of Δ.

Assume that $q(A)$ and $p(B)$ are invertible, define $Y := q(A)^{-1} \Delta p(B)^{-1}$ and observe that $X - Y = r(A) X r(B)^{-1}$. In particular

$$\|X - Y\|_2 = \|r(A) X r(B)^{-1}\|_2 \leqslant \|X\|_2 \max_E |r(x)| \max_F |r(x)|^{-1} = \|X\|_2 \frac{\max_E |r(x)|}{\min_F |r(x)|}.$$

Since $\sigma_{k+1}(X)$ coincides with the minimum of $\|X - W\|_2$ taken over all the matrices W of rank k, and since $\operatorname{rank}(Y) = \operatorname{rank}(\Delta) \leqslant l \cdot d$, we find that

$$\sigma_{l \cdot d + 1}(X) \leqslant \|X - Y\|_2 \leqslant \|X\|_2 \frac{\max_E |r(x)|}{\min_F |r(x)|}.$$

Taking the infimum over the set of rational functions of degree (d, d) completes the proof. □

The normality hypothesis can be relaxed by replacing it with the diagonalizability of A and B. The price to pay is a larger constant depending on the conditioning of the eigenvector matrices as stated by the following.

Corollary 4.6.4. *Let $X \in \mathbb{C}^{m \times m}$ and suppose that there exist two diagonalizable matrices $A, B \in \mathbb{C}^{m \times m}$ such that $\Delta_{A,B}(X) = d$, that is $A = V_A D_A V_A^{-1}$, $B = V_B D_B V_B^{-1}$ with D_A and D_B diagonal matrices. Then, indicating with E and F the spectrum of A and B respectively, it holds:*

$$\sigma_{1 + l \cdot d}(X) \leqslant Z_l(E, F) \cdot \|X\|_2 \cdot \kappa(V_A) \cdot \kappa(V_B)$$

where $\kappa(W) = \|W\|_2 \|W^{-1}\|_2$ denotes the condition number of W.

The bounds provided by Theorem 4.6.3 and Corollary 4.6.4 are informative if and only if the two sets in the Zolotarev problem are disjoint. We are going to use these tools in Chapter 6 where E and F are contained in the unit disc and in its complementary, respectively.

The case where E and F are disjoint subsets of the real line, has been extensively studied by Zolotarev [101] and is one of the few cases in which explicit estimates for $Z_l(E,F)$ have been found. The result we are going to quote is adapted to the settings of Chapter 6 and can be found in [52]. See also [1, 92, 77] for classical references.

Theorem 4.6.5 (Zolotarev)**.** *Let $\delta \in (0,1)$, $E := [-\infty, -\delta^{-1}] \cup [\delta^{-1}, +\infty]$ and $F = [-\delta, \delta]$. Then*

$$Z_{2l}(E,F) \leqslant \frac{2\rho^l}{1-2\rho^l},$$

where

$$\rho := \exp\left(-\frac{\pi K(\sqrt{1-\delta^4})}{2K(\delta^2)}\right), \qquad K(x) := \int_0^1 \frac{1}{\sqrt{(1-t^2)(1-x^2t^2)}}dt.$$

Moreover, if $\delta \approx 1$ then $K(\delta^2) \approx \log\left(\frac{4}{\sqrt{1-\delta^4}}\right)$ and $K(\sqrt{1-\delta^4}) \approx \frac{\pi}{2}$, yielding

$$Z_{2l}(E,F) \leqslant \frac{2\rho^l}{1-2\rho^l} \approx \frac{2\widetilde{\rho}^l}{1-2\widetilde{\rho}^l}, \qquad \widetilde{\rho} := \exp\left(-\frac{\pi^2}{2\log\left(\frac{16}{1-\delta^4}\right)}\right).$$

Chapter 5

Numerical quasiseparable preservation in matrix functions

Matrix functions are an evergreen topic in matrix algebra due to their diffusion in applications [45, 69, 62, 59, 49]. In the latter we often have to deal with structured matrices which can be exploited for speeding up algorithms to reduce storage costs. Then, it is not hard to imagine why the interaction of structure with matrix functions is an intriguing subject.

Studies concerning the numerical preservation of data-sparse patterns were carried out in some recent papers [10, 9, 32, 11, 85]. Regarding the quasiseparable structure, in [46, 47, 54] Gavrilyuk, Hackbusch and Khoromskij addressed the issue of approximating some matrix functions using the hierarchical format [26]. In these works the authors prove that —given a quasiseparable matrix A and a holomorphic function $f(z)$— computing $f(A)$ via a quadrature formula applied to the contour integral definition, yields an approximation of the result with a low quasiseparable rank. Employing the HODLR arithmetic in this procedure provides an algorithm for approximating $f(A)$ with almost linear complexity. The feasibility of this approach is equivalent to the existence of a rational function $r(z) = \frac{p(z)}{q(z)}$ which well-approximates the holomorphic function $f(z)$ on the spectrum of the argument A. More precisely, since the quasiseparable rank is invariant under inversion and sub-additive with respect to matrix addition and multiplication, if $r(z)$ is a good approximation of $f(z)$ with low degree then the matrix $r(A)$ is an accurate approximation of $f(A)$ with low quasiseparable rank. For example, in [68] the authors show the quasiseparable preservation when computing spectral projectors of Hermitian matrices exploiting the best rational approximant of the sign function. This argument explains the preservation of the structure, but still needs a deeper analysis if one wants to provide estimates of the quasiseparable rank of $f(A)$, for a general f.

In this chapter we deal with this issue by studying the interplay between the off-diagonal singular values of the matrices A and B such that $B = f(A)$. Our intent is to

understand which parameters of the model come into play in the numerical preservation of the structure and to extend the analysis to functions with singularities.

In Section 5.2 we see how the integral definition of a matrix function enables us to study the structure of the off-diagonal blocks in $f(A)$. In Section 5.2.2 we use the tool developed in Section 4.4 to deriving bounds for the off-diagonal singular values of matrix functions.

In Section 5.3 we adapt the approach to treat functions with singularities.

The key role is played by Theorem 5.3.1 which extends the Dunford-Cauchy formula to the case of some singularities inside the contour of integration. In Section 5.4 we comment on computational aspects and we perform some experiments for validating the theoretical results. Finally, in Section 5.5 we give some concluding remarks.

5.1 DEFINITIONS OF MATRIX FUNCTION

In [59] —which we indicate as a reference for this topic— the author focuses on three equivalent definitions of matrix function. For our purposes we recall only two of them: one based on the Jordan canonical form of the argument and the other which is a generalization of the Cauchy integral formula.

Definition 5.1.1. *Let $A \in \mathbb{C}^{m\times m}$ and $f(z)$ be a function holomorphic in a set containing the spectrum of A. Indicating with $J = diag(J_1, \ldots, J_p) = V^{-1}AV$ the Jordan canonical form of A, we define $f(A) := V \cdot f(J) \cdot V^{-1} = V \cdot \mathrm{diag}(f(J_k)) \cdot V^{-1}$ where J_k is a $m_k \times m_k$ Jordan block and*

$$J_k = \begin{bmatrix} \lambda_k & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{bmatrix}, \quad f(J_k) = \begin{bmatrix} f(\lambda_k) & f'(\lambda_k) & \ldots & \frac{f^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ & \ddots & \ddots & \vdots \\ & & \ddots & f'(\lambda_k) \\ & & & f(\lambda_k) \end{bmatrix}.$$

Definition 5.1.2 (Dunford-Cauchy integral formula)**.** *Let $f(z)$ be a holomorphic function in $\mathcal{D} \subseteq \mathbb{C}$ and $A \in \mathbb{C}^{m\times m}$ be a matrix whose spectrum is contained in $\Omega \subset \mathcal{D}$. Then we define*

$$f(A) := \frac{1}{2\pi \boldsymbol{i}} \int_{\partial\Omega} (zI - A)^{-1} f(z) dz. \tag{5.1}$$

The matrix-valued function $\mathfrak{R}(z) := (zI - A)^{-1}$ is called resolvent.

Suppose that the spectrum of A is contained in a disc $\Omega = B(z_0, r) := \{|z - z_0| < r\}$ where the function is holomorphic. Then, it is possible to write $f(A)$ as an integral (5.1) along $\mathbb{T} := \partial B(0, 1)$ for a matrix with spectral radius less than 1. In fact,

$$\frac{1}{2\pi \boldsymbol{i}} \int_{\{|z-z_0|=r\}} (zI - A)^{-1} f(z) dz \quad = \quad \frac{1}{2\pi \boldsymbol{i}} \int_{\mathbb{T}} (wI - \tilde{A})^{-1} f(rw + z_0) dw$$

where $\tilde{A} = r^{-1}(A - z_0 I)$ has the spectrum contained in $B(0,1)$. Given the above remark it is not restrictive to consider only the case of A having spectral radius less than 1.

Remark 5.1.3. *In the following we will often require, besides the non singularity of $(zI - A)$, also that $(zI - D)$ is invertible along the path of integration for any trailing diagonal block D. This is not restrictive because the set of points that make at least one of these matrices singular is finite. Moreover —given a sufficiently large domain of analyticity for f— one can choose r large enough to guarantee this property. As an example, any r such that $r \geqslant \|A\|$ is a valid choice for any induced norm.*

5.2 OFF-DIAGONAL ANALYSIS OF $f(A)$

The aim of this section is characterizing the structure of the off-diagonal blocks by means of the integral definition of $f(A)$.

5.2.1 *Structure of an off-diagonal block*

Consider the Dunford-Cauchy integral formula (5.1) in the case $\partial\Omega = \mathbb{T}$ and A with the spectrum strictly contained in the unit disc. In this case the spectral radius of A is less than 1 and we can expand the resolvent as $\mathfrak{R}(z) = (zI - A)^{-1} = \sum_{n\geqslant 0} z^{-(n+1)} A^n$.

Applying component-wise the residue theorem we find that the result of the integral in (5.1) coincides with the coefficient of degree -1 in the Laurent expansion of $(zI - A)^{-1} f(z)$. Thus, examining the Laurent expansion of an off-diagonal block, we can derive a formula for the corresponding block in $f(A)$. Partitioning A as follows

$$A = \begin{bmatrix} \bar{A} & \bar{B} \\ \bar{C} & \bar{D} \end{bmatrix} \quad \Rightarrow \quad \mathfrak{R}(z) = \begin{bmatrix} zI - \bar{A} & -\bar{B} \\ -\bar{C} & zI - \bar{D} \end{bmatrix}^{-1}$$

and supposing that the spectral radius of $\bar{D}$ is less than 1 (which is not restrictive thanks to Remark 5.1.3) we get

$$\mathfrak{R}(z) = \begin{bmatrix} S_{zI-\bar{D}}^{-1} & * \\ (zI - \bar{D})^{-1}\bar{C} S_{zI-\bar{D}}^{-1} & * \end{bmatrix},$$

where $S_{zI-\bar{D}} = zI - \bar{A} - \bar{B}(zI - \bar{D})^{-1}\bar{C}$ is the Schur complement of the bottom right block and $*$ denotes blocks which are not relevant for our analysis. We can write the Laurent expansion of the two inverse matrices:

$$(zI - \bar{D})^{-1} = \sum_{j\geqslant 0} z^{-(j+1)} \bar{D}^j, \qquad S_{zI-\bar{D}}^{-1} = \begin{bmatrix} I & 0 \end{bmatrix} \cdot \left(\sum_{j\geqslant 0} z^{-(j+1)} A^j \right) \cdot \begin{bmatrix} I \\ 0 \end{bmatrix},$$

where for deriving the expansion of $S_{zI-\bar{D}}^{-1}$ we used the fact that it corresponds to the upper left block in $\mathfrak{R}(z)$.

Let $f(z) = \sum_{n\geqslant 0} a_n z^n$ be the Laurent expansion of f in $\mathbb{T}$ and let $\mathfrak{R}(z)\cdot f(z) := \begin{bmatrix} * & * \\ G(z) & * \end{bmatrix}$, then

$$G(z) = \sum_{n\geqslant 0} a_n \sum_{j\geqslant 0} \bar{D}^j \bar{C}\cdot [I \quad 0]\cdot \sum_{s\geqslant 0} A^s z^{n-j-s-2}\cdot [I \quad 0]^t. \tag{5.2}$$

Exploiting this relation we can prove the following.

Lemma 5.2.1. *Let* $A = \begin{bmatrix} \bar{A} & \bar{B} \\ \bar{C} & \bar{D} \end{bmatrix}$ *be a square matrix with square diagonal blocks,* $\bar{C} = uv^*$ *and suppose that the spectrum of* A *and* $\bar{D}$ *is contained in* $B(0,1)$*. Consider* $f(z) = \sum_{n\geqslant 0} a_n z^n$ *for* $|z|\leqslant 1$ *and let* $f(A) = \begin{bmatrix} * & * \\ \tilde{C} & * \end{bmatrix}$ *be partitioned according to* A*. Then*

$$\tilde{C} = \sum_{n\geqslant 1} a_n \left[u \,\middle|\, \bar{D}\cdot u \,\middle|\, \dots \,\middle|\, \bar{D}^{n-1}\cdot u\right]\cdot \left[(A^*)^{n-1}\tilde{v} \,\middle|\, \dots \,\middle|\, A^*\tilde{v} \,\middle|\, \tilde{v}\right]^* [I\ 0]^t$$

with $\tilde{v} = [I\ 0]^t v$*.*

Proof. By the Dunford-Cauchy formula, the subdiagonal block $\tilde{C}$ is equal to $\int_{\mathbb{T}} G(z)dz$. By means of the residue theorem we can write the latter as the coefficient of degree -1 in (5.2), that is

$$\tilde{C} = \sum_{n\geqslant 1} a_n \sum_{j=0}^{n-1} \bar{D}^j uv^*\cdot [I\ 0] A^{n-j-1} [I\ 0]^t = \sum_{n\geqslant 1} a_n \sum_{j=0}^{n-1} \bar{D}^j u\tilde{v}^* A^{n-j-1} [I\ 0]^t,$$

which is in the sought form. □

Remark 5.2.2. *The expression that we obtained for* $\tilde{C}$ *in the previous Lemma is a sum of outer products of vectors of the form* $\bar{D}^j u$ *with* $(A^*)^{n-j-1}\tilde{v}$*, where the spectral radii of* A *and* $\bar{D}$ *are both less than* 1*. This implies that the addends become negligible for a sufficiently large* n*. So, in order to derive bounds for the singular values, we will focus on the truncated sum*

$$\sum_{n=1}^{s} a_n \left[u \,\middle|\, \bar{D}\cdot u \,\middle|\, \dots \,\middle|\, \bar{D}^{n-1}\cdot u\right]\cdot \left[(A^*)^{n-1}\tilde{v} \,\middle|\, \dots \,\middle|\, A^*\tilde{v} \,\middle|\, \tilde{v}\right]^* [I\ 0]^t \tag{5.3}$$

which can be rewritten as:

$$\left[u \,\middle|\, \bar{D}\cdot u \,\middle|\, \dots \,\middle|\, \bar{D}^{s-1}\cdot u\right]\cdot \left[\sum_{n=0}^{s-1} a_{n+1}(A^*)^n\tilde{v} \,\middle|\, \dots \,\middle|\, (a_s A^* + a_{s-1}I)\tilde{v} \,\middle|\, a_s\tilde{v}\right]^* [I\ 0]^t. \tag{5.4}$$

Using the notation introduced in Section 4.5 we can rewrite (5.4) *as*

$$\mathcal{KM}_s(\bar{D}, u)\cdot J_s \cdot \mathcal{HM}_p(A^*, \tilde{v})^*\cdot [I\ 0]^t,$$

where $p(x) = \sum_{i=0}^{s-1} a_{i+1} z^i$*.*

5.2.2 *Decay in the off-diagonal singular values of $f(A)$*

We are ready to study the off-diagonal singular values in function of matrices using the results of Section 4.4.

We prefer to begin by stating a simpler result which holds for matrices with spectrum contained in $B(0,1)$ and function holomorphic on a larger disc. In the following corollaries it is shown how to adapt this result to more general settings.

Theorem 5.2.3. *Let $A \in \mathbb{C}^{m\times m}$ be quasiseparable of rank k and such that A and all its trailing submatrices are enclosed in $(\rho, R_A, \mathcal{V}_A)$ and diagonalizable. Consider $f(z)$ holomorphic on $\overline{B(0,R)}$ with $R > 1$. Then, we can bound the singular values of a generic off-diagonal block $\tilde{C}$ in $f(A)$ with*

$$\sigma_l(\tilde{C}) \leqslant \gamma e^{-\frac{(\alpha+\alpha')l}{k}}, \quad \alpha = \log\left(\frac{R_A}{\rho}\right), \quad \alpha' = \log(R),$$

where $\gamma := \max_{|z|=R} |f(z)| \cdot \kappa_{max}^2 \cdot \|A\|_2 \cdot \Lambda(\rho, R_A, \mathcal{V}_A, R) \cdot \frac{k\cdot\rho}{RR_A-\rho}$ and κ_{max} is the maximum among the spectral condition numbers of the trailing submatrices of A.

Proof. Consider the partitioning $A = \left[\begin{smallmatrix} \bar{A} & \bar{B} \\ \bar{C} & \bar{D} \end{smallmatrix}\right]$ and for simplicity the case $k = 1$, $\bar{C} = uv^*$. The general case is obtained by linearity summing k objects of this kind coming from the SVD of $\bar{C}$ and applying Lemma 4.3.3. We rewrite the Dunford-Cauchy formula for $f(A)$

$$f(A) = \frac{1}{2\pi \boldsymbol{i}} \int_{\mathbb{T}} (zI - A)^{-1} f(z) dz.$$

Let $f(z) = \sum_{n\geqslant 0} a_n z^n$ be the Taylor expansion of $f(z)$ in $B(0,R)$. The corresponding off-diagonal block $\tilde{C}$ in $f(A)$ can be written as the outer product in Remark 5.2.2

$$\mathcal{KM}_s(\bar{D}, u) \cdot J_s \cdot \mathcal{HM}_p(A^*, \tilde{v})^* \cdot [I\ 0]^t + g_s(A), \tag{5.5}$$

where $\bar{v} = [I\ 0]^t v$ and $g_s(A)$ is the remainder of the truncated Taylor series at order s. Since $f(z)$ is holomorphic in $\overline{B(0,R)}$, Theorem 1.4.1 ensures that

$$|a_j| \leqslant \max_{|z|=R} |f(z)| \cdot R^{-j}.$$

Applying Theorem 4.5.18 we get that $\forall r \in (\rho, R_A)$

$$\sigma_l(\tilde{C} - g_s(A)) \leqslant \gamma e^{-(\alpha+\alpha')l},$$

with $\alpha, \alpha', \delta, \kappa_{max}$ as in the thesis and $\gamma = \max_{|z|=R} |f(z)| \cdot \kappa_{max}^2 \|A\|_2 \cdot \Lambda(\rho, R_A, \mathcal{V}_A)$. Observing that this bound is independent on s and $\lim_{s\to\infty} g_s(A) = 0$, we get the thesis. □

Corollary 5.2.4. *Let $A \in \mathbb{C}^{m\times m}$ be a k-quasiseparable matrix, $z_0 \in \mathbb{C}$ and $R' \in \mathbb{R}^+$ such that $R'^{-1}(A - z_0 I)$ is enclosed in $(\rho, R_A, \mathcal{V}_A)$. Then, for any holomorphic function $f(z)$ in $\overline{B(z_0, R)}$ with $R > R'$, any off-diagonal block $\tilde{C}$ in $f(A)$ has singular values bounded by*

$$\sigma_l(\tilde{C}) \leqslant \gamma e^{-\frac{(\alpha+\alpha')l}{k}}, \quad \alpha = \log\left(\frac{R_A}{\rho}\right), \quad \alpha' = \log\left(\frac{R}{R'}\right),$$

where $\gamma := \max_{|z-z_0|=R} |f(z)| \cdot \kappa_{max}^2 \cdot \|A - z_0 I\|_2 \cdot \Lambda(\rho, R_A, \mathcal{V}_A, R) \cdot \frac{k\cdot\rho}{RR_A - \rho R'}$ and κ_{max} is the maximum among the spectral condition numbers of the trailing submatrices of $R'^{-1}(A - z_0 I)$.

Proof. Define $g(z) = f(R'z + z_0)$ which is holomorphic on $B(0, \frac{R}{R'})$. Observing that $f(A) = g(R'^{-1}(A - z_0 I))$ we can conclude by applying Theorem 5.2.3. □

Remark 5.2.5. *If we can find $z_0 \in \mathbb{C}$ such that $\|A - z_0 I\|_2 < R$ then it is always possible to find $(\rho, R_A, \mathcal{V}_A)$ with $\rho \in (0, 1)$ which satisfies the hypothesis of the previous corollary. A worst case estimate for $\frac{\rho}{R_A}$ is $\frac{\|A - z_0 I\|_2}{R}$ since this is the radius of a circle containing the spectrum of the rescaled matrix and — given that the Riemann map for a ball centered in 0 is the identity — $R_A = 1$.*

Example 5.2.6 (Real spectrum)**.** *Here, we want to estimate the quantity $\frac{R_A}{\rho}$ in the case of a real spectrum for the matrix A. Suppose that — possibly after a scaling — the latter is contained in the symmetric interval $[-a, a]$ with $a \in (0, 1)$. The logarithmic capacity of this set is $\frac{a}{2}$ and the inverse of the associated Riemann map is $\psi(z) = z + \frac{a^2}{4}$. This follows by observing that the function $z + z^{-1}$ maps the circle of radius 1 into $[-2, 2]$, so it is sufficient to compose the latter with two homothetic transformations to get $\psi(z)$. Moreover, observe that — given $r \geqslant \frac{a}{2}$ — the function ψ maps the circle of radius r into an ellipse of foci $[-a, a]$. Therefore, in order to get R_A it is sufficient to compute for which r we have $\psi(r) = 1$. This corresponds to find the solution of $r + \frac{a^2}{4r} = 1$ which is greater than $\frac{a}{2}$. This yields*

$$R_A = \frac{1 + \sqrt{1 - a^2}}{2} \quad \Rightarrow \quad \frac{R_A}{\rho} = \frac{1 + \sqrt{1 - a^2}}{a}.$$

5.3 FUNCTIONS WITH SINGULARITIES

If some singularities of f lie inside $B(z_0, R)$ then $f(A) \neq \frac{1}{2\pi i} \int_{\partial B(z_0,R)} f(z)(zI - A)^{-1} dz$. However, since the coefficients of the Laurent expansion of f with negative degrees in (5.2) do not affect the result, the statement of Theorem 5.2.3 holds for the matrix $\frac{1}{2\pi i} \int_{\partial B(z_0,R)} f(z)(zI - A)^{-1} dz$. In this section we prove that — under mild conditions — the difference of the above two terms still has a quasiseparable structure. This numerically preserves the quasiseparability of $f(A)$.

5.3.1 *An extension of the Dunford-Cauchy integral formula*

The main tool we are going to use to overcome the difficulties in case of singularities is the following result, which extends the integral formula in Definition 5.1.1.

Theorem 5.3.1. *Let $f(z)$ be a meromorphic function with a discrete set of poles $\mathcal{P}$ and $A \in \mathbb{C}^{m\times m}$ with spectrum $\mathcal{S}$ such that $\mathcal{S} \cap \mathcal{P} = \emptyset$. Moreover, consider Γ simple closed curve in the complex plane which encloses $\mathcal{S}$ and $T := \{z_1, \dots, z_t\} \subseteq \mathcal{P}$ subset of poles with orders $d_1, \dots, d_t$, respectively. Then*

$$\frac{1}{2\pi \boldsymbol{i}} \int_\Gamma (zI - A)^{-1} f(z) dz = f(A) + \sum_{j=1}^{t} R_j(z_j I - A),$$

where R_j is the rational function

$$R_j(z) := \sum_{l=1}^{d_j} (-1)^{l+1} \frac{f_j^{(d_j-l)}(z_j)}{(d_j - l)!} z^{-l}$$

and $f_j(z) = (z - z_j)^{d_j} f(z)$, extended to the limit in z_j. In particular if the poles in T are simple then

$$\frac{1}{2\pi \boldsymbol{i}} \int_\Gamma (zI - A)^{-1} f(z) dz = f(A) + \sum_{j=1}^{t} f_j(z_j) \cdot (z_j I - A)^{-1} = f(A) + \sum_{j=1}^{t} f_j(z_j) \mathfrak{R}(z_j).$$

Proof. We first prove the statement for A diagonalizable. Assume that $V^{-1}AV = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, then

$$\frac{1}{2\pi \boldsymbol{i}} \int_\Gamma (zI - A)^{-1} f(z) dz = V^{-1} \begin{bmatrix} \frac{1}{2\pi \boldsymbol{i}} \int_\Gamma \frac{f(z)}{z-\lambda_1} & & \\ & \ddots & \\ & & \frac{1}{2\pi \boldsymbol{i}} \int_\Gamma \frac{f(z)}{z-\lambda_m} \end{bmatrix} V. \tag{5.6}$$

Applying the Residue theorem we arrive at

$$\frac{1}{2\pi \boldsymbol{i}} \int_\Gamma \frac{f(z)}{z - \lambda_p} = \operatorname{Res}\left(\frac{f}{z-\lambda_p}, \lambda_p\right) + \sum_{j=1}^{t} \operatorname{Res}\left(\frac{f}{z-\lambda_p}, z_j\right), \qquad p = 1, \dots, m.$$

Since λ_p is a simple pole of $\frac{f}{z-\lambda_p}$ the first summand is equal to $f(\lambda_p)$.

On the other hand z_j is a pole of order d_j of $\frac{f}{z-\lambda_p}$, therefore its residue is

$$\begin{aligned} \operatorname{Res}\left(\frac{f}{z-\lambda_p}, z_j\right) &= \frac{1}{(d_j-1)!} \lim_{z\to z_j} \frac{\partial^{d_j-1}}{\partial z^{d_j-1}} \left((z-z_j)^{d_j} \frac{f}{z-\lambda_p}\right) \\ &= \frac{1}{(d_j-1)!} \frac{\partial^{d_j-1}}{\partial z^{d_j-1}} \left(\frac{f_j}{z-\lambda_p}\right)(z_j). \end{aligned}$$

One can prove by induction (see Appendix A, Proposition A.0.1) that, given a sufficiently differentiable $f_j(z)$, it holds

$$\frac{\partial^{d-1}}{\partial z^{d-1}}\left(\frac{f_j(z)}{z-\lambda_p}\right)=\sum_{l=1}^{d}(-1)^{l+1}\frac{(d-1)!}{(d-l)!}f_j^{(d-l)}(z)(z-\lambda_p)^{-l},\quad d\in\mathbb{Z}^+. \tag{5.7}$$

Setting $d=d_j$ in (5.7) we derive

$$\operatorname{Res}\left(\frac{f}{z-\lambda_p},z_j\right)=R_j(z_j-\lambda_p).$$

To conclude it is sufficient to rewrite the diagonal matrix in (5.6) as

$$\begin{bmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_m)\end{bmatrix}+\sum_{j=1}^{t}\begin{bmatrix} R_j(z_j-\lambda_1) & & \\ & \ddots & \\ & & R_j(z_j-\lambda_m)\end{bmatrix}.$$

We now prove the thesis for

$$A=\begin{bmatrix}\lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1\\ & & & \lambda\end{bmatrix},$$

because the general non diagonalizable case can be decomposed in sub-problems of that kind. We have that

$$\frac{1}{2\pi\boldsymbol{i}}\int_\Gamma(zI-A)^{-1}f(z)dz=\frac{1}{2\pi\boldsymbol{i}}\begin{bmatrix}\int_\Gamma\frac{f(z)}{z-\lambda} & \int_\Gamma\frac{f(z)}{(z-\lambda)^2} & \cdots & \int_\Gamma\frac{f(z)}{(z-\lambda)^m}\\ & \ddots & \ddots & \vdots\\ & & \ddots & \int_\Gamma\frac{f(z)}{(z-\lambda)^2}\\ & & & \int_\Gamma\frac{f(z)}{z-\lambda}\end{bmatrix}.$$

In order to reapply the previous argument it is sufficient to prove that

(i) $\operatorname{Res}(\frac{f}{(z-\lambda)^{h+1}},\lambda)=\frac{f_j^{(h)}(\lambda)}{h!}\ h=1,\dots,m-1,$

(ii) $\operatorname{Res}(\frac{f}{(z-\lambda)^{h+1}},z_j)=\frac{R_j^{(h)}(z_j-\lambda)}{h!}\ h=1,\dots,m-1.$

The point (i) is a direct consequence of the fact that λ is a pole of order $h+1$ of the function $\frac{f(z)}{(z-\lambda)^{h+1}}$. Concerning (ii) observe that z_j is again a pole of order d_j for the function $\frac{f(z)}{(z-\lambda)^{h+1}}$ so

$$\operatorname{Res}\left(\frac{f}{(z-\lambda)^{h+1}},z_j\right)=\frac{1}{(d_j-1)!}\frac{\partial^{d_j-1}}{\partial z^{d_j-1}}\left(\frac{f_j(z)}{(z-\lambda)^{h+1}}\right)(z_j).$$

One can prove by induction (see Appendix A, Proposition A.0.1) that, for each $d \in \mathbb{Z}^+$, $h \in \mathbb{N}$:

$$\frac{\partial^{d-1}}{\partial z^{d-1}}\left(\frac{f_j(z)}{(z-\lambda)^{h+1}}\right) = \frac{(d-1)!}{h!}\sum_{l=1}^{d}(-1)^{l+h+1}\frac{(l+h-1)!}{(d-l)!(l-1)!}f_j^{(d-l)}(z)(z-\lambda)^{-(h+l)}. \tag{5.8}$$

Successive derivation of R_j repeated h times yields:

$$R_j^{(h)}(z) = \sum_{l=1}^{d_j}(-1)^{l+h+1}\frac{(l+h-1)!}{(d_j-l)!(l-1)!}f_j^{(d_j-l)}(z_j)z^{-(h+l)},$$

and by setting $d = d_j$ in (5.8) we finally get (ii). □

5.3.2 *Functions with poles*

Using Theorem 5.3.1 we can extend Corollary 5.2.4, giving a concise statement in the case of simple poles.

Corollary 5.3.2. *Let $A \in \mathbb{C}^{m\times m}$ be a quasiseparable matrix with rank k, $z_0 \in \mathbb{C}$ and $R' \in \mathbb{R}^+$ such that $R'^{-1}(A - z_0 I)$ is enclosed in $(\rho, R_A, \mathcal{V}_A)$. Consider $R > R'$ and a function $f(z)$ holomorphic on the annulus $\overline{\mathbb{A}(R', R)}$. If the ball $B(z_0, R')$ contains t simple poles of f then any off-diagonal block $\tilde{C}$ in $f(A)$ has singular values bounded by*

$$\sigma_l(\tilde{C}) \leqslant \gamma e^{-\frac{(\alpha+\alpha')(l-tk)}{k}}, \quad \alpha = \log\left(\frac{R_A}{\rho}\right), \quad \alpha' = \log\left(\frac{R}{R'}\right),$$

where $\gamma := \max_{|z-z_0|=R}|f(z)| \cdot \kappa_{max}^2 \cdot \|A - z_0 I\|_2 \cdot \Lambda(\rho, R_A, \mathcal{V}_A, R) \cdot \frac{k\cdot\rho}{RR_A - \rho R'}$ and κ_{max} is the maximum among the spectral condition numbers of the trailing submatrices of $R'^{-1}(A - z_0 I)$.

Proof. Let $f(z) = \sum_{n\in\mathbb{Z}} a_n z^n$ be the series expansion of f in $\mathcal{A}$ and $z_1, \ldots, z_t$ be the simple poles of f inside $B(z_0, R')$. Theorem 1.4.1 implies that

$$|a_j| \leqslant \|f(z)\|_{\infty,\partial B(z_0,R)} \cdot \left(\frac{R'}{R}\right)^j, \qquad n \geqslant 0.$$

According to what we observed at the beginning of Section 5.3 we can apply Corollary 5.2.4 to the off-diagonal singular values of $B := \int_{\partial B(z_0,R')} f(z)(zI - A)^{-1}dz$. Moreover, using Theorem 5.3.1 we get

$$f(A) = B - \sum_{j=1}^{t} f_j(z_j) \cdot (z_j I - A)^{-1}.$$

Observing that the right summand has at most quasiseparable rank tk we can conclude, using Lemma 4.3.4, that the bound on the singular values of $f(A)$ is the same which holds for B, but shifted by the quantity $t \cdot k$. □

5.3.3 *Functions with essential singularities*

Consider the case of a function $f(z)$ holomorphic in $\mathbb{C} \setminus \{a\}$ with an essential singularity in a. Moreover, suppose that a is not an eigenvalue of the argument $A \in \mathbb{C}^{m\times m}$. In a suited punctured disc $B(a, R) \setminus \{a\}$ — which contains the spectrum of A — we can expand f as

$$f(z) := \sum_{n\in\mathbb{Z}} a_n (z-a)^n.$$

In particular we can decompose f as $f_1(z-a) + f_2((z-a)^{-1})$ with f_i holomorphic on $B(0, R)$ for $i = 1, 2$. Therefore

$$f(A) = f_1(A - aI) + f_2((A - aI)^{-1}).$$

Since f_1 and f_2 are both holomorphic and the operations of shift and inversion preserve the quasiseparable rank we can apply Theorem 5.2.3 for f_1 and f_2 computed on different arguments. Finally, use Lemma 4.3.3 to get estimates on the off-diagonal singular values of $f(A)$.

One can use this approach in the case of finite order poles and find equivalent bounds to Corollary 5.3.2, although in a less explicit form.

5.3.4 *Functions with branches*

We conclude this section describing how to readapt the approach in the case of functions with multiple branches. The same trick can be used to deal with other scenarios, such as the presence of singularities that has been described previously.

The main idea is that, in the integral definition of a matrix function, the path Γ does not need to be a single Jordan curve, but can be defined as a union of a finite number of them. The only requirement is that the function is analytic in the Jordan regions, and that the spectrum is contained in their union.

In our settings, it might happen that we cannot enclose the spectrum in a single ball without capturing also the branching point. However, it is always possible to cover it with the union of a finite number of such balls. In this context, assuming that the path Γ is split as the borders of t balls, denoted by $\Gamma_1, \ldots, \Gamma_t$, one has

$$f(A) = \frac{1}{2\pi \boldsymbol{i}} \sum_{i=1}^{t} \int_{\Gamma_i} f(z)\mathfrak{R}(z)dz.$$

Assuming that the number t is small enough, we can obtain the numerical quasiseparability of $f(A)$ by the quasiseparability of each of the addends and then relying on Lemma 4.3.3. Inside each $\Gamma_i = B(z_i, r_i)$ we can perform the change of variable $\tilde{z} := r_i(z - z_i)$ and

write the resolvent as (here the coefficient D will be different by scaling and translation in every Γ_i):

$$\mathfrak{R}(\tilde{z}) = \begin{bmatrix} * & * \\ (\tilde{z}I - D)^{-1}C(\tilde{z})S_D(\tilde{z})^{-1} & * \end{bmatrix}, \qquad \begin{cases} (\tilde{z}I - D)^{-1} = \sum_{j\in\mathbb{Z}} D_j \tilde{z}^j \\ S_D^{-1}(\tilde{z}) = \sum_{s\in\mathbb{Z}} H_s \tilde{z}^s \end{cases}$$

The construction of the coefficients D_j can be done by writing D in Jordan canonical from as

$$V^{-1}DV = \begin{bmatrix} J_{\text{in}} & \\ & J_{\text{out}} \end{bmatrix},$$

where J_{in} refers to the part of the spectrum inside Γ_i, and J_{out} to the one outside. Thanks to the change of variable in the integral, this corresponds to asking that the spectrum of J_{in} is inside the unit disc, and the one of J_{out} outside. Then, one has the following definition for D_j:

$$D_j = \begin{cases} V \begin{bmatrix} J_{\text{in}}^{-j-1} & 0 \\ 0 & 0 \end{bmatrix} V^{-1} & j < 0 \\ \\ -V \begin{bmatrix} 0 & 0 \\ 0 & J_{\text{out}}^{-j-1} \end{bmatrix} V^{-1} & j \geqslant 0 \end{cases},$$

and an analogous formula holds for the coefficients H_s. This provides the Laurent expansion of the off-diagonal block in the integrand. A similar analysis to the one carried out in the previous sections can be used to retrieve the decay on the singular values of this block.

5.4 COMPUTATIONAL ASPECTS AND VALIDATION OF THE BOUNDS

In the previous sections we have proved that the numerical quasiseparable structure is often present in $f(A)$. This property can be used to efficiently evaluate $f(A)$ by means of contour integration. We briefly describe the strategy in the next subsection and we refer the reader to [54] for more details. In Section 5.4.2 we compare our bounds with the actual decay in some concrete cases.

5.4.1 *Contour integration*

The Cauchy integral formula (5.1) can be exploited for approximating $f(A)$ by means of a numerical integration scheme. Recall that, given a complex valued function $g(x)$ defined on an interval $[a, b]$ one can approximate its integral by

$$\int_a^b g(x)dx \;\approx\; \sum_{k=1}^{N} w_k \cdot g(x_k) \tag{5.9}$$

where w_k are the *weights* and x_k are the *nodes*. Since we are interested in integrating a function on $\mathbb{T}$ we can write

$$\frac{1}{2\pi \boldsymbol{i}} \int_{\mathbb{T}} f(z)(zI - A)^{-1} dz = \frac{1}{2\pi} \int_0^{2\pi} e^{\boldsymbol{i}x} f(e^{\boldsymbol{i}x})(e^{\boldsymbol{i}x} I - A)^{-1} dx,$$

where we have parametrized $\mathbb{T}$ by means of e^{ix}. The right-hand side can be approximated by means of (5.9), so we obtain:

$$f(A) \approx \frac{1}{2\pi} \sum_{k=1}^{N} w_k \cdot e^{\boldsymbol{i}x_k} f(e^{\boldsymbol{i}x_k}) \mathfrak{R}(e^{\boldsymbol{i}x_k}). \tag{5.10}$$

This approach has already been explored [46], mainly for the computation of $f(A)b$ due to the otherwise high cost of the inversions in the general case. The pseudocode of the procedure is reported in Algorithm 1.

Algorithm 1 — based on (5.10) — can be carried out cheaply when A is represented as a HODLR-matrix, since the inversion only requires $O(m \log^2(m))$ flops. Moreover, not only the resolvent $\mathfrak{R}(e^{ix_k})$ is representable as a HODLR-matrix, but the same holds for the final result $f(A)$ in view of Theorem 5.2.3. This guarantees the applicability of the above strategy even when dealing with large dimensions.

Algorithm 1 Pseudocode for the evaluation of a contour integral on $\mathbb{T}$

procedure ContourIntegral(f, A) ▷ Evaluate $\frac{1}{2\pi \boldsymbol{i}} \int_{\mathbb{T}} f(z)(zI - A)^{-1} dz$
  $N \leftarrow 1$
  $M \leftarrow f(1) \cdot (I - A)^{-1}$
  $err \leftarrow \infty$
  while $err > \sqrt{u}$ **do**
    $M_{\text{old}} \leftarrow M$
    $M \leftarrow \frac{1}{2} M$ ▷ The new weights are applied to the old evaluations
    $N \leftarrow 2N$
    for $j = 1, 3, \ldots, N - 1$ **do** ▷ Sum the evaluations on the new nodes
      $z \leftarrow e^{\frac{2\pi \boldsymbol{i} j}{N}}$
      $M \leftarrow M + \frac{z f(z)}{N} \cdot (zI - A)^{-1}$
    end for
    $err \leftarrow \|M - M_{\text{old}}\|_2$
  end while
  return M
end procedure

The results in Section 5.3 enable us to deal with functions having poles inside the domain of integration. The only additional step that is required is to compute the correction term described in Theorem 5.3.1. Notice that this step just requires additional

Size	t_{inv}	Res_{inv}	t_{sum}	Res_{sum}
128	2.95 s	$1.33 \cdot 10^{-13}$	1.51 s	$3.3 \cdot 10^{-14}$
256	9.78 s	$4.58 \cdot 10^{-12}$	4.84 s	$1.2 \cdot 10^{-12}$
512	24.6 s	$5.55 \cdot 10^{-11}$	12.2 s	$3.02 \cdot 10^{-12}$
1,024	57 s	$5.87 \cdot 10^{-11}$	23.5 s	$3.92 \cdot 10^{-11}$
2,048	132 s	$6.01 \cdot 10^{-11}$	48.1 s	$3.99 \cdot 10^{-11}$
4,096	245 s	$6.59 \cdot 10^{-11}$	127 s	$5.69 \cdot 10^{-10}$

Table 5.1.: Timing and accuracy on the computation of the matrix function $f(z) = e^z \sin(z)^{-1}$ on a Hermitian matrix A with spectrum contained the unit disc. The residues are measured relatively to the norm of the computed matrix function $f(A)$.

evaluations of the resolvent and so does not change the asymptotic complexity of the whole procedure.

