

**OPTIMAL
HANKEL STRUCTURED
RANK-1 APPROXIMATION**

HANNA ELISABETH KNIRSCH

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

OPTIMAL HANKEL STRUCTURED RANK-1 APPROXIMATION

DISSERTATION

for the award of the degree

DOCTOR RERUM NATURALIUM

of the Georg-August-Universität Göttingen

within the doctoral program *Mathematical Sciences*

of the Georg-August University School of Science (GAUSS)

submitted by

HANNA ELISABETH KNIRSCH

from Erlangen

Göttingen, 2021

Thesis Committee

Prof. Dr. Gerlind PLONKA-HOCH

Institute of Numerical and Applied Mathematics, Georg-August-Universität Göttingen

Prof. Dr. D. Russell LUKE

Institute of Numerical and Applied Mathematics, Georg-August-Universität Göttingen

Dr. Matthew K. TAM

School of Mathematics and Statistics, University of Melbourne

Members of the Examination Board

Reviewer

Prof. Dr. Gerlind PLONKA-HOCH

Institute of Numerical and Applied Mathematics, Georg-August-Universität Göttingen

Second Reviewer

Prof. Dr. Dirk A. LORENZ

Institute of Analysis und Algebra, Technische Universität Braunschweig

Further Members of the Examination Board

Prof. Dr. D. Russell LUKE

Institute of Numerical and Applied Mathematics, Georg-August-Universität Göttingen

Prof. Dr. Dominic SCHUHMACHER

Institute for Mathematical Stochastics, Georg-August-Universität Göttingen

Jun.-Prof. Dr. Anne WALD

Institute of Numerical and Applied Mathematics, Georg-August-Universität Göttingen

Prof. Dr. Damaris SCHINDLER

Mathematical Institute, Georg-August-Universität Göttingen

Date of the oral examination: 16 February 2022

*For there is always light
if only we're brave enough to see it,
if only we're brave enough to be it.*

AMANDA GORMAN

THANKS

I would like to express my heartfelt gratitude to all the people who have, with their support, contributed to the success of this dissertation. Special thanks are due to everyone who with their thorough proofreading and invaluable advice helped improve my writing.

In the first place, I would like to thank my adviser Gerlind Plonka-Hoch. I deeply admire the quiet enthusiasm and perseverance with which she conducts her research. Without her continuous encouragement and support, this dissertation would not exist. My sincere thanks also go to my co-advisers Russell Luke and Matthew Tam for their keen interest in my work. Moreover, I am much obliged to Dirk Lorenz who invited me to give a talk in his seminar and readily agreed to co-review this dissertation.

During my doctoral studies, I was a member of the Research Training Group (RTG) 2088 *Discovering Structure in Complex Data: Statistics meets Optimization and Inverse Problems* funded by the German Research Foundation (DFG). I gratefully acknowledge their generous financial support as well as the possibility to attend scientific conferences and to meet with fascinating people. In this context, I want to express my thanks to Diana Sieber who managed and organized the above, and who always was a reliable source of information.

I am deeply grateful for all my wonderful colleagues and friends from the *Mathematical Signal and Image Processing* group, the RTG 2088, and beyond. I especially thank Inge Keller, Markus Petz, Raha Razavi, and Benjamin Kocurov for the unforgettably pleasant working environment and the fun at lunch breaks. Furthermore, I particularly thank Katharina Müller and Moritz Wemheuer who have been valuable companions and friends ever since we set foot in this university.

My profound thanks also goes to the Ultimate Frisbee team *Göttinger 7* for the compensation provided by practice sessions and tournaments, and the many friends I have found

Thanks

there. I want to particularly mention Wieland and Filippo who never fail to cheer me up and supported me and my moods during the past years.

I am forever grateful to my family, particularly my father Ralf and my brothers Lukas and Peter for always pushing me beyond my comfort zone, but also for unquestioningly being my safe haven when needed. Moreover, I thank my grandparents for sending special treats and keeping their fingers crossed during the final weeks of writing this dissertation. Special thanks also go to my great aunt Barb, who spontaneously stepped in with her English expertise.

I thank Johannes who has my back—and my heart—for everything.

CONTENTS

| | |
|--|-------------|
| Thanks | vii |
| Contents | xi |
| List of Figures | xiii |
| List of Tables | xiv |
| List of Algorithms | xv |
| Notation | xvi |
| | |
| Introduction | 1 |
| | |
| I Optimal Rank-1 Hankel Approximation | 9 |
| | |
| 1 Matrix Approximations | 11 |
| 1.1 Matrix Rank and Matrix Norms | 11 |
| 1.2 Low-Rank Approximation | 14 |
| 1.3 Hankel Structured Approximation | 16 |
| | |
| 2 Characterization of Hankel Matrices Depending on Their Rank | 21 |
| 2.1 Rank-1 Hankel Matrices | 21 |
| 2.2 Hankel Matrices of Higher Rank | 24 |

| | | |
|-----------|--|------------|
| 3 | Optimal Rank-1 Hankel Approximation in the Frobenius Norm | 37 |
| 3.1 | Complex Rank-1 Hankel Approximation | 38 |
| 3.2 | Real Rank-1 Hankel Approximation | 43 |
| 4 | Optimal Rank-1 Hankel Approximation in the Spectral Norm | 53 |
| 4.1 | Definiteness of Diagonal-Plus-Rank-1 Matrices | 56 |
| 4.2 | The Optimal Approximation Error | 60 |
| 4.2.1 | Isolated Largest Eigenvalue | 60 |
| 4.2.2 | Multiple Largest Eigenvalue | 70 |
| 4.3 | Computation of the Optimal Approximation | 77 |
| II | Benchmarking Structured Low-Rank Approximation Methods | 83 |
| 5 | General Affine Matrix Structure | 85 |
| 6 | Local Optimization | 89 |
| 6.1 | Kernel Representation of the Rank Constraint | 89 |
| 6.2 | Image Representation of the Rank Constraint | 95 |
| 7 | Alternating Projections | 103 |
| 7.1 | Basic Observations | 106 |
| 7.2 | Our New Convergence Result | 109 |
| 7.3 | Numerical Assessment of Convergence | 123 |
| 8 | Convex Relaxation | 125 |
| 9 | Numerical Examples and Comparisons | 135 |
| 9.1 | Revisiting Some Examples | 135 |
| 9.2 | More Comparisons | 139 |
| 9.2.1 | Approximation with Respect to the Frobenius Norm | 140 |
| 9.2.2 | Approximation with Respect to the Spectral Norm | 147 |
| | Conclusion and Outlook | 151 |
| | Bibliography | 155 |

LIST OF FIGURES

| | | |
|-----|---|-----|
| 3.1 | Optimal r1H errors for real and complex parameters | 51 |
| 4.1 | Optimal parameters \tilde{c} and \tilde{z} for Example 4.20 | 76 |
| 7.1 | Average number of iterations needed for convergence | 124 |
| 8.1 | A generic trade-off curve between approximation error and rank | 128 |
| 8.2 | Trade-off curve for Example 8.3 | 129 |
| 8.3 | Rank of the solution matrix for different regularization parameters in Example 8.3 | 130 |
| 8.4 | Rank of the solution matrix for a refined range of regularization parameters in Example 8.3 | 130 |
| 8.5 | Trade-off curve for Example 8.3 with refined range of regularization parameters | 130 |
| 8.6 | Trade-off curve for Example 8.5 with weighted norm | 132 |
| 8.7 | Trade-off curve for Example 8.6 | 133 |
| 9.1 | Relative approximation errors in the Frobenius norm for (4×4) matrices . | 141 |
| 9.2 | Relative approximation errors in the Frobenius norm for (10×10) matrices | 143 |
| 9.3 | Ranks of the approximation matrices | 144 |
| 9.4 | Development of mean relative errors for rectangular matrices | 144 |
| 9.5 | Relative approximation errors in the spectral norm for (4×4) matrices . . | 148 |
| 9.6 | Relative approximation errors in the spectral norm for (10×10) matrices | 148 |

LIST OF TABLES

| | | |
|-----|---|-----|
| 9.1 | Parameters for optimal rank-1 Hankel approximation of the matrix (9.1) w.r.t. the Frobenius and the spectral norm | 136 |
| 9.2 | Absolute and relative approximation errors from the Examples 3.8, 6.2, 6.8 and 7.15 | 137 |
| 9.3 | Parameters for optimal rank-1 Hankel approximation of the matrix (9.2) w.r.t. the Frobenius and the spectral norm | 138 |
| 9.4 | Absolute and relative approximation errors from Examples 6.1, 6.7 and 8.3 | 139 |
| 9.5 | Mean relative errors and deviations from the minimal error for (4×4) matrices | 142 |
| 9.6 | Mean relative errors in the Frobenius norm and deviations from the minimal error for (10×10) matrices | 142 |
| 9.7 | Numerical complexities of each method | 146 |
| 9.8 | Mean relative errors in the spectral norm and mean deviations from the minimal error | 147 |

LIST OF ALGORITHMS

| | | |
|-----|---|-----|
| 4.1 | Optimal rank-1 Hankel approximation w.r.t. the spectral norm | 77 |
| 4.2 | Optimal rank-1 Hankel approximation w.r.t. the spectral norm for isolated largest eigenvalue | 79 |
| 4.3 | Optimal rank-1 Hankel approximation w.r.t. the spectral norm for multiple largest eigenvalue | 81 |
| 6.1 | Structured low-rank approximation by factorization | 98 |
| 7.1 | Cadzow's algorithm for general structured low-rank approximation | 104 |
| 7.2 | Cadzow's algorithm for rank-1 Hankel approximation | 105 |

NOTATION

Numbers and Sets

| | |
|--------------------------------|---|
| \mathbb{N}, \mathbb{N}_0 | natural numbers and natural numbers explicitly including zero |
| $\mathbb{R}, \bar{\mathbb{R}}$ | real numbers and extended real numbers $\bar{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ |
| $\mathbb{C}, \bar{\mathbb{C}}$ | complex numbers and extended complex numbers $\bar{\mathbb{C}} := \mathbb{C} \cup \{\infty\}$ |
| M, N | matrix dimensions, always assuming $M, N \geq 2$ |

Matrices and Vectors

| | |
|--|---|
| $\mathbf{I} = \mathbf{I}_N$ | identity matrix of size $N \times N$ |
| $\mathbf{J} = \mathbf{J}_N$ | counter-identity matrix of size $N \times N$, see (1.10) |
| \mathbf{e}_k | k -th vector of the standard basis $\mathbf{e}_k = (\delta_{k-1,j})_{j=0}^{N-1}$, $k = 1, \dots, N$, with the Kronecker delta $\delta_{kj} = \begin{cases} 1, & j=k \\ 0, & j \neq k \end{cases}$ |
| \mathbf{A} | initial matrix of size $(M \times N)$ for the approximation problems |
| $\mathbf{A}^\top, \mathbf{A}^*, \mathbf{z}^\top, \mathbf{z}^*$ | transpose, and complex conjugate and transpose of a matrix or vector |
| $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ | singular value decomposition of the matrix \mathbf{A} , see Definition 1.1 |
| $\mathbf{H} = \mathbf{H}_{M,N}$ | Hankel matrix of size $M \times N$, see Definition 1.10 |
| $\hat{\mathbf{z}} = \hat{\mathbf{z}}_N(z)$ | structured vector $\hat{\mathbf{z}}_N(z) = \left(1 \quad z \quad z^2 \quad \dots \quad z^{N-1}\right)^\top \in \mathbb{C}^N$, see (2.1) |
| $\mathbf{z} = \mathbf{z}_N(z)$ | <i>normalized</i> structured vector $\mathbf{z} = \hat{\mathbf{z}}/\ \hat{\mathbf{z}}\ _2 \in \mathbb{C}^N$, see (2.5) |
| \mathbf{P} | block-diagonal Hankel projection matrix, see (1.6) |

Norms and Operators

| | |
|-----------------------------------|---|
| $\ \cdot\ _F$ | Frobenius norm for matrices, see Definition 1.5 |
| $\ \cdot\ _2$ | Euclidean vector norm or matrix spectral norm, see Definition 1.6 |
| $\ \cdot\ _\infty$ | elementwise maximum norm of a vector or matrix, see Definition 1.7 |
| \sum' | summation where terms of the form $\frac{0}{0}$ are omitted, see Chapter 4 |
| $\text{vec}(\mathbf{A})$ | vectorization of a matrix along its columns |
| $\text{diagvec}(\mathbf{A})$ | vectorization of a matrix along its counter-diagonals, see (1.8) |
| \mathcal{S} | general structure specification map, see Chapter 5 |
| $\mathcal{P}_{\mathcal{S}}$ | orthogonal projection onto the affine space of \mathcal{S} -structured matrices |
| $\mathcal{H} = \mathcal{H}_{M,N}$ | specification map for Hankel structure of size $M \times N$, see (5.1) |
| \mathcal{P} | orthogonal projection onto the linear space of Hankel matrices |

Abbreviations

| | |
|------------|--|
| LRA | low-rank approximation |
| SLRA | structured low-rank approximation |
| r1H | rank-1 Hankel approximation |
| RMP | rank minimization problem |
| AAK theory | theory of Adamjan, Arov, and Kreĭn |
| SVD | singular value decomposition |
| RE | relative error |
| MRE | mean relative error (9.3) |
| MAD | mean absolute deviation from optimum (9.4) |
| MSD | mean squared deviation from optimum (9.5) |

INTRODUCTION

The world is full of data and has been long before *big data* became a slogan, ages before computers were even invented. Yet without numerical processing, the data were used in the form of generation-spanning experience. Our ancestors knew when to seed and when to harvest their crop, just to name one example.

With the invention and further development of computers, scientists are able to acquire and store ever larger data sets. Examples are bio-medical or financial data, or data arising in social studies or engineering. Because of the immense amount and complexity of the data, mere experience does not suffice to gain an insight from them. Mathematical models are necessary to outline, analyze, and explain the data. Only then can they serve a purpose, such as drawing a conclusion or making a prediction.

It is common consensus—often entitled Occam’s razor—that among all available models analyzing a data set for the same purpose, the simplest one should be used.

“Everything should be made as simple as possible, but no simpler.”

—Albert Einstein.

There are different kinds of mathematical models; one popular example is the linear model. Linear models are described by linear operators which—in finite dimensions—can be represented by matrices. Often the complexity of a model is related to the rank of this matrix: a simple model corresponds to a matrix with low rank. Thus, fitting a simple model to given data amounts to low-rank approximation (LRA) of a matrix constructed from the data. Some even believe that

“behind every linear data modeling problem there is a (hidden) low-rank approximation problem”

—Ivan Markovsky [Mar08].

While simple unstructured LRA is equivalent to fitting a linear static model to the data, other models may impose further requirements on the matrix. For example, non-linear or dynamic models lead to *structured* low-rank approximation (SLRA) of the data matrix. That is, the data matrix is to be approximated by a low-rank matrix that additionally exhibits a certain structure.

Among all matrix structures, the Hankel structure is especially relevant because of its connection to dynamic linear time-invariant models. Such models—and therefore Hankel structured low-rank approximations—are widely used in areas like system theory, signal processing, computer algebra, or machine learning. Concrete applications include:

- errors-in-variables identification [Mar08; MWV⁺05],
- frequency estimation [AAK71; AC11a; AC19; PP19],
- approximation by finite-rate-of-innovation signals [MV05; VMB02], and
- finding approximate common divisors of polynomials [CFP03].

Hankel structured low-rank approximation also occurs in Prony’s method [PT14; Pot17], and its modifications [BM86; OS95; ZP19] which are used for the recovery of structured functions [Kel21] or sparse phase retrieval in one dimension [BP17].

General Problem Formulation

We state the structured low-rank approximation (SLRA) problem. Given an initial matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$ with $M, N \geq 2$, find

$$\begin{aligned} \min \|\mathbf{A} - \mathbf{H}\| \quad & \text{such that } \text{rank } \mathbf{H} \leq r, \\ & \text{and } \mathbf{H} \text{ has a certain structure,} \end{aligned} \tag{1}$$

where the norm is usually a (weighted) Frobenius norm, and sometimes the spectral norm. With $r < \min\{M, N\}$ we specify an upper bound for the desired rank of the approximating matrix \mathbf{H} . The SLRA problem may also be termed structured rank- r approximation problem when explicit information about the rank is desired. In this dissertation, we will always consider the SLRA problem with Hankel structure.

Within problem (1) there are two subproblems, namely unstructured rank- r approximation (LRA) and structured approximation without rank constraint. It is a well-known fact that the solution of the unstructured LRA problem can be given in terms of the singular value decomposition (SVD) of \mathbf{A} . However, the low-rank approximation is in general not structured; it usually does not even inherit any potential structure from \mathbf{A} . The structured

approximation problem without the rank constraint is convex and the solution can easily be obtained. In case of Hankel structure this is done by an averaging procedure. In contrast, the above minimization problem (1) with both constraints combined is not convex and highly non-trivial, especially for the spectral norm.

Related Solution Approaches

A very simple approach to find an approximate solution to problem (1) is Cadzow's algorithm [Cad88; Gil10], and its modifications presented in [AC13; CFP03]. Cadzow's algorithm is based on alternating projections between the subspace of structured matrices and the set of low-rank matrices, and can be applied for the Frobenius and the spectral norm. Alternating projection methods are widely popular because of their simplicity and broad applicability. Despite its common use, there are no results on convergence of Cadzow's algorithm known to the author that can reliably be applied to the general low-rank Hankel approximation setting (see also Chapter 7).

Unlike Cadzow's algorithm, the following methods engage in solving the SLRA problem exclusively for the (weighted) Frobenius norm and real matrices $\mathbf{A} \in \mathbb{R}^{M \times N}$.

One such approach relies on local optimization techniques. For the (weighted) Frobenius norm, problem (1) can be written as a non-linear eigenvalue problem, see [BM86; OS95; ZP19]. In addition, there are a wide variety of publications treating (1) as a non-linear structured least squares problem [IUM14; ZG20], or a structured total least squares problem, see [DeM94; DeM93; GZ11; LMV00; LV01; MVP05; MWV⁺05].

Another approach is to reformulate problem (1) as polynomial or rational function optimization. Details can be found in [OSS14; UM12].

A completely different approach to the SLRA problem is the concept of convex relaxation. The optimization problem (1) is not convex due to the non-convexity of the rank constraint. Therefore, a straightforward convex relaxation is achieved by replacing the rank by the nuclear norm [Faz02; FPS⁺13]. Going further, one can minimize the convex envelope of the Frobenius norm and the rank constraint [AC19; ACO17; GRG18; GG18; LO16] instead of solving problem (1). Additionally connected to convex relaxations, there are subspace based and hybrid methods [LV10; VD96].

Yet another attempt to solve problem (1) involves randomized alternating projections and backtracking. Such an algorithm is proposed by [GZ15; GZ13]. In [GZ15], the authors even claim guaranteed convergence to the optimal solution. Unfortunately, the corresponding

software could not be located on the web for comparison.

As can be perceived by the above summary, there are a variety of methods designed for the Frobenius norm. Regarding the SLRA problem in the spectral norm in contrast, the literature is scarce. It was formulated in [Ant98] but almost no related work is to be found. Two exceptions are given by [Ant97; Rum03a] for the following special problems.

For regular real ($N \times N$) matrices, the minimal spectral norm distance to a singular (i.e., rank-deficient) structured matrix is studied in [Rum03a]. (The same topic is considered for the Frobenius norm in [Rum03b].) A result in [Rum03a] can indeed be exploited to construct a rank- $(N - 1)$ Hankel approximation for a given full-rank Hankel matrix which is optimal with regard to the spectral norm. Unfortunately, this approach cannot be extended to construct Hankel approximations of lower (than $N - 1$) rank.

In [Ant97] rank-1 Hankel approximation in the spectral norm is considered for real initial matrices with Hankel structure. More precisely, the existence of a rank-1 Hankel approximation that achieves the same error as the unstructured low-rank approximation is investigated.

The theory of Adamjan, Arov, and Kreĭn (AAK theory) should also be mentioned in the context of SLRA. It is not exactly concerned with problem (1) but with a similar *infinite* problem. The AAK theory deals with low-rank Hankel approximation with respect to the ℓ_2 -operator norm, where the initial matrix \mathbf{A} is an infinite Hankel matrix whose entries obey a certain decay property. The main theorem states that the infinite low-rank Hankel approximation always attains the same error as the infinite unstructured low-rank approximation. Optimal infinite low-rank Hankel approximations can be computed numerically [BM05; PP16] and have been used to compute adaptive Fourier series with exponential decay for large classes of functions [PP19]. Unfortunately, the AAK theory cannot be transferred to finite matrices, see [BM05].

Contributions of this Work

In this dissertation, we consider a special case of SLRA, namely the rank-1 Hankel approximation (r1H) problem

$$\begin{aligned} \min \|\mathbf{A} - \mathbf{H}\| \quad & \text{such that } \text{rank } \mathbf{H} = 1, \\ & \text{and } \mathbf{H} \text{ has Hankel structure} \end{aligned} \tag{2}$$

for a given initial matrix \mathbf{A} of size $M \times N$.

The vast majority of related approaches only achieves *approximate* solutions to the SLRA problem with respect to the Frobenius norm. We are interested in provably *optimal* solutions of the r1H problem (2) for both the Frobenius and the spectral norm. Optimal solution is meant in the sense that there exists no better solution, that is, with smaller error and fulfilling both constraints. Note that all our results can easily be transferred to the Toeplitz structure as outlined in Chapter 1.

In the formulation of (2) we require the rank of the approximating matrix \mathbf{H} to be equal to one instead of the smaller or equal we used to describe the general SLRA problem (1). This is because the only matrix with rank smaller than one is the trivial zero matrix. Therefore, we exclude it from our considerations.

Our key idea to solve the r1H problem (2) is analytically reformulating it. This reformulation enables us not only to prove the characterization of an optimal solution, but also allows us to develop a numerical algorithm for its computation.

For the Frobenius norm, our results hold for any initial matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$. Then the optimal solution itself is a complex Hankel matrix $\mathbf{H} \in \mathbb{C}^{M \times N}$ of rank one. In this case, we characterize the optimal solution by a maximization problem for a rational function.

With respect to the spectral norm, problem (2) is much more delicate. Paying tribute to this fact, we restrict our research to real symmetric initial matrices $\mathbf{A} \in \mathbb{R}^{N \times N}$. Also in this case, we characterize the optimal rank-1 Hankel approximation by a maximization problem for a rational function. This rational function is however of a completely different structure than the one for the Frobenius norm.

Thus, it is not surprising that the optimal rank-1 Hankel approximations for the Frobenius norm and the spectral norm usually differ. They only coincide in the trivial case, when the generically unstructured rank-1 approximation happens to have Hankel structure.

Furthermore, we not only identify optimal solutions to the r1H problem (2), we also assess them in comparison to the unstructured rank-1 approximation. More precisely, we give necessary and sufficient conditions identifying initial matrices \mathbf{A} for which the rank-1 Hankel approximation error is as small as the unstructured rank-1 approximation error. In the case of the spectral norm, we thereby extend the work in [Ant97].

In contrast to the methods described as related approaches, our characterizations provide guaranteed *optimal* solutions and we can compute them numerically. Our optimal solutions can therefore serve as benchmarks for different methods engaging in the r1H problem.

While benchmarking different methods for rank-1 Hankel approximation, we also deal with Cadzow's algorithm. There are only partial convergence results for Cadzow's algorithm

in the SLRA setting [ZG17]. For the special case of rank-1 Hankel approximation, we give a complete proof for the convergence of Cadzow’s algorithm. The resulting limit point however does not usually coincide with the optimal solution of the r1H problem (2)—neither for the Frobenius norm nor for the spectral norm.

Our main publication [KPP21a] is a consequence of the results summarized above. Additional examples complementing [KPP21a] are published in [KPP21b]. In this dissertation, we also present further results extending [KPP21a], of which some are published in [Kni21]. These manuscripts were issued prior to the composition of this thesis. Thus naturally, some sections of this thesis will be largely very similar to parts of them. This will always be explicitly stated at the beginning of the concerned sections.

Organization of this Dissertation

This thesis is split into two main parts. In Part I, we develop analytical characterizations and numerical algorithms for computation of optimal rank-1 Hankel approximations in the Frobenius and the spectral norm. Then in Part II, we use these optimal solutions as benchmarks for the comparison of different SLRA methods.

We begin Part I with some basic definitions and concepts presented in Chapter 1. Chapter 2 is devoted to the characterization of Hankel matrices depending on their rank. Using the characterization of rank-1 Hankel matrices, we analytically reformulate problem (2) with respect to the Frobenius and the spectral norm in Chapters 3 and 4, respectively. Based on these reformulations, we develop algorithms to compute the optimal solutions to problem (2) numerically.

Part II is dedicated to the benchmarking of different methods from the literature against our optimal solutions from Part I. In order to understand these methods, we give a characterization of general structured matrices in Chapter 5. The methods themselves are presented in Chapters 6 to 8. They are based on local optimization (Chapter 6), alternating projections (Chapter 7), and convex relaxation (Chapter 8). In Chapter 7 we also give a new proof of convergence for the special r1H setting. The benchmarking is done by means of small examples and on a broader basis in Chapter 9.

Finally, we draw conclusions and give an outlook on possible topics for future research.



OPTIMAL RANK-1 HANKEL APPROXIMATION

The structured low-rank approximation (SLRA) problem is generally accepted to be an important and interesting one. Nevertheless, the methods that are usually used to solve it only do so approximately (in the case of local optimization techniques or convex relaxations), or do not have guaranteed convergence, let alone to the optimal solution (in the case of Cadzow's algorithm). For the special case of the rank-1 Hankel approximation (r1H) problem,

$$\min \|\mathbf{A} - \mathbf{H}\| \quad \text{such that } \text{rank } \mathbf{H} = 1, \\ \text{and } \mathbf{H} \text{ has Hankel structure,}$$

we are able to characterize and exactly compute *optimal* solutions both with respect to the Frobenius norm and the spectral norm.

For a deeper understanding of the problem, we give some basic definitions and notions in Chapter 1. In Sections 1.2 and 1.3, we also summarize the long solved individual problems of unstructured low-rank approximation and Hankel structured approximation without rank constraint.

Our approach to the r1H problem crucially depends on the special structure of Hankel matrices of low rank that we explain in Chapter 2. Thus, more important is Section 2.1, where we characterize Hankel matrices of rank one. Nevertheless, with Section 2.2, we give an excursus to Hankel matrices of higher rank.

In Chapters 3 and 4, we solve the r1H problem for the Frobenius norm and the spectral norm, respectively. In Chapter 3, we develop a reformulation of the r1H problem as rational function maximization. This formulation enables us to identify and compute the optimal solution. First, we deal with very general initial matrices $\mathbf{A} \in \mathbb{C}^{M \times N}$ in Section 3.1. For such, the optimal solution will be a complex Hankel matrix $\mathbf{H} \in \mathbb{C}^{M \times N}$ of rank one. In Section 3.2, some special results are added for real initial matrices $\mathbf{A} \in \mathbb{R}^{M \times N}$ and their real rank-1 Hankel approximations $\mathbf{H} \in \mathbb{R}^{M \times N}$.

Due to the more complicated nature of the spectral norm, in Chapter 4, we restrict ourselves to real symmetric initial matrices \mathbf{A} and real rank-1 Hankel approximations. The key to the optimal solution of the r1H problem with respect to the spectral norm is the interdependence of optimal approximation and optimal approximation error. Therefore, we examine the optimal error more closely in Section 4.2. One must differentiate if the by modulus largest eigenvalue of the initial matrix \mathbf{A} is isolated or occurs with higher multiplicity. In both cases, we can again characterize the optimal rank-1 Hankel approximation by a maximization problem for a rational function. This rational function depends on the optimal approximation error as opposed to the rational function found for the Frobenius norm. In Section 4.3, we translate our theoretical results to executable algorithms for computation of the optimal rank-1 Hankel approximation in the spectral norm.

When comparing the characterizations of the optimal rank-1 Hankel approximation for the Frobenius and the spectral norm, we observe that they are different. In fact, the optimal solutions to the r1H problem are usually not the same for the Frobenius and the spectral norm. They only coincide in the trivial case, when the generically unstructured rank-1 approximation does, by chance, have Hankel structure.

For both the Frobenius norm and the spectral norm, we also give necessary and sufficient conditions to ensure that the optimal rank-1 Hankel approximation achieves the same error as the unstructured rank-1 approximation. For the spectral norm, we thereby extend the result from [Ant97].

Moreover, we illustrate our results by small but insightful examples.

1

MATRIX APPROXIMATIONS

In this chapter we revisit some basics that are essential throughout this thesis.

In Section 1.1, we introduce the singular value decomposition (SVD) and the rank of a matrix. The latter is needed to understand the concept of low-rank approximation. Furthermore, we give formal definitions of different matrix norms, which will be used to measure the approximation error.

In Sections 1.2 and 1.3, we explicitly introduce the subproblems related to low-rank Hankel approximation, namely the unstructured low-rank approximation (LRA) problem and the Hankel structured approximation problem without rank constraint.

First, let us fix the notation $\mathbf{a}^* := \bar{\mathbf{a}}^\top$ for the conjugate transpose of a vector or a matrix.

1.1 Matrix Rank and Matrix Norms

Any matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$ can be decomposed into one diagonal matrix $\mathbf{\Sigma}$ and two unitary matrices \mathbf{U} and \mathbf{V} .

Definition 1.1 (Singular value decomposition) Let $\mathbf{A} \in \mathbb{C}^{M \times N}$ and assume $M \leq N$. We define an economic version of the singular value decomposition (SVD) as the factorization

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*,$$

where $\mathbf{U} \in \mathbb{C}^{M \times M}$ and $\mathbf{V} \in \mathbb{C}^{N \times M}$ have orthonormal columns, that is, $\mathbf{U}^*\mathbf{U} = \mathbf{V}^*\mathbf{V} = \mathbf{I}_M$. The columns \mathbf{u}_j of \mathbf{U} and \mathbf{v}_j of \mathbf{V} , $j = 0, \dots, M - 1$, are also called left and right

singular vectors of \mathbf{A} , respectively. The matrix $\mathbf{\Sigma} = \text{diag}(\sigma_0, \dots, \sigma_{M-1}) \in \mathbb{R}^{M \times M}$ is a diagonal matrix containing the singular values of \mathbf{A} . Note that the singular values are uniquely determined by the matrix \mathbf{A} , and we may assume them to be ordered $\sigma_0 \geq \sigma_1 \geq \dots \geq \sigma_{M-1} \geq 0$ largest to smallest. The tuple $\{\sigma_j, \mathbf{u}_j, \mathbf{v}_j\}$ is also called j -th singular triple of \mathbf{A} .

Remark 1.2 1. In the usual definition of the SVD, the matrix \mathbf{V} is an $(N \times N)$ unitary matrix, and the diagonal matrix $\mathbf{\Sigma}$ is padded with zeros to an $(M \times N)$ matrix.

2. For a real matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, \mathbf{U} and \mathbf{V} can be guaranteed to be real orthogonal matrices.

3. The SVD of \mathbf{A} is connected to the eigendecompositions of $\mathbf{A}^* \mathbf{A}$ and $\mathbf{A} \mathbf{A}^*$ by

$$\mathbf{A}^* \cdot \mathbf{A} = \mathbf{V} \mathbf{\Sigma}^* \mathbf{U}^* \cdot \mathbf{U} \mathbf{\Sigma} \mathbf{V}^* = \mathbf{V} \cdot (\mathbf{\Sigma}^* \mathbf{\Sigma}) \cdot \mathbf{V}^*$$

and

$$\mathbf{A} \cdot \mathbf{A}^* = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^* \cdot \mathbf{V} \mathbf{\Sigma}^* \mathbf{U}^* = \mathbf{U} \cdot (\mathbf{\Sigma} \mathbf{\Sigma}^*) \cdot \mathbf{U}^*,$$

respectively. Thus, the non-zero singular values of \mathbf{A} are the square roots of the non-zero eigenvalues of $\mathbf{A}^* \mathbf{A}$ or $\mathbf{A} \mathbf{A}^*$. In particular, if $\mathbf{A} \in \mathbb{R}^{N \times N}$ is a real symmetric matrix with eigenvalues $\lambda_0, \dots, \lambda_{N-1}$, then we have the relation $|\lambda_j| = \sigma_j, j = 0, \dots, N - 1$, between eigenvalues and singular values.

One fundamental characteristic of a matrix is its rank, which is defined as follows. Besides Definition 1.3, it can also be expressed via the SVD, see Lemma 1.4 (5).

Definition 1.3 (Rank of a matrix) The column rank of a matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$ is the dimension of the column space of \mathbf{A} , which is the same as the number of linearly independent column vectors in \mathbf{A} . Analogously, the row rank of \mathbf{A} is the dimension of the row space of \mathbf{A} , that is, the number of linearly independent row vectors in \mathbf{A} .

It is a well-known elementary result in linear algebra that row rank and column rank are always equal. This number is simply called the rank of \mathbf{A} .

We list some useful essential characterizations and properties of the rank in the following lemma. All of them can be found in [HJ13, Chap. 0].

Lemma 1.4 Let $\mathbf{A} \in \mathbb{C}^{M \times N}$ be an arbitrary matrix.

- (1) The rank of \mathbf{A} is always smaller than or equal to the smaller one of the matrix dimensions:
 $\text{rank } \mathbf{A} \leq \min\{M, N\}$.

- (2) The rank is subadditive: Let $\mathbf{B} \in \mathbb{C}^{M \times N}$ be a second matrix of the same size as \mathbf{A} . Then we have $\text{rank}(\mathbf{A} + \mathbf{B}) \leq \text{rank} \mathbf{A} + \text{rank} \mathbf{B}$.
- (3) The rank of \mathbf{A} plus the dimension of its kernel equals the number of columns of \mathbf{A} . Therefore, we can express $\text{rank} \mathbf{A} = N - \dim(\ker \mathbf{A})$. This fact is known as the rank-nullity theorem.
- (4) The rank of \mathbf{A} is the smallest number r such that \mathbf{A} can be written as the product $\mathbf{A} = \mathbf{P}\mathbf{L}$ of two matrices $\mathbf{P} \in \mathbb{C}^{M \times r}$ and $\mathbf{L} \in \mathbb{C}^{r \times N}$.
- (5) Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be the singular value decomposition (SVD) of \mathbf{A} . Then the rank of \mathbf{A} is equal to the number of its non-zero singular values.

Next, we give the formal definitions of the Frobenius norm and the spectral norm, which are used to measure the approximation error in our matrix approximations. Moreover, we provide alternative representations of these norms, which will be of use later on.

Definition 1.5 (Frobenius norm) The Frobenius norm—named after the German mathematician Ferdinand Georg Frobenius (1849-1917)—is a matrix norm related to the Euclidean vector norm. For a matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$, the Frobenius norm is given by

$$\|\mathbf{A}\|_F := \left(\sum_{j=0}^{M-1} \sum_{k=0}^{N-1} |a_{jk}|^2 \right)^{1/2} = \|\text{vec } \mathbf{A}\|_2,$$

where $\text{vec } \mathbf{A} \in \mathbb{C}^{MN}$ is a vectorization of \mathbf{A} . Different representations of the Frobenius norm involve the singular values of \mathbf{A} or the diagonal elements of $\mathbf{A}^* \mathbf{A}$:

$$\|\mathbf{A}\|_F = \left(\sum_{j=0}^{r-1} \sigma_j^2 \right)^{1/2} = \text{tr}(\mathbf{A}^* \mathbf{A})^{1/2},$$

where $r = \text{rank}(\mathbf{A})$ and $\text{tr}(\mathbf{A}^* \mathbf{A}) = \sum_{j=0}^{N-1} (\mathbf{A}^* \mathbf{A})_{jj}$ is the trace (sum of diagonal entries) of the positive semidefinite Hermitian matrix $\mathbf{A}^* \mathbf{A}$.

Definition 1.6 (Spectral norm) The spectral norm is the operator norm induced by the Euclidean vector norm. For a matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$, the spectral norm is defined as

$$\|\mathbf{A}\|_2 := \max_{\substack{\mathbf{x} \in \mathbb{C}^N \\ \mathbf{x} \neq \mathbf{0}}} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \max_{\substack{\mathbf{x} \in \mathbb{C}^N \\ \|\mathbf{x}\|_2=1}} \|\mathbf{A}\mathbf{x}\|_2.$$

The name of spectral norm originates from the fact that it is the same as the spectral radius (largest eigenvalue) of the positive semidefinite Hermitian matrix $\mathbf{A}^* \mathbf{A}$,

$$\|\mathbf{A}\|_2^2 = \lambda_0(\mathbf{A}^* \mathbf{A}) = \sigma_0^2(\mathbf{A}).$$

The spectral radius $\lambda_0(\mathbf{A}^* \mathbf{A})$, in turn, is the squared largest singular value of \mathbf{A} , see Remark 1.2. In the case where \mathbf{A} admits an eigendecomposition, the spectral norm is equal to the largest absolute value of its eigenvalues, $\|\mathbf{A}\|_2 = |\lambda_0(\mathbf{A})|$.

There is the well-known inequality $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F$ between Frobenius and spectral norm. This inequality is clear by inspecting the respective second representations of the two norms. Furthermore, both the Frobenius and the spectral norm are unitarily invariant. This means that $\|\mathbf{A}\| = \|\mathbf{U} \cdot \mathbf{A} \cdot \mathbf{V}\|$ for all matrices \mathbf{A} and for all unitary matrices \mathbf{U} and \mathbf{V} with $\mathbf{U}\mathbf{U}^* = \mathbf{I}_M$ and $\mathbf{V}\mathbf{V}^* = \mathbf{I}_N$, and can easily be seen from the representations of the norms using the singular values of \mathbf{A} .

The following norm is not used to measure the approximation error in this thesis. Nevertheless, it plays an important role in the crucial lemma of Chapter 7.

Definition 1.7 (Elementwise maximum norm) For a matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$ with entries $\mathbf{A} = (a_{jk})_{j,k=0}^{M-1, N-1}$ we define the elementwise maximum norm as

$$\|\mathbf{A}\|_\infty := \max_{\substack{0 \leq j \leq M-1 \\ 0 \leq k \leq N-1}} |a_{jk}|.$$

It is identical to the vector maximum norm of the vectorized matrix, $\|\mathbf{A}\|_\infty = \|\text{vec } \mathbf{A}\|_\infty$.

1.2 Low-Rank Approximation

Given $\mathbf{A} \in \mathbb{C}^{M \times N}$, we state the low-rank approximation (LRA) problem

$$\min \|\mathbf{A} - \mathbf{B}\| \quad \text{such that } \text{rank } \mathbf{B} \leq r, \quad (1.1)$$

where the norm can be either the Frobenius norm or the spectral norm.

This problem can be solved by truncating the singular value decomposition of \mathbf{A} in the manner of the following theorem.

Theorem 1.8 (Eckart-Young-Mirsky [EY36]) *Let $\mathbf{A} \in \mathbb{C}^{M \times N}$ and let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be its SVD. Recall that the singular values are ordered $\sigma_0 \geq \sigma_1 \geq \dots \geq \sigma_{M-1} \geq 0$ largest to smallest.*

Then the best rank- r approximation of \mathbf{A} with respect to the Frobenius norm and the spectral norm is given by

$$\mathbf{A}_r := \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^* := \sum_{j=0}^{r-1} \sigma_j \cdot \mathbf{u}_j \mathbf{v}_j^*,$$

where \mathbf{u}_j and \mathbf{v}_j are the j -th columns of \mathbf{U} and \mathbf{V} , respectively. The truncated matrices $\mathbf{U}_r := (\mathbf{u}_0 \dots \mathbf{u}_{r-1})$ and $\mathbf{V}_r := (\mathbf{v}_0 \dots \mathbf{v}_{r-1})$ consist of the first r left and right singular vectors, respectively. Finally, $\mathbf{\Sigma}_r := \text{diag}(\sigma_0, \dots, \sigma_{r-1})$ is the diagonal matrix containing the r largest singular values of \mathbf{A} .

The resulting minimal approximation errors are given by

$$\|\mathbf{A} - \mathbf{A}_r\|_F^2 = \left\| \sum_{j=r}^{M-1} \sigma_j \cdot \mathbf{u}_j \mathbf{v}_j^* \right\|_F^2 = \sum_{j=r}^{M-1} \sigma_j^2$$

in the Frobenius norm, and by

$$\|\mathbf{A} - \mathbf{A}_r\|_2^2 = \left\| \sum_{j=r}^{M-1} \sigma_j \cdot \mathbf{u}_j \mathbf{v}_j^* \right\|_2^2 = \sigma_r^2$$

in the spectral norm.

Remark 1.9 1. Note that for the Frobenius norm, the rank- r approximation \mathbf{A}_r obtained by truncating the SVD is unique if and only if $\sigma_{r-1} > \sigma_r$. For the spectral norm, the best rank- r approximation usually is not unique even if $\sigma_{r-1} > \sigma_r$, see [Mar19, Thm. 4.5 and Rem. 4.7].

2. The Eckart-Young-Mirsky theorem does not only hold for the Frobenius and the spectral norm. Rather it holds for any unitarily invariant matrix norm, see [VWD05; Mar19, Rem. 4.6].

Specifically, for the rank-1 approximation problem

$$\min \|\mathbf{A} - \mathbf{B}\| \quad \text{such that } \text{rank } \mathbf{B} = 1,$$

the Eckart-Young-Mirsky Theorem provides the solution $\mathbf{B} = \mathbf{A}_1 = \sigma_0 \cdot \mathbf{u}_0 \mathbf{v}_0^*$.

The resulting minimal approximation errors are

$$\min_{\text{rank } \mathbf{B}=1} \|\mathbf{A} - \mathbf{B}\|_F^2 = \|\mathbf{A} - \sigma_0 \cdot \mathbf{u}_0 \mathbf{v}_0^*\|_F^2 = \sum_{j=1}^{M-1} \sigma_j^2 \quad (1.2)$$

for the Frobenius norm, and

$$\min_{\text{rank } \mathbf{B}=1} \|\mathbf{A} - \mathbf{B}\|_2^2 = \|\mathbf{A} - \sigma_0 \cdot \mathbf{u}_0 \mathbf{v}_0^*\|_2^2 = \sigma_1^2 \quad (1.3)$$

for the spectral norm.

Clearly, these optimal rank-1 approximation errors cannot be undercut by the solution of a *structured* rank-1 approximation problem. Therefore we will compare our solutions of the r1H problem (2) to these errors of the unstructured rank-1 approximation problem. We will refer to (1.2) and (1.3) as optimal error bounds.

1.3 Hankel Structured Approximation

Hankel structured matrices are especially important for mathematical modeling because of their connection to linear time-invariant systems. In this section we rigorously define Hankel matrices. Furthermore, we state the Hankel structured approximation problem and explain how to solve it.

Definition 1.10 (Hankel matrix) An $(M \times N)$ matrix \mathbf{H} is called a Hankel matrix if it is constant along each of its counter-diagonals. Phrased differently, the entries of \mathbf{H} only depend on the sum of their indices. A generic Hankel matrix is given by

$$\mathbf{H} = (h_{j,k})_{j,k=0}^{M-1,N-1} = (h_{j+k})_{j,k=0}^{M-1,N-1} = \begin{pmatrix} h_0 & h_1 & h_2 & \cdots & h_{M-1} & h_M & \cdots & h_{N-1} \\ h_1 & h_2 & & & & & & h_N \\ h_2 & & & & & & & \vdots \\ \vdots & & & & & & & \vdots \\ h_{M-1} & h_M & \cdots & h_{N-1} & h_N & \cdots & h_{M+N-2} & \vdots \end{pmatrix}, \quad (1.4)$$

where for the second line, we have assumed that $M \leq N$. This type of structured matrices is named after the German mathematician Hermann Hankel (1839-1873).

For a matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$, we have the Hankel structured approximation problem

$$\min \|\mathbf{A} - \mathbf{H}\| \quad \text{such that } \mathbf{H} \text{ has Hankel structure,} \quad (1.5)$$

where we consider either the Frobenius norm or the spectral norm.

The set of Hankel matrices is a linear subspace of $\mathbb{C}^{M \times N}$. Thus, there exists an orthogonal projection \mathcal{P} onto this subspace with respect to the Frobenius inner product, which is the standard inner product of vectorized matrices. This orthogonal projection onto the subspace of Hankel matrices can be stated explicitly as follows, see also [Buc94; GNZ01]. For a general matrix $\mathbf{A} = (a_{jk})_{j,k=0}^{M-1, N-1} \in \mathbb{C}^{M \times N}$, this projection is obtained by averaging the matrix elements along its counter-diagonals. Assuming $M \leq N$, we have

$$\mathcal{P}(\mathbf{A}) := (h_{j+k})_{j,k=0}^{M-1, N-1} \in \mathbb{C}^{M \times N} \quad (1.6)$$

with

$$h_k := \begin{cases} \frac{1}{k+1} \cdot \sum_{j=0}^k a_{j, k-j} & \text{for } k = 0, \dots, M-1, \\ \frac{1}{M} \cdot \sum_{j=0}^{M-1} a_{j, k-j} & \text{for } k = M, \dots, N-1, \\ \frac{1}{M+N-1-k} \cdot \sum_{j=k+1-N}^{M-1} a_{j, k-j} & \text{for } k = N, \dots, M+N-2. \end{cases} \quad (1.7)$$

For $M > N$ we simply take the transpose twice; namely, $\mathcal{P}(\mathbf{A}) = \mathcal{P}(\mathbf{A}^\top)^\top$.

The Hankel approximation problem (1.5) in the Frobenius norm is solved by this projection. More precisely, we have

$$\min_{\mathbf{H} \text{ Hankel}} \|\mathbf{A} - \mathbf{H}\|_F = \|\mathbf{A} - \mathcal{P}(\mathbf{A})\|_F$$

This solution to problem (1.5) is unique since, as a linear subspace, the set of rank one Hankel matrices is in fact convex.

In Lemma 7.2, we will see that \mathcal{P} is a projection also with respect to the spectral norm, albeit without the notion of orthogonality.

The Hankel projection $\mathcal{P}(\mathbf{A})$ in (1.6) and (1.7) can also be written as a linear mapping of the vector associated to the matrix \mathbf{A} . We define the vectorization of a matrix along its

Assuming $M \leq N$ we can write explicitly

$$\mathbf{T} = \begin{pmatrix} t_0 & t_{-1} & \cdots & t_{M-N} & \cdots & t_{-N+1} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ t_1 & \vdots & \ddots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ t_{M-1} & \cdots & t_1 & t_0 & t_{-1} & \cdots & t_{M-N} \end{pmatrix}.$$

Toeplitz matrices are named after the German-jewish mathematician Otto Toeplitz (1881-1940).

The close relation between Hankel and Toeplitz matrices is manifested via the counter-identity matrix

$$\mathbf{J}_N := \begin{pmatrix} 0 & \cdots & 0 & 1 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{N \times N}. \quad (1.10)$$

Any Toeplitz matrix $\mathbf{T} \in \mathbb{C}^{M \times N}$ can be represented by a Hankel matrix \mathbf{H} of the same size as

$$\mathbf{T} = \mathbf{H} \cdot \mathbf{J}_N.$$

For the individual entries we have $t_j = h_{j+N-1}$, $j = -N + 1, \dots, M - 1$, when indexed as in Definitions 1.10 and 1.11.

With this relation we can write the Toeplitz structured approximation of \mathbf{A} as a Hankel structured approximation of $\mathbf{A} \cdot \mathbf{J}_N$:

$$\begin{aligned} & \min \|\mathbf{A} - \mathbf{T}\|^2 && \text{such that } \mathbf{T} \text{ has Toeplitz structure} \\ & = \min \|\mathbf{A} - \mathbf{H} \cdot \mathbf{J}_N\|^2 && \text{such that } \mathbf{H} \text{ has Hankel structure} \\ & = \min \|\mathbf{A} \cdot \mathbf{J}_N - \mathbf{H}\|^2 && \text{such that } \mathbf{H} \text{ has Hankel structure,} \end{aligned}$$

where the norm can either be the Frobenius or the spectral norm. The last equality holds since both the Frobenius and the spectral norm are invariant under unitary transformations.

The rank of a matrix (such as \mathbf{T} or \mathbf{H}) is not touched when multiplied with an invertible matrix (such as \mathbf{J}_N). Thus, the above equalities also hold for the Toeplitz and Hankel SLRA problems (1). Therefore, while we will always consider Hankel structured approximations, all our results can easily be transferred to the Toeplitz structure.

2

CHARACTERIZATION OF HANKEL MATRICES DEPENDING ON THEIR RANK

2.1 Rank-1 Hankel Matrices

Before considering rank- r Hankel matrices very generally in Section 2.2, we first study the simplest case of rank-1 Hankel matrices individually and in a very detailed manner. With these considerations, we lay the foundation for our main results, which follow in Chapters 3 and 4.

For any number $z \in \mathbb{C}$, we define the structured vector

$$\hat{\mathbf{z}}_N(z) := (z^k)_{k=0}^{N-1} = \left(1 \quad z \quad z^2 \quad \dots \quad z^{N-1}\right)^T \in \mathbb{C}^N. \quad (2.1)$$

When there is no risk of confusion about the specific parameter z or the dimension N , we will omit these specifications and write $\hat{\mathbf{z}}_N$ or $\hat{\mathbf{z}}$ instead of $\hat{\mathbf{z}}_N(z)$.

With the definition of the structured vector (2.1), we can now characterize Hankel matrices of rank one.

Lemma 2.1 *A complex (possibly rectangular) rank-1 matrix $\mathbf{H} \in \mathbb{C}^{M \times N}$ has Hankel structure if and only if it is either of the form*

$$\mathbf{H} = c \cdot \hat{\mathbf{z}}_M \hat{\mathbf{z}}_N^T = c \cdot (z^{j+k})_{j,k=0}^{M-1, N-1} \quad (2.2)$$

or

$$\mathbf{H} = c \cdot \mathbf{e}_M \mathbf{e}_N^\top = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & c \end{pmatrix}, \quad (2.3)$$

where $c \in \mathbb{C} \setminus \{0\}$, $z \in \mathbb{C}$, and \mathbf{e}_N is the N -th vector of the standard basis in \mathbb{C}^N .

The idea of a parametric representation of a Hankel matrix of rank one has already been studied. It has appeared for example in [Ant97], where however the sparse rank-1 Hankel matrix (2.3) was omitted. Nevertheless, we give a proof for the sake of completeness.

Proof. Obviously, the two matrices $\mathbf{H} = c \cdot \hat{\mathbf{z}}_M \hat{\mathbf{z}}_N^\top$ and $\mathbf{H} = c \cdot \mathbf{e}_M \mathbf{e}_N^\top$ are rank-1 matrices with Hankel structure.

To prove the converse, recall that by Lemma 1.4 (4), any rank-1 matrix of size $(M \times N)$ can be represented as the outer product $\mathbf{a}\mathbf{b}^\top$ of two vectors $\mathbf{a} = (a_0 \cdots a_{M-1})^\top \in \mathbb{C}^M$ and $\mathbf{b} = (b_0 \cdots b_{N-1})^\top \in \mathbb{C}^N$. Then, the Hankel structure imposed on a rank-1 matrix implies the conditions

$$a_j \cdot b_k = a_m \cdot b_n, \quad \text{for } j + k = m + n, \quad (2.4)$$

where $j, m = 0, \dots, M-1$ and $k, n = 0, \dots, N-1$.

Assuming that $h_0 = a_0 b_0 \neq 0$, we can define $z := a_1/a_0$. It follows from (2.4) with $j + k = 1$ (i.e., from $a_0 \cdot b_1 = a_1 \cdot b_0$), that also $b_1/b_0 = z$. Thus we have $a_1 = z \cdot a_0$ and $b_1 = z \cdot b_0$.

We will show

$$a_j = z^j \cdot a_0, \quad \text{for } j = 1, \dots, M-1$$

and

$$b_j = z^j \cdot b_0, \quad \text{for } j = 1, \dots, N-1$$

by induction. For $0 < j < M-1$, we obtain from (2.4) that

$$a_{j+1} \cdot b_0 = a_j \cdot b_1 = (z^j \cdot a_0) \cdot (z \cdot b_0) = z^{j+1} \cdot a_0 \cdot b_0.$$

Since $b_0 \neq 0$, this implies $a_j = z^j \cdot a_0$ for $j = 1, \dots, M-1$. Analogously, for $0 < j < N-1$, we have

$$a_0 \cdot b_{j+1} = a_1 \cdot b_j = (z \cdot a_0) \cdot (z^j \cdot b_0) = z^{j+1} \cdot a_0 \cdot b_0,$$

which, since $a_0 \neq 0$, implies $b_j = z^j \cdot b_0$ for $j = 1, \dots, N-1$. Thus, \mathbf{H} has the desired structure (2.2) with $z = a_1/a_0$ and $c = a_0 \cdot b_0$.

