A Canonical Form for Weighted Automata and Applications to Approximate Minimization

Borja Balle, Prakash Panangaden, and Doina Precup
School of Computer Science, McGill University
Montreal, Canada
{bballe|prakash|dprecup}@cs.mcgill.ca

***Abstract*—We study the problem of constructing approximations to a weighted automaton. Weighted finite automata (WFA) are closely related to the theory of rational series. A rational series is a function from strings to real numbers that can be computed by a WFA. Among others, this includes probability distributions generated by hidden Markov models and probabilistic automata. The relationship between rational series and WFA is analogous to the relationship between regular languages and ordinary automata. Associated with such rational series are infinite matrices called Hankel matrices which play a fundamental role in the theory of minimal WFA. Our contributions are: (1) an effective procedure for computing the singular value decomposition (SVD) of such *infinite* Hankel matrices based on their finite representation in terms of WFA; (2) a new *canonical form* for WFA based on this SVD decomposition; and, (3) an algorithm to construct *approximate minimizations* of a given WFA. The goal of our approximate minimization algorithm is to start from a minimal WFA and produce a smaller WFA that is close to the given one in a certain sense. The desired size of the approximating automaton is given as input. We give bounds describing how well the approximation emulates the behavior of the original WFA.**

The study of this problem is motivated by the analysis of machine learning algorithms that synthesize weighted automata from spectral decompositions of finite Hankel matrices. It is known that when the number of states of the target automaton is correctly guessed, these algorithms enjoy consistency and finite-sample guarantees in the probably approximately correct (PAC) learning model. It has also been suggested that asking the learning algorithm to produce a model smaller than the true one will still yield useful models with reduced complexity. Our results in this paper vindicate these ideas and confirm intuitions provided by empirical studies. Beyond learning problems, our techniques can also be used to reduce the complexity of any algorithm working with WFA, at the expense of incurring a small, controlled amount of error.

***Index Terms*—weighted automata; canonical form; Hankel matrices; approximate minimization**

I. Introduction

We address a relatively new issue for the logic and computation community: the *approximate* minimization of transition systems or automata. This concept is appropriate for systems that are quantitative in some sense: weighted automata, probabilistic automata of various kinds and timed automata. This paper focuses on weighted automata where we are able to make a number of contributions that combine ideas from duality with ideas from the theory of linear operators and their spectrum. Our new contributions are

- An algorithm for the SVD decomposition of *infinite* Hankel matrices based on their representation in terms of weighted automata.
- A new canonical form for weighted automata arising from the SVD of its corresponding Hankel matrix.
- An algorithm to construct approximate minimizations of given weighted automata by truncating the canonical form.

Minimization of automata has been a major subject since the 1950s, starting with the now classical work of the pioneers of automata theory. Recently there has been activity on novel algorithms for minimization based on duality [1], [2] which are ultimately based on a remarkable algorithm due to Brzozowski from the 1960s [3]. The general co-algebraic framework permits one to generalize Brzozowski's algorithm to other classes of automata like weighted automata.

Weighted automata are very useful in a variety of practical settings, such as machine learning (where they are used to represent predictive models for time series data and text), but also in the general theory of quantitative systems. There has also been interest in this type of representation, for example, in concurrency theory [4] and in semantics [5]. We discuss the machine learning motivations at greater length, as they are the main driver for the present work. However, we emphasize that the genesis of one set of key ideas came from previous work on a co-algebraic view of minimization.

Spectral techniques for learning latent variable models have recently drawn a lot of attention in the machine learning community. Following the significant milestone papers [6], [7], in which an efficient spectral algorithm for learning hidden Markov models (HMM) and stochastic rational languages was given, the field has grown very rapidly. The original algorithm, which is based on singular value decompositions of Hankel matrices, has been extended to reduced-rank HMM [8], predictive state representations (PSR) [9], finite-state transducers [10], [11], and many other classes of functions on strings [12], [13], [14]. Although each of these papers works with slightly different problems and analysis techniques, the key ingredient turns out to be always the same: parametrize the target model as a weighted finite automaton (WFA) and learn this WFA from the SVD of a finite sub-block of its Hankel matrix [15]. Therefore, it is possible (and desirable) to study all these learning algorithms from the point of view of rational series, which are exactly the class of real-valued functions on strings that can be computed by WFA. In addition to their use in spectral learning algorithms, weighted automata are also commonly used in other areas of pattern recognition for sequences, including: speech recognition [16], image compression [17], natural language processing [18], model checking [19], and machine translation [20].

Part of the appeal of spectral learning techniques comes from their computational superiority when compared to iterative algorithms like Expectation–Maximization (EM) [21]. Another very attractive property of spectral methods is the possibility of proving rigorous statistical guarantees about the learned hypothesis. For example, under a realizability assumption, these methods are known to be consistent and amenable to finite-sample analysis in the PAC sense [6]. An important detail is that, in addition to realizability, these results work under the assumption that the user correctly guesses the number of latent states of the target distribution. Though this is not a real caveat when it comes to using these algorithms in practice – the optimal number of states can be identified using a model selection procedure [22] – it is one of the barriers in extending the statistical analysis of spectral methods to the non-realizable setting.

Tackling the non-realizability question requires, as a special case, dealing with the situation in which data is generated from a WFA with n states and the learning algorithm is asked to produce a WFA with $\hat{n} < n$ states. This case is already a non-trivial problem which – barring the noisiness introduced by the use of statistical data instead of the original WFA – can be easily interpreted as an approximate minimization of WFA. From this point of view, the possibility of using spectral learning algorithms for approximate minimization of a small class of hidden Markov models has been recently considered in [23]. This paper also presents some restricted theoretical results bounding the error between the original and minimized HMM in terms of the total variation distance. Though incomparable to ours, these bounds are the closest work in the literature to our approach[1]. Another paper on which the issue of approximate minimization of weighted automata is considered in a tangential manner is [25]. In this case the authors again focus on an ℓ^1-like accuracy measure to compare two automata: an original one, and another one obtained by removing transitions with small weights occurring during an exact minimization procedure. Though the removal operation is introduced as a means of obtaining a numerically stable minimization algorithm, the paper also presents some experiments exploring the effect of removing transitions with larger weights. With the exception of these timid results, the problem of approximate minimization remains largely unstudied. In the present paper we set out to initiate the systematic study of approximate minimization of WFA. We believe our results – beyond their intrinsic automata-theoretic interest – will also provide tools for addressing important problems in learning theory, including the robust statistical analysis of spectral learning algorithms.

Let us conclude this introduction by mentioning the potential wide applicability of our results in the field of algorithms for manipulating, combining, and operating with quantitative systems. In particular, the possibility of obtaining reduced-size models incurring a small, controlled amount of error might provide a principled way for speeding up a number of such algorithms.

The content of the paper is organized as follows. Section II defines the notation that will be used throughout the paper and reviews a series of well-known results that will be needed. Section III establishes the existence of a canonical form for WFA and provides a polynomial-time algorithm for computing it (the first major contribution of this work). The computation of this canonical form lies at the heart of our approximate minimization algorithm, which is described and analyzed in Section IV. Our main theoretical result in this section is to establish bounds describing how well the approximation obtained by the algorithm emulates the behavior of the original WFA. In Section V we discuss two technical aspects of our work: its relation and consequences with the mathematical theory of low-rank approximation of rational series; and the (ir)relevance of an assumption made in our results from Sections III and IV. We conclude with Section VI, where we point out interesting future research directions.

II. Background

A. Notation for Matrices

Given a positive integer d, we denote $[d] = \{1, \ldots, d\}$. We use bold letters to denote vectors $\mathbf{v} \in \mathbb{R}^d$ and matrices $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$. Unless explicitly stated, all vectors are column vectors. We write $\mathbf{I}$ for the identity matrix, $\operatorname{diag}(a_1, \ldots, a_n)$ for a diagonal matrix with $a_1, \ldots, a_n$ in the diagonal, and $\operatorname{diag}(\mathbf{M}_1, \ldots, \mathbf{M}_n)$ for the block-diagonal matrix containing the square matrices $\mathbf{M}_i$ along the diagonal. The ith coordinate vector $(0, \ldots, 0, 1, 0, \ldots, 0)^\top$ is denoted by $\mathbf{e}_i$. For a matrix $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$, $i \in [d_1]$, and $j \in [d_2]$, we use $\mathbf{M}(i, :)$ and $\mathbf{M}(:, j)$ to denote the ith row and the jth column of $\mathbf{M}$ respectively. Given a matrix $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ we can consider the vector $\operatorname{vec}(\mathbf{M}) \in \mathbb{R}^{d_1 \cdot d_2}$ obtained by concatenating the columns of $\mathbf{M}$ so that $\operatorname{vec}(\mathbf{M})((i-1)d_2 + j) = \mathbf{M}(i, j)$. Given two matrices $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ and $\mathbf{M}' \in \mathbb{R}^{d_1' \times d_2'}$ we denote their Kronecker (or tensor) product by $\mathbf{M} \otimes \mathbf{M}' \in \mathbb{R}^{d_1 d_1' \times d_2 d_2'}$, with entries given by $(\mathbf{M} \otimes \mathbf{M}')((i-1)d_1' + i', (j-1)d_2' + j') = \mathbf{M}(i, j)\mathbf{M}'(i', j')$, where $i \in [d_1]$, $j \in [d_2]$, $i' \in [d_1']$, and $j' \in [d_2']$. For simplicity, we will sometimes write $\mathbf{M}^{\otimes 2} = \mathbf{M} \otimes \mathbf{M}$, and similarly for vectors. A *rank factorization* of a rank n matrix $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ is an expression of the form $\mathbf{M} = \mathbf{Q}\mathbf{R}$ where $\mathbf{Q} \in \mathbb{R}^{d_1 \times n}$ and $\mathbf{R} \in \mathbb{R}^{n \times d_2}$ are full-rank matrices.

