Towards an AAK Theory Approach to Approximate Minimization in the Multi-Letter Case

Anonymous

Abstract

We study the approximate minimization problem of weighted finite automata (WFAs): given a WFA, we want to compute its optimal approximation when restricted to a given size. We reformulate the problem as a rank-minimization task in the spectral norm, and propose a framework to apply Adamyan-Arov-Krein (AAK) theory to the approximation problem. This approach has already been successfully applied to the case of WFAs and language modelling black boxes over one-letter alphabets (Balle et al., 2021; Lacroce et al., 2021). Extending the result to multi-letter alphabets requires solving the following two steps. First, we need to reformulate the approximation problem in terms of noncommutative Hankel operators and noncommutative functions, in order to apply results from multivariable operator theory. Secondly, to obtain the optimal approximation we need a version of noncommutative AAK theory that is constructive. In this paper, we successfully tackle the first step, while the second challenge remains open.

Keywords: Approximate minimization, Hankel matrices, AAK theory, weighted finite automata, language modelling.

1. Introduction

The problem of minimizing an automaton has been well studied over the past seventy years. When dealing with quantitative models, like weighted or probabilistic automata, it becomes possible to define quantitative notions of model similarity, and to find *approximately* minimal approximations. In particular, given a minimal weighted finite automaton (WFA), the *approximate minimization problem* consists in finding a WFA, smaller than the minimal one, that mimics its behaviour. We are interested in quantifying and minimizing the approximation error. The approximate minimization problem is strictly related to knowledge distillation and extraction tasks (Weiss et al., 2019; Okudono et al., 2020; Eyraud and Ayache, 2020; Ayache et al., 2018; Rabusseau et al., 2019). When the solution of the problem is optimal, this approach has a clear advantage compared to other methods, as it allows us to search for the best WFA among those of a predefined size. This is particularly useful when dealing with limited computing resources, or to improve interpretability.

Several norms can be considered to estimate the error. The approximate minimization problem was formalized by Balle et al. (2019), and the error measured with respect to the ℓ^2 norm. In this paper, we reformulate the problem in terms of its Hankel matrix $\mathbf{H}$ and look for a low-rank approximation in the spectral norm. This norm has the advantage that it can be used to compare different classes of models. Moreover, it is possible to find (and compute) a global minimum for the error in polynomial time (Balle et al., 2021). In fact, the celebrated Adamyan-Arov-Krein (AAK) theory provides a way, based on properties of Hankel operators and complex functions, to find the optimal approximation of $\mathbf{H}$ within the class of Hankel matrices. We lay out a framework for the application of this theory

to the approximate minimization problem, analyzing the case of one-letter and multi-letter alphabet separately. In the first case, standard AAK theory can be applied, and the proof of the AAK theorem tells us how to construct the optimal approximation. This setting has been studied by Balle et al. (2021) to obtain an algorithm, based on AAK theory, returning the optimal approximation of a class of WFAs. Lacroce et al. (2021) generalize this approach to find an (asymptotically) optimal approximation of a general black-box model trained for language modelling on sequential data, still under the one-letter assumption. Extending the work to the multi-letter case requires a noncommutative (NC) version of AAK theory. Tackling this problem is fundamental for the application of these results and to experimentally compare the performance of the spectral norm against other norms (behavioral metrics, word error rate, or normalized discounted cumulative gain). To achieve this, two challenges need to be addressed. First, to apply the AAK theorem it is necessary to reformulate the approximation problem in terms of NC Hankel operators, defined in an appropriate NC space. Second, we need a constructive version of this theorem to find the optimal approximation. In this paper, we tackle the first challenge and make the following contributions. We start by summarizing the approach used by Balle et al. (2021); Lacroce et al. (2021) in the one-letter case. Then, we reformulate the approximate minimization problem of models over multi-letter alphabets in terms of NC Hankel operators. Finally, we suggest a way to link the Hankel matrix of a WFA to a NC rational function. While the second challenge remains open, this constitutes a first, encouraging step towards its solution, since the rational function is key in the construction of the optimal approximation.

2. Background

Let $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$ be the sets of natural, integers and real numbers, respectively. Given $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$, $\mathbf{N} \in \mathbb{R}^{d_1' \times d_2'}$ we denote their *Kronecker product* by $\mathbf{M} \otimes \mathbf{N} \in R^{d_1 d_1' \times d_2 d_2'}$ with entries given by $(\mathbf{M} \otimes \mathbf{N})((i-1)d_1'+i', (j-1)d_2'+j') = \mathbf{M}(i,j)\mathbf{N}(i',j')$. Let $\mathbf{M} \in \mathbb{R}^{p \times q}$ of rank n, the compact *singular value decomposition* SVD of $\mathbf{M}$ is the factorization $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, 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 $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{1}$, where $\mathbf{1}$ denotes the identity matrix, and $\mathbf{D}$ is a diagonal matrix. The columns of $\mathbf{U}$ and $\mathbf{V}$ are called left and right *singular vectors*, while the diagonal entries of $\mathbf{D}$ are the *singular values*. The *spectral radius* $\rho(\mathbf{M})$ of $\mathbf{M}$ is the largest modulus among its eigenvalues. Let ℓ^2 be the space of square-summable sequences over $\mathbb{N}$. Let $\mathcal{L}^p(\mathbb{T})$ be the space of measurable functions on $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$ for which the p-th power of the absolute value is Lebesgue integrable.