If now $h_0 = a_0 \cdot b_0 = 0$, then either $a_0 = 0$ or $b_0 = 0$. Consequently, either the complete first row or the complete first column of $\mathbf{H} = \mathbf{a}\mathbf{b}^\top$ contains only zeros. Obeying the Hankel structure we inductively obtain that all the entries of $\mathbf{a}\mathbf{b}^\top$ are zero except for the last one. That last entry then has to be non-zero, $c := a_{M-1} \cdot b_{N-1} \neq 0$, since otherwise \mathbf{H} would be the zero matrix with rank zero, and thereby violate the rank-1 condition. Hence, in this case, we have \mathbf{H} as in (2.3). \square

Remark 2.2 Similarly to (2.1), we define the reversed structured vector

$$\hat{\mathbf{w}}_N(z) := \left(z^{N-1-k} \right)_{k=0}^{N-1} = \left(z^{N-1} \quad z^{N-2} \quad \dots \quad z \quad 1 \right)^\top \in \mathbb{C}^N.$$

Then, analogously to Lemma 2.1, we can show that a rank-1 Hankel matrix $\mathbf{H} \in \mathbb{C}^{M \times N}$ is of the form

$$\mathbf{H} = c \cdot \hat{\mathbf{w}}_M \hat{\mathbf{w}}_N^\top \quad \text{or} \quad \mathbf{H} = c \cdot \mathbf{e}_1 \mathbf{e}_1^\top.$$

Note that, here, we slightly abuse notation and write \mathbf{e}_1 for the first vector of the standard basis in both \mathbb{C}^M and \mathbb{C}^N .

This characterization of rank-1 Hankel matrices is equivalent to the previous one given in Lemma 2.1. Indeed, using the counter-identity matrix \mathbf{J}_N from (1.10), for $z \neq 0$, we have the connection

$$\hat{\mathbf{z}}_N(z) = \mathbf{J}_N \cdot \hat{\mathbf{w}}_N(z) = z^{N-1} \cdot \hat{\mathbf{w}}_N(1/z)$$

between the structured vectors $\hat{\mathbf{z}}_N$ and $\hat{\mathbf{w}}_N$. For the sparse rank-1 Hankel matrix (2.3) in Lemma 2.1 we have $c \cdot \mathbf{e}_M \mathbf{e}_N^\top = c \cdot \hat{\mathbf{w}}_M(0) \cdot \hat{\mathbf{w}}_N(0)^\top$.

In the following, we strive to harmonize Remark 2.2 and Lemma 2.1. To this end, we consider the *normalized* structured vectors

$$\mathbf{z}_N(z) := \frac{\hat{\mathbf{z}}_N(z)}{\|\hat{\mathbf{z}}_N(z)\|_2} = \left(\sum_{k=0}^{N-1} |z|^{2k} \right)^{-1/2} \cdot \left(1 \quad z \quad z^2 \quad \dots \quad z^{N-1} \right)^\top, \quad (2.5)$$

and $\mathbf{w}_N(z) := \hat{\mathbf{w}}_N(z) / \|\hat{\mathbf{w}}_N(z)\|_2$ defined analogously. Note that this normalization is always possible since $\|\hat{\mathbf{z}}_N(z)\|_2 \geq 1$ for all $z \in \mathbb{C}$; for $z = 0$ we invoke the convention $0^0 = 1$. Again, we will only write \mathbf{z}_N or \mathbf{z} instead of $\mathbf{z}_N(z)$ when we do not risk confusion about the parameters.

We find that $\mathbf{z}_N(z) = \mathbf{w}_N(1/z)$. In particular for real z , we obtain the first and last vector of the standard basis as limit cases

$$\mathbf{e}_1 = \mathbf{z}_N(0) = \lim_{z \rightarrow \infty} \mathbf{w}_N(z) \quad \text{and} \quad \mathbf{e}_N = \mathbf{w}_N(0) = \lim_{z \rightarrow \infty} \mathbf{z}_N(z)$$

for $z \rightarrow \infty$. For complex z , we understand $z \rightarrow \infty$ as $|z| \rightarrow \infty$ and $\mathcal{I}m(z) \rightarrow 0$ and use the limits introduced above in the same way.

With this notion we do not have to deal with the sparse rank-1 Hankel matrix (2.3) separately. Instead we allow the structure parameter z to assume the value infinity. Thus, we can reformulate Lemma 2.1 as follows.

Lemma 2.3 *A complex rank-1 matrix $\mathbf{H} \in \mathbb{C}^{M \times N}$ has Hankel structure if and only if it is of the form*

$$\mathbf{H} = \frac{c}{\|\hat{\mathbf{z}}_M\|_2 \|\hat{\mathbf{z}}_N\|_2} \cdot \hat{\mathbf{z}}_M \hat{\mathbf{z}}_N^T, = c \cdot \mathbf{z}_M \mathbf{z}_N^T, \quad (2.6)$$

where $c \in \mathbb{C} \setminus \{0\}$ and $z \in \overline{\mathbb{C}} := \mathbb{C} \cup \{\infty\}$.

As another consequence of allowing $z \in \overline{\mathbb{C}}$, we do not need to distinguish between the structured vector \mathbf{z} and the reversed structured vector \mathbf{w} anymore.

Lemma 2.3 is in accordance with the model for rank- r Hankel matrices from [HR84], see also the next section.

2.2 Hankel Matrices of Higher Rank

This section is an excursus to Hankel matrices of higher rank. Analogously to Section 2.1, we establish a complete characterization of rank- r Hankel matrices. Thereby we build upon the model from [HR84].

If any $(M \times N)$ matrix has rank r , it must possess exactly r linearly independent rows or columns, see Definition 1.3. Throughout this section assume that $M \leq N$. Since the rank is smaller than or equal to the smaller matrix dimension (i.e., $r \leq M$), we will work with rows

rather than columns. Now, consider a Hankel matrix

$$\mathbf{H} = \begin{pmatrix} h_0 & h_1 & h_2 & \cdots & h_{M-1} & h_M & \cdots & h_{N-1} \\ h_1 & h_2 & & & & & & h_N \\ h_2 & & & & & & & \vdots \\ \vdots & & & & & & & \vdots \\ h_{M-1} & h_M & \cdots & h_{N-1} & h_N & \cdots & h_{M+N-2} \end{pmatrix}$$

as in (1.4) and suppose $\text{rank } \mathbf{H} = r$.

Assume for the moment that already the first r rows of \mathbf{H} are linearly independent.

Then the $(r + 1)$ -st row must be a linear combination of the first r rows. Denoting the r -th row $(h_{r-1} \cdots h_{r+N-2})$ of \mathbf{H} by \mathbf{h}_{r-1} , we thus obtain

$$\mathbf{h}_r = a_1 \mathbf{h}_{r-1} + a_2 \mathbf{h}_{r-2} + \cdots + a_r \mathbf{h}_0$$

for some coefficients a_1, \dots, a_r . Due to the Hankel structure of \mathbf{H} , the same relation can be stated for the entries of \mathbf{H} instead of the columns. Furthermore, it remains true for all indices larger than or equal to r . We obtain

$$\begin{aligned} h_n &= a_1 h_{n-1} + a_2 h_{n-2} + \cdots + a_r h_{n-r} \\ \Leftrightarrow 0 &= h_n - a_1 h_{n-1} - a_2 h_{n-2} - \cdots - a_r h_{n-r} \end{aligned} \quad (2.7)$$

for $n = r, \dots, M + N - 2$ and the same coefficients a_1, \dots, a_r as above. In other words, the entries of \mathbf{H} must satisfy the homogeneous linear recurrence relation of order r described by (2.7). Without loss of generality, we assume here $a_r \neq 0$, otherwise the order of the recurrence would be smaller than r .

Obviously, the recurrence (2.7) is trivially solved by the zero-sequence, $h_k = 0$ for all k . This trivial solution is not important for us since the matrix whose entries are all zero has rank zero instead of r . It is only mentioned here for completeness' sake.

There is extensive literature on recurrence relations and their solutions, see [AN07; Ber86; GKP94; KP11; Wil06] just to name a few examples. We proceed here according to [Ber86].

Associated to the recurrence relation (2.7), there is the characteristic polynomial

$$\pi(x) = x^r - a_1 x^{r-1} - \cdots - a_{r-1} x - a_r,$$

whose degree r matches the order of the recurrence. The coefficients a_1, \dots, a_r are the same as in (2.7). Recall that $a_r \neq 0$ by assumption; consequently, zero cannot be a root of the characteristic polynomial. Let thus $z_1, \dots, z_m \neq 0$ be the distinct complex roots of $\pi(x)$ with corresponding multiplicities r_1, \dots, r_m satisfying $r_1 + \dots + r_m = r$. More formally, we have

$$\pi(x) = \prod_{\mu=1}^m (x - z_\mu)^{r_\mu},$$

where $z_\mu \neq 0$ and $r_1 + \dots + r_m = r$.

Remark 2.4 Since \mathbb{C} is algebraically closed, the respective roots' multiplicities always add up to the polynomial degree.

Having the characteristic polynomial and its roots set up, we can now state the following theorem, which provides the general solution to (2.7).

Theorem 2.5 ([Ber86, Thms. 1.1 & 1.2]) *Let z_1, z_2, \dots, z_m be the roots of the characteristic polynomial of a homogeneous linear recurrence relation (2.7) of order r . Furthermore, let r_1, r_2, \dots, r_m be their corresponding multiplicities, satisfying $r_1 + r_2 + \dots + r_m = r$. Then the general solution of the recurrence is given by*

$$h_n = \sum_{\mu=1}^m \sum_{\nu=0}^{r_\mu-1} c_{\mu,\nu} \cdot \binom{n}{\nu} \cdot z_\mu^n \quad (2.8)$$

for $n \geq r$. Every particular solution of (2.7) can be written in the form (2.8) with appropriate coefficients $c_{\mu,\nu}$, which are determined by the initial values h_0, h_1, \dots, h_{r-1} . Moreover, there are no further solutions than the ones mentioned in (2.8).

Due to the close connection between Hankel matrices of rank r and homogeneous linear recurrence relations of order r , the entries of a rank- r Hankel matrix can be of the form (2.8) for some coefficients $c_{\mu,\nu}$. As can be seen by (2.8), the structure of the entries h_n crucially depends on the multiplicities r_μ of the roots z_μ of the characteristic polynomial. Indeed, the entries' structure fundamentally differs depending on the multiplicities of the roots as the following example illustrates.

Example 2.6 Consider on the one hand a characteristic polynomial of degree two with distinct roots z_1 and z_2 , namely, $\pi(x) = (x - z_1) \cdot (x - z_2)$. The corresponding $(N \times N)$

Hankel matrix of rank two is given by

$$\begin{pmatrix} c_1 + c_2 & c_1 z_1 + c_2 z_2 & \cdots & c_1 z_1^{N-1} + c_2 z_2^{N-1} \\ c_1 z_1 + c_2 z_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ c_1 z_1^{N-1} + c_2 z_2^{N-1} & \cdots & \cdots & c_1 z_1^{2N-2} + c_2 z_2^{2N-2} \end{pmatrix},$$

which can be decomposed into the sum of two rank-1 Hankel matrices

$$= c_1 \cdot \begin{pmatrix} 1 & z_1 & \cdots & z_1^{N-1} \\ z_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ z_1^{N-1} & \cdots & \cdots & z_1^{2N-2} \end{pmatrix} + c_2 \cdot \begin{pmatrix} 1 & z_2 & \cdots & z_2^{N-1} \\ z_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ z_2^{N-1} & \cdots & \cdots & z_2^{2N-2} \end{pmatrix}.$$

Both of these rank-1 Hankel matrices have the same structure as seen in (2.2). Due to the simple form of the characteristic polynomial, we always have $\nu = 0$ in the solution (2.8). For better readability, we have omitted this index in the coefficients and only write $c_{\mu,0} = c_{\mu}$, $\mu = 1, 2$, in the above equation.

On the other hand, assume the characteristic polynomial has only one root z with multiplicity two, namely, $\pi(x) = (x - z)^2$. Then the corresponding rank-2 Hankel matrix of size $(N \times N)$ is of a different structure. Specifically, it is given by

$$\begin{pmatrix} c_0 & (c_0 + c_1)z & (c_0 + 2c_1)z^2 & \cdots & (c_0 + (N-1)c_1)z^{N-1} \\ (c_0 + c_1)z & (c_0 + 2c_1)z^2 & \ddots & \ddots & \vdots \\ (c_0 + 2c_1)z^2 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ (c_0 + (N-1)c_1)z^{N-1} & \cdots & \cdots & \cdots & (c_0 + (2N-2)c_1)z^{2N-2} \end{pmatrix}$$

and can be decomposed into

$$= c_0 \cdot \begin{pmatrix} 1 & z & \cdots & z^{N-1} \\ z & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ z^{N-1} & \cdots & \cdots & z^{2N-2} \end{pmatrix} + c_1 \cdot \begin{pmatrix} 0 & z & \cdots & (N-1)z^{N-1} \\ z & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ (N-1)z^{N-1} & \cdots & \cdots & (2N-2)z^{2N-2} \end{pmatrix}.$$

Here, we have $\mu = 1$. Therefore, we drop this index and write $c_{1,\nu} = c_\nu$, $\nu = 0, 1$, for short. Note that c_1 has a different meaning in the two example matrices.

For the second example matrix, the structure of the second summand differs notably from the one of the first summand. Indeed, the second summand already is of rank two, and adding the first summand (a rank-1 matrix as in (2.2)) does not alter the overall rank.

There is a slightly different representation of (2.8), see for example [Mar16]. They suggest

$$h_n = \sum_{\mu=1}^m \sum_{\nu=0}^{r_\mu-1} c_{\mu,\nu} \cdot \binom{n}{\nu} \cdot z_\mu^n = \sum_{\mu=1}^m \sum_{\nu=0}^{r_\mu-1} c'_{\mu,\nu} \cdot \nu! \cdot \binom{n}{\nu} \cdot z_\mu^{n-\nu}, \quad (2.9)$$

with modified constants $c'_{\mu,\nu} = c_{\mu,\nu} \cdot \frac{z_\mu^\nu}{\nu!}$. The advantage of this representation is its connection to the derivatives of the polynomial entries of the structured vector $\hat{\mathbf{z}}_N(z)$ defined in (2.1). Following [HR84], we define

$$\mathbf{l}_N^r(z) := \left(r! \cdot \binom{k}{r} \cdot z^{k-r} \right)_{k=0}^{N-1} = \left(\frac{d^r}{dz^r} z^k \right)_{k=0}^{N-1} \in \mathbb{C}^N \quad (2.10)$$

for $r = 0, \dots, N-1$ and $z \in \mathbb{C} \setminus \{0\}$. The binomial coefficient is $\binom{k}{r} = 0$ whenever $k < r$ by convention. Similarly as before in Section 2.1, we omit the index and write $\mathbf{l}^r(z) = \mathbf{l}_N^r(z)$ when there is no risk of confusing the dimensions. Note that for $r = 0$, we obtain exactly $\mathbf{l}_N^0(z) = (z^k)_{k=0}^{N-1} = \hat{\mathbf{z}}_N(z)$, the structured vector from (2.1).

In order to further investigate rank- r Hankel matrices, consider the matrix corresponding to one single summand in (2.9). Such a Hankel matrix is given by

$$\mathcal{H}_{M,N}(\mathbf{l}_{M+N-2}^r(z)) = \left(r! \cdot \binom{k+\ell}{r} \cdot z^{k+\ell-r} \right)_{k,\ell=0}^{M-1,N-1} = \left(\frac{d^r}{dz^r} z^{k+\ell} \right)_{k,\ell=0}^{M-1,N-1} \quad (2.11)$$

and is called *elementary* Hankel matrix by [HR84]. The operator $\mathcal{H}_{M,N}$ forms a Hankel matrix of size $(M \times N)$ from a parameter vector $\mathbf{p} \in \mathbb{C}^{M+N-2}$ of different entries, see the formal introduction thereof in Chapter 5. Whenever the dimensions are clear from the context, we may write $\mathcal{H}^r(z)$ instead of $\mathcal{H}_{M,N}(\mathbf{l}_{M+N-2}^r(z))$.

Remark 2.7 For $r = 0$, we obtain the rank-1 Hankel matrix $\mathcal{H}_{M,N} \mathbf{l}_{M+N-2}^0(z) = \hat{\mathbf{z}}_M \hat{\mathbf{z}}_N^\top$ as in (2.2) analogously to Section 2.1. However, for $r \geq 1$, the elementary Hankel matrix $\mathcal{H}_{M,N} \mathbf{l}_{M+N-2}^r(z)$ cannot be written as the outer product of $\mathbf{l}_M^r(z)$ and $\mathbf{l}_N^r(z)$.

Theorem 2.5 establishes all the solutions of a homogeneous linear recurrence relation of order r . Nevertheless, there are more rank- r Hankel matrices than those based on superpositions of (2.9) as the following example clarifies. The reason for this fact is the finiteness of the matrices.

Example 2.8 Consider (3×3) Hankel matrices of rank two. For the sake of simplicity, we set all coefficients $c_{\mu,\nu} = 1$. The two Hankel structures that fit into the framework of linear recurrence relations are given by

1. $z_1 \neq z_2, z_1, z_2 \in \mathbb{C} \setminus \{0, \infty\}$ resulting in $\mathbf{H} = \begin{pmatrix} 1+1 & z_1+z_2 & z_1^2+z_2^2 \\ z_1+z_2 & z_1^2+z_2^2 & z_1^3+z_2^3 \\ z_1^2+z_2^2 & z_1^3+z_2^3 & z_1^4+z_2^4 \end{pmatrix}$ and
2. $z_1 = z_2 = z \in \mathbb{C} \setminus \{0, \infty\}$ resulting in $\mathbf{H} = \begin{pmatrix} 1 & 2z & 3z^2 \\ 2z & 3z^2 & 4z^3 \\ 3z^2 & 4z^3 & 5z^4 \end{pmatrix}$,

compare also Example 2.6.

However, there are more Hankel matrices that are clearly of rank two but do not fit into the recurrence framework. These are precisely

3. $\mathbf{H} = \begin{pmatrix} 1+1 & z_1 & z_1^2 \\ z_1 & z_1^2 & z_1^3 \\ z_1^2 & z_1^3 & z_1^4 \end{pmatrix}$, which we will shortly see as the limit case $z_1 \in \mathbb{C} \setminus \{0, \infty\}, z_2 = 0$,
4. $\mathbf{H} = \begin{pmatrix} 1 & z_1 & z_1^2 \\ z_1 & z_1^2 & z_1^3 \\ z_1^2 & z_1^3 & z_1^4+1 \end{pmatrix}$, which will correspond to $z_1 \in \mathbb{C} \setminus \{0, \infty\}, z_2 = \infty$,
5. $\mathbf{H} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, corresponding to $z_1 = z_2 = 0$,
6. $\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, corresponding to $z_1 = 0$ and $z_2 = \infty$, and finally
7. $\mathbf{H} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$, corresponding to $z_1 = z_2 = \infty$.

By means of Example 2.8, we have demonstrated that not all possible rank- r Hankel matrices are based on superpositions of the elementary Hankel matrices (2.11). In the following, we analyze the special cases in items 3 to 7 more closely.

We examine the case when the $(r + 1)$ -st row of \mathbf{H} is not a linear combination of the r preceding ones. Consider the double triangular matrix with constant counter-diagonals

$$\mathbf{H}_{\text{sparse}} = \begin{pmatrix} s_0 & \cdots & \cdots & \cdots & s_{p-1} \\ \vdots & & & & \vdots \\ s_{p-1} & & & & \\ & & & & s'_{q-1} \\ & & & & \vdots \\ & & & s'_{q-1} & \cdots & \cdots & \cdots & s'_0 \end{pmatrix} \in \mathbb{C}^{M \times N}, \quad (2.12)$$

where the empty spaces in the matrix stand for the appropriate number of zeros. We assume $s_{p-1} \neq 0$ and $s'_{q-1} \neq 0$, and $p + q \leq M$. Then clearly, this is a Hankel matrix with rank $\mathbf{H}_{\text{sparse}} = p + q$. But it does not fit into the framework of recurrence relations of order $r = p + q$ as defined before.

However, if we allow $z = 0$ in (2.10) and (2.11), we obtain additional elementary matrices. First, consider the analogue of (2.10) with $z = 0$. Since the binomial coefficient is $\binom{k}{r} = 0$ for $k < r$, this is

$$\mathbf{l}_N^r(0) = \left(r! \cdot \binom{k}{r} \cdot z^{k-r} \Big|_{z=0} \right)_{k=0}^{N-1} = \left(\frac{d^r}{dz^r} z^k \Big|_{z=0} \right)_{k=0}^{N-1} = (\delta_{kr})_{k=0}^{N-1} = \mathbf{e}_{r+1},$$

the $(r + 1)$ -st vector of the standard basis. For the elementwise representation of the latter we use the Kronecker delta $\delta_{kr} = \begin{cases} 1, & k=r \\ 0, & k \neq r \end{cases}$. Note that this is in accordance with Section 2.1, where for $r = 0$ we have $\mathbf{l}^0(0) = \hat{\mathbf{z}}(0) = \mathbf{e}_1$.

The elementary Hankel matrix corresponding to $\mathbf{l}^r(0)$ is

$$\mathcal{H}_{M,N}(\mathbf{l}_{M+N-2}^r(0)) = (\delta_{k+\ell,r})_{k,\ell=0}^{M-1,N-1} = \begin{pmatrix} & & & & 1 \\ & & & & \vdots \\ & & & & \vdots \\ & & & & \vdots \\ 1 & & & & \end{pmatrix},$$

the $(r + 1)$ -st counter-diagonal matrix (starting to count in the upper right corner). The empty spaces in this matrix stand for the appropriate number of zeros.

Recall that in Section 2.1, we interpreted the last vector of the standard basis as structured vector indexed by $z = \infty$, more precisely, $\mathbf{e}_N = (0 \cdots \cdots 0 \ 1)^\top = \hat{\mathbf{z}}_N(\infty)$. In analogy,

2. The canonical representation of a Hankel matrix is not unique. One and the same Hankel matrix may even have up to infinitely many canonical representations, see [HR84, Thm. 8.1] and Example 2.12.

3. The rank of the representation ρ as in Definition 2.10 is not to be confused with the rank of the matrix from Definition 1.3. In fact, ρ can be strictly larger than the rank of the represented Hankel matrix, see Example 2.12. In representations (2.13) of minimal rank ρ , both ranks coincide.

In the upcoming example we illustrate the existence of canonical representations whose rank ρ exceeds not only the rank of the matrix but even its dimensions.

Example 2.12 Consider the following (3×3) Hankel matrix

$$\mathbf{H} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 6 \\ 0 & 6 & 24z \end{pmatrix}$$

with $z \in \mathbb{C}$. This matrix has a representation as a single elementary matrix. Namely, $\mathbf{H} = \mathcal{H}l^3(z)$ since

$$\frac{d^3}{dz^3} z^3 = 6 \quad \text{and} \quad \frac{d^3}{dz^3} z^4 = 24z,$$

compare (2.11). This is not a simple canonical representation, let alone a representation of minimal rank. In fact, the canonical representation using only $\mathcal{H}l^3(z)$ is given by (2.13) with $m = 1$, $\rho = 4$, and $c_{1,\nu} = 0$, for $\nu = 0, 1, 2$ and $c_{1,3} = 1$. So the rank of the representation is $\rho = 4$ while the matrix dimensions are only $M = N = 3$, and clearly, we have $\text{rank } \mathbf{H} = 2$.

The canonical representation of minimal rank is given by

$$\mathbf{H} = 24z \cdot \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} + 6 \cdot \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} = 24z \cdot \mathcal{H}l^0(\infty) + 6 \cdot \mathcal{H}l^1(\infty).$$

In terms of (2.13), we have $m = 2$, $r_\mu = 1$ for $\mu = 1, 2$, $c_{1,0} = 24z$ and $c_{2,0} = 6$. So for this representation we actually have $\rho = 2 = \text{rank } \mathbf{H}$.

Canonical representations of singular Hankel matrices are useful for Hankel SLRA problems. In this case, that is, when $\text{rank } \mathbf{H} = r < M$, the canonical representation of minimal rank ρ is unique, and we have $\text{rank } \mathbf{H} = r = \rho < M$, see [HR84, Cor. 8.1]. Hence, rank- r

Hankel matrices for $r < M$ are characterized by their canonical representation of minimal rank ρ .

We slightly reformulate the canonical representation (2.13) and state the following theorem, which is a combination of several statements from [HR84].

Theorem 2.13 *A Hankel matrix $\mathbf{H} \in \mathbb{C}^{M \times N}$ of rank $r < M$ is completely characterized by its canonical representation of minimal rank. This characterization is given by*

$$\mathbf{H} = \sum_{\nu=0}^{r_0-1} c_{0,\nu} \cdot \mathcal{H}\mathbf{U}^\nu(0) + \sum_{\mu=1}^m \sum_{\nu=0}^{r_\mu-1} c_{\mu,\nu} \cdot \mathcal{H}\mathbf{U}^\nu(z_\mu) + \sum_{\nu=0}^{r_\infty-1} c_{\infty,\nu} \cdot \mathcal{H}\mathbf{U}^\nu(\infty),$$

with $m, r_\mu \in \mathbb{N}_0$, and non-zero coefficients $c_{\mu,r_\mu-1} \neq 0$ for $\mu = 0, 1, \dots, m, \infty$. For the rank constraint we require $\rho = \sum_{\mu} r_\mu = r$.

Now that we have found a characterization of rank- r Hankel matrices, we want to answer the following question.

How many different types of rank- r Hankel matrices are there for fixed r ? In order to answer this question, we will count through all the possibilities of different canonical representations from Theorem 2.13 of rank $\rho = r < M$. Before starting to count, recall that the vectors $\mathbf{l}^r(0)$ and $\mathbf{l}^r(\infty)$ do not satisfy a homogeneous linear recurrence relation. Therefore, the elementary Hankel matrices with $z = 0$ and $z = \infty$ take a special position, as indicated by Theorem 2.13, and are treated separately in the sequel.

We first consider the sparse Hankel matrices of size $(M \times N)$

$$\mathbf{H}_{\text{sparse}} := \sum_{\nu=0}^{r_0-1} c_{0,\nu} \cdot \mathcal{H}\mathbf{U}^\nu(0) + \sum_{\nu=0}^{r_\infty-1} c_{\infty,\nu} \cdot \mathcal{H}\mathbf{U}^\nu(\infty)$$

as in (2.12). For fixed rank $r < M$, there are $r + 1$ possibilities to distribute non-zero entries between $z = 0$ and $z = \infty$, that is, between upper left and lower right corner of the matrix.

Next we examine dense $(M \times N)$ Hankel matrices of rank r . They are of the form

$$\mathbf{H}_{\text{dense}} := \sum_{\mu=1}^m \sum_{\nu=0}^{r_\mu-1} c_{\mu,\nu} \cdot \mathcal{H}\mathbf{U}^\nu(z_\mu),$$

where $\sum_{\mu=1}^m r_\mu = r$. As seen in Example 2.6, the structure of a dense rank- r Hankel matrix crucially depends on the partition of the rank r into the summands r_μ , $\mu = 1, \dots, m$. The

number of possibilities to partition a number r into positive integer summands is given by the partition function $p(r)$, see [AS92, Sec. 24.2.1].

Now we have to combine sparse and dense Hankel matrices to a generic Hankel matrix of rank r . We count the number of differently structured rank- r Hankel matrices

$$\# \{ \mathbf{H} \text{ Hankel and } \text{rank } \mathbf{H} = r \} = \sum_{k=0}^r p(r-k) \cdot (k+1). \quad (2.14)$$

In this equation, $p(r-k)$ is the number of differently structured dense Hankel matrices of rank $r-k$; and $k+1$ is the number of possibilities to split the remaining rank k between upper left and lower right corner of a sparse Hankel matrix.

Although there is no closed form of the partition function $p(r)$ known, there are approximation formulas for large r [HR18; Rad37]. In particular, the number of digits of $p(r)$ is approximately proportional to \sqrt{r} for large r . This illustrates how fast the partition function alone grows with its argument. In addition to its fast growth, in (2.14) the partition function is summed over all lower ranks.

Example 2.14 We calculate the number of different structures for a few exemplary ranks. For a rank-2 Hankel matrix there are

$$p(2) \cdot 1 + p(1) \cdot 2 + p(0) \cdot 3 = 7$$

different structures possible. See Example 2.8 for an explicit list of these structures.

For higher ranks, say $r = 5$ and $r = 10$ for example, we obtain

$$\sum_{k=0}^5 p(5-k) \cdot (k+1) = 45 \quad \text{and} \quad \sum_{k=0}^{10} p(10-k) \cdot (k+1) = 432,$$

respectively. These examples demonstrate the enormous increase in the number of possible types of Hankel matrices even for moderate rank.

Remark 2.15 When counting the number of different possible rank- r Hankel structures, we treated the sparse Hankel matrices separately. This makes sense not only because of their exceptional behavior with respect to recurrence relations. Besides, there are computational reasons to consider $z = 0$ and $z = \infty$ individually. Especially $z_\mu = \infty$ is difficult to handle in numerical computations.

Even if we didn't count the sparse Hankel matrices separately in (2.14), we would end up with $p(r)$ different rank- r Hankel structures, which grows fast enough on its own.

In Chapters 3 and 4, we find optimal rank-1 Hankel approximations by identifying the optimal parameters c and z from Section 2.1. The exact same procedure for rank- r Hankel approximation quickly becomes infeasible due to the enormous number (2.14) of different rank- r Hankel structures, compare Example 2.8. On the other hand, those Hankel structures that cannot be written as a sum of rank-1 Hankel matrices can be viewed as limit cases of the latter, where some of the structure parameters tend toward each other, or to zero or infinity. Heuristically, it is easy to accept that these limit cases only take up a small part in the set of rank- r Hankel matrices. In [AC11a; AC11b], the notion of a thin set is employed to describe this small part and to argue that it may be disregarded when solving the Hankel SLRA problem.

3

OPTIMAL RANK-1 HANKEL APPROXIMATION IN THE FROBENIUS NORM

In this chapter we examine the rank-1 Hankel approximation (r1H) problem (2) in the Frobenius norm, namely,

$$\min \|\mathbf{A} - \mathbf{H}\|_F \quad \text{such that } \text{rank } \mathbf{H} = 1, \\ \text{and } \mathbf{H} \text{ has Hankel structure}$$

for an initial matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$.

We analytically reformulate the r1H problem such that numerical computation of its optimal solution is feasible. This chapter is mostly based on and is in parts identical with our paper [KPP21a, Sec. 3]. In Section 3.2 we give some additional results that are not contained in [KPP21a] and have not been published otherwise yet.

Using the structure of a rank-1 Hankel matrix from Lemma 2.3 we can formulate the minimization problem above as

$$\min_{\substack{z \in \overline{\mathbb{C}} \\ c \in \mathbb{C} \setminus \{0\}}} \|\mathbf{A} - c \cdot \mathbf{z}_M \mathbf{z}_N^\top\|_F. \quad (3.1)$$

This means, we need to find coefficients $c \in \mathbb{C} \setminus \{0\}$ and $z \in \overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ such that the error $\|\mathbf{A} - c \cdot \mathbf{z}_M \mathbf{z}_N^\top\|_F$ in the Frobenius norm is minimized. In this context we will, without loss of generality, assume $|a_{0,0}| \geq |a_{M-1,N-1}|$ for the top left and bottom right

entry of the matrix $\mathbf{A} = (a_{jk})_{j,k=0}^{M-1,N-1} \in \mathbb{C}^{M \times N}$. This assumption implies the relation

$$\begin{aligned} \|\mathbf{A} - c \cdot \mathbf{z}_M(\infty)\mathbf{z}_N(\infty)^\top\|_F^2 &= \|\mathbf{A} - c \cdot \mathbf{e}_M \mathbf{e}_N^\top\|_F^2 \\ &\geq \|\mathbf{A} - c \cdot \mathbf{e}_1 \mathbf{e}_1^\top\|_F^2 = \|\mathbf{A} - c \cdot \mathbf{z}_M(0)\mathbf{z}_N(0)^\top\|_F^2 \end{aligned}$$

for all $c \in \mathbb{C}$. The vectors \mathbf{e}_1 and \mathbf{e}_M or \mathbf{e}_N are the first and last vectors of the standard basis in \mathbb{C}^M or \mathbb{C}^N , respectively. We use the notation \mathbf{e}_1 for the first vector of the standard basis of both \mathbb{C}^M and \mathbb{C}^N as in Chapter 2.

As a consequence of the above inequality, the parameter $z = \infty$ will not generate the (only) desired optimal rank-1 Hankel approximation of \mathbf{A} and can therefore be disregarded. The fact that there is no need to deal with the value $z = \infty$ makes up the advantage of the assumption $|a_{0,0}| \geq |a_{M-1,N-1}|$. Furthermore, it does not pose any restriction on \mathbf{A} since it can simply be replaced by the flipped matrix $\mathbf{J}_M \mathbf{A} \mathbf{J}_N$.

3.1 Complex Rank-1 Hankel Approximation

First, we consider the minimization problem (3.1) in full generality for complex matrices \mathbf{A} and complex parameters c and z . In Section 3.2 we will examine the special case of real matrices \mathbf{A} more closely. The main results of this chapter are stated in the upcoming Theorem 3.1 and in Theorem 3.4.

Theorem 3.1 *Let $\mathbf{A} = (a_{jk})_{j,k=0}^{M-1,N-1} \in \mathbb{C}^{M \times N}$ with $|a_{0,0}| \geq |a_{M-1,N-1}|$, and assume $\text{rank } \mathbf{A} \geq 1$. Then an optimal rank-1 Hankel approximation $\tilde{c} \cdot \tilde{\mathbf{z}}_M \tilde{\mathbf{z}}_N^\top$ of \mathbf{A} is determined by*

$$\tilde{z} \in \underset{z \in \mathbb{C}}{\text{argmax}} |\mathbf{z}_M^* \mathbf{A} \bar{\mathbf{z}}_N| \quad \text{and} \quad \tilde{c} := \tilde{\mathbf{z}}_M^* \mathbf{A} \tilde{\mathbf{z}}_N, \quad (3.2)$$

where the vectors $\tilde{\mathbf{z}}_M$ and $\tilde{\mathbf{z}}_N$ are the normalized structured vectors defined by the parameter \tilde{z} via (2.5).

Proof. Using Definition 1.5 of the Frobenius norm, we obtain

$$\begin{aligned} \|\mathbf{A} - c \cdot \mathbf{z}_M \mathbf{z}_N^\top\|_F^2 &= \text{tr} \left((\mathbf{A} - c \cdot \mathbf{z}_M \mathbf{z}_N^\top)^* \cdot (\mathbf{A} - c \cdot \mathbf{z}_M \mathbf{z}_N^\top) \right) \\ &= \text{tr} \left(\mathbf{A}^* \mathbf{A} - c \cdot \mathbf{A}^* \cdot \mathbf{z}_M \mathbf{z}_N^\top - \bar{c} \cdot \bar{\mathbf{z}}_N \mathbf{z}_M^* \cdot \mathbf{A} + |c|^2 \cdot \bar{\mathbf{z}}_N \mathbf{z}_M^* \cdot \mathbf{z}_M \mathbf{z}_N^\top \right) \\ &= \|\mathbf{A}\|_F^2 - c \cdot \mathbf{z}_M^\top \bar{\mathbf{A}} \mathbf{z}_N - \bar{c} \cdot \mathbf{z}_M^* \mathbf{A} \bar{\mathbf{z}}_N + |c|^2. \end{aligned} \quad (3.3)$$

Here we have used that the trace of the outer product of two vectors equals their inner product (i.e., $\text{tr}(\mathbf{b}\mathbf{a}^\top) = \mathbf{a}^\top\mathbf{b}$), and that the vectors \mathbf{z}_M and \mathbf{z}_N are normalized.

In order to solve the minimization problem (3.1), we first assume z to be fixed. For fixed z , we consider the derivatives of (3.3) with respect to the real and imaginary parts of c . More precisely, we form the partial derivatives with respect to c_1 and c_2 , where $c = c_1 + ic_2$ and $c_1, c_2 \in \mathbb{R}$. Thereby we obtain the necessary conditions

$$\begin{aligned} \frac{\partial}{\partial c_1} \|\mathbf{A} - c \cdot \mathbf{z}_M \mathbf{z}_N^\top\|_F^2 &= -2 \cdot \text{Re}(\mathbf{z}_M^* \mathbf{A} \bar{\mathbf{z}}_N) + 2c_1 = 0, \\ \frac{\partial}{\partial c_2} \|\mathbf{A} - c \cdot \mathbf{z}_M \mathbf{z}_N^\top\|_F^2 &= -2 \cdot \text{Im}(\mathbf{z}_M^* \mathbf{A} \bar{\mathbf{z}}_N) + 2c_2 = 0 \end{aligned}$$

for the optimal parameter $\tilde{c} = \tilde{c}_1 + i\tilde{c}_2$. These put together yield $\tilde{c} = \mathbf{z}_M^* \mathbf{A} \bar{\mathbf{z}}_N$. After substituting this \tilde{c} into (3.3), it remains to solve

$$\min_{z \in \mathbb{C}} \left(\|\mathbf{A}\|_F^2 - 2 \cdot |\mathbf{z}_M^* \mathbf{A} \bar{\mathbf{z}}_N|^2 + |\mathbf{z}_M^* \mathbf{A} \bar{\mathbf{z}}_N|^2 \right) = \min_{z \in \mathbb{C}} \left(\|\mathbf{A}\|_F^2 - |\mathbf{z}_M^* \mathbf{A} \bar{\mathbf{z}}_N|^2 \right).$$

Thus, we obtain

$$\tilde{z} \in \operatorname{argmax}_{z \in \mathbb{C}} |\mathbf{z}_M^* \mathbf{A} \bar{\mathbf{z}}_N|^2 = \operatorname{argmax}_{z \in \mathbb{C}} |\mathbf{z}_M^\top \bar{\mathbf{A}} \mathbf{z}_N|$$

as claimed. \square

Remark 3.2 1. The optimal structure parameter \tilde{z} in (3.2) may not be unique; that is, the maximum $\max_{z \in \mathbb{C}} |F(z)|$ may be attained for several different values \tilde{z} . If for example the initial matrix $\mathbf{A} = (a_{jk})_{j,k=0}^{M-1, N-1} \in \mathbb{C}^{M \times N}$ has itself Hankel structure, with $a_{jk} = 0$ whenever $j+k$ is odd, then $\tilde{z} \in \operatorname{argmax}_{z \in \mathbb{C}} |F(z)|$ implies that also $-\tilde{z} \in \operatorname{argmax}_{z \in \mathbb{C}} |F(z)|$. Any of the values $\tilde{z} \in \operatorname{argmax}_{z \in \mathbb{C}} |F(z)|$ leads to an optimal rank-1 Hankel approximation, see also Example 3.8 for an explicit precedent.

2. By Theorem 3.1, the computation of the optimal rank-1 Hankel approximation reduces to finding a position $z \in \mathbb{C}$ where the maximum of the complex function $|F(z)|$ with

$$F(z) := \mathbf{z}_M^\top \bar{\mathbf{A}} \mathbf{z}_N$$

is attained. Since $\|\mathbf{z}_M\|_2 = \|\mathbf{z}_N\|_2 = 1$ for all $z \in \mathbb{C}$, the function $F(z)$ has no poles. The absolute value $|F(z)|$ is bounded by $\|\mathbf{A}\|_2$ which follows from the proof of the next theorem. If additionally $\mathbf{A} \in \mathbb{R}^{N \times N}$ is symmetric or $\mathbf{A} \in \mathbb{C}^{N \times N}$ is Hermitian, then $F(z)$ is a Rayleigh quotient. Hence $\lambda_{N-1} \leq F(z) \leq \lambda_0$, where λ_{N-1} and λ_0 are the smallest and

largest eigenvalue of \mathbf{A} , respectively, see [H]13, Sec. 4.2].

3. For the function $F(z)$, we observe that

$$F(z) = \mathbf{z}_M^\top \overline{\mathbf{A}} \mathbf{z}_N = \frac{\sum_{\ell=0}^{M+N-2} \left(\sum_{j+k=\ell} \bar{a}_{jk} \right) \cdot z^\ell}{\left(\sum_{j=0}^{M-1} |z|^{2j} \right)^{1/2} \cdot \left(\sum_{j=0}^{N-1} |z|^{2j} \right)^{1/2}},$$

where $\mathbf{A} = (a_{jk})_{j,k=0}^{M-1, N-1}$. Therefore, without loss of generality, \mathbf{A} can be replaced by the Hankel matrix $\mathcal{P}(\mathbf{A})$ from (1.6).

The numerator of F ,

$$\sum_{\ell=0}^{M+N-2} \left(\sum_{j+k=\ell} \bar{a}_{j,k} \right) \cdot z^\ell,$$

does not change if the Hankel matrix $\mathcal{P}(\mathbf{A})$ is reshaped into a Hankel matrix of size $2 \times (M + N - 2)$ with entries $h_\ell = \sum_{j+k=\ell} \bar{a}_{jk}$, $\ell = 0, \dots, M + N - 2$, as in Definition 1.10. This observation gives the link to the rational function approach in [UM12]. Note that such reshaping of a rank-1 Hankel matrix into the size $2 \times (M + N - 2)$ does not alter its rank, see [HR84]. However, it does change the solution of (3.1) since the denominator of F ,

$$\left(\sum_{j=0}^{M-1} |z|^{2j} \right)^{1/2} \cdot \left(\sum_{j=0}^{N-1} |z|^{2j} \right)^{1/2},$$

strongly depends on the shape of the Hankel matrix.

4. In [CGM⁺11; KL98], a similar rational approximation problem as (3.2) appears in the context of finding an approximate greatest common divisor. The denominator there has exactly the same structure as the denominator of F if $M = N$. In [CGM⁺11], a subdivision method on squares in the complex plane is proposed to solve that problem.

With Theorem 3.1 we find an optimal approximation in the space of rank-1 Hankel matrices. We want to compare its error to the optimal error bound (1.2) from the unstructured rank-1 approximation.

Recall from Section 1.2 that the optimal unstructured rank-1 approximation is given by the matrix $\sigma_0 \cdot \mathbf{u}_0 \mathbf{v}_0^*$ formed from the first singular triple of \mathbf{A} . Thereby, σ_0 is the largest singular value of \mathbf{A} . The corresponding left and right singular vectors \mathbf{u}_0 and \mathbf{v}_0 are characterized

by the following set of equations (see also Remark 1.2.3)

$$\mathbf{A}\mathbf{A}^* \cdot \mathbf{u}_0 = \sigma_0 \cdot \mathbf{u}_0, \quad \mathbf{A}^* \mathbf{A} \cdot \mathbf{v}_0 = \sigma_0 \cdot \mathbf{v}_0 \quad \text{and} \quad \mathbf{u}_0 = \frac{1}{\sigma_0} \cdot \mathbf{A}\mathbf{v}_0, \quad \mathbf{v}_0 = \frac{1}{\sigma_0} \cdot \mathbf{A}^* \mathbf{u}_0. \quad (3.4)$$

The following theorem answers the question in which cases the optimal solution to the r1H problem is as good as the unstructured rank-1 approximation, that is, in which cases the optimal error bound is attained. It turns out that the optimal error bound (1.2) can only be attained if the rank-1 approximation $\sigma_0 \cdot \mathbf{u}_0 \mathbf{v}_0^*$ already happens to have Hankel structure.

Theorem 3.3 *Let $\mathbf{A} = (a_{jk})_{j,k=0}^{M-1,N-1} \in \mathbb{C}^{M \times N}$ with $|a_{0,0}| \geq |a_{M-1,N-1}|$. The optimal rank-1 Hankel approximation $\tilde{c} \cdot \tilde{\mathbf{z}}_M \tilde{\mathbf{z}}_N^\top$ from Theorem 3.1 attains the optimal error bound (1.2), or in other words, satisfies the following equation*

$$\|\mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}_M \tilde{\mathbf{z}}_N^\top\|_F^2 = \sum_{j=1}^{N-1} \sigma_j^2 = \|\mathbf{A}\|_F^2 - \|\mathbf{A}\|_2^2, \quad (3.5)$$

if and only if the singular vectors \mathbf{u}_0 and \mathbf{v}_0 of \mathbf{A} corresponding to the largest singular value σ_0 are of the structured form (2.5); more precisely, if we have $\mathbf{u}_0 = \tilde{\mathbf{z}}_M$ and $\mathbf{v}_0 = \tilde{\mathbf{z}}_N$ for some $\tilde{z} \in \mathbb{C}$.

Proof. Truncating the SVD of \mathbf{A} , we obtain an optimal unstructured rank-1 approximation $\sigma_0 \cdot \mathbf{u}_0 \mathbf{v}_0^*$ by Theorem 1.8. The corresponding error is given by

$$\|\mathbf{A} - \sigma_0 \cdot \mathbf{u}_0 \mathbf{v}_0^*\|_F^2 = \sum_{j=1}^{N-1} \sigma_j^2 = \|\mathbf{A}\|_F^2 - \sigma_0^2,$$

which we named optimal error bound (1.2) in Section 1.2.

Assume now that $\mathbf{u}_0 = \tilde{\mathbf{z}}_M$ and $\mathbf{v}_0 = \tilde{\mathbf{z}}_N$ are structured with $\mathbf{z}_M, \mathbf{z}_N$ as in (2.5). With $\tilde{c} = \tilde{\mathbf{z}}_M^* \mathbf{A} \tilde{\mathbf{z}}_N$ from Theorem 3.1, it follows that

$$\tilde{c} \cdot \tilde{\mathbf{z}}_M \tilde{\mathbf{z}}_N^\top = \tilde{\mathbf{z}}_M^* \mathbf{A} \tilde{\mathbf{z}}_N \cdot \tilde{\mathbf{z}}_M \tilde{\mathbf{z}}_N^\top = \mathbf{u}_0^* \mathbf{A} \mathbf{v}_0 \cdot \mathbf{u}_0 \mathbf{v}_0^* = \sigma_0 \cdot \mathbf{u}_0 \mathbf{v}_0^*,$$

that is, the unstructured and the structured rank-1 approximation coincide.

For the converse, assume that the optimal structured rank-1 approximation $\tilde{c} \cdot \tilde{\mathbf{z}}_M \tilde{\mathbf{z}}_N^\top$ attains the optimal error bound; i.e., satisfies (3.5). According to equation (3.3) from the

proof of Theorem 3.1, we have

$$\|\mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}_M \tilde{\mathbf{z}}_N^\top\|_F^2 = \|\mathbf{A}\|_F^2 - |\tilde{\mathbf{z}}_M^* \mathbf{A} \bar{\tilde{\mathbf{z}}}_N|^2.$$

From this equation and (3.5) we conclude, on the one hand, the relation

$$\sigma_0^2 = \|\mathbf{A}\|_2^2 = |\tilde{\mathbf{z}}_M^* \mathbf{A} \bar{\tilde{\mathbf{z}}}_N|^2.$$

On the other hand, the Theorem of Rayleigh-Ritz (see [HJ13, Sec. 4.2]) implies

$$|\tilde{\mathbf{z}}_M^* \mathbf{A} \bar{\tilde{\mathbf{z}}}_N|^2 \leq \|\bar{\tilde{\mathbf{z}}}_N \tilde{\mathbf{z}}_N^\top\|_2 \cdot \|\mathbf{A}^* \tilde{\mathbf{z}}_M\|_2^2 = \|\mathbf{A}^* \tilde{\mathbf{z}}_M\|_2^2 \leq \|\mathbf{A}\|_2^2.$$

Here, the first inequality is tight if and only if $\mathbf{A}^* \tilde{\mathbf{z}}_M$ is an eigenvector of the matrix $\bar{\tilde{\mathbf{z}}}_N \tilde{\mathbf{z}}_N^\top$ to its non-zero eigenvalue $\|\tilde{\mathbf{z}}_N\|_2^2$. The second inequality is tight if moreover $\tilde{\mathbf{z}}_M$ is an eigenvector of $\mathbf{A} \mathbf{A}^*$ to the largest eigenvalue $\sigma_0^2 = \|\mathbf{A}\|_2^2$. The assertion now follows from comparison with (3.4). \square

In the remainder of this chapter, we derive further properties of the optimal parameter \tilde{z} in (3.2). Thereby we aim at finding an efficient algorithm providing optimal parameters \tilde{c} and \tilde{z} . First, we consider the possible range of \tilde{z} .

Theorem 3.4 *Let $\mathbf{A} \in \mathbb{C}^{M \times N}$ with $\text{rank } \mathbf{A} \geq 1$. For $z \in \bar{\mathbb{C}}$ define*

$$F(z) := \mathbf{z}_M^* \mathbf{A} \bar{\mathbf{z}}_N \quad \text{and} \quad G(z) := \mathbf{z}_M^* \mathbf{J}_M \mathbf{A} \mathbf{J}_N \bar{\mathbf{z}}_N,$$

with the counter-identity matrix \mathbf{J}_N as in (1.10). Denote by $F_{max} := \max_{|z| \leq 1} |F(z)|$ and $G_{max} := \max_{|z| \leq 1} |G(z)|$ the absolute maxima of the functions F and G , respectively. Then the optimal structure parameter \tilde{z} leading to an optimal rank-1 Hankel approximation $\tilde{c} \cdot \tilde{\mathbf{z}}_M \tilde{\mathbf{z}}_N^\top$ of \mathbf{A} is determined by

$$\tilde{z} \in \begin{cases} \operatorname{argmax}_{|z| \leq 1} |F(z)| & \text{if } F_{max} \geq G_{max}, \\ \left(\operatorname{argmax}_{|z| < 1} |G(z)| \right)^{-1} & \text{if } F_{max} < G_{max}. \end{cases}$$

Proof. The proof is based on the fact that $G(z) = F(1/z)$. Once this is established, the assertion follows from Theorem 3.1.

Recall from Section 2.1 that we have the relation

$$\mathbf{J}_N \cdot \mathbf{z}_N(z) = \mathbf{z}_N(1/z)$$

for $z \in \mathbb{C} \setminus \{0\}$. For $z \in \{0, \infty\}$ the structured vector $\mathbf{z}_N(z)$ is the first, respectively last vector of the standard basis, and the above relation indeed holds for all $z \in \overline{\mathbb{C}}$. Thus, we conclude that

$$G(z) = (\mathbf{J}_M \mathbf{z}_M(z))^* \cdot \mathbf{A} \cdot (\mathbf{J}_N \bar{\mathbf{z}}_N(z)) = \mathbf{z}_M(1/z)^* \cdot \mathbf{A} \cdot \bar{\mathbf{z}}_N(1/z) = F(1/z).$$

Now Theorem 3.1 yields the assertion. \square

Remark 3.5 Using Theorem 3.4 and the two functions F and G , we can restrict the search for an optimal parameter \tilde{z} to the unit disc $\{z: |z| \leq 1\}$. This elegantly spares us the rather uncomfortable handling of $z = \infty$ without invoking the assumption $|a_{0,0}| \geq |a_{M-1,N-1}|$.

3.2 Real Rank-1 Hankel Approximation

In the following, we consider real matrices $\mathbf{A} \in \mathbb{R}^{M \times N}$ and their real optimal rank-1 Hankel approximations; that is, we restrict our search to real coefficients $\tilde{c} \in \mathbb{R} \setminus \{0\}$ and structure parameters $\tilde{z} \in \overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. In this case we can derive additional conditions on \tilde{z} that simplify its computation. A similar approach for approximation in a weighted Frobenius norm has been presented in [DeM94].

Theorem 3.6 *Let $\mathbf{A} = (a_{jk})_{j,k=0}^{M-1,N-1} \in \mathbb{R}^{M \times N}$ be a matrix with $|a_{0,0}| \geq |a_{M-1,N-1}|$, and $\text{rank } \mathbf{A} \geq 1$. If $\tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top$ is an optimal rank-1 Hankel approximation of \mathbf{A} , then*

$$Q(\tilde{z}) := a'(\tilde{z}) \cdot p(\tilde{z}) - a(\tilde{z}) \cdot p'(\tilde{z}) = 0,$$

where $a(z)$ and $p(z)$ are the functions

$$a(z) := \sum_{j=0}^{M-1} \sum_{k=0}^{N-1} a_{jk} \cdot z^{j+k} \quad \text{and} \quad p(z) := \left(\sum_{j=0}^{M-1} z^{2j} \right)^{1/2} \cdot \left(\sum_{j=0}^{N-1} z^{2j} \right)^{1/2} \geq 1.$$

Besides, $a'(z)$ and $p'(z)$ denote the first derivatives of $a(z)$ and $p(z)$, respectively.

