PAPER

Optimal approximate minimization of one-letter weighted finite automata

Clara Lacroce[1,2] (iD), Borja Balle[3], Prakash Panangaden[1,2] (iD) and Guillaume Rabusseau[2,4,5] (iD)

[1]School of Computer Science, McGill University, Montréal, Canada, [2]Mila, Montréal, Canada, [3]Department of Research, Google DeepMind, London, UK, [4]DIRO, Université de Montréal, Montréal, Canada and [5]CIFAR AI Chair, Canada
Corresponding author: Clara Lacroce; Email: clara.lacroce@mail.mcgill.ca

Abstract

In this paper, we study the approximate minimization problem of weighted finite automata (WFAs): to compute the best possible approximation of a WFA given a bound on the number of states. By reformulating the problem in terms of Hankel matrices, we leverage classical results on the approximation of Hankel operators, namely the celebrated Adamyan-Arov-Krein (AAK) theory. We solve the optimal spectral-norm approximate minimization problem for irredundant WFAs with real weights, defined over a one-letter alphabet. We present a theoretical analysis based on AAK theory and bounds on the quality of the approximation in the spectral norm and ℓ^2 norm. Moreover, we provide a closed-form solution, and an algorithm, to compute the optimal approximation of a given size in polynomial time.

1. Introduction

Weighted finite automata (WFAs) are an expressive class of models representing functions defined over sequences. The *approximate minimization problem* is concerned with finding an automaton that approximates the behavior of a given minimal WFA, while being smaller in size. Clearly, the two automata compute different languages, so the objective is to minimize the approximation error (Balle et al., 2015, 2019). Approximate minimization can be particularly useful in the context of spectral learning algorithms (Bailly et al., 2009; Hsu et al., 2012; Balle et al., 2014a; Hsu et al., 2012). When applied to a learning task, such algorithms can be viewed as working in two steps. First, they compute a minimal WFA that explains the training data exactly. Then, they obtain a model that generalizes to the unseen data by producing a smaller approximation to the minimal WFA, thus preventing overfitting of the data.

A key point in solving approximation tasks is to choose how to quantify the error. We propose to rewrite the problem in terms of Hankel matrices, mathematical objects strictly related to WFAs, and to measure the error in terms of the *spectral norm*. This allows us to exploit the work of Adamyan, Arov, and Krein which has come to be known as AAK theory (Adamyan et al., 1971): a series of results connecting the theory of complex functions to Hankel matrices. The core of this theory provides us with theoretical guarantees for the exact computation of the spectral norm of the error and a method to construct the optimal approximation. We show that the spectral norm of the Hankel matrix of a WFA can be computed accurately in polynomial time (cubic in the

number of states of the automaton). This is a great advantage compared, for example, to behavioral norms, which are easier to interpret but harder to compute (Balle et al., 2017, 2022). The spectral norm has another advantage over WFA-specific behavioral metrics. In fact, an important extension of this work is the application of the method to other classes of models. In the one-letter case, a similar algorithm can be found to approximate a black-box model over sequential data using a WFA (Lacroce et al., 2021). With this in mind, we think that it is preferable to consider a norm defined on the input-output function – or the Hankel matrix – rather than the parameters of the specific model considered.

We summarize our main contributions:

- We apply AAK theory to the approximate minimization problem of WFAs by establishing a correspondence between the parameters of a WFA and the coefficients of a complex function on the unit circle. To the best of our knowledge, this paper represents the first attempt to apply AAK theory to WFAs.
- We present a theoretical analysis of the optimal spectral-norm approximate minimization problem of 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 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).
- We propose a self-contained algorithm that returns the unique optimal spectral-norm approximation of a given size in polynomial time.
- We tighten the connection, made in (Balle et al. (2019)), between WFAs and discrete dynamical systems, by adapting some of the control theory concepts to this setting, *for example* the *all-pass system* (Glover, 1984).

In this paper, we present and expand the results of our previous work (Balle et al., 2021). The contents of this paper are organized as follows. In Section 2, we define the notation that will be used throughout the paper and review a series of well-known results from the theory of automata and from functional analysis. In Section 3, we establish the framework to reformulate the approximate minimization problem in terms of Hankel operators and AAK theory. Section 4 presents the theoretical foundation of our contribution and a closed-form solution for our problem. Section 5 shows how to implement the algorithm derived from the solution obtained in the previous section, while in Section 6 we provide an example and compute the optimal approximation of a given WFA. Section 7 discusses the related work in approximate minimization and control theory. Finally, in Sections 8 and 9 we highlight possible directions for future work, analyze the limitations of this approach, and summarize our contribution.

2. Background

In this section, we recall the fundamental definitions and preliminary results that are used throughout the paper. After defining weighted finite automata and Hankel matrices, we will provide an overview of AAK theory. We will see in the next section that our objective is to rewrite the approximate minimization problems as low-rank approximation of a Hankel matrix. In the paper, We use AAK theory to solve the low-rank approximation problem while preserving the Hankel property.

2.1 Preliminaries

We denote with $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{C}$ the set of natural, integers, real and complex numbers, respectively. We use bold letters for vectors and matrices; all vectors considered are column vectors. We denote

with **1** the identity matrix, specifying its dimension only when not clear from the context. We refer to the i-th row and the j-th column of **M** by $\mathbf{M}(i,:)$ and $\mathbf{M}(:,j)$. 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 $\mathrm{rank}(\mathbf{M}) = \mathrm{rank}(\mathbf{P}) = \mathrm{rank}(\mathbf{Q}) = n$. Let $\mathbf{M} \in \mathbb{R}^{p \times q}$ of rank n, the compact *singular value decomposition* SVD of **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 **D** is a diagonal matrix. The columns of **U** and **V** are called left and right *singular vectors*, while the entries of **D** are the *singular values*. The *Moore-Penrose pseudo-inverse* $\mathbf{M}^{+}$ of **M** is the unique matrix such that $\mathbf{MM}^{+}\mathbf{M} = \mathbf{M}$, $\mathbf{M}^{+}\mathbf{MM}^{+} = \mathbf{M}^{+}$, with $\mathbf{M}^{+}\mathbf{M}$ and $\mathbf{MM}^{+}$ Hermitian (Zhu, 1990). The *spectral radius* $\rho(\mathbf{M})$ of a matrix **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. A linear operator $T : X \to Y$ between Hilbert spaces is *bounded* if it has finite operator norm, *that is* $\|T\|_{op} = \sup_{\|g\|_X \le 1} \|Tg\|_Y < \infty$. We denote by **T** the (infinite) matrix associated with T by some (canonical) orthonormal basis on H. An operator is *compact* if the image of the unit ball in X is relatively compact. Given Hilbert spaces X, Y and a compact operator $T : X \to Y$, we denote its adjoint by T^{*}. The *singular numbers* $\{\sigma_n\}_{n \ge 0}$ of T are the square roots of the eigenvalues of the self-adjoint operator $T^{*}T$, arranged in decreasing order. A σ-*Schmidt pair* $\{\boldsymbol{\xi}, \boldsymbol{\eta}\}$ for T is a couple of norm 1 vectors such that: $\mathbf{T}\boldsymbol{\xi} = \sigma\boldsymbol{\eta}$ and $\mathbf{T}^{*}\boldsymbol{\eta} = \sigma\boldsymbol{\xi}$. The Hilbert-Schmidt decomposition provides a generalization of the compact SVD for the infinite matrix of a compact operator T using singular numbers and orthonormal Schmidt pairs: $\mathbf{Tx} = \sum_{n \ge 0} \sigma_n \langle \mathbf{x}, \boldsymbol{\xi}_n \rangle \boldsymbol{\eta}_k$ (Zhu, 1990). The *spectral norm* $\|\mathbf{T}\|$ of the matrix representing the operator T is the largest singular number of T. Note that the spectral norm of **T** corresponds to the operator norm of T.

Let ℓ^2 be the Hilbert space of square-summable sequences over Σ^{*}, with norm $\|f\|_2^2 = \sum_{x \in \Sigma^{*}} |f(x)|^2$ 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 $1 < p < \infty$, $\mathscr{L}^p(\mathbb{T})$ 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 $\mathscr{L}^{\infty}(\mathbb{T})$ the space of measurable functions that are bounded, with norm $\|f\|_{\infty} = \sup\{|f(x)| : x \in \mathbb{T}\}$.

2.2 Hankel matrix and weighted automata

Let Σ be a fixed finite alphabet and Σ^{*} be the set of all finite strings with symbols in Σ. We denote with ε the empty string. Given $p, s \in \Sigma^{*}$, we denote with ps the string obtained by their concatenation. Let $f : \Sigma^{*} \to \mathbb{R}$ be a function defined on sequences, we consider the 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 1. *A 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* $\mathbf{H}(p, s) = \mathbf{H}(p', s')$.

Given a Hankel matrix $\mathbf{H} \in \mathbb{R}^{\Sigma^{*} \times \Sigma^{*}}$, there is a unique function $f : \Sigma^{*} \to \mathbb{R}$ such that $\mathbf{H}_f = \mathbf{H}$. Intuitively, the Hankel property tells us that each entry of the matrix only depends on the composition of the coordinates. Since rows and columns are indexed using strings, then the value stored in each entry only depends on the string obtained by concatenating the coordinates.

Weighted finite automata are a class of models defined over sequential data. A *weighted finite automaton* (WFA) of n states over Σ is a tuple $A = \langle \boldsymbol{\alpha}, \{A_a\}_{a \in \Sigma}, \boldsymbol{\beta} \rangle$, where $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$ are the vector of initial and final weights, respectively, and $\mathbf{A}_a \in \mathbb{R}^{n \times n}$ is the matrix containing the transition weights associated with each symbol $a \in \Sigma$. Every WFA A with real weights realizes (or computes) a function $f_A : \Sigma^{*} \to \mathbb{R}$, *that is* given a string $x = x_1 \cdots x_t \in \Sigma^{*}$, it returns $f_A(x) = \boldsymbol{\alpha}^{\top}\mathbf{A}_{x_1} \cdots \mathbf{A}_{x_t}\boldsymbol{\beta} = \boldsymbol{\alpha}^{\top}\mathbf{A}_x\boldsymbol{\beta}$. A function $f : \Sigma^{*} \to \mathbb{R}$ is called *rational* if there exists a WFA A that

realizes it. Given $f : \Sigma^* \to \mathbb{R}$, we can use the Hankel matrix $\mathbf{H}_f \in \mathbb{R}^{\Sigma^* \times \Sigma^*}$ to recover information about the weighted automaton computing f.