Now, we show an example where Theorem 5.3.1 can be used to derive an alternative algorithm for the evaluation of matrix functions with poles inside the domain.

More precisely, we consider a matrix A with spectrum contained in the unit disc, and the evaluation of the matrix function $f(A)$ with $f(z) = \frac{e^z}{\sin(z)}$. The application of Theorem 5.3.1 yields

$$f(A) = \frac{1}{2\pi \boldsymbol{i}} \int_{\mathbb{T}} f(z)\mathfrak{R}(z)dz + A^{-1}.$$

Then, one can choose to obtain $f(A)$ by computing $e^A \cdot (\sin A)^{-1}$, which requires the evaluation of two integrals and one inverse, or using the above formula, which only requires one integral, one inverse and a sum.

We used an adaptive doubling strategy for the number of nodes i.e., starting with N-th roots of the unit for a small value of N. We apply the quadrature rule (5.10) and we double N until the quality of the approximation is satisfying. In order to check this, we require that the norm of the difference between two consecutive approximations is smaller than a certain threshold. Since the quadrature rule is quadratically convergent [95] and the magnitude of the distance between the approximations at step k and $k+1$ is a heuristic estimate for the error at step k we choose as treshold $\sqrt{u}$ where u is the unit roundoff. In this way we should get an error of the order of u.

We show in Table 5.1, where the approach relying on Theorem 5.3.1 and on computing the function separately are identified by the labels “sum” and “inv”, respectively, that the first choice is faster (due to the reduced number of inversions required) and has a similar accuracy. The matrices in this example have been chosen to be 1-quasiseparable, Hermitian with spectrum in $(-1, 1)$. We have verified the accuracy of the results computing the 2-norm of the residue with respect to the direct application of Definition 5.1.1 to the argument.

5.4.2 *Validation of the bounds*

This section is devoted to check the accuracy of the estimates for the singular values that we have shown in the paper. In order to do so we compute some matrix function on quasiseparable matrices and verify the singular value decay in one of the off-diagonal block. In particular, for a matrix of order m — m even — we consider the off-diagonal block with row indices from $\frac{m}{2}+1$ to m and column indices from 1 to $\frac{m}{2}$. Then, we compare the obtained result with the theoretical bound coming from Theorem 5.2.3. Notice that Theorem 5.2.3 provides a family of bounds depending on a parameter R which can be chosen as long as $f(z)$ is holomorphic in $B(0,R)$. So, in every experiment we estimated the l-th singular value by choosing the parameter R which provides the tighter bound, among the admissible values for the function f under consideration.

We choose two particular classes of 1-quasiseparable matrices for the tests, since we can easily determine the bounds on them:

HERMITIAN TRIDIAGONAL MATRICES These matrices are generated with elements taken from a random Gaussian distribution $N(0,1)$, and are then scaled and shifted so that their spectrum is contained in a ball of center 0 and radius $\frac{3}{4}$. These matrices are normal and the same holds for their submatrices, so we can avoid the computation of the constants $\kappa_s(\cdot)$ which are all equal to 1.

HESSENBERG (SCALED) UNITARY MATRICES We consider a random unitary matrix which is also upper Hessenberg, and so in particular it is 1-quasiseparable (since unitary matrices are rank symmetric - the rank of the lower off-diagonal blocks is equal to the corresponding block above). Then, we scale the matrices multiplying by $\frac{3}{4}$, in order to keep the spectrum on the circle of radius $\frac{3}{4}$. We obtain these matrices in MATLAB by running the command `[A,~] = .75 * qr(hess(randn(N)));` where N is the chosen dimension.

As a first example we consider the matrix exponential e^A which can be easily computed by means of `expm`. We have computed it for many random tridiagonal matrices of size 1000×1000, and the measured and theoretical decays in the submatrix $e^A(501:1000, 1:500)$ are reported in Figure 5.1.

Similarly, in Figure 5.2 we have reported the results of the analogous experiment concerning the function $\log(4I+A)$. In fact, in order for the logarithm to be well defined, we need to make sure that the spectrum of the matrix inside the logarithm does not have any negative value.

As a last example for the tridiagonal matrices we have considered the case of the function $\sqrt{4I+A}$, where the matrix has been shifted again in order to obtain a reasonable estimate by moving the spectrum away from the branching point. The result for this experiment are reported in Figure 5.3.

In the same figures we have reported also the experiments in the case of the scaled unitary Hessenberg matrix. In this case the variance in the behavior of the singular

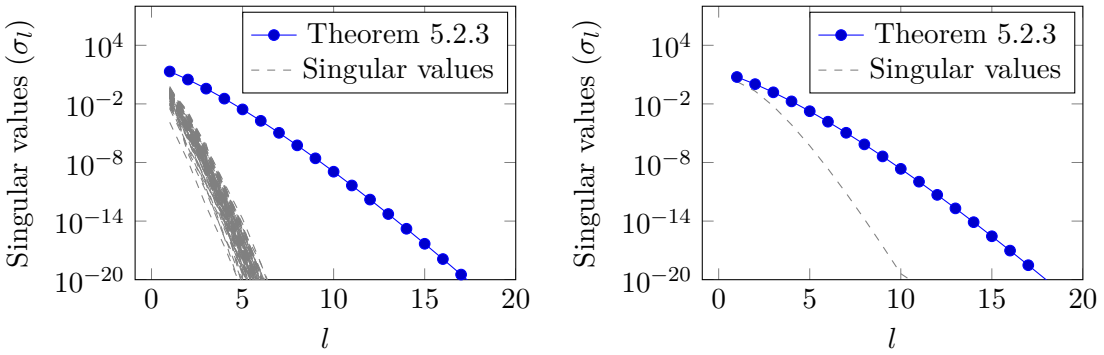

Figure 5.1.: On the left, the bound on the singular values of the off-diagonal matrices of e^A for 100 random Hermitian tridiagonal matrices scaled in order to have spectral radius $\frac{3}{4}$ are shown. In the right picture the same experiment with a scaled upper Hessenberg unitary matrix is reported (with 1 matrix only).

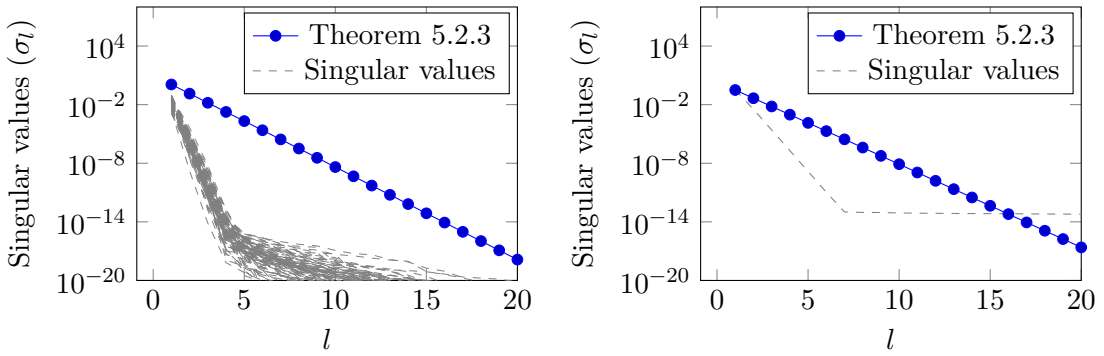

Figure 5.2.: The picture reports the same experiment of Figure 5.1, with the logarithm in place of the exponential. The matrices have however been shifted by $4I$ in order to make the function well-defined. Since this corresponds to evaluating the function $\log(z+4)$ on the original matrix, one can also find a suitable ball centered in 0 where the function is analytic.

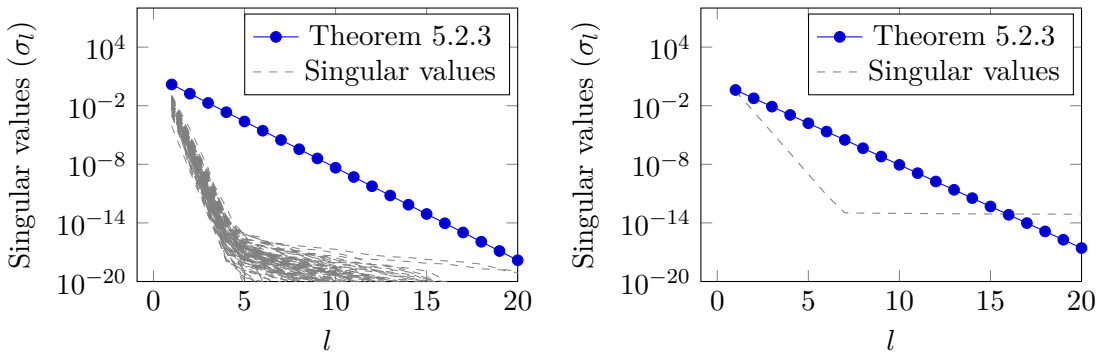

Figure 5.3.: In the left picture the bounds on the singular values of the off-diagonal matrices of $\sqrt{4I + A}$ for 100 random Hermitian tridiagonal matrix scaled in order to have spectral radius $\frac{3}{4}$ are shown. In the right picture the same experiment is repeated for a scaled and shifted upper Hessenberg unitary matrix.

values was very small in the experiments, and so we have only reported one example for each case.

Notice that while in the symmetric (or Hermitian) case every trailing diagonal submatrix is guaranteed to be normal, this is not true anymore for the scaled unitary Hessenberg matrices. Nevertheless, one can verify in practice that these matrices are still not far from normality, and so the bounds that we obtain do not degrade much.

5.5 CONCLUSIONS AND RESEARCH LINES

The numerical preservation of the quasiseparable structure when computing a matrix function is an evident phenomenon. Theoretically, this can be explained with the existence of accurate rational approximants of the function over the spectrum of the argument. In this chapter we have given a closer look to the off-diagonal structure of $f(A)$ providing concrete bounds for its off-diagonal singular values. The off-diagonal blocks have been described as a product between structured matrices with a strong connection with Krylov spaces. This —combined with polynomial interpolation techniques— is the key for proving the bounds.

Moreover, we have developed new tools to deal with the difficulties arising in the treatment of singularities and branching points. In particular, the formula of Corollary 5.3.2 can be employed with the technology of Hierarchical matrices for efficiently computing matrix functions with singularities. An example of this strategy has been provided along with the numerical validation of the bounds.

It would be interesting to see if the analysis can be extended to multivariate functions of matrices [66]. The understanding of this topic could shed some lights on when to expect the quasiseparable structure in several applications as the computation of geometric mean of matrices and the solution of matrix equations.

Chapter 6

Numerical quasiseparable preservation in cyclic reduction

Cyclic reduction, CR for short, is an algorithm originally introduced by G. H. Golub and R. W. Hockney in [60, 31] for the solution of certain block tridiagonal linear systems coming from the finite difference discretization of elliptic PDEs. It has been later generalized and extended to other contexts, like for instance to the solution of polynomial matrix equations, and has been proven to be a successful method for solving a large class of queuing problems and infinite Markov Chains. We refer the reader to the books [17], [16] and to the survey paper [23] for more details and for the many references to the literature.

In this chapter we address the problem of whether a quasiseparable structure in the input data is preserved by the iterative scheme of the algorithm. The positive answer to this question leads us to a version of CR with a high computational efficiency by relying on the HODLR-matrix arithmetic. In Section 6.1 and 6.2 the algorithm is introduced as iterative method for solving quadratic matrix equations and as a direct method for solving tridiagonal block Toeplitz linear systems, respectively. In particular, the computational complexity in the case of a low quasiseparable rank preservation are emphasized. In Section 6.3 it is described the functional interpretation of the algorithm, which plays an important role in the theoretical analysis of its properties. In Section 6.4 a study of the exact quasiseparable rank is performed in the case of starting banded blocks. In Section 6.5 and 6.6 the numerical preservation of the structure is analyzed with different approaches. Finally, in Section 6.7 we report the numerical results of some experiments involving the CR with HODLR representation.

6.1 SOLVING QUADRATIC MATRIX EQUATIONS

Keeping in mind its application in the study of QBD processes given in Chapter 2, we consider the quadratic matrix equation

$$A_{-1} + A_0 X + A_1 X^2 = 0 \qquad A_i \in \mathbb{R}^{m\times m} \quad i = -1, 0, 1, \tag{6.1}$$

and we indicate with $\xi_1, \ldots, \xi_{2m}$ the roots of $\det(A_{-1} + zA_0 + z^2 A_1) = 0$.

Observe that if the solution X of (6.1) exists then it is such that

$$\begin{bmatrix} A_0 & A_1 & & \\ A_{-1} & A_0 & A_1 & \\ & \ddots & \ddots & \ddots \end{bmatrix} \begin{bmatrix} X \\ X^2 \\ X^3 \\ \vdots \end{bmatrix} = \begin{bmatrix} -A_{-1} \\ 0 \\ 0 \\ \vdots \end{bmatrix}.$$

Applying an even-odd permutation to both block-columns and block-rows we get

$$\left[\begin{array}{ccc|ccc} A_0 & & & A_1 & A_{-1} & \\ & A_0 & & & A_1 & \ddots \\ & & \ddots & & & \ddots \\ \hline A_{-1} & & & A_0 & & \\ A_1 & A_{-1} & & & A_0 & \\ & \ddots & \ddots & & & \ddots \end{array}\right] \left[\begin{array}{c} X^2 \\ X^4 \\ \vdots \\ \hline X \\ X^3 \\ \vdots \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ \vdots \\ \hline -A_{-1} \\ 0 \\ \vdots \end{array}\right].$$

If A_0 is not singular then one step of block Gaussian elimination is performed in order to vanish the south-western block, yielding

$$\left[\begin{array}{ccc|ccc} A_0 & & & A_1 & A_{-1} & \\ & A_0 & & & A_1 & \ddots \\ & & \ddots & & & \ddots \\ \hline & & & \widehat{A}_0^{(1)} & A_1^{(1)} & \\ & & & A_{-1}^{(1)} & A_0^{(1)} & \ddots \\ & & & & \ddots & \ddots \end{array}\right] \left[\begin{array}{c} X^2 \\ X^4 \\ \vdots \\ \hline X \\ X^3 \\ \vdots \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ \vdots \\ \hline -A_{-1} \\ 0 \\ \vdots \end{array}\right],$$

where

$$\begin{aligned} \widehat{A}_0^{(1)} &= A_0 - A_{-1}(A_0)^{-1}A_1, \\ A_0^{(1)} &= A_0 - A_{-1}(A_0)^{-1}A_1 - A_1(A_0)^{-1}A_{-1}, \\ A_1^{(1)} &= -A_1(A_0)^{-1}A_1, \\ A_{-1}^{(1)} &= -A_{-1}(A_0)^{-1}A_{-1}. \end{aligned}$$

The crucial thing to notice is that the lower right block is again tridiagonal block Toeplitz unless for the block in position $(1,1)$. This is the mechanism which underlies the cyclic reduction. In fact, iterating this procedure h-times —assuming the non singularity of the sequence $\{A_0^{(n)}\}$— and looking at the south-eastern block, we obtain the system

$$\begin{bmatrix} \widehat{A}_0^{(h+1)} & A_1^{(h+1)} & \\ A_{-1}^{(h+1)} & A_0^{(h+1)} & \ddots \\ & \ddots & \ddots \end{bmatrix} \begin{bmatrix} X \\ X^{2^h+1} \\ X^{2\cdot 2^h+1} \\ \vdots \end{bmatrix} = \begin{bmatrix} -A_{-1} \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

where

$$\begin{aligned} \widehat{A}_0^{(h+1)} &= \widehat{A}_0^{(h)} - A_{-1}^{(h)}(A_0^{(h)})^{-1}A_1^{(h)}, \\ A_0^{(h+1)} &= A_0^{(h)} - A_{-1}^{(h)}(A_0^{(h)})^{-1}A_1^{(h)} - A_1^{(h)}(A_0^{(h)})^{-1}A_{-1}^{(h)}, \\ A_1^{(h+1)} &= -A_1^{(h)}(A_0^{(h)})^{-1}A_1^{(h)}, \\ A_{-1}^{(h+1)} &= -A_{-1}^{(h)}(A_0^{(h)})^{-1}A_{-1}^{(h)}. \end{aligned} \tag{6.2}$$

It has been proven in [23] that if both (6.1) and $A_{-1}X^2 + A_0X + A_1 = 0$ admit solutions with spectral radius less than 1 and the splitting property

$$|\xi_1| \leqslant |\xi_2| \leqslant \ldots \leqslant |\xi_m| < 1 < |\xi_{m+1}| \leqslant \ldots \leqslant |\xi_{2m}|, \tag{6.3}$$

holds then $A_{-1}^{(h)}, A_1^{(h)} \to 0$ and the sequence $(\widehat{A}_0^{(h)})^{-1}A_{-1}$ converges to the minimal nonnegative solution of (6.1). As we will see in Section 6.5.4 condition 6.3 can be relaxed assuming $|\xi_m| < |\xi_{m+1}|$ and scaling the coefficients.

Without any further assumption on the structure of the blocks, each step of CR requires a small number of matrix multiplications and one matrix inversion for the resulting computational cost of $O(m^3)$ arithmetic operations (ops) per step. Assuming starting blocks with a low quasiseparable rank and the numerical preservation of the structure we get —relying on the HODLR-matrix representation— an iterative method with cost $O(m\log(m)^2)$ per step.

6.2 SOLVING FINITE TRIDIAGONAL BLOCK TOEPLITZ SYSTEMS

We consider a block tridiagonal linear system of the kind $\mathcal{A}_n x = b$ where $\mathcal{A}_n = \mathrm{trid}_n(A_{-1}, A_0, A_1)$ and the blocks A_i are $m \times m$ matrices such that CR can be carried out with no breakdown.

$$\begin{bmatrix} A_0 & A_1 & & & \\ A_{-1} & A_0 & A_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & A_1 \\ & & & A_{-1} & A_0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ \vdots \\ b_n \end{bmatrix}, \qquad x_i, b_i \in \mathbb{R}^m. \tag{6.4}$$

For simplicity, assume $n = 2^q - 1$ so that the description of CR is simpler, for more details in the general case we refer the reader to [23].

An odd-even permutation of block rows and columns yields

$$\left[\begin{array}{cccc|ccc} A_0 & & & & A_1 & & \\ & A_0 & & & A_{-1} & \ddots & \\ & & \ddots & & & \ddots & A_1 \\ & & & A_0 & & & A_{-1} \\ \hline A_{-1} & A_1 & & & A_0 & & \\ & \ddots & \ddots & & & \ddots & \\ & & A_{-1} & A_1 & & & A_0 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_3 \\ \vdots \\ x_n \\ \hline x_2 \\ x_4 \\ \vdots \\ x_{n-1} \end{array}\right] = \left[\begin{array}{c} b_1 \\ b_3 \\ \vdots \\ b_n \\ \hline b_2 \\ b_4 \\ \vdots \\ b_{n-1} \end{array}\right].$$

Again, one step of block Gaussian elimination is performed to vanish the south-western block, yielding

$$\left[\begin{array}{cccc|ccc} A_0 & & & & A_1 & & \\ & A_0 & & & A_{-1} & \ddots & \\ & & \ddots & & & \ddots & \ddots \\ & & & \ddots & & & \ddots \\ & & & A_0 & & & A_{-1} \\ \hline & & & & A_0^{(1)} & A_1^{(1)} & \\ & & & & A_{-1}^{(1)} & \ddots & \ddots \\ & & & & & \ddots & \ddots \\ & & & & & & A_{-1}^{(1)} \end{array}\right.$$
$$\left.\begin{array}{cc} A_1 & \\ \ddots & \\ \ddots & A_1 \\ \ddots & \\ & A_0^{(1)} \\ \ddots & \\ \ddots & A_1^{(1)} \\ A_{-1}^{(1)} & A_0^{(1)} \end{array}\right] \left[\begin{array}{c} x_1 \\ x_3 \\ \vdots \\ x_n \\ \hline x_2 \\ x_4 \\ \vdots \\ x_{n-1} \end{array}\right] = \left[\begin{array}{c} b_1 \\ b_3 \\ \vdots \\ b_n \\ \hline b_1^{(1)} \\ b_2^{(1)} \\ \vdots \\ b_{\frac{n-1}{2}}^{(1)} \end{array}\right]$$

with

$$\begin{aligned} A_0^{(1)} &= A_0 - A_{-1}A_0^{-1}A_1 - A_1A_0^{-1}A_{-1}, \\ A_{-1}^{(1)} &= -A_{-1}A_0^{-1}A_{-1}, \quad A_1^{(1)} = -A_1A_0^{-1}A_1, \\ b_i^{(1)} &= b_{2i} - A_{-1}A_0^{-1}b_{2i-1} - A_1A_0^{-1}b_{2i+1}, \quad i = 1, \dots \frac{n-1}{2}. \end{aligned} \tag{6.5}$$

The south-eastern block yields the system of the kind $\mathcal{A}_{\frac{n-1}{2}} x_{even} = b^{(1)}$ with $\mathcal{A}_{\frac{n-1}{2}} = \mathrm{trid}_{\frac{n-1}{2}}(A_{-1}^{(1)}, A_0^{(1)}, A_1^{(1)})$, where x_{even} denotes the subvector of x formed with the even block components, whose solution can be obtained by cyclically applying CR. Once the even block components of the block vector x have been computed, they can be substituted in the first part of the linear equations so that the odd block components of x are recovered. The hierarchical quasiseparability of the block matrices makes each operation of low cost.

Thus, the first (as well as the generic) step of CR performs the following steps

(i) Given the $m \times m$ matrices A_{-1}, A_0, A_1 compute the matrices $A_{-1}^{(1)}, A_0^{(1)}, A_1^{(1)}$.

(ii) Given the m-vectors b_i, $i = 1, \ldots, n$, compute $b_i^{(1)}$, $i = 1, \ldots, \frac{n-1}{2}$ by means of (6.5).

(iii) Recursively solve the system $\mathrm{trid}_{\frac{n-1}{2}} x_{even} = b^{(1)}$ by means of CR.

(iv) Compute the odd components of x with back substitution:

$$\begin{aligned} x_1 &= A_0^{-1}(b_1 - A_1 x_2), \\ x_i &= A_0^{-1}(b_i - A_{-1}x_{i-1} - A_1 x_{i+1}), \qquad i = 3, 5, \ldots, n-2, \\ x_n &= A_0^{-1}(b_1 - A_{-1}x_{n-1}), \end{aligned}$$

If the starting blocks A_{-1}, A_0, A_1 are quasiseparable, e.g., they are tridiagonal and the structure is numerically preserved by the iterative scheme then —relying on the HODLR-matrix representation— the cost of step (i) is $O(k^2 m \log^2 m)$, while the costs of steps (ii) and (iv) is $O(k^2 nm \log m)$ where k is an upper bound for the quasiseparable rank of the matrices during the iteration. Therefore, indicating with $T(m, n)$ the asymptotic computational complexity of the whole algorithm with $n = 2^q - 1$, we have

$$T(m, n) = T\left(m, \frac{n-1}{2}\right) + O(k^2 m \log^2 m) + O(k^2 nm \log m).$$

Since $T(m, 1) = O(k^2 m \log^2 m)$, we obtain $T(m, n) = O(k^2 mn \log m) + O(k^2 m \log^2 m \log n)$. For $m = n$ this yields $T(m, m) = O(k^2 m^2 \log m) + O(k^2 m \log^3 m)$.

It is interesting to remark that if $\mathcal{A}_m$ is the discrete Laplacian where $A_{-1} = A_1 = -I$, $A_0 = \mathrm{trid}_m(-1, 4, -1)$, then CR has a cost of $O(m^2 \log m)$ ops [93] while the fast Poisson solvers based on the combination of Fourier analysis and CR [57] have a cost of $O(m^2 \log \log m)$ ops. So this approach has a slightly higher cost but covers a wider range of cases including tridiagonal block Toeplitz matrices with banded (not necessarily Toeplitz) blocks.

Observe that CR preserves slightly more general structures than the block tridiagonal block Toeplitz. In particular it is possible to handle the case where the first and last blocks in the main diagonal differ from the other blocks on the same diagonal, see [23].

6.3 FUNCTIONAL INTERPRETATION

We recall the functional interpretation of the cyclic reduction introduced in the Markov chains framework [17] and generalized in order to prove applicability and convergence properties of this algorithm [23].

Associate the matrices $A_i^{(h)}$, $i = -1, 0, 1$ defined in (6.2) with the matrix Laurent polynomial

$$\varphi^{(h)}(z) := z^{-1} A_{-1}^{(h)} + A_0^{(h)} + z A_1^{(h)}, \tag{6.6}$$

starting with $\varphi^{(0)}(z) = \varphi(z) = z^{-1}A_{-1} + A_0 + zA_1$. Moreover, we define the matrix rational function $\psi^{(h)}(z) = \varphi^{(h)}(z)^{-1}$. The matrix function $\psi^{(h)}(z)$ turns out to enjoy the following recurrence property

$$\begin{cases} \psi(z) = \psi^{(0)}(z) := \varphi(z)^{-1}, \\ \psi^{(h+1)}(z^2) := \frac{1}{2}(\psi^{(h)}(z) + \psi^{(h)}(-z)), \end{cases}$$

for every $z \in \mathbb{C}$ such that $\det(\varphi^{(j)}(z)) \neq 0$, $j = 0, \ldots, h$. In particular, expanding the recurrence relation in the sequence $\{\psi^{(h)}\}_{h\in\mathbb{N}}$, we find that

$$\psi^{(h)}(z^{2^h}) = \frac{1}{2^h} \sum_{j=0}^{2^h-1} \psi^{(0)}(\omega^j z) \tag{6.7}$$

where $\omega = e^{\frac{2\pi}{N}\boldsymbol{i}}$ is a principal N-th root of unity for $N = 2^h$, and $\boldsymbol{i}$ denotes the imaginary unit, so that

$$\varphi^{(h)}(z^{2^h}) = \left(\frac{1}{2^h} \sum_{j=0}^{2^h-1} \psi^{(0)}(\omega^j z) \right)^{-1}. \tag{6.8}$$

Observe that in the case where A_{-1}, A_0 and A_1 are tridiagonal, then $\varphi(z)$ is tridiagonal as well, so that for any value of z such that $\det \varphi(z) \neq 0$, the matrix $\psi(z)$ is semi-separable, that is, $\mathrm{tril}(\psi(z)) = \mathrm{tril}(L)$, $\mathrm{triu}(\psi(z)) = \mathrm{triu}(U)$, where L and U are matrices of rank 1.

In the next sections we will exploit this tool for studying the exact and numerical quasiseparable rank of the three sequences $A_{-1}^{(h)}$, $A_0^{(h)}$ and $A_1^{(h)}$.

6.4 STUDY OF THE EXACT QUASISEPARABLE RANK IN THE BANDED CASE

In this subsection we consider the case in which the starting blocks A_i, $i = -1, 0, 1$ are tridiagonal.

A tridiagonal matrix enjoys the property of having all submatrices strictly contained either under or above the main diagonal, of rank 1. We aim to show the theoretical growth of this rank structure during the CR iterations for the blocks A_i, $i = -1, 0, 1$.

6.4.1 *Upper bounds for the tridiagonal case*

We start this subsection stating a technical result which will be useful later.

Lemma 6.4.1. *Let $p, q \in \mathbb{Z}$, L be the bidiagonal matrix*

$$\begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix} \tag{6.9}$$

and $A \in \mathbb{R}^{m\times m}$ be such that $r_{lw}^{(p)}(A) \leqslant k_l$ and $r_{up}^{(q)}(A) \leqslant k_u$. Then

$$\begin{array}{ll} r_{lw}^{(p)}(AL^t) \leqslant k_l, & r_{up}^{(q+1)}(AL^t) \leqslant k_u, \\ r_{lw}^{(p-1)}(AL) \leqslant k_l, & r_{up}^{(q)}(AL) \leqslant k_u, \\ r_{lw}^{(p)}(L^tA) \leqslant k_l, & r_{up}^{(q+1)}(L^tA) \leqslant k_u, \\ r_{lw}^{(p-1)}(LA) \leqslant k_l, & r_{up}^{(q)}(LA) \leqslant k_u, \\ r_{lw}^{(p-1)}(L^tAL) \leqslant k_l, & r_{up}^{(q+1)}(L^tAL) \leqslant k_u, \\ r_{lw}^{(p-1)}(LAL^t) \leqslant k_l, & r_{up}^{(q+1)}(LAL^t) \leqslant k_u. \end{array}$$

Moreover if A is generator representable, i.e.

$$tril(A,p) = tril(UV^*,p) \quad and \quad triu(A,q) = triu(WZ^*,q),$$

with $U, V \in \mathbb{R}^{m\times k_l}$ and $W, Z \in \mathbb{R}^{m\times k_u}$, then even the matrices previously considered are generator representable.

Proof. In order to prove all the inequalities it is sufficient to observe the four possible effects of the multiplication by L or L^t:

$\boxed{\cdot L^t}$ Replaces the i-th column with the difference between the i-th and the $i-1$-th.

$\boxed{\cdot\ L}$ Replaces the i-th column with the difference between the i-th and the $i+1$-th.

$\boxed{L^t\cdot}$ Replaces the i-th row with the difference between the i-th and the $i+1$-th.

$\boxed{L\ \cdot}$ Replaces the i-th row with the difference between the i-th and the $i-1$-th.

Concerning the representation properties, we prove the statement only for the lower part of L^tAL because the other verifications are made in the same way.

We know that

$$a_{ij} = \sum_{r=1}^{k_l} u_{ir}v_{jr} \quad \forall i \geqslant j-p.$$

Let i, j be such that $i \geqslant j-p+1$. Observe that the j-th column of the matrix $A \cdot L$ is the difference between the j-th and the $(j+1)$-th column of A (unless the m-th, which coincides with the m-th column of A). Hence

$$(AL)_{ij} = \sum_{r=1}^{k_l} u_{ir}(v_{jr} - v_{j+1r}) \qquad (v_{n+1r} = 0\ \forall r = 1, \ldots, k_l).$$

We repeat the argument observing that the i-th row of L^tAL is the difference between the i-th and the $(i+1)$-th row of AL. Therefore

$$(L^tAL)_{ij} = \sum_{r=1}^{k_l}(u_{ir} - u_{i+1r})(v_{jr} - v_{j+1r}) \qquad (u_{n+1r} = v_{n+1r} = 0 \ \forall r = 1, \dots, k_l).$$

So taking

$$(\tilde{U})_{ir} = u_{ir} - u_{i+1r} \quad \text{and} \quad (\tilde{V})_{ir} = v_{ir} - v_{i+1r}$$

we get $\mathrm{tril}(L^tAL, p-1) = \tilde{U}\tilde{V}^*$. □

Remark 6.4.2. *If A is the inverse of an element in $\mathbb{B}_p^p$ (so $r_{lw}^{(p-1)}(A) \leqslant p$ and $r_{up}^{(1-p)}(A) \leqslant p$) then $s_{rank}((L^t)^kAL^k) \leqslant (p,p)$ $\forall k = 0, \dots, p-1$.*

Moreover if A is a strict band matrix then $(L^t)^kAL^k$ is extended (p,p)-generator representable semiseparable $\forall k = 0, \dots, p-1$.

Moreover, in [63] a practical formula, involving the shift operator L, for the inverse of a matrix with a generated part plus a band correction is provided.

Theorem 6.4.3 (Theorem 3.3 in [63])**.** *Let $B \in \mathbb{B}_l^u$ matrix and $S \in \mathbb{G}_{k_l}^{k_u}$ be two $m \times m$ matrices. Then the inverse of their sum has this multiplicative structure:*

$$(B+S)^{-1} = (D_1'L^t) \cdot \ldots \cdot (D_{k_u}'L^t) \cdot \widetilde{B}^{-1} \cdot (LD_{k_l}) \cdot \ldots \cdot (LD_1),$$

where D_i and D_i' are diagonal $m \times m$ matrices, $\widetilde{B} \in \mathbb{B}_{l+k_l}^{u+k_u}$ and L is the bidiagonal matrix (6.9).

We are ready to prove the following upper bound.

Theorem 6.4.4. *If the Cyclic Reduction algorithm starts with A_{-1}, A_0, A_1 irreducible tridiagonal then the matrices $A_{-1}^{(h)}, A_0^{(h)}$ and $A_1^{(h)}$ of the iteration scheme verify*

$$\begin{aligned}
&q_{rank}(A_{-1}^{(h)}) \leqslant (2^h, 2^h), && s_{rank}(A_{-1}^{(h)}) \leqslant (2^h+1, 2^h+1),\\
&q_{rank}(A_1^{(h)}) \leqslant (2^h, 2^h), && s_{rank}(A_1^{(h)}) \leqslant (2^h+1, 2^h+1),\\
&q_{rank}(A_0^{(h)}) \leqslant (2^{h+1}-1, 2^{h+1}-1), && s_{rank}(A_0^{(h)}) \leqslant (2^{h+1}, 2^{h+1}).
\end{aligned}$$

Proof. We define $\mathcal{X} := \{z \in \mathbb{C} : \quad \det(\varphi^{(j)}(z)) \neq 0, \ j = 0, \dots, h\}$. First observe that $\forall z \in \mathcal{X} \quad \psi^{(0)}(z) = z\varphi^{(0)}(z)^{-1} \in \mathbb{G}_1^1$ since is the inverse of an irreducible tridiagonal matrix. This and formula (6.8) imply that $\varphi^{(h)}(z)$ is the inverse of an element in $\mathbb{G}_{2^h}^{2^h}$. Since the quasiseparable rank is invariant under inversion we get

$$q_{rank}(\varphi^{(h)}(z)) \leqslant (2^h, 2^h), \qquad s_{rank}(\varphi^{(h)}(z)) \leqslant (2^h+1, 2^h+1), \qquad \forall z \in \mathcal{X}.$$

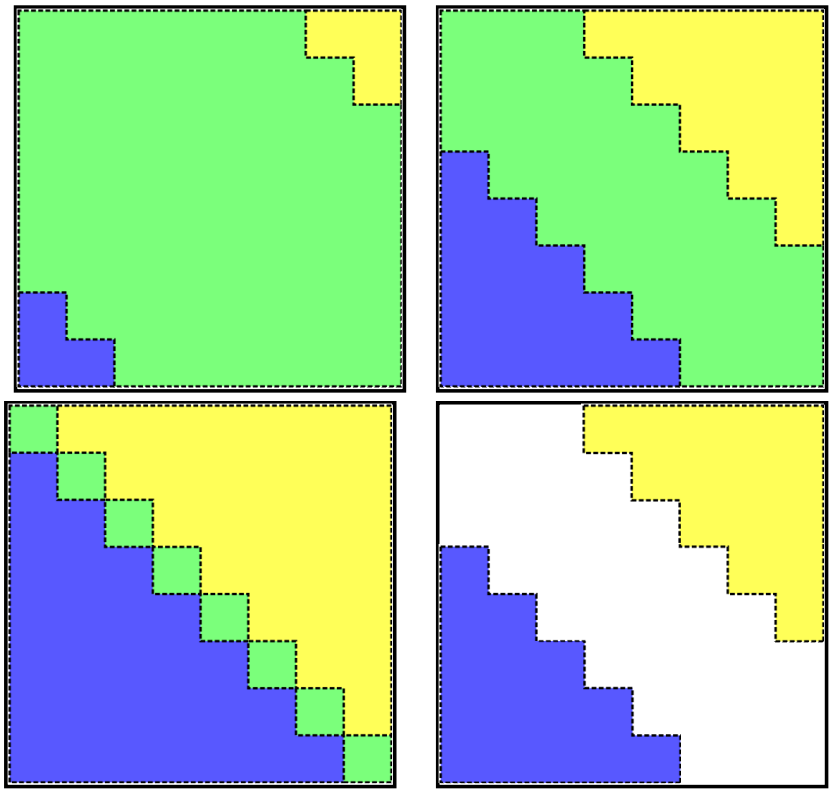

Figure 6.1.: The trend of the representation property under the transformation $A \to (L^t)^k A L^k$ at the variance of k. The considered A is the inverse of an element in $\mathbb{B}_p^p$. The parts represented as tril or triu of a p-rank matrix are filled in yellow and blue respectively, while their intersection is filled in green. The different images refer to the cases $k = 0$, $0 < k < p-1$, $k = p-1$ and $k > p-1$, respectively.

Observe that the following relations hold:

$$\begin{aligned} A_{-1}^{(h)} &= \lim_{z\to 0} \varphi^{(h)}(z), \\ A_{1}^{(h)} &= \lim_{z\to +\infty} \tfrac{\varphi^{(h)}(z)}{z^2}, \\ A_{0}^{(h)} &= \tfrac{1}{2}\left(\tfrac{\varphi^{(h)}(z)}{z} + \tfrac{\varphi^{(h)}(-z)}{-z}\right). \end{aligned}$$

Since for each fixed h the complementary of $\mathcal{X}$ is finite we can say, due to the lower semi-continuity of the quasiseparable and semiseparable rank, that the bounds can be extended to the limit and so the thesis for $A_{-1}^{(h)}$ and $A_{1}^{(h)}$ is proved.

Concerning $A_{0}^{(h)}$ we have to study the structure of $\varphi^{(h)}(z)$ a little deeper. We know that $\varphi^{(h)}(z)$ is the inverse of a sum of elements in $\mathbb{G}_1^1$. We can choose a $z \in \mathcal{X}$ such that

each component of the generators of the inverse of $\varphi^{(h)}(z)$ is nonzero and we write (we omit the dependence on h and z to ease the notation)

$$(\varphi^{(h)}(z))^{-1} = \psi_1 + \psi_2$$

where $\psi_1 \in \mathbb{G}_1^1$ with generator (u_1, v_1, w_1, z_1) and $\psi_2 \in \mathbb{G}_{2^h-1}^{2^h-1}$.

