The approximate minimization problem of weighted finite automata and applications to language modelling: an approach based on Adamyan-Arov-Krein theory

Clara Lacroce

School of Computer Science
McGill University, Montreal
April 2022

A thesis submitted to McGill University
in partial fulfillment of the requirements of the
degree of Doctor of Philosophy

Abstract

In this thesis, we leverage classical results from the theory of Hankel operators to tackle the approximate minimization problem and its applications to language modelling. In this context, we apply our analysis to weighted finite automata (WFAs) as well as black-box models on sequential data. Given one of these models, we are concerned with finding an approximately minimal realization of the language that it is computing. In particular, we want to construct a weighted finite automaton that fits within a given size constraint and mimics the behaviour of the original model while minimizing the approximation error. We reformulate the problem in terms of low-rank approximation of infinite Hankel matrices and apply Adamyan-Arov-Krein (AAK) approximation theory to solve it.

We first solve the optimal spectral-norm approximate minimization problem for irredundant WFAs over a one-letter alphabet. We present a theoretical analysis based on AAK theory, and provide a closed-form solution, and an algorithm, to compute the optimal approximation of a given size in polynomial time. We then extend these results to black boxes trained for language modelling. We study the conditions under which AAK theory can be applied to find the optimal approximation of a black-box model, without accessing the training data. Moreover, we prove that the proposed method returns an asymptotically-optimal approximation and allows us to use the spectral norm to measure the distance between the black box and the WFA. Finally, we present a framework to apply noncommutative multivariable operator theory to the study of models defined over multi-letter alphabets. We highlight the main obstacles towards a generalization of AAK methods to the multi-letter setting and we conclude by providing possible directions for future work.

Abrégé

Dans cette thèse, nous nous appuyons sur des résultats classiques de la théorie des opérateurs de Hankel pour étudier le problème de la minimisation approximative et ses applications à la modélisation du langage. Dans ce contexte, nous appliquons notre analyse aux automates finis pondérés (WFA) ainsi qu'aux modèles de boîtes noires entrainés sur des données séquentielles. Étant donné un de ces modèles, nous cherchons à obtenir une réalisation approximativement minimale du langage qu'il reconnait. Plus précisément, nous voulons construire un automate fini pondéré qui respecte une contrainte de taille donnée et imite le comportement du modèle original tout en minimisant l'erreur d'approximation. Nous reformulons le problème en termes d'approximation à faible rang de matrices de Hankel infinies et appliquons la théorie d'approximation d'Adamyan-Arov-Krein (AAK) pour le résoudre.

Nous résolvons tout d'abord optimalement le problème de minimisation approximative dans la norme spectrale pour les WFA non-redondants sur un alphabet d'une lettre. Nous présentons une analyse théorique basée sur la théorie AAK, et fournissons une solution analytique, ainsi qu'un algorithme, pour calculer l'approximation optimale d'une taille donnée en temps polynomial. Nous étendons ensuite ces résultats aux boîtes noires entraînées pour la modélisation du langage. Nous étudions les conditions sous lesquelles la théorie AAK peut être appliquée pour trouver l'approximation optimale d'un modèle de boîte noire, sans accéder aux données d'entraînement. De plus, nous prouvons que la méthode proposée fournit une approximation asymptotiquement optimale et nous permet d'utiliser la norme spectrale pour mesurer la distance entre la boîte noire et le WFA. Enfin, nous présentons un cadre permettant d'appliquer la théorie des opérateurs multivariables non-commutatifs à l'étude de modèles définis sur des alphabets multi-lettres. Nous soulignons les principaux obstacles à la généralisation des méthodes AAK au cadre multi-lettres et nous concluons en proposant des directions possibles pour les travaux futurs.

Contribution to Original Knowledge

This thesis contributes to the understanding of the approximate minimization problem in language modelling. To the best of our knowledge, this thesis represents the first attempt to apply the AAK theory to solve the approximate minimization problem of WFAs and black-box models trained for language modelling on sequential data.

Specifically, we make the following contributions.

- A framework to reformulate the approximate minimization problem of models over one-letter or multi-letter alphabets in terms of functional analysis and noncommutative multivariable operator theory, respectively. In both cases, we suggest a way to link the Hankel matrix of a WFA to a Hankel operator and to a complex rational function [LPR22].

- An application of the AAK theory to study the approximate minimization problem of WFAs over a one-letter alphabet [BLP$^+$21].

  - We present a theoretical analysis of the optimal spectral-norm approximate minimization problem for WFAs, based on their connection with finite-rank infinite Hankel matrices. We provide a closed form solution for real weighted automata $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ over a one-letter alphabet, under the assumption $\rho(\mathbf{A}) < 1$ on the spectral radius.

  - We propose a self-contained algorithm that returns the unique optimal spectral-norm approximation of a given size in polynomial time.

  - We bound the approximation error, both in terms of the Hankel matrix (spectral norm) and of the rational function computed by the WFA (ℓ^2 norm).

- An application of the AAK theory to study the approximate minimization problem of black boxes trained computing a function with bounded ℓ^1 norm [LPR21].

  - We propose an algorithm that, given a black-box model $\mathcal{M}$ trained for language modelling on a one-letter alphabet and a target size k, returns a WFA with k states corresponding to an asymptotically-optimal spectral approximation of $\mathcal{M}$. We do not assume any knowledge of the internal structure of the black box, nor of the training data.
  - We use tools from signal processing and arguments from random matrix theory to provide an asymptotic analysis of the approximation problem in the case of infinite-rank infinite Hankel matrices.
  - We propose a new way to compute the distance between a black box and the extracted WFA, based on the AAK theory. We provide bounds on the approximation error in terms of spectral and ℓ^2 norm, and strategies to improve precision when the rank is infinite.

- We lay out possible approaches that can be used to try to address the question of whether or not the proof of the noncommutative AAK theorem can be made constructive.

Contribution of Authors

- Chapters 1, 7, 8, 9 were written by the author specifically for this thesis.
- Chapter 2 and the Appendix present the technical background of this thesis, and were written by the author. Part of this content was already presented in the background section and Appendix of the paper [BLP+21].
- Chapter 3 and 6 were written by the author specifically for this thesis. Part of the content was included in the paper *"Towards an AAK Theory Approach to Approximate Minimization in the Multi-Letter Case"* [LPR22] that was presented at Learnaut 2022 (the 4th edition of the workshop Learning and Automata).
- Chapter 4 is based on the paper *"Optimal Spectral-Norm Approximate Minimization of Weighted Finite Automata"* [BLP+21] that was published at ICALP 2021 (the 48th International Colloquium on Automata, Languages, and Programming). The authors are listed in alphabetical order. The author of this thesis led the project, wrote the paper and leveraged the existing literature in control theory to come up with the theoretical solution proposed. In particular, the construction of the WFA E is based on the all-pass system from the theory of dynamical systems, and the use of the Bartels-Stewart algorithm was proposed by Glover in his solution of the analogous continuous problem [Glo84, CC97, Gu05, Ant05]. Borja Balle had the original idea of applying AAK theory to the approximate minimization problem of WFAs, and gave detailed feedback on the written paper. Guillaume Rabusseau reviewed all the proofs, and provided comments on all the mathematical details. Prakash Panangaden helped with the theoretical study and understanding of the problem, and provided technical direction and supervision of the paper. Doina Precup funded the research and provided feedback on the results.

- Chapter 5 is based on the paper *"Extracting Weighted Automata for Approximate Minimization in Language Modelling"* [LPR21] that was published at ICGI 2020/2021 (the 15th International Conference on Grammatical Inference). The authors are listed in alphabetical order. The author of this thesis led the project, formulated the problem, came up with the solution and wrote the paper. Prakash Panangaden helped with the theoretical understanding of the problem and provided technical direction and supervision of the paper. Guillaume Rabusseau reviewed the mathematical details, in particular the noise analysis and the formulation of the problem.

Acknowledgments

I would like to start by acknowledging my supervisors, Prakash Panangaden and Doina Precup, for giving me this opportunity and for their guidance. This thesis would not have been possible if it weren't for Prakash's unparalleled breadth of knowledge and eclectic interests. Thank you, Prakash, for welcoming me into your group, for the countless hours you dedicated to me and my education, for always replying to my emails within minutes, and for laughing at every single one of my jokes. This is some serious commitment! Thank you, Doina, for the freedom to explore that you left me during the PhD, for the encouragement and the unconditional trust you gifted me from the first day: it was of invaluable help. I would like to thank my external examiner James Worrell, and external member Rémi Eyraud, for taking the time to read my thesis and for all the interesting suggestions. I also would like to thank my internal examiner, Robert Robere, whose detailed comments greatly improved the presentation of this thesis.

I am extremely grateful to my collaborators, Borja Balle and Guillaume Rabusseau. Thank you, Borja, for being such an endless source of ideas, and for always being patient, kind and encouraging, even when I was clearly wrong (the free group, *sigh!*). Thank you, Guillaume, for being the proof that not all heroes wear capes, for always caring, and for your precious feedback (even the night of a deadline).

My deepest gratitude goes to Harsh Satija: a colleague, a friend, a pillar during these years. Thank you for your irreplaceable friendship, and for always supporting me, my work, and my improbable hobbies. Being your deskmate was the biggest stroke of luck I could hope for. Special thanks to Alessandro Sordoni, for being such a great friend, inspiring researcher, and persevering knitter. I also would like to thank my friends in the RL Lab, in particular Tianyu, Tristan, Nicolas, Di, Pascale, Ira, Lianna, Jean, Yue, Nadeem, Roger and Zaf.

I am profoundly grateful to all the people that filled my free time with joy. I would like to thank the staff of the Montreal Children's Hospital Volunteering Services, for welcoming me and offering me a chance to feel at home while in Canada. Shout-out also to the teachers

of the McGill daycare: volunteering with toddlers taught me more about conflict resolution than a decade in university did. Thank you, Sandro and Malorie, for your friendship and all the laughs during the weekends and the Christmas breaks spent at the bakery.

My gratitude goes to my friends in Montréal and Italy: Jack, Audrey, Giulia, Alice, Patrick, Al, Alvaro, Marta, Mattia, Loris, Baldo, Madda, Corbi, and Hofer, for their encouragement and support. Thank you, Maria Elena, because your voice messages are what got me through this degree, 20 minutes at the time. Thank you, Maxime, for your infinite kindness and constant support. Thank you for all the math, the brainstorming, the proofreading, and for always challenging me to be a better researcher and a better person. I am extremely lucky to have you. Finally, I want to acknowledge my family, who never asked me when I would be done with university: this kind of support is hard to come by.

Contents

List of Tables

List of Figures

List of Abbreviations

AAK	Adamyan, Arov and Krein Theory
CNN	Convolutional Neural Network
DFA	Deterministic Finite Automaton
GPA	Generative Probabilistic Automaton
LM	Language Modelling
LM-RNN	Recurrent Neural Network trained for Language Modelling
LSTM	Long Short Term Memory Networks
MIMO	Multi-Input-Multi-Output System
NC	Noncommutative
RNN	Recurrent Neural Network
SISO	Single-Input-Single-Output System
SVA	Singular Value Automaton
SVD	Singular Value Decomposition
WFA	Weighted Finite Automaton

List of Symbols

$\mathbf{1}$	Identity matrix
$\mathbf{M}^+$	Moore-Penrose pseudo-inverse
T^*	Adjoint of the matrix T
$\rho(M)$	Spectral radius of M
$\otimes$	Kronecker product
$\overline{\text{Span}}(\mathcal{M})$	topological closure of the linear span of $\mathcal{M}$
ℓ^2	Space of square summable sequences
$\mathbb{D}$	Unit Disc
$\mathbb{T}$	Unit Circle
$\mathcal{L}^2(\mathbb{T})$	Space of square integrable functions
$\mathcal{R}_k$	Set of strictly proper rational functions of rank k
$\deg(\theta)$	Degree of θ
$\widehat{f}(n)$	Fourier coefficient of f
$\mathcal{H}^2$	Hardy space
$\mathcal{H}^2_-$	Negative Hardy space
$\mathcal{H}^\infty$	Infinite Hardy space
M_f	Operator of left multiplication by f
$\mathbb{B}^d_{\mathbb{N}}$	Noncommutative unit ball
$\rho_{NC}(z)$	Joint spectral radius of $f(z)$

$\mathbb{F}_n$	Free monoid on n generators
F^2	Fock Space
$\mathcal{H}^2(\mathbb{F}_d)$	Noncommutative Hardy Space
$\mathcal{H}^\infty_{\mathrm{NC}}$	Uniformly bounded free noncommutative functions

Chapter 1

Introduction

In the rapidly evolving, performance-oriented field of deep learning, lack of interpretability and high computational costs remain two of the biggest obstacles towards safer and more accessible models [DVK17]. For machine learning systems to be used safely, it is of paramount importance to improve their interpretability. However, neural networks are often black boxes that, while trustworthy, might fail without giving any insight on why that happened [EA20]. On the other hand, the performance of deep learning models has been more and more dependent on increasing computational resources. This rapid escalation in the computational requirements is becoming economically and environmentally unsustainable [TGLM20]. The need to address these issues is at the root of the increasing number of works focusing on knowledge distillation [HVD15, EA20], and on approximating neural networks with other kinds of models, that are easier to interpret.

We are particularly interested in models learning functions defined over sequences of observations. Sequential data is ubiquitous, and is at the basis of many tasks, from natural language processing to reinforcement learning and computational biology. With this class of data, particular attention has been given to the problem of extracting, from a Recurrent Neural Network (RNN) [Elm90, HS97] a weighted finite automaton (WFA) [AEG18, RLP19, WGY19, OWSH20, EA20, SRRS21, ZDX+21]. Weighted finite automata are an

expressive and efficient class of models that are suited for sequence modelling and prediction [DE08, CHM04]. In particular, they can compute any probability distribution defined by a Hidden Markov Model [DE08], and can model the transition and observation behavior of partially observable Markov decision processes [TJ15]. They are also connected to other reinforcement learning frameworks, such as predictive state representations [LSS01, TJ15], and encompass probabilistic automata of various kinds. The advantage of employing WFAs over neural networks is that they are generally easier to interpret, thanks to their graphical representation [HVLS16], and they are faster to compute. In the literature, automata have already been applied in several ways, to improve the understanding, interpretability, and verification of RNNs. For example, Dong et al. [DWS+20] propose a way to analyze a RNN using a probabilistic finite automaton extracted from it, based on the symbolic encoding of RNN hidden state vectors. While weighted automata provide more expressiveness, simpler models like deterministic finite automata have been successfully used in the context of formal verification. In Klep et al. [KNR+21] the authors use active learning to learn a deterministic finite automaton from a given RNN, and use model checking for verification. Deterministic finite automata are applied to the verification problem also by Wang et al. [WZLG18], in the context of adversarial perturbations. Ma et al. [MDL+22] propose an adversarial sequence generation approach for RNNs using symbolic weighted finite automata, an extension of WFAs that can handle strings over infinite alphabets more efficiently [SHYS21].

It is important to remark that there are situations in which even a WFA might become too expensive to compute, or too difficult to interpret, and therefore needs to be approximated using a smaller weighted automaton. This generally occurs when the WFA is already minimal but still has too many states, or when we are required to test for a given property multiple times. If we are willing to trade off some of the accuracy for a faster processing time, a possible approach is to use *approximate minimization* techniques. The approximate minimization problem is concerned with finding an "approximately" minimal approximation of a given model. In the case of automata, given a minimal WFA, the objective is to find

a smaller automaton that mimics its behaviour [BPP19]. The two WFAs compute different functions, so the process produces an approximation error. This can be assessed in several ways: the norm that we choose to measure the error impacts greatly the minimization process.

The approximate minimization problem is related to knowledge distillation and extraction tasks. When the solution of the problem is optimal, this approach has a clear advantage compared to the other methods: it allows us to search for the best WFA among those of a predefined size. This can be particularly useful when dealing with limited computing resources. Moreover, bounding the number of states of a WFA can help improve the interpretability of its graphical representation [HVLS16]. Thus, we can choose the size according to predefined computational or interpretability constraints, and then search for the best approximation of that size. Another advantage of approximate minimization is that it allows for a broader theoretical analysis of the problem. For example, most results in the extraction literature are obtained under the assumption that the RNN is trained over a regular language (an exception can be found in the work of Okudono et al. [OWSH20] and Eyraud and Ayache [EA20]). The experiments performed by Eyraud and Ayache suggest that, even when the RNN is trained on data not corresponding to a WFA, it seems to be computing approximations of rational series [EA20]. Nonetheless, the regular-language setting remains a strong assumption to make without theoretical guarantees of convergence. Tackling this problem using approximate minimization can provide an avenue to investigate "asymptotic" characterizations of the problem. We remark that approximate minimization can be convenient also in the context of spectral learning algorithms [BDR09, BCLQ14, BHP14, HKZ12]. When applied to a learning task, such algorithms start by computing a minimal WFA that explains the training data exactly. Then, they obtain a model that generalizes to unseen data by producing a smaller approximation to the minimal WFA. The size reduction is a crucial step: the exact machine might overfit the data and generalize poorly. In this context, approximate minimization can be particularly useful when the learning algorithm is asked

to produce a WFA smaller than the one that is generating the data.

1.1 Rationale and Objectives

The approximate minimization problem for weighted finite automata was first formalized by Balle et al. [BPP19]. In their paper, the authors provide an algorithm that, given a first WFA, computes a new WFA that is approximately minimal, and obtain bounds in the ℓ^2 norm. The authors conclude the paper highlighting the following interesting research question:

Q1: *Given a bound on the size, can we construct the best possible approximation to a WFA?*

Our first objective is to answer this question, and in this sense, the paper of Balle et al. can be considered as the starting point for this thesis. We then broaden the class of models we consider, and address the following point:

Q2: *Given a bound on the size, can we construct the best possible approximation to a black-box model computing a function with bounded ℓ^1 norm?*

We remark that black box models trained for language modelling are computing functions belonging to this class. The choice of these kind of models, which encompasses RNNs, is in line with the literature of WFA extraction from neural networks. We believe that answering this question is relevant to the field. Indeed, all the methods proposed in the literature provide approximations that are qualitatively good, but hard to compare, and lack theoretical guarantees.

In the case of one-letter alphabets, both questions can be reformulated and solved in terms of functional analysis. In particular, a collection of results on the theory of Hankel operators called the Adamyan-Arov-Krein (AAK) theory can be applied. This set of results has been widely exploited in control theory, in the context of model reduction of linear time-invariant dynamical systems [Glo84]. AAK theory provides us with a way to compute the

optimal low-rank approximation in the spectral norm of the Hankel matrix of a (compact) Hankel operator. Mathematically, these two questions correspond to finding the optimal low-rank approximation in the spectral norm of the finite-rank and infinite-rank infinite Hankel matrices associated with a WFA and a black-box model, respectively.

We evaluate the quality of the approximation in the spectral norm. The first reason why we choose this norm is because of its link with the AAK theory, which allows us to find a global minimum for the approximation error. At the same time, the spectral norm enjoys other interesting properties. For example, it can be computed precisely, and it can be minimized efficiently (in polynomial time). Moreover, it can be used to compare different classes of models, since it is defined on the Hankel matrix and not on the specific architecture considered. For example, it allows to define and to precisely compute the distance between a WFA and a black-box model. Therefore, solving the approximate minimization problem with respect to the spectral norm can help to quantitatively analyze different approximations.

Since the correspondence between the approximate minimization problem and the problem solved by AAK theory is direct only in the one-letter case, the last question that we try to answer is:

Q3: *Can we generalize these results to multi-letter alphabets?*

Addressing this question is central for future applications of this work. The main challenge arises from the fact that the mathematical theory necessary to study the case of multi-letter alphabets is noncommutative. In this setting, the few existing results related to AAK theory are obtained using non-constructive arguments.

1.2 Overview of the Results

In this thesis, we study the approximate minimization problem, with a particular emphasis on the task of language modelling. We focus our attention first on weighted finite automata, and second on black-box models computing a function with bounded ℓ^1 norm. We reformu-

late the approximation problem as a rank minimization problem for Hankel matrices and choose to measure the error in terms of the spectral norm. Our main objective is to obtain a framework to apply tools from AAK theory to the approximation problem, in order to answer the research questions stated before. We divide the problem into two cases, depending on whether the model is computing a function over a one-letter or a multi-letter alphabet. In the first case, the advantage of choosing the spectral norm becomes clear. The AAK theory tells us how to compute the optimal approximation of a given size, and the corresponding approximation error. We propose a method, based on Fourier analysis, to adapt the formalism of AAK theory to the setting of WFAs and black boxes over one-letter alphabets. For both classes of models, we provide theoretical guarantees and an algorithm that returns the optimal approximation of a given size in the spectral norm. In the multi-letter setting, the situation is more complicated, as it requires a noncommutative version of AAK theory. We provide a novel framework for the application of noncommutative multivariable operator theory to the setting of WFAs and black-box models. Moreover, we propose a way to associate a noncommutative rational function to a given WFA. Unfortunately, the proof of the noncommutative version of the AAK theorem is not constructive, so it is not possible to easily obtain an algorithm from it. We illustrate possible approaches and obstacles that need to be overcome in order to extract an algorithm from the theorem.

1.3 Outline of the Thesis

This thesis is organized as follows. In Chapter 2, we set the notation and we present the necessary background notions used throughout the thesis: weighted finite automata, language modelling and AAK theory. The main contributions of this thesis are detailed in chapters 3, 4, 5, 6 and 7. Chapter 3 is concerned with the methodological core of this thesis. After a brief analysis of the salient properties of the spectral norm, we introduce the approximate minimization problem in the broader context of low-rank approximation of infinite Hankel

matrices. We illustrate the high-level approach to solve the problem in the case of one-letter alphabets, where AAK theory can be directly applied. In Chapter 4, we address question **Q1** in the case of one-letter alphabets. In particular, we analyze and solve the approximate minimization problem for irredundant weighted finite automata with weights in $\mathbb{R}$ defined over a one-letter alphabet. Following the method illustrated in Chapter 3, we adapt AAK theory to this setting, and we provide theoretical guarantees, an algorithm for optimal approximation, and bounds on the quality of the approximation in the spectral and ℓ^2 norms. In Chapter 5, we study the approximate minimization problem for black boxes computing a function $f \in \ell^1$ over one-letter alphabets, thus addressing part of question **Q2**. We show that the task of language modelling is encompassed in this setting. We provide an algorithm that returns a weighted finite automaton that fits within a given size constraint and which mimics the behaviour of the original model while minimizing the spectral norm. In particular, we show that minimizing the approximation error between a WFA and a black-box model can be solved optimally in a tractable way. We provide theoretical guarantees and an asymptotic analysis to study the potentially infinite-rank infinite Hankel matrix of the black box, without accessing the training data. In Chapter 6 and Chapter 7 we take on question **Q3** and focus on the approximate minimization problem in the case of multi-letter alphabets. In Chapter 6 we review fundamental results from noncommutative functions theory and introduce a new framework to study the problem in this setting. We propose a way to associate a noncommutative Hankel operator and a noncommutative rational function to a given model. In Chapter 7 we overview several ways to approach the problem of finding an optimal solution in this setting, and highlight where they fall short. In Chapter 8 we present a detailed review of the literature concerned with the approximate minimization problem. In particular, we present a summary of the growing line of work in WFA extraction from RNNs. The approximate minimization problem is tightly related to the task of model reduction of linear time-invariant dynamical systems, as the impulse-response of a discrete time-invariant Single-Input-Single-Output SISO system can be parametrized using

a WFA. Specifically, the line of work on "Hankel-norm approximation" is the closest to our setting, and employs methods from AAK theory. Thus, we provide a literature review of relevant methods studied by the control theory and signal processing communities. Finally, in Chapter 9 we summarize our contributions and detail limitations and possible directions for future work.

Chapter 2

Technical Background

In this chapter, we recall the fundamental definitions and preliminary results that are used throughout the thesis.

2.1 Notation

We denote with $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$ the sets of natural numbers, integers and real numbers, respectively. We use bold letters for vectors and matrices; all vectors considered are column vectors unless otherwise specified. We denote with $\mathbf{1}$ the identity matrix, specifying its dimension only when it is not clear from the context. We denote with $\mathbf{v}(i)$, $\mathbf{M}(i,:)$ and $\mathbf{M}(:,j)$ the i-th component of the vector $\mathbf{v}$, and the i-th row and j-th column of the matrix $\mathbf{M}$, respectively. Given two matrices $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$, $\mathbf{N} \in \mathbb{R}^{d_1' \times d_2'}$ we denote their *Kronecker product* by $\mathbf{M} \otimes \mathbf{N} \in R^{d_1 d_1' \times d_2 d_2'}$ with entries given by $(\mathbf{M} \otimes \mathbf{N})((i-1)d_1' + i', (j-1)d_2' + j') = \mathbf{M}(i,j)\mathbf{N}(i',j')$.

Given a matrix $\mathbf{M} \in \mathbb{R}^{p \times q}$ of rank n, a *rank factorization* is a factorization $\mathbf{M} = \mathbf{PQ}$, where $\mathbf{P} \in \mathbb{R}^{p \times n}$, $\mathbf{Q} \in \mathbb{R}^{n \times q}$ and $\mathrm{rank}(\mathbf{P}) = \mathrm{rank}(\mathbf{Q}) = n$. Let $\mathbf{M} \in \mathbb{R}^{p \times q}$ of rank n, the *singular value decomposition* SVD of $\mathbf{M}$ is the factorization $\mathbf{M} = \mathbf{UDV}^\top$, where $\mathbf{U} \in \mathbb{R}^{p \times n}$, $\mathbf{D} \in \mathbb{R}^{n \times n}$, $\mathbf{V} \in \mathbb{R}^{q \times n}$ are such that $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{1}$, and $\mathbf{D}$ is a diagonal matrix with entries $\sigma_0 \geq \cdots \geq \sigma_{n-1} > 0$. If the previous inequalities are strict, the SVD is unique. The columns of $\mathbf{U}$ and $\mathbf{V}$ are called left and right *singular vectors*, while the entries

$\sigma_0 \geq \cdots \geq \sigma_{n-1} > 0$ of $\mathbf{D}$ are the *singular values*. The *Moore-Penrose pseudo-inverse* $\mathbf{M}^+$ of $\mathbf{M}$ is the unique matrix such that $\mathbf{M}\mathbf{M}^+\mathbf{M} = \mathbf{M}$, $\mathbf{M}^+\mathbf{M}\mathbf{M}^+ = \mathbf{M}^+$, with $\mathbf{M}^+\mathbf{M}$ and $\mathbf{M}\mathbf{M}^+$ Hermitian. We remark that the pseudo-inverse can be computed using the SVD of $\mathbf{M}$: $\mathbf{M}^+ = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^\top$. The *Choleski decomposition* is the factorization of a Hermitian matrix into the product of a lower triangular matrix and its conjugate transpose. The *spectral radius* $\rho(\mathbf{M})$ of a matrix $\mathbf{M}$ is the largest modulus among its eigenvalues.

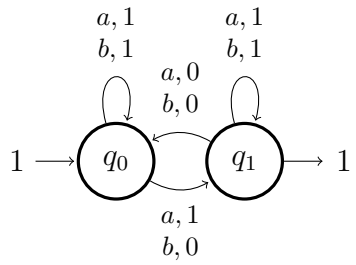
A *Hilbert space* is a complete normed vector space where the norm arises from an inner product. We denote with $\ell^p := \ell^p(\mathbb{N})$ the standard sequence space. Let $\ell^2(\Sigma^*)$ be the Hilbert space of square-summable sequences over Σ^*, with norm defined as $\|f\|_2^2 = \sum_{x\in\Sigma^*} |f(x)|^2$, and inner product $\langle f, g\rangle = \sum_{x\in\Sigma^*} f(x)g(x)$ for $f, g \in \mathbb{R}^{\Sigma^*}$. Let $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$ be the complex unit circle, $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$ the (open) complex unit disc. Let $1 < p < \infty$, $\mathcal{L}^p(\mathbb{T})$ be the space of measurable functions on $\mathbb{T}$ for which the p-th power of the absolute value is Lebesgue integrable. For $p = \infty$, we denote with $\mathcal{L}^\infty(\mathbb{T})$ the space of measurable functions that are bounded, with norm $\|f\|_\infty = \sup\{|f(x)| : x \in \mathbb{T}\}$.

2.2 Weighted Finite Automata

In this section, we provide an overview of weighted finite automata. For a more detailed presentation we refer the reader to [BR11, Ber79, Moh09, Fli74]. We remark that weighted automata can be defined over arbitrary semi-rings. In this thesis we focus only on automata with real weights and the usual addition and multiplication operations. Anytime we mention WFAs, we imply that they have real weights. We follow the notation of Balle, Panangaden and Precup [BPP15], and use a linear-algebraic representation.

Let Σ be a fixed finite alphabet, Σ^* the set of all finite strings with symbols in Σ. We use ε to denote the empty string. Given two letters $p, s \in \Sigma$, we denote with ps their concatenation (this operation directly extends to words in Σ^*).

Definition 2.2.1. *A **weighted finite automaton** (WFA) of dimension n over Σ is a tuple*



$$\boldsymbol{\alpha} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \mathbf{A}_a = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \mathbf{A}_b = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$f_A = \boldsymbol{\alpha}^\top \mathbf{A}_x \boldsymbol{\beta}$$

Figure 2.1: Weighted automaton with two states q_0 and q_1. The initial weights are denoted using arrows pointing to each state, the final weights are point out and the transition weights are given by arrows between states. In particular, this automaton count the number of occurrences of the letter a in a string. On the right side of the figure we can see the corresponding initial vector $\boldsymbol{\alpha}$, final vector $\boldsymbol{\beta}$, and transition matrices $\mathbf{A}_a$ and $\mathbf{A}_b$.

$A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}_{a\in\Sigma}, \boldsymbol{\beta} \rangle$, *where:*

- $\boldsymbol{\alpha} \in \mathbb{R}^n$ *is the vector of initial weights,*
- $\mathbf{A}_a \in \mathbb{R}^{n\times n}$ *is the matrix defined for each symbol* a *and containing the transition weights associated with it,*
- $\boldsymbol{\beta} \in \mathbb{R}^n$ *is the vector of final weights.*

Every WFA A computes a function $f_A : \Sigma^* \to \mathbb{R}$, *i.e.* given a string $x = x_1 \cdots x_t \in \Sigma^*$, it returns $f_A(x) = \boldsymbol{\alpha}^\top \mathbf{A}_{x_1} \cdots \mathbf{A}_{x_t} \boldsymbol{\beta} = \boldsymbol{\alpha}^\top \mathbf{A}_x \boldsymbol{\beta}$. An example of weighted automaton in its linear-algebraic representation is given in Figure 2.1.

Definition 2.2.2. *A function* $f : \Sigma^* \to \mathbb{R}$ *is called* ***rational*** *if there exists a WFA* A *that computes it,* i.e. *such that* $f = f_A$*, where* $f_A(x) = \boldsymbol{\alpha}^\top \mathbf{A}_{x_1} \cdots \mathbf{A}_{x_t} \boldsymbol{\beta} = \boldsymbol{\alpha}^\top \mathbf{A}_x \boldsymbol{\beta}$*, with* $x = x_1 \cdots x_t \in \Sigma^*$*. The* ***rank*** *of* f *is the dimension of the smallest WFA realizing* f*.*

We can consider different classes of automata, computing different kind of functions.

Definition 2.2.3. *We say that a WFA* $A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}_{a\in\Sigma}, \boldsymbol{\beta} \rangle$ *is a* ***generative probabilistic automaton*** *(GPA) if* $f_A(x) \geq 0$ *for every* x*, and* $\sum_{x\in\Sigma^*} f_A(x) = 1$, i.e. *if* f_A *computes a probability distribution over* Σ^**.*

In general, this class of automata can contain pathological examples with states not connected to any final state. To avoid these cases, we introduce the following property on the spectral radius of the transition matrix.

Definition 2.2.4. *Given a WFA* $A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}_{a\in\Sigma}, \boldsymbol{\beta}\rangle$*, let* $\mathbf{A} = \sum_{a\in\Sigma} \mathbf{A}_a$*. The WFA* A *is* ***irredundant*** *if* $\rho(\mathbf{A}) < 1$*.*

Given a function $f : \Sigma^* \to \mathbb{R}$, we can consider a bi-infinite matrix $\mathbf{H}_f \in \mathbb{R}^{\Sigma^*\times\Sigma^*}$ having rows and columns indexed by strings and defined by $\mathbf{H}_f(p, s) = f(ps)$ for $p, s \in \Sigma^*$.

Definition 2.2.5. *A (bi-infinite) matrix* $\mathbf{H} \in \mathbb{R}^{\Sigma^*\times\Sigma^*}$ *is* ***Hankel*** *if, for all* $p, p', s, s' \in \Sigma^*$ *such that* $ps = p's'$*, we have* $\mathbf{H}(p, s) = \mathbf{H}(p', s')$*.*

We remark that, given a Hankel matrix $\mathbf{H} \in \mathbb{R}^{\Sigma^*\times\Sigma^*}$, there exists a unique function $f : \Sigma^* \to \mathbb{R}$ such that $\mathbf{H}_f = \mathbf{H}$.

The following theorem characterizes all Hankel matrices of finite rank.

Theorem 2.2.1 ([CP71, Fli74])**.** *A function* $f : \Sigma^* \to \mathbb{R}$ *can be computed by a WFA if and only if the Hankel matrix* $\mathbf{H}_f$ *has finite rank* n*. In that case,* n *is the minimal number of states of any WFA* A *computing* f*.*

Given a WFA $A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}_{a\in\Sigma}, \boldsymbol{\beta}\rangle$ realizing a rational function f, it is possible to define the forward and backward matrices associated with its Hankel matrix $\mathbf{H}_f$.

Definition 2.2.6. *The* ***forward matrix*** *of* A *is the infinite matrix* $\mathbf{F}_A \in \mathbb{R}^{\Sigma^*\times n}$ *with entries given by* $\mathbf{F}_A(p, :) = \boldsymbol{\alpha}^\top \mathbf{A}_p$ *for any* $p \in \Sigma^*$*, while the* ***backward matrix*** *of* A *is* $\mathbf{B}_A \in \mathbb{R}^{\Sigma^*\times n}$ *given by* $\mathbf{B}_A(s, :) = (\mathbf{A}_s\boldsymbol{\beta})^\top$ *for any* $s \in \Sigma^*$*.*

Let $\mathbf{H}_f$ be the Hankel matrix of f, its forward-backward FB factorization is: $\mathbf{H}_f = \mathbf{F}\mathbf{B}^\top$. The forward and backward matrices are related to two important properties of an automaton.

Definition 2.2.7. *A WFA with* n *states is* ***reachable*** *if* $\operatorname{rank}(\mathbf{F}_A) = n$*, while it is* ***observable*** *if* $\operatorname{rank}(\mathbf{B}_A) = n$*. A WFA is* ***minimal*** *if it is reachable and observable.*

We remark that, if the WFA A is minimal, the FB factorization is a rank factorization [BCLQ14].

2.2.1 Singular Value Automaton

We recall the definition of the singular value automaton [BPP15, BPP19], a canonical form for WFAs.

Definition 2.2.8. *Let f be a rational function and suppose $\mathbf{H}_f$ admits a SVD, $\mathbf{H}_f = \mathbf{U}\mathbf{D}\mathbf{V}^\top$. A **singular value automaton** (SVA) for f is the minimal WFA A realizing f such that $\mathbf{F}_A = \mathbf{U}\mathbf{D}^{1/2}$ and $\mathbf{B}_A = \mathbf{V}\mathbf{D}^{1/2}$.*

Suppose that $\mathbf{F}$ is such that the inner products of its columns

$$\langle \mathbf{F}(:,i), \mathbf{F}(:,j) \rangle = \sum_{x \in \Sigma^*} \mathbf{F}(x,i)\mathbf{F}(x,j) \tag{2.1}$$

are finite for any $i, j = 1, \ldots, n$. Then the positive semi-definite matrix $\mathbf{P} = \mathbf{F}^\top\mathbf{F}$ is well defined. The SVA can be computed with an efficient algorithm relying on the following matrices [BPP19].

Definition 2.2.9. *Let f be a rational function, $\mathbf{H}_f = \mathbf{F}\mathbf{B}^\top$ a FB factorization. If the matrices $\mathbf{P} = \mathbf{F}^\top\mathbf{F}$ and $\mathbf{Q} = \mathbf{B}^\top\mathbf{B}$ are well defined, we call $\mathbf{P}$ the **reachability Gramian** and $\mathbf{Q}$ the **observability Gramian**.*

The Gramians can alternatively be characterized (and computed [BPP19]) using fixed point equations, corresponding to Lyapunov equations when $|\Sigma| = 1$ [Lya50].

Theorem 2.2.2. *Let $A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}_{a\in\Sigma}, \boldsymbol{\beta} \rangle$ be a WFA with n states and well-defined Grami-*

ans $\mathbf{P}$, $\mathbf{Q}$. *Then* $X = \mathbf{P}$ *and* $Y = \mathbf{Q}$ *solve the following equations:*

$$X - \sum_{a \in \Sigma} \mathbf{A}_a X \mathbf{A}_a^\top = \boldsymbol{\beta}\boldsymbol{\beta}^\top \tag{2.2}$$

$$Y - \sum_{a \in \Sigma} \mathbf{A}_a^\top X \mathbf{A}_a = \boldsymbol{\alpha}\boldsymbol{\alpha}^\top. \tag{2.3}$$

In particular, for $|\Sigma| = 1$, the Gramians of $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ are the solutions of:

$$X - \mathbf{A} X \mathbf{A}^\top = \boldsymbol{\beta}\boldsymbol{\beta}^\top \tag{2.4}$$

$$Y - \mathbf{A}^\top Y \mathbf{A} = \boldsymbol{\alpha}\boldsymbol{\alpha}^\top. \tag{2.5}$$

We recall an important property relating the Gramian matrices of a WFA to the singular values of its Hankel matrix.

Lemma 2.2.3. *Let* A *be a minimal WFA with* n *states realizing a rational function* $f \in \ell^2$ *with reachability and observability Gramians* $\mathbf{P}$, $\mathbf{Q}$. *Suppose* $\mathbf{P} = \mathbf{Q} = \mathbf{D}$, *where* $\mathbf{D}$ *is a diagonal matrix having ordered entries* $\sigma_0 \geq \sigma_1 \geq \ldots \sigma_{n-1} \geq 0$. *Then* A *is a singular value automaton, and* $\mathbf{D}$ *is the matrix of singular values of the Hankel matrix* $\mathbf{H}_f$.

We remark that, thanks to this result, we are able to precisely compute the singular values of the Hankel matrix associated to a WFA, despite the fact that this matrix is infinite.

Finally, we can use the Gramians to derive an expression and compute in polynomial time the SVA of a minimal WFA A [BPP19].

Theorem 2.2.4. *Let* A *be a minimal WFA with* n *states realizing a rational function* $f \in \ell^2$ *with reachability and observability Gramians* $\mathbf{P}$, $\mathbf{Q}$. *Let* $\mathbf{Q} = \mathbf{L}_Q \mathbf{L}_Q^\top$ *and* $\mathbf{P} = \mathbf{L}_P \mathbf{L}_P^\top$ *be their Cholesky decomposition. Suppose* $\mathbf{L}_P^\top \mathbf{L}_Q$ *has singular value decomposition* $\mathbf{U}\mathbf{D}\mathbf{V}^\top$. *Then the WFA* $A_{\mathbf{S}} = \langle \mathbf{S}^\top \boldsymbol{\alpha}, \mathbf{S}^{-1} \mathbf{A}_a \mathbf{S}, \mathbf{S}^{-1} \boldsymbol{\beta} \rangle$ *with* $\mathbf{S} = \mathbf{L}_P^{-\top} \mathbf{U} \mathbf{D}^{1/2}$ *is a SVA for* A.

We consider the following example of WFA represented in its SVA form.

Example 2.2.5. Let $A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}, \boldsymbol{\beta} \rangle$ be a weighted finite automaton over a one-letter alphabet $\Sigma = \{a\}$, with:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{\alpha} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} 1 \\ 1 \\ 1, \end{pmatrix}$$

and with Gramians:

$$\mathbf{P} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 2 \\ 0 & 2 & 5 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} 6 & 3 & 1 \\ 3 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Its SVA form is:

$$\mathbf{A} = \begin{pmatrix} 0.579 & 0.461 & 0.046 \\ -0.461 & -0.192 & 0.225 \\ 0.046 & -0.225 & -0.387 \end{pmatrix}, \quad \boldsymbol{\alpha} = \begin{pmatrix} 1.650 \\ -0.851 \\ 0.038 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} 1.650 \\ 0.851 \\ 0.038 \end{pmatrix},$$

with Gramians:

$$\mathbf{P} = \mathbf{Q} = \begin{pmatrix} 4.67 & 0 & 0 \\ 0 & 1.79 & 0 \\ 0 & 0 & 0.12 \end{pmatrix}.$$

The SVA form of a WFA is unique up to the same set of conditions for which the singular value decomposition is unique. Therefore, the use of the SVA form allows us to obtain results and approximation bounds that are representation independent.

2.2.2 Spectral Method

Given a Hankel matrix $\mathbf{H}_f$ of rank n, we can recover the minimal WFA A realizing f by using the method proposed in [BCLQ14].

Let $\Sigma' = \Sigma \cup \{\varepsilon\}$. We consider a basis $\mathcal{B} = (\mathcal{P}, \mathcal{S})$, with $\mathcal{P}, \mathcal{S} \subset \Sigma^*$. Let $\mathbf{H}_\mathcal{B}$ be a

sub-block of $\mathbf{H}_f$ defined over $\mathcal{B}$.

Definition 2.2.10. *The basis* $\mathcal{B} = (\mathcal{P}, \mathcal{S})$ *is* prefix-closed *and* complete *if* $\mathcal{P} = \mathcal{P}' \cdot \Sigma'$ *for some* $\mathcal{P}'$*, and* $\mathbf{H}_\mathcal{B}$ *has rank* n*.*

When $\mathcal{B} = (\mathcal{P}, \mathcal{S})$ is prefix-closed and complete, we can consider the following elements:

- a sub-block $\mathbf{H}_a$ defined over $\mathcal{B}$ by $\mathbf{H}_a(u, v) = \mathbf{H}(u \cdot a, v)$ for each $a \in \Sigma'$;
- a vector $\mathbf{h}_{\mathcal{P},\varepsilon}$ having coordinates $\mathbf{h}_{\mathcal{P},\varepsilon}(u) = \mathbf{H}(u, \varepsilon)$;
- a vector $\mathbf{h}_{\varepsilon,\mathcal{S}}$ having coordinates $\mathbf{h}_{\varepsilon,\mathcal{S}}(v) = \mathbf{H}(\varepsilon, v)$.