Given a matrix $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ of rank n, its *singular value decomposition* (SVD)[2] is a decomposition of the form $\mathbf{M} =$

[1] After the submission of this manuscript we became aware of the concurrent work [24], where a problem similar to the one considered here is addressed, albeit different methods are used and the results are not directly comparable.

[2] To be more precise, this is a *reduced* singular value decomposition, since the inner dimensions of the decomposition are all equal to the rank. In this paper we shall always use the term SVD to mean reduced SVD.

$\mathbf{UDV}^\top$ where $\mathbf{U} \in \mathbb{R}^{d_1 \times n}$, $\mathbf{D} \in \mathbb{R}^{n \times n}$, and $\mathbf{V} \in \mathbb{R}^{d_2 \times n}$ are such that: $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}$, and $\mathbf{D} = \operatorname{diag}(\mathfrak{s}_1, \dots, \mathfrak{s}_n)$ with $\mathfrak{s}_1 \geq \cdots \geq \mathfrak{s}_n > 0$. The columns of $\mathbf{U}$ and $\mathbf{V}$ are called left and right singular vectors respectively, and the $\mathfrak{s}_i$ are its singular values. The SVD is unique (up to sign changes in associate singular vectors) whenever all inequalities between singular values are strict. A similar spectral decomposition exists for bounded operators between separable Hilbert spaces. In particular, for finite-rank bounded operators one can write the infinite matrix corresponding to the operator in a fixed basis, and recover a concept of reduced SVD decomposition for such infinite matrices which shares the same properties described above for finite matrices [26].

For $1 \leq p \leq \infty$ we will write $\|\mathbf{v}\|_p$ for the ℓ^p norm of vector $\mathbf{v}$. The corresponding *induced norm* on matrices is $\|\mathbf{M}\|_p = \sup_{\|\mathbf{v}\|_p = 1} \|\mathbf{M}\mathbf{v}\|_p$. In addition to induced norms, we will also need to define Schatten norms. If $\mathbf{M}$ is a rank-n matrix with singular values $\mathbf{s} = (\mathfrak{s}_1, \dots, \mathfrak{s}_n)$, the *Schatten p-norm* of $\mathbf{M}$ is given by $\|\mathbf{M}\|_{\mathrm{S},p} = \|\mathbf{s}\|_p$. Most of these norms have given names: $\|\cdot\|_2 = \|\cdot\|_{\mathrm{S},\infty} = \|\cdot\|_{\mathrm{op}}$ is the *operator (or spectral) norm*; $\|\cdot\|_{\mathrm{S},2} = \|\cdot\|_{\mathrm{F}}$ is the *Frobenius norm*; and $\|\cdot\|_{\mathrm{S},1} = \|\cdot\|_{\mathrm{tr}}$ is the *trace (or nuclear) norm*. For a matrix $\mathbf{M}$ the *spectral radius* is the largest modulus $\rho(\mathbf{M}) = \max_i |\lambda_i(\mathbf{M})|$ among the eigenvalues of $\mathbf{M}$. For a square matrix $\mathbf{M}$, the series $\sum_{k \geq 0} \mathbf{M}^k$ converges if and only if $\rho(\mathbf{M}) < 1$, in which case the sum yields $(\mathbf{I} - \mathbf{M})^{-1}$.

Sometimes we will name the columns and rows of a matrix using ordered index sets $\mathcal{I}$ and $\mathcal{J}$. In this case we will write $\mathbf{M} \in \mathbb{R}^{\mathcal{I} \times \mathcal{J}}$ to denote a matrix of size $|\mathcal{I}| \times |\mathcal{J}|$ with rows indexed by $\mathcal{I}$ and columns indexed by $\mathcal{J}$.

B. Weighted Automata, Rational Series, and Hankel Matrices

Let Σ be a fixed finite alphabet with $|\Sigma| = k$ symbols, and $\Sigma^\star$ the set of all finite strings with symbols in Σ. We use λ to denote the empty string. Given two strings $p, s \in \Sigma^\star$ we write $w = ps$ for their concatenation, in which case we say that p is a prefix of w and s is a suffix of w. We denote by $|w|$ the length (number of symbols) in a string $w \in \Sigma^\star$. Given a set of strings $X \subseteq \Sigma^\star$ and a function $f : \Sigma^\star \to \mathbb{R}$, we denote by $f(X)$ the summation $\sum_{x \in X} f(x)$ if defined. For example, we will write $f(\Sigma^t) = \sum_{|x|=t} f(x)$ for any $t \geq 0$.

Now we introduce our notation for weighted automata. We want to note that we will not be dealing with weights on arbitrary semi-rings; this paper only considers automata with real weights, with the usual addition and multiplication operations. In addition, instead of resorting to the usual description of automata as directed graphs with labelled nodes and edges, we will use a linear-algebraic representation, which is more convenient. A *weighted finite automata* (WFA) of dimension n over Σ is a tuple $A = \langle \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\infty, \{\mathbf{A}_\sigma\}_{\sigma \in \Sigma} \rangle$ where $\boldsymbol{\alpha}_0 \in \mathbb{R}^n$ is the vector of *initial weights*, $\boldsymbol{\alpha}_\infty \in \mathbb{R}^n$ is the vector of *final weights*, and for each symbol $\sigma \in \Sigma$ the matrix $\mathbf{A}_\sigma \in \mathbb{R}^{n \times n}$ contains the *transition weights* associated with σ. Note that in this representation a fixed initial state is given by $\boldsymbol{\alpha}_0$ (as opposed to formalisms that only specify a transition structure), and the transition endomorphisms $\mathbf{A}_\sigma$ and the final linear form $\boldsymbol{\alpha}_\infty$ are given in a fixed basis on $\mathbb{R}^n$ (as opposed to abstract descriptions where these objects are represented as basis-independent elements over some n-dimensional vector space).

We will use $\dim(A)$ to denote the dimension of a WFA. The state-space of a WFA of dimension n is identified with the integer set $[n]$. Every WFA A *realizes* a function $f_A : \Sigma^\star \to \mathbb{R}$ which, given a string $x = x_1 \cdots x_t \in \Sigma^\star$, produces

$$f_A(x) = \boldsymbol{\alpha}_0^\top \mathbf{A}_{x_1} \cdots \mathbf{A}_{x_t} \boldsymbol{\alpha}_\infty = \boldsymbol{\alpha}_0^\top \mathbf{A}_x \boldsymbol{\alpha}_\infty \ ,$$

where we defined the shorthand notation $\mathbf{A}_x = \mathbf{A}_{x_1} \cdots \mathbf{A}_{x_t}$ that will be used throughout the paper. A function $f : \Sigma^\star \to \mathbb{R}$ is called *rational* if there exists a WFA A such that $f = f_A$. The *rank* of a rational function f is the dimension of the smallest WFA realizing f. We say that a WFA is *minimal* if $\dim(A) = \operatorname{rank}(f_A)$.

An important operation on WFA is *conjugation* by an invertible matrix. Suppose A is a WFA of dimension n and $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is invertible. Then we can define the WFA

$$A' = \mathbf{Q}^{-1} A \mathbf{Q} = \langle \mathbf{Q}^\top \boldsymbol{\alpha}_0, \mathbf{Q}^{-1} \boldsymbol{\alpha}_\infty, \{\mathbf{Q}^{-1} \mathbf{A}_\sigma \mathbf{Q}\} \rangle \ . \quad (1)$$

It is immediate to check that $f_A = f_{A'}$. This means that the function computed by a WFA is invariant under conjugation, and that given a rational function f, there exist infinitely many WFA realizing f. In addition, the following result characterizes all minimal WFA realizing a particular rational function.

Theorem 1 ([27]). *If A and B are minimal WFA realizing the same function, then $B = \mathbf{Q}^{-1} A \mathbf{Q}$ for some invertible $\mathbf{Q}$.*

A function $f : \Sigma^\star \to \mathbb{R}$ can be trivially identified with an element from the free vector space $\mathbb{R}^{\Sigma^\star}$. This vector space contains several subspaces which will play an important role in the rest of the paper. One is the subspace of all rational functions, which we denote by $\mathcal{R}(\Sigma)$. Note that $\mathcal{R}(\Sigma)$ is a linear subspace, because if $f, g \in \mathcal{R}(\Sigma)$ and $c \in \mathbb{R}$, then cf and $f + g$ are both rational [27]. Another important family of subspaces of $\mathbb{R}^{\Sigma^\star}$ are the ones containing all functions with *finite* p-norm for some $1 \leq p \leq \infty$, which is given by $\|f\|_p^p = \sum_{x \in \Sigma^\star} |f(x)|^p$ for finite p, and $\|f\|_\infty = \sup_{x \in \Sigma^\star} |f(x)|$; we denote this space by $\ell^p(\Sigma)$. Note that like in the usual theory of Banach spaces of sequences, we have $\ell^p(\Sigma) \subset \ell^q(\Sigma)$ for $p < q$. Of these, $\ell^2(\Sigma)$ can be endowed with the structure of a separable Hilbert space with the inner product $\langle f, g \rangle = \sum_{x \in \Sigma^\star} f(x) g(x)$. Recall that in this case we have the *Cauchy–Schwarz inequality* $\langle f, g \rangle^2 \leq \|f\|_2^2 \, \|g\|_2^2$. In addition, we have its generalization, *Hölder's inequality*: given $f \in \ell^p(\Sigma)$ and $g \in \ell^q(\Sigma)$ with $p^{-1} + q^{-1} \leq 1$, then $\|f \cdot g\|_1 \leq \|f\|_p \|g\|_q$, where $(f \cdot g)(x) = f(x) g(x)$. By intersecting any of the previous subspaces with $\mathcal{R}(\Sigma)$ one obtains $\ell^p_{\mathcal{R}}(\Sigma) = \mathcal{R}(\Sigma) \cap \ell^p(\Sigma)$, the normed vector space containing all rational functions with finite p-norm. In most cases the alphabet Σ will be clear from the context and we will just write $\mathcal{R}$, ℓ^p, and $\ell^p_{\mathcal{R}}$.