2.1. Hankel matrix and Weighted Automata

Let Σ be a fixed finite alphabet, Σ^* the set of all finite strings with symbols in Σ, and ε the empty string. Given $p, s \in \Sigma^*$, we denote with ps their concatenation. Let $f : \Sigma^* \to \mathbb{R}$, we can consider a 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 (bi-infinite) matrix $\mathbf{H} \in \mathbb{R}^{\Sigma^* \times \Sigma^*}$ is **Hankel** if for all $p, p', s, s' \in \Sigma^*$ such that $ps = p's'$, we have $\mathbf{H}(p,s) = \mathbf{H}(p',s')$. Given a Hankel matrix $\mathbf{H} \in \mathbb{R}^{\Sigma^* \times \Sigma^*}$, there exists a unique function $f : \Sigma^* \to \mathbb{R}$ such that $\mathbf{H}_f = \mathbf{H}$.*

A **weighted finite automaton** (WFA) of n states over Σ is a tuple $A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}, \boldsymbol{\beta} \rangle$, 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 transition matrix associated with each symbol $a \in \Sigma$. In this paper, we only consider automata with real weights. In this case, every WFA A realizes a function $f_A : \Sigma^* \to \mathbb{R}$, *i.e.*, given a string $x = x_1 \cdots x_t \in \Sigma^*$, it returns $f_A(x) = \boldsymbol{\alpha}^\top \mathbf{A}_{x_1} \cdots \mathbf{A}_{x_t} \boldsymbol{\beta} = \boldsymbol{\alpha}^\top \mathbf{A}_x \boldsymbol{\beta}$. Note that f can be realized by a WFA if and only if $\mathbf{H}_f$ has finite rank n, in which case n is the minimal number of states of any WFA realizing f (Carlyle and Paz, 1971; Fliess, 1974).

2.2. Hankel Operators and AAK Theory

Given a function $f : \mathbb{N} \to \mathbb{R}$, we consider the Hankel matrix $\mathbf{H}_f$ defined by $\mathbf{H}_f(i, j) = f(i + j)$. This matrix can be interpreted as the expression of a linear **Hankel operator** $H_f : \ell^2 \to \ell^2$ in terms of the canonical basis of the sequence space. Alternatively, using the Fourier isomorphism, Hankel operators can be defined in a complex function space. In fact, ℓ^2 can be embedded into $\ell^2(\mathbb{Z})$, which is isomorphic to $\mathcal{L}^2(\mathbb{T})$. Therefore, to each sequence $\boldsymbol{\mu} = (\mu_0, \mu_1, \dots) \in \ell^2$ we can associate two functions, $\mu^- = \sum_{j=0}^{\infty} \boldsymbol{\mu}_j z^{-j-1}$ and $\mu^+ = \sum_{j=0}^{\infty} \boldsymbol{\mu}_j z^j$. Conversely, we can associate any given function $\phi \in \mathcal{L}^2(\mathbb{T})$ with the sequence of its Fourier coefficients $\widehat{\phi}(n)$. The function space $\mathcal{L}^2(\mathbb{T})$ can be partitioned into two orthogonal subspaces, the **Hardy space** $\mathcal{H}^2$ and the negative Hardy space $\mathcal{H}^2_-$, containing functions that have only positive or negative Fourier coefficients, respectively. Note that $\mathcal{H}^2$ is isomorphic to the set of square-integrable functions analytic on the disc. For a detailed presentation of these results we refer the reader to Nikol'Skii (2002).

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

We remark that the symbol is not unique, and that we have $\|H_\phi\| \leq \|\phi\|_\infty$ (Nehari, 1957).

Every Hankel matrix $\mathbf{H}$ satisfies the Hankel property $\mathbf{H}(j, k) = \{\alpha_{j+k}\}_{j,k \geq 0}$. Another way to express this property is to rephrase it as an operator identity. We consider the **shift operator** S, with $S(x_0, x_1, \dots) = (0, x_0, x_1, \dots)$ and denote its left inverse by S^*. An operator H is Hankel if and only if the following **Hankel equation** is satisfied:

$$HS = S^* H. \tag{1}$$

Alternatively, we can consider the shift operator S in the function space, and generalize Equation (1) for operators $H : \mathcal{H}^2 \to \mathcal{H}^2_-$:

$$HS = \mathbb{P}_- S H. \tag{2}$$

We can now introduce the main result of Adamyan et al. (1971).

Theorem 3 (AAK Theorem) *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: $\|H_\phi - H_g\| = \|\mathbf{H} - \mathbf{G}\| = \sigma_k$. We say that $\mathbf{G}$ is the optimal approximation of size k of $\mathbf{H}$.*

The proof of this theorem relies on ϕ and g, the symbols of the original operator and of the best approximation, respectively, and on the following fundamental inequality:

$$\|H_\phi - H_g\| \leq \|\phi - g\|_\infty \leq \sigma_k. \tag{3}$$

The symbol of a finite-rank Hankel operator is a rational function (Kronecker, 1881).