Then, from the rank factorization $\mathbf{H}_\mathcal{B} = \mathbf{F}\mathbf{B}^\top$, we can compute a minimal weighted automaton $A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}, \boldsymbol{\beta} \rangle$ for f:

$$\boldsymbol{\alpha}^\top = \mathbf{h}_{\varepsilon,\mathcal{S}}^\top \mathbf{B}^{\top +}, \quad \boldsymbol{\beta} = \mathbf{F}^+ \mathbf{h}_{\mathcal{P},\varepsilon}, \quad \mathbf{A}_a = \mathbf{F}^+ \mathbf{H}_a \mathbf{B}^{\top +}. \tag{2.6}$$

If we denote with $O(k)$ the time required to compute the Hankel matrix $\mathbf{H}_f$, and n is the rank of $\mathbf{H}_f$, we have that the total running time of the spectral algorithm is $O(k + n|\mathcal{P}||\mathcal{S}| + n^2|\mathcal{P}||\Sigma|)$.

2.3 Language Modelling

In this section we briefly introduce the task of language modelling. We also provide a very high level overview of recurrent neural networks and black-box models computing functions assigning a real value to sequential data. The content of this chapter is based on several sources, including [GBC16, Low20, EA20].

2.3.1 Recurrent Neural Networks

Recurrent neural networks [Elm90, HS97], or RNNs, are a class of neural networks designed to process sequential data. The main difference with other kind of neural networks, like

feedforward neural networks, is that RNNs are characterized by feedback loops. These serve to maintain an internal memory based on history information through the hidden states, and allow us to model data that is correlated in time by using sequences as input of the RNN. At each time step, a RNN receives an input and returns a new state vector, depending on the input and on the sequence obtained so far. The hidden states are updated at each time step according to a function whose parameters are the same for all time steps. More specifically, at each time step t we have the following update equations:

$$\begin{cases} h_t = g(\mathbf{W}h_{t-1} + \mathbf{U}x_t + b) \\ y_t = \text{softmax}(\mathbf{V}h_t + c) \end{cases} \tag{2.7}$$

where $\mathbf{W}$ is the matrix of parameters shared between hidden states, $\mathbf{U}$ and $\mathbf{V}$ are the input and output matrices, respectively, and b and c represent input and output bias. The activation function g represents the non-linear component of the RNN. Examples of activation functions are sigmoid, tanh, ReLU.

This class of models can be employed in a variety of tasks, thanks to the several types of existing architectures [WGY18b, MWG+20]. For example, a RNN with a single output at the last time step can be used to make a categorical prediction, while one with an output at each time step can be used as a language model. The most commonly used class of RNNs is that of Long Short Term Memory networks (LSTMs). A LSTM is a recurrent neural network with a gating mechanism determining how information is passed between time steps. We won't analyze in detail the different types of architectures, as it is outside the scope of this thesis. For an overview of their expressive power we refer the reader to the work of Weiss et al. and Merrill et al. [WGY18b, MWG+20].

2.3.2 The Task of Language Modelling

The task of language modelling (LM) consists of predicting the next element in a sequence, given the previous ones. Given a sequence, the joint probability over all sequences is factorized into the product of conditional probabilities. The way the model compute these conditional probabilities vary according to the LM technique that is employed. Language models can be implemented using RNNs [MKB+10]. In particular, a language modelling RNN (LM-RNN) receives as input a sentence, one word at the time, and updates the hidden state h_t at time step t. Then, the model outputs a probability distribution over all words in the vocabulary. This way, the prediction of the next word is dependent on all previous words seen by the model. When computing the probability associated to a string, a LM-RNN defines a distribution over sequences that can be represented by a Hankel matrix.

In this thesis, we are particularly interested in studying the broader case of LM black-box models, where we do not require access to the inner representation of the networks. In addition to RNNs, other important models are encompassed by the general setting of black boxes for language modelling, for example transformers [VSP+17].

2.4 AAK Theory

In this section, we introduce the work of Adamyan, Arov and Krein which has come to be known as AAK theory [AAK71]. This theory applies to Hankel operators and complex functions. We start by recalling some fundamental mathematical definitions from functional analysis [Zhu90]. Then, we define Hankel operators in sequence spaces, and use Fourier analysis to reformulate this theory in function spaces. More details can be found in Appendix A. A comprehensive presentation of the concepts recalled in this section can be found in the books of Nikolski and Peller [Nik02, Pel12], which we follow closely for the exposition of this topic.

2.4.1 Operator Theory

Given a Hilbert space, and $0 < p < \infty$, we write $\|\mathbf{v}\|_p$ to denote the ℓ^p norm of a vector $\mathbf{v}$. We can then define the corresponding induced norm on linear operators.

Definition 2.4.1. *Let $T : X \to Y$ be a linear operator between Hilbert spaces, the **operator norm** of T is defined as:*

$$\|T\| = \sup\{\|Tx\|_Y : \|x\|_X = 1\}.$$

We denote with $\mathbf{T}$ the (infinite) matrix associated with the operator T by some (canonical) orthonormal basis.

We define the following set of properties for linear operators between Hilbert spaces.

Definition 2.4.2. *Let $T : X \to Y$ be a linear operator between Hilbert spaces.*

- *T is **bounded** if it has finite operator norm;*
- *T has **finite rank** if its range has finite dimension;*
- *T is **compact** if the image of the unit ball in X is relatively compact.*

We remark that compact operators between Hilbert spaces can be alternatively defined in terms of converging sequences: an operator is compact if it is the limit of finite-rank operators in the operator norm. If the operator T is the limit in the operator norm of the sequence of finite-rank operators $\{T^i\}_{i\geq 0}$, we denote with $T^i \to T$ the limit:

$$\lim_{i\to\infty} \left\|T^i - T\right\| = 0.$$

Finally, note that a sufficient condition to show compactness is to require that the Hankel operator considered is bounded and of finite rank.

Let $\langle \cdot, \cdot \rangle$ denote the inner product of a given Hilbert space. Let $T : X \to Y$ be a compact operator, the *adjoint* operator T^* is the linear operator $T^* : Y \to X$ such that

$\langle Tx, y\rangle_Y = \langle x, T^*y\rangle_X$, for $x \in X$, $y \in Y$. Given the adjoint operator, it is possible to introduce two definitions that will play a fundamental role in our analysis.

Definition 2.4.3. *The* ***singular numbers*** $\{\sigma_n\}_{n\geq 0}$ *of a compact operator* T *are the square roots of the eigenvalues of the self-adjoint operator* T^*T*, arranged in decreasing order. We say that a singular number is simple if it is not repeated.*

By convention, the singular numbers are arranged in decreasing order and according to their multiplicities. In particular, we have that $\sigma_0(T) = \|(T^*T)\|^{1/2} = \|T\|$.

Definition 2.4.4. *Let* σ *be a singular number for* T*. A* σ-***Schmidt pair*** $\{\boldsymbol{\xi}, \boldsymbol{\eta}\}$ *for* T *is a couple of norm* 1 *vectors such that:*

$$\begin{cases} \mathbf{T}\boldsymbol{\xi} = \sigma\boldsymbol{\eta} \\ \mathbf{T}^*\boldsymbol{\eta} = \sigma\boldsymbol{\xi} \end{cases}$$

Using singular numbers and Schmidt pairs it is possible to define the *Hilbert-Schmidt decomposition*, a generalization of the compact SVD for the infinite matrix of a compact operator T:

$$\mathbf{Tx} = \sum_{n\geq 0} \sigma_n \langle \mathbf{x}, \boldsymbol{\xi}_n\rangle \boldsymbol{\eta}_k. \tag{2.8}$$

Using this decomposition, it is clear that the singular numbers of a compact operator T can also be characterized in relation to $\mathcal{F}_k$, the set of continuous linear operators R with $\text{rank}(R) \leq k$:

$$\sigma_k(T) = \min\{\|T - R\| : R \in \mathcal{F}_k\}.$$

This definition has the advantage that it can be generalized to arbitrary bounded operators, by simply defining:

$$\sigma_k(T) = \text{dist}(T, \mathcal{F}_k). \tag{2.9}$$

Definition 2.4.5. *The* ***spectral norm*** $\|\mathbf{T}\|$ *of the matrix representing the operator* T *is the largest singular number of* T*.*

It is important to note that the spectral norm of $\mathbf{T}$ corresponds to the operator norm of T.

2.4.2 Hankel Operators on Sequences

Given a function $f : \mathbb{N} \to \mathbb{R}$, we consider the Hankel matrix $\mathbf{H}_f$ defined by the Hankel property $\mathbf{H}_f(i,j) = f(i+j)$. This matrix can be interpreted as the expression of a linear (Hankel) operator $H_f : \ell^2 \to \ell^2$ in terms of the canonical basis of the sequence space. The property on the Hankel matrix can be rephrased as an operator identity. Defining the shift operator by $S(x_0, x_1, \dots) = (0, x_0, x_1, \dots)$ and denoting its left inverse by S^*, with $S^*(y_0, y_1, \dots) = (y_1, y_2, \dots)$, we have that H is a Hankel operator if and only if the following **Hankel equation** is satisfied:

$$HS = S^*H. \tag{2.10}$$

The following theorem characterizes bounded Hankel operators on the space of square summable sequences.

Theorem 2.4.1 (Nehari [Neh57])**.** *Let $H : \ell^2 \to \ell^2$ be a Hankel operator with Hankel matrix defined as $\mathbf{H}(j,k) = \{\alpha_{j+k}\}_{j,k\geq 0}$:*

$$\mathbf{H} = \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \dots \\ \alpha_1 & \alpha_2 & \alpha_3 & \dots \\ \alpha_2 & \alpha_3 & \alpha_4 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{2.11}$$

The operator H is bounded on ℓ^2 if and only if there exists a function $\psi \in \mathcal{L}^\infty(\mathbb{T})$ such that:

$$\alpha_m = \widehat{\psi}(m), \quad m \geq 0. \tag{2.12}$$

In this case:

$$\|H\| = \inf\{\|\psi\|_\infty : \widehat{\psi}(n) = \alpha(n),\, n \geq 0\}, \tag{2.13}$$

where $\widehat{\psi}(n)$ is the n-th Fourier coefficient of ψ.

Example 2.4.2. Let $\alpha_n = \frac{1}{n+1}$, $n \geq 0$. The matrix:

$$\mathbf{H} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \dots \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \dots \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

is a Hankel matrix, and is often referred to as the Hilbert matrix. We can apply Theorem 2.4.1 to show that the Hilbert matrix is bounded. In particular, it is possible to show that the function on $\mathbb{T}$:

$$\psi(e^i t) = ie^{-it}(\pi - t), \quad t \in [0, 2\pi)$$

is such that $||\psi||_\infty = \pi$, and:

$$\widehat{\psi}(n) = \frac{1}{n+1}, \quad n \geq 0.$$

2.4.3 Hankel Operators on Hardy Spaces

Using the Fourier isomorphism we can introduce an alternative characterization of Hankel operators with respect to complex function spaces. The first step needed to reformulate the definition is to embed ℓ^2 into $\ell^2(\mathbb{Z})$. This way, we can use the Fourier isomorphism to associate a complex function to each sequence in $\ell^2(\mathbb{Z})$. In particular, given a sequence $\boldsymbol{\mu} = (\mu_0, \mu_1, \dots) \in \ell^2$, we can define two functions in $\mathcal{L}^2(\mathbb{T})$:

$$\mu^- = \sum_{j=0}^{\infty} \boldsymbol{\mu}_j z^{-j-1}, \tag{2.14}$$

$$\mu^+ = \sum_{j=0}^{\infty} \boldsymbol{\mu}_j z^j. \tag{2.15}$$

On the other hand, a function $\phi \in \mathcal{L}^2(\mathbb{T})$ in the complex variable z can be represented by its Fourier expansion:

$$\phi = \sum_{n\in\mathbb{Z}} \widehat{\phi}(n) z^n \tag{2.16}$$

and can be identified, using the orthonormal basis $\{z^n\}_{n\in\mathbb{Z}}$, with the sequence of its Fourier coefficients

$$\widehat{\phi}(n) = \int_{\mathbb{T}} \phi(z)\bar{z}^n dz,\ n \in \mathbb{Z}. \tag{2.17}$$

We can partition the function space $\mathcal{L}^2(\mathbb{T})$ into two subspaces.

Definition 2.4.6. *For $0 < p \leq \infty$, the **Hardy space** $\mathcal{H}^p$ on $\mathbb{T}$ is the subspace of $\mathcal{L}^p(\mathbb{T})$ defined as:*

$$\mathcal{H}^p = \{\phi \in \mathcal{L}^p(\mathbb{T}) : \widehat{\phi}(n) = 0, n < 0\}, \tag{2.18}$$

*while the **negative Hardy space** on $\mathbb{T}$ is the subspace of $\mathcal{L}^p(\mathbb{T})$*

$$\mathcal{H}^p_- = \{\phi \in \mathcal{L}^p(\mathbb{T}) : \widehat{\phi}(n) = 0, n \geq 0\}. \tag{2.19}$$

Hardy spaces can also be defined on the open unit disc $\mathbb{D}$.

Definition 2.4.7. *The **Hardy space** $\mathcal{H}^p(\mathbb{D})$ on $\mathbb{D}$ for $0 < p < \infty$ consists of functions ϕ analytic in $\mathbb{D}$ and such that:*

$$\|\phi\|_p := \sup_{0<r<1} \left(\int_{\mathbb{T}} |\phi(r\xi)|^p dm(\xi) \right)^{1/p} < \infty \tag{2.20}$$

and it is equipped with the norm $\|\cdot\|_p$. For $p = \infty$, $\mathcal{H}^\infty(\mathbb{D})$ is the space of bounded analytic functions in $\mathbb{D}$ with norm:

$$\|\phi\|_\infty := \sup_{\xi\in\mathbb{D}} |\phi(\xi)|. \tag{2.21}$$

Interestingly, $\mathcal{H}^p(\mathbb{D})$ and $\mathcal{H}^p$ can be canonically identified by associating a function ϕ defined in the disc with its limit on the boundary, which is a function in $\mathcal{H}^p$ (a proof can be

found in Nikolski [Nik02]). Thus, we will identify such functions in the unit disc with their boundary value on the unit circle.

Now, we can reformulate the definition of Hankel operators in Hardy spaces (see [Nik02] for a proof that those definitions are equivalent). We denote with $\mathbb{P}_- : \mathcal{L}^2(\mathbb{T}) \to \mathcal{H}^2_-$ the orthogonal projection on the negative Hardy space.

Definition 2.4.8. *Let ϕ be a function in the space $\mathcal{L}^2(\mathbb{T})$. A **Hankel operator** is an operator $H_\phi : \mathcal{H}^2 \to \mathcal{H}^2_-$ defined by $H_\phi g = \mathbb{P}_- \phi g$. The function ϕ is called a **symbol** of the Hankel operator H_ϕ.*

The Hankel matrix $\mathbf{H}$ associated to the Hankel operator $H_\phi : \mathcal{H}^2 \to \mathcal{H}^2_-$ is:

$$\mathbf{H} = \begin{pmatrix} \widehat{\phi}(-1) & \widehat{\phi}(-2) & \widehat{\phi}(-3) & \dots \\ \widehat{\phi}(-2) & \widehat{\phi}(-3) & \widehat{\phi}(-4) & \dots \\ \widehat{\phi}(-3) & \widehat{\phi}(-4) & \widehat{\phi}(-5) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{2.22}$$

We remark that to each Hankel operator we can associate more than one symbol: the function ϕ is a symbol, together with any other function having the same component in the negative Hardy space. This can be seen using a reformulation of Theorem 2.4.1.

Theorem 2.4.3 (Nehari [Neh57])**.** *Let $\phi \in \mathcal{L}^2(\mathbb{T})$ be a symbol for the Hankel operator on Hardy spaces $H_\phi : \mathcal{H}^2 \to \mathcal{H}^2_-$. Then, H_ϕ is bounded on $\mathcal{H}^2$ if and only if there exists $\psi \in \mathcal{L}^\infty(\mathbb{T})$ such that $\widehat{\psi}(m) = \widehat{\phi}(m)$ for all $m < 0$. If the conditions above are satisfied, then:*

$$\|H_\phi\| = \inf\{\|\psi\|_\infty : \widehat{\psi}(m) = \widehat{\phi}(m),\, m < 0\}. \tag{2.23}$$

As a consequence, if H_ϕ is a bounded operator, we can consider without loss of generality $\phi \in \mathcal{L}^\infty(\mathbb{T})$.

An alternative characterization of Hankel operators relies on the use of the shift operator S in the function space, and generalises Equation 2.10. In fact, for an operator $H : \mathcal{H}^2 \to \mathcal{H}^2_-$

we have the following operator identity:

$$HS = \mathbb{P}_- SH. \tag{2.24}$$

This equation, which is also referred to as **Hankel equation**, corresponds to the Hankel property seen in Equation 2.10. In the case of Hardy spaces, the operator of multiplication by z acts as right shift S on the space of Fourier coefficients, in fact $\widehat{(z\phi)}(n) = \widehat{\phi}(n-1)$. Analogously, the left inverse S^* corresponds to the truncated multiplication operator. We remark that the standard proof of Nehari theorem relies on the fact that the shift and its inverse are isometries.

We recall the definition of an important class of complex functions.

Definition 2.4.9. *The complex function $g \in \mathcal{L}^p(\mathbb{D})$, $p > 0$, is **rational** if $g = t/q$, where t and q are polynomials. The rank of g is the maximum between the degrees of t and q. A rational function is **strictly proper** if the degree of t is strictly smaller than that of q.*

Finite rank Hankel operators are closely related to the theory of rational functions.

Theorem 2.4.4 (Kronecker [Kro81])**.** *Let H_ϕ be a bounded Hankel operator with matrix $\mathbf{H}$. Then $\mathbf{H}$ has finite rank if and only if $\mathbb{P}_-\phi$ is a strictly proper rational function. Moreover the rank of $\mathbf{H}$ is equal to the number of poles (counted with multiplicities) of $\mathbb{P}_-\phi$ inside the unit disc.*

We conclude this section by mentioning a few important results the theory of shift-invariant spaces.

Definition 2.4.10. *A function $f \in \mathcal{H}^2$ is called **inner** if it is unimodular almost everywhere on the unit circle, whereas it is called **outer** if $\overline{Span}\{z^n f : n \geq 0\} = \mathcal{H}^2$.*

The following theorem, known as Beurling's Theorem, states that shift-invariant spaces can be characterized by means of inner functions [Beu49].

Theorem 2.4.5 (Beurling's Theorem [Beu49]). *Let E be a closed shift-invariant subspace of H^2, $SE \subset E$. Then, there exists a unique inner function $\Theta \in H^2$ such that:*

$$E = \Theta H^2 = \{\Theta f : f \in H^2\}. \tag{2.25}$$

2.4.4 AAK Theorem

We can now introduce the main result of Adamyan, Arov and Krein [AAK71]. The theorem, stated for Hankel operators over Hardy spaces, shows that for infinite dimensional Hankel matrices the constraint of preserving the Hankel property does not affect the achievable approximation error.

Theorem 2.4.6 (AAK Theorem [AAK71]). *Let H_ϕ be a compact Hankel operator of rank n, matrix $\mathbf{H}$ and singular numbers $\sigma_0 \geq \cdots \geq \sigma_{n-1} > 0$. Then there exists a unique Hankel operator H_g with matrix $\mathbf{G}$ of rank $k < n$ such that:*

$$\|H_\phi - H_g\| = \|\mathbf{H} - \mathbf{G}\| = \sigma_k. \tag{2.26}$$

We say that $\mathbf{G}$ is the optimal approximation of size k of $\mathbf{H}$.

We denote with $\mathcal{R}_k \subset \mathcal{H}_-^\infty$ the set of strictly proper rational functions of rank k, and we consider the set of functions:

$$\mathcal{H}_k^\infty = \{\psi \in \mathcal{L}^\infty(\mathbb{T}) : \ \exists g \in \mathcal{R}_k, \exists l \in \mathcal{H}^\infty, \ \psi = g + l\}. \tag{2.27}$$

The proof of this theorem is directly connected with the problem of approximating a bounded function defined on the unit circle. In fact, we can reformulate AAK theorem in terms of the symbols associated with the Hankel operators.

Theorem 2.4.7 ([AAK71]). *If $\phi \in \mathcal{L}^\infty(\mathbb{T})$ then there exists a complex function $\psi \in \mathcal{H}_k^\infty$*

such that:

$$\|\phi - \psi\|_\infty = \sigma_k(H_\phi). \tag{2.28}$$

This theorem provides us with an alternative interpretation of singular numbers, relating them to the "smoothness" of the corresponding operator (or symbol). The advantage of this second formulation is that its proof is constructive, and tells us how to find the function ψ.

We state as corollary the key point of the proof of AAK Theorem, that provides us with a practical way to find the best approximating symbol.

Corollary 2.4.7.1. *Let ϕ and $\{\boldsymbol{\xi}_k, \boldsymbol{\eta}_k\}$ be a symbol and a σ_k-Schmidt pair for H_ϕ. A function $\psi \in \mathcal{L}^\infty(\mathbb{T})$ is the best AAK approximation according to Theorem 2.4.7, if and only if:*

$$(\phi - \psi)\xi_k^+ = \sigma_k \eta_k^-. \tag{2.29}$$

Moreover, the function ψ does not depend on the particular choice of the pair $\{\boldsymbol{\xi}_k, \boldsymbol{\eta}_k\}$.

Note that the solutions of Theorem 2.4.6 and 2.4.7 are related.

Corollary 2.4.7.2. *Let $\psi \in \mathcal{H}_k^\infty$, with $\psi = l + g$, $g \in \mathcal{R}_k$, $l \in \mathcal{H}^\infty$. If ψ solves Equation 2.28, then $G = H_g$ is the unique Hankel operator from Theorem 2.4.6.*

Proof. Let H_ϕ be a Hankel operator with symbol $\phi(z) \in \mathcal{L}^\infty(\mathbb{T})$ and matrix $\mathbf{H}$.

Let $\psi(z) = g(z) + l(z) \in \mathcal{H}_k^\infty$ be the solution of Equation 2.28. We have:

$$\|H_\phi - H_\psi\| = \|H_{\phi-\psi}\| \tag{2.30}$$

$$= \left\| H_{\sigma_k \eta_k^-(z)/\xi_k^+(z)} \right\| \tag{2.31}$$

$$\leq \sigma_k \left\| \eta_k^-(z)/\xi_k^+(z) \right\|_\infty = \sigma_k \tag{2.32}$$

where first we used Corollary 2.4.7.1 and then Nehari's Theorem. Now, using the definition of Hankel operator, we have:

$$\|H_\phi - H_\psi\| = \|H_\phi - H_g\| = \|\mathbf{H} - \mathbf{G}\| \leq \sigma_k. \tag{2.33}$$

Since $\|\mathbf{H} - \mathbf{G}\| \geq \sigma_k$ (from Eckart-Young theorem [EY36]), it follows that $\|\mathbf{H} - \mathbf{G}\| = \sigma_k$. Note that $\mathbf{G}$ has rank k, as required, because $g \in \mathcal{R}_k$(Theorem 2.4.4). □

2.4.5 Generalized AAK Theory

Within the functional analysis community, there have been various attempt to generalize the results of AAK theory [TV94, Car09, ACP15, Pop03]. As noted in the previous sections, there are several ways to define Hankel operators, depending on whether one is considering the characterization in terms of shifts or of symbol. We are going to consider the characterization in terms of shift operators, and extend the definition of Hankel operator according to the work of Treil and Volberg [TV94]. In their paper, they provide a generalization of Hankel operators and of the Hankel equation (Equation 2.24), where the role of the shift operators is played by two distinct operators, that are not necessarily isometries. This generalized version of AAK theory will be used in Chapter 7.

Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two Hilbert spaces. Let S_1 be an expanding operator in $\mathcal{H}_1$, and S_2 be a contractive operator in $\mathcal{H}_2$. Let $\mathcal{H}_2 = \mathcal{H}_2^+ \oplus \mathcal{H}_2^-$ be an orthogonal decomposition of $\mathcal{H}_2$ such that $S_2\mathcal{H}_2^+ \subset \mathcal{H}_2^+$. Let $\mathbb{P}_+$ and $\mathbb{P}_-$ be the orthogonal projections of $\mathcal{H}_2$ onto $\mathcal{H}_2^+$ and $\mathcal{H}_2^-$ respectively. We recall the definitions from Treil and Volberg [TV94].

Definition 2.4.11. *A **generalized Hankel operator** $\Gamma : \mathcal{H}_1 \to \mathcal{H}_2^-$ is a bounded linear operator satisfying the following relation:*

$$\Gamma S_1 = \mathbb{P}_- S_2 \Gamma. \tag{2.34}$$

The generalization of the concept of symbol in this setting is the multiplier.

Definition 2.4.12. *A **multiplier** is a bounded operator $T : \mathcal{H}_1 \to \mathcal{H}_2$ satisfying*

$$TS_1 = S_2 T. \tag{2.35}$$

Classical Hankel operator H	Generalized Hankel operator Γ
$H : \mathcal{H}^2 \to \mathcal{H}^2_-$	$\Gamma : \mathcal{H}_1 \to \mathcal{H}_2^-$
$HS = \mathbb{P}_- SH$	$\Gamma S_1 = \mathbb{P}_- S_2 \Gamma$
$S : \mathcal{H}^2 \to \mathcal{H}^2$, $S(f) = zf$	$S_1 : \mathcal{H}_1 \to \mathcal{H}_1$,$S_2 : \mathcal{H}_2 \to \mathcal{H}_2$
S isometry	S_1 expansive, S_2 contractive
$\mathcal{H}^2$ Hardy space	$\mathcal{H}_1$
$\mathcal{H}^2_-$ Neg Hardy space	$\mathcal{H}_2^-$
$\mathcal{L}^2(\mathbb{T}) = \mathcal{H}^2 \oplus \mathcal{H}^2_-$	$\mathcal{H}_2 = \mathcal{H}_2^+ \oplus \mathcal{H}_2^-$

Table 2.1: Comparison between classical and generalized Hankel operators.

Given a multiplier T, is it possible to define a Hankel operator Γ_T such that, for $f \in \mathcal{H}_1$:

$$\Gamma_T f = \mathbb{P}_- T f.$$

In Table 2.1 we compare the classical definition of Hankel operator with that of the generalized Hankel operator. As noted at the beginning of this section, the main conceptual difference is that in the second case the shift is replaced by two distinct operators, that do not need to be isometries. By setting $S_1 = S_2 = S$, $\mathcal{H}_1 = \mathcal{H}^2$ and $\mathcal{H}_2 = \mathcal{L}^2(\mathbb{T})$ the generalized definition reduces to the classical one.

Let A be a self-adjoint operator on a separable Hilbert space $\mathcal{H}$, and let $\mathcal{P}_+$ and $\mathcal{P}_-$ be the orthogonal projections onto the non negative and strictly negative part of its spectrum. We denote $\mathcal{H}_\mp = \mathcal{P}_\mp \mathcal{H}$ and $A_- = A|_{\mathcal{H}_-}$. We assume that A_- is invertible. Let

$$\mathcal{K}_+ = \{\mathbf{v} \in \mathcal{H} : (A\mathbf{v}, \mathbf{v}) \geq 0\} \tag{2.36}$$

be the cone of A-nonnegative vectors.

Theorem 2.4.8 ([Iok64])**.** *Let A_- be invertible and let S be a bounded operator in $\mathcal{H}$ such that $S\mathcal{K}_+ \subset \mathcal{K}_+$ and $\mathcal{P}_+ S \mathcal{P}_-$ is a compact operator. Then there exists a subspace $\mathcal{M}$ of $\mathcal{K}_+$ which is maximal (by inclusion) and S-invariant,* i.e. *$S\mathcal{M} \subset \mathcal{M}$.*

We recall that the singular numbers of a Hankel operator $\Gamma : \mathcal{H}_1 \to \mathcal{H}_2^-$ on a Hilbert

space can be characterized in terms of subspaces $\mathcal{M} \subset \mathcal{H}_1$ in the following way:

$$\sigma_n(\Gamma) = \inf\{\|\Gamma|_{\mathcal{M}}\| : \text{codim}\mathcal{M} \leq n\}.$$

Using the fixed point theorem stated above, we obtain the following version of the AAK theorem:

Theorem 2.4.9 (Generalized AAK Theorem [TV94])**.** *Let Γ be a generalized Hankel operator with respect to the operators S_1 (expansive) ans S_2 (contractive). Let $\{\sigma_i\}_{i\geq 0}$ be the sequence of its singular numbers. Then:*

$$\sigma_n(\Gamma) = \inf\{\|\Gamma|_{\mathcal{M}}\| : \mathcal{M},\ \textit{codim}\mathcal{M} \leq n,\ S_1\mathcal{M} \subset \mathcal{M}\}$$

and the infimum is attained.

It is important to remark that this version of the AAK theorem is not constructive. When restricting to the setting of a classical Hankel operator, it is possible to make this proof constructive by leveraging Theorem 2.4.5 and Theorem 2.4.3. In fact, when we are considering the shift operator defined over Hardy spaces, *i.e.* $S_1 = S_2 = S$, the S-invariant space $\mathcal{M}$ considered in Theorem 2.4.9 is completely determined by the zeros of the singular functions associated to σ_n.

Chapter 3

An AAK Theory Approach to Approximate Minimization

In this chapter, we show how the approximate minimization problem can be studied in terms of Hankel matrices. We then illustrate the connections with the problem solved by AAK theory, and the advantages in the use of the spectral norm.

Then, we analyze the case of models over one-letter alphabets. We provide a framework to reformulate the approximate minimization problem using the formalism of AAK theory. We show how we can associate a complex rational function to a given WFA over a one-letter alphabet. Finally, we conclude the chapter by providing the "recipe" to apply AAK theory that will be used in the following two chapters.

3.1 Low-Rank Approximation

The approach we chose to tackle the approximate minimization problem is based on the study of the Hankel matrix of the original, given model. In particular, Theorem 2.2.1 highlight a fundamental connection between the size of an automaton and the rank of the corresponding Hankel matrix. This result motivates our choice to reformulate the approximation problem as a low-rank approximation of the Hankel matrix. The matrix obtained has rank smaller

than the original one, so by Theorem 2.2.1 the WFA associated with it has fewer states and is, therefore, a candidate solution for the approximate minimization problem.

The task of approximating a matrix with one of smaller rank is a well-studied minimization problem. An analytical solution in terms of the singular value decomposition is due to Schmidt, Eckart and Young, and Mirsky [EY36, Mir60, Sch89].

Theorem 3.1.1 ([EY36, Mir60, Sch89])**.** *Let* $\mathbf{H}$ *be a matrix of rank* n*, and let* $\{\sigma_i\}_{i\geq 0}$ *be its singular numbers, with* $\sigma_0 \geq \cdots \geq \sigma_{n-1} > 0$*. Then, if* $\mathbf{R}$ *is a matrix of rank* k*, we have:*

$$\|\mathbf{H} - \mathbf{R}\| \geq \sigma_k. \tag{3.1}$$

The equality is attained when $\mathbf{R}$ *corresponds to the truncated SVD of* $\mathbf{H}$*.*

Note that a low-rank approximation obtained by truncating the singular value decomposition is not in general a Hankel matrix. This can be seen in the following example.

Example 3.1.2. We consider the Hankel matrix $\mathbf{M} \in \mathbb{R}^{3\times 3}$,

$$\mathbf{M} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{pmatrix}.$$

The singular value decomposition of $\mathbf{M}$ is $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, with

$$\mathbf{U} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \sqrt{\frac{2}{3}} & 0 \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} 6 & 0 & 0 \\ 0 & \sqrt{3} & 0 \\ 0 & 0 & \sqrt{3} \end{pmatrix} \quad \mathbf{V} = \begin{pmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \sqrt{\frac{2}{3}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \end{pmatrix}.$$

The rank 2 matrix $\overline{\mathbf{M}}$ obtained by truncating the SVD is not Hankel:

$$\overline{\mathbf{M}} = \begin{pmatrix} 1 & 2 & 3 \\ \frac{5}{2} & 2 & \frac{3}{2} \\ \frac{5}{2} & 2 & \frac{3}{2} \end{pmatrix}.$$

The Hankel property is necessary to extract a WFA from the matrix. Thus, we aim to find a method that preserves it. As shown in Section 2.4, AAK theory provides a (partial) answer to this problem. In particular, Theorem 2.4.6 states that, under proper assumptions, we can obtain a result comparable to the one of Theorem 3.1.1, while searching for the best approximation within the class of Hankel matrices. This means that it is possible to find a Hankel matrix performing as well as the truncated SVD while approximating with respect to the spectral norm.

3.1.1 The Significance of the Spectral Norm

A key point in solving approximation tasks is to decide how to quantify the error between the original and the approximate model. It is therefore natural to wonder if there are norms that are preferable to others. There are a few fundamental properties that we consider to be desirable in our setting.

- We think that being *"architecture independent"* is an important feature for the chosen norm, as it allows us to compare different classes of models using the same norm. For example, one can think of the literature on approximating RNNs using WFAs. In that case, the objective is to extract from a trained RNN an automaton that accurately mimics its behaviour (we refer the reader to Chapter 8 for a literature review). Using the spectral norm, we can measure the distance between a given RNN and the extracted WFA. This is particularly valuable, especially in light of the paper of Marzouk and de la Higuera [MdlH20], where the authors show that the general equivalence

problem between classes of WFAs and RNNs is not decidable. Choosing the spectral norm has the advantage that it allows us to analyze different models through their Hankel matrices, independently of the specific architecture considered. This means that addressing the approximate minimization problem using the spectral norm can facilitate the comparison between different classes of models, and the development of a distance that can be precisely computed (and minimized).

- The chosen norm should be *computationally reasonable to minimize*. Indeed, as shown in our first result [BLP$^+$21], which will be discussed in Chapter 4, the spectral norm of the Hankel matrix of a WFA can be computed in polynomial time. Similarly, we show in our second result [LPR21] (detailed in Chapter 5) that minimizing the approximation error between a WFA and a black box model can be (asymptotically) solved optimally in a tractable way. This is not the case for every metric. For instance, the distance between two WFAs can be computed using behavioural metrics, a powerful technique for understanding approximate behavioural equivalence. A metric based on bisimulation is presented in the work of Balle, Gourdeau and Panangaden [BGP17, BGP22]. While exploring this approach can still be of interest, the fact that this metric is hard to compute makes it unsuitable for our purposes. Moreover, this metric is specifically designed for weighted automata, so it is not directly applicable to other classes of models dealing with sequential data.

- It is important to choose a norm that *can be computed accurately*. The spectral norm is a good candidate for the task. For example, it is possible to precisely compute the singular values of a WFA in its SVA form using its Gramian matrices. Moreover, the choice of the spectral norm allows us to exploit the results from AAK theory illustrated in Chapter 2. This is a great advantage compared, for example, to solving the same rank-constrained optimization problem for the ℓ^2 norm of the original function (which is non convex) [BPP19]. Note that it is still possible to obtain interesting results in the

ℓ^2 norm using balanced truncation methods [BPP19], but the approximation recovered is not optimal.

- Ideally, we would like to find a *global minimum* of the chosen norm. We would also like to solve the approximate minimization problem optimally. Measuring the error using the spectral norm has the advantage that, under appropriate assumptions, it is possible not only to find a global minimum for the approximation error, but also to recover (at least asymptotically) the model corresponding to the optimal approximation.

- Finally, it would be interesting to choose a norm that can be *theoretically compared* with other norms. Theorem 2.4.7 helps us connect the spectral norm to other important norms, like ℓ^2 norm and $\mathcal{L}^\infty$ norm. While this is not of particular use in the context of our current setting, the $\mathcal{L}^\infty$ norm has proven to be very fruitful in related areas like control theory [Ant05] and could be an interesting direction for future work.

It is important to notice that three of the points made above directly derive from the application of AAK theory to the approximation problem. In fact, Theorem 2.4.6 provides us with a way to find, precisely and efficiently, the matrix of the minimal error and the matrix of the best approximation. It is important to remark that, while AAK theory is a theory for Hankel operators using tools from harmonic analysis, the approximate minimization problem is formulated in the context of functions over strings. Therefore, we need a way to transfer the results from one setting to the other. To apply the results of AAK theory to the approximate minimization problem of models computing functions on sequential data, we divide the problem into two cases, depending on the size of the alphabet.

3.2 A Framework for One-Letter Alphabets

In this section, we analyze the case of alphabets Σ with the property that $|\Sigma| = 1$. In this case, the set of strings Σ^* can be identified with the set of natural numbers $\mathbb{N}$ by

associating to each string its length. For example, if $\Sigma = \{a\}$, the empty string corresponds to 0, while the string $'a'$ corresponds to 1, $'aa'$ to 2, $'aaa'$ to 3, and so on. This way, the rational function $f : \Sigma^* \to \mathbb{R}$ can be interpreted as a function $f : \mathbb{N} \to \mathbb{R}$. Moreover, since the natural numbers can be canonically embedded into $\mathbb{Z}$, we can consider a more general function $f : \mathbb{Z} \to \mathbb{R}$. Note that the identification of Σ^* with $\mathbb{N}$ and the embedding into $\mathbb{Z}$ are the fundamental steps allowing us to apply the Fourier isomorphism to reformulate the problem in the Hardy space, where it can be solved using Theorem 2.4.6. Unfortunately, this idea cannot be directly generalized to bigger alphabets, since in this case, the corresponding embedding yields a non-abelian structure. We will discuss this in detail later in the thesis.

3.2.1 From Hankel Matrix to Hankel Operator

We consider a Hankel matrix arising from a model over a one-letter alphabet and analyze how it relates to the Hankel operators used in AAK theory. For simplicity, we assume that the model is a WFA, but the reasoning can easily be generalized to the broader setting of language modelling black boxes (where the Hankel matrix does not necessarily have a finite rank). Let $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ be a weighted automaton with n states over a one-letter alphabet. Let $f_A : \Sigma^* \to \mathbb{R}$ be the rational function computed by the WFA $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$, and let $\mathbf{H}$ be its Hankel matrix. Following the theory presented in Section 2.4, we can associate to a rank n Hankel matrix $\mathbf{H}$ two different operators, which we will denote with H_f and H_ϕ. On the one hand, we can consider the Hankel operator acting over sequences $H_f : \ell^2 \to \ell^2$, associated with the function $f_A : \Sigma^* \to \mathbb{R}$ computed by the WFA. Thus, the Hankel matrix is defined by:

$$\mathbf{H}(i, j) = f_A(i + j) \quad \text{for } i, j \geq 0, \tag{3.2}$$

where the identification of Σ^* and $\mathbb{N}$ allows us to interpret the concatenation of two strings as the sum of their length. Note that in this case we call f "rational" because it is realized

by a WFA of size n. The matrix $\mathbf{H}$ can then be represented as:

$$\mathbf{H} = \mathbf{H}_f = \begin{pmatrix} f_A(0) & f_A(1) & f_A(2) & \dots \\ f_A(1) & f_A(2) & f_A(3) & \dots \\ f_A(2) & f_A(3) & f_A(4) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{3.3}$$

On the other hand, we can interpret the matrix $\mathbf{H}$ as $\mathbf{H}_\phi$, the matrix associated with a Hankel operator over complex (Hardy) spaces $H_\phi : \mathcal{H}^2 \to \mathcal{H}^2_-$. We recall that, in this case, the operator and matrix are related to a complex function $\phi \in \mathcal{L}^2(\mathbb{T})$, the symbol, satisfying the following property: $H_\phi f = \mathbb{P}_- \phi f$. The entries of the matrix are now defined by means of the Fourier coefficients of ϕ as:

$$\mathbf{H}(j,k) = \widehat{\phi}(-j-k-1) \quad \text{for } j,k \geq 0. \tag{3.4}$$

Now, the function $\mathbb{P}_-\phi = \widehat{\phi}(-j-k-1)$ is "rational" in the sense of Definition 2.4.9. The matrix $\mathbf{H}$ can be represented as:

$$\mathbf{H} = \mathbf{H}_\phi = \begin{pmatrix} \widehat{\phi}(-1) & \widehat{\phi}(-2) & \widehat{\phi}(-3) & \dots \\ \widehat{\phi}(-2) & \widehat{\phi}(-3) & \widehat{\phi}(-4) & \dots \\ \widehat{\phi}(-3) & \widehat{\phi}(-4) & \widehat{\phi}(-5) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{3.5}$$

Now, we can derive the relationship between f_A and ϕ simply by looking at the entries of the Hankel matrix. Since we have $\mathbf{H} = \mathbf{H}_f = \mathbf{H}_\phi$, the two representations of the Hankel

matrix need to coincide. We have:

$$\mathbf{H} = \begin{pmatrix} f_A(0) & f_A(1) & f_A(2) & \dots \\ f_A(1) & f_A(2) & f_A(3) & \dots \\ f_A(2) & f_A(3) & f_A(4) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} \widehat{\phi}(-1) & \widehat{\phi}(-2) & \widehat{\phi}(-3) & \dots \\ \widehat{\phi}(-2) & \widehat{\phi}(-3) & \widehat{\phi}(-4) & \dots \\ \widehat{\phi}(-3) & \widehat{\phi}(-4) & \widehat{\phi}(-5) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{3.6}$$

from which we obtain:

$$f(n) = \widehat{\phi}(-n-1). \tag{3.7}$$

3.2.2 Symbols and Rational Functions

The framework we just described provides us with a straightforward way to associate a symbol to a given WFA.

We consider a minimal WFA $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ with n states defined over a one-letter alphabet Σ, and Hankel matrix $\mathbf{H}$. Let $f : \Sigma^* \to \mathbb{R}$ be the function realized by A. Since Σ^* is isomorphic to $\mathbb{N}$, the function computed by A is $f(x) = \boldsymbol{\alpha}^\top \mathbf{A}^x \boldsymbol{\beta}$. To determine the symbol ϕ of H, we use Equation 3.7.

We obtain:

$$\mathbb{P}_- \phi = \sum_{k \geq 0} f(k) z^{-k-1} = \sum_{k \geq 0} \boldsymbol{\alpha}^\top \mathbf{A}^k \boldsymbol{\beta} z^{-k-1} = \boldsymbol{\alpha}^\top (z\mathbf{1} - \mathbf{A})^{-1} \boldsymbol{\beta} \tag{3.8}$$

where the last equality holds if A is irredundant. This is the component of the symbol that belongs to $\mathcal{H}^2_-$. It is easy to see that the complex function obtained is a rational function, with number of poles (counted with multiplicity) equal to the size of the WFA.

Representing a symbol in terms of the parameters of a WFA has the great advantage that allows us to leverage at the same time properties from functional analysis and from the theory of weighted automata. This description will prove to be particularly useful in Chapter 4.

3.2.3 Recipe to Apply AAK Theory

In order to tackle approximate minimization using AAK theory, we need to reformulate the problem in terms of Hankel operators and functions in the complex space. In the one-letter case, this is possible thanks to the framework introduced in the previous sections. In particular, a direct correspondence between a symbol and the function computed by a WFA or black-box model is obtained in Equation 3.7. In the next two chapters we study the approximate minimization problem of WFAs and black-box models over one letter alphabets. To solve it, we apply AAK theory according to the following steps:

1. Given a model on a one-letter alphabet, consider its Hankel matrix $\mathbf{H}$ and the function f that it is computing. By applying the second interpretation of the Hankel matrix, it is possible to derive the negative Fourier coefficients of a symbol ϕ from the entries of the Hankel matrix. In the case of WFAs, that we study in Chapter 4, a symbol can be directly computed using Equation 3.8.