The space $\ell^1_{\mathcal{R}}$ of *absolutely convergent rational series* will play a central role in the theory to be developed in this paper.

An important example of functions in $\ell^1_{\mathcal{R}}$ is that of probability distributions over $\Sigma^\star$ realized by WFA, also known as rational stochastic languages. Formally speaking, these are rational functions $f \in \mathcal{R}$ satisfying the constraints $f(x) \geq 0$ and $\sum_x f(x) = 1$. This implies that $\ell^1_{\mathcal{R}}$ includes all functions realized by probabilistic automata with stopping probabilities [28], hidden Markov models with absorbing states [29], and predictive state representations for dynamical systems with discounting or finite horizon [30]. Note that given a WFA A, the membership problem $f_A \in \ell^1_{\mathcal{R}}$ is known to be semi-decidable [31].

Let $\mathbf{H} \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$ be a bi-infinite matrix whose rows and columns are indexed by strings. We say that $\mathbf{H}$ is *Hankel*[3] if for all strings $p, p', s, s' \in \Sigma^\star$ such that $ps = p's'$ we have $\mathbf{H}(p, s) = \mathbf{H}(p', s')$. Given a function $f : \Sigma^\star \to \mathbb{R}$ we can associate with it a Hankel matrix $\mathbf{H}_f \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$ with entries $\mathbf{H}_f(p, s) = f(ps)$. Conversely, given a matrix $\mathbf{H} \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$ with the Hankel property, there exists a unique function $f : \Sigma^\star \to \mathbb{R}$ such that $\mathbf{H}_f = \mathbf{H}$. The following well-known theorem characterizes all Hankel matrices of finite rank.

Theorem 2 ([27]). *For any function $f : \Sigma^\star \to \mathbb{R}$, the Hankel matrix $\mathbf{H}_f$ has finite rank n if and only if f is rational with $\mathrm{rank}(f) = n$. In other words, $\mathrm{rank}(f) = \mathrm{rank}(\mathbf{H}_f)$ for any function $f : \Sigma^\star \to \mathbb{R}$.*

III. A Canonical Form for WFA

In this section we discuss the existence and computation of a canonical form for WFA realizing absolutely convergent rational functions. Our canonical form is strongly related to the singular value decomposition of infinite Hankel matrices. In particular, its existence and uniqueness is a direct consequence of the existence and uniqueness of SVD for Hankel matrices of functions in $\ell^1_{\mathcal{R}}$, as we shall see in the first part of this section. Furthermore, the algorithm given in Section III-B for computing the canonical form can also be interpreted as a procedure for computing the SVD of an infinite Hankel matrix.

A. Existence of the Canonical Form

A matrix $\mathbf{T} \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$ can be interpreted as the expression of a (possibly unbounded) linear operator $T : \ell^2 \to \ell^2$ in terms of the canonical basis $\{\mathbf{e}_x\}_{x \in \Sigma^\star}$. In the case of a Hankel matrix $\mathbf{H}_f$, the associated operator H_f is called a *Hankel operator*, and corresponds to the convolution-like operation $(H_f g)(x) = \sum_y f(xy) g(y)$ (assuming the series converges).

Recall the *operator norm* of $T : \ell^2 \to \ell^2$ is defined as $\|T\|_{\mathrm{op}} = \sup_{\|f\|_2 \leq 1} \|Tf\|_2$. An operator is *bounded* if $\|T\|_{\mathrm{op}}$ is finite. Although not all Hankel operators are bounded, next lemma gives a sufficient condition for H_f to be bounded.

Lemma 3. *If $f \in \ell^1$, then H_f is bounded.*

[3] In real analysis a matrix $\mathbf{M}$ is Hankel if $\mathbf{M}(i, j) = \mathbf{M}(k, l)$ whenever $i + j = k + l$, which implies that $\mathbf{M}$ is symmetric. In our case we have $\mathbf{H}(p, s) = \mathbf{H}(p', s')$ whenever $ps = p's'$, but $\mathbf{H}$ is not symmetric because string concatenation is not commutative whenever $|\Sigma| > 1$.

Proof: Let $h(x) = 1 + |x|$ and note that $f \in \ell^1$ implies $\sup_x |f(x)|(1 + |x|) < \infty$; i.e. $f \cdot h \in \ell^\infty$. Now let $g \in \ell^2$ with $\|g\|_2 = 1$ and for any $x \in \Sigma^\star$ define the function $f_x(y) = f(xy)$. Then we have

$$\begin{aligned}
\|H_f g\|_2^2 &= \sum_x \left(\sum_y f(xy) g(y) \right)^2 = \sum_x \langle f_x, g \rangle^2 \\
&\leq \|g\|_2^2 \sum_x \|f_x\|_2^2 = \sum_x \sum_y f(xy)^2 \\
&= \sum_z (1 + |z|) f(z)^2 = \sum_z |f(z)| |(1 + |z|) f(z)| \\
&\leq \|f\|_1 \|f \cdot h\|_\infty < \infty \ ,
\end{aligned}$$

where we used Cauchy–Schwarz inequality, that the number different ways to split a string z into a prefix and a suffix equals $1 + |z|$, and Hölder's inequality. This concludes the proof. ∎

Theorem 2 and Lemma 3 imply that, for any $f \in \ell^1_{\mathcal{R}}$, the Hankel matrix $\mathbf{H}_f$ represents a bounded finite-rank linear operator H_f on the Hilbert space ℓ^2. Hence, $\mathbf{H}_f$ admits a reduced singular value decomposition $\mathbf{H}_f = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{\Sigma^\star \times n}$ and $\mathbf{D} \in \mathbb{R}^{n \times n}$ with $n = \mathrm{rank}(f)$. The *Hankel singular values* of a rational function $f \in \ell^1_{\mathcal{R}}$ are defined as the singular values of the Hankel matrix $\mathbf{H}_f$. These singular values can be used to define a new set of norms on $\ell^1_{\mathcal{R}}$: the *Schatten–Hankel p-norm* of $f \in \ell^1_{\mathcal{R}}$ is given by $\|f\|_{\mathrm{H},p} = \|\mathbf{H}_f\|_{\mathrm{S},p} = \|(\mathfrak{s}_1, \ldots, \mathfrak{s}_n)\|_p$. It is straightforward to verify that $\|\cdot\|_{\mathrm{H},p}$ satisfies the properties of a norm.

Note an SVD of $\mathbf{H}_f$ yields a rank factorization given by $\mathbf{H}_f = (\mathbf{U}\mathbf{D}^{1/2})(\mathbf{V}\mathbf{D}^{1/2})^\top$. But SVD is not the only way to obtain rank factorizations for Hankel matrices. In fact, if f is rational, then every minimal WFA A realizing f induces a rank factorization of $\mathbf{H}_f$ as follows. Let $\mathbf{P}_A \in \mathbb{R}^{\Sigma^\star \times n}$ be the *forward matrix* of A given by $\mathbf{P}_A(p, :) = \boldsymbol{\alpha}_0^\top \mathbf{A}_p$ for any string $p \in \Sigma^\star$. Similarly, let $\mathbf{S}_A \in \mathbb{R}^{\Sigma^\star \times n}$ be the *backward matrix* of A given by $\mathbf{S}_A(s, :) = (\mathbf{A}_s \boldsymbol{\alpha}_\infty)^\top$ for any string $s \in \Sigma^\star$. Since $H_f(p, s) = f(ps) = \boldsymbol{\alpha}_0^\top \mathbf{A}_p \mathbf{A}_s \boldsymbol{\alpha}_\infty = \mathbf{P}_A(p, :) \mathbf{S}_A^\top(:, s)$, we obtain $\mathbf{H}_f = \mathbf{P}_A \mathbf{S}_A^\top$. This is known as the *forward–backward* (FB) rank factorization of $\mathbf{H}_f$ induced by A [15]. The following result shows that among the infinity of minimal WFA realizing a given rational function $f \in \ell^1_{\mathcal{R}}$, there exists one whose induced FB rank factorization coincides with $\mathbf{H}_f = (\mathbf{U}\mathbf{D}^{1/2})(\mathbf{V}\mathbf{D}^{1/2})^\top$.

Theorem 4. *Let $f \in \ell^1_{\mathcal{R}}$ and suppose $\mathbf{H}_f = (\mathbf{U}\mathbf{D}^{1/2})(\mathbf{V}\mathbf{D}^{1/2})^\top$ is a rank factorization induced by SVD. Then there exists a minimal WFA A for f inducing the same rank factorization. That is, A induces a FB rank factorization of $\mathbf{H}_f$ given by $\mathbf{P}_A = \mathbf{U}\mathbf{D}^{1/2}$ and $\mathbf{S}_A = \mathbf{V}\mathbf{D}^{1/2}$.*

Since we have already established the existence of an SVD for $\mathbf{H}_f$ whenever $f \in \ell^1_{\mathcal{R}}$, the theorem is just a direct application of the following lemma.