Proof. Employing Theorem 3.1 we obtain the optimal structure parameter \tilde{z} as

$$\tilde{z} \in \operatorname{argmax}_{z \in \mathbb{R}} |F(z)| \quad \text{with} \quad F(z) = \mathbf{z}_M^\top \mathbf{A} \mathbf{z}_N = \frac{a(z)}{p(z)},$$

compare also Remark 3.2.3. In other words, $F(\tilde{z})$ is an extremal value of F and as such has vanishing first derivative. The first derivative of F is given by

$$F'(z) = \frac{a'(z) \cdot p(z) - a(z) \cdot p'(z)}{p(z)^2}.$$

Since $p(z) \geq 1$ for all $z \in \mathbb{R}$, we obtain the necessary condition

$$a'(\tilde{z}) \cdot p(\tilde{z}) - a(\tilde{z}) \cdot p'(\tilde{z}) = 0$$

on the optimal parameter \tilde{z} as claimed. \square

Considering the monomial representation of the polynomial

$$a(z) = \sum_{j=0}^{M-1} \sum_{k=0}^{N-1} a_{jk} z^{j+k} =: \sum_{\ell=0}^{M+N-1} h_\ell z^\ell \quad (3.6)$$

with $h_\ell = \sum_{j+k=\ell} a_{jk}$, $\ell = 0, \dots, M+N-1$, as in Remark 3.2.3, we can conclude even more.

Corollary 3.7 *Let $\mathbf{A} \in \mathbb{R}^{M \times N}$ with $|a_{0,0}| \geq |a_{M-1,N-1}|$, and $\operatorname{rank} \mathbf{A} \geq 1$. Further, let $a(z)$ be given as in (3.6), and let*

$$\tilde{z} \in \operatorname{argmax}_{z \in \mathbb{R}} (\mathbf{z}_M^\top \mathbf{A} \mathbf{z}_N)^2 = \operatorname{argmax}_{z \in \mathbb{R}} \left(\frac{a(z)}{p(z)} \right)^2$$

denote an optimal structure parameter for the approximation of \mathbf{A} .

- (1) *If $h_\ell \geq 0$ for $\ell = 0, \dots, M+N-2$ and $h_0 \geq h_{M+N-2}$, then there exists a non-negative optimal parameter $\tilde{z} \geq 0$.*
- (2) *If $h_\ell \geq 0$ for ℓ even and $h_\ell \leq 0$ for ℓ odd, then there exists a non-positive optimal parameter $\tilde{z} \leq 0$.*

Proof. The first assertion follows directly from the observation that $a(z) \geq a(-z)$ for $z \geq 0$,

while $p(z) = p(-z)$ is an even function. In the second case, we have $a(-z) \geq a(z)$ for all $z \geq 0$, and the assertion follows similarly. \square

We illustrate the application of Theorem 3.1 together with the above results of Theorem 3.6 and Corollary 3.7 in the following small example.

Example 3.8 Consider the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \\ 1/2 & 0 & 1 \end{pmatrix}$$

with corresponding polynomials

$$a(z) = 1 + 3/2 \cdot z^2 + z^4 \quad \text{and} \quad p(z) = 1 + z^2 + z^4.$$

Inspecting $a(z)$ in light of Corollary 3.7, it is immediately clear that there is a non-negative optimal structure parameter.

In order to find precise optimal rank-1 Hankel approximations, we employ Theorems 3.1 and 3.6. First we form

$$\begin{aligned} Q(z) &= a'(z) \cdot p(z) - a(z) \cdot p'(z) \\ &= (3z + 4z^3) \cdot (1 + z^2 + z^4) - (1 + 3/2 \cdot z^2 + z^4) \cdot (2z + 4z^3) \\ &= z - z^5 = z \cdot (1 - z^4), \end{aligned}$$

whose roots we have to find according to Theorem 3.6. We identify the real roots $z = 0$, $z = 1$, $z = -1$ as candidates for optimal structure parameters.

By Theorem 3.1, the optimal structure parameters must satisfy $\tilde{z} \in \operatorname{argmax}_{z \in \mathbb{C}} a(z)/p(z)$, see also Corollary 3.7. So by

$$\frac{a(z)}{p(z)} = \frac{1 + 3/2 \cdot z^2 + z^4}{1 + z^2 + z^4} = \begin{cases} 1, & z = 0 \\ 7/6, & z = \pm 1 \end{cases}$$

we obtain $\tilde{z} = \pm 1$ as optimal structure parameters. Note that the existence of a negative optimal parameter $\tilde{z} < 0$ does not contradict Corollary 3.7.

Finally, the optimal coefficient is given by $\tilde{c} = a(\tilde{z})/p(\tilde{z}) = 7/6$ for both $\tilde{z} = 1$ and

$\tilde{z} = -1$. Thus we get the two real optimal approximation matrices

$$7/6 \cdot \tilde{\mathbf{z}}(1) \tilde{\mathbf{z}}(1)^\top = \frac{7}{18} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \quad \text{and} \quad 7/6 \cdot \tilde{\mathbf{z}}(-1) \tilde{\mathbf{z}}(-1)^\top = \frac{7}{18} \cdot \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \end{pmatrix}.$$

The approximation error for both is

$$\|\mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}} \tilde{\mathbf{z}}^\top\|_F = \left\| \frac{1}{18} \cdot \begin{pmatrix} 11 & \pm 7 & 2 \\ \pm 7 & 2 & \pm 7 \\ 2 & \pm 7 & 11 \end{pmatrix} \right\|_F = \frac{\sqrt{450}}{18} \approx 1.1785,$$

which is rounded to four digits.

For a comparison of Example 3.8 with the optimal rank-1 Hankel approximation with respect to the spectral norm see Section 9.1.

We affiliate some further results on special cases of the real r1H problem. In fact, often we have the case that $\mathbf{A} \in \mathbb{R}^{N \times N}$ is a real square matrix with non-negative components. If the coefficients h_ℓ of the corresponding polynomial $a(z)$ from (3.6) are non-negative and monotonically decreasing, then we find the structure parameter \tilde{z} —and thus generate the optimal rank-1 Hankel approximation—very easily.

Theorem 3.9 *Let $\mathbf{A} \in \mathbb{R}^{N \times N}$ with $\text{rank } \mathbf{A} \geq 1$. Let $a(z)$ be given as in (3.6) with non-negative coefficients $h_j \geq 0$ for $j = 0, \dots, 2N - 2$. Assume that the two sequences $(h_{2j})_{j=0}^{N-1}$ and $(h_{2j+1})_{j=0}^{N-2}$ are monotonically decreasing with $h_0 > h_{2N-2}$ and $h_1 > 0$.*

Then there is an optimal structure parameter $\tilde{z} \in \arg\max_{z \in \mathbb{R}} |F(z)|$ located in the open interval $(0, 1)$. Moreover, this $\tilde{z} \in (0, 1)$ is the only positive root of $Q(z) = a'(z)p(z) - a(z)p'(z)$. Note that here $p(z) = \mathbf{z}_N^\top \mathbf{z}_N = \sum_{j=0}^{N-1} z^{2j}$ is a polynomial.

Proof. By Theorem 3.6 the desired value \tilde{z} is a root of the polynomial $Q(z)$. Corollary 3.7 yields the existence of $\tilde{z} \geq 0$ maximizing the function $\left(\frac{a(z)}{p(z)}\right)^2$. To prove this theorem, it therefore suffices to show that $Q(z)$ possesses only one non-negative root, which is, more precisely, located in $(0, 1)$.

First, we observe that

$$Q(0) = a'(0)p(0) - a(0)p'(0) = h_1 > 0$$

since $p(0) = 1$ and $p'(0) = 0$ as well as $h_1 > 0$ by assumption.

Further, with $p(1) = N$ and $p'(1) = N(N - 1)$, we have

$$\begin{aligned}
 Q(1) &= a'(1)p(1) - a(1)p'(1) \\
 &= \sum_{j=0}^{2N-2} jh_j \cdot N - \sum_{j=0}^{2N-2} h_j \cdot N(N - 1) \\
 &= N \cdot \sum_{j=0}^{2N-2} (j - N + 1) \cdot h_j \\
 &= N \cdot \sum_{j=0}^{N-1} j \cdot (h_{N-1+j} - h_{N-1-j}) < 0,
 \end{aligned}$$

where we have used that both sequences $(h_{2j})_{j=0}^{N-1}$ and $(h_{2j+1})_{j=0}^{N-2}$ are monotonically decreasing by assumption. Furthermore, we have exploited the assumption that $h_0 > h_{2N-2}$. Thus, $Q(z)$ has at least one root inside the open interval $(0, 1)$.

In the next step, we show that $Q(z)$ possesses only this one positive root and no further ones. For this purpose, we will consider the polynomial $R(z) := Q(z) \cdot (1 - z^2)^2$.

Note that by the formula for the geometric partial sum, for any $z \in (0, 1)$, we have

$$p(z) = \sum_{j=0}^{N-1} z^{2j} = \frac{1 - z^{2N}}{1 - z^2},$$

and

$$p'(z) = \frac{(2N - 2) \cdot z^{2N+1} - 2N \cdot z^{2N-1} + 2z}{(1 - z^2)^2}$$

by the quotient rule. Thus, we obtain

$$\begin{aligned}
 R(z) &= Q(z) \cdot (1 - z^2)^2 \\
 &= (a'(z)p(z) - a(z)p'(z)) \cdot (1 - z^2)^2 \\
 &= \sum_{j=0}^{2N-2} jh_j \cdot z^{j-1} \cdot (1 - z^{2N}) \cdot (1 - z^2) \\
 &\quad - \sum_{j=0}^{2N-2} h_j \cdot z^j \cdot \left((2N - 2) \cdot z^{2N+1} - 2N \cdot z^{2N-1} + 2z \right)
 \end{aligned}$$

$$\begin{aligned}
 &= \sum_{j=0}^{2N-2} \left(j h_j \cdot z^{j-1} - (j+2) \cdot h_j \cdot z^{j+1} + (2N-j) \cdot h_j \cdot z^{2N-1+j} \right. \\
 &\quad \left. - (2N-2-j) \cdot h_j \cdot z^{2N+1+j} \right).
 \end{aligned}$$

An index shift $j' = j + 2$ in some of the summands yields

$$\begin{aligned}
 R(z) &= \sum_{j=0}^{2N-2} \left(j h_j \cdot z^{j-1} + (2N-j) \cdot h_j \cdot z^{2N-1+j} \right) \\
 &\quad - \sum_{j'=2}^{2N} \left(j' h_{j'-2} \cdot z^{j'-1} - (2N-j') \cdot h_{j'-2} \cdot z^{2N-1+j'} \right) \\
 &= \sum_{j=2}^{2N-2} \left(j \cdot (h_j - h_{j-2}) \cdot z^{j-1} \right) \\
 &\quad + h_1 - (2N-1) \cdot h_{2N-3} \cdot z^{2N-2} - 2N \cdot h_{2N-2} \cdot z^{2N-1} \\
 &\quad + \sum_{j=2}^{2N-2} \left((2N-j) \cdot (h_j - h_{j-2}) \cdot z^{2N-1+j} \right) \\
 &\quad + 2N \cdot h_0 \cdot z^{2N-1} + (2N-1) \cdot h_1 \cdot z^{2N} + h_{2N-3} \cdot z^{4N-2} \\
 &= h_1 \cdot z^0 + \sum_{j=2}^{2N-2} \left(j \cdot (h_j - h_{j-2}) \cdot z^{j-1} \right) - (2N-1) \cdot h_{2N-3} \cdot z^{2N-2} \\
 &\quad + 2N \cdot (h_0 - h_{2N-2}) \cdot z^{2N-1} + (2N-1) \cdot h_1 \cdot z^{2N} \\
 &\quad + \sum_{j=2}^{2N-2} \left((2N-j) \cdot (h_j - h_{j-2}) \cdot z^{2N-1+j} \right) - h_{2N-3} \cdot z^{4N-2},
 \end{aligned}$$

where finally, the summands are ordered according to the exponent of z .

By the assumptions on the polynomial coefficients h_j , we have $h_1 > 0$, $h_{2N-3} \geq 0$, and $(h_0 - h_{2N-2}) > 0$, as well as $(h_j - h_{j-2}) \leq 0$ for $j = 0, \dots, 2N-2$. Thus, the sequence of coefficients of the polynomial $R(z) = Q(z) \cdot (1 - z^2)^2$ exhibits exactly three changes of sign. In this case, the rule of Descartes [Hen74, Sec. 6.2] implies that $R(z)$ has either one or three positive real roots (counted according to their multiplicity). Since the polynomial factor $(1 - z^2)^2$ already has the positive root 1 with multiplicity two, we conclude that $Q(z)$ has exactly one positive real root. By the considerations in the beginning of this proof, this root is contained in the interval $(0, 1)$, and the proof is complete. \square

Remark 3.10 In the special case where Theorem 3.9 applies, the optimal parameter \tilde{z} can be found efficiently by employing a Newton type method, for example with starting value $z_0 = 0.5$.

A result similar to the one of Theorem 3.9 is obtained for increasing sequences of polynomial coefficients.

Corollary 3.11 *Let $\mathbf{A} \in \mathbb{R}^{N \times N}$ with $\text{rank } \mathbf{A} \geq 1$. Let $a(z)$ be given as in (3.6) with $h_\ell \geq 0$ for $\ell = 0, \dots, 2N - 2$. Assume that the two sequences $(h_{2\ell})_{\ell=0}^{N-1}$ and $(h_{2\ell+1})_{\ell=0}^{N-2}$ are monotonically increasing with $h_0 < h_{2N-2}$ and $h_{2N-3} > 0$.*

Then there is an optimal structure parameter $\tilde{z} \in \text{argmax}_{z \in \mathbb{R}} |F(z)|$ located in the interval $(1, \infty)$. Moreover \tilde{z} is the only positive root of $Q(z) = a'(z)p(z) - a(z)p'(z)$.

Proof. For $z \in (1, \infty)$ we observe that

$$\frac{a(1/z)}{p(1/z)} = \frac{a(1/z)}{z^{-2N+2} \cdot p(z)} = \frac{1}{p(z)} \cdot \sum_{j=0}^{2N-2} h_j z^{2N-2-j} =: \frac{\hat{a}(z)}{p(z)},$$

where the polynomial $\hat{a}(z)$ has the coefficients $\hat{h}_j := h_{2N-2-j}$, $j = 0, \dots, 2N - 2$. These are the same coefficients as the ones of $a(z)$ but in reverse order.

Thus, the newly defined sequences $(\hat{h}_{2j})_{j=0}^{N-1}$ and $(\hat{h}_{2j+1})_{j=0}^{N-2}$ are monotonically decreasing, and we have additionally $\hat{h}_0 > \hat{h}_{2N-2}$ and $\hat{h}_1 = h_{2N-3} > 0$. The assertion now follows from Theorem 3.9 applied to the polynomial $\hat{a}(z)$. \square

It is also possible to locate unique negative optimal parameters $\tilde{z} < 0$ in the fashion of Theorem 3.9 and Corollary 3.11. However, in order to achieve the upcoming results, the assumptions on the sequences of polynomial coefficients have to be slightly more complicated.

Corollary 3.12 *Let $\mathbf{A} \in \mathbb{R}^{N \times N}$ with $\text{rank } \mathbf{A} \geq 1$. Let $a(z)$ be given as in (3.6). Further, let $Q(z)$ be given as in Theorem 3.6.*

- (1) *Assume that $(h_{2j})_{j=0}^{N-1}$ is a non-negative, monotonically decreasing sequence with $h_0 > h_{2N-2} \geq 0$, and $(h_{2j+1})_{j=0}^{N-2}$ is a non-positive, monotonically increasing sequence with $h_0 < 0$. Then there exists $\tilde{z} \in (-1, 0)$ generating an optimal rank-1 Hankel approximation of \mathbf{A} . Moreover, this is the only negative root of $Q(z)$.*

(2) Assume that $(h_{2j})_{j=0}^{N-1}$ is a non-positive, monotonically decreasing sequence with $0 \geq h_0 > h_{2N-2}$, and $(h_{2j+1})_{j=0}^{N-2}$ is a non-negative, monotonically increasing sequence with $h_{2N-3} > 0$. Then there is $\tilde{z} \in (-\infty, -1)$ generating an optimal rank-1 Hankel approximation of \mathbf{A} . Moreover, this is the only negative root of $Q(z)$.

Proof. The idea is to apply Theorem 3.9 in order to prove the first assertion. To this end, define the auxiliary coefficients $\hat{h}_{2j+1} := -h_{2j+1}$ for $j = 0, \dots, N-2$. Then $(\hat{h}_{2j+1})_{j=0}^{N-1}$ is a non-negative, monotonically decreasing sequence with $\hat{h}_1 > 0$. Now, we can write

$$a(z) = a_0(z) + a_1(z) \quad \text{with} \quad a_0(z) = \sum_{j=0}^{N-1} h_{2j} z^{2j}, \quad a_1(z) = \sum_{j=0}^{N-2} h_{2j+1} z^{2j+1}$$

and introduce

$$\hat{a}(z) := a_0(z) - a_1(z) = \sum_{j=0}^{N-1} h_{2j} z^{2j} - \sum_{j=0}^{N-2} h_{2j+1} z^{2j+1} = \sum_{j=0}^{N-1} h_{2j} z^{2j} + \sum_{j=0}^{N-2} \hat{h}_{2j+1} z^{2j+1}.$$

Further, let $\hat{Q}(z) := \hat{a}'(z)p(z) - \hat{a}(z)p'(z)$. By Theorem 3.9, $\hat{Q}(z)$ possesses only one positive real root which is located in $(0, 1)$. For the polynomial $Q(z)$, it follows that

$$\begin{aligned} Q(z) &= a'(z) \cdot p(z) - a(z) \cdot p'(z) \\ &= (a'_0(z) + a'_1(z)) \cdot p(z) - (a_0(z) + a_1(z)) \cdot p'(z) \\ &= (-a'_0(-z) + a'_1(-z)) \cdot p(-z) - (a_0(-z) - a_1(-z)) \cdot (-p'(-z)) \\ &= -\hat{a}'(-z) \cdot p(-z) + \hat{a}(-z) \cdot p'(-z) \\ &= -\hat{Q}(-z), \end{aligned}$$

since both $a_0(z)$ and $p(z)$ are even polynomials while $a_1(z)$ is odd. Thus, the polynomial $Q(z)$ has exactly one negative root, which is located in $(-1, 0)$.

The second assertion follows analogously from Corollary 3.11. \square

We conclude this chapter by drawing a line back to complex rank-1 Hankel approximation with the following question: Given a real matrix \mathbf{A} , can we always restrict the search for the optimal parameters \tilde{c} and \tilde{z} to real numbers, or is it possible to achieve better results by allowing complex parameters? The following example shows that indeed complex parameters may provide better approximations.

Example 3.13 We want to find an optimal rank-1 Hankel approximation with respect to the Frobenius norm for the initial matrix

$$\mathbf{A} = \begin{pmatrix} 1 & -1/2 & -1 \\ -1/2 & -1 & -1/2 \\ -1 & -1/2 & 1 \end{pmatrix}.$$

Using Theorem 3.6 we find the solution for the optimal real parameters $\tilde{c}_{\text{real}} \approx 1.0635$ and $\tilde{z}_{\text{real}} \approx -0.1291$. The resulting Frobenius norm error is $\|\mathbf{A} - \tilde{c}_{\text{real}} \cdot \tilde{\mathbf{z}}_{\text{real}} \tilde{\mathbf{z}}_{\text{real}}^T\|_F \approx 2.2066$.

Allowing complex values, the somewhat nicely chosen parameters $c_{\text{complex}} = 5/3$ and $z_{\text{complex}} = \pm i$ yield the smaller error $\|\mathbf{A} - c_{\text{complex}} \cdot \mathbf{z}_{\text{complex}} \mathbf{z}_{\text{complex}}^T\|_F = \sqrt{261}/9 \approx 1.7951$. Actually, with an implementation of Theorem 3.4 we find two optimal pairs of complex parameters $\tilde{c}_{\text{complex}} \approx 1.5312 \pm 0.8472i$ and $\tilde{z}_{\text{complex}} \approx 0.2500 \pm 0.9682i$ to produce the minimal approximation error $\|\mathbf{A} - \tilde{c}_{\text{complex}} \cdot \tilde{\mathbf{z}}_{\text{complex}} \tilde{\mathbf{z}}_{\text{complex}}^T\|_F \approx 1.7139$. All parameter values and errors have been rounded to four digits.

We also run the comparison between real and complex rank-1 Hankel approximation on a broader basis. For this purpose, we randomly generate ten real initial matrices $\mathbf{A} \in \mathbb{R}^{10 \times 10}$ with entries in $[-50, 50]$. (The same matrices will be used for the comparison of different r1H methods in Section 9.2.)

For each of these initial matrices we solve the r1H problem once for real parameters c and z , and once allowing them to be complex. Then we compute the relative approximation error $\text{RE} := \|\mathbf{A} - \mathbf{H}\|_F / \|\mathbf{A}\|_F$ for each initial matrix and both (real and complex) approximations. These relative approximation errors are depicted in Figure 3.1 for each initial matrix. We conclude that Example 3.13 is not an isolated case. Indeed, the complex rank-1 Hankel approximation frequently outperforms the purely real one.

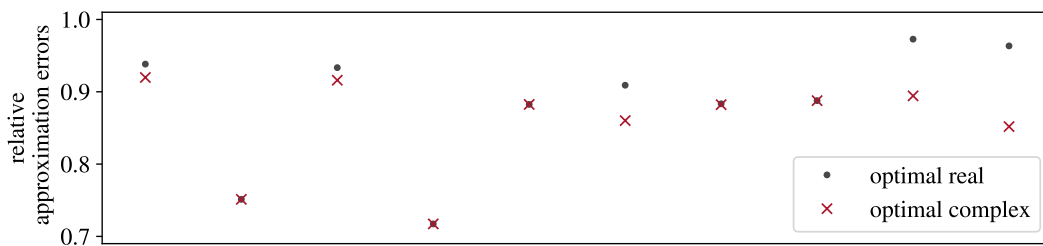


Figure 3.1 Optimal r1H errors for only real and allowing complex parameters.

Remark 3.14 For matrices \mathbf{A} with non-negative entries satisfying $a_{0,0} \geq a_{M-1,N-1}$ there is always an optimal rank-1 Hankel approximation with real non-negative parameters \tilde{c} and \tilde{z} . Consider the polynomial $a(z)$ from (3.6), which, in this case, has only non-negative coefficients h_ℓ . Thus, we obtain

$$|\mathbf{z}_M^\top \mathbf{A} \mathbf{z}_N| = \frac{|a(z)|}{p(|z|)} = \frac{|\sum_{\ell=0}^{M+N-2} h_\ell z^\ell|}{p(|z|)} \leq \frac{\sum_{\ell=0}^{M+N-2} h_\ell |z|^\ell}{p(|z|)}$$

with $p(z)$ as in Theorem 3.6. This term can be maximized by a real value $\tilde{z} \geq 0$. The optimal coefficient $\tilde{c} = \tilde{\mathbf{z}}_M^\top \mathbf{A} \tilde{\mathbf{z}}_N$ is consequently also real and non-negative.

4

OPTIMAL RANK-1 HANKEL APPROXIMATION IN THE SPECTRAL NORM

In this chapter we deal with the rank-1 Hankel approximation (r1H) problem (2) in the spectral norm:

$$\min \|\mathbf{A} - \mathbf{H}\|_2 \quad \text{such that } \text{rank } \mathbf{H} = 1, \\ \text{and } \mathbf{H} \text{ has Hankel structure.}$$

Unless indicated otherwise, this chapter is based on our paper [KPP21a, Sec. 4] and is in parts similar with the representations therein.

By Definition 1.6 it is apparent that the spectral norm of a matrix usually cannot be expressed in terms of the matrix' entries. This makes the minimization problem (2) for the spectral norm much more delicate than for the Frobenius norm. In order to still obtain some results for the spectral norm, we restrict our considerations to real symmetric matrices and their real optimal rank-1 Hankel approximations.

Lemma 2.3 suggests that for a symmetric matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ we can equivalently write the real r1H problem as

$$\min_{\substack{z \in \overline{\mathbb{R}} \\ c \in \mathbb{R} \setminus \{0\}}} \|\mathbf{A} - c \cdot \mathbf{z}\mathbf{z}^T\|_2, \quad (4.1)$$

with $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ and $\mathbf{z} = \mathbf{z}_N(z)$ as in (2.5).

As a real symmetric matrix, \mathbf{A} possesses an eigendecomposition $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$. In this decomposition, $\mathbf{V} = (\mathbf{v}_0 \cdots \mathbf{v}_{N-1})$ is the orthogonal matrix whose columns \mathbf{v}_j are

the normalized eigenvectors of \mathbf{A} . The diagonal matrix $\mathbf{\Lambda} = \text{diag}(\lambda_0, \dots, \lambda_{N-1})$ contains the corresponding eigenvalues of \mathbf{A} , which are ordered by absolute value $|\lambda_0| \geq |\lambda_1| \geq \dots \geq |\lambda_{N-1}| \geq 0$ largest to smallest. Without loss of generality, we assume $\lambda_0 = |\lambda_0| > 0$. We have the correspondence $\mathbf{A}\mathbf{v}_j = \lambda_j\mathbf{v}_j$, for $j = 0, \dots, N-1$, between eigenvalues and eigenvectors.

With the matrix \mathbf{V} we can transfer the r1H problem (4.1) into the eigenbasis of \mathbf{A} . To this end let $\boldsymbol{\mu}(z) := \mathbf{V}^\top \mathbf{z}(z) = (\mu_0 \cdots \mu_{N-1})^\top$. In other words,

$$\mathbf{z}(z) = \mathbf{V}\boldsymbol{\mu}(z) = \sum_{j=0}^{N-1} \mu_j \mathbf{v}_j, \quad \text{with } \mu_j = \mathbf{v}_j^\top \mathbf{z}, \quad j = 0, \dots, N-1,$$

is the representation of $\mathbf{z}(z)$ in the orthogonal eigenbasis $\{\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_{N-1}\}$ of \mathbf{A} . We simply write $\boldsymbol{\mu}$ instead of $\boldsymbol{\mu}(z)$ when no confusion about the structure parameter z is risked.

With $\boldsymbol{\mu}$ introduced, now problem (4.1) can be rewritten as

$$\begin{aligned} \min_{\substack{z \in \overline{\mathbb{R}} \\ c \in \mathbb{R} \setminus \{0\}}} \|\mathbf{A} - c \cdot \mathbf{z}\mathbf{z}^\top\|_2 &= \min_{\substack{z \in \overline{\mathbb{R}} \\ c \in \mathbb{R} \setminus \{0\}}} \|\mathbf{V}^\top \mathbf{A} \mathbf{V} - c \cdot \mathbf{V}^\top \mathbf{z}\mathbf{z}^\top \mathbf{V}\|_2 \\ &= \min_{\substack{z \in \overline{\mathbb{R}} \\ c \in \mathbb{R} \setminus \{0\}}} \|\mathbf{\Lambda} - c \cdot \boldsymbol{\mu}\boldsymbol{\mu}^\top\|_2. \end{aligned} \quad (4.2)$$

The first equality of the above is true because the spectral norm is invariant under orthogonal transformation.

Our goal is to find necessary and sufficient conditions for optimal parameters \tilde{c} and \tilde{z} such that the matrix $\tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top$ solves the minimization problem (4.1), respectively (4.2). For the following considerations, we assume that such a minimizer exists. The minimizer will however not always be unique. Let now $\tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top$ denote such an optimal solution and let

$$\tilde{\lambda} := \tilde{\lambda}_{\tilde{c}, \tilde{z}} := \|\mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top\|_2 = \|\mathbf{\Lambda} - \tilde{c} \cdot \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^\top\|_2 \quad (4.3)$$

denote the corresponding optimal approximation error. Here, $\tilde{\boldsymbol{\mu}} = \mathbf{V}^\top \tilde{\mathbf{z}}$ is the transformed structured vector corresponding to the optimal parameter \tilde{z} .

Recall that the spectral norm of a real symmetric matrix is equal to the modulus of its largest eigenvalue, see also Definition 1.6. Therefore, by (4.3) either $\tilde{\lambda}$ is the largest or $-\tilde{\lambda}$ is the smallest eigenvalue of the difference matrix $\mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top$, respectively $\mathbf{\Lambda} - \tilde{c} \cdot \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^\top$. Note that the largest eigenvalue of the difference matrix strongly depends on the parameters \tilde{c}

and \tilde{z} . This is indicated by the subscript in the above definition (4.3) of $\tilde{\lambda}$. We tend to omit the subscript since usually there is no risk of confusion.

Note that we can shift the eigenvalues of a symmetric matrix by adding a multiple of the identity. This can be seen by

$$\mathbf{V}^\top(\mathbf{A} + \lambda\mathbf{I})\mathbf{V} = \mathbf{V}^\top\mathbf{A}\mathbf{V} + \lambda\mathbf{V}^\top\mathbf{V} = \mathbf{\Lambda} + \lambda\mathbf{I} = \text{diag}(\lambda_0 + \lambda, \dots, \lambda_{N-1} + \lambda).$$

In light of this fact we introduce the two auxiliary matrices

$$\mathbf{M}_1(\lambda, c, z) := \lambda\mathbf{I} - \mathbf{\Lambda} + c \cdot \boldsymbol{\mu}(z)\boldsymbol{\mu}(z)^\top \quad (4.4a)$$

and

$$\mathbf{M}_2(\lambda, c, z) := \lambda\mathbf{I} + \mathbf{\Lambda} - c \cdot \boldsymbol{\mu}(z)\boldsymbol{\mu}(z)^\top. \quad (4.4b)$$

These are exactly the difference matrix $\mathbf{\Lambda} - c \cdot \boldsymbol{\mu}\boldsymbol{\mu}^\top$ where $-\lambda\mathbf{I}$ and $\lambda\mathbf{I}$ have been added, respectively. In the case of $\mathbf{M}_1(\lambda, c, z)$, the shifted difference matrix has additionally been multiplied by -1 .

Of special interest for us is the case when the optimal difference matrix in (4.3) is shifted by its largest eigenvalue $\tilde{\lambda}$. More precisely, we are especially interested in the matrices

$$\mathbf{M}_1(\tilde{\lambda}) := \mathbf{M}_1(\tilde{\lambda}, \tilde{c}, \tilde{z}) = \tilde{\lambda}\mathbf{I} - \mathbf{\Lambda} + \tilde{c} \cdot \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^\top \quad (4.5a)$$

and

$$\mathbf{M}_2(\tilde{\lambda}) := \mathbf{M}_2(\tilde{\lambda}, \tilde{c}, \tilde{z}) = \tilde{\lambda}\mathbf{I} + \mathbf{\Lambda} - \tilde{c} \cdot \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^\top. \quad (4.5b)$$

For simplicity we usually omit the optimal parameters \tilde{c} and \tilde{z} when referring to $\mathbf{M}_1(\tilde{\lambda}, \tilde{c}, \tilde{z})$ and $\mathbf{M}_2(\tilde{\lambda}, \tilde{c}, \tilde{z})$. Nevertheless, keep in mind that the optimal error $\tilde{\lambda} = \tilde{\lambda}_{\tilde{c}, \tilde{z}}$ depends on the optimal parameters. Thus the auxiliary matrices $\mathbf{M}_1(\tilde{\lambda})$ and $\mathbf{M}_2(\tilde{\lambda})$ depend on the optimal parameters \tilde{c} and \tilde{z} through $\tilde{\lambda}$.

Since $\tilde{\lambda}$ is the modulus of the largest eigenvalue of the difference matrix $\mathbf{\Lambda} - \tilde{c} \cdot \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^\top$, both $\mathbf{M}_1(\tilde{\lambda})$ and $\mathbf{M}_2(\tilde{\lambda})$ are positive semidefinite. If $\tilde{\lambda}$ is the largest eigenvalue of the difference matrix, then $\mathbf{M}_1(\tilde{\lambda})$ actually possesses the eigenvalue zero. On the other hand, if $-\tilde{\lambda}$ is the smallest eigenvalue of the difference matrix, then the smallest eigenvalue of $\mathbf{M}_2(\tilde{\lambda})$ is zero. By construction (4.3) at least one of these cases occurs. Therefore, (4.3) is equivalent to both $\mathbf{M}_1(\tilde{\lambda})$ and $\mathbf{M}_2(\tilde{\lambda})$ being positive semidefinite and at least one of them actually possessing the eigenvalue zero.

Note that the auxiliary matrices \mathbf{M}_1 and \mathbf{M}_2 have a special structure; namely, they can be decomposed into the sum of a diagonal matrix and a symmetric rank-1 matrix. For this reason we will investigate conditions for the definiteness of such matrices in detail in the next section.

4.1 Definiteness of Diagonal-Plus-Rank-1 Matrices

For this section, let \mathbf{B} be the sum of a diagonal matrix $\mathbf{D} = \text{diag}(d_0, \dots, d_{N-1}) \in \mathbb{R}^{N \times N}$ and a symmetric rank-1 matrix; that is, let

$$\mathbf{B} := \mathbf{D} + c \cdot \mathbf{b}\mathbf{b}^\top,$$

where $\mathbf{b} = (b_0 \dots b_{N-1})^\top \in \mathbb{R}^N$ and $c \in \mathbb{R}$.

Before diving deeper into the definiteness of \mathbf{B} , we start by giving the following observation on its determinant.

Lemma 4.1 *The determinant of the matrix $\mathbf{B} = \mathbf{D} + c \cdot \mathbf{b}\mathbf{b}^\top$ is given by*

$$\det \mathbf{B} = \det \mathbf{D} + c \cdot \sum_{j=0}^{N-1} b_j^2 \cdot \left(\prod_{\substack{k=0 \\ k \neq j}}^{N-1} d_k \right).$$

If, additionally, the diagonal matrix \mathbf{D} is invertible, we have

$$\det \mathbf{B} = \det \mathbf{D} \cdot \left(1 + c \cdot \sum_{j=0}^{N-1} \frac{b_j^2}{d_j} \right).$$

Proof. We employ the rule for computing determinants of block matrices [Sil00],

$$\det \begin{pmatrix} \mathbf{D} & -\mathbf{b} \\ c \cdot \mathbf{b}^\top & 1 \end{pmatrix} = \det(1 \cdot \mathbf{D} + c \cdot \mathbf{b}\mathbf{b}^\top) = \det \mathbf{B}.$$

Then we expand the determinant on the left with respect to the last column in order to obtain the claimed formula. For the case where \mathbf{D} is invertible, see also [Dem97]. \square

Remark 4.2 The above lemma also works if the rank-1 matrix is not symmetric. Let $\mathbf{B} = \mathbf{D} + c \cdot \mathbf{a}\mathbf{b}^\top$, with $\mathbf{a} = (a_0 \dots a_{N-1})^\top \in \mathbb{R}^N$, and $\mathbf{b} = (b_0 \dots b_{N-1})^\top \in \mathbb{R}^N$

as before. In this case, we have

$$\det \mathbf{B} = \det \mathbf{D} + c \cdot \sum_{j=0}^{N-1} a_j b_j \cdot \left(\prod_{\substack{k=0 \\ k \neq j}}^{N-1} d_k \right).$$

Recall that we are interested in the two auxiliary matrices $\mathbf{M}_1(\tilde{\lambda})$ and $\mathbf{M}_2(\tilde{\lambda})$ from (4.5). Their diagonal parts are

$$\tilde{\lambda} \mathbf{I} - \mathbf{\Lambda} \quad \text{and} \quad \tilde{\lambda} \mathbf{I} + \mathbf{\Lambda},$$

respectively. As will become clear later in Proposition 4.6, there are exactly two different types of these diagonal matrices. Either

- the diagonal part is positive semidefinite, i.e., \mathbf{D} has only non-negative entries, or
- the diagonal part has exactly one negative entry while all the other entries are non-negative.

Therefore, we examine positive definiteness of the matrix $\mathbf{B} = \mathbf{D} + c \cdot \mathbf{b}\mathbf{b}^\top$ for exactly these two types of diagonal parts. For each of the two cases, we give necessary and sufficient conditions on c and \mathbf{b} such that the matrix \mathbf{B} is positive semidefinite, see Lemmas 4.3 and 4.4 correspondingly.

We need to fix some notation first. For the remainder of this chapter, in the case $b_j = d_j = 0$, we just omit the term b_j^2/d_j whenever it would appear. This omission is consistent with the widely used convention $\frac{0}{0} = 0$. To remind the reader that such terms may occur in a sum, we will use the notation \sum' instead of \sum .

Lemma 4.3 *Let $c > 0$ and assume that \mathbf{D} has exactly one negative eigenvalue, say $d_0 < 0$ and $d_1, \dots, d_{N-1} \geq 0$. Then the matrix $\mathbf{B} = \mathbf{D} + c \cdot \mathbf{b}\mathbf{b}^\top$ is positive semidefinite if and only if the vector \mathbf{b} and coefficient c satisfy*

$$\sum_{j=0}^{N-1} \frac{b_j^2}{(-d_j)} \geq \frac{1}{c}, \quad (4.6)$$

where $b_j = 0$ whenever $d_j = 0$. Moreover, if $d_j > 0$ for $j = 1, \dots, N-1$, and the inequality (4.6) is strict, then \mathbf{B} is in fact positive definite.

Proof. According to [Pru86], a matrix is positive semidefinite if and only if all its principal minors (i.e., the determinants of all possible $(r \times r)$ principal submatrices for $r = 1, \dots, N$)

are non-negative. We observe that all the principal submatrices of \mathbf{B} are of the same form as \mathbf{B} itself.

Now consider the (2×2) submatrices $\mathbf{B}_{\{0,j\}}$ obtained from the rows and columns of \mathbf{B} with indices 0 and j , where the corresponding $d_j = 0$. By Lemma 4.1, and accounting for all the vanishing terms (or by direct computation), its determinant is given by

$$\det \mathbf{B}_{\{0,j\}} = \det \begin{pmatrix} d_0 + c \cdot b_0^2 & c \cdot b_0 b_j \\ c \cdot b_0 b_j & c \cdot b_j^2 \end{pmatrix} = c \cdot b_j^2 \cdot d_0.$$

Since $d_0 < 0$, this determinant is negative as long as $b_j \neq 0$. Hence, it is imperative that $b_j = 0$ for all indices j with $d_j = 0$ in order for \mathbf{B} to be positive semidefinite.

Let J be the index set containing all indices corresponding to non-zero entries of \mathbf{D} . Observe that J is non-empty because $0 \in J$ since $d_0 < 0$ by assumption. Denote by \mathbf{B}_J and \mathbf{D}_J the corresponding principal submatrices of \mathbf{B} and \mathbf{D} , respectively. By Lemma 4.1, we have

$$\det \mathbf{B}_J = \det \mathbf{D}_J \cdot \left(1 + c \cdot \sum_{j \in J} \frac{b_j^2}{d_j} \right),$$

where $\det \mathbf{D}_J < 0$. Thus, $\det \mathbf{B}_J$ being non-negative is equivalent to

$$\left(1 + c \cdot \sum_{j \in J} \frac{b_j^2}{d_j} \right) \leq 0 \quad \Leftrightarrow \quad \sum_{j \in J} \frac{b_j^2}{(-d_j)} \geq \frac{1}{c}.$$

Clearly, these conditions are already sufficient for all principal minors of \mathbf{B} corresponding to subsets of J to be non-negative. By adding the zero terms corresponding to indices not in J to the above inequality, the first claim follows.

If $d_j > 0$ for $j = 1, \dots, N-1$, and the above inequality for $J = \{0, 1, \dots, N-1\}$ is strict, then already all leading principal minors of \mathbf{B} are strictly positive. It is a well-known fact that this is equivalent to \mathbf{B} being positive definite. \square

Lemma 4.4 *Let $c > 0$, and assume that \mathbf{D} is positive semidefinite with at least one strictly positive eigenvalue, say $d_0 > 0$ and $d_1, \dots, d_{N-1} \geq 0$. Then the matrix $\mathbf{B} = \mathbf{D} - c \cdot \mathbf{b}\mathbf{b}^\top$ is positive semidefinite if and only if the vector \mathbf{b} and coefficient c satisfy*

$$\sum_{j=0}^{N-1} \frac{b_j^2}{d_j} \leq \frac{1}{c}, \tag{4.7}$$

where $b_j = 0$ whenever $d_j = 0$. Moreover, if $d_j > 0$ for all $j = 0, \dots, N - 1$ and the inequality (4.7) is strict, then \mathbf{B} is in fact positive definite.

Proof. Similarly as in the proof of Lemma 4.3, we consider the principal minors [Pru86] of \mathbf{B} . First, we examine at the (2×2) submatrices $\mathbf{B}_{\{0,j\}}$ obtained from the rows and columns of \mathbf{B} with indices 0 and j , where the corresponding $d_j = 0$. By Lemma 4.1, and accounting for all the vanishing terms (or by direct computation), its determinant is given by

$$\det \mathbf{B}_{\{0,j\}} = \det \begin{pmatrix} d_0 - c \cdot b_0^2 & -c \cdot b_0 b_j \\ -c \cdot b_0 b_j & -c \cdot b_j^2 \end{pmatrix} = -c \cdot b_j^2 \cdot d_0,$$

which is negative unless $b_j = 0$. Hence, we conclude as before, that necessarily $b_j = 0$ whenever $d_j = 0$.

Let again J be the index set containing all indices corresponding to non-zero entries of \mathbf{D} , and denote by \mathbf{B}_J and \mathbf{D}_J the corresponding principal submatrices of \mathbf{B} and \mathbf{D} , respectively. By Lemma 4.1, we have

$$\det \mathbf{B}_J = \det \mathbf{D}_J \cdot \left(1 - c \cdot \sum_{j \in J} \frac{b_j^2}{d_j} \right),$$

where $\det \mathbf{D}_J > 0$. Thus, $\det \mathbf{B}_J$ being non-negative is equivalent to

$$\left(1 - c \cdot \sum_{j \in J} \frac{b_j^2}{d_j} \right) \geq 0 \quad \Leftrightarrow \quad \sum_{j \in J} \frac{b_j^2}{d_j} \leq \frac{1}{c}.$$

These conditions are already sufficient for all principal minors of \mathbf{B} to be non-negative. By adding the zero terms corresponding to indices not in J to the above inequality, the first claim follows.

If $d_j > 0$ for all $j = 0, \dots, N - 1$ and the above inequality is strict, then all leading principal minors are positive and thus \mathbf{B} is positive definite. \square

With these Lemmas we can now come back to our original problem, namely, optimal rank-1 Hankel approximation in the spectral norm. In the next section we closely examine the optimal approximation error. Ensuing from there, we derive conditions on the optimal parameters \tilde{c} and \tilde{z} .

4.2 The Optimal Approximation Error

The optimal parameters with respect to the Frobenius norm from Chapter 3 can be calculated independently of the approximation error, see Theorem 3.1. In contrast, the spectral norm error is very much interlinked with the parameters c and z constituting a rank-1 Hankel approximation. As a consequence, the optimal parameters \tilde{c} and \tilde{z} with respect to the spectral norm cannot be calculated without computing the optimal approximation error $\tilde{\lambda}$ at the same time. Therefore, we examine the optimal spectral norm error more closely. Only the precise understanding of its behavior enables us to find conditions on the optimal parameters \tilde{c} and \tilde{z} .

First, in Section 4.2.1, we assume that the by modulus largest eigenvalue of \mathbf{A} is strictly larger than the by modulus second largest eigenvalue. The case where several eigenvalues have maximal absolute value is treated separately in Section 4.2.2.

4.2.1 Isolated Largest Eigenvalue

In this section, let $\mathbf{A} \in \mathbb{R}^{N \times N}$ have an isolated largest eigenvalue; that is, we assume $\lambda_0 = \|\mathbf{A}\|_2 > |\lambda_1|$. Recall that, without loss of generality, we assume $\lambda_0 > 0$ for all of this chapter. When the largest eigenvalue is isolated, then we will see that we can always find an optimal rank-1 Hankel approximation of \mathbf{A} with respect to the spectral norm.

We start by presenting upper and lower bounds for the resulting optimal approximation error. For the proof of these bounds we need the following lemma.

Lemma 4.5 *Let some values $\lambda > 0$, $c \in \mathbb{R} \setminus \{0\}$, and $z \in \overline{\mathbb{R}}$ be given. If both auxiliary matrices $\mathbf{M}_1(\lambda, c, z)$ and $\mathbf{M}_2(\lambda, c, z)$ from (4.4) are positive definite, then λ is larger than the spectral norm of the difference matrix $\mathbf{A} - c \cdot \boldsymbol{\mu}\boldsymbol{\mu}^\top$; that is, we have $\lambda > \|\mathbf{A} - c \cdot \boldsymbol{\mu}\boldsymbol{\mu}^\top\|_2$.*

Proof. Denote the eigenvalues of the difference matrix $\mathbf{A} - c \cdot \boldsymbol{\mu}\boldsymbol{\mu}^\top$ by η_i , $i = 0, \dots, N-1$, ordered by absolute value. Thus, we have $|\eta_0| = \|\mathbf{A} - c \cdot \boldsymbol{\mu}\boldsymbol{\mu}^\top\|_2$.

We prove this lemma by contraposition. Assume that $|\eta_0| \geq \lambda$. In the eigenbasis of the difference matrix, $\mathbf{A} - c \cdot \boldsymbol{\mu}\boldsymbol{\mu}^\top$, the auxiliary matrix $\mathbf{M}_1(\lambda, c, z) = \lambda\mathbf{I} - \mathbf{A} + c \cdot \boldsymbol{\mu}\boldsymbol{\mu}^\top$ becomes $\lambda\mathbf{I} - \text{diag}(\eta_i)_{i=0}^{N-1}$. Analogously, $\mathbf{M}_2(\lambda, c, z) = \lambda\mathbf{I} + \mathbf{A} - c \cdot \boldsymbol{\mu}\boldsymbol{\mu}^\top$ becomes $\lambda\mathbf{I} + \text{diag}(\eta_i)_{i=0}^{N-1}$.

Now we distinguish two cases based on the sign of η_0 . If $\eta_0 > 0$, then $\lambda - \eta_0 \leq 0$ and thus $\mathbf{M}_1(\lambda, c, z)$ is not positive definite. On the other hand, if $\eta_0 < 0$, then $\lambda + \eta_0 \leq 0$ and thus $\mathbf{M}_2(\lambda, c, z)$ is not positive definite. This proves the claim. \square

Proposition 4.6 *Let $\tilde{\lambda} = \|\mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top\|_2$ be the optimal r1H error (4.3). Then we have*

$$|\lambda_1| \leq \tilde{\lambda} < \lambda_0,$$

where λ_0 and λ_1 are the largest and (by modulus) second largest eigenvalue of \mathbf{A} , respectively.

Proof. By the Eckart-Young-Mirsky Theorem (Theorem 1.8), the lower bound $|\lambda_1| \leq \tilde{\lambda}$ follows immediately.

Consider now some $\lambda \geq \lambda_0$. We show that we can always find parameters c and z such that both $\mathbf{M}_1(\lambda, c, z)$ and $\mathbf{M}_2(\lambda, c, z)$ in (4.4) are strictly positive definite. Then, by combining Lemma 4.5 and the definition (4.3) of the optimal approximation error $\tilde{\lambda}$, we obtain $\tilde{\lambda} \leq \|\mathbf{A} - c \cdot \boldsymbol{\mu}\boldsymbol{\mu}^\top\|_2 < \lambda$.

For instance, choose some $z \in \overline{\mathbb{R}}$ such that $\mu_0 = \mathbf{v}_0^\top \mathbf{z} \neq 0$. Further set $c = \frac{\lambda - |\lambda_1|}{2} > 0$. Then, on the one hand, both the diagonal part $\lambda \mathbf{I} - \mathbf{A}$ and the rank-1 part $c \cdot \boldsymbol{\mu}\boldsymbol{\mu}^\top$ of the first auxiliary matrix $\mathbf{M}_1(\lambda)$ are positive semidefinite. As a sum of two positive semidefinite matrices, $\mathbf{M}_1(\lambda)$ is itself positive semidefinite. Further, Lemma 4.1 yields

$$\begin{aligned} \det \mathbf{M}_1(\lambda) &= \det \left(\text{diag}(\lambda - \lambda_j)_{j=0}^{N-1} \right) + c \cdot \sum_{j=0}^{N-1} \mu_j^2 \cdot \left(\prod_{\substack{k=0 \\ k \neq j}}^{N-1} (\lambda - \lambda_k) \right) \\ &\geq 0 + c \cdot \mu_0^2 \cdot \prod_{k=1}^{N-1} (\lambda - \lambda_k) > 0, \end{aligned}$$

thus $\mathbf{M}_1(\lambda)$ is indeed strictly positive definite.

On the other hand, the diagonal part $\lambda \mathbf{I} + \mathbf{A}$ of the second auxiliary matrix $\mathbf{M}_2(\lambda)$ is positive definite with smallest possible eigenvalue $\lambda - |\lambda_1| > 0$. Since $\lambda - |\lambda_1| \leq \lambda + \lambda_j$ for all $j = 0, \dots, N-1$, we find

$$\sum_{j=0}^{N-1} \frac{\mu_j^2}{\lambda + \lambda_j} \leq (\lambda - |\lambda_1|)^{-1} \cdot \sum_{j=0}^{N-1} \mu_j^2 = \frac{\|\boldsymbol{\mu}\|_2^2}{\lambda - |\lambda_1|} < \frac{2}{\lambda - |\lambda_1|} = \frac{1}{c},$$

where the last inequality holds because $\|\boldsymbol{\mu}\|_2 = \|\mathbf{z}\|_2 = 1$. Thus by Lemma 4.4, the second auxiliary matrix $\mathbf{M}_2(\lambda)$ is strictly positive definite, too.

Hence, the optimal error $\tilde{\lambda}$ is bounded from below and above by $|\lambda_1| \leq \tilde{\lambda} < \lambda_0$. \square

Remark 4.7 In the above proposition, for $\lambda \geq \lambda_0$ and $z \in \overline{\mathbb{R}}$ satisfying $\mu_0 = \mathbf{v}_0^\top \mathbf{z} \neq 0$,

the coefficient c can in fact be chosen such that $0 < c < \left(\sum_{j=0}^{N-1} \frac{\mu_j^2}{\lambda + \lambda_j} \right)^{-1}$. This range is sufficient to ensure that both $\mathbf{M}_1(\lambda)$ and $\mathbf{M}_2(\lambda)$ are strictly positive definite according to Lemmas 4.1 and 4.4.

In order to state the main theorem of this chapter, we introduce the following function. Recalling that $\lim_{z \rightarrow \infty} \mathbf{z}(z) = \mathbf{z}(\infty) = \mathbf{e}_N$, let

$$f(z, \lambda^2) := \sum'_{j=0}^{N-1} \frac{(\mathbf{v}_j^\top \mathbf{z})^2}{\lambda_j^2 - \lambda^2} = \sum'_{j=0}^{N-1} \frac{\mu_j^2}{\lambda_j^2 - \lambda^2} \quad (4.8)$$

for $z \in \overline{\mathbb{R}}$ and $\lambda^2 \in [\lambda_1^2, \lambda_0^2)$. The range of λ^2 is chosen according to Proposition 4.6.

Note that for $\lambda^2 = \lambda_1^2$ the function f in (4.8) is only well-defined if $\mathbf{v}_j^\top \mathbf{z} = \mu_j = 0$ for all j with $\lambda_j^2 = \lambda_1^2$. This is in analogy to Lemmas 4.3 and 4.4, and we keep using the notation \sum' introduced in Section 4.1.

Remark 4.8 As stated above, for any fixed $z \in \overline{\mathbb{R}}$,

$$f(z, \lambda_1^2) = \lim_{\lambda^2 \rightarrow \lambda_1^2} f(z, \lambda^2)$$

is bounded if and only if $\mathbf{v}_j^\top \mathbf{z} = \mu_j = 0$ for all j with $\lambda_j^2 = \lambda_1^2$. Similarly, for any fixed $z \in \overline{\mathbb{R}}$,

$$f(z, \lambda_0^2) = \lim_{\lambda^2 \rightarrow \lambda_0^2} f(z, \lambda^2)$$

is bounded if and only if $\mathbf{v}_0^\top \mathbf{z} = \mu_0 = 0$. The conditions for boundedness of $f(z, \lambda_1^2)$ and $f(z, \lambda_0^2)$ are satisfied in our setting. So we can extend the domain of f to the closed interval $\lambda^2 \in [\lambda_1^2, \lambda_0^2]$, using the convention $\frac{0}{0} = 0$ and the notation \sum' as before.