Theorem 2 (Carlyle and Paz (1971); Fliess (1974)). *A function $f : \Sigma^* \to \mathbb{R}$ is realized by a WFA A if and only if $\mathbf{H}_f$ has finite rank. In that case, the rank of $\mathbf{H}_f$ corresponds to the minimal number of states of any automaton realizing f.*

Given a WFA $A = \langle \boldsymbol{\alpha}, \{A_a\}_{a \in \Sigma}, \boldsymbol{\beta} \rangle$, the *forward matrix* of A is the infinite matrix $\mathbf{F}_A \in \mathbb{R}^{\Sigma^* \times n}$ 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: $\mathbf{H}_f = \mathbf{F} \mathbf{B}^\top$. 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. If A is minimal, the (unique) FB factorization is a rank factorization (Balle et al., 2014a).

We recall the definition of the singular value automaton, a canonical form for WFAs (Balle et al., 2015).

Definition 3. *Let $f : \Sigma^* \to \mathbb{R}$ be a rational function and suppose $\mathbf{H}_f$ admits an 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 $\mathbf{B}_A = \mathbf{V} \mathbf{D}^{1/2}$.*

The SVA can be computed with an efficient algorithm relying on the following matrices (Balle et al., 2019).

Definition 4. *Let $f : \Sigma^* \to \mathbb{R}$ be a rational function, $\mathbf{H}_f = \mathbf{F} \mathbf{B}^\top$ the FB factorization. If the matrices $\mathbf{P} = \mathbf{F}^\top \mathbf{F}$ and $\mathbf{Q} = \mathbf{B}^\top \mathbf{B}$ are well defined (i.e., the inner products of their columns are finite for any column), we call $\mathbf{P}$ the **reachability Gramian** and $\mathbf{Q}$ the **observability Gramian**.*

If A is in its SVA form, the Gramians associated with its FB factorization satisfy $\mathbf{P}_A = \mathbf{Q}_A = \mathbf{D}$, where $\mathbf{D}$ is the matrix of singular values of the corresponding Hankel matrix. The Gramians can alternatively be characterized and computed (Balle et al., 2019)) using fixed point equations, corresponding to Lyapunov equations when $|\Sigma| = 1$ (Lyapunov, 1950).

Theorem 5. *Let $|\Sigma| = 1$, $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ a WFA with n states and well-defined Gramians $\mathbf{P}, \mathbf{Q}$. Then $X = \mathbf{P}$ and $Y = \mathbf{Q}$ solve the equations $X - \mathbf{A} X \mathbf{A}^\top = \boldsymbol{\beta} \boldsymbol{\beta}^\top$ and $Y - \mathbf{A}^\top Y \mathbf{A} = \boldsymbol{\alpha} \boldsymbol{\alpha}^\top$.*

Finally, we recall the definition of *generative probabilistic automata* (GPA). A WFA $A = \langle \boldsymbol{\alpha}, \{A_a\}_{a \in \Sigma}, \boldsymbol{\beta} \rangle$ is a GPA if $f_A(x) \geq 0$ for every x and $\sum_{x \in \Sigma^*} f_A(x) = 1$, *that is if f_A computes a probability distribution over Σ^**. In general, this class of automata can contain pathological examples having 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 6. *Given a WFA $A = \langle \boldsymbol{\alpha}, \{A_a\}_{a \in \Sigma}, \boldsymbol{\beta} \rangle$, let $\mathbf{A} = \sum_{a \in \Sigma} \mathbf{A}_a$. The WFA A is **irredundant** if $\rho(\mathbf{A}) < 1$.*

2.3 AAK theory

In this section, we introduce the theory of optimal approximation for Hankel operators and complex functions known as AAK theory (Adamyan et al., 1971). A comprehensive presentation of the concepts recalled in this section can be found in Nikol'Skii (2002); Peller (2012).

We consider the space of square-integrable complex functions on the unit circle $\mathscr{L}^2(\mathbb{T})$. To avoid any confusion with functions defined over sequences, when dealing with complex function we make explicit the dependence on the complex variable $z = e^{it}$. Note that a function $\phi(z) \in \mathscr{L}^2(\mathbb{T})$ can be represented, using the orthonormal basis $\{z^n\}_{n \in \mathbb{Z}}$, by means of its Fourier series: $\phi(z) = \sum_{n \in \mathbb{Z}} \widehat{\phi}(n) z^n$, with Fourier coefficients $\widehat{\phi}(n) = \int_{\mathbb{T}} \phi(z) \bar{z}^n dz$, $n \in \mathbb{Z}$. Thus, we can partition the function space $\mathscr{L}^2(\mathbb{T})$ into two subspaces.

Definition 7. *The **Hardy space** $\mathscr{H}^2$ and the **negative Hardy space** $\mathscr{H}_-^2$ on $\mathbb{T}$ are the subspaces of $\mathscr{L}^2(\mathbb{T})$ defined as:*

$$\mathscr{H}^2 = \left\{ \phi(z) \in \mathscr{L}^2(\mathbb{T}) : \widehat{\phi}(n) = 0, n < 0 \right\},$$

$$\mathscr{H}_-^2 = \left\{ \phi(z) \in \mathscr{L}^2(\mathbb{T}) : \widehat{\phi}(n) = 0, n \geq 0 \right\}.$$

Interestingly, the elements of the Hardy space can be canonically identified with the set of functions analytic in the unit disc $\mathbb{D}$, with the property that the square of their absolute value is integrable on $\mathbb{T}$ (a proof can be found in Nikol'Skii (2002)). Thus, we will make no difference between these functions in the unit disc and their boundary value on the circle. Moreover, we remark that the definition of Hardy space can be generalized for any p-th power of the functions' absolute value, for $0 < p \leq \infty$.

We define Hankel operators in the Hardy spaces.

Definition 8. *Let $\phi(z)$ be a function in the space $\mathscr{L}^2(\mathbb{T})$. A **Hankel operator** is an operator $H_\phi : \mathscr{H}^2 \to \mathscr{H}_-^2$ defined by $H_\phi f(z) = \mathbb{P}_- \phi f(z)$, where $\mathbb{P}_-$ is the orthogonal projection from $\mathscr{L}^2(\mathbb{T})$ onto $\mathscr{H}_-^2$. The function $\phi(z)$ is called a **symbol** of the Hankel operator H_ϕ.*

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

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

Note that this matrix satisfies the Hankel property, as each entry only depends on the composition of the corresponding coordinates.

Nehari's theorem (Nehari, 1957), characterizes bounded Hankel operators and their norm.

Theorem 9 (Nehari (1957)). *Let $\phi \in \mathscr{L}^2(\mathbb{T})$ be a symbol of the Hankel operator on Hardy spaces $H_\phi : \mathscr{H}^2 \to \mathscr{H}_-^2$. Then, H_ϕ is bounded on $\mathscr{H}^2$ if and only if there exists $\psi \in \mathscr{L}^\infty(\mathbb{T})$ such that $\widehat{\psi}(m) = \widehat{\phi}(m)$ for all $m < 0$. If the conditions above are satisfied, then:*

$$\|H_\phi\| = \inf \left\{ \|\psi\|_\infty : \widehat{\psi}(m) = \widehat{\phi}(m), \ m < 0 \right\}. \tag{2}$$

As a consequence of Theorem 9, if H_ϕ is a bounded operator, we can consider without loss of generality $\phi(z) \in \mathscr{L}^\infty(\mathbb{T})$. We remark that a Hankel operator has infinitely many different symbols, since $H_\phi = H_{\phi+\psi}$ for $\psi(z) \in \mathscr{H}^\infty$.

Definition 10. *The complex function $\phi(z)$ is **rational** if $\phi(z) = p(z)/q(z)$, with $p(z)$ and $q(z)$ polynomials. The rank of $\phi(z)$ is the maximum between the degrees of $p(z)$ and $q(z)$. A rational function is **strictly proper** if the degree of $p(z)$ is strictly smaller than that of $q(z)$.*

The following result of Kronecker relates finite-rank infinite Hankel matrices to rational functions.

Theorem 11 (Kronecker (1881)). *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 (with multiplicities) of $\mathbb{P}_-\phi$ inside the unit disc.*

We are ready to state the main result of Adamyan et al. (1971). The theorem shows that for infinite dimensional Hankel matrices the constraint of preserving the Hankel property does not affect the achievable approximation error.

Theorem 12 (Adamyan et al. (1971)). *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:*

$$\|H_\phi - H_g\| = \|\mathbf{H} - \mathbf{G}\| = \sigma_k. \tag{3}$$

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

$$\mathscr{H}_k^\infty = \left\{ \psi \in \mathscr{L}^\infty(\mathbb{T}) : \ \exists g \in \mathscr{R}_k, \exists l \in \mathscr{H}^\infty, \ \psi = g + l \right\}. \tag{4}$$

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

Theorem 13 (Adamyan et al. (1971)). *Let $\phi \in \mathscr{L}^\infty(\mathbb{T})$. Then there exists a complex function $\psi \in \mathscr{H}_k^\infty$ such that:*

$$\|\phi - \psi\|_\infty = \sigma_k(H_\phi). \tag{5}$$

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 a corollary the critical steps of the proof, that allows us to find the best approximating symbol.