2. Now, we want to apply AAK theory to solve the approximation problem. This can be done in different ways. In Chapter 4, where an expression for ϕ has been obtained, we leverage Theorem 2.4.7 to find $\psi \in \mathcal{L}^\infty(\mathbb{T})$, the best approximation of ϕ in the infinity norm. In Chapter 5, where an expression for ϕ is not readily available, we use a different (equivalent) version of this theorem. In both cases, this step is fundamental to obtain an expression for the function ψ.

3. From ψ, we are interested in extracting the rational component. To achieve this, we study the position of the poles of ψ, and apply Theorem 2.4.4.

4. The last step consists in finding a WFA representation for the optimal approximation. This is implemented very differently in the two chapters. In the case of Chapter 4, where we are minimizing a WFA, using Equation 3.8 allows us to effectively represent complex functions in terms of the parameters of weighted automata. Thus, the rational

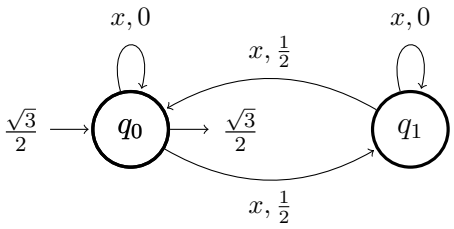


Figure 3.1: Graphical representation of the generative probabilistic automaton described in Example 3.2.1.

function obtained in the previous step directly reveals the parameters of the optimal approximation, which can be found precisely. In the case of Chapter 5, instead, we first obtain the Hankel matrix corresponding to the optimal approximation, and then recover a WFA using the spectral method.

It is important to remark that most of the results we use are obtained in the context of operators, while the setting we are interested in is the one of matrices. To apply AAK theory, we choose to work with the basis of the Hardy spaces. This allows us to alternate between matrix and operator and to directly transfer results from one interpretation to the other. Thus, while we keep the notations distinct to remain faithful to the original definitions (*e.g.*, compactness is a property defined for the operator H, not for the matrix $\mathbf{H}$), to have an intuition of the results it is always possible to think in terms of Hankel matrices.

We consider the following example, from the paper [BLP+21].

Example 3.2.1. Let $|\Sigma| = 1$, $\Sigma = \{x\}$, we consider the WFA $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ represented in Figure 3.1, with:

$$\mathbf{A} = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}, \quad \boldsymbol{\alpha} = \begin{pmatrix} \frac{\sqrt{3}}{2} \\ 0 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \frac{\sqrt{3}}{2} \\ 0 \end{pmatrix},$$

Note that A is a generative probabilistic automaton. Indeed, we have that

- $f_A(x) \geq 0$
- $\sum_{x \in \Sigma^*} f_A(x) = 1$,

since the rational function realized by the WFA is defined as:

$$f_A(x\cdots x) = f_A(k) = \boldsymbol{\alpha}^\top \mathbf{A}^k \boldsymbol{\beta} = \begin{cases} 0 & \text{if } k \text{ is odd} \\ \frac{3}{4}2^{-k} & \text{if } k \text{ is even} \end{cases}$$

where k corresponds to the string where x is repeated k-times. We remark that A is minimal and already in its SVA form, with Gramians

$$\mathbf{P} = \mathbf{Q} = \begin{pmatrix} \frac{4}{5} & 0 \\ 0 & \frac{1}{5} \end{pmatrix}. \tag{3.9}$$

The corresponding Hankel matrix, with entries defined as $\mathbf{H}(i,j) = f(i+j)$, has rank 2:

$$\mathbf{H} = \begin{pmatrix} f_A(0) & f_A(1) & f_A(2) & \dots \\ f_A(1) & f_A(2) & f_A(3) & \dots \\ f_A(2) & f_A(3) & f_A(4) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} \frac{3}{4} & 0 & \frac{3}{16} & \dots \\ 0 & \frac{3}{16} & 0 & \dots \\ \frac{3}{16} & 0 & \frac{3}{64} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{3.10}$$

Now, we can apply the second interpretation of the Hankel matrix and look at it with respect to the symbol, using the definition $\mathbf{H}(j,k) = \widehat{\phi}(-j-k-1)$. We have:

$$\mathbf{H} = \begin{pmatrix} \frac{3}{4} & 0 & \frac{3}{16} & \dots \\ 0 & \frac{3}{16} & 0 & \dots \\ \frac{3}{16} & 0 & \frac{3}{64} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} \widehat{\phi}(-1) & \widehat{\phi}(-2) & \widehat{\phi}(-3) & \dots \\ \widehat{\phi}(-2) & \widehat{\phi}(-3) & \widehat{\phi}(-4) & \dots \\ \widehat{\phi}(-3) & \widehat{\phi}(-4) & \widehat{\phi}(-5) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

We can recover the rational component of a symbol, *i.e.* the projection of ϕ on the negative

Hardy space.

$$\mathbb{P}_{-}\phi = \sum_{n\geq 0} \widehat{\phi}(-n-1)z^{-n-1} = \sum_{n\geq 0} \frac{3}{4} 4^{-n} z^{-2n-1} = \frac{3z}{4z^2-1}.$$

Note that this is a complex rational function having degree 2, and it has two poles inside the unit disc at $z = \pm\frac{1}{2}$ (as predicted by Theorem 2.4.4). It is important to remark that from the Hankel matrix we can only recover the negative Fourier coefficients of ϕ, meaning only the component of the symbol that belongs to the negative Hardy space.

Chapter 4

Weighted Finite Automata on One-Letter Alphabets

In this chapter, we study the approximate minimization problem of a class of weighted automata over a one-letter alphabet. We start by formulating the problem, with particular emphasis on the hypothesis required. In Section 4.2, we analyze the main theoretical results, leveraging the framework introduced in Chapter 3. This allows us to obtain a closed-form solution for the optimal approximation problem. We then design an approximation algorithm, presented in Section 4.3. We end the chapter with a derivation and analysis of the approximation error (Section 4.4), and with a few concluding remarks (Section 4.6). The content of this chapter is presented in the work "Optimal Spectral-Norm Approximate Minimization of Weighted Finite Automata" [BLP+21], that was published and presented at ICALP 2021 (the 48th International Colloquium on Automata, Languages, and Programming).

4.1 Problem Formulation

Let $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ be a minimal WFA with n states and real weights, defined over a one-letter alphabet. We assume that A is in its SVA form, to obtain representation-independent approximation bounds. Let $\mathbf{H}$ be the Hankel matrix of A, we denote with σ_i, for $0 \leq i < n$,

the singular numbers. Given a target number of states $k < n$, we say that a WFA $\widehat{A}_k$ with k states solves the *optimal spectral-norm approximate minimization* problem if the Hankel matrix $\mathbf{G}$ of $\widehat{A}_k$ satisfies:

$$\|\mathbf{H} - \mathbf{G}\| = \sigma_k(\mathbf{H}). \tag{4.1}$$

4.1.1 Assumptions

We list and justify our assumptions.

Size of the Alphabet

We restrict our focus on alphabets of one letter, so we assume $|\Sigma| = 1$. As noted in Section 3.2, this implies that the set of strings Σ^* can be identified with $\mathbb{N}$, and embedded into $\mathbb{Z}$. This step is needed to be able to apply Fourier theory and to reformulate the problem in terms of complex functions and Hardy spaces. This assumption is fundamental and will hold for the rest of the chapter.

Compactness of the Operator

Theorem 2.4.6 requires the Hankel operator H to be compact. To ensure that this condition is satisfied, we consider operators that have finite rank and are bounded. As noted in Theorem 2.2.1, the first condition is automatically satisfied by all WFAs, and the rank of the Hankel matrix is equal to the number of states. Moreover, since the rank is finite, the singular values can be computed exactly using the Gramian matrices, introduced in Definition 2.2.9. To guarantee that the minimal WFA $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ is associated to a bounded operator, we assume that A is irreduntant, *i.e.* that $\rho(\mathbf{A}) < 1$, where ρ is the spectral radius of $\mathbf{A}$. As a matter of fact, to guarantee boundness it is enough that the WFA being considered computes a function $f \in \ell^2$ (a proof of this statement can be found in the paper of Balle, Panangaden and Precup [BPP19]). However, the stricter assumption on the spectral radius is needed when computing the symbol associated to a WFA. This condition

directly implies the existence of the SVA, and of the Gramian matrices $\mathbf{P}$ and $\mathbf{Q}$, where $\mathbf{P} = \mathbf{Q}$ and are diagonal matrices [BPP19]. For example, we note that a minimal GPA computes a function $f \in \ell^1$, so the condition on $\rho(\mathbf{A})$ is automatically satisfied by this class of WFAs [BPP19]. We will briefly investigate the possibility of relaxing this assumption in Section 4.5.

SVA Form

We assume that the WFA $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ is in SVA form. This assumption is not necessary, as the SVA can be efficiently computed from a WFA satisfying the set of assumptions stated above. Starting from a WFA in SVA form is particularly useful, as it allows us to obtain results that are representation independent. Since the alphabet has size one, the Hankel matrix $\mathbf{H}$ is symmetric. Therefore, if we denote with λ_i the i-th non-zero eigenvalue of $\mathbf{H}$, and we consider the components of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we have that $\boldsymbol{\alpha}_i = \operatorname{sgn}(\lambda_i)\boldsymbol{\beta}_i$, where $\operatorname{sgn}(\lambda_i) = \lambda_i/|\lambda_i|$.

4.2 Approximate Minimization

In this section, we present the theoretical aspects of the proposed method for approximate minimization. In particular, we provide a closed-form solution for the parameters of the optimal WFA, and obtain the equations that will be used for the approximation algorithm in the next section.

4.2.1 Outline

Solving the approximate minimization problem corresponds to applying Theorem 2.4.6. Our objective is to find the optimal WFA, so we will focus on representing the inputs and outputs of the problem effectively by means of WFAs. We proceed using the following steps:

1. *Compute a symbol ϕ for H using the dual interpretation of Hankel matrices.* Following what was described in Chapter 3, we use the Hankel matrix to obtain the negative Fourier coefficients of ϕ, and derive its Fourier series.

2. *Compute the optimal symbol ψ using Corollary 2.4.7.1.* The objective at this point is to derive the optimal symbol. In particular, we want to find a suitable representation for the function ψ described in Theorem 2.4.7. The idea is that we define this function in terms of the parameters of an auxiliary WFA $\widehat{A}$. This allows us to leverage the properties of weighted automata, while still keeping the formulation general. Then, we are going to look for the parameters of $\widehat{A}$ that make ψ the solution of Theorem 2.4.7.

3. *Extract the rational component by solving for g in Corollary 2.4.7.2.* This step is arguably the most conceptually challenging, as it requires to identify the position of the function's poles. In fact, we know from Theorem 2.4.4 that g has k poles, all inside the unit disc. Thanks to the representation chosen for the functions, finding the poles corresponds to studying the eigenvalues of the transition matrix of $\widehat{A}$.

4. *Obtain the optimal approximation.* The extracted rational component g directly reveals the parameters of the WFA $\widehat{A}_k$.

4.2.2 Finding a Symbol for the WFA

We follow the setting presented in Section 4.1 and consider a minimal WFA $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ with n states in SVA form, defined over a one-letter alphabet $\Sigma = \{a\}$, its Hankel matrix $\mathbf{H}$, corresponding to the bounded operator H, and the singular numbers σ_i, for $0 \leq i < n$. Let $f : \Sigma^* \to \mathbb{R}$ be the function realized by A. We denote by x the string where a is repeated x times, so we have $f(x) = \boldsymbol{\alpha}^\top \mathbf{A}^x \boldsymbol{\beta}$.

In the Section 3.2.2 we obtained in Equation 3.8 a representation for the negative Fourier expansion of a symbol. Note that the series converges because A is irredundant. Since we are dealing with a bounded Hankel operator, from Nehari's Theorem we know that it is possible

to consider a bounded symbol without losing any generality. Therefore, we can set

$$\phi = \boldsymbol{\alpha}^\top (z\mathbf{1} - \mathbf{A})^{-1}\boldsymbol{\beta} \tag{4.2}$$

as a symbol for H, associated to the WFA A.

4.2.3 Finding the Optimal Symbol

The second step to solve the approximate minimization problem is to find a proper expression for the complex functions ψ and $e = \phi - \psi$ described in Theorem‘2.4.7. The key idea is to define the functions using the parameters of two auxiliary WFAs. This allows us to leverage at the same time properties of WFAs and of complex functions. We define the WFAs as follows. Let k be the target number of states of the best approximation, we consider a WFA $\widehat{A} = \langle \widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}} \rangle$ with $j \geq k$ states, satisfying the following two properties:

1. 1 is not an eigenvalue of $\widehat{\mathbf{A}}$
2. the automaton $E = \langle \boldsymbol{\alpha}_e, \mathbf{A}_e, \boldsymbol{\beta}_e \rangle$ is minimal, where

$$\mathbf{A}_e = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{A}} \end{pmatrix}, \quad \boldsymbol{\alpha}_e = \begin{pmatrix} \boldsymbol{\alpha} \\ -\widehat{\boldsymbol{\alpha}} \end{pmatrix}, \quad \boldsymbol{\beta}_e = \begin{pmatrix} \boldsymbol{\beta} \\ \widehat{\boldsymbol{\beta}} \end{pmatrix}. \tag{4.3}$$

Using the parameters of the automaton $\widehat{A}$, and a constant C, we define a function

$$\psi = \widehat{\boldsymbol{\alpha}}^\top (z\mathbf{1} - \widehat{\mathbf{A}})^{-1}\widehat{\boldsymbol{\beta}} + C. \tag{4.4}$$

Analogously, we use the parameters of E to define the function

$$e = \phi - \psi = \boldsymbol{\alpha}_e^\top (z\mathbf{1} - \mathbf{A}_e)^{-1}\boldsymbol{\beta}_e - C. \tag{4.5}$$

Now that we have a general expression for the functions, we want to find the parameters of

$\widehat{A}$ that make ψ the solution of Theorem 2.4.7. The first important condition which needs to be satisfied is the boundness around the unit circle: by definition, ψ is the sum of two component, one that is bounded and one that has poles only inside the unit disc. Therefore, there cannot be poles on the unit circle. Following our definition, the poles of ψ correspond to the eigenvalues of $\widehat{\mathbf{A}}$, counted with their multiplicities. By assumption, 1 is not an eigenvalue of $\widehat{A}$, so ψ does not have any poles on the unit circle, and therefore $\psi \in \mathcal{L}^\infty(\mathbb{T})$. The same holds for the function $e = \boldsymbol{\alpha}_e^\top (z\mathbf{1} - \mathbf{A}_e)^{-1}\boldsymbol{\beta}_e - C$. Since the first condition is satisfied, we can use Corollary 2.4.7.1 to find the triple $\widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}}$ such that ψ satisfies Equation 2.29. Note that, with this purpose, the constant term $C \in H^\infty$ becomes necessary to characterize ψ. In fact, while the H^∞-component of the symbol does not affect the Hankel norm, it plays a role in the computation of the $\mathcal{L}^\infty$-norm (in Equation 2.28) according to Nehari's theorem (Theorem 2.4.1), so it cannot be dismissed. The assumption on the minimality of the automaton E will be used later in this section.

Finding the functions in the Hardy space corresponding to a σ_k-Schmidt pair is relativity straightforward.

Theorem 4.2.1 ([BLP+21])**.** *Let σ_k be a singular number of the Hankel operator H. The singular functions associated with the σ_k-Schmidt pair $\{\boldsymbol{\xi}_k, \boldsymbol{\eta}_k\}$ of H are:*

$$\xi_k^+(z) = \sigma_k^{-1/2}\boldsymbol{\beta}^\top(\mathbf{1} - z\mathbf{A})^{-1}\mathbf{e}_k \tag{4.6}$$

$$\eta_k^-(z) = \sigma_k^{-1/2}\boldsymbol{\alpha}^\top(z\mathbf{1} - \mathbf{A}^\top)^{-1}\mathbf{e}_k. \tag{4.7}$$

*If ψ is the best approximation to the symbol, then $\sigma_k^{-1}e$ has modulus 1 almost everywhere on the unit circle (*i.e. *it is unimodular).*

Proof. Let $\mathbf{F}$ and $\mathbf{B}$ be the forward and backward matrices, respectively, with $\mathbf{H} = \mathbf{F}\mathbf{B}^\top$, $\mathbf{P} = \mathbf{F}^\top\mathbf{F}, \mathbf{Q} = \mathbf{B}^\top\mathbf{B}$. We consider the σ_k-Schmidt pair $\{\boldsymbol{\xi}_k, \boldsymbol{\eta}_k\}$. By definition, $\mathbf{H}^\top\mathbf{H}\boldsymbol{\xi}_k =$

$\sigma_k^2 \boldsymbol{\xi}_k$. By rewriting in terms of the FB factorization, we obtain:

$$\mathbf{H}^\top \mathbf{H} \boldsymbol{\xi}_k = \sigma_k^2 \boldsymbol{\xi}_k \tag{4.8}$$

$$\mathbf{B}\mathbf{F}^\top \mathbf{F}\mathbf{B}^\top \boldsymbol{\xi}_k = \sigma_k^2 \boldsymbol{\xi}_k \tag{4.9}$$

$$\mathbf{B}\mathbf{P}\mathbf{B}^\top \boldsymbol{\xi}_k = \sigma_k^2 \boldsymbol{\xi}_k \tag{4.10}$$

$$\mathbf{B}\mathbf{P}\mathbf{e}_k = \sigma_k^2 \boldsymbol{\xi}_k \tag{4.11}$$

where in the last step we set $\mathbf{e}_k = \mathbf{B}^\top \boldsymbol{\xi}_k$, to reduce the SVD problem of $\mathbf{H}$ to the one of $\mathbf{QP}$. Note that, since $\mathbf{P}$ and $\mathbf{Q}$ are diagonal, $\mathbf{e}_k$ is the k-th coordinate vector $(0, \ldots, 0, 1, 0, \ldots, 0)^\top$. Since $\mathbf{e}_k$ is an eigenvector of $\mathbf{QP}$ for σ_k^2, we get:

$$\mathbf{B}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{P}\mathbf{e}_k = \sigma_k^2 \boldsymbol{\xi}_k \tag{4.12}$$

$$\mathbf{B}\mathbf{Q}^{-1}\mathbf{e}_k = \boldsymbol{\xi}_k. \tag{4.13}$$

Moreover, $\mathbf{H}$ is symmetric, so we have that the singular vectors $\boldsymbol{\eta}_k$ and $\boldsymbol{\xi}_k$ have the same coordinates up to the sign of the corresponding eigenvalues. We obtain:

$$\xi_k^+(z) = \sum_{j=0}^{\infty} \sigma_k^{-1/2} \boldsymbol{\beta}^\top \mathbf{A}^j \mathbf{e}_k z^j = \sigma_k^{-1/2} \boldsymbol{\beta}^\top (\mathbf{1} - z\mathbf{A})^{-1} \boldsymbol{e}_k \tag{4.14}$$

$$\eta_k^-(z) = \sum_{j=0}^{\infty} \sigma_k^{-1/2} \boldsymbol{\alpha}^\top \mathbf{A}^{j\top} \mathbf{e}_k z^{-j-1} = \sigma_k^{-1/2} \boldsymbol{\alpha}^\top (z\mathbf{1} - \mathbf{A}^\top)^{-1} \boldsymbol{e}_k \tag{4.15}$$

where the singular functions have been computed following Equation 2.14. If r is the multiplicity of σ_k, from Corollary 2.4.7.1 we get the following fundamental equation:

$$(\phi - \psi) \boldsymbol{\beta}^\top (\mathbf{1} - z\mathbf{A})^{-1} \mathbf{V} = \sigma_k \boldsymbol{\alpha}^\top (z\mathbf{1} - \mathbf{A}^\top)^{-1} \mathbf{V}$$

where $\mathbf{V} = \begin{pmatrix} \mathbf{0} & \mathbf{1}_r \end{pmatrix}^\top$ is a $n \times r$ matrix. Consequently, we obtain the function:

$$\sigma_k^{-1} e = \frac{\boldsymbol{\alpha}^\top (z\mathbf{1} - \mathbf{A}^\top)^{-1} \mathbf{V}}{\boldsymbol{\beta}^\top (\mathbf{1} - z\mathbf{A})^{-1} \mathbf{V}}$$

which is unimodular, since $\boldsymbol{\alpha}_i = \operatorname{sgn}(\lambda_i)\boldsymbol{\beta}_i$, and $\mathbf{A} = \operatorname{sgn}(\lambda_i)\mathbf{A}^\top$. □

We tied functions to WFAs by means of their parameters, so now we can investigate how the fact that $\sigma_k^{-1} e$ is unimodular reflects on the structure of the WFA $E = \langle \boldsymbol{\alpha}_e, \mathbf{A}_e, \boldsymbol{\beta}_e \rangle$ associated with it. We remark that a parallel can be drawn between dynamical systems and automata, by noting that the impulse-response of a discrete time-invariant Single-Input-Single-Output SISO system can be parametrized as a WFA over a one-letter alphabet (more details about this correspondence can be found in Section 8.3). This allows us to apply a theorem from the control theory literature, and to find two matrices, $\mathbf{P}_e$ and $\mathbf{Q}_e$, satisfying properties similar to those of the Gramians [CC97]. It is important to notice that, *a priori*, the controllability and observability Gramians of E might not be well defined. The proof of the following theorem relies on the minimality of the WFA E [Sch00], and builds on the maximum modulus principle, according to which the maximum modulus of an holomorphic function is attained on the boundary of the domain. We refer the reader to Appendix B.1 for a sketch of the proof, while the detailed, original version, can be found in the book of Chui and Chen [CC97, Theorem 6.3].

Theorem 4.2.2 ([CC97])**.** *Consider the function* $e = \boldsymbol{\alpha}_e^\top (z\mathbf{1} - \mathbf{A}_e)^{-1} \boldsymbol{\beta}_e - C$ *and the corresponding minimal WFA* $E = \langle \boldsymbol{\alpha}_e, \mathbf{A}_e, \boldsymbol{\beta}_e \rangle$ *associated with it. If* $\sigma_k^{-1} e$ *is unimodular, then there exists a unique pair of symmetric invertible matrices* $\mathbf{P}_e$ *and* $\mathbf{Q}_e$ *satisfying:*

(a) $\mathbf{P}_e - \mathbf{A}_e \mathbf{P}_e \mathbf{A}_e^\top = \boldsymbol{\beta}_e \boldsymbol{\beta}_e^\top$

(b) $\mathbf{Q}_e - \mathbf{A}_e^\top \mathbf{Q}_e \mathbf{A}_e = \boldsymbol{\alpha}_e \boldsymbol{\alpha}_e^\top$

(c) $\mathbf{P}_e \mathbf{Q}_e = \sigma_k^2 \mathbf{1}$

We can now derive the parameters of the WFA $\widehat{A} = \langle \widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}} \rangle$ that make ψ the solution of Theorem 2.4.7.

Theorem 4.2.3 ([BLP+21])**.** *Let $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ be a minimal WFA with n states in its SVA form, and let $\phi = \boldsymbol{\alpha}^\top (z\mathbf{1} - \mathbf{A})^{-1}\boldsymbol{\beta}$ be a symbol for its Hankel operator H. Let σ_k be a singular number of multiplicity r for H, with:*

$$\sigma_0 \geq \cdots > \sigma_k = \cdots = \sigma_{k+r-1} > \sigma_{k+r} \geq \cdots \geq \sigma_{n-1} > 0. \tag{4.16}$$

We can partition the Gramian matrices $\mathbf{P}$, $\mathbf{Q}$ as follows:

$$\mathbf{P} = \mathbf{Q} = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \sigma_k \mathbf{1}_r \end{pmatrix}, \tag{4.17}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{(n-r)\times(n-r)}$ is the diagonal matrix containing the remaining singular numbers, and partition $\mathbf{A}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to conform with the Gramians:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \quad \boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}. \tag{4.18}$$

Let $\mathbf{R} = \sigma_k^2 \mathbf{1}_{n-r} - \boldsymbol{\Sigma}^2$, we denote by $(\cdot)^+$ the Moore-Penrose pseudo-inverse. If the function $\psi = \widehat{\boldsymbol{\alpha}}^\top (z\mathbf{1} - \widehat{\mathbf{A}})^{-1}\widehat{\boldsymbol{\beta}} + C$ is the best approximation of ϕ, then:

- *If $\boldsymbol{\alpha}_2 \neq \mathbf{0}$:*

$$\begin{cases} \widehat{\boldsymbol{\beta}} = -\widehat{\mathbf{A}}\mathbf{A}_{21}^\top (\boldsymbol{\beta}_2^\top)^+ \\ \widehat{\boldsymbol{\alpha}} = \widehat{\mathbf{A}}^\top \mathbf{R}\mathbf{A}_{12} (\boldsymbol{\alpha}_2^\top)^+ \\ \widehat{\mathbf{A}}(\mathbf{A}_{11}^\top - \mathbf{A}_{21}^\top (\boldsymbol{\beta}_2^\top)^+ \boldsymbol{\beta}_1^\top) = \mathbf{1} \end{cases} \tag{4.19}$$

- *If* $\boldsymbol{\alpha}_2 = \mathbf{0}$*:*

$$\begin{cases} \widehat{\boldsymbol{\beta}} = (\mathbf{1} - \widehat{\mathbf{A}}\mathbf{A}_{11}^\top)(\boldsymbol{\beta}_1^\top)^+ \\ \widehat{\boldsymbol{\alpha}} = -(\mathbf{R} - \widehat{\mathbf{A}}^\top \mathbf{R}\mathbf{A}_{11})(\boldsymbol{\alpha}_1^\top)^+ \\ \widehat{\mathbf{A}}\mathbf{A}_{21}^\top = \mathbf{0} \end{cases} \tag{4.20}$$

Proof. Since $\sigma^{-1}e = \phi - \psi$ is unimodular, from Theorem 4.2.2 there exist two symmetric nonsingular matrices $\mathbf{P}_e$, $\mathbf{Q}_e$ satisfying the fixed point equations:

$$\mathbf{P}_e - \mathbf{A}_e \mathbf{P}_e \mathbf{A}_e^\top = \boldsymbol{\beta}_e \boldsymbol{\beta}_e^\top \tag{4.21}$$

$$\mathbf{Q}_e - \mathbf{A}_e^\top \mathbf{Q}_e \mathbf{A}_e = \boldsymbol{\alpha}_e \boldsymbol{\alpha}_e^\top \tag{4.22}$$

and such that $\mathbf{P}_e\mathbf{Q}_e = \sigma_k^2\mathbf{1}$. We can partition $\mathbf{P}_e$ and $\mathbf{Q}_e$ according to the definition of $\mathbf{A}_e$ (see Equation 4.3):

$$\mathbf{P}_e = \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{12}^\top & \mathbf{P}_{22} \end{pmatrix}, \quad \mathbf{Q}_e = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{12}^\top & \mathbf{Q}_{22} \end{pmatrix}.$$

From Equation 4.21 and 4.22, we note that $\mathbf{P}_{11}$ and $\mathbf{Q}_{11}$ correspond to the controllability and observability Gramians of A:

$$\mathbf{P}_{11} = \mathbf{Q}_{11} = \mathbf{P} = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \sigma_k\mathbf{1} \end{pmatrix}.$$

Moreover, since $\mathbf{P}_e\mathbf{Q}_e = \sigma_k^2\mathbf{1}$, we get $\mathbf{P}_{12}\mathbf{Q}_{12}^\top = \sigma_k^2\mathbf{1} - \mathbf{P}^2$. It follows that $\mathbf{P}_{12}\mathbf{Q}_{12}^\top$ has rank $n - r$. Without loss of generality we can set $\dim \widehat{\mathbf{A}} = j = n - r$, and choose an appropriate basis for the state space such that $\mathbf{P}_{12} = \begin{pmatrix} \mathbf{1} & \mathbf{0} \end{pmatrix}^\top$ and $\mathbf{Q}_{12} = \begin{pmatrix} \mathbf{R} & \mathbf{0} \end{pmatrix}^\top$, with $\mathbf{R} = \sigma_k^2\mathbf{1} - \boldsymbol{\Sigma}^2$. Once $\mathbf{P}_{12}$ and $\mathbf{Q}_{12}$ are fixed, the values of $\mathbf{P}_{22}$ and $\mathbf{Q}_{22}$ are automatically determined. We

obtain:

$$\mathbf{P}_e = \begin{pmatrix} \mathbf{\Sigma} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \sigma_k \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & -\mathbf{\Sigma}\mathbf{R}^{-1} \end{pmatrix}, \quad \mathbf{Q}_e = \begin{pmatrix} \mathbf{\Sigma} & \mathbf{0} & \mathbf{R} \\ \mathbf{0} & \sigma_k \mathbf{1} & \mathbf{0} \\ \mathbf{R} & \mathbf{0} & -\mathbf{\Sigma}\mathbf{R} \end{pmatrix}. \tag{4.23}$$

Now that we have an expression for the matrices $\mathbf{P}_e$ and $\mathbf{Q}_e$ of Theorem 4.2.2, we can rewrite the fixed point equations to derive the parameters $\widehat{\boldsymbol{\alpha}}$, $\widehat{\mathbf{A}}$ and $\widehat{\boldsymbol{\beta}}$. We obtain the following systems:

$$\begin{cases} \mathbf{P} - \mathbf{A}\mathbf{P}\mathbf{A}^\top = \boldsymbol{\beta}\boldsymbol{\beta}^\top \\ \mathbf{N} - \mathbf{A}\mathbf{N}\widehat{\mathbf{A}}^\top = \boldsymbol{\beta}\widehat{\boldsymbol{\beta}}^\top \\ -\mathbf{\Sigma}\mathbf{R}^{-1} + \widehat{\mathbf{A}}\mathbf{\Sigma}\mathbf{R}^{-1}\widehat{\mathbf{A}}^\top = \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^\top \end{cases} \qquad \begin{cases} \mathbf{P} - \mathbf{A}^\top\mathbf{P}\mathbf{A} = \boldsymbol{\alpha}\boldsymbol{\alpha}^\top \\ \mathbf{M} - \mathbf{A}^\top\mathbf{M}\widehat{\mathbf{A}} = -\boldsymbol{\alpha}\widehat{\boldsymbol{\alpha}}^\top \\ -\mathbf{\Sigma}\mathbf{R} + \widehat{\mathbf{A}}^\top\mathbf{\Sigma}\mathbf{R}\widehat{\mathbf{A}} = \widehat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\alpha}}^\top \end{cases} \tag{4.24}$$

where $\mathbf{N} = \begin{pmatrix} \mathbf{1} \\ \mathbf{0} \end{pmatrix}$ and $\mathbf{M} = \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix}$.

We can rewrite the second equation of each system as follows:

$$\begin{cases} \mathbf{1} - \mathbf{A}_{11}\widehat{\mathbf{A}}^\top = \boldsymbol{\beta}_1\widehat{\boldsymbol{\beta}}^\top \\ -\mathbf{A}_{21}\widehat{\mathbf{A}}^\top = \boldsymbol{\beta}_2\widehat{\boldsymbol{\beta}}^\top \end{cases} \qquad \begin{cases} \mathbf{R} - \mathbf{A}_{11}^\top\mathbf{R}\widehat{\mathbf{A}} = -\boldsymbol{\alpha}_1\widehat{\boldsymbol{\alpha}}^\top \\ \widehat{\mathbf{A}}^\top\mathbf{R}\mathbf{A}_{12} = \widehat{\boldsymbol{\alpha}}\boldsymbol{\alpha}_2^\top \end{cases} \tag{4.25}$$

If $\boldsymbol{\alpha}_2 \neq \mathbf{0}$, then also $\boldsymbol{\beta}_2 \neq \mathbf{0}$ (recall that $\boldsymbol{\alpha}_i = \mathrm{sgn}(\lambda_i)\boldsymbol{\beta}_i$), and we have:

$$\begin{cases} \widehat{\boldsymbol{\beta}} = -\widehat{\mathbf{A}}\mathbf{A}_{21}^\top(\boldsymbol{\beta}_2^\top)^+ \\ \widehat{\boldsymbol{\alpha}} = \widehat{\mathbf{A}}^\top\mathbf{R}\mathbf{A}_{12}(\boldsymbol{\alpha}_2^\top)^+ \\ \widehat{\mathbf{A}}(\mathbf{A}_{11}^\top - \mathbf{A}_{21}^\top(\boldsymbol{\beta}_2^\top)^+\boldsymbol{\beta}_1^\top) = \mathbf{1} \end{cases} \tag{4.26}$$

with $(\boldsymbol{\alpha}_2^\top)^+ = \frac{\boldsymbol{\alpha}_2}{\boldsymbol{\alpha}_2^\top\boldsymbol{\alpha}_2}$ and $(\boldsymbol{\beta}_2^\top)^+ = \frac{\boldsymbol{\beta}_2}{\boldsymbol{\beta}_2^\top\boldsymbol{\beta}_2}$.

If $\boldsymbol{\alpha}_2 = \mathbf{0}$, we have $\widehat{\mathbf{A}}\mathbf{A}_{21}^\top = \mathbf{0}$. We remark that $\widehat{\mathbf{A}}$ has size $(n-r)\times(n-r)$, while $\mathbf{A}_{21}^\top$ is $(n-r)\times r$, so the system of equations corresponding to $\widehat{\mathbf{A}}\mathbf{A}_{21}^\top = \mathbf{0}$ is underdetermined if

$r < \frac{n}{2}$, in which case we can find an alternative set of solutions:

$$\begin{cases} \widehat{\boldsymbol{\beta}} = (\mathbf{1} - \widehat{\mathbf{A}}\mathbf{A}_{11}^{\top})(\boldsymbol{\beta}_1^{\top})^{+} \\ \widehat{\boldsymbol{\alpha}} = -(\mathbf{R} - \widehat{\mathbf{A}}^{\top}\mathbf{R}\mathbf{A}_{11})(\boldsymbol{\alpha}_1^{\top})^{+} \\ \widehat{\mathbf{A}}\mathbf{A}_{21}^{\top} = \mathbf{0} \end{cases} \tag{4.27}$$

with $\widehat{\mathbf{A}} \neq \mathbf{0}$. On the other hand, if $r \geq \frac{n}{2}$, *i.e.* if the multiplicity of the singular number σ_k is more than half the size of the original WFA, the system might not have any solution unless $\widehat{\mathbf{A}} = \mathbf{0}$ (or unless $\mathbf{A}_{21}$ was zero to begin with). In this setting the method proposed returns $\widehat{\mathbf{A}} = \mathbf{0}$. □

We remark that in the (rare) case in which the algorithm returns $\widehat{\mathbf{A}} = \mathbf{0}$, an alternative and preferable approach is to search for an approximation of size $k-1$ or $k+r$. This way, the multiplicity $\bar{r}$ of the new singular number is such that $\bar{r} < \frac{n}{2}$, and the system in Equation 4.27 is underdetermined.

Theorem 4.2.3 provides us with a way to compute the coefficients of the function ψ solving Theorem 2.4.7. It is important to notice that the WFA $\widehat{A}$ is not the best approximation. Intuitively, the problem is that it is potentially too big and not necessarily irredundant. From the AAK theorem we know that the optimal approximation corresponds to a bounded Hankel operator, so the resulting WFA has to be irredundant. We need to "extract" from $\widehat{A}$ a smaller WFA of size k. We do this by extracting the component of the function ψ that belongs to the negative Hardy space.

4.2.4 Extracting the Rational Component

The objective of this section is to "isolate" the function $g \in \mathcal{R}_k$, *i.e.* the *rational component* of ψ. To do this, we study the position of the poles of ψ. In fact, we know from Theorem 2.4.4 that the poles of a strictly proper rational function lie inside the unit disc. As noted before, the key to solving our problem is the way we parametrized the functions. We defined ψ so

that its poles correspond to the eigenvalues of $\widehat{A}$. Therefore, we study the eigenvalues of $\widehat{\mathbf{A}}$ using the following auxiliary result from Ostrowski [OS62]. A proof of this theorem can be found in [Wim73].

Theorem 4.2.4 ([OS62])**.** *Let $|\Sigma| = 1$, and let $\mathbf{P}$ be a solution to the fixed point equation $X - \mathbf{A}X\mathbf{A}^\top = \boldsymbol{\beta}\boldsymbol{\beta}^\top$ for the WFA $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$. If A is reachable, then:*

- *The number of eigenvalues λ of $\mathbf{A}$ such that $|\lambda| < 1$ is equal to the number of positive eigenvalues of $\mathbf{P}$.*
- *The number of eigenvalues λ of $\mathbf{A}$ such that $|\lambda| > 1$ is equal to the number of negative eigenvalues of $\mathbf{P}$.*

After a change of basis (that we detail in Section 4.3 with the approximation algorithm), we can rewrite $\widehat{\mathbf{A}}$ in block-diagonal form:

$$\widehat{\mathbf{A}} = \begin{pmatrix} \widehat{\mathbf{A}}_+ & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{A}}_- \end{pmatrix} \tag{4.28}$$

where the modulus of the eigenvalues of $\widehat{\mathbf{A}}_+$ (resp. $\widehat{\mathbf{A}}_-$) is smaller (resp. greater) than one. We then apply the same change of coordinates on $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$.

We can finally find the rational component of the function ψ, *i.e.* the function g from Corollary 2.4.7.2 necessary to solve that approximate minimization problem.

Theorem 4.2.5 ([BLP+21])**.** *Let $\widehat{\mathbf{A}}_+, \widehat{\boldsymbol{\alpha}}_+, \widehat{\boldsymbol{\beta}}_+$ be as in Equation 4.28. The rational component of ψ is the function $g = \widehat{\boldsymbol{\alpha}}_+^\top (z\mathbf{1} - \widehat{\mathbf{A}}_+)^{-1} \widehat{\boldsymbol{\beta}}_+$.*

Proof. Clearly $\psi = g + l$, with $l = \widehat{\boldsymbol{\alpha}}_-^\top (z\mathbf{1} - \widehat{\mathbf{A}}_-)^{-1} \widehat{\boldsymbol{\beta}}_-$, $l \in \mathcal{H}^\infty$. To conclude the proof we need to show that g has k poles inside the unit disc, and that therefore it has rank k. We do this by studying the modulus of the eigenvalues of $\widehat{\mathbf{A}}_+$.

Since E is minimal, $\widehat{A}$ is reachable by definition, so we can use Theorem 4.2.4 and solve the problem by directly examining the eigenvalues of $-\boldsymbol{\Sigma}\mathbf{R}$. From the proof of Theorem 4.2.3

we have $-\mathbf{\Sigma R} = \mathbf{\Sigma}(\mathbf{\Sigma}^2 - \sigma_k^2\mathbf{1})$, where $\mathbf{\Sigma}$ is the diagonal matrix having as elements the singular numbers of H different from σ_k. It follows that $-\mathbf{\Sigma R}$ has only k strictly positive eigenvalues, and $\widehat{\mathbf{A}}$ has k eigenvalues with modulus smaller than 1. Thus, $\widehat{\mathbf{A}}_+$ has k eigenvalues, corresponding to the poles of g. □

4.2.5 Solving the Approximation Problem

Now that we have found the rational function g, a symbol for the operator that solves Theorem 2.4.6, we need to find the parameters of $\widehat{A}_k$, the WFA corresponding to the optimal approximation. These are directly revealed by the expression of g, due to the way we parametrized the function.

Theorem 4.2.6. *Let $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ be a minimal WFA with n states over a one-letter alphabet. Let A be in its SVA form. The optimal spectral-norm approximation of rank k is given by the WFA $\widehat{A}_k = \langle \widehat{\boldsymbol{\alpha}}_+, \widehat{\mathbf{A}}_+, \widehat{\boldsymbol{\beta}}_+ \rangle$.*

Proof. From Corollary 2.4.7.2 we know that g is the rational function associated with the Hankel matrix of the best approximation. Given the correspondence between the Fourier coefficients of g and the entries of the matrix, we have:

$$g = \widehat{\boldsymbol{\alpha}}_+^\top (z\mathbf{1} - \widehat{\mathbf{A}}_+)^{-1} \widehat{\boldsymbol{\beta}}_+ = \sum_{k \geq 0} \widehat{\boldsymbol{\alpha}}_+^\top \widehat{\mathbf{A}}_+^k \widehat{\boldsymbol{\beta}}_+ z^{-k-1} = \sum_{k \geq 0} \bar{f}(k) z^{-k-1} \tag{4.29}$$

where $\bar{f} : \Sigma^* \to \mathbb{R}$ is the function computed by $\widehat{A}_k$ and $\widehat{\boldsymbol{\alpha}}_+, \widehat{\mathbf{A}}_+, \widehat{\boldsymbol{\beta}}_+$ are the parameters. □

4.2.6 Example

We consider the WFA in SVA form introduced in Example 2.2.5:

$$\mathbf{A} = \begin{pmatrix} 0.579 & 0.461 & 0.046 \\ -0.461 & -0.192 & 0.225 \\ 0.046 & -0.225 & -0.387 \end{pmatrix}, \quad \boldsymbol{\alpha} = \begin{pmatrix} 1.650 \\ -0.851 \\ 0.038 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} 1.650 \\ 0.851 \\ 0.038 \end{pmatrix},$$

Note that $\mathbf{A}$ has spectral radius strictly smaller than 1, having eigenvales:

$$\lambda_{1,2} = 0.0162324 \pm 0.0297233i \qquad \lambda_3 = 0.0324648. \tag{4.30}$$

We want to find the optimal approximation of rank 2 in the spectral norm. According to the partition in Equation 4.17, we have

$$\mathbf{P} = \mathbf{Q} = \begin{pmatrix} 4.67 & 0 & 0 \\ 0 & 1.79 & 0 \\ 0 & 0 & 0.12 \end{pmatrix},$$

so that $\sigma_2^2 = 0.12$ and:

$$\mathbf{\Sigma} = \begin{pmatrix} 4.67 & 0 \\ 0 & 1.79 \end{pmatrix}.$$

$$\mathbf{A}_{11} = \begin{pmatrix} 0.579 & 0.461 \\ -0.461 & -0.192 \end{pmatrix}, \quad \mathbf{A}_{1,2} = \begin{pmatrix} 0.046 \\ 0.225 \end{pmatrix},$$

$$\mathbf{A}_{2,1}^\top = \begin{pmatrix} 0.046 \\ -0.225 \end{pmatrix}, \quad \mathbf{A}_{22} = -0.387$$

$$\boldsymbol{\alpha}_1 = \begin{pmatrix} 1.650 \\ -0.851 \end{pmatrix}, \quad \boldsymbol{\beta}_1 = \begin{pmatrix} 1.650 \\ 0.851 \end{pmatrix}, \quad \boldsymbol{\alpha}_2 = \boldsymbol{\beta}_2 = 0.038$$

Since $\boldsymbol{\alpha}_2 \neq 0$, we can use Equation 4.19 to find the coefficients of the WFA $\widehat{A} = \langle \widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}} \rangle$.