Let L be the bidiagonal matrix of Lemma 6.4.1, $D_1 = \operatorname{diag}(u_1)^{-1}$, $D_1' = \operatorname{diag}(z_1)^{-1}$. It is easy to see that

$$LD_1\psi_1 D_1' L^t \in \mathbb{B}_0^0 \quad \text{(is diagonal)}$$

and the matrix $LD_1\psi_2 D_1' L^t$, thanks to Lemma 6.4.1, has the strictly lower and strictly upper triangular part representable with a generator of rank $2^h - 1$. Since

$$\varphi^{(h)}(z) = D_1' L^t (LD_1(\psi_1 + \psi_2)D_1' L^t)^{-1} LD_1$$

applying Theorem 6.4.3 we obtain that there exist $D_1, D_1', D_2, D_2', \dots D_{2^h}, D_{2^h}'$ invertible diagonal matrices and $E \in \mathbb{B}_{2^h-1}^{2^h-1}$ such that

$$\varphi^{(h)}(z) = \left(D_1' L^t\right) \dots \left(D_{2^h}' L^t\right) E^{-1} \left(LD_{2^h}\right) \dots (LD_1)\,.$$

The matrix E^{-1} is $(2^h - 1, 2^h - 1)$-semiseparable. Observing that multiplying on the left or on the right for invertible diagonal matrices does not modify the rank structures, applying iteratively Lemma 6.4.1 and Remark 6.4.2 to the matrix E^{-1} we can write

$$\varphi^{(h)}(z) = \varphi_1 + \varphi_2$$

where $r_{lw}^{(-2)}(\varphi_1) \leqslant 2^h - 1$ and $r_{up}^{(2)}(\varphi_1) \leqslant 2^h - 1$ and $\varphi_2 \in \mathbb{B}_1^1$. This means that $\forall z \in \mathcal{X}$ such that $-z \in \mathcal{X}$ the matrix

$$\frac{1}{2}\left(\frac{\varphi^{(h)}(z)}{z} + \frac{\varphi^{(h)}(-z)}{-z}\right)$$

can be written as a matrix with $r_{lw}^{(-2)} \leqslant 2^{h+1} - 2$ and $r_{up}^{(2)} \leqslant 2^{h+1} - 2$ plus a tridiagonal correction. This implies the thesis. □

6.4.2 *Extension to general banded matrices*

Looking closely at the arguments that prove Theorem 6.4.4, we can see that everything relies on two facts:

(i) If the matrices A_{-1}, A_0 and A_1 are tridiagonal then $\varphi^{(0)}(z)$ is tridiagonal.

(ii) The inverse of a (irreducible/strict) tridiagonal matrix is a (generator representable) $(1, 1)$-semiseparable matrix.

The immediate generalization of these properties leads us to consider banded matrices.

Theorem 6.4.5. *If the Cyclic Reduction algorithm starts with* $A_i \in \mathbb{B}_{l^{(i)}}^{u^{(i)}}$ *for* $i = -1, 0, 1$, *then the matrices* $A_{-1}^{(h)}, A_0^{(h)}$ *and* $A_1^{(h)}$ *verify*

$$\begin{aligned} q_{rank}(A_{-1}^{(h)}) &\leqslant (l \cdot 2^h \ , \ u \cdot 2^h), 1 \\ q_{rank}(A_1^{(h)}) &\leqslant (l \cdot 2^h \ , \ u \cdot 2^h), \\ q_{rank}(A_0^{(h)}) &\leqslant (l \cdot (2^{h+1} - 1) \ , \ u \cdot (2^{h+1} - 1)), \end{aligned}$$

with $l := \max_i \ l^{(i)}$ *and* $u := \max_i \ u^{(i)}$.

Proof. Concerning the first two inequalities we observe that since $\varphi^{(0)}(z) \in \mathbb{B}_l^u$, we can do similar considerations to the proof of the tridiagonal case.

Even for $A_0^{(h)}$ we emulate what we have done getting that $\varphi^{(0)}(z) \in \mathbb{B}_l^u$ implies

$$(\varphi^{(h)}(z))^{-1} = \psi_1 + \psi_2$$

where ψ_1 (l, u)-semiseparable and $\psi_2 \in \mathbb{G}_{l(2^h-1)}^{u(2^h-1)}$. In particular ψ_2 is the sum of $2^h - 1$ inverse of elements in $\mathbb{B}_l^u$ therefore $r_{lw}^{(l-1)}(\psi_2) \leqslant l(2^h - 1)$ and $r_{up}^{(1-u)}(\psi_2) \leqslant u(2^h - 1)$. Since we have the freedom of choice on z we can assume that the matrices involved have generators with non zero components.

Our aim is to show that $(\varphi^{(h)}(z))^{-1}$ is the sum of an (l, u)-band matrix plus a $(l \cdot (2^h - 1), u \cdot (2^h - 1))$ quasiseparable matrix. In order to do this we prove individually that the upper and lower part of $(\varphi^{(h)}(z))^{-1}$ have the right structure, i.e a generated part plus a band. Actually we do that explicitly only for the upper part, because the lower is analogous.

Suppose that $\mathrm{tril}(\psi_1, l-1) = UV^*$ and $\mathrm{triu}(\psi_1, 1-u) = WZ^*$. We call

$$D_1' = \mathrm{diag}(w^{(1)})^{-1}$$

where $w^{(j)}$ indicate the j-th column of W. Then with a direct verification we have that

$$r_{lw}^{(l-1)}(L^t D_1' \psi_1) \leqslant l, \quad r_{up}^{(2-u)}(L^t D_1' \psi_1) \leqslant u, \quad r_{up}^{(1)}(L^t D_1' \psi_1) \leqslant u - 1.$$

We can iterate the process until we run out all the generators of the upper part, i.e. there exist $D_1', \ldots, D_u'$ invertible diagonal matrices such that

$$\begin{aligned} (L^t D_u') \ldots (L^t D_1') \psi_1 &\in \mathbb{B}_n^0 \quad \text{(is lower triangular)}, \\ r_{lw}^{(0)} \left((L^t D_u') \ldots (L^t D_1') \psi_1 \right) &\leqslant l \end{aligned}$$

and

$$\begin{aligned} r_{up}^{(1)}((L^t D_u') \ldots (L^t D_1') \psi_2) &\leqslant u \cdot (2^h - 1), \\ r_{lw}^{(0)}((L^t D_u') \ldots (L^t D_1') \psi_2) &\leqslant l \cdot (2^h - 1). \end{aligned}$$

Using again Theorem 6.4.3 we get that $\exists D_1, \ldots, D_{u\cdot(2^h-1)}$, $D'_1, \ldots, D'_{u+l\cdot 2^h}$ invertible diagonal matrices and $E \in \mathbb{B}_{l\cdot 2^h}^{u\cdot(2^h-1)}$ such that

$$\varphi^{(h)}(z) = \left(D'_1 L^t\right) \ldots \left(D'_{u\cdot(2^h-1)} L^t\right) E^{-1} \left(L D_{u+l\cdot 2^h}\right) \ldots \left(L D_{u+1}\right) \left(L^t D_u\right) \ldots \left(L^t D_1\right).$$

Using again Lemma 6.4.1, we have that $r_{up}^{(u)}(\varphi^{(h)}(z)) \leqslant u \cdot (2^h - 1)$ therefore the upper triangular part of $\varphi^{(h)}(z)$ has the desired structure. Similarly we get $r_{lw}^{(-l)}(\varphi^{(h)}(z)) \leqslant l \cdot (2^h - 1)$, therefore

$$\varphi^{(h)}(z) = \varphi_1 + \varphi_2$$

where $r_{lw}^{(-l)}(\varphi_1) \leqslant l \cdot (2^h - 1)$, $r_{up}^{(u)}(\varphi_1) \leqslant u \cdot (2^h - 1)$ and $\varphi_2 \in \mathbb{B}_l^u$.

We conclude exploiting the final argument of the proof of Theorem 6.4.4. □

6.5 NUMERICAL PRESERVATION: QUEUEING THEORY FRAMEWORK

In this section we assume the additional hypotheses of (2.2). To be precise, we assume $A_{-1} = -P_{-1}$, $A_0 = I - P_0$ and $A_1 = -P_1$ where the P_is are non negative $m \times m$-matrices with a low quasiseparable rank and such that $P_{-1} + P_0 + P_1$ is substochastic.

Looking at the results of the previous section we see that, as far as we know, the quasiseparable rank can grow exponentially with respect to the number of iterations. Despite that, plotting the singular values of the off-diagonal blocks of the matrices $A_i^{(h)}$ shows an interesting behavior as reported in Figure 6.5.

It is evident that, even though the number of nonzero singular values grows at each step of CR, the number of singular values above the machine precision – denoted by a horizontal line in Figure 6.5– is bounded by a moderate constant. Moreover, the singular values seem to stay below a straight-line which constitutes an asymptotic bound. That is, they get closer to this line as $h \to \infty$. The logarithm scale suggests that the computed singular values $\sigma_l^{(h)}$ decay exponentially with l and the basis of the exponential grows with h but has a limit less than 1.

In this section we will prove this property relating the basis of the exponential decay to the width of the domain of analyticity of the matrix function $\psi(z) = \varphi(z)^{-1}$.

6.5.1 *Exponential decay of the singular values in $\psi^{(h)}(z)$*

It is clear that, if the blocks A_i $i = -1, 0, 1$ have an off-diagonal rank structure, then the matrix $\varphi^{(0)}(z)$ also enjoys this property. We will show that this fact implies the exponential decay of the singular values of the off-diagonal blocks of $\varphi^{(h)}(z^{2^h})$ for every h and for any $z \in \mathbb{T}$.

Given an integer $N > 0$, let $\omega_N = e^{2\pi \boldsymbol{i}/N}$ and observe that

$$\frac{1}{N} \sum_{j=0}^{N-1} (z\omega_N^j)^k = \begin{cases} z^k & k \equiv 0 \mod N \\ 0 & \text{otherwise} \end{cases}.$$

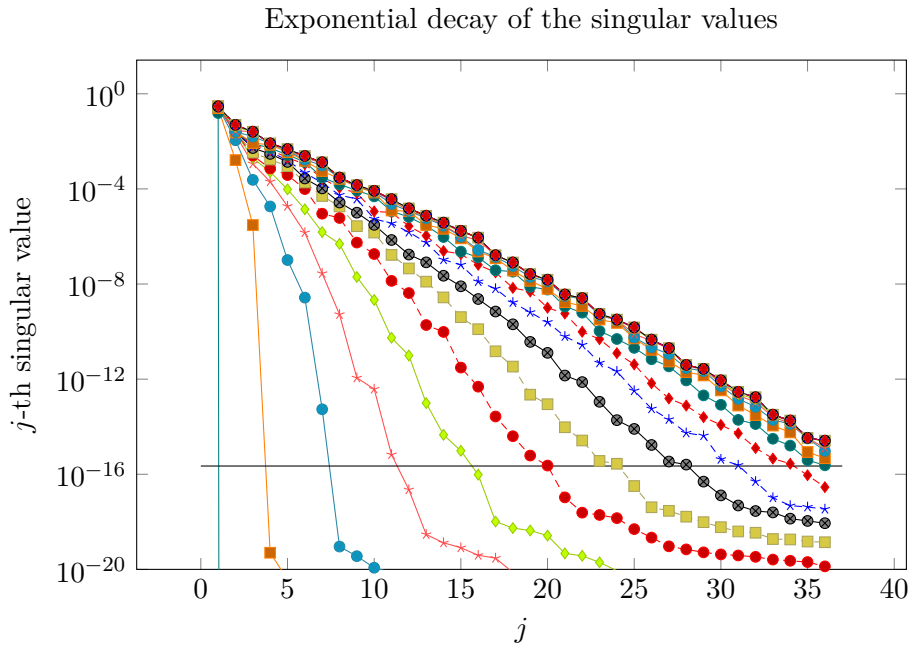

Figure 6.2.: Log-scale plot of the most significant singular values of the largest south-western submatrix of $A_0^{(h)}$ contained in the lower triangular part, for $m = 1600$ and $h = 1, \dots, 15$. The horizontal line denotes the machine precision threshold. Matrices are randomly generated so that $-A_{-1}, I - A_0, -A_1$ are non negative tridiagonal matrices and $I - A_{-1} - A_0 - A_1$ is stochastic.

This way, if $A(z) = \sum_{i\in\mathbb{Z}} z^i A_i$ is a matrix Laurent series analytic on the annulus $\mathbb{A}(r_1, r_2)$ for $0 < r_1 < 1 < r_2$, then

$$\frac{1}{N}\sum_{j=0}^{N-1} A(\omega_n^j z) = \sum_{i\in\mathbb{Z}} z^{Ni} A_{Ni} = \widehat{A}(z^N)$$

where $\widehat{A}(z) := \sum_{i\in\mathbb{Z}} z^i A_{Ni}$ is analytic on $\mathbb{A}(r_1^N, r_2^N)$.

We denote by I_N the operator which maps $A(z)$ into $\widehat{A}(z)$ and write $\widehat{A}(z) = I_N(A(z))$. Observe that I_N is linear and continuous on the space of analytic functions on $\mathbb{A}(r_1, r_2)$.

Moreover, in view of (6.8), we have $\psi^{(h)} = I_N(\psi^{(0)})$ for $N = 2^h$. This way, if we prove that any off-diagonal submatrix $\tilde{C}(z)$ of $\psi^{(0)}(z)$ is such that $I_N(\tilde{C}(z))$ has the exponential decay property for its singular values, then we have shown this property also for $\psi^{(h)}(z)$.

Partition $\varphi(z)$ and $\varphi(z)^{-1}$ as follows

$$\varphi(z) = \begin{bmatrix} I - E(z) & -B(z) \\ -C(z) & I - D(z) \end{bmatrix}, \qquad \psi(z) := \varphi(z)^{-1} = \begin{bmatrix} \tilde{E}(z) & \tilde{B}(z) \\ \tilde{C}(z) & \tilde{D}(z) \end{bmatrix}, \tag{6.10}$$

where $E(z)$ and $D(z)$ are square matrices of any compatible size.

Theorem 6.5.1. *Let $\varphi(z) = z^{-1}A_{-1} + A_0 + zA_1$ be an $m \times m$ matrix function such that*

(i) The matrices $-A_{-1}, I - A_0$ and $-A_1$ are non-negative and $I - \varphi(z)$ has spectral radius smaller than 1 for any $z \in \mathbb{T}$.

(ii) The blocks A_i are k-quasiseparable, $\|I - A_0\|_2 \leqslant L$ and $\|A_i\|_2 \leqslant L$, $i = -1, 1$.

(iii) There exist $t > 1$ and $\delta > 0$ such that $\det \varphi(z) \neq 0$ and $\|\varphi(z)^{-1}\|_2 \leqslant \delta$ for $z \in \mathbb{A}(t^{-1}, t)$.

Then $\rho(I - \varphi(z)) < 1$ for any $z \in \mathbb{A}(t^{-1}, t)$ and in the partitioning (6.10), both blocks $I - E(z)$ and $I - D(z)$ are invertible for any $z \in \mathbb{A}(t^{-1}, t)$. Moreover, for any $z \in \mathbb{T}$ and for any h, the singular values of $\tilde{C}^{(h)} := I_N(\tilde{C}(z))$, with $N = 2^h$, are such that

$$\sigma_l(\tilde{C}^{(h)}(z)) \leqslant 3Me^{-\frac{l-3k}{6k}\log t}, \qquad M = \frac{4L\delta^2}{(1 - e^{-N\log t})(1 - t^{-1})}. \tag{6.11}$$

Moreover, if A_{-1}, A_0, A_1 are tridiagonal then the above bound turns into

$$\sigma_l(\tilde{C}^{(h)}(z)) \leqslant Me^{-\frac{l}{2}\log t}. \tag{6.12}$$

Proof. Let us prove that $\rho(I - \varphi(z)) < 1$ for any $z \in \mathbb{A}(t^{-1}, t)$. By contradiction, assume that there exists $\xi \in \mathbb{A}(t^{-1}, t)$ such that $\rho(I - \varphi(\xi)) \geqslant 1$. Since $I - A_0$ and $-A_i \geqslant 0$ for $i = -1, 1$ then $|I - \varphi(\xi)| \leqslant I - \varphi(|\xi|)$, and by the monotonicity of the spectral radius we get $1 \leqslant \rho(I - \varphi(\xi)) \leqslant \rho(I - \varphi(|\xi|))$. Thus, since $\rho(I - \varphi(1)) < 1 \leqslant \rho(I - \varphi(|\xi|))$ and ρ is a continuous function, then there exists $1/t < \hat{\xi} < t$ such that $\rho(I - \varphi(\hat{\xi})) = 1$. Since $I - \varphi(\hat{\xi})$ is nonnegative, then by the Perron-Frobenius theorem there exists an eigenvalue of $I - \varphi(\hat{\xi})$ equal to 1, that is $\varphi(\hat{\xi})$ would be singular, which contradicts the assumptions.

Now we prove that $I - D(z)$ and $I - E(z)$ are invertible for any $z \in \mathbb{A}(t^{-1}, t)$. Since $|D(z)| \leqslant D(|z|)$, for the monotonicity of the spectral radius, we have $\rho(D(z)) \leqslant \rho(|D(z)|) \leqslant \rho(D(|z|))$. On the other hand, $D(|z|)$ is a principal submatrix of the nonnegative matrix $I - \varphi(|z|)$ so that $\rho(D(|z|)) \leqslant \rho(I - \varphi(|z|))$ which is less than 1 since $|z| \in \mathbb{A}(t^{-1}, t)$. We conclude that $\rho(D(z)) < 1$ for any $z \in \mathbb{A}(t^{-1}, t)$ so that $I - D(z)$ is nonsingular. The same argument can be used to deduce that $I - E(z)$ is nonsingular.

Now we prove the bound (6.11) on the singular values. For simplicity we assume that $k = 1$, the general case can be treated similarly. Since the off-diagonal blocks of A_i have rank at most 1 then $C_i = u_i v_i^*$, $i = -1, 0, 1$, for suitable vectors u_i, v_i where we assume that $\|u_i\|_2 = \|C_i\|_2$, $\|v_i\|_2 = 1$. Thus, we have $C(z) = \sum_{i=-1}^{1} z^i u_i v_i^*$. Since $I - D(z)$ is invertible on $\mathbb{A}(t^{-1}, t)$, we have

$$\tilde{C}(z) = H(z) \sum_{i=-1}^{1} z^i u_i v_i^* K(z)$$

where $H(z) = (I - D(z))^{-1}$, $K(z) = S_D(z)^{-1} = \tilde{E}(z)$, and $H(z)$, $K(z)$ are analytic for $z \in \mathbb{A}(t^{-1}, t)$. Consider the Fourier series of $H(z)$ and $K(z)$, that is, $H(z) = \sum_{s\in\mathbb{Z}} z^s H_s$, $K(z) = \sum_{s\in\mathbb{Z}} z^s K_s$, and recall that the coefficients H_s, K_s have an exponential decay, Theorem 1.4.1, that is,

$$|(H_s)_{i,j}| \leqslant \max_{z\in\mathbb{A}(t^{-1},t)} |(H(z))_{i,j}| e^{-|s|\log t}, \quad |(K_s)_{i,j}| \leqslant \max_{z\in\mathbb{A}(t^{-1},t)} |(K(z))_{i,j}| e^{-|s|\log t}.$$

Since for any matrix norm induced by an absolute norm $\|\cdot\|$ and for any matrix A it holds that $|a_{i,j}| \leqslant \|A\|$ so that we may write

$$\|H_s\| \leqslant \max_{z\in\mathbb{A}(t^{-1},t)} \|H(z)\| e^{-|s|\log t}, \quad \|K_s\| \leqslant \max_{z\in\mathbb{A}(t^{-1},t)} \|K(z)\| e^{-|s|\log t}, \tag{6.13}$$

Now recall that $\tilde{C} = H(z) \sum_{i=-1,0,1} z^i u_i v_i^* K(z)$, set $z \in \mathbb{T}$ and consider the generic ith term $z^i H(z) u_i v_i^* K(z)$ in the above summation. We have

$$z^i H(z) u_i v_i^* K(z) = \sum_{s,h\in\mathbb{Z}} z^{s+h+i} H_s u_i v_i K_h = \sum_{s\in\mathbb{Z}} H_s u_i \sum_{p\in\mathbb{Z}} z^{p+i} v_i^* K_{p-s},$$

where we have set $p = s + h$. Now, applying the operator I_N to the above matrix cancels out the terms in z^{p+i} such that $p + i$ is not multiple of N, so that we are left with the terms where $p + i = Nq$ and we get

$$I_N(z^i H(z) u_i v_i^* K(z)) = \sum_{s\in\mathbb{Z}} H_s u_i \sum_{q\in\mathbb{Z}} z^q v_i^* K_{Nq-i-s} =: \sum_{s\in\mathbb{Z}} \hat{u}_s^{(i)} \hat{v}_s^{(i)}(z),$$

for $\hat{u}_s^{(i)} = H_s u_i$, $\hat{v}_s^{(i)}(z) = \sum_{q\in\mathbb{Z}} z^q v_i^* K_{Nq-i-s}$. Thus we may write

$$I_N(\tilde{C}) = \sum_{s\in\mathbb{Z}} \hat{U}_s \hat{V}_s(z)^*, \quad \hat{U}_s = \left[\hat{u}_s^{(-1)}, \hat{u}_s^{(0)}, \hat{u}_s^{(1)}\right], \quad \hat{V}_s(z) = \left[\hat{v}_s^{(-1)}(z), \hat{v}_s^{(0)}(z), \hat{v}_s^{(1)}(z)\right].$$

To complete the proof, recall that $z \in \mathbb{T}$ and apply Lemma 4.3.1 with $k = 3$ to the series $\sum_{s\in\mathbb{Z}} \hat{U}_s V_s^*$. In order to do this, we have to provide upper bounds to $\|\hat{U}_s \hat{V}_s(z)^*\|_2$ for $z \in \mathbb{T}$. We have $\|\hat{U}_s \hat{V}_s(z)^*\|_2 \leqslant \|\hat{U}_s\|_2 \|\hat{V}_s(z)\|_2$. Concerning $\|\hat{U}_s\|_2$, since $\hat{U}_s = H_s [u_{-1}, u_0, u_1]$, we have

$$\|\hat{U}_s\|_2 \leqslant \|H_s\|_2 \| [u_{-1}, u_0, u_1] \|_2 \leqslant \sqrt{3} \|H_s\|_2 \max_i \|C_i\|_2,$$

where the latter inequality follows from the fact that $\|u_i\|_2 = \|C_i\|_2$ and that consequently, $\| [u_{-1}, u_0, u_1] \|_2 \leqslant \sqrt{3} \max_i \|C_i\|_2$. Thus from (6.13) we get

$$\|\hat{U}_s\|_2 \leqslant \sqrt{3} L \max_{z\in\mathbb{A}(t^{-1},t)} \|H(z)^{-1}\|_2 e^{-|s|\log t}.$$

Similarly, since $\|v_i\|_2 = 1$ and $|z| = 1$, we have

$$\|\hat{v}_s^{(i)}\|_2 \leqslant \sum_{q\in\boldsymbol{Z}} \|K_{Nq-i-s}\|_2 \leqslant \max_{z\in\mathbb{A}(t^{-1},t)} \|K(z)\|_2 \sum_{q\in\boldsymbol{Z}} e^{-|Nq-i-s|\log t},$$

where the last inequality follows from (6.13). Define r the remainder of the division of $i+s$ by N, so that $i+s = N\hat{q}+r$, and get

$$\begin{aligned}\sum_{q\in\boldsymbol{Z}} e^{-|Nq-i-s|\log t} &= \sum_{q\in\boldsymbol{Z}} e^{-|N(q-\hat{q})+r|\log t} = \sum_{q\in\boldsymbol{Z}} e^{-|Nq+r|\log t}\\ &= e^{-r\log t} + \sum_{q\geqslant 1} e^{-(Nq-r)\log t} + \sum_{q\geqslant 1} e^{-(Nq+r)\log t}\\ &= e^{-r\log t} + (e^{r\log t}+e^{-r\log t})\left(\frac{1}{1-e^{-N\log t}}-1\right) \leqslant \frac{2}{1-e^{-N\log t}}.\end{aligned}$$

Whence we deduce that

$$\|\hat{V}_s\|_2 \leqslant \frac{2\sqrt{3}}{1-e^{-N\log t}} \max_{z\in\mathbb{A}} \|K(z)\|_2.$$

Combining the two bounds yields

$$\|\hat{U}_s\hat{V}_s(z)\|_2 \leqslant \frac{6L}{1-e^{-N\log t}} \max_{z\in\mathbb{A}} \|K(z)\|_2 \cdot \max_{z\in\mathbb{A}} \|H(z)\|_2 \cdot e^{-|s|\log t}. \tag{6.14}$$

It remains to estimate $\|K(z)\|_2$ and $\|H(z)\|_2$. Concerning $K(z) = \tilde{E}(z)$, observe that this is a principal submatrix of $\psi(z)$ so that $\|K(z)\|_2 \leqslant \|\psi(z)\|_2$. Concerning $H(z) = (I-D(z))^{-1}$, observe that from the condition $I-A_0$ and $-A_i \geqslant 0$ for $i=-1,1$ it follows that $|D(z)| \leqslant D(|z|)$ and that

$$\rho(D(z)) \leqslant \rho(|D(z)|) \leqslant \rho(D(|z|)) \leqslant \rho(\varphi(|z|)) < 1$$

since $I-D(z)$ is a principal submatrix of $\varphi(z)$. Thus we may write $(I-D(z))^{-1} = \sum_{j=0}^{\infty} D(z)^j$ and $|(I-D(z))^{-1}| \leqslant (I-D(|z|))^{-1}$. Now, since $A_i \geqslant 0$ for $i=-1,0,1$, then

$$\begin{aligned}\tilde{D}(|z|) &= (I-D(|z|))^{-1} + \underbrace{(I-D(|z|))^{-1}}_{\geqslant 0}\underbrace{C(|z|)}_{\leqslant 0}\underbrace{S_{I-D(|z|)}^{-1}}_{\geqslant 0}\underbrace{B(|z|)}_{\leqslant 0}\underbrace{(I-D(|z|))^{-1}}_{\geqslant 0}\\ &\geqslant (I-D(|z|))^{-1},\end{aligned}$$

so that $\|(I-D(z))^{-1}\|_2 \leqslant \|(I-D(|z|))^{-1}\|_2 \leqslant \|\tilde{D}(|z|)\|_2 \leqslant \max_{z\in\mathbb{A}} \|\psi(z)\|_2$. Thus, applying Lemma 4.3.1 together with the bound (6.14) and rank of the blocks 3 yields

$$\sigma_l(\tilde{C}^{(h)}(z)) \leqslant \frac{12L\delta^2}{(1-e^{-N\log t})(1-t^{-1})} e^{-\frac{l-3}{6}\log t}.$$

If the blocks A_i are k-quasiseparable, then Lemma 4.3.1 is applied with rank of the blocks $3k$ so that the exponent $(l-3)/6$ is replaced by $(l-3k)/(6k)$, If $\varphi(z)$ is tridiagonal, then $u_{-1} = u_0 = u_1$ and $v_{-1} = v_0 = v_1$, so that $\hat{U}_j$ and $\hat{V}_j$ are formed by a single column, i.e., Lemma 4.3.1 is applied with rank of the blocks 1. This provides (6.12). □

6.5.2 *Exponential decay of the singular values in $\varphi^{(h)}(z)$*

In this section, we prove the decay property of the singular values in the off-diagonal submatrices of $\varphi^{(h)}(z)$ when $|z| = 1$. The proof is obtained by combining the decay property for the matrix function $\psi^{(h)}$, stated in Theorem 6.5.1, with a suitable lemma which allows to extend this property to the matrix inverse.

Lemma 6.5.2. *Let $\varphi^{(h)}(z) = z^{-1}A_{-1}^{(h)} + A_0^{(h)} + zA_1^{(h)}$ be the $m \times m$-matrix Laurent polynomial obtained at the h-th step of CR. Under the hypotheses of Theorem 6.5.1, for every $z \in \mathbb{T}$ we have the following bound:*

$$\sigma_l(C^{(h)}) \leqslant K(L_h, \varphi) \cdot \sigma_l(\tilde{C}^{(h)}), \qquad K(L_h, \varphi) = (1 + 3L_h)(1 + L_h + L_h^2 \|\varphi(1)^{-1}\|_2)$$

where $\varphi^{(h)}(z)$ and $\varphi^{(h)}(z)^{-1}$ are partitioned as in (6.10) *and L_h is such that $\|I - A_0^{(h)}\|_2 \leqslant L_h$ and $\|A_i^{(h)}\|_2 \leqslant L_h$ $i = -1, 1$.*

Proof. With the notation of the partitioning (6.10) applied to $\varphi^{(h)}(z)$, from Lemma 4.2.1 applied to $\varphi^{(h)}(z)$ we have

$$\sigma_l(C^{(h)}) \leqslant \|I - E^{(h)}(z)\|_2 \|S_{I-E^{(h)}}(z)\|_2 \sigma_l(\tilde{C}^{(h)}).$$

Thus, since $z \in \mathbb{T}$ and $I - E^{(h)}(z)$ is a submatrix of $\varphi^{(h)}(z)$, we have

$$\|I - E^{(h)}(z)\|_2 \leqslant \|\varphi^{(h)}(z)\|_2 \leqslant 1 + 3L_h.$$

Taking the norms in $S_{I-E^{(h)}}(z) = I - D^{(h)}(z) - C^{(h)}(z)(I - E^{(h)}(z))^{-1}B^{(h)}(z)$ we get

$$\|S_{I-E^{(h)}}(z)\|_2 \leqslant 1 + L_h + L_h^2 \|(I - E^{(h)}(z))^{-1}\|_2.$$

Moreover, for $z \in \mathbb{T}$ we have $|(I - E^{(h)}(z))^{-1}| \leqslant \sum_{i=0}^{\infty} E^{(h)}(1)^i$ so that

$$\begin{aligned}\|(I - E^{(h)}(z))^{-1}\|_2 &\leqslant \|(I - E^{(h)}(1))^{-1}\|_2 = \|\sum_{i=0}^{\infty} E^{(h)}(1)^i\|_2 \\ &\leqslant \|\sum_{i=0}^{\infty} A^{(h)}(1)^i\|_2 = \|\varphi^{(h)}(1)^{-1}\|_2,\end{aligned}$$

where we have set $A^{(h)}(z) = -z^{-1}A_{-1}^{(h)} + I - A_0^{(h)} - zA_1^{(h)}$. Here, we have used the property that the conditions $-A_{-1}^{(h)}, I - A_0^{(h)}, -A_1^{(h)} \geqslant 0$ and $\rho(I - A_{-1}^{(h)} - A_0^{(h)} - A_1^{(h)}) < 1$ are preserved at each step of CR (see [17]). Finally, since $\varphi^{(h)}(1)^{-1} = \psi^{(h)}(1) = \frac{1}{N}\sum_{i=0}^{N-1} \psi(\omega_N^i)$, for $N = 2^h$ (see Section 6.3), we have $\|\varphi^{(h)}(1)^{-1}\|_2 \leqslant \|\psi(1)\|_2$. □

Remark 6.5.3. *Note that the previous bound still holds with $\|\varphi^{(h)}(1)^{-1}\|_2$, in place of $\|\varphi(1)^{-1}\|_2$. Experimentally, $\|\varphi^{(h)}(1)^{-1}\|_2$ is much smaller than $\|\varphi(1)^{-1}\|_2$ just after few steps h.*

Observe that L_h depends on the step h of CR. However, since under the assumptions of Theorem 6.5.1, the sequences generated by CR are such that $\lim_k A_i^{(h)} = 0$, for $i = 1, -1$ while $\lim_h A_0^{(h)}$ is finite (see [17]), then there exists L such that $L \geqslant L_h$. Thus, Combining Lemma 6.5.2 and Theorem 6.5.1 we obtain the following result.

Corollary 6.5.4. *Let $\varphi^{(h)}(z) = z^{-1}A_{-1}^{(h)} + A_0^{(h)} + zA_1^{(h)}$ be the $m \times m$-matrix Laurent polynomial obtained at the h-th step of CR and assume the hypothesis of Theorem 6.5.1. Then for any off-diagonal submatrix $C^{(h)}(z)$ of $\varphi^{(h)}(z)$ we have*

$$\sigma_l(C^{(h)}) \leqslant 3MK \cdot e^{\frac{l-3k}{6k}\log t},$$

where $K = (1+3L)(1+L+L^2\|\varphi(1)^{-1}\|_2)$, M is the constant defined in Theorem 6.5.1 and $L \geqslant \|A_i^{(h)}\|_2$, for $i = -1, 0, 1$. In particular, if A_i is tridiagonal for $i = -1, 0, 1$ then $\sigma_s(C^{(h)}) \leqslant MK \cdot e^{-(\frac{s}{2})\log t}$

6.5.3 *Exponential decay of the singular values in $A_i^{(h)}$*

To prove the decay of the singular values in the off-diagonal submatrices of $A_i^{(h)}$ for $i = -1, 0, 1$ we rely on the following result of which we omit the elementary proof.

Lemma 6.5.5. *Let $\varphi(z) = z^{-1}A_{-1} + A_0 + zA_1$ and let ξ be a primitive 6-th root of the unity. Then*

$$\begin{aligned}
A_{-1} &= \frac{1}{3}\left(\xi\varphi(\xi) + \xi^5\varphi(\xi^5) - \varphi(-1)\right),\\
A_0 &= \frac{1}{2}\left(\varphi(z) + \varphi(-z)\right),\\
A_1 &= \frac{1}{3}\left(\xi^5\varphi(\xi) + \xi\varphi(\xi^5) - \varphi(-1)\right).
\end{aligned}$$

We may conclude with the decay property for the singular values of the off-diagonal submatrices of $A_i^{(h)}$, for $i = -1, 0, 1$.

Lemma 6.5.6. *Let $\varphi^{(h)}(z)$ be the matrix function generated at the hth step of CR with the property that every off-diagonal submatrix $B(z)$ of $\varphi^{(h)}(z)$ has decaying singular values such that $\sigma_s(B(z)) \leqslant \gamma e^{-\alpha s}$. Then every coefficient B_i of $B(z) = z^{-1}B_{-1} + B_0 + zB_1$ is such that $\sigma_s(B_0) \leqslant \gamma e^{-\alpha\frac{j-2}{2}}$, $\sigma_s(B_i) \leqslant \gamma e^{-\alpha\frac{j-3}{3}}$, for $i = 1, -1$.*

Proof. By Lemma 6.5.5, we have an expression for B_i based on evaluations of $B(z)$. In particular, we have $A_0 = \frac{1}{2}(B(i) + B(-i))$, $\quad A_{\pm 1} = \frac{1}{3}(\xi^{\mp 1}B(\xi) + \xi^{\mp 5}B(\xi^5) - B(-1))$, where ξ is a primitive 6-th root of the unity. Applying Lemma 4.3.3 completes the proof. □

6.5.4 *The Markovian case*

As we saw in Chapter 2, one of the application of the CR is in the Markovian framework, where applicability and convergence properties are guaranteed. In that case, the matrix function φ satisfies almost all the hypotheses made in the previous subsections but it is singular at $z = 1$ since 1 is always an eigenvalue of $\varphi(z)$. Nevertheless we will show that Corollary 6.5.4 can still be applied considering a rescaled version of $\varphi(z)$.

When the coefficients A_i for $i = -1, 0, 1$ represent the blocks of the transition matrix of an irreducible not null recurrent QBD process, the eigenvalues of $\varphi(z)$ enjoy the following properties [44, 23]:

(i) $|\lambda_1| \leqslant |\lambda_2| \leqslant \ldots \leqslant |\lambda_{m-1}| \leqslant \lambda_m < \lambda_{m+1} \leqslant |\lambda_{m+2}| \leqslant \ldots \leqslant |\lambda_{2m}|$, with $\lambda_m, \lambda_{m+1} \in \mathbb{R}$ and one of the two equal to 1.

(ii) In the annulus $\{\lambda_m < |z| < \lambda_{m+1}\}$ φ is invertible and the spectral radius of $I - \varphi(z)$ is strictly less than 1.

Hence we consider the rescaled version of φ, that is, $\varphi_\theta(z) := \varphi(\theta z)$, and we choose $\theta = \sqrt{\lambda_m \lambda_{m+1}}$. We obtain a matrix function invertible on $\mathbb{A}(t^{-1}, t)$ where $t = \sqrt{\frac{\lambda_{m+1}}{\lambda_m}}$.

Observe that $\varphi_\alpha^{(h)}(z) := \varphi^{(h)}(\alpha^{2^h} z)$ so applying CR to φ_α one obtains the same matrix sequences up to a rescaling factor. In particular the exponential decay of the singular values is left unchanged as shown in the following.

Theorem 6.5.7. *For given $t > 1$ and $\delta \geqslant 0$, consider the following class of matrix functions associated with QBD stochastic processes with k-quasiseparable blocks:*

$$\chi_{\delta,t} := \left\{ \varphi(z) : \ \|\varphi^{-1}(z)\|_2 \leqslant \delta \quad t^{-1} \leqslant |z| \leqslant t, \quad t < \lambda_{m+1}/\lambda_m \right\}.$$

Then there exists a uniform constant $\gamma(\delta, t)$ such that for any off-diagonal block $C^{(h)}(z)$ of $\varphi^{(h)}(z)$, with $\varphi \in \chi_{\delta,t}$, its l-th singular value is bounded by

$$\sigma_l(C^{(h)}(z)) \leqslant \gamma(\delta, t) \cdot e^{-\frac{l-3k}{6k} \log t}.$$

Remark 6.5.8. *Observe that in the case of null-recurrent QBD processes one has $\lambda_m = \lambda_{m+1} = 1$, so that there is no open annulus including $\mathbb{T}$ where $\varphi(z)$ is nonsingular and we can not apply Theorem 6.5.1. This drawback can be partially overcome by applying the shift technique of [17, 23]. This technique allows to construct a new matrix function $\tilde{\varphi}(z)$ which has the same eigenvalues of $\varphi(z)$ except for the eigenvalue 1 which is shifted to 0. So that $\tilde{\varphi}(z)$ has an open annulus containing $\mathbb{T}$ where it is nonsingular. Moreover, applying CR to $\tilde{\varphi}(z)$ generates matrix sequences which easily allow to recover the corresponding matrix sequences obtained by applying CR to $\varphi(z)$. The sequences associated with $\varphi(z)$ differ from the sequences associated with $\tilde{\varphi}(z)$ by a rank-1 correction. This way, if the exponential decay of the singular values holds for the latter sequences, it*

holds also for the former ones. The difficulty that still remains is that the nonnegativity of the blocks $-A_{-1}$, $I - A_0$ and $-A_1$ is not generally satisfied by the function $\tilde{\varphi}(z)$ so that in principle Theorem 6.5.1 cannot be applied and a different version specific for this case should be formulated.

6.6 REFINEMENT OF THE ANALYSIS

In the results of Section 6.5, the rate of decay is related to the width of the annulus where $\varphi(z)$ is invertible. This analysis provides an under estimate of the decay properties of the off-diagonal singular values of $A_{-1}^{(h)}$, $A_0^{(h)}$ and $A_1^{(h)}$. In fact, it turns out that, in many cases where the matrix polynomial $z^2A_1 + zA_0 + A_{-1}$ is singular at some point just outside a thin annulus $\mathbb{A}(t^{-1}, t)$ obtained with some t very close to 1, the observed exponential decay of the singular values is still evident with a basis of the exponential much smaller than the given theoretical bound t^{-1}.

A typical example is given by the discrete Laplacian matrix where $A_0 = \mathrm{trid}(-1, 4, -1)$, $A_{-1} = A_1 = -I$ so that $t^{-1} = 1 - 1/(n+1) + O(1/(n+1)^2)$. For moderately large values of n, the bound t^{-j} is still close to 1 for values of j as large as n. As a consequence, the plot of the upper bounds —coming from Section 6.5— to the singular values would be an almost horizontal line. On the other hand from the numerical experiments it turns out that the decay of the singular values is still exponential despite the width of the annulus collapses to zero, and the basis of the exponential is much less than t and almost independent of n.

In this section we provide a different theoretical explanation of the fast decay of the singular values which relies on a more detailed off-diagonal analysis of $\psi^{(z)}(z)$ and on the results of Section 4.6. Moreover, we do not require require additional hypotheses on the sign of the blocks A_is, as in the queuing problems setting.

6.6.1 *Some preliminaries*

This time, we only focus on proving that the matrix function $\psi^{(h)}(z)$, has off-diagonal blocks with singular values which decay *exponentially* to zero. The decay property can be extended to $\varphi^{(h)}(z)$ by inversion and finally to the blocks $A_{-1}^{(h)}, A_0^{(h)}, A_1^{(h)}$ whenever it is possible to provide results analogous to those of Section 6.5.2 and Section 6.5.3.

We define the following class of problems.