Corollary 14. *Let ϕ and $\{\boldsymbol{\xi}_k, \boldsymbol{\eta}_k\}$ be a symbol and a σ_k-Schmidt pair for H_ϕ. A function $\psi \in \mathscr{L}^\infty(\mathbb{T})$ is the best AAK approximation according to Theorem 13, if and only if:*

$$(\phi - \psi)\xi_k^+ = \sigma_k \eta_k^-. \tag{6}$$

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 12 and 13 are strictly related.

Corollary 15. *Let $\psi \in \mathscr{H}_k^\infty$, with $\psi = l + g$, $g \in \mathscr{R}_k$, $l \in \mathscr{H}^\infty$. If ψ solves Equation 5, then H_g is the unique Hankel operator from Theorem 12.*

In particular, this means that to find the Hankel operator H_g corresponding to the optimal approximation, we can first obtain ψ by applying Corollary 14. Then, we can extract the rational component g of ψ: this will correspond to a symbol for $\mathbf{H}_g$.

3. AAK Theory and Approximate Minimization

Theorem 2 establishes a correspondence between a given minimal WFA A with n states and a Hankel matrix $\mathbf{H}$ of rank n. The relation between rank and number of states is what motivates our choice to reformulate the approximate minimization problem as low-rank approximation of the Hankel matrix. The approach that we propose to approximate A is to find the minimal WFA corresponding to the Hankel matrix that minimizes $\mathbf{H}$ optimally in the spectral norm. We recall the fundamental result of Eckart and Young (1936).

Theorem 16 (Eckart and Young (1936)). *Let $\mathbf{H}$ be a Hankel matrix corresponding to a compact Hankel operator of rank n, and σ_m, with $0 \leq m < n$ and $\sigma_0 \geq \cdots \geq \sigma_{n-1} > 0$, its singular numbers. Then, if $\mathbf{R}$ is a matrix of rank k, we have: $\|\mathbf{H} - \mathbf{R}\| \geq \sigma_k$. The equality is attained when $\mathbf{R}$ corresponds to the truncated SVD of $\mathbf{H}$.*

In the following example, we compute the low-rank approximation of a finite Hankel matrix using the truncated SVD.

Example 1. *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:

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

It is easy to see that the low-rank approximation of a Hankel matrix obtained by truncating its SVD is not in general a Hankel matrix. This is problematic, since the low-rank approximation needs to be a Hankel matrix in order to correspond to a WFA. On the other hand, we have seen that, by applying AAK theory, we can find the optimal Hankel matrix minimizing the (Hankel) matrix of a Hankel operator in the Hardy spaces. Our objective is to find a way to apply AAK theory to solve the approximate minimization problem of WFAs. To do this, we need an appropriate framework to reformulate this task in terms of Hankel operators and complex functions.

3.1 Defining a Hankel operator: the one-letter assumption

As a first step, we want to understand whether or not a Hankel operator on the Hardy space can be associated with the Hankel matrix of a weighted automaton. To do so, we compare the Hankel matrix $\mathbf{H}_f$ of a WFA realizing a function f over an alphabet Σ, to the Hankel matrix $\mathbf{H}_\phi$ of a

Hankel operator in the Hardy space:

$$\mathbf{H}_f = \begin{pmatrix} f(\varepsilon) & f(a) & f(b) & \dots \\ f(a) & f(aa) & f(ab) & \dots \\ f(b) & f(ba) & f(bb) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \qquad \mathbf{H}_\phi = \begin{pmatrix} \widehat{\phi}(-1) & \widehat{\phi}(-2) & \widehat{\phi}(-3) & \dots \\ \widehat{\phi}(-2) & \widehat{\phi}(-3) & \widehat{\phi}(-4) & \dots \\ \widehat{\phi}(-3) & \widehat{\phi}(-4) & \widehat{\phi}(-5) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{7}$$

We remark that the columns and rows of $\mathbf{H}_f$ are indexed using the letters of the alphabet Σ:

$$\mathbf{H}_f(p, s) = f(ps) \quad \text{for } p, s \in \Sigma,$$

while in the case of $\mathbf{H}_\phi$, the entries are indexed using natural numbers

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

If we think of the intuitive definition of the Hankel property presented in the previous section, we have that it holds in both cases the entries of the matrices only depend on the composition of the coordinate. Note that "composition" means concatenation of letters in the first case, and sum of numbers in the second one. One fundamental difference is that adding natural numbers is a commutative operation, while concatenating letters is not. For example, while for the matrix corresponding to a Hankel operator in the Hardy space we have:

$$\mathbf{H}_\phi(0, 1) = \widehat{\phi}(-2) = \mathbf{H}_\phi(1, 0),$$

in the case of the WFA's matrix, this is not true:

$$\mathbf{H}_f(a, b) = f(ab) \neq f(ba) = \mathbf{H}_f(b, a).$$

This fact reflects in the much stronger structural property satisfied by Hankel of matrices in the Hardy spaces, where the Hankel property implies that the anti-diagonals have constant entries. This property is not reflected by the matrix of an arbitrary WFA, so it is not always possible to associate a Hankel operator to an automaton over an alphabet of arbitrary size, and AAK theory cannot be generally applied. The only case in which concatenation of strings is commutative is when we are restricting our focus on alphabets of one letter. In particular, when $|\Sigma| = 1$, the set of strings Σ^* can be identified with $\mathbb{N}$. Therefore, the function $f : \Sigma^* \to \mathbb{R}$ recognized by a minimal WFA can be rewritten as $f : \mathbb{N} \to \mathbb{R}$, and the Hankel matrix $\mathbf{H}_f$ associated with it can be interpreted as the matrix of a Hankel operator between sequences $H_f : \ell^2 \to \ell^2$. In this case, the Hankel matrix is defined by $\mathbf{H}(i, j) = f(i + j)$, for $i, j \geq 0$:

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

Using the Fourier isomorphism, we can interpret $\mathbf{H}_f$ as the matrix $\mathbf{H}_\phi$ of a Hankel operator over Hardy spaces, associated with a complex function $\phi \in \mathscr{L}^2(\mathbb{T})$. In particular, we can embed the sequence space ℓ^2 into $\ell^2(\mathbb{Z})$ by "duplicating" each vector, that is by associating $\boldsymbol{\mu} = (\mu_0, \mu_1, \dots) \in \ell^2$ to $\boldsymbol{\mu}^{(2)} = (\dots, \mu_1, \mu_0, \mu_1, \dots) \in \ell^2(\mathbb{Z})$. Then, we can use the Fourier isomorphism to map the vector $\boldsymbol{\mu}^{(2)} \in \ell^2(\mathbb{Z})$ to the complex function space $\mathscr{L}^2(\mathbb{T})$. In this way, each
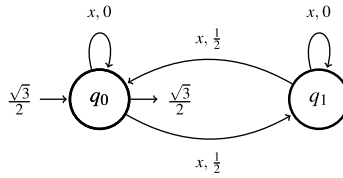
Figure 1. Graphical representation of the generative probabilistic automaton described in Example 2.

vector $\boldsymbol{\mu} \in \ell^2$ corresponds to two functions in the Hardy spaces:

$$\mu^-(z) = \sum_{j=0}^{\infty} \boldsymbol{\mu}_j z^{-j-1} \in \mathscr{H}_-^2, \tag{8}$$

$$\mu^+(z) = \sum_{j=0}^{\infty} \boldsymbol{\mu}_j z^j \in \mathscr{H}^2.$$

Moreover, we can derive the relationship between f and ϕ:

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

from which we obtain:

$$f(n) = \widehat{\phi}(-n-1). \tag{9}$$

Since we know how to express the function f with respect to the parameters of the WFA, we can explicitly compute the rational component of the symbol:

$$\mathbb{P}_-\phi = \sum_{k\geq 0} f(k) z^{-k-1} = \sum_{k\geq 0} \boldsymbol{\alpha}^\top \mathbf{A}^k \boldsymbol{\beta} z^{-k-1} = \boldsymbol{\alpha}^\top (z\mathbf{1} - \mathbf{A})^{-1} \boldsymbol{\beta}, \tag{10}$$

where the last equality holds only if $\rho(A) < 1$.

The correspondence between symbol and function computed by a model allows us to reformulate the approximation problem in terms of Hankel operators and functions in the complex space and to apply AAK theory.

We consider the following example, from Balle et al. (2021).

Example 2. *Let $|\Sigma| = 1$, $\Sigma = \{x\}$, we consider the WFA $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ represented in Figure 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

$$\mathbf{P} = \mathbf{Q} = \begin{pmatrix} \frac{4}{5} & 0 \\ 0 & \frac{1}{5} \end{pmatrix}. \tag{11}$$

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

$$\mathbf{H} = \begin{pmatrix} f_A(0) & f_A(1) & f_A(2) & \cdots \\ f_A(1) & f_A(2) & f_A(3) & \cdots \\ f_A(2) & f_A(3) & f_A(4) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} \frac{3}{4} & 0 & \frac{3}{16} & \cdots \\ 0 & \frac{3}{16} & 0 & \cdots \\ \frac{3}{16} & 0 & \frac{3}{64} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{12}$$

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:

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

We can recover the rational component of a symbol, that is 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 11). 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.

4. Solving the Approximate Minimization Problem

In this section, we present the theoretical contribution of this paper, a closed-form solution for the approximate minimization problem.

4.1 Assumptions

We briefly list and analyze the assumptions made to solve the approximate minimization problem. A class of automata that automatically satisfies the following properties is that of generative probabilistic automata.

4.1.1 One-letter alphabet

We tackle the approximate minimization problem in the case of automata with real weights, defined over a one-letter alphabet. As discussed before, this assumption is needed in order to apply AAK theory and will hold for the rest of the paper (see Section 3.1 for more details).

4.1.2 SVA form

We assume that the minimal 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 (Balle et al., 2019). Starting from a WFA in SVA form 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 coordinates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we have that $\boldsymbol{\alpha}_i = \mathrm{sgn}\,(\lambda_i)\boldsymbol{\beta}_i$, where $\mathrm{sgn}\,(\lambda_i) = \lambda_i/|\lambda_i|$.