2.3. Multivariable Operator Theory

The ideal formalism to extend the previous results to a NC setting is provided by Fock spaces. Let $\mathcal{H}_n$ be a Hilbert space, $\mathcal{H}_n^{\otimes k}$ the tensor product of k copies of $\mathcal{H}_n$, and $\mathcal{H}_n^{\otimes 0} := \mathbb{C}$.

Definition 4 *Let $\mathcal{H}_n$ be a n-dimensional Hilbert space. The **Fock space** F^2 of $\mathcal{H}_n$ is:*

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

Let $\mathbb{F}_n$ be the free monoid on n generators $g_1, \dots g_n$, with identity element g_0. Given $\alpha \in \mathbb{F}_n$, with $\alpha = g_{i_1} g_{i_2} \cdots g_{i_k}$, we define its length by $|\alpha| = k$, and $|g_0| = 0$. Analogously, we can define an element of the Fock space $e_\alpha = e_{i_1} \otimes e_{i_2} \otimes \cdots \otimes e_{i_k}$ and $e_{i_0} = 1$. Note that $\mathcal{B} = \{e_{g_i} : g_i \in \mathbb{F}_n\}$ is an orthonormal basis for the Fock space F^2. The Fock space is isomorphic to the Hilbert space of square summable sequences indexed by $\mathbb{F}_n$. The Fock space can be also identified with $\mathcal{H}^2(\mathbb{F}_n)$, a canonical NC analogue of the Hardy space. Given a collection of n NC variables (matrices or operators) $z = [z_1, \dots, z_n]$, with $z^\alpha := z_{i_1} \cdot z_{i_2} \cdots z_{i_k}$, we can consider $f \in F^2$ and represent it as a formal power series: $f(z) = \sum_{\alpha \in \mathbb{F}_n} \widehat{f}_\alpha z^\alpha$, converging for $\sum_i ||z_i z_i^*|| < 1$. We define the **NC Hardy space** as:

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

This means that we can choose between a "sequence" interpretation (F^2) or a "functional" interpretation ($\mathcal{H}^2(\mathbb{F}_n)$) of the results. We can now use sequences of operators to extend the definition of a Hankel operator in a way meaningful for NC spaces (Popescu, 2003).

Definition 5 *Let $X = [X_1, \dots, X_n]$, $X_i \in B(\mathcal{Y})$ be an arbitrary sequence of bounded operators on a Hilbert space $\mathcal{Y}$, and let $T = [T_1, \dots, T_n]$, $T_i \in B(\mathcal{H})$. Suppose $\mathcal{H} = \mathcal{H}_- \oplus \mathcal{H}_+$, with $\mathcal{H}_+$ invariant with respect to each $T_i \in B(\mathcal{H})$. Let $\mathbb{P}_-$ be the orthogonal projection on $\mathcal{H}_-$. A **NC Hankel operator** is a bounded linear operator $\Gamma : \mathcal{Y} \to \mathcal{H}_-$ such that:*

$$\Gamma X_i = \mathbb{P}_- T_i \Gamma \qquad \text{for any } i = 1, \dots, n. \tag{4}$$

The definition of symbol provided in the commutative case can be generalized as follows. A **multiplier** is a bounded linear operator $A : \mathcal{Y} \to \mathcal{H}$ such that $AX_i = T_i A$ for $i = 1, \dots, n$. Given a multiplier, it is always possible to associate with it a Hankel operator such that $||\Gamma_A|| \leq ||A||$, and defined as $\Gamma_A y = \mathbb{P}_- A y$ for $y \in \mathcal{Y}$ (Popescu, 2003).

We have the following noncommutative version of AAK theorem (Popescu, 2003).

Theorem 6 (NC AAK Theorem) *Let $X = [X_1, \dots, X_n]$, $X_i \in B(\mathcal{Y})$, and let $T = [T_1, \dots, T_n]$, $T_i \in B(\mathcal{H})$, be such that: $\|X_1 y_1 + \cdots + X_n y_n\|^2 \geq \|y_1\|^2 + \cdots + \|y_n\|^2$ and $\|T_1 h_1 + \cdots + T_n h_n\|^2 \leq \|h_1\|^2 + \cdots + \|h_n\|^2$ for $y_i \in \mathcal{Y}$ and $h_i \in \mathcal{H}$. Let $\Gamma_A : \mathcal{Y} \to \mathcal{H}_-$ be a NC Hankel operator, with $\Gamma X_i = \mathbb{P}_- T_i \Gamma$ for any $i = 1, \dots, n$. Then there exists an optimal approximation of Γ of size at most k.*

We conclude this section with a quick overview of NC rational functions (Jury et al., 2021a). A NC rational expression is any syntactically valid expressions involving several

NC variables, scalars, $+$, $\cdot$, $^{-1}$ and parentheses. A **NC rational function** is an equivalence class between rational expressions, where we say that r_1 and r_2 belong to the same equivalence class if r_1 can be transformed into r_2 by algebraic manipulations. Unlike the commutative case, a NC rational function does not admit a canonical coprime fraction representation (Kaliuzhnyi-Verbovetskyi and Vinnikov, 2009). A "canonical" way to represent NC rational functions comes from the theory of formal languages (Fliess, 1974; Berstel, 1979; Schützenberger, 1961). In particular, every NC rational function containing 0 in its domain admits a **minimal realization** of size n. If each $\mathbf{A}_j$ is a square matrix of size n, $\mathbf{b}$, $\mathbf{c}$ are vectors of size n and each complex variable z_j is a square matrix of size m, we have:

$$r(z) = \mathbf{c}^* \otimes \mathbf{1}_m \left(\mathbf{1}_n \otimes \mathbf{1}_m - \sum \mathbf{A}_j \otimes z_j\right)^{-1} \mathbf{b} \otimes \mathbf{1}_m = \mathbf{c}^* \left(\mathbf{1} - \sum \mathbf{A}_j z_j\right)^{-1} \mathbf{b}. \tag{5}$$

The last equality can be used as a more concise representation of the rational function.