We have:

$$\begin{cases} \widehat{\boldsymbol{\beta}} = -\widehat{\mathbf{A}}\mathbf{A}_{21}^{\top}(\boldsymbol{\beta}_2^{\top})^{+} \\ \widehat{\boldsymbol{\alpha}} = \widehat{\mathbf{A}}^{\top}\mathbf{R}\mathbf{A}_{12}(\boldsymbol{\alpha}_2^{\top})^{+} \\ \widehat{\mathbf{A}}(\mathbf{A}_{11}^{\top} - \mathbf{A}_{21}^{\top}(\boldsymbol{\beta}_2^{\top})^{+}\boldsymbol{\beta}_1^{\top}) = \mathbf{1} \end{cases}$$

$$\begin{cases} \widehat{\boldsymbol{\beta}} = -\widehat{\mathbf{A}} \begin{pmatrix} 0.046 \\ \\ -0.225 \end{pmatrix} (0.038)^{-1} \\ \widehat{\boldsymbol{\alpha}} = \widehat{\mathbf{A}}^{\top} \left(\begin{pmatrix} 0.12 & 0 \\ \\ 0 & 0.12 \end{pmatrix} - \begin{pmatrix} 4.67 & 0 \\ \\ 0 & 1.79 \end{pmatrix}^{2} \right) \begin{pmatrix} 0.046 \\ \\ 0.225 \end{pmatrix} (0.038)^{-1} \\ \widehat{\mathbf{A}} \left(\begin{pmatrix} 0.579 & 0.461 \\ \\ -0.461 & -0.192 \end{pmatrix}^{\top} - \begin{pmatrix} 0.046 \\ \\ -0.225 \end{pmatrix} (0.038)^{-1} \begin{pmatrix} 1.650 \\ \\ 0.851 \end{pmatrix}^{\top} \right) = \mathbf{1} \end{cases}$$

so we get:

$$\widehat{\mathbf{A}} = \begin{pmatrix} 0.578 & 0.178 \\ \\ -1.221 & -0.169 \end{pmatrix}, \quad \widehat{\boldsymbol{\alpha}} = \begin{pmatrix} 7.105 \\ \\ -1.579 \end{pmatrix}, \quad \widehat{\boldsymbol{\beta}} = \begin{pmatrix} 0.353 \\ \\ 0.474 \end{pmatrix}.$$

Now, we want to extract the rational component. To do so, we look at the modulus of the eigenvalues of $\widehat{A}$. We have:

$$\lambda_{1,2} = 0.204593 \pm 0.278322i.$$

As we can see, both eigenvalues have modulus smaller than one. Following the notation introduced in the previous section, we obtain: $\widehat{A}_k = \langle \widehat{\boldsymbol{\alpha}}_+, \widehat{\mathbf{A}}_+, \widehat{\boldsymbol{\beta}}_+ \rangle = \langle \widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}} \rangle$.

Algorithm 1: `AAKapproximation`

input : A minimal WFA A, with $\boldsymbol{\alpha}_2 \neq 0$, n states and in SVA form, its Gramian $\mathbf{P}$, a target number of states $k < n$

output: A WFA $\widehat{A}_k$ with k states

1 Let $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{A}_{11}, \mathbf{A}_{12}, \mathbf{A}_{22}, \boldsymbol{\Sigma}$ be the blocks defined in Eq. 4.17

2 Let $(\boldsymbol{\alpha}_2^\top)^+ = \frac{\boldsymbol{\alpha}_2}{\boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_2}$, $(\boldsymbol{\beta}_2^\top)^+ = \frac{\boldsymbol{\beta}_2}{\boldsymbol{\beta}_2^\top \boldsymbol{\beta}_2}$

3 Let $\mathbf{R} = \sigma_k^2 \mathbf{1} - \boldsymbol{\Sigma}^2$

4 Let $\widehat{\mathbf{A}} = (\mathbf{A}_{11}^\top - \mathbf{A}_{21}^\top (\boldsymbol{\beta}_2^\top)^+ \boldsymbol{\beta}_1^\top)^{-1}$

5 Let $\widehat{\boldsymbol{\alpha}} = \widehat{\mathbf{A}}^\top \mathbf{R} \mathbf{A}_{12} (\boldsymbol{\alpha}_2^\top)^+$

6 Let $\widehat{\boldsymbol{\beta}} = -\widehat{\mathbf{A}} \mathbf{A}_{21}^\top (\boldsymbol{\beta}_2^\top)^+$

7 Let $\widehat{A} = \langle \widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}} \rangle$

8 Let $\widehat{A}_k \leftarrow$ `BlockDiagonalize`$(\widehat{A})$

9 **return** $\widehat{A}_k$

4.3 Algorithm

We now use the results obtained in the previous section to define Algorithm 1, that we call `AAKapproximation`.

The algorithm takes as input a target number of states $k < n$, a minimal irredundant WFA A n states and in SVA form, and its Gramian $\mathbf{P}$. We assume $\boldsymbol{\alpha}_2 \neq 0$. If $\boldsymbol{\alpha}_2 = 0$, it is enough to substitute the Steps $4, 5, 6$ with the analogues from Equation 4.20. As mentioned in Section 4.1.1, the constraints on the WFA A to be minimal and in SVA form are not essential. In fact a WFA with n states can be minimized in time $O(n^3)$ [BR11], and the SVA computed in $O(n^3)$ [BPP19]. The algorithm applies the results of Theorem 4.2.3 in order to derive the parameters of the optimal WFA. The output of the algorithm is the WFA $\widehat{A}_k$ corresponding to the unique optimal spectral-norm approximation of A.

Block Diagonalization The algorithm involves a call to Algorithm 2, `BlockDiagonalize`. This algorithm corresponds to the steps necessary to derive the WFA $\widehat{A}_k$ associated to the rational function g. One way to solve the problem is to compute the Jordan form of the matrix. Unfortunately, this problem is ill-conditioned, so it is not suitable for our algorithmic purposes. Following an idea of Glover [Glo84], we compute the Schur decomposition, *i.e.*

we find an orthogonal matrix $\mathbf{U}$ such that the matrix $\mathbf{U}^\top \widehat{\mathbf{A}} \mathbf{U}$ is upper triangular, with the eigenvalues of $\widehat{\mathbf{A}}$ on the diagonal. We obtain:

$$\mathbf{T} = \mathbf{U}^\top \widehat{\mathbf{A}} \mathbf{U} = \begin{pmatrix} \widehat{\mathbf{A}}_+ & \widehat{\mathbf{A}}_{12} \\ \mathbf{0} & \widehat{\mathbf{A}}_- \end{pmatrix} \tag{4.31}$$

where the eigenvalues are arranged in increasing order of modulus, and the modulus of those in $\widehat{\mathbf{A}}_+$ (resp. $\widehat{\mathbf{A}}_-$) is smaller (resp. greater) than one. To transform this upper triangular matrix into a block-diagonal one, we use the following result.

Theorem 4.3.1 ([Rot52])**.** *Let* $\mathbf{T}$ *be the matrix defined in Equation 4.31. The matrix* $\mathbf{X}$ *is a solution of the equation* $\widehat{\mathbf{A}}_+ \mathbf{X} - \mathbf{X} \widehat{\mathbf{A}}_- + \widehat{\mathbf{A}}_{12} = \mathbf{0}$ *if and only if the matrices*

$$\mathbf{M} = \begin{pmatrix} \mathbf{1} & \mathbf{X} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}, \quad \textit{and} \quad \mathbf{M}^{-1} = \begin{pmatrix} \mathbf{1} & -\mathbf{X} \\ \mathbf{0} & \mathbf{1} \end{pmatrix} \tag{4.32}$$

satisfy:

$$\mathbf{M}^{-1} \mathbf{T} \mathbf{M} = \begin{pmatrix} \widehat{\mathbf{A}}_+ & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{A}}_- \end{pmatrix}, \tag{4.33}$$

where $\mathbf{T}$ *is the matrix defined in Equation 4.31.*

Setting $\mathbf{\Gamma} = \begin{pmatrix} \mathbf{1}_k & \mathbf{0} \end{pmatrix}$ we can now derive the rational component of the WFA:

$$\widehat{\mathbf{A}}_+ = \mathbf{\Gamma} \mathbf{M}^{-1} \mathbf{U}^\top \widehat{\mathbf{A}} \mathbf{U} \mathbf{M} \mathbf{\Gamma}^\top \tag{4.34}$$

$$\widehat{\boldsymbol{\alpha}}_+ = \mathbf{\Gamma} \mathbf{M}^\top \mathbf{U}^\top \widehat{\boldsymbol{\alpha}} \tag{4.35}$$

$$\widehat{\boldsymbol{\beta}}_+ = \mathbf{\Gamma} \mathbf{M}^{-1} \mathbf{U}^\top \widehat{\boldsymbol{\beta}}. \tag{4.36}$$

The algorithm `BlockDiagonalize` corresponds to the implementation of this procedure, and Step 2 can be performed using the Bartels-Stewart algorithm [BS72].

Algorithm 2: `BlockDiagonalize`

input : A WFA $\widehat{A}$
output: A WFA $\widehat{A}_k$ wit $\rho < 1$

Compute the Schur decomposition of $\widehat{\mathbf{A}} = \mathbf{U}\mathbf{T}\mathbf{U}^\top$, where $|T_{11}| \leq |T_{22}| \leq \dots$
Solve $\widehat{\mathbf{A}}_{11}\mathbf{X} - \mathbf{X}\widehat{\mathbf{A}}_{22} + \widehat{\mathbf{A}}_{12} = \mathbf{0}$ for $\mathbf{X}$
Let $\mathbf{M} = \begin{pmatrix} \mathbf{1} & \mathbf{X} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}$ and $\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{1} & -\mathbf{X} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}$
Let $\mathbf{\Gamma} = \begin{pmatrix} \mathbf{1}_k & \mathbf{0} \end{pmatrix}$
Let $\widehat{\mathbf{A}}_+ = \mathbf{\Gamma}\mathbf{M}^{-1}\mathbf{U}^\top\widehat{\mathbf{A}}\mathbf{U}\mathbf{M}\mathbf{\Gamma}^\top$
Let $\widehat{\boldsymbol{\alpha}}_+ = \mathbf{\Gamma}\mathbf{M}^\top\mathbf{U}^\top\widehat{\boldsymbol{\alpha}}$
Let $\widehat{\boldsymbol{\beta}}_+ = \mathbf{\Gamma}\mathbf{M}^{-1}\mathbf{U}^\top\widehat{\boldsymbol{\beta}}$
Let $\widehat{A}_k = \langle \widehat{\boldsymbol{\alpha}}_+, \widehat{\mathbf{A}}_+, \widehat{\boldsymbol{\beta}}_+ \rangle$
return $\widehat{A}_k$

4.3.1 Computational Cost

The running time of `BlockDiagonalize` with input a WFA $\widehat{A}$ with $(n-r)$ states is thus in $O((n-r)^3)$, where r is the multiplicity of the singular value considered. The running time of `AAKapproximation` for an input WFA $\widehat{A}$ with n states is in $O((n-r)^3)$. In particular, it is possible to analyze the cost associated to each step of the algorithms [TBI97]:

- The product of two $n \times n$ matrices can be computed in time $O(n^3)$ using a standard iterative algorithm.
- The inversion of a $n \times n$ matrix can be computed in time $O(n^3)$ using Gauss-Jordan elimination.
- The computation of the Schur decomposition of a $n \times n$ matrix can be done with a two-step algorithm, where each step takes $O(n^3)$, using the Hessenberg form of the matrix.
- The Bartels-Stewart algorithm applied to upper triangular matrices to find a matrix of size $m \times n$ takes $O(mn^2 + nm^2)$.

4.4 Error Analysis

Thanks to the use of AAK theory, the method outlined in the previous sections is guaranteed to return the rank k optimal spectral-norm approximation of a WFA satisfying our assumptions, and the singular number σ_k provides the error. As noticed before, since the Hankel matrix has finite rank and we can derive the Gramian matrices of the WFA, the singular number corresponding to the error can be computed precisely, even though the Hankel matrix is infinite.

Similarly to the case of SVA truncation [BPP19], owing to the ordering of the singular numbers, the error decreases when k increases, meaning that allowing $\widehat{A}_k$ to have more states guarantees a better approximation of A. Note that the solution we propose is optimal in the spectral norm, but it might not be the case in other norms. Nonetheless, we have the following bound between ℓ^2 norm and spectral norm.

Theorem 4.4.1. *Let A be a minimal WFA computing $f : \Sigma^* \to \mathbb{R}$, with matrix $\mathbf{H}$. Let $\widehat{A}_k$ be its optimal spectral-norm approximation, computing $g : \Sigma^* \to \mathbb{R}$, with matrix $\mathbf{G}$. Then:*

$$\|f - g\|_{\ell^2} \leq \|\mathbf{H} - \mathbf{G}\| = \sigma_k. \tag{4.37}$$

Proof. Let $\mathbf{e}_0 = \begin{pmatrix} 1 & 0 & \cdots \end{pmatrix}^\top$, $f : \Sigma^* \to \mathbb{R}$, $g : \Sigma^* \to \mathbb{R}$ with Hankel matrices $\mathbf{H}$ and $\mathbf{G}$, respectively. We have:

$$\begin{aligned} \|f - g\|_{\ell^2} &= \left(\sum_{n=0}^{\infty} |f_n - g_n|^2 \right)^{1/2} \\ &= \|(\mathbf{H} - \mathbf{G})\mathbf{e}_0\|_{\ell^2} \\ &\leq \sup_{\|\mathbf{x}\|_{\ell^2}=1} \|(\mathbf{H} - \mathbf{G})\mathbf{x}\|_{\ell^2} \\ &= \|\mathbf{H} - \mathbf{G}\| = \sigma_k \end{aligned}$$

where the second equation follows by definition and by observing that matrix difference is

computed entry-wise. □

4.5 Relaxing the Spectral Radius Assumption

In this section we examine the possibility of extending our method by relaxing one of the hypothesis made at the beginning of the chapter. In particular, we consider a WFA over a one-letter alphabet with $\rho(\mathbf{A}) \neq 1$, *i.e.* not necessarily irredundant. In this case, the method can be extended and the quality of the approximation can be estimated, but the result is not optimal in the spectral norm. Once again, we draw inspiration from the control theory literature, where some theoretical work has been done to study an analogous approach for continuous time systems and their approximation error [Glo84].

The key idea is to block-diagonalize $\mathbf{A}$ like we did in Section 4.2.4. This way, we obtain two components, $\mathbf{A}_+$ and $\mathbf{A}_-$, with the property that $\rho < 1$ and $\rho > 1$, respectively. We tackle each component separately. The case of $A_+ = \langle \boldsymbol{\alpha}_+, \mathbf{A}_+, \boldsymbol{\beta}_+ \rangle$, the component having $\rho(\mathbf{A}) < 1$, can be dealt with in the way presented in the previous sections. This means that we can find an optimal spectral-norm approximation of the desired size for A_+. Then, we can consider the second component, $A_- = \langle \boldsymbol{\alpha}_-, \mathbf{A}_-, \boldsymbol{\beta}_- \rangle$. In this case, we apply the transformation

$$z^{j-1} \mapsto z^{-j} \quad \text{for } j \geq 1$$

to the symbol $\phi'(z)$ associated to A_-. Then, the function

$$\phi'(z^{-1}) = \sum_{k \geq 0} \boldsymbol{\alpha}_-^\top \mathbf{A}_-^k z^k \boldsymbol{\beta}_- = \boldsymbol{\alpha}_-^\top (\mathbf{1} - z\mathbf{A}_-)^{-1} \boldsymbol{\beta}_- \tag{4.38}$$

is well defined, as the series converges for z with small enough modulus. The use of this transformation allows us to obtain a function having poles only inside the unit disc, and to apply the method presented in this chapter. We remark that in this case an important choice to make is the size of the target approximation of A_-, as it can influence the quality of the

result. Analyzing the effects of this parameter on the approximation error is an interesting direction for future work, both on the theoretical and experimental side.

4.6 Discussion

In this chapter, we proposed an algorithm based on AAK theory for the approximate minimization problem of weighted finite automata with real weights over a one-letter alphabet. Leveraging the results from operator theory and complex analysis presented in Section 2.4, we constructed the best possible approximation to an automaton given a bound on the size. In particular, under the assumption that the WFA is irredundant, we provided theoretical guarantees and an algorithm to find the parameters of the best WFA approximation in the spectral norm. Moreover, we provided bounds on the error in the spectral and ℓ^2 norms. The proposed method applies to real WFAs $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$, defined over a one-letter alphabet, with the property that the spectral radius of the transition matrix is strictly smaller than 1. One-letter alphabets have proven to be of independent interest when dealing with automata, as in this case the classes of regular and of context-free languages collapse [Pig15]. While this setting is certainly restricted, we believe that this result constitutes a first fundamental step towards optimal approximation. Furthermore, to the best of out knowledge it constitutes the first attempt to use AAK techniques in the setting of formal languages and automata theory. These methods have been very fruitful in related areas like control theory and signal processing, so we think that automata theory can also benefit from it. The result presented in this chapter highlights and strengthens the interesting connections between functional analysis, automata theory and control theory, unifying tools from different domains in one formalism.

Chapter 5

Black-Box Models for Language Modelling on One-Letter Alphabets

In this chapter, we study the approximate minimization problem of black boxes computing a function $f \in \ell^1$ over a one-letter alphabet. Note that this setting encompasses the case of models trained for language modelling on sequential data over a one-letter alphabet. Analogously to what done in the previous chapter, we start by formulating the problem and emphasizing the hypotheses necessary to solve it. In Section 5.2 we outline the idea behind the approximate minimization method, and we propose the application of a well-known result in signal processing to test for compactness [CL94]. Then, in Section 5.3, we present and justify the building blocks of the proposed approximation algorithm. We end the chapter with an analysis of the approximation error (Section 5.4) and with concluding remarks (Section 5.5). The content of this chapter was presented in the paper "Extracting Weighted Automata for Approximate Minimization in Language Modelling" [LPR21], that was published and presented at ICGI 2020/2021 (the 15th International Conference on Grammatical Inference).

5.1 Problem Formulation

To properly formulate our problem we need a few preliminary definitions. We start by introducing the notion of asymptotic sequence. We say that the sequence of matrices $\{\mathbf{G}_k^i\}_{i\geq 0}$ is an *asymptotic sequence* for the matrix $\mathbf{G}_k$, if the corresponding sequence of operators $\{G_k^i\}_{i\geq 0}$ converges to the operator G_k in the operator norm, *i.e.*, if:

$$\lim_{i\to\infty} \left\| G_k - G_k^i \right\| = 0. \tag{5.1}$$

In this case, we can use the notation $G_k^i \to G_k$.

Let $|\Sigma| = 1$, $\Sigma^* = \mathbb{N}$. We consider a black-box model trained for language modelling and computing a function $f : \mathbb{N} \to \mathbb{R}$, with Hankel matrix $\mathbf{H}$ corresponding to the operator H. Let k be the target size of the approximation, and $n > k$. We denote with G_k the optimal approximation of H of rank k. We say that a WFA $\widehat{A}_k^n$ with k states is an *asymptotically-optimal* (n, k)*-approximation* for a LM black box if the Hankel matrix $\mathbf{G}_k^n$ of $\widehat{A}_k^n$ belongs to an asymptotic sequence for $\mathbf{G}_k$. In the notation, we will omit the specification (n, k) whenever the parameters are clear from the context. We remark that this asymptotically-optimal analysis does not return a unique optimal approximation, ad the result is dependent on the chosen converging sequence.

Note that, if $\{\sigma_j\}_{j\geq 0}$ are the singular numbers of H, for an asymptotic sequence we have:

$$\lim_{i\to\infty} \left\| H - G_k^i \right\| = \sigma_k. \tag{5.2}$$

Alternative Formulation

We briefly remark that the approximate minimization problem can be formulated in a different way, that was explored by Kung and Lin [KL81] in the context of linear time-invariant dynamical systems. In the problem definition illustrated above, we fixed the size of the approximation, and searched for the solution producing the smallest possible error. Alter-

natively, we could have set the tolerance ρ allowed for the approximation error, with the objective of finding the smallest minimal WFA such that the spectral norm of the approximation error is smaller than ρ. In this case, if $\rho \in (\sigma_k, \sigma_{k-1})$, then the best approximation has size at least $k-1$, and can be found following the same solution as the standard approximation problem. Note that this formulation can be particularly interesting for applications. In fact, it allows us to reduce the computational cost and minimize the system while having control of the approximation error.

5.1.1 Assumptions

We list and justify our main assumptions.

Size of the Alphabet

Analogously to what done in the previous chapter, we restrict the setting to one-letter alphabets. This restriction, necessary to apply the Fourier isomorphism and use AAK theory (as noted in Section 3.2), is the main limitation of this approach.

Compactness of the Operator

The proof of Theorem 2.4.6 is constructive only for compact operators. We show in Section 5.2.2 that compactness is automatically respected by black boxes for language modelling (Theorem 5.2.2). Under these assumptions, the proposed algorithm can be applied to any black-box model trained for language modelling on a one-letter alphabet, for example RNNs [HS97] and transformers [VSP+17]. The necessary condition is actually less restrictive, and allows us to potentially apply the method beyond the task of language modelling: if f is the function computed by the black box considered, it is enough that $f \in \ell^1$.

Rank of the Matrix

We assume that the rank r of $\mathbf{H}$ is infinite. The finite-rank problem encompasses the case in which the black box is a WFA, or it is assumed to be computing a function that could be recognized by a WFA. This case could be reformulated in the framework of the algorithm illustrated in the previous chapter. In that case, it is important to make sure that the WFA obtained using the spectral method from the Hankel matrix of the black box is also computing a bounded function (by checking the property on the spectral radius of the transition matrix). Note that the algorithm proposed in this chapter can be easily adapted to finite-rank Hankel matrices. We require this additional assumption just to simplify the exposition of our solution, which specifically extends the previous chapter in the case of infinite-rank Hankel matrices. This is a clear improvement with respect to most of the works in the literature, where the RNN or the black box is generally trained over a regular language (therefore producing a finite-rank Hankel matrix).

5.2 Approximate Minimization

In this section, we present the theoretical details of the proposed method for approximate minimization. We first outline our approach and provide an intuition of the asymptotic method. Then, we analyze the problem of testing for compactness, and define asymptotic sequences to link the approximate minimization task to AAK theory.

5.2.1 Outline

Given the Hankel matrix $\mathbf{H}$ of a black box, we would like to use Theorem 2.4.6 to find the optimal approximation. When the rank is not finite, this might not be algorithmically possible. For infinite-rank Hankel matrices, the asymptotically-optimal (n, k)-approximation is the closest we can get to an optimal approximation. As a matter of fact, we can look at

the at the sequence of finite rank operators $\{H^i\}_{i\geq 0}$ converging to H:

$$H^0 \quad H^1 \quad \dots \quad H^n \quad \dots \quad H. \tag{5.3}$$

We will show that H is a compact operator, so this sequence is guaranteed to exist by definition. Each element of the sequence has finite rank, therefore we can use AAK theory to find its best approximation in the spectral norm. This way, we find the optimal approximation G_k^i of size k of each finite-rank Hankel operator H^i. Therefore, we obtain a second sequence of operators $\{G_k^i\}_{i\geq 0}$:

$$\begin{array}{ccccccc} H^0 & H^1 & \cdots & H^n & \cdots & H \\ \downarrow & \downarrow & & \downarrow & & \downarrow \\ AAK & AAK & \cdots & AAK & \dots & AAK \\ \downarrow & \downarrow & & \downarrow & & \downarrow \\ G_k^0 & G_k^1 & \cdots & G_k^n & \cdots & G_k \end{array} \tag{5.4}$$

If $\{\mathbf{G}_k^i\}_{i\geq 0}$ is an asymptotic sequence, then it is enough to search for the best rank-k approximation of H_n to find an asymptotically-optimal (n,k)-approximation.

In the next section, we study the convergence of asymptotic sequences. We prove that a solution for the asymptotically-optimal problem can be obtained from Theorem 2.4.6, but it is not unique, since the result is dependent on the chosen sequence $\{H^i\}_{i\geq 0}$. Nonetheless, we show that we can get arbitrarily close to the optimal solution.

For an intuition of the problem formulation, we can think about the Hankel matrices in terms of WFAs. Given a sequence of finite rank matrices $\{\mathbf{H}^i\}_{i\geq 0}$ converging to $\mathbf{H}$, we can associate to each of them a WFA (Theorem 2.4.4). Therefore we have a sequence of WFAs of increasing size that "converges" to the LM black box. The matrix $\mathbf{G}_k$ of rank k corresponds to the optimal approximation for $\mathbf{H}$, *i.e.*, it is the WFA $\widehat{A}_k$ with k states that best approximate the LM black box. Finally, the sequence of matrices $\mathbf{G}_k^i$ of optimal approximations of rank k, corresponds to a second sequence of WFAs $\widehat{A}_k^i$, all having size

k. When $\{\mathbf{G}_k^i\}_{i\geq 0}$ is an asymptotic sequence for $\mathbf{G}_k$, the corresponding sequence of WFAs "converges" to $\widehat{A}_k$.

We solve the approximation problem using the following steps:

1. *Extract the Hankel matrix from the black box.* Given a LM black box, we use it to fill the entries of a truncated Hankel matrix H^n, which belongs to the sequence of finite rank operators $\{H^i\}_{i\geq 0}$ converging to H.

2. *Compute the optimal symbol ψ.* The objective at this point is to derive the optimal symbol. To do so, we apply the formula in Theorem 5.3.1.

3. *Extract the rational component.* We apply Theorem 2.4.4 and isolate the part of the function with poles inside the unit disc using partial fraction decomposition.

4. *Obtain the optimal approximation.* We leverage the recursive structure of Hankel matrices and use the rational function's coefficients to fill the entries of the Hankel matrix of the asymptotically-optimal approximation. Then, we use the spectral method to extract a WFA corresponding to the Hankel matrix obtained through the previous steps.

5.2.2 Testing for Compactness

In this section, we focus on the problem of establishing the compactness of the Hankel operator associated to the LM black box. This is a necessary step, as compactness is required in the constructive proof of Theorem 2.4.6 and Theorem 5.3.1. In the case of a finite rank matrix, it is enough to test if $f \in \ell^2$ [BPP19]. Then, the problem can be rewritten in terms of finite matrices, the Gramians, and it is possible to find the exact error and an algorithm returning the optimal approximation. In the infinite-rank case considered in this chapter, this is not possible. Furthermore, there is no guarantee that the problem can be solved algorithmically. Thankfully, a Hankel operator does not need to have finite rank in order to be compact. We can see this in the following example.

Example 5.2.1. We consider the Hankel operator associated to the matrix:

$$\mathbf{H} = \begin{pmatrix} 1 & \frac{1}{2^2} & \frac{1}{3^2} & \cdots \\ \frac{1}{2^2} & \frac{1}{3^2} & \frac{1}{4^2} & \cdots \\ \frac{1}{3^2} & \frac{1}{4^2} & \frac{1}{5^2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The operator is bounded and compact even though it doesn't have finite rank.

Therefore we need to tackle two important challenges. First, we need a new criterion to test for compactness. Second, we need to find an alternative way to make our approach algorithmic, since we cannot rely anymore on the Gramian matrices.

We address the first problem using the definition of compactness: if we can find a sequence $\{H^i\}_{i\geq 0}$ of bounded and finite rank Hankel operators converging to H, then H is compact. To start, we can consider the matrix of the LM black box, indexed using natural numbers (we leverage the fact that the set of strings is isomorphic to $\mathbb{N}$), so we have $\mathbf{H}(i,j) = f(i+j)$. Then, we build by truncation a converging sequence. Let $t \geq 0$, we consider the sequence of Hankel matrices defined as:

$$\mathbf{H}^t(i,j) = \begin{cases} f(i+j) & \text{if } i+j \leq t \\ 0 & \text{otherwise} \end{cases}. \tag{5.5}$$

For example, the element at position n is:

$$\mathbf{H}^n = \begin{pmatrix} f_0 & f_1 & \dots & f_{n-1} & 0 & \dots \\ f_1 & & \iddots & \iddots & \vdots & \\ \vdots & \iddots & \iddots & & \vdots & \\ f_{n-1} & \iddots & & & \vdots & \\ 0 & \dots & \dots & \dots & 0 & \\ \vdots & & & & & \ddots \end{pmatrix}. \tag{5.6}$$

In the following theorem we prove that the sequence converges.

Theorem 5.2.2. *Let $|\Sigma| = 1$. Let $f : \mathbb{N} \to \mathbb{R}$ be the function computed by a black box for language modelling, and let $\mathbf{H}$ be its Hankel matrix. Let $\{\mathbf{H}^t\}_{t\geq 0}$ be the sequence of matrices defined in Equation 5.5. Then, since $f \in \ell^1$, we have that the sequence of the corresponding Hankel operators $\{H^t\}_{t\geq 0}$ converges to H.*

Proof. Let $f : \mathbb{N} \to \mathbb{R}$ be the function computed by the black box. We have:

$$\left\|H - H^t\right\| \leq \left\| \sum_{i=0}^{\infty} f(i) z^{-i-1} - \sum_{i=0}^{t} f(i) z^{-i-1} \right\|_{\infty} \tag{5.7}$$

$$\leq \left\| \sum_{i=t+1}^{\infty} f(i) z^{-i-1} \right\|_{\infty} \tag{5.8}$$

$$\leq \sum_{i=t+1}^{\infty} |f(i)| \tag{5.9}$$

where the first inequality follows from Theorem 2.4.3. Since the black box is trained for language modelling, we have that $\sum_{k\geq 0} |f(k)| = 1$. Thus, $f \in \ell^1$, and it follows directly that $H^t \to H$. □

We remark that the proof relies only on $f \in \ell^1$, so the result can hold for tasks different from language modelling.

We have found a sequence of finite-rank operators converging to H, therefore the operator

is compact. Note that the definition of truncated sequence presented in Equation 5.5 directly addresses the question of finding an algorithmic implementation for the problem. In fact, this definition reduces each matrix to a non-zero finite sub-block. This allows us to discard the infinite zero-part and to work only with the $n \times n$ sub-block of the matrix $\mathbf{H}^n$.

5.2.3 Asymptotic Sequences

The last theoretical step to address is the continuity of the approximation: if G_k and G_k^i are the optimal approximations of H and of H^i, respectively, we want $\{\mathbf{G}_k^i\}_{i\geq 0}$ to be an asymptotic sequence for $\mathbf{G}_k$, so that the sequence of operators $\{G_k^i\}_{i\geq 0}$ converges to G_k. This problem has been extensively studied, in the context of signal processing, in the fundamental work of Chui and Li [CL94, CLW91]. We recall the following result.

Theorem 5.2.3 ([CL94])**.** *Let H be a bounded Hankel operator, $\{\sigma_i\}_{i\geq 0}$, its singular numbers. Suppose to have a sequence $\{H^i\}_{i\geq 0}$ of bounded Hankel operators converging to H. Let G_k and G_k^i be the unique optimal approximations of rank k of H and of H^i for any i, respectively. If $\sigma_{k-1} \neq \sigma_k$, then the sequence $\{G_k^i\}_i \geq 0$ converges to G_k.*

This theorem gives us conditions under which we can solve the approximation problem (at least asymptotically) for the black box. To apply the theorem, we need to test that the property $\sigma_{k-1} \neq \sigma_k$ on the singular numbers of the bounded Hankel operator H holds when k is the size of the best approximation. This condition cannot be tested experimentally, since we don't have access to the infinite Hankel matrix $\mathbf{H}$, and we cannot compute precisely the singular numbers. Instead, we can address the problem by using arguments from random matrix theory, using the relation between singular numbers and eigenvalues. In fact, a random matrix has only distinct eigenvalues with probability one [TV14]. Therefore, up to at worst a small perturbation, we can view any $\mathbf{H}^t$ for $t > 0$ as a random matrix having only simple singular values with probability one, and this property holds (in the limit) also for $\mathbf{H}$ [vNW93, TV14]. Compact operators have a simple spectrum after arbitrarily small

perturbations, and the spectrum of symmetric matrices is very stable, so the quality of the result is not affected [HM94, Kat13, Tao12]. In practice, for most settings the Hankel matrix $\mathbf{H}$ will satisfy the condition of Theorem 5.2.3 with probability one. This is the case, for example, of RNNs trained using a gradient-based method with a random initialization. Thus, it remains to consider the case of an adversarial setting, in which the black box to approximate is specifically engineered to have $\sigma_{k-1} = \sigma_k$. To address this situation we can add some random noise to the matrix $\mathbf{H}$ post training. The matrix of noise $\mathbf{N}$ has to be chosen appropriately, if we want to apply our method. In particular, we need to preserve compactness and the Hankel property. This means that $\mathbf{N}$ needs to be a Hankel matrix, and that $H + N$ needs to be a compact operator. For instance, we can choose $\mathbf{N}$ to be a Hankel matrix, with first row $\mathbf{N}(0, j)$ sampled uniformly in the interval $[-(j+2)^{-p}, (j+2)^{-p}]$, with $p \geq 2$ fixed, so that the operator N is compact. This way, we are guaranteed to have a random Hankel matrix, and the condition for compactness is respected. Moreover, for every $\varepsilon > 0$, we can find an exponent $p \geq 2$ such that $\|\mathbf{N}\| \leq \varepsilon$, so the perturbation can be chosen to be arbitrarily small. Thus, $\mathbf{H}+\mathbf{N}$ is a random matrix corresponding to a compact Hankel operator, and satisfies the conditions of Theorem 5.2.3 with probability one. We will address the additional error due to small perturbations in Section 5.4.

We are finally ready to show that if $\mathbf{H}^n$ belongs to the sequence of bi-infinite truncation Hankel matrices $\{\mathbf{H}^t\}_{t\geq 0}$ introduced in Equation 5.5, then by solving the problem described by Theorem 2.4.6 for $\mathbf{H}^n$ we can find an asymptotically-optimal (n, k)-approximation.

Theorem 5.2.4. *Let $\mathbf{H}$ and $\mathbf{H}^n$ be as above, and assume $\sigma_k \neq \sigma_{k-1}$. If $\mathbf{G}_k^n$ is the optimal approximation of $\mathbf{H}^n$ according to Theorem 2.4.6, then a WFA having Hankel matrix $\mathbf{G}_k^n$ is an asymptotically-optimal (n, k)-approximation, and we have:*

$$\sigma_k \leq \|\mathbf{H} - \mathbf{G}_k^n\| \leq \sigma_k + 2\left(1 - \sum_{i=0}^{n} f(i)\right). \tag{5.10}$$

Proof. Let σ_k^n be the singular number $k+1$ of the operator H^n, and let $\mathbf{G}_k^n$ be the optimal

approximation described by Theorem 2.4.6, *i.e.*:

$$\|\mathbf{H}^n - \mathbf{G}_k^n\| = \sigma_k^n.$$

We have:

$$\begin{aligned}\|\mathbf{H} - \mathbf{G}_k^n\| &\leq \|\mathbf{H} - \mathbf{H}^n\| + \|\mathbf{H}^n - \mathbf{G}_k^n\| \\ &= \|\mathbf{H} - \mathbf{H}^n\| + \sigma_k^n.\end{aligned}$$

From Theorem 3.1.1 we know that $\|\mathbf{H} - \mathbf{G}_k^n\| \geq \sigma_k^n$. On the other hand, using Lemma B.2.1 and Cauchy's interlace theorem [Hwa04] (both recalled in Appendix B.2), we obtain:

$$\sigma_k^n \leq \sigma_k + \|\mathbf{H} - \mathbf{H}^n\|.$$

It follows that:

$$\sigma_k \leq \|\mathbf{H} - \mathbf{G}_k^n\| \leq \sigma_k + 2\|\mathbf{H} - \mathbf{H}^n\|. \tag{5.11}$$

Now, $\mathbf{H}^n$ belongs to the sequence of truncation matrices $\{\mathbf{H}^t\}_{t\geq 0}$, and the sequence converges to $\mathbf{H}$ (Theorem 5.2.2). Since $\sigma_k \neq \sigma_{k-1}$, the conditions of Theorem 5.2.3 hold. Therefore, the sequence of matrices of best approximations $\{\mathbf{G}_k^t\}_{t\geq 0}$ is an asymptotic sequence for $\mathbf{G}_k$, and $\mathbf{G}_n^k$ belongs to it. Thus, the WFA having matrix $\mathbf{G}_n^k$ is an asymptotically-optimal (n, k)-approximation, and Equation 5.2 holds. Moreover, from Equation 5.9, we have:

$$\|\mathbf{H} - \mathbf{G}_k^n\| \leq \sigma_k + 2\|\mathbf{H} - \mathbf{H}^n\| \leq \sigma_k + 2\sum_{i=n+1}^{\infty} f(i) = \sigma_k + 2\left(1 - \sum_{i=0}^{n} f(i)\right). \tag{5.12}$$

□

The bound clearly shows that, as n increases, we approach the optimal approximation.

5.3 Algorithm

In this algorithm, we propose to use an alternative version of AAK theorem, from Chui and Chen [CC97, Theorem 4.7]. Leveraging this result it is possible to find a symbol for the best approximation. We first recall that a *Toeplitz matrix* is a matrix $\mathbf{T}$ with entries defined by $\mathbf{T}(j,k) = t_{j-k}$ for $j,k \geq 0$:

$$\mathbf{T} = \begin{pmatrix} t_0 & t_{-1} & \dots & t_{-n} & t_{-n+1} & \dots \\ t_1 & t_0 & t_{-1} & \ddots & t_{-n} & \ddots \\ t_2 & t_1 & t_0 & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & t_{-1} & \\ \vdots & & & & & \ddots \end{pmatrix}. \tag{5.13}$$

Theorem 5.3.1 ([CC97])**.** *Let $\{\boldsymbol{\xi}_k, \boldsymbol{\eta}_k\}$ be any σ_k-Schmidt pair for H. We consider a bi-infinite upper triangular Toeplitz matrix $\mathbf{T}$, defined as follows:*

- $\mathbf{T}$ *has only zeros on the main diagonal*
- *the first row is defined by $\mathbf{T}(0,k) = \mathbf{H}(0,k-1)$ for $k > 0$*
- *the remaining entries are defined by $\mathbf{T}(j,k) = \mathbf{T}(j+1,k+1)$.*

Let $\mathbf{z} = \begin{pmatrix} 1 & z & z^2 & \dots \end{pmatrix}^\top$ where z is the complex variable. Then, the rational function $r(z)$ corresponding to a symbol for the best approximation of rank k is:

$$r(z) = \mathbb{P}_- \left(\frac{\mathbf{z}^\top \mathbf{T} \boldsymbol{\xi}}{\mathbf{z}^\top \boldsymbol{\xi}} \right). \tag{5.14}$$

To simplify the notation across this section, we set $f_i = f(i)$. We recall the two bi-infinite

matrices necessary to compute the best approximation:

$$\mathbf{H}^n = \begin{pmatrix} f_0 & f_1 & \dots & f_{n-1} & 0 & \dots \\ f_1 & & \iddots & \iddots & \vdots & \\ \vdots & \iddots & \iddots & & \vdots & \\ f_{n-1} & \iddots & & & \vdots & \\ 0 & \dots & \dots & \dots & 0 & \\ \vdots & & & & & \ddots \end{pmatrix}, \mathbf{T} = \begin{pmatrix} 0 & f_0 & \dots & f_{n-1} & 0 & \dots \\ \vdots & \ddots & \ddots & & f_{n-1} & \\ \vdots & & \ddots & \ddots & \vdots & \\ \vdots & & & \ddots & f_0 & \\ 0 & \dots & \dots & \dots & 0 & \\ \vdots & & & & & \ddots \end{pmatrix}. \tag{5.15}$$

While theoretically we are dealing with infinite matrices, the truncation proposed allows us to work with finite sub-blocks of them. For example, we consider only the $n \times n$ sub-block of $\mathbf{H}^n$, but we will still denote it with $\mathbf{H}^n$ for the sake of simplicity. Analogously, if $\mathbf{z}$ and $\mathbf{T}$ are the infinite vector and matrix defined in Theorem 5.3.1 for $\mathbf{H}$, in the algorithm we only consider the truncation:

$$\mathbf{z}^n(i) = \mathbf{z}(i), \quad \mathbf{T}^n(i,j) = \mathbf{T}(i,j) \quad \text{for } i,j < n, \qquad \mathbf{z}^n \in \mathbb{R}^{n\times 1}, \mathbf{T}^n \in \mathbb{R}^{n\times n} \tag{5.16}$$

where the discarded entries are irrelevant, being multiplied by zeros in the infinite case.

Note that, since we are assuming that $\mathbf{H}$ has infinite rank (see Section 5.1.1), the truncation $\mathbf{H}^n$ has full rank: if this was not the case, since $\mathbf{H}^n$ is the leading principal sub-matrix of $\mathbf{H}$, we would have $r = \text{rank}(\mathbf{H}^n)$ (it follows from a result of [Al'17], that we recall in Appendix B.2).

The algorithm takes as input a black box $\mathcal{M}$ trained for language modelling on a one-letter alphabet, a target number of states k, the size of the truncation $n > k$, and a perturbation matrix $\mathbf{N}^n$ defined as in Section 5.2.2, with $\mathbf{N}^n$ Hankel and $H^n + N^n$ compact. After obtaining the Hankel matrix of the black box and computing its singular values, we apply Theorem 5.3.1 in order to find its symbol. The main challenge is then to extract the rational

Algorithm 3: `AAKmethod`

input : A trained LM black box $\mathcal{M}$ of unknown rank,
a target number of states k, the size of the truncation $n > k$,
a perturbation matrix $\mathbf{N}^n$ as in Section 5.2.2

output: A WFA $\widehat{A}_k^n$ of size k

1 Let $\widetilde{\mathbf{H}}^n \leftarrow$ `GetHankel(`$\mathcal{M}, n, \mathbf{N}^n$`)`
2 Let σ_k^n, $\boldsymbol{\xi}^n \leftarrow$ `ComputeEigenpair(`$\widetilde{\mathbf{H}}^n$`)`
3 Let $\mathbf{T}^n$, $\mathbf{z}^n$ defined as in Equation 5.16
4 Let $\psi = \frac{(\mathbf{z}^n)^\top \mathbf{T}^n \boldsymbol{\xi}^n}{(\mathbf{z}^n)^\top \boldsymbol{\xi}^n}$
5 Let $r \leftarrow$ `ExtractRational(`ψ`)`
6 Let $\mathbf{G}_k^n \leftarrow$ `RecoverMatrix(`$r, k+1$`)`
7 Let $\widehat{A}_k^n \leftarrow$ `SpectralMethod(`$\mathbf{G}_k^n, \mathcal{B}$`)`
8 **return** $\widehat{A}_k^n$

component. We already tackled this problem in Chapter 4, but since in this case we don't want to extract the parameters of the WFA, we will follow a different approach. We use the method of partial fraction decomposition and obtain a rational function. It is then immediate to compute the entries of the corresponding Hankel matrix, and to use spectral learning to extract a WFA from it. The algorithm then returns a WFA $\widehat{A}_k^n$ having k states, with Hankel matrix corresponding to an asymptotically-optimal (n,k)-approximation of $\mathbf{H}$.