4.1.3 Compactness of the operator

To apply Theorem 12, we need the Hankel operator H to be compact. To ensure that this condition is satisfied, we study the respective Hankel matrix. In our setting, the Hankel matrix has finite rank (equal to the number of states of the minimal WFA that we are considering). Moreover, the singular values can be computed exactly using the Gramian matrices introduced in Definition 4. A finite-rank operator is compact if it is bounded. Therefore, we just need to check that the Hankel operator is bounded. To this extent, Balle et al. (2019) show that it is enough that the WFA being considered computes a function $f \in \ell^2$. We make the slightly stronger assumption that the transition matrix A is irredundant, *that is* that $\rho(\mathbf{A}) < 1$, where ρ is the spectral radius. This condition directly implies boundness and the existence of the SVA and the Gramian matrices $\mathbf{P}$ and $\mathbf{Q}$, where $\mathbf{P} = \mathbf{Q}$ and are diagonal matrices (Balle et al., 2019). Moreover, it allows us to compute a closed form for the symbol of a WFA, as seen in Equation 10.

4.2 Problem formulation

Let $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ be a minimal irredundant WFA with n states and real weights, defined over a one-letter alphabet and represented in its SVA form. 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:

$$\|\mathbf{H} - \mathbf{G}\| = \sigma_k(\mathbf{H}). \tag{13}$$

Note that the content of the "optimal spectral-norm approximate minimization" is equivalent to the problem solved by Theorem 12, with the exception that here we represent the inputs and outputs of the problem effectively by means of WFAs.

Based on the AAK theory sketched in Section 2.3, we draw the following steps:

1. *Compute a symbol for the WFA.* Given an irredundant WFA on a one-letter alphabet, we consider its Hankel matrix $\mathbf{H}$ and the function f that it is computing. We use Equation 10 to associate a complex rational function to the WFA.

2. *Compute the optimal symbol $\psi(z)$ using Corollary 14.* The main challenge here is to find a suitable representation for the functions $\psi(z)$ and $e(z) = \phi(z) - \psi(z)$. We define them in terms of two auxiliary WFAs. The key point is to select constraints on their parameters to leverage the properties of weighted automata, while still keeping the formulation general.

3. *Extracting the rational component by solving for g(z) in Corollary 15.* 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 11 that $g(z)$ has k poles, all inside the unit disc.

4. *Find a WFA representation for g(z).* Since in Step 2 we parametrized the functions using WFAs, the expression of $g(z)$ directly reveals the WFA $\widehat{A}_k$.

4.3 Finding the optimal approximation

We analyze each of the steps detailed above.

4.3.1 Finding a symbol for the WFA

Let $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ be a minimal irredundant WFA with n states, already represented in SVA form. A realizes a function $f : \Sigma^* \to \mathbb{R}$, defined over a one-letter alphabet $\Sigma = \{a\}$. Let $\mathbf{H}$ be its Hankel matrix, with corresponding bounded Hankel operator H, and singular numbers σ_i, for $0 \leq i < n$.

As seen in sub-section 3.1, we can associate a complex function to the WFA. In particular, since we are assuming that A is irredundant, from Equation 10 we obtain an expression for the rational component of the symbol:

$$\mathbb{P}_- \phi = \boldsymbol{\alpha}^\top (z\mathbf{1} - \mathbf{A})^{-1} \boldsymbol{\beta}. \tag{14}$$

4.3.2 Finding the optimal symbol

To find the solution to Theorem 12, we need to first derive the function ψ from Theorem 13. Therefore, 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 13. Since our objective is to find the WFA corresponding to the optimal approximation, we focus on representing these functions using the parameters of two auxiliary WFAs. We consider a WFA $\widehat{A} = \langle \widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}} \rangle$ with more than k states, such that the automaton $E = \langle \boldsymbol{\alpha}_e, \mathbf{A}_e, \boldsymbol{\beta}_e \rangle$ computing the difference between A and $\widehat{A}$ is minimal, with:

$$\mathbf{A}_e = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{A}} \end{pmatrix}, \quad \boldsymbol{\alpha}_e = \begin{pmatrix} \boldsymbol{\alpha} \\ -\widehat{\boldsymbol{\alpha}} \end{pmatrix}, \quad \boldsymbol{\beta}_e = \begin{pmatrix} \boldsymbol{\beta} \\ \widehat{\boldsymbol{\beta}} \end{pmatrix}. \tag{15}$$

Now, given $C \in \mathscr{H}^\infty$, we consider the complex functions:

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

The idea is that we want to find the parameters of $\widehat{A}$ that make ψ the solution of Theorem 13. By definition, ψ is the sum of two components, one that is bounded around the unit circle and one that has k poles inside the unit disc (where k is the size of the sought approximation). Therefore, there cannot be poles on the unit circle. By looking at the way we defined the function ψ, we can see that its poles correspond to the eigenvalues of $\widehat{\mathbf{A}}$, counted with their multiplicities. Thus, in order for ψ to be the solution of Theorem 13, 1 cannot be an eigenvalue of $\widehat{\mathbf{A}}$, and the WFA $\widehat{A}$ needs to have at least k states.

As remarked in the previous section, the parameters of the automaton A only encode the negative Fourier coefficients of the symbol. We add C to the definition of ψ to account for the $\mathscr{H}^\infty$ component when considering the difference $\phi - \psi$. In fact, while this component of the symbol does not affect the spectral norm, it plays a role in the computation of the $\mathscr{L}^\infty$-norm (in Equation 5), so it cannot be entirely dismissed. Nonetheless, we won't need to find the value of C, as ultimately we are only interested in the WFA's parameters.

Now that we have an expression for ψ and e, we can look back at Theorem 13. From this theorem, we know that by definition, $\sigma_k^1 e$ is a unimodular function. This property of e can be used to derive a set of constraints on the parameters of the WFA $E = \langle \boldsymbol{\alpha}_e, \mathbf{A}_e, \boldsymbol{\beta}_e \rangle$. In particular, it is possible to use the maximum modulus principle, according to which the maximum modulus of an holomorphic function is attained on the boundary of the domain. To do so, we leverage a result from the control theory literature (Chui and Chen, 1997), that can be easily applied to our setting. In fact, 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. This allows us to apply a theorem from Chui and Chen (Chui and Chen (1997), Theorem 6.3) to find two matrices, $\mathbf{P}_e$ and $\mathbf{Q}_e$, satisfying properties similar to those of the Gramians. It is important to notice that, *a priori*, the controllability and observability Gramians of E might not be well defined.

Theorem 17 (Chui and Chen (1997)). *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. $\sigma_k^{-1} e$ is unimodular if and only if there exists a unique pair of symmetric invertible matrices $\mathbf{P}_e$ and $\mathbf{Q}_e$ satisfying:*

(1) $\mathbf{P}_e - \mathbf{A}_e \mathbf{P}_e \mathbf{A}_e^\top = \boldsymbol{\beta}_e \boldsymbol{\beta}_e^\top$
(2) $\mathbf{Q}_e - \mathbf{A}_e^\top \mathbf{Q}_e \mathbf{A}_e = \boldsymbol{\alpha}_e \boldsymbol{\alpha}_e^\top$
(3) $\mathbf{P}_e \mathbf{Q}_e = \sigma_k^2 \mathbf{1}$

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 13.

Theorem 18. *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:*

$$\sigma_0 \geq \ldots > \sigma_k = \cdots = \sigma_{k+r-1} > \sigma_{k+r} \geq \cdots \geq \sigma_{n-1} > 0. \tag{16}$$

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

$$\mathbf{P} = \mathbf{Q} = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \sigma_k \mathbf{1}_r \end{pmatrix}, \tag{17}$$

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:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \quad \boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}. \tag{18}$$

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

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

$$\begin{cases} \widehat{\boldsymbol{\beta}} = -\widehat{\mathbf{A}} \mathbf{A}_{21}^\top (\boldsymbol{\beta}_2^\top)^+ \\ \widehat{\boldsymbol{\alpha}} = \widehat{\mathbf{A}}^\top \mathbf{R} \mathbf{A}_{12} (\boldsymbol{\alpha}_2^\top)^+ \\ \widehat{\mathbf{A}} (\mathbf{A}_{11}^\top - \mathbf{A}_{21}^\top (\boldsymbol{\beta}_2^\top)^+ \boldsymbol{\beta}_1^\top) = \mathbf{1} \end{cases} \tag{19}$$

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

$$\begin{cases} \widehat{\boldsymbol{\beta}} = (\mathbf{1} - \widehat{\mathbf{A}}\mathbf{A}_{11}^{\top})(\boldsymbol{\beta}_1^{\top})^{+} \\ \widehat{\boldsymbol{\alpha}} = -(\mathbf{R} - \widehat{\mathbf{A}}^{\top}\mathbf{R}\mathbf{A}_{11})(\boldsymbol{\alpha}_1^{\top})^{+} \\ \widehat{\mathbf{A}}\mathbf{A}_{21}^{\top} = \mathbf{0} \end{cases} \tag{20}$$

Proof. We prove the first implication of this proof by applying Theorem 17 and by obtaining from it a set of equations allowing us to derive the parameters of the WFA $\widehat{A}$. The other direction of the theorem can be easily proved by direct computation.