Lemma 5. *Suppose $f \in \ell^1_{\mathcal{R}}$ and $\mathbf{H}_f = \mathbf{P}\mathbf{S}^\top$ is a rank factorization. Then there exists a minimal WFA A realizing f which induces this factorization.*

Proof: Let B be any minimal WFA realizing f and denote $n = \operatorname{rank}(f)$. Then we have two rank factorizations $\mathbf{P}\mathbf{S}^\top = \mathbf{P}_B\mathbf{S}_B^\top$ for the Hankel matrix $\mathbf{H}_f$. Therefore, the columns of $\mathbf{P}$ and $\mathbf{P}_B$ both span the same n-dimensional sub-space of $\mathbb{R}^{\Sigma^\star}$, and there exists a change of basis $\mathbf{Q} \in \mathbb{R}^{n\times n}$ such that $\mathbf{P}_B\mathbf{Q} = \mathbf{P}$. This implies we must also have $\mathbf{S}^\top = \mathbf{Q}^{-1}\mathbf{S}_B^\top$. It follows that $A = \mathbf{Q}^{-1}B\mathbf{Q}$ is a minimal WFA for f inducing the desired rank factorization. ■

The results above leads us to our first contribution: the definition of a canonical form for WFA realizing functions in $\ell^1_{\mathcal{R}}$.

Definition 6. *Let $f \in \ell^1_{\mathcal{R}}$. A* singular value automaton *(SVA) for f is a minimal WFA A realizing f such that the FB rank factorization of $\mathbf{H}_f$ induced by A has the form given in Theorem 4.*

Note the SVA provided by Theorem 4 is unique up to the same conditions in which SVD is unique. In particular, it is easy to verify that if the Hankel singular values of $f \in \ell^1_{\mathcal{R}}$ satisfy the strict inequalities $\mathfrak{s}_1 > \cdots > \mathfrak{s}_n$, then the transition weights of the SVA A of f are uniquely defined, and the initial and final weights are uniquely defined up to sign changes.

Then next subsection gives a polynomial-time algorithm for computing the SVA of a function $f \in \ell^1_{\mathcal{R}}$ starting from a WFA realizing f.

B. Computing the Canonical Form

As we have seen above, a bi-infinite Hankel matrix $\mathbf{H}_f$ of rank n can actually be represented with the $n(2 + kn)$ parameters needed to specify the initial, final and transition weights of a minimal WFA A realizing f. Though in principle A contains enough information to reconstruct $\mathbf{H}_f$, a priori it is not clear that A provides an efficient representation for operating on $\mathbf{H}_f$. Luckily, it turns out WFA possess a rich algebraic structure allowing many operations on rational functions and their corresponding Hankel matrices to be performed in "compressed" form by operating directly on WFA representing them [27]. In this section we show it is also possible to compute the SVD of $\mathbf{H}_f$ by operating on a minimal WFA realizing f; that is, we give an algorithm for computing SVA representations.

We start with a simple linear algebra fact showing how to leverage a rank factorization of a given matrix in order to compute its reduced SVD. Let $\mathbf{M} \in \mathbb{R}^{p\times s}$ be a matrix of rank n and suppose $\mathbf{M} = \mathbf{P}\mathbf{S}^\top$ is a rank factorization. Let $\mathbf{G}_p = \mathbf{P}^\top\mathbf{P} \in \mathbb{R}^{n\times n}$ be the Gram matrix of the columns of $\mathbf{P}$. Since $\mathbf{G}_p$ is positive definite, it admits a spectral decomposition $\mathbf{G}_p = \mathbf{V}_p\mathbf{D}_p\mathbf{V}_p^\top$. Similarly, we have $\mathbf{G}_s = \mathbf{S}^\top\mathbf{S} = \mathbf{V}_s\mathbf{D}_s\mathbf{V}_s^\top$. With this notation we have the following.

Lemma 7. *Let $\tilde{\mathbf{M}} = \mathbf{D}_p^{1/2}\mathbf{V}_p^\top\mathbf{V}_s\mathbf{D}_s^{1/2}$ with reduced SVD $\tilde{\mathbf{M}} = \tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}^\top$. If $\mathbf{Q}_p = \mathbf{V}_p\mathbf{D}_p^{-1/2}\tilde{\mathbf{U}}$, $\mathbf{U} = \mathbf{P}\mathbf{Q}_p$, $\mathbf{Q}_s = \mathbf{V}_s\mathbf{D}_s^{-1/2}\tilde{\mathbf{V}}$, $\mathbf{V} = \mathbf{S}\mathbf{Q}_s$, and $\mathbf{D} = \tilde{\mathbf{D}}$, then $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ is a reduced SVD for $\mathbf{M}$.*

Proof: We just need to check the columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal, and $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$:

$$
\begin{aligned}
\mathbf{U}^\top\mathbf{U} &= \mathbf{Q}_p^\top\mathbf{P}^\top\mathbf{P}\mathbf{Q}_p \\
&= \tilde{\mathbf{U}}^\top\mathbf{D}_p^{-1/2}\mathbf{V}_p^\top\mathbf{G}_p\mathbf{V}_p\mathbf{D}_p^{-1/2}\tilde{\mathbf{U}} \\
&= \tilde{\mathbf{U}}^\top\mathbf{D}_p^{-1/2}\mathbf{V}_p^\top\mathbf{V}_p\mathbf{D}_p\mathbf{V}_p^\top\mathbf{V}_p\mathbf{D}_p^{-1/2}\tilde{\mathbf{U}} \\
&= \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}} \\
&= \mathbf{I} \ , \\
\mathbf{V}^\top\mathbf{V} &= \mathbf{Q}_s^\top\mathbf{S}^\top\mathbf{S}\mathbf{Q}_s \\
&= \tilde{\mathbf{V}}^\top\mathbf{D}_s^{-1/2}\mathbf{V}_s^\top\mathbf{G}_s\mathbf{V}_s\mathbf{D}_s^{-1/2}\tilde{\mathbf{V}} \\
&= \tilde{\mathbf{V}}^\top\mathbf{D}_s^{-1/2}\mathbf{V}_s^\top\mathbf{V}_s\mathbf{D}_s\mathbf{V}_s^\top\mathbf{V}_s\mathbf{D}_s^{-1/2}\tilde{\mathbf{V}} \\
&= \tilde{\mathbf{V}}^\top\tilde{\mathbf{V}} \\
&= \mathbf{I} \ , \\
\mathbf{U}\mathbf{D}\mathbf{V}^\top &= \mathbf{P}\mathbf{Q}_p\mathbf{D}\mathbf{Q}_s^\top\mathbf{S}^\top \\
&= \mathbf{P}\mathbf{V}_p\mathbf{D}_p^{-1/2}\tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}^\top\mathbf{D}_s^{-1/2}\mathbf{V}_s^\top\mathbf{S}^\top \\
&= \mathbf{P}\mathbf{V}_p\mathbf{D}_p^{-1/2}\tilde{\mathbf{M}}\mathbf{D}_s^{-1/2}\mathbf{V}_s^\top\mathbf{S}^\top \\
&= \mathbf{P}\mathbf{V}_p\mathbf{D}_p^{-1/2}\mathbf{D}_p^{1/2}\mathbf{V}_p^\top\mathbf{V}_s\mathbf{D}_s^{1/2}\mathbf{D}_s^{-1/2}\mathbf{V}_s^\top\mathbf{S}^\top \\
&= \mathbf{P}\mathbf{S}^\top \\
&= \mathbf{M} \ .
\end{aligned}
$$

■

Note the above result we does not require p and s to be finite. In particular, when $\mathbf{M}$ is an infinite matrix associated with a finite-rank bounded operator, the computation of $\mathbf{Q}_p$ and $\mathbf{Q}_s$ can still be done efficiently as long as $\mathbf{G}_p$ and $\mathbf{G}_s$ are available.

Our goal now is to leverage this result in order to compute the SVD of the bi-infinite Hankel matrix $\mathbf{H}_f$ associated with a rational function $f \in \ell^1_{\mathcal{R}}$. The key step will be to compute the Gram matrices associated with the rank factorization induced by a minimal WFA for f. We start with a lemma showing how to compute the inner product between two rational functions.

Lemma 8. *Let $A = \langle\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\infty, \{\mathbf{A}_\sigma\}\rangle$ and $B = \langle\boldsymbol{\beta}_0, \boldsymbol{\beta}_\infty, \{\mathbf{B}_\sigma\}\rangle$ be minimal WFA realizing functions $f_A, f_B \in \ell^2_{\mathcal{R}}$. Suppose the spectral radius of the matrix $\mathbf{C} = \sum_\sigma \mathbf{A}_\sigma \otimes \mathbf{B}_\sigma$ satisfies $\rho(\mathbf{C}) < 1$. Then the inner product between f_A and f_B can be computed as:*

$$\langle f_A, f_B\rangle = (\boldsymbol{\alpha}_0 \otimes \boldsymbol{\beta}_0)^\top (\mathbf{I} - \mathbf{C})^{-1} (\boldsymbol{\alpha}_\infty \otimes \boldsymbol{\beta}_\infty) \ .$$

Proof: First note $g(x) = f_A(x)f_B(x)$ is in ℓ^1 by Hölder's inequality. Therefore $\langle f_A, f_B\rangle = \sum_x g(x) \leq \sum_x |g(x)| < \infty$. In addition, g is rational [27] and can be computed by the WFA $C = \langle\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_\infty, \{\mathbf{C}_\sigma\}\rangle$ given by

$$
\begin{aligned}
\boldsymbol{\gamma}_0 &= \boldsymbol{\alpha}_0 \otimes \boldsymbol{\beta}_0 \ , \\
\boldsymbol{\gamma}_\infty &= \boldsymbol{\alpha}_\infty \otimes \boldsymbol{\beta}_\infty \ , \\
\mathbf{C}_\sigma &= \mathbf{A}_\sigma \otimes \mathbf{B}_\sigma \ .
\end{aligned}
$$

Now note one can use a simple induction argument to show that for any finite $t \geq 0$ we have

$$s_t = \sum_{x\in\Sigma^t} \boldsymbol{\gamma}_0^\top\mathbf{C}_x\boldsymbol{\gamma}_\infty = \boldsymbol{\gamma}_0^\top\mathbf{C}^t\boldsymbol{\gamma}_\infty \ .$$

Because $g \in \ell^1$, the series $\sum_{t \geq 0} s_t$ is absolutely convergent. Thus we must have $\lim_{k \to \infty} \sum_{t \leq k} s_t = L$ for some finite $L \in \mathbb{R}$. Since $\rho(\mathbf{C}) < 1$ implies the identity $\sum_{t \geq 0} \mathbf{C}^t = (\mathbf{I} - \mathbf{C})^{-1}$, we must necessarily have $L = \boldsymbol{\gamma}_0^\top (\mathbf{I} - \mathbf{C})^{-1} \boldsymbol{\gamma}_\infty$. ∎

Note the assumption $\rho(\mathbf{C}) < 1$ is an essential part of our calculations. We shall make similar assumptions in the remaining of this section. See Section V-B for a discussion on this assumption and how to remove it.