3. A Framework for the Approximate Minimization Problem

Given a minimal WFA, we consider its Hankel matrix $\mathbf{H}$, with rank n equal to the number of states. We propose to reformulate the approximation problem as a low-rank approximation of $\mathbf{H}$. Thus, we can find a matrix of rank $k < n$, approximating $\mathbf{H}$ in the spectral norm, and recover a WFA having k states using the spectral method (Balle et al., 2014). A well known theorem by Eckart and Young (1936) states that the optimal approximation of $\mathbf{H}$ is obtained by truncating its SVD, but the resulting matrix is not necessarily Hankel. This is a problem, as we want to extract from the matrix a WFA. Leveraging AAK theory, it is possible to find a Hankel matrix attaining the same bound as the optimal approximation. In the next two sections, we show how to associate to $\mathbf{H}$ a Hankel operator and a symbol in the case of one-letter and multi-letter alphabets. This is a necessary step in order to apply Theorem 3, since its constructive proof relies on the definition of a symbol.

3.1. One-letter Alphabets

Let $|\Sigma| = 1$. Then, Σ^* can be identified with $\mathbb{N}$ by associating to each string its length. Since $\mathbb{N}$ can be embedded into $\mathbb{Z}$, we can interpret $f : \Sigma^* \to \mathbb{R}$ as $f : \mathbb{Z} \to \mathbb{R}$. This fundamental step allows us to apply the Fourier isomorphism to reformulate the problem in the Hardy space, where it can be solved using Theorem 3. This setting has been studied by Balle et al. (2021) and Lacroce et al. (2021) in the context of WFAs and black-box models, respectively.

3.1.1. Defining a Hankel Operator and a symbol

Let $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ be a minimal WFA with n states over a one-letter alphabet, computing $f_A : \Sigma^* \to \mathbb{R}$ with Hankel matrix $\mathbf{H}$. To apply AAK theory, we need to associate with $\mathbf{H}$ a Hankel operator. We can define two different operators, H_f and H_ϕ. On the one hand, we can consider the Hankel operator acting over sequences $H_f : \ell^2 \to \ell^2$, associated with the function $f_A : \Sigma^* \to \mathbb{R}$ with Hankel matrix defined by $\mathbf{H}(i, j) = f_A(i+j)$, for $i, j \geq 0$. On the other hand, we can interpret $\mathbf{H}$ as the matrix $\mathbf{H}_\phi$ associated with a Hankel operator over Hardy spaces, $H_\phi : \mathcal{H}^2 \to \mathcal{H}^2_-$. Now, the operator and matrix are related (by definition) to a complex function $\phi \in \mathcal{L}^2(\mathbb{T})$, the symbol. The entries of the matrix are defined by

means of the Fourier coefficients of ϕ as $\mathbf{H}(j,k) = \widehat{\phi}(-j-k-1)$ for $j,k \geq 0$. Note that the function $\mathbb{P}_-\phi = \widehat{\phi}(-j-k-1)$ is a complex rational function (Kronecker, 1881).

We can derive the relationship between f_A and ϕ: since we have $\mathbf{H} = \mathbf{H}_f = \mathbf{H}_\phi$, the two representations of the Hankel matrix need to coincide. We obtain: $f(n) = \widehat{\phi}(-n-1)$. Therefore, in the case of a WFA $A = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$, we obtain the rational function:

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

where the last equality holds if $\rho(A) < 1$. Now that we have the WFA's symbol, we can find the best approximation using the constructive proof of Theorem 3 (Balle et al., 2021).

3.2. Multi-letter Alphabets

In this section, we consider a WFA over Σ, with $|\Sigma| = d > 1$. In this case, Σ^* can be identified with $\mathbb{F}_d$, the free monoid generated by d elements. $\mathbb{F}_d$ is not abelian, so it cannot be embedded into $\mathbb{Z}$, and we cannot directly apply Fourier analysis like in the previous section. We first find a noncommutative version of Equation (2), and suitable transformations to play the roles of the shifts. We then find an appropriate generalization of the Hardy spaces. This allows us to define an equivalent of Definition 5 in the case of Hankel matrices arising from a WFA. Finally, we associate the NC Hankel operator with a NC rational function by leveraging a property of the multipliers.