If ψ is the optimal approximation of ϕ, the function $\sigma^{-1}e = \phi - \psi$ is unimodular and from Theorem 17 there exist two symmetric nonsingular matrices $\mathbf{P}_e$, $\mathbf{Q}_e$ satisfying the fixed point equations:

$$\mathbf{P}_e - \mathbf{A}_e\mathbf{P}_e\mathbf{A}_e^{\top} = \boldsymbol{\beta}_e\boldsymbol{\beta}_e^{\top} \tag{21}$$

$$\mathbf{Q}_e - \mathbf{A}_e^{\top}\mathbf{Q}_e\mathbf{A}_e = \boldsymbol{\alpha}_e\boldsymbol{\alpha}_e^{\top} \tag{22}$$

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 15):

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

From Equations 21 and 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:

$$\mathbf{P}_e = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \sigma_k\mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & -\boldsymbol{\Sigma}\mathbf{R}^{-1} \end{pmatrix}, \quad \mathbf{Q}_e = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \mathbf{R} \\ \mathbf{0} & \sigma_k\mathbf{1} & \mathbf{0} \\ \mathbf{R} & \mathbf{0} & -\boldsymbol{\Sigma}\mathbf{R} \end{pmatrix}. \tag{23}$$

Now that we have an expression for the matrices $\mathbf{P}_e$ and $\mathbf{Q}_e$ of Theorem 17, 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:

$$\begin{cases} \mathbf{P} - \mathbf{A}\mathbf{P}\mathbf{A}^{\top} = \boldsymbol{\beta}\boldsymbol{\beta}^{\top} \\ \mathbf{N} - \mathbf{A}\mathbf{N}\widehat{\mathbf{A}}^{\top} = \boldsymbol{\beta}\widehat{\boldsymbol{\beta}}^{\top} \\ -\boldsymbol{\Sigma}\mathbf{R}^{-1} + \widehat{\mathbf{A}}\boldsymbol{\Sigma}\mathbf{R}^{-1}\widehat{\mathbf{A}}^{\top} = \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^{\top} \end{cases} \qquad \begin{cases} \mathbf{P} - \mathbf{A}^{\top}\mathbf{P}\mathbf{A} = \boldsymbol{\alpha}\boldsymbol{\alpha}^{\top} \\ \mathbf{M} - \mathbf{A}^{\top}\mathbf{M}\widehat{\mathbf{A}} = -\boldsymbol{\alpha}\widehat{\boldsymbol{\alpha}}^{\top} \\ -\boldsymbol{\Sigma}\mathbf{R} + \widehat{\mathbf{A}}^{\top}\boldsymbol{\Sigma}\mathbf{R}\widehat{\mathbf{A}} = \widehat{\boldsymbol{\alpha}}\widehat{\boldsymbol{\alpha}}^{\top} \end{cases} \tag{24}$$

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:

$$\begin{cases} \mathbf{1} - \mathbf{A}_{11}\widehat{\mathbf{A}}^\top = \boldsymbol{\beta}_1\widehat{\boldsymbol{\beta}}^\top \\ -\mathbf{A}_{21}\widehat{\mathbf{A}}^\top = \boldsymbol{\beta}_2\widehat{\boldsymbol{\beta}}^\top \end{cases} \qquad \begin{cases} \mathbf{R} - \mathbf{A}_{11}^\top\mathbf{R}\widehat{\mathbf{A}} = -\boldsymbol{\alpha}_1\widehat{\boldsymbol{\alpha}}^\top \\ \widehat{\mathbf{A}}^\top\mathbf{R}\mathbf{A}_{12} = \widehat{\boldsymbol{\alpha}}\boldsymbol{\alpha}_2^\top \end{cases} \tag{25}$$

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

$$\begin{cases} \widehat{\boldsymbol{\beta}} = -\widehat{\mathbf{A}}\mathbf{A}_{21}^\top(\boldsymbol{\beta}_2^\top)^+ \\ \widehat{\boldsymbol{\alpha}} = \widehat{\mathbf{A}}^\top\mathbf{R}\mathbf{A}_{12}(\boldsymbol{\alpha}_2^\top)^+ \\ \widehat{\mathbf{A}}(\mathbf{A}_{11}^\top - \mathbf{A}_{21}^\top(\boldsymbol{\beta}_2^\top)^+\boldsymbol{\beta}_1^\top) = \mathbf{1} \end{cases} \tag{26}$$

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 $\widehat{\mathbf{A}}\mathbf{A}_{21}^\top = \mathbf{0}$, in which case we obtain the following set of solutions:

$$\begin{cases} \widehat{\boldsymbol{\beta}} = (\mathbf{1} - \widehat{\mathbf{A}}\mathbf{A}_{11}^\top)(\boldsymbol{\beta}_1^\top)^+ \\ \widehat{\boldsymbol{\alpha}} = -(\mathbf{R} - \widehat{\mathbf{A}}^\top\mathbf{R}\mathbf{A}_{11})(\boldsymbol{\alpha}_1^\top)^+ \\ \widehat{\mathbf{A}}\mathbf{A}_{21}^\top = \mathbf{0} \end{cases} . \tag{27}$$

This completes the proof of the first implication. The other direction of the proof can be verified by direct computation by first obtaining the matrices $\mathbf{P}_e$ and $\mathbf{Q}_e$, and then showing that they are symmetric, invertible, and satisfy the set of equations of Theorem 17, hence proving that $\sigma_k^{-1}e$ is unimodular. □

We remark that, when $\boldsymbol{\alpha}_2 = \mathbf{0}$, the solution returned by the algorithm will depend on the size of the original automaton and the target approximation. Specifically, $\widehat{\mathbf{A}}$ 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 obtain the solutions in Equation 20, with $\widehat{\mathbf{A}} \neq \mathbf{0}$. On the other hand, if $r \geq \frac{n}{2}$, *that is* 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 $\widehat{\mathbf{A}} = \mathbf{0}$. 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 $k+1$. This way, the multiplicity r of the singular number σ_k is such that $r < \frac{n}{2}$, and the system in Equation 27 is underdetermined.

Theorem 18 provides us with a way to compute the coefficients of the function ψ solving Theorem 13. It is important to notice that the WFA A_k is not necessarily the best approximation we are looking for. Intuitively, the problem is that it might be too big, as irredundancy is not guaranteed by the system of equations (while we know from AAK theory that the best approximation corresponds to a bounded operator). Therefore, in these cases we need to "extract" from A_k a smaller WFA of size k. We do this by extracting the component of the function ψ that belongs to the negative Hardy space.

4.3.3 Extracting the rational component

The objective of this section is to "isolate" the function $g \in \mathscr{R}_k$, *that is* the *rational component* of ψ. To do this, we study the position of the poles of ψ. In fact, we know from Theorem 11 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 and Schneider (1962). A proof of this theorem can be found in Wimmer (1973).

Theorem 19 (Ostrowski and Schneider ([1962](#))). *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 5 with the approximation algorithm), we can rewrite $\widehat{\mathbf{A}}$ in block-diagonal form:

$$\widehat{\mathbf{A}} = \begin{pmatrix} \widehat{\mathbf{A}}_+ & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{A}}_- \end{pmatrix} \tag{28}$$

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, *that is* the function g from Corollary 15 necessary to solve that approximate minimization problem.

Theorem 20. *Let $\widehat{\mathbf{A}}_+$ be as in Equation 28, and $\widehat{\boldsymbol{\alpha}}_+, \widehat{\boldsymbol{\beta}}_+$ obtained applying the same change of basis. 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 $l = \widehat{\boldsymbol{\alpha}}_-^\top (z\mathbf{1} - \widehat{\mathbf{A}}_-)^{-1} \widehat{\boldsymbol{\beta}}_-, l \in \mathscr{H}^\infty$. 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 19 and solve the problem by directly examining the eigenvalues of $-\boldsymbol{\Sigma}\mathbf{R}$. From the proof of Theorem 18, we have $-\boldsymbol{\Sigma}\mathbf{R} = \boldsymbol{\Sigma}(\boldsymbol{\Sigma}^2 - \sigma_k^2 \mathbf{1})$, where $\boldsymbol{\Sigma}$ is the diagonal matrix having as elements the singular numbers of H different from σ_k. It follows that $-\boldsymbol{\Sigma}\mathbf{R}$ 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. □

4.3.4 Solving the approximation problem

Now that we have found the rational function g, a symbol for the operator that solves Theorem 12, 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 function's parametrization.

Theorem 21. *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 15, 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:

$$g = \widehat{\boldsymbol{\alpha}}_+^\top (z\mathbf{1} - \widehat{\mathbf{A}}_+)^{-1} \widehat{\boldsymbol{\beta}}_+ = \sum_{k \geq 0} \widehat{\boldsymbol{\alpha}}_+^\top \widehat{\mathbf{A}}_+^k \widehat{\boldsymbol{\beta}}_+ z^{-k-1} = \sum_{k \geq 0} \bar{f}(k) z^{-k-1} \tag{29}$$

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. □

4.4 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 (Balle et al., 2019), 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 22. *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:*

$$\|f - g\|_{\ell^2} \leq \|\mathbf{H} - \mathbf{G}\| = \sigma_k. \tag{30}$$

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. $\qquad\square$

5. Algorithm

We now use the results obtained in the previous sections 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 20. As mentioned in Section 4.1, 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 $O(n^3)$ (Berstel and Reutenauer, 2011), and the SVA computed in $O(n^3)$ (Balle et al., 2019). The algorithm applies the results of Theorem 18 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.

5.1 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 with 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 the steps of Glover (1984), we compute the Schur decomposition, *that is* we find an orthogonal matrix $\mathbf{U}$ such that

---

Algorithm 1: `AAK approximation`

---

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. 17 and 18

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 $\widehat{A}_k \leftarrow \texttt{BlockDiagonalize}(\widehat{A})$

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

---

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:

$$\mathbf{T} = \mathbf{U}^\top \widehat{\mathbf{A}} \mathbf{U} = \begin{pmatrix} \widehat{\mathbf{A}}_+ & \widehat{\mathbf{A}}_{12} \\ \mathbf{0} & \widehat{\mathbf{A}}_- \end{pmatrix} \tag{31}$$

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 23 (Roth (1952)). *Let $\mathbf{T}$ be the matrix defined in Equation 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*

$$\mathbf{M} = \begin{pmatrix} \mathbf{1} & \mathbf{X} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}, \quad \text{and} \quad \mathbf{M}^{-1} = \begin{pmatrix} \mathbf{1} & -\mathbf{X} \\ \mathbf{0} & \mathbf{1} \end{pmatrix} \tag{32}$$

satisfy:

$$\mathbf{M}^{-1} \mathbf{T} \mathbf{M} = \begin{pmatrix} \widehat{\mathbf{A}}_+ & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{A}}_- \end{pmatrix}, \tag{33}$$

where $\mathbf{T}$ is the matrix defined in Equation 31.