The following result shows how to efficiently compute the Gram matrices associated with the rank factorization induced by a minimal WFA for a function $f \in \ell^1_{\mathcal{R}}$.

Lemma 9. *Let $f \in \ell^1_{\mathcal{R}}$ with $\operatorname{rank}(f) = n$, and $A = \langle \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\infty, \{\mathbf{A}_\sigma\} \rangle$ be a minimal WFA for f inducing the FB rank factorization $\mathbf{H}_f = \mathbf{P}\mathbf{S}^\top$. Let $\mathbf{A}^\otimes = \sum_\sigma \mathbf{A}_\sigma^{\otimes 2}$. If $\rho(\mathbf{A}^\otimes) < 1$, then the Gram matrices $\mathbf{G}_p, \mathbf{G}_s \in \mathbb{R}^{n \times n}$ associated with the factorization induced by A satisfy $\operatorname{vec}(\mathbf{G}_p)^\top = (\boldsymbol{\alpha}_0^{\otimes 2})^\top (\mathbf{I} - \mathbf{A}^\otimes)^{-1}$ and $\operatorname{vec}(\mathbf{G}_s) = (\mathbf{I} - \mathbf{A}^\otimes)^{-1} \boldsymbol{\alpha}_\infty^{\otimes 2}$.*

Proof: For $i \in [n]$ let $\mathbf{p}_i = \mathbf{P}(:, i) \in \mathbb{R}^{\Sigma^\star}$ be the ith column of $\mathbf{P}$. The key observation is that the function $p_i : \Sigma^\star \to \mathbb{R}$ defined by $p_i(x) = \mathbf{p}_i(x)$ is in $\ell^2_{\mathcal{R}}$. To show rationality one just needs to check p_i is the function realized by the WFA $A_{p,i} = \langle \boldsymbol{\alpha}_0, \mathbf{e}_i, \{\mathbf{A}_\sigma\} \rangle$ by construction of the induced rank factorization. The fact that $\|p_i\|_2$ is finite follows from Theorem 4 by noting that $\mathbf{p}_i$ is a linear combination of left singular vectors of $\mathbf{H}_f$, which belong to ℓ^2 by definition. Thus, $\mathbf{G}_p(i,j) = \mathbf{p}_i^\top \mathbf{p}_j$ is well-defined and corresponds to the inner product $\langle p_i, p_j \rangle$ which, by Lemma 8 can be computed as

$$(\boldsymbol{\alpha}_0^{\otimes 2})^\top \left(\mathbf{I} - \mathbf{A}^\otimes \right)^{-1} (\mathbf{e}_i \otimes \mathbf{e}_j) \ .$$

Since $\mathbf{e}_i \otimes \mathbf{e}_j = \mathbf{e}_{(i-1)n+j}$, we obtain the desired expression for $\operatorname{vec}(\mathbf{G}_p)$. The expression for $\operatorname{vec}(\mathbf{G}_s)$ follows from an identical argument using automata $\mathbf{A}_{s,i} = \langle \mathbf{e}_i, \boldsymbol{\alpha}_\infty, \{\mathbf{A}_\sigma\} \rangle$. ∎

Combining the results above we now show it is possible to compute an SVA for $f \in \ell^1_{\mathcal{R}}$ starting from a minimal WFA realizing f. The procedure is called `ComputeSVA` and its description is given in Algorithm 1.

Algorithm 1: `ComputeSVA`

Input: A minimal WFA A with n states for $f \in \ell^1_{\mathcal{R}}$
Output: An SVA A' for f
Compute $\mathbf{G}_p$ and $\mathbf{G}_s$ `/* cf. Lemma 9 */`
Compute $\mathbf{Q}_p$, $\mathbf{Q}_s$, and $\mathbf{D}$ `/* cf. Lemma 7 */`
Let $A' = \mathbf{D}^{1/2} \mathbf{Q}_s^\top A \mathbf{Q}_p \mathbf{D}^{1/2}$ `/* cf. Eq. (1) */`
return A'

Theorem 10. *Let $A = \langle \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\infty, \{\mathbf{A}_\sigma\} \rangle$ be a minimal WFA for f such that $\mathbf{A}^\otimes = \sum_\sigma \mathbf{A}_\sigma^{\otimes 2}$ satisfies $\rho(\mathbf{A}^\otimes) < 1$. Then the WFA A' computed by `ComputeSVA`(A) is an SVA for f.*

Proof: Let $\mathbf{Q} = \mathbf{Q}_p \mathbf{D}^{1/2}$. Our first observation is that $\mathbf{Q}^{-1} = \mathbf{D}^{1/2} \mathbf{Q}_s^\top$ and thus A and A' are minimal WFA for f. Indeed, we already showed in the proof of Lemma 7 that

$$\mathbf{Q}_p \mathbf{D}^{1/2} \mathbf{D}^{1/2} \mathbf{Q}_s^\top = \mathbf{Q}_p \mathbf{D} \mathbf{Q}_s^\top = \mathbf{I} \ .$$

In addition, it is immediate to check that if A induces the rank factorization $\mathbf{H}_f = \mathbf{P}\mathbf{S}^\top$, then A' induces the rank factorization $\mathbf{H}_f = (\mathbf{P}\mathbf{Q}_p\mathbf{D}^{1/2})(\mathbf{D}^{1/2}\mathbf{Q}_s^\top \mathbf{S}^\top)$, which by Lemma 7 satisfies $\mathbf{P}\mathbf{Q}_p\mathbf{D}^{1/2} = \mathbf{U}\mathbf{D}^{1/2}$ and $\mathbf{D}^{1/2}\mathbf{Q}_s^\top\mathbf{S}^\top = \mathbf{D}^{1/2}\mathbf{V}^\top$. ∎

We conclude this section by mentioning that it is possible to modify `ComputeSVA` to take as input a *non-minimal* WFA A realizing a function $f \in \ell^1_{\mathcal{R}}$ under the same assumption on the spectral radius of the matrix $\mathbf{A}^\otimes$ as we have here. We shall present the details of this modification somewhere else. Nonetheless, we note that if given a non-minimal WFA A, one always has the option to minimize A (e.g. using the WFA minimization algorithm in [27]) before attempting the SVA computation.

C. Computational Complexity

To bound the running time of `ComputeSVA`(A) we recall the following facts from numerical linear algebra (see e.g. [32]):

- The SVD of $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ $(d_1 \geq d_2)$ can be computed in time $O(d_1 d_2^2)$.
- The spectral decomposition of a symmetric matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ can be computed in time $O(d^3)$.
- The inverse of an invertible matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ can be computed in time $O(d^3)$.

Now note that according to Lemma 9, computing the Gram matrices requires $O(kn^4)$ operations to obtain $\mathbf{I} - \mathbf{A}^\otimes$, plus the inversion of this $n^2 \times n^2$ matrix, which can be done in time $O(n^6)$. From Lemma 7 we see that once the $n \times n$ Gram matrices $\mathbf{G}_p$ and $\mathbf{G}_s$ are given, then computing the singular values $\mathbf{D}$ and the change of basis matrices $\mathbf{Q}_p$ and $\mathbf{Q}_s$ can be done in time $O(n^3)$. Finally, the cost of conjugating the WFA A into A' takes time $O(kn^3)$, where $k = |\Sigma|$ and $n = \dim(A)$. Hence, the overall running time of `ComputeSVA`(A) is $O(n^6 + kn^4)$. Of course, this is a rough estimate which does not take into account improvements that might be possible in practice, especially in those cases where the transition matrices of A are sparse – in such case the complexity of most operations could be bounded in terms of the number of non-zeros.

D. A Fundamental Property of SVA

This section gives a fundamental property of SVA. Namely, a bounds on the transition coefficients of an SVA in terms of the Hankel singular values of the function it realizes. In the next section we shall exploit this fundamental property of SVA for designing and analysing approximate minimization algorithms for WFA.

Lemma 11. *Let $A = \langle \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\infty, \{\mathbf{A}_\sigma\} \rangle$ be an SVA with n states realizing a function $f \in \ell^1_{\mathcal{R}}$ with Hankel singular values $\mathfrak{s}_1 \geq \cdots \geq \mathfrak{s}_n$. Then the following are satisfied:*

1) *For all $j \in [n]$, $\sum_i \mathfrak{s}_i \sum_\sigma \mathbf{A}_\sigma(i,j)^2 = \mathfrak{s}_j - \boldsymbol{\alpha}_0(j)^2$,*

2) *For all* $i \in [n]$, $\sum_j \mathfrak{s}_j \sum_\sigma \mathbf{A}_\sigma(i,j)^2 = \mathfrak{s}_i - \boldsymbol{\alpha}_\infty(i)^2$.