Definition 6.6.1. *Let $\varphi(z) = z^{-1}A_{-1} + A_0 + zA_1$, where A_{-1}, A_0, A_1 are $m \times m$ matrices with entries in $\mathbb{C}$, be such that CR can be applied with no breakdown by means of* (6.2)*. Let $f(l)$ be a positive function in $\ell^1(\mathbb{N})$. We say that $\varphi(z)$ is* f-decaying-quasiseparable *if, $\forall h \in \mathbb{N}$, $\forall z \in \mathbb{T}$ and for every off-diagonal block $\widetilde{C}^{(h)}(z)$ of $\psi^{(h)}(z)$, we have*

$$\sigma_l(\widetilde{C}^{(h)}(z)) \leqslant \|\psi^{(h)}(z)\|_2 \cdot f(l),$$

where $\sigma_l(\widetilde{C}^{(h)}(z))$ denotes the l-th singular value of the matrix $\widetilde{C}^{(h)}(z)$. We indicate the set of such matrix functions $\varphi(z)$ as $\mathrm{DQ}(f)$.

6.6.2 *Laurent coefficients of an off-diagonal block*

Let us consider the matrix Laurent series expansion of $\psi(z)$, that is, $\psi(z) = \sum_{i=-\infty}^{+\infty} z^i H_i$ for $z \in \mathbb{A}(t^{-1}, t)$, which exists and is convergent since $\psi(z)$ is analytic in the domain $\mathbb{A}(t^{-1}, t)$ because $\varphi(z)$ is analytic and nonsingular. We are going to analyze the properties of the coefficients of an off-diagonal block of this Laurent series.

Consider the following partitioning of $\psi(z)$ and $\varphi(z)$

$$\varphi(z) = \begin{pmatrix} A(z) & B(z) \\ C(z) & D(z) \end{pmatrix}, \ \psi(z) = \begin{pmatrix} \widetilde{A}(z) & \widetilde{B}(z) \\ \widetilde{C}(z) & \widetilde{D}(z) \end{pmatrix} = \begin{pmatrix} S_D(z)^{-1} & * \\ -D(z)^{-1}C(z)S_D(z)^{-1} & * \end{pmatrix},$$

where the diagonal blocks are square, $S_D(z) = A(z) - B(z)D(z)^{-1}C(z)$ is the Schur complement of $D(z)$, and $*$ denotes blocks which are not relevant for our analysis.

Moreover, suppose that the splitting (6.3) holds also for the eigenvalues of $D(z)$ –this is true for problems from stochastic models which are ruled by M-matrices– and assume that the matrix coefficients A_i have quasiseparable rank k for $i = -1, 0, 1$. This guarantees that the matrix functions $\varphi(z)$ and $D(z)$ are invertible in the annulus $\mathbb{A}(t^{-1}, t)$ for some $t > 1$.

Observe that, since the off-diagonal blocks of A_i have rank at most k for $i = -1, 0, 1$, then any off-diagonal block $C(z)$ of $\varphi(z)$ can be written as

$$C(z) = z^{-1}U_{-1}V_{-1}^* + U_0V_0^* + zU_1V_1^*, \qquad \|U_i\| = \|A_i\|, \quad \|V_i\| = 1,$$

where U_i and V_i have k columns and the superscript t denotes transposition.

Defining

$$U = \begin{bmatrix} U_{-1} & | & U_0 & | & U_1 \end{bmatrix}, \quad V(z) = \begin{bmatrix} z^{-1}V_{-1} & | & V_0 & | & zV_1 \end{bmatrix},$$

we can write $\widetilde{C}(z) = -\widetilde{U}(z)\widetilde{V}(z)^*$, where $\widetilde{U}(z) = D(z)^{-1}U$ and $\widetilde{V}(z) = S_D(z)^{-t}V(z)$. Observe that $S_D(z)^{-1}$ is the upper left diagonal block of $\psi(z)$. This gives us a crucial information on the coefficients of the matrix Laurent series expansion of $D(z)^{-1}$ and $S_D(z)^{-1}$. In order to perform this analysis we have to recall a general result which provides an explicit expression of the coefficients H_i of the Laurent expansion of $\psi(z)$.

Theorem 6.6.2 (Part of Theorem 3.20 in [17])**.** *Let $\varphi(z) = z^{-1}A_{-1} + A_0 + zA_1$ with $A_i \in \mathbb{R}^{m\times m}$, $i = -1, 0, 1$ and assume that the eigenvalues ξ_i, $i = 1, \ldots, 2m$ of $\varphi(z)$ satisfy* (6.3)*. Moreover suppose that there exist R and $\widehat{R}$ with spectral radius less than* 1 *which solve the matrix equations*

$$A_1 + XA_0 + X^2A_{-1} = 0, \tag{6.15}$$

$$X^2A_1 + XA_0 + A_{-1} = 0, \tag{6.16}$$

respectively. Then there exist G and $\widehat{G}$ solutions of the reversed matrix equations

$$A_1X^2 + A_0X + A_{-1} = 0, \tag{6.17}$$
$$A_1 + A_0X + A_{-1}X^2 = 0, \tag{6.18}$$

respectively, with spectral radius less than 1. Moreover, expanding $\varphi(z)^{-1} = \sum_{j=-\infty}^{+\infty} z^j H_j$ yields

$$H_j = \begin{cases} H_0\widehat{R}^{-j} = G^{-j}H_0 & j \leqslant 0, \\ H_0R^j = \widehat{G}^jH_0 & j \geqslant 0, \end{cases} \qquad \begin{matrix} \widehat{R} = H_0^{-1}H_{-1}, \\ R = H_0^{-1}H_1, \end{matrix} \qquad \begin{matrix} G = H_{-1}H_0^{-1}, \\ \widehat{G} = H_1H_0^{-1}. \end{matrix}$$

The spectrum of G and $\widehat{R}$ is formed by the eigenvalues of $\varphi(z)$ inside the unit disc, the spectrum of $\widehat{G}$ and R is formed by the reciprocals of the eigenvalues of $\varphi(z)$ outside the unit disc.

This result, applied with $\varphi(z) = D(z)$ and combined with what said previously, tells us that the Laurent coefficients of $\widetilde{U}(z)$ are of the form

$$D^{-1}(z) = \sum_{j\in\mathbb{Z}} z^j H_{D,j}, \qquad H_{D,j} = \begin{cases} G_D^{-j}H_{D,0} & j \leqslant 0, \\ \widehat{G}_D^{j}H_{D,0} & j \geqslant 0, \end{cases}$$

where G_D and $\widehat{G}_D$ are the solutions of the matrix equations associated with $D(z)$ of the kind (6.17) and

$$S_D(z)^{-1} = \sum_{j\in\mathbb{Z}} z^j H_{S,j}, \qquad H_{S,j} = \begin{cases} [I\ 0]H_0\widehat{R}^{-j}[I\ 0]^t & j \leqslant 0, \\ [I\ 0]H_0R^{j}[I\ 0]^t & j \geqslant 0, \end{cases}$$

where the latter equation is obtained by applying Theorem 6.6.2 to the original matrix Laurent polynomial $\varphi(z)$.

Consider the simpler case where $k = 1$ and the decomposition of each off-diagonal block $C(z)$ of $\varphi(z)$ can be written as $C(z) = uv^*$ (a constant dyad). This is not restrictive since, in the other cases, we can write $C(z)$ as a linear combination of at most $3k$ terms of the above form with coefficients z^j, $j = -1, 0, 1$.

In view of Theorem 6.6.2, for $z \in \mathbb{T}$ we can write each off-diagonal block $\widetilde{C}(z)$ of $\psi(z)$ as

$$\widetilde{C}(z) = \widetilde{u}(z)\widetilde{v}(z)^*, \qquad \widetilde{u}(z) = \sum_{j\geqslant 0} \widehat{G}_D^{j}H_{D,0}uz^j + \sum_{j<0} G_D^{-j}H_{D,0}uz^j,$$

where $\widetilde{v}(z) = S_D(z)^{-t}v$, the matrix function $S_D(z)^{-1}$ is the inverse of the Schur complement of $D(z)$ and $\|v\|_2 \leqslant 1$. Observe that the Laurent coefficients of $\widetilde{u}(z)$ corresponding to positive powers of z lie in the Krylov subspace $\mathcal{K}_j(\widehat{G}_D, H_{D,0}u)$, while the coefficients corresponding to the negative powers are in $\mathcal{K}_j(G_D, H_{D,0}u)$.

Analogously we know that

$$v^* S_D(z)^{-1} = \left(\sum_{j \geqslant 0} \widehat{v}^* H_{\psi,0} R^j z^j + \sum_{j<0} \widehat{v}^* H_{\psi,0} \widehat{R}^{-j} z^j \right) \begin{bmatrix} I \\ 0 \end{bmatrix}, \qquad \widehat{v} := \begin{bmatrix} v \\ 0 \end{bmatrix},$$

therefore

$$\begin{aligned} -\widetilde{C}(z) = &\left(\sum_{j \geqslant 0} \widehat{G}_D^j H_{D,0} u z^j + \sum_{j<0} G_D^{-j} H_{D,0} u z^j \right) \\ &\cdot \left(\sum_{j>0} \widehat{v}^* H_{\psi,0} R^j z^j + \sum_{j \leqslant 0} \widehat{v}^* H_{\psi,0} \widehat{R}^{-j} z^j \right) \begin{bmatrix} I \\ 0 \end{bmatrix}. \end{aligned} \tag{6.19}$$

Denoting by $\widetilde{C}^{(h)}(z^{2^h})$ the corresponding off-diagonal sub-block in $\psi^{(h)}$, from (6.7) we have

$$\widetilde{C}^{(h)}(z^{2^h}) = \frac{1}{2^h} \sum_{j=1}^{2^h} \widetilde{C}(z \zeta_{2^h}^j). \tag{6.20}$$

Relying on (6.19) we can prove the following result.

Lemma 6.6.3. *If $C(z) = uv^*$, then $-\widetilde{C}^{(h)}(z^{2^h})$ is the sum of the following four outer products:*

$$\begin{aligned} -\widetilde{C}^{(h)}(z^{2^h}) = \Big[&\mathcal{KM}_{2^h}(\widehat{G}_D, \widehat{a}) \cdot \mathcal{KM}_{2^h}(\widehat{R}^*, \widehat{b})^* \\ &+ z^{2^h-1} \cdot \mathcal{KM}_{2^h}(\widehat{G}_D, \widehat{a}) \cdot J_{2^h} \cdot \mathcal{KM}_{2^h}(R^*, b)^* \\ &+ z^{1-2^h} \cdot \mathcal{KM}_{2^h}(G_D, a) \cdot J_{2^h} \cdot \mathcal{KM}_{2^h}(\widehat{R}^*, \widehat{b})^* \\ &+ \mathcal{KM}_{2^h}(G_D, a) \cdot \mathcal{KM}_{2^h}(R^*, b)^* \Big] \begin{bmatrix} I \\ 0 \end{bmatrix}, \end{aligned} \tag{6.21}$$

where

$$a = \left(\sum_{s \in 2^h \mathbb{Z} \cap \mathbb{N}} z^{-s-1} G_D^{s+1} \right) H_{D,0} u, \qquad b = \left(\sum_{s \in 2^h \mathbb{Z} \cap \mathbb{N}} z^{s+1} R^{s+1} \right)^* H_{\psi,0}^* \widehat{v},$$

$$\widehat{a} = \left(\sum_{s \in 2^h \mathbb{Z} \cap \mathbb{N}} z^s \widehat{G}_D^s \right) H_{D,0} u, \qquad \widehat{b} = \left(\sum_{s \in 2^h \mathbb{Z} \cap \mathbb{N}} z^{-s} \widehat{R}^s \right)^* H_{\psi,0}^* \widehat{v}$$

and $J_{2^h} = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix} \in \mathbb{R}^{2^h \times 2^h}$ *is the counter identity.*

Proof. Thanks to (6.19) we may write $-\widetilde{C}(z)$ as the sum of four outer products. By the linearity of (6.20) we can consider them separately. Take for example

$$\left(\sum_{j\geqslant 0}\widehat{G}_D^j H_{D,0}uz^j\right)\cdot\left(\sum_{j\leqslant 0}\widehat{v}^*H_{\psi,0}\widehat{R}^{-j}z^j\right)=\sum_{j\geqslant 0}\widehat{G}_D^j H_{D,0}u\widehat{v}^*H_{\psi,0}\sum_{s\geqslant 0}\widehat{R}^s z^{j-s},$$

where we have ignored $[I\ 0]^t$ because it can be factored on the right. The block $\widetilde{C}^{(h)}(z)$ of $\psi^{(h)}(z)$ corresponding to $\widetilde{C}(z)$ in $\psi(z)$ verifies the relation $\widetilde{C}^{(h)}(z^{2^h})=\frac{1}{2^h}\sum_{l=1}^{2^h}\widetilde{C}(z\zeta_{2^h}^l)$ so that

$$\begin{aligned}
&\frac{1}{2^h}\sum_{l=1}^{2^h}\sum_{j\geqslant 0}\widehat{G}_D^j H_{D,0}u\widehat{v}^*H_{\psi,0}\sum_{s\geqslant 0}\widehat{R}^s(z\zeta_{2^h}^l)^{j-s}\\
&\quad=\sum_{j\geqslant 0}\widehat{G}_D^j H_{D,0}u\widehat{v}^*H_{\psi,0}\sum_{s\in(2^h\mathbb{Z}+j)\cap\mathbb{N}}\widehat{R}^s z^{j-s}\\
&\quad=\sum_{j=0}^{2^h-1}\left(\sum_{s\in(2^h\mathbb{Z}+j)\cap\mathbb{N}}z^s\widehat{G}_D^s\right)H_{D,0}u\widehat{v}^*H_{\psi,0}\left(\sum_{s\in(2^h\mathbb{Z}+j)\cap\mathbb{N}}\widehat{R}^s z^{-s}\right),
\end{aligned}$$

where $2^h\mathbb{Z}+j:=\{s\in\mathbb{Z}\mid s\equiv j \mod 2^h\}$. Observe that the $(j+1)$-st term of the previous sum is equal to the j-th term multiplied on the left by $z\widehat{G}_d$ and on the right by $z^{-1}\widehat{R}$. In particular we can rewrite it as

$$\left[a\,\middle|\, z\widehat{G}_D\cdot a\,\middle|\,\ldots\,\middle|\,(z\widehat{G}_D)^{2^h-1}\cdot a\right]\cdot\left[\widehat{b}\,\middle|\,(z^{-1}\widehat{R}^*)\cdot\widehat{b}\,\middle|\,\ldots\,\middle|\,(z^{-1}\widehat{R}^*)^{2^h-1}\cdot\widehat{b}\right]^*,$$

that is, $\mathcal{KM}_{2^h}(z\widehat{G}_D,a)\cdot\mathcal{KM}_{2^h}(z^{-1}\widehat{R}^*,\widehat{b})$.

The variables z in the above factors cancel out, and we obtain one of the addends in the statement of the theorem.

Then consider $\left(\sum_{j\geqslant 0}\widehat{G}_D^j H_{D,0}uz^j\right)\cdot\left(\sum_{j>0}\widehat{v}^*H_{\psi,0}R^{-j}z^j\right)$ for which we arrive at the expression

$$\sum_{j=0}^{2^h-1}\left(\sum_{s\in(2^h\mathbb{Z}+j)\cap\mathbb{N}}z^s\widehat{G}_D^s\right)H_{D,0}u\widehat{v}^*H_{\psi,0}\left(\sum_{s\in(2^h\mathbb{Z}-j)\cap\mathbb{N}}R^s z^s\right).$$

This time we have a product of the form

$$\left[a\,\middle|\, z\widehat{G}_D\cdot a\,\middle|\ldots\middle|(z\widehat{G}_D)^{2^h-1}\cdot a\,\right]\cdot\left[(zR^*)^{2^h-1}\cdot b\,\middle|\ldots\middle|(zR^*)\cdot b\,\middle|\,b\,\right]^*,$$

that is $z^{2^h-1}\cdot\mathcal{KM}_{2^h}(\widehat{G}_D,a)\cdot J_{2^h}\cdot\mathcal{KM}_{2^h}(\widehat{R}^*,\widehat{b})^*$. The other two relations are obtained in a similar manner. □

In the case $C(z)=z^s uv^*$ with $s=-1,1$ one can recover the same behavior just taking into account a shift in the powers of z in (6.19) that modifies the powers of z in the outer products accordingly.

6.6.3 *Decay in the singular values of $\psi^{(h)}(z)$*

Proposition 6.6.4. *Under the assumptions and the notation of Lemma 6.6.3 we have*

$$\widetilde{C}^{(h)}(z^{2^h}) = [I\ I] \cdot X^{(h)}(z) Y^{(h)}(z)^* \cdot [I\ I]^t \cdot [I\ 0]^t$$

where

$$X^{(h)}(z) := \begin{bmatrix} \mathcal{KM}_{2^h}(\widehat{G}_D, \widehat{a}) \\ z^{1-2^h} \mathcal{KM}_{2^h}(G_D, a) J_{2^h} \end{bmatrix}, \qquad Y^{(h)}(z) := \begin{bmatrix} z^{2^h-1} \mathcal{KM}_{2^h}(R^t, b) J_{2^h} \\ \mathcal{KM}_{2^h}(\widehat{R}^t, \widehat{b}) \end{bmatrix}.$$

Moreover, we have the following displacement relations:

$$\rho_{W_D,\Pi}(X^{(h)}) = 1, \qquad \rho_{W,\Pi}(Y^{(h)}) = 1,$$

with

$$W_D := \begin{bmatrix} \widehat{G}_D & 0 \\ 0 & G_D^\dagger \end{bmatrix}, \qquad W := \begin{bmatrix} (R^\dagger)^* & 0 \\ 0 & \widehat{R}^* \end{bmatrix}, \qquad \Pi = \begin{bmatrix} 0 & & & 1 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}$$

and the super-script † indicates the Moore-Penrose pseudoinverse.

Proof. The first claim simply follows by expanding the expression for $\widetilde{C}^{(h)}(z^{2^h})$ and by comparing it with equation (6.21). The displacement relations can be retrieved as in Example 4.6.2. □

The above result allows us to give a bound to the singular values of $\widetilde{C}^{(h)}(z)$.

Theorem 6.6.5. *Let $\varphi(z) = z^{-1}A_{-1} + A_0 + zA_1$ be an $m \times m$ matrix Laurent polynomial such that the CR —given by* (6.2)*— can be carried out with no breakdown, the splitting property* (6.3) *is verified, and $\varphi(z)$ has quasiseparable rank* 1 *for every $z \in \mathbb{T}$. Assume that the matrices R and $\widehat{R}$ which solve the matrix equations* (6.15) *are diagonalizable by means of the two eigenvector matrices V_R and $V_{\widehat{R}}$, respectively. Then $\varphi(z) \in \mathrm{DQ}(f)$ where*

$$f(l) := \gamma \cdot Z_l(E, \mathbb{T}),$$

with γ a multiple of $\max\{\kappa(V_R), \kappa(V_{\widehat{R}})\}$ and E contains the eigenvalues of $\varphi(z)$.

Proof. Notice that a generic off-diagonal matrix $\widetilde{C}^{(h)}(z)$ in $\psi^{(h)}(z)$ can be seen as a submatrix of

$$[I\ I] X^{(h)}(z) Y^{(h)}(z)^* \begin{bmatrix} I \\ I \end{bmatrix}.$$

In view of Proposition 6.6.4 we know that $Y^{(h)}(z)$ has displacement rank 1. The displacement relation for $Y^{(h)}(z)$ involves the matrices W and Π whose eigenvalues

correspond to those of $\varphi(z)$ and to the roots of the unity of order 2^h, respectively. Moreover, W is diagonalizable by means of $V_W := \mathrm{diag}(V_R^{-*}, V_{\widehat{R}}^{-*})$. Therefore, applying Corollary 4.6.4 we can write

$$\sigma_{1+l}(Y^{(h)}(z)) \leqslant Z_l(E, \mathbb{T}) \cdot \|Y^{(h)}(z)\|_2 \cdot \kappa(V_W).$$

Since W is block-diagonal we have $\kappa(V_W) = \max\{\kappa(V_R), \kappa(V_{\widehat{R}})\}$. In particular we can bound the singular values of $\widetilde{C}^{(h)}(z)$ with the quantity

$$\sigma_{1+l}(\widetilde{C}^{(h)}(z)) \leqslant 2 \cdot Z_l(E, \mathbb{T}) \cdot \|X^{(h)}(z)\|_2 \cdot \|Y^{(h)}(z)\|_2 \cdot \kappa(V_W).$$

Defining $\gamma := 2 \cdot \kappa(V_W) \cdot \max\limits_{h \in \mathbb{N}, z \in \mathbb{T}} \frac{2\|X^{(h)}(z)\|_2 \cdot \|Y^{(h)}(z)\|_2}{\|\psi^{(h)}(z)\|_2}$ we get the thesis. □

The constant γ in the previous theorem is an index of how much the factorization $X^{(h)}(z)Y^{(h)}(z)^*$ is unbalanced. This limitation is not present in the following result which describes the asymptotic behavior as $h \to \infty$. It is possible to show that the block diagonal terms in $W^{(h)}(z) = X^{(h)}(z)Y^{(h)}(z)^*$ quickly decay to 0 in practice, making the following bounds numerically accurate after a few steps.

Theorem 6.6.6. *Let $W^{(h)}(z) = X^{(h)}(z)Y^{(h)}(z)^*$, where $X^{(h)}(z)$ and $Y^{(h)}(z)$ are the matrices defined in Proposition 6.6.4. Then $\lim_{h\to\infty} W^{(h)}(z) = W^{(\infty)}$ has the following block partitioning*

$$W^{(\infty)} = \begin{bmatrix} 0 & B_1 \\ B_2 & 0 \end{bmatrix}$$

where the diagonal blocks are square and the off-diagonal blocks are independent of z. Moreover, we have $\rho_{V_D,V}(W^{(\infty)}) = 2$, where

$$V_D := \begin{bmatrix} \widehat{G}_D & 0 \\ 0 & G_D \end{bmatrix} \quad \text{and} \quad V := \begin{bmatrix} R^\dagger & 0 \\ 0 & \widehat{R}^\dagger \end{bmatrix}.$$

If the matrices $G_D, \widehat{G}_D, R$ and $\widehat{R}$ are diagonalizable by means of $V_{G_D}, V_{\widehat{G}_D}, V_R$ and $V_{\widehat{R}}$, respectively, then, indicating with $\widetilde{C}$ the off-diagonal block in $H_{\psi,0}$ corresponding to $\widetilde{C}(z)$ we have the following bounds to its singular values

$$\sigma_{1+2l}(\widetilde{C}) \leqslant \gamma \cdot Z_l(E, F), \qquad \gamma := 2 \cdot \max\{\kappa(V_G), \kappa(V_{\widehat{G}})\} \cdot \max\{\kappa(V_R), \kappa(V_{\widehat{R}})\} \cdot \|\widetilde{C}\|_2,$$

where E contains the eigenvalues of $\varphi(z)$ and $D(z)$ inside the unit disc while F contains those outside.

Proof. From the definition of $X^{(h)}$ and $Y^{(h)}$ one has

$$W^{(h)}(z) = \begin{bmatrix} z^{2^h-1}\mathcal{KM}_{2^h}(\widehat{G}_D, \widehat{a})J_{2^h}(\mathcal{KM}_{2^h}(R^*, b))^* & \mathcal{KM}_{2^h}(\widehat{G}_D, \widehat{a})(\mathcal{KM}_{2^h}(\widehat{R}^*, \widehat{b}))^* \\ \mathcal{KM}_{2^h}(G_D, a)(\mathcal{KM}_{2^h}(R^*, b))^* & z^{1-2^h}\mathcal{KM}_{2^h}(G_D, a)J_{2^h}(\mathcal{KM}_{2^h}(\widehat{R}^*, \widehat{b}))^* \end{bmatrix}.$$

Since the spectral radii of the matrices R, $\widehat{R}$, G_D and $\widehat{G}_D$ are less than 1 and in view of Theorem 4.4.3, the block diagonal entries of $W^{(h)}$ tend to zero as $h \to \infty$. Instead, the two off-diagonal blocks have limits B_1 and B_2, respectively. More precisely

$$W^{(\infty)} = \begin{bmatrix} 0 & B_1 \\ B_2 & 0 \end{bmatrix}, \quad B_1 = \sum_{i \geqslant 0} \widehat{G}_D^i \widehat{a}\widehat{b}^* \widehat{R}^i, \quad B_2 = \sum_{i \geqslant 0} G_D^i a b^* R^i.$$

Thus, tracing the argument of Example 4.6.2, we have

$$\widehat{G}_D B_1 - B_1 \widehat{R}^\dagger = \widehat{a}\widehat{b}^* \widehat{R}^\dagger.$$

An analogous computation for B_2 gives the rank-2 displacement. The matrix $\widetilde{C}$ can be written as $\widetilde{C} = [I\ I] \cdot W^{(\infty)} \cdot [I\ I]^t \cdot [I\ 0]$, which corresponds to an off-diagonal block of $\lim_{h\to\infty} \psi^{(h)}(z)$. Due to the recurrence relation $\psi^{(h+1)}(z^2) = \frac{1}{2}(\psi^{(h)}(z) + \psi^{(h)}(-z))$, this limit is equal to the central coefficient $H_{\psi,0}$ in the series expansion of $\psi(z)$. The thesis follows by applying Corollary 4.6.4. □

6.6.4 *Experimental validation of the results*

This section is devoted to verify the previous results by means of numerical experiments. We do that by computing numerical estimates of the bound given in Theorem 6.6.6 together with the singular values of the off-diagonal blocks of $H_{\psi,0}$. The actual bounds are obtained by choosing a particular family of rational functions that suit the considered problem. We will see that, even if our choices are relatively simple, and not optimal, they already provide sharp decay bounds in practice.

As a first example, we consider instances of the problem coming from the framework of Markov chains i.e., the sum $I - A_{-1} - A_0 - A_1$ is sub-stochastic, that is, it has non-negative entries and the sum along each row is at most 1. In particular, the matrices $-A_{-1}, I - A_0$ and $-A_1$ have non negative entries and are scaled in order to satisfy the splitting assumption (6.3) (see Section 6.5.4).

We select dense 300×300-blocks generated at random and such that $\varphi(z)$ is of quasiseparable rank 1. For satisfying the latter hypothesis we impose that the strictly triangular parts of the blocks are the restrictions of dyads with the same left vectors.

We divide the resulting distribution of the eigenvalues in three cluster. One is contained in a neighborhood of 0, another is in the complement of the disc of radius 4 and finally we have two eigenvalues close to 1, λ_1 and λ_2, inside and outside the unit circle, respectively.

Motivated by this, we choose the sequence of rational function

$$r_l(z) := \frac{z - \lambda_1}{z - \lambda_2} z^{l-1},$$

for roughly estimating the Zolotarev problem. The results are shown in Figure 6.3.

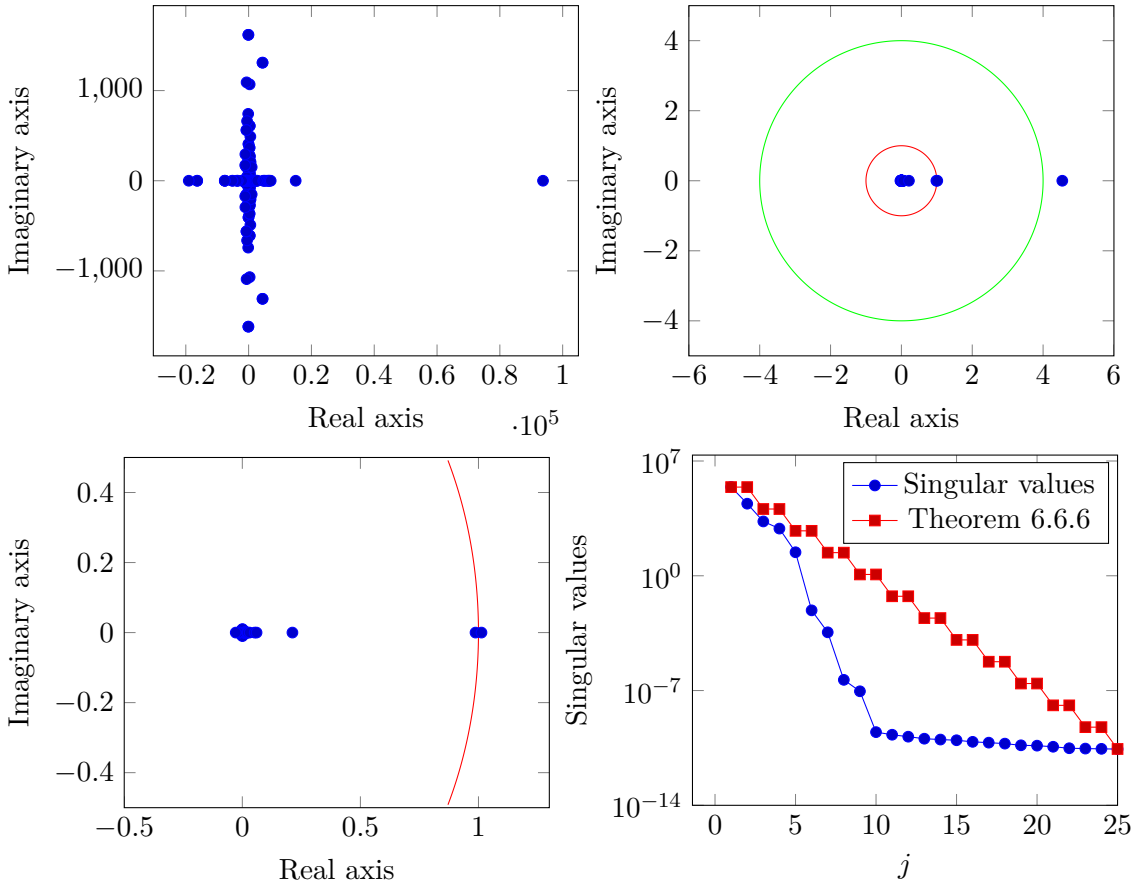

Figure 6.3.: In the top and in the lower left figures the distribution of the eigenvalues. In the lower right image the singular values decay in $H_0(151:300, 1:150)$ and the bound given by Theorem 6.6.6.

As a second example, we consider the linear system arising from the discretization of a 2D Poisson equation, whose matrix is block tridiagonal with the following form:

$$T = \begin{bmatrix} \tilde{A} & \tilde{C} & & & \\ B & A & C & & \\ & \ddots & \ddots & \ddots & \\ & & B & A & C \\ & & & \hat{B} & \hat{A} \end{bmatrix}, \qquad A = \begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{bmatrix}, \qquad B = C = -I.$$

The above system can be solved by means of the cyclic reduction. The eigenvalues of the associated $\varphi(z)$ can be computed explicitly and one can easily check that they are real positive and provide a splitting $t^{-1} = 1 - 1/(n+1) + O(1/(n+1)^2)$. The matrices A, B and C are very special instances of 1-quasiseparable matrices, so we can state a refined version of Theorem 6.6.6, which gives a smaller displacement rank for the limit case.

Proposition 6.6.7. *Let A, B and C as above, and $W^{(\infty)}(z)$ as defined in Theorem 6.6.6. If $\tilde{C}$ is one off-diagonal block of $H_{\psi,0}$ then*

$$\sigma_{1+l}(\widetilde{C}) \leqslant \gamma \cdot Z_l(E, F), \qquad \gamma := 2 \cdot \|\widetilde{C}\|_2.$$

Proof. Due to the symmetry properties of the coefficients A, B and C, and to the palindromicity of $\phi(z)$ and $D(z)$, we have

$$G_D = \hat{G}_D, \qquad R = \hat{R}, \qquad a = \hat{a}, \qquad b = \hat{b}.$$

In this way, we find that the matrix $\begin{bmatrix} I & I \end{bmatrix} W^{(\infty)} \begin{bmatrix} I \\ I \end{bmatrix}$ satisfies a displacement relation of rank 1 with the same matrices of Theorem 6.6.6. Therefore, the bound on the singular values holds with l instead of $2l$. Moreover, since the matrices $G, \hat{G}, R$ and $\hat{R}$ can be diagonalized by means of orthogonal matrices, the maximum of their spectral conditioning is 1. □

In order to verify the bound for this example we have carried out CR until convergence on a 200×200 example and we have plotted the singular values of an off-diagonal block of the computed H_0. Then, we have estimated the bound coming from Proposition 6.6.7 using a rational function of this form:

$$r_l(z) := (z - \delta) \prod_{j=1}^{l-1} \frac{z - q_j}{z - p_j}.$$

The points q_j and p_j are chosen with a greedy approach as the maximizer and minimizer of $r_{l-1}(z)$ in the sets E and F respectively. The point δ is the rightmost eigenvalue of $\varphi(z)$ inside the unit disc. The bound is compared with the one coming from Theorem 4.6.5 and with the one from Theorem 6.5.1. The results are reported in Figure 6.4.

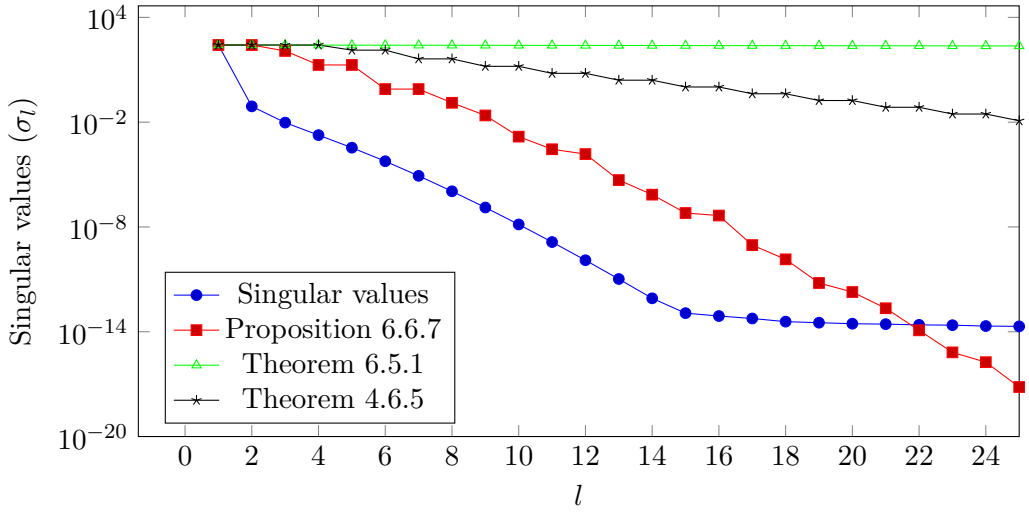

Figure 6.4.: Decay of the singular values in one of the off-diagonal blocks of H_0 in the Laurent expansion of $\psi(z)$, computed by means of the CR for the Poisson matrix. We have reported the actual decay and the bounds obtained by means of Proposition 6.6.7, the results in Theorem 6.5.1, and Theorem 4.6.5.

In this case the bound from Theorem 6.5.1 is useless since the approach used there relies on a wide splitting of the eigenvalues of $\varphi(z)$. It is also interesting to note that even if the bound of Theorem 4.6.5 is optimal for real intervals an ad hoc choice for the approximant in a discrete set can deliver better results.

As a last example, we analyze a case where our bounds do not indicate that the intermediate $\varphi^{(h)}(z)$ are numerically quasiseparable. Consider the coefficients A_i defined as follows:

$$A_{-1} := \rho\Pi, \quad A_0 := I + \rho^2\Pi^2, \quad A_1 := \rho\Pi.$$

In this case the coefficients A_i are 1-quasiseparable and we know that $\varphi(z)$ can be factored as $\varphi(z) = (zI - \rho\Pi)(z^{-1}I - \rho\Pi)$. Therefore, the eigenvalues of $\varphi(z)$ all lie on the circles of radii ρ and ρ^{-1}. Moreover, the solution $G = \Pi$ is itself 1-quasiseparable. Choosing a value for $\rho \approx 1$ yields a very slow decay in our theoretical bounds. As shown in Figure 6.5, where we have chosen $n = 500$, and $\rho = 1 - 10^{-6}$, the numerical quasiseparable structure is not present in the intermediate $\varphi^{(h)}(z)$. In fact, after a few steps, the off-diagonal blocks of $A_i^{(h)}$ have almost full rank.

In this case, the use of HODLR matrices is not convenient, even though the original coefficients and the solution are efficiently representable in this format.

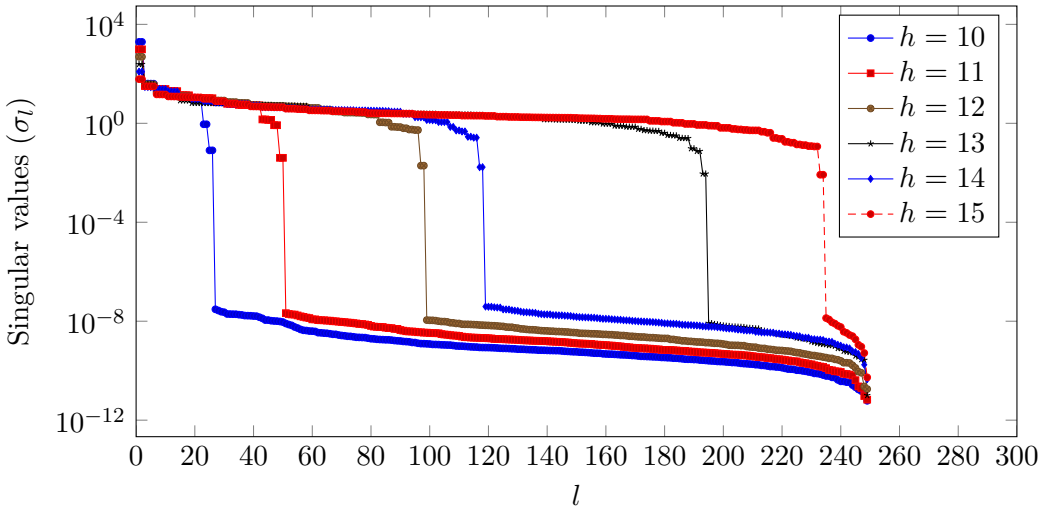

Figure 6.5.: Decay of the singular values in the off-diagonal block $A_0^{(h)}(\frac{n}{2}+1:n, 1:\frac{n}{2})$ in a case with a slow guaranteed decay. Here n is equal to 500. The decay is reported for different intermediate steps h.

6.7 USING CR WITH THE HODLR REPRESENTATION

In this section we test the CR with the hierarchical representation which for notational simplicity we refer to as Quasiseparable Cyclic Reduction (QCR for short).

For the implementation of this algorithm we relied on the open source library `H2Lib` [25]. The library has been wrapped in MEX files for use in MATLAB, where the numerical experiments have been run. The code developed in this context is freely available at [75]. The bindings developed in the testing of the algorithm are only a partial mapping of all the routines available in the original `H2Lib` library but we feel that it is worth making them public so they can be used as a base for a further extension.

6.7.1 *Solving quadratic matrix equations*

Here, we address the problem of solving the quadratic matrix equation

$$A_{-1} + A_0 X + A_1 X^2 = 0 \tag{6.22}$$

where the $m \times m$-matrices A_is have a low quasiseparable rank. For simplicity, we consider the queuing problems settings in which applicability and convergence of the CR are guaranteed. To be precise we have $-A_{-1}, I - A_0$ and A_1 non negative and the matrix $I - A_{-1} - A_0 - A_1$ stochastic.