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

$$\widehat{\mathbf{A}}_+ = \boldsymbol{\Gamma} \mathbf{M}^{-1} \mathbf{U}^\top \widehat{\mathbf{A}} \mathbf{U} \mathbf{M} \boldsymbol{\Gamma}^\top \tag{34}$$

$$\widehat{\boldsymbol{\alpha}}_+ = \boldsymbol{\Gamma} \mathbf{M}^\top \mathbf{U}^\top \widehat{\boldsymbol{\alpha}} \tag{35}$$

$$\widehat{\boldsymbol{\beta}}_+ = \boldsymbol{\Gamma} \mathbf{M}^{-1} \mathbf{U}^\top \widehat{\boldsymbol{\beta}}. \tag{36}$$

The algorithm `BlockDiagonalize` corresponds to the implementation of this procedure, and Step 5 can be performed using the Bartels-Stewart algorithm (Bartels and Stewart, 1972).

Algorithm 2: `BlockDiagonalize`

---

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

1 **if** dim $\widehat{\mathbf{A}} = k$ **then**
2 | **return** $\widehat{A}_k$
3 **else**
4 | Compute the Schur decomposition of $\widehat{\mathbf{A}} = \mathbf{U}\mathbf{T}\mathbf{U}^\top$, where $|T_{11}| \leq |T_{22}| \leq \ldots$
5 | Solve $\widehat{\mathbf{A}}_{11}\mathbf{X} - \mathbf{X}\widehat{\mathbf{A}}_{22} + \widehat{\mathbf{A}}_{12} = \mathbf{0}$ for $\mathbf{X}$
6 | Let $\mathbf{M} = \begin{pmatrix} 1 & \mathbf{X} \\ 0 & 1 \end{pmatrix}$ and $\mathbf{M}^{-1} = \begin{pmatrix} 1 & -\mathbf{X} \\ 0 & 1 \end{pmatrix}$
7 | Let $\mathbf{\Gamma} = \begin{pmatrix} \mathbf{1}_k & \mathbf{0} \end{pmatrix}$
8 | Let $\widehat{\mathbf{A}}_+ = \mathbf{\Gamma}\mathbf{M}^{-1}\mathbf{U}^\top\widehat{\mathbf{A}}\mathbf{U}\mathbf{M}\mathbf{\Gamma}^\top$
9 | Let $\widehat{\boldsymbol{\alpha}}_+ = \mathbf{\Gamma}\mathbf{M}^\top\mathbf{U}^\top\widehat{\boldsymbol{\alpha}}$
10 | Let $\widehat{\boldsymbol{\beta}}_+ = \mathbf{\Gamma}\mathbf{M}^{-1}\mathbf{U}^\top\widehat{\boldsymbol{\beta}}$
11 | Let $\widehat{A}_k = \langle \widehat{\boldsymbol{\alpha}}_+, \widehat{\mathbf{A}}_+, \widehat{\boldsymbol{\beta}}_+ \rangle$
12 | **return** $\widehat{A}_k$

5.2 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 with each step of the algorithms (Trefethen and Bau III, 1997):

- 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 $O(mn^2 + nm^2)$.

6. Example

We consider the following weighted finite automaton with three states over a one-letter alphabet, represented in SVA form:

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

The objective is to find the WFA with two states solving the approximate minimization problem optimally.

We first note that $\mathbf{A}$ has spectral radius strictly smaller than 1, having eigenvalues:

$$\lambda_{1,2} = 0.0162324 \pm 0.0297233i \qquad \lambda_3 = 0.0324648. \tag{37}$$

Therefore, the assumptions listed Section 4.1 are satisfied, and we can apply Theorem 18. We compute the Gramian matrices and obtain, according to the partition in Equation 17, the following matrix:

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

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

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

We then proceed by partitioning $\mathbf{A}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and obtain:

$$\mathbf{A}_{11} = \begin{pmatrix} 0.579 & 0.461 \\ -0.461 & -0.192 \end{pmatrix}, \quad \mathbf{A}_{1,2} = \begin{pmatrix} 0.046 \\ 0.225 \end{pmatrix}, \quad \mathbf{A}_{2,1}^\top = \begin{pmatrix} 0.046 \\ -0.225 \end{pmatrix}, \quad \mathbf{A}_{22} = -0.387.$$

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

Since $\boldsymbol{\alpha}_2 \neq 0$, we can use Equation 19 to find the coefficients of the auxiliary WFA $\widehat{A} = \langle \widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}} \rangle$.

We have:

$$\begin{cases} \widehat{\boldsymbol{\beta}} = -\widehat{\mathbf{A}} \mathbf{A}_{21}^\top (\boldsymbol{\beta}_2^\top)^+ \\ \widehat{\boldsymbol{\alpha}} = \widehat{\mathbf{A}}^\top \mathbf{R} \mathbf{A}_{12} (\boldsymbol{\alpha}_2^\top)^+ \\ \widehat{\mathbf{A}}(\mathbf{A}_{11}^\top - \mathbf{A}_{21}^\top (\boldsymbol{\beta}_2^\top)^+ \boldsymbol{\beta}_1^\top) = \mathbf{1} \\ \widehat{\boldsymbol{\beta}} = -\widehat{\mathbf{A}} \begin{pmatrix} 0.046 \\ -0.225 \end{pmatrix} (0.038)^{-1} \\ \widehat{\boldsymbol{\alpha}} = \widehat{\mathbf{A}}^\top \left(\begin{pmatrix} 0.12 & 0 \\ 0 & 0.12 \end{pmatrix} - \begin{pmatrix} 4.67 & 0 \\ 0 & 1.79 \end{pmatrix}^2 \right) \begin{pmatrix} 0.046 \\ 0.225 \end{pmatrix} (0.038)^{-1} \\ \widehat{\mathbf{A}} \left(\begin{pmatrix} 0.579 & 0.461 \\ -0.461 & -0.192 \end{pmatrix}^\top - \begin{pmatrix} 0.046 \\ -0.225 \end{pmatrix} (0.038)^{-1} \begin{pmatrix} 1.650 \\ 0.851 \end{pmatrix}^\top \right) = \mathbf{1} \end{cases}$$

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 in order to find the optimal approximation. To do so, we block-diagonalize the transition matrix $\widehat{\mathbf{A}}$ and look at the modulus of its eigenvalues.

We have:

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

As we can see, both eigenvalues have modulus smaller than one. This means that the WFA $\widehat{A}$ is exactly the optimal approximation of size two that we are looking for, and there aren't any components that need to be discarded. Following the notation introduced in the previous section, we have: $\widehat{A}_k = \langle \widehat{\boldsymbol{\alpha}}_+, \widehat{\mathbf{A}}_+, \widehat{\boldsymbol{\beta}}_+ \rangle = \langle \widehat{\boldsymbol{\alpha}}, \widehat{\mathbf{A}}, \widehat{\boldsymbol{\beta}} \rangle$.

7. Related Work

The problem of minimizing automata has been an important subject of research since the 1950s. There is a remarkable algorithm due to Brzozowski (Brzozowski, 1962, 1964) 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 (Droste et al., 2009). This line of algorithms is based on a new understanding of Brzozowski's algorithm from the point of view of duality (Bonchi et al., 2012b;a,2014; Bezhanishvili et al., 2012) and extends readily to other settings. In the context of quantitative systems, like weighted or probabilistic automata, it becomes meaningful to investigate different kinds of quantitative approximations and in particular of the approximate minimization problem. In Balle et al., (2017, 2022), the authors propose a bisimulation pseudometric to express the notion of behavioral proximity between states of WFAs. For example, this becomes particularly useful when a small perturbation has been applied to the parameters of the WFAs, and bisimulation alone would fail to capture their proximity. Similarly, in the context of finding a robust way of approximating dynamical systems despite the high sensitivity to the choice of parameters, Cardelli et al. (2018) introduce a notion of approximate bisimulation for ordinary differential equations with polynomial derivatives. Given a model and a parameter ϵ, they propose a method to compute an ϵ bisimulation over the ODEs variables and then perturb minimally the model to obtain a new one where the ϵ-bisimulation is an exact one. We remark that in this case the author generalizes the notion of differential equivalence that, while remaining very close to the concept of bisimulation defined for automata and dynamical systems, applies to ODE variables rather than the state space. The study of the approximate minimization 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 (Kulesza et al., 2015). 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 (Kulesza et al., 2014). 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 (Kotsalis and Shamma, 2015). We remark that the framework of Hidden Markov models is encompassed by weighted automata (Denis and Esposito, 2008). Balle, Panangaden, and Precup are the first authors to formalize the approximate minimization problem for WFAs (Balle et al., 2015, 2019). 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. Their method can be viewed as a generalization to multi-letter alphabets of the balanced realization approach from control theory (Antoulas, 2005). 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 in Balle and Rabusseau (2020). The main limitation of these approaches based on SVA truncation is that the approximation obtained is not optimal in any norm. We partially address this point in this work,

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. Part of this results were presented in (Balle et al., 2021). In Lacroce et al. (2021), we extend this results to the more general setting of black-box models trained for language modeling over one-letter alphabets. In Lacroce et al. (2022); Lacroce (2022), we analyze the problem of extending the method presented in this paper to the case of multi-letter alphabets.