Proof: Recall that A induces the rank factorization $\mathbf{H}_f = \mathbf{P}\mathbf{S}^\top = (\mathbf{U}\mathbf{D}^{1/2})(\mathbf{D}^{1/2}\mathbf{V}^\top)$ corresponding to the SVD of $\mathbf{H}_f$. Let $\mathbf{p}_j$ be the jth column of $\mathbf{P} = [\mathbf{p}_1 \cdots \mathbf{p}_n]$ and note we have $\|\mathbf{p}_j\|_2^2 = \mathfrak{s}_j$. By appropriately decomposing the sum in $\|\mathbf{p}_j\|_2^2$ we get the following[4]:

$$\mathfrak{s}_j = \mathbf{p}_j(\lambda)^2 + \sum_{\sigma \in \Sigma} \sum_{x \in \Sigma^\star} \mathbf{p}_j(x\sigma)^2 \ . \tag{2}$$

Let us write $\mathbf{p}_j^\sigma$ for the element of $\ell^2(\Sigma)$ given by $\mathbf{p}_j^\sigma(x) = \mathbf{p}_j(x\sigma)$. Note that by construction we have $\mathbf{p}_j^\sigma = \mathbf{P} \cdot \mathbf{A}_\sigma(:,j) = \sum_{i \in [n]} \mathbf{p}_i \mathbf{A}_\sigma(i,j)$. Since A is an SVA, the columns of $\mathbf{P}$ are orthogonal and therefore we have

$$\begin{aligned}
\|\mathbf{p}_j^\sigma\|_2^2 &= \left\langle \sum_i \mathbf{p}_i \mathbf{A}_\sigma(i,j), \sum_{i'} \mathbf{p}_{i'} \mathbf{A}_\sigma(i',j) \right\rangle \\
&= \sum_{i,i'} \mathbf{A}_\sigma(i,j)\mathbf{A}_\sigma(i',j)\langle \mathbf{p}_i, \mathbf{p}_{i'} \rangle \\
&= \sum_i \mathfrak{s}_i \mathbf{A}_\sigma(i,j)^2 \ .
\end{aligned}$$

Plugging this into (2) and noting that $\mathbf{p}_j(\lambda) = \boldsymbol{\alpha}_0(j)$, we obtain the first claim. The second claim follows from applying the same argument to the columns of $\mathbf{S}$. ■

To see the importance of this lemma for approximate minimization, let us consider the following simple consequence which can be derived by combining the bounds for $\mathbf{A}_\sigma(i,j)$ obtained from considering it belongs to the ith row and the jth column of $\mathbf{A}_\sigma$:

$$|\mathbf{A}_\sigma(i,j)| \leq \min\left\{ \sqrt{\frac{\mathfrak{s}_i}{\mathfrak{s}_j}}, \sqrt{\frac{\mathfrak{s}_j}{\mathfrak{s}_i}} \right\} = \sqrt{\frac{\min\{\mathfrak{s}_i, \mathfrak{s}_j\}}{\max\{\mathfrak{s}_i, \mathfrak{s}_j\}}} \ .$$

This bound is telling us that in an SVA, transition weights further away from the diagonals of the $\mathbf{A}_\sigma$ are going to be small whenever there is a wide spread between the largest and smallest singular values; for example, $|\mathbf{A}_\sigma(1,n)| \leq \sqrt{\mathfrak{s}_n/\mathfrak{s}_1}$. Intuitively, this means that in an SVA the last states are very weakly connected to the first states, and therefore removing these connections should not affect the output of the WFA too much. Theorem 12 below exploits this intuition and turns it into a definite quantitative statement.

IV. Approximate Minimization of WFA

In this section we describe and analyse an approximate minimization algorithm for WFA. The algorithm takes as input a minimal WFA A with n states and a target number of states $\hat{n}$, and outputs a new WFA $\hat{A}$ with $\hat{n}$ states approximating the original WFA A. To obtain $\hat{A}$ we first compute the SVA A' associated to A, and then remove the $n - \hat{n}$ states associated with the smallest singular values of $\mathbf{H}_{f_A}$. We call this algorithm `SVATruncation` (see Algorithm 2 for details). Since the algorithm only involves a call to `ComputeSVA` and a simple algebraic manipulation of the resulting WFA, the running time of `SVATruncation` is dominated by the complexity of `ComputeSVA`, and hence is polynomial in $|\Sigma|$, $\dim(A)$ and $\hat{n}$.

[4]Here we are implicitly using the fact that $\sum_x \mathbf{p}_j(x)^2$ is absolutely (and therefore unconditionally) convergent, which implies that any rearrangement of its terms will converge to the same value.

Algorithm 2: `SVATruncation`

Input: A minimal WFA A with n states, a target number of states $\hat{n} < n$
Output: A WFA $\hat{A}$ with $\hat{n}$ states

1 Let $A' \leftarrow$ `ComputeSVA`(A)
2 Let $\mathbf{\Pi} = [\mathbf{I}_{\hat{n}} \ \mathbf{0}] \in \mathbb{R}^{\hat{n} \times n}$
3 Let $\hat{\mathbf{A}}_\sigma = \mathbf{\Pi}\mathbf{A}'_\sigma\mathbf{\Pi}^\top$ for all $\sigma \in \Sigma$
4 Let $\hat{\boldsymbol{\alpha}}_0 = \mathbf{\Pi}\boldsymbol{\alpha}'_0$
5 Let $\hat{\boldsymbol{\alpha}}_\infty = \mathbf{\Pi}\boldsymbol{\alpha}'_\infty$
6 Let $\hat{A} = \langle \hat{\boldsymbol{\alpha}}_0, \hat{\boldsymbol{\alpha}}_\infty, \{\hat{\mathbf{A}}_\sigma\} \rangle$
7 **return** $\hat{A}$

Roughly speaking, the rationale behind `SVATruncation` is that given an SVA, the states corresponding to the smallest singular values are the ones with less influence on the Hankel matrix, and therefore should also be the ones with less influence on the associated rational function. However, the details are more tricky than this simple intuition. The reason being that a low rank approximation to $\mathbf{H}_f$ obtained by truncating its SVD is not in general a Hankel matrix, and therefore does not correspond to any rational function. In particular, the Hankel matrix of the function $\hat{f}$ computed by $\hat{A}$ is not obtained by truncating the SVD of $\mathbf{H}_f$. This makes our analysis non-trivial.

The main result of this section is the following theorem, which bounds the ℓ^2-distance between the rational function f realized by the original WFA A, and the rational function $\hat{f}$ realized by the output WFA $\hat{A}$. The principal attractive of our bound is that it only depends on intrinsic quantities associated with the function f; that is, the final error bound is independent of which WFA A is given as input. To comply with the assumptions made in the previous section, we shall assume like in previous section that the input WFA A satisfies $\rho(\mathbf{A}^\otimes) < 1$. The same precepts about this assumption discussed in Section V-B apply here.

Theorem 12. *Let* $f \in \ell^1_{\mathcal{R}}$ *with* $\operatorname{rank}(f) = n$ *and fix* $0 < \hat{n} < n$. *If* A *is a minimal WFA realizing* f *and such that* $\rho(\mathbf{A}^\otimes) < 1$, *then the WFA* $\hat{A} =$ `SVATruncation`$(A, \hat{n})$ *realizes a function* $\hat{f}$ *satisfying*

$$\|f - \hat{f}\|_2^2 \leq C_f \sqrt{\mathfrak{s}_{\hat{n}+1} + \cdots + \mathfrak{s}_n} \ , \tag{3}$$

where C_f *is a positive constant depending only on* f.

A few remarks about this result are in order. The first is to observe that because $\mathfrak{s}_1 \geq \cdots \geq \mathfrak{s}_n$, the error decreases when $\hat{n}$ increases, which is the desired behavior: the more states $\hat{A}$ has, the closer it is to A. The second is that (3) does not depend on which representation A of f is given as input to `SVATruncation`. This is a consequence of first obtaining the corresponding SVA A' before truncating. Obviously, one could obtain another approximate minimization by truncating

A directly. However, in that case the final error would depend on the initial A and in general it does not seem possible to use this approach for providing *representation independent* bounds on the quality of approximation.

A. Proof Sketch for Theorem 12

A full detailed proof of Theorem 12 can be found in the technical report [33]. Here we will only give an outline of how the main ideas behind the proof, and sketch the key technical lemmas.

The first step in the proof is to combine A' and $\hat{A}$ into a single WFA B computing $f_B = (f - \hat{f})^2$, and then decompose the error as

$$\|f - \hat{f}\|_2^2 = \sum_{t \geq 0} \left(\sum_{x \in \Sigma^t} f_B(x) \right) .$$

One can then proceed to bound $f_B(\Sigma^t)$ for all $t \geq 0$ in terms of the weights of A'; this involves lengthy algebraic manipulations with many intermediate steps exploiting a variety of properties of matrix norms and Kronecker products. The reader can consult [34] for a comprehensive account of these properties. The last and key step is to exploit the internal structure of the SVA canonical form in order to turn these bounds into representation independent quantities. This part of the analysis is based on the powerful Lemma 11.

Step 1: We start by noting that, without loss of generality, we can assume the automaton A given as input to `SVATruncation` is in SVA form (in which case $A' = A$).

Now we introduce some notation by splitting the weights conforming A into a block corresponding to states 1 to $\hat{n}$, and another block containing states $\hat{n}+1$ to n. With this, we write the following:

$$\boldsymbol{\alpha}_0 = \left[\boldsymbol{\alpha}_0^{(1)} \ \boldsymbol{\alpha}_0^{(2)}\right] \ ,$$
$$\boldsymbol{\alpha}_\infty = \begin{bmatrix} \boldsymbol{\alpha}_\infty^{(1)} \\ \boldsymbol{\alpha}_\infty^{(2)} \end{bmatrix} \ ,$$
$$\mathbf{A}_\sigma = \begin{bmatrix} \mathbf{A}_\sigma^{(11)} & \mathbf{A}_\sigma^{(12)} \\ \mathbf{A}_\sigma^{(21)} & \mathbf{A}_\sigma^{(22)} \end{bmatrix} \ .$$

It is immediate to check that $\hat{A} = \texttt{SVATruncation}(A, \hat{n})$ is given by $\hat{\boldsymbol{\alpha}}_0 = \boldsymbol{\alpha}_0^{(1)}$, $\hat{\boldsymbol{\alpha}}_\infty = \boldsymbol{\alpha}_\infty^{(1)}$, and $\hat{\mathbf{A}}_\sigma = \mathbf{A}_\sigma^{(11)}$. For simplicity of notation, we assume here and throughout the rest of the proof that initial weights of WFA are given as *row* vectors.