3.2.1. Defining a Hankel Operator and a symbol

A WFA A over Σ, with $|\Sigma| = d$, computes a function $f : \mathbb{F}_d \to \mathbb{R}$, with Hankel matrix $\mathbf{H}$. This function can be interpreted as an element in the Fock space F^2 (see Appendix B). We consider the shift operators defined on the Fock space. For $i = 1, \ldots, d$, the **NC left shift** $S = (S_1, \ldots, S_d)$ and **NC right shift** $R = (R_1, \ldots, R_d)$ are defined by:

$$S_i(e_\alpha) := e_i \otimes e_\alpha = e_{i\alpha}, \qquad R_i(e_\alpha) := e_\alpha \otimes e_i = e_{\alpha i}.$$

We can express the right shift in terms of the left one by using a unitary operator U, the *flipping operator*: $R_i = U^* S_i U$, where $U(e_{i_1} \otimes e_{i_2} \otimes \cdots \otimes e_{i_k}) = e_{i_k} \otimes \cdots \otimes e_{i_2} \otimes e_{i_1}$. We obtain a NC version of Equation (1). A proof can be found in Appendix A.

Theorem 7 *Let $|\Sigma| = d$, and let $\mathbf{H}$ be a WFA's Hankel matrix. Let S and R be the NC left and right shifts on F^2, S^* and R^* their adjoints. Then, the following equation holds:*

$$\mathbf{H} S_i = R_i^* \mathbf{H} \qquad \text{for } i = 1, \ldots, d. \tag{6}$$

To extend Definition 5, we need to find appropriate spaces $\mathcal{Y}$, $\mathcal{H}_-$, and $\mathcal{H} = \mathcal{H}_- \oplus \mathcal{H}_+$. It has become clear that the natural noncommutative generalization of $\ell^2(\mathbb{N})$ is the Fock space, and that the NC Hardy space generalizes the Hardy space, so we set $\mathcal{Y} = \mathcal{H}_+ = F^2$ (or $\mathcal{Y} = \mathcal{H}_+ = \mathcal{H}^2(\mathbb{F}_d)$). In the one-letter case, the role of $\mathcal{H}$ was played by $\mathcal{L}^2(\mathbb{T}) \cong \ell^2(\mathbb{Z})$. A function $f \in \mathcal{L}^2(\mathbb{T})$ can be represented using the sequence of its Fourier coefficients, indexed by powers of the complex variable z. Analogously, we can set $\mathcal{H} = F_0^2 \oplus F^2$, and interpret it as the set of infinite sequences that are indexed by negative and nonnegative powers of the

NC variables $z_1, \dots z_d$ (the F_0^2 and F^2 components, respectively). In Appendix B we present an example of the application of this mathematical framework to a WFA over a 2-letter alphabet. The following theorem (proof in Appendix A) shows that the formalization we chose is not only suitable to describe the Hankel matrix of a WFA, but it also leads to an appropriate definition of NC Hankel operator.

Theorem 8 *Let $S = (S_1, \dots, S_d)$, $R = (R_1, \dots, R_d)$ be the left and right shifts on F^2, S^* and R^* their adjoints. Let $\mathcal{Y} = F^2$, $\mathcal{H} = F_0^2 \oplus F^2$, where $F_0^2 = \bigoplus_{k>0} (\mathbb{R}^d)^{\otimes k}$. We set $\mathcal{H}_- = F_0^2$ and $\mathcal{H}_+ = F^2$, and we define, for $i = 1, \dots d$, a bilateral shift on $\mathcal{H}$:*

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

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

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

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

is a NC Hankel operator according to Definition 5.

In the next theorem (proof in Appendix A), we show that the properties needed in Theorem 6 hold in our setting, *i.e.* that NC AAK theory can be applied to the study of WFAs.

Theorem 9 *Let $H : \mathcal{Y} \to \mathcal{H}_-$, with $HS_i = \mathbb{P}_- \overline{R}_i H$, be the NC Hankel operator defined in the previous theorem. Then:*

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

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

While the choice of the Fock space seems pretty natural, other spaces containing F^2 could have played the role of $\mathcal{H}$, such as the free group over d elements. We show in Appendix C why this choice is not ideal in our setting.

As seen in Section 2.3, in the NC case a role similar to that of the symbol is played by an operator, the multiplier. We want a functional representation of the multiplier depending on the original Hankel matrix (like the symbol in the one-letter case). To achieve this, we first analyze the multiplier and find that, with minimal manipulations, we can get a functional description of it. Then, we show that this description is strictly related to the original Hankel operator, and can be used to rewrite Equation (3) in the NC case.

We start by noting that, using the flipping operator, we can rewrite the property of the multiplier as: $UAS_a = S_a UA$. The operators commuting with the left shift are called S-analytic operators, and can be represented using a function θ. An S-analytic operator G has NC symbol θ if, for every v, $GS_a v = S_a \theta v$ (Popescu, 1993, 1995). Concretely, this means that we can represent the operator UA in terms of its NC symbol θ, which corresponds to the multiplication by the first column of the matrix of UA (this follows from Popescu (1993, Theorem 1.6)). Moreover, it is easy to show that $\|UA\| = \|U\theta\|_\infty$. Thus, if H is a NC Hankel operator with multiplier A, and θ is the NC symbol of UA, we have:

$$\|UH\| \leq \|UA\| = \|U\theta\|_\infty. \tag{7}$$

We refer to $U\theta$ as the **NC flipped symbol** of H. Note that it can be written as $U\theta = \phi + c$, with $\phi \in H_0^2(\mathbb{F}_d)$ and $c \in H^2(\mathbb{F}_d)$. By construction, ϕ corresponds to the multiplication by the first column of $\mathbf{H}$. Note that if R is a bounded operator, $\|UH - R\| = \|H - U^*R\|$ and since UH is also a NC Hankel operator, it makes sense to search for its optimal approximation. If we denote with UG the best approximation of UH, we have that G is the best approximation of H. We have a NC generalization of Equation (3):

$$\|UH - UG\| \leq \|UA - UB\| \leq \|\phi + c - \psi - d\|_\infty. \tag{8}$$

We conclude by deriving an expression for the NC flipped symbol associated to a WFA. Let $A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}, \boldsymbol{\beta} \rangle$ be a WFA computing a function f, let $\mathbf{H}$ be its Hankel matrix and H the NC Hankel operator. The NC flipped symbol associated with H is defined using the entries of the first column of $\mathbf{H}$. Its series expression is:

$$\mathbb{P}_-(\phi + c) = \sum_{a \in \mathbb{F}_n} f(a) z^a = \sum_{a \in \mathbb{F}_n} \boldsymbol{\alpha}^\top \mathbf{A}^a \boldsymbol{\beta} z^\alpha = \boldsymbol{\alpha}^\top (\mathbf{1} - \sum \mathbf{A}_j z_j)^{-1} \boldsymbol{\beta}.$$

Note that ϕ is a rational function: in the noncommutative case also, there is a tight connection between WFAs and (NC) rational functions. This is very relevant, since in the one-letter case rewriting Equation (3) in terms of the WFA's parameters is the key step to find the best approximation (Balle et al., 2021). At this stage, it is not clear if the proof of Theorem 8 can be made constructive. Nonetheless, by obtaining a noncommutative counterpart of this equation, expressed using the parameters of a WFA, we have built the machinery necessary to attack the problem in the case of multi-letter alphabets.

4. Conclusion

In this paper, we propose a way to associate a Hankel operator and a complex rational function to the Hankel matrix of a given WFA. This allows us to highlight the connections between approximate minimization and AAK theory. The application of AAK theory to the approximate minimization problem in the one-letter setting has been studied by Balle et al. (2021) and Lacroce et al. (2021) for WFAs and black boxes, respectively. To the best of our knowledge, this is the first attempt to apply AAK theory to the multi-letter case.

The approximate minimization problem is an interesting alternative to extraction when trying to approximate a black box (like RNNs) with a WFA. It allows to find the best approximation of a given size, directly improving interpretability and reducing the computational cost. The results in this paper can be easily generalized to the black-box setting.

The framework we proposed is a key step towards solving the approximate minimization problem, as it allows us to rephrase it in terms of noncommutative AAK theory, where we know that a solution exists (Adamyan et al., 1971; Popescu, 2003). In the commutative setting, this is enough to construct the optimal approximation of a given size. Unfortunately, in the noncommutative setting AAK theorem is not constructive, so the problem of finding the best approximation remains open. Recent progress in the field of noncommutative multivariable operator theory (Jury et al., 2021b; Ball and Bolotnikov, 2021) leaves us hopeful that this challenge can be addressed. We think that the problem of constructing the optimal approximation is very relevant, as solving it would allow us to find a provable algorithm for the approximate minimization problem of black boxes, and provide us with a metric between different classes of models.

References

Vadim M. Adamyan, Damir Zyamovich Arov, and Mark Grigorievich Krein. Analytic Properties of Schmidt Pairs for a Hankel Operator and the Generalized Schur–Takagi problem. *Mathematics of The Ussr-sbornik*, 15:31–73, 1971.

Stéphane Ayache, Rémi Eyraud, and Noé Goudian. Explaining Black Boxes on Sequential Data Using Weighted Automata. In *Proceedings of the 14th International Conference on Grammatical Inference, ICGI 2018, Wrocław, Poland, September 5-7, 2018*, volume 93 of *Proceedings of Machine Learning Research*, pages 81–103. PMLR, 2018. URL http://proceedings.mlr.press/v93/ayache19a.html.

Joseph A. Ball and Vladimir Bolotnikov. *Noncommutative Function-Theoretic Operator Theory and Applications*. Cambridge Tracts in Mathematics. Cambridge University Press, 2021.

Borja Balle, Xavier Carreras, Franco M. Luque, and Ariadna Quattoni. Spectral Learning of Weighted Automata - A Forward–Backward Perspective. *Mach. Learn.*, 96(1-2):33–63, 2014. doi: 10.1007/s10994-013-5416-x. URL https://doi.org/10.1007/s10994-013-5416-x.

Borja Balle, Prakash Panangaden, and Doina Precup. Singular Value Automata and Approximate Minimization. *Math. Struct. Comput. Sci.*, 29(9):1444–1478, 2019. doi: 10.1017/S0960129519000094. URL https://doi.org/10.1017/S0960129519000094.