The control theory community has largely studied approximate minimization in the context of linear time-invariant systems (Antoulas, 2005). A parallel with these results can be drawn by noting that the impulse response of a discrete Single-Input-Single-Output SISO system can be parametrized as a WFA over a one-letter alphabet. Glover (1984) presents a state-space solution for the case of continuous Multi-Input-Multi-Output MIMO systems. His method led to a widespread application of these results, thanks to its computational and theoretical simplicity. This stems from the structure of the continuous Lyapunov equations. For discrete systems, though, the quadratic nature of the Lyapunov equations does not allow for a simple closed-form formula for the state-space solution (Chui and Chen, 1997). Thus, most of the results for the discrete case work with a suboptimal version of the problem (Ball and Ran, 1987; Al-Hussari et al., 1993; Ionescu and Oara, 2001). A solution for the SISO case can be found using a polynomial approach, but it does not provide an explicit representation of the state space nor it generalizes to the MIMO setting. The first to actually extend Glover results is Gu, who provides an elegant solution for the MIMO discrete problem (Gu, 2005). Glover and Gu's solutions rely on building an *all-pass system*, equivalent to the WFA E in our case. Part of our contribution is the adaptation of some of the control theory tools to WFAs.

8. Extensions and Future Work

In this section, we examine possible extensions of our method by relaxing some of the hypothesis.

8.1 Removing the finite-rank assumption

The proof of Theorem 12 is constructive for any compact Hankel operator. In the setting of this paper, compactness is guaranteed, as the operator corresponding to an irredundant WFA has finite rank and is bounded. While boundness is necessary for compactness, the finite-rank hypothesis is not. Therefore, an interesting extension of this work is to investigate other classes of models by relaxing the finite-rank (or finite state) assumption. An example of models corresponding to infinite-rank Hankel matrices is recurrent neural networks (RNNs) (Hochreiter and Schmidhuber, 1997). Recently, particular attention has been given to the problem of extracting, from an RNN, a weighted finite automaton (Ayache et al., 2018; Rabusseau et al., 2019; Weiss et al., 2019; Theertha Suresh et al., 2021; Okudono et al., 2020; Eyraud and Ayache 2024; Zhang et al., 2021). In this sense, the knowledge distillation task (Hinton et al., 2015) is very similar to an approximate minimization problem, since WFAs are a less expensive alternative to RNNs, while still being expressive and suited for sequence modeling and prediction (Cortes et al., 2004; Denis and Esposito, 2008). In Lacroce et al. (2021), we investigated the use of AAK theory on black-box models trained for language modeling on sequential data. In particular, we showed that compactness is automatically respected by black boxes for language modeling, and proposed an algorithm for the one-letter setting, based on AAK theory. This particular extension of the method presented in this paper constitutes a first fundamental step towards developing provable approximation algorithms for black-box models.

8.2 Removing the spectral radius assumption

One could consider a WFA over a one-letter alphabet with $\rho(\mathbf{A}) \neq 1$, *that is* not necessarily irredundant. In this case, the method proposed in the previous sections 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 (Glover, 1984).

The key idea is to block-diagonalize $\mathbf{A}$ like we did in Section 4.3.3. 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

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

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.

8.3 Removing the one-letter assumption

The most pressing direction for future work is undoubtedly to extend our results to a multi-letter setting. The work of Adamyan, Arov and Krein provides us with a powerful theory connecting sequences to the study of complex functions. Unfortunately, this approach cannot be directly generalized to the multi-letter case, when Σ^* is a noncommutative monoid, as it requires to generalize standard harmonic analysis results to the non-abelian case. A recent line of work in multivariable operator theory has been centered around extending results of standard operator theory to the case of noncommutative operators defined on Fock spaces Frazho (1982); Bunce (1984); Arias and Popescu (1995); Popescu (1989, 1992, 1993, 1995, 2003, 2006, 2010, 2013); Ball and Bolotnikov (2021); Jury et al. (2021). In particular, a noncommutative definition of Hankel operator, and a noncommutative version of the AAK theorem are presented in a recent work of Popescu Popescu (2003), 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 noncommutative 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 the Hardy space and of the symbol.

- *Making AAK constructive:* the proof of the noncommutative version of the AAK theorem does not provide us with an expression for the optimal approximation. An interesting direction would be to explore ways to extend the proof to a constructive one.

In Lacroce et al. (2022), we proposed a framework to associate a noncommutative Hankel operator (defined on a noncommutative version of the Hardy space) and a noncommutative rational function to the Hankel matrix computed by a model on sequential data, solving the first point listed above. In the one-letter case, obtaining the framework allowed us to reformulate the approximation problem in terms of functional analysis, and to solve it using the constructive proof of AAK theorem. In Lacroce (2022), we tried to address the question of whether or not the proof of the noncommutative AAK theorem can be made constructive. While we did not manage to provide a definitive answer, we laid out possible approaches that can be used

to tackle the problem of making the proof of the noncommutative version of AAK theorem constructive.

9. Conclusion

In this paper, we applied the AAK theory for Hankel operators and complex functions with the framework of WFAs in order to construct the optimal approximation to an automaton given a bound on the size. We propose an algorithm to find the parameters of the best WFA approximation in the spectral norm and derive bounds on the error. Our method applies to real irredundant WFAs defined over a one-letter alphabet. These alphabets have proven to be of independent interest when dealing with automata, as in this case the classes of regular and context-free languages collapse (Pighizzini, 2015).

We think the spectral norm has desirable characteristics, making it a solid candidate for the approximate minimization task. For example, it can be minimized in polynomial time and a global minimum for the error can be computed accurately. Moreover, the fact that this norm is independent on the specific architecture or model considered facilitates future applications of this method, as it can be used to compare different classes of models. Nonetheless, a 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 behavioral metrics, or other metrics coming from natural language processing (e.g., word error rate and normalized discounted cumulative gain). To some extent, this problem is a collateral effect of the size of the alphabet: the comparison between spectral norm and other kind of norms is possible only in the multi-letter setting. Obtaining algorithms for the multi-letter case will thus open the possibility of evaluating the quality of the spectral norm.

While the one-letter setting is certainly restricted, we believe that this work constitutes a first fundamental step in the direction of optimal approximation. Furthermore, the use of AAK techniques has proven to be very fruitful in related areas like control theory; we think that automata theory can also benefit from it. The use of such methods can help deepen the understanding of the behavior of rational functions. This paper highlights and strengthens the interesting connections between functional analysis, automata theory, and control theory, unifying tools from different domains in one formalism.

Competing interests.   The authors declare none.

References

Adamyan, V. M., Arov, D. Z. and Krein, M. G. (1971). Analytic properties of schmidt pairs for a Hankel operator and the generalized Schur–Takagi problem. *Mathematics of the Ussr-Sbornik*. **15** (1) 31–73.

Al-Hussari, M., Jaimoukha, I. and Limebeer, D. (1993). A descriptor approach for the solution of the one-block distance problem. In: *Proceedings of the IFAC World Congress*.

Antoulas, A. C. (2005). *Approximation of Large-Scale Dynamical Systems*, SIAM.

Arias, A. and Popescu, G. (1995). Factorization and reflexivity on Fock spaces. *Integral Equations and Operator Theory*. **23** (3) 268–286.

Ayache, S., Eyraud, R. and Goudian, N. (2018). Explaining Black Boxes on Sequential Data Using Weighted Automata, In: *Proceedings of the 14th International Conference on Grammatical Inference, ICGI. 2018*, volume 93 of Proceedings of Machine Learning Research (PMLR) Wrocław, Poland, September 5-7, 201881–103.

Bailly, R., Denis, F. and Ralaivola, L. (2009). Grammatical inference as a principal component analysis problem. In: *Proceedings of the 26th Annual International Conference on Machine Learning (ICML), Association for Computing Machinery*, NY, USA, 33–40.

Ball, J. A. and Bolotnikov, V. (2021). *Noncommutative Function-Theoretic Operator Theory and Applications*, Cambridge Tracts in Mathematics, Cambridge University Press.

Ball, J. A. and Ran, A. C. (1987). Optimal Hankel norm model reductions and Wiener–Hopf factorization I: the canonical case. *SIAM Journal On Control and Optimization* **25** (2) 362–382.

Balle, B., Carreras, X., Luque, F. M. and Quattoni, A. (2014*a*). Spectral learning of weighted automata - a forward–Backward perspective. *Machine Learning* **96** (1-2) 33–63.

Balle, B., Gourdeau, P. and Panangaden, P. (2017) Bisimulation metrics and norms for weighted finite automata, In: Chatzigiannakis, I., Indyk, P., Kuhn, F. and Muscholl, A. (eds.) *44th International Colloquium On Automata, Languages, and Programming, ICALP, volume 80 of LIPIcs*, Warsaw, Poland, 103:1–103:14, Schloss Dagstuhl - Leibniz-Zentrum für Informatik. 2017,

Balle, B., Gourdeau, P. and Panangaden, P. (2022). Bisimulation metrics and norms for real-weighted automata. *Information and Computation* **282** 104649.

Balle, B., Hamilton, W. L. and Pineau, J. (2014*b*). Methods of Moments for Learning Stochastic Languages: Unified Presentation and Empirical Comparison, In: *Proceedings of the 31th International Conference on Machine Learning (ICML), volume 32 of JMLR Workshop and Conference Proceedings*, Beijing, China, 1386–1394.

Balle, B., Lacroce, C., Panangaden, P., Precup, D. and Rabusseau, G. (2021) Optimal spectral-norm approximate minimization of weighted finite automata, In:Bansal, N., Merelli, E. and Worrell, J. (eds.) *48th International Colloquium On Automata, Languages, and Programming, ICALP, vol.198*, Glasgow, Scotland (Virtual Conference), 118, LIPIcs

Balle, B., Panangaden, P. and Precup, D. (2015). A Canonical Form for Weighted Automata and Applications to Approximate Minimization. In: *30th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, Kyoto, Japan: IEEE Computer Society. 701–712.

Balle, B., Panangaden, P. and Precup, D. (2019). Singular value automata and approximate minimization. *Mathematical Structures in Computer Science* **29** (9) 1444–1478.