Although $\hat{A}$ has $\hat{n}$ states, it is convenient to define a WFA with n states computing the same function as $\hat{A}$. We call this WFA $\tilde{A}$, and its construction is explained in the following claim, whose proof is a simple calculation exploiting the structure of the matrices $\tilde{\mathbf{A}}_\sigma$.

Claim 1. The WFA $\tilde{A} = \langle \tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\alpha}}_\infty, \{\tilde{\mathbf{A}}_\sigma\}\rangle$ with n states given below satisfies $\tilde{f} = f_{\tilde{A}} = \hat{f}$:

$$\tilde{\boldsymbol{\alpha}}_0 = \boldsymbol{\alpha}_0 \ ,$$
$$\tilde{\boldsymbol{\alpha}}_\infty = \begin{bmatrix} \boldsymbol{\alpha}_\infty^{(1)} \\ \mathbf{0} \end{bmatrix} \ ,$$
$$\tilde{\mathbf{A}}_\sigma = \begin{bmatrix} \mathbf{A}_\sigma^{(11)} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_\sigma^{(22)} \end{bmatrix} = \operatorname{diag}(\mathbf{A}_\sigma^{(11)}, \mathbf{A}_\sigma^{(22)}) \ .$$

By combining A and $\tilde{A}$ we can obtain a WFA computing squares of differences between f and $\hat{f}$. The construction is given in the following claim, which follows from the same argument used in the proof of Lemma 8.

Claim 2. Let $B = \langle \boldsymbol{\beta}_0^{\otimes 2}, \boldsymbol{\beta}_\infty^{\otimes 2}, \{\mathbf{B}_\sigma^{\otimes 2}\}\rangle$ be the WFA with $4n^2$ states where

$$\boldsymbol{\beta}_0 = [\boldsymbol{\alpha}_0 \ -\tilde{\boldsymbol{\alpha}}_0] \ ,$$
$$\boldsymbol{\beta}_\infty = \begin{bmatrix} \boldsymbol{\alpha}_\infty \\ \tilde{\boldsymbol{\alpha}}_\infty \end{bmatrix} \ ,$$
$$\mathbf{B}_\sigma = \begin{bmatrix} \mathbf{A}_\sigma & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{A}}_\sigma \end{bmatrix} = \operatorname{diag}(\mathbf{A}_\sigma, \tilde{\mathbf{A}}_\sigma) \ .$$

Then $f_B = (f - \tilde{f})^2$.

From the weights of automaton B we define the following vectors and matrices:

$$\boldsymbol{\gamma}_0 = \boldsymbol{\alpha}_0 \otimes [\boldsymbol{\alpha}_0 \ -\tilde{\boldsymbol{\alpha}}_0] = \boldsymbol{\alpha}_0 \otimes \boldsymbol{\beta}_0 \ ,$$
$$\boldsymbol{\gamma}_\infty = \boldsymbol{\alpha}_\infty \otimes \begin{bmatrix} \boldsymbol{\alpha}_\infty \\ \tilde{\boldsymbol{\alpha}}_\infty \end{bmatrix} = \boldsymbol{\alpha}_\infty \otimes \boldsymbol{\beta}_\infty \ ,$$
$$\tilde{\boldsymbol{\gamma}}_\infty = \tilde{\boldsymbol{\alpha}}_\infty \otimes \begin{bmatrix} \boldsymbol{\alpha}_\infty \\ \tilde{\boldsymbol{\alpha}}_\infty \end{bmatrix} = \tilde{\boldsymbol{\alpha}}_\infty \otimes \boldsymbol{\beta}_\infty \ ,$$
$$\mathbf{C}_\sigma = \mathbf{A}_\sigma \otimes \begin{bmatrix} \mathbf{A}_\sigma & \mathbf{0} \ , \\ \mathbf{0} & \tilde{\mathbf{A}}_\sigma \end{bmatrix} = \mathbf{A}_\sigma \otimes \mathbf{B}_\sigma \ ,$$
$$\tilde{\mathbf{C}}_\sigma = \tilde{\mathbf{A}}_\sigma \otimes \begin{bmatrix} \mathbf{A}_\sigma & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{A}}_\sigma \end{bmatrix} = \tilde{\mathbf{A}}_\sigma \otimes \mathbf{B}_\sigma \ .$$

We will also write $\mathbf{C} = \sum_\sigma \mathbf{C}_\sigma$ and $\tilde{\mathbf{C}} = \sum_\sigma \tilde{\mathbf{C}}_\sigma$.

Step 2: The notations defined about let us state the following claim, which will be the starting point of our bounds. The result follows from the same calculations used in the proof of Lemma 8 and the observation that $\boldsymbol{\beta}_0^{\otimes 2} = [\boldsymbol{\gamma}_0 \ -\boldsymbol{\gamma}_0]$.

Claim 3. For any $t \geq 0$ we have

$$\Delta_t = \sum_{x \in \Sigma^t} (f(x) - \hat{f}(x))^2 = \boldsymbol{\gamma}_0 \left(\mathbf{C}^t \boldsymbol{\gamma}_\infty - \tilde{\mathbf{C}}^t \tilde{\boldsymbol{\gamma}}_\infty \right) \ .$$

Note that the error that we need to bound in order to proof Theorem 12 can be written as $\|f - \hat{f}\|_2^2 = \sum_{t \geq 0} \Delta_t$. Our strategy is to obtain a separate bound for each Δ_t and then sum them all together. We start by bounding $|\Delta_t|$ in terms of the norms of the matrices and vectors defined above.

Lemma 13. *For any $t \geq 0$ the following bound holds:*

$$|\Delta_t| \leq \|\boldsymbol{\gamma}_0\| \|\tilde{\mathbf{C}}\|^t \|\boldsymbol{\gamma}_\infty - \tilde{\boldsymbol{\gamma}}_\infty\| + t \|\boldsymbol{\gamma}_0\| \|\boldsymbol{\gamma}_\infty\| \max\{\|\mathbf{C}\|, \|\tilde{\mathbf{C}}\|\}^{t-1} \|\mathbf{C} - \tilde{\mathbf{C}}\| \ .$$

Proof (sketch): We proceed by induction on t. The base case is easy. For the inductive step, first use the triangle inequality to show that

$$\begin{aligned}|\Delta_{t+1}| \leq &\|\boldsymbol{\gamma}_0\|\|\tilde{\mathbf{C}}\|^{t+1}\|\boldsymbol{\gamma}_\infty - \tilde{\boldsymbol{\gamma}}_\infty\| \\ &+ \|\boldsymbol{\gamma}_0\|\|\mathbf{C}\|^t\|\mathbf{C} - \tilde{\mathbf{C}}\|\|\boldsymbol{\gamma}_\infty\| \\ &+ |\Delta'_t| \ ,\end{aligned}$$

where $\Delta'_t = \boldsymbol{\gamma}_0 \left(\mathbf{C}^t\tilde{\mathbf{C}}\boldsymbol{\gamma}_\infty - \tilde{\mathbf{C}}^t\tilde{\mathbf{C}}\boldsymbol{\gamma}_\infty\right)$. Next we note that our inductive hypothesis can be seen to imply the bound

$$|\Delta'_t| \leq t\|\boldsymbol{\gamma}_0\|\|\boldsymbol{\gamma}_\infty\| \max\{\|\mathbf{C}\|, \|\tilde{\mathbf{C}}\|\}^t\|\mathbf{C} - \tilde{\mathbf{C}}\| \ .$$

A simple algebraic simplification combining the two bounds yields the desired result. ■

Now we proceed to derive individual bounds for all the terms that appear in previous lemma. We start with the following claim which gives two bounds that are clear by definition.

Claim 4. We have $\|\boldsymbol{\gamma}_0\| = \sqrt{2}\|\boldsymbol{\alpha}_0\|^2$ and $\|\boldsymbol{\gamma}_\infty\| \leq \sqrt{2}\|\boldsymbol{\alpha}_\infty\|^2$.

The next step is to bound the term involving the norms $\|\mathbf{C}\|$ and $\|\tilde{\mathbf{C}}\|$. This leads to the definition of a representation independent parameter we call ρ_f.