In the case when λ is confined to the open interval $\lambda^2 \in (\lambda_1^2, \lambda_0^2)$ we can give explicit upper and lower bounds for the value of f . First, note that for $\lambda^2 \in (\lambda_1^2, \lambda_0^2)$, the matrix $\mathbf{A}^2 - \lambda^2 \mathbf{I}$ is invertible, and its inverse has the same eigenvectors as \mathbf{A} and reciprocal eigenvalues, in formulas that is, $(\mathbf{A}^2 - \lambda^2 \mathbf{I})^{-1} \cdot \mathbf{v}_j = (\lambda_j^2 - \lambda^2)^{-1} \cdot \mathbf{v}_j$. Recalling the representation $\mathbf{z} = \sum_{j=0}^{N-1} \mu_j \cdot \mathbf{v}_j$ in the basis of eigenvectors, we find

$$f(z, \lambda^2) = \sum_{j=0}^{N-1} \frac{\mu_j^2}{\lambda_j^2 - \lambda^2} = \mathbf{z}^\top \cdot (\mathbf{A}^2 - \lambda^2 \mathbf{I})^{-1} \cdot \mathbf{z},$$

that means f can be seen as a Rayleigh quotient for the matrix $(\mathbf{A}^2 - \lambda^2 \mathbf{I})^{-1}$. Therefore, f

is bounded by the largest and smallest eigenvalue of $(\mathbf{A}^2 - \lambda^2 \mathbf{I})^{-1}$

$$\min \{(\lambda^2 - \lambda_{N-1}^2)^{-1}, (\lambda_0^2 - \lambda^2)^{-1}\} \leq f(z, \lambda^2) \leq \max \{(\lambda^2 - \lambda_{N-1}^2)^{-1}, (\lambda_0^2 - \lambda^2)^{-1}\},$$

see [HJ13, Sec. 4.2].

The upcoming main theorem consists of two parts. In the first part we specify necessary and sufficient conditions to ensure that the optimal rank-1 Hankel approximation achieves the same error as the unstructured rank-1 approximation. By Theorem 1.8 and Proposition 4.6, this optimal error bound is given by $|\lambda_1|$, see also (1.3).

On that account, part (1) of Theorem 4.9 constitutes a generalization of the result from [Ant97]. The statement from [Ant97] is exactly Theorem 4.9 (1) but with the assumption that \mathbf{A} be a real Hankel matrix whose largest and second largest eigenvalue occur with multiplicity one. We relax these assumptions to \mathbf{A} being a more general real symmetric matrix whose largest eigenvalue only is restricted in its multiplicity. The latter restriction can even be lifted, see Theorem 4.15.

The second part of the theorem deals with the case when the optimal error bound $|\lambda_1|$ from (1.3) cannot be attained. We characterize the parameters that constitute the optimal rank-1 Hankel approximation in this case as well.

Theorem 4.9 enables us to develop an algorithm to compute an optimal rank-1 Hankel approximation numerically, see Section 4.3 and the algorithms therein.

Theorem 4.9 *Let $\mathbf{A} \in \mathbb{R}^{N \times N}$ be symmetric with $\text{rank}(\mathbf{A}) > 1$. Assume that $\lambda_0 = \|\mathbf{A}\|_2 > |\lambda_1|$. Let $\tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top$ be an optimal rank-1 Hankel approximation of \mathbf{A} in the spectral norm. Further, let f be defined as in (4.8).*

(1) *The optimal error bound $\|\mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top\|_2^2 = \lambda_1^2$ is attained if and only if there is $\tilde{z} \in \overline{\mathbb{R}}$ such that*

$$\mathbf{v}_j^\top \tilde{\mathbf{z}} = 0 \quad \text{for all } j \text{ with } |\lambda_j| = |\lambda_1| \quad \text{and} \quad f(\tilde{z}, \lambda_1^2) \geq 0, \quad (4.9)$$

and if moreover \tilde{c} is chosen such that

$$\sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^\top \tilde{\mathbf{z}})^2}{\lambda_j + |\lambda_1|} \leq \frac{1}{\tilde{c}} \leq \sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^\top \tilde{\mathbf{z}})^2}{\lambda_j - |\lambda_1|}. \quad (4.10)$$

(2) If there is no $\tilde{z} \in \overline{\mathbb{R}}$ satisfying (4.9), the best approximation error that can be achieved by a rank-1 Hankel matrix is the minimal value $\tilde{\lambda}$ in $(|\lambda_1|, \lambda_0)$ satisfying the relation

$$\max_{z \in \overline{\mathbb{R}}} f(z, \tilde{\lambda}^2) = 0. \quad (4.11)$$

In this case we have

$$\tilde{z} \in \operatorname{argmax}_{z \in \overline{\mathbb{R}}} f(z, \tilde{\lambda}^2) \quad \text{and} \quad \tilde{c} = \left(\sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^\top \tilde{\mathbf{z}})^2}{\lambda_j - \tilde{\lambda}} \right)^{-1} > 0. \quad (4.12)$$

Proof. Throughout this proof let

$$\tilde{\lambda} = \|\mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top\|_2$$

denote the optimal approximation error; in other words, the parameters \tilde{c} and \tilde{z} generate an optimal rank-1 Hankel approximation of \mathbf{A} . Recall from the beginning of this chapter that this holds if and only if the symmetric matrices $\mathbf{M}_1(\tilde{\lambda})$ and $\mathbf{M}_2(\tilde{\lambda})$ from (4.5) are both positive semidefinite and at least one of them actually possesses the eigenvalue zero.

First note that, the optimal parameter \tilde{z} necessarily satisfies $\tilde{\mu}_0 = \mathbf{v}_0^\top \tilde{\mathbf{z}} \neq 0$. Otherwise, if $\mu_0 = 0$, we would find $(\mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top) \cdot \mathbf{v}_0 = \mathbf{A} \cdot \mathbf{v}_0 = \lambda_0 \cdot \mathbf{v}_0$, and therefore $\tilde{\lambda} = \|\mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top\|_2 \geq \lambda_0$. This contradicts the upper bound $\tilde{\lambda} < \lambda_0$ from Proposition 4.6.

Furthermore, we find the necessary condition $\tilde{c} > 0$. Otherwise, for $c \leq 0$, we would add a positive semidefinite matrix to \mathbf{A} , thereby enlarging the spectral norm

$$\begin{aligned} \|\mathbf{A} - c \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top\|_2 &= \max_{\|\mathbf{v}\|=1} |\mathbf{v}^\top (\mathbf{A} - c \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top) \mathbf{v}| \\ &\geq \mathbf{v}_0^\top \mathbf{A} \mathbf{v}_0 - c \cdot (\mathbf{v}_0^\top \tilde{\mathbf{z}})^2 \\ &= \lambda_0 + |c| \cdot (\mathbf{v}_0^\top \tilde{\mathbf{z}})^2 \geq \lambda_0. \end{aligned}$$

This would again contradict the upper bound from Proposition 4.6.

We derive necessary and sufficient conditions on the optimal parameters $\tilde{c} > 0$ and $\tilde{z} \in \overline{\mathbb{R}}$, and the optimal approximation error $\tilde{\lambda} \in [|\lambda_1|, \lambda_0)$ by inspecting the matrices $\mathbf{M}_1(\tilde{\lambda})$ and $\mathbf{M}_2(\tilde{\lambda})$. Thereby we prove part (1) of the theorem.

Note that for the entries of the diagonal part $\tilde{\lambda}\mathbf{I} - \mathbf{A}$ of the first auxiliary matrix $\mathbf{M}_1(\tilde{\lambda})$ we have $\tilde{\lambda} - \lambda_0 < 0$ while $\tilde{\lambda} - \lambda_j \geq 0$ for $j = 1, \dots, N-1$. Thus, we can apply Lemma 4.3

to $\mathbf{M}_1(\tilde{\lambda})$. The diagonal part $\tilde{\lambda}\mathbf{I} + \mathbf{\Lambda}$ of the second auxiliary matrix $\mathbf{M}_2(\tilde{\lambda})$ is positive semidefinite. In fact, we have $\tilde{\lambda} + \lambda_j \geq 0$ for all $j = 0, \dots, N-1$, and actually $\tilde{\lambda} + \lambda_0 > 0$. So, $\mathbf{M}_2(\tilde{\lambda})$ meets the prerequisites of Lemma 4.4. From Lemmas 4.3 and 4.4, it follows that $\mathbf{M}_1(\tilde{\lambda})$ and $\mathbf{M}_2(\tilde{\lambda})$ are simultaneously positive semidefinite if and only if

$$\sum_{j=0}^{N-1} \frac{\mu_j^2}{\lambda_j + \tilde{\lambda}} \leq \frac{1}{\tilde{c}} \leq \sum_{j=0}^{N-1} \frac{\mu_j^2}{\lambda_j - \tilde{\lambda}} \quad (4.13)$$

and if in case of $\tilde{\lambda} = |\lambda_1|$ moreover $\mu_j = \mathbf{v}_j^T \tilde{\mathbf{z}} = 0$ holds for all j with $|\lambda_j| = |\lambda_1| = \tilde{\lambda}$. Obviously, such a parameter \tilde{c} can only exist if

$$\sum_{j=0}^{N-1} \frac{\mu_j^2}{\lambda_j + \tilde{\lambda}} - \sum_{j=0}^{N-1} \frac{\mu_j^2}{\lambda_j - \tilde{\lambda}} = 2\tilde{\lambda} \cdot \sum_{j=0}^{N-1} \frac{\mu_j^2}{\lambda_j^2 - \tilde{\lambda}^2} = 2\tilde{\lambda} \cdot f(\tilde{\mathbf{z}}, \tilde{\lambda}^2) \geq 0,$$

with $f(\tilde{\mathbf{z}}, \tilde{\lambda}^2)$ as in (4.8). Observe that $\tilde{\lambda} \geq |\lambda_1| > 0$ since we have assumed $\text{rank } \mathbf{A} > 1$. Hence, the condition $f(\tilde{\mathbf{z}}, \tilde{\lambda}^2) \geq 0$ follows, and we conclude (4.9) and (4.10) for $\tilde{\lambda} = |\lambda_1|$.

In order to prove part (2) of the theorem, assume that the condition (4.9) is not satisfied for any $z \in \overline{\mathbb{R}}$; that is, assume $\tilde{\lambda} > |\lambda_1|$. Inspecting the two sums in (4.13) we observe that the left-hand sum increases for decreasing $\tilde{\lambda}$ while the sum on the right-hand side decreases with $\tilde{\lambda}$. Thus, (4.13) implies the equalities

$$\sum_{j=0}^{N-1} \frac{\mu_j^2}{\lambda_j + \tilde{\lambda}} = \frac{1}{\tilde{c}} = \sum_{j=0}^{N-1} \frac{\mu_j^2}{\lambda_j - \tilde{\lambda}} \quad (4.14)$$

for the minimal error $\tilde{\lambda}$. Otherwise, we could find a parameter \tilde{c} such that both inequalities in (4.13) are strict. Then, Lemmas 4.3 and 4.4 yield that the two auxiliary matrices $\mathbf{M}_1(\tilde{\lambda})$ and $\mathbf{M}_2(\tilde{\lambda})$ are strictly positive definite. This implies that there is some λ with $|\lambda_1| \leq \lambda < \tilde{\lambda}$ such that $\mathbf{M}_1(\lambda)$ and $\mathbf{M}_2(\lambda)$ are still positive semidefinite. The existence of such λ contradicts our assumption (4.3) on the optimality of $\tilde{\lambda}$. The expression for \tilde{c} thus follows from (4.14).

Further, we conclude

$$\sum_{j=0}^{N-1} \frac{\mu_j^2}{\lambda_j + \tilde{\lambda}} - \sum_{j=0}^{N-1} \frac{\mu_j^2}{\lambda_j - \tilde{\lambda}} = 2\tilde{\lambda} \cdot \sum_{j=0}^{N-1} \frac{\mu_j^2}{\lambda_j^2 - \tilde{\lambda}^2} = 2\tilde{\lambda} \cdot f(\tilde{\mathbf{z}}, \tilde{\lambda}^2) = 0.$$

Since $\tilde{\lambda} > |\lambda_1| > 0$, this shows that $f(\tilde{\mathbf{z}}, \tilde{\lambda}^2) = 0$.

Finally, for the fixed optimal error $\tilde{\lambda}$, we consider $f(z, \tilde{\lambda}^2)$ as a polynomial in z . We show that $f(z, \tilde{\lambda}) \leq 0$ for all $z \in \overline{\mathbb{R}}$. Assume to the contrary that there is some z with $f(z, \tilde{\lambda}) > 0$. With the same arguments as before, we obtain a range for the choice of \tilde{c} . But then \tilde{c} can be taken such that the two matrices $\mathbf{M}_1(\tilde{\lambda})$ and $\mathbf{M}_2(\tilde{\lambda})$ are both strictly positive definite. In that case, $\tilde{\lambda}$ is no longer the optimal error, contradicting our assumption. Thus, we have shown the assertion (4.11) and the characterization of \tilde{z} in (4.12), which completes the proof. \square

Remark 4.10 1. The conditions (4.9) in Theorem 4.9 are particularly satisfied if the eigenvector \mathbf{v}_0 corresponding to the largest eigenvalue λ_0 already is of the form $\mathbf{v}_0 = \tilde{\mathbf{z}}$ for some $\tilde{z} \in \overline{\mathbb{R}}$. In this case, since $\mathbf{v}_j^\top \tilde{\mathbf{z}} = \mathbf{v}_j^\top \mathbf{v}_0 = 0$ for $j = 1, \dots, N-1$, the non-negativity condition on f simplifies to

$$f(z, \lambda_1^2) = \frac{(\mathbf{v}_0^\top \tilde{\mathbf{z}})^2}{\lambda_0^2 - \lambda_1^2} = \frac{\|\tilde{\mathbf{z}}\|^2}{\lambda_0^2 - \lambda_1^2} = \frac{1}{\lambda_0^2 - \lambda_1^2} \geq 0.$$

This is certainly satisfied since $\lambda_0^2 - \lambda_1^2 > 0$ by assumption.

2. The optimal parameters \tilde{c} and \tilde{z} determining the optimal rank-1 Hankel approximation with respect to the spectral norm need not be unique. If the optimal error bound $|\lambda_1|$ from (1.3) is attained and the inequalities (4.10) are strict, then there are multiple possible choices for the optimal coefficient \tilde{c} , see Example 4.11. But still, if the optimal error bound cannot be attained (i.e., $\tilde{\lambda} > |\lambda_1|$) and \tilde{c} is determined uniquely by (4.12), it may happen that $\tilde{z} \in \operatorname{argmax}_{z \in \overline{\mathbb{R}}} f(z, \tilde{\lambda}^2)$ is not unique, see Example 4.12.

3. The relation (4.14) in the proof of Theorem 4.9 directly implies that $\det \mathbf{M}_1(\tilde{\lambda}) = \det \mathbf{M}_2(\tilde{\lambda}) = 0$. Equivalently, both $\tilde{\lambda}$ and $-\tilde{\lambda}$ are eigenvalues of the difference matrix $\mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top$.

We give two small examples to illustrate Theorem 4.9. In the first one, the optimal error bound $\tilde{\lambda} = |\lambda_1|$ from (1.3) is attained and non-uniqueness of the optimal solution is impressively demonstrated. It has been published almost identically in our conference paper [KPP21b].

Example 4.11 Consider the symmetric matrix

$$\mathbf{A} = \begin{pmatrix} 12 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & 9 \end{pmatrix},$$

its eigenvalues $\lambda_0 = 12$, $\lambda_1 = 11$, and $\lambda_2 = 1$ and corresponding normalized eigenvectors

$$\mathbf{v}_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{v}_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}, \quad \mathbf{v}_2 = \frac{1}{\sqrt{5}} \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}.$$

Following part (1) of Theorem 4.9, we first search for structured vectors $\mathbf{z} \in \mathbb{R}^3$ that are orthogonal to the second eigenvector \mathbf{v}_1 by solving the equation

$$\mathbf{v}_1^\top \mathbf{z} = 0 \quad \Leftrightarrow \quad z + 2z^2 = z \cdot (1 + 2z) = 0.$$

We obtain the solutions $z = 0$ and $z = -1/2$, which provide the normalized structured vectors $\mathbf{z}(0) = (1 \ 0 \ 0)^\top$ and $\mathbf{z}(-1/2) = \frac{4}{\sqrt{21}} \cdot (1 \ -1/2 \ 1/4)^\top$. Note that the structured vector $\mathbf{z}(\infty) = (0 \ 0 \ 1)^\top$ is not a solution to $\mathbf{v}_1^\top \mathbf{z} = 0$.

For each solution we check the non-negativity condition $f(z, \lambda_1^2) \geq 0$, according to (4.9). We obtain

$$f(0, 11^2) = \frac{1}{12^2 - 11^2} + 0 = \frac{1}{144 - 121} = \frac{1}{23} \geq 0$$

and

$$f(-1/2, 11^2) = \frac{16/21 \cdot 1}{12^2 - 11^2} + \frac{1/5 \cdot 16/21 \cdot (1 + 1/4)^2}{1^2 - 11^2} \approx 0.03 \geq 0.$$

So both $\tilde{z} = 0$ and $\tilde{z} = -1/2$ are optimal structure parameters.

Thus, we obtain two different optimal rank-1 Hankel approximation matrices

$$\tilde{c} \cdot \tilde{\mathbf{z}}(0) \tilde{\mathbf{z}}(0)^\top = \tilde{c} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and

$$\tilde{c} \cdot \tilde{\mathbf{z}}(-1/2) \tilde{\mathbf{z}}(-1/2)^\top = \tilde{c} \cdot \begin{pmatrix} 1 & -1/2 & 1/4 \\ -1/2 & 1/4 & -1/8 \\ 1/4 & -1/8 & 1/16 \end{pmatrix},$$

where the coefficients \tilde{c} lie within the range

$$\frac{1}{23} = \frac{1}{12 + 11} \leq \frac{1}{\tilde{c}} \leq \frac{1}{12 - 11} = 1 \quad \Leftrightarrow \quad 1 \leq \tilde{c} \leq 23 \quad \text{for } \tilde{z} = 0,$$

and analogously

$$1.355 \approx \frac{42}{31} \leq \tilde{c} \leq \frac{5796}{307} \approx 18.880 \quad \text{for } \tilde{z} = -1/2,$$

according to (4.10). For any such approximation, the error $\|\mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top\|_2 = \lambda_1 = 11$ is exactly the optimal error bound. Especially consider the boundaries of \tilde{c} for $\tilde{z} = 0$

$$\|\mathbf{A} - 1 \cdot \tilde{\mathbf{z}}(0)\tilde{\mathbf{z}}(0)^\top\|_2 = \left\| \begin{pmatrix} 11 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & 9 \end{pmatrix} \right\|_2 = 11$$

and

$$\|\mathbf{A} - 23 \cdot \tilde{\mathbf{z}}(0)\tilde{\mathbf{z}}(0)^\top\|_2 = \left\| \begin{pmatrix} -11 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & 9 \end{pmatrix} \right\|_2 = 11.$$

In this particular example we easily see that for $\tilde{c} = 1$ and $\tilde{c} = 23$ the by modulus largest eigenvalue $\tilde{\lambda} = |\pm 11|$ of the difference matrix occurs with multiplicity two. All values $1 < \tilde{c} < 23$ in between cause one of these two eigenvalues to decrease in absolute value, while the other one remains untouched.

Next, we consider an example for which the optimal error bound (1.3) cannot be attained (i.e., $\tilde{\lambda} > |\lambda_1|$). We find the optimal rank-1 Hankel approximation by applying part (2) of Theorem 4.9. In this particular case, the optimal solution can, in fact, be calculated completely analytically.

Example 4.12 Consider the symmetric matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \\ 1/2 & 0 & 1 \end{pmatrix}$$

with normalized eigenvectors

$$\mathbf{v}_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{v}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix},$$

and corresponding eigenvalues $\lambda_0 = 3/2$ and $\lambda_1 = \lambda_2 = 1/2$. Thus the optimal approximation error $\tilde{\lambda}$ lies within the interval $[1/2, 3/2)$. Since the system of equations

$$\begin{aligned} \mathbf{v}_1^\top \mathbf{z} &= 0 & \Leftrightarrow & \quad z = 0 \\ \mathbf{v}_2^\top \mathbf{z} &= 0 & \Leftrightarrow & \quad 1 - z^2 = 0 \end{aligned}$$

does not have a common root in $\overline{\mathbb{R}}$, the optimal error bound cannot be attained. This means, we have $\tilde{\lambda} > |\lambda_1| = 1/2$. Therefore, we need to find $\tilde{\lambda} \in (1/2, 3/2)$ and \tilde{z} such that $f(\tilde{z}, \tilde{\lambda}^2)$ satisfies (4.11), namely, $\max_{z \in \overline{\mathbb{R}}} f(z, \tilde{\lambda}^2) = 0$ and $\tilde{z} \in \operatorname{argmax}_{z \in \overline{\mathbb{R}}} f(z, \tilde{\lambda}^2)$. Filling in the respective eigenvalues in (4.8), we obtain

$$f(z, \lambda^2) = \frac{(\mathbf{v}_0^\top \mathbf{z})^2}{9/4 - \lambda^2} + \frac{(\mathbf{v}_1^\top \mathbf{z})^2}{1/4 - \lambda^2} + \frac{(\mathbf{v}_2^\top \mathbf{z})^2}{1/4 - \lambda^2}.$$

Multiplying with the normalization factor $(1 + z + z^2)$ of the structured vector, this results in

$$\begin{aligned} & (1 + z^2 + z^4) \cdot f(z, \lambda^2) \\ &= \frac{1/2 + z^2 + z^4/2}{9/4 - \lambda^2} + \frac{z^2}{1/4 - \lambda^2} + \frac{1/2 - z^2 + z^4/2}{1/4 - \lambda^2} \\ &= \frac{1}{(9/4 - \lambda^2)(1/4 - \lambda^2)} \cdot \left((5/4 - \lambda^2) z^4 + (1/4 - \lambda^2) z^2 + (5/4 - \lambda^2) \right) \\ &= \frac{5/4 - \lambda^2}{(9/4 - \lambda^2)(1/4 - \lambda^2)} \cdot \left(\left(z^2 - \frac{\lambda^2 - 1/4}{2 \cdot (5/4 - \lambda^2)} \right)^2 + 1 - \left(\frac{\lambda^2 - 1/4}{2 \cdot (5/4 - \lambda^2)} \right)^2 \right), \end{aligned}$$

where in the last line we assume $\lambda^2 \neq 5/4$. A direct inspection of the last expression provides that $\max_{z \in \overline{\mathbb{R}}} f(z, \tilde{\lambda}^2) = 0$ if and only if

$$1 - \left(\frac{\lambda^2 - 1/4}{2 \cdot (5/4 - \lambda^2)} \right)^2 = 0,$$

that is, if $\tilde{\lambda}^2 = 11/12$.

We then obtain the optimal parameters from (4.12) as

$$\tilde{z}^2 = \frac{\tilde{\lambda}^2 - 1/4}{2 \cdot (5/4 - \tilde{\lambda}^2)} = \frac{11/12 - 1/4}{2 \cdot (5/4 - 11/12)} = 1 \quad \Rightarrow \quad \tilde{z} = \pm 1$$

and

$$\tilde{c} = \left(\frac{2/3}{3/2 - \sqrt{11/12}} + \frac{1/3}{1/2 - \sqrt{11/12}} + 0 \right)^{-1} = 2.$$

Taking the normalization of $\tilde{\mathbf{z}}$ into account, we obtain for the approximation error

$$\left\| \mathbf{A} - 2 \cdot \frac{1}{3} \begin{pmatrix} 1 & \pm 1 & 1 \\ \pm 1 & 1 & \pm 1 \\ 1 & \pm 1 & 1 \end{pmatrix} \right\|_2 = \left\| \frac{1}{6} \begin{pmatrix} 2 & \pm 4 & -1 \\ \pm 4 & -1 & \pm 4 \\ -1 & \pm 4 & 2 \end{pmatrix} \right\|_2 = \sqrt{\frac{11}{12}} \approx 0.9574,$$

rounded to four digits.

The same matrix as in Example 4.12 has been approximated in Example 3.8 for the Frobenius norm. For a comparison of the results see Section 9.1. Especially Table 9.1 provides a neat arrangement of the respective optimal parameters and optimal approximation errors.

In most examples, especially ones of larger dimensions, the optimal error may not be deduced directly by inspection of the function $f(z, \lambda^2)$ as in Example 4.12. The optimal rank-1 Hankel approximation can still be computed numerically in those cases, see the bisection procedure in Section 4.3 for this purpose.

4.2.2 Multiple Largest Eigenvalue

Now let the by modulus largest eigenvalue of $\mathbf{A} \in \mathbb{R}^{N \times N}$ occur with higher multiplicity; that is, we have $\lambda_0 = \|\mathbf{A}\|_2 = |\lambda_1|$. In this case, an optimal Hankel structured approximation of true rank one does not always exist as can be deduced from Proposition 4.13. In fact, either there is no solution to problem (4.1) or the optimal error bound (1.3) is attained. In Theorem 4.15, we give necessary and sufficient conditions for the latter case to occur.

Partial results and examples from this section have appeared in our main publication [KPP21a] and in compressed form in [Kni21].

We start this section with an adaptation of Proposition 4.6 to matrices \mathbf{A} with multiple largest eigenvalue.

Proposition 4.13 *Assume that the largest eigenvalue of \mathbf{A} occurs with higher multiplicity $\lambda_0 = \|\mathbf{A}\|_2 = |\lambda_1|$, without loss of generality $\lambda_0 > 0$. Then any number λ strictly larger than the largest eigenvalue λ_0 of \mathbf{A} cannot be the optimal approximation error for rank-1 Hankel approximation in the spectral norm.*

Proof. Similarly to the proof of Proposition 4.6, we find that for $\lambda > \lambda_0$ there are parameters $z \in \overline{\mathbb{R}}$ and $c \in \mathbb{R} \setminus \{0\}$ such that both $\mathbf{M}_1(\lambda, c, z)$ and $\mathbf{M}_2(\lambda, c, z)$ are strictly positive definite.

First, note that for $\lambda > \lambda_0$ both the diagonal parts $\lambda\mathbf{I} - \mathbf{\Lambda}$ and $\lambda\mathbf{I} + \mathbf{\Lambda}$ are strictly positive definite. Choose any $z \in \overline{\mathbb{R}}$ and set $c = \frac{\lambda - \lambda_0}{2}$. Then, for these parameters, $\mathbf{M}_1(\lambda, c, z)$ is strictly positive definite since it is the sum of a positive definite and a positive semidefinite matrix. The second auxiliary matrix $\mathbf{M}_2(\lambda, c, z)$ is strictly positive definite by the same arguments as in the proof of Proposition 4.6. \square

Note that by the Eckart-Young-Mirsky Theorem (Theorem 1.8) the error of a rank-1 approximation cannot be smaller than $|\lambda_1| = \lambda_0$. So, for a matrix with multiple largest absolute eigenvalue, two situations can occur. Either the optimal error bound λ_0 is attained by the optimal rank-1 Hankel approximation, or there exists no Hankel matrix of true rank one optimally approximating the matrix \mathbf{A} . In the latter case, the r1H problem (4.1) does not have a solution. When relaxing (4.1) slightly to

$$\begin{aligned} \min \|\mathbf{A} - \mathbf{H}\|_2 \quad & \text{such that } \text{rank } \mathbf{H} \leq 1, \\ & \text{and } \mathbf{H} \text{ has Hankel structure} \\ = \min_{\substack{z \in \overline{\mathbb{R}} \\ c \in \mathbb{R}}} \|\mathbf{A} - c \cdot \mathbf{z}\mathbf{z}^\top\|_2, \end{aligned} \quad (4.15)$$

allowing the rank of the approximating matrix to be smaller than or equal to one, we only obtain the trivial solution of a zero matrix (i.e., $c = 0$). See Example 4.18 for an illustration of this incidence.

Remark 4.14 Of course, the optimal error bound λ_0 is always attained for $c = 0$. However with $c = 0$ the matrix $c \cdot \mathbf{z}\mathbf{z}^\top$ is not a rank-1 matrix but the zero matrix. Therefore, this case is deliberately excluded.

In the upcoming theorem, we give necessary and sufficient conditions under which the optimal error bound $\tilde{\lambda} = \lambda_0 = |\lambda_1|$ from (1.3) is attained for a Hankel matrix of true rank one.

Theorem 4.15 *Let $\mathbf{A} \in \mathbb{R}^{N \times N}$ be symmetric with $\text{rank } \mathbf{A} > 1$. Assume that the largest eigenvalue occurs with higher multiplicity $\lambda_0 = \|\mathbf{A}\|_2 = |\lambda_1|$, without loss of generality $\lambda_0 > 0$. The optimal error bound $\|\mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top\|_2 = |\lambda_1| = \lambda_0$ can be attained by a rank-1*

Hankel matrix $\tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top$ if and only if $\tilde{z} \in \overline{\mathbb{R}}$ satisfies either

$$\mathbf{v}_j^\top \tilde{\mathbf{z}} = 0 \quad \text{for all } j \text{ with } \lambda_j = -\lambda_0, \quad (4.16a)$$

or

$$\mathbf{v}_j^\top \tilde{\mathbf{z}} = 0 \quad \text{for all } j \text{ with } \lambda_j = \lambda_0. \quad (4.16b)$$

Then \tilde{c} chosen as

$$\tilde{c} = \left(\sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^\top \tilde{\mathbf{z}})^2}{\lambda_0 + \lambda_j} \right)^{-1} > 0 \quad \text{in the first case (4.16a),} \quad (4.17a)$$

respectively

$$\tilde{c} = \left(- \sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^\top \tilde{\mathbf{z}})^2}{\lambda_0 - \lambda_j} \right)^{-1} < 0 \quad \text{in the second case (4.16b)} \quad (4.17b)$$

ensures that the optimal error bound λ_0 is attained by $\tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top$.

Proof. First, assume that z does not satisfy (4.16); that is, z satisfies neither (4.16a) nor (4.16b). This means that there are indices j_1 and j_2 contradicting (4.16a) and (4.16b), respectively. More precisely, we have $\lambda_{j_1} = -\lambda_0$ and $\mathbf{v}_{j_1}^\top \mathbf{z} = \mu_{j_1} \neq 0$, as well as $\lambda_{j_2} = \lambda_0$ and $\mathbf{v}_{j_2}^\top \mathbf{z} = \mu_{j_2} \neq 0$. We show that for such z , the two auxiliary matrices $\mathbf{M}_1(\lambda_0, c, z)$ and $\mathbf{M}_2(\lambda_0, c, z)$ cannot both be positive semidefinite. Thus z as assumed cannot be an optimal parameter attaining the optimal error λ_0 .

In order to examine the definiteness of $\mathbf{M}_1(\lambda_0, c, z)$ and $\mathbf{M}_2(\lambda_0, c, z)$, define the vectors $\mathbf{x}_1 := (\delta_{j_1, j})_{j=0}^{N-1}$ and $\mathbf{x}_2 := (\delta_{j_2, j})_{j=0}^{N-1}$. Therein, $\delta_{j, k}$ denotes the Kronecker symbol; that is, x_{j_1} and x_{j_2} are the only non-zero entries of \mathbf{x}_1 and \mathbf{x}_2 , respectively. Consider, on the one hand,

$$\mathbf{x}_1^\top \cdot \mathbf{M}_2(\lambda_0, c, z) \cdot \mathbf{x}_1 = (\lambda_0 + \lambda_{j_1}) \cdot x_{j_1} - c \cdot (x_{j_1} \cdot \mu_{j_1})^2 = -c \cdot (x_{j_1} \cdot \mu_{j_1})^2 < 0 \quad \text{for } c > 0.$$

Thus, we need $c < 0$ in order for $\mathbf{M}_2(\lambda_0, c, z)$ to possibly be positive semidefinite. On the other hand, we have

$$\mathbf{x}_2^\top \cdot \mathbf{M}_1(\lambda_0, c, z) \cdot \mathbf{x}_2 = (\lambda_0 - \lambda_{j_2}) \cdot x_{j_2} + c \cdot (x_{j_2} \cdot \mu_{j_2})^2 = c \cdot (x_{j_2} \cdot \mu_{j_2})^2 < 0 \quad \text{for } c < 0.$$

All in all, for z not satisfying (4.16), there exists no $c \neq 0$ such that both $\mathbf{M}_1(\lambda_0, c, z)$ and

$\mathbf{M}_2(\lambda_0, c, z)$ are simultaneously positive semidefinite.

Second, let $\tilde{z} \in \overline{\mathbb{R}}$ such that $\mathbf{v}_j^\top \tilde{z} = 0$ for all j with $\lambda_j = -\lambda_0$ and \tilde{c} chosen accordingly (i.e., \tilde{c} as in (4.17a)). Then, $\mathbf{M}_1(\lambda_0, \tilde{c}, \tilde{z})$ is positive semidefinite as a sum of two positive semidefinite matrices. Furthermore, $\mathbf{M}_2(\lambda_0, \tilde{c}, \tilde{z})$ is positive semidefinite by Lemma 4.4. The choice of \tilde{c} ensures that it has zero as an eigenvalue.

The argument is analogous for $\tilde{z} \in \overline{\mathbb{R}}$ such that $\mathbf{v}_j^\top \tilde{z} = 0$ for all j with $\lambda_j = \lambda_0$ and \tilde{c} as in (4.17b). In this case, $\mathbf{M}_2(\lambda_0)$ is the sum of two positive semidefinite matrices and therefore positive semidefinite. Replacing c by $-c$ in Lemma 4.4, we find that $\mathbf{M}_1(\lambda_0)$ is also positive semidefinite with eigenvalue zero. Hence, $\|\mathbf{A} - \tilde{c} \cdot \tilde{z}\tilde{z}^\top\|_2 = \lambda_0 = |\lambda_1|$ in both cases. \square

Remark 4.16 The precise choices of \tilde{c} in (4.17) are sufficient for the optimal error bound but not necessary. The optimal error bound being attained only implies that the optimal coefficient \tilde{c} is contained in a suitable interval. If we have $\|\mathbf{A} - \tilde{c} \cdot \tilde{z}\tilde{z}^\top\|_2 = \lambda_0 = |\lambda_1|$, then either

$$\mathbf{v}_j^\top \tilde{z} = 0 \quad \text{for all } j \text{ with } \lambda_j = -\lambda_0, \quad \text{and} \quad 0 < \tilde{c} \leq \left(\sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^\top \tilde{z})^2}{\lambda_0 + \lambda_j} \right)^{-1}, \quad (4.18a)$$

or

$$\mathbf{v}_j^\top \tilde{z} = 0 \quad \text{for all } j \text{ with } \lambda_j = \lambda_0, \quad \text{and} \quad 0 > \tilde{c} \geq \left(- \sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^\top \tilde{z})^2}{\lambda_0 - \lambda_j} \right)^{-1}. \quad (4.18b)$$

This is because, from the prerequisites of Theorem 4.15, we cannot determine whether the matrix $\mathbf{M}_1(\lambda_0)$ in the case (4.16a), or $\mathbf{M}_2(\lambda_0)$ in the case (4.16b) possesses the eigenvalue zero or is in fact strictly positive definite. More precisely, if for example the positive largest eigenvalue $+\lambda_0$ is a single one (i.e., $\lambda_j \neq \lambda_0$ for $j = 1, \dots, N-1$), we have no means of knowing whether $\mathbf{M}_1(\lambda_0)$ has a zero-eigenvalue or not.

However, if for example the positive largest eigenvalue occurs with higher multiplicity itself (i.e., $\lambda_0 = \lambda_1 > 0$), then $\mathbf{M}_1(\lambda_0)$ does have the eigenvalue zero and we may choose \tilde{c} in the range (4.18b). The analogue holds for the negative largest eigenvalue $-\lambda_0$ and $\mathbf{M}_2(\lambda_0)$. See also the proof of Corollary 4.19 on that matter.

Remark 4.17 Of course, the condition (4.16a) (respectively (4.16b)) is fulfilled for a structured vector \tilde{z} belonging to the eigenspace spanned by those \mathbf{v}_j with $\lambda_j = -\lambda_0$ (respectively $\lambda_j = \lambda_0$). But this is not necessary as opposed to the case with isolated largest eigenvalue, compare Theorem 4.9.

The requirements (4.16) can be viewed as a set of polynomial equations in the variable z . It might happen that neither the set of polynomial equations corresponding to (4.16a) nor the one corresponding to (4.16b) has a joint solution in $\overline{\mathbb{R}}$. In this case, problem (4.1) does not have a solution and the relaxed problem (4.15) is only solved by the zero matrix. We now give an example of this phenomenon.

Example 4.18 Consider the following symmetric matrix

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{5 \times 5}.$$

This matrix only has two distinct eigenvalues, namely $\lambda_0 = \lambda_1 = \lambda_2 = 1$ and $\lambda_3 = \lambda_4 = -1$, which additionally have the same absolute value. The corresponding normalized eigenvectors are

$$\mathbf{v}_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{v}_3 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{v}_4 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

respectively.

We check for the conditions (4.16) from Theorem 4.15. For $z \in \mathbb{R}$, neither of the two systems of equations

$$\begin{aligned} \mathbf{v}_3^T \mathbf{z} = 0 &\Leftrightarrow z + z^3 = 0 & \text{and} & & \mathbf{v}_0^T \mathbf{z} = 0 &\Leftrightarrow 1 - z^4 = 0 \\ \mathbf{v}_4^T \mathbf{z} = 0 &\Leftrightarrow 1 + z^4 = 0 & & & \mathbf{v}_1^T \mathbf{z} = 0 &\Leftrightarrow z - z^3 = 0 \\ & & & & \mathbf{v}_2^T \mathbf{z} = 0 &\Leftrightarrow z^2 = 0 \end{aligned}$$

corresponding to (4.16a) and (4.16b), respectively, has a joint solution. Further note that $\mathbf{z}(\infty) = (0 \ 0 \ 0 \ 0 \ 1)^T$ is not a solution to either of the systems. Thus, for this matrix \mathbf{A} , a solution to problem (4.1) of true rank one does not exist. The only solution to the relaxed problem (4.15) is the zero matrix.

Example 4.18 demonstrates that, for some initial matrices, the condition (4.16) from Theorem 4.15 are impossible to fulfill. However, they are especially satisfied if the largest eigenvalues all occur with the same sign.

Corollary 4.19 *Let $\mathbf{A} \in \mathbb{R}^{N \times N}$ be symmetric with rank $\mathbf{A} > 1$. Assume that the largest eigenvalue occurs with higher multiplicity $\lambda_0 = \|\mathbf{A}\|_2 = |\lambda_1|$. Further suppose that all by modulus largest eigenvalues of \mathbf{A} , $|\lambda_j| = \|\mathbf{A}\|_2 = |\lambda_0|$, actually have the same sign, without loss of generality, let $\lambda_0 = \lambda_1 = \dots > 0$.*

Then every rank-1 Hankel matrix $\tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top$ with arbitrary $\tilde{\mathbf{z}} \in \overline{\mathbb{R}}$ and \tilde{c} chosen in the range

$$0 < \tilde{c} \leq \left(\sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^\top \tilde{\mathbf{z}})^2}{\lambda_0 + \lambda_j} \right)^{-1}$$

solves the r1H problem (4.1). Thereby the optimal error bound

$$\|\mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top\|_2 = \|\mathbf{A}\|_2 = \lambda_0 = \lambda_1$$

is always attained.

Proof. Additionally to the proof of Theorem 4.15, we have to show that the auxiliary matrix $\mathbf{M}_1(\lambda_0, \tilde{c}, \tilde{\mathbf{z}})$ already has the eigenvalue zero. Then we do not need to ensure that $\mathbf{M}_2(\lambda_0, \tilde{c}, \tilde{\mathbf{z}})$ has zero as an eigenvalue, which was done by choosing \tilde{c} as in (4.17a) in Theorem 4.15. For $\mathbf{M}_2(\lambda_0, \tilde{c}, \tilde{\mathbf{z}})$ to be positive semidefinite, it is sufficient to choose \tilde{c} in the range given in the corollary, see also Remark 4.16.

Consider the determinant of $\mathbf{M}_1(\lambda_0, \tilde{c}, \tilde{\mathbf{z}})$. By Lemma 4.1 we have

$$\det \mathbf{M}_1(\lambda_0, \tilde{c}, \tilde{\mathbf{z}}) = \det(\lambda_0 \mathbf{I} - \mathbf{A}) + \tilde{c} \cdot \sum_{j=0}^{N-1} (\mathbf{v}_j^\top \tilde{\mathbf{z}})^2 \cdot \left(\prod_{\substack{k=0 \\ k \neq j}}^{N-1} (\lambda_0 - \lambda_k) \right) = 0,$$

where clearly $\det(\lambda_0 \mathbf{I} - \mathbf{A}) = 0$. Furthermore, by the assumption $\lambda_0 = \lambda_1$, the product is zero in every summand. Therefore, both $\mathbf{M}_1(\lambda_0, \tilde{c}, \tilde{\mathbf{z}})$ and $\mathbf{M}_2(\lambda_0, \tilde{c}, \tilde{\mathbf{z}})$ are positive semidefinite for \tilde{c} in the given range and in any case $\mathbf{M}_1(\lambda_0, \tilde{c}, \tilde{\mathbf{z}})$ has eigenvalue zero. This yields the claim. \square

We end this section with a small example where Corollary 4.19 is applicable.

Example 4.20 Consider the following symmetric matrix and its eigenvectors

$$\mathbf{A} = \begin{pmatrix} 11 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & 9 \end{pmatrix}, \quad \mathbf{v}_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{v}_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}, \quad \mathbf{v}_2 = \frac{1}{\sqrt{5}} \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}.$$

Its eigenvalues are $\lambda_0 = 11$, $\lambda_1 = 11$, and $\lambda_2 = 1$, so both of the by modulus largest eigenvalues $\lambda_0 = \lambda_1 = 11$ have the same sign. Hence, according to Corollary 4.19, for any $\tilde{z} \in \overline{\mathbb{R}}$ we can choose \tilde{c} anywhere in the range

$$0 < \tilde{c} \leq \left(\frac{1}{1 + \tilde{z}^2 + \tilde{z}^4} \cdot \left(\frac{1}{11 + 11} + \frac{\frac{1}{5} \cdot (\tilde{z} + 2\tilde{z}^2)^2}{11 + 11} + \frac{\frac{1}{5} \cdot (-2\tilde{z} + \tilde{z}^2)^2}{11 + 1} \right) \right)^{-1}$$

in order to obtain an optimal pair of parameters (\tilde{c}, \tilde{z}) .

The relation between \tilde{z} and matching \tilde{c} is depicted in Figure 4.1 for $\tilde{z} \in [-100, 100]$. Any pair of parameters in the colored area of Figure 4.1 constitutes an optimal rank-1 Hankel approximation attaining the optimal error bound $\|\mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top\|_2 = 11$.

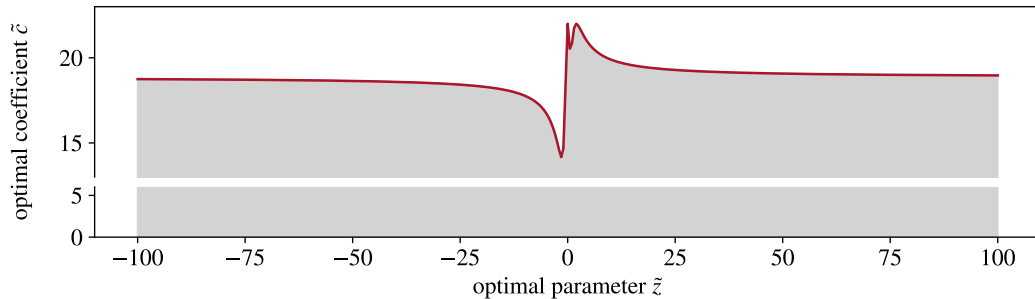


Figure 4.1 Dark red line: upper bound on the optimal coefficient \tilde{c} depending on the structure parameter \tilde{z} . Light grey area: admissible pairs (\tilde{c}, \tilde{z}) generating optimal rank-1 Hankel approximations $\tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top$ of \mathbf{A} from Example 4.20.

Note that the upper bound on \tilde{c} depends continuously on the optimal \tilde{z} . In particular, we obtain

$$0 < \tilde{c} \leq \left(0 + \frac{\frac{1}{5} \cdot 2^2}{11 + 11} + \frac{\frac{1}{5} \cdot 1}{11 + 1} \right)^{-1} = \frac{132}{7} \approx 18.8571,$$

for the limit $\tilde{\mathbf{z}}(\infty) = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \end{pmatrix}^\top$. This value fits nicely with the margins of Figure 4.1 (dark red line).

4.3 Computation of the Optimal Approximation

Theorems 4.9 and 4.15 can be exploited to implement an algorithm for the numerical computation of the optimal rank-1 Hankel approximation with respect to the spectral norm. The parts of this section concerning the optimal rank-1 Hankel approximation when the largest eigenvalue is isolated follow closely along the lines of the respective subsection of our paper [KPP21a, Sec. 4.3]. The remaining ones for multiple largest eigenvalue are new.

The natural first step in Algorithm 4.1 is to determine whether the largest eigenvalue is isolated or occurs with higher multiplicity. Depending on which is the case, the decision on the further procedure is made.

Algorithm 4.1 Optimal rank-1 Hankel approximation w.r.t. the spectral norm

Input: Symmetric matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$.

Compute the eigendecomposition of \mathbf{A} to obtain the eigenvalues

$\lambda_0 \geq |\lambda_1| \geq \dots \geq |\lambda_{N-1}| \geq 0$ and the corresponding normalized eigenvectors

$\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_{N-1}$;

if the largest eigenvalue is isolated, **then**

compute the optimal rank-1 Hankel approximation according to Theorem 4.9, see Algorithm 4.2;

if the largest eigenvalue is not isolated, **then**

compute the optimal rank-1 Hankel approximation according to Theorem 4.15, see Algorithm 4.3.

Output: Parameters \tilde{c} and \tilde{z} generating an optimal rank-1 Hankel approximation of \mathbf{A} with the optimal approximation error $\tilde{\lambda}$ in the spectral norm.

In case of isolated largest eigenvalue, we have to verify whether the conditions (4.9) from part (1) of Theorem 4.9 can be satisfied. If this is the case for some $\tilde{z} \in \overline{\mathbb{R}}$, we choose \tilde{c} according to (4.10). Keep in mind that we have to check for the structured vector $\mathbf{z} = \left(0 \dots \dots 0 \ 1\right)^\top$ indexed by $z = \infty$ separately. This can be easily done by examining the last entry of the eigenvectors \mathbf{v}_j corresponding to the eigenvalues λ_j with $|\lambda_j| = |\lambda_1|$. Only if all of these last entries are zero, the structured vector $\mathbf{z} = \left(0 \dots \dots 0 \ 1\right)^\top$ is orthogonal to all the \mathbf{v}_j in question, as demanded by (4.9).

If (4.9) cannot be satisfied, then we have to employ the relations (4.11) and (4.12) from part (2) of Theorem 4.9 to determine the optimal parameters \tilde{c} and \tilde{z} . To this end, we make

an important observation about the function f from (4.8). Recall the definition

$$f(z, \lambda^2) := \sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^\top \mathbf{z})^2}{\lambda_j^2 - \lambda^2} = \sum_{j=0}^{N-1} \frac{\mu_j^2}{\lambda_j^2 - \lambda^2} \quad (4.19)$$

for $z \in \overline{\mathbb{R}}$ and $\lambda^2 \in [\lambda_1^2, \lambda_0^2)$.

For fixed $z \in \overline{\mathbb{R}}$, the function $f(z, \lambda^2)$ is strictly monotonically increasing in λ^2 . This fact can be used to track down the optimal error $\tilde{\lambda}$ in (4.3) when part (2) of Theorem 4.9 applies (i.e., when $\tilde{\lambda} \in (|\lambda_1|, \lambda_0)$). For any fixed $\lambda^2 \in (\lambda_1^2, \lambda_0^2)$, define $f_\lambda(\cdot) := f(\cdot, \lambda^2)$. Then relation (4.11) and the monotonicity of $f(z, \lambda^2)$ for fixed z imply

- if $\max_{z \in \overline{\mathbb{R}}} f_\lambda(z) > 0$, then the optimal error satisfies $\tilde{\lambda}^2 < \lambda^2$,
- if $\max_{z \in \overline{\mathbb{R}}} f_\lambda(z) < 0$, then the optimal error satisfies $\tilde{\lambda}^2 > \lambda^2$,
- if $\max_{z \in \overline{\mathbb{R}}} f_\lambda(z) = 0$, then the optimal error satisfies $\tilde{\lambda}^2 = \lambda^2$, and the optimal rank-1 Hankel approximation is generated by $\tilde{z} \in \operatorname{argmax}_{z \in \overline{\mathbb{R}}} f(z, \lambda)$ and the corresponding coefficient \tilde{c} from (4.12).

Therefore, a bisection iteration on $\lambda \in (|\lambda_1|, \lambda_0)$ can be used to find the optimal approximation error $\tilde{\lambda}$. At each step of the bisection iteration we have to determine the sign of the maximum of f_λ .

In order to find a simpler range for z in which to find that maximum, we apply an idea similar to the one used in Theorem 3.4 for the Frobenius norm. Let

$$g_\lambda(z) := \sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^\top \cdot \mathbf{J}\mathbf{z})^2}{\lambda_j^2 - \lambda^2}, \quad (4.20)$$

where $\mathbf{J} = \mathbf{J}_N$ denotes the counter-identity matrix (1.10). Because of $\mathbf{z}(1/z) = \mathbf{J}\mathbf{z}(z)$, we have

$$f_\lambda(1/z) = \sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^\top \cdot \mathbf{z}(1/z))^2}{\lambda_j^2 - \lambda^2} = \sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^\top \cdot \mathbf{J}\mathbf{z}(z))^2}{\lambda_j^2 - \lambda^2} = g_\lambda(z)$$

for $z \neq 0$. Moreover, we have $g_\lambda(0) = \lim_{z \rightarrow \infty} f_\lambda(z)$. Thus, in order to find the maximum of $f_\lambda(z)$, we can split our search between f_λ and g_λ . We search for the maximum of $f_\lambda(z)$ merely inside the interval $[-1, 1]$. To account for the rest of the extended real numbers $\overline{\mathbb{R}} \setminus [-1, 1]$, we search for the maximum of $g_\lambda(z)$ inside the open interval $(-1, 1)$.

Algorithm 4.2 Optimal rank-1 Hankel approximation w.r.t. the spectral norm for isolated largest eigenvalue

Input: Eigenvalues $\lambda_0 > |\lambda_1| \geq \dots \geq |\lambda_{N-1}| \geq 0$ and the corresponding normalized eigenvectors $\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_{N-1}$, threshold $\varepsilon > 0$.

Compute the set Σ of joint real roots of the polynomials $v_j(z) = \mathbf{v}_j^T \mathbf{z}$ corresponding to the eigenvalues λ_j with $|\lambda_j| = |\lambda_1|$, and with \mathbf{z} as in (2.5);

for $z \in \Sigma$ **do**

compute

$$f_{\lambda_1}(z) = f(z, \lambda_1^2) = \sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^T \mathbf{z})^2}{\lambda_j^2 - \lambda_1^2}$$

if $f_{\lambda_1}(z) \geq 0$, **then set**

$$\tilde{\lambda} = |\lambda_1|, \quad \tilde{z} = z, \quad \tilde{c} = \left(\sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^T \mathbf{z})^2}{\lambda_j - |\lambda_1|} \right)^{-1}$$

add $(\tilde{\lambda}, \tilde{c}, \tilde{z})$ to Solution;

if $\Sigma = \emptyset$ or $f_{\lambda_1}(z) < 0$ for all $z \in \Sigma$, **then**

 apply the following bisection iteration: set $a = |\lambda_1|$ and $b = \lambda_0$,

while $b - a > \varepsilon$ **do**

compute $\lambda = \frac{a+b}{2}$, and find the maximal value

$$W = \max \left\{ \max_{z \in [-1,1]} f_\lambda(z), \max_{z \in (-1,1)} g_\lambda(z) \right\},$$

 with f_λ and g_λ defined in (4.19) and (4.20), respectively,

if $W > 0$, **then set** $b = \lambda$,

if $W < 0$, **then set** $a = \lambda$,

if $W = 0$, **then** we have found the optimal error $\tilde{\lambda} = \lambda$, set $a = b$

if W was the maximum of $f_{\tilde{\lambda}}$ at $z_f \in [-1, 1]$, **then set** $\tilde{z} = z_f$,

else if W was the maximum of $g_{\tilde{\lambda}}$ at $z_g \in (-1, 1)$, **then set** $\tilde{z} = 1/z_g$

compute

$$\tilde{c} = \left(\sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^T \tilde{\mathbf{z}})^2}{\lambda_j - \tilde{\lambda}} \right)^{-1}$$

add $(\tilde{\lambda}, \tilde{c}, \tilde{z})$ to Solution;

return Solution.

Output: Optimal approximating error $\tilde{\lambda}$ and parameters \tilde{c} and \tilde{z} generating an optimal rank-1 Hankel approximation of \mathbf{A} .

Depending on the sign of the larger maximum

$$W := \max \left\{ \max_{z \in [-1,1]} f_\lambda(z), \max_{z \in (-1,1)} g_\lambda(z) \right\},$$

we narrow down the range in which to search for the optimal error $\tilde{\lambda}$. The bisection procedure stops when $W = 0$ for a certain λ , or when a sufficient accuracy for λ is reached.