	CR		$QCR_{10^{-16}}$		$QCR_{10^{-12}}$		$QCR_{10^{-8}}$	
Size	Time (s)	Residue	Time (s)	Residue	Time (s)	Residue	Time (s)	Residue
100	$6.04e-02$	$1.91e-16$	$2.21e-01$	$1.79e-15$	$2.04e-01$	$8.26e-14$	$1.92e-01$	$7.40e-10$
200	$1.88e-01$	$2.51e-16$	$5.78e-01$	$1.39e-14$	$5.03e-01$	$1.01e-13$	$4.29e-01$	$2.29e-09$
400	$1.61e+01$	$2.09e-16$	$3.32e+00$	$1.41e-14$	$2.60e+00$	$1.33e-13$	$1.98e+00$	$1.99e-09$
800	$2.63e+01$	$2.74e-16$	$4.55e+00$	$1.94e-14$	$3.49e+00$	$2.71e-13$	$2.63e+00$	$2.69e-09$
1600	$8.12e+01$	$3.82e-12$	$1.18e+01$	$3.82e-12$	$8.78e+00$	$3.82e-12$	$6.24e+00$	$3.39e-09$
3200	$6.35e+02$	$5.46e-08$	$3.12e+01$	$5.46e-08$	$2.21e+01$	$5.46e-08$	$1.51e+01$	$5.43e-08$
6400	$5.03e+03$	$3.89e-08$	$7.83e+01$	$3.89e-08$	$5.38e+01$	$3.89e-08$	$3.58e+01$	$3.87e-08$
12800	$4.06e+04$	$1.99e-08$	$1.94e+02$	$1.99e-08$	$1.29e+02$	$1.99e-08$	$8.37e+01$	$1.97e-08$

Table 6.1.: Timings and accuracy for 15 iterations of CR at the increasing of the size of the blocks.

Numerical results

For a fair comparison, we have compiled `H2Lib` with the LAPACK library used by MATLAB. Moreover, we have disabled the parallelism in the Intel MKL library to obtain more accurate results. It is important to notice that running with parallelism enabled in the MKL library leads to improved performance both for `H2LIb` and for MATLAB, but the improvement is more relevant in the latter. This is due to the fact that the library is optimized for the multiplication of large matrices, such as in the full CR implementation (when full matrices of large size are multiplied together). The multiplication of the small rectangular matrices involved in the hierarchical representation, instead, benefit less from this implementation. Anyway, also in this case we see that our implementation is more efficient even if starting from larger dimension. For example, on a Xeon server with 24 threads available our implementation is faster than the standard one approximately for $n > 500$.

Table 6.1 reports the results of some numerical experiments, where in each column we have reported: the size of the blocks from $m = 100$ up to $m = 12800$, the CPU time, in seconds, required by standard CR and the residual error, then from column 3 to column 5 we reported the CPU time, in seconds, and the residual error of our implementation with values of $\epsilon = 10^{-16}$, 10^{-12}, 10^{-8}, respectively. It is interesting to observe that the precision of the result does not deteriorate much for large values of m. Moreover, the speed-up that we get goes beyond two orders of magnitude.

In Table 6.2 we repeat the experiment fixing the size to 1600 and letting the band of the starting blocks to increase exponentially from 2 up to 128. It should be noted that the gain of time of our implementation seems to deteriorate linearly with respect to the increase of the band.

In Figures 6.6-6.7 we give a graphic description, in logarithmic scale, of the growth of the CPU time in the latter experiments. The test problems are generated randomly.

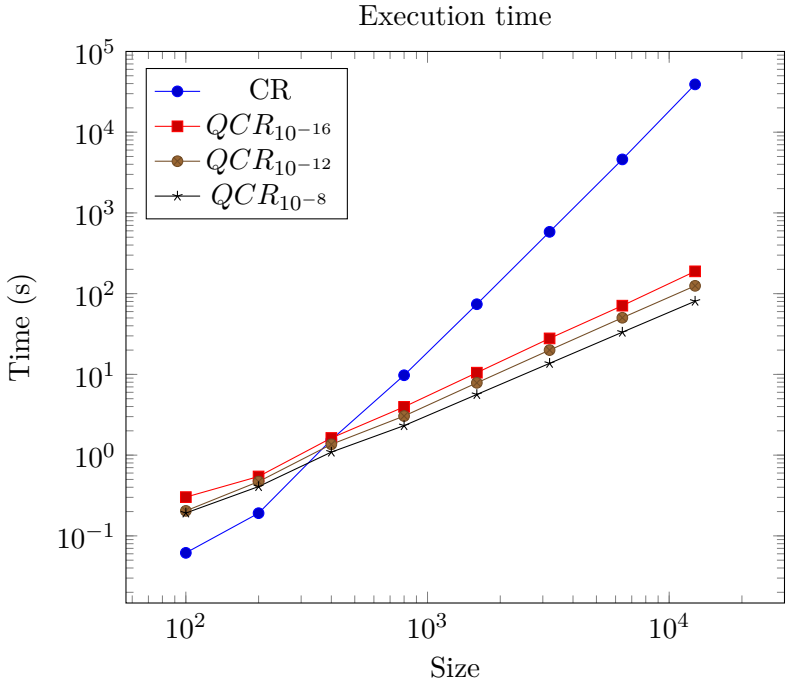

Figure 6.6.: Timings of the different implementations of CR. The algorithms are applied to tridiagonal blocks with increasing size.

6.7.2 *Solving certain generalized Sylvester equations*

For an $m \times n$ matrix X denote $x = \mathrm{vec}(X)$ the mn-vector obtained by stacking the columns of X. Then, for any pair of matrices A, B of compatible sizes, one has $\mathrm{vec}(AB) = (I \otimes A)\mathrm{vec}(B) = (B^t \otimes I)\mathrm{vec}(A)$.

Consider the linear matrix equation

$$\sum_{i=1}^{s} A_i X B_i = C, \qquad A_i \in \mathbb{R}^{m\times m},\ B_i \in \mathbb{R}^{n\times n},\ X, C \in \mathbb{R}^{m\times n}, \tag{6.23}$$

and suppose that B_i, $i = 1, \ldots, s$ are tridiagonal Toeplitz matrices.

Applying the vec operator on both sides of (6.23) we get the $mn \times mn$ linear system

$$Wx = c, \qquad W = \sum_{i=1}^{s} B_i^t \otimes A_i, \quad x = \mathrm{vec}(X), \quad c = \mathrm{vec}(C). \tag{6.24}$$

Since each term $B_i^t \otimes A_i$ is block tridiagonal and block Toeplitz, then the coefficient matrix of (6.24) is block tridiagonal, block Toeplitz as well. If the matrices A_i are k_i-quasiseparable then the blocks of W are k-quasiseparable with $k = \sum_{i=1}^{s} k_i$. If k is

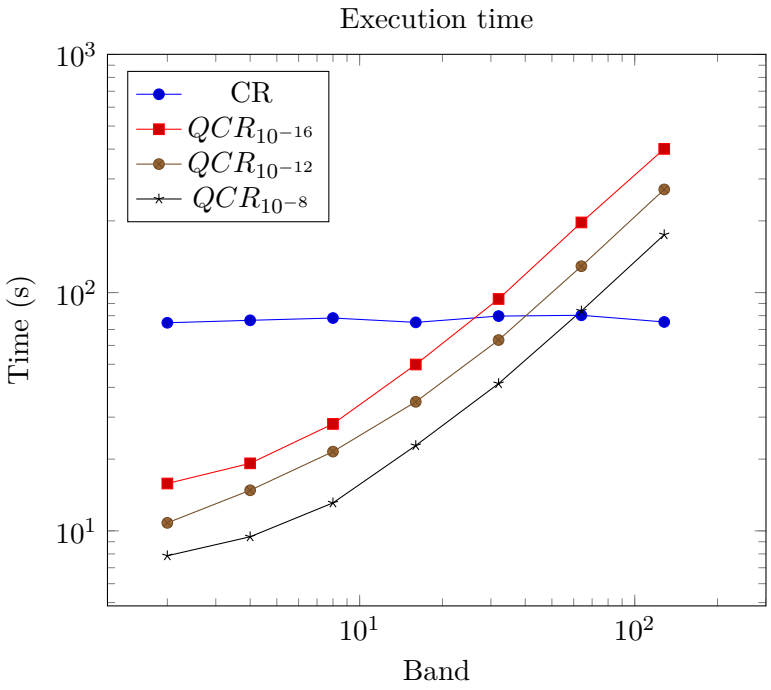

Figure 6.7.: Timings of the different implementations of CR. The algorithms are applied to band blocks with increasing band and size 1600.

	CR		$QCR_{10^{-16}}$		$QCR_{10^{-12}}$		$QCR_{10^{-8}}$	
Band	Time (s)	Residue	Time (s)	Residue	Time (s)	Residue	Time (s)	Residue
2	$7.47e+01$	$2.11e-16$	$1.58e+01$	$6.95e-15$	$1.08e+01$	$2.62e-13$	$7.86e+00$	$2.57e-09$
4	$7.65e+01$	$1.66e-16$	$1.92e+01$	$4.88e-15$	$1.48e+01$	$2.36e-13$	$9.44e+00$	$3.15e-09$
8	$7.82e+01$	$1.48e-16$	$2.81e+01$	$6.11e-15$	$2.15e+01$	$2.08e-13$	$1.31e+01$	$2.10e-09$
16	$7.50e+01$	$1.35e-16$	$4.99e+01$	$4.98e-15$	$3.48e+01$	$2.29e-13$	$2.28e+01$	$2.08e-09$
32	$7.97e+01$	$1.33e-16$	$9.40e+01$	$5.79e-15$	$6.32e+01$	$2.01e-13$	$4.15e+01$	$2.28e-09$
64	$8.03e+01$	$1.31e-16$	$1.97e+02$	$6.79e-15$	$1.29e+02$	$1.99e-13$	$8.37e+01$	$2.01e-09$
128	$7.53e+01$	$1.28e-16$	$4.01e+02$	$5.89e-15$	$2.71e+02$	$2.02e-13$	$1.75e+02$	$2.15e-09$

Table 6.2.: Timings and accuracy for 15 iterations of CR on blocks with size 1600 with different bands.

negligible with respect to m then we may solve the generalized Sylvester equation by means of quasiseparable CR.

Numerical results

A possible application of this algorithm is solving discretized partial differential equations coming from convection diffusion problems of the form

$$-\epsilon\Delta u(x,y) + \mathbf{w}\cdot\nabla u(x,y) = f(x,y), \qquad \Omega \subset \mathbb{R}^2 \tag{6.25}$$

where $u(x,y)$ is the unknown function, and we assume that the convection vector $\mathbf{w}$ depends only on one of the two coordinates. For simplicity we assume that it only depends on x. According to [83] we can discretize the above problem obtaining the following Sylvester equation in the matrix unknown U:

$$\epsilon T_1 U + \epsilon U T_2 + \Phi_1 B_1 U + \Phi_2 U B_2 = F.$$

The independence on y of the convection vector ensures that all the right factors in the previous equation are almost Toeplitz. The matrices Φ_i are diagonal while T_i and B_i arise from the discretization of the differential operators and they are all tridiagonal and Toeplitz with the exceptions of the first and last rows (due to the boundary conditions). The matrix F contains the evaluations of the function f on the discretized grid. We refer to [83] for an in depth analysis.

We performed some numerical tests on one of the example in [83] namely (6.25) with $\epsilon = 0.0333$ and $\mathbf{w} = (1 + \frac{(x+1)^2}{4}, 0)$. Since in this case $\Phi_2 = 0$ the problem is reduce to solving the Sylvester equation

$$(\epsilon T_1 + \Phi_1 B_1)U + U\epsilon T_2 = F.$$

In Figure 6.8 we compare the timings and the residue with those of the function `lyap` from the control toolbox of MATLAB R2013a. Note that our approach can be applied even if the second coordinate of $\mathbf{w}$ is non zero and dependent on x. In fact, in this way we retrieve a generalized Sylvester equation that can be solved with this algorithm.

6.8 CONCLUSIONS AND RESEARCH LINES

In this chapter we have provided different perspectives about the preservation of quasiseparability in the iterative scheme of CR, analyzing both the exact and the approximate structure.

The connection between the phenomenon and the existence of accurate solutions of certain discrete rational approximation problems have been pointed out.

The application to solve large scale unilateral quadratic matrix equations arising in the study of QBD processes, has been presented. Also examples related to the solution of Sylvester equations arising in the discretization of elliptic PDEs, have been shown.

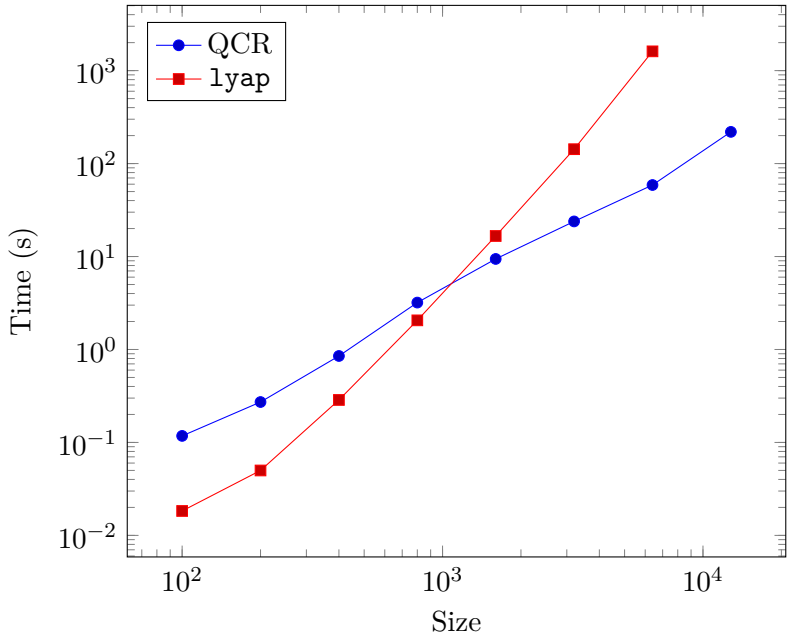

Figure 6.8.: Timings of the quasiseparable cyclic reduction (QCR) and the Sylvester solver implemented in the `lyap` function in MATLAB.

Size	T_{QCR} (s)	Res_{QCR}	T_{lyap} (s)	Res_{lyap}
100	0.12	$2.16 \cdot 10^{-13}$	$1.83 \cdot 10^{-2}$	$1.18 \cdot 10^{-12}$
200	0.27	$1.54 \cdot 10^{-12}$	$4.99 \cdot 10^{-2}$	$5.56 \cdot 10^{-12}$
400	0.85	$5.53 \cdot 10^{-12}$	0.29	$5.17 \cdot 10^{-11}$
800	3.2	$4.19 \cdot 10^{-11}$	2.06	$9.04 \cdot 10^{-11}$
1,600	9.42	$1.25 \cdot 10^{-10}$	16.63	$5.64 \cdot 10^{-10}$
3,200	23.86	$6.78 \cdot 10^{-10}$	142.78	$2.06 \cdot 10^{-9}$
6,400	58.79	$2.41 \cdot 10^{-9}$	1,612	$2.98 \cdot 10^{-8}$
12,800	219.27	$7.8 \cdot 10^{-9}$	—	—

Table 6.3.: Timings and residues of the Sylvester equation solved by means of the quasiseparable cyclic reduction (QCR) and the Sylvester solver implemented in the `lyap` function in MATLAB. The residues are computed by evaluating $\|\epsilon T_1 U + \epsilon U T_2 + \Phi_1 B_1 U - D\|_2$.

The use of CR, together with HODLR representation, has been tested on such examples, confirming a substantial speed up in retrieving the solution.

Moreover, applications to solving certain generalized Sylvester equations, of the form

$$\sum_{i=1}^{k} A_i X B_i = D,$$

have been analyzed in the case where all the blocks B_i are tridiagonal Toeplitz (possibly with only the first and last row with different entries), and the A_is have a low quasiseparable rank. Under these hypotheses, and the assumption that the sum of the quasiseparable ranks of the blocks A_i is negligible compared to m, the complexity of the method is $O(m^2 \log m)$.

A strictly related theme is the solution of non-symmetric Riccati equations (NAREs)

$$C + XA + DX - XBX = 0,$$

with rank structured coefficients. In fact, by means of the Cayley transformation, it is possible to rephrase such issue as solving a unilateral quadratic matrix equation and employ the CR [16][Section 2.7]. This topic will be subject of future research.

Chapter 7

Semi-infinite quasi-Toeplitz matrix computation

Semi-infinite matrices can be viewed as linear operators on Banach spaces of one-sided infinite sequences. Classical examples of those sequences are the spaces $\ell^p(\mathbb{C})$ of complex-valued sequences $\{c_n\}_{n\in\mathbb{N}}$ such that $\sum_{n\in\mathbb{N}} |c_n|^p \leqslant \infty$, with $p \in \mathbb{Z}^+$.

The typical approaches for treating linear algebra issue involving infinite data structures, rely on truncation. For example, a strategy to solve an infinite linear system consists in selecting a finite section of the coefficient matrix and of the right-hand side and solving the finite linear system associated. Then, hope that the outcome well approximates a finite part of the solution of the original problem, assuming a sufficiently large initial section. In [73] the author analyzes when this approach is feasible. Similar techniques can be adopted for solving matrix equations or computing matrix functions, but —in general— there is no guarantee of success. In [72, 6, 70] bad effects of truncation are highlighted when solving infinite quadratic matrix equations arising in the Markov chains framework.

Here we want to consider an alternative perspective: keep the infinite size of the data and look for structures that allow a finite representation at arbitrary precision.

7.1 Dealing with an infinite amount of data

Obviously, the subset of infinite matrices that we can handle in an exact way are those that can be represented with a finite number of parameters. Among the latter, we start by considering banded Toeplitz matrices, i.e., matrices of the kind $T = (t_{i,j})$ such that $t_{i,j} = a_{j-i}$ for some sequence $\{a_k\}_{k\in\mathbb{Z}}$ with only finitely many nonzero entries. This structure is ubiquitous in the applications where some sort of shift invariance property is satisfied by the underlying mathematical model.

However, even with this class the implementative part is problematic because performing arithmetic operations between Toeplitz matrices, for instance computing the inverse, causes the loss of sparsity and of the Toeplitz structure. Therefore, despite in the initial stage the data can be represented with a finite number of parameters, we apparently need to store an infinite amount of entries in order to carry on any algorithm based on elementary matrix operations. A way out of this drawback is to relax our requests by asking the existence of a finitely generated approximation of the outcome, at any arbitrary precision. In particular, we focus on matrices that can be decomposed as the sum of a Toeplitz matrix associated with a sequence $\{a_k\}_{k\in\mathbb{Z}}$ such that $\lim_{k\to\pm\infty} a_k = 0$ and a semi-infinite matrix with a decay in the modulus of its entries, along every direction. If the decay in the two addends is sufficiently fast we get an object that is well approximated —in some norm— by a banded Toeplitz matrix plus a matrix with only a finite number of nonzero entries.

7.2 PRELIMINARIES

Recall that if $a(z) = \sum_{i\in\mathbb{Z}} a_i z^i$ is analytic in the annulus $\mathbb{A}(r, R)$, for some $r < 1 < R$, then for any $\epsilon > 0$ there exists a constant $\gamma > 0$ such that

$$|a_i| \leqslant \gamma(R-\epsilon)^{-i}, \quad |a_{-i}| \leqslant \gamma(r+\epsilon)^i, \quad i \in \mathbb{Z}^+ \tag{7.1}$$

(see Theorem 1.4.1). The exponential decay of the bounds (7.1) implies that $\sum_{i\in\mathbb{Z}} |a_i| < +\infty$, that is $a(z) \in \mathcal{W}$, and also that

$$\sum_{i\in\mathbb{Z}^+} i|a_i| < +\infty, \quad \sum_{i\in\mathbb{Z}^+} i|a_{-i}| < +\infty. \tag{7.2}$$

The latter inequalities will be particularly useful in the next section where we define the class of quasi-Toeplitz matrices.

Notice that, if $a(z)$ is analytic and nonzero over $\mathbb{A}(r, R)$, then the reciprocal function $a(z)^{-1}$ is well defined and analytic over $\mathbb{A}(r, R)$ so that $a^{-1}(z) = \sum_{i\in\mathbb{Z}} \widetilde{a}_i z^i$ and the analogous of equation (7.2) holds true for the Fourier coefficients of $a(z)^{-1}$.

In the following, we denote by $a^+(z)$ and by $a^-(z)$ the power series defined by the coefficients of $a(z)$ with positive and with negative powers, respectively, that is, $a^+(z) = \sum_{i\in\mathbb{Z}^+} a_i z^i$ and $a^-(z) = \sum_{i\in\mathbb{Z}^+} a_{-i} z^i$, so that $a(z) = a_0 + a^+(z) + a^-(z^{-1})$.

We associate with the functions $a(z)$, $a^+(z)$ and $a^-(z)$ the following semi-infinite matrices

$$\begin{aligned} &T(a) = (t_{i,j})_{i,j}, && t_{i,j} = a_{j-i}, \\ &H(a^+) = (h^+_{i,j})_{i,j}, && h^+_{i,j} = a_{i+j-1}, \qquad i,j \in \mathbb{Z}^+, \\ &H(a^-) = (h^-_{i,j})_{i,j}, && h^-_{i,j} = a_{-i-j+1}, \end{aligned}$$

i.e., $T(a)$ is the Toeplitz matrix associated with the function $a(z)$, while $H(a^+)$ and $H(a^-)$ are the Hankel matrices associated with the functions $a^+(z)$ and $a^-(z)$, respectively. The function $a(z)$ is called the *symbol* of the Toeplitz matrix $T(a)$.

Finally, denote by $\mathcal{F}$ the class of semi-infinite matrices $F = (f_{i,j})_{i,j\in\mathbb{Z}^+}$ such that $\|F\|_{\mathcal{F}} := \sum_{i,j\in\mathbb{Z}^+} |f_{i,j}|$ is finite. The norm that we use in this case is just the 1-norm if we look at the matrix F as an infinite vector.

Observe that $\mathcal{F}$ is a vector space, closed under rows-by-columns multiplication, and $\|F\|_{\mathcal{F}}$ is a norm over $\mathcal{F}$ which is endowed of the sub-multiplicative property. In the following, we write $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ to denote the linear space $\mathcal{F}$ endowed with the norm $\|\cdot\|_{\mathcal{F}}$. We have the following:

Lemma 7.2.1. *$(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ equipped with matrix sum and multiplication is a Banach algebra over $\mathbb{C}$.*

Proof. We need to show that given $E, F \in \mathcal{F}$ and $\alpha \in \mathbb{C}$ it holds

(i) $\alpha E \in \mathcal{F}$,

(ii) $E + F \in \mathcal{F}$,

(iii) $EF \in \mathcal{F}$ and in particular $\|EF\|_{\mathcal{F}} \leqslant \|E\|_{\mathcal{F}}\|F\|_{\mathcal{F}}$,

(iv) $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ is a complete metric space.

Clearly, $\sum_{i,j\in\mathbb{Z}^+} |\alpha e_{i,j}| = |\alpha| \sum_{i,j\in\mathbb{Z}^+} |e_{i,j}| < +\infty$ which proves (i). By the triangular inequality one obtains that $\sum_{i,j\in\mathbb{Z}^+} |e_{i,j} + f_{i,j}| \leqslant \sum_{i,j\in\mathbb{Z}^+} |e_{i,j}| + \sum_{i,j\in\mathbb{Z}^+} |f_{i,j}| < +\infty$ which implies (ii). If $H = EF = (h_{i,j})$ then $h_{i,j} = \sum_{r\in\mathbb{Z}^+} e_{i,r} f_{r,j}$ so that, defining $\alpha_r = \sum_{i\in\mathbb{Z}^+} |e_{i,r}|$, and $\beta_r = \sum_{j\in\mathbb{Z}^+} |f_{r,j}|$, for the quantity $\|EF\|_{\mathcal{F}} = \sum_{i,j\in\mathbb{Z}^+} |h_{i,j}|$ we have

$$\|EF\|_{\mathcal{F}} \leqslant \sum_{i,j,r\in\mathbb{Z}^+} |e_{i,r}| \cdot |f_{r,j}| = \sum_{r\in\mathbb{Z}^+} \alpha_r \beta_r \leqslant \left(\sum_{r\in\mathbb{Z}^+} \alpha_r\right)\left(\sum_{r\in\mathbb{Z}^+} \beta_r\right) = \|E\|_{\mathcal{F}} \cdot \|F\|_{\mathcal{F}},$$

which shows (iii). Finally, we observe that any matrix $E \in \mathcal{F}$ can be viewed as a vector $v = (v_i)_{i\in\mathbb{Z}^+}$ obtained by ordering the entries $e_{i,j}$ along the anti-diagonals, starting from $e_{1,1}$, followed by the entries $e_{i,j}$ with indices that have sum $3, 4, 5, \ldots$, and so on. Moreover, the norm $\|\cdot\|_{\mathcal{F}}$ corresponds to the ℓ^1 norm in the space of infinite sequences having finite sum of their moduli. This way, the space $\mathcal{F}$ actually coincides with ℓ^1, which is a Banach space. Thus, we get (iv). □

Observe that the condition $\|F\|_{\mathcal{F}} < +\infty$ implies that for any $\epsilon > 0$ there exists an integer $k > 0$ such that $\sum_{i,j\geqslant k} |f_{i,j}| < \epsilon$, that is, the entries of the matrix F decay to zero as $i, j \to \infty$ so that F can be approximated with an arbitrarily small error by a finite matrix. This property ensures that we can represent F with a finite number of parameters up to an error which is negligible with respect to the roundoff error.

Any semi-infinite matrix $S = (s_{i,j})_{i,j\in\mathbb{Z}^+}$ can be viewed as a linear operator, acting on semi-infinite vectors $v = (v_i)_{i\in\mathbb{Z}^+}$, which maps the vector v onto the vector u such that $u_i = \sum_{j\in\mathbb{Z}^+} s_{i,j} v_j$, provided that the results of the summations are finite.

Indeed, the matrices $F \in \mathcal{F}$ define linear operators on the space ℓ^1 of semi-infinite vectors $v = (v_i)$ such that $\|v\|_1 = \sum_{i\in\mathbb{Z}^+} |v_i|$ is finite, since

$$\sum_{i\in\mathbb{Z}^+} |\sum_{j\in\mathbb{Z}^+} f_{i,j}v_j| \leqslant \sum_{i,j\in\mathbb{Z}^+} |f_{i,j}v_j| \leqslant \sum_{i,j\in\mathbb{Z}^+} |f_{i,j}| \cdot \sup_k |v_k|$$

which is finite as the product of two finite terms.

For any integer $p \geqslant 1$, we may wonder if also the matrices $T(a)$, $H(a^+)$ and $H(a^-)$ define linear operators acting on the Banach space ℓ^p formed by vectors v such that the ℓ^p norm $\|v\|_p = (\sum_{i\in\mathbb{Z}^+} |v_i|^p)^{1/p}$ is finite. In this case we may evaluate the p-norm of the operator S (operator norm) as $\|S\|_p := \sup_{\|v\|_p=1} \|Sv\|_p$. The answer to this question is given by the following result of [28] which relates the matrix $T(a)T(b)$ with $T(ab)$, $H(a^-)$ and $H(a^+)$.

Theorem 7.2.2. *For $a(z), b(z) \in \mathcal{W}$ let $c(z) = a(z)b(z)$. Then we have*

$$T(a)T(b) = T(c) - H(a^-)H(b^+).$$

Moreover, for any $a(z) \in \mathcal{W}$ and for any $p \geqslant 1$, including $p = \infty$, we have

$$\|T(a)\|_p \leqslant \|a\|_{\mathcal{W}}, \quad \|H(a^-)\|_p \leqslant \|a^-\|_{\mathcal{W}}, \quad \|H(a^+)\|_p \leqslant \|a^+\|_{\mathcal{W}}.$$

A direct consequence of the above result is that the product of two Toeplitz matrices can be written as a Toeplitz matrix plus a correction whose ℓ^p-norm is bounded by $\|a\|_{\mathcal{W}}\|b\|_{\mathcal{W}}$.

A similar property holds for matrix inversion in the case where the function $a(z)$ is nonzero for $|z| = 1$ and its winding number is zero. In fact, in this case we may apply another classical result (we refer to the book [27] for more details) which relates the invertibility of the operator $T(a)$ to the winding number of $a(z)$, that is, the (integer) number of times that the complex number $a(\cos\theta + \boldsymbol{i}\sin\theta)$ winds around the origin as θ moves fro 0 to 2π.

Theorem 7.2.3 (Gohberg 1952)**.** *Let $a(z)$ be a continuous function from $\mathbb{T}$ in $\mathbb{C}$. Then the linear operator $T(a)$ is invertible if and only if the winding number of $a(z)$ is zero and $a(z)$ does not vanishes on $\mathbb{T}$.*

Thus, under the assumptions of the above theorem, it follows that $T(a)$ is invertible and we have [28, Proposition 1.18]

$$T(a)^{-1} = T(a^{-1}) + E,$$

where $\|E\|_p$ is bounded from above by a constant.

In the analysis that we are going to perform in the next section, the above properties concerning the ℓ^p norms are very useful, but are not enough to arrive at an algorithmic implementation concerning Toeplitz and quasi-Toeplitz matrices. In fact, our request

is to write the product and the inverse of Toeplitz matrices as a Toeplitz matrix plus a correction whose entries have a decay along every direction. Mathematically, this means to give conditions under which $E = H(a^-)H(b^+) \in \mathcal{F}$.

Finally, we recall a result concerning the Wiener-Hopf factorization of $a(z)$ which will be useful next.

Theorem 7.2.4. *Let $a(z) \in \mathcal{W}$ be a function which does not vanish for $z \in \mathbb{T}$ and such that its winding number is κ. Then $a(z)$ admits the Wiener-Hopf factorization*

$$a(z) = u(z)z^{\kappa}\ell(z),$$

where $u(z) = \sum_{i=0}^{\infty} u_i z^i$, $\ell(z) = \sum_{i=0}^{\infty} \ell_i z^{-i}$ are in $\mathcal{W}$ and $u(z)$, $\ell(z^{-1})$ do not vanish in the closed unit disc. If $k = 0$ the factorization is said canonical.

7.3 QUASI-TOEPLITZ MATRICES

In this section we introduce the classes of quasi-Toeplitz matrices and analyze their properties.

Definition 7.3.1. *We say that the semi-infinite matrix A is a* quasi-Toeplitz matrix (QT-matrix) *if it can be written in the form*

$$A = T(a) + E,$$

where $a(z) = \sum_{i=-\infty}^{+\infty} a_i z^i$ is in the Wiener class, and $E = (e_{i,j}) \in \mathcal{F}$. We refer to $T(a)$ as the Toeplitz part of A, and to E as the correction. We denote by $\mathcal{QT}$ the class of QT-matrices. Moreover we define the following norm on $\mathcal{QT}$

$$\|T(a) + E\|_{\mathcal{QT}} := \|a\|_{\mathcal{W}} + \|E\|_{\mathcal{F}}.$$

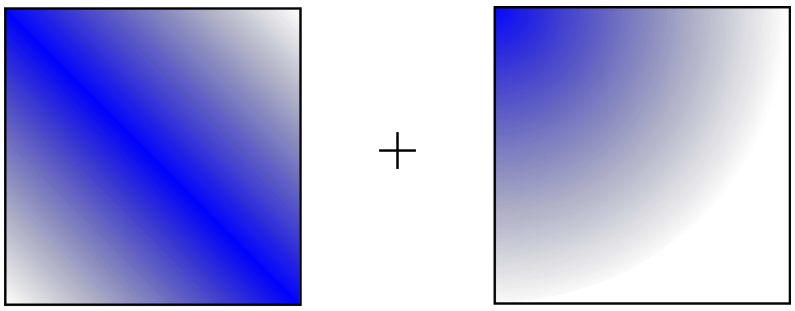

Figure 7.1.: Graphic description of the $\mathcal{QT}$ structure; the intensity of the color indicates the magnitude of the absolute values of the entry

Observe that given $A \in \mathcal{QT}$ there is a unique way to decompose it in the sense of Definition 7.3.1. In fact, suppose by contradiction that there exist $a_1(z), a_2(z) \in \mathcal{W}$ and $E_1, E_2 \in \mathcal{F}$ with $a_1 \neq a_2$ and $E_1 \neq E_2$ such that

$$A = T(a_1) + E_1 = T(a_2) + E_2.$$

Then we should have $E_1 - E_2 = T(a_2) - T(a_1) = T(a_2 - a_1)$, hence $\|E_1 - E_2\|_{\mathcal{F}} = \|T(a_2 - a_1)\|_{\mathcal{F}}$. On the other hand, since $T(a_2 - a_1) \neq 0$ we have $\|T(a_2 - a_1)\|_{\mathcal{F}} = \infty$, which contradicts the fact that $E_1 - E_2 \in \mathcal{F}$.

Lemma 7.3.2. *The set $\mathcal{QT}$ endowed with the norm $\|\cdot\|_{\mathcal{QT}}$ is a Banach space.*

Proof. The set of quasi-Toeplitz matrices is clearly isomorphic to the direct sum $\mathcal{QT} \simeq \mathcal{W} \oplus \mathcal{F}$. Since both $\mathcal{W}$ and $\mathcal{F}$ are Banach spaces the composition of the 1-norm of $\mathbb{R}^2$ with the vector valued function $T(a) + E \to (\|a\|_{\mathcal{W}}, \|E\|_{\mathcal{F}})$ makes $\mathcal{W} \oplus \mathcal{F}$ a complete metric space. □

The class $\mathcal{QT}$ clearly includes all the matrices encountered in QBD processes, formed by a banded Toeplitz part, and by a correction E such that $e_{i,j} = 0$ for $i, j > k$ for some integer k.

The goal of this section is to prove that the subclass of QT-matrices associated with continuous symbols $a(z)$ such that $a'(z) \in \mathcal{W}$ form a normed matrix algebra, i.e., a vector space closed under matrix multiplication. To this end, it is useful to introduce the following sub-algebra of $\mathcal{W}$.

Definition 7.3.3. *We denote $\mathcal{W}_1 = \{a(z) \in \mathcal{W} : \quad a(z)$ continuous, and $a'(z) \in \mathcal{W}\}$, and define the norm*

$$\|a\|_{\mathcal{W}_1} = \|a\|_{\mathcal{W}} + \|a'\|_{\mathcal{W}}.$$

We recall that $\mathcal{W}_1$ is a Banach algebra with the norm $\|a\|_{\mathcal{W}_1}$, see [29].

Definition 7.3.4. *We call CQT-matrix, any matrix $T(a) + E \in \mathcal{QT}$ such that the symbol $a(z) \in \mathcal{W}_1$. We denote by $\mathcal{CQT}$ the subset of $\mathcal{QT}$ formed by CQT-matrices. Moreover, we define the following norm in $\mathcal{CQT}$:*

$$\|T(a) + E\|_{\mathcal{CQT}} := \|a\|_{\mathcal{W}_1} + \|E\|_{\mathcal{F}}.$$

Definition 7.3.5. *We call AQT-matrices the subset of CQT matrices whose symbol is analytic. We denote this set with $\mathcal{AQT}$.*

Next, we provide a few results which are useful to prove that $\mathcal{CQT}$ is a Banach algebra. The following lemma shows that the product of two semi-infinite Toeplitz matrices associated with symbols in $\mathcal{W}_1$ belongs to $\mathcal{CQT}$.

Lemma 7.3.6. *Let $a(z), b(z) \in \mathcal{W}_1$ and set $c(z) = a(z)b(z)$. Then $T(a)T(b) = T(c) + E_c$ where $E_c \in \mathcal{F}$, moreover,*

$$\|E_c\|_{\mathcal{F}} \leqslant \|H(a^-)\|_{\mathcal{F}} \cdot \|H(b^+)\|_{\mathcal{F}} = \sum_{i\in\mathbb{Z}^+} i|a_{-i}| \sum_{i\in\mathbb{Z}^+} i|b_i|.$$

Proof. From Theorem 7.2.2 we deduce that $T(a)T(b) = T(c) + E_c$ where we set $E_c = -H(a^-)H(b^+)$. Let us prove that $H(a^-), H(b^+) \in \mathcal{F}$. We have $\|H(b^+)\|_{\mathcal{F}} = \sum_{i,j\in\mathbb{Z}^+} |b_{i+j-1}|$. Setting $k = i + j - 1$ we may write $\|H(b^+)\|_{\mathcal{F}} = \sum_{k\in\mathbb{Z}^+} k|b_k|$ which is finite since $b(z) \in \mathcal{W}_1$. The same argument applies to $H(a^-)$. In view of Lemma 7.2.1, $\mathcal{F}$ is a normed matrix algebra therefore $\|E_c\|_{\mathcal{F}} \leqslant \|H(a^-)\|_{\mathcal{F}} \cdot \|H(b^+)\|_{\mathcal{F}} < +\infty$. □

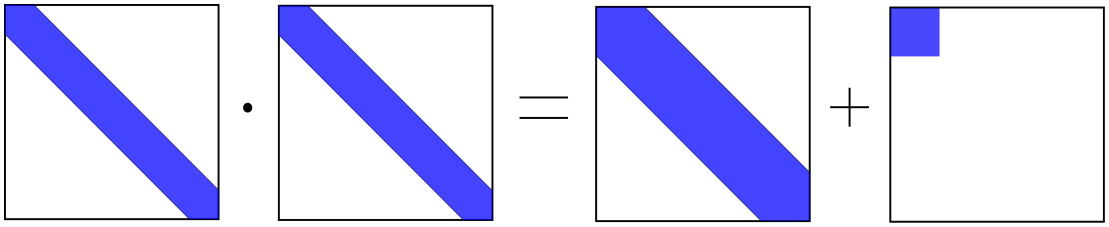

Figure 7.2.: Multiplication of two semi-infinite banded Toeplitz matrices

Remark 7.3.7. Observe that the quantities $\sum_{i\in\mathbb{Z}^+} i|a_{-i}|$ and $\sum_{i\in\mathbb{Z}^+} i|b_i|$ coincide with the $\mathcal{W}$-norms of the first derivatives of the functions $a^-(z)$ and $b^+(z)$, respectively. This way we may rewrite the bound given in Lemma 7.3.6 as

$$\|E_c\|_{\mathcal{F}} \leqslant \|(a^-)'\|_{\mathcal{W}} \|(b^+)'\|_{\mathcal{W}} \leqslant \|a'\|_{\mathcal{W}} \|b'\|_{\mathcal{W}}. \tag{7.3}$$

The condition $a(z), b(z) \in \mathcal{W}_1$ is needed to prove Lemma 7.3.6, as it is demonstrated by the following example. Consider the case where $a(z) = \sum_{i=0}^{+\infty} a_{-i} z^{-i}$, $b(z) = \sum_{i=0}^{+\infty} b_i z^i$, $a_{-i} = b_i = i^{-3/2}$. Clearly $a(z), b(z) \in \mathcal{W}$ but $a(z)'$ and $b(z)'$ are not in $\mathcal{W}$ since $\sum_{i\in\mathbb{Z}^+} i a_{-i}$ and $\sum_{i\in\mathbb{Z}^+} i b_i$ are not convergent. Moreover,

$$\|H(a^-)H(b^+)\|_{\mathcal{F}} = \sum_{i,j\in\mathbb{Z}^+} \sum_{r=0}^{+\infty} \frac{1}{(i+r)^{3/2}} \frac{1}{(r+j)^{3/2}} = \sum_{r=0}^{+\infty} \left(\sum_{k=r+1}^{+\infty} \frac{1}{k^{3/2}} \right)^2.$$

This is the sum of the squares of the remainders of the series $\sum_{i=1}^{+\infty} \frac{1}{i^{3/2}}$. This sum diverges since these remainders behave like $\int_r^{+\infty} \frac{1}{x^{3/2}} dx = \frac{2}{\sqrt{r}}$.

Now we can prove the main result of this section which states that $\mathcal{CQT}$ is closed under multiplication.