Borja Balle, Clara Lacroce, Prakash Panangaden, Doina Precup, and Guillaume Rabusseau. Optimal Spectral-Norm Approximate Minimization of Weighted Finite Automata. In Nikhil Bansal, Emanuela Merelli, and James Worrell, editors, *48th International Colloquium on Automata, Languages, and Programming, ICALP 2021, July 12-16, 2021, Glasgow, Scotland (Virtual Conference)*, volume 198 of *LIPIcs*, pages 118:1–118:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. doi: 10.4230/LIPIcs.ICALP.2021.118. URL https://doi.org/10.4230/LIPIcs.ICALP.2021.118.

J. Berstel. Transductions and Context-Free Languages. In *Teubner Studienbücher : Informatik*, 1979.

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

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936. doi: 10.1007/BF02288367. URL https://doi.org/10.1007/BF02288367.

Rémi Eyraud and Stéphane Ayache. Distillation of Weighted Automata from Recurrent Neural Networks Using a Spectral Approach. *CoRR*, abs/2009.13101, 2020. URL https://arxiv.org/abs/2009.13101.

Michel Fliess. Matrice de Hankel. *Journal de Mathématique Pures et Appliquées*, 5:197–222, 1974.

Michael Jury, Robert Martin, and Eli Shamovich. Non-commutative rational functions in the full fock space. *Transactions of the American Mathematical Society*, 2021a.

Michael T Jury, Robert TW Martin, and Eli Shamovich. Blaschke–singular–outer factorization of free non-commutative functions. *Advances in Mathematics*, 384:107720, 2021b.

Dmitry S. Kaliuzhnyi-Verbovetskyi and Victor Vinnikov. Singularities of rational functions and minimal factorizations: The noncommutative and the commutative setting. *Linear Algebra and its Applications*, 430(4):869–889, 2009. ISSN 0024-3795. doi: https://doi.org/10.1016/j.laa.2008.08.027. URL https://www.sciencedirect.com/science/article/pii/S0024379508003893.

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

Clara Lacroce, Prakash Panangaden, and Guillaume Rabusseau. Extracting weighted automata for approximate minimization in language modelling. In Jane Chandlee, Rémi Eyraud, Jeff Heinz, Adam Jardine, and Menno van Zaanen, editors, *Proceedings of the Fifteenth International Conference on Grammatical Inference*, volume 153 of *Proceedings of Machine Learning Research*, pages 92–112. PMLR, 23–27 Aug 2021. URL https://proceedings.mlr.press/v153/lacroce21a.html.

Zeev Nehari. On Bounded Bilinear Forms. *Annals of Mathematics*, 65(1):153–162, 1957.

Nikolai K. Nikol'Skii. *Operators, Functions and Systems: An Easy Reading*, volume 92 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2002.

Takamasa Okudono, Masaki Waga, Taro Sekiyama, and Ichiro Hasuo. Weighted Automata Extraction from Recurrent Neural Networks via Regression on State Spaces. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 5306–5314. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/5977.

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

Gelu Popescu. Multi-Analytic Operators on Fock Spaces. *Mathematische Annalen*, 303(1): 31–46, 1995.

Gelu Popescu. Multivariable Nehari Problem and Interpolation. *Journal of Functional Analysis*, 200:536–581, 2003. ISSN 0022-1236. doi: 10.1016/S0022-1236(03)00078-8. URL https://doi.org/10.1016/S0022-1236(03)00078-8.

Guillaume Rabusseau, Tianyu Li, and Doina Precup. Connecting Weighted Automata and Recurrent Neural Networks through Spectral Learning. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of

Proceedings of Machine Learning Research, pages 1630–1639. PMLR, 2019. URL http://proceedings.mlr.press/v89/rabusseau19a.html.

M.P. Schützenberger. On the definition of a family of automata. *Information and Control*, 4(2):245–270, 1961. ISSN 0019-9958. doi: https://doi.org/10.1016/S0019-9958(61)80020-X. URL https://www.sciencedirect.com/science/article/pii/S001999586180020X.

Gail Weiss, Yoav Goldberg, and Eran Yahav. Learning Deterministic Weighted Automata with Queries and Counterexamples. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8558–8569, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/d3f93e7766e8e1b7ef66dfdd9a8be93b-Abstract.html.

Appendix A. Proofs

Proof [Theorem 7] For this proof, we leverage the functional representation of the Fock space. We recall that in the NC Hardy space, the left shift is equivalent to left multiplication by one of the noncommutative variables: $S_i f = z_i f$. Moreover, the function $f : \Sigma^* \to \mathbb{R}$ associated to the Hankel matrix can be represented by means of a formal power series in the NC Hardy space $f = \sum_{\alpha \in \Sigma^*} f(\alpha) z^\alpha$. This function corresponds to the first column of the Hankel matrix. Analogously, it is easy to see that the column at index α is: $\mathbf{H}e_\alpha = \sum_{\beta \in \Sigma^*} f(\beta\alpha) z^\beta$. Therefore:

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

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

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

Thus, for any $i = 1, \ldots, d$, we have: $\mathbf{H}S_i = R_i^* \mathbf{H}$, which concludes the proof. This shows that the Hankel matrix arising from a WFA defined over a multi-letter alphabets satisfies the NC version of the Hankel equation. ■

Proof [Theorem 8] In order to prove the theorem we need to verify that $\mathcal{H}_- \subset \mathcal{H}$, and that if $\mathcal{H} = \mathcal{H}_- \oplus \mathcal{H}_+$, then $\mathcal{H}_+$ is invariant under each $\overline{R}_i$. In particular, we want to show that these properties are satisfied when $\mathcal{Y} = F^2$, $\mathcal{H}_- = F_0^2$ and $\mathcal{H} = F_0^2 \oplus F^2$. The first property follows directly from the definition of $\mathcal{H}$: $F_0^2 \subset F_0^2 \oplus F^2$. As for the second property, we note that $\mathcal{H}_- = F_0^2$, it follows by definition that $\mathcal{H}_+ = F^2$.