To better understand Algorithm 3, we analyze in detail its building blocks.

5.3.1 From Black Box to Hankel matrix

The step of extracting a Hankel matrix from a given black-box model corresponds to the first two lines of the algorithm and to the functions `GetHankel` and `ComputeEigenpair`.

Following [AEG18], we consider a black box trained for language modelling, and use it to fill the entries of a Hankel matrix $\mathbf{H}^n$. In particular, we query the black box on each string of length smaller than n, and fill the first line in $\mathbf{H}^n$ with the answer f_n. Since the Hankel property holds, we obtain a $n \times n$ Hankel matrix $\mathbf{H}^n$, having entries f_n on the first n anti-diagonals, and zeroes everywhere else. As mentioned in Section 5.2.3, we add a perturbation matrix $\mathbf{N}^n$ to $\mathbf{H}^n$, which can be set to zero when the singular numbers σ_k and σ_{k-1} of $\mathbf{H}$

are known to be distinct. The output of the function `GetHankel` is the perturbed matrix $\widetilde{\mathbf{H}}^n = \mathbf{H}^n + \mathbf{N}^n$.

Now that we have the Hankel matrix, we want to compute a Schmidt pair, since it will be needed in the next step of the algorithm. The function `ComputeEigenpair` takes the matrix $\widetilde{\mathbf{H}}^n$ and returns the singular number σ_k^n of $\widetilde{\mathbf{H}}^n$, and a corresponding singular vector. Since $\widetilde{\mathbf{H}}^n$ has finite rank and is symmetric, its singular numbers are the absolute values of the eigenvalues, *i.e.* $\sigma_k^n = |\lambda_k|$. Analogously, given the eigenvalue λ_k and a corresponding eigenvector $\mathbf{v}_k^n$, a Schmidt pair is given by $(\boldsymbol{\xi}^n, \boldsymbol{\eta}^n)$, with $\boldsymbol{\xi}^n = \mathbf{v}_k^n$, $\boldsymbol{\eta}^n = \operatorname{sgn}(\lambda_k)\mathbf{v}_k^n$, and $\operatorname{sgn}(\lambda_k) = \lambda_k/|\lambda_k|$. Once again, while the actual eigenvectors should be infinite, we consider only the first n coordinates.

5.3.2 Applying AAK Theory

Now we have obtained a singular vector, and we can compute the matrix $\mathbf{T}^n$ and the vector $\mathbf{z}^n$ (defined in Equation 5.16). These are all the elements necessary to apply AAK theory, as we can compute the function ψ by applying directly the formula of Theorem 5.3.1.

Once we have obtained ψ, the objective becomes to extract its rational component. From Theorem 2.4.4 we know that finite-rank Hankel matrices correspond to strictly proper rational functions, with all the poles inside the complex unit disc. From Equation 5.14 in Theorem 5.3.1 we obtain, before applying the projection, a function $\psi = \frac{a}{b}$. We are only interested in $r = \mathbb{P}_-\psi$, as ψ might contain poles outside the unit disc. By definition, the poles of ψ correspond to the zeros of b, so we can isolate the part of the function with poles inside the unit disc using partial fraction decomposition. This method allows us to rewrite the rational function $\psi = \frac{a}{b}$ as:

$$\psi = \frac{a}{b} = c + \sum_i \frac{a_i}{b_i}, \tag{5.17}$$

where each $\frac{a_i}{b_i}$ is a strictly proper rational function, and each factor b_i of the denominator is

a power of an irreducible polynomial. Now, we can analyze the position of the zero of each b_i: if it is outside or on the complex unit disc, then we discard the term $\frac{a_i}{b_i}$. The output of `ExtractRational` is the sum of the remaining terms, corresponding to the component in $\mathcal{H}^2_-$ of ψ. We remark that the general problem of computing a partial fraction decomposition may be ill-conditioned, in particular in presence of high-order poles [You83]. Nonetheless, in the setting analyzed in this chapter, the function considered is unlikely to have high order poles (this can be shown following the same random-matrix approach shown in the previous section). Therefore, the partial fraction decomposition can be computed efficiently, with the naive implementation having complexity $O(n^3)$ for a fraction with n poles [KT77, MYW14].

Using `ExtractRational` we obtained a strictly proper rational function:

$$r = \frac{p}{q}, \qquad \text{where} \quad p = \sum_{i=1}^{k} p_i z^{k-i}, \quad q = z^k + \sum_{i=1}^{k} q_i z^{k-i}, \tag{5.18}$$

and p and q are relatively prime, with q having degree k. As seen in Section 2.4, if $r = \sum_{n\geq 0} g_n z^{-n-1}$, then $\mathbf{G}^n_k(j,k) = g_{j+k}$. The coefficients g_i of the Hankel matrix can be recovered from the following set of equations, obtained from the constructive proof of Theorem 2.4.4 [CC97]:

$$\begin{cases} g_0 = p_1 \\ g_1 = p_2 - g_0 q_1 \\ \dots \\ g_{k-1} = p_k - g_{k-2}q_1 - \dots - g_0 q_{k-1} \end{cases} \qquad \begin{cases} g_k + \sum_{i=1}^{k} q_i g_{k-i} = 0 \\ g_{k+1} + \sum_{i=1}^{k} q_i g_{k+1-i} = 0 \\ \dots \end{cases} . \tag{5.19}$$

These equations form a linear system, which can be easily solved to derive the matrix $\mathbf{G}^n_k$ of rank k having entries $\mathbf{G}^n_k(i,j) = g_{i+j}$. It is important to note that we don't need to compute all the coefficients of the matrix in order to extract a WFA using the spectral method. We show in the next paragraph that the first $k+1$ coefficients are enough to retrieve the WFA.

5.3.3 From Hankel Matrix to WFA

In this last block we call the function `SpectralMethod` to extract a WFA from the Hankel matrix computed in the previous step. In particular, we use the function to recover the minimal WFA $\widehat{A}_k^n = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ with k states computing the function $g : \Sigma^* \to \mathbb{R}$ such that $\mathbf{G}_k^n(i,j) = g_{i+j}$. We use the spectral method outlined in Section 2.2. The first step of the algorithm is to select a prefix-closed and complete basis $\mathcal{B}$. As noted before, since we are working with a one-letter alphabet, the Hankel matrix $\mathbf{G}_k^n$ is symmetric. In this case, if $\mathbf{G}_k^n$ has rank k, then the size of the biggest leading principal submatrix is $k \times k$ [Al'17]. Consequently, the natural choice for $\mathcal{B} = (\mathcal{P}, \mathcal{S})$ is to have $\mathcal{P} = \mathcal{S}$, with $\mathcal{P}$ containing all the strings having size strictly smaller than k. Following the notation of Section 2.2, $\mathbf{H}_\varepsilon$ corresponds to the $k \times k$ leading principal submatrix of $\mathbf{G}_k^n$, and $\mathbf{h}_{\mathcal{P},\varepsilon}$, $\mathbf{h}_{\varepsilon,\mathcal{S}}$ are its first column and row, respectively. Finally, $\mathbf{H}_a$ is the sub-block of $\mathbf{G}_k^n$ having the same rows as $\mathbf{H}_\varepsilon$, and the columns obtained by shifting each individual column of $\mathbf{H}_\varepsilon$ by one column. Using Equation 2.6 we obtain the WFA $\widehat{A}_k^n$.

5.3.4 Computational Cost

The running time of the algorithm `AAKmethod` with input a target number of states k, and size of the truncation of the Hankel matrix $n > k$ is $O(n^3)$. In particular, it is possible to analyze the cost associated to each step of the algorithms [TBI97]:

- Given the structure of the Hankel matrix, the cost of filling the truncation of the Hankel matrix in Line 1 of the algorithm is $O(n)$.
- The cost of computing the singular value and the singular vector in Line 2 is at most $O(n^3)$.
- The product of two $n \times n$ matrices can be computed in time $O(n^3)$.

- The partial fraction decomposition in Line 5 can be computed efficiently, with the naive implementation having complexity $O(n^3)$ for a fraction with n poles [KT77, MYW14].
- Solving a linear system of $k+1$ equations (Line 6) has complexity $O((k+1)^3)$.
- Given the matrix $\mathbf{G}_k^n$ of rank k, we have that the total running time of the spectral algorithm in Line 7 is $O(k^3)$.

5.4 Error Analysis

As seen in the previous chapter, the minimal error that can be attained when doing spectral-norm approximation is given by the singular number σ_k, where k is the size of the approximation. When the rank is not finite, we can only recover an asymptotically-optimal solution, and the error cannot be exactly computed. We provide a bound on the error to estimate the quality of the approximation with respect to the optimal case. From Equation 5.12, we obtain the following bound:

$$\|\mathbf{H} - \mathbf{G}_k^n\| \leq \sigma_k + 2\left(1 - \sum_{i=0}^{n} f(i)\right). \tag{5.20}$$

It is important to remark that, since $f \in \ell^1$, we have information about the asymptotic behaviour of the function. In particular, $f(n) \to 0$ when $n \to \infty$, meaning that "little" probability is allocated to very long strings. Thus, a direct way to reduce the error is to select the biggest possible n as size of the truncation matrix.

We can estimate the value of σ_k in terms of σ_k^n using Lemma B.2.1 in Appendix B.2:

$$|\sigma_k - \sigma_k^n| \leq 1 - \sum_{i=0}^{n} f(i). \tag{5.21}$$

Note that σ_k^n can be precisely computed, since it is the singular value of a matrix having finite rank. An alternative way to reduce the error when additional information is available

is to explore other types of truncations. The use of this approach to improve the convergence rate has been explored by Chui and Li [CL94] in the context of signal processing.

Error with Noise

As discussed in Section 5.2.3, there are adversarial settings in which it is desirable to add a matrix of noise $\mathbf{N}$ to the Hankel matrix $\mathbf{H}$. When this happens, it is necessary to consider the effect that this process has on the approximation error. While theoretically we would like to add an infinite matrix of noise $\mathbf{N}$, practically we consider the finite sub-block of the infinite matrix $\mathbf{N}^n$ obtained by truncation in a way analogous to Equation 5.5. As before, we consider only the $n \times n$ sub-block of $\mathbf{N}^n$, but we still denote it with $\mathbf{N}^n$ for the sake of simplicity. We obtain the following bound.

Theorem 5.4.1. *Let $\mathbf{N}^n$ be the finite matrix of noise obtained by truncating the infinite matrix $\mathbf{N}$ as defined above, and let $\widetilde{\mathbf{G}}_k^n$ be an asymptotically-optimal (n,k)-approximation of the matrix $\widetilde{\mathbf{H}}^n = \mathbf{H}^n + \mathbf{N}^n$. Then the approximation error is bounded by:*

$$\left\|\mathbf{H} - \widetilde{\mathbf{G}}_k^n\right\| \leq \|\mathbf{H} - \mathbf{G}_k^n\| + 2\|\mathbf{N}^n\|. \tag{5.22}$$

Proof. Let $\widetilde{\mathbf{G}}_k^n$ and $\widetilde{\sigma}_k^n$ be the optimal approximation and the $(k+1)$-th singular number of $\mathbf{H}^n + \mathbf{N}^n$, respectively. From Theorem 2.4.6 we have:

$$\left\|\mathbf{H}^n + \mathbf{N}^n - \widetilde{\mathbf{G}}_k^n\right\| = \widetilde{\sigma}_k^n. \tag{5.23}$$

Then:

$$\begin{aligned}\left\|\mathbf{H}-\widetilde{\mathbf{G}}_k^n\right\| &\leq \|\mathbf{H}-\mathbf{H}^n-\mathbf{N}^n\|+\left\|\mathbf{H}^n+\mathbf{N}^n-\widetilde{\mathbf{G}}_k^n\right\| \\ &\leq \|\mathbf{H}-\mathbf{H}^n\|+\|\mathbf{N}^n\|+\widetilde{\sigma}_k^n \\ &\leq \|\mathbf{H}-\mathbf{H}^n\|+2\|\mathbf{N}^n\|+\sigma_k^n \\ &\leq \sigma_k+2\|\mathbf{H}-\mathbf{H}^n\|+2\|\mathbf{N}^n\|\end{aligned}$$

where we used Equation 5.23 in the second step, and in the last two steps we applied Lemma B.2.1, with:

$$|\widetilde{\sigma}_k^n-\sigma_k^n|\leq\|\mathbf{N}^n\|,$$

and

$$|\sigma_k^n-\sigma_k|\leq\|\mathbf{H}-\mathbf{H}^n\|.$$

□

This means that the additional error depends only on the norm of the matrix of noise. We have already remarked that this can be chosen to be arbitrarily small. Moreover, since only a finite sub-block of $\mathbf{N}^n$ is different from zero, the norm can be precisely computed.

5.5 Discussion

In this chapter, we presented the approximate minimization problem for black boxes trained for language modelling of sequential data over a one-letter alphabet. To solve this problem, we applied the AAK theory for Hankel operators [AAK71] and continuity results from the signal processing literature [CL94, CLW91]. Given a language model and a target size as input, we provided an algorithm to extract a WFA corresponding to an asymptotically-optimal approximation in the spectral norm. The algorithm can be applied to black box models like RNNs or transformers. In particular, we showed that minimizing the approximation er-

ror between a WFA and a black box model can be (asymptotically) solved optimally in a tractable way. This is a first step towards developing provable approximation algorithms for black-box models. Using this method, we can measure the distance between a given RNN and the extracted WFA. This is particularly valuable, especially given that the general equivalence problem between classes of WFAs and RNNs is undecidable [MdlH20].

The use of approximate minimization over regular extraction has the advantage that it allows us to choose the size of the approximation and search the optimal WFA within this constraint. This is particularly useful when the extracted WFA is used for interpretability. In fact, every WFA has a graphical representation, but this is helpful only when the number of states is small enough to actually make it readable. Moreover, approximate minimization can be used to reduce the computational cost of the task considered, as the new model is smaller and often easier to compute compared to the original one. Note that, even though the spectral algorithm used in our approach does not guarantee that the extracted WFA will preserve the probabilistic nature of the function considered, there are methods that can partially address this issue [Bai11, AGH$^+$14].

A point deserving further investigation is to understand how the approximation method that we proposed performs with respect to more popular metrics, such as word error rate or normalized discounted cumulative gain. This could help evaluate how meaningful (and accurate) it is to use the spectral norm in an experimental setting, but the comparison is possible only for multi-letter alphabets. Nonetheless, we think that the choice of the spectral norm to evaluate the approximation error constitutes an interesting way to approach the problem of approximating black boxes with WFAs.

Analogously to the case of WFAs presented in Chapter 4, the main drawback in this approach remains the restriction to one-letter alphabets. As noted in the previous chapter, the application of this rich mathematical theory has shown to be very effective in areas like control theory or signal processing, and this work further highlights fruitful connections with these fields.

Chapter 6

A Framework for Approximate Minimization in the Multi-Letter Case

In this chapter, we focus our attention on models defined over multi-letter alphabets. In the first section, we recall some fundamental results needed to understand our contribution. After a brief overview of noncommutative function theory and Fock spaces, we briefly illustrate the noncommutative version of the AAK theorem proposed by Popescu [Pop03]. Then, we generalize the approach proposed in Chapter 3 and present a framework to reformulate the approximate minimization problem in terms of multivariable noncommutative operator theory in the Fock space. Moreover, we suggest a way to link the Hankel matrix of a WFA to a symbol and a noncommutative rational function. Given the fundamental role played by these two mathematical objects in constructing the optimal approximation in the one-letter case, we consider this a fundamental step towards solving the problem in the multi-letter setting.

Part of this chapter was included in the work "Towards an AAK Theory Approach to Approximate Minimization in the Multi-Letter Case" [LPR22], presented at Learnaut 2022 (the 4th edition of the workshop Learning and Automata).

6.1 Preliminaries

In this section, we recall some notions from noncommutative function theory. This is not a comprehensive presentation of the subject, as we merely review the part of the theory that is strictly needed to approach our problem. For a more thorough exposition we refer the reader to [KVV14, KVV09, Pop93, BB19]. We follow the definitions and notations used by Popescu [Pop89b, Pop89c, Pop95b, Pop89a, Pop06a, AP95, Pop10, Pop06b, Pop13, Pop92], Ball and Bolotnikov [BB19] and Jury, Martin and Shamovich [JMS21b, JMS21a, Jur21]. We start by recalling the definition of the Fock space. We then define NC rational functions and NC Hankel operators, and conclude with a noncommutative analogue of the AAK theorem.

6.1.1 Fock Spaces and NC Functions

As a first step towards NC multivariable theory, we define NC functions. To gain some intuition about the reasoning behind the definitions, we start by providing an overview of noncommutative polynomials. It is natural to evaluate this type of polynomials over matrices, so we don't have to worry about preserving the noncommutative structure after evaluation. This means that, for each polynomial p, we can actually consider a family of polynomials p_n, where n is the size of the matrices considered. For example given a polynomial p, we denote with p_2 the same polynomial evaluated over matrices of size 2×2, with p_3 if it is evaluated over matrices of size 3×3, and so on. Under this representation, a NC polynomial defines a polynomial for each "level" determined by the size of the matrices. Therefore, we can consider the graded set

$$M^d = \bigsqcup_{i=1}^{\infty} M_n^d \tag{6.1}$$

where we denote with M_n the set of matrices of size $n \times n$, and with M_n^d the set of d-tuples with elements in M_n. By interpreting $n = \infty$ as the level of bounded operators, we can unify polynomials over matrices and over operators under the same formalism. A thorough

description of the functional calculus for the noncommutative case can be found in the work of Popescu [Pop95a]. In this thesis, we denote general tuples of NC variables with lower case letters, instead of bold capital letters, even if they may take values over matrices. We use bold capital letters or capital letters only when a result is stated specifically for matrices or operators, respectively. When the tuple is composed by operators, we refer to it as **row operator**.

To summarize, a NC polynomial is associated to a set of polynomials $p_n : M_n^d \to M_n$ for $n > 0$, where each p_n respects the size of the matrices. The levels of matrices are related, as p respects direct sums. In fact, if we consider $x \in M_n^d$, $y \in M_m^d$, we can define their direct sum to be the d-tuple with elements in M_{n+m}^d:

$$\begin{pmatrix} x_i & 0 \\ 0 & y_i \end{pmatrix} \qquad 0 < i \leq d. \tag{6.2}$$

Another property of interest is conjugation by similarities. Let $x = [x_1, \ldots, x_n] \in M_n^d$ and let $\mathbf{S}$ be an invertible $n \times n$ matrix. We say that a polynomial p respects similarities if:

$$p(\mathbf{S}^{-1} x \mathbf{S}) = \mathbf{S}^{-1} p(x) \mathbf{S}, \tag{6.3}$$

where $\mathbf{S}^{-1} x \mathbf{S} = [\mathbf{S}^{-1} x_1 \mathbf{S}, \ldots, \mathbf{S}^{-1} x_n \mathbf{S}]$.

More generally, we have the following definition.

Definition 6.1.1. *A **NC function** f is a function defined on tuples of matrices of all sizes such that f is graded and respects direct sums and similarities.*

In general, NC functions are defined over NC sets, namely subsets of M_d which are graded, respect direct sums and are closed under unitary transformations. An example of a

NC set is the **NC row ball**:

$$\mathbb{B}^d_{\mathbb{N}} = \bigsqcup_{i=1}^{\infty} \mathbb{B}^d_i\,; \qquad \mathbb{B}^d_n = \{x \in M^d_n : \sum_i ||x_i x_i^*|| < 1\}. \tag{6.4}$$

We fix a row structure on $\mathbb{B}^d_{\mathbb{N}}$.

Definition 6.1.2. *A d-tuple of matrices (or of bounded operators) $x = [x_1, \dots x_d] \in M^d_n$, with $x_i \in M_n$, is a **row contraction** if*

$$\sum_{i \le k} ||x_i x_i^*|| < 1. \tag{6.5}$$

The NC row ball consists of all the matrices (or operators) that are strict row contractions. Alternatively one can consider as NC set the column ball, where we require $\sum_i ||x_i^* x_i|| < 1$, or the polydisc, where each matrix is a contraction. While we won't analyze this in detail, NC functions have also interesting analytical properties. For example, if a NC function is continuous at each level n, then each level is holomorphic. Moreover, if a NC function is locally bounded, then each level is continuous (and therefore differentiable and holomorphic).

Similarly, one can introduce NC power series using the formalism provided by Fock spaces. Let $n \in \mathbb{N}$, we consider a n-dimensional Hilbert space $\mathcal{H}_n$. We denote with $\mathcal{H}_n^{\otimes k}$ the tensor product of k copies of $\mathcal{H}_n$, and $\mathcal{H}_n^{\otimes 0} := \mathbb{C}$. We define the full Fock space as follows.

Definition 6.1.3. *The **full Fock space** F^2 of $\mathcal{H}_n$ is the space:*

$$F^2 = F^2(\mathcal{H}_n) = \bigoplus_{k \ge 0} \mathcal{H}_n^{\otimes k} = \mathbb{C} \oplus \mathcal{H}_n \oplus (\mathcal{H}_n \otimes \mathcal{H}_n) \oplus \dots\ . \tag{6.6}$$

We can then consider the free monoid $\mathbb{F}_n$ on n generators $g_1, \dots g_n$, with identity element g_0. Given an element $\alpha \in \mathbb{F}_n$, with $\alpha = g_{i_1} g_{i_2} \cdots g_{i_k}$, we define its length by setting $|\alpha| = k$, and $|\alpha| = 0$ if $\alpha = g_0$. Analogously, we can define an element of the Fock space $e_\alpha = e_{i_1} \otimes e_{i_2} \otimes \cdots \otimes e_{i_k}$ and $e_{i_0} = 1$. Note that $\mathcal{B} = \{e_{g_i} : g_i \in \mathbb{F}_n\}$ is an orthonormal basis for

the Fock space F^2. Therefore, it is easy to see that the Fock space over $\mathbb{C}^n$ is isomorphic to the Hilbert space of square summable sequences indexed by $\mathbb{F}_n$.

Example 6.1.1. We consider the free monoid generated by two elements $\mathbb{F}_2$, where the generators are $g_1 = a$ and $g_2 = b$. Then, the word $\alpha = aba$ is an element of $\mathbb{F}_2$, with $\alpha = g_1 g_2 g_1$. The corresponding element in the Fock space $F^2(\mathcal{H}_2)$ is $e_\alpha = e_1 \otimes e_2 \otimes e_1$.

Note that we can define shift operators on the Fock space.

Definition 6.1.4. *The **left shift**, also called left creation operator, $S_i : F^2 \to F^2$ is the operator defined, for $i = 1, \ldots, d$, by:*

$$S_i(e_\alpha) := e_i \otimes e_\alpha = e_{i\alpha}. \tag{6.7}$$

*The **right shift**, or right creation operator, $R_i : F^2 \to F^2$ is the operator defined, for $i = 1, \ldots, d$, by:*

$$R_i(e_\alpha) := e_\alpha \otimes e_i = e_{\alpha i}. \tag{6.8}$$

Note that these noncommutative shift operators are row operators: $S = (S_1, \ldots, S_d)$, $R = (R_1, \ldots, R_d)$.

The Fock space can be also identified with $\mathcal{H}^2(\mathbb{F}_n)$, a canonical NC analogue of the Hardy space we defined on the unit disc. The elements of $\mathcal{H}^2(\mathbb{F}_n)$ are defined on the NC open unit row ball $\mathbb{B}_{\mathbb{N}}^d$. Given a collection of n formal NC variables $z = [z_1, \ldots, z_n]$, with $z^\alpha := z_{i_1} \cdot z_{i_2} \cdots z_{i_k}$, we can consider an element $f \in F^2$ and represent it as a formal power series:

$$f(z) = \sum_{\alpha \in \mathbb{F}_n} \widehat{f}_\alpha z^\alpha. \tag{6.9}$$

We define the **NC Hardy space** as:

$$\mathcal{H}^2(\mathbb{F}_n) = \left\{ \sum_{\alpha \in \mathbb{F}_n} \widehat{f}_\alpha z^\alpha : \sum_{\alpha \in \mathbb{F}_n} ||f_\alpha||^2 < \infty \right\}. \tag{6.10}$$

We remark that alternatively, we could have defined the shift operators in the NC Hardy space, interpreting them as multiplication for a noncommutative variable:

$$S_i = M_{z_i}^S \qquad R_i = M_{z_i}^R \tag{6.11}$$

where:

$$M_{z_i}^S f(z) = z_i f(z) \qquad M_{z_i}^R f(z) = f(z) z_i. \tag{6.12}$$

The convergence of a NC power series can be established by estimating its joint spectral radius.

Definition 6.1.5. *The **joint spectral radius** $\rho_{NC}(z)$ of a NC power series f, with $f(z) = \sum_{\alpha \in \mathbb{F}_n} \widehat{f}_\alpha z^\alpha$, is defined as:*

$$\rho_{NC}(z) = \limsup_{N \to \infty} \left[\sup_{\sum_{|\alpha|=N} |\widehat{f}_\alpha|^2 = 1} \left\| \sum_{|\alpha|=N} \widehat{f}_\alpha z^\alpha \right\|^{1/N} \right]. \tag{6.13}$$

Analogously, we could have defined the joint spectral radius for a d-tuple of $n \times n$ matrices $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_d]$ as follows:

$$\rho_{NC}(\mathbf{X}) = \lim \sqrt[2k]{(\mathbf{X}(\mathbf{1}_d \otimes \mathbf{1}_n)\mathbf{X}^*)^k}. \tag{6.14}$$

It is possible to show that $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_d]$ is a d-tuple of matrices with $\rho_{NC}(\mathbf{X}) < 1$ if and only if $\mathbf{X}$ is jointly similar to a strict row contraction in the NC row ball. Moreover, a NC power series converges whenever $\rho_{NC} < 1$. Note that, given a strict row contraction $z = [z_1, \ldots z_d]$ on a Hilbert space, the formal series defined in equation 6.9 converges absolutely in the operator norm when evaluated at z [Pop06a]. Therefore, any $f \in \mathcal{H}^2(\mathbb{F}_n)$ can be interpreted as a noncommutative function on strict row contractions in the NC open unit row ball [KVV14], and the NC Hardy space $\mathcal{H}^2(\mathbb{F}_n)$ is the Hilbert space of all analytic free NC functions on $\mathbb{B}_{\mathbb{N}}^d$ having square-summable Taylor coefficients [JMS21b]. We remark that

it is also possible to extend to the noncommutative case the definition of $\mathcal{H}^\infty$, the space of uniformly bounded analytic functions in the disk. In fact, in the NC setting, we can define $\mathcal{H}^\infty_{\mathrm{NC}}$ as the algebra of uniformly bounded free NC functions in $\mathbb{B}^d_{\mathbb{N}}$.

$$\mathcal{H}^\infty_{\mathrm{NC}} = \left\{ f \in \mathrm{Hol}(\mathbb{B}^d_{\mathbb{N}}) : \sum_{\alpha \in \mathbb{F}_n} \widehat{f}_\alpha z^\alpha, \ \sup_{z \in \mathbb{B}^d_{\mathbb{N}}} ||f(z)|| < \infty \right\}. \tag{6.15}$$

It is worth recalling that in the one-variable case, the space $\mathcal{H}^\infty$ can be alternatively viewed as the algebra of all functions which multiply $\mathcal{H}^2$ into itself (multiplier algebra). This means that given any $f \in \mathcal{H}^\infty$ we can consider a bounded multiplication operator $M_f : \mathcal{H}^2 \to \mathcal{H}^2$ such that $M_f g = fg$ for any $g \in \mathcal{H}^2$. M_f is called **multiplier**. This reasoning extends to the noncommutative setting, where we can similarly identify $\mathcal{H}^\infty_{\mathrm{NC}}$ with the algebra of left multipliers of $\mathcal{H}^2(\mathbb{F}_n)$, *i.e.* the weakly-closed algebra generated by the left creation operators on the full Fock space, and the identity [JMS21a, Pop06a, SSS18]. To avoid introducing too many symbols and to keep the exposition clear, we will mostly use the first interpretation of $\mathcal{H}^\infty_{\mathrm{NC}}$, and make explicit the use of the multipliers notation only when it is not clear from the context (mainly in Chapter 7).

6.1.2 NC Rational Functions

The free algebra of noncommutative polynomials over a field $\mathbb{K}$ admits a universal division ring of fractions [Coh95]. In particular, we have the following definition:

Definition 6.1.6. *A **NC rational expression** is any syntactically valid expressions involving several NC variables, scalars from $\mathbb{K}$, the operations $+$, $\cdot$, $^{-1}$ and the parentheses.*

An important challenge in defining rational functions is to decide if two rational expressions represent the same functions. There are several approaches to deal with this challenge, but the most common approach is to define NC rational functions in the following way.

Definition 6.1.7. *A **NC rational function** is an equivalence class between rational expressions, where we say that r_1 and r_2 belong to the same equivalence class if r_1 can be*

transformed into r_2 by algebraic manipulations.

Note that dividing for zero is not a valid operation, and deciding whether or not a given rational expression (or rational function) is zero might not always be obvious (for example, when dealing with nested inversions).

Definition 6.1.8. *The **domain** of a rational expression r is the set of all the tuple of matrices X for which $r(z)$ is defined. The domain of a NC rational function is the union of the domains of its rational expressions.*

We remark that Definition 6.1.7 is equivalent to requiring that two rational expressions r_1 and r_2 give the same result when evaluated over matrices of arbitrary size belonging to the intersection of their domains [HMS17]. In this thesis we are concerned with applying rational functions to operators belongings to some fixed algebra. It is not trivial that being zero when evaluated on the division ring of NC rational functions directly implies also being zero when applied to operators. For example, we could have different rational expressions representing the same rational functions that don't agree when evaluated over the same operator. Thankfully, in this thesis we will only work with the weakly-closed algebra generated by the left creation operators over the Fock space. In this setting, we don't have to worry about this issue, and we can assume that this operation is well defined [HMS17, Coh95].

It is also important to remark that, unlike the commutative case, a NC rational function does not admit a canonical coprime fraction representation [KVV09]. Therefore, it is of paramount importance to find a "canonical" way to represent NC rational functions. A subset of NC rational functions particularly relevant in our analysis is the class or **regular** rational functions, *i.e.* functions that contains zero in their domain. An important property of rational functions in the classical (commutative) Hardy space H^2 is that if a rational function r is regular near zero and can be represented as power series having square-summable coefficients, then it cannot have any poles in the closed unit disk, and is therefore regular in a disk of radius greater than one. These facts extend naturally when the NC variable

is defined on the multivariable row ball, but do not generalize to the NC polydisk [Sch61, Coh95, JMS21a]. Interestingly, the main tools used to study regular NC rational functions come from control theory and from the theory of formal languages [Fli74, Ber79, Sch61]. In fact, it is a well known result from realization theory [KPV17] that every regular NC rational function can be represented in terms of matrices of linear polynomials, and admits a minimal realization of size n:

$$r(z) = \mathbf{c}^* \mathbf{L}_{\mathbf{A}}^{-1}(z)\mathbf{b} \tag{6.16}$$

where $\mathbf{A}_j$ are square matrices of size n, $\mathbf{b}$, $\mathbf{c}$ are vectors of size n and $\mathbf{L}_{\mathbf{A}}(z)$ is a linear pencil (or resolvent)

$$\mathbf{L}_{\mathbf{A}}(z) = \mathbf{1} - \sum \mathbf{A}_j z_j. \tag{6.17}$$

We have chosen this notation to have a more concise representation of the rational function, but to be precise these products are actually Kronecker tensor products. If we unpack this notation, we have that for a d-tuple of variable $z = [z_1, \dots z_d]$, where we can assume that each z_j is a $N \times N$ matrix, the rational function $r(z)$ can be represented in the following way:

$$r(z) = \mathbf{c}^* \otimes \mathbf{1}_N \left(\mathbf{1}_n \otimes \mathbf{1}_N - \sum \mathbf{A}_j \otimes z_j \right)^{-1} \mathbf{b} \otimes \mathbf{1}_N. \tag{6.18}$$

Interestingly, the minimal realization has the maximal domain of all the rational expressions that it represents [Vol18]. Note that we can use the formalism of minimal realizations to represent polynomials (in this case it is enough to choose $\mathbf{A}_j$ that are jointly nilpotent). We conclude this section with a result from Jury et al. characterizing rational functions in the NC Hardy space [JMS21a].

Theorem 6.1.2 ([JMS21a]). *Let $\mathbf{A} = [\mathbf{A}_1, \dots, \mathbf{A}_d]$, and let r be a NC rational function with minimal realization:*

$$r(z) = \mathbf{c}^* \left(\mathbf{1} - \mathbf{A}z \right)^{-1} \mathbf{b}. \tag{6.19}$$

Then $r \in \mathcal{H}^2(\mathbb{F}_n)$ if and only if $\rho_{NC}(\mathbf{A}) < 1$.

6.1.3 NC Hankel Operators

We recall Popescu's definition of NC Hankel operator [Pop03]. Let $X = [X_1, \dots, X_n]$, $X_i \in B(\mathcal{Y})$ be an arbitrary sequence of bounded operators on a Hilbert space $\mathcal{Y}$, and let $T = [T_1, \dots, T_n]$, $T_i \in B(\mathcal{H})$. Suppose to have an orthogonal decomposition $\mathcal{H} = \mathcal{H}_- \oplus \mathcal{H}_+$ such that $\mathcal{H}_+$ is invariant with respect to each $T_i \in B(\mathcal{H})$, for $i = 1, \dots n$. We denote with $\mathbb{P}_-$ and $\mathbb{P}_+$ the orthogonal projections on $\mathcal{H}_-$ and $\mathcal{H}_+$ respectively.

Definition 6.1.9. *A **NC Hankel operator** is a bounded linear operator $\Gamma : \mathcal{Y} \to \mathcal{H}_-$ such that:*

$$\Gamma X_i = \mathbb{P}_- T_i \Gamma \qquad \text{for any } i = 1, \dots, n. \tag{6.20}$$

Similarly to what seen in Section 2.4.5, the definition of symbol can be generalized using the notion of multiplier.

Definition 6.1.10. *A **multiplier** is bounded linear operator $A : \mathcal{Y} \to \mathcal{H}$ such that:*

$$AX_i = T_i A \quad \text{for any } i = 1, \dots, n. \tag{6.21}$$

Following Popescu [Pop03] we remark that, given a multiplier, it is always possible to associate with it a Hankel operator defined as:

$$\Gamma_A y = \mathbb{P}_- A y \quad \text{for } y \in \mathcal{Y}, \tag{6.22}$$

and $||\Gamma_A|| \leq ||A||$. In fact it is easy to see that:

$$\begin{aligned} \Gamma_A X_i y &= \mathbb{P}_- A X_i y \\ &= \mathbb{P}_- T_i \mathbb{P}_- A y + \mathbb{P}_- T_i \mathbb{P}_+ A y \\ &= \mathbb{P}_- T_i \mathbb{P}_- A y \\ &= \mathbb{P}_- T_i \Gamma y. \end{aligned}$$

The converse of this statement is the subject of the multivariable Nehari theorem.

Theorem 6.1.3 (NC Nehari Theorem [Pop03]). *Let $X = [X_1, \dots, X_n]$, $X_i \in B(\mathcal{Y})$ and $T = [T_1, \dots, T_n]$, with $T_i \in B(\mathcal{H})$, be such that:*

$$\|X_1 y_1 + \cdots + X_n y_n\|^2 \geq \|y_1\|^2 + \cdots + \|y_n\|^2, \quad y_i \in \mathcal{Y} \tag{6.23}$$

$$\|T_1 h_1 + \cdots + T_n h_n\|^2 \leq \|h_1\|^2 + \cdots + \|h_n\|^2, \quad h_i \in \mathcal{H}. \tag{6.24}$$

Then, given a generalized Hankel operator $\Gamma_A : \mathcal{Y} \to \mathcal{H}_-$, with $\Gamma X_i = \mathbb{P}_- T_i \Gamma$ for any $i = 1, \dots, n$, there exists a multiplier $A : \mathcal{Y} \to \mathcal{H}$ such that $\Gamma = \Gamma_A$ and $||\Gamma|| = ||A||$.

The following corollary is a direct consequence of Nehari's Theorem.

Corollary 6.1.3.1 ([Pop03]). *Let $\Gamma_A : \mathcal{Y} \to \mathcal{H}_-$ be a NC Hankel operator, then:*

$$\|\Gamma_A\| = \inf\{\|A - F\| : F : \mathcal{Y} \to \mathcal{H} \text{ is a multiplier with } F(\mathcal{Y}) \subset \mathcal{H}_+\}. \tag{6.25}$$

Moreover, there exists an optimal multiplier F^* such that $\|\Gamma_A\| = \|A - F^*\|$.

6.1.4 NC AAK Theorem

In this section we state a NC version of the AKK theorem. In particular, we consider the theorem obtained by Popescu [Pop03], which generalizes Theorem 2.4.9 from Treil and Volberg [TV94]. Note that, unlike Theorem 2.4.6, this version of the theorem is not constructive.

Let $B \in B(\mathcal{Y})$ be a self-adjoint operator on a Hilbert space $\mathcal{Y}$, and let $\mathcal{P}_+$ and $\mathcal{P}_-$ be the orthogonal projections onto the non negative and strictly negative spectral subspaces. We denote $\mathcal{Y}^\mp = \mathcal{P}_\mp \mathcal{Y}$. It is possible to show that if $\dim \mathcal{Y}^- < \infty$, then it is equal to the number of negative eigenvalues of B, counted with their multiplicities. We assume that the operator $B_- := \mathcal{P}_A | \mathcal{Y}^-$ is invertible. Let

$$\mathcal{K}_+ = \{\mathbf{v} \in \mathcal{H} : (B\mathbf{v}, \mathbf{v}) \geq 0\} \tag{6.26}$$

be the cone of B-nonnegative vectors. We recall from [Pop03] the multivariable version Theorem 2.4.8 from Iokhvidov and Fan [Iok64].

Theorem 6.1.4 ([Pop03]). *Let $B \in B(\mathcal{Y})$ be a self-adjoint operator on a Hilbert space $\mathcal{Y}$, with B_- be invertible. Let $X_i \in B(\mathcal{Y})$, for $i = 1, dots, n$ be such that:*

$$X_1\mathcal{K}_+ + \cdots + X_n\mathcal{K}_+ \subset \mathcal{K}_+ \tag{6.27}$$

and $\mathcal{P}_+X_i\mathcal{P}_-$ is a compact operator. Then there exists a maximal subspace $\mathcal{M}$ of $\mathcal{K}_+$ which is maximal (by inclusion) and X_i-invariant.

We recall that, analogously to what done in Section 2.4.5, the singular numbers of a Hankel operator $\Gamma_A : \mathcal{Y} \to \mathcal{H}_-$ can be characterized in terms of subspaces $\mathcal{M} \subset \mathcal{Y}$ in the following way:

$$\sigma_n(\Gamma) = \inf\{\|\Gamma|_{\mathcal{M}}\| : \mathrm{codim}\mathcal{M} \leq n\}.$$

Applying Theorem 6.1.4 to the operator $B = \sigma_n(\Gamma)1 - \Gamma^*\Gamma$, $B \in B(\mathcal{Y})$, and to the cone of its nonnegative vectors

$$\mathcal{K}_+ = \{\mathbf{v} \in \mathcal{Y} : (B\mathbf{v}, \mathbf{v}) \geq 0\} = \{\mathbf{v} \in \mathcal{Y} : ||\Gamma\mathbf{v}|| \leq \sigma_n\mathbf{v}\} \tag{6.28}$$

we obtain the following version of the AAK theorem:

Theorem 6.1.5 (NC AAK Theorem [Pop03]). *Let $X = [X_1, \ldots, X_n]$, $X_i \in B(\mathcal{Y})$, and let $T = [T_1, \ldots, T_n]$, $T_i \in B(\mathcal{H})$, be such that:*

$$\|X_1y_1 + \cdots + X_ny_n\|^2 \geq \|y_1\|^2 + \cdots + \|y_n\|^2, \quad y_i \in \mathcal{Y} \tag{6.29}$$

$$\|T_1h_1 + \cdots + T_nh_n\|^2 \leq \|h_1\|^2 + \cdots + \|h_n\|^2, \quad h_i \in \mathcal{H}. \tag{6.30}$$

Let $\Gamma_A : \mathcal{Y} \to \mathcal{H}_-$ be a NC Hankel operator, with $\Gamma X_i = \mathbb{P}_-T_i\Gamma$ for any $i = 1, \ldots, n$. Let

$\{\sigma_i\}_{i\geq 0}$ be the sequence of its singular numbers. Then:

$$\sigma_n(\Gamma) = \inf\{\|\Gamma|_{\mathcal{M}}\| : \mathcal{M} \subset \mathcal{Y},\ codim\mathcal{M} \leq n,\ X_i\mathcal{M} \subset \mathcal{M}\}$$

and the infimum is attained.

6.2 A Framework for Multi-Letter Alphabets

In this section, we consider alphabets Σ with the property that $|\Sigma| = d$, with $d > 1$. In this case, the set of strings Σ^* can be identified with $\mathbb{F}_d$, the free monoid generated by d elements. Therefore, a WFA $A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}, \boldsymbol{\beta} \rangle$ over a multi-letter alphabet Σ can be seen as computing a function $f_A : \mathbb{F}_d \to \mathbb{R}$. Note that when $d > 1$, $\mathbb{F}_d$ has a noncommutative structure, so it cannot be embedded into $\mathbb{Z}$. Concretely, this means that the approximate minimization problem cannot be addressed using AAK theory on Hardy spaces. To extend these results, it is thus necessary to find a way to adapt the standard methods from harmonic analysis to the nonabelian setting. As noted in Section 6.1.1, a recent line of work in multivariable operator theory has been centered around extending the results of standard operator theory to the case of noncommutative operators defined on Fock spaces [Fra82, Bun84, Pop93, AP95, Pop10, Pop06a, Pop89c, Pop95b, Pop03, Pop89b, Pop13, Pop92]. In particular, a NC version of the AAK theorem is presented in a recent work of Popescu [Pop03], but its proof is not constructive. Therefore, solving the approximate minimization problem for multi-letter alphabets using AAK theory comes with two distinct challenges:

- *Finding a NC Hankel operator:* given a WFA and its Hankel matrix, we need to find a way to reformulate the approximation problem using multivariable operators. In particular, we need to find a noncommutative analogue of Hardy spaces, the shift operator, and of the symbol.

- *Making AAK constructive again:* the proof of the noncommutative version of the AAK

theorem does not provide us with an expression for the optimal approximation. Ideally, we would like to rework the proof so that it becomes constructive.

In the next sections we analyze the first point, while the second will be the focus of the next chapter.