Theorem 7.3.8. *Let $A, B \in \mathcal{CQT}$, where $A = T(a) + E_a$, $B = T(b) + E_b$. Then we have $C = AB = T(c) + E_c \in \mathcal{CQT}$ with $c(z) = a(z)b(z)$. Moreover,*

$$\|E_c\|_{\mathcal{F}} \leqslant \|H(a^-)\|_{\mathcal{F}} \cdot \|H(b^+)\|_{\mathcal{F}} + \|a\|_{\mathcal{W}} \|E_b\|_{\mathcal{F}} + \|b\|_{\mathcal{W}} \|E_a\|_{\mathcal{F}} + \|E_a\|_{\mathcal{F}} \cdot \|E_b\|_{\mathcal{F}}.$$

Proof. We have $C = AB = (T(a) + E_a)(T(b) + E_b)$. Applying Theorem 7.2.2 yields

$$C = T(c) - H(a^-)H(b^+) + T(a)E_b + E_aT(b) + E_aE_b =: T(c) + E_c,$$

where

$$E_c = -H(a^-)H(b^+) + T(a)E_b + E_aT(b) + E_aE_b. \tag{7.4}$$

Therefore, it is sufficient to prove that $\|E_c\|_{\mathcal{F}}$ is finite. From Lemmas 7.3.6 and 7.2.1 it follows that both $\|H(a^-)H(b^+)\|_{\mathcal{F}}$ and $\|E_aE_b\|_{\mathcal{F}}$ are finite. It remains to show that $\|E_aT(b)\|_{\mathcal{F}}$ and $\|T(a)E_b\|_{\mathcal{F}}$ are finite. We prove this property only for $\|T(a)E_b\|_{\mathcal{F}}$ since the boundedness of the other matrix norm follows by transposition. In fact, for any $F \in \mathcal{F}$ one has $\|F\|_{\mathcal{F}} = \|F^t\|_{\mathcal{F}}$ and $T(a)^t = T(\widehat{a})$ where $\widehat{a}(z) = a(z^{-1})$ and $\|a\|_{\mathcal{W}} = \|\widehat{a}\|_{\mathcal{W}}$. Denote $H = T(a)E_b = (h_{i,j})$ and $E_b = (e_{i,j})$. We have $h_{i,j} = \sum_{r=1}^{+\infty} a_{r-i}e_{r,j}$ so that

$$\|H\|_{\mathcal{F}} = \sum_{i,j\in\mathbb{Z}^+} |h_{i,j}| \leqslant \sum_{i,j\in\mathbb{Z}^+} \sum_{r=1}^{+\infty} |a_{r-i}e_{r,j}|.$$

Substituting $k = r - i$ yields

$$\|H\|_{\mathcal{F}} \leqslant \sum_{k\in\mathbb{Z}} |a_k| \sum_{j=1}^{+\infty} \sum_{i=-k+1}^{+\infty} |e_{k+i,j}|.$$

Since $\sum_{j=1}^{+\infty}\sum_{i=-k+1}^{+\infty} |e_{k+i,j}| = \sum_{j=1}^{+\infty}\sum_{i=1}^{+\infty} |e_{i,j}| = \|E_b\|_{\mathcal{F}}$ for any k, we have

$$\|H\|_{\mathcal{F}} \leqslant \sum_{k\in\mathbb{Z}} |a_k| \|E_b\|_{\mathcal{F}} = \|a\|_{\mathcal{W}}\|E_b\|_{\mathcal{F}} < +\infty.$$

Thus, taking norms in (7.4) yields

$$\|E_c\|_{\mathcal{F}} \leqslant \|H(a^-)\|_{\mathcal{F}} \cdot \|H(b^+)\|_{\mathcal{F}} + \|a\|_{\mathcal{W}}\|E_b\|_{\mathcal{F}} + \|E_a\|_{\mathcal{F}} \cdot \|b\|_{\mathcal{W}} + \|E_a\|_{\mathcal{F}} \cdot \|E_b\|_{\mathcal{F}}$$

which completes the proof. □

Observe that in view of Remark 7.3.7 we may write

$$\|E_c\|_{\mathcal{F}} \leqslant \|a'\|_{\mathcal{W}}\|b'\|_{\mathcal{W}} + \|a\|_{\mathcal{W}}\|E_b\|_{\mathcal{F}} + \|E_a\|_{\mathcal{F}} \cdot \|b\|_{\mathcal{W}} + \|E_a\|_{\mathcal{F}} \cdot \|E_b\|_{\mathcal{F}}. \tag{7.5}$$

Now, our next goal is to prove that the class $\mathcal{CQT}$ is a Banach algebra.

Theorem 7.3.9. *The class $\mathcal{CQT}$ equipped with the norm $\|\cdot\|_{\mathcal{CQT}}$ is a Banach algebra over $\mathbb{C}$. Moreover $\|AB\|_{\mathcal{CQT}} \leqslant \|A\|_{\mathcal{CQT}}\|B\|_{\mathcal{CQT}}$ for any matrices $A, B \in \mathcal{CQT}$.*

Proof. Theorem 7.3.8 ensures the closure of $\mathcal{CQT}$ under matrix multiplication. To prove the sub-multiplicative property of the norm, i.e.,

$$\|AB\|_{\mathcal{CQT}} \leqslant \|A\|_{\mathcal{CQT}} \cdot \|B\|_{\mathcal{CQT}}$$

for any $A, B \in \mathcal{CQT}$, $A = T(a) + E_a$, $B = T(b) + E_b$, observe that

$$\begin{aligned} \|ab\|_{\mathcal{W}_1} =& \|ab\|_{\mathcal{W}} + \|(ab)'\|_{\mathcal{W}} = \|ab\|_{\mathcal{W}} + \|a'b + ab'\|_{\mathcal{W}} \\ \leqslant& \|a\|_{\mathcal{W}}\|b\|_{\mathcal{W}} + \|a'\|_{\mathcal{W}}\|b\|_{\mathcal{W}} + \|a\|_{\mathcal{W}}\|b'\|_{\mathcal{W}}. \end{aligned} \tag{7.6}$$

Since $\|AB\|_{\mathcal{CQT}} = \|ab\|_{\mathcal{W}_1} + \|E_c\|_{\mathcal{F}}$, for $c(z) = a(z)b(z)$, and where E_c is defined as in Theorem 7.3.8, by applying (7.5) and (7.6) we obtain

$$\begin{aligned} \|AB\|_{\mathcal{CQT}} &\leqslant \|ab\|_{\mathcal{W}_1} + \|a'\|_{\mathcal{W}}\|b'\|_{\mathcal{W}} + \|a\|_{\mathcal{W}}\|E_b\|_{\mathcal{F}} + \|b\|_{\mathcal{W}}\|E_a\|_{\mathcal{F}} + \|E_a\|_{\mathcal{F}}\|E_b\|_{\mathcal{F}} \\ &\leqslant \|a\|_{\mathcal{W}}\|b\|_{\mathcal{W}} + \|a'\|_{\mathcal{W}}\|b\|_{\mathcal{W}} + \|a\|_{\mathcal{W}}\|b'\|_{\mathcal{W}} + \|a'\|_{\mathcal{W}}\|b'\|_{\mathcal{W}} + \|a\|_{\mathcal{W}}\|E_b\|_{\mathcal{F}} \\ &\quad + \|b\|_{\mathcal{W}}\|E_a\|_{\mathcal{F}} + \|E_a\|_{\mathcal{F}}\|E_b\|_{\mathcal{F}} \\ &= (\|a\|_{\mathcal{W}} + \|a'\|_{\mathcal{W}})(\|b\|_{\mathcal{W}} + \|b'\|_{\mathcal{W}}) + \|a\|_{\mathcal{W}}\|E_b\|_{\mathcal{F}} + \|b\|_{\mathcal{W}}\|E_a\|_{\mathcal{F}} + \|E_a\|_{\mathcal{F}}\|E_b\|_{\mathcal{F}} \\ &\leqslant (\|a\|_{\mathcal{W}_1} + \|E_a\|_{\mathcal{F}})(\|b\|_{\mathcal{W}_1} + \|E_b\|_{\mathcal{F}}) \\ &= \|A\|_{\mathcal{CQT}}\|B\|_{\mathcal{CQT}}. \end{aligned}$$

Concerning the completeness, observe that the set of CQT matrices is isomorphic to the direct sum $\mathcal{CQT} \simeq \mathcal{W}_1 \oplus \mathcal{F}$. Since both $\mathcal{W}_1$ and $\mathcal{F}$ are Banach spaces, the composition of the 1-norm of $\mathbb{R}^2$ with the vector valued function $T(a) + E \to (\|a\|_{\mathcal{W}_1}, \|E\|_{\mathcal{F}})$ makes $\mathcal{W}_1 \oplus \mathcal{F}$ a complete metric space.

□

Remark 7.3.10. It is interesting to notice that $\mathcal{AQT}$ with the norm $\|\cdot\|_{\mathcal{CQT}}$ is not Banach. In fact, consider the sequence of semi-infinite Toeplitz matrices $\{T(a_n)\}$ with $a_n(z) = \sum_{j=1}^{n} \frac{1}{j^3} z^j$, and observe that this is a Cauchy sequence in $\mathcal{AQT}$ with the norm $\|\cdot\|_{\mathcal{CQT}}$, but its limit does not belong to $\mathcal{AQT}$ because the corresponding symbol $a(z) = \sum_{j=1}^{\infty} \frac{1}{j^3} z^j$ is not analytic. On the other hand, the completeness of $\mathcal{CQT}$ implies that any Cauchy sequence in $\mathcal{AQT}$ admits limit in $\mathcal{CQT}$. Therefore, we can claim that the limit of a Cauchy sequence in $\mathcal{AQT}$ can still be represented —at an arbitrary precision— with a finite number of parameters.

Since $\mathcal{CQT}$ is a normed matrix algebra, if $A \in \mathcal{CQT}$ and B is an infinite matrix such that $BA = AB = I$, then $B \in \mathcal{CQT}$. In the next section we represent the inverse matrix of an infinite Toeplitz matrix $T(a)$ in terms of the Wiener-Hopf factorization of $a(z)$.

7.3.1 *Inverse of a CQT-matrix*

Assume that $a(z) \in \mathcal{W}_1$ does not vanishes on the unit circle and its winding number is zero, so that in view of Theorem 7.2.4 there exists the canonical Wiener-Hopf factorization $a(z) = u(z)\ell(z)$. From this factorization we deduce the following matrix factorization

$$T(a) = T(u)T(\ell),$$

where $T(\ell)$ is lower triangular and $T(u)$ is upper triangular. Since $u(z)$ and $\ell(z^{-1})$ do not vanish in the unit disc, the functions $u(z)$ and $\ell(z)$ have inverse in $\mathcal{W}_1$, by Theorem 7.2.2 are such that $T(u)T(u^{-1}) = T(u^{-1})T(u) = I$, and $T(\ell)T(\ell^{-1}) = T(\ell^{-1})T(\ell) = I$, so that

$$T(a)^{-1} = T(\ell)^{-1}T(u)^{-1} = T(\ell^{-1})T(u^{-1}).$$

In view of Lemma 7.3.6, we have

$$T(a)^{-1} = T(a^{-1}) - H((\ell^{-1})^{-})H((u^{-1})^{+}) = T(a^{-1}) - H(\ell^{-1})H(u^{-1}) \in \mathcal{CQT}. \quad (7.7)$$

That is, a semi-infinite Toeplitz matrix associated with a symbol $a(z) \in \mathcal{W}_1$, with null winding number, which does not annihilates in $\mathbb{T}$, is invertible and its inverse is a CQT-matrix.

This fact, together with the available algorithms to compute the Wiener-Hopf factorization of $a(z)$, enables us to implement the inversion of CQT-matrices in a very efficient manner. We will see this in the next section.

7.4 CQT MATRIX ARITHMETIC

The properties that we have described in the previous sections imply that any finite computation which takes as input a set of CQT-matrices and that performs matrix additions, multiplications, inversions, and multiplications by a scalar, generates results that belong to $\mathcal{CQT}$. If the computation can be carried out with no breakdown, say caused by singularity, then the output still belongs to $\mathcal{CQT}$.

This observation makes it possible to compute functions of semi-infinite CQT-matrices in an efficient way or to solve quadratic matrix equations where the coefficients are CQT-matrices. In order to do that, we have to provide a simple and effective way of representing, up to an arbitrarily small error, CQT-matrices by means of a finite number of parameters. This is done in this section.

Given a QT-matrix $A = T(a) + E_a$, since the symbol $a(z)$ belongs to the Wiener class, and since the correction matrix E_a has entries with finite sum of their moduli, we may write A through its *truncated form* $\widetilde{A} = \mathrm{trunc}(A)$. That is, for any $\epsilon > 0$ there exist integers n_-, n_+, k_-, k_+ such that

$$\begin{aligned} &A = \widetilde{A} + \mathcal{E}_a, \quad \|\mathcal{E}_a\|_{\mathcal{QT}} \leqslant \epsilon, \\ &\widetilde{A} = T(\widetilde{a}) + \widetilde{E}_a, \\ &\widetilde{a}(z) = \sum_{i=-n_-}^{n_+} a_i z^i, \end{aligned} \quad (7.8)$$

where $\widetilde{E}_a = (\widetilde{e}_{i,j})$, is such that $\widetilde{e}_{i,j} = e_{i,j}$ for $i = 1, \ldots, k_-$, $j = 1, \ldots, k_+$, while $\widetilde{e}_{i,j} = 0$ elsewhere.

In this way, we can approximate any given QT-matrix A, to any desired precision, with a CQT-matrix $\widetilde{A}$ where the Toeplitz part is banded and the correction $\widetilde{E}_a$ has a finite dimensional nonzero part. The CQT-matrix $\widetilde{A}$ can be easily stored with a finite number of memory locations. The "finite approximation" $\widetilde{A}$ of a QT-matrix A is the computational counterpart with which we are going to work in practice.

Observe that, if $A \in \mathcal{CQT}$ and the symbol $a(z)$ is analytic, for the exponential decay of the coefficients $|a_i|$, the values of $n_\pm$ are $O(\log \epsilon^{-1})$. Concerning the values of $k_\pm$, unless we make additional assumptions on the decay of the entries $|e_{i,j}|$ as i, j tend to infinity, the values that $k_\pm$ can assume are as large as $1/\epsilon$. Think for instance to the case where $e_{i,j} = 1/(i+j)^p$ for $p > 2$ where $k_\pm$ are of the order of $1/\epsilon^{p-1}$. The same qualitative bounds hold for the coefficients a_i if we simply assume that $a(z) \in \mathcal{W}_1$.

Here and in the sequel, we do not care much to give *a priori* bounds to the values of $n_\pm$ and $k_\pm$ since these values can be determined automatically at run time during the computation.

Another observation concerns the truncated correction $\widetilde{E}_a$. In fact, from the computational point of view, it is convenient to express the matrix $\widetilde{E}_a$ by means of a factorization of the kind $\widetilde{E}_a = F_a G_a^t$, where matrices F_a and G_a have a number of columns given by the rank of $\widetilde{E}_a$ and infinitely many rows. In this way, in presence of low-rank corrections, the storage is reduced together with the computational cost for performing matrix arithmetic. This representation in product form can be obtained by means of SVD up to some error which can be controlled at run time and which can be included in $\mathcal{E}_a$. Observe also that the truncation operates both on the function $a(z)$ and in the correction E_a by means of compression.

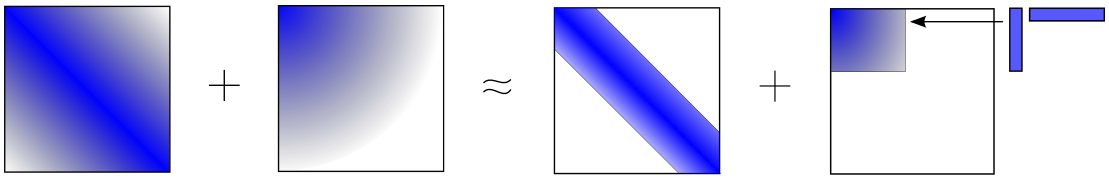

Figure 7.3.: Pictorial description of the representation of QT-matrices

In the following, we represent a QT-matrix $A = T(a) + E_a$ in the form (7.8) with $\widetilde{E}_a = F_a G_a^t$ where F_a has f_a nonzero rows and k_a columns, G_a has g_a nonzero rows and k_a columns, and the error $\mathcal{E}_a$ has a sufficiently small norm. This way, $\widetilde{E}_a$ has f_a nonzero rows, g_a nonzero columns and rank at most k_a.

With this notation we may easily implement the operations of addition, subtraction, multiplication and inversion of two CQT-matrices $\widetilde{A}$, $\widetilde{B}$ which are the truncated representations of two QT matrices A and B i.e.,

$$A = \widetilde{A} + \mathcal{E}_a, \quad \widetilde{A} = \mathrm{trunc}(A) = T(\widetilde{a}) + \widetilde{E}_a$$
$$B = \widetilde{B} + \mathcal{E}_b, \quad \widetilde{B} = \mathrm{trunc}(B) = T(\widetilde{b}) + \widetilde{E}_b,$$

denote by $\star$ any arithmetic operation, define $C = A \star B$, $\widehat{C} = \widetilde{A} \star \widetilde{B}$ and $\widetilde{C} = \mathrm{trunc}(\widehat{C})$. We define *total error* in the operation $\star$ as $\mathcal{E}_c^{tot} = C - \widetilde{C}$, the *local error* as $\mathcal{E}_c^{loc} = \widehat{C} - \widetilde{C}$ and the *inherent error* as $\mathcal{E}_c^{in} = C - \widehat{C}$, so that $\mathcal{E}_c^{tot} = \mathcal{E}_c^{in} + \mathcal{E}_c^{loc}$. Observe that the inherent error is the result of $\mathcal{E}_a$ and $\mathcal{E}_b$ through the performed matrix operation, the local error is generated by the truncation of the matrix arithmetic operation $\widetilde{A} \star \widetilde{B}$, while the total error is the sum of the two errors. Formally, these errors behave like the inherent error and the round-off error in the standard floating point arithmetic.

In our study we do not analyze the growth of the inherent error in each arithmetic operation, but rather we limit ourselves to operate the truncation and compression in such a way that the norm of the local error is bounded by a given value ϵ, say the machine precision. Moreover, we do not consider the errors generated by the floating point arithmetic.

7.4.1 *Addition*

Let $A = \widetilde{A} + \mathcal{E}_a$ and $B = \widetilde{B} + \mathcal{E}_b$ be CQT matrices where $\widetilde{A} = T(\widetilde{a}) + \widetilde{E}_a$, $\widetilde{B} = T(\widetilde{b}) + \widetilde{E}_b$ with $\widetilde{a}(z)$, $\widetilde{b}(z)$ Laurent polynomials of degrees $n_a^{\pm}$ and $n_b^{\pm}$ respectively, and $\widetilde{E}_a = F_a G_a^t$, $\widetilde{E}_b = F_b G_b^t$.

If A and B have the above representation, then, for the matrix $C = A + B$ we have the representation

$$C = \widetilde{A} + \widetilde{B} + \mathcal{E}_a + \mathcal{E}_b,$$

from which we deduce that the inherent error is $\mathcal{E}_c^{in} = \mathcal{E}_a + \mathcal{E}_b$. On the other hand, concerning $\widehat{C} = \widetilde{A} + \widetilde{B}$ we have

$$\widehat{C} = T(\widetilde{a} + \widetilde{b}) + \widetilde{E}_a + \widetilde{E}_b,$$

where $\widetilde{a}(z) + \widetilde{b}(z)$ is a Laurent polynomial of degrees $n_c^- = \max(n_a^-, n_b^-)$, $n_c^+ = \max(n_a^+, n_b^+)$. while

$$E_c = \widetilde{E}_a + \widetilde{E}_b = F_c G_c^t,$$
$$F_c = [F_a, F_b], \quad G_c = [G_a, G_b],$$

where $f_c = \max(f_a, f_b)$ and $g_c = \max(g_a, g_b)$ are the number of nonzero rows of F_c and G_c, respectively, and $k_c = k_a + k_b$ is the number of columns of F_c and G_c.

The Laurent polynomial $\widetilde{a}(z) + \widetilde{b}(z)$ can be truncated and replaced by a Laurent polynomial $\widetilde{c}(z)$ of possibly less degree. Also the value of k_c, can be reduced and the matrices F_c, G_c can be compressed, by using a compression technique which guarantees a local error with norm bounded by a given ϵ. This technique, based on computing SVD and QR factorization is explained in the next section. Denoting by $\widetilde{F}_c$, $\widetilde{G}_c$ the matrices obtained after compressing F_c and G_c, respectively, we have

$$\widetilde{C} = \mathrm{trunc}(\widehat{C}) = T(\widetilde{c}) + \widetilde{E}_c + \mathcal{E}_c^{loc}, \quad \widetilde{E}_c = \widetilde{F}_c \widetilde{G}_c^t,$$

where $\mathcal{E}_c^{loc}$ denotes the local error due to truncation and compression, i.e. $\mathcal{E}_c^{loc} = \widetilde{A} + \widetilde{B} - \mathrm{trunc}(\widetilde{A} + \widetilde{B})$. This way we have

$$A + B = T(\widetilde{c}) + \widetilde{E}_c + \mathcal{E}_c^{loc} + \mathcal{E}_c^{in}.$$

7.4.2 *Multiplication*

A similar expression holds for multiplication. For the product $C = AB$ we have the equation

$$AB = \widetilde{A}\widetilde{B} + \widetilde{A}\mathcal{E}_b + \mathcal{E}_a\widetilde{B} + \mathcal{E}_a\mathcal{E}_b$$

from which we deduce that the inherent error is $\mathcal{E}_c^{in} = \widetilde{A}\mathcal{E}_b + \mathcal{E}_a\widetilde{B} + \mathcal{E}_a\mathcal{E}_b$. Moreover we have

$$\begin{aligned}\widehat{C} = \widetilde{A}\widetilde{B} =& T(\widetilde{a})T(\widetilde{b}) + T(\widetilde{a})\widetilde{E}_b + \widetilde{E}_a T(\widetilde{b}) + \widetilde{E}_a\widetilde{E}_b \\ =& T(\widetilde{a}\widetilde{b}) - H(\widetilde{a}^-)H(\widetilde{b}^+) + T(\widetilde{a})\widetilde{E}_b + \widetilde{E}_a T(\widetilde{b}) + \widetilde{E}_a\widetilde{E}_b \\ =:& T(\widetilde{a}\widetilde{b}) + E_c.\end{aligned}$$

Observe that, since $\widetilde{a}^-(z)$ and $\widetilde{b}^+(z)$ are polynomials, the matrices $H(\widetilde{a}_-)$ and $H(\widetilde{b}_+)$ have a finite number of nonzero entries. Therefore, we may factorize the product $H(\widetilde{a}^-)H(\widetilde{b}^+)$ in the form FG^t. Thus, we find that the matrix E_c can be written as $E_c = F_c G_c^t$ where

$$F_c = [F, T(\widetilde{a})F_b, F_a], \quad G_c = [G, G_b, T(\widetilde{b})^t G_a + G_b(F_b^t G_a)].$$

This provides the finite representation of the product $\widehat{C} = \widetilde{A}\widetilde{B}$ where $n_c^- = n_a^- + n_b^-$, $n_c^+ = n_a^+ + n_b^+$, $f_c = \max(f_b + n_a^-, f_a)$, $g_c = \max(n_b^+, g_b, g_a + n_b^-)$, and $k_c = k_a + k_b + n_b^+$.

Also in this case we may apply a compression technique, based on SVD for reducing the memory storage of the correction and for reducing the degree of the Laurent polynomial $\widetilde{a}(z)\widetilde{b}(z)$. Operating in this way, we introduce a local error $\mathcal{E}_c^{loc} = \widetilde{A}\widetilde{B} - \mathrm{trunc}(\widetilde{A}\widetilde{B})$. Denoting by $\widetilde{c}(z)$ the truncation of the Laurent polynomial $\widetilde{a}(z)\widetilde{b}(z)$ and with $\widetilde{F}_c\widetilde{G}_c^t$ the compression of $F_c G_c^t$, we have

$$\widehat{C} = \widetilde{A}\widetilde{B} = T(\widetilde{c}) + \widetilde{F}_c\widetilde{G}_c^t + \mathcal{E}_c^{loc}.$$

This way we have

$$C = AB = T(\widetilde{c}) + \widetilde{F}_c\widetilde{G}_c^t + \mathcal{E}_c^{loc} + \mathcal{E}_c^{in},$$

which expresses the result C of the multiplication in terms of the approximated value $\widetilde{C} = T(\widetilde{c}) + \widetilde{E}_c$, the local error $\mathcal{E}_c^{loc}$ and the inherent error $\mathcal{E}_c^{in}$. The overall error is given by $\mathcal{E}_c = \mathcal{E}_c^{loc} + \mathcal{E}_c^{in}$.

7.4.3 *Matrix inversion*

It is worth paying a particular attention to the operation of matrix inversion since it is less immediate than multiplication and addition.

First, we consider the problem of inverting the matrix $A = T(a)$, i.e., we assume that $E_a = 0$. The general case will be treated afterwords.

Recall that, if $a(z) \in \mathcal{W}_1$ does not vanish in the unit circle and if it has a zero winding number, then Theorem 7.2.3 implies that the matrix $T(a)$ is invertible and, in view of Theorem 7.2.4, there exists the canonical Wiener-Hopf factorization $a(z) = u(z)\ell(z)$ so that (7.7) holds. Thus, a finite representation of A^{-1} is obtained by truncating the Laurent series of $1/a(z)$ to a Laurent polynomial and by approximating the Hankel matrices $H((\ell^{-1})^-)$ and $H((u^{-1})^+)$ by means of matrices having a finite number of nonzero entries, an infinite number of rows and the same finite number of columns. The latter operation can be achieved by truncating the power series $\ell^{-1}(z)$ and $u^{-1}(z)$ to polynomials and by numerically compressing the product of the Hankel matrices obtained this way. This operation can be effectively performed by reducing the Hankel matrices to tridiagonal form by means of Lanczos method with orthogonalization. This procedure takes advantage of the Hankel structure since the matrix-vector product can be computed by means of FFT in $O(n \log n)$ operations where n is the size of the Hankel matrix. The advantage of this compression is that the cost grows as $O(r^2 n \log n)$ where r is the numerical rank of the matrix.

If $a(z)$ is analytic in the annulus $\mathbb{A}(r_a, R_a) \supset \mathbb{T}$, then its coefficients have an exponential decay so that $|a_i^+| \leqslant \gamma\lambda_+^i$, $|a_i^-| \leqslant \gamma\lambda_-^i$, $|u_i| \leqslant \gamma\lambda_+^i$, $|\ell_i^-| \leqslant \gamma\lambda_-^i$, for some positive γ and for $1/R_a < \lambda_+ < 1$, $r_a < \lambda_- < 1$. Thus, we find that for the truncated approximation of the matrix A the values of n^+, n^-, f, g are bounded by $\log(\gamma^{-1}\epsilon^{-1})/\log(\lambda_\pm^{-1})$.

Performing numerical experiments it turns out that the singular values of the principal submatrices of the Hankel matrices $H(\ell^-)$ and $H(u^+)$ associated with power series having coefficients with an exponential decay, have an exponential decay themselves. So that also the truncation on the value of the numerical rank k of $H(\ell^-)H(u^+)$ can be performed efficiently.

The analysis of the inherent error due to inversion is related to the analysis of the condition number of semi-infinite Toeplitz matrices. We do not carry out this analysis, we refer the reader to the books [28], [29] on this regard.

Now consider the more general case of the matrix $A = T(a) + F_a G_a^t$ which we assume already in its truncated form. Assume $T(a)$ invertible and write $A = T(a)(I + T(a)^{-1}F_a G_a^t)$. Denoting for simplicity $U = T(u)$, $L = T(\ell)$ we have

$$
\begin{aligned}
&(T(a) + F_a G_a^t)^{-1} = T(a)^{-1} - L^{-1}(U^{-1}F_a)Y^{-1}(G_a^t L^{-1})U^{-1},\\
&Y = I + G_a^t L^{-1} U^{-1} F_a,
\end{aligned}
$$

where Y is a finite matrix which is invertible if and only if A is invertible. This way, the algorithm for computing A^{-1} in its finite QT-matrix representation is given by the following steps:

1. compute the spectral factorization $a(z) = u(z)\ell(z)$;
2. compute the coefficients of the power series $\widetilde{u}(z) = 1/u(z)$ and $\widetilde{\ell}(z) = 1/\ell(z)$, so that $L^{-1} = T(\widetilde{\ell})$, $U^{-1} = T(\widetilde{u})$;
3. represent the matrix $H = L^{-1}U^{-1}$ as $T(c) + F_h G_h^t$, where $c(z) = \widetilde{\ell}(z)\widetilde{u}(z)$ by means of Theorem 7.2.2;
4. compute the products: $G_1 = T(\widetilde{\ell})G_a$, $F_1 = T(\widetilde{u})F_a$;
5. compute $Y = I + G_1^t F_1$, $F_2 = F_1 Y^{-1}$, $F_3 = T(\widetilde{\ell})F_2$, $G_2 = T(\widetilde{u})G_1$;
6. output the coefficients of $c(z)$ and the matrices $F_c = [F_h, F_3]$, $G_c = [G_h, G_2]$.

For computing the spectral factorization of $a(z)$ we rely on the algorithm of [14] which employs evaluation/interpolation techniques at the Fourier points, see AppendixB.

7.4.4 *Compression*

The algorithms that implement the CQT-arithmetic have to deal with two issues of compression.

The first one concerns the compression of the finite correction in the outcome of an arithmetic operation. That is, given the matrix E in the form $E = FG^t$ where F and G are matrices of size $m \times k$ and $n \times k$, respectively, we aim to reduce the size k and to approximate E in the form $\widetilde{F}\widetilde{G}^t$ where $\widetilde{F}$ and $\widetilde{G}$ are matrices of size $m \times \widetilde{k}$ and $n \times \widetilde{k}$, respectively, with $\widetilde{k} < k$. We can treat this problem as in Section 3.4.1 by means of truncating the reduced SVD of E.

The second one regards the compression of the product of two Hankel matrices and occurs when multiplying or inverting CQT-matrices. If the size of the latter is big, e.g., when we multiply two CQT-matrices with a large Toeplitz bandwidth, the use of SVD can be too expensive. In such cases we rely on the two sided Lanczos method [90], see Appendix C. The latter enable us to find adaptively a low-rank approximation. Other approaches, that we want to test in the future for handling this task, concern randomized techniques of compression [55].

7.5 FINITE QUASI-TOEPLITZ ARITHMETIC

Given a symbol $a(z)$ and $m \in \mathbb{Z}^+$ we indicate with $T_m(a)$ the finite $m \times m$-Toeplitz matrix obtained by selecting the first m rows and columns of $T(a)$. Instead, with $H_m(a^-)$

and $H_m(a^+)$ we denote the $m \times m$-Hankel anti-triangular matrices generated by the first $m-1$ negative and positive coefficients of $a(z)$, respectively.

The approach that we have followed in this chapter can be easily adapted to retrieve an arithmetic for quasi-Toeplitz matrices of finite size. The crucial tool —for doing this extension— is a version of Theorem 7.2.2 in the finite case.

Theorem 7.5.1. *For $a(z), b(z) \in \mathcal{W}$ let $c(z) = a(z)b(z)$. Then we have*

$$T_m(a)T_m(b) = T_m(c) - H_m(a^-)H_m(b^+) - J_m H_m(a_+)H_m(b_-)J_m,$$

where J_m is the flip matrix having 1 on the anti-diagonal and zeros elsewhere.

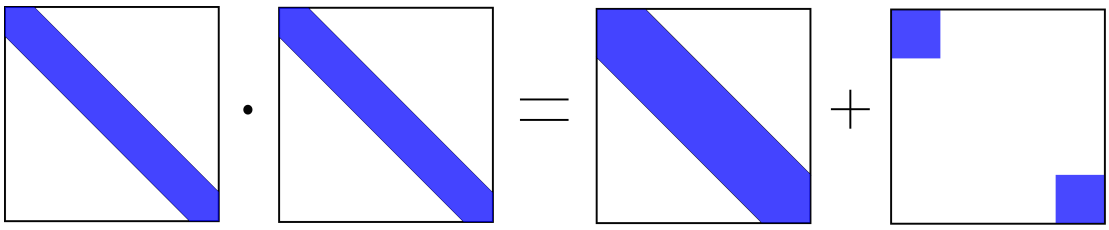

Figure 7.4.: Multiplication of two finite banded Toeplitz matrices

If $a(z) = \sum_{i=-k}^{k} a_i z^i$, $b(z) = \sum_{i=-k}^{k} b_i z^i$ with k much smaller than m, then the matrices $H_m(a^-)H_m(b^+)$ and $J_m H_m(a^+)H_m(b^-)J_m$ have disjoint supports located in the upper leftmost corner and in the lower rightmost corner, respectively. Thus, $T_m(a)T_m(b)$ can be represented as the sum of the Toeplitz matrix associated with the Laurent polynomial $c(z)$ and of two correction matrices E^+ and E^- which collect the finite number of nonzero entries located in the upper leftmost and in the lower rightmost corners, respectively.

An immediate byproduct of Theorem 7.5.1 is the following corrected canonical decomposition for a finite Toeplitz matrix.

Corollary 7.5.2. *Let $a(z) \in \mathcal{W}$ be non vanishing on $\mathbb{T}$ and with winding number 0. If $a(z) = u(z)l(z)$ is the canonical factorization of $a(z)$ then it holds*

$$T_m(a) = T_m(u)T_m(l) - J_m H_m(u_+)H_m(l_-)J_m.$$

This paves the way for computing the inverse of a finite Toeplitz matrix by means of the technique based on the Sherman-Morrison-Woodbury formula used in Section 7.4.3. In particular, we get

$$T_m(a)^{-1} = T_m(a^{-1}) + E$$

and the matrix E —if m is large compared to the bandwidth of the symbol— has nonzero entries only in the upper left and lower right corners.

Algorithms for dealing with the finite quasi-Toeplitz matrices can be easily obtained from those presented in Section 7.4 just by taking into account the additional lower

rightmost corner correction. In the case of a sufficiently large gap between the size m and the bandwidth k of the symbols that come into play, the two corner corrections behave independently of each other and the finite CQT matrix arithmetic becomes more effective. The cost of these operations essentially depends on the Toeplitz bandwidth of the outcome and on the sizes and ranks of the correction matrices. The cost remains small as long as the bandwidth and the size of the corrections E^+, E^- remain small together with their rank. Whether this condition is not satisfied, the two corrections may spread and overlap. This may cause a slowdown due to the additional operations which are needed in the computation.

7.6 SOLVING SEMI-INFINITE QUADRATIC MATRIX EQUATIONS

Consider the quadratic matrix equation (2.2):

$$A_{-1} + A_0 X + A_1 X^2 = 0.$$

The arithmetic developed in Section 7.4 paves the way to the use of CR when $A_i \in \mathcal{CQT}$, $i = -1, 0, 1$. Observe that, since $\mathcal{CQT}$ is an algebra, all the matrices generated by CR belong to $\mathcal{CQT}$. Moreover, the Toeplitz part of these matrices have associated symbols $a_{-1}^{(h)}(z)$, $a_0^{(h)}(z)$, $a_1^{(h)}(z)$, $\widetilde{a}^{(h)}(z)$, $\widehat{a}^{(h)}(z)$, which satisfy the same recurrence equations as (6.2). More precisely, we have the *scalar* functional relations

$$\begin{aligned}
a_0^{(h+1)}(z) &= a_0^{(h)}(z) - 2a_1^{(h)}(z)a_{-1}^{(h)}(z)/a_0^{(h)}(z),\\
a_1^{(h+1)}(z) &= -a_1^{(h)}(z)^2/a_0^{(h)}(z), \quad a_{-1}^{(h+1)}(z) = -a_{-1}^{(h)}(z)^2/a_0^{(h)}(z),\\
\widetilde{a}^{(h+1)}(z) &= \widetilde{a}^{(h)}(z) - a_1^{(h)}(z)a_{-1}^{(h)}(z)/a_0^{(h)}(z),
\end{aligned}$$

with $h = 0, 1, \ldots$, where $a_i^{(0)}(z) = a_i(z)$, $i = -1, 0, 1$ and $\widetilde{a}^{(0)}(z) = a_0(z)$. Observe that since all the quantities in the above recurrence are scalar functions, they commute so that $\widehat{a}^{(h)}(z)$ coincides with $\widetilde{a}^{(h)}(z)$.

As pointed out in [15], [23], in the scalar case CR reduces to the celebrated Graeffe iteration whose properties have been investigated in [82]. Thus, in order to analyze the convergence of the sequences defined above, we rely on the convergence properties of the Graeffe iteration applied to quadratic polynomials. In particular, we know that if, for a given $z \in \mathbb{T}$ the polynomial $p_z(x) := a_1(z)x^2 + a_0(z)x + a_{-1}(z)$ associated with the triple $(a_{-1}(z), a_0(z), a_1(z))$, has one root inside the unit disc and one root outside, then the sequence $-(a_{-1}(z)/\widetilde{a}^{(h)}(z))$ has a limit $g(z)$ which coincides with the root of the polynomial $p_z(x)$ inside the unit disc. More precisely, pointwise $g(z)$ corresponds either to $\frac{-a_0(z)+\sqrt{\Delta(z)}}{2a_1(z)}$ or $\frac{-a_0(z)-\sqrt{\Delta(z)}}{2a_1(z)}$ with $\Delta(z) = a_0(z)^2 - 4a_1(z)a_{-1}(z)$.

The following theorem provides mild conditions which ensure the above properties, and are generally satisfied in the applications.

Theorem 7.6.1. *Let* $a_i(z) = a_{i,-1}z^{-1} + a_{i,0} + a_{i,1}z$, *for* $i = -1, 0, 1$, *be such that* $\sum_{i,j=-1}^{1} a_{i,j} = 0$, $a_{0,0} < 0$, $a_{i,j} \geqslant 0$, *otherwise. If*

(i) $a_{-1,0} > 0$ *or* $a_{1,0} > 0$,

(ii) $a_{ij} \neq 0$ *for at least a pair* (i, j), *with* $j \neq 0$,

then for any $z \in \mathbb{T}$, $z \neq 1$, *the quadratic polynomial* $p_z(x) = a_1(z)x^2 + a_0(z)x + a_{-1}(z)$, *has a root of modulus less than 1 and a root of modulus greater than 1.*

Proof. Without loss of generality we may assume that the entries $a_{i,j}$ belong to the interval $[-1, 1]$. If not, we may scale equation (2.2) by a suitable constant and reduce it to this case. As a first step we show that there are no roots of modulus 1. Assume by contradiction that x is a root of modulus 1. Obviously, we have $p_z(x) = 0$ if and only $p_z(x) + x = x$. Observe that, if $z \in \mathbb{T}$, the left hand-side of the previous equation is a convex combination of the points in the discrete set $\mathcal{C}_{x,z} := \{x^i z^j,\ i = 0, 1, 2,\ j = -1, 0, 1\} \subset \mathbb{T}$. If $z \neq 1$, condition (i) and the fact that $-1 \leqslant a_{0,0} < 0$ ensure that the convex combination involves at least two different points of the unit circle, either x and 1 or x and x^2. Therefore, this convex combination $p_z(x) + x$ is equal to a point which belongs to the interior of the unit disc. This contradicts the fact that $|p_z(x) + x| = |x| = 1$. This argument excludes roots on $\mathbb{T}$ for $z \in \mathbb{T} \setminus \{1\}$. We conclude by showing that there is exactly one root of modulus less than 1. In order to prove this, we first show that $|a_0(z)| > |a_{-1}(z) + a_1(z)|$ holds for any $z \in \mathbb{T} \setminus \{1\}$. Therefore, by applying the Rouché Theorem one finds that the functions $f(x) = a_0(z)x$ and $p_z(x)$ have the same number of zeros in the open unit disc. To prove the inequality $|a_0(z)| > |a_{-1}(z) + a_1(z)|$ we observe that

$$\begin{aligned} |a_{0,-1}z^{-1} + a_{0,0} + a_{0,1}z| &\geqslant |a_{0,0}| - |a_{0,-1}z^{-1}| - |a_{0,1}z| = -a_{0,0} - a_{0,-1} - a_{0,1} \\ &= a_{-1,-1} + a_{-1,0} + a_{-1,1} + a_{1,-1} + a_{1,0} + a_{1,1} \\ &\geqslant |a_{-1,-1}z^{-1} + a_{-1,0} + a_{-1,1}z + a_{1,-1}z^{-1} + a_{1,0} + a_{1,1}z| \end{aligned}$$

where at least one of the two above inequalities is strict because of condition (ii). □

Corollary 7.6.2. *Under the conditions of Theorem 7.6.1, if* $a_1(z) \neq 0$ *for any* $z \in \mathbb{T}$ *and* $a_{-1}(1) \neq a_1(1)$, *then* $g(z) = \lim_h -a_1(z)/\widetilde{a}^{(h)}(z)$ *is an analytic function.*

Proof. We recall that the roots of a monic polynomial are analytic functions of the coefficients, on the set where the polynomial has not multiple roots [30]. Thus, in order to prove the analyticity of $g(z)$, it is sufficient to show that that $p_z(x)$ has no multiple root $\forall z \in \mathbb{T}$. This follows from Theorem 7.6.1 if $z \in \mathbb{T} \setminus \{1\}$. Moreover, observe that for $z = 1$, $p_1(x)$ has roots 1 and $\frac{a_{-1}(1)}{a_1(1)}$ where the latter is real, non negative and different from 1 by assumption. □

With the information that we have collected so far, we cannot yet say if the matrix G belongs to $\mathcal{CQT}$. In fact, in principle, writing $G = T(g) + E_g$, it is not ensured that

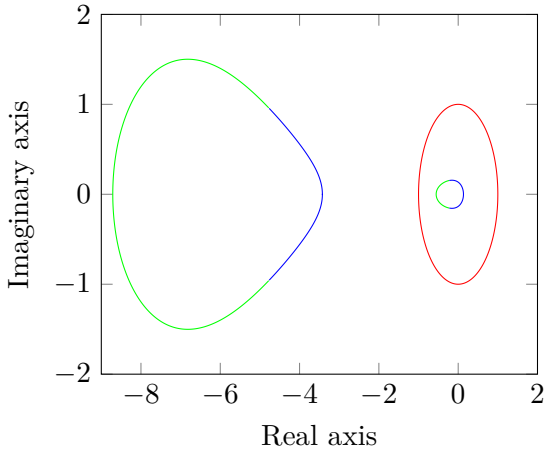

Figure 7.5.: Roots of $p_z(x)$ as z varies on $\mathbb{T}$; In red the unit circle, in green the plot of $\frac{-a_0(z)+\sqrt{\Delta(z)}}{2a_1(z)}$ and in blue the plot of $\frac{-a_0(z)-\sqrt{\Delta(z)}}{2a_1(z)}$. These expressions parametrize piecewise the roots of $p_z(x)$ inside and outside the unit circle.