Upon termination, we have to distinguish which one of the functions f and g contributes the overall maximum W . If $W = \max_{z \in [-1,1]} f_\lambda(z) =: f_\lambda(z_f)$, we set $\tilde{z} = z_f$. If however $W = \max_{z \in (-1,1)} g_\lambda(z) =: g_\lambda(z_g)$, then we have to proceed with the reciprocal of its maximizer $\tilde{z} = 1/z_g$. In the latter case, we have to check if $z_g = 0$ and if necessary manually set $\tilde{z} = \infty$ instead of $\tilde{z} = 1/z_g$. Besides the smaller range, the use of both functions f and g has the advantage that $\tilde{z} = \infty$ does not have to be found as a maximizer by the built-in functions. Instead, $\tilde{z} = \infty$ is equivalent to $z_g = 0$.

We summarize our deductions in Algorithm 4.2.

When the largest eigenvalue is not isolated, we invoke Theorem 4.15. We have to find structured vectors \mathbf{z} that are orthogonal to the eigenspace of $-\lambda_0$ or $+\lambda_0$. We do so by interpreting the inner product $\mathbf{v}_j^\top \mathbf{z}$ as a polynomial $v_j(\cdot)$ in z . The joint real roots of the set of polynomials corresponding to either $-\lambda_0$ or $+\lambda_0$ are the optimal structure parameters \tilde{z} . Again, $z = \infty$ can be an optimal structure parameter and has to be accounted for individually. As before, when checking condition (4.9), this can be done via the last entry of the eigenvectors. The structure parameter $z = \infty$ generates an optimal rank-1 Hankel approximation if and only if the last entry of all eigenvectors corresponding to $-\lambda_0$ (respectively $+\lambda_0$) are zero.

Once an optimal structure parameter \tilde{z} is determined, the corresponding optimal coefficient \tilde{c} is computed according to (4.17), depending on which eigenspace $\tilde{\mathbf{z}}$ is orthogonal to. Algorithm 4.3 summarizes the computation of the optimal rank-1 Hankel approximation in case of multiple largest eigenvalue.

Remark 4.21 Obviously, the optimal rank-1 Hankel approximation in the spectral norm depends on the distribution of all eigenvalues of \mathbf{A} , as well as the structure of its eigenvectors. In particular, the optimal parameters \tilde{c} and \tilde{z} generating the optimal rank-1 Hankel approximation with respect to the spectral norm usually do not coincide with those parameters found for the Frobenius norm. This fact can be seen by means of Examples 3.8 and 4.12. See

Table 9.1 for a neat overview of the respective optimal parameters.

Remark 4.22 The theory of Adamjan, Arov, and Kreĭn (AAK theory) deals with the structured low-rank approximation problem for infinite matrices. From this theory, we learn that the optimal parameter \tilde{z} should be a root of the so-called Laurent series obtained from the infinite singular vector corresponding to the second largest singular value σ_1 , see for example [BM05; Pot17]. This is similar to what we do when checking whether the optimal error bound can be achieved: we inspect all roots of $v_1(z) = \mathbf{v}_1^\top \mathbf{z}$, which can be interpreted as the finite Laurent polynomial corresponding to the second singular vector.

Algorithm 4.3 Optimal rank-1 Hankel approximation w.r.t. the spectral norm for multiple largest eigenvalue

Input: Eigenvalues $\lambda_0 = |\lambda_1| \geq \dots \geq |\lambda_{N-1}| \geq 0$ and the corresponding normalized eigenvectors $\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_{N-1}$, threshold $\varepsilon > 0$.

Compute the set Σ^- of joint real roots \tilde{z} of the polynomials $v_j(z) = \mathbf{v}_j^\top \mathbf{z} = 0$ for all j with $\lambda_j = -\lambda_0$;

for $\tilde{z} \in \Sigma^-$ **do**

compute

$$\tilde{c} = \left(\sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^\top \tilde{\mathbf{z}})^2}{\lambda_0 + \lambda_j} \right)^{-1} > 0$$

add $(\tilde{\lambda} = |\lambda_0|, \tilde{c}, \tilde{z})$ to Solution;

compute the set Σ^+ of joint real roots of the polynomials $v_j(z) = \mathbf{v}_j^\top \mathbf{z} = 0$ for all j with $\lambda_j = +\lambda_0$;

for $\tilde{z} \in \Sigma^+$ **do**

compute


$$\tilde{c} = - \left(\sum_{j=0}^{N-1} \frac{(\mathbf{v}_j^\top \tilde{\mathbf{z}})^2}{\lambda_0 - \lambda_j} \right)^{-1} < 0$$

add $(\tilde{\lambda} = |\lambda_0|, \tilde{c}, \tilde{z})$ to Solution;

return Solution.

Output: List of tuples, each containing the optimal approximation error $\tilde{\lambda} = |\lambda_0|$ and parameters \tilde{c} and \tilde{z} generating an optimal rank-1 Hankel approximation of the matrix \mathbf{A} ,

empty when there is no optimal solution of true rank one, only the zero matrix.



BENCHMARKING STRUCTURED LOW-RANK APPROXIMATION METHODS

There are a variety of different optimization approaches that engage in the structured low-rank approximation (SIRA) problem for

“Different methods for solving the problem can be obtained by choosing different combinations of rank representation and optimization method”

—Ivan Markovsky [Mar19].

This part is to compare rank-1 Hankel approximations produced by different SLRA methods found in the literature. Thereby, the optimal solutions to the r1H problem with respect to the Frobenius and the spectral norm serve as benchmarks.

Before we come to the major comparison, we review three main approaches to deal with the structured low-rank approximation problem. These are

- approaches based on local optimization in a neighborhood of an initial value,
- Cadzow’s method, which is a heuristic application of alternating projections, and
- convex relaxation of the low-rank constraint achieved by the nuclear norm.

These methods are more universally applicable—both in terms of structure and desired rank of the approximation—than the ones presented in Part I of this thesis.

However, they exclusively find real structured approximations for real initial matrices. Methods explicitly suited for complex structured low-rank approximation of complex (or even real) matrices are unknown to the author. Therefore, complex approximations are excluded from this part.

Except for Cadzow’s method, which does not change for different norms (see Remark 7.3), all of the aforementioned focus on approximation with respect to the (weighted) Frobenius norm. Hence, the comparison of approximations in the spectral norm only takes up a small portion of this part.

Furthermore, in most cases, the relaxed rank constraint $\text{rank } \mathbf{H} \leq r$ is used in order to circumvent the issue of non-existence of a solution with fixed rank $\text{rank } \mathbf{H} = r$, see [CFP03]. For the relaxed rank constraint and linear matrix structures, the feasible set is non-empty and closed, thus a solution always exists.

In Chapter 5, we introduce a general representation of structured matrices. This representation will enable us to understand the SLRA methods in the subsequent chapters. Chapters 6 to 8 each discuss one of the three main approaches—local optimization, alternating projections, and convex relaxation. Chapter 6 actually contains two methods that are both based on local optimization but use different ways of expressing the rank constraint, see Sections 6.1 and 6.2.

We adapt all methods mentioned above to explicitly fit the r1H problem. This adaption is straightforward for the local optimization approaches (Chapter 6) and Cadzow’s method (Chapter 7). For Cadzow’s method, besides the mere adaption, we give a new convergence result in the setting of rank-1 Hankel approximation, see Section 7.2. The modification of the convex relaxation method for rank-1 Hankel approximation in Chapter 8 is more involved than for the other methods. It turns out that it is less suited to the task, too.

At the end of each chapter, we affiliate small examples by which we compare the method at hand to the optimal one.

Finally, Chapter 9 is dedicated to overall comparisons. In Section 9.1, we again summarize the small examples encountered throughout this dissertation. In Section 9.2, we run both more and larger-scale comparisons of the methods from Chapters 6 to 8. Our optimal rank-1 Hankel approximations from Part I always serve as benchmarks.

5

GENERAL AFFINE MATRIX STRUCTURE

The methods we are about to present and compare in this part are designed to deal with a vast variety of affine matrix structures. In order to be able to handle general affine matrix structures, we need the following definition from [Mar19] (earlier versions in [MVP05; MWV⁺06]). Proceeding from this general definition, we will always come back to Hankel structured matrices.

From now on in this part, we always assume that $2 \leq M \leq N$ without loss of generality. Otherwise we can consider the transpose of the matrix.

Definition 5.1 An affine matrix structure is given by the structure specification map \mathcal{S} from a parameter space \mathbb{R}^{n_p} to the space of affinely structured matrices in $\mathbb{R}^{M \times N}$. The structure specification map is defined by

$$\mathcal{S}: \mathbb{R}^{n_p} \rightarrow \mathbb{R}^{M \times N}, \quad \mathcal{S}(\mathbf{p}) = \mathbf{S}_{\text{affine}} + \sum_{k=0}^{n_p-1} p_k \cdot \mathbf{S}_k,$$

where p_k denotes the k -th entry of the parameter vector \mathbf{p} . The matrices $\mathbf{S}_{\text{affine}}$ and \mathbf{S}_k , $k = 0, \dots, n_p - 1$, are $(M \times N)$ matrices and form a basis of the structure in question. The number of parameters n_p is required to be minimal, where minimal is meant in the sense that $\text{image } \mathcal{S} = \{\mathcal{S}(\mathbf{p}) : \mathbf{p} \in \mathbb{R}^{n_p}\}$ cannot be represented with fewer than n_p parameters.

The Hankel structure can be represented in form of a structure specification map as introduced in Definition 5.1. This rather abstract definition becomes clearer when we

formulate it specifically for a Hankel matrix

$$\mathbf{H} = \begin{pmatrix} h_0 & \cdots & h_{M-1} & \cdots & h_{N-1} \\ \vdots & & \ddots & & \vdots \\ h_{M-1} & \cdots & h_{N-1} & \cdots & h_{M+N-2} \end{pmatrix} \in \mathbb{R}^{M \times N}.$$

Any Hankel matrix \mathbf{H} is completely determined by the entries of its first row and last column. Thus, our parameter vector

$$\mathbf{p} := \left(h_0 \quad h_1 \quad \cdots \quad h_{M+N-2} \right)^\top \in \mathbb{R}^{n_p},$$

is the vector of generically different entries of \mathbf{H} . Consequently, the dimension of the parameter space is given by $n_p = M + N - 1$. Furthermore, the Hankel structure depends linearly on the parameters, so that we have $\mathbf{S}_{\text{affine}} = \mathbf{0}$. The matrices \mathbf{S}_k , $k = 0, \dots, n_p - 1$ are taken as a basis for the space of Hankel matrices. More precisely, each \mathbf{S}_k is the matrix with ones on the k -th counter-diagonal (starting to count by zero in the top left corner of the matrix), see (5.1). Together we obtain the explicit Hankel structure specification $\mathcal{H} := \mathcal{H}_{M,N} := \mathcal{S}_{\text{Hankel}}$ of size $M \times N$ as the map $\mathcal{H}: \mathbb{R}^{n_p} \rightarrow \mathbb{R}^{M \times N}$,

$$\mathcal{H}(\mathbf{p}) = h_0 \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & \\ 0 & & \end{pmatrix} + h_1 \cdot \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & \\ 0 & & \end{pmatrix} + h_1 \cdot \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & \\ 1 & & \end{pmatrix} + \cdots = \mathbf{H}, \quad (5.1)$$

where the empty spaces in the matrices stand for the appropriate number of zeros.

Let $\text{vec } \mathbf{A}$ be the vectorization of a matrix obtained by stacking its columns on top of each other. It is convenient to define the structure specification matrix

$$\mathbf{S} := \left(\text{vec } \mathbf{S}_0 \quad \text{vec } \mathbf{S}_1 \quad \cdots \quad \text{vec } \mathbf{S}_{n_p-1} \right) \in \mathbb{R}^{MN \times n_p}.$$

Then, the structure specification map \mathcal{S} can also be written as a multiplication of the parameter vector with the structure specification matrix \mathbf{S} , namely

$$\text{vec } \mathcal{S}(\mathbf{p}) = \text{vec } \mathbf{S}_{\text{affine}} + \mathbf{S} \cdot \mathbf{p}, \quad \text{respectively} \quad \mathcal{S}(\mathbf{p}) = \mathbf{S}_{\text{affine}} + \text{vec}^{-1}(\mathbf{S} \cdot \mathbf{p}). \quad (5.2)$$

Remark 5.2 The minimality of n_p (see Definition 5.1) is necessary and sufficient for \mathbf{S}

having full column rank. In the case of Hankel structure, the matrix \mathbf{S} moreover consists only of zeros and ones, and there is at most one non-zero element in each row of \mathbf{S} . This means that any one entry of the Hankel matrix $\mathcal{H}(\mathbf{p})$ corresponds to at most one entry of the parameter vector \mathbf{p} .

The structure specification matrix \mathbf{S} is also useful to express the projection onto the space of structured matrices as the following lemma from [IUM14] shows.

Lemma 5.3 ([IUM14, Lemma 2.1]) *For a structure specification \mathcal{S} whose structure specification matrix \mathbf{S} has at most one non-zero element in each row, the orthogonal projection $\mathcal{P}_{\mathcal{S}}(\mathbf{A})$ of a matrix \mathbf{A} onto image \mathcal{S} is given by*

$$\mathcal{P}_{\mathcal{S}}(\mathbf{A}) = \mathcal{S}(\mathbf{S}^{\dagger} \cdot \text{vec } \mathbf{A}),$$

where $\mathbf{S}^{\dagger} := (\mathbf{S}^{\top} \mathbf{S})^{-1} \mathbf{S}^{\top}$ is the Moore-Penrose pseudoinverse of \mathbf{S} .

Remark 5.4 1. The above lemma can be interpreted as follows: Multiplying the vectorized version of an arbitrary matrix \mathbf{A} by the pseudoinverse \mathbf{S}^{\dagger} extracts the parameter vector of its projection onto the structured subspace. More precisely, we have $\mathbf{S}^{\dagger} \cdot \text{vec } \mathbf{A} = \mathbf{p}_{\mathbf{A}}$ if and only if $\mathcal{P}_{\mathcal{S}}(\mathbf{A}) = \mathcal{S}(\mathbf{p}_{\mathbf{A}})$ for some parameter vector $\mathbf{p}_{\mathbf{A}} \in \mathbb{R}^{n_p}$.

2. Combining (5.2) and Lemma 5.3 we also have

$$\text{vec}(\mathcal{P}_{\mathcal{S}}(\mathbf{A})) = \text{vec } \mathbf{S}_{\text{affine}} + \mathbf{\Pi}_{\mathbf{S}} \cdot \text{vec } \mathbf{A},$$

where $\mathbf{\Pi}_{\mathbf{S}} = \mathbf{S} \cdot \mathbf{S}^{\dagger} = \mathbf{S} \cdot (\mathbf{S}^{\top} \mathbf{S})^{-1} \mathbf{S}^{\top}$ is the orthogonal projector onto the image of \mathbf{S} . Thus, $\mathbf{\Pi}_{\mathbf{S}}$ is related to the Hankel projection matrix \mathbf{P} from (1.9), except that for the definition of \mathbf{P} we have used a different diagonalization pattern, see Section 1.3.

With the representation of structured matrices introduced in Definition 5.1, it makes sense to consider the structured low-rank approximation (SLRA) problem entirely in terms of the parameter vectors. Unfortunately, in terms of the parameter vectors, we can only express (weighted) norms that are calculated from the matrix (or vector) entries, such as the Frobenius norm. The spectral norm is not accessible that way.

Nonetheless, we state the SLRA problem (1) with respect to some weighted norm as follows. For a structured initial matrix $\mathbf{A} := \mathcal{S}(\mathbf{p}_{\mathbf{A}})$, find

$$\min_{\mathbf{p} \in \mathbb{R}^{n_p}} \|\mathbf{p}_{\mathbf{A}} - \mathbf{p}\|_W^2 \quad \text{subject to } \text{rank } \mathcal{S}(\mathbf{p}) \leq r. \quad (5.3)$$

Note that with this problem formulation, only structured matrices can be approximated as opposed to the general matrix \mathbf{A} from Chapter 3. This is not a limitation, however, because we can replace a general unstructured matrix \mathbf{A} with its Hankel projection $\mathcal{P}(\mathbf{A})$ in a preprocessing step and define $\mathbf{p}_A := \mathbf{S}^\dagger \cdot \text{vec } \mathbf{A}$, see Remark 5.4. As observed in Remark 3.2.3, it does not change the optimal solution from Chapter 3 whether we approximate an arbitrary matrix \mathbf{A} or its Hankel projection $\mathcal{P}(\mathbf{A})$.

In (5.3), $\|\cdot\|_W$ is the weighted norm

$$\|\mathbf{p}\|_W^2 := \mathbf{p}^\top \cdot \mathbf{W} \cdot \mathbf{p},$$

where $\mathbf{W} \in \mathbb{R}^{n_p \times n_p}$ is a positive definite weight matrix.

We find the following connection between the weighted norm of the parameter vectors and the Frobenius norm of the structured matrices

$$\|\mathcal{S}(\mathbf{p}_A) - \mathcal{S}(\mathbf{p})\|_F^2 = \|\mathbf{S} \cdot (\mathbf{p}_A - \mathbf{p})\|_2^2 = (\mathbf{p}_A - \mathbf{p})^\top \mathbf{S}^\top \cdot \mathbf{S} (\mathbf{p}_A - \mathbf{p}) = \|\mathbf{p}_A - \mathbf{p}\|_W^2,$$

where in the last step we take $\mathbf{W} = \mathbf{S}^\top \mathbf{S}$. Since \mathbf{S} has full column rank (see Remark 5.2), thus defined \mathbf{W} is indeed positive definite. In particular, the Frobenius norm of a Hankel matrix $\mathcal{H}_{M,N}(\mathbf{p}) \in \mathbb{R}^{M \times N}$ is calculated with the diagonal weight matrix

$$\mathbf{W} = \text{diag}(1, 2, \dots, M, \dots, M, \dots, 2, 1) \in \mathbb{R}^{n_p \times n_p}, \quad (5.4)$$

where the term M occurs $N - M + 1$ times.

The rank-1 Hankel approximation (r1H) problem (3.1) with respect to the Frobenius norm can thus equivalently be formulated as follows. For an initial Hankel matrix $\mathbf{A} = \mathcal{H}_{M,N}(\mathbf{p}_A) \in \mathbb{R}^{M \times N}$, find

$$\min_{\mathbf{p} \in \mathbb{R}^{n_p}} \|\mathbf{p}_A - \mathbf{p}\|_W^2 \quad \text{subject to } \text{rank } \mathcal{H}_{M,N}(\mathbf{p}) = 1, \quad (5.5)$$

with the Hankel structure specification (5.1) and weight matrix (5.4).

Remark 5.5 Employing the structure specification from Definition 5.1, various structures other than the Hankel structure can be expressed. Among those, Toeplitz matrices readily come to mind. However, much more complex matrix structures also belong to that pattern, including block Hankel, block Toeplitz, and mosaic Hankel structures. The interested reader may be referred to [Hei95; UM14].

6

LOCAL OPTIMIZATION

We consider the structured low-rank approximation (SLRA) problem with respect to the (weighted) Frobenius norm. Here, only real structured initial matrices \mathbf{A} are approximated by real matrices \mathbf{H} of the same structure and lower rank. Recall in contrast Section 3.1, where we dealt with arbitrary complex initial matrices and their complex rank-1 Hankel approximations.

In this chapter, we address two different approaches to solve the real SLRA problem, which are both based on local optimization techniques. The main difference lies in their respective manner of expressing the low-rank constraint. This is also how this chapter splits into sections.

As mentioned earlier, the methods in this chapter exhibit great variability in terms of the targeted rank, matrix structure, and norm weight. Besides, they can deal with missing elements in the data as well as fixed elements in the approximating matrix, see [IUM14; UM14; UM19]. We will, however, concentrate on plain rank-1 Hankel approximation with respect to the Frobenius norm in order to enable a useful comparison with our results from Chapter 3.

6.1 Kernel Representation of the Rank Constraint

This section is based on [Mar19; UM14]. The elaborations therein are summarized and simplified for scalar matrix structures (opposed to block/mosaic matrix structures). For the examples we used the implementations [MU14; UM19] in MATLAB.

In this section, we use the following representation of the low-rank constraint. As introduced in Lemma 1.4 part (3), we have the equivalence

$$\text{rank } \mathbf{H} \leq r \quad \Leftrightarrow \quad \mathbf{R} \cdot \mathbf{H} = \mathbf{0} \quad \text{for some matrix } \mathbf{R} \in \mathbb{R}^{d \times M} \text{ with full row rank,}$$

where $d := M - r$ is the rank reduction, and $\mathbf{H} \in \mathbb{R}^{M \times N}$ with $2 \leq M \leq N$. This formulation is called the *kernel representation* of the rank constraint.

For future reference, the local optimization method using the kernel representation of the rank constraint, which we present here, will also simply be called kernel method. This abbreviation is done in order to ease both notation and readability.

Using the kernel representation and the notations from Chapter 5, the SLRA problem (5.3) reads

$$\begin{aligned} \min_{\substack{\mathbf{p} \in \mathbb{R}^{n_p} \\ \mathbf{R} \in \mathbb{R}^{d \times M}}} \|\mathbf{p}_A - \mathbf{p}\|_W^2 & \quad \text{subject to } \text{rank } \mathbf{R} = d \\ & \quad \text{and } \mathbf{R} \cdot \mathcal{S}(\mathbf{p}) = \mathbf{0}, \end{aligned} \quad (6.1)$$

where $r = M - d$ is the desired rank of the approximation.

Problem (6.1) can equivalently be written as a double minimization problem

$$\min_{\mathbf{R} \in \mathbb{R}^{d \times M}} f(\mathbf{R}) \quad \text{subject to } \text{rank } \mathbf{R} = d, \quad (6.2a)$$

$$f(\mathbf{R}) = \left(\min_{\mathbf{p} \in \mathbb{R}^{n_p}} \|\mathbf{p}_A - \mathbf{p}\|_W^2 \quad \text{subject to } \mathbf{R} \cdot \mathcal{S}(\mathbf{p}) = \mathbf{0} \right), \quad (6.2b)$$

where (6.2a) is called the outer minimization and (6.2b) is called inner minimization.

With a change of variables, the inner minimization problem (6.2b) can be rewritten as a so-called least-norm problem with respect to the Euclidean vector norm. We introduce as a new variable the weighted difference vector

$$\Delta \mathbf{p} := \mathbf{W}^{1/2} \cdot (\mathbf{p}_A - \mathbf{p}). \quad (6.3)$$

For diagonal weight matrices, $\mathbf{W}^{1/2}$ is defined as the diagonal matrix of square roots of the original entries. If the weight matrix is a more general positive definite matrix, $\mathbf{W}^{1/2}$ shall denote a Cholesky factor of the weight matrix $\mathbf{W} \in \mathbb{R}^{n_p \times n_p}$. Summarized we can say $\mathbf{W}^{1/2}$ is a matrix satisfying $\mathbf{W} = (\mathbf{W}^{1/2})^\top \cdot \mathbf{W}^{1/2}$.

With (6.3) the weighted norm of the difference between the parameter vectors becomes

$$\|\mathbf{p}_A - \mathbf{p}\|_W^2 = (\mathbf{p}_A - \mathbf{p})^\top \cdot \mathbf{W} \cdot (\mathbf{p}_A - \mathbf{p}) = (\mathbf{p}_A - \mathbf{p})^\top \left(\mathbf{W}^{1/2} \right)^\top \cdot \mathbf{W}^{1/2} (\mathbf{p}_A - \mathbf{p}) = \|\Delta \mathbf{p}\|_2^2,$$

the ordinary Euclidean vector norm of $\Delta \mathbf{p}$.

For an $(m \times n)$ matrix $\mathbf{A} = (a_{jk})_{j,k=0}^{m-1,n-1}$ and a $(p \times q)$ matrix \mathbf{B} , the Kronecker product is defined as the $(mp \times nq)$ block matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{0,0} \cdot \mathbf{B} & \cdots & a_{0,n-1} \cdot \mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m-1,0} \cdot \mathbf{B} & \cdots & a_{m-1,n-1} \cdot \mathbf{B} \end{pmatrix}.$$

With this definition, we further define

$$\mathbf{s} := \mathbf{s}(\mathbf{R}) := \text{vec}(\mathbf{R} \cdot \mathcal{S}(\mathbf{p}_A)) = (\mathbf{I}_N \otimes \mathbf{R}) \cdot \text{vec} \mathcal{S}(\mathbf{p}_A) \in \mathbb{R}^{dN}$$

and

$$\begin{aligned} \mathbf{G} &:= \mathbf{G}(\mathbf{R}) := (\mathbf{I}_N \otimes \mathbf{R}) \cdot \mathbf{S} \cdot \mathbf{W}^{-1/2} \\ &= \left(\text{vec}(\mathbf{R} \cdot \mathbf{S}_0) \cdots \text{vec}(\mathbf{R} \cdot \mathbf{S}_{n_p-1}) \right) \cdot \mathbf{W}^{-1/2} \in \mathbb{R}^{dN \times n_p}, \end{aligned}$$

where $\mathbf{W}^{-1/2}$ denotes the inverse of $\mathbf{W}^{1/2}$. This inverse exists since \mathbf{W} is positive definite.

Now the inner minimization problem (6.2b) can be rewritten as

$$\min_{\Delta \mathbf{p} \in \mathbb{R}^{n_p}} \|\Delta \mathbf{p}\|_2^2 \quad \text{subject to } \mathbf{G} \cdot \Delta \mathbf{p} = \mathbf{s}. \quad (6.4)$$

This is a least-norm problem in standard form [BV04, Chap. 6]. As such, it has an analytic closed-form solution.

If the matrix $\mathbf{G} = \mathbf{G}(\mathbf{R})$ has full row rank, then

$$\mathbf{\Gamma} := \mathbf{\Gamma}(\mathbf{R}) := \mathbf{G}(\mathbf{R}) \cdot \mathbf{G}(\mathbf{R})^\top \in \mathbb{R}^{dN \times dN}$$

is invertible, and the solution of the inner problem (6.2b) is given by

$$\Delta \mathbf{p}_{\text{ker}} := \mathbf{G}^\top \cdot \mathbf{\Gamma}^{-1} \cdot \mathbf{s} \quad (6.5)$$

and

$$f(\mathbf{R}) = \|\Delta \mathbf{p}_{\text{ker}}\|_2^2 = \mathbf{s}^\top \cdot \mathbf{\Gamma}^{-1} \cdot \mathbf{s}. \quad (6.6)$$

Therein, the solution to problem (6.1), $\mathbf{p}_{\text{ker}} = \mathbf{p}_{\text{ker}}(\mathbf{R})$, is defined in terms of \mathbf{R} since the matrices \mathbf{G} , $\mathbf{\Gamma}$, and the vector \mathbf{s} depend on \mathbf{R} . More precisely, the solution parameter vector

is given by

$$\mathbf{p}_{\text{ker}} := \mathbf{p}_A - \mathbf{W}^{-1/2} \cdot \Delta \mathbf{p}_{\text{ker}} = \mathbf{p}_A - \mathbf{W}^{-1/2} \cdot \mathbf{G}^\top \cdot \mathbf{\Gamma}^{-1} \cdot \mathbf{s}$$

via (6.3) and (6.5).

Thus, the variable \mathbf{p} (respectively $\Delta \mathbf{p}$) is eliminated from problem (6.1). As a consequence, the double minimization problem (6.2) reduces to a problem in the matrix variable \mathbf{R} only:

$$\min_{\mathbf{R} \in \mathbb{R}^{d \times M}} f(\mathbf{R}) \quad \text{subject to } \text{rank } \mathbf{R} = d, \quad (6.7)$$

where $f(\mathbf{R})$ is given by (6.6). For more details on the equivalence of problems (5.3), (6.1), and (6.7) see [MWV⁺06, Chap. 4; Mar19, Chap. 4; UM14].

The reduced outer problem (6.7) can now be solved by standard local optimization methods such as MATLAB's `fmincon`. All local optimization methods need an initial point to begin with. In the software presented in [MU14; UM19], the unstructured low-rank approximation \mathbf{A}_r of \mathbf{A} from Theorem 1.8 is used as a starting point. The resulting initial value for \mathbf{R} is a full row rank matrix $\mathbf{R}_{\text{lra}} \in \mathbb{R}^{d \times M}$ such that $\mathbf{R}_{\text{lra}} \cdot \mathbf{A}_r = \mathbf{0}$. By default, the full row rank condition on \mathbf{R} is imposed as the constraint $\mathbf{R} \cdot \mathbf{R}^\top = \mathbf{I}_d$, see [MU14].

In our case, namely the case of the r1H problem, the rank reduction $d = M - 1$ is very large. Hence, we have $n_p = M + N - 1 \leq dN$ for $M > 2$, which implies that $\mathbf{G} \in \mathbb{R}^{dN \times n_p}$ cannot have full row rank. Consequently, $\mathbf{\Gamma}$ is not invertible. If in that case the least-norm problem (6.4) is feasible nevertheless, it still has an analytic solution. This solution is given by replacing the inverse of $\mathbf{\Gamma}$ in (6.5) and (6.6) by its Moore-Penrose pseudoinverse $\mathbf{\Gamma}^\dagger$, see [UM14].

However, the implementation presented in [MU14; UM19] requires $\mathbf{\Gamma}$ to actually be invertible. In [MU14, Note 3], it is explicitly noted that for Hankel structured low-rank approximation, the rank reduction d can be at most one. In Examples 6.1 and 6.2 we perform rank-1 Hankel approximation on full rank matrices of sizes (4×4) and (3×3) , respectively. So in both cases, we clearly have $d \geq 2$ for the rank reduction. Indeed, when running the unaltered MATLAB code [MU14; UM19] on our example matrices, it returns approximation matrices with entries about 10^{-14} for Example 6.1; and even 10^{-16} for Example 6.2. This order of magnitude is so close to zero that we cannot consider the rank-1 Hankel approximation successful.

The restriction that the rank reduction d for Hankel matrices can be at most one, is quite severe. A workaround to overcome this limitation is given by [HR84, Prop. 5.4]. For

$r < M \leq N$ and $n_p = M + N - 1$, the equivalence

$$\text{rank } \mathcal{H}_{M,N}(\mathbf{p}) \leq r \quad \Leftrightarrow \quad \text{rank } \mathcal{H}_{r+1,n_p-r}(\mathbf{p}) \leq r$$

follows from said proposition.

Thus, the rank constraint of any Hankel structured low-rank approximation can be recast into an equivalent constraint on a reshaped Hankel matrix of size $(r + 1) \times (n_p - r)$. Thereby, the rank reduction is reduced to $d = 1$. The r1H problem in the Frobenius norm with reshaped rank constraint is

$$\min_{\mathbf{p} \in \mathbb{R}^{n_p}} \|\mathbf{p}_A - \mathbf{p}\|_W^2 \quad \text{subject to } \text{rank } \mathcal{H}_{2,n_p-1}(\mathbf{p}) = 1. \quad (6.8)$$

Here, the norm weight remains $\mathbf{W} = \text{diag}(1, 2, \dots, M, \dots, M, \dots, 2, 1)$ from (5.4) with the term M occurring $N - M + 1$ times. This is the weight matrix corresponding to the Frobenius norm of the Hankel matrix $\mathcal{H}_{M,N}$ in the original shape. Leaving the weight matrix unchanged with respect to the original problem (5.5) is essential in order to maintain equivalence of (6.8) and the original problem. Now, in the formulation (6.8), the r1H problem can be solved by the MATLAB code [MU14; UM19].

We conclude this section with two small examples illustrating the accuracy of the rank-1 Hankel approximation in the Frobenius norm obtained by the kernel method.

The approximation errors from these examples are again summarized in Tables 9.2 and 9.4 in Chapter 9. There, they are compared to the errors produced by the approximation methods presented in Section 6.2 and Chapters 7 and 8. As benchmarks, the minimal approximation errors with respect to the Frobenius norm from Chapter 3 are shown in that comparison as well.

Example 6.1 Consider the following Hankel matrix and its corresponding parameter vector

$$\mathbf{A} = \begin{pmatrix} 3 & 2 & 1 & 1 \\ 2 & 1 & 1 & 2 \\ 1 & 1 & 2 & 5 \\ 1 & 2 & 5 & 2 \end{pmatrix} \quad \text{and} \quad \mathbf{p}_A = (3 \ 2 \ 1 \ 1 \ 2 \ 5 \ 2)^\top.$$

We use the MATLAB implementation [MU14; UM19]. At that, we impose the rank constraint on the reshaped Hankel matrix $\mathcal{H}_{2,6}(\mathbf{p})$ instead of $\mathcal{H}_{4,4}(\mathbf{p})$.

The resulting output is the parameter vector

$$\mathbf{p}_{\text{ker}} \approx \left(1.02 \quad 1.25 \quad 1.53 \quad 1.88 \quad 2.30 \quad 2.82 \quad 3.46 \right)^\top,$$

from which we form the Hankel matrix

$$\mathbf{H}_{\text{ker}} = \mathcal{H}_{4,4}(\mathbf{p}_{\text{ker}}) \approx \begin{pmatrix} 1.02 & 1.25 & 1.53 & 1.88 \\ 1.25 & 1.53 & 1.88 & 2.30 \\ 1.53 & 1.88 & 2.30 & 2.82 \\ 1.88 & 2.30 & 2.82 & 3.46 \end{pmatrix}$$

as solution of the r1H problem (6.8). The entries of the solution vector and matrix have been rounded to two decimal digits. The rank of the solution matrix is indeed $\text{rank } \mathbf{H}_{\text{ker}} = 1$.

We find that this solution is very close to the optimal one found by the method we developed in Chapter 3. Indeed, a closer examination reveals that the maximal elementwise deviation of \mathbf{p}_{ker} from the optimal parameter vector $\tilde{\mathbf{p}} := \mathbf{S}^\dagger \cdot \text{vec}(\tilde{\mathbf{c}} \cdot \tilde{\mathbf{z}}_M \tilde{\mathbf{z}}_N^\top)$ only lies in the sixth decimal digit, in fact, $\|\mathbf{p}_{\text{ker}} - \tilde{\mathbf{p}}\|_\infty \approx 10^{-6}$.

The approximation errors produced by the kernel method are $\|\mathbf{p}_A - \mathbf{p}_{\text{ker}}\|_2 \approx 3.5355$ in terms of parameter vectors, and $\|\mathbf{A} - \mathbf{H}_{\text{ker}}\|_F \approx 4.5685$ in the Frobenius norm. The respective relative approximation errors are given by $\|\mathbf{p}_A - \mathbf{p}_{\text{ker}}\|_2 / \|\mathbf{p}_A\|_2 \approx 0.5103$ and $\|\mathbf{A} - \mathbf{H}_{\text{ker}}\|_F / \|\mathbf{A}\|_F \approx 0.4816$. All errors have been rounded to four decimal digits and are thus indistinguishable from the minimal errors, see Table 9.4.

Example 6.2 Consider the matrix known from Examples 3.8 and 4.12 and its corresponding parameter vector

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \\ 1/2 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{p}_A = \left(1 \quad 0 \quad 1/2 \quad 0 \quad 1 \right)^\top.$$

With [MU14; UM19] we obtain the parameter vector and corresponding approximation matrix

$$\mathbf{p}_{\text{ker}} = \left(1 \quad 0 \quad 0 \quad 0 \quad 0 \right)^\top \quad \text{and} \quad \mathbf{H}_{\text{ker}} = \mathcal{H}_{3,3}(\mathbf{p}_{\text{ker}}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Of course, the approximation matrix \mathbf{H}_{ker} is of rank one.

Let us compare the above parameter vector \mathbf{p}_{ker} to the optimal parameter vector $\tilde{\mathbf{p}} = \mathbf{S}^\dagger \cdot \text{vec}(\tilde{\mathbf{c}} \cdot \tilde{\mathbf{z}}_M \tilde{\mathbf{z}}_N^\top) = 7/18 \cdot (1 \ 1 \ 1 \ 1 \ 1)^\top$ from Example 3.8. The structural difference is quite pronounced even at first glance. More rigorously, the deviation $\|\mathbf{p}_{\text{ker}} - \tilde{\mathbf{p}}\|_\infty = 7/18 \approx 0.3889$ is substantially larger than in Example 6.1.

The approximation errors (rounded to four digits) are $\|\mathbf{p}_A - \mathbf{p}_{\text{ker}}\|_2 \approx 1.1180$ in terms of parameter vectors and $\|\mathbf{A} - \mathbf{H}_{\text{ker}}\|_F \approx 1.3229$ in the Frobenius norm. The respective relative approximation errors (also rounded to four digits) are given by $\|\mathbf{p}_A - \mathbf{p}_{\text{ker}}\|_2 / \|\mathbf{p}_A\|_2 \approx 0.7454$ and $\|\mathbf{A} - \mathbf{H}_{\text{ker}}\|_F / \|\mathbf{A}\|_F \approx 0.7977$.

The absolute and relative approximation errors of both examples can also be found in Tables 9.2 and 9.4. There, they are arranged in order to facilitate the comparison to the approximation errors produced by the other approximation methods presented in the sequel.

6.2 Image Representation of the Rank Constraint

In this section, another formulation of the rank constraint is used. It is given by Lemma 1.4 part (4) as follows. For $\mathbf{H} \in \mathbb{R}^{M \times N}$, we have the equivalence

$$\text{rank } \mathbf{H} \leq r \quad \Leftrightarrow \quad \mathbf{H} = \mathbf{P} \cdot \mathbf{L} \quad \text{for some matrices } \mathbf{P} \in \mathbb{R}^{M \times r}, \mathbf{L} \in \mathbb{R}^{r \times N}.$$

The factorization of a low-rank matrix into a product of two matrices with smaller dimensions is called *image representation* of the low-rank constraint. It is widely used in methods for unstructured low-rank approximation, see for example [FXG18; Gol65; HC03; Ste99]. However, adopting it to structured low-rank approximation is not easy [CFP03]. The main difficulty is to impose the structure via the factors \mathbf{P} and \mathbf{L} .

Using the image representation of the rank constraint, the SLRA problem (5.3) reads

$$\min_{\substack{\mathbf{P} \in \mathbb{R}^{M \times r} \\ \mathbf{L} \in \mathbb{R}^{r \times N}}} \|\mathbf{A} - \mathbf{P}\mathbf{L}\|_F^2 \quad \text{subject to } \mathbf{P}\mathbf{L} = \mathcal{S}(\mathbf{p}) \quad (6.9)$$

for some parameter vector $\mathbf{p} \in \mathbb{R}^{n_p}$.

Problem (6.9) is addressed in [IUM14], where the image representation of the rank constraint is also termed matrix factorization approach. In the remainder of this thesis, we will also call it the image method for short. With this name, we distinguish it from the kernel method presented in Section 6.1.

Remark 6.3 In the factorization of the low-rank matrix, the image representation approach is similar to our approach using the structured vectors \mathbf{z}_M and \mathbf{z}_N , compare Chapters 2 and 3. The striking difference, however, is that here, the structure is not directly encoded in the matrix factors but left as a constraint.

Now, we will summarize the factorization approach (or image method) given in [IUM14] and apply it to rank-1 Hankel approximation. The proposal is to use regularized minimization in order to iteratively satisfy both the low-rank and the structure constraint. One of the constraints (i.e., low-rank or structure) is imposed directly and is thus satisfied at each iteration step. The deviation from the remaining requirement is included in the problem as regularization term. This requirement will be achieved only upon convergence.

With the regularized formulation, the constrained minimization problem (5.3), respectively (6.9), becomes an unconstrained problem. As such, for each fixed regularization parameter γ the problem can be solved easily. But solvability comes at the cost of one constraint being satisfied only approximately.

There are two possibilities as of which constraint to regularize:

- penalize the deviation from the desired structure by solving

$$\min_{\substack{\mathbf{P} \in \mathbb{R}^{M \times r} \\ \mathbf{L} \in \mathbb{R}^{r \times N}} \|\mathbf{A} - \mathbf{P}\mathbf{L}\|_F^2 + \gamma \cdot \|\mathbf{P}\mathbf{L} - \mathcal{P}_S(\mathbf{P}\mathbf{L})\|_F^2, \quad (6.10)$$

where the iterate $\mathbf{P}\mathbf{L}$ is always of the desired rank, while the structure is achieved gradually as γ increases, or

- penalize the deviation from the low-rank requirement by

$$\min_{\substack{\mathbf{P} \in \mathbb{R}^{M \times r} \\ \mathbf{L} \in \mathbb{R}^{r \times N}} \|\mathbf{A} - \mathcal{P}_S(\mathbf{P}\mathbf{L})\|_F^2 + \gamma \cdot \|\mathcal{P}_S(\mathbf{P}\mathbf{L}) - \mathbf{P}\mathbf{L}\|_F^2, \quad (6.11)$$

where the iterate $\mathcal{P}_S(\mathbf{P}\mathbf{L})$ possesses the desired structure in each step but the low-rank constraint is satisfied only upon convergence as γ increases.

Note that what is considered the current iterate differs between (6.10) and (6.11).

With \mathcal{P}_S , we denote the projection onto the affine space of structured matrices, where the structure is given by the specification \mathcal{S} , see Definition 5.1. The coefficient γ is a regularization parameter (or penalization parameter) which is increased for higher iterations

in order to have both requirements (approximately) fulfilled. Particularly, for $\gamma = \infty$ the regularization term $\|\mathbf{PL} - \mathcal{P}_S(\mathbf{PL})\|_F^2$ has to be zero and problems (6.9)–(6.11) are equivalent. The specific choice of γ throughout the iterations is discussed below.

Assume now that $\mathbf{A} = \mathcal{S}(\mathbf{p}_A)$ is itself a structured matrix, or that \mathbf{A} has been projected onto the space of structured matrices in a preprocessing step, that is, $\mathbf{p}_A = \mathbf{S}^\dagger \cdot \text{vec } \mathbf{A}$, see Lemma 5.3 and Remark 5.4. Then problem (6.11) can equivalently be formulated only using parameter vectors

$$\min_{\substack{\mathbf{P} \in \mathbb{R}^{M \times r} \\ \mathbf{L} \in \mathbb{R}^{r \times N}} \left\| \mathbf{p}_A - \mathbf{S}^\dagger \cdot \text{vec}(\mathbf{PL}) \right\|_W^2 + \gamma \cdot \|\mathbf{PL} - \mathcal{P}_S(\mathbf{PL})\|_F^2, \quad (6.12)$$

where for the weight matrix we have $\mathbf{W} = \mathbf{S}^\top \mathbf{S}$; for example, choose \mathbf{W} as in (5.4) for Hankel matrices.

Remark 6.4 Originally, the regularized factorization approach [IUM14] was proposed for a more general weighted matrix norm $\|\mathbf{A}\|_W^2 = \text{vec } \mathbf{A} \cdot \mathbf{W}_M \cdot \text{vec } \mathbf{A}$, where \mathbf{W}_M is a positive definite weight matrix. But we are only interested in the Frobenius norm which coincides with the weighted norm for $\mathbf{W}_M = \mathbf{I}$.

Equivalence of (6.10) and a vectorized version such as (6.12) only holds if the weight matrix satisfies $\mathbf{W}_M = (\mathbf{S}^\dagger)^\top \cdot \mathbf{W} \cdot \mathbf{S}^\dagger$. This is not the case for the Frobenius norm ($\mathbf{W}_M = \mathbf{I}$), see [IUM14] for more details.

In contrast, equivalence of (6.11) and its vectorized version (6.12) holds for any weighted vectorized matrix norm as long as \mathbf{W}_M and \mathbf{W} comply with the relation $\mathbf{W} = \mathbf{S}^\top \cdot \mathbf{W}_M \cdot \mathbf{S}$. That is why, from now on, we focus on (6.11), respectively (6.12).

The key element in the approach [IUM14] is to alternately improve the approximations for \mathbf{L} and \mathbf{P} while the respective other one remains fixed. More precisely, one has to solve the following two optimization problems.

For fixed \mathbf{P} solve

$$\min_{\mathbf{L} \in \mathbb{R}^{r \times N}} \left\| \mathbf{p}_A - \mathbf{S}^\dagger \cdot \text{vec}(\mathbf{PL}) \right\|_W^2 + \gamma \cdot \|\mathbf{PL} - \mathcal{P}_S(\mathbf{PL})\|_F^2 \quad (6.13a)$$

and for fixed \mathbf{L} solve

$$\min_{\mathbf{P} \in \mathbb{R}^{M \times r}} \left\| \mathbf{p}_A - \mathbf{S}^\dagger \cdot \text{vec}(\mathbf{PL}) \right\|_W^2 + \gamma \cdot \|\mathbf{PL} - \mathcal{P}_S(\mathbf{PL})\|_F^2. \quad (6.13b)$$

These minimization problems in one variable are equivalent to least squares problems in standard Euclidean norm as the next lemma demonstrates.

Lemma 6.5 ([IUM14, Lemma 4.1]) *The minimization problems (6.13) are equivalent to the least squares problems*

$$(6.13a) \Leftrightarrow \min_{\mathbf{L} \in \mathbb{R}^{r \times N}} \left\| \begin{pmatrix} \mathbf{W}^{1/2} \cdot \mathbf{S}^\dagger \\ \sqrt{\gamma} \cdot \mathbf{\Pi}_{\mathbf{S}_\perp} \end{pmatrix} (\mathbf{I}_N \otimes \mathbf{P}) \cdot \text{vec}(\mathbf{L}) - \begin{pmatrix} \mathbf{W}^{1/2} \cdot \mathbf{p}_A \\ \sqrt{\gamma} \cdot \text{vec}(\mathbf{S}_{\text{affine}}) \end{pmatrix} \right\|_2^2, \quad (6.14a)$$

$$(6.13b) \Leftrightarrow \min_{\mathbf{P} \in \mathbb{R}^{M \times r}} \left\| \begin{pmatrix} \mathbf{W}^{1/2} \cdot \mathbf{S}^\dagger \\ \sqrt{\gamma} \cdot \mathbf{\Pi}_{\mathbf{S}_\perp} \end{pmatrix} (\mathbf{L}^\top \otimes \mathbf{I}_M) \cdot \text{vec}(\mathbf{P}) - \begin{pmatrix} \mathbf{W}^{1/2} \cdot \mathbf{p}_A \\ \sqrt{\gamma} \cdot \text{vec}(\mathbf{S}_{\text{affine}}) \end{pmatrix} \right\|_2^2, \quad (6.14b)$$

where \otimes denotes the Kronecker product. The matrix $\mathbf{\Pi}_{\mathbf{S}_\perp} := \mathbf{I}_{MN} - \mathbf{S}\mathbf{S}^\dagger$ is the orthogonal projector onto the left kernel of \mathbf{S} . As before in Section 6.1, $\mathbf{W}^{1/2} \in \mathbb{R}^{n_p \times n_p}$ denotes a matrix satisfying $\mathbf{W} = (\mathbf{W}^{1/2})^\top \cdot \mathbf{W}^{1/2}$.

For the proof see [IUM14].

The advantage of formulating (6.13) as the least squares problems (6.14) is that the latter have closed-form solutions. This results in the double iteration that is summarized in Algorithm 6.1. An implementation can be found in [UM19].

Algorithm 6.1 Structured low-rank approximation by factorization

Input: Parameter vector $\mathbf{p}_A \in \mathbb{R}^{n_p}$ of an initial Hankel matrix \mathbf{A} , desired structure \mathcal{S} , desired rank r , initial value $\mathbf{P}_0 \in \mathbb{R}^{M \times r}$ for the left factor \mathbf{P} .

Set $\mathbf{P} = \mathbf{P}_0$, $\gamma_1 = 1$;

for $j = 1, 2, \dots$ *until a stopping criterion is satisfied* **do**

for $k = 1, 2, \dots$ *until a stopping criterion is satisfied* **do**

 update \mathbf{L} from (6.14a),

 update \mathbf{P} from (6.14b).

 Set γ_{j+1} such that $\gamma_{j+1} > \gamma_j$.

Output: Factors $\mathbf{P} \in \mathbb{R}^{M \times r}$ and $\mathbf{L} \in \mathbb{R}^{r \times N}$ corresponding to a structured rank- r approximation of \mathbf{A} .

As any local optimization method, Algorithm 6.1 needs an initial value. Per default [UM19], the matrix consisting of the first r left singular values of \mathbf{A} is used as initial value \mathbf{P}_0 . More precisely, if $\mathbf{U}_r \Sigma_r \mathbf{V}_r^\top$ is the optimal unstructured rank- r approximation of \mathbf{A} given by Theorem 1.8, then $\mathbf{P}_0 := \mathbf{U}_r$ is taken as initialization. This is a sensible choice since

for $\gamma = 0$ problems (6.11) and (6.12) are equivalent to the unstructured LRA problem (1.1). For the latter, the truncated SVD provides a globally optimal solution.

Recall that for the theoretical value $\gamma = \infty$, problem (6.12) is exactly the SLRA problem (5.3), respectively (6.9). Practically we can fix γ to a large value and obtain a solution $\mathcal{P}_S(\mathbf{P}\mathbf{L})$ that is approximately a low-rank matrix. But since fixed large values of γ may lead to numerical issues, an adaptive updating scheme for γ is applied. Roughly, it operates as follows: If the inner loop in Algorithm 6.1 has been expensive to carry out, increase γ only moderately. If, on the other hand, carrying out the inner loop has been cheap, increase γ more ambitiously. For more theoretical details see [IUM14; NW06], and [UM19] for the practical implementation.

In the adaptive updating scheme for γ we start with a small regularization parameter. Then each further iteration is initialized with the solution of the previous iteration. This procedure provides a good initial value for each step.

Concerning the stopping criteria, the iteration of the inner loop in Algorithm 6.1 is terminated and γ increased when there is only little change in the column space of \mathbf{P} . The outer loop is stopped when γ reaches a specified large threshold. See [IUM14; UM19] for more details.

Remark 6.6 Unlike the kernel method from Section 6.1, the factorization approach presented here should not have issues with a small target rank (or large rank reduction). In fact, in [IUM14] it is claimed that the proposed factorization approach is even more efficient for problems with small target rank r . We assess this claim by means of the two small examples that we used already in Section 6.1.

Example 6.7 Consider the following Hankel matrix and its corresponding parameter vector

$$\mathbf{A} = \begin{pmatrix} 3 & 2 & 1 & 1 \\ 2 & 1 & 1 & 2 \\ 1 & 1 & 2 & 5 \\ 1 & 2 & 5 & 2 \end{pmatrix} \quad \text{and} \quad \mathbf{p}_A = \begin{pmatrix} 3 & 2 & 1 & 1 & 2 & 5 & 2 \end{pmatrix}^\top.$$

We use the MATLAB implementation [UM19] with Hankel structure, desired rank $r = 1$ and the default initialization for \mathbf{P} given by the truncated SVD.

As solution we obtain the parameter vector

$$\mathbf{p}_{\text{im}} \approx \begin{pmatrix} 1.01 & 1.24 & 1.52 & 1.87 & 2.30 & 2.83 & 3.48 \end{pmatrix}^\top$$

and, consequently, the rank-1 Hankel matrix

$$\mathbf{H}_{\text{im}} = \mathcal{H}(\mathbf{p}_{\text{im}}) \approx \begin{pmatrix} 1.01 & 1.24 & 1.52 & 1.87 \\ 1.24 & 1.52 & 1.87 & 2.30 \\ 1.52 & 1.87 & 2.30 & 2.83 \\ 1.87 & 2.30 & 2.83 & 3.48 \end{pmatrix},$$

rounded to two decimal digits.

Similarly as in Example 6.1, this result is quite close to the optimal parameter vector from Chapter 3. However, with a deviation from the optimal parameter vector of $\|\tilde{\mathbf{p}} - \mathbf{p}_{\text{im}}\|_{\infty} \approx 0.0224$ it is not as close. This is also noticeable in the approximation errors, which are larger than for the kernel method, see also Table 9.4.

The approximation errors produced by the image method are $\|\mathbf{p}_A - \mathbf{p}_{\text{im}}\|_2 \approx 3.5460$ in terms of the parameter vectors and $\|\mathbf{A} - \mathbf{H}_{\text{im}}\|_F \approx 4.5687$ in the Frobenius norm. The corresponding relative errors are $\|\mathbf{p}_A - \mathbf{p}_{\text{im}}\|_2 / \|\mathbf{p}_A\|_2 \approx 0.5118$ and $\|\mathbf{A} - \mathbf{H}_{\text{im}}\|_F / \|\mathbf{A}\|_F \approx 0.4816$. All errors have been rounded to four digits.

Example 6.8 Now also consider again the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \\ 1/2 & 0 & 1 \end{pmatrix} \quad \text{with} \quad \mathbf{p}_A = \begin{pmatrix} 1 & 0 & 1/2 & 0 & 1 \end{pmatrix}^{\top}.$$

Applying the implementation [UM19], we obtain the parameter vector

$$\mathbf{p}_{\text{im}} \approx 10^{-12} \cdot \begin{pmatrix} 0.5 & 0 & 0.3 & 0 & 0.5 \end{pmatrix}^{\top} \approx \mathbf{0}.$$

Thus, the approximating matrix is practically the zero matrix with rank $\mathbf{H}_{\text{im}} \approx 0$.