6.2.1 From Hankel Matrix to Hankel Operator

We consider a model over a multi-letter alphabet Σ, with $|\Sigma| = d$, and the Hankel matrix associated with it. As noted in the previous section, the model is computing a function $f : \mathbb{F}_d \to \mathbb{R}$. It is important to notice that a function of this type can be interpreted as an element in the Fock space F^2. We use a running example to better illustrate the properties and connections between the approximation problem and multivariable operator theory.

Example 6.2.1. We consider an alphabet $\Sigma = \{a, b\}$ with two letters, and we denote with ε the empty string. The set of strings Σ^* can be associated with the free monoid generated by two elements $\mathbb{F}_2$, where the generators are $g_1 = a$ and $g_2 = b$. As shown in Example 6.1.1, a word $\alpha = aba$ can be seen as an element in $\mathbb{F}_2$, with $\alpha = aba = g_1 g_2 g_1$, and the corresponding element in the Fock space F^2 is $e_\alpha = e_1 \otimes e_2 \otimes e_1$. Now, we can consider a function:

$$\begin{aligned} f : \Sigma^* &\to \mathbb{R} \\ \alpha &\mapsto f(\alpha) \end{aligned}$$

This function can be viewed either as an element in the Fock space F^2, using a sequence interpretation:

$$(f(\varepsilon), f(a), f(b), f(aa), f(ab), f(ba), f(bb), f(aaa), \dots) \in F^2 = \bigoplus_{k \geq 0} (\mathbb{R}^2)^{\otimes k}, \tag{6.31}$$

Classical Hankel op H	Generalized Hankel op Γ	NC Hankel op Γ
$H : \mathcal{H}^2 \to \mathcal{H}^2_-$	$\Gamma : \mathcal{H}_1 \to \mathcal{H}_2^-$	$\Gamma : \mathcal{Y} \to \mathcal{H}_-$
$HS = \mathbb{P}_- SH$	$\Gamma S_1 = \mathbb{P}_- S_2 \Gamma$	$\Gamma X_i = \mathbb{P}_- T_i \Gamma$
$S : \mathcal{H}^2 \to \mathcal{H}^2$, $S(f) = zf$	$S_1 : \mathcal{H}_1 \to \mathcal{H}_1$	$X = [X_1, \dots, X_n]$, $X_i \in B(\mathcal{Y})$
$S : \mathcal{H}^2 \to \mathcal{H}^2$, $S(f) = zf$	$S_2 : \mathcal{H}_2 \to \mathcal{H}_2$	$T = [T_1, \dots, T_n]$, $T_i \in B(\mathcal{H})$
S isometry	S_1 exp., S_2 contract.	X exp., T contract.
$\mathcal{H}^2$ Hardy space	$\mathcal{H}_1$	$\mathcal{Y} = ?$
$\mathcal{H}^2_-$ Neg Hardy space	$\mathcal{H}_2^-$	$\mathcal{H}_- = ?$
$\mathcal{L}^2(\mathbb{T}) = \mathcal{H}^2 \oplus \mathcal{H}^2_-$	$\mathcal{H}_2 = \mathcal{H}_2^+ \oplus \mathcal{H}_2^-$	?

Table 6.1: Comparison between classical, generalized and NC Hankel operators

or as a power series in the NC Hardy space $\mathcal{H}^2(\mathbb{F}_2)$, using a functional interpretation:

$$f(\varepsilon) + f(a)z_1 + f(b)z_2 + f(aa)z_1^2 + f(ab)z_1z_2 + f(ba)z_2z_1 + \cdots = \sum_{\alpha \in \Sigma^*} f(\alpha) z^\alpha. \qquad (6.32)$$

In fact, we recall that the Fock space is isomorphic to the Hilbert space of square summable sequences indexed by Σ^*, which is in turn isomorphic to the set of functions $f : \Sigma^* \to \mathbb{R}$.

Our objective is to leverage the correspondence between a function computed by a model and its Fock space interpretation, in order to solve the approximate minimization problem using the theory developed in Section 6.1.1. In particular, we want to adapt Definition 6.1.9, the definition of NC Hankel operator introduced by Popescu [Pop03], to the language modelling setting.

As a first step, we report in table 6.1 a comparison between the different classes of Hankel operators introduced in Chapter 2. In the first column, we have the classical Hankel operator defined between Hardy spaces, and the shift corresponds to the multiplication by the complex variable z (see Definition 2.4.8). In the second column, we describe the generalized Hankel operator from Treil and Volberg [TV94]. The main difference in this definition is that the "shift" considered does not have to be an isometry, but is substituted by a contractive and an expansive operator (see Definition 2.4.11). Finally, in the last column we recall the definition of NC Hankel operator from Popescu [Pop03], which was introduced

in Section 6.1.3 (see Definition 6.1.9). In this case, instead of simple operators, we consider row operators. Similarly to the generalized case, the shift is replaced by two operators that are not necessarily isometries.

To apply multivariable operator theory (and the NC AAK theorem) to the approximate minimization problem, we need to understand how to reformulate Equation 2.24 (the Hankel equation), and to find suitable transformations that will play the roles of the row operators T and X. Moreover, we need to choose appropriate generalizations of the Hardy spaces. Specifically, we need to find a noncommutative analogue of the function space $\mathcal{L}^2(\mathbb{T})$: a space containing $\mathcal{H}_-$, and in which the noncommutative equivalent of the backward shift is defined.

We start by considering the shift operators defined on the Fock space F^2. We recall that $\mathcal{B} = \{e_{g_i} : g_i \in \mathbb{F}_n\}$ is an orthonormal basis for F^2, where $e_\alpha = e_{i_1} \otimes e_{i_2} \otimes \cdots \otimes e_{i_k}$ if $\alpha = g_{i_1} g_{i_2} \cdots g_{i_k}$, and $e_{i_0} = 1$. The left shift $S_i : F^2 \to F^2$ is defined, for $i = 1, \ldots, d$, by:

$$S_i(e_\alpha) := e_i \otimes e_\alpha = e_{i\alpha}, \tag{6.33}$$

while the right shift $R_i : F^2 \to F^2$ is defined, for $i = 1, \ldots, d$, by:

$$R_i(e_\alpha) := e_\alpha \otimes e_i = e_{\alpha i}. \tag{6.34}$$

To avoid using different notations for the same operator, we will always denote the right and left shifts as R and S, respectively, independently on whether they are defined on F^2 or on the NC Hardy space. Note that the NC shifts are row operators: $S = (S_1, \ldots, S_d)$, $R = (R_1, \ldots, R_d)$.

Example 6.2.2. We continue with the setting of Example 6.2.1. In this case, the right shift is defined as $S = (S_1, S_2)$, with:

$$S_1(e_\alpha) = e_{a\alpha} \tag{6.35}$$

and

$$S_2(e_\alpha) = e_{b\alpha}. \tag{6.36}$$

The adjoint of S is defined as:

$$S_1^*(e_\alpha) = \begin{cases} e_{\alpha'} & \text{if } \alpha = a\alpha' \\ 0 & \text{otherwise.} \end{cases} \tag{6.37}$$

The right shift and its adjoint can be defined in a similar way.

It is possible to express the right shift in terms of the left one by using a unitary operator, the *flipping operator* U:

$$R_i = U^* S_i U \tag{6.38}$$

with

$$U(e_{i_1} \otimes e_{i_2} \otimes \cdots \otimes e_{i_k}) = e_{i_k} \otimes \cdots \otimes e_{i_2} \otimes e_{i_1}. \tag{6.39}$$

We use the left and right shifts to rewrite Equation 2.10 in the case of multi-letter alphabets.

Theorem 6.2.3. *Let $|\Sigma| = d$, and let $\mathbf{H}$ be the Hankel matrix associated to a function $f : \Sigma^* \to \mathbb{R}$,* i.e. *$\mathbf{H}(i,j) = f(ij)$. Let $S = (S_1, \ldots, S_d)$, $R = (R_1, \ldots, R_d)$ be the left and right shifts on F^2, S^* and R^* their adjoints. Then, the following Hankel equation holds:*

$$\mathbf{H}S_i = R_i^*\mathbf{H} \qquad \text{for } i = 1, \ldots, d. \tag{6.40}$$

Proof. For this proof, we leverage the functional representation of the Fock space. We recall that in the NC Hardy space, the left shift is equivalent to left multiplication by one of the noncommutative variables: $S_i f = z_i f$. Moreover, the function $f : \Sigma^* \to \mathbb{R}$ associated to the Hankel matrix can be represented by means of a formal power series in the NC Hardy space:

$$f = \sum_{\alpha \in \Sigma^*} f(\alpha) z^\alpha. \tag{6.41}$$

Furthermore, this function corresponds to the first column of the Hankel matrix. Analogously, it is easy to see that the column at index α is:

$$\mathbf{H}e_\alpha = \sum_{\beta \in \Sigma^*} f(\beta\alpha) z^\beta. \tag{6.42}$$

Therefore:

$$\mathbf{H}S_i(e_\alpha) = \sum_{\beta \in \Sigma^*} f(\beta i\alpha) z^\beta. \tag{6.43}$$

On the other hand, we can consider the adjoint of the right shift:

$$R_i^* \mathbf{H}e_\alpha = R_i^* \sum_{\beta \in \Sigma^*} f(\beta\alpha) z^\beta. \tag{6.44}$$

In this case, after applying R_i^* the terms in the series for which $\beta \neq \beta' i$ for any β', become zero. We obtain:

$$R_i^* \mathbf{H}e_\alpha = \sum_{\beta' \in \Sigma^*} f(\beta' i\alpha) z^{\beta'}, \tag{6.45}$$

Thus, for any $i = 1, \ldots, d$, we have the equation:

$$\mathbf{H}S_i = R_i^* \mathbf{H} \tag{6.46}$$

which concludes the proof. □

We use the recurring example to better illustrate the theorem above and the new Hankel equation.

Example 6.2.4. We consider the Hankel matrix $\mathbf{H}$ associated to the rational function f computed by a WFA $A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}, \boldsymbol{\beta} \rangle$ defined over an alphabet $\Sigma = \{a, b\}$. We have the

following bi-infinite matrix:

$$\mathbf{H} = \begin{pmatrix} f(\varepsilon) & f(a) & \dots & f(ba) & \dots & f(aba) & \dots \\ f(a) & f(aa) & \dots & f(aba) & \dots & f(aaba) & \dots \\ f(b) & f(ba) & \dots & f(bba) & \dots & f(baba) & \dots \\ f(aa) & f(aaa) & \dots & f(aaba) & \dots & f(aaaba) & \dots \\ f(ab) & f(aba) & \dots & f(abba) & \dots & f(ababa) & \dots \\ f(ba) & f(baa) & \dots & f(baba) & \dots & f(baaba) & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}. \tag{6.47}$$

It is easy to see that:

$$\begin{aligned} \mathbf{H}S_a(e_{ba}) &= \mathbf{H}e_{aba} \\ &= \sum_{\beta\in\Sigma^*} f(\beta aba)z^\beta \\ &= f(aba)z_0 + f(aaba)z_1 + f(baba)z_2 + \dots \end{aligned} \tag{6.48}$$

On the other hand, if we consider the adjoint of the right shift, we have the following relation:

$$\begin{aligned} R_a^*\mathbf{H}(e_{ba}) = R_a^* \sum_{\beta\in\Sigma^*} f(\beta ba)z^\beta &= \sum_{\beta'\in\Sigma^*} f(\beta' aba)z^{\beta'} \\ &= f(aba)z_0 + f(aaba)z_1 + f(baba)z_2 + \dots . \end{aligned} \tag{6.49}$$

Following this reasoning, we can obtain the same results for S_b and R_b^*, so the Hankel equation holds.

Equation 6.40 can be rewritten using only the left shift by leveraging the flipping operator U. In fact, by setting $\mathbf{J} = U\mathbf{H}$, we obtain:

$$\mathbf{J}S_\alpha = S_\alpha^*\mathbf{J} \qquad \forall\alpha\in\Sigma. \tag{6.50}$$

Now that we have a noncommutative version of the Hankel equation, we need to find appropriate definitions for the spaces $\mathcal{Y}$, $\mathcal{H}_-$, and $\mathcal{H} = \mathcal{H}_- \oplus \mathcal{H}_+$. To achieve this, we draw inspiration from the one-letter setting. It has become clear from the examples above and from the previous theorem that, for a sequence-to-sequence interpretation of the Hankel matrix, the natural noncommutative generalization of $\ell^2(\mathbb{N})$ is the Fock space F^2. For the functional interpretation we consider instead the NC Hardy space. Therefore, we propose to set $\mathcal{Y} = \mathcal{H}_+ = F^2$ (or $\mathcal{Y} = \mathcal{H}_+ = \mathcal{H}^2(\mathbb{F}_d)$ in the functional interpretation). For the definition of $\mathcal{H}$, we decided to set $\mathcal{H} = F_0^2 \oplus F^2$. To have a better intuition of the reasoning behind this definition, we can once again resort to the one-letter case and to the analogy with the Hardy spaces. In the one-letter case, the role of $\mathcal{H}$ was played by $\mathcal{L}^2(\mathbb{T}) \cong \ell^2(\mathbb{Z})$. In this case, we represented each function f in $\mathcal{L}^2(\mathbb{T})$ using the bi-infinite sequence of its Fourier coefficients, and $\mathcal{L}^2(\mathbb{T}) = \mathcal{H}^2_- \oplus \mathcal{H}^2$. In particular, the elements of $\mathcal{H}^2_-$ are functions with only negative Fourier coefficients, so they can be seen as sequences indexed by negative powers of the complex variable z, while elements of $\mathcal{H}^2$ correspond to sequences indexed by nonnegative powers of z. Therefore, a function $f \in \mathcal{L}^2(\mathbb{T})$ can be represented as:

$$\begin{array}{cccccccc} (\dots, & \widehat{f}(-2), & \widehat{f}(-1), & \widehat{f}(0), & \widehat{f}(1), & \widehat{f}(2), & \dots) \\ \dots & z^{-2} & z^{-1} & z^0 & z^1 & z^2 & \dots \end{array}$$

Analogously, we can think of $\mathcal{H} = F_0^2 \oplus F^2$ as a set of bi-infinite sequences that are indexed by negative powers (the F_0^2 - component), and nonnegative powers of the NC variables $z_1, \dots, z_d$ (the F^2-component). We can see below an example of the indexing in the setting of Example 6.2.1:

$$\begin{array}{cccccccccc} (\dots, & f(a^{-2}), & f(b^{-1}), & f(a^{-1}), & f(\varepsilon), & f(a), & f(b), & f(aa), & f(ab), & \dots) \\ \dots & z_1^{-2} & z_2^{-1} & z_1^{-1} & z_1^0 z_2^0 & z_1^1 & z_2^1 & z_1^2 & z_1^1 z_2^1 & \dots \end{array}$$

It is important to remark that the correspondence between $\mathcal{L}^2(\mathbb{T}) = \mathcal{H}^2_- \oplus \mathcal{H}^2$ and $\mathcal{H} = F_0^2 \oplus F^2$ is not perfect. While it is certainly helpful to think about the indexing in terms of negative and nonnegative exponents, the functional interpretation of the Fourier

expansion is lost in the noncommutative case, where the function cannot be easily analyzed at the boundary of its domain.

To simplify the exposition, we start by considering the sequence interpretation. As noted before, the Fock space and the NC Hardy space are isomorphic, so the results obtained in one setting can be easily transposed to the other one.

Theorem 6.2.5. *Let $S = (S_1, \dots, S_d)$, $R = (R_1, \dots, R_d)$ be the left and right shifts on F^2, S^* and R^* their adjoints. Let $\mathcal{Y} = F^2$, $\mathcal{H} = F_0^2 \oplus F^2$, where $F_0^2 = \bigoplus_{k>0} (\mathbb{R}^d)^{\otimes k}$. We set $\mathcal{H}_- = F_0^2$ and $\mathcal{H}_+ = F^2$, and we define, for $i = 1, \dots, d$, a bilateral shift on $\mathcal{H}$:*

$$\begin{cases} \overline{R}_i(e_\alpha) = R_i^*(e_\alpha) & \text{for } e_\alpha \in \mathcal{H}_- \\ \overline{R}_i(e_\alpha) = R_i(e_\alpha) & \text{for } e_\alpha \in \mathcal{H}_+ \end{cases} . \tag{6.51}$$

Let $\mathbb{P}_-$ be the orthogonal projection on $\mathcal{H}_-$.

Then, the operator $H : \mathcal{Y} \to \mathcal{H}_-$ defined by the following property:

$$HS_i = \mathbb{P}_- \overline{R}_i H \qquad \text{for any } i = 1, \dots, n \tag{6.52}$$

is a NC Hankel operator according to Definition 6.1.9

Proof. In order to prove the theorem we need to verify the following two properties:

(a) $\mathcal{H}_- \subset \mathcal{H}$

(b) If $\mathcal{H} = \mathcal{H}_- \oplus \mathcal{H}_+$, then $\mathcal{H}_+$ is invariant under each $\overline{R}_i$.

In particular, we want to show that these properties are satisfied when $\mathcal{Y} = F^2$, $\mathcal{H}_- = F_0^2$ and $\mathcal{H} = F_0^2 \oplus F^2$.

(a) The first property follows directly from the definition of $\mathcal{H}$: $F_0^2 \subset F_0^2 \oplus F^2$.

(b) Since $\mathcal{H}_- = F_0^2$, it follows by definition that $\mathcal{H}_+ = F^2$.

We want to show that:

$$\overline{R}_1\mathcal{H}_+ + \cdots + \overline{R}_d\mathcal{H}_+ \subseteq \mathcal{H}_+ \tag{6.53}$$

i.e. that for any $h_i \in \mathcal{H}_+$ we have $\overline{R}_1 h_1 + \cdots + \overline{R}_d h_d \in \mathcal{H}_+$. Since $\overline{R}_i(e_\alpha) = R_i(e_\alpha)$ for $e_\alpha \in F^2$, the condition above can be reformulated as:

$$R_i(e_{\alpha_1}) + \cdots + R_d(e_{\alpha_n}) \in F^2, \tag{6.54}$$

which holds by definition of R, since $R_i(e_\alpha) = e_{\alpha i} \in F^2$ for any α, and the linear combination of elements in F^2 is an element in F^2.

□

While the choice of the Fock space seems natural, the definition of $\mathcal{H}$ is arbitrary, and other spaces containing F^2 could have been employed. For example, one can think of the free group over d elements, or of Cohn's free field [Coh95]. The reasoning behind our choice will become clearer after the following theorem, which states that the row operators $\overline{R}$ and S respect the properties required to apply the noncommutative version of Nehari's Theorem (Theorem 6.1.3).

Theorem 6.2.6. *Let $H : \mathcal{Y} \to \mathcal{H}_-$, with $HS_i = \mathbb{P}_-\overline{R}_i H$, be the NC Hankel operator defined in the previous theorem. Then:*

(a) $\|S_1 y_1 + \cdots + S_d y_d\|^2 \geq \|y_1\|^2 + \cdots + \|y_d\|^2$ *for* $y_i \in \mathcal{Y}$

(b) $\left\|\overline{R}_1 h_1 + \cdots + \overline{R}_d h_d\right\|^2 \leq \|h_1\|^2 + \cdots + \|h_d\|^2$ *for* $h_i \in \mathcal{H}$

and the hypothesis of Theorem 6.1.3 are satisfied.

Proof. (a) We want to show that $\|S_1 y_1 + \cdots + S_d y_d\|^2 \geq \|y_1\|^2 + \cdots + \|y_d\|^2$ for $y_i \in \mathcal{Y}$ $(= F^2)$. Leveraging the fact that the shifts have pairwise orthogonal ranges, so

$S_i^* S_j = \mathbf{1}\delta_{i,j}$, and that each S_i is an isometry, we obtain:

$$\begin{aligned}\|S_1 y_1 + \cdots + S_d y_d\|^2 &= \langle S_1 y_1, S_1 y_1\rangle + \langle S_1 y_1, S_2 y_2\rangle + \cdots + \langle S_d y_d, S_d y_d\rangle \\ &= \langle S_1 y_1, S_1 y_1\rangle + \langle S_2 y_2, S_2 y_2\rangle + \cdots + \langle S_n y_n, S_n y_n\rangle \\ &= \|y_1\|^2 + \cdots + \|y_2\|^2.\end{aligned}$$

(b) We want to show that $\left\|\overline{R}_1 h_1 + \cdots + \overline{R}_d h_d\right\|^2 \leq \|h_1\|^2 + \cdots + \|h_d\|^2$ for any $h_i \in \mathcal{H}$ ($= F_0^2 \oplus F^2$). Similarly to the previous point, we have that the shifts have orthogonal ranges, so the result holds with the equality.

□

The Free Group It is now easy to see why the free group is not a suitable choice for our setting. We denote with $\mathbb{F}_d^*$ the free group on d elements, and with $\ell(\mathbb{F}_d^*)$ the set of sequences indexed by elements in the free group. By setting $\mathcal{H} = \ell(\mathbb{F}_d^*)$, the conditions of Theorem 6.2.5 are satisfied, but the ones of Theorem 6.2.6 are not. It is easy to see that $\mathcal{H}_- \subset \mathcal{H}$, $\mathcal{H}_+$ is invariant under the bilateral shift, and that property (a) of Theorem 6.2.6 is satisfied. On the other hand, property (b) does not hold anymore. The components of the bilateral shifts don't have orthogonal ranges, as $\overline{R}_i^* \overline{R}_j \in \mathcal{H}$ even when $i \neq j$. Intuitively the space is "too big" for property (b) to hold. If we consider the intuition provided earlier in the section, about indexing the elements of $\mathcal{H} = F_0^2 \oplus F^2$ using negative and nonnegative exponents, we have that in the case of the free group any combination of positive and negative exponents is allowed. Therefore, when defined on $\ell(\mathbb{F}_d^*)$, the adjoint of the shift is:

$$\overline{R}_i^*(e_\alpha) = \begin{cases} e_{\alpha'} & \text{if } \alpha = \alpha' i \\ e_{\alpha i^{-1}} & \text{otherwise.} \end{cases} \tag{6.55}$$

6.2.2 NC Symbols and NC Rational Functions

In Section 6.1.3, we have seen that the multiplier (see Definition 6.1.10) plays, in the non-commutative case, a role similar to that of the symbol. On the other hand, unlike in the commutative case, the multiplier is an operator, not a function. Ideally, we would like to obtain a functional representation of the multiplier. To achieve this, we first analyze the multiplier and find that, with minimal manipulations, we can get a functional description of it. In particular, using the flipping operator U defined in Equation 6.39, we can rewrite the NC Hankel equation for the NC Hankel operator H as:

$$HS_a = U^* S_a^* UH, \tag{6.56}$$

or

$$UHS_a = S_a^* UH. \tag{6.57}$$

Analogously, the property satisfied by the multiplier becomes:

$$UAS_a = S_a UA. \tag{6.58}$$

An operator commuting with the shift is said to be NC S-analytic [Pop93]. NC S-analytic operators can be characterized by means of a function, called the **NC symbol**. The full details of this definition can be found in the work of Popescu [Pop95b]. In particular, a NC symbol ϕ for the operator UA respects the following property:

$$UAS_a v = S_a \theta v. \tag{6.59}$$

Concretely, this means that we can use the NC symbol θ to represent the operator UA (which is the multiplier up to a unitary transformation). By construction, the NC symbol θ corresponds to the multiplication by the first column of the matrix of UA (this statement

follows from [Pop93, Theorem 1.6]). It is now easy to see that this functional description is actually strictly related to the original Hankel operator. In fact, we can apply again the unitary transformation to obtain a function $U\theta$. Note that this function can be written as the sum of two components, $U\theta = \phi + c$, with $\phi \in H_0^2(\mathbb{F}_d)$ and $c \in H^2(\mathbb{F}_d)$ and, by construction, ϕ corresponds to the multiplication by the first column of $\mathbf{H}$. We refer to $U\theta$ as the **NC flipped symbol** of H.

Now that we have obtained a noncommutative generalization of the symbol and of Equation 7.1, it becomes interesting to investigate if we can derive an expression for the NC flipped symbol of the NC Hankel operator arising from a WFA. This would allow us to describe a given minimal WFA using a (noncommutative multivariable) complex function, in a way similar to what we did in Section 3.2.2 for the one-letter case. Let $A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}, \boldsymbol{\beta} \rangle$ be a minimal WFA computing a function f over a multi-letter alphabet, let $\mathbf{H}$ be its Hankel matrix and H the NC Hankel operator. The NC flipped symbol associated with H can be expressed using the entries of the first column of $\mathbf{H}$. Its series expression is:

$$\mathbb{P}_-(\phi + c) = \sum_{a \in \mathbb{F}_n} f(a) z^a \tag{6.60}$$

$$= \sum_{a \in \mathbb{F}_n} \boldsymbol{\alpha}^\top \mathbf{A}^a \boldsymbol{\beta} z^\alpha \tag{6.61}$$

$$= \boldsymbol{\alpha}^\top (\mathbf{1} - \sum \mathbf{A}_j z_j)^{-1} \boldsymbol{\beta} \tag{6.62}$$

We recall that the equation above can be unpacked using the Kronecker product as in Equation 6.18. Note that ϕ is a NC rational function: once again we have a tight connection between the rational functions studied by Fliess, Berstel and Reutenauer [Fli74, Ber79] and (NC) rational functions defined in (noncommutative multivariable) operator theory. This is very relevant, since in the one-letter case rewriting Equation 7.1 in terms of the WFA's parameters is the key step to algorithmically find the best approximation [BLP+21]. Therefore, assuming that there is a way to make the noncommutative proof constructive,

we have built the machinery necessary to attack the problem in the case of multi-letter alphabets.

Chapter 7

Tackling the Multi-Letter Case: Approaches and Obstacles

In the previous chapter, we proposed a way to associate a NC Hankel operator and a NC rational function to the Hankel matrix computed by a model on sequential data. In the one-letter case, this allowed us to reformulate the approximation problem in terms of functional analysis, and to solve it using AAK theory. Ideally, we would like to proceed in a similar way in the noncommutative case, and leverage the results in the previous chapter in order to apply techniques from noncommutative multivariable operator theory. In particular, we want to apply Theorem 6.1.5, the NC version of the AAK theorem proved by Popescu [Pop03]. As mentioned before, the main obstacle in this approach is that the proof of this theorem is not constructive. This means that, while we are guaranteed the existence of (at least) one optimal approximation, we do not have any information on how to construct it. This makes the multi-letter setting fundamentally different from the one-letter case, where the power of AAK theory lies in the possibility of easily calculating the best approximation.

We tried to extend this proof in several ways, but ultimately did not succeed. In this chapter, we summarize the two main approaches that we attempted in order to obtain the desired constructive proof:

1. We tried to make the constructive proof noncommutative. Looking at the standard constructive proof of the AAK theorem (which can be found in Appendix A.2), we analyzed the building blocks of the constructive proof, to understand whether or not they could be generalized to the noncommutative setting.

2. We tried a system-theory approach. Leveraging the NC rational function obtained in Section 6.2.2, and its expression in terms of the parameters of the WFA, we tried to proceed similarly to Chapter 4 and to derive a set of constraints necessary for the result to hold.

Both approaches come with a specific set of challenges, that cannot (yet) be completely overcome. We highlight the main obstacles we encountered, and provide possible directions for future work in the noncommutative setting.

In the first section, we briefly review the three main blocks of the proof of Theorem 2.4.6. Then, we analyze the role played by the symbol in this proof, and show how some of the results can be generalized to the noncommutative setting. Next, we provide a high-level overview of the key role played by the singular vectors in obtaining a description of the shift-invariant space which lies at the heart of the proof. In particular, we focus on the relation between shift-invariant spaces, inner functions and the optimal approximation. In the second section, we summarize an alternative possible approach (and its obstacles), based on the connections between multivariable operator theory and system theory. We conclude the chapter with a discussion and a summary of the questions that remain open.

7.1 Towards a Constructive Proof

We start with a brief overview of the three main steps in the proof of Theorem 2.4.6. A more detailed version of this proof can be found in Appendix A.2:

1. Let $\xi_k \in \ker(H_\phi^* H_\phi - \sigma_k^2 1)$, set $\psi = \phi - \frac{H_\phi \xi_k}{\xi_k}$.

2. Show that H_ψ is the optimal approximation: $\|H_\phi - H_\psi\| = \sigma_k$.

3. Show that H_ψ has the right size: $\text{rank}(H_\psi) \leq k$.

Following these three steps, we are able to construct the optimal approximation of a given Hankel operator. While the result itself is about operators, the proof has a more functional flavour. In particular, leveraging the relationship between a Hankel operator and its symbol, we can reframe this setting as the approximation problem of a bounded function on the unit circle (Theorem 2.4.7). This functional approach is also reflected in the last step of the proof, which relies heavily on the two main ingredients of Beurling's theorem: the shift-invariant space and the inner-outer factorization.

7.1.1 Symbols and Norm Inequalities

The first two steps of the constructive proof of the AAK theorem follow from properties of the symbol of a Hankel operator. In particular, we recall that, given a Hankel operator $H_\phi : \mathcal{H}^2 \to \mathcal{H}^2_-$, a symbol is a function $\phi \in \mathcal{L}^2(\mathbb{T}) = \mathcal{H}^2_- \oplus \mathcal{H}^2$ satisfying:

- $H_\phi f = \mathbb{P}_- \phi f$,
- $\|H_\phi\| = \inf\{\|\psi\|_\infty : \widehat{\psi}(m) = \widehat{\phi}(m),\ m < 0\}$, from which it follows: $\|H_\phi\| \leq \|\phi\|_\infty$.

In the proof of the AAK theorem, these properties are used to show the following set of inequalities:

$$\|H_\phi - H_\psi\| \leq \|\phi - \psi\|_\infty \leq \sigma_k. \tag{7.1}$$

In the noncommutative case, similar properties are satisfied by the multiplier. In particular, given a NC Hankel operator $H_A : \mathcal{H}^2(\mathbb{F}_d) \to \mathcal{H}^2_0(\mathbb{F}_d)$, with $H_A S_a = \mathbb{P}_- R_a H_A$, a multiplier $A : \mathcal{H}^2(\mathbb{F}_d) \to \mathcal{H}^2_0(\mathbb{F}_d) \oplus \mathcal{H}^2(\mathbb{F}_d)$ satisfies the following properties:

- $H_A y = \mathbb{P}_- A y$,
- $||H_A|| \leq ||A||$.

Note that an operator A is a multiplier for a NC Hankel operator, together with any other operator assuming the same values over $\mathcal{H}_0^2(\mathbb{F}_d)$. In Section 6.2.2 we have seen how we can find a functional representation for the multiplier in terms of the NC flipped symbol.

Now, we would like to find a noncommutative version of Equation 7.1. To do that, we first recall the following properties of NC S-analytic operators and their NC symbols.

Theorem 7.1.1 ([Pop93])**.** *Let T_θ be a NC S-analytic operator, and let θ be its NC symbol. The following properties hold:*

- *If $U\theta \in \mathcal{H}_{NC}^\infty$, then $\|T_\theta\| = \|U\theta\|_\infty$*
- *If $U\theta_1, U\theta_2 \in \mathcal{H}_{NC}^\infty$, then:*
  - $T_{\theta_1} + T_{\theta_2} = T_{\theta_1+\theta_2}$
  - $T_{\lambda\theta_1} = \lambda T_{\theta_1}$ *for* $\lambda \in \mathbb{C}$
  - $T_{\theta_1} T_{\theta_2} = T_{\theta_2 \otimes \theta_1}$.

Applying this theorem, we have that if H is a NC Hankel operator with multiplier A, and if θ is the NC symbol of UA:

$$\|UH\| \leq \|UA\| = \|U\theta\|_\infty. \tag{7.2}$$

Note that UH is also a NC Hankel operator, so it makes sense to search for its optimal approximation. We start by remarking the following property, where R is a bounded operator:

$$||UH - R|| = \langle UH - R, UH - R\rangle \tag{7.3}$$

$$= \langle U(H - U^*R), U(H - U^*R)\rangle \tag{7.4}$$

$$= \|H - U^*R\|. \tag{7.5}$$

Therefore, if we denote with UG the optimal approximation of UH, we have that $G = U^*UG$ is the optimal approximation of H. The following chain of inequalities directly generalizes Equation 7.1 in the noncommutative case:

$$\|UH - UG\| \leq \|UA - UB\| \leq \|\phi + c - \psi - d\|_\infty. \tag{7.6}$$

where H is the NC Hankel operator that we want to approximate, A is its multiplier and $\phi + c = U\theta$, where θ the NC symbol of UA.

Note that we have found a way to rephrase one of the key steps of the commutative proof of AAK theorem using the new noncommutative framework.

7.1.2 Shift-Invariant Spaces and Inner Functions

We now want to understand the fundamental step that makes the proof in the classical case constructive. Indeed, all three versions of the AAK theorem (classical, generalized and noncommutative) relies on the existence of a shift-invariant space. What makes the classical proof constructive is an interesting property that relates this shift-invariant space with the singular vectors. For simplicity, we assume that σ_k is a simple singular number, *i.e.* $\sigma_{k-1} > \sigma_k > \sigma_{k+1}$. Interestingly, any singular vector ξ_k corresponding to the k-th singular value σ_k has precisely k roots $\{z_i\}_{0 \leq i < k}$ (counted with their multiplicities) inside the unit disc [Cla68]. The proof of the classical AAK theorem shows that the rational symbol of the rank-k optimal approximation has poles located at $\{1/z_i\}_{0 \leq i < k}$. Since the location of the symbol's poles can be directly calculated using the singular vectors, it is relatively easy to build the optimal approximation. More specifically, the theorem is proven by analyzing the shift-invariant space defined by the functions generated by a σ_k-Schmidt pair. Therefore, it would be useful to investigate if a relation between shift-invariant spaces and the singular vectors still exists in the (commutative) generalized AAK theorem and the NC AAK theorem.

In particular, we can consider the proof of Theorem 2.4.9, the commutative generalized

AAK theorem. The theorem states that, given a generalized Hankel operator Γ with respect to the operators S_1 (expansive) and S_2 (contractive), we have:

$$\sigma_n(\Gamma) = \inf\{\|\Gamma|_{\mathcal{M}}\| : \mathcal{M}, \operatorname{codim}\mathcal{M} \leq n, S_1\mathcal{M} \subset \mathcal{M}\},$$

and the infimum is attained. In this case, a fixed point theorem is used to show the existence of a shift-invariant space $\mathcal{M}$ of codimension k. To understand the connection between this proof and the constructive one, it is enough to consider the shift-invariant space $\mathcal{N} = \overline{\operatorname{Span}}\{S^j f : j \geq 0\}$ generated by $f \in H^2$, where $\overline{\operatorname{Span}}$ denotes the topological closure of the linear span of the set and S is the unilateral shift on the Hardy space. If $\mathcal{N}$ has finite codimension k, using Beurling's theorem (Theorem 2.4.5) and the properties of inner functions, it is possible to prove that f has precisely k zeroes inside the unit disc. Moreover, $\mathcal{N}$ consists of those functions that share the zeroes of f. This means that $g \in \mathcal{N}$ if and only if any zero of f is also a zero of g, with greater or equal multiplicity. Now, if we reformulate Theorem 2.4.9 in the setting of classical Hankel operators, it is easy to see that:

$$\mathcal{M} = \overline{\operatorname{Span}}\{S^j \xi_k : j \geq 0\}, \tag{7.7}$$

where ξ_k is the singular vector corresponding to the k-th singular value σ_k. Being able to describe the shift-invariant space in terms of the singular vectors is fundamental in the constructive proof. However, when considering the proof in the generalized commutative case, it is not clear if this reasoning can be extended, or if $\mathcal{M}$ is determined at all by the zeroes of ξ_k. Andersson et al. [Car09] note that this statement cannot hold in its full generality, and provide a constructive version of Theorem 2.4.9 for a more restricted class of Hilbert spaces (to which the Dirichlet space belongs). Also in this case, while the inclusion $\overline{\operatorname{Span}}\{S^j \xi_k : j \geq 0\} \subset \mathcal{M}$ trivially holds, the reverse is still an open problem.

In the noncommutative case, we are in a somehow similar situation. The NC AAK theorem (Theorem 6.1.5) directly extends the generalized AAK theorem (Theorem 2.4.9),

and its proof also relies on a fixed-point argument. Therefore, it is unclear whether or not $\mathcal{M}$ is determined by the zeroes of ξ_k. Moreover, it is not obvious *a priori* that there is a noncommutative counterpart of the results related to Beurling's theorem and the inner-outer factorization (Theorem A.1.1) that are used to describe the shift-invariant space in the classical case.

To better understand how the definition of inner-outer factorization can be extended in the noncommutative case, we analyze in more details the last passage of the commutative proof. This step is concerned with proving that the rank of H_ψ (the candidate optimal approximation of the Hankel operator H_ϕ) is indeed k. By construction, $\ker(H_\phi^* H_\phi - \sigma_k^2 1) \subset \ker H_\psi$. Moreover, since $\ker H_\psi$ is a shift-invariant subspace, by Beurling's theorem we have that $\theta\mathcal{H}^2 \subset \ker H_\psi$, where θ is the greatest common divisor of the inner parts of non-zero functions $f \in \ker(H_\phi^* H_\phi - \sigma_k^2 1)$. From Theorem A.1.3 in Appendix A, it follows that $\text{rank} H_\psi \leq \dim(\theta\mathcal{H}^2)^\perp = \deg\theta$. This key passage in the proof ties the problem of establishing the rank of the operator to that of factorizing a complex function.

Interestingly, there is a noncommutative counterpart of most of the results described in the previous paragraph. In particular, Popescu provides a characterization of inner and outer NC functions and a description of shift-invariant spaces similar to that of Beurling's theorem [Pop93]. These results are a natural extension of the commutative case, expressed in terms of multipliers in the multiplier algebra $\mathcal{H}^\infty_{\text{NC}}$ (that we introduced at the end of Section 6.1.1). In fact, in the commutative case, we could have expressed the properties defining Blaschke, singular inner and outer functions in terms of multipliers. Given a function $f \in \mathcal{H}^\infty$, we say that f is inner if the associated multiplier M_f is an isometry, and outer if M_f has dense range. Moreover, an inner function θ is Blaschke if:

$$\theta\mathcal{H}^2 = S(\theta) = \left\{ f \in H^2 : \frac{f}{\theta} \in \text{Hol}(\mathbb{D}) \right\}. \tag{7.8}$$

Popescu's definition of inner and outer function is equivalent to that of the commutative

case: a NC function is inner if the corresponding NC left multiplier is an isometry, it is outer if the multiplier has dense range. Analogously, Popescu extends Beurling's theorem: a subspace $\mathcal{M} \subset \mathcal{H}^2(\mathbb{F}_n)$ is invariant for each S_α (for $\alpha \in \mathbb{F}_n$) if and only if there exists an inner multiplier Θ such that $\mathcal{M} = \Theta\mathcal{H}^2(\mathbb{F}_n)$ [Pop93]. Note that the range of a NC inner left multiplier is a subspace that is invariant with respect to the right shift [DP99]. In recent work, Jury, Martin and Shamovich [JMS21b] provide a noncommutative generalization of the definition of Blaschke product and of Theorem A.1.2. The definition of NC Blaschke left multiplier relies on the concepts of NC varieties and NC singularity spaces. In particular, given any $F \in \mathcal{H}^\infty_{\mathrm{NC}}$, the left NC variety of F is:

$$\mathrm{Sing}(F) = \bigsqcup_{n\geq 0} \mathrm{Sing}_n(F)\,; \quad \mathrm{Sing}_n(F) := \{(z, y) \in \mathbb{B}_n^d \times \mathbb{C}^n : y^* F(z) = 0\}. \tag{7.9}$$

The left NC singularity space of F is the shift-invariant space defined as:

$$\mathcal{S}(F) = \{h \in \mathcal{H}^2(\mathbb{F}_n) : y^* h(z) = 0\, \forall (z, y) \in \mathrm{Sing}(F)\}. \tag{7.10}$$

We say that a NC inner function $B \in \mathcal{H}^\infty_{\mathrm{NC}}$ is NC Blaschke if $\mathrm{Ran}(M_B) = \mathcal{S}(B)$, where $\mathcal{S}(B)$ is the singularity space of B. Therefore, like in the classical case, NC Blaschke inner functions are completely determined by their noncommutative zeros.

7.1.3 Challenges

Having managed to obtain a noncommutative counterpart of the definition of symbol, and that the concept of inner and outer functions and Blaschke product have already been generalized, is certainly encouraging. The main challenge with this approach arises when we try to establish the dimension of the shift-invariant subspace (the third and last step of the constructive proof). The result in the commutative case relies on counting the inner divisors of the inner function generating the shift-invariant subspace, and on Theorem A.1.3 (whose

proof follows from the definition of Blaschke factors and the relation between them and the reproducing kernels defined on the space). Interestingly, the Fock space is a reproducing kernel Hilbert space, and a connection can be established between reproducing kernels and NC rational functions (the set of reproducing kernels for $\mathcal{H}^2(\mathbb{F}_n)$ is equal to the set of rational functions on the same space [JMS21a]). However, we do not have yet a satisfying characterization of NC Blaschke factors, and this is a big obstacle towards establishing the noncommutative result in a way similar to the commutative case. In particular, the question of whether or not we can characterize NC Blaschke factors in terms of their minimal realizations is still an open question in the field of noncommutative multivariable operator theory [JMS21b]. It is important to consider that the problem of factoring polynomials is much harder in the NC setting, therefore some of the arguments that rely on polynomial factorization in the one-variable case are unlikely to succeed in the multi-variable setting.

7.2 An Alternative Approach from System Theory

The last approach that we consider is based on the solution we proposed in Chapter 4 to solve the approximate minimization problem for WFAs. The method is inspired by system theory, and the state-space approach used to solve the model reduction problem. An overview of these results can be found in the next chapter, in Section 8.3. The connection between discrete systems and operator theory is discussed in the work of Helton [Hel74] and has been extended to the multi-linear case of Fornasini-Marchesini systems by Ball and Bolotnikov [BB19, BB21]. In particular, these two authors highlight the connection between contractive multipliers and dissipative linear systems.