$\|E_g\|_{\mathcal{F}} < \infty$. The boundedness of $\|E_g\|_{\mathcal{F}}$ can be proved if E_g has all entries with the same sign. This analysis is part of the subject of our future research. On this regard, it is worth citing the paper [91] where, relying on probabilistic arguments, it is proved that the matrices G and R asymptotically share the Toeplitz structure.

7.6.1 *Numerical results*

In order to validate our analysis, we consider ten instances of the two-node Jackson network, analyzed in [80]. In details, we assume

$$A_{-1} = \begin{bmatrix} (1-q)\mu_2 & q\mu_2 & & \\ & (1-q)\mu_2 & q\mu_2 & \\ & & \ddots & \ddots \end{bmatrix},$$

$$A_0 = \begin{bmatrix} -(\lambda_1+\lambda_2+\mu_2) & \lambda_1 & & \\ (1-p)\mu_1 & -(\lambda_1+\lambda_2+\mu_1+\mu_2) & \lambda_1 & \\ & \ddots & \ddots & \ddots \end{bmatrix},$$

$$A_1 = \begin{bmatrix} \lambda_2 & & \\ p\mu_1 & \lambda_2 & \\ & \ddots & \ddots \end{bmatrix},$$

where the parameters $p, q, \lambda_1, \lambda_2, \mu_1, \mu_2$ are chosen according to Table 7.1. These examples are also studied in [88] where it is shown the bad effect of truncation in approximating the

Case	λ_1	λ_2	μ_1	μ_2	p	q
1	1	0	1.5	2	1	0
2	1	0	2	1.5	1	0
3	0	1	1.5	2	0	1
4	0	1	2	1.5	0	1
5	1	1	2	2	0.1	0.8
6	1	1	2	2	0.8	0.1
7	1	1	2	2	0.4	0.4
8	1	1	10	10	0.5	0.5
9	1	5	10	15	0.4	0.9
10	5	1	15	10	0.9	0.4

Table 7.1.: Parameters values of the test examples for the two node Jackson tandem network

stationary distribution. Different decay properties of the invariant probability distribution correspond to the different values of the parameters.

We have applied CR in all the 10 cases and computed the minimal non-negative solution G represented in the CQT form as $T(g) + U_g V_g^t$. In the results of the tests that we have performed, we report, besides the CPU time in seconds, also the norm of the residual error $E = A_1 G^2 + A_0 G + A_{-1}$ where we used both the infinity norm $\|E\|_\infty$ and the CQT norm $\|E\|_{\mathcal{CQT}}$.

In order to analyze the intrinsic complexity of the problem, we also report the band width of the matrix $T(g)$, that is the number of non-negligible coefficients of the Laurent series $\sum_{i\in\mathbb{Z}} g_i z^i$, the number of the nonzero rows of the matrices U_g and V_g and the number of their columns that is their rank.

All this information is reported in Table 7.2. We may observe that a high CPU time, like for instance in the case of Problem 7, corresponds to large values of the band width in the matrix $T(g)$ or to large sizes of the correction. The large values of these two components of the CQT representation of G imply that the entries $g_{i,j}$ have a low decay speed as $i, j \to \infty$.

7.7 FUNCTIONS OF FINITE AND SEMI-INFINITE QUASI-TOEPLITZ MATRICES

Once that a certain fast arithmetic is provided one can try to speed up the computation of matrix functions. Here we address this issue both theoretically and practically. We prove that, under certain conditions on the function $f(z) : \mathbb{C} \to \mathbb{C}$, we can provide the definition of $f(A)$ for any $A \in \mathcal{CQT}$ and we show that $f(A) \in \mathcal{CQT}$. These conditions include the case of the exponential function and of the main functions which are encountered in the applications.

Case	CPU time	Res_∞	$Res_{\mathcal{CQT}}$	Band	Rows	Columns	Rank
1	2.61 s	$8.63 \cdot 10^{-16}$	$5.98 \cdot 10^{-13}$	561	541	138	8
2	2.91 s	$1.49 \cdot 10^{-15}$	$7.88 \cdot 10^{-13}$	561	555	145	8
3	0.29 s	$1.11 \cdot 10^{-16}$	$2.67 \cdot 10^{-14}$	143	89	66	8
4	2.32 s	$6.77 \cdot 10^{-16}$	$6 \cdot 10^{-13}$	463	481	99	9
5	0.48 s	$1.23 \cdot 10^{-15}$	$1.07 \cdot 10^{-13}$	233	108	148	9
6	7.96 s	$1.92 \cdot 10^{-14}$	$6.65 \cdot 10^{-13}$	455	462	153	10
7	29 s	$4.29 \cdot 10^{-15}$	$6.87 \cdot 10^{-12}$	1,423	1,543	247	13
8	1.01 s	$1.14 \cdot 10^{-15}$	$4.34 \cdot 10^{-13}$	366	348	40	6
9	0.3 s	$5.44 \cdot 10^{-16}$	$2.48 \cdot 10^{-14}$	157	81	86	8
10	1.25 s	$1.09 \cdot 10^{-15}$	$3.4 \cdot 10^{-14}$	268	241	107	8

Table 7.2.: Features of the computed solutions by means of CR

Another case of interest concerns matrices associated with an analytic symbol $a(z)$ where the coefficients of the Toeplitz part have an exponential decay. This situation is very convenient from the computational point of view. However, the class of matrices that we obtain this way, which we called analytically quasi-Toeplitz (AQT), is still a matrix algebra, but is not a Banach space with the norm $\|\cdot\|_{\mathcal{CQT}}$. In the analysis that we carry out, we point out the cases where the result of the computation is still in the class of AQT matrices.

We consider two possible extensions: the case where $f(z)$ is assigned as a Laurent series, for instance $\exp(z) = \sum_{i=0}^{\infty} \frac{1}{i!} z^i$, so that the matrix extension is formally given by $\sum_{i=0}^{\infty} a_i A^i$ and then the case where $f(z)$ is defined by means of the Dunford-Cauchy formula (5.1). The computational strategy using the latter definition is analogous to the one used for computing functions of quasiseparable matrices in Section 5.4.

Concerning the recent literature in this research area, it is worth citing [51] where the computation of functions of Hermitian Toeplitz matrices is addressed. In [13] the exponential function of a block-triangular block-Toeplitz matrix is analyzed with application to solving certain fluid queues. In the recent paper [67] the problem of computing the exponential function of finite Toeplitz matrices is investigated and several applications are presented. In [22] the case of the exponential of a semi-infinite CQT matrix is analyzed in depth.

7.7.1 *Function of a CQT matrix: power series representation*

In this section we give conditions under which a function $f(x)$, expressed in terms of a power series or a Laurent series, can be applied to matrices A in the class $\mathcal{CQT}$, and prove that under these conditions $f(A)$ still belongs to $\mathcal{CQT}$.

Let $a(z) \in \mathcal{W}_1$ and $A = T(a) + E \in \mathcal{CQT}$. Assume we are given a complex valued function $f(x) = \sum_{i=0}^{+\infty} f_i x^i$ which is analytic on the open disc $B(0,\rho) = \{x \in \mathbb{C} : \quad |x| < \rho\}$. Observe that, if $a(\mathbb{T}) \subseteq B(0,\rho)$, then the composed function $f(a(z))$ belongs to $\mathcal{W}_1$.

Define $\varphi_k(x) = \sum_{i=0}^{k} f_i x^i$ and observe that for any integers h, k such that $h > k$ one has $\varphi_h(A) - \varphi_k(A) = \sum_{i=k+1}^{h} f_i A^i$. Thus,

$$\|\varphi_h(A) - \varphi_k(A)\|_{\mathcal{CQT}} \leqslant \sum_{i=k+1}^{h} |f_i| \cdot \|A\|_{\mathcal{CQT}}{}^i. \tag{7.9}$$

This inequality implies the following result.

Theorem 7.7.1. *Let $A = T(a) + E_a \in \mathcal{CQT}$ and let $f(x) = \sum_{i=0}^{+\infty} f_i x^i$ be analytic in $\mathbb{D}(\rho)$. If $\|A\|_{\mathcal{CQT}} < \rho$ then $f(A) = \sum_{i=0}^{+\infty} f_i A^i$ is well defined, belongs to $\mathcal{CQT}$, and*

$$f(A) = T(f(a)) + E_{f(a)}, \quad E_{f(a)} \in \mathcal{F}.$$

Furthermore, if $A \in \mathcal{AQT}$ then $f(A) \in \mathcal{AQT}$. More precisely, there exists an annulus $\mathbb{A}(r,R)$ containing $\mathbb{T}$, such that $f(a(z))$ is well defined and analytic for $z \in \mathbb{A}(r,R)$.

Proof. We prove that the sequence $\varphi_k(A) = \sum_{i=0}^{k} f_i A^i$ is a Cauchy sequence in $(\mathcal{CQT}, \|\cdot\|_{\mathcal{CQT}})$. In fact, since $\|A\|_{\mathcal{CQT}} < \rho$ there exists $0 < \delta < \rho$ such that $\|A\|_{\mathcal{CQT}} = \rho - \delta$. Thus, from (7.9), for $h > k$ we have $\|\varphi_h(A) - \varphi_k(A)\|_{\mathcal{CQT}} \leqslant \sum_{i=k+1}^{h} |f_i|(\rho-\delta)^i$. On the other hand, in view of equation (7.1) with $\epsilon = \delta/2$, there exists γ such that $|f_i| \leqslant \gamma(\rho - \delta/2)^{-i}$. This implies that $\|\varphi_h(A) - \varphi_k(A)\|_{\mathcal{CQT}} \leqslant \gamma \sum_{i=k+1}^{h} \lambda^i$, $\lambda = (\rho-\delta)/(\rho-\delta/2) < 1$. Thus for sufficiently large values of h and k, the latter summation is smaller than any given $\epsilon > 0$ so that the sequence $\varphi_k(A)$ is Cauchy. Since the space $\mathcal{CQT}$ is Banach, there exists $F \in \mathcal{CQT}$ such that $\lim_k \|\varphi_k(A) - F\|_{\mathcal{CQT}} = 0$. That is, $F := f(A)$ is well defined and belongs to $\mathcal{CQT}$. Thus, $f(A)$ can be written as $f(A) = T(g) + E_g$ for a suitable $g(z) \in \mathcal{W}_1$ and $E_g \in \mathcal{F}$. Observe that $\varphi_k(A)$ can be written in the form $\varphi_k(A) = T(\varphi_k(a)) + E_k$ for a suitable $E_k \in \mathcal{F}$. Thus, the convergence of $\varphi_k(A)$ to $T(g) + E_g$ in the norm $\|\cdot\|_{\mathcal{CQT}}$ implies that $\lim_k \|E_k - E_g\|_{\mathcal{F}} = 0$ and $\lim_k \|\varphi_k(a) - g\|_{\mathcal{W}} = 0$. Thus we deduce that $g(z) = f(a(z))$. In the case $A \in \mathcal{AQT}$, in order to show that $F \in \mathcal{AQT}$, it is sufficient to prove that $g(z) = f(a(z))$ is analytic over some annulus $\mathbb{A}(r,R)$. From the condition $\|a\|_{\mathcal{W}} \leqslant \|A\|_{\mathcal{CQT}} < \rho$ it follows that for $|z| = 1$, we have $|a(z)| \leqslant \sum_{i\in\mathbb{Z}} |a_i| \cdot |z|^i = \|a\|_{\mathcal{W}} < \rho$. By continuity of $a(z)$ there exists an open annulus $\mathbb{A}(r,R)$ which includes the unit circle $\mathbb{T}$, such that $|a(z)| < \rho$ for $z \in \mathbb{A}(r,R)$. This way, the function $f(a(z))$ is well defined and analytic in $\mathbb{A}(r,R)$ since composition of analytic functions. This shows that $f(A) \in \mathcal{AQT}$ and the proof is complete. □

Now we consider the problem of determining bounds to $\|E_{f(a)}\|_{\mathcal{F}}$. These bounds are useful from the computational point of view since they provide an indication of the mass of information which is stored in the correction part of $f(A)$. Equivalently, they tell us how much the matrix $f(A)$ differs from a Toeplitz matrix. For simplicity, we deal

with the case where $A = T(a)$ is Toeplitz. Then we treat the general case of a matrix $A = T(a) + E_a$.

Since $\|a\|_{\mathcal{W}} < \rho$, for the analyticity of $f(x)$ in the disc $B(0, \rho)$, we may write

$$\|f(a)\|_{\mathcal{W}} \leqslant \sum_{i=0}^{+\infty} |f_i| \cdot {\|a\|_{\mathcal{W}}}^i < \sum_{i=0}^{+\infty} |f_i| \rho^i < \infty.$$

Let $A^k = T(a^k) + E_k$ and decompose $\varphi_k(A)$ as $\varphi_k(A) = G_k + F_k$, where $G_k = \sum_{i=0}^{k} f_i T(a^i)$, $F_k = \sum_{i=0}^{k} f_i E_i$. Then we have $G_k = T(\sum_{i=0}^{k} f_i a^i)$ so that, $\lim_k G_k = T(f(a))$ and $\lim_k F_k = f(A) - T(f(a)) = E_{f(a)}$.

The following result from [22] provides a representation of the matrices $T(a)^i$ and $(T(a) + E)^i$.

Theorem 7.7.2. *If $a(z) \in \mathcal{W}_1$ then $T(a)^i = T(a^i) + E_i$, where $E_1 = 0$ and $E_i = T(a)E_{i-1} - H(a^-)H((a^{i-1})^+)$, $\quad i \geqslant 2$. Moreover,*

$$\|E_i\|_{\mathcal{F}} \leqslant \frac{i(i-1)}{2} {\|a'\|_{\mathcal{W}}}^2 {\|a\|}^{i-2}_{\mathcal{W}}.$$

If $A = T(a) + E \in \mathcal{CQT}$ then $A^i = T(a^i) + D_i$, where $D_0 = E$ and

$$D_i = AD_{i-1} - H(a^-)H((a^{i-1})^+) + ET(a^{i-1}), \quad i \geqslant 1.$$

Moreover, for $\alpha = {\|a'\|}^2_{\mathcal{W}} + \|E\|_{\mathcal{F}}$, $\quad \beta = {\|a'\|_{\mathcal{W}}}^2$ we have

$$\|D_i\|_{\mathcal{F}} \leqslant \frac{1}{\|E\|_{\mathcal{F}}} \left(\alpha \frac{(\|a\|_{\mathcal{W}} + \|E\|_{\mathcal{F}})^i - {\|a\|}^i_{\mathcal{W}}}{\|E\|_{\mathcal{F}}} - \beta i {\|a\|}^{i-1}_{\mathcal{W}} \right).$$

Now we can provide upper bounds for $\|E_{f(a)}\|_{\mathcal{F}}$ in the case of an almost general function $f(z)$.

Theorem 7.7.3. *Assume that the function $f(x) = \sum_{i \in \mathbb{Z}^+} f_i x^i$ is analytic on $B(0, \rho)$, that $a(z) \in \mathcal{W}_1$ and is such that $\|a\|_{\mathcal{W}} < \rho$. Let $A = T(a)$ and $f(A) = T(f(a)) + E_{f(a)}$. Then*

$$\|E_{f(a)}\|_{\mathcal{F}} \leqslant \frac{1}{2} {\|a'\|_{\mathcal{W}}}^2 g''(\|a\|_{\mathcal{W}})$$

where $g(z) = \sum_{i=0}^{\infty} |f_i| z^i$.

Proof. Recall from Theorem 7.7.1 that $f(a(z)) \in \mathcal{W}_1$. From Theorem 7.7.2 we have the bound

$$\|E_i\|_{\mathcal{F}} \leqslant \frac{i(i-1)}{2} {\|a'\|_{\mathcal{W}}}^2 {\|a\|_{\mathcal{W}}}^{i-2}$$

so that for the matrix $E_{f(a)} = \sum_{i=0}^{\infty} f_i E_i$ we have

$$\|E_{f(a)}\|_{\mathcal{F}} \leqslant \sum_{i=0}^{\infty} |f_i| \cdot \|E_i\|_{\mathcal{F}} \leqslant \frac{1}{2} {\|a'\|_{\mathcal{W}}}^2 \sum_{i=0}^{\infty} i(i-1)|f_i| \cdot {\|a\|_{\mathcal{W}}}^{i-2} = \frac{1}{2} {\|a'\|_{\mathcal{W}}}^2 g''(\|a\|_{\mathcal{W}}),$$

where $g''(\|a\|_{\mathcal{W}})$ is well defined and finite since $\|a\|_{\mathcal{W}} < \rho$ and $f(z)$ is analytic for $|z| \leqslant \rho$. This completes the proof. □

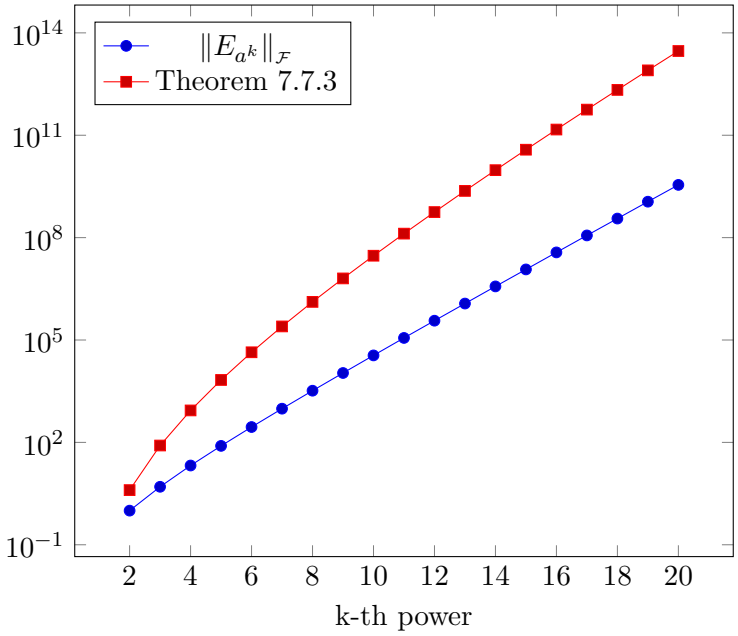

Figure 7.6.: Comparison between the norm of the non Toeplitz part of $T(a)^k$, where $a = z^{-1} + 1 + z$, with the upper bound provided by Theorem 7.7.3

In Figure 7.6 we show the values of $\|E_{a^k}\|_{\mathcal{F}}$ and of the bound provided by Theorem 7.7.3, where $a = z^{-1} + 1 + z$, as k increases.

Observe that in the case of a power series with non-negative coefficients f_i we have $g(x) = f(x)$. In particular, for $f(x) = e^x$ we get

$$\|E_{\exp(a)}\|_{\mathcal{F}} \leqslant \frac{1}{2}\|a'\|_{\mathcal{W}}{}^2 \exp(\|a\|_{\mathcal{W}}), \tag{7.10}$$

which coincides with the bound given in [22].

In the case where $A = T(a) + E_a$, we may prove a similar bound relying on Theorem 7.7.2 as expressed by the following

Theorem 7.7.4. *Assume that the function $f(x) = \sum_{i\in\mathbb{Z}^+} f_i x^i$ is analytic on $B(0,\rho)$, that $a(z) \in \mathcal{W}_1$, and is such that $\|a\|_{\mathcal{W}} < \rho$. Let $A = T(a) + E_a$ and $f(A) = T(f(a)) + E_{f(a)}$. Then*

$$\|E_{f(a)}\|_{\mathcal{F}} \leqslant \frac{1}{\|E_a\|_{\mathcal{F}}}\left(\alpha\frac{g(\|a\|_{\mathcal{W}} + \|E_a\|_{\mathcal{F}}) - g(\|a\|_{\mathcal{W}})}{\|E_a\|_{\mathcal{F}}} - \beta g'(\|a\|_{\mathcal{W}})\right)$$

where $g(z) = \sum_{i=0}^{\infty} |f_i| z^i$ and $\alpha = \|a'\|^2_{\mathcal{W}} + \|E_a\|_{\mathcal{F}}, \quad \beta = \|a'\|_{\mathcal{W}}{}^2$.

Proof. Recall from Theorem 7.7.1 that $f(a(z)) \in \mathcal{W}_1$ and that $E_{f(a)} = \lim_k F_k$, $F_k = \sum_{i=0}^{k} f_i D_i$ for $A^i = T(a^i) + D_i$. From Theorem 7.7.2 we have the bound

$$\|D_i\|_{\mathcal{F}} \leqslant \frac{1}{\|E_a\|_{\mathcal{F}}} \left(\alpha \frac{(\|a\|_{\mathcal{W}} + \|E_a\|_{\mathcal{F}})^i - \|a\|^i{}_{\mathcal{W}}}{\|E_a\|_{\mathcal{F}}} - \beta i \|a\|^{i-1}{}_{\mathcal{W}} \right)$$

with $\alpha = \|a'\|^2{}_{\mathcal{W}} + \|E_a\|_{\mathcal{F}}$, $\beta = \|a'\|_{\mathcal{W}}{}^2$ so that for the matrix $E_{f(a)} = \sum_{i=0}^{\infty} f_i D_i$ we get the bound $\|E_{f(a)}\|_{\mathcal{F}} \leqslant \sum_{i=0}^{\infty} |f_i| \cdot \|D_i\|_{\mathcal{F}}$ which leads to

$$\|E_{f(a)}\|_{\mathcal{F}} \leqslant \frac{1}{\|E_a\|_{\mathcal{F}}} \left(\alpha \frac{g(\|a\|_{\mathcal{W}} + \|E_a\|_{\mathcal{F}}) - g(\|a\|_{\mathcal{W}})}{\|E_a\|_{\mathcal{F}}} - \beta g'(\|a\|_{\mathcal{W}}) \right).$$

This completes the proof. □

Observe that, taking the limit for $\|E_a\|_{\mathcal{F}} \to 0$ in the bound given in the above theorem yields the bound of Theorem 7.7.3.

Next, we consider the case where the function $f(x)$ is assigned as a Laurent series in the form $f(x) = \sum_{i\in\mathbb{Z}} f_i x^i$ analytic over the open annulus $\mathbb{A}(r_f, R_f)$ for $r_f < R_f$. We recall from Theorem 1.4.1 the following decay property of the coefficients f_i:

$$\forall\ \epsilon > 0,\ \epsilon < R_f,\ \exists \gamma > 0\ :\ |f_i| \leqslant \gamma (R_f - \epsilon)^{-i}, \quad |f_{-i}| \leqslant \gamma (r_f + \epsilon)^i, \quad i > 0. \tag{7.11}$$

Concerning the existence of $f(A)$ for $A \in \mathcal{AQT}$ we have the following

Theorem 7.7.5. *Let $f(x) = \sum_{i\in\mathbb{Z}} a_i x^i$ be an analytic function in the open annulus $\mathbb{A}(r_f, R_f)$. Let $a(z) \in \mathcal{W}_1$ and consider a matrix $A = T(a) + E_a \in \mathcal{CQT}$. If $a(\mathbb{T}) \subset \mathbb{A}(r_f, R_f)$, $\|A^{-1}\|_{\mathcal{CQT}} < r_f^{-1}$ and $\|A\|_{\mathcal{CQT}} < R_f$ then*

$$f(A) := \sum_{i\in\mathbb{Z}} a_i A^i = T(f(a)) + E_{f(a)} \in \mathcal{CQT}.$$

Moreover if $A \in \mathcal{AQT}$ then $f(A) \in \mathcal{AQT}$.

Proof. The proof follows the same line as the one of Theorem 7.7.1. We consider $\varphi_k(x) = \sum_{i=-k}^{k} f_i x^i$ and show that $\varphi_k(A)$ is a Cauchy sequence in $\mathcal{CQT}$. Since $\|A^{-1}\|_{\mathcal{CQT}} < r_f^{-1}$ and $\|A\|_{\mathcal{CQT}} < R_f$, there exists $0 < \delta < R_f$ such that $\|A^{-1}\|_{\mathcal{CQT}} \leqslant (r_f + \delta)^{-1}$ and $\|A\|_{\mathcal{CQT}} \leqslant R_f - \delta$. Thus, applying the inequality (7.11) with $\epsilon = \delta/2$, for $h > k > 0$ we get

$$\begin{aligned} \|\varphi_k(A) - \varphi_h(A)\|_{\mathcal{QT}} &\leqslant \sum_{i=k-1}^{h} \left(|f_i| \cdot \|A\|_{\mathcal{CQT}}{}^i + |f_{-i}| \cdot \|A^{-1}\|^i{}_{\mathcal{CQT}} \right) \\ &\leqslant \gamma \sum_{i=k-1}^{h} \left(\left(\frac{R_f - \delta}{R_f - \delta/2} \right)^i + \left(\frac{r_f + \delta/2}{r_f + \delta} \right)^i \right). \end{aligned}$$

The latter quantity converges to 0 for $k \to \infty$ so that the sequence $\varphi_k(A)$ is Cauchy in $\mathcal{CQT}$ and thus there exists $F \in \mathcal{CQT}$ such that $\lim_{k\to\infty} \|\varphi_k(A) - F\|_{\mathcal{QT}} = 0$. Therefore the matrix F has the form $F = T(g) + E_g$ for some function g in the Wiener class and for $E_g \in \mathcal{F}$. By using the same argument as in the proof of Theorem 7.7.1, we obtain that $g(z) = f(a(z))$.

Now, consider the case $a(z)$ analytic. Since $a(\mathbb{T}) \subset \mathbb{A}(r_f, R_f)$, then there exists an open annulus $\mathbb{A}(r, R)$, which includes the unit circle, such that $a(\mathbb{A}(r, R)) \subseteq \mathbb{A}(r_f, R_f)$ so that $f(a(z))$ is well defined and analytic for $z \in \mathbb{A}(r, R)$. Thus $g(z) = f(a(z))$ is analytic for $z \in \mathbb{A}(r, R)$. Therefore we may conclude that $F \in \mathcal{AQT}$. □

Observe that the two technical hypotheses

$$\|A^{-1}\|_{\mathcal{CQT}} < r_f^{-1}, \quad \|A\|_{\mathcal{CQT}} < R_f,$$

given in Theorem 7.7.5, are not needed if $f(x)$ is a Laurent polynomial, i.e., a function of the form $f(x) = \sum_{i=-n_1}^{n_2} f_i x^i$. If the function is entire on $\mathbb{C}$, we need no additional assumption. For example we can claim that the exponential function of a $\mathcal{CQT}$-matrix is again a $\mathcal{CQT}$-matrix.

Computational aspects

Observe that, if $f(x) = \sum_{i=0}^{\infty} f_i x^i$ and $A = T(a)$, the combination of the two expressions $\varphi_k(A) = \sum_{i=0}^{k} f_i A^i$ and $A^i = T(a^i) + E_i$, enables one to compute the quantity $\varphi_k(A)$ at a low computational effort. In fact, decomposing $\varphi_k(A)$ as $\varphi_k(A) = T(\varphi_k(a)) + F_k$, from $\varphi_{k+1}(A) = \varphi_k(A) + f_{k+1}A^{k+1}$ we deduce the equation

$$F_{k+1} = F_k + f_{k+1}E_k$$

for updating the correction part F_{k+1} in $\varphi_{k+1}(A)$. The above equation is easily implementable, moreover, representing F_k in the form $F_k = Y_k W_k^t$, where Y_k and W_k are matrices with infinitely many rows and a finite number of columns, and providing the same representation for E_k as $E_k = U_k V_k^t$, we may use the updating equation

$$Y_{k+1} = [Y_k \mid f_{k+1}U_k], \quad W_{k+1} = [W_k \mid V_k]. \tag{7.12}$$

Moreover, in order to keep low the number of columns in the matrices Y_{k+1} and W_{k+1}, one can apply a compression procedure based on the rank-revealing QR factorization and on SVD, to the two matrices in the right hand sides of (7.12). This strategy has been successfully used in [22] in the case of the exponential function.

Updating the Toeplitz part in $\varphi_k(A)$, that is, computing the coefficients of $\varphi_{k+1}(a(z))$ given those of $\varphi_k(a(z))$, can be performed by means of the evaluation/interpolation technique using as knots the roots of the unity of sufficiently large order. In fact, in this case we may rely on FFT to carry out the computation at a low cost.

k	time	band	rows	columns	rank	$\|E_{exp(a)}\|_{\mathcal{F}}$	bound (7.10)
1	$2.52 \cdot 10^{-2}$	35	17	17	7	3.58	40.17
2	$3.22 \cdot 10^{-2}$	55	32	37	8	14.4	436.79
3	$3.4 \cdot 10^{-2}$	78	48	59	8	38.7	3,636.12
4	$3.77 \cdot 10^{-2}$	104	47	85	8	92.4	24,407.44
5	$4.1 \cdot 10^{-2}$	133	48	114	8	214	$1.4 \cdot 10^5$
6	$4.34 \cdot 10^{-2}$	165	49	144	9	497	$7.21 \cdot 10^5$
7	$4.11 \cdot 10^{-2}$	199	53	178	9	1,170	$3.41 \cdot 10^6$
8	$4.52 \cdot 10^{-2}$	236	55	216	9	2,780	$1.51 \cdot 10^7$
9	$4.91 \cdot 10^{-2}$	274	54	252	9	6,720	$6.33 \cdot 10^7$
10	$5.16 \cdot 10^{-2}$	315	55	299	9	16,400	$2.55 \cdot 10^8$

Table 7.3.: Computation of $\exp(T(a))$ where $a(z) = \sum_{i=-1}^{k} z^i$.

A similar computational strategy can be used if $f(x)$ is assigned as a Laurent series in the form $\sum_{i\in\mathbb{Z}} f_i x^i$ so that $f(A)$ takes the form $f(A) = f_0 I + \sum_{i=1}^{\infty}(f_i A^i + f_{-i}A^{-i})$. Thus, once the matrix A^{-1} has been written in the form $A^{-1} = T(a^{-1}) + E_{a^{-1}}$, one can apply the above technique. Similar equations can be given in the case the Toeplitz matrix is finite and has a sufficiently large size.

As an example to show the effectiveness of our approach, we performed two numerical experiments. In the first one, we applied the above machinery to compute the exponential of the semi-infinite Toeplitz matrix $T(a)$ associated with the symbol $a(z) = \sum_{i=-1}^{k} z^i$ for $k = 1, 2, \ldots, 10$ corresponding to a Toeplitz matrix in Hessenberg form. In table 7.3 we report, besides the CPU time in seconds, the values of the numerical bandwidth of the exponential function, the dimension of the non-negligible part of the correction $E_{\exp}$ and its rank.

We point out that the approximation of $\exp(T(a))$ represented in the AQT form is quite good and that the CPU time needed for this computation is particularly low. We observe also that he rank of the correction has a moderate growth with respect to the band of $T(a)$.

In the second experiment, we consider matrices of finite size extending the AQT-arithmetic as pointed out in Section 7.5. More precisely, we applied the power series definition for computing $\exp(A)$, where $A = H^{10}$ and H is the $m \times m$ matrix $\mathrm{trid}(1, 2, 1)/(2 + 2\cos(\frac{\pi}{m+1}))$. In the numerical test we have chosen increasing values of m as integer powers of 10. Observe that, the matrix A is diagonalizable by means of the sine transform. Therefore, for all the matrices in the algebra generated by A and for any function f, it is possible to retrieve a particular column of $f(A)$ with linear cost. In order to validate the results, we report —as residual error— the Euclidean norm of the difference between the first column of the outcome and the first column of $\exp(A)$ computed by means of the sine transform. Table 7.4 shows the execution time in seconds,

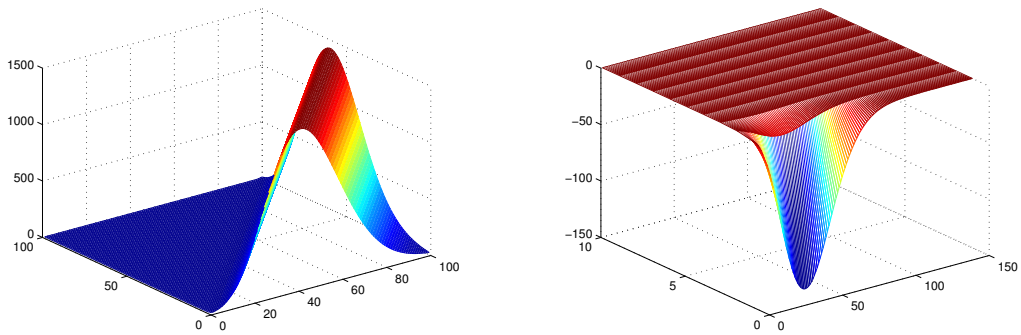

Figure 7.7.: Three-dimensional plot of $\exp(T(a))$ where $a(z) = \sum_{i=-1}^{10} z^i$; on the left the Toeplitz matrix $T(e^a)$; on the right the matrix $E_{exp(a)}$

Size	time	error	band	rows	columns	rank
100	$5.58 \cdot 10^{-2}$	$8.51 \cdot 10^{-16}$	87	81	49	15
1,000	$5.45 \cdot 10^{-2}$	$2.08 \cdot 10^{-15}$	87	65	45	15
10,000	$6.54 \cdot 10^{-2}$	$8.04 \cdot 10^{-16}$	87	65	45	15
$1 \cdot 10^5$	$8.22 \cdot 10^{-2}$	$1.87 \cdot 10^{-15}$	87	73	65	15
$1 \cdot 10^6$	$7.94 \cdot 10^{-2}$	$1.45 \cdot 10^{-15}$	87	46	65	15
$1 \cdot 10^7$	$8.18 \cdot 10^{-2}$	$1.04 \cdot 10^{-15}$	87	46	67	15

Table 7.4.: Computation of $\exp(A)$, with $A = H^{10}$ where $H = \text{trid}(1, 2, 1)/(2 + 2\cos(\frac{\pi}{m+1}))$ is an $m \times m$ matrix.

the residual errors, the Toeplitz bandwidth and the features of the correction. Note that, the features of only one correction are reported because, due to the symmetry of A, the upper left and lower right corner corrections are equal.

7.7.2 *Function of a CQT matrix: the Dunford-Cauchy integral*

The definition of $f(A)$ based on the contour integral can be easily extended to infinite matrices which represent bounded operators [89, 48].

Definition 7.7.6. *Let A be a semi-infinite matrix which represents a bounded linear operator on $\ell^2(\mathbb{Z}^+)$ and let $\Lambda = \{z \in \mathbb{C} : \quad zI - A$ is not invertible$\}$ be its spectrum. Given an analytic function $f(x)$ defined on a compact domain $\Omega \supseteq \Lambda$ having boundary $\partial\Omega$, $f(A)$ is defined as*

$$f(A) := \frac{1}{2\pi \boldsymbol{i}} \int_{\partial\Omega} f(z)\mathfrak{R}(z)dz \tag{7.13}$$

where $\mathfrak{R}(z) = (zI - A)^{-1}$ is the resolvent.

The integral formula (7.13) allows us to approximate $f(A)$ through a numerical integration scheme. That is, given a differentiable arc-length parametrization $\gamma : [a, b] \to \mathbb{C}$ of $\partial\Omega$ we can write

$$\frac{1}{2\pi \boldsymbol{i}} \int_{\partial\Omega} f(z)\mathfrak{R}(z)dz = \int_a^b g(x)dx$$

where $g(x) := \frac{1}{2\pi \boldsymbol{i}}\gamma'(x)f(\gamma(x))\mathfrak{R}(\gamma(x))$ is a matrix valued function. The above integral can be approximated by means of a quadrature formula with nodes x_k and weights w_k, i.e.,

$$\int_a^b g(x)dx \;\approx\; \sum_{k=1}^{N} w_k \cdot g(x_k). \tag{7.14}$$

The approximation schemes are determined with the strategy of increasing the number of nodes until the required precision is reached. If the weights are non-negative, the approximation (7.14) converges for $N \to \infty$ to $f(A)$.

We consider the trapezoidal approximation scheme with a doubling strategy for the nodes. That is, we consider the double indexed family $\{x_k^{(n)}, w_k^{(n)}\}$ such that:

- $n \in \mathbb{Z}^+$ and $k = 1, \ldots, 2^n + 1$,
- $a = x_1^{(n)} < x_2^{(n)} < \cdots < x_{2^n+1}^{(n)} = b$ are equally spaced points in $[a, b]$ $\forall n \in \mathbb{Z}^+$,
- $w_1^{(n)} = w_{2^n+1}^{(n)} = \frac{b-a}{2^{n+1}}$ and $w_k^{(n)} = \frac{b-a}{2^n}$, $k = 2, \ldots, 2^n$.

In particular, observe that the nodes at a certain step n correspond to those with odd indices at step $n + 1$.

Using a trapezoidal approximation of the integral (7.13) we can prove that the function of a CQT-matrix is again a CQT-matrix.

Theorem 7.7.7. *Let $A = T(a) + E_a$ be a* CQT*-matrix with spectrum Λ and symbol $a(z) \in \mathcal{W}_1$. Let $f(z)$ be an analytic function defined on the domain $\Omega \subset \mathbb{C}$ which encloses Λ such that $a(\mathbb{T}) \subset \Omega$. Assume that $\partial\Omega$ admits a differentiable arc length parametrization $\gamma : [a, b] \to \partial\Omega$. Then $f(A)$ is a* CQT*-matrix.*

Moreover, if $A \in \mathcal{AQT}$ then $f(A) \in \mathcal{AQT}$.