Naturally, the approximation errors are just $\|\mathbf{p}_A - \mathbf{p}_{\text{im}}\|_2 = \|\mathbf{p}_A\|_2 = 1.5$ along with $\|\mathbf{A} - \mathbf{H}_{\text{im}}\|_F = \|\mathbf{A}\|_F \approx 1.6583$, which is rounded to four digits.

As explained in [CFP03], in some cases, the best approximation is given by a matrix with rank smaller than the desired rank. However, this is not the case here as Table 9.2 shows.

The approximation errors computed in the above examples are again tabled in Section 9.1. Tables 9.2 and 9.4 enable an easy comparison of the different r1H methods presented in this part of the thesis.

Remark 6.9 Recall that in Remark 6.6, we cited the claim from [IUM14] that the image method is especially suited for approximations with small target rank. Now we can inspect this in light of Examples 6.7 and 6.8. We realize that these examples do not confirm that claim. To the contrary, compared to the results in Examples 6.1 and 6.2 from Section 6.1 the above examples exhibit notably worse approximation errors, compare also Tables 9.2 and 9.4.

7

ALTERNATING PROJECTIONS

It is a well known fact that a point in the intersection of two closed and convex sets can be found by alternately projecting first onto one set and then onto the other—starting from an initial point. This alternating projections method is popular because of its simplicity. It has been rediscovered repeatedly for different applications, see [BB96] and the references therein for an overview. Moreover, it ignites ongoing interest, see for example the recent publication [WCW⁺21].

The simplest example of closed convex sets are affine spaces. Alternating projections onto such were analyzed in [vNeu49; vNeu50]. In this case, the alternating projections onto the respective sets converge to the orthogonal projection onto their intersection (assuming that it is non-empty). In other words, for any given initial point, the closest point in the intersection of two affine spaces can be found by alternately projecting onto the respective sets separately.

One advantage of closed convex (and especially affine) constraint sets is that the separate projection subproblems are easy to solve. Any point has a unique nearest point in a closed convex set, in other words, the projection onto this set is well-defined. Moreover, usually this projection is computationally manageable. For its computational and conceptual simplicity, the method of alternating projections undoubtedly has an intuitive appeal.

Therefore, it is tempting to use the analogous idea also for non-convex constraint sets. The first proposal to apply alternating projections to non-convex sets—and specifically to the set of low-rank matrices—is attributed to James A. Cadzow [Cad88]. For this reason, the alternating projections method is also often called Cadzow algorithm or Cadzow’s method

in the context of signal processing.

Let us consider the low-rank Hankel approximation problem in the light of alternating projections. While the Hankel structure—as well as any other matrix structure covered by Chapter 5—constitutes an affine subspace of $\mathbb{C}^{M \times N}$, this is not the case for the low-rank constraint. For one, the set of matrices with rank equal to r might be open, see [CFP03]. While this deficiency can be easily overcome by considering the closed set of matrices with rank smaller than or equal to r , still it is certainly not convex. Nonetheless, a closest rank- r approximation to a given matrix is readily computed via the singular value decomposition, see Theorem 1.8. Thus, the premise of easily solved subproblems is satisfied.

Cadzow’s method for general structured low-rank approximation is summarized in Algorithm 7.1, and in Algorithm 7.2 distinctly for rank-1 Hankel approximation. Thereby, we use the notation from Chapters 1 and 5.

Algorithm 7.1 Cadzow’s algorithm for general structured low-rank approximation

Input: Matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$, target rank r , and a structure specification \mathcal{S} .

Initialize $\mathbf{A}^{(0)} = \mathbf{A}$;

for $j = 0, 1, 2, \dots$ *until a stopping criterion is reached* **do**

compute the optimal (unstructured) rank- r approximation $\mathbf{A}_r^{(j)}$ via truncated SVD;

compute the projection onto the space of structured matrices $\mathbf{A}^{(j)} = \mathcal{P}_{\mathcal{S}}(\mathbf{A}_r^{(j)})$.

Output: A structured low-rank approximation $\mathbf{A}^{(\infty)}$ of \mathbf{A} .

Remark 7.1 1. As a stopping criterion, the relative change in the iterate can be used, see [WCW⁺21]. Accordingly, the algorithm terminates when $\|\mathbf{A}^{(j+1)} - \mathbf{A}^{(j)}\| / \|\mathbf{A}^{(j)}\|$ is small enough.

2. It may happen that the rank- r approximation $\mathbf{A}_r^{(j)}$ in Algorithm 7.1 is not unique, see also Remark 1.9. In that case we just take any such approximation that is given by the first r singular vectors as ordered by the SVD procedure employed. Note that floating point precision errors in numerical algorithms usually prevent such occasions.

3. Also note that there might be a difference in the resulting approximation depending on whether we start Algorithm 7.1 with the one or the other projection, see [Cad88]. We decided to work with the succession of approximations as indicated in Algorithms 7.1 and 7.2 for comparability reasons. The methods from both Chapters 6 and 8 will unquestionably produce Hankel matrices. In the setup of Algorithms 7.1 and 7.2, Cadzow’s method does the same. While the limit itself may depend on which approximation is applied first, our

convergence result from Section 7.2 does not.

Although Cadzow’s algorithm seems to converge in applications, no theoretical convergence results are given in [Cad88]. This gap was partly bridged by [AC13; LM08] for alternating projections on manifolds. However, these results cannot be applied to the case of low-rank Hankel approximation, see also Remark 7.14. Thus, the alternating projections method remains a heuristic approach to solve the Hankel SLRA problem for general low rank. For the r1H problem, however, we will give a complete proof of convergence in the course of this chapter.

We consider Cadzow’s alternating projection method as detailed in Algorithm 7.2, that is, expressly for the r1H problem. In this specific setting, we can show that the sequence of alternating projections always converges to a fixed point. This is a new result that has been published in [KPP21a, Sec. 5]. This whole chapter is based on our paper [KPP21a, Sec. 5] and is partly similar with the derivations therein.

Algorithm 7.2 Cadzow’s algorithm for rank-1 Hankel approximation

Input: Matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$ with single largest singular value.

Initialize $\mathbf{H}^{(0)} = \mathbf{A}$;

for $j = 0, 1, 2, \dots$ *until a stopping criterion is reached* **do**

compute the optimal rank-1 approximation of $\mathbf{H}^{(j)} = \sum_{k=0}^{M-1} \sigma_k^{(j)} \cdot \mathbf{u}_k^{(j)} (\mathbf{v}_k^{(j)})^*$,

$$\mathbf{A}^{(j)} := \sigma_0^{(j)} \cdot \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*,$$

where $\sigma_0^{(j)}$ is the largest singular value of $\mathbf{H}^{(j)}$ with corresponding normalized singular vectors $\mathbf{u}_0^{(j)}$ and $\mathbf{v}_0^{(j)}$;

compute the optimal Hankel approximation

$$\mathbf{H}^{(j+1)} := \mathcal{P}(\mathbf{A}^{(j)})$$

of $\mathbf{A}^{(j)}$, where \mathcal{P} is given in (1.6).

Output: $\mathbf{H}^{(\infty)} = \mathbf{0}$ if $\sigma := \lim_{j \rightarrow \infty} \sigma_0^{(j)} = 0$, or

$\mathbf{H}^{(\infty)} = \sigma \cdot \mathbf{u}\mathbf{v}^*$ with $\mathbf{u}\mathbf{v}^* := \lim_{j \rightarrow \infty} \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*$ if $\sigma := \lim_{j \rightarrow \infty} \sigma_0^{(j)} > 0$.

Having stated Algorithm 7.2, we make some basic observations thereon in Section 7.1. Then, in Section 7.2, we show a series of lemmas that will enable us to prove that—at least for rank-1 Hankel approximation—Cadzow’s algorithm always converges, see Theorem 7.12.

It usually does not, however, converge to the optimal solution neither with respect to the Frobenius nor the spectral norm. This is in accordance with earlier results [CFP03; DeM94]. On that point, see the comparisons with the optimal results from Chapters 3 and 4 in Chapter 9.

7.1 Basic Observations

In this section we lay the groundwork for this chapters' main theorem by making some basic observations on Algorithm 7.2. We start with the following general result on the Hankel projection.

Lemma 7.2 *Let $2 \leq M \leq N$ and $\mathbf{A} \in \mathbb{C}^{M \times N}$. Then the Hankel projection \mathcal{P} in (1.6) satisfies*

$$\|\mathcal{P}(\mathbf{A})\|_F \leq \|\mathbf{A}\|_F,$$

and equality holds if and only if \mathbf{A} already is a Hankel matrix.

Moreover, if \mathbf{A} is a rank-1 matrix $\mathbf{A} = \mathbf{a}\mathbf{b}^*$ with $\mathbf{a} \in \mathbb{C}^M$ and $\mathbf{b} \in \mathbb{C}^N$, then we even have

$$\|\mathcal{P}(\mathbf{a}\mathbf{b}^*)\|_2 \leq \|\mathcal{P}(\mathbf{a}\mathbf{b}^*)\|_F \leq \|\mathbf{a}\mathbf{b}^*\|_F = \|\mathbf{a}\mathbf{b}^*\|_2 = \|\mathbf{a}\|_2 \cdot \|\mathbf{b}\|_2.$$

The equalities $\|\mathcal{P}(\mathbf{a}\mathbf{b}^*)\|_F = \|\mathbf{a}\mathbf{b}^*\|_F$ and $\|\mathcal{P}(\mathbf{a}\mathbf{b}^*)\|_2 = \|\mathbf{a}\mathbf{b}^*\|_2$ hold if and only if there is $z \in \overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ such that $\mathbf{a} = \mathbf{z}_M$ and $\mathbf{b} = \overline{\mathbf{z}}_N$ are structured vectors as given in (2.5).

Proof. Recall the representation of the Hankel projection as matrix-vector multiplication

$$\text{diagvec } \mathcal{P}(\mathbf{A}) = \mathbf{P} \cdot \text{diagvec}(\mathbf{A}),$$

where $\text{diagvec}(\mathbf{A}) \in \mathbb{C}^{MN}$ is the vectorization of \mathbf{A} along its counter-diagonals as defined in (1.8), and \mathbf{P} is the block-diagonal projection matrix

$$\mathbf{P} = \text{diag} \left(1, \frac{1}{2} \mathbf{1}_2, \dots, \frac{1}{M} \mathbf{1}_M, \dots, \frac{1}{M} \mathbf{1}_M, \dots, \frac{1}{2} \mathbf{1}_2, 1 \right) \in \mathbb{C}^{MN \times MN},$$

which was defined in (1.9). Note that each block $\frac{1}{n} \mathbf{1}_n$, $n = 1, \dots, M$, of the matrix \mathbf{P} has the eigenvalues one with multiplicity one and zero with multiplicity $n - 1$. Hence, the Hankel projection matrix \mathbf{P} possesses the same eigenvalues, with multiplicity $M + N - 1$

and $MN - M - N + 1$, respectively, and thus operator norm $\|\mathbf{P}\|_2 = 1$. Therefore we have

$$\|\mathcal{P}(\mathbf{A})\|_F = \|\mathbf{P} \cdot \text{diagvec}(\mathbf{A})\|_2 \leq \|\mathbf{P}\|_2 \cdot \|\text{diagvec}(\mathbf{A})\|_2 = \|\mathbf{A}\|_F,$$

by consistency of the operator norm. The equalities hold by Definitions 1.5 and 1.6 of the Frobenius and spectral norm, respectively. The above inequality is tight if only if $\text{diagvec}(\mathbf{A}) = \mathbf{P} \cdot \text{diagvec}(\mathbf{A})$, that is, if \mathbf{A} has Hankel structure.

If $\mathbf{A} = \mathbf{a}\mathbf{b}^*$, then Lemma 2.3 states that $\mathbf{a}\mathbf{b}^*$ has Hankel structure if and only if $\mathbf{a} = \mathbf{z}_M$ and $\mathbf{b} = \bar{\mathbf{z}}_N$ for some $z \in \bar{\mathbb{C}}$. The equalities $\|\mathbf{a}\mathbf{b}^*\|_F = \|\mathbf{a}\mathbf{b}^*\|_2 = \|\mathbf{a}\|_2 \cdot \|\mathbf{b}\|_2$ are obvious by the characterization of the Frobenius and spectral norm via singular values, see Definitions 1.5 and 1.6. \square

Remark 7.3 Restricted to rank-1 matrices, the operator \mathcal{P} is a projection also with respect to the spectral norm since clearly $\mathcal{P}^2 = \mathcal{P}$ and $\|\mathcal{P}(\mathbf{a}\mathbf{b}^*)\|_2 \leq \|\mathbf{a}\mathbf{b}^*\|_2$ by Lemma 7.2. Hence, the two components of Algorithm 7.2 are irrespective of the norm. The operator \mathcal{P} performs a projection onto the space of Hankel matrices, and the truncated SVD provides an optimal rank-1 approximation for both the Frobenius and the spectral norm. Therefore, it lies in the nature of Cadzow's method for the r1H problem (Algorithm 7.2) to be indifferent of the norm in which the approximation error is measured.

Recall that the projection matrix $\mathbf{P} \in \mathbb{C}^{MN \times MN}$ (1.9) used in Lemma 7.2 has only two distinct eigenvalues (zero and one). Moreover, it is real and symmetric, so any two eigenvectors corresponding to different eigenvalues are orthogonal. Consequently, any vector $\mathbf{w} \in \mathbb{C}^{MN}$ can be decomposed into the direct sum $\mathbf{w} = \mathbf{w}_0 \oplus \mathbf{w}_1$, where $\mathbf{P} \cdot \mathbf{w}_0 = \mathbf{0}$ and $\mathbf{P} \cdot \mathbf{w}_1 = \mathbf{w}_1$, and $\mathbf{w}_0^* \mathbf{w}_1 = 0$. In other words, \mathbf{w}_0 and \mathbf{w}_1 belong to the orthogonal eigenspaces of \mathbf{P} corresponding to the eigenvalue zero and one, respectively.

The following observations make use of this decomposition and will form the basis for the proof of Theorem 7.12. We begin by inspecting the second part of the iteration in Algorithm 7.2, that is, the projection onto the space of Hankel matrices. For simplicity, we usually consider only the outer product matrix $\mathbf{u}_0^{(j)}(\mathbf{v}_0^{(j)})^*$ without the factor $\sigma_0^{(j)}$ instead of the whole matrix $\mathbf{A}^{(j)} = \sigma_0^{(j)} \cdot \mathbf{u}_0^{(j)}(\mathbf{v}_0^{(j)})^*$.

We define $\text{diagvec}(\mathbf{u}_0^{(j)}(\mathbf{v}_0^{(j)})^*) =: \mathbf{w}^{(j)} \in \mathbb{C}^{MN}$, which is decomposed into the direct sum $\mathbf{w}^{(j)} = \mathbf{w}_0^{(j)} \oplus \mathbf{w}_1^{(j)}$ with $\mathbf{P} \cdot \mathbf{w}_0^{(j)} = \mathbf{0}$, $\mathbf{P} \cdot \mathbf{w}_1^{(j)} = \mathbf{w}_1^{(j)}$, and $(\mathbf{w}_0^{(j)})^* \mathbf{w}_1^{(j)} = 0$ as

explained above. Then, for each $j \in \mathbb{N}$ we find

$$\text{diagvec } \mathcal{P} \left(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) = \mathbf{P} \cdot \left(\mathbf{w}_0^{(j)} \oplus \mathbf{w}_1^{(j)} \right) = \mathbf{w}_1^{(j)}.$$

Furthermore, we specify the distance

$$\delta_j := \left\| \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* - \mathcal{P} \left(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) \right\|_F = \left\| \mathbf{w}^{(j)} - \mathbf{w}_1^{(j)} \right\|_2 = \left\| \mathbf{w}_0^{(j)} \right\|_2. \quad (7.1)$$

This can be seen as the distance between the current iterate matrix $\mathbf{A}^{(j)}$ and its Hankel projection $\mathbf{H}^{(j+1)}$ divided by the factor $\sigma_0^{(j)} > 0$. In the case $\sigma_0^{(j)} = 0$, this interpretation is neither valid nor interesting.

With the δ_j specified in (7.1), we obtain

$$\left\| \mathcal{P} \left(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) \right\|_F^2 = \left\| \mathbf{w}_1^{(j)} \right\|_2^2 = 1 - \delta_j^2, \quad (7.2)$$

where we have used that $1 = \left\| \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right\|_F^2 = \left\| \mathbf{w}^{(j)} \right\|_2^2 = \left\| \mathbf{w}_0^{(j)} \right\|_2^2 + \left\| \mathbf{w}_1^{(j)} \right\|_2^2$. It will later become clear that δ_j is decreasing with $\lim_{j \rightarrow \infty} \delta_j = 0$ as a consequence of the convergent sequence of Frobenius norms $\lim_{j \rightarrow \infty} \left\| \mathcal{P} \left(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) \right\|_F = 1$, see Corollary 7.7.

Next, we consider the first part of the subsequent iteration of Algorithm 7.2. More precisely, we consider the singular value decomposition

$$\mathcal{P} \left(\sigma_0^{(j)} \cdot \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) = \mathbf{H}^{(j+1)} = \sum_{k=0}^{M-1} \sigma_k^{(j+1)} \cdot \mathbf{u}_k^{(j+1)} (\mathbf{v}_k^{(j+1)})^*. \quad (7.3)$$

We assume $\sigma_0^{(j)} > 0$ for the largest singular value of $\mathbf{H}^{(j)}$ from the previous iteration step. Dividing both sides of the above equation (7.3) by this value, we obtain

$$\mathcal{P} \left(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) = \sum_{k=0}^{M-1} \frac{\sigma_k^{(j+1)}}{\sigma_0^{(j)}} \cdot \mathbf{u}_k^{(j+1)} (\mathbf{v}_k^{(j+1)})^* =: \sum_{k=0}^{M-1} s_k^{(j+1)} \cdot \mathbf{u}_k^{(j+1)} (\mathbf{v}_k^{(j+1)})^*,$$

where $s_k^{(j+1)} := \sigma_k^{(j+1)} / \sigma_0^{(j)}$. On the one hand, equation (7.2) gives us

$$\sum_{k=0}^{M-1} (s_k^{(j+1)})^2 = \left\| \mathcal{P} \left(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) \right\|_F^2 = 1 - \delta_j^2. \quad (7.4)$$

On the other hand, the Eckart-Young-Mirsky Theorem 1.8 yields

$$\begin{aligned} \sum_{k=1}^{M-1} (s_k^{(j+1)})^2 &= \left\| \mathcal{P}(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*) - s_0^{(j+1)} \cdot \mathbf{u}_0^{(j+1)} (\mathbf{v}_0^{(j+1)})^* \right\|_F^2 \\ &< \left\| \mathcal{P}(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*) - \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right\|_F^2 = \delta_j^2 \end{aligned} \quad (7.5)$$

since $s_0^{(j+1)} \cdot \mathbf{u}_0^{(j+1)} (\mathbf{v}_0^{(j+1)})^*$ is the optimal rank-1 approximation of $\mathcal{P}(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*)$. Note that the above is an estimate for the distance between the current Hankel matrix $\mathbf{H}^{(j+1)}$ and its rank-1 approximation $\mathbf{A}^{(j+1)}$ divided by the largest singular value of the previous iteration step, $\sigma_0^{(j)} > 0$. In the case $\sigma_0^{(j)} = 0$, we do not need to consider the above derivations in order to obtain a convergence result, see the proof of Theorem 7.12.

Combining equations (7.4) and (7.5), we find

$$1 - 2 \cdot \delta_j^2 < (s_0^{(j+1)})^2 \leq 1 - \delta_j^2. \quad (7.6)$$

Remark 7.4 As a follow-up to Remark 7.1.2, we note that the ambiguity in the largest eigenvalues does not occur in the later course of Algorithm 7.2. More precisely, the inequality (7.6) together with $\lim_{j \rightarrow \infty} \delta_j = 0$ (see Corollary 7.7) implies the existence of some $j_0 \in \mathbb{N}$ such that for all $j > j_0$ the value $\sigma^{(j)} < \sigma^{(j_0)}$ is small enough to ensure that the largest singular value $s_0^{(j+1)}$ of $\mathcal{P}(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*)$ is isolated.

7.2 Our New Convergence Result

In the course of this section, we will show that the Cadzow's algorithm for rank-1 Hankel approximation (Algorithm 7.2) always converges to a unique fixed point. This fixed point usually is a rank-1 Hankel approximation of \mathbf{A} . But it may also happen that it is the zero matrix, and thus not a rank-1 Hankel approximation, see Example 7.15. Either way, we will show that the sequence of matrices $\mathbf{A}^{(j)} = \sigma_0^{(j)} \cdot \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*$ generated by Algorithm 7.2 converges to a Hankel matrix (zero or rank-1). So upon convergence, the last Hankel projection step in Algorithm 7.2 is redundant.

In order to rigorously analyze the convergence properties of Algorithm 7.2 we start with a series of lemmas.

Lemma 7.5 *Let $\mathbf{A} \in \mathbb{C}^{M \times N}$ with $2 \leq M \leq N$ and $\text{rank } \mathbf{A} \geq 1$. The sequence $(\sigma_0^{(j)})_{j \in \mathbb{N}}$ of from Algorithm 7.2 converges, and we define the limit point $\sigma := \lim_{j \rightarrow \infty} \sigma_0^{(j)}$.*

Proof. If the first singular vectors $\mathbf{u}_0 = \mathbf{u}_0^{(0)}$ and $\mathbf{v}_0 = \mathbf{v}_0^{(0)}$ of \mathbf{A} are of the structured form $\mathbf{z}_M(z)$ and $\bar{\mathbf{z}}_N(z)$ from (2.5) for some $z \in \bar{\mathbb{C}}$, respectively, then the optimal rank-1 approximation $\sigma_0 \cdot \mathbf{u}_0 \mathbf{v}_0^*$ of \mathbf{A} already has Hankel structure. Therefore, by definition of \mathcal{P} , we find that $(\sigma_0^{(j)})_{j \in \mathbb{N}} \equiv (\sigma_0)_{j \in \mathbb{N}}$ is a constant sequence.

Assume now that $\mathbf{A}^{(0)} = \sigma_0 \cdot \mathbf{u}_0 \mathbf{v}_0^*$ does not have Hankel structure. Then by Lemma 7.2, we find for any $j \in \mathbb{N}_0$ we that

$$\sigma_0^{(j+1)} = \left\| \mathcal{P} \left(\sigma_0^{(j)} \cdot \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) \right\|_2 \leq \sigma_0^{(j)} \cdot \left\| \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right\|_2 = \sigma_0^{(j)}. \quad (7.7)$$

For the last equality, we have used that the singular vectors are normalized.

This inequality is strict as long as the rank-1 matrix $\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*$ does not have Hankel structure, see Lemma 7.2. In particular, we have $\sigma_0^{(1)} < \sigma_0^{(0)}$ for $j = 0$ by assumption.

Thus, the sequence of singular values $(\sigma_0^{(j)})_{j \in \mathbb{N}}$ is monotonically decreasing. Since $\sigma_0^{(j)}$ is bounded from below for all j , convergence follows and we write $\sigma := \lim_{j \rightarrow \infty} \sigma_0^{(j)}$. \square

Combining Lemmas 7.2 and 7.5, we obtain convergence of the norms $\left\| \mathcal{P}(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*) \right\|_F$ and $\left\| \mathcal{P}(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*) \right\|_2$ for positive limit σ .

Corollary 7.6 *Let $\mathbf{A} \in \mathbb{C}^{M \times N}$ with $2 \leq M \leq N$ and $\text{rank } \mathbf{A} \geq 1$. Assume that the limiting singular value from Algorithm 7.2 is positive (i.e., $\sigma = \lim_{j \rightarrow \infty} \sigma_0^{(j)} > 0$). Consider the Hankel matrices $\mathcal{P}(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*)$, $j \in \mathbb{N}$, where the vectors $\mathbf{u}_0^{(j)}$ and $\mathbf{v}_0^{(j)}$ are generated by Algorithm 7.2.*

Then both the sequence of their Frobenius norms and the sequence of their spectral norms converge to one, that is, we have

$$\lim_{j \rightarrow \infty} \left\| \mathcal{P}(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*) \right\|_2 = \lim_{j \rightarrow \infty} \left\| \mathcal{P}(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*) \right\|_F = 1.$$

Proof. For the spectral norm, we have

$$\begin{aligned} \sigma &= \lim_{j \rightarrow \infty} \sigma_0^{(j+1)} = \lim_{j \rightarrow \infty} \left\| \mathcal{P} \left(\sigma_0^{(j)} \cdot \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) \right\|_2 \\ &= \lim_{j \rightarrow \infty} \sigma_0^{(j)} \cdot \lim_{j \rightarrow \infty} \left\| \mathcal{P} \left(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) \right\|_2. \end{aligned}$$

Dividing this by $\sigma = \lim_{j \rightarrow \infty} \sigma_0^{(j)}$ on both ends, we obtain

$$\lim_{j \rightarrow \infty} \left\| \mathcal{P} \left(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) \right\|_2 = 1.$$

From Lemma 7.2, we have for any $j \in \mathbb{N}$

$$\begin{aligned}\sigma_0^{(j+1)} &= \left\| \mathcal{P}(\sigma_0^{(j)} \cdot \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*) \right\|_2 \\ &\leq \left\| \mathcal{P}(\sigma_0^{(j)} \cdot \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*) \right\|_F \\ &\leq \left\| \sigma_0^{(j)} \cdot \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right\|_2 = \sigma_0^{(j)}.\end{aligned}$$

Taking the limit and dividing all parts of this inequality by $\sigma = \lim_{j \rightarrow \infty} \sigma_0^{(j)}$ we also obtain

$$\lim_{j \rightarrow \infty} \left\| \mathcal{P}(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*) \right\|_F = 1$$

in the Frobenius norm, as was claimed. \square

Note that convergence of the norms as shown in Corollary 7.6 does not imply convergence of the sequence of matrices. But, we obtain another convergence result, namely convergence of the distance δ_j defined in (7.1).

Corollary 7.7 *Consider the distance δ_j from (7.1), and assume that $\sigma = \lim_{j \rightarrow \infty} \sigma_j > 0$. Then the sequence $(\delta_j)_{j \in \mathbb{N}}$ decreases to zero, we have $\lim_{j \rightarrow \infty} \delta_j = 0$.*

Proof. The assertion follows directly by combining the observation (7.2) from Section 7.1 and Corollary 7.6. \square

The next lemma yields a convergent subsequence of $(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*)_{j \in \mathbb{N}}$.

Lemma 7.8 *Let $\mathbf{A} \in \mathbb{C}^{M \times N}$ with $2 \leq M \leq N$ and $\text{rank } \mathbf{A} \geq 1$. Assume that the limiting singular value from Algorithm 7.2 is positive (i.e., $\sigma = \lim_{j \rightarrow \infty} \sigma_0^{(j)} > 0$).*

Then there is a subsequence $(\mathbf{u}_0^{(j_k)} (\mathbf{v}_0^{(j_k)})^)_{k \in \mathbb{N}}$ of $(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*)_{j \in \mathbb{N}}$ from Algorithm 7.2 that converges to a limit $\mathbf{u}\mathbf{v}^*$. Furthermore, this limit $\mathbf{u}\mathbf{v}^*$ is a rank-1 Hankel matrix; this means, there is $z \in \overline{\mathbb{C}}$ such that*

$$\mathbf{u}\mathbf{v}^* = \lim_{k \rightarrow \infty} \mathbf{u}_0^{(j_k)} (\mathbf{v}_0^{(j_k)})^* = \mathbf{z}_M(z) \mathbf{z}_N(z)^\top,$$

where \mathbf{z}_M and \mathbf{z}_N are the normalized structured vectors from (2.5).

Proof. If the first singular vectors of \mathbf{A} are structured, that is, $\mathbf{u}_0 = \mathbf{z}_M(z)$ and $\mathbf{v}_0 = \bar{\mathbf{z}}_N(z)$, then the optimal rank-1 approximation of \mathbf{A} already has Hankel structure. In that case,

we have $\sigma_0^{(j)} \equiv \sigma_0$ and $\mathbf{u}_0^{(j)}(\mathbf{v}_0^{(j)})^* \equiv \mathbf{u}_0\mathbf{v}_0^*$ for all $j \in \mathbb{N}$; in other words $(\sigma_0^{(j)})_{j \in \mathbb{N}}$ and $(\mathbf{u}_0^{(j)}(\mathbf{v}_0^{(j)})^*)_{j \in \mathbb{N}}$ are constant sequences and there is nothing left to prove.

Assume now that $\mathbf{u}_0\mathbf{v}_0^*$ does not have Hankel structure. Recall that

$$\lim_{j \rightarrow \infty} \left\| \mathcal{P} \left(\mathbf{u}_0^{(j)}(\mathbf{v}_0^{(j)})^* \right) \right\|_2 = 1$$

by Corollary 7.6. Since the vectors $\mathbf{u}_0^{(j)}$ and $\mathbf{v}_0^{(j)}$ are normalized, the sequence of matrices $(\mathbf{u}_0^{(j)}(\mathbf{v}_0^{(j)})^*)_{j \in \mathbb{N}}$ is bounded. More precisely, $\|\mathbf{u}_0^{(j)}(\mathbf{v}_0^{(j)})^*\|_2 = 1$ for all $j \in \mathbb{N}$.

We conclude that there exists a subsequence $(\mathbf{u}_0^{(j_k)}(\mathbf{v}_0^{(j_k)})^*)_{k \in \mathbb{N}}$ that converges to an accumulation point $\mathbf{u}\mathbf{v}^*$. By Lemma 7.2 and Corollary 7.6 this point is a fixed point of the Cadzow algorithm (i.e., $\mathcal{P}(\mathbf{u}\mathbf{v}^*) = \mathbf{u}\mathbf{v}^*$) and thus a rank-1 Hankel matrix. \square

Remark 7.9 Note that in [ZG20] a similar result has been shown for low-rank Hankel approximation by Cadzow's algorithm. What was not studied in [ZG20] is the question whether the partial sequence indeed converges to a matrix with the desired rank.

With Theorem 7.12, we will prove that in fact the full sequence $(\mathbf{u}_0^{(j)}(\mathbf{v}_0^{(j)})^*)_{j \in \mathbb{N}}$ in Algorithm 7.2 converges to the fixed point $\mathbf{u}\mathbf{v}^*$ found in Lemma 7.8. The proof essentially relies on the observation that a rank-1 matrix $\mathbf{a}\mathbf{b}^* \in \mathbb{C}^{M \times N}$, which is close to the space of Hankel matrices, is also close to the set of rank-1 Hankel matrices.

Lemma 7.10 For vectors $\mathbf{a} = (a_j)_{j=0}^{M-1} \in \mathbb{C}^M$ and $\mathbf{b} = (b_j)_{j=0}^{N-1} \in \mathbb{C}^N$ with $\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2 = 1$, assume that their outer product is elementwisely close to some Hankel matrix; in rigorous terms, assume

$$\|\mathbf{a}\mathbf{b}^* - \mathcal{P}(\mathbf{a}\mathbf{b}^*)\|_\infty \leq \delta \tag{7.8}$$

with δ small enough, e.g. $\delta < \frac{1}{6\sqrt{MN}}$. The norm $\|\cdot\|_\infty$ denotes the elementwise maximum as given in Definition 1.7.

Then the outer product $\mathbf{a}\mathbf{b}^*$ is also close to a rank-1 Hankel matrix in the Frobenius norm,

$$\min_{c \in \mathbb{C}, z \in \overline{\mathbb{C}}} \left\| \mathbf{a}\mathbf{b}^* - c \cdot \mathbf{z}_M \mathbf{z}_N^T \right\|_F < C \cdot \delta, \tag{7.9}$$

where the constant C only depends on the dimensions M and N .

Proof. Let a_μ and b_ν denote the by modulus largest entries of \mathbf{a} and \mathbf{b} , respectively. Then, since \mathbf{a} and \mathbf{b} are normalized, we have $|a_\mu| \geq \frac{1}{\sqrt{M}}$ and $|b_\nu| \geq \frac{1}{\sqrt{N}}$. Without loss of

generality, assume that $\mu < M - 1$. Otherwise consider the flipped matrix $\mathbf{J}_M \cdot \mathbf{a}\mathbf{b}^* \cdot \mathbf{J}_N$ instead of $\mathbf{a}\mathbf{b}^*$, where \mathbf{J}_N is the counter-identity matrix of size N from (1.10).

Choose $z := \frac{a_{\mu+1}}{a_\mu}$. Using the Hankel structure of $\mathcal{P}(\mathbf{a}\mathbf{b}^*)$ and the assumption (7.8), we obtain for any $\ell = 1, \dots, N - 1$

$$a_\mu \bar{b}_\ell = a_{\mu+1} \bar{b}_{\ell-1} + \delta_{\mu,\ell} = z \cdot a_\mu \bar{b}_{\ell-1} + \delta_{\mu,\ell},$$

with some correction term $\delta_{\mu,\ell}$. This correction term is bounded by $|\delta_{\mu,\ell}| \leq 2\delta$ for $\ell = 0, \dots, N - 1$ because of assumption (7.8).

By rearranging the above equality we obtain

$$\bar{b}_\ell = \bar{b}_{\ell-1} \cdot z + \frac{\delta_{\mu,\ell}}{a_\mu}, \quad (7.10)$$

and inductively

$$\bar{b}_\ell = \bar{b}_0 \cdot z^\ell + \frac{1}{a_\mu} \cdot \sum_{j=0}^{\ell-1} \delta_{\mu,\ell-j} \cdot z^j$$

for $\ell = 1, \dots, N - 1$. In fact, this relation remains true for $\ell = 0$ if the sum over an empty index set is assigned the value zero. This assignment is common practice.

Note that the first summands, $\bar{b}_0 \cdot z^\ell$ for $\ell = 0, \dots, N - 1$, are exactly the entries of $\bar{b}_0 \cdot \hat{\mathbf{z}}_N(z)$, where $\hat{\mathbf{z}}_N(z)$ is the non-normalized structured vector from (2.1). Thus, it follows that

$$\|\bar{\mathbf{b}} - \bar{b}_0 \cdot \hat{\mathbf{z}}_N(z)\|_\infty = \max_{\ell=0,\dots,N-1} \left| \frac{1}{a_\mu} \cdot \sum_{j=0}^{\ell-1} \delta_{\mu,\ell-j} \cdot z^j \right| \leq 2 \cdot \sqrt{MN} \cdot \delta. \quad (7.11)$$

For the inequality, we used that $|z| \leq 1$ by the choice of z , as well as the aforementioned estimates $|a_\mu| \leq \frac{1}{\sqrt{M}}$ for the by modulus largest entry of \mathbf{a} and $|\delta_{\mu,\ell-j}| \leq 2\delta$ for the correction term.

Similarly, if $\nu < N - 1$, we find for all $\ell = 1, \dots, M - 1$

$$a_\ell \bar{b}_\nu = a_{\ell-1} \bar{b}_{\nu+1} + \delta_{\nu,\ell} = a_{\ell-1} \cdot \left(\bar{b}_\nu \cdot z + \frac{\delta_{\mu,\nu+1}}{a_\mu} \right) + \delta_{\nu,\ell},$$

where, in the second step, we have replaced $\bar{b}_{\nu+1}$ by the expression in (7.10). For the occurring correction term we have again $|\delta_{\nu,\ell}| \leq 2\delta$ by assumption (7.8).

Rearranging this equation yields

$$a_\ell = a_{\ell-1} \cdot z + \frac{\delta_{\mu,\nu+1} \cdot a_{\ell-1}}{a_\mu \bar{b}_\nu} + \frac{\delta_{\nu,\ell}}{\bar{b}_\nu} = a_{\ell-1} \cdot z + \frac{\tilde{\delta}_{\mu,\nu+1}}{\bar{b}_\nu} + \frac{\delta_{\nu,\ell}}{\bar{b}_\nu}, \quad (7.12)$$

with a modified correction term $\tilde{\delta}_{\mu,\nu+1} := \frac{a_{\ell-1}}{a_\mu} \cdot \delta_{\mu,\nu+1}$. This modified correction term is again bounded by $|\tilde{\delta}_{\mu,\nu+1}| \leq 2\delta$ since $|\frac{a_{\ell-1}}{a_\mu}| \leq 1$ by construction. As before, we obtain inductively

$$a_\ell = a_0 \cdot z^\ell + \frac{\tilde{\delta}_{\mu,\nu+1}}{\bar{b}_\nu} \cdot \sum_{j=0}^{\ell-1} z^j + \frac{1}{\bar{b}_\nu} \cdot \sum_{j=0}^{\ell-1} \delta_{\nu,\ell-j} \cdot z^j$$

for $\ell = 0, 1, \dots, M-1$.

Now, $a_0 \cdot z^\ell$, $\ell = 0, \dots, M-1$, are the entries of $a_0 \cdot \hat{\mathbf{z}}_M(z)$ with $\hat{\mathbf{z}}_M(z)$ from (2.1). Hence, analogously to (7.11) we obtain

$$\|\mathbf{a} - a_0 \cdot \hat{\mathbf{z}}_M(z)\|_\infty \leq 4 \cdot M \sqrt{N} \cdot \delta, \quad (7.13)$$

where we have used the estimates $|z| \leq 1$, $|\tilde{\delta}_{\mu,N-1}| \leq 2\delta$, $|\delta_{N-2,\ell-j}| \leq 2\delta$, and $|b_\nu| \geq \frac{1}{\sqrt{N}}$.

If $\nu = N-1$, meaning that b_{N-1} is the by modulus largest entry of \mathbf{b} , we cannot carry out the above steps to obtain (7.13). But, we can use the assumption (7.8) to show that $|b_{N-2}| \geq \frac{1}{\sqrt{N}}$. Then we can replace \bar{b}_ν by \bar{b}_{N-2} in the expression (7.12) for a_ℓ in order to obtain a similar estimate as (7.13) for sufficiently small δ .

To see this in more detail, consider

$$a_\mu \bar{b}_{N-1} = a_{\mu+1} \bar{b}_{N-2} + \delta_{\mu,N-1}.$$

Note that $a_{\mu+1}$ does exist since we have assumed that $\mu < M-1$ without loss of generality. This equation is equivalent to

$$a_{\mu+1} \bar{b}_{N-2} = a_\mu \bar{b}_{N-1} - \delta_{\mu,N-1},$$

where we have again used the correction term $\delta_{\mu,N-1}$ with $|\delta_{\mu,N-1}| \leq 2\delta$.

Thus, we obtain the estimate

$$|a_{\mu+1} \bar{b}_{N-2}| = |a_\mu \bar{b}_{N-1} - \delta_{\mu,N-1}| \geq |a_\mu \bar{b}_{N-1}| - 2\delta.$$

We divide this inequality by $|a_\mu|$, and since $\frac{|a_{\mu+1}|}{|a_\mu|} \leq 1$, it follows that

$$|b_{N-2}| \geq \frac{|a_{\mu+1}\bar{b}_{N-2}|}{|a_\mu|} \geq \frac{|a_\mu\bar{b}_{N-1}| - 2\delta}{|a_\mu|} = |b_{N-1}| - \frac{2\delta}{|a_\mu|}.$$

Using the bounds $|a_\mu| \geq \frac{1}{\sqrt{M}}$ and $|b_\nu| = |b_{N-1}| \geq \frac{1}{\sqrt{N}}$ on the largest entries of \mathbf{a} and \mathbf{b} , respectively, we can further estimate

$$|b_{N-2}| \geq |b_{N-1}| - \frac{2\delta}{|a_\mu|} \geq \frac{1}{\sqrt{N}} - 2\delta \cdot \sqrt{M}.$$

With δ small enough, e.g. $\delta < \frac{1}{6\sqrt{MN}}$, we obtain

$$|b_{N-2}| \geq \frac{1}{\sqrt{N}} - 2\delta \cdot \sqrt{M} \geq \frac{1}{\sqrt{N}} - 2 \cdot \frac{1}{6\sqrt{MN}} \cdot \sqrt{M} = \frac{1}{\sqrt{N}} - \frac{1}{3N} \geq \frac{1}{N},$$

where the last inequality holds since $N \geq 2$.

Now, replacing ν by $N - 2$ in (7.12), we obtain a similar expression. Namely

$$a_\ell = a_{\ell-1} \cdot z + \frac{\tilde{\delta}_{\mu, N-1}}{\bar{b}_{N-2}} + \frac{\delta_{N-2, \ell}}{\bar{b}_{N-2}},$$

from which inductively follows

$$a_\ell = a_0 \cdot z^\ell + \frac{\tilde{\delta}_{\mu, N-1}}{\bar{b}_{N-2}} \cdot \sum_{j=0}^{\ell-1} z^j + \frac{1}{\bar{b}_{N-2}} \cdot \sum_{j=0}^{\ell-1} \delta_{N-2, \ell-j} \cdot z^j.$$

Recall that $a_0 \cdot z^\ell$ are the entries of $a_0 \cdot \hat{\mathbf{z}}_M(z)$ with the structured vector $\hat{\mathbf{z}}_M(z)$ from (2.1). Thus, we conclude a similar estimate as (7.13)

$$\|\mathbf{a} - a_0 \cdot \hat{\mathbf{z}}_M(z)\|_\infty \leq 4 \cdot MN \cdot \delta, \quad (7.14)$$

where we have used the same estimates as for (7.13) except for $|b_\nu| \geq \frac{1}{\sqrt{N}}$ which is replaced by $|b_{N-2}| \geq \frac{1}{N}$.

The claim (7.9) now follows with the ensuing chain of inequalities. Note that (7.11) and (7.13) can further be bounded by $4 \cdot MN \cdot \delta$ like (7.14). In line 3 we use this bound $4 \cdot MN \cdot \delta$ on (7.11), (7.13), and (7.14), and in line 4 we use the original bound (7.11). Furthermore, we

employ the fact that \mathbf{a} and \mathbf{b} are normalized. Altogether, we have

$$\begin{aligned}
\|\mathbf{ab}^* - a_0\bar{b}_0 \cdot \hat{\mathbf{z}}_M \hat{\mathbf{z}}_N^\top\|_F^2 &= \sum_{j=0}^{M-1} \sum_{k=0}^{N-1} |a_j \bar{b}_k - a_0 \bar{b}_0 z^{j+k}|^2 \\
&= \sum_{j=0}^{M-1} \sum_{k=0}^{N-1} \left| a_j \cdot (\bar{b}_k - \bar{b}_0 z^k) + \bar{b}_0 z^k \cdot (a_j - a_0 z^j) \right|^2 \\
&\leq (4 \cdot MN \cdot \delta)^2 \cdot \sum_{j=0}^{M-1} \sum_{k=0}^{N-1} \left(|a_j| + |\bar{b}_0 z^k| \right)^2 \\
&\leq (4 \cdot MN \cdot \delta)^2 \cdot \sum_{j=0}^{M-1} \sum_{k=0}^{N-1} \left(|a_j| + |b_k| + 2 \cdot \sqrt{MN} \cdot \delta \right)^2 \\
&\leq (4 \cdot MN \cdot \delta)^2 \cdot \sum_{j=0}^{M-1} \sum_{k=0}^{N-1} 3 \cdot \left(|a_j|^2 + |b_k|^2 + (2 \cdot \sqrt{MN} \cdot \delta)^2 \right) \\
&\leq 3 \cdot (4 \cdot MN \cdot \delta)^2 \cdot (M + N + 4 \cdot M^2 N^3 \cdot \delta^2) \\
&\leq 48 \cdot (MN)^2 \cdot (M + N + 4 \cdot M^2 N^3) \cdot \delta^2.
\end{aligned}$$

The last inequality holds for $\delta \leq 1$ which is clearly satisfied. Thus, the constant C may be chosen as

$$C < 48 \cdot (MN)^2 \cdot (M + N + 4 \cdot M^2 N^3),$$

which only depends on the dimensions M and N . Then the assertion (7.9) is valid for $z = \frac{a_{\mu+1}}{a_\mu}$ as chosen in the beginning of this proof and $c = a_0 b_0 \cdot (\|\hat{\mathbf{z}}_M\|_2 \cdot \|\hat{\mathbf{z}}_N\|_2)$. \square

Remark 7.11 The estimates derived in the above proof of Lemma 7.10 are very coarse and could presumably be refined in several places. Nevertheless, they suffice for our purposes.

With these preliminaries, we can now show this chapter's main theorem on the convergence of Cadzow's rank-1 Hankel approximation to one fixed point. Part (1) of Theorem 7.12 covers the case when the limiting singular value σ is zero and is quite trivial. The case when the limiting singular value σ is positive is dealt with in part (2).

The proof of Theorem 7.12 is done by applying the lemmas and observations of this and the previous section in the right way. The idea of the proof can be summarized as follows: First, we apply Lemma 7.10 to establish the distance between the rank-1 matrix $\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*$ and some rank-1 Hankel matrix. Second, we show that all further rank-1 iterates lie within

the same distance of the rank-1 Hankel matrix established in the first step. We conclude that the limit $\mathbf{u}\mathbf{v}^*$ of the subsequence from Lemma 7.8 has to lie inside the ball around this rank-1 Hankel matrix, too. The proof is completed by invoking Corollary 7.7, which states that the radius of this ball tends to zero as the iteration advances.

Theorem 7.12 *Let $\mathbf{A} \in \mathbb{C}^{M \times N}$ with $2 \leq M \leq N$ and $\text{rank } \mathbf{A} \geq 1$.*

(1) *If $\sigma = \lim_{j \rightarrow \infty} \sigma_0^{(j)} = 0$, then the sequence of matrices $(\sigma_0^{(j)} \cdot \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*)_{j \in \mathbb{N}}$ from Algorithm 7.2 converges to the zero matrix.*

(2) *If $\sigma = \lim_{j \rightarrow \infty} \sigma_0^{(j)} > 0$, then the sequence of matrices $(\sigma_0^{(j)} \cdot \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*)_{j \in \mathbb{N}}$ from Algorithm 7.2 converges to a rank-1 matrix $\sigma \cdot \mathbf{u}\mathbf{v}^*$.*

Moreover, there exists $z \in \overline{\mathbb{C}}$ such that

$$\mathbf{u}\mathbf{v}^* := \lim_{j \rightarrow \infty} \mathbf{u}_j \mathbf{v}_j^* = \mathbf{z}_M \mathbf{z}_N^\top,$$

with $\mathbf{z}_M = \mathbf{z}_M(z)$ and $\mathbf{z}_N = \mathbf{z}_N(z)$ structured as in (2.5). This means that Algorithm 7.2 provides the rank-1 Hankel approximation $\sigma \cdot \mathbf{u}\mathbf{v}^$.*

Remark 7.13 Theorem 7.12 states that Cadzow's algorithm converges to a rank-1 Hankel approximation of \mathbf{A} if $\sigma > 0$. This does not mean that Cadzow's algorithm converges to the optimal solution of the r1H problem.

Proof. As shown in Lemma 7.5, the sequence of largest singular values produced by Algorithm 7.2 always converges to a limit $\sigma \geq 0$. The singular vectors $\mathbf{u}_0^{(j)}$ and $\mathbf{v}_0^{(j)}$ are normed for all $j \in \mathbb{N}$, thus the sequence of their outer products $(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*)_{j \in \mathbb{N}}$ is bounded.

If now $\sigma = \lim_{j \rightarrow \infty} \sigma_0^{(j)} = 0$, then the sequence $(\sigma_0^{(j)} \cdot \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*)_{j \in \mathbb{N}}$ converges to the zero matrix. In this case the outer products $\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*$ may or may not converge to a matrix of Hankel structure. This concludes the proof of part (1) of the theorem.

For part (2), assume $\sigma = \lim_{j \rightarrow \infty} \sigma_0^{(j)} > 0$. Combining the distance δ_j defined in (7.1) and Lemma 7.10, there exist $c_j \in \mathbb{C}$ and $z_j \in \overline{\mathbb{C}}$ such that

$$\left\| \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* - c_j \cdot \mathbf{z}_M(z_j) \mathbf{z}_N(z_j)^\top \right\|_F < C \cdot \delta_j,$$

where the constant C only depends on the dimensions M and N . Since all the vectors in this inequality are normalized and thus $\left\| \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right\|_F = \left\| \mathbf{z}_M(z_j) \mathbf{z}_N(z_j)^\top \right\|_F = 1$, we

obtain $|1 - |c_j|| < C \cdot \delta_j$. It follows that the parameters z_j chosen above satisfy

$$\left\| \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* - \mathbf{z}_M(z_j) \mathbf{z}_N(z_j)^\top \right\|_F < 2C \cdot \delta_j \quad (7.15)$$

for each $j \in \mathbb{N}$.

Recall that we have convergence $\lim_{j \rightarrow \infty} \delta_j = 0$ by Corollary 7.7. Thus, we may choose δ_j as small as we like. Proceeding from there, we show that all further iteration matrices $\mathbf{u}_0^{(k)} (\mathbf{v}_0^{(k)})^*$, for $k > j$, also fulfill the estimate (7.15). In other words, all matrices $\mathbf{u}_0^{(k)} (\mathbf{v}_0^{(k)})^*$, for $k > j$, are contained in the ball of radius $2C \cdot \delta_j$ around $\mathbf{z}_M(z_j) \mathbf{z}_N(z_j)^\top$.

It is sufficient to show that the ensuing iterate satisfies (7.15). The argument can then be repeated for any $k > j + 1$. We want to show

$$\left\| \mathbf{u}_0^{(j+1)} (\mathbf{v}_0^{(j+1)})^* - \mathbf{z}_M(z_j) \mathbf{z}_N(z_j)^\top \right\|_F < 2C \cdot \delta_j.$$

Observe that the Frobenius norm can be expressed via the trace (see Definition 1.5) as follows

$$\begin{aligned} & \left\| \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* - \mathbf{z}_M(z_j) \mathbf{z}_N(z_j)^\top \right\|_F^2 \\ &= \text{tr} \left(\left(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* - \mathbf{z}_M \mathbf{z}_N^\top \right)^* \cdot \left(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* - \mathbf{z}_M \mathbf{z}_N^\top \right) \right) \\ &= \text{tr} \left(\mathbf{v}_0^{(j)} (\mathbf{v}_0^{(j)})^* + \bar{\mathbf{z}}_N \mathbf{z}_N^\top - (\mathbf{u}_0^{(j)})^* \mathbf{z}_M \cdot \mathbf{v}_0^{(j)} \mathbf{z}_N^\top - \mathbf{z}_M^* \mathbf{u}_0^{(j)} \cdot \bar{\mathbf{z}}_N (\mathbf{v}_0^{(j)})^* \right) \\ &= 2 - 2 \cdot \text{Re} \left((\mathbf{u}_0^{(j)})^* \mathbf{z}_M \cdot (\mathbf{v}_0^{(j)})^\top \mathbf{z}_N \right), \end{aligned} \quad (7.16)$$

where from the second line on we have omitted the argument z_j in the structured vectors $\mathbf{z}_M = \mathbf{z}_M(z_j)$ and $\mathbf{z}_N = \mathbf{z}_N(z_j)$.

With this expression, (7.15) is equivalent to

$$\text{Re} \left((\mathbf{u}_0^{(j)})^* \mathbf{z}_M \cdot (\mathbf{v}_0^{(j)})^\top \mathbf{z}_N \right) > 1 - 2 \cdot (C\delta_j)^2$$

and we strive to show that also

$$\text{Re} \left((\mathbf{u}_0^{(j+1)})^* \mathbf{z}_M \cdot (\mathbf{v}_0^{(j+1)})^\top \mathbf{z}_N \right) > 1 - 2 \cdot (C\delta_j)^2. \quad (7.17)$$

We consider the distance between the current outer product $\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*$ and the rank-1 Hankel matrix $\mathbf{z}_M(z_j) \mathbf{z}_N(z_j)^\top$, compare Lemma 7.10. Therefore, recall the decomposition $\text{diagvec}(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*) = \mathbf{w}^{(j)} = \mathbf{w}_0^{(j)} \oplus \mathbf{w}_1^{(j)}$ into a direct sum according to the eigenvalues

of the block-diagonal projection matrix \mathbf{P} from (1.9), see Section 7.1. Using the projection property $\mathbf{P} \cdot \text{diagvec}(\mathbf{z}_M(z_j)\mathbf{z}_N(z_j)^\top) = \text{diagvec}(\mathbf{z}_M(z_j)\mathbf{z}_N(z_j)^\top)$, we conclude

$$\begin{aligned} & \left\| \mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* - \mathbf{z}_M(z_j)\mathbf{z}_N(z_j)^\top \right\|_F^2 \\ &= \left\| \text{diagvec} \left(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) - \text{diagvec}(\mathbf{z}_M(z_j)\mathbf{z}_N(z_j)^\top) \right\|_2^2 \\ &= \left\| \mathbf{P} \cdot \left(\mathbf{w}_1^{(j)} - \text{diagvec}(\mathbf{z}_M(z_j)\mathbf{z}_N(z_j)^\top) \right) \oplus \mathbf{w}_0^{(j)} \right\|_2^2 \\ &= \left\| \mathbf{P} \cdot \mathbf{w}_1^{(j)} - \text{diagvec}(\mathbf{z}_M(z_j)\mathbf{z}_N(z_j)^\top) \right\|_2^2 + \left\| \mathbf{w}_0^{(j)} \right\|_2^2 \\ &= \left\| \mathcal{P} \left(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) - \mathbf{z}_M(z_j)\mathbf{z}_N(z_j)^\top \right\|_F^2 + \delta_j^2. \end{aligned}$$

The distance $\delta_j = \left\| \mathbf{w}_0^{(j)} \right\|_2$ has been defined in (7.1).