Balle, B. and Rabusseau, G. (2020). Approximate minimization of weighted tree automata. *Information and Computation* **282** 104654.

Bartels, R. H. and Stewart, G. W. (1972). Solution of the matrix equation ax + xb = c [f4]. *Communications of the ACM* **15** (9) 820–826.

Berstel, J. and Reutenauer, C. (2011). *Noncommutative Rational Series with Applications*, Cambridge University Press.

Bezhanishvili, N., Kupke, C. and Panangaden, P. (2012) Minimization via duality, *Logic, Language, Information and Computation - 19th International Workshop, WoLLIC. 2012, Buenos Aires Proceedings, of Lecture Notes in Computer Science, vol. 7456*, Argentina: Springer, 191–205.

Bonchi, F., Bonsangue, M., Boreale, M., Rutten, J. and Silva, A. (2012a). A coalgebraic perspective on linear weighted automata. *Information and Computation* **211** 77–105.

Bonchi, F., Bonsangue, M. M., Hansen, H. H., Panangaden, P., Rutten, J. and Silva, A. (2014). Algebra-coalgebra duality in Brzozowski's minimization algorithm. *ACM Transactions On Computational Logic* **15** (1) 1–29.

Bonchi, F., Bonsangue, M., Rutten, J. and Silva, A. (2012b) Brzozowski's algorithm (co)algebraically, Constable, R. and Silva, A., (eds.) *Logics and Program Semantics: Essays Dedicated to Dexter Kozen, Lecture Notes In Computer Science, vol. 7230*, Springer-Verlag, 12–23.

Brzozowski, J. A. (1962). Canonical regular expressions and minimal state graphs for definite events. In: Fox, J., (ed.) *Proceedings of the Symposium on Mathematical Theory of Automata, in MRI Symposia Series*, Polytechnic Press of the Polytechnic Institute of Brooklyn, **12** 529–561.

Brzozowski, J. A. (1964). Derivatives of regular expressions. *Journal of the ACM* **11** (4) 481–494.

Bunce, J. W. (1984). Models for n-tuples of noncommuting operators. *Journal of Functional Analysis* **57** (1) 21–30.

Cardelli, L., Tribastone, M., Tschaikowski, M. and Vandin, A. (2018). Guaranteed error bounds on approximate model abstractions through reachability analysis. In:McIver, A. and Horváth, A., (eds.) *Quantitative Evaluation of Systems - 15th International Conference, (QEST),* Proceeding volume 11024 of Lecture Notes in Computer Science, Beijing, China, September 4-7, 2018, Springer, 104–121.

Carlyle, J. and Paz, A. (1971). Realizations by stochastic finite automata. *Journal of Computer and System Sciences* **5** (1) 26–40.

Chui, C. K. and Chen, G. (1997). *Discrete H^∞ Optimization With Applications in Signal Processing and Control Systems*, Springer-Verlag.

Cortes, C., Haffner, P. and Mohri, M. (2004). Rational kernels: theory and Algorithms. *Journal of Machine Learning Research (JMLR)* **5** 1035–1062.

Denis, F. and Esposito, Y. (2008). On rational stochastic languages. *Fundamenta Informaticae* **86** (1, 2) 41–77.

Droste, M., Kuich, W. and Vogler, H. (2009). *Handbook of Weighted Automata*, Springer Science & Business Media.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika* **1** (3) 211–218.

Eyraud, R. and Ayache, S. (2024). Distillation of weighted automata from recurrent neural networks using a spectral approach. *Mach. Learn.* **113** (5) 3233–3266.

Fliess, M. (1974). Matrice de Hankel. *Journal De Mathématique Pures et Appliquées* **5** 197–222.

Frazho, A. E. (1982). Models for noncommuting operators. *Journal of Functional Analysis* **48** (1) 1–11.

Glover, K. (1984). All optimal Hankel–Nnorm approximations of linear multivariable systems and their $\mathscr{L}^{\infty}$–Error bounds. *International Journal of Control* **39** (6) 1115–1193.

Gu, G. (2006). All optimal Hankel-norm approximations and their error bounds in discrete-time. *International Journal of Control* **78** (6) 408–423.

Hinton, G., Vinyals, O. and Dean, J. (2015). Distilling the knowledge in a neural network, *CoRR*, abs/1503.02531.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation* **9** (8) 1735–1780.

Hsu, D., Kakade, S. M. and Zhang, T. (2012). A spectral algorithm for learning Hidden Markov Models. *Journal of Computer and System Sciences* **78** (5) 1460–1480.

Ionescu, V. and Oara, C. (2001) The four-block adamjan-arov-kein problem for discrete-time systems, *Linear Algebra and Its Application*. Elsevier, 95–119.

Jury, M., Martin, R. and Shamovich, E. (2021). *Non-Commutative Rational Functions in the Full Fock Space*, Transactions of the American Mathematical Society.

Kotsalis, G. and Shamma, J. S. (2015). Limits of performance for the model reduction problem of hidden markov models. In: *2015 54th IEEE Conference on Decision and Control (CDC)*, 4674–4679.

Kronecker, L. (1881). Zur Theorie der elimination einer Variablen aus zwei algebraischen Gleichungen. *Montasber. Königl. Preussischen Acad Wies* 535–600.

Kulesza, A., Jiang, N. and Singh, S. (2015) Low-rank spectral learning with weighted loss functions, In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research* (PMLR), 517–525.

Kulesza, A., Rao, N. R. and Singh, S. (2014). Low-Rank Spectral Learning. In: Kaski, S. and Corander, J. (eds.) *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research (PMLR)*, Reykjavik, Iceland, **33**, 522–530.

Lacroce, C. (2022). The Approximate Minimization Problem of Weighted Finite Automata and Applications to Language Modelling: An Approach Based on Adamyan-Arov-Krein Theory. *PhD thesis*. McGill University.

Lacroce, C., Panangaden, P. and Rabusseau, G. (2021). Extracting weighted automata for approximate minimization in language modelling. In: Chandlee, J., Eyraud, R., Heinz, J., Jardine, A. and van Zaanen, M. (eds.) *Proceedings of the Fifteenth International Conference on Grammatical Inference*, volume 153 of Proceedings of Machine Learning Research (PMLR), **153**, 92–112.

Lacroce, C., Panangaden, P. and Rabusseau, G. (2022). Towards an AAK theory approach to approximate minimization in the multi-letter case. *CoRR*. abs/2206.00172.

Lyapunov, A. M. (1950). The general problem of the stability of motion [in Russian]. Gostekhizdat, Moscow, Elsevier.

Nehari, Z. (1957). On bounded bilinear forms. *The Annals of Mathematics* **65** (1) 153–162.

Nikol'Skii, N. K. (2002) Operators, functions and systems: an easy reading, *volume 92 of Mathematical Surveys and Monographs*, American Mathematical Society.

Okudono, T., Waga, M., Sekiyama, T. and Hasuo, I. (2020). 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, AAAI Press. 5306–5314.

Ostrowski, A. and Schneider, H. (1962). Some theorems on the inertia of general matrices. *Journal of Mathematical Analysis and Applications* **4** (1) 72–84.

Peller, V. (2012). *Hankel Operators and Their Applications*, Springer Science & Business Media.

Pighizzini, G. (2015). Investigations on automata and languages over a unary alphabet. *International Journal of Foundations of Computer Science* **26** (07) 827–850.

Popescu, G. (1989a). Isometric dilations for infinite sequences of noncommuting operators. *Transactions of the American Mathematical Society* **316** (2) 523–536.

Popescu, G. (1989*b*). Models for infinite sequences of noncommuting operators. *Acta Scientiarum Mathematicarum* **53** 355–285.

Popescu, G. (1992). On intertwining dilations for sequences of noncommuting operators. *Journal of Mathematical Analysis and Applications* **167** (2) 382–402.

Popescu, G. (1993). Noncommutative dilation theory on Fock spaces. *PhD thesis*. Texas A&M University.

Popescu, G. (1995). Multi-analytic operators on Fock spaces. *Mathematische Annalen* **303** (1) 31–46.

Popescu, G. (2003). Multivariable Nehari problem and interpolation. *Journal of Functional Analysis* **200** (2) 536–581.

Popescu, G. (2006). Free holomorphic functions on the unit ball of b(h)n. *Journal of Functional Analysis* **241** (1) 268–333.

Popescu, G. (2010). *Operator Theory On Noncommutative Domains*, American Mathematical Society.

Popescu, G. (2013). Noncommutative multivariable operator theory. *Integral Equations and Operator Theory* **75** (1) 87–133.

Rabusseau, G., Li, T. and Precup, D. (2019). Connecting Weighted Automata and Recurrent Neural Networks through Spectral Learning. In: Chaudhuri, K. and Sugiyama, M. (eds.) *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS. 2019, volume 89 of Proceedings of Machine Learning Research (PMLR)*, Naha, Okinawa, Japan, 1630–1639.

Roth, W. E. (1952). The equations ax - yb = c and ax - xb = c in matrices. *Proceedings of the American Mathematical Society* **3** (3) 392–396.

Theertha Suresh, A., Roark, B., Riley, M. and Schogol, V. (2021). Approximating probabilistic models as weighted finite automata. *Comput. Linguistics*, **47**(2) 221–254.

Trefethen, L. N. and Bau, D.III. (1997). *Numerical Linear Algebra*, volume 50, Siam, 1–235.

Weiss, G., Goldberg, Y. and Yahav, E. (2019). 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)*, December 8-14, 2019, Vancouver, BC, Canada, 8558–8569.

Wimmer, H. K. (1973). On the Ostrowski-Schneider inertia theorem. *Journal of Mathematical Analysis and Applications* **41** (1) 164–169.

Zhang, X., Du, X., Xie, X., Ma, L., Liu, Y. and Sun, M. (2021). 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*, AAAI Press. 11699–11707, Virtual Event, February 2-9, 2021.

Zhu, K. (1990). *Operator Theory in Function Spaces*, Volume 138, American Mathematical Society.