We start by briefly recalling the main steps we followed in Chapter 4 to solve the problem in that single-letter setting:

- We define $\widehat{A} = \langle \widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}} \rangle$ with $j \geq k$ states such that 1 is not an eigenvalue of $\widehat{\mathbf{A}}$ and

$E = \langle \boldsymbol{\alpha}_e, \mathbf{A}_e, \boldsymbol{\beta}_e \rangle$ is minimal, with:

$$\mathbf{A}_e = \begin{bmatrix} \mathbf{A} & 0 \\ 0 & \widehat{\mathbf{A}} \end{bmatrix}, \quad \boldsymbol{\alpha}_e = \begin{bmatrix} \boldsymbol{\alpha} \\ -\widehat{\boldsymbol{\alpha}} \end{bmatrix}, \quad \boldsymbol{\beta}_e = \begin{bmatrix} \boldsymbol{\beta} \\ \widehat{\boldsymbol{\beta}} \end{bmatrix}.$$

- We consider the function: $e = \phi - \psi = \boldsymbol{\alpha}_e^\top (z\mathbf{1} - \mathbf{A}_e)^{-1} \boldsymbol{\beta}_e + C_e$.
- From the AAK theorem we know that: $\|e\|_\infty = \sigma_k$. Therefore, from the maximum modulus principle for holomorphic functions, we have $e(z)e^*(\bar{z}^{-1}) = \sigma_k^2 \mathbf{1}$.
- From the condition above we derive the following set of equations:

  (a) $\mathbf{P}_e - \mathbf{A}_e \mathbf{P}_e \mathbf{A}_e^\top = \boldsymbol{\beta}_e \boldsymbol{\beta}_e^\top$

  (b) $\mathbf{Q}_e - \mathbf{A}_e^\top \mathbf{Q}_e \mathbf{A}_e = \boldsymbol{\alpha}_e \boldsymbol{\alpha}_e^\top$

  (c) $\mathbf{P}_e \mathbf{Q}_e = \sigma_k^2 \mathbf{1}$

  from which we obtain $\widehat{\mathbf{A}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}$.

- We extract the rational component of ψ by locating the poles.

A noncommutative version of these steps would allows us to build a WFA that, by construction, is an optimal approximation. It is important to note that, in the one-letter case, we had theoretical guarantees on the existence of the solution, derived from Theorem 2.4.7. This is not guaranteed in the multi-letter case, where we don't have an equivalent of the AAK theorem for the symbols. Therefore, following this approach, we might not be able to find a solution at all.

Following the method presented in Section 6.2.2, we can associate a NC rational function

to a given WFA. In particular, we consider the following NC rational functions:

$$\phi = \boldsymbol{\alpha}^*(\mathbf{1} - \sum \mathbf{A}_j z_j)^{-1}\boldsymbol{\beta} \tag{7.11}$$

$$\psi = \widehat{\boldsymbol{\alpha}}^*(\mathbf{1} - \sum \widehat{\mathbf{A}}_j z_j)^{-1}\widehat{\boldsymbol{\beta}} \tag{7.12}$$

$$e = \boldsymbol{\alpha}_e^*(\mathbf{1} - \sum \mathbf{A}_{ej} z_j)^{-1}\boldsymbol{\beta}_e, \tag{7.13}$$

where we did not write explicitly the tensor products, and where an additional constant needs to be added to each function to obtain a NC flipped symbol. The functions above are defined using the coefficients of three WFAs: $A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}, \boldsymbol{\beta}\rangle$, $\widehat{A} = \langle \widehat{\boldsymbol{\alpha}}, \{\widehat{\mathbf{A}}_a\}, \widehat{\boldsymbol{\beta}}\rangle$, and $E = \langle \boldsymbol{\alpha}_e, \{(\mathbf{A}_e)_a\}, \boldsymbol{\beta}_e\rangle$. From Equation 7.6, we can see that it is still reasonable to hope that the following inequality holds:

$$\|e\|_\infty \leq \sigma_k. \tag{7.14}$$

Using the language of multipliers introduced in section 6.1.1, this means that $\sigma_k^{-1}e$ (or, more precisely, $M_{\sigma_k^{-1}e}$) is a contractive multiplier. Ball and Bolotnikov show that it is possible to characterize contractive multipliers between Fock spaces in terms of state-space realizations of discrete systems [BB19, Theorem 3.8]. We now reformulate this result in the framework of weighted automata.

Theorem 7.2.1 ([BB19])**.** *Let $e = \boldsymbol{\alpha}_e^*(\mathbf{1} - \sum \mathbf{A}_{ej} z_j)^{-1}\boldsymbol{\beta}_e + c$ be a NC function associated to an automaton E defined over an alphabet Σ, with $|\Sigma| = d$. Then, $\sigma_k^{-1}e$ is a contractive multiplier if and only if the following properties are satisfied:*

(a) $\mathbf{A}_e = [\mathbf{A}_{e1}, \dots \mathbf{A}_{ed}]$ *is strongly stable, i.e. for all x:*

$$\lim_{N\to\infty} \sum_{\alpha\in\mathbb{F}_d : |\alpha| = N} \|\mathbf{A}_e^\alpha x\|^2 = 0 \tag{7.15}$$

(b) $\mathbf{Q}_e - \mathbf{A}_e^\top \mathbf{Q}_e \mathbf{A}_e \geq \boldsymbol{\alpha}_e \boldsymbol{\alpha}_e^\top$

$$(b)\quad \begin{bmatrix} \boldsymbol{\beta}_e \\ c \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_e^* & c^* \end{bmatrix} = \begin{bmatrix} \mathbf{Q}^{-1} \otimes \mathbf{1} & 0 \\ 0 & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{A}_e \\ \boldsymbol{\alpha}_e \end{bmatrix} \mathbf{Q}_e^{-1} \begin{bmatrix} \mathbf{A}_e^* & \boldsymbol{\alpha}_e^* \end{bmatrix}.$$

This theorem can be seen as a noncommutative version of Theorem 4.2.2. Also in this case, there exists a matrix $\mathbf{Q}_e$ satisfying the fixed-point equation characterizing the Gramians (property (b)). The first property corresponds to the requirement we imposed on the spectral radius of $\mathbf{A}_e$ in the one-letter case, and is analogous to the condition on the joint spectral radius used by Balle et al. [BPP19].

7.2.1 Challenges

Theorem 7.2.1, paired with the definition of NC flipped symbol, is extremely encouraging. It shows that, indeed, a parallel can be drawn between the one-letter and the multi-letter case. Nonetheless, there are two major challenges that need to be faced before we can hope to obtain a result.

The first challenge is linked to the computation of the infinity norm in the multi-variable case. In the one-letter setting, the maximum modulus principle for holomorphic functions allows us to obtain the infinity norm of a function by computing its value at the boundary. In particular, when we consider the constraint $\|e\|_\infty = \sigma_k$ from the AAK theorem, this can be reformulated as a problem on the unit circle, and obtain $e(z)e^*(\bar{z}^{-1}) = \sigma_k^2\mathbf{1}$. Being able to easily compute the infinity norm is at the core of the result in Chapter 4, and it is necessary to obtain the set of constraints in Theorem 4.2.2. Some of these constraints have been translated to the multi-letter case in Theorem 7.2.1, but we are still missing a property similar to that in point (c) of Theorem 4.2.2. Concretely, we do not have enough constraints to derive the desired result. Obtaining a bound in the infinity norm is not as immediate in the multi-letter case. In his paper *Free holomorphic functions on the unit ball of* $B(H)^n$, Popescu develops a theory for noncommutative multivariable holomorphic functions, including a version of the maximum modulus principle [Pop06a, Theorem 3.3]. The theory

pioneered by Popescu constitutes a nice and direct extension of the single-variable case, but it is still not clear how it can be adapted to our setting. Nonetheless, it is a direction that is worth pursuing.

The second challenge seems harder to overcome. The issue lies with what, in Chapter 4, we had referred to as the extraction of the rational component. More specifically, in the one-letter case, we heavily applied Theorem 2.4.7. This theorem told us precisely to which class of functions the symbol of the best approximation belongs, as we knew that $\psi \in \mathcal{H}_k^\infty$. This implied that the rational component had k zeros inside the unit disc. Therefore, to extract it, it was enough to block-diagonalize the transition matrix. At this stage, it is unclear whether or not this holds in the noncommutative case. Even assuming that we can solve the first challenge, we do not necessarily know if we have obtained the best approximation, or if we need to extract a component from it. Once again, we are back to the problem of factorizing a NC rational function.

7.3 Discussion

In this chapter, we provided an overview of the main challenges and obstacles that need to be overcome to make the proof of the NC AAK theorem constructive, and presented two possible approaches to solve the problem. The first approach entails finding the appropriate shift-invariant space and proving that it has the right codimension. From the theory of Krein spaces, we know that there might be more than one maximal subspace (in which case they all have the same dimension). Thus, choosing the right shift-invariant space could be challenging. The alternative to using Krein methods is to find a way to describe the NC Blaschke factors. This is strictly related to the problem of factorizing a noncommutative polynomial, and it is an open question in the field of noncommutative multivariable operator theory. Addressing it is fundamental for this approach to work. The system-theory approach has the advantage that can be formulated in terms of minimal realizations. Indeed, several

methods from functional analysis that rely on factorizing a polynomial have been successfully extended to the noncommutative case using minimal realizations. Even though there are still several challenges that need to be addressed, we believe that at this stage the system-theory approach is the most promising one. To proceed further, the most pressing problem is to analyze the boundary values of a NC rational function. The noncommutative theory needed to solve our task has deep connections to noncommutative algebra, free real algebraic geometry, and free probability; the problem can therefore be tackled using different approaches. The field of noncommutative multivariable operator theory is still very young, and there have been recently several signs of progress in this direction. Addressing them is a necessary step towards obtaining the strong theoretical guarantees, and closed-form solutions, that we have in the one-letter case.

Chapter 8

Related Work

In this chapter, we review works in the literature that are relevant to the results presented in this thesis. We start by briefly mentioning recent papers related to the (approximate) minimization problem for weighted finite automata. Then, we provide an overview of the growing literature concerned with the extraction of deterministic and weighted finite automata from neural networks. Finally, we explore the connections with the fields of control theory and signal processing, where the approximate minimization problem is explored in the context of model reduction.

8.1 Approximate Minimization of Automata

The problem of minimizing automata has been an important subject of research since the 1950s. There is a remarkable algorithm due to Brzozowski [Brz62, Brz64] that reduces a DFA to a minimal one. However, its worst-case running time is exponential in the number of states. Despite this shortcoming, this algorithm has seen a resurgence recently, mainly because it can be generalized to new models, such as weighted automata [DKV09]. This line of algorithms is based on a new understanding of Brzozowski's algorithm from the point of view of duality [BBRS12, BBB$^+$12, BBH$^+$14, BKP12] and extend readily to other settings. In the context of quantitative systems, like weighted or probabilistic automata, it becomes

meaningful to investigate the approximate minimization problem. The study of this problem and of its applications are fairly recent, and only a few works have been published on the subject. A problem analogous to approximate minimization is addressed by Kulesza, Jiang, and Singh for the spectral algorithm. The authors provide a bound on the loss of the learned low-rank model in terms of the singular values that are discarded during training [KJS15]. In a previous work, the same group of authors connected spectral learning to the approximation problem of a small class of Hidden Markov models, bounding the error in terms of the total variation distance [KRS14]. Still in the context of Hidden Markov models, Kotsalis and Shamma provide bounds for the model reduction problem using the spectral norm as a measure of the error [KS15]. We remark that the framework of Hidden Markov models is encompassed by weighted automata [DE08]. Balle, Panangaden, and Precup are the first authors to formalize the approximate minimization problem for WFAs [BPP15, BPP19]. The technique presented in their paper relies on the construction (and truncation) of the singular value automaton, a canonical expression for WFAs arising from the singular value decomposition of the corresponding Hankel matrix (see Section 2.2.1 for more details and for a formal definition). Their method can be viewed as a generalization to multi-letter alphabets of the balanced realization approach from control theory [Ant05]. The authors conclude their analysis by providing bounds on the approximation error in the ℓ^2 norm. The result is supported by strong theoretical guarantees and applies to a large class of WFAs. This method has later been extended to the setting of weighted tree automata by Balle and Rabusseau [BR20]. The main limitation of these approaches based on SVA truncation is that the approximation obtained is not optimal in any norm. We address this point in our first work, detailed in Chapter 4, where we obtain an algorithm for the optimal approximation in the spectral norm for the same class of WFAs considered by Balle, Panangaden, and Precup, but restricted to a one-letter alphabet [BLP$^+$21].

8.2 Extraction of Automata from Neural Networks

Shortly after their first definition by Elman [Elm90], Cleeremans et al. note that, when learning a regular language, simple recurrent neural networks tend to cluster their states in a way similar to the automaton for that language [CSSM89]. More recently, Oliva and Lago-Fernández expand this line of research by studying the ability of recurrent neural networks to model and recognize simple regular languages [OLF19]. In particular, they show that under proper levels of noise and regularization, the RNNs can obtain high accuracy, and the hidden units display activation patterns similar to the discrete states in a deterministic finite automaton. In subsequent work, the same authors perform an empirical study of the stability of RNNs trained to recognize regular languages [OL20]. When a small amount of noise is introduced into the activation function, an analysis of the network activations shows a set of clusters resembling discrete states in a finite state machine, with stable and deterministic transitions between them.

There have been various attempts in the literature to quantitatively analyze the relation between RNNs (or more generally neural networks) and formal models. This has been done mainly via structural correspondences or extraction. For the former, the work of Rabusseau, Li, and Precup [RLP19] shows that WFAs are expressively equivalent to second-order RNNs with linear activation functions and that there exists a one-to-one mapping between the two classes. This result leads to a natural extension of weighted finite automata for sequences of continuous vectors. Moreover, by extending the spectral learning algorithm for WFA, the authors provide a first provable learning algorithm for linear second order RNNs. This is notable, particularly in light of the experimental results presented by Quattoni and Carreras [QC19], which suggest that several forms of non-linearities can be approximated by linear models. Finally, another attempt to bridge the gap between RNNs and WFAs can be found in the paper of Recasens and Quattoni [RQ13], where an extension of probabilistic transducers to continuous inputs is proposed, and in the work of Li, Rabusseau, and Precup [LRP18], that introduces nonlinear weighted automata.

Recent works investigate the relationship between convolutional neural networks CNNs, rational recurrences and finite automata. Quattoni et al. compare a classical CNN architecture for sequence classification against a simple model based on WFAs [QC20]. Their experiments suggest that, despite the apparent simplicity of WFA models and training algorithms, the performance of WFAs is comparable to that of CNNs. In a previous work, Schwartz et al. present a new model combining neural representation learning with WFAs, and showing that these are more expressive than one-layer CNNs [STS18]. Building on this work, Peng et al. define the concept of rational recurrence with the objective of tightening the relationship between WFAs and RNNs for better interpretability [PSTS18]. Specifically, a RNN is rationally recurrent if its recurrent state can be computed by a WFA [PSTS18]. In a parallel line of work, Weiss et al. and Merrill et al. explore the expressive power of different RNN architectures [WGY18b, MWG$^+$20]. They also inspect the computational feasibility of checking for equivalence and computing the distance between languages recognized by these models. In particular, the expressive power of rational and non-rational RNNs is compared, analyzing both state and language expressiveness (the amount of information that RNN states can capture, and which languages can be recognized when the state is passed to a classifier). Finally, studying the expressive power of RNNs, Marzouk and de la Higuera show that the general equivalence problem between WFAs and weighted first-order LM-RNNs is undecidable [MdlH20].

An alternative and well-studied approach consists in extracting an automaton from a trained RNN by discretizing and clustering the hidden state space [GMC$^+$92, OG96a, OG96b, WGY18a, WGY19, OWSH20]. In their works, Giles and Omlin propose a quantization algorithm on the internal state space of a second-order RNN to extract a deterministic finite automaton [OG96a, OG96b]. In [OG96a], a clustering algorithm is first used to parse data into a RNN. The different states obtained by parsing are then stored and the states of the DFA are obtained using the k-means algorithm. Finally, the transitions between states are obtained by parsing the needed elements into the RNN. For a survey on approaches to

clustering in the context of DFA extraction from RNNs, we refer the reader to Jacobsson [Jac05] and Wang et al. [WZI+18]. Merrill et al. propose a method to extract a DFA from a RNN inspired by the state merging paradigm from grammatical inference [MT22]. While performing experiments with their method, the authors noted that continuing to train an RNN after it has perfectly learned the target language improves the extraction performance. Weiss et al. achieve state of the art accuracy when extracting a DFA from RNNs trained over regular languages [WGY18a]. The key idea of the paper is to combine a clustering of the internal space with an exact learning algorithm to extract a DFA from a given RNN. In particular, given a RNN-acceptor R trained over a finite alphabet Σ, the objective is to extract a DFA A that classifies sequences in a way observably equivalent to R. To achieve this, the authors apply Angluin's L^* algorithm [Ang87] using the trained RNN as an oracle. To make the problem more tractable, a finite abstraction $A^{R,p}$ of the RNN is considered. $A^{R,p}$ and the DFA A obtained using the L^* algorithm need to be equivalent to each other. Whenever they disagree on a sample (equivalence queries) the RNN is used to find the true classification of the string and decide whether to return it as a counterexample (A was wrong) or to refine the partition ($A^{R,p}$ was not accurate enough). The membership queries are instead addressed by using directly the RNN. It is important to note that, even if $A^{R,p}$ and the DFA A converge, the equivalence of R and A is not guaranteed. In subsequent work, the same authors extend their result to extract Probabilistic Deterministic Finite Automata from a RNN language model, using conditional probabilities and a local tolerance to compare observations [WGY19]. Analogously, Okudono et al. use spectral learning to extract a WFA from a RNN trained on rational languages, addressing equivalence queries using regression methods [OWSH20]. Moreover, while most of the literature on extraction focuses on RNNs modelling languages that are already recognizable by a WFA, the authors briefly explore the performance of their algorithm over RNNs more expressive than WFAs. Even though it is relevant to note that the extracted WFA still exposes interpretable structures hidden in the RNN, the theoretical guarantees necessary to generalize the result are missing. Zhang et al.

propose a new method to extract WFAs from RNNs using quantization, achieving accurate approximation even on large-scale tasks [ZDX+21]. The problem of verification of a RNN is investigated by Wang et al. through rule extraction and k-means clustering [WZOI+18]. In particular, they propose to apply a small perturbation to check robustness to adversarial examples, using a deterministic finite automaton as an oracle.

A common feature of the extraction algorithms presented above is that they require access to the inner representation of the RNN in order to find a finite partition of the latent representations generated and used by the model. In contrast, Ayache et al. and Eyraud et al. propose a spectral algorithm to extract a WFA from a black-box model that assigns numerical values to symbolic sequential data, without access to the training samples [AEG18, EA20]. To achieve this, a complete and prefixed-closed basis is generated, and the trained RNN is used as an oracle to fill the required sub-blocks of the Hankel matrix. Finally, a WFA is extracted from the sub-blocks using low-rank factorization. The experiments show that the quality of the extracted WFA quickly increases as its rank grows, until a relatively small rank is reached. This behaviour is observed even when the RNN is trained on data extracted from a language that is not recognizable by a WFA, like in the case of the SPiCe dataset. Interestingly, this seems to suggest that the RNNs considered could be computing approximations of rational series.

With the exception of the work of Suresh et al. [SRRS21], where a WFA is extracted from a given probabilistic model over sequences, with the objective of minimizing the Kullback-Leibler divergence with the source model, the majority of the works mentioned in this section evaluate the quality of the extracted WFA in terms of predictive performance. Specifically, the most commonly used metrics are the word error rate and the normalized discounted cumulative gain. Note that probabilistic measures such as perplexity are generally not used when the spectral algorithm is employed, as there are no guarantees that the returned automaton is probabilistic. While evaluating the predictive performance of a model is an important step in its validation, it does not provide theoretical guarantees. This makes it

difficult to compare quantitatively different approximations. In our second contribution, detailed in Chapter 5, we proceed in a manner similar to Ayache et al. [AEG18], and consider a generic black-box model trained for language modelling [LPR21]. Moreover, we work specifically in the general setting of infinite-rank infinite Hankel matrices. This means that the model considered in our solution is not assumed to be computing a rational function. We also provide a theoretical analysis of the problem, and guarantees for the convergence of the approximation operator. We measure the error using the spectral norm, which allows us to compare quantitatively different classes of models, unlike most of the previous work.

8.3 Control Theory and Signal Processing

The control theory community has largely studied approximate minimization in the context of linear time-invariant dynamical systems, and several methods have been proposed. We refer the reader to [Ant05, Mei83] for a complete survey on the subject. A parallel can be drawn between dynamical systems and automata by noting that the impulse-response of a discrete time-invariant Single-Input-Single-Output SISO system can be parametrized as a WFA over a one-letter alphabet. In fact, we can consider the state-variable description for a discrete-time system, given by the dynamical equation:

$$\begin{cases} \mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \end{cases} \tag{8.1}$$

where $\mathbf{A} \in \mathbb{C}^{n\times n}$, $\mathbf{B} \in \mathbb{C}^{n\times m}$, $\mathbf{C} \in \mathbb{C}^{p\times n}$, $\mathbf{D} \in \mathbb{C}^{p\times m}$ are constant matrices, while $\mathbf{x} \in \mathbb{C}^n$, $\mathbf{u} \in \mathbb{C}^m$, $\mathbf{y} \in \mathbb{C}^p$ are the vectors of states, inputs and outputs, respectively, and have for components square-summable sequences of complex numbers. Note that in a SISO system we have $p = m = 1$. When the matrix $\mathbf{A}$ is asymptotically stable (*i.e.* all its eigenvalues lie inside the unit circle) we can derive explicit equations and an expression for the impulse-

response:

$$G(t) := (\mathbf{C}\mathbf{A}^{t-1}\mathbf{B}) + \delta(t)\mathbf{D} \quad \text{for } t \geq 0 \tag{8.2}$$

where $\delta(t) = 0$ if $t > 0$ and $\delta(t) = 1$ otherwise. It is now clear that, for $t > 0$ and $p = m = 1$, the impulse-response corresponds to the function computed by a WFA over a one-letter alphabet, and t denotes the length of a string.

The problem of optimal approximation of control systems with respect to the spectral norm (also referred to as Hankel norm) has been a major line of research in the 1980s. Kung and Lin first applied AAK theory to the control theory setting to obtain an approximation algorithm [Kun80, KL81]. The first state-space solution to the problem is due to Glover, in the setting of continuous Multi-Input-Multi-Output MIMO systems [Glo84]. The construction proposed by Glover relies on embedding the initial stable system into an all-pass dilation of it, having both stable and antistable components. Glover's method led to a widespread application of these results, thanks to its computational efficiency and theoretical clarity. In fact, while a deep understanding of AAK theory requires fundamental tools from functional and harmonic analysis, the approach proposed by Glover is very algebraic. Another important aspect influencing the simplicity of the result stems from the structure of the Lyapunov equations of continuous systems. This simplicity however does not extend to discrete control systems, where the Lyapunov equations have a quadratic form. As noted by Chui and Chen [CC97], a simple closed-form formula for the state space solution of a discrete system does not exist. Therefore, most of the solutions proposed for the discrete case rely on stricter assumptions or are sub-optimal. A first state space solution in the discrete setting can be found in the work of Ball and Ran, where the authors attack a slightly modified (and easier) version of the problem, seeking a sub-optimal solution [BR87]. These results have been reformulated in a descriptor-system framework by Al-Hussari and subsequently by Ionescu and Oara [AHJL93, IO01]. In the first case, the optimal problem is solved under the assumption that the singular value σ_k has multiplicity one [AHJL93], while in the latter the authors propose an alternative solution for the sub-optimal problem [IO01]. The approach

most resembling Glover's method is due to Gu, who provides an elegant solution for the MIMO discrete problem under the assumption that the singular number has multiplicity one [Gu05]. This setting (and therefore the one of Glover) is the closest to the problem presented in Chapter 4. In the chapter, we adapt the concept of all-pass system to the theory of WFAs, in order to leverage a result from Chui and Chen [CC97] (Theorem 4.2.2). We remark that the solution proposed in this thesis is optimal, and there is no additional requirement on the multiplicity of the singular number considered for the error. Finally, we note that a solution for the SISO case can also be found by using a polynomial approach [Ant05]. The result, in this case, holds without any additional assumption but does not provide an explicit representation of the state space of the solution. Moreover, it does not generalize well to the MIMO setting, while a state-space approach does.

A partial solution for the approximation problem of infinite-dimensional systems can be found in the work of Glover, Curtain and Partington [GCP88]. The extensive work of Chui [CLW91, CLW92, CL94] analyzes the continuity of approximation and truncation methods in signal processing. In particular, part of the analysis presented in Chapter 5 is backed by results from Chui [CL94], and our contribution builds upon its work. In the paper, Chui studies the conditions guaranteeing the continuity of the approximation operator. He shows that the approximation operator converges whenever its singular numbers are simple and provides examples of converging sequences of Hankel operators. His analysis is missing a clear criterion to check the condition on the singular numbers, that we provide using tools from random matrix theory.

We conclude this section by mentioning that AAK theory has more recently been applied to the problem of sparse approximation of structured signals in signal processing, providing interesting results in the ℓ^p norm [ACd11, PP16].

Chapter 9

Conclusion

This chapter is dedicated to the concluding remarks of this thesis. After summarizing our contributions and the implications of our work, we discuss the main limitations of the proposed techniques. Finally, we provide possible directions for future work.

9.1 Summary of Contributions

This thesis makes several contributions to the approximate minimization problem of models on sequential data. While tackling this problem, we tightened the connections between the study of weighted finite automata and discrete dynamical systems. Chapter 3 and Chapter 6 provide a framework to reformulate the approximate minimization problem of models over one-letter or multi-letter alphabets in terms of functional analysis and noncommutative multivariable operator theory, respectively. While a parallel can be drawn between the one-letter framework and certain results in control theory, the multi-letter setting can be seen as a completely novel approach to this problem. We remark that the proposed frameworks are relevant beyond the approximate minimization problem: we consider this to be an interesting contribution on its own, as it opens the doors to new potential applications of these mathematical theories to the field of language modelling. This would not be the first time that functional analysis methods are applied to solve machine learning problems. For

example, one can think of the use of Reproducing kernel Hilbert spaces in machine learning [HSS08], or the recent application of Fourier analysis to study the spectral bias of neural networks [RBA+19]. Chapter 4 and Chapter 5 focus on the case of one-letter alphabets, and address in this setting the research questions **Q1** and **Q2** stated in Chapter 1. We highlight the parallel between this type of approximation problem and the model reduction task in control theory and signal processing. We adapt some of the formalism of these fields to the case of WFA and black boxes. In particular, in Chapter 4 we study the approximate minimization problem of irredundant WFAs over a one-letter alphabet. To do so, we use the WFAs' parameters to define complex functions on the unit circle. This allows us to apply the results from AAK theory and to provide a closed-form solution for the optimal approximate minimization problem in the spectral norm. While the underlying theory of complex functions supporting our method can be a bit involved, the proposed algorithm relies only on simple algebraic manipulations and can be run in polynomial time. In Chapter 5, we extend these results to more general black-box models trained on sequential data for a language modelling task. Unlike the majority of works in the literature, that focus on specific models (for example RNNs) computing a rational function, we do not assume any knowledge on the internal structure of the model, nor on the training data. Moreover, we specifically study models computing a function that cannot be recognized by a WFA. We also provide theoretical guarantees on the convergence of our approximation. We then apply AAK theory and obtain an algorithm returning a WFA of fixed size that approximates the model asymptotically optimally. This is particularly important, as it constitutes a first step towards developing provable approximation algorithms for black-box models. Finally, we use the spectral norm as a measure of the error between a given black box and its asymptotically-optimal approximation. This allows us to compute the distance between WFAs and black-box models, and to obtain bounds in the ℓ^2 norm. Even though this result relies on strong assumptions on the size of the alphabet, it still constitutes the first algorithmic attempt to find the optimal approximating WFA while providing strong theoretical

guarantees and a precise estimate of the error. Chapter 6 and Chapter 7 are devoted to the case of multi-letter alphabets, and partially address the question **Q3** from Chapter 1. We propose a way, similar to the one-letter case, to associate a noncommutative rational function to a given WFA. Then, we try to address the question of whether or not the proof of the noncommutative AAK theorem can be made constructive. While we don't manage to provide a definitive answer, we lay out possible approaches that can be used to tackle the problem. This overview bridges methods from multivariable noncommutative operator theory and system theory. Given the variety of sources, and the depth of the results that we analyze, we believe that providing a unified source for these methods is of intrinsic value.

9.2 Limitations and Directions for Future Work

In this section, we list the main limitations of this thesis, and suggest ways to address them as possible directions for future work.

Size of the Alphabet

The main limitation of this work is that the proposed method is constructive only in the case of models computing functions over one-letter alphabets. Generalization to a multi-letter setting has proven to be rather challenging, and it is hard to predict when (or whether at all) it will be possible to obtain an algorithm for this case. Therefore, while the results still bears theoretical relevance and novelty, the possible direct applications are very limited.

The most compelling direction for future work is to extend the results to the case of multi-letter alphabets. Obtaining a constructive proof of the noncommutative AAK theorem would unlock several directions worth investigating. The first, most direct step, would be to generalize the results of Chapter 4 and obtain an algorithm for WFA extraction from a WFA of a bigger size in the multi-letter case. It would then be interesting to investigate if the convergence results from Chui and li [CL94], that we applied to study the approximation

of infinite-rank infinite Hankel matrix, still hold for the multi-letter setting. This step is necessary if we want to use the method to study the approximation of black boxes on sequential data.

Evaluating the Spectral Norm

In Section 3.1.1, we listed relevant properties of the spectral norm. We think the spectral norm has desirable characteristics, that make it a solid candidate for the approximate minimization task. Nonetheless, a big limitation of this work is that we do not have a clear picture of how effective it is to use the spectral norm to evaluate the approximation of WFAs and black boxes. Concretely, we do not know how the spectral norm performs with respect to more popular metrics such as behavioral metrics, word error rate, or normalized discounted cumulative gain. This problem is a collateral effect deriving from the size of the alphabet. The comparison between spectral norm and other kind of norms is possible only in the multi-letter setting, making it impossible to test in the one-letter case. Obtaining algorithms for the multi-letter case will thus directly open the possibility of evaluating the quality of the spectral norm. The algorithm for black-box models presented in Chapter 5 provides other avenues that would be interesting to explore. In particular, the sequence of truncated Hankel matrices could be defined in alternative ways. Exploring experimentally how different truncations influence the convergence rate is an interesting direction for future work.

Beyond Language Modelling

To apply the AAK theorem, we required the Hankel operator considered to be compact. To guarantee that this property is satisfied, we restricted our focus to models computing functions $f \in \ell^1$. For WFAs, this means considering only automata that are irredundant. For black-box models, we instead required a language modelling task. It is important to remark that, while this kind of task is encompassed in the hypothesis $f \in \ell^1$, the class of functions

that satisfy this property is actually bigger. Nonetheless, it would be interesting to explore ways to extend the method beyond this hypothesis. In Chapter 4 we illustrated a possible approach in the case of WFAs. In future work, it would be interesting to investigate how these methods can be applied to other machine learning settings, for example reinforcement learning.

Bibliography

[AAK71] Vadim M. Adamyan, Damir Zyamovich Arov, and Mark Grigorievich Krein. Analytic Properties of Schmidt Pairs for a Hankel Operator and the Generalized Schur–Takagi problem. *Mathematics of The Ussr-sbornik*, 15:31–73, 1971.

[ACd11] Fredrik Andersson, Marcus Carlsson, and Maarten V. de Hoop. Sparse Approximation of Functions Using Sums of Exponentials and AAK Theory. *Journal of Approximation Theory*, 163(2):213–248, 2011. URL: https://www.sciencedirect.com/science/article/pii/S0021904510001620, doi:https://doi.org/10.1016/j.jat.2010.09.005.

[ACP15] Fredrik Andersson, Marcus Carlsson, and Karl-Mikael Perfekt. AAK-Type Theorems for Hankel Operators on Weighted Spaces. *Bulletin des Sciences Mathématiques*, 139(2):184–197, 2015. URL: https://www.sciencedirect.com/science/article/pii/S000744971400061X, doi:https://doi.org/10.1016/j.bulsci.2014.08.008.

[AEG18] Stéphane Ayache, Rémi Eyraud, and Noé Goudian. Explaining Black Boxes on Sequential Data Using Weighted Automata. In *Proceedings of the 14th International Conference on Grammatical Inference, ICGI 2018, Wrocław, Poland, September 5-7, 2018*, volume 93 of *Proceedings of Machine Learning Research*, pages 81–103. PMLR, 2018. URL: http://proceedings.mlr.press/v93/ayache19a.html.

[AGH+14] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor Decompositions for Learning Latent Variable Models. *Journal of Machine Learning Research*, 15(80):2773–2832, 2014. URL: `http://jmlr.org/papers/v15/anandkumar14b.html`.

[AHJL93] M.M Al-Hussari, I.M. Jaimoukha, and D.J.N. Limebeer. A Descriptor Approach for the Solution of the One-Block Distance Problem. In *In Proceedings of the IFAC World Congress*, 1993.

[Al'17] Yu.A. Al'pin. The Hankel Matrix Rank Theorem Revisited. *Linear Algebra and its Applications*, 534:97–101, 2017. URL: `https://www.sciencedirect.com/science/article/pii/S002437951730486X`, `doi:https://doi.org/10.1016/j.laa.2017.08.010`.

[Ang87] Dana Angluin. Learning Regular Sets from Queries and Counterexamples. *Information and Computation*, 75:87–106, 1987.

[Ant05] Athanasios C. Antoulas. *Approximation of Large-Scale Dynamical Systems*. SIAM, 2005.

[AP95] Alvaro Arias and Gelu Popescu. Factorization and Reflexivity on Fock Spaces. *Integral equations and operator theory*, 23(3):268–286, 1995.

[Bai11] Raphael Bailly. Quadratic Weighted Automata: Spectral Algorithm and Likelihood Maximization. In Chun-Nan Hsu and Wee Sun Lee, editors, *Proceedings of the Asian Conference on Machine Learning*, volume 20 of *Proceedings of Machine Learning Research*, pages 147–163, South Garden Hotels and Resorts, Taoyuan, Taiwain, 14–15 Nov 2011. PMLR. URL: `http://proceedings.mlr.press/v20/bailly11.html`.

[BB19] Joseph A. Ball and Vladimir Bolotnikov. Hardy-Space Function Theory, Operator Model Theory, and Dissipative Linear Systems: the multivariable, Free-Noncommutative, Weighted bergman-Space Setting, 2019. `arXiv:1906.02814`.

[BB21] Joseph A. Ball and Vladimir Bolotnikov. *Noncommutative Function-Theoretic Operator Theory and Applications*. Cambridge Tracts in Mathematics. Cambridge University Press, 2021.

[BBB+12] Filippo Bonchi, Marcello Bonsangue, Michele Boreale, Jan Rutten, and Alexandra Silva. A Coalgebraic Perspective on Linear Weighted Automata. *Information and Computation*, 211:77–105, 2012.

[BBH+14] Filippo Bonchi, Marcello M. Bonsangue, Helle Hvid Hansen, Prakash Panangaden, Jan Rutten, and Alexandra Silva. Algebra-Coalgebra Duality in Brzozowski's Minimization Algorithm. *ACM Transactions on Computational Logic*, 2014.

[BBRS12] Filippo Bonchi, Marcello Bonsangue, Jan Rutten, and Alexandra Silva. Brzozowski's algorithm (co)algebraically. In Robert Constable and Alexandra Silva, editors, *Logics and Program Semantics: Essays Dedicated to Dexter Kozen*, volume 7230 of *Lecture Notes In Computer Science*, pages 12–23. Springer-Verlag, 2012.

[BCLQ14] Borja Balle, Xavier Carreras, Franco M. Luque, and Ariadna Quattoni. Spectral Learning of Weighted Automata - A Forward–Backward Perspective. *Mach. Learn.*, 96(1-2):33–63, 2014. `doi:10.1007/s10994-013-5416-x`.

[BDR09] Raphaël Bailly, François Denis, and Liva Ralaivola. Grammatical inference as a principal component analysis problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 33—40, New

York, NY, USA, 2009. Association for Computing Machinery. doi:10.1145/1553374.1553379.

[Ber79] J. Berstel. Transductions and Context-Free Languages. In *Teubner Studienbücher : Informatik*, 1979.

[Beu49] Arne Beurling. On Two Problems Concerning Linear Transformations in Hilbert Space. *Acta Mathematica*, 81(none):239 – 255, 1949. doi:10.1007/BF02395019.

[BGP17] Borja Balle, Pascale Gourdeau, and Prakash Panangaden. Bisimulation Metrics and Norms for Weighted Finite Automata. In Ioannis Chatzigiannakis, Piotr Indyk, Fabian Kuhn, and Anca Muscholl, editors, *44th International Colloquium on Automata, Languages, and Programming, ICALP 2017, July 10-14, 2017, Warsaw, Poland*, volume 80 of *LIPIcs*, pages 103:1–103:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017. doi:10.4230/LIPIcs.ICALP.2017.103.

[BGP22] Borja Balle, Pascale Gourdeau, and Prakash Panangaden. Bisimulation Metrics and Norms for Real-Weighted Automata. *Inf. Comput.*, 282:104649, 2022. doi:10.1016/j.ic.2020.104649.

[BHP14] Borja Balle, William L. Hamilton, and Joelle Pineau. Methods of Moments for Learning Stochastic Languages: Unified Presentation and Empirical Comparison. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1386–1394. JMLR.org, 2014. URL: http://proceedings.mlr.press/v32/balle14.html.

[BKP12] Nick Bezhanishvili, Clemens Kupke, and Prakash Panangaden. Minimization via Duality. In *Logic, Language, Information and Computation - 19th International Workshop, WoLLIC 2012, Buenos Aires, Argentina, September 3-6,*

2012. Proceedings, volume 7456 of *Lecture Notes in Computer Science*, pages 191–205. Springer, 2012.

[BLP+21] Borja Balle, Clara Lacroce, Prakash Panangaden, Doina Precup, and Guillaume Rabusseau. Optimal Spectral-Norm Approximate Minimization of Weighted Finite Automata. In Nikhil Bansal, Emanuela Merelli, and James Worrell, editors, *48th International Colloquium on Automata, Languages, and Programming, ICALP 2021, July 12-16, 2021, Glasgow, Scotland (Virtual Conference)*, volume 198 of *LIPIcs*, pages 118:1–118:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPIcs.ICALP.2021.118.

[BPP15] Borja Balle, Prakash Panangaden, and Doina Precup. A Canonical Form for Weighted Automata and Applications to Approximate Minimization. In *30th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2015, Kyoto, Japan, July 6-10, 2015*, pages 701–712. IEEE Computer Society, 2015. doi:10.1109/LICS.2015.70.

[BPP19] Borja Balle, Prakash Panangaden, and Doina Precup. Singular Value Automata and Approximate Minimization. *Math. Struct. Comput. Sci.*, 29(9):1444–1478, 2019. doi:10.1017/S0960129519000094.

[BR87] Joseph A. Ball and Andre CM. Ran. Optimal Hankel Norm Model Reductions and Wiener–Hopf Factorization I: The Canonical Case. *SIAM Journal on Control and Optimization*, 25(2):362–382, 1987.

[BR11] Jean Berstel and Christophe Reutenauer. *Noncommutative Rational Series with Applications*, volume 137. Cambridge University Press, 2011.

[BR20] Borja Balle and Guillaume Rabusseau. Approximate Minimization of Weighted Tree Automata. *Information and Computation*, page 104654,

2020. URL: https://www.sciencedirect.com/science/article/pii/S0890540120301425, doi:https://doi.org/10.1016/j.ic.2020.104654.

[Brz62] Janusz A. Brzozowski. Canonical Regular Expressions and Minimal State Graphs for Definite Events. In J. Fox, editor, *Proceedings of the Symposium on Mathematical Theory of Automata*, number 12 in MRI Symposia Series, pages 529–561. Polytechnic Press of the Polytechnic Institute of Brooklyn, April 1962. Book appeared in 1963.

[Brz64] Janusz A. Brzozowski. Derivatives of Regular Expressions. *J. ACM*, 11(4):481–494, 1964.

[BS72] R. H. Bartels and G. W. Stewart. Solution of the Matrix Equation AX + XB = C [f4]. *Commun. ACM*, 15(9):820–826, 1972.

[Bun84] John W Bunce. Models for n-Tuples of Noncommuting Operators. *Journal of Functional Analysis*, 57(1):21–30, 1984. URL: https://www.sciencedirect.com/science/article/pii/0022123684900983, doi:https://doi.org/10.1016/0022-1236(84)90098-3.

[Car09] Marcus Carlsson. AAK-Theory on Weighted Spaces. In *Proceedings of the Project Review, Geo-Mathematical Imaging Group (Purdue University, West Lafayette IN)*, volume 1, pages 27–48, 2009.

[CC97] Charles K. Chui and Guanrong Chen. *Discrete H^∞ Optimization With Applications in Signal Processing and Control Systems*. Springer-Verlag, 1997.

[CHM04] Corinna Cortes, Patrick Haffner, and Mehryar Mohri. Rational Kernels: Theory and Algorithms. *Journal of Machine Learning Research (JMLR)*, 5:1035–1062, 2004. URL: http://www.cs.nyu.edu/~mohri/postscript/jmlr.pdf.

[CL94] Charles K. Chui and Xin Li. Continuity of Best Hankel Approximation and Convergence of Near–Best Approximants. *SIAM J. Control Optim.*, 32(6):1769—-1781, November 1994. doi:10.1137/S0363012992232245.

[Cla68] Douglas N. Clark. On the Spectra of Bounded, Hermitian, Hankel Matrices. *American Journal of Mathematics*, 90(2):627–656, 1968. URL: http://www.jstor.org/stable/2373546.

[CLW91] Charles K. Chui, Xin Li, and Joseph D Ward. System Reduction Via Truncated Hankel Matrices. *Mathematics of Control, Signals and Systems*, 4(2):161–175, 1991.

[CLW92] Charles K. Chui, Xin Li, and Joseph D. Ward. Rate of Convergence of Schmidt Pairs and Rational Functions Corresponding to Best Approximants of Truncated Hankel Operators. *Math. Control. Signals Syst.*, 5(1):67–79, 1992. doi:10.1007/BF01211976.

[Coh95] Paul M. Cohn. *Skew Fields: Theory of General Division Rings*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1995. doi:10.1017/CBO9781139087193.

[CP71] J.W. Carlyle and A. Paz. Realizations by Stochastic Finite Automata. *Journal of Computer and System Sciences*, 5(1):26–40, 1971.

[CSSM89] Axel Cleeremans, David Servan-Schreiber, and James L McClelland. Finite state automata and simple recurrent networks. *Neural computation*, 1(3):372–381, 1989.

[DE08] François Denis and Yann Esposito. On Rational Stochastic Languages. *Fundamenta Informaticae*, 86(1, 2):41–77, 2008.

[DKV09] Manfred Droste, Werner Kuich, and Heiko Vogler. *Handbook of weighted automata*. Springer Science & Business Media, 2009.

[DP99] Kenneth R. Davidson and David R. Pitts. Invariant Subspaces and Hyper-Reflexivity for Free Semigroup Algebras. *Proceedings of the London Mathematical Society*, 78(2):401–430, 1999. doi:10.1112/S002461159900180X.

[DVK17] Finale Doshi-Velez and Been Kim. Towards a Rigorous Science of Interpretable Machine Learning, 2017. arXiv:1702.08608.