Taking (7.15) into account, we obtain

$$\left\| \mathcal{P} \left(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) - \mathbf{z}_M(z_j)\mathbf{z}_N(z_j)^\top \right\|_F^2 < (2C \cdot \delta_j)^2 - \delta_j^2. \quad (7.18)$$

Let the SVD of $\mathcal{P}(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*)$ be given by

$$\mathcal{P} \left(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) = \sum_{k=0}^{M-1} s_k^{(j+1)} \cdot \mathbf{u}_k^{(j+1)} (\mathbf{v}_k^{(j+1)})^*$$

as in (7.3) with $s_k^{(j+1)} = \sigma_k^{(j+1)}/\sigma_0^{(j)}$. We use the orthogonal matrices $(\mathbf{u}_0^{(j+1)} \dots \mathbf{u}_{M-1}^{(j+1)})$ and $(\mathbf{v}_0^{(j+1)} \dots \mathbf{v}_{N-1}^{(j+1)})$ occurring in this SVD to transform the normalized structured vectors $\mathbf{z}_M(z_j)$ and $\bar{\mathbf{z}}_N(z_j)$ into

$$\boldsymbol{\alpha} := \boldsymbol{\alpha}(z_j) := \left((\mathbf{u}_0^{(j+1)})^* \mathbf{z}_M(z_j) \dots (\mathbf{u}_{M-1}^{(j+1)})^* \mathbf{z}_M(z_j) \right)^\top$$

and

$$\boldsymbol{\beta} := \boldsymbol{\beta}(z_j) := \left((\mathbf{v}_0^{(j+1)})^* \bar{\mathbf{z}}_N(z_j) \dots (\mathbf{v}_{M-1}^{(j+1)})^* \bar{\mathbf{z}}_N(z_j) \right)^\top.$$

In particular, this basis transform does not affect the normalization (i.e., $\|\boldsymbol{\alpha}\|_2 = \|\boldsymbol{\beta}\|_2 = 1$). The first entries of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are of special interest for us as $\mathcal{R}e(\alpha_0 \bar{\beta}_0)$ occurs in (7.17).

The same orthogonal transform by $(\mathbf{u}_0^{(j+1)} \dots \mathbf{u}_{M-1}^{(j+1)})$ and $(\mathbf{v}_0^{(j+1)} \dots \mathbf{v}_{N-1}^{(j+1)})$ can be applied to the norm in (7.18) because of the orthogonal invariance of the Frobenius

norm. We obtain

$$\begin{aligned}
(2C \cdot \delta_j)^2 - \delta_j^2 &> \left\| \mathcal{P} \left(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^* \right) - \mathbf{z}_M(z_j) \mathbf{z}_N(z_j)^\top \right\|_F^2 \\
&= \left\| \text{diag} \left(s_k^{(j+1)} \right)_{k=0}^{M-1} - \boldsymbol{\alpha}(z_j) \boldsymbol{\beta}(z_j)^* \right\|_F^2 \\
&= 1 + \sum_{k=0}^{M-1} (s_k^{(j+1)})^2 - 2 \cdot \sum_{k=0}^{M-1} s_k^{(j+1)} \cdot \Re(\alpha_k(z_j) \overline{\beta_k(z_j)}),
\end{aligned}$$

where $s_k^{(j+1)} = \sigma_k^{(j+1)} / \sigma_0^{(j)}$ are the the singular values of $\mathcal{P}(\mathbf{u}_0^{(j)} (\mathbf{v}_0^{(j)})^*)$, compare (7.3).

From Section 7.1, recall the observation (7.4), namely $\sum_{k=0}^{M-1} (s_k^{(j+1)})^2 = 1 - \delta_j^2$. Using this observation while rearranging the last inequality, we have

$$\sum_{k=0}^{M-1} s_k^{(j+1)} \cdot \Re(\alpha_k(z_j) \overline{\beta_k(z_j)}) > 1 - 2 \cdot (C\delta_j)^2. \quad (7.19)$$

By Hölder's inequality,

$$\sum_{k=0}^{M-1} \left| \Re(\alpha_k(z_j) \overline{\beta_k(z_j)}) \right| \leq \sum_{k=0}^{M-1} \left| \alpha_k(z_j) \overline{\beta_k(z_j)} \right| \leq \|\boldsymbol{\alpha}(z_j)\|_2 \cdot \|\boldsymbol{\beta}(z_j)\|_2 = 1,$$

and (7.5), it follows that

$$\begin{aligned}
\sum_{k=1}^{M-1} \left| s_k^{(j+1)} \cdot \Re(\alpha_k(z_j) \overline{\beta_k(z_j)}) \right| &\leq \left(\sum_{j=1}^{M-1} (s_k^{(j+1)})^2 \right)^{1/2} \cdot \left(\sum_{k=1}^{M-1} \Re(\alpha_k(z_j) \overline{\beta_k(z_j)}) \right)^{1/2} \\
&< \delta_j \cdot \sum_{k=1}^{M-1} \left| \Re(\alpha_k(z_j) \overline{\beta_k(z_j)}) \right| \\
&\leq \delta_j \cdot \left(1 - \Re(\alpha_0(z_j) \overline{\beta_0(z_j)}) \right).
\end{aligned}$$

Thus, (7.19) implies that

$$\begin{aligned}
1 - 2 \cdot (C\delta_j)^2 &< \sum_{k=0}^{M-1} s_k^{(j+1)} \cdot \Re(\alpha_k(z_j) \overline{\beta_k(z_j)}) \\
&= s_0^{(j+1)} \cdot \Re(\alpha_0(z_j) \overline{\beta_0(z_j)}) + \sum_{k=1}^{M-1} s_k^{(j+1)} \cdot \Re(\alpha_k(z_j) \overline{\beta_k(z_j)})
\end{aligned}$$

$$\begin{aligned}
&\leq s_0^{(j+1)} \cdot \mathcal{R}e(\alpha_0(z_j)\overline{\beta_0(z_j)}) + \delta_j \cdot \left(1 - \mathcal{R}e(\alpha_0(z_j)\overline{\beta_0(z_j)})\right) \\
&= (s_0^{(j+1)} - \delta_j) \cdot \mathcal{R}e(\alpha_0(z_j)\overline{\beta_0(z_j)}) + \delta_j,
\end{aligned}$$

and finally

$$\mathcal{R}e(\alpha_0(z_j)\overline{\beta_0(z_j)}) > \frac{1 - 2 \cdot (C\delta_j)^2 - \delta_j}{s_0^{(j+1)} - \delta_j} \geq \frac{1 - 2 \cdot (C\delta_j)^2 - \delta_j}{1 - \delta_j^2/2 - \delta_j} > 1 - 2 \cdot (C\delta_j)^2$$

as desired, compare (7.17). In the above line of inequalities, we have used the estimate $s_0^{(j+1)} \leq \sqrt{1 - \delta_j^2} \leq 1 - \delta_j^2/2$ from (7.6). The last inequality holds for δ_j small enough and can be verified by an easy calculation.

This shows that the $(j + 1)$ -st iterate is close the rank-1 Hankel matrix $\mathbf{z}_M(z_j)\mathbf{z}_N(z_j)^\top$, more precisely,

$$\left\| \mathbf{u}_0^{(j+1)}(\mathbf{v}_0^{(j+1)})^* - \mathbf{z}_M(z_j)\mathbf{z}_N(z_j)^\top \right\|_F < 2C \cdot \delta_j$$

because of (7.16).

As final step of this proof, we recall the limit $\mathbf{u}\mathbf{v}^* = \lim_{k \rightarrow \infty} \mathbf{u}_0^{(j_k)}(\mathbf{v}_0^{(j_k)})^*$ of the subsequence from Lemma 7.8. We conclude that this limit $\mathbf{u}\mathbf{v}^*$ also has to lie inside the ball around $\mathbf{z}_M(z_j)\mathbf{z}_N(z_j)^\top$ with radius $2C \cdot \delta_j$ for any j and thus

$$\left\| \mathbf{u}_0^{(j)}(\mathbf{v}_0^{(j)})^* - \mathbf{u}\mathbf{v}^* \right\|_F < 4C \cdot \delta_j.$$

Now, since $\lim_{j \rightarrow \infty} \delta_j = 0$ by Corollary 7.7, we have convergence of $\mathbf{u}_0^{(j)}(\mathbf{v}_0^{(j)})^*$ to $\mathbf{u}\mathbf{v}^*$ and the proof is complete. \square

Remark 7.14 There are several attempts in the literature to show convergence of Cadzow's algorithm in broad generality. However, to our understanding, none of them are really reliable. For example, note that the results from [LM08] cannot be applied to our setting since the manifolds in question do not satisfy the required so-called transversality condition. In [AC11b; AC13], the transversality condition is relaxed and replaced by the weaker condition of existence of non-tangential intersection points. The convergence results then rely on the assumption that the angle in these intersection points between the considered manifolds is bounded away from zero, or equivalently, that the value $\sigma(A)$ in [AC13, Def. 3.1] is smaller than one. This assumption is not easy to show in the setting of rank-1 Hankel approximation and possibly not even satisfied.

We give an example, that can be calculated analytically by hand. It confirms that indeed there are initial matrices for which Cadzow's method converges to the zero matrix despite the existence of an optimal solution of true rank one, compare Examples 3.8 and 4.12. We also refer to the summary in Table 9.2.

Example 7.15 Consider again the example matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \\ 1/2 & 0 & 1 \end{pmatrix}$$

with eigenvalues $\lambda_0 = \lambda_0^{(0)} = 3/2$ and $\lambda_1 = \lambda_1^{(0)} = \lambda_2 = \lambda_2^{(0)} = 1/2$ and corresponding normalized eigenvectors

$$\mathbf{v}_0 = \mathbf{v}_0^{(0)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{v}_1 = \mathbf{v}_1^{(0)} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{v}_2 = \mathbf{v}_2^{(0)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}.$$

Since \mathbf{A} is real symmetric we use the eigendecomposition instead of the SVD in this example.

Otherwise following Algorithm 7.2, we find the unstructured rank-1 approximation of \mathbf{A} from the largest eigenvalue λ_0 and corresponding eigenvector \mathbf{v}_0

$$\mathbf{A}^{(0)} = \lambda_0 \cdot \mathbf{v}_0 \mathbf{v}_0^T = \frac{3}{2} \cdot \frac{1}{2} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

Averaging along the counter-diagonals, we obtain the first Hankel matrix of the iteration

$$\mathbf{H}^{(1)} = \mathcal{P}(\mathbf{A}^{(0)}) = \frac{3}{2} \cdot \frac{1}{2} \begin{pmatrix} 1 & 0 & 2/3 \\ 0 & 2/3 & 0 \\ 2/3 & 0 & 1 \end{pmatrix} = \frac{3}{2} \cdot \begin{pmatrix} 1/2 & 0 & 1/3 \\ 0 & 1/3 & 0 \\ 1/3 & 0 & 1/2 \end{pmatrix} = \lambda_0 \cdot \mathcal{P}(\mathbf{v}_0 \mathbf{v}_0^T).$$

Now, $\mathbf{v}_0^{(1)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 1 \end{pmatrix}^T = \mathbf{v}_0^{(0)}$ is the eigenvector of $\mathcal{P}(\mathbf{v}_0 \mathbf{v}_0^T)$ corresponding to its largest eigenvalue $\lambda_0^{(1)} = 5/6$.

Further iterations yield

$$\mathbf{v}_0^{(j)} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad \text{and} \quad \lambda_0^{(j)} = \frac{3}{2} \cdot \left(\frac{5}{6}\right)^j$$

Obviously, the sequence of largest eigenvectors, $(\mathbf{v}_0^{(j)})_{j \in \mathbb{N}}$, is constant, and the largest eigenvalues $\lambda_0^{(j)}$, $j \in \mathbb{N}$ form a null sequence, $\lim_{j \rightarrow \infty} \lambda_0^{(j)} = 0$. In other words, the Cadzow algorithm fails to converge to a rank-1 matrix.

Example 7.15 demonstrates that Cadzow's Algorithm 7.2 may converge to the zero matrix. Even if this is not the case and the Cadzow iteration converges to a rank-1 Hankel matrix, this is usually not the optimal solution to the r1H problem. For both the Frobenius and the spectral norm, we usually see significant gaps between Cadzow's r1H error and the minimal one, compare Chapter 9, especially Figures 9.2 and 9.6. This behavior confirms previous results on alternating projection algorithms for the r1H setting, see e.g. [CFP03; DeM94].

7.3 Numerical Assessment of Convergence

Recall Remark 7.1.1 on the stopping criterion for Algorithm 7.1. In Theorem 7.12, we have proven convergence of Cadzow's algorithm for rank-1 Hankel approximation. So instead of the relative change mentioned in Remark 7.1.1, we can simply use convergence as stopping criterion for Algorithm 7.2. That is, the algorithm does not terminate until for some j the rank-1 matrix $\mathbf{A}^{(j)}$ has Hankel structure and the last step $\mathbf{H}^{(j+1)} = \mathcal{P}(\mathbf{A}^{(j)})$ is redundant.

In this section, we want to numerically investigate the number of iterations that is needed until convergence. In order to do so, we generate ten parameter vectors $\mathbf{p}_A \in \mathbb{R}^{n_p}$ with $n_p = 19$ and entries between -50 and 50 . (Actually, we use the same parameter vectors $\mathbf{p}_A \in \mathbb{R}^{19}$ as in Section 9.2.) From each parameter vector we assemble the Hankel matrices $\mathcal{H}_{M,N}(\mathbf{p}_A)$, see Chapter 5, of different shapes $M \times N$ while $M + N = n_p + 1 = 20$ is fixed. Then for each shape, we average the number of iterations needed until convergence over the ten different parameter vectors. Our findings are depicted in Figure 7.1.

Contemplating Figure 7.1, we notice that for square and "moderately rectangular" input matrices, Cadzow's algorithm is relatively efficient. The mean number of iterations needed for $M = 10, 9, 8, 7$ is well below 100 (actually it is around 65).

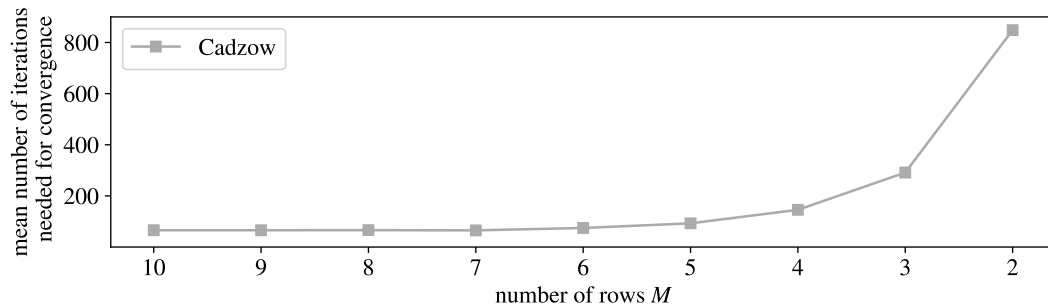


Figure 7.1 Average number of iterations needed for convergence of Algorithm 7.2 for decreasing number of rows M while $M + N = n_p + 1 = 20$ remains constant.

For greater differences between M and N however, the iteration count increases drastically. In the extreme case, when approximating matrices of size (2×18) , Algorithm 7.2 converges after an average of 848.3 iterations.

We draw the conclusion that Algorithm 7.2 does not only converge in theory, but also always converges in practical applications. However, it makes sense to maintain a stopping criterion based on the relative change in the iterate as in Remark 7.1. This will still lead to very good results while not exhausting the number of iterations. Furthermore, one might want to use a more efficient implementation of Cadzow's algorithm, as for example [WCW⁺21], instead of the rather naive one we used here.

As a sidenote, we remark that convergence to a zero matrix does not happen in any of the examples considered for Figure 7.1.

The development of the r1H error in the Frobenius norm for decreasing number of rows with fixed $n_p = 19$ is assessed in Section 9.2.1. From Figure 9.4 we observe that Cadzow's mean relative error (MRE, see (9.3)) largely mimics the behavior of the optimal MRE. Cadzow's MRE is noticeably larger than the optimal one, but the difference between them approximately remains the same for the different matrix shapes.

8

CONVEX RELAXATION

The rank of a matrix is not a convex function. This is the main reason why structured low-rank approximation problems are so difficult to solve exactly. Another heuristic method to tackle this obstacle—besides Cadzow’s algorithm from Chapter 7—is based on convex relaxation of the rank.

One way to characterize the rank of a matrix is via its singular values. The rank of a matrix is equal to the number of its non-zero singular values, see Lemma 1.4 part (5). Let $\mathbf{A} \in \mathbb{C}^{M \times N}$ have the singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$, where $\mathbf{\Sigma} = \text{diag}(\sigma_0, \dots, \sigma_{M-1})$. Denote by $\boldsymbol{\sigma} = (\sigma_0 \cdots \cdots \sigma_{M-1})^\top$ the vector of singular values. Then the rank of \mathbf{A} can be expressed as

$$\text{rank } \mathbf{A} = r = \#\{j: \sigma_j \neq 0\} =: \|\boldsymbol{\sigma}\|_0,$$

where $\|\cdot\|_0$ is called the ℓ_0 -quasi-norm. It should be emphasized that this is not a true vector norm because it is not homogeneous.

The notion of ℓ_0 -quasi-norm is often used in the context of compressed sensing, signal processing, and harmonic analysis where sparse approximations (i.e., with low ℓ_0 -quasi-norm) of vectors are of interest. However, the problem of minimizing the ℓ_0 -quasi-norm of a vector is not convex.

Therefore, when trying to find the sparsest vector, often its ℓ_1 -norm is minimized instead, see for example [CDS01; Don06a; Don06b]. The convex relaxation of the ℓ_0 -quasi-norm by the ℓ_1 -norm is known to be an efficient heuristic actually yielding sparse solutions [Don06a; Don06b]. This fact calls for a similar approach in terms of matrix (quasi-)norms.

Definition 8.1 (Nuclear norm) Let $\mathbf{A} \in \mathbb{C}^{M \times N}$, $M \leq N$, be a matrix with singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$, where $\mathbf{\Sigma} = \text{diag } \boldsymbol{\sigma} = \text{diag}(\sigma_0, \dots, \sigma_{M-1})$. The nuclear norm of \mathbf{A} is defined as

$$\|\mathbf{A}\|_* := \sum_{j=0}^{M-1} \sigma_j = \|\boldsymbol{\sigma}\|_1,$$

the sum of its singular values. The nuclear norm belongs to the Ky-Fan-norms, a family of matrix norms named after the Chinese-American mathematician Ky Fan.

Remark 8.2 1. Unlike the ℓ_0 -quasi-norm for vectors, the nuclear norm is indeed a matrix norm, see [HJ13, Sec. 7.4].

2. The nuclear norm is the dual of the spectral norm, that is,

$$\|\mathbf{A}\|_* = \sup\{\text{tr}(\mathbf{B}^\top \mathbf{A}) : \|\mathbf{B}\|_2 \leq 1\}.$$

3. For symmetric and positive semidefinite matrices, the nuclear norm is equal to the trace (sum of diagonal elements) of the matrix [FHB01; Faz02].

4. The nuclear norm constitutes a convex envelope for the rank of a matrix, see [Faz02; FHB04].

The nuclear norm was introduced in [FHB01; Faz02] as a convex heuristic for the rank minimization problem (RMP),

$$\min \text{rank}(\mathbf{A}) \quad \text{subject to } \mathbf{A} \in \mathcal{C}, \quad (8.1)$$

where \mathcal{C} is a convex set of constraints.

In engineering and computational sciences, often the simplest model that satisfies certain constraints is of interest. Since the rank can be a measure for model complexity, it is not surprising that the RMP has a wide range of applications in its own right, see [Faz02; FHB04; RFP10]. However, it is non-convex, indeed it is NP-hard in general [VB96].

Instead of problem (8.1), in [FHB01] it was proposed to solve

$$\min \|\mathbf{A}\|_* \quad \text{subject to } \mathbf{A} \in \mathcal{C}, \quad (8.2)$$

which is in fact convex and can therefore be easily solved. For example, problem (8.2) can be formulated as semidefinite program and solved by general convex optimization methods [FHB01; Faz02; FHB04; LV10]. Based on interior-point methods, an efficient implementation

has been developed in [LV10] by exploiting the particular problem structure of (8.2). The corresponding software is available as part of the `cvxopt`-package [ADV21] for Python.

Similar to ℓ_1 -norm minimization, which often yields sparse solutions, problem (8.2) often has solutions of low rank. A theoretical characterization of circumstances when the nuclear norm heuristic produces a solution of *minimal* rank is given in [RFP10].

Although loosely related, the RMP (8.1) is quite different from our r1H problem (2). We are interested in the (optimal) approximation of a given matrix \mathbf{A} by a Hankel matrix \mathbf{H} that is forced to possess rank one. By contrast, in (8.1) the rank is to be minimized subject to some convex set of constraints. The convex relaxation (8.2) does not diminish this discrepancy. Thus unfortunately, the heuristic (8.2) is not directly applicable to our problem (2).

The key ideas for conciliating problems (2) and (8.2) have been introduced in Chapters 5 and 6. We fall back on the parameter vectors $\mathbf{p} \in \mathbb{R}^{n_p}$, and then

- impose the Hankel structure directly with $\mathbf{H} = \mathcal{H}(\mathbf{p})$, and
- regularize the deviation from the input matrix \mathbf{A} in terms of the parameter vectors $\|\mathbf{p}_A - \mathbf{p}\|_W^2$,

where $\|\cdot\|_W$ is the weighted vector norm introduced in Chapter 5. Usually, the occurring weight matrix $\mathbf{W} \in \mathbb{R}^{n_p \times n_p}$ is either the identity matrix corresponding to the Euclidean norm of the parameter vectors, or the positive definite matrix defined in (5.4) corresponding to the Frobenius norm of Hankel matrices.

The spectral norm cannot be expressed while using the problem representation with parameter vectors $\mathbf{p} \in \mathbb{R}^{n_p}$. Furthermore, the initial matrix \mathbf{A} needs to have Hankel structure itself or it has to be replaced by its Hankel projection $\mathcal{P}(\mathbf{A})$, as mentioned earlier in Chapter 5.

Using the above ideas, our aim is to solve the regularized nuclear norm minimization problem

$$\min \left(\|\mathcal{H}(\mathbf{p})\|_* + \gamma \cdot \|\mathbf{p}_A - \mathbf{p}\|_W^2 \right), \quad (8.3)$$

as a convex relaxation for the regularized rank minimization problem

$$\min \left(\text{rank } \mathcal{H}(\mathbf{p}) + \gamma \cdot \|\mathbf{p}_A - \mathbf{p}\|_W^2 \right),$$

where the regularization parameter γ admits the following interpretation: For $\gamma = 0$, the approximation condition is dropped and the resulting Hankel matrix with minimal nuclear

norm (or minimal rank) is the zero matrix. On the opposite side, for $\gamma = \infty$, the variable \mathbf{p} has to be equal to the parameter vector \mathbf{p}_A , and $\mathcal{H}(\mathbf{p}) = \mathbf{A}$ is generically not of low rank, let alone rank one.

In this respect, the regularized optimization problem (8.3) is different from the regularized optimization in Section 6.2 where in the limit $\gamma \rightarrow \infty$ the desired solution is reached. Here, a good balance has to be found between the rank minimization and the approximation property of the parameter vector \mathbf{p} .

We seek for this balance by solving problem (8.3) with different values for the regularization parameter. For the resulting set of solutions we plot the trade-off curve between the approximation error and the actual rank of the Hankel matrix $\mathcal{H}(\mathbf{p})$. A generic trade-off curve is shown in Figure 8.1. Since the rank always is a natural number, this plot results in a step function. From the trade-off curve, the smallest approximation error for which the solution is indeed a rank-1 matrix can be determined.

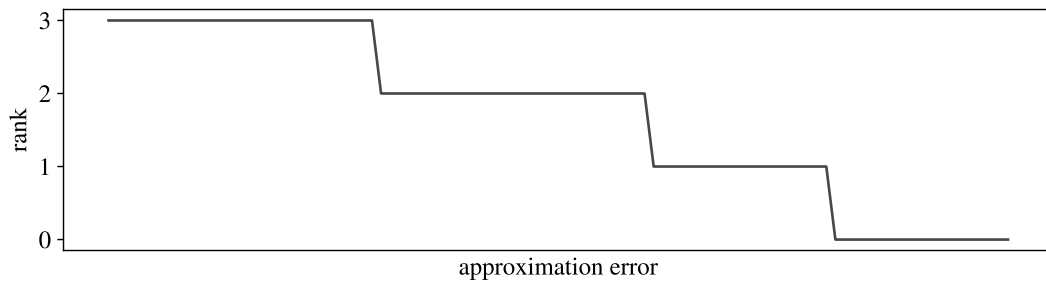


Figure 8.1 A generic trade-off curve between the approximation error $\|\mathbf{p}_A - \mathbf{p}\|_W^2$ and the rank of the approximating Hankel matrix $\text{rank } \mathcal{H}(\mathbf{p})$. The approximation error is small on the left-hand side of the horizontal axis and grows to the right.

In order to be able to test the nuclear norm heuristic for rank-1 Hankel approximation, we have a closer look at the implementation [ADV21; LV10]. The software is designed to solve the more general nuclear norm minimization problem

$$\min \left(\|\mathcal{S}(\mathbf{p})\|_* + 1/2 \cdot \mathbf{p}^\top \mathbf{B} \mathbf{p} + \mathbf{d}^\top \mathbf{p} \right),$$

where \mathcal{S} is the structure specification map from Definition 5.1. We set $\mathbf{B} := 2\gamma \cdot \mathbf{W}$ and $\mathbf{d} := -2\gamma \cdot \mathbf{W} \cdot \mathbf{p}_A$, where \mathbf{W} is the weight matrix corresponding to the norm used (either the identity or given in (5.4)). Then, the above is equivalent to our regularized nuclear norm approximation problem (8.3).

We test the software [ADV21] on the same examples that we considered before in Chapters 6 and 7. The errors produced by the convex relaxation method from this chapter are compared to the other results in Section 9.1, see Tables 9.2 and 9.4.

Example 8.3 Consider again the matrix

$$\mathbf{A} = \begin{pmatrix} 3 & 2 & 1 & 1 \\ 2 & 1 & 1 & 2 \\ 1 & 1 & 2 & 5 \\ 1 & 2 & 5 & 2 \end{pmatrix}.$$

For this example, let $\mathbf{W} = \mathbf{I}$, which corresponds to the Euclidean norm of the parameter vectors $\|\mathbf{p}_A - \mathbf{p}\|_W = \|\mathbf{p}_A - \mathbf{p}\|_2$.

We run the software [ADV21] to solve problem (8.3) for different regularization parameters γ . At first we let $0 \leq \gamma < 1$ with 0.05 as step size. For each such γ we plot the approximation error and nuclear norm of the resulting approximation matrix $\mathcal{H}(\mathbf{p})$ in Figure 8.2. It is apparent that the rank indeed decreases as a by-product of nuclear norm minimization.

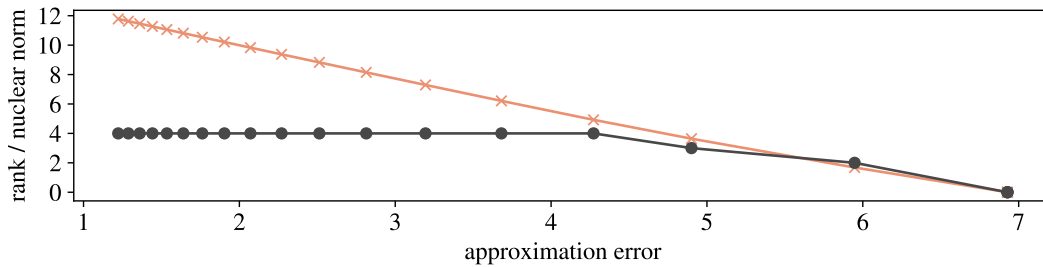


Figure 8.2 Trade-off curve between the approximation error $\|\mathbf{p}_A - \mathbf{p}\|_2$ and rank (dark grey line), respectively nuclear norm (light red line) of $\mathcal{H}(\mathbf{p})$ for Example 8.3. The regularization parameters used are $\gamma \in [0, 1)$ with 0.05 as step size.

In Figure 8.2, we also see that no Hankel matrix of rank one was computed for the adopted set of regularization parameters. In order to learn how to better choose γ , we also plot rank $\mathcal{H}(\mathbf{p})$ for each regularization parameter, see Figure 8.3.

An inspection thereof suggests that we refine the step size of the regularization parameters in the area between $\gamma = 0.1$ and $\gamma = 0.15$. Figure 8.4 shows rank $\mathcal{H}(\mathbf{p})$ for γ in this range and with the smaller step size of 0.005. The corresponding trade-off curve is plotted in Figure 8.5.

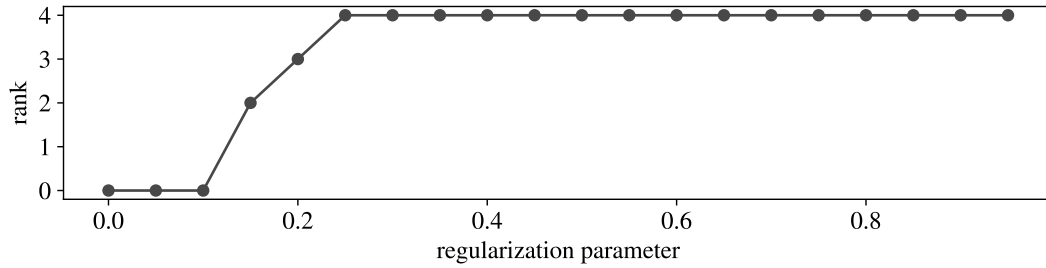


Figure 8.3 Rank of the solution matrix $\mathcal{H}(\mathbf{p})$ for different regularization parameters $\gamma \in [0, 1)$ with step size 0.05 in Example 8.3.

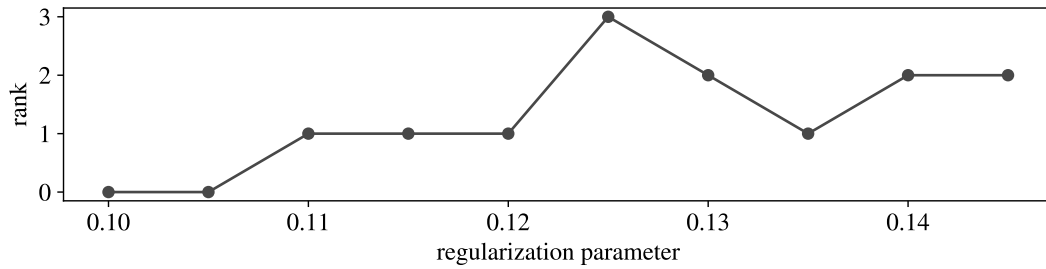


Figure 8.4 Rank of the solution matrix $\mathcal{H}(\mathbf{p})$ for regularization parameters $\gamma \in [0.1, 0.15)$ with the refined step size of 0.005 in Example 8.3.

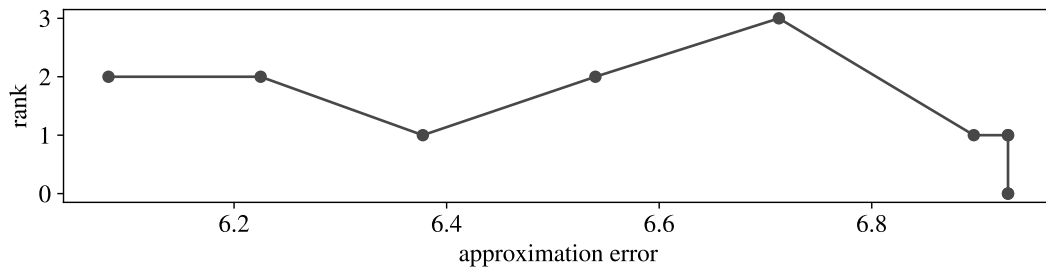


Figure 8.5 Trade-off curve between the approximation error $\|\mathbf{p}_A - \mathbf{p}\|_2$ and the rank of $\mathcal{H}(\mathbf{p})$ in Example 8.3. Calculations are made for γ in the range $[0.1, 0.15)$ with refined stepsize 0.005. The smallest error $\|\mathbf{p}_A - \mathbf{p}_{\text{nuc}}\|_2 \approx 6.3776$ is achieved for $\gamma = 0.135$.

Indeed “zooming in” on the regularization parameter like this, reveals several points where the solution $\mathcal{H}(\mathbf{p})$ of problem (8.3) has rank one. Among those, we select the rank-1 Hankel approximation $\mathbf{H}_{\text{nuc}} := \mathcal{H}(\mathbf{p}_{\text{nuc}})$ with the smallest approximation error. It is given by

$$\mathbf{p}_{\text{nuc}} = \left(0.1518 \quad 0.1933 \quad 0.2461 \quad 0.3133 \quad 0.3990 \quad 0.5080 \quad 0.6468 \right)^\top$$

and is the solution of (8.3) for $\gamma = 0.135$.

The approximation errors achieved by this parameter vector \mathbf{p}_{nuc} are $\|\mathbf{p}_A - \mathbf{p}_{\text{nuc}}\|_2 \approx 6.3776$ in terms of the parameter vectors, and $\|\mathbf{A} - \mathbf{H}_{\text{nuc}}\|_F \approx 8.6885$ in the Frobenius norm. The respective relative approximation errors $\|\mathbf{p}_A - \mathbf{p}_{\text{nuc}}\|_2 / \|\mathbf{p}_A\|_2 \approx 0.9205$ and $\|\mathbf{A} - \mathbf{H}_{\text{nuc}}\|_F / \|\mathbf{A}\|_F \approx 0.9158$ arise as a result. All errors as well as the parameter vector \mathbf{p}_{nuc} have been rounded to four digits.

Remark 8.4 1. The Figures 8.4 and 8.5 do not exhibit the expected monotonicity, as for example Figure 8.1 does. This is because in (8.3), the nuclear norm is minimized instead of the rank. The nuclear norm is in fact monotonically decreasing in the approximation error and monotonically increasing in the regularization parameter as expected. However, small nuclear norms may still correspond to higher values of the rank when there are several small but non-zero singular values.

2. As is evident in Figure 8.4, several distinct regularization parameters γ may lead to Hankel matrices $\mathcal{H}(\mathbf{p})$ of rank one. Among those, the largest regularization parameter places the most emphasis on the approximation error $\|\mathbf{p}_A - \mathbf{p}\|_W^2$. Thus, this one contributes the best approximate solution to the r1H problem.

Example 8.5 For this example we stick with the same matrix \mathbf{A} as in Example 8.3 and perform the same procedure again. The difference is that, this time, we use the non-trivial weight matrix \mathbf{W} from (5.4). Then the nuclear norm minimization problem (8.3) features the Frobenius norm $\|\mathbf{A} - \mathcal{H}(\mathbf{p})\|_F$ as regularization term.

We directly jump to the interesting range of regularization parameters. The resulting trade-off curve for γ between 0 and 0.1 with a step size of 0.005 is shown in Figure 8.6.

The smallest errors obtained by a rank-1 Hankel matrix are $\|\mathbf{A} - \mathcal{H}(\mathbf{p}_{\text{nuc}})\|_F \approx 8.0818$ and $\|\mathbf{p}_A - \mathbf{p}_{\text{nuc}}\|_2 \approx 5.9672$ with corresponding relative errors $\|\mathbf{A} - \mathcal{H}(\mathbf{p}_{\text{nuc}})\|_F / \|\mathbf{A}\|_F \approx 0.8519$ and $\|\mathbf{p}_A - \mathbf{p}_{\text{nuc}}\|_2 / \|\mathbf{p}_A\|_2 \approx 0.8613$.

The parameter vector producing these errors is

$$\mathbf{p}_{\text{nuc}} \approx \left(0.3745 \quad 0.4590 \quad 0.5625 \quad 0.6894 \quad 0.8450 \quad 1.0356 \quad 1.2692 \right)^\top,$$

which is computed for $\gamma = 0.075$. Errors and parameter vector are rounded to four decimal digits.

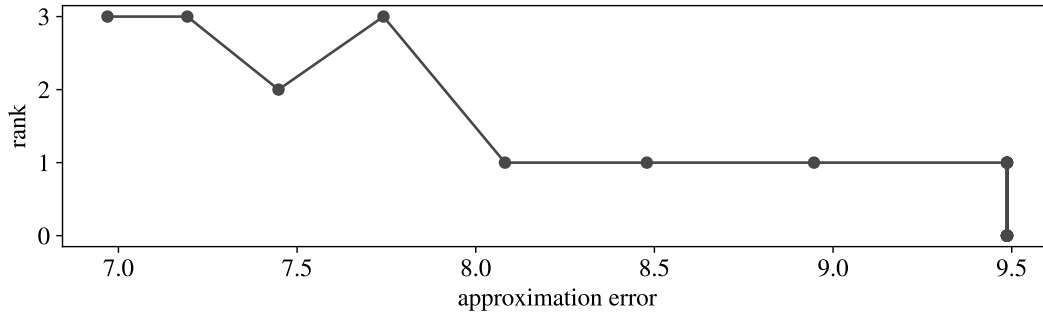


Figure 8.6 Trade-off curve between the Frobenius norm error $\|\mathbf{A} - \mathcal{H}(\mathbf{p})\|_F$ and the rank of $\mathcal{H}(\mathbf{p})$ for Example 8.5. The curve is shown for regularization parameters $\gamma \in [0, 0.1)$ with 0.005 as step size.

Comparing the results from Examples 8.3 and 8.5 we observe that in the latter both errors are smaller. In Example 8.3 the vector norm $\|\mathbf{p}_A - \mathbf{p}_{\text{nuc}}\|_2$ is minimized besides the nuclear norm, whereas in Example 8.5 the Frobenius norm $\|\mathbf{A} - \mathcal{H}(\mathbf{p}_{\text{nuc}})\|_F$ is included in the regularized minimization problem (8.3). Therefore, the described behavior is expected for the Frobenius norm error $\|\mathbf{A} - \mathcal{H}(\mathbf{p}_{\text{nuc}})\|_F$, the contrary is anticipated for the vector norm error $\|\mathbf{p}_A - \mathbf{p}_{\text{nuc}}\|_2$.

A possible explanation is that the solution of (8.3) highly depends on the regularization parameter. It might be possible to find an equally good solution in Example 8.3 when further refining the step size for the regularization parameter. We refrain from trying so since in both examples the step size has been the same, and thus results should be comparable. More importantly, we are mostly interested in the Frobenius norm error after all. For this reason, we only include the results from Example 8.5 for comparison in Table 9.4.

Finally we come to the second example matrix that is always invoked in this part of the thesis.

Example 8.6 As in the previous chapters we also consider the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \\ 1/2 & 0 & 1 \end{pmatrix} \quad \text{with parameter vector} \quad \mathbf{p}_A = \begin{pmatrix} 1 & 0 & 1/2 & 0 & 1 \end{pmatrix}^\top.$$

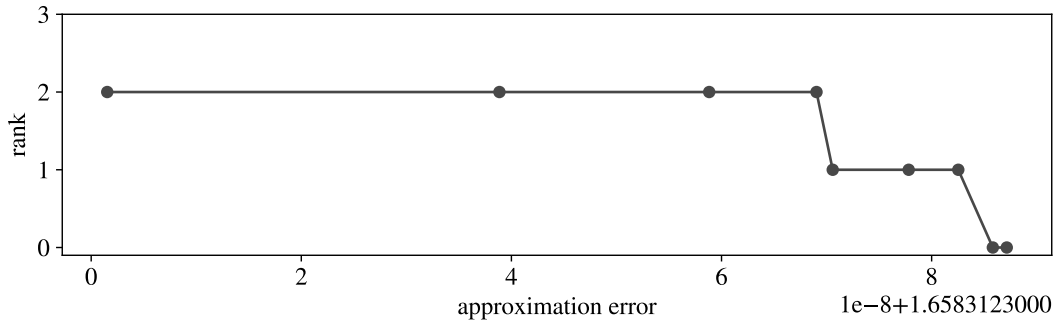


Figure 8.7 Trade-off curve between the approximation error $\|\mathbf{p}_A - \mathbf{p}\|_W = \|\mathbf{A} - \mathcal{H}(\mathbf{p})\|_F$ and the rank of $\mathcal{H}(\mathbf{p})$ for Example 8.6. The regularization parameters γ are in $[0.3, 0.4]$ with 0.01 as step size.

For this example we directly use the Frobenius norm in the penalization term of (8.3); that is, we use \mathbf{W} from (5.4). Furthermore, we immediately jump to the interesting range of regularization parameters. With γ between 0.3 and 0.4 with step size 0.01 we obtain the trade-off curve depicted in Figure 8.7.

At first glance, this trade-off curve might look like expected, compare Figure 8.1. We have to observe, however, that the approximation errors are all given in the order of $\|\mathbf{p}_A - \mathbf{p}\|_W \approx 1.6583 + 10^{-8}$. This means that the approximation errors are suspiciously close (deviation only in the eighth decimal digit) to the norm of the input matrix $\|\mathbf{A}\|_F \approx 1.6583$.

Nevertheless, we consider the output with smallest approximation error. It is calculated by [ADV21] for $\gamma = 0.33$, and is given by

$$\mathbf{p}_{\text{nuc}} \approx \left(1.252 \cdot 10^{-8} \quad 9.298 \cdot 10^{-25} \quad 1.220 \cdot 10^{-8} \quad 4.565 \cdot 10^{-24} \quad 1.252 \cdot 10^{-8} \right)^T.$$

Here, we display only three significant places due to space limitation.

Sure enough, this parameter vector is almost zero. The approximation errors are given by $\|\mathbf{A} - \mathbf{H}_{\text{nuc}}\|_F \approx \|\mathbf{A}\|_F - 2.4592 \cdot 10^{-8}$ in the Frobenius norm, and $\|\mathbf{p}_A - \mathbf{p}_{\text{nuc}}\|_2 \approx 1.5 - 2.0606 \cdot 10^{-8}$ in the Euclidean norm. Of course, the relative errors only differ very slightly from one: $\|\mathbf{A} - \mathbf{H}_{\text{nuc}}\|_F / \|\mathbf{A}\|_F \approx 1 - 1.4830 \cdot 10^{-8} \approx 1.0000$ and $\|\mathbf{p}_A - \mathbf{p}_{\text{nuc}}\|_2 / \|\mathbf{p}_A\|_2 \approx 1 - 1.3737 \cdot 10^{-8} \approx 1.0000$. We display four significant digits of the errors.

In Table 9.2, the errors are given less accurately than here. This is due to reasons of better comparability with the errors produced by other methods.

Remark 8.7 Note that we always have to compute the rank numerically. The numerical rank is the number of singular values that are larger than a certain threshold (compare Lemma 1.4). In Python and MATLAB, this threshold defaults to

$$\text{tol} := \sigma_0 \cdot \max\{M, N\} \cdot \text{eps},$$

where `eps` is the machine precision. It is approximately given by $\text{eps} \approx 2.22 \cdot 10^{-16}$ for standard floating point arithmetic in Python and MATLAB, see also [Hig02; PTV⁺07]. The threshold also depends on the scale of the matrix via its largest singular value σ_0 and its dimensions M and N .

This default threshold is too sensitive for our purposes here. Using it, we do not obtain a single rank-1 matrix for any regularization parameter. For all three examples in this chapter, the solution to problem (8.3) would be either a full rank matrix or the zero matrix.

Hence, we have to set a custom threshold, below which the singular values are considered zero. In the examples in this chapter, this threshold is set to 10^{-8} by trial and error. Setting the threshold smaller than 10^{-8} yields parameter vectors in the order of 10^{-10} which have to be considered zero. A larger threshold is not favourable since it would lead to very inaccurate numerical ranks. However, one might expect that a larger threshold in Example 8.6 might lead to a more useful parameter vector. This is not the case. To the contrary, even a really large threshold of 10^{-3} does not lead to a rank-1 Hankel approximation for Example 8.6 that is comparable to the optimal solution from Chapter 3.

9

NUMERICAL EXAMPLES AND COMPARISONS

So far in this part, we have reviewed a variety of different methods that can be used for rank-1 Hankel approximation. Finally, we want to compare the approximation accuracy of these methods. The minimal error achieved by an optimal approximation from Part I shall always serve as benchmark.

The majority of the methods from this part, as well as the optimal rank-1 Hankel approximation in the spectral norm, only deal with real input matrices and their real approximations. Therefore, we limit our comparisons to real rank-1 Hankel approximation of real matrices. For a comparison between real and complex optimal approximations with respect to the Frobenius norm, see Figure 3.1.

The methods based on local optimization (Chapter 6) and convex relaxation (Chapter 8) make extensive use of the concept of parameter vector introduced in Chapter 5. This is why these methods can only be used to approximate matrices that already feature Hankel structure. Consequently, in this chapter we only approximate Hankel matrices and compare the resulting approximation errors.

9.1 Revisiting Some Examples

In this section, we take up two of the small example matrices that we have encountered repeatedly throughout this thesis. For these matrices, we exemplarily contrast the optimal rank-1 Hankel approximations with respect to the Frobenius and the spectral norm obtained by the methods from Chapters 3 and 4.

Furthermore, we compare the Frobenius norm errors produced by the different SLRA approximation methods from Chapters 6 to 8 by means of these examples. The errors are compared to the minimal error computed by our method from Chapter 3.

The matrix most often used as an illustrative example is

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \\ 1/2 & 0 & 1 \end{pmatrix}, \quad (9.1)$$

which is known from Examples 3.8, 4.12, 6.2, 6.8 and 7.15.

First, we compare the optimal rank-1 Hankel approximations that we obtain for the Frobenius and the spectral norm. For this simple matrix the optimal solutions have been calculated analytically by hand, see Examples 3.8 and 4.12. We list the respective optimal coefficients \tilde{c} and optimal structure parameters \tilde{z} in Table 9.1. Besides, we also register the resulting approximation errors in both the Frobenius and the spectral norm.

| Optimal approximation | optimal \tilde{c} | optimal \tilde{z} | $\ \mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^T\ _2$ | $\ \mathbf{A} - \tilde{c} \cdot \tilde{\mathbf{z}}\tilde{\mathbf{z}}^T\ _F$ |
|------------------------|---------------------|---------------------|---|---|
| for the Frobenius norm | 7/6 | ± 1 | 1.0458 | 1.1785 |
| for the spectral norm | 2 | ± 1 | 0.9574 | 1.4434 |

Table 9.1 Optimal parameters constituting optimal rank-1 Hankel approximations with respect to the Frobenius norm (Example 3.8) and the spectral norm (Example 4.12). The errors are rounded to four decimal digits and minimal errors are displayed in boldface.

In this particular example, the optimal structure parameter for the approximation with respect to the Frobenius norm and the one for the spectral norm coincide. This is usually not the case, see for example Table 9.3. Nonetheless, it is apparent from the different values of \tilde{c} , that the optimal solutions for Frobenius and spectral norm differ.

Note that the spectral norm of a matrix is always smaller than its Frobenius norm. In order to find the minimal r1H error, we have to compare the errors within the same column of Table 9.1. We see that for both norms the approximation error is significantly smaller when the approximation method is specifically designed for the respective norm.

Remark 9.1 As all the entries in the matrix (9.1) are real and non-negative, there is an optimal rank-1 Hankel approximation with respect to the Frobenius norm that is also real,

see Remark 3.14. Thus, there is no need to search for complex optimal parameters \tilde{c} and \tilde{z} for the Frobenius norm. The same holds true for the second example matrix (9.2) considered in this section.

The matrix \mathbf{A} in (9.1) has been approximated by a rank-1 Hankel matrix using each of the introduced methods, see Examples 3.8, 4.12, 6.2, 6.8 and 7.15. The respective resulting approximation errors in the Frobenius norm are listed in Table 9.2.

As mentioned before, the rank-1 Hankel approximation for the matrix (9.1) has also been computed with respect to the spectral norm. However, since most of the methods presented in Part II are designed for the (weighted) Frobenius norm only, it makes little sense to compare the errors in the spectral norm. This is why it does not occur in Table 9.2.

In the table, the method “optimal” refers to the optimal rank-1 Hankel approximation in the Frobenius norm from Chapter 3. The terms “kernel” and “image” are short notation for the local optimization methods using the kernel and the image representation of the rank constraint from Sections 6.1 and 6.2, respectively. “Cadzow” naturally stands for Cadzow’s algorithm from Chapter 7. Note that the result for Cadzow’s algorithm has been calculated analytically such that the respective values in the bottom three rows of Table 9.2 are exact. The convex relaxation heuristic using the nuclear norm from Chapter 8 is abbreviated by “nucnorm” in Table 9.2. These notations will be used throughout this chapter.

| Method | optimal | kernel | image | Cadzow | nucnorm |
|---|---------------|--------|--------|--------|---------|
| $\ \mathbf{A} - \mathcal{H}(\mathbf{p})\ _F$ | 1.1785 | 1.3229 | 1.6583 | 1.6583 | 1.6583 |
| $\ \mathbf{p}_A - \mathbf{p}\ _2$ | 1.0304 | 1.1180 | 1.5000 | 1.5 | 1.5000 |
| $\ \mathbf{A} - \mathcal{H}(\mathbf{p})\ _F / \ \mathbf{A}\ _F$ | 0.7107 | 0.7977 | 1.0000 | 1 | 1.0000 |
| $\ \mathbf{p}_A - \mathbf{p}\ _2 / \ \mathbf{p}_A\ _2$ | 0.6869 | 0.7454 | 1.0000 | 1 | 1.0000 |

Table 9.2 Absolute and relative approximation errors rounded to four digits in the Frobenius norm and the Euclidean norm of parameter vectors for Examples 3.8, 6.2, 6.8 and 7.15. The different columns correspond to different SLRA methods. The minimal errors are displayed in boldface.

This is a peculiar example since for three of the methods (the image method, Cadzow’s algorithm, and the nuclear norm heuristic) the approximation error is just the norm of the matrix \mathbf{A} itself. This implies that these methods only return the zero matrix rather than a matrix of true rank one. In contrast, even compared to the optimal errors, the kernel method yields quite good results.

The second example matrix we want to examine more closely is

$$\mathbf{A} = \begin{pmatrix} 3 & 2 & 1 & 1 \\ 2 & 1 & 1 & 2 \\ 1 & 1 & 2 & 5 \\ 1 & 2 & 5 & 2 \end{pmatrix}, \quad (9.2)$$

see also Examples 6.1, 6.7 and 8.5.

We compute the optimal rank-1 Hankel approximations of this matrix with respect to the Frobenius norm and the spectral norm via the methods established in Chapters 3 and 4, respectively. The respective optimal parameters and resulting approximation errors are listed in Table 9.3. In this example, the optimal solutions for the Frobenius and the spectral norm differ in both the optimal coefficient \tilde{c} and the optimal structure parameter \tilde{z} . This behavior, in contrast to the first example matrix (9.1), is expected and is the more generic one.

| Optimal approximation | optimal \tilde{c} | optimal \tilde{z} | $\ \mathbf{A} - \tilde{c} \cdot \tilde{z}\tilde{z}^T\ _2$ | $\ \mathbf{A} - \tilde{c} \cdot \tilde{z}\tilde{z}^T\ _F$ |
|------------------------|---------------------|---------------------|---|---|
| for the Frobenius norm | 8.3144 | 1.2256 | 3.2085 | 4.5685 |
| for the spectral norm | 9.9621 | 1.1431 | 3.1595 | 4.9325 |

Table 9.3 Optimal parameters constituting optimal rank-1 Hankel approximations of the matrix (9.2) with respect to the Frobenius norm and the spectral norm. The errors are rounded to four decimal digits. Minimal errors are displayed in boldface.

Rank-1 Hankel approximations of the matrix (9.2) have been calculated using local optimization techniques and the nuclear norm heuristic in Examples 6.1, 6.7 and 8.5. The resulting approximation errors are summarized in Table 9.4. Although not calculated in a specific example, we have also added the error produced by Cadzow's algorithm. Besides, we have of course included the optimal result from Chapter 3.