We want to show that: $\overline{R}_1 \mathcal{H}_+ + \ldots \overline{R}_d \mathcal{H}_+ \subseteq \mathcal{H}_+$, *i.e.* that for any $h_i \in \mathcal{H}_+$ we have $\overline{R}_1 h_1 + \ldots \overline{R}_d h_d \in \mathcal{H}_+$. Since $\overline{R}_i(e_\alpha) = R_i(e_\alpha)$ for $e_\alpha \in F^2$, the condition can be reformulated as: $R_i(e_{\alpha_1}) + \ldots R_d(e_{\alpha_n}) \in F^2$, which holds by definition of R, since $R_i(e_\alpha) = e_{\alpha i} \in F^2$ for any α, and the linear combination of elements in F^2 is an element in F^2. ■

Proof [Theorem 9]

(a) Leveraging the fact that the shifts have pairwise orthogonal ranges, so $S_i^* S_j = \mathbf{1}\delta_{i,j}$, and that each S_i is an isometry, we obtain:

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

(b) The shifts have orthogonal ranges, so the result holds with the equality.

■

Appendix B. Example

Example 1 *Let $\Sigma = \{a, b\}$, ε the empty string. Σ^* corresponds to the free monoid generated by two elements $\mathbb{F}_2$, where the generators are $g_1 = a$ and $g_2 = b$. A word $\alpha = aba$ can be seen as an element in $\mathbb{F}_2$, with $\alpha = aba = g_1 g_2 g_1$, and the corresponding element in the Fock space F^2 is $e_\alpha = e_1 \otimes e_2 \otimes e_1$. A function $f : \Sigma^* \to \mathbb{R}$ can be viewed either as an element in the Fock space F^2, using a sequence interpretation:*

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

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

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

As we can see, we obtain a bi-infinite sequence, indexed by the powers of the NC variables:

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

Now, we can consider the right shift $S = (S_1, S_2)$, with: $S_1(e_\alpha) = e_{a\alpha}$ and $S_2(e_\alpha) = e_{b\alpha}$. The adjoint of S is defined as: $S_1^*(e_\alpha) = e_{\alpha'}$ if $\alpha = a\alpha'$, zero otherwise. The right shift and its adjoint can be defined in a similar way.

Let $A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}, \boldsymbol{\beta}\rangle$ be a WFA computing a function f, with Hankel matrix $\mathbf{H}$.

$$\mathbf{H} = \begin{pmatrix} f(\varepsilon) & f(a) & \dots & f(ba) & \dots & f(aba) & \dots \\ f(a) & f(aa) & \dots & f(aba) & \dots & f(aaba) & \dots \\ f(b) & f(ba) & \dots & f(bba) & \dots & f(baba) & \dots \\ f(aa) & f(aaa) & \dots & f(aaba) & \dots & f(aaaba) & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}.$$

It is easy to see that:

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

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

$$R_a^*\mathbf{H}(e_{ba}) = R_a^* \sum_{\beta\in\Sigma^*} f(\beta ba)z^\beta = \sum_{\beta'\in\Sigma^*} f(\beta' aba)z^{\beta'} = f(aba)z_0 + f(aaba)z_1 + \dots.$$

We can obtain the same results for S_b and R_b^*, so we can see that the Hankel equation holds.

Appendix C. The Free Group

We denote with $\mathbb{F}_d^*$ the free group on d elements, and with $\ell(\mathbb{F}_d^*)$ the set of sequences indexed by elements in the free group. We can now show that, by setting $\mathcal{H} = \ell(\mathbb{F}_d^*)$, the conditions of Theorem 8 are satisfied, but the ones of Theorem 9 are not. It is easy to see that $\mathcal{H}_- \subset \mathcal{H}$, $\mathcal{H}_+$ is invariant under the bilateral shift, and that property (a) of Theorem 9 is satisfied. On the other hand, property (b) does not hold anymore. The components of the bilateral shifts don't have orthogonal ranges, as $\overline{R}_i^*\overline{R}_j \in \mathcal{H}$ even when $i \neq j$. Intuitively the space is "too big" for property (b) to hold. If we consider the intuition provided earlier about indexing the elements of $\mathcal{H} = F_0^2 \oplus F^2$ using negative and nonnegative exponents, we have that in the case of the free group any combination of positive and negative exponents is allowed. Therefore, when defined on $\ell(\mathbb{F}_d^*)$, the adjoint of the shift is:

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

Our objective is to be able to apply Theorem 6, to conclude that it is possible to find a optimal approximation of the NC Hankel operator associated to a WFA. For this to happen, we need Theorem 9 to hold. Thus, the free group is not a viable option in our setting.