Proof. Given the family $\{x_k, w_k\}$ of nodes and weights of the trapezoidal approximation scheme for (7.13) we consider the sequence of rational functions in A:

$$\{r_n(A)\}_{n\in\mathbb{Z}^+} = \left\{ \sum_{k=1}^{2^n+1} w_k^{(n)} g(x_k^{(n)}) \right\}_{n\in\mathbb{Z}^+} = \left\{ \frac{b-a}{2^n} \sum_{k=1}^{2^n} g(x_k^{(n)}) \right\}_{n\in\mathbb{Z}^+}$$

where g is defined according to (7.14) and the latter equality follows from the fact that $\partial\Omega$ is a closed simple curve, thus $\gamma(x_1^{(n)}) = \gamma(x_{2^{n+1}}^{(n)})$. This sequence is formed by CQT-matrices whose limit, if it exists, has a Toeplitz part with symbol $f(a(z))$. Therefore, it is sufficient to show that this sequence is Cauchy with respect to the norm $\|\cdot\|_{\mathcal{CQT}}$.

Consider the difference

$$r_{n+1}(A) - r_n(A) = \frac{b-a}{2^{n+1}} \sum_{k=1}^{2^n} \left(g(x_{2k}^{(n+1)}) - g(x_{2k-1}^{(n+1)}) \right)$$

and observe that (for notational simplicity we omit the superscript $(n+1)$ in the nodes)

$$g(x_{2k}) - g(x_{2k-1}) = l(x_{2k})\mathfrak{R}(\gamma(x_{2k})) - l(x_{2k-1})\mathfrak{R}(\gamma(x_{2k-1}))$$

where $l : [a,b] \to \mathbb{C}$, $l(x) = \frac{1}{2\pi \boldsymbol{i}} \gamma'(x) f(\gamma(x))$. Assuming that $\gamma(x)$ has continuous second derivative, then $l(x)$ is a Lipschitz function. Indicating with L the Lipschitz constant of l and defining $M := \max_{\partial\Omega} \|\mathfrak{R}(z)\|_{\mathcal{CQT}}$, $G := \max_{[a,b]} |l(x)|$ we get

$$\begin{aligned}
\|g(x_{2k}) - g(x_{2k-1})\|_{\mathcal{CQT}} \leqslant\ & |l(x_{2k}) - l(x_{2k-1})| \cdot \|\mathfrak{R}(\gamma(x_{2k}))\|_{\mathcal{CQT}} \\
& + |l(x_{2k-1})| \cdot \|\mathfrak{R}(\gamma(x_{2k})) - \mathfrak{R}(\gamma(x_{2k-1}))\|_{\mathcal{CQT}} \\
\leqslant\ & L|x_{2k} - x_{2k-1}| \cdot \|\mathfrak{R}(\gamma(x_{2k}))\|_{\mathcal{CQT}} \\
& + |l(x_{2k-1})| \cdot |\gamma(x_{2k}) - \gamma(x_{2k-1})| \cdot \|\mathfrak{R}(\gamma(x_{2k}))\|_{\mathcal{CQT}} \|\mathfrak{R}(\gamma(x_{2k-1}))\|_{\mathcal{CQT}} \\
\leqslant\ & \frac{LM(b-a)}{2^{n+1}} + \frac{GM^2(b-a)}{2^{n+1}}
\end{aligned}$$

where we used $|\gamma(x_{2k}) - \gamma(x_{2k-1})| \leqslant |x_{2k} - x_{2k-1}|$ and the identity $\mathfrak{R}(z_1) - \mathfrak{R}(z_2) = (z_2 - z_1)\mathfrak{R}(z_1)\mathfrak{R}(z_2)$. In particular, we can write

$$\|r_{n+1}(A) - r_n(A)\|_{\mathcal{CQT}} \leqslant \frac{b-a}{2^{n+1}} \sum_{k=1}^{2^n} \frac{(LM + GM^2)(b-a)}{2^{n+1}} = c \cdot 2^{-(n+2)}$$

where $c := (b-a)^2(LM + GM^2)$ is independent of n. Therefore, given $n_2 > n_1$, we have

$$\begin{aligned}
\|r_{n_2}(A) - r_{n_1}(A)\|_{\mathcal{CQT}} &\leqslant \sum_{j=n_1}^{n_2-1} \|r_{j+1}(A) - r_j(A)\|_{\mathcal{CQT}} \\
&\leqslant c \sum_{j=n_1}^{n_2-1} 2^{-(j+2)} \leqslant c \cdot 2^{-(n_1+1)},
\end{aligned}$$

which proves that $\{r_n(A)\}_{n \in \mathbb{Z}^+}$ is a Cauchy sequence in the Banach algebra of $\mathcal{CQT}$-matrices. By relying on the same arguments used in the proof of Theorem 7.7.1 we deduce that $g(z) = f(a(z))$. So if $a(z)$ is analytic in a certain annulus $\mathbb{A}(r_a, R_a)$ containing $\mathbb{T}$ then there exists $\mathbb{A}(r, R) \subset \mathbb{A}(r_a, R_a)$ such that $a(\mathbb{A}(r, R)) \subset \mathbb{A}(r_f, R_f)$. Thus the composed function $g(z) = f(a(z))$ is analytic in $\mathbb{A}(r, R)$. This completes the proof. □

Computational aspects

Numerical integration based on the trapezoidal rule at the roots of unity can be easily implemented to approximate a matrix function assigned in terms of a Dunford-Cauchy

Size	time	error	band	rows	columns	rank
100	$5.83 \cdot 10^{-2}$	$8.51 \cdot 10^{-16}$	155	89	90	15
1,000	$7.93 \cdot 10^{-2}$	$2.08 \cdot 10^{-15}$	79	89	90	15
10,000	$8.15 \cdot 10^{-2}$	$8.04 \cdot 10^{-16}$	79	89	90	15
$1 \cdot 10^5$	$6.87 \cdot 10^{-2}$	$1.87 \cdot 10^{-15}$	79	89	89	15
$1 \cdot 10^6$	$8.37 \cdot 10^{-2}$	$1.45 \cdot 10^{-15}$	79	89	89	15
$1 \cdot 10^7$	$8.27 \cdot 10^{-2}$	$1.04 \cdot 10^{-15}$	79	89	90	15

Table 7.5.: Computation of $\sqrt{A}$, with $A = I + H^{10}$ where $H = \text{trid}(1, 2, 1)/(2 + 2\cos(\frac{\pi}{m+1}))$ is an $m \times m$ matrix.

Size	time	error	band	rows	columns	rank
100	1.9	$5.57 \cdot 10^{-14}$	87	89	90	15
1,000	1.88	$5.5 \cdot 10^{-14}$	159	90	90	15
10,000	1.53	$5.57 \cdot 10^{-14}$	159	89	90	15
$1 \cdot 10^5$	1.62	$5.56 \cdot 10^{-14}$	159	89	89	15
$1 \cdot 10^6$	1.99	$5.54 \cdot 10^{-14}$	159	90	89	15
$1 \cdot 10^7$	1.65	$5.56 \cdot 10^{-14}$	159	89	90	15

Table 7.6.: Computation of $\log(A)$, with $A = I + H^{10}$ where $H = \text{trid}(1, 2, 1)/(2 + 2\cos(\frac{\pi}{m+1}))$ is an $m \times m$ matrix.

integral. In fact all the operations involved in the computation reduce to performing matrix additions, multiplication of a matrix by a scalar and matrix inversion. The latter is the one with the highest computational cost.

We applied the contour integral definition for computing $\sqrt{I + H^{10}}$ and $\log(I + H^{10})$ where H is the $m \times m$ matrix $H = \text{trid}(1, 2, 1)/(2 + 2\cos(\frac{\pi}{m+1}))$ considered in Section 7.7.1. We used the trapezoidal rule with a doubling strategy for the nodes for integrating on a disc which contains the spectrum of $I + H^{10}$. Since H is rescaled to have spectrum in $[0, 1]$, we selected as center of the disc 1.5 and radius 1. Table 7.5-7.6 report the execution time, the residuals, the Toeplitz bandwidth and the features of the correction as the size of the argument increases exponentially. Once again, we reported only the features of one correction because, due to the symmetry of A, the upper left and lower right corner corrections are equal.

7.8 CONCLUSIONS AND RESEARCH LINES

We have introduced the class of semi-infinite quasi-Toeplitz matrices and proved that it is a Banach space with a suitable norm. Then we have considered the subspace formed by quasi-Toeplitz matrices associated with a continuous symbol $a(z)$ such that $a'(z) \in \mathcal{W}$,

and proved that it is a Banach algebra where the norm is sub-multiplicative. These properties have been used to define a matrix arithmetic on the algebra of semi-infinite CQT matrices. We are currently working on a MATLAB toolbox for handling such data structures. The beta version of this tool has been used to design methods for solving quadratic matrix equations with semi-infinite matrix coefficients encountered in QBD stochastic processes. In particular, this paves the way to design a procedure able to retrieve the stationary distribution of a level independent QBD in the positive quadrant of the plane (see Chapter 2). This should be —as far as we know— the first numerical algorithm for solving such issue in general hypotheses and will be part of future research.

We have extended the concept of matrix function to CQT matrices, i.e., infinite matrices of the form $A = T(a) + E$, by showing that, under suitable mild assumptions, for a CQT matrix A and for a function $f(x)$ expressed either in terms of a power (Laurent) series, or in terms of the Dunford-Cauchy integral, the matrix function $f(A)$ is still a CQT matrix. We have outlined algorithms for the computation of $f(A)$. This approach has been adapted to the case of $f(A_m)$ where A_m is the $m \times m$ leading principal submatrix of A.

Among the open issues that will be part of our research interests, it would be interesting to analyze the behavior of the singular values $\sigma_i^{(k)}$ of the $m \times m$ truncation of the correction E_k such that $(T(a) + E)^k = T(a^k) + E_k$, and relate the decay of these values for $i = 1, 2, \ldots, m$ and for $k = 1, 2, \ldots$, to the qualitative properties of the function $a(z)$. In fact, from the numerical experiments that we have performed with several functions $a(z)$, it turns out that the numerical rank of E_k remains bounded by a constant independent of k.

Finally, a somewhat natural extension of the tools developed in this chapter is the use of multivariate symbols $a(z_1, \ldots, z_s)$. The development of an analogous arithmetic for this framework means the management of finite and semi-infinite multilevel Toeplitz structures. We think that this deserves further investigations.

Chapter 8

Concluding remarks

In this work we have seen a number of aspects that are taken into account when a class of structured matrices is studied:

(i) efficient representation,

(ii) fast arithmetic,

(iii) preservation of the structure,

(iv) computing functions of structured matrices,

(v) solving matrix equations with structured coefficients.

One of our main contributions is a framework for analyzing the numerical preservation of quasiseparability in matrix computations. Using this tool, we managed to state and proof bounds on the growth of the quasiseparable rank when computing a matrix function and executing the cyclic reduction algorithm (CR). It deserves to be pointed out that we often retrieved a connection between the quasiseparable preservation and issues of approximation theory.

We focused on the HODLR representation for exploiting the structure and we tested its effectiveness in the CR. This yielded fast procedures for solving certain linear and quadratic structured matrix equations.

Another aspect that would be interesting to deepen is the use of the $\mathcal{H}^2$ format in place of HODLR matrices. In theory this approach can remove the logarithmic factors in the complexity of the matrix operations by means of nested basis techniques, see [54].

Motivated by the applications to stochastic processes, we introduced a new class of structured matrices —the CQT matrices— which can model finite and infinite data structures. Issues (i)-(v) have been addressed both from the theoretical and practical point of view. The resulting fast arithmetic of CQT matrices is an original contribution. It has two important benefits:

- it allows to carry on otherwise not feasible computations because of the infinite matrices involved,
- in the finite size case, it exploits the Toeplitz structure without using displacement properties.

The first property paves the way to deal with infinite version of problems which are well-studied in the finite case, e.g., finding the stationary distribution of QBD processes.

We used the second property to provide alternative methods for the fast computation of functions of large scale Toeplitz and quasi-Toeplitz matrices.

Many other questions and ideas have been briefly summarized at the end of each chapter. We look forward to explore these research lines.

Appendix A

A technical result

Proposition A.0.1. *Let $f \in \mathcal{C}^\infty(\mathbb{C})$ and $\lambda \in \mathbb{C}$ then $\forall d \in \mathbb{Z}^+, h \in \mathbb{N}$*

$$\frac{\partial^{d-1}}{\partial z^{d-1}}\left(\frac{f(z)}{(z-\lambda)^{h+1}}\right) = \frac{(d-1)!}{h!}\sum_{l=1}^{d}(-1)^{l+h+1}\frac{(l+h-1)!}{(d-l)!(l-1)!}f^{(d-l)}(z)(z-\lambda)^{-(h+l)}.$$

Proof. For every fixed $h \in \mathbb{N}$ we proceed by induction on d. For $d = 1$ we get

$$\frac{f(z)}{(z-\lambda)^{h+1}} = \frac{0!}{h!}(-1)^2\frac{h!}{0!0!}\frac{f(z)}{(z-\lambda)^{h+1}}.$$

For the inductive step, let $d > 1$ and observe that

$$\begin{aligned}
\frac{\partial^d}{\partial z^d}\left(\frac{f(z)}{(z-\lambda)^{h+1}}\right) &= \frac{\partial}{\partial z}\left(\frac{\partial^{d-1}}{\partial z^{d-1}}\left(\frac{f(z)}{(z-\lambda)^{h+1}}\right)\right)\\
&= \frac{\partial}{\partial z}\left(\frac{(d-1)!}{h!}\sum_{l=1}^{d}(-1)^{l+h+1}\frac{(l+h-1)!}{(d-l)!(l-1)!}f^{(d-l)}(z)(z-\lambda)^{-(h+l)}\right)\\
&= \frac{(d-1)!}{h!}\sum_{l=1}^{d}(-1)^{l+h+1}\frac{(l+h-1)!}{(d-l)!(l-1)!}f^{(d+1-l)}(z)(z-\lambda)^{-(h+l)}\\
&+ \frac{(d-1)!}{h!}\sum_{l=1}^{d}(-1)^{l+h+2}(h+l)\frac{(l+h-1)!}{(d-l)!(l-1)!}f^{(d-l)}(z)(z-\lambda)^{-(h+l+1)}\\
&= \frac{(d-1)!}{h!}\sum_{l=1}^{d}(-1)^{l+h+1}\frac{(l+h-1)!}{(d-l)!(l-1)!}f^{(d+1-l)}(z)(z-\lambda)^{-(h+l)}\\
&+ \frac{(d-1)!}{h!}\sum_{l=2}^{d+1}(-1)^{l+h+1}(h+l-1)\frac{(l+h-2)!}{(d+1-l)!(l-2)!}f^{(d+1-l)}(z)(z-\lambda)^{-(h+l)}\\
&= \frac{d!}{h!}\sum_{l=1}^{d+1}(-1)^{l+h+1}\frac{(l+h-1)!}{(d+1-l)!(l-1)!}f^{(d+1-l)}(z)(z-\lambda)^{-(h+l)}.
\end{aligned}$$

□

Appendix B

Computing the spectral factorization

Let $p(z)$ be a polynomial of degree n with complex coefficients and such that its roots $\xi_1, \ldots, \xi_n$ verify

$$|\xi_1| \leqslant \ldots \leqslant |\xi_m| < 1 < |\xi_{m+1}| \leqslant \ldots |\xi_n| \tag{B.1}$$

and define

$$u(z) := \prod_{i=1}^{m}(z - \xi_i) = \sum_{i=0}^{m} u_{m-i} z^i, \qquad l(z) := \prod_{i=m+1}^{n}(z - \xi_i) = \sum_{i=0}^{n-m} l_i z^i,$$

such that $p(z) = p_n \cdot u(z)l(z)$. Observe that the factors $u(z)$ and $l(z)$ are strictly linked to the spectral factorization of $a(z) := z^{-m}p(z)$, in fact:

$$a(z) = z^{-m}p_n \cdot u(z)l(z) = p_n \cdot u_R(z^{-1})l(z)$$

where $u_R(z) := \sum_{i=0}^{m} u_i z^i$. The strategy used in [14] for computing the coefficients of $u(z)$ and $l(z)$ rely on the following result.

Theorem B.0.2 ([14])**.** *Under the assumption* (B.1) *there exists a Laurent series* $x(z) := \sum_{i=-\infty}^{+\infty} x_i z^i$ *such that* $a(z)x(z) = 1$ *and* $x(z) \in \mathcal{W}$*. Moreover for every* $q > \max\{m, m - n\}$ *the* $q \times q$*-Toeplitz matrix* $T = (x_{i-j})_{i,j=1,\ldots q}$ *is such that*

$$T\mathbf{l} = (a_{n-m}u_0)^{-1}e_1, \qquad T^t\mathbf{u} = (a_{n-m}l_0)^{-1}e_1,$$

where $\mathbf{l} := (l_0, \ldots, l_{n-m}, 0, \ldots, 0)^t \in \mathbb{C}^q$ *and* $\mathbf{u} := (u_0, \ldots, u_m, 0, \ldots, 0)^t \in \mathbb{C}^q$*.*

The previous theorem suggests this general scheme:

(i) Choose $q > \max\{m, m - n\}$ and compute the central coefficients $x_{-q}, \ldots, x_q$ of the Laurent series $x(z)$ such that $a(z)x(z) = 1$.

(ii) Define $T = (x_{i-j})_{i,j=1,\ldots q}$ and solve the two linear systems $T\mathbf{l} = e_1$ and $T^t\mathbf{u} = e_1$.

We deal with the problem of performing step (i) in the next section. Step (ii) consists of solving two finite linear systems with a Toeplitz coefficient matrix. This task can be handled by means of the customary algorithms, like the fast, superfast, or iterative techniques [64] with a cost ranging form $q \log q$ to q^2.

B.1 COMPUTE THE INVERSE OF A LAURENT POLYNOMIAL

We want to compute the central coefficients of the inverse of a Laurent polynomial $a(z) = z^{-m}p(z)$ whose zeros verify (B.1). We indicate with x_i the exact coefficient of $a(z)^{-1}$ and with $\tilde{x}_i$ the approximation of the latter, computed using the following evaluation/interpolation strategy [14].

1. Choose N a large enough power of 2,
2. Evaluate $a(z)$ at the N-th roots of 1 getting $w_i = a(\zeta_N^i),\ \ i = 0, \ldots, N-1$,
3. Compute $t_i = \frac{1}{w_i},\ \ i = 0, \ldots, N-1$,
4. Interpolate $(\zeta_N^i, t_i),\ \ i = 0, \ldots, N-1$ with the inverse DFT and obtain the coefficients $s_j = \frac{1}{N}\sum_{i=0}^{N-1} \zeta_N^{-ij} t_i,\ \ j = 0, \ldots, N-1$,
5. Return $\tilde{x}_j = s_{j \bmod N},\ \ j = -\frac{N}{2}, \ldots, \frac{N}{2} - 1$.

The choice of N affects the accuracy of the coefficients and can be performed in an adaptive way. This consists in doubling the number of nodes and comparing the coefficients obtained with N and $2N$ nodes, respectively. If the variation (evaluated in a certain norm) of the coefficients with indeces $-\frac{N}{2}, \ldots \frac{N}{2} - 1$ is under a given threshold then we stop, otherwise we double the nodes and we repeat the procedure.

An alternative way to compute the central coefficients of $a(z)^{-1}$ relies on the Graeffe algorithm. Once again, we refer to [14] for a complete description and a comparison between the evaluation/interpolation method and the Graeffe iteration.

Appendix C

Two-sided Lanczos method

The aim of the two sided Lanczos method [90] is to compute an approximation of rank k of a matrix $A \in \mathbb{C}^{m\times n}$, exploiting the characterization of the singular vectors as eigenvectors of A^tA and AA^t, respectively. The idea is to generate orthonormal bases of the Krylov sub-spaces

$$\text{span}\{u_1, AA^tu_1, \dots, (AA^t)^{k-1}u_1\}, \qquad \text{span}\{v_1, A^tAv_1, \dots, (A^tA)^{k-1}v_1\}, \tag{C.1}$$

where u_1 is a starting guess with unit Euclidean norm and $v_1 := \frac{A^tu_1}{\|A^tu_1\|_2}$. In order to retrieve such bases, a Grahm Schmidt process is carried out, see lines 2-6 of Algorithm 2. Then, we get two matrices $U_k \in \mathbb{R}^{m\times k}$ and $V_k \in \mathbb{R}^{n\times k}$ whose columns form orthonormal bases of the Krylov sub-spaces (C.1).

Algorithm 2 Pseudocode for the two sided Lanczos algorithm

1: **procedure** TwoSidedLanczos(A, u_1, k) ▷ Compute U_k, Σ_k, V_k such that $A \approx U_k\Sigma_kV_k^t$
2: $\tilde{v} \leftarrow A^tu_1, \quad \alpha_1 \leftarrow \|\tilde{v}\|_2, \quad v_1 \leftarrow \frac{\tilde{v}}{\alpha_1}$
3: **for** $j = 1, \dots, k$ **do**
4: $\tilde{u} \leftarrow Av_j - \alpha_ju_j, \quad \beta_{j+1} \leftarrow \|\tilde{u}\|_2, \quad u_{j+1} \leftarrow \frac{\tilde{u}}{\beta_{j+1}}$
5: $\tilde{v} \leftarrow A^tu_{j+1} - \beta_{j+1}v_j, \quad \alpha_{j+1} \leftarrow \|\tilde{v}\|_2, \quad v_{j+1} \leftarrow \frac{\tilde{v}}{\alpha_{j+1}}$
6: **end for**
7: Set $U_k = (u_1, \dots, u_k)$ and $V_k = (v_1, \dots, v_k)$
8: Compute the SVD $(\widehat{U}, \widehat{\Sigma}, \widehat{V})$ of the matrix B_k defined as in (C.2)
9: $U_k \leftarrow U_k\widehat{U}, \quad \Sigma_k \leftarrow \widehat{\Sigma}, \quad V_k \leftarrow V_k\widehat{V}$
10: **return** U_k, Σ_k, V_k
11: **end procedure**

Moreover, the relation between u_1 and v_1 implies that

$$A^tU_k = V_kB_k^t, \qquad AV_k = U_kB_k + \beta_{k+1}u_{k+1}e_k^t,$$

where e_k is the k-th unit vector of length k and

$$B_k := \begin{bmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \beta_k & \alpha_k \end{bmatrix}. \tag{C.2}$$

Finally, the rank-k approximation of A is computed as

$$A_k = U_k B_k V_k^t,$$

so for retrieving an outer product representation it is sufficient to compute the SVD of B_k. Note that, in order to avoid loss of orthogonality, one needs to re-orthogonalize the vectors $\tilde{u}$ and $\tilde{v}$ —in lines 3 and 4— with respect to the previously computed $u_1, \ldots, u_{j-1}$ and $v_1, \ldots, v_{j-1}$, respectively.

It is possible to choose adaptively the rank k of the approximation using a stopping criterion which depends on the computed α_j and β_j. The heuristic choice we usually made in our experiments is $\max\{|\alpha_j|, |\beta_j|\}$ less than a given threshold.

Bibliography

[1] N. I. Akhiezer. *Elements of the theory of elliptic functions*, volume 79 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1990. Translated from the second Russian edition by H. H. McFaden.

[2] K. M. R. Audenaert. A singular value inequality for Heinz means. *Linear Algebra Appl.*, 422(1):279–283, 2007.

[3] K. M. R. Audenaert and F. Kittaneh. Problems and conjectures in matrix and operator inequalities. arXiv preprint arXiv:1201.5232, 2012.

[4] W. W. Barrett. A theorem on inverses of tridiagonal matrices. *Linear Algebra Appl.*, 27:211–217, 1979.

[5] W. W. Barrett and P. J. Feinsilver. Inverses of banded matrices. *Linear Algebra Appl.*, 41:111–130, 1981.

[6] N. Bean and G. Latouche. Approximations to quasi-birth-and-death processes with infinite blocks. *Adv. in Appl. Probab.*, 42(4):1102–1125, 2010.

[7] B. Beckermann. Singular values of small displacement rank matrices. Talk at conference Structured Numerical Linear Algebra Problems: Algorithms and Applications, Cortona, 2004.

[8] B. Beckermann and A. Townsend. On the singular values of matrices with displacement structure. arXiv preprint arXiv:1609.09494, 2016.

[9] M. Benzi and P. Boito. Decay properties for functions of matrices over C^*-algebras. *Linear Algebra Appl.*, 456:174–198, 2014.

[10] M. Benzi, P. Boito, and N. Razouk. Decay properties of spectral projectors with applications to electronic structure. *SIAM Rev.*, 55(1):3–64, 2013.

[11] M. Benzi and V. Simoncini. Decay bounds for functions of Hermitian matrices with banded or Kronecker structure. *SIAM J. Matrix Anal. Appl.*, 36(3):1263–1282, 2015.

[12] A. Berman and R. J. Plemmons. *Nonnegative matrices in the mathematical sciences*, volume 9 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994. Revised reprint of the 1979 original.

[13] D. A. Bini, S. Dendievel, G. Latouche, and B. Meini. Computing the exponential of large block-triangular block-Toeplitz matrices encountered in fluid queues. *Linear Algebra Appl.*, 502:387–419, 2016.

[14] D. A. Bini, G. Fiorentino, L. Gemignani, and B. Meini. Effective fast algorithms for polynomial spectral factorization. *Numer. Algorithms*, 34(2-4):217–227, 2003. International Conference on Numerical Algorithms, Vol. II (Marrakesh, 2001).

[15] D. A. Bini, L. Gemignani, and B. Meini. Computations with infinite Toeplitz matrices and polynomials. *Linear Algebra Appl.*, 343/344:21–61, 2002. Special issue on structured and infinite systems of linear equations.

[16] D. A. Bini, B. Iannazzo, and B. Meini. *Numerical solution of algebraic Riccati equations*, volume 9 of *Fundamentals of Algorithms*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2012.

[17] D. A. Bini, G. Latouche, and B. Meini. *Numerical methods for structured Markov chains*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2005. Oxford Science Publications.

[18] D. A. Bini, S. Massei, and B. Meini. On Functions of quasi Toeplitz matrices. arXiv preprint arXiv:1611.06406, 2016.

[19] D. A. Bini, S. Massei, and B. Meini. Semi-Infinite Quasi-Toeplitz Matrices with Applications to QBD Stochastic Processes. arXiv preprint arXiv:1611.06337, 2016.

[20] D. A. Bini, S. Massei, and L. Robol. Efficient cyclic reduction for Quasi-Birth-Death problems with rank structured blocks. *Appl. Numer. Math.*, 2016.

[21] D. A. Bini, S. Massei, and L. Robol. On the decay of the off-diagonal singular values in cyclic reduction. arXiv preprint arXiv:1608.01567, 2016.

[22] D. A. Bini and B. Meini. On the exponential of semi-infinite quasi Toeplitz matrices. arXiv preprint arXiv:1611.06380.

[23] D. A. Bini and B. Meini. The cyclic reduction algorithm: from Poisson equation to stochastic processes and beyond. In memoriam of Gene H. Golub. *Numer. Algorithms*, 51(1):23–60, 2009.

[24] P. Boito, Y. Eidelman, L. Gemignani, and I. Gohberg. Implicit QR with compression. *Indag. Math. (N.S.)*, 23(4):733–761, 2012.

[25] S. Börm. H2lib. Available from GitHub at `https://github.com/H2Lib/H2Lib`, 2015.

[26] S. Börm, L. Grasedyck, and W. Hackbusch. Hierarchical matrices. *Lecture notes*, 21:2003, 2003.

[27] A. Böttcher and S. M. Grudsky. *Toeplitz matrices, asymptotic linear algebra, and functional analysis*. Birkhäuser Verlag, Basel, 2000.

[28] A. Böttcher and S. M. Grudsky. *Spectral properties of banded Toeplitz matrices*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005.

[29] A. Böttcher and B. Silbermann. *Introduction to large truncated Toeplitz matrices*. Universitext. Springer-Verlag, New York, 1999.

[30] D. R. Brillinger. The analyticity of the roots of a polynomial as functions of the coefficients. *Math. Mag.*, 39:145–147, 1966.

[31] B. L. Buzbee, G. H. Golub, and C. W. Nielson. On direct methods for solving Poisson's equations. *SIAM J. Numer. Anal.*, 7:627–656, 1970.

[32] C. Canuto, V. Simoncini, and M. Verani. On the decay of the inverse of matrices that are sum of Kronecker products. *Linear Algebra Appl.*, 452:21–39, 2014.

[33] S. Chandrasekaran, M. Gu, J. Xia, and J. Zhu. A fast QR algorithm for companion matrices. In *Recent advances in matrix and operator theory*, volume 179 of *Oper. Theory Adv. Appl.*, pages 111–143. Birkhäuser, Basel, 2008.

[34] S. Chang. On the distribution of the characteristic values and singular values of linear integral equations. *Trans. Amer. Math. Soc.*, 67:351–367, 1949.

[35] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT Press, Cambridge, MA, third edition, 2009.

[36] M. Crouzeix. Numerical range and functional calculus in Hilbert space. *J. Funct. Anal.*, 244(2):668–690, 2007.

[37] S. W. Drury. On a question of Bhatia and Kittaneh. *Linear Algebra Appl.*, 437(7):1955–1960, 2012.

[38] Y. Eidelman and I. Gohberg. On generators of quasiseparable finite block matrices. *Calcolo*, 42(3-4):187–214, 2005.

[39] Y. Eidelman, I. Gohberg, and L. Gemignani. On the fast reduction of a quasiseparable matrix to Hessenberg and tridiagonal forms. *Linear Algebra Appl.*, 420(1):86–101, 2007.

[40] Y. Eidelman, I. Gohberg, and I. Haimovici. *Separable type representations of matrices and fast algorithms. Vol. 1*, volume 234 of *Operator Theory: Advances and Applications*. Birkhäuser/Springer, Basel, 2014. Basics. Completion problems. Multiplication and inversion algorithms.

[41] S. W. Ellacott. Computation of Faber series with application to numerical polynomial approximation in the complex plane. *Math. Comp.*, 40(162):575–587, 1983.

[42] K. Fan. Maximum properties and inequalities for the eigenvalues of completely continuous operators. *Proc. Nat. Acad. Sci., U. S. A.*, 37:760–766, 1951.

[43] M. Fiedler and T. L. Markham. Completing a matrix when certain entries of its inverse are specified. *Linear Algebra Appl.*, 74:225–237, 1986.

[44] H. R. Gail, S. L. Hantler, and B. A. Taylor. Matrix-geometric invariant measures for $G/M/1$ type Markov chains. *Comm. Statist. Stochastic Models*, 14(3):537–569, 1998.

[45] F. R. Gantmacher. *The theory of matrices. Vols. 1, 2.* Translated by K. A. Hirsch. Chelsea Publishing Co., New York, 1959.

[46] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij. $\mathcal{H}$-matrix approximation for the operator exponential with applications. *Numer. Math.*, 92(1):83–111, 2002.

[47] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij. Data-sparse approximation to a class of operator-valued functions. *Math. Comp.*, 74(250):681–708, 2005.

[48] M. I. Gil'. Estimates for entries of matrix valued functions of infinite matrices. *Math. Phys. Anal. Geom.*, 11(2):175–186, 2008.

[49] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.

[50] W. H. Gustafson. A note on matrix inversion. *Linear Algebra Appl.*, 57:71–73, 1984.

[51] J. Gutiérrez-Gutiérrez, P. M. Crespo, and A. Böttcher. Functions of the banded Hermitian block Toeplitz matrices in signal processing. *Linear Algebra Appl.*, 422(2-3):788–807, 2007.

[52] S. Güttel, E. Polizzi, P. T. P. Tang, and G. Viaud. Zolotarev quadrature rules and load balancing for the FEAST eigensolver. *SIAM J. Sci. Comput.*, 37(4):A2100–A2122, 2015.

[53] W. Hackbusch. A sparse matrix arithmetic based on $\mathcal{H}$-matrices. I. Introduction to $\mathcal{H}$-matrices. *Computing*, 62(2):89–108, 1999.

[54] W. Hackbusch. *Hierarchical matrices: algorithms and analysis*, volume 49 of *Springer Series in Computational Mathematics*. Springer, Heidelberg, 2015.

[55] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, 2011.

[56] Q. He. *Fundamentals of matrix-analytic methods*. Springer, New York, 2014.

[57] D. Heller. Some aspects of the cyclic reduction algorithm for block tridiagonal linear systems. *SIAM J. Numer. Anal.*, 13(4):484–496, 1976.

[58] P. Henrici. Computational complex analysis. In *The influence of computing on mathematical research and education (Proc. Sympos. Appl. Math., Vol. 20, Univ. Montana, Missoula, Mont., 1973)*, pages 79–86. Amer. Math. Soc., Providence, R.I., 1974.

[59] N. J. Higham. *Functions of matrices*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008. Theory and computation.

[60] R. W. Hockney. A fast direct solution of Poisson's equation using Fourier analysis. *J. Assoc. Comput. Mach.*, 12:95–113, 1965.

[61] A. Horn. On the singular values of a product of completely continuous operators. *Proc. Nat. Acad. Sci. U. S. A.*, 36:374–375, 1950.

[62] R. A. Horn and C. R. Johnson. *Topics in matrix analysis*. Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.

[63] J. Jain, H. Li, C. Koh, and V. Balakrishnan. $O(n)$ algorithms for banded plus semiseparable matrices. In *Numerical methods for structured matrices and applications*, volume 199 of *Oper. Theory Adv. Appl.*, pages 347–358. Birkhäuser Verlag, Basel, 2010.

[64] T. Kailath and A. H. Sayed, editors. *Fast reliable algorithms for matrices with structure*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999.

[65] M. Kobayashi and M. Miyazawa. Revisiting the tail asymptotics of the double QBD process: refinement and complete solutions for the coordinate and diagonal directions. In *Matrix-analytic methods in stochastic models*, volume 27 of *Springer Proc. Math. Stat.*, pages 145–185. Springer, New York, 2013.

[66] D. Kressner. Bivariate matrix functions. *Oper. Matrices*, 8(2):449–466, 2014.

[67] D. Kressner and R. Luce. Fast computation of the matrix exponential for a Toeplitz matrix. arXiv preprint arXiv:1607.01733, 2016.

[68] D. Kressner and A. Susnjara. Fast computation of spectral projectors of banded matrices. arXiv preprint arXiv:1608.01164, 2016.

[69] P. Lancaster and M. Tismenetsky. *The theory of matrices*. Computer Science and Applied Mathematics. Academic Press, Inc., Orlando, FL, second edition, 1985.

[70] G. Latouche, G. T. Nguyen, and P. G. Taylor. Queues with boundary assistance: the effects of truncation. *Queueing Syst.*, 69(2):175–197, 2011.

[71] G. Latouche and V. Ramaswami. *Introduction to matrix analytic methods in stochastic modeling*. ASA-SIAM Series on Statistics and Applied Probability. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; American Statistical Association, Alexandria, VA, 1999.

[72] G. Latouche and P. Taylor. Truncation and augmentation of level-independent QBD processes. *Stochastic Process. Appl.*, 99(1):53–80, 2002.

[73] M. Lindner. *Infinite matrices and their finite sections*. Frontiers in Mathematics. Birkhäuser Verlag, Basel, 2006. An introduction to the limit operator method.

[74] A. I. Markushevich. *Theory of functions of a complex variable. Vol. I, II, III*. Chelsea Publishing Co., New York, english edition, 1977. Translated and edited by Richard A. Silverman.

[75] S. Massei and L. Robol. H2lib's matlab bindings. Available from GitHub at `https://github.com/robol/h2lib-matlab`, 2015.

[76] S. Massei and L. Robol. Decay bounds for the numerical quasiseparable preservation in matrix functions. *Linear Algebra and its Applications*, 2016.

[77] A. A. Medovikov and V. I. Lebedev. Variable time steps optimization of L_ω-stable Crank-Nicolson method. *Russian J. Numer. Anal. Math. Modelling*, 20(3):283–303, 2005.

[78] M. Miyazawa. Tail decay rates in double QBD processes and related reflected random walks. *Math. Oper. Res.*, 34(3):547–575, 2009.

[79] M. Miyazawa. Light tail asymptotics in multidimensional reflecting processes for queueing networks. *TOP*, 19(2):233–299, 2011.

[80] A. J. Motyer and P. G. Taylor. Decay rates for quasi-birth-and-death processes with countably many phases and tridiagonal block generators. *Adv. in Appl. Probab.*, 38(2):522–544, 2006.

[81] M. F. Neuts. *Matrix-geometric solutions in stochastic models*, volume 2 of *Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins University Press, Baltimore, Md., 1981. An algorithmic approach.

[82] A. Ostrowski. Recherches sur la méthode de Graeffe et les zéros des polynomes et des séries de Laurent. Chapitres III et IV. *Acta Math.*, 72:157–257, 1940.

[83] D. Palitta and V. Simoncini. Matrix-equation-based strategies for convection-diffusion equations. *BIT*, 56(2):751–776, 2016.

[84] G. Polya. Remark on Weyl's note "Inequalities between the two kinds of eigenvalues of a linear transformation.". *Proc. Nat. Acad. Sci. U. S. A.*, 36:49–51, 1950.

[85] S. Pozza and V. Simoncini. Decay bounds for non-hermitian matrix functions. arXiv preprint arXiv:1605.01595, 2016.

[86] Y. Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2003.

[87] E. B. Saff. Logarithmic potential theory with applications to approximation theory. *Surv. Approx. Theory*, 5:165–200, 2010.

[88] Y. Sakuma and M. Miyazawa. On the effect of finite buffer truncation in a two-node Jackson network. *J. Appl. Probab.*, 42(1):199–222, 2005.

[89] M. Shao. On the finite section method for computing exponentials of doubly-infinite skew-Hermitian matrices. *Linear Algebra Appl.*, 451:65–96, 2014.

[90] H. D. Simon and H. Zha. Low-rank matrix approximation using the Lanczos bidiagonalization process with applications. *SIAM J. Sci. Comput.*, 21(6):2257–2274 (electronic), 2000.

[91] D. Stanford, W. Horn, and G. Latouche. Tri-layered QBD processes with boundary assistance for service resources. *Stoch. Models*, 22(3):361–382, 2006.

[92] I. A. Stegun, editor. *Pocketbook of mathematical functions*. Verlag Harri Deutsch, Thun, 1984. Abridged edition of ıt Handbook of mathematical functions edited by Milton Abramowitz and Irene A. Stegun, Material selected by Michael Danos and Johann Rafelski.

[93] P. N. Swarztrauber and R. A. Sweet. Vector and parallel methods for the direct solution of Poisson's equation. *J. Comput. Appl. Math.*, 27(1-2):241–263, 1989. Reprinted in ıt Parallel algorithms for numerical linear algebra, 241–263, North-Holland, Amsterdam, 1990.

[94] Y. Tao. More results on singular value inequalities of matrices. *Linear Algebra Appl.*, 416(2-3):724–729, 2006.

[95] L. N. Trefethen and J. A. C. Weideman. The exponentially convergent trapezoidal rule. *SIAM Rev.*, 56(3):385–458, 2014.

[96] E. Van Camp. *Diagonal-plus-semiseparable matrices and their use in numerical linear algebra*. PhD thesis, PhD thesis, Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan, 2005.

[97] R. Vandebril, M. Van Barel, and N. Mastronardi. *Matrix computations and semiseparable matrices. Vol. 1.* Johns Hopkins University Press, Baltimore, MD, 2008. Linear systems.

[98] R. Vandebril, M. Van Barel, and N. Mastronardi. *Matrix computations and semiseparable matrices. Vol. II.* Johns Hopkins University Press, Baltimore, MD, 2008. Eigenvalue and singular value methods.

[99] H. Weyl. Inequalities between the two kinds of eigenvalues of a linear transformation. *Proc. Nat. Acad. Sci. U. S. A.*, 35:408–411, 1949.

[100] X. Zhan. Some research problems on the Hadamard product and singular values of matrices. *Linear and Multilinear Algebra*, 47(2):191–194, 2000.

[101] E. I. Zolotarev. Application of elliptic functions to questions of functions deviating least and most from zero. *Zap. Imp. Akad. Nauk St. Petersburg*, 21:30:1–59, 1877.

Index