[DWS+20] Guoliang Dong, Jingyi Wang, Jun Sun, Yang Zhang, Xinyu Wang, Ting Dai, Jin Song Dong, and Xingen Wang. Towards Interpreting Recurrent Neural Networks through Probabilistic Abstraction. In *35th IEEE/ACM International Conference on Automated Software Engineering, ASE 2020, Melbourne, Australia, September 21-25, 2020*, pages 499–510. IEEE, 2020. doi:10.1145/3324884.3416592.

[EA20] Rémi Eyraud and Stéphane Ayache. Distillation of Weighted Automata from Recurrent Neural Networks Using a Spectral Approach. *CoRR*, abs/2009.13101, 2020. URL: https://arxiv.org/abs/2009.13101, arXiv:2009.13101.

[Elm90] Jeffrey L Elman. Finding Structure in Time. *Cognitive science*, 14(2):179–211, 1990.

[EY36] C. Eckart and G. Young. The Approximation of one Matrix by Another of Lower Rank. *Psychometrika*, 1:211–218, 1936. doi:10.1007/BF02288367.

[Fli74] Michel Fliess. Matrice de Hankel. *Journal de Mathématique Pures et Appliquées*, 5:197–222, 1974.

[Fra82] Arthur E Frazho. Models for Noncommuting operators. *Journal of Functional Analysis*, 48(1):1–11, 1982. doi:https://doi.org/10.1016/0022-1236(82)90057-X.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[GCP88] Keith Glover, Ruth F Curtain, and Jonathan R Partington. Realisation and Approximation of Linear Infinite–Dimensional Systems with Error Bounds. *SIAM Journal on Control and Optimization*, 26(4):863–898, 1988.

[Glo84] Keith Glover. All Optimal Hankel–Nnorm Approximations of Linear Multivariable Systems and their $\mathcal{L}^\infty$–Error Bounds. *International Journal of Control*, 39(6):1115–1193, 1984. doi:10.1080/00207178408933239.

[GMC+92] Clyde Lee Giles, Clifford B. Miller, Dong Chen, Hsing-Hen Chen, Guo-Zheng Sun, and Yee-Chun Lee. Learning and Extracting Finite State Automata with Second-Order Recurrent Neural Networks. *Neural Comput.*, 4(3):393–405, 1992. doi:10.1162/neco.1992.4.3.393.

[Gu05] Guoxiang Gu. All Optimal Hankel-Norm Approximations and their Error Bounds in Discrete-Time. *International Journal of Control*, 78(6):408–423, 2005. doi:10.1080/00207170500110988.

[Hel74] J William Helton. Discrete Time Systems, Operator Models, and Scattering Theory. *Journal of Functional analysis*, 16(1):15–38, 1974.

[Her11] Gustav Herglotz. Über Potenzreihen mit positivem, Reellen Teil im Einheitskreis. In *Berichte der Sächsischen Gesellschaft der Wissenschaften zu Leipzig*, volume 63 of *Mathematics, Physics*, pages 501–511, 1911.

[HKZ12] Daniel Hsu, Sham M. Kakade, and Tong Zhang. A Spectral Algorithm for Learning Hidden Markov Models. *J. Comput. Syst. Sci.*, 78(5):1460–1480, September 2012. doi:10.1016/j.jcss.2011.12.025.

[HM94] Lars Hörmander and Anders Melin. A Remark on Perturbations of Compact Operators. *Mathematica Scandinavica*, 75(2):255–262, 1994. URL: http://www.jstor.org/stable/24491887.

[HMS17] J. William Helton, Tobias Mai, and Roland Speicher. Applications of Realizations (aka Linearizations) to Free Probability, 2017. arXiv:1511.05330.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[HSS08] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel Methods in Machine Learning. *The annals of statistics*, 36(3):1171–1220, 2008.

[HVD15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015.

[HVLS16] Christian Albert Hammerschmidt, Sicco Verwer, Qin Lin, and Radu State. Interpreting Finite Automata for Sequential Data, 2016. arXiv:1611.07100.

[Hwa04] Suk-Geun Hwang. Cauchy's Interlace Theorem for Eigenvalues of Hermitian Matrices. *The American Mathematical Monthly*, 111(2):157–159, 2004. arXiv:https://doi.org/10.1080/00029890.2004.11920060, doi:10.1080/00029890.2004.11920060.

[IO01] Vlad Ionescu and Cristian Oara. The four-block Adamjan-Arov-Kein problem for Discrete-Time Systems. In *Linear Algebra and its Application*, pages 95–119. Elsevier, 2001.

[Iok64] I. S. Iokhvidov. On a lemma of K. Fan generalizing Tyckhonov's fixed point principle. *Dokl Akad Nauk SSSR, (Russian)*, 159:501–504, 1964.

[Jac05] Henrik Jacobsson. Rule Extraction from Recurrent Neural Networks: A Taxonomy and Review. *Neural Comput.*, 17(6):1223–1263, 2005. doi:10.1162/0899766053630350.

[JMS21a] Michael Jury, Robert Martin, and Eli Shamovich. Non-Commutative Rational Functions in the Full Fock Space. *Transactions of the American Mathematical Society*, 2021.

[JMS21b] Michael T Jury, Robert TW Martin, and Eli Shamovich. Blaschke–Singular–Outer Factorization of Free Non-Commutative Functions. *Advances in Mathematics*, 384:107720, 2021.

[Jur21] Michael Jury. Mini-Course on Non-Commutative Function Theory. Mini-course and Workshop on the Non-commutative Function Theory, The Fields Institute, Toronto, 16–19 Nov 2021. URL: http://www.fields.utoronto.ca/activities/21-22/function-non-commutative.

[Kat13] Tosio Kato. *Perturbation Theory for Linear Operators*, volume 132. Springer Science & Business Media, 2013.

[KG69] M. Kreĭn and I. Gohberg. *Introduction to the Theory of Linear Nonselfadjoint Operators in Hilbert Space*, volume 18 of *Translations of Mathematical Monographs*. American Mathematical Society, 1969.

[KJS15] Alex Kulesza, Nan Jiang, and Satinder Singh. Low-Rank Spectral Learning with Weighted Loss Functions. In *Artificial Intelligence and Statistics*, pages 517–525. PMLR, 2015.

[KL81] Sun-Yuan Kung and David W. Lin. Optimal Hankel-Norm Model Reductions: Multivariable Systems. *IEEE Transactions Automation Control*, 26:832–852, 1981.

[KNR+21] Igor Khmelnitsky, Daniel Neider, Rajarshi Roy, Xuan Xie, Benoît Barbot, Benedikt Bollig, Alain Finkel, Serge Haddad, Martin Leucker, and Lina Ye. Property-Directed Verification and Robustness Certification of Recurrent Neural Networks. In Zhe Hou and Vijay Ganesh, editors, *Automated Technology for Verification and Analysis - 19th International Symposium, ATVA 2021, Gold Coast, QLD, Australia, October 18-22, 2021, Proceedings*, volume 12971 of *Lecture Notes in Computer Science*, pages 364–380. Springer, 2021. doi:10.1007/978-3-030-88885-5_24.

[KPV17] Igor Klep, James Eldred Pascoe, and Jurij Volčič. Regular and Positive Noncommutative Rational Functions. *Journal of the London Mathematical Society*, 95(2):613—632, Feb 2017. URL: http://dx.doi.org/10.1112/jlms.12030, doi:10.1112/jlms.12030.

[Kro81] L. Kronecker. Zur Theorie der Elimination einer Variablen aus zwei algebraischen Gleichungen. *Montasber. Königl. Preussischen Acad Wies*, pages 535 – 600, 1881.

[KRS14] Alex Kulesza, N. Raj Rao, and Satinder Singh. Low-Rank Spectral Learning. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 522–530, Reykjavik, Iceland, 22–25 April 2014. PMLR. URL: http://proceedings.mlr.press/v33/kulesza14.html.

[KS15] Georgios Kotsalis and Jeff S. Shamma. Limits of Performance for the Model Reduction Problem of Hidden Markov Models. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 4674–4679, 2015. doi:10.1109/CDC.2015.7402948.

[KT77] H. T. Kung and D. M. Tong. Fast Algorithms for Partial Fraction Decomposition. *SIAM J. Comput.*, 6(3):582–593, 1977. doi:10.1137/0206042.

[Kun80] Sun-Yuan Kung. Optimal Hankel-Nnorm Model Reductions: Scalar Systems. In *Proceedings of the 1980 Joint Automation Control Conference,San Francisco, CA*, page Paper FA8.A, 1980.

[KVV09] Dmitry S. Kaliuzhnyi-Verbovetskyi and Victor Vinnikov. Singularities of Rational Functions and Minimal Factorizations: the Noncommutative and the Commutative Setting. *Linear Algebra and its Applications*, 430(4):869–889, 2009. URL: https://www.sciencedirect.com/science/article/pii/S0024379508003893, doi:https://doi.org/10.1016/j.laa.2008.08.027.

[KVV14] Dmitry S Kaliuzhnyi-Verbovetskyi and Victor Vinnikov. *Foundations of Free Noncommutative Function Theory*, volume 199. American Mathematical Soc., 2014.

[Low20] Ryan Lowe. Learning and Evaluating Neural Network Models for Human-Machine Communication. *McGill University*, 2020.

[LPR21] Clara Lacroce, Prakash Panangaden, and Guillaume Rabusseau. Extracting Weighted Automata for Approximate Minimization in Language Modelling. In Jane Chandlee, Rémi Eyraud, Jeff Heinz, Adam Jardine, and Menno van Zaanen, editors, *Proceedings of the Fifteenth International Conference on Grammatical Inference*, volume 153 of *Proceedings of Machine Learning Research*, pages

92–112. PMLR, 23–27 Aug 2021. URL: https://proceedings.mlr.press/v153/lacroce21a.html.

[LPR22] Clara Lacroce, Prakash Panangaden, and Guillaume Rabusseau. Towards an AAK Theory Approach to Approximate Minimization in the Multi-Letter Case. *CoRR*, abs/2206.00172, 2022. arXiv:2206.00172, doi:10.48550/arXiv.2206.00172.

[LRP18] Tianyu Li, Guillaume Rabusseau, and Doina Precup. Nonlinear Weighted Finite Automata. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 679–688. PMLR, 09–11 Apr 2018. URL: http://proceedings.mlr.press/v84/li18a.html.

[LSS01] Michael L. Littman, Richard S. Sutton, and Satinder P. Singh. Predictive Representations of State. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 1555–1561. MIT Press, 2001. URL: https://proceedings.neurips.cc/paper/2001/hash/1e4d36177d71bbb3558e43af9577d70e-Abstract.html.

[Lya50] Aersity M Lyapunov. The General Problem of the Stability of Motion [in Russian]. *Gostekhizdat, Moscow*, 1950.

[MDL+22] Mingjun Ma, Dehui Du, Yuanhao Liu, Yanyun Wang, and Yiyang Li. Efficient Adversarial Sequence Generation for RNN with Symbolic Weighted Finite Automata. *CoRR*, 2022.

[MdlH20] Reda Marzouk and Colin de la Higuera. Distance and Equivalence Between Finite State Machines and Recurrent Neural Networks: Computational Results. *CoRR*, abs/2004.00478, 2020. URL: `https://arxiv.org/abs/2004.00478`, `arXiv:2004.00478`.

[Mei83] Jean Meinguet. A Simplified Presentation of the Adamjan-Arov-Krein Approximation Theory. In H. Werner, L. Wuytack, E. Ng, and H. J. Bünger, editors, *Computational Aspects of Complex Analysis*, pages 217–248, Dordrecht, 1983. Springer Netherlands. `doi:10.1007/978-94-009-7121-9_9`.

[Mir60] L. Mirsky. Symmetric Gauge Functions and Unitarily Invariant Norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, January 1960. `doi:10.1093/qmath/11.1.50`.

[MKB+10] Tomás Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent Neural Network Based Language Model. In Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, editors, *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048. ISCA, 2010. URL: `http://www.isca-speech.org/archive/interspeech_2010/i10_1045.html`.

[Moh09] Mehryar Mohri. *Weighted Automata Algorithms*. Springer-Verlag, 2009.

[MT22] William Merrill and Nikolaos Tsilivis. Extracting Finite Automata from RNNs Using State Merging, 2022. `arXiv:2201.12451`.

[MWG+20] William Merrill, Gail Weiss, Yoav Goldberg, Roy Schwartz, Noah A. Smith, and Eran Yahav. A Formal Hierarchy of RNN Architectures. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*

2020, Online, July 5-10, 2020, pages 443–459. Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.acl-main.43.

[MYW14] Youneng Ma, Jinhua Yu, and Yuanyuan Wang. Efficient Recursive Methods for Partial Fraction Expansion of General Rational Functions. *J. Appl. Math.*, 2014:895036:1–895036:18, 2014. doi:10.1155/2014/895036.

[Neh57] Zeev Nehari. On Bounded Bilinear Forms. *Annals of Mathematics*, 65(1):153–162, 1957.

[Nik02] Nikolai K. Nikol'Skii. *Operators, Functions and Systems: An Easy Reading*, volume 92 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2002.

[OG96a] Christian W. Omlin and Clyde Lee Giles. Constructing Deterministic Finite-State Automata in Recurrent Neural Networks. *J. ACM*, 43(6):937–972, 1996. doi:10.1145/235809.235811.

[OG96b] Christian W. Omlin and Clyde. Lee Giles. Extraction of Rules From Discrete-Time Recurrent Neural Networks. *Neural Networks*, 9(1):41–52, 1996. doi:10.1016/0893-6080(95)00086-0.

[OL20] Christian Oliva and Luis Fernando Lago-Fernández. Stability of Internal States in Recurrent Neural Networks Trained on Regular Languages. *CoRR*, abs/2006.10828, 2020. URL: https://arxiv.org/abs/2006.10828, arXiv:2006.10828.

[OLF19] Christian Oliva and Luis F. Lago-Fernández. Interpretability of Recurrent Neural Networks Trained on Regular Languages. In Ignacio Rojas, Gonzalo Joya, and Andreu Catala, editors, *Advances in Computational Intelligence*, pages 14–25, Cham, 2019. Springer International Publishing.

[OS62] Alexander Ostrowski and Hans Schneider. Some Theorems on the Inertia of General Matrices. *J. Math. Anal. Appl*, 4(1):72–84, 1962.

[OWSH20] Takamasa Okudono, Masaki Waga, Taro Sekiyama, and Ichiro Hasuo. Weighted Automata Extraction from Recurrent Neural Networks via Regression on State Spaces. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 5306–5314. AAAI Press, 2020. URL: `https://aaai.org/ojs/index.php/AAAI/article/view/5977`.

[Pel12] Vladimir Peller. *Hankel Operators and their Applications*. Springer Science & Business Media, 2012.

[Pig15] G. Pighizzini. Investigations on Automata and Languages Over a Unary Alphabet. *Int. J. Found. Comput. Sci.*, 26:827–850, 2015.

[Pop89a] Gelu Popescu. Characteristic Functions for Infinite Sequences of Noncommuting Operators. *Journal of Operator Theory*, pages 51–71, 1989.

[Pop89b] Gelu Popescu. Isometric Dilations for Infinite Sequences of Noncommuting Operators. *Transactions of the American Mathematical Society*, 316(2):523–536, 1989.

[Pop89c] Gelu Popescu. Models for Infinite Sequences of Noncommuting Operators. *Acta Sei. Math*, 53:355–285, 1989.

[Pop92] Gelu Popescu. On Intertwining Dilations for Sequences of Noncommuting Operators. *Journal of mathematical analysis and applications*, 167(2):382–402, 1992.

[Pop93] Gelu Popescu. *Noncommutative Dilation Theory on Fock Spaces*. PhD thesis, Texas A&M University, 1993.

[Pop95a] Gelu Popescu. Functional Calculus for Noncommuting Operators. *Michigan Mathematical Journal*, 42(2):345 – 356, 1995. doi:10.1307/mmj/1029005232.

[Pop95b] Gelu Popescu. Multi-Analytic Operators on Fock Spaces. *Mathematische Annalen*, 303(1):31–46, 1995.

[Pop03] Gelu Popescu. Multivariable Nehari Problem and Interpolation. *Journal of Functional Analysis*, 200:536–581, 2003. doi:10.1016/S0022-1236(03)00078-8.

[Pop06a] Gelu Popescu. Free Holomorphic Functions on the Unit Ball of B(H)n. *Journal of Functional Analysis*, 241(1):268–333, 2006. URL: https://www.sciencedirect.com/science/article/pii/S0022123606003028, doi:https://doi.org/10.1016/j.jfa.2006.07.004.

[Pop06b] Gelu Popescu. Operator Theory on Noncommutative Varieties. *Indiana University mathematics journal*, pages 389–442, 2006.

[Pop10] Gelu Popescu. *Operator Theory on Noncommutative Domains*. American Mathematical Soc., 2010.

[Pop13] Gelu Popescu. Noncommutative Multivariable Operator Theory. *Integral Equations and Operator Theory*, 75(1):87–133, 2013.

[PP16] Gerlind Plonka and Vlada Pototskaia. Application of the AAK Theory for Sparse Approximation of Exponential Sums, 2016. arXiv:1609.09603.

[PSTS18] Hao Peng, Roy Schwartz, Sam Thomson, and Noah A. Smith. Rational Recurrences. *CoRR*, abs/1808.09357, 2018. URL: http://arxiv.org/abs/1808.09357, arXiv:1808.09357.

[QC19] Ariadna Quattoni and Xavier Carreras. Interpolated Spectral NGram Language Models. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5926–5930. Association for Computational Linguistics, 2019. doi:10.18653/v1/p19-1594.

[QC20] Ariadna Quattoni and Xavier Carreras. A Comparison Between CNNs and WFAs for Sequence Classification. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 159–163, Online, November 2020. Association for Computational Linguistics. URL: https://aclanthology.org/2020.sustainlp-1.21, doi:10.18653/v1/2020.sustainlp-1.21.

[RBA+19] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron C. Courville. On the Spectral Bias of Neural Networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 2019. URL: http://proceedings.mlr.press/v97/rahaman19a.html.

[Rie22] Freidrich Riesz. Über die Randwerte einer analytischen Funktion. *Mathematische Zeitschrift*, 18:87–95, 1922.

[RLP19] Guillaume Rabusseau, Tianyu Li, and Doina Precup. Connecting Weighted Automata and Recurrent Neural Networks through Spectral Learning. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*,

pages 1630–1639. PMLR, 2019. URL: http://proceedings.mlr.press/v89/rabusseau19a.html.

[Rot52] William E. Roth. The Equations AX - YB = C and AX - XB = C in Matrices. *Proceedings of the American Mathematical Society*, 3(3):392–396, 1952.

[RQ13] Adria Recasens and Ariadna Quattoni. Spectral Learning of Sequence Taggers over Continuous Sequences. In *Proc. ECML 2019*, pages 289–304, 2013.

[RSN55] Frigyes Riesz and Béla Szökefalvi-Nagy. *Functional Analysis [by] Frigyes Riesz and Béla Sz.-Nagy. Translated from the 2nd French ed. by Leo F. Boron.* F. Ungar Pub. Co., New York, 1955.

[Sch61] M.P. Schützenberger. On the Definition of a Family of Automata. *Information and Control*, 4(2):245–270, 1961. URL: https://www.sciencedirect.com/science/article/pii/S001999586180020X, doi:https://doi.org/10.1016/S0019-9958(61)80020-X.

[Sch89] Erhard Schmidt. Zur Theorie der linearen und nichtlinearen Integralgleichungen. In *Integralgleichungen und Gleichungen mit unendlich vielen Unbekannten*, pages 190–233. Springer, 1989.

[Sch00] B.De Schutter. Minimal State-Space Realization in Linear System Theory: an Overview. *Journal of Computational and Applied Mathematics*, 121(1):331–354, 2000. doi:10.1016/S0377-0427(00)00341-1.

[SHYS21] Kaito Suzuki, Diptarama Hendrian, Ryo Yoshinaka, and Ayumi Shinohara. Query Learning Algorithm for Symbolic Weighted Finite Automata. In Jane Chandlee, Rémi Eyraud, Jeff Heinz, Adam Jardine, and Menno van Zaanen, editors, *Proceedings of the Fifteenth International Conference on Grammatical Inference*, volume 153 of *Proceedings of Machine Learning Research*, pages 202–

216. PMLR, 23–27 Aug 2021. URL: https://proceedings.mlr.press/v153/suzuki21a.html.

[Smi29] V. Smirnoff. Sur les Valeurs Limites des Fonctions, Régulières à l'Intérieur d'un Cercle. *J. Soc. Phys.-Math. Léningrade*, 2(2):22–37, 1929.

[SRRS21] Ananda Theertha Suresh, Brian Roark, Michael Riley, and Vlad Schogol. Approximating Probabilistic Models as Weighted Finite Automata. *Comput. Linguistics*, 47(2):221–254, 2021. doi:10.1162/coli_a_00401.

[SSS18] Guy Salomon, Orr Shalit, and Eli Shamovich. Algebras of Bounded Noncommutative Analytic Functions on Subvarieties of the Noncommutative Unit Ball. *Transactions of the American Mathematical Society*, 370(12):8639–8690, 2018.

[STS18] Roy Schwartz, Sam Thomson, and Noah A. Smith. SoPa: Bridging CNNs, RNNs, and Weighted Finite-State Machines. *CoRR*, abs/1805.06061, 2018. URL: http://arxiv.org/abs/1805.06061, arXiv:1805.06061.

[Tao12] Terence Tao. *Topics in Random Matrix Theory*, volume 132. American Mathematical Soc., 2012.

[TBI97] Lloyd N Trefethen and David Bau III. *Numerical Linear Algebra*, volume 50. Siam, 1997.

[TGLM20] Neil C. Thompson, Kristjan H. Greenewald, Keeheon Lee, and Gabriel F. Manso. The Computational Limits of Deep Learning. *CoRR*, abs/2007.05558, 2020. URL: https://arxiv.org/abs/2007.05558, arXiv:2007.05558.

[TJ15] Michael R. Thon and Herbert Jaeger. Links between Multiplicity Automata, Observable Operator Models and Predictive State Representations: a Unified Learning Framework. *J. Mach. Learn. Res.*, 16:103–147, 2015. URL: http://dl.acm.org/citation.cfm?id=2789276.

[TV94] Sergei Treil and Alexander Volberg. A Fixed Point Approach to Nehari's Problem and its Applications. In E. L. Basor and I. Gohberg, editors, *Toeplitz Operators and Related Topics*, pages 165–186, Basel, 1994. Birkhäuser Basel.

[TV14] Terence Tao and Van Vu. Random Matrices Have Simple Spectrum, 2014. `arXiv:1412.1438`.

[vNW93] J. von Neumann and E. P. Wigner. *Über das Verhalten von Eigenwerten bei adiabatischen Prozessen*, pages 294–297. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993. `doi:10.1007/978-3-662-02781-3_20`.

[Vol18] Jurij Volčič. Matrix Coefficient Realization Theory of Noncommutative Rational Functions. *Journal of Algebra*, 499:397–437, 2018. URL: `https://www.sciencedirect.com/science/article/pii/S0021869317306634`, `doi:https://doi.org/10.1016/j.jalgebra.2017.12.009`.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All You Need. *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[WGY18a] Gail Weiss, Yoav Goldberg, and Eran Yahav. Extracting Automata from Recurrent Neural Networks Using Queries and Counterexamples. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5244–5253. PMLR, 2018. URL: `http://proceedings.mlr.press/v80/weiss18a.html`.

[WGY18b] Gail Weiss, Yoav Goldberg, and Eran Yahav. On the Practical Computational Power of Finite Precision RNNs for Language Recognition. *CoRR*, 2018. URL: http://arxiv.org/abs/1805.04908, arXiv:1805.04908.

[WGY19] Gail Weiss, Yoav Goldberg, and Eran Yahav. Learning Deterministic Weighted Automata with Queries and Counterexamples. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8558–8569, 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/d3f93e7766e8e1b7ef66dfdd9a8be93b-Abstract.html.

[Wim73] H.K Wimmer. On the Ostrowski-Schneider Inertia Theorem. *Journal of Mathematical Analysis and Applications*, 41(1):164–169, 1973. doi:10.1016/0022-247X(73)90190-X.

[WZI+18] Qinglong Wang, Kaixuan Zhang, Alexander G. Ororbia II, Xinyu Xing, Xue Liu, and C. Lee Giles. An Empirical Evaluation of Rule Extraction from Recurrent Neural Networks. *Neural Comput.*, 30(9), 2018. doi:10.1162/neco_a_01111.

[WZLG18] Qinglong Wang, Kaixuan Zhang, Xue Liu, and C. Lee Giles. Verification of Recurrent Neural Networks Through Rule Extraction. *CoRR*, abs/1811.06029, 2018. URL: http://arxiv.org/abs/1811.06029, arXiv:1811.06029.

[WZOI+18] Qinglong Wang, Kaixuan Zhang, Alexander G. Ororbia II, Xinyu Xing, Xue Liu, and Clyde Lee Giles. An Empirical Evaluation of Rule Extraction From Recurrent Neural Networks. *Neural computation*, 30(9):2568–2591, 2018.

[You83] N.J. Young. The Singular-Value Decomposition of an Infinite Hankel Matrix. *Linear Algebra and its Applications*, 50:639–656, 1983. URL: https://www.sciencedirect.com/science/article/pii/0024379583900733, doi:https://doi.org/10.1016/0024-3795(83)90073-3.

[ZDX+21] Xiyue Zhang, Xiaoning Du, Xiaofei Xie, Lei Ma, Yang Liu, and Meng Sun. Decision-Guided Weighted Automata Extraction from Recurrent Neural Networks. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 11699–11707. AAAI Press, 2021. URL: `https://ojs.aaai.org/index.php/AAAI/article/view/17391`.

[Zhu90] Kehe Zhu. *Operator Theory in Function Spaces*, volume 138. American Mathematical Society, 1990.

Appendix A

Elements of Functional Analysis

A.1 Inner-Outer Factorization

Every function in the Hardy space can be factorized using inner and outer functions.

Theorem A.1.1 (Inner-Outer Factorization [Smi29])**.** *Let $f \in \mathcal{H}^2$, $f \neq 0$. Then there exist an inner function $\theta \in \mathcal{H}^2$ and an outer function $g \in \mathcal{H}^2$ such that:*

$$f = \theta g. \tag{A.1}$$

Moreover, such factorization is unique uo to a constant factor, and $\overline{Span}\{z^n f : n \geq 0\} = \theta\mathcal{H}^2$.

We recall the definition of Blaschke products. The sequence $\{\lambda_i\}_{i\geq 0}$ of points inside the unit disk is said to satisfy the Blaschke condition if $\sum_{n\geq 0}(1 - |\lambda_n|)$. For $\lambda \in \mathbb{D}$, we define a Blaschke factor by:

$$b_\lambda = \frac{|\lambda|}{\lambda}\frac{\lambda - z}{1 - \overline{\lambda}z}. \tag{A.2}$$

Definition A.1.1. *A **Blaschke product** is defined, given a sequence $\{\lambda_i\}_{i\geq 0}$ of points*

inside the unit disk satisfying the Blaschke condition, as the infinite product:

$$B = \Pi_{n\geq 0} b_{\lambda_n} = \lim_{r\to 1} \Pi_{|\lambda_n|<r} b_{\lambda_n}. \tag{A.3}$$

To summarize, a Blaschke product is a bounded analytic function in the open unit disc constructed to have zeros at a sequence of prescribed complex numbers. In the case of uniformly bounded analytic functions in the disc, the inner-outer factorization can be further refined by breaking the inner part into two factors, an inner function without any zero (a **singular inner** function), and an inner function with zeros (a Blaschke product).

Theorem A.1.2 ([Her11, Rie22, Beu49])**.** *Let f be a bounded analytic function in $\mathbb{D}$. Then f admits a Blaschke-singular-outer factorization:*

$$f = B \cdot s \cdot g, \tag{A.4}$$

where B is a Blaschke product, s is singular inner and g is an outer function.

We recall a last result, which follows from Beurling's Theorem. The proof, which relies on properties of the reproducing kernel associated to $\lambda \in \mathbb{D}$, can be found in the book of Nikolski [Nik02].

Theorem A.1.3. *Let $\theta \in \mathcal{H}^2$ be an inner function and let $\mathcal{K}_\theta = \mathcal{H}^2 \ominus \theta\mathcal{H}^2$ be the orthogonal complement of a shift-invariant subspace. The following are equivalent:*

- *$\mathcal{K}_\theta$ is finite dimensional*
- *$\theta = B = \Pi_{n\geq 0} b_{\lambda_n}$ is a finite Blaschke product.*

Moreover $\dim \mathcal{K}_\theta = deg\theta$*, where $deg\theta$ stands for the degree of θ, if θ is a finite Blaschke product, and $deg\theta = \infty$ otherwise.*

A.2 Proof of AAK Theorem

In this section, we summarize the main components of the proof of the AAK theorem [AAK71]. The complete proof can be found in the original article of Adamyan, Arov and Krein [AAK71], or in the book of Nikolski [Nik02, Theorem 7.2.1].

We recall the statement of the AAK theorem.

Theorem A.2.1 (AAK Theorem [AAK71])**.** *Let H_ϕ be a compact Hankel operator of rank n, matrix $\mathbf{H}$ and singular numbers $\sigma_0 \geq \cdots \geq \sigma_{n-1} > 0$. Then there exists a unique Hankel operator H_ψ of rank $k < n$ such that:*

$$\|H_\phi - H_\psi\| = \sigma_k.$$

Before proceeding with proof of the theorem, we need the following lemmas (the proofs can be found in the book of Nikolski [Nik02]).

Lemma A.2.2. *[Nik02, Lemma 7.2.5] Let $\{\boldsymbol{\xi_k}, \boldsymbol{\eta_k}\}$ be a σ_k-Schmidt pair of H_ϕ. Then, the function:*

$$\phi_s = \frac{\eta_k}{\xi_k}$$

is unimodular almost everywhere on the unit circle and does not depend on the choice of the Schmidt pair $\{\boldsymbol{\xi}, \boldsymbol{\eta}\}$.

Lemma A.2.3. *[Nik02, Lemma 7.2.6] Let $\xi_k \in \ker(H_\phi^* H_\phi - \sigma_k^2 1)$ and let Θ be the greatest common divisor of the inner parts of nonzero functions $\xi_k \in \ker(H_\phi^* H_\phi - \sigma_k^2 1)$. Then $\overline{\theta}\xi_k \in \ker(H_{\overline{\theta}\phi}^* H_{\overline{\theta}\phi} - \sigma_k^2 1)$ for every inner divisor θ of Θ.*

Lemma A.2.4. *[Nik02, Lemma 7.2.4] Let A be an operator and let $\lambda > \sigma_\infty(A)$ be a singular number with corresponding multiplicity $\mu + 1$, so that $\lambda = \sigma_k(A) = \sigma_{k+1}(A) = \cdots = \sigma_{k+\mu}(A) > \sigma_{k+\mu+1}(A)$. If $B = UAV$, where U and V are contractions, then:*

$$\dim \ker(B^* B - \lambda^2 1) \leq k + \mu + 1.$$

Lemma A.2.5. *[Nik02, Lemma 7.2.7] Following the notation of Lemma A.2.3, we have:*

$$\dim\ker(H_\phi^* H_\phi - \sigma_k^2 1) + \deg\Theta \leq \dim\overline{Span}\{\overline{\theta}\ker(H_\phi^* H_\phi - \sigma_k^2 1) : \theta \textit{ is an inner divisor of } \Theta\}.$$

Proof of the AAK Theorem. We assume $\sigma_k > \sigma_\infty$. The proof can be divided in three steps.

(1) *Define the symbol of the best approximation.*

Let $\xi_k \in \ker(H_\phi^* H_\phi - \sigma_k^2 1)$ be a σ_k function. We set:

$$\psi = \phi - \frac{H_\phi \xi_k}{\xi_k}.$$

It is possible to prove that the function $\psi \in \mathcal{L}^\infty$ does not depend on the particular choice of ξ_k.

(2) *Show that it is the optimal approximation:* $\|H_\phi - H_\psi\| \leq \sigma_k$

It is easy to see that:

$$\|H_\phi - H_\psi\| = \left\| H_{\frac{H_\phi \xi_k}{\xi_k}} \right\| \leq \sigma_k \left\| \frac{H_\phi \xi_k}{\sigma_k \xi_k} \right\|_\infty \leq \sigma_k$$

where the last inequality follows from the fact that the function $\frac{H_\phi \xi_k}{\sigma_k \xi_k}$ is unimodular almost everywhere on the unit circle (Lemma A.2.2).

(3) *Show that it has the right size:* $\operatorname{rank}(H_\psi) \leq k$.

We start by noting that, for any nonzero $\xi_k \in \ker(H_\phi^* H_\phi - \sigma_k^2 1)$, we have:

$$H_\psi \xi_k = \mathbb{P}_- \left(\phi \xi_k - \frac{H_\phi \xi_k}{\xi_k} \xi_k \right) = \mathbb{P}_- \phi \xi_k - H_\phi \xi_k = 0.$$

It follows that $\ker(H_\phi^* H_\phi - \sigma_k^2 1) \subset \ker(H_\psi)$. Since the kernel of a Hankel operator is a

shift-invariant space, by Theorem 2.4.5 follows that:

$$\Theta\mathcal{H}^2 \subset \ker H_\psi$$

where Θ is the greatest common divisor of the inner parts of nonzero functions $\xi_k \in \ker(H_\phi^* H_\phi - \sigma_k^2 1)$. Therefore, we have:

$$\operatorname{rank} H_\psi = \dim H_\psi(\mathcal{H}^2) = \dim(\ker H_\psi)^\perp \leq \dim(\Theta\mathcal{H}^2)^\perp = \deg\Theta.$$

To conclude the proof it is enough to show that $\deg\Theta \leq k$. Let $A = H_\phi$, let $U = 1$ and $V = \Theta$ be the operator of multiplication by Θ. Note that $UAV = H_\phi\Theta = H_{\phi\Theta}$, so applying Lemma A.2.4 we obtain:

$$\dim\ker(H_{\phi\Theta}^* H_{\phi\Theta} - \sigma_k^2 1) \leq k + \mu + 1$$

where $\mu + 1 = \dim\ker((H_\phi^* H_\phi - \sigma_k^2 1)$. This means that $H_{\phi\Theta}$ and H_ϕ both have singular value σ_k, and they coincide in the shift-invariant space spanned by $\ker((H_\phi^* H_\phi - \sigma_k^2 1)$. Now, by Lemma A.2.3 we get:

$$\mathcal{L} := \overline{\operatorname{Span}}\{\overline{\theta}\ker(H_\phi^* H_\phi - \sigma_k^2 1) : \theta \text{ is an inner divisor of } \Theta\} \subset \ker(H_{\phi\Theta}^* H_{\phi\Theta} - \sigma_k^2 1).$$

Applying Lemma A.2.5 we obtain:

$$\dim\ker(H_\phi^* H_\phi - \sigma_k^2 1) + \deg\Theta \leq \dim\mathcal{L} \leq \dim\ker(H_{\phi\Theta}^* H_{\phi\Theta} - \sigma_k^2 1)$$

which proves that $\deg\Theta \leq k$.

□

Appendix B

Proofs

B.1 Proofs of Chapter 4

Proof of Theorem 4.2.2. In order to prove Theorem 4.2.2 we need an auxiliary lemma [CC97, Lemma 6.1]. These are the analogous of some control theory results, rephrased in terms of WFAs. The original theorem and lemma, together with the corresponding proofs, can be found in [CC97]. Hence, we only provide a sketch of the proofs.

Lemma B.1.1 ([CC97])**.** *Let* $E = \langle \boldsymbol{\alpha}_e, \mathbf{A}_e, \boldsymbol{\beta}_e \rangle$ *be a minimal WFA. Let* $e(z) = \boldsymbol{\alpha}_e^\top (z\mathbf{1} - \mathbf{A}_e)^{-1}\boldsymbol{\beta}_e - C$*, if* $\sigma_k^{-1} e(z)$ *is unimodular, then there exist a unique invertible symmetric matrix* $\mathbf{T}$ *satisfying:*

(a) $\mathbf{A}_e^\top \mathbf{T} \boldsymbol{\beta}_e = \boldsymbol{\alpha}_e C$

(b) $\sigma_k^2 \boldsymbol{\alpha}_e^\top \mathbf{T}^{-1} \mathbf{A}_e^\top = C \boldsymbol{\beta}_e^\top$

(c) $\mathbf{A}_e^\top \mathbf{T} \mathbf{A}_e - C^{-1} \mathbf{A}_e^\top \mathbf{T} \boldsymbol{\beta}_e \boldsymbol{\alpha}_e^\top = \mathbf{T}$

Proof. Since $\sigma_k^{-1} e(z)$ is unimodular, we have that:

$$e(z) e^*(\bar{z}^{-1}) = \sigma_k^2 \mathbf{1} \tag{B.1}$$

where we denote with e^* the adjoint function. From the equation above, we obtain:

$$e^*(\bar{z}^{-1}) = \sigma_k^2 e^{-1}(z) = \sigma_k^2(-C + \boldsymbol{\alpha}_e^\top(z\mathbf{1} - \mathbf{A}_e)^{-1}\boldsymbol{\beta}_e)^{-1} \tag{B.2}$$

$$= -\sigma_k^2 C^{-1} - \sigma_k^2 C^{-1}\boldsymbol{\alpha}_e^\top((z\mathbf{1} - (\mathbf{A}_e + C^{-1}\boldsymbol{\beta}_e\boldsymbol{\alpha}_e))^{-1}\boldsymbol{\beta}_e C^{-1} \tag{B.3}$$

where we used the matrix inversion lemma. On the other hand we have:

$$e^*(\bar{z}^{-1}) = -C + \boldsymbol{\beta}_e^\top(z^{-1}\mathbf{1} - \mathbf{A}_e^\top)^{-1}\boldsymbol{\alpha}_e \tag{B.4}$$

$$= -C + \boldsymbol{\beta}_e^\top(-\mathbf{A}_e^{-\top}(\mathbf{1} - z\mathbf{A}_e^\top) + \mathbf{A}_e^{-\top})(\mathbf{1} - z\mathbf{A}_e^\top)^{-1}\boldsymbol{\alpha}_e \tag{B.5}$$

$$= -(C - \boldsymbol{\beta}_e^\top\mathbf{A}_e^{-\top}\boldsymbol{\alpha}_e) - \boldsymbol{\beta}_e^\top\mathbf{A}_e^{-\top}(z\mathbf{1} - \mathbf{A}_e^{-\top})^{-1}\mathbf{A}_e^{-\top}\boldsymbol{\alpha}_e \tag{B.6}$$

where we used again the matrix inversion lemma before grouping the terms. If the quantities in Equation B.3 and Equation B.6 have to be equal, we need their constant term to be the same. Then, we want the $\mathcal{H}_-^\infty$-components to correspond, so we consider the corresponding Hankel matrices. It is easy to see that we can once again associate the coefficients of these complex functions to the parameters of a WFA. From the minimality of E we obtain:

$$\begin{cases} \sigma_k^2 C^{-1}\boldsymbol{\alpha}_e^\top = \boldsymbol{\beta}_e^\top\mathbf{A}_e^{-\top}\mathbf{T} \\ \mathbf{A}_e + C^{-1}\boldsymbol{\beta}_e\boldsymbol{\alpha}_e = \mathbf{T}^{-1}\mathbf{A}_e^{-\top}\mathbf{T} \\ \boldsymbol{\beta}_e C^{-1} = \mathbf{T}^{-1}\mathbf{A}_e^{-\top}\boldsymbol{\alpha}_e \end{cases} \tag{B.7}$$

where $\mathbf{T}$ is an invertible matrix [BCLQ14]. This system is equivalent to:

$$\begin{cases} \sigma_k^2\boldsymbol{\alpha}_e^\top\mathbf{T}^{-1}\mathbf{A}_e^\top = C\boldsymbol{\beta}_e^\top \\ \mathbf{A}_e^\top\mathbf{T}\mathbf{A}_e - C^{-1}\mathbf{A}_e^\top\mathbf{T}\boldsymbol{\beta}_e\boldsymbol{\alpha}_e^\top = \mathbf{T} \\ \mathbf{A}_e^\top\mathbf{T}\boldsymbol{\beta}_e = \boldsymbol{\alpha}_e C \end{cases} \tag{B.8}$$

To conclude the proof it remains to check that $\mathbf{T}$ is symmetric, and this can be done by

direct computations. □

Proof of Theorem 4.2.2. This proof follows easily from Lemma B.1.1 by setting $\mathbf{P} = -\sigma_k^2 \mathbf{T}^{-1}$ and $\mathbf{Q} = -\mathbf{T}$. We obtain point (c) by direct multiplication. Then, we substitute the last equation in B.8 into the second one, and we obtain:

$$\mathbf{A}_e^\top \mathbf{T} \mathbf{A}_e - \boldsymbol{\alpha}_e \boldsymbol{\alpha}_e^\top = \mathbf{T} \tag{B.9}$$

which verifies point (b) with $\mathbf{Q} = -\mathbf{T}$. Point (a) can be obtained analogously combining the first and second equations in B.8. □

B.2 Proofs of Chapter 5

Riesz Eigenvalues Inequality

The following result is used in the proof of Theorem 5.2.4 and to establish the relation between the singular value of the original Hankel matrix and the one of the truncation. This result is due to [RSN55] (see [KG69] for the proof and for a more general version of this theorem).

Lemma B.2.1 ([RSN55])**.** *Let T, S be two self-adjoint compact operators, and let σ_k^T, σ_k^S for $k \geq 0$ be their singular numbers. Then:*

$$|\sigma_k^T - \sigma_k^S| \leq \|\mathbf{S} - \mathbf{T}\|. \tag{B.10}$$

We remark that, while the original statement is about eigenvalues, in our setting this holds automatically for singular values.

Cauchy's Interlace Theorem

Cauchy's Interlace Theorem is used in the proof of Theorem 5.2.4. We refer the reader to [Hwa04] for a proof this theorem.

Theorem B.2.2 (Cauchy's Interlace Theorem)**.** *Let* $\mathbf{A}$ *be a* $n \times n$ *Hermitian matrix, let* $\mathbf{B}$ *be the principal submatrix of* $\mathbf{A}$ *of order* $(n-1) \times (n-1)$*. If* $\lambda_0 \geq \cdots \geq \lambda_{n-1}$ *are the eigenvalues of* $\mathbf{A}$*, and* $\mu_0 \geq \cdots \geq \mu_{n-2}$ *are the eigenvalues of* $\mathbf{B}$*, then:*

$$\lambda_0 \geq \mu_0 \geq \lambda_1 \geq \mu_1 \geq \ldots \lambda_{n-2} \geq \mu_{n-2} \geq \lambda_{n-1}. \tag{B.11}$$

Rank of the Principal Submatrix

We conclude by recalling the result of [Al'17], which shows the relation between the rank of the Hankel matrix and the one of its leading principal submatrix.

Theorem B.2.3 ([Al'17])**.** *A Hankel matrix* $\mathbf{H}$ *has a finite rank* r *if and only if the first* r *rows of* $\mathbf{H}$ *are linearly independent, and generate the row* $r+1$ *as a linear combination.*