The matrix (9.2) turns out to be much more benign than the one in (9.1) as all of the relative approximation errors are well below one. In other words, none of the methods employed returns the zero matrix. The kernel method even yields a solution that is hardly distinguishable from the optimum, see Example 6.1 for more details. The image method, too, generates very good results in this case. Its relative approximation error in the Frobenius norm is as small as the optimal error in the displayed precision of four decimal digits.

Recall that approximating the matrix (9.1), the nuclear norm heuristic from Chapter 8 fails

| Method | optimal | kernel | image | Cadzow | nucnorm |
|---|---------------|---------------|---------------|--------|---------|
| $\ \mathbf{A} - \mathcal{H}(\mathbf{p})\ _F$ | 4.5685 | 4.5685 | 4.5687 | 4.5748 | 8.0818 |
| $\ \mathbf{p}_A - \mathbf{p}\ _2$ | 3.5355 | 3.5355 | 3.5460 | 3.6099 | 5.9672 |
| $\ \mathbf{A} - \mathcal{H}(\mathbf{p})\ _F / \ \mathbf{A}\ _F$ | 0.4816 | 0.4816 | 0.4816 | 0.4822 | 0.8519 |
| $\ \mathbf{p}_A - \mathbf{p}\ _2 / \ \mathbf{p}_A\ _2$ | 0.5103 | 0.5103 | 0.5118 | 0.5210 | 0.8613 |

Table 9.4 Absolute and relative approximation errors rounded to four digits in the Frobenius norm and the Euclidean norm of parameter vectors for different SLRA methods, see Examples 6.1, 6.7 and 8.5. The minimal value for each error is displayed in boldface.

to deliver a Hankel approximation of true rank one. Admittedly, so do Cadzow’s algorithm and the image method for that peculiar example. But from Table 9.4, we see that also for the matrix (9.2) the nuclear norm heuristic yields decidedly worse results than the other methods. We conclude that it is not really suited for the r1H problem after all. Additionally, finding the critical range, where the regularization parameter provides a good balance involves a lot of manual work, compare Examples 8.3 and 8.6. For these reasons we exclude the nuclear norm heuristic from the broader comparisons that we conduct in the next section.

9.2 More Comparisons

In this section, we want to compare the different methods applied to the r1H problem on a broader basis. To this end, we generate ten matrices $\mathbf{A}^{(i)}$, $i = 1, \dots, 10$, containing random entries in the interval $[-50, 50]$. Given this wide range of entries, the norms of the input matrices $\|\mathbf{A}^{(i)}\|$ may vary greatly. Thus, the absolute approximation errors $\|\mathbf{A}^{(i)} - \mathbf{H}_{\text{method}}^{(i)}\|$ are not comparable. This is why in this section, we only use the relative errors

$$\text{RE}_{\text{method}}^{(i)} := \frac{\|\mathbf{A}^{(i)} - \mathbf{H}_{\text{method}}^{(i)}\|}{\|\mathbf{A}^{(i)}\|}, \quad i = 1, \dots, 10$$

to compare the different methods’ approximation accuracies. The relative errors are, by nature, contained in the interval $[0, 1]$ and therefore much more useful for comparison. Here $\mathbf{H}_{\text{method}}$ is the rank-1 Hankel matrix obtained by the respective method (optimal, kernel, image, or Cadzow).

In addition to the individual relative approximation errors $\text{RE}^{(i)}$, we also evaluate the

mean

$$\text{MRE}_{\text{method}} := \frac{1}{10} \cdot \sum_{i=1}^{10} \frac{\|\mathbf{A}^{(i)} - \mathbf{H}_{\text{method}}^{(i)}\|}{\|\mathbf{A}^{(i)}\|} = \frac{1}{10} \cdot \sum_{i=1}^{10} \text{RE}_{\text{method}}^{(i)} \quad (9.3)$$

over all ten relative approximation errors.

In order to better assess the approximation accuracies of the methods from Part II, we furthermore determine mean deviations from the respective optimal relative approximation error. More precisely, we calculate the mean absolute deviation (MAD) from the optimal relative error

$$\text{MAD}_{\text{method}} := \frac{1}{10} \cdot \sum_{i=1}^{10} |\text{RE}_{\text{opt}}^{(i)} - \text{RE}_{\text{method}}^{(i)}|, \quad (9.4)$$

and the mean squared deviation (MSD) from the optimal relative error

$$\text{MSD}_{\text{method}} := \frac{1}{10} \sum_{i=1}^{10} (\text{RE}_{\text{opt}}^{(i)} - \text{RE}_{\text{method}}^{(i)})^2. \quad (9.5)$$

In these definitions, $\text{RE}_{\text{opt}}^{(i)}$ is the minimal relative approximation error obtained by the optimal methods from Chapters 3 and 4. Correspondingly, $\text{RE}_{\text{method}}^{(i)}$ is the relative approximation error produced by one of the methods (kernel, image, Cadzow) from Part II.

First, we compare approximation errors resulting from different r1H methods in the Frobenius norm in Section 9.2.1. Then, in Section 9.2.2, we add comparisons of approximation errors in the spectral norm.

9.2.1 Approximation with Respect to the Frobenius Norm

All of the methods presented in Part II of this thesis can be used to solve the r1H problem in the Frobenius norm. However, these methods are known to yield only approximative solutions. In this section, we want to compare the approximation accuracies of kernel, image, and Cadzow's method to the optimal rank-1 Hankel approximation with respect to the Frobenius norm from Chapter 3. Using each of the aforementioned methods, we compute rank-1 Hankel approximations for matrices of different sizes. Ensuing, we compare the resulting individual relative approximation errors as well as the mean errors (9.3)–(9.5).

For reasons explained in the end of Section 9.1, the rank-1 Hankel approximation via nuclear norm minimization does not appear here.

For the first comparison, we use ten randomly generated (4×4) Hankel matrices with entries in the interval $[-50, 50]$. Specifically, we randomly generate ten parameter vectors $\mathbf{p}_A \in \mathbb{R}^{n_p}$ with $n_p = 7$. From these parameter vectors we assemble ten (4×4) Hankel matrices according to (5.1). We use Hankel matrices to begin with because the approximation methods based on local optimization from Chapter 6 can only deal with structured input. This does not constitute a severe restriction because of the special structure of the optimal rank-1 Hankel approximation, see Remark 3.2, and the intrinsic nature of Cadzow’s method, see Algorithm 7.2.

Figure 9.1 shows the relative approximation errors $\text{RE}_{\text{method}} = \|\mathbf{A} - \mathbf{H}_{\text{method}}\|_F / \|\mathbf{A}\|_F$, for rank-1 approximation of each method and each input matrix. The methods compared are the optimal one, the local optimization techniques using kernel and image representation of the rank constraint, and Cadzow’s method.

In most of the depicted examples, the errors lie relatively close together. Nevertheless, we see that the local optimization methods (kernel and image) rarely produce a visibly larger error than the optimal solution. Cadzow’s method does sometimes provide notably worse approximations than the others. The errors contributed by Cadzow’s, kernel and image method are larger than the optimal errors in all cases.

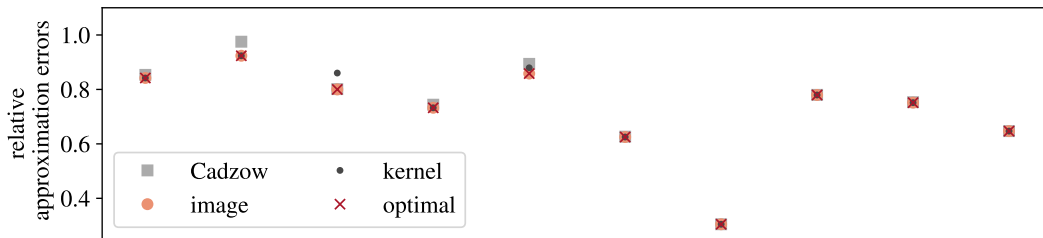


Figure 9.1 Relative rank-1 Hankel approximation errors in the Frobenius norm obtained by different methods for ten randomly generated (4×4) Hankel matrices with entries in $[-50, 50]$.

This is reflected in Table 9.5 where the mean relative approximation errors (MREs) (9.3) in the Frobenius norm are listed for each method. The optimal rank-1 Hankel approximation clearly produces the smallest MRE. The next best approximations on average are achieved by the image method, followed by the kernel method and finally Cadzow’s algorithm. The mean absolute deviations (MADs) from the minimal error (9.4) validate this ranking of the methods. As a consequence of the overall very similar approximation errors, the mean squared deviations (MSDs) are inconclusive in this example.

| Method | optimal | kernel | image | Cadzow |
|-------------------------|---------|--------|--------|--------|
| mean relative error | 0.7264 | 0.7346 | 0.7315 | 0.7379 |
| mean absolute deviation | | 0.0081 | 0.0051 | 0.0115 |
| mean squared deviation | | 0.0004 | 0.0003 | 0.0004 |

Table 9.5 Mean relative errors (9.3) in the Frobenius norm, and mean absolute and squared deviations from the minimal relative error (9.4) and (9.5) for different methods. Input matrices are the same ten randomly generated (4×4) Hankel matrices with entries in $[-50, 50]$ as used for Figure 9.1. All values are rounded to four decimal digits.

Next, we want to test the methods on larger matrices. We generate ten parameter vectors $\mathbf{p}_A \in \mathbb{R}^{n_p}$ with $n_p = 19$ and random entries between -50 and 50 . From these parameter vectors we synthesize ten (10×10) Hankel matrices. For these we conduct the same comparison as before for the (4×4) matrices.

Table 9.6 lists the mean relative error (9.3) for each method as well as the mean deviations (9.4) and (9.5) from the minimal relative error. Our optimal solution provides the smallest MRE.

As before, we derive a ranking of the remaining methods according to their MREs. This ranking is the same as for the (4×4) matrices. More explicitly, the image method yields the second best mean relative error after the optimal solution. Next in line is the kernel method, and Cadzow's iteration comes last. Note that the differences between the methods in both MRE and mean deviations from the optimal error are more noticeable than in Table 9.5.

| Method | optimal | kernel | image | Cadzow |
|-------------------------|---------|--------|--------|--------|
| mean relative error | 0.8839 | 0.9021 | 0.8932 | 0.9130 |
| mean absolute deviation | | 0.0181 | 0.0092 | 0.0291 |
| mean squared deviation | | 0.0012 | 0.0005 | 0.0019 |

Table 9.6 Mean relative errors in the Frobenius norm, and mean absolute and squared deviations from the minimal relative error. Input matrices are ten randomly generated (10×10) Hankel matrices with entries in $[-50, 50]$. All values are rounded to four decimal digits.

The same impression is also gained when considering the relative approximation errors for each random matrix individually. We depict the relative approximation error for each

method and each random matrix in Figure 9.2. The marks for the errors of the different methods are wide-spread compared to Figure 9.1. This supports the conclusion that the optimal approximation from Chapter 3 can deal better with larger matrices than the other methods.

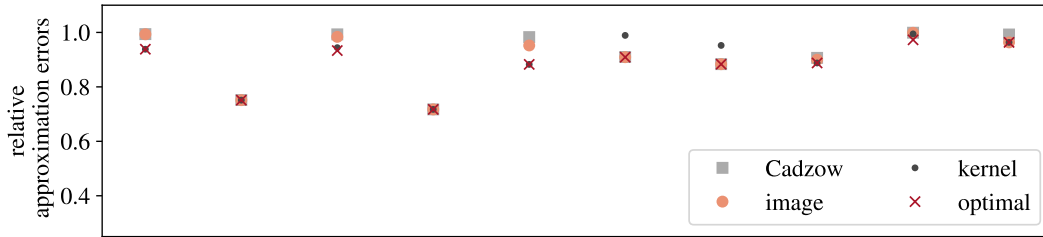


Figure 9.2 Relative rank-1 Hankel approximation errors obtained by different methods for ten randomly generated (10×10) Hankel matrices with entries in $[-50, 50]$. The same initial matrices are at the basis of Table 9.6.

Besides the relative approximation errors, we also want to inspect the actual ranks of the supposed rank-1 Hankel approximations. Here, we are confronted with the problem of numerical rank computation. Similarly as in Chapter 8 (see especially Remark 8.7), the default threshold (used in both Python and MATLAB) of

$$\text{tol} = \sigma_0 \cdot \max\{M, N\} \cdot \text{eps}$$

with $\text{eps} \approx 2.22 \cdot 10^{-16}$ seems too sensitive for the local optimization methods. Using this default threshold, we obtain approximation matrices with a numerical rank of around four and up to eight for the kernel method. For the image method, we even obtain approximation matrices of full numerical rank.

For these two methods, we choose to replace the default by the significantly larger thresholds of $\text{tol}_{\text{ker}} = 10^{-12}$ for the kernel method and $\text{tol}_{\text{img}} = 10^{-10}$ for the image method. The resulting numerical ranks are shown in Figure 9.3. Computed with the custom threshold of $\text{tol}_{\text{ker}} = 10^{-12}$, the kernel method actually produces rank-1 Hankel matrices as demanded. The image method, however, still repeatedly contributes matrices of rank two or three although computed with the larger threshold $\text{tol}_{\text{img}} = 10^{-10}$.

This information on the actual ranks puts the good approximation results (Tables 9.5 and 9.6 and Figures 9.1 and 9.2) of the image method into a different perspective. While granting really good approximation errors, it repeatedly does not obey the rank-1 constraint.

This fact also contradicts the claim cited in Remark 6.6, compare also Remark 6.9. Thus, the image method is not as well suited for rank-1 Hankel approximation after all.

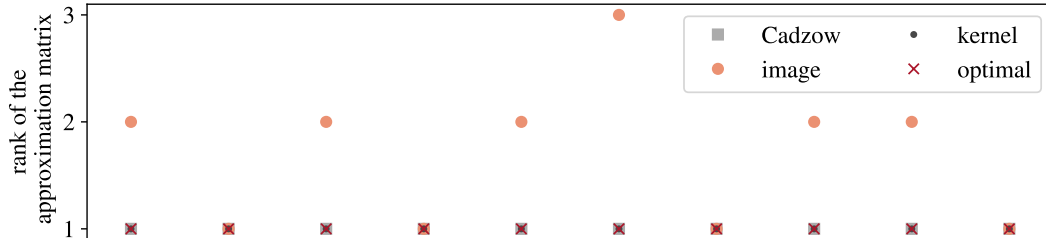


Figure 9.3 Actual ranks of the Hankel matrices approximating the same (10×10) Hankel matrices from Figure 9.2. The ranks are computed using the following thresholds: $\tau_{01_{ker}} = 10^{-12}$ for the kernel method, $\tau_{01_{img}} = 10^{-10}$ for the image method, and the default threshold τ_{01} for Cadzow’s algorithm and the optimal solution.

Now, we also examine the approximation behavior on rectangular matrices since so far, we have only considered the approximation of square matrices. We reuse the same entries, that is, the same ten parameter vectors $\mathbf{p}_A \in [-50, 50]^{19}$ generated for the (10×10) matrices. Now we assemble them to Hankel matrices of different shapes. In other words, we apply the Hankel structure operator $\mathcal{H}_{M,N}$ from (5.1) to the ten parameter vectors for different values of M and N such that $M + N - 1 = n_p = 19$ and $M \leq N$.

For each shape and each approximation method, we compute the MRE over ten input matrices. Its development for decreasing number of rows M and constant number of parameters $n_p = 19$ is depicted in Figure 9.4.

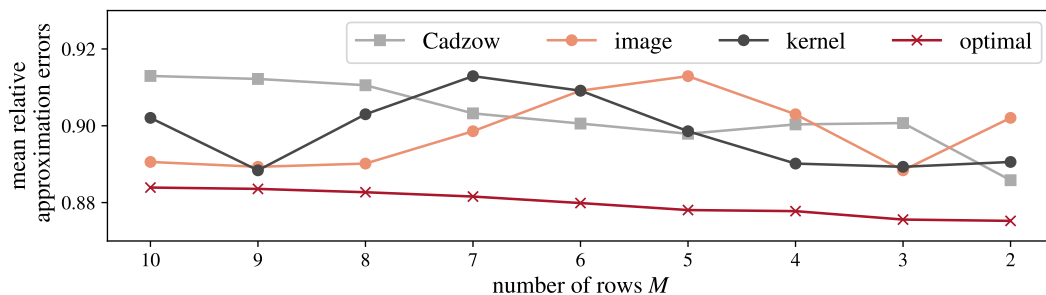


Figure 9.4 Development of the mean relative approximation errors for increasing difference between M and N while $n_p = 19$ remains constant.

We perceive that the optimal error decreases slightly for decreasing number of rows. This is not surprising since for decreasing number of rows M , the rank of the initial matrix, $\text{rank } \mathbf{A} \leq M$, necessarily decreases, too. Thus, the rank reduction becomes less severe and smaller relative errors can be expected for decreasing M .

Among the non-optimal low-rank approximation methods, only Cadzow's algorithm seems to exhibit this decaying behavior. The methods based on local optimization both feature an increase in the mean relative approximation error before eventually decreasing.

We briefly study the numerical complexity of the different SLRA methods. Recall the necessary condition for an optimal real rank-1 Hankel approximation of a real initial matrix $\mathbf{A} = (a_{jk})_{j,k=0}^{M-1,N-1}$ from Theorem 3.6. More precisely, if \tilde{z} is an optimal structure parameter, then we must have

$$Q(\tilde{z}) = a'(\tilde{z}) \cdot p(\tilde{z}) - a(\tilde{z}) \cdot p'(\tilde{z}) = 0,$$

where $a(z)$ and $p(z)$ are the functions

$$a(z) = \sum_{j=0}^{M-1} \sum_{k=0}^{N-1} a_{jk} \cdot z^{j+k} \quad \text{and} \quad p(z) = \left(\sum_{j=0}^{M-1} z^{2j} \right)^{1/2} \cdot \left(\sum_{j=0}^{N-1} z^{2j} \right)^{1/2}.$$

In the case when the initial matrix \mathbf{A} is square (i.e., $M = N$), both functions $a(z)$ and $p(z)$ are polynomials of degree $2N - 2$. Consequently, $Q(z)$ is a polynomial of degree $4N - 5$. Candidates for optimal structure parameters \tilde{z} can now be found by computing the roots of $Q(z)$. For rectangular initial matrices (i.e., $M \neq N$), similar arguments reduce the r1H problem to finding the roots of a polynomial of degree $3M + 3N - 1$. We then have to determine which roots correspond to the maximum of the function $|F(z)| = \frac{|a(z)|}{p(z)}$.

The root finding can be done via an eigendecomposition of the $(4N - 5) \times (4N - 5)$ companion matrix of $Q(z)$, which comes at the cost of roughly $\mathcal{O}((M + N)^3)$ operations, see [HJ13, Sec. 3.3; Dem97, Sec. 4.5]. Clearly, the computation of the roots is the dominating step in terms of numerical complexity.

With $M \leq N$, one iteration step of the kernel method has a numerical complexity of $\mathcal{O}((M - r)^3 \cdot N^3)$, and the numerical complexity of the image method is given by $\mathcal{O}(M \cdot N^3 \cdot r^2)$ per iteration, see [IUM14]. For $r = 1$, these complexities simplify to $\mathcal{O}((M - 1)^3 \cdot N^3) = \mathcal{O}(M^3 \cdot N^3)$ for the kernel method and to $\mathcal{O}(M \cdot N^3)$ for the image

method.

Taking the special problem structure into account, the complexity for the kernel method has been improved to about $\mathcal{O}((M-r)^3 \cdot MN)$ in [UM14]. For rank-1 approximation, this becomes $\mathcal{O}(M^4 \cdot N)$.

In terms of numerical complexity, Cadzow's algorithm is dominated by the SVD, which has to be performed in each iteration. The numerical cost of performing one SVD is about $\mathcal{O}(N^3)$, see [Dem97, Sec. 5.4]. According to [WCW⁺21], the complexity of Cadzow's algorithm can be reduced to $\mathcal{O}((M+N) \cdot r^2 + (M+N) \cdot r \cdot \log(M+N) + r^3)$ per iteration. This gives $\mathcal{O}((M+N) \cdot \log(M+N))$ in the considered rank-1 setting.

We summarize all method's numerical complexities and their improved versions (if available) in Table 9.7.

| Method | optimal | kernel | image | Cadzow |
|----------------|--------------------------|------------------------------|--------------------------------------|--------------------|
| cost/iteration | * $\mathcal{O}((M+N)^3)$ | $\mathcal{O}(M^3 \cdot N^3)$ | $\mathcal{O}(M \cdot N^3)$ | $\mathcal{O}(N^3)$ |
| improved | | $\mathcal{O}(M^4 \cdot N)$ | $\mathcal{O}((M+N) \cdot \log(M+N))$ | |

Table 9.7 Numerical complexities per iteration for each method in the setting of rank-1 Hankel approximation. *Our optimal method is not an iterative procedure and its total cost is displayed.

As a conclusion, we may say that the local optimization methods, kernel and image, come at the highest numerical cost per iteration. Perhaps, this is because of their versatility in terms of the structure and desired rank r . We observe that the kernel method is better suited for rectangular matrices while the image method is cheaper for (nearly) square matrices.

Taking the more general complexities for the rank- r Hankel approximation problem into account, we see that the kernel method is essentially better for large target ranks. In contrast, the image method (compare Remark 6.6) and Cadzow's algorithm come at lower numerical cost for small target ranks. In general, the complexity of Cadzow's algorithm from [WCW⁺21] is surprisingly low. It will depend on the number of iterations needed if it is in fact cheaper to perform than our method from Chapter 3, since our method computes an optimal rank-1 Hankel approximation in a single step.

In our (10×10) examples, the kernel method needed 10.7 iterations on average with a minimum of 5 and a maximum of 17 iterations. The image method takes at least 14 and at most 29 iterations with an average of 18 iterations. In the same examples, Cadzow's method needs a number of iterations between 30 and 86 until convergence yielding an average of

65.8 iterations. Note that, as the stopping criterion for Cadzow’s algorithm, we have used convergence to a rank-1 Hankel matrix, where the numerical rank is computed with the sensitive default threshold tol . With a less sensitive threshold, the number of iterations can be reduced. However, still more than double the iterations of the kernel or image method are needed when using the larger threshold $\text{tol}_{\text{img}} = 10^{-10}$. Hence, at least in the examples considered here, the optimal method is cheaper than Cadzow’s algorithm.

9.2.2 Approximation with Respect to the Spectral Norm

Recall that opposed to the local optimization techniques, Cadzow’s Algorithm 7.1 is not designed specifically for the Frobenius norm. To the contrary, it consists of the very same steps when computing a structured low-rank approximation with respect to the spectral norm, see also Remark 7.3. Thus it makes sense to compare the results of Cadzow’s method to the optimal solutions in the spectral norm.

We take up the same randomly generated Hankel matrices of sizes (4×4) and (10×10) that we have used for comparison of the r1H errors in the Frobenius norm. For these matrices we compute both the optimal rank-1 Hankel approximations from Chapter 4 and Cadzow’s approximations from Chapter 7.

The resulting relative approximation errors are depicted in Figures 9.5 and 9.6, respectively. Both of them confirm what we expect. Cadzow’s method yields acceptable rank-1 Hankel approximations, but it does not achieve the minimal error.

| Method | (4×4) matrices | | (10×10) matrices | |
|-------------------------|-------------------------|--------|---------------------------|--------|
| | optimal | Cadzow | optimal | Cadzow |
| mean relative error | 0.7411 | 0.7638 | 0.8155 | 0.8379 |
| mean absolute deviation | | 0.0227 | | 0.0223 |
| mean squared deviation | | 0.0013 | | 0.0010 |

Table 9.8 Mean relative errors in the spectral norm, and mean absolute and squared deviations from the minimal relative error. Initial matrices are the same ten randomly generated (4×4) and (10×10) Hankel matrices with entries in $[-50, 50]$ as used for the Frobenius norm, see Tables 9.5 and 9.6. All values are rounded to four decimal digits.

Recall that for the Frobenius norm, the differences in the approximation accuracies are more pronounced for larger matrices. Comparing Figures 9.5 and 9.6, we cannot say the

same for the spectral norm. Therefore, we list the mean relative errors and deviations of Cadzow's error from the minimal errors in Table 9.8. Indeed, MAD and MSD differ only slightly for (4×4) and (10×10) initial matrices. This suggests that Cadzow's method is not sensitive to the size of the initial matrix when we measure the r1H error in the spectral norm.

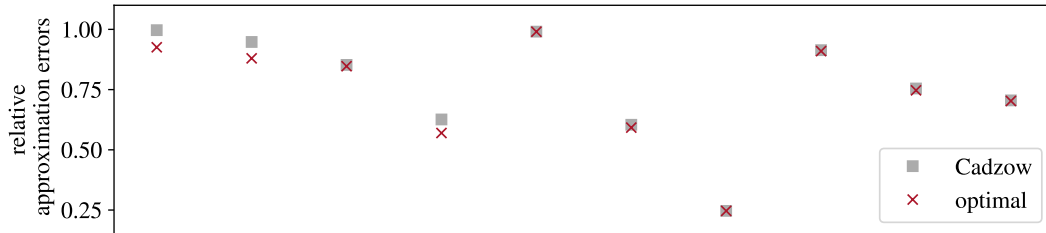


Figure 9.5 Relative errors in the spectral norm for rank-1 Hankel approximation using Cadzow's and the optimal method. Input matrices are the same (4×4) Hankel matrices as in Figure 9.1.

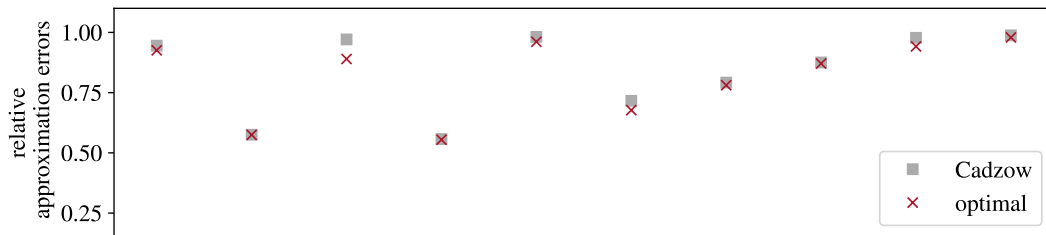


Figure 9.6 Relative errors in the spectral norm for rank-1 Hankel approximation using Cadzow's and the optimal method. Approximated are the same ten (10×10) Hankel matrices as used for Figure 9.2.

Remark 9.2 The optimal rank-1 Hankel approximation in the spectral norm has been derived for real symmetric matrices only. Therefore, we cannot consider the development of the mean relative approximation error for changing numbers of rows and columns as was done in Section 9.2.1.

CONCLUSION AND OUTLOOK

In this dissertation, we have dealt with the rank-1 Hankel approximation (r1H) problem

$$\min \|\mathbf{A} - \mathbf{H}\| \quad \text{such that } \text{rank } \mathbf{H} = 1, \quad (2)$$

and \mathbf{H} has Hankel structure

with respect to the Frobenius norm and the spectral norm. This problem is a special case of the structured low-rank approximation (SLRA) problem (1).

Conclusion

Part I of this dissertation is dedicated to characterizing the *optimal* solutions of the r1H problem (2). This characterization of the optimal solutions is based on the characterization of rank-1 Hankel matrices given in Chapter 2. Depending on the considered norm, the optimal solutions are of different nature. Therefore, we have devoted Chapters 3 and 4 to optimally solving the r1H problem with regard to the Frobenius norm and the spectral norm, respectively.

For each norm, we analytically transform the r1H problem (2) to a maximization problem of a rational function. Using this transformation, we are able to develop an algorithm for the numerical computation of an optimal solution for each norm. For the Frobenius norm, we observe that an optimal solution to the r1H problem always exists. For the spectral norm, however, this is not the case, see Example 4.18. In general, the optimal approximation matrices for the Frobenius norm and the spectral norm usually differ.

We also assess the minimal r1H errors in comparison with the optimal error bounds (1.2) and (1.3) given by the unstructured rank-1 approximation. For the Frobenius norm, the optimal error bound (1.2) is only attained if the rank-1 matrix formed by the first singular triple of \mathbf{A} already has Hankel structure, see Theorem 3.3. For the spectral norm, the conditions for achievement of the optimal error bound (1.3) are less restrictive, see Theorems 4.9 and 4.15.

Part II contains summaries of four SLRA methods from three categories: two methods based on local optimization, an alternating projections procedure called Cadzow’s algorithm, and a convex relaxation heuristic based on the nuclear norm.

In the case of Cadzow’s method, we devise a thorough proof of convergence in the r1H setting. Its limit point, however, is usually not the optimal solution—neither for the Frobenius norm nor for the spectral norm. We conjecture that Cadzow’s limit point and the optimal solution only coincide in the trivial case; namely, when the first singular triple of the initial matrix already forms a Hankel matrix.

All of the aforementioned methods have been adapted to the r1H problem. Then their resulting rank-1 Hankel approximations and errors are compared by means of small examples. In Chapter 9, we conduct comparisons on more, randomly generated initial matrices. The comparisons confirm our theoretical results: our methods from Chapters 3 and 4 indeed always lead to smaller approximation errors than the methods reviewed in Part II. Thus, our optimal solutions to the r1H problem serve as benchmarks for the other methods.

Furthermore, we explain the numerical complexity of our method for real rank-1 Hankel approximation in the Frobenius norm. Then, we compare our method, the two local optimization methods, and Cadzow’s algorithm in terms of numerical complexity. Our optimal method and Cadzow’s algorithm are significantly more efficient than the local optimization methods. The comparison between Cadzow’s and our optimal method depends on how many iterations Cadzow’s method needs until convergence.

Outlook

A first extension of our work could be to evaluate the numerical complexity of our method for the spectral norm. Moreover, it might be interesting to see if it is possible to improve the numerical complexity of our methods from Chapters 3 and 4.

Our approach to solve the r1H problem is based on the characterization of rank-1 Hankel matrices from Chapter 2. One aspect of future research could be to extend this approach to

more general matrix structures, such as Sylvester or block Hankel structures.

Another extension of our work could be to investigate analytical characterizations of the solutions to the rank- r Hankel approximation problem for $r > 1$. Unfortunately, for increasing target rank r , the number of different types of Hankel matrices increases greatly, see Section 2.2. However, there is one generic type of a rank- r Hankel matrix; namely, the sum of r rank-1 Hankel matrices; the remaining types may be viewed as limit cases, see also Section 2.2. Even when disregarding the limit cases, optimal rank- r Hankel approximation in the Frobenius norm leads to the optimization of multivariate polynomials, see [OSS14], and therefore to high numerical complexity. For the spectral norm, an analytical transformation to a manageable problem is not yet known for higher rank.

In order to obtain a Hankel structured approximation of rank- r , another idea could be to iteratively apply our optimal r1H method. While briefly checking this idea, we realized that this strategy does not work out because the iterative application of optimal rank-1 Hankel approximations leads to worse approximations than any SLRA method described in Part II. This is even the case if the initial matrix can be decomposed into the sum of r rank-1 Hankel matrices; that is, if the initial matrix already has the desired structure and rank. Heuristically speaking, the first optimal rank-1 Hankel approximation already incorporates too much of the information that should be encoded in the subsequent ones. It would be interesting to investigate whether an adaptive iterative application of our method can overcome this issue.

BIBLIOGRAPHY

- [AS92] M. Abramowitz and I. A. Stegun, eds. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Reprint of the 1972 edition. Dover Publications, Inc., New York, 1992.
- [AAK71] V. M. Adamjan, D. Z. Arov, and M. G. Krein. Analytic properties of the Schmidt pairs of a Hankel operator and the generalized Schur-Takagi problem. In: *Mathematics of the USSR-Sbornik* 86.128 (1971), pp. 34–75.
- [AN07] J. M. Alongi and G. S. Nelson. *Recurrence and topology*. Vol. 85. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2007. doi: 10.1090/gsm/085.
- [ADV21] M. Andersen, J. Dahl, and L. Vandenberghe. *CVXOPT: Python Software for Convex Optimization*. 2021. URL: <http://cvxopt.org/index.html>.
- [AC11a] F. Andersson and M. Carlsson. A Fast Alternating Projection Method for Complex Frequency Estimation. In: *Proceedings of the Project Review, Geo-Mathematical Imaging Group*. 2011. URL: <https://gmig.science.purdue.edu/pdfs/2011/11-02.pdf>.
- [AC11b] F. Andersson and M. Carlsson. Alternating Projections on Low-Dimensional Manifolds. In: *Proceedings of the Project Review, Geo-Mathematical Imaging Group*. 2011. URL: <https://gmig.science.purdue.edu/pdfs/2011/11-03.pdf>.
- [AC13] F. Andersson and M. Carlsson. Alternating Projections on Nontangential Manifolds. In: *Constructive Approximation* 38.3 (2013), pp. 489–525. doi: 10.1007/s00365-013-9213-3.
- [AC19] F. Andersson and M. Carlsson. Fixed-point algorithms for frequency estimation and structured low rank approximation. In: *Applied and Computational Harmonic Analysis* 46.1 (2019), pp. 40–65. doi: 10.1016/j.acha.2017.03.004.
- [ACO17] F. Andersson, M. Carlsson, and C. Olsson. Convex envelopes for fixed rank approximation. In: *Optimization Letters* 11.8 (2017), pp. 1783–1795. doi: 10.1007/s11590-017-1146-5.

- [Ant98] A. C. Antoulas. Approximation of linear operators in the 2-norm. In: *Linear Algebra and its Applications* 278.1 (1998), pp. 309–316. doi: 10.1016/S0024-3795(97)10087-8.
- [Ant97] A. C. Antoulas. On the Approximation of Hankel Matrices. In: U. Helmke, D. Prätzel-Wolters, and E. Zerz. *Operators, systems, and linear algebra*. 1997.
- [BB96] H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. In: *SIAM Review* 38.3 (1996), pp. 367–426. doi: 10.1137/S0036144593251710.
- [BP17] R. Beinert and G. Plonka. Sparse Phase Retrieval of One-Dimensional Signals by Prony’s Method. In: *Frontiers in Applied Mathematics and Statistics* 3.5 (2017). doi: 10.3389/fams.2017.00005.
- [Ber86] L. Berg. *Lineare Gleichungssysteme mit Bandstruktur und ihr asymptotisches Verhalten*. VEB Deutscher Verlag der Wissenschaften, Berlin, 1986.
- [BM05] G. Beylkin and L. Monzón. On approximation of functions by exponential sums. In: *Applied and Computational Harmonic Analysis* 19.1 (2005), pp. 17–48. doi: 10.1016/j.acha.2005.01.003.
- [BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004. doi: 10.1017/CBO9780511804441.
- [BM86] Y. Bresler and A. Macovski. Exact maximum likelihood parameter estimation of superimposed exponential signals in noise. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 34.5 (1986), pp. 1081–1089. doi: 10.1109/TASSP.1986.1164949.
- [Buc94] V. M. Buchstaber. Time series analysis and Grassmannians. In: *Applied problems of Radon transform*. Vol. 162. American Mathematical Society Translations: Series 2. Providence, RI: American Mathematical Society, 1994, pp. 1–17. doi: 10.1090/trans2/162/01.
- [Cad88] J. A. Cadzow. Signal enhancement—a composite property mapping algorithm. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 36.1 (1988), pp. 49–62. doi: 10.1109/29.1488.
- [CDS01] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. In: *SIAM Review* 43.1 (2001). Reprinted from *SIAM Journal on Scientific Computing* 20.1, pp. 129–159. doi: 10.1137/S003614450037906X.
- [CGM⁺11] G. Chèze, A. Galligo, B. Mourrain, et al. A subdivision method for computing nearest gcd with certification. In: *Theoretical Computer Science* 412.35 (2011), pp. 4493–4503. doi: 10.1016/j.tcs.2011.04.018.
- [CFP03] M. T. Chu, R. E. Funderlic, and R. J. Plemmons. Structured low rank approximation. In: *Linear Algebra and its Applications* 366 (2003), pp. 157–172. doi: 10.1016/S0024-3795(02)00505-0.

- [DeM94] B. De Moor. Total least squares for affinely structured matrices and the noisy realization problem. In: *IEEE Transactions on Signal Processing* 42.11 (1994), pp. 3104–3113.
- [DeM93] B. De Moor. Structured total least squares and L_2 approximation problems. In: *Linear Algebra and its Applications* 188-189 (1993), pp. 163–205. DOI: 10.1016/0024-3795(93)90468-4.
- [Dem97] J. W. Demmel. *Applied numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997. DOI: 10.1137/1.9781611971446.
- [Don06a] D. L. Donoho. Compressed sensing. In: *IEEE Transactions on Information Theory* 52.4 (2006), pp. 1289–1306. DOI: 10.1109/TIT.2006.871582.
- [Don06b] D. L. Donoho. For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution. In: *Communications on Pure and Applied Mathematics* 59 (2006), pp. 797–829.
- [EY36] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. In: *Psychometrika* 1.3 (1936), pp. 211–218. DOI: 10.1007/BF02288367.
- [FHB01] M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In: *Proceedings of the 2001 American Control Conference*. Vol. 6. 2001, pp. 4734–4739. DOI: 10.1109/ACC.2001.945730.
- [Faz02] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis. Stanford University, 2002.
- [FHB04] M. Fazel, H. Hindi, and S. P. Boyd. Rank minimization and applications in system theory. In: *Proceedings of the 2004 American Control Conference*. Vol. 4. 2004, pp. 3273–3278. DOI: 10.1109/ACC.2004.182896.
- [FPS⁺13] M. Fazel, T. K. Pong, D. Sun, et al. Hankel matrix rank minimization with applications to system identification and realization. In: *SIAM Journal on Matrix Analysis and Applications* 34.3 (2013), pp. 946–977. DOI: 10.1137/110853996.
- [FXG18] Y. Feng, J. Xiao, and M. Gu. Flip-Flop Spectrum-Revealing QR Factorization and Its Applications on Singular Value Decomposition. In: *ETNA - Electronic Transactions on Numerical Analysis* 51 (2018). DOI: 10.1553/etna_vol51s469.
- [GZ15] J. Gillard and A. Zhigljavsky. Stochastic algorithms for solving structured low-rank matrix approximation problems. In: *Communications in Nonlinear Science and Numerical Simulation* 21.1-3 (2015), pp. 70–88. DOI: 10.1016/j.cnsns.2014.08.023.
- [Gil10] J. Gillard. Cadzow’s basic algorithm, alternating projections and singular spectrum analysis. In: *Statistics and its Interface* 3.3 (2010), pp. 335–343. DOI: 10.4310/SII.2010.v3.n3.a7.

- [GZ11] J. Gillard and A. Zhigljavsky. Analysis of structured low rank approximation as an optimization problem. In: *Informatica* 22.4 (2011), pp. 489–505. DOI: 10.1007/s12583-011-0202-9.
- [GZ13] J. Gillard and A. Zhigljavsky. Optimization challenges in the structured low rank approximation problem. In: *Journal of Global Optimization* 57.3 (2013), pp. 733–751. DOI: 10.1007/s10898-012-9962-8.
- [Gol65] G. Golub. Numerical methods for solving linear least squares problems. In: *Numerische Mathematik* 7.3 (1965), pp. 206–216. DOI: 10.1007/BF01436075.
- [GNZ01] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky. *Analysis of time series structure. SSA and related techniques*. Vol. 90. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, FL, 2001. xii+305. DOI: 10.1201/9781420035841.
- [GKP94] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete mathematics*. A foundation for computer science. 2nd ed. Addison-Wesley Publishing Company, Reading, MA, 1994.
- [GRG18] C. Grussler, A. Rantzer, and P. Giselsson. Low-Rank Optimization With Convex Constraints. In: *IEEE Transactions on Automatic Control* 63.11 (2018), pp. 4000–4007. DOI: 10.1109/TAC.2018.2813009.
- [GG18] C. Grussler and P. Giselsson. Low-rank inducing norms with optimality interpretations. In: *SIAM Journal on Optimization* 28.4 (2018), pp. 3057–3078. DOI: 10.1137/17M1115770.
- [HR18] G. H. Hardy and S. Ramanujan. Asymptotic Formulæ in Combinatory Analysis. In: *Proceedings of the London Mathematical Society* 17.1 (1918), pp. 75–115. DOI: 10.1112/plms/s2-17.1.75.
- [Hei95] G. Heinig. Generalized inverses of Hankel and Toeplitz mosaic matrices. In: *Linear Algebra and its Applications* 216 (1995), pp. 43–59. DOI: 10.1016/0024-3795(93)00097-J.
- [HR84] G. Heinig and K. Rost. *Algebraic methods for Toeplitz-like matrices and operators*. Vol. 13. Operator Theory: Advances and Applications. Birkhäuser Verlag, Basel, 1984. DOI: 10.1007/978-3-0348-6241-7.
- [Hen74] P. Henrici. *Applied and computational complex analysis*. Vol. 1. Pure and Applied Mathematics: Power series—integration—conformal mapping—location of zeros. Wiley-Interscience, New York-London-Sydney, 1974.
- [Hig02] N. J. Higham. *Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA*. 2nd ed. Society for Industrial and Applied Mathematics, 2002. DOI: 10.1137/1.9780898718027.
- [HJ13] R. A. Horn and C. R. Johnson. *Matrix analysis*. 2nd ed. Cambridge University Press, Cambridge, 2013.

- [HC03] D. A. Huckaby and T. F. Chan. On the Convergence of Stewart’s QLP Algorithm for Approximating the SVD. In: *Numerical Algorithms* 32.2 (2003), pp. 287–316. DOI: 10.1023/A:1024082314087.
- [IUM14] M. Ishteva, K. Usevich, and I. Markovskiy. Factorization Approach to Structured Low-Rank Approximation with Applications. In: *SIAM Journal on Matrix Analysis and Applications* 35.3 (2014), pp. 1180–1204. DOI: 10.1137/130931655.
- [KL98] N. K. Karmarkar and Y. N. Lakshman. On approximate GCDs of univariate polynomials. In: vol. 26. 6. 1998, pp. 653–666. DOI: 10.1006/jSCO.1998.0232.
- [KP11] M. Kauers and P. Paule. *The concrete tetrahedron. Symbolic sums, recurrence equations, generating functions, asymptotic estimates*. Texts and Monographs in Symbolic Computation. SpringerWienNewYork, Vienna, 2011. DOI: 10.1007/978-3-7091-0445-3.
- [Kel21] I. M. Keller. *Modifications of Prony’s Method for the Reconstruction of Structured Functions*. PhD thesis. Georg-August-Universität Göttingen, 2021. URL: <http://hdl.handle.net/21.11130/00-1735-0000-0008-59C9-2>.
- [Kni21] H. Knirsch. Optimal Rank-1 Hankel Approximation in the Spectral Norm for Matrices with Multiple Largest Eigenvalue. In: *Proceedings in Applied Mathematics and Mechanics* 21.1 (2021). DOI: 10.1002/pamm.202100012.
- [KPP21a] H. Knirsch, M. Petz, and G. Plonka. Optimal rank-1 Hankel approximation of matrices: Frobenius norm and spectral norm and Cadzow’s algorithm. In: *Linear Algebra and its Applications* (2021). DOI: 10.1016/j.laa.2021.07.004.
- [KPP21b] H. Knirsch, M. Petz, and G. Plonka. The Difference between Optimal Rank-1 Hankel Approximations in the Frobenius Norm and the Spectral Norm. In: *Proceedings in Applied Mathematics and Mechanics* 20.1 (2021). DOI: 10.1002/pamm.202000085.
- [LO16] V. Larsson and C. Olsson. Convex low rank approximation. In: *International Journal of Computer Vision* 120.2 (2016), pp. 194–214. DOI: 10.1007/s11263-016-0904-7.
- [LMV00] P. Lemmerling, N. Mastronardi, and S. Van Huffel. Fast algorithm for solving the Hankel/Toeplitz structured total least squares problem. In: *Numerical Algorithms* 23.4 (2000), pp. 371–392. DOI: 10.1023/A:1019116520737.
- [LV01] P. Lemmerling and S. Van Huffel. Analysis of the Structured Total Least Squares Problem for Hankel/Toeplitz Matrices. In: *Numerical Algorithms* 27.1 (2001), pp. 89–114. DOI: 10.1023/A:1016775707686.
- [LM08] A. S. Lewis and J. Malick. Alternating Projections on Manifolds. In: *Mathematics of Operations Research* 33.1 (2008), pp. 216–234. URL: <http://www.jstor.org/stable/25151848>.
- [LV10] Z. Liu and L. Vandenberghe. Interior-Point Method for Nuclear Norm Approximation with Application to System Identification. In: *SIAM Journal on Matrix Analysis and Applications* 31.3 (2010), pp. 1235–1256. DOI: 10.1137/090755436.

- [MV05] I. Maravić and M. Vetterli. Sampling and reconstruction of signals with finite rate of innovation in the presence of noise. In: *IEEE Transactions on Signal Processing* 53.8 (2005), pp. 2788–2805. doi: 10.1109/TSP.2005.850321.
- [Mar08] I. Markovsky. Structured low-rank approximation and its applications. In: *Automatica. A Journal of IFAC, the International Federation of Automatic Control* 44.4 (2008), pp. 891–909. doi: 10.1016/j.automatica.2007.09.011.
- [Mar19] I. Markovsky. *Low-Rank Approximation. Algorithms, implementation, applications*. 2nd ed. Communications and Control Engineering Series. Springer, Cham, 2019. doi: 10.1007/978-3-319-89620-5.
- [MU14] I. Markovsky and K. Usevich. Software for weighted structured low-rank approximation. In: *Journal of Computational and Applied Mathematics* 256 (2014), pp. 278–292. doi: 10.1016/j.cam.2013.07.048.
- [MVP05] I. Markovsky, S. Van Huffel, and R. Pintelon. Block-Toeplitz/Hankel structured total least squares. In: *SIAM Journal on Matrix Analysis and Applications* 26 (4 2005), pp. 1083–1099. doi: 10.1137/S089579803434902.
- [MWV⁺05] I. Markovsky, J. C. Willems, S. Van Huffel, et al. Application of structured total least squares for system identification and model reduction. In: *IEEE Transactions on Automatic Control* 50.10 (2005), pp. 1490–1500. doi: 10.1109/TAC.2005.856643.
- [MWV⁺06] I. Markovsky, J. C. Willems, S. Van Huffel, et al. *Exact and Approximate Modeling of Linear Systems. A Behavioral Approach*. Society for Industrial and Applied Mathematics, 2006. doi: 10.1137/1.9780898718263.
- [Mar16] H. Martin. *Rekursive Folgen mit besonderem Fokus auf die Fibonaccifolge*. Diplomarbeit. Universität Wien, 2016. URL: <http://othes.univie.ac.at/43167/1/44899.pdf>.
- [NW06] J. Nocedal and S. Wright. *Numerical Optimization*. 2nd ed. Springer Series in Operations Research and Financial Engineering. Springer, New York, NY, 2006. doi: 10.1007/978-0-387-40065-5.
- [OS95] M. R. Osborne and G. K. Smyth. A modified Prony algorithm for exponential function fitting. In: *SIAM Journal on Scientific Computing* 16.1 (1995), pp. 119–138. doi: 10.1137/0916008.
- [OSS14] G. Ottaviani, P.-J. Spaenlehauer, and B. Sturmfels. Exact solutions in structured low-rank approximation. In: *SIAM Journal on Matrix Analysis and Applications* 35.4 (2014), pp. 1521–1542. doi: 10.1137/13094520X.
- [PP16] G. Plonka and V. Pototskaia. *Application of the AAK theory for sparse approximation of exponential sums*. 2016. doi: 1609.09603v1.
- [PP19] G. Plonka and V. Pototskaia. Computation of adaptive Fourier series by sparse approximation of exponential sums. In: *The Journal of Fourier Analysis and Applications* 25.4 (2019), pp. 1580–1608. doi: 10.1007/s00041-018-9635-1.

- [PT14] G. Plonka and M. Tasche. Prony methods for recovery of structured functions. In: *GAMM-Mitteilungen* 37.2 (2014), pp. 239–258. doi: 10.1002/gamm.201410011.
- [Pot17] V. Pototskaia. *Application of AAK Theory for Sparse Approximation*. PhD thesis. Georg-August-Universität Göttingen, 2017. URL: <http://hdl.handle.net/11858/00-1735-0000-0023-3F4B-1>.
- [PTV⁺07] W. H. Press, S. A. Teukolsky, W. T. Vetterling, et al. *Numerical recipes. The art of scientific computing*. 3rd ed. Cambridge University Press, Cambridge, 2007.
- [Pru86] J. E. Prussing. The principal minor test for semidefinite matrices. In: *Journal of Guidance, Control, and Dynamics* 9.1 (1986), pp. 121–122. doi: 10.2514/3.20077.
- [Rad37] H. Rademacher. On the Partition Function $p(n)$. In: *Proceedings of the London Mathematical Society* 43.4 (1937), pp. 241–254. doi: 10.1112/plms/s2-43.4.241.
- [RFP10] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. In: *SIAM Review* 52.3 (2010), pp. 471–501. doi: 10.1137/070697835.
- [Rum03a] S. M. Rump. Structured Perturbations Part I: Normwise Distances. In: *SIAM Journal on Matrix Analysis and Applications* 25.1 (2003), pp. 1–30. doi: 10.1137/S0895479802405732.
- [Rum03b] S. M. Rump. Structured Perturbations Part II: Componentwise Distances. In: *SIAM Journal on Matrix Analysis and Applications* 25.1 (2003), pp. 31–56. doi: 10.1137/S0895479802405744.
- [Sil00] J. R. Silvester. Determinants of Block Matrices. In: *The Mathematical Gazette* 84.501 (2000), pp. 460–467. doi: 10.2307/3620776.
- [Ste99] G. W. Stewart. The QLP Approximation to the Singular Value Decomposition. In: *SIAM Journal on Scientific Computing* 20.4 (1999), pp. 1336–1348. doi: 10.1137/S1064827597319519.
- [UM12] K. Usevich and I. Markovsky. Structured low-rank approximation as a rational function minimization. In: *IFAC Proceedings Volumes* 45.16 (2012), pp. 722–727. doi: 10.3182/20120711-3-BE-2027.00143.
- [UM14] K. Usevich and I. Markovsky. Variable projection for affinely structured low-rank approximation in weighted 2-norms. In: *Journal of Computational and Applied Mathematics* 272 (2014), pp. 430–448. doi: 10.1016/j.cam.2013.04.034.
- [UM19] K. Usevich and I. Markovsky. Software package for mosaic-Hankel structured low-rank approximation. In: *IEEE 58th Conference on Decision and Control*. 2019, pp. 7165–7170. doi: 10.1109/CDC40024.2019.9028867.
- [VD96] P. Van Overschee and B. De Moor. *Subspace identification for linear systems. Theory—implementation—applications*. Kluwer Academic Publishers, Boston, MA, 1996. xiv+254. doi: 10.1007/978-1-4613-0465-4.

- [VB96] L. Vandenberghe and S. Boyd. Semidefinite programming. In: *SIAM Review* 38.1 (1996), pp. 49–95. doi: 10.1137/1038003.
- [VWD05] B. Vanluyten, J. C. Willems, and B. De Moor. Model Reduction of Systems with Symmetries. In: *Proceedings of the 44th IEEE Conference on Decision and Control*. 2005, pp. 826–831. doi: 10.1109/CDC.2005.1582259.
- [VMB02] M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. In: *IEEE Transactions on Signal Processing* 50.6 (2002), pp. 1417–1428. doi: 10.1109/TSP.2002.1003065.
- [vNeu49] J. von Neumann. On Rings of Operators. Reduction Theory. In: *Annals of Mathematics* 50.2 (1949), pp. 401–485. URL: <http://www.jstor.org/stable/1969463>.
- [vNeu50] J. von Neumann. *Functional Operators. The Geometry of Orthogonal Spaces*. Vol. 2. Annals of Mathematics Studies, No. 22. Princeton University Press, Princeton, N. J., 1950.
- [WCW⁺21] H. Wang, J.-F. Cai, T. Wang, et al. Fast Cadzow’s algorithm and a gradient variant. In: *Journal of Scientific Computing* 88.2 (2021). doi: 10.1007/s10915-021-01550-8.
- [Wil06] H. S. Wilf. *generatingfunctionology*. 3rd ed. A K Peters, Ltd., Wellesley, MA, 2006.
- [ZP19] R. Zhang and G. Plonka. Optimal approximation with exponential sums by a maximum likelihood modification of Prony’s method. In: *Advances in Computational Mathematics* 45.3 (2019), pp. 1657–1687. doi: 10.1007/s10444-019-09692-y.
- [ZG20] N. Zvonarev and N. Golyandina. *Image space projection for low-rank signal estimation: Modified Gauss-Newton method*. 2020. URL: <https://arxiv.org/abs/1803.01419v3>.
- [ZG17] N. Zvonarev and N. Golyandina. Iterative algorithms for weighted and unweighted finite-rank time-series approximations. In: *Statistics and its Interface* 10.1 (2017), pp. 5–18. doi: 10.4310/SII.2017.v10.n1.a1.