Lemma 14. *Let $\rho_f = \|\sum_\sigma \mathbf{A}_\sigma \otimes \mathbf{A}_\sigma\|$. Then ρ_f is a positive constant depending only on f which satisfies:*

$$\|\tilde{\mathbf{C}}\| \leq \|\mathbf{C}\| = \rho_f \ .$$

Proof: We start by noting that $\|\mathbf{C}\| = \max\{\|\sum_\sigma \mathbf{A}_\sigma \otimes \mathbf{A}_\sigma\|, \|\sum_\sigma \mathbf{A}_\sigma \otimes \tilde{\mathbf{A}}_\sigma\|\}$. Then we use $\|\sum_\sigma \mathbf{A}_\sigma \otimes \tilde{\mathbf{A}}_\sigma\| = \|\sum_\sigma \tilde{\mathbf{A}}_\sigma \otimes \mathbf{A}_\sigma\| = \max\{\|\sum_\sigma \mathbf{A}_\sigma^{(11)} \otimes \mathbf{A}_\sigma\|, \|\sum_\sigma \mathbf{A}_\sigma^{(22)} \otimes \mathbf{A}_\sigma\|\}$ to show that $\|\mathbf{C}\| = \|\sum_\sigma \mathbf{A}_\sigma \otimes \mathbf{A}_\sigma\|$, since the rest of terms in the maximum correspond to norms of sub-matrices of $\sum_\sigma \mathbf{A}_\sigma \otimes \mathbf{A}_\sigma$. Now a similar argument can be used to show that

$$\begin{aligned}\|\tilde{\mathbf{C}}\| = \max\{&\|\sum_\sigma \mathbf{A}_\sigma^{(11)} \otimes \mathbf{A}_\sigma\|, \|\sum_\sigma \mathbf{A}_\sigma^{(11)} \otimes \tilde{\mathbf{A}}_\sigma\|, \\ &\|\sum_\sigma \mathbf{A}_\sigma^{(22)} \otimes \mathbf{A}_\sigma\|, \|\sum_\sigma \mathbf{A}_\sigma^{(22)} \otimes \tilde{\mathbf{A}}_\sigma\|\} \\ &\leq \|\sum_\sigma \mathbf{A}_\sigma \otimes \mathbf{A}_\sigma\| \ .\end{aligned}$$

Note ρ_f is representation independent because it only depends on the transition weights of the SVA form of f, which is unique. ■

In the remaining of the proof we will assume the following holds.

Assumption 1. The SVA A is such that $\rho_f = \|\sum_\sigma \mathbf{A}_\sigma^{\otimes 2}\| < 1$.

This assumption is not essential, and is only introduced to simplify the calculations involved in the proof. See the remarks at the end of this appendix for a discussion on how to remove the assumption.

In order to bound $\|\boldsymbol{\gamma}_\infty - \tilde{\boldsymbol{\gamma}}_\infty\|$ and $\|\mathbf{C} - \tilde{\mathbf{C}}\|$ we make an extensive use of the properties of SVA given in Lemma 11. This allows us to obtain bounds that only depend on the Hankel singular values of f, which are intrinsic representation-independent quantities associated with f.

Lemma 15.

$$\|\boldsymbol{\gamma}_\infty - \tilde{\boldsymbol{\gamma}}_\infty\| \leq \sqrt{2}\|\boldsymbol{\alpha}_\infty\|\sqrt{\mathfrak{s}_{\hat{n}+1} + \cdots + \mathfrak{s}_n} \ .$$

Proof: We start by unwinding the definitions of $\boldsymbol{\gamma}_\infty$ and $\tilde{\boldsymbol{\gamma}}_\infty$ to obtain the bound:

$$\begin{aligned}\|\boldsymbol{\gamma}_\infty - \tilde{\boldsymbol{\gamma}}_\infty\| &= \|(\boldsymbol{\alpha}_\infty - \tilde{\boldsymbol{\alpha}}_\infty) \otimes \boldsymbol{\beta}_\infty\| \\ &= \|\boldsymbol{\alpha}_\infty - \tilde{\boldsymbol{\alpha}}_\infty\|\|\boldsymbol{\beta}_\infty\| \\ &= \|\boldsymbol{\alpha}_\infty^{(2)}\|\sqrt{\|\boldsymbol{\alpha}_\infty\|^2 + \|\tilde{\boldsymbol{\alpha}}_\infty\|^2} \\ &\leq \sqrt{2}\|\boldsymbol{\alpha}_\infty\|\|\boldsymbol{\alpha}_\infty^{(2)}\| \ ,\end{aligned}$$

where we used that $\|\tilde{\boldsymbol{\alpha}}_\infty\| \leq \|\boldsymbol{\alpha}_\infty\|$. Now note that Lemma 11 yields the crude estimate $\boldsymbol{\alpha}_\infty(i)^2 \leq \mathfrak{s}_i$ for all $i \in [n]$. We use this last observation to obtain the following bound and complete the proof:

$$\|\boldsymbol{\alpha}_\infty^{(2)}\| = \sqrt{\sum_{i=\hat{n}+1}^{n} \boldsymbol{\alpha}_\infty(i)^2} \leq \sqrt{\mathfrak{s}_{\hat{n}+1} + \cdots + \mathfrak{s}_n} \ .$$

■

Lemma 16.

$$\|\mathbf{C} - \tilde{\mathbf{C}}\| \leq \sqrt{\sum_\sigma \|\mathbf{A}_\sigma\|^2}\sqrt{\frac{\mathfrak{s}_{\hat{n}+1} + \cdots + \mathfrak{s}_n}{\mathfrak{s}_{\hat{n}}}} \ .$$

Proof (sketch): Let $\boldsymbol{\Gamma} = \mathbf{C} - \tilde{\mathbf{C}} = \sum_\sigma (\mathbf{A}_\sigma - \tilde{\mathbf{A}}_\sigma) \otimes \mathbf{B}_\sigma$. By expanding $\mathbf{A}_\sigma - \tilde{\mathbf{A}}_\sigma$ in this expression one can see that

$$\boldsymbol{\Gamma} = \begin{bmatrix} \mathbf{0} & \boldsymbol{\Gamma}^{(12)} \\ \boldsymbol{\Gamma}^{(21)} & \mathbf{0} \end{bmatrix} \ ,$$

where $\boldsymbol{\Gamma}^{(ij)} = \sum_\sigma \mathbf{A}_\sigma^{(ij)} \otimes \mathbf{B}_\sigma$ for $ij \in \{12, 21\}$. Since both $\boldsymbol{\Gamma}^{(12)}$ and $\boldsymbol{\Gamma}^{(21)}$ are unitarily equivalent to block-diagonals matrices, $\boldsymbol{\Gamma}$ is also unitarily equivalent to a block-diagonal matrix. Thus, after a calculation we get

$$\|\boldsymbol{\Gamma}\| = \max\left\{\|\sum_\sigma \mathbf{A}_\sigma^{(12)} \otimes \mathbf{A}_\sigma\|, \|\sum_\sigma \mathbf{A}_\sigma^{(21)} \otimes \mathbf{A}_\sigma\|\right\} \ . \tag{4}$$

Finally one bounds the two terms in the maximum above as follows. First use the triangle and Cauchy–Schwarz inequalities to show that:

$$\|\sum_\sigma \mathbf{A}_\sigma^{(12)} \otimes \mathbf{A}_\sigma\| \leq \sqrt{\sum_\sigma \|\mathbf{A}_\sigma^{(12)}\|^2}\sqrt{\sum_\sigma \|\mathbf{A}_\sigma\|^2} \ .$$

Then use Lemma 11 to obtain the bound

$$\sum_\sigma \|\mathbf{A}_\sigma^{(12)}\|^2 \leq \sum_\sigma \|\mathbf{A}_\sigma^{(12)}\|_F^2 \leq \frac{\mathfrak{s}_{\hat{n}+1} + \cdots + \mathfrak{s}_n}{\mathfrak{s}_{\hat{n}}} \ .$$

A similar bound on the second term in (4) follows mutatis mutandis from a mimetic argument. ■

Step 3: Now we can put all the above together in order to obtain the bound in Theorem 12. We start with the following bound for $|\Delta_t|$ for some fixed t, which follows from simple algebraic manipulations:

$$|\Delta_t| \leq \left(C_1' \rho_f^t + C_2' \frac{t \rho_f^{t-1}}{\sqrt{\mathfrak{s}_{\hat{n}}}} \right) \sqrt{\mathfrak{s}_{\hat{n}+1} + \cdots + \mathfrak{s}_n} \enspace ,$$

where $C_1' = 2\|\boldsymbol{\alpha}_0\|^2 \|\boldsymbol{\alpha}_\infty\|$ and $C_2' = 2\|\boldsymbol{\alpha}_0\|^2 \|\boldsymbol{\alpha}_\infty\|^2 \left(\sum_\sigma \|\mathbf{A}_\sigma\|^2\right)^{1/2}$ are constants depending only on f.

Finally, using that $\|f - \hat{f}\|_2^2 = \sum_{t \geq 0} \Delta_t$, we can sum all the bounds on Δ_t and obtain

$$\begin{aligned} \|f - \hat{f}\|_2^2 &\leq \left(C_1 + \frac{C_2}{\sqrt{\mathfrak{s}_{\hat{n}}}} \right) \sqrt{\mathfrak{s}_{\hat{n}+1} + \cdots + \mathfrak{s}_n} \\ &\leq C_f \sqrt{\mathfrak{s}_{\hat{n}+1} + \cdots + \mathfrak{s}_n} \enspace , \end{aligned}$$

with $C_1 = C_1'/(1-\rho_f)$, $C_2 = C_2'/(1-\rho_f)^2$, and $C_f = C_1 + C_2/\sqrt{\mathfrak{s}_n}$. This concludes the proof of Theorem 12.

Note that Assumption 1 played a crucial role in asserting the convergence of $\sum_{t \geq 0} \rho_f^t$. Although the assumption may not hold in general, it can be shown using the results from [35, Section 2.3.4] that there exists a minimal WFA D with $f_D = f^2$ such that $\|\sum_\sigma \mathbf{D}_\sigma\| < 1$. Thus, by Theorem 1 we have $\mathbf{D}_\sigma = \mathbf{Q}^{-1} \mathbf{A}_\sigma^{\otimes 2} \mathbf{Q}$ for some $\mathbf{Q}$. It is then possible to rework the proof of Lemma 13 using D and obtain a similar bound involving $\|\sum_\sigma \mathbf{D}_\sigma\|$ instead of $\|\mathbf{C}\|$ and $\|\tilde{\mathbf{C}}\|$. The details of this approach will be developed in a longer version of this paper, but it suffices to say here that in the end one obtains the same bound given in Theorem 12 with a different constant C_f'.

V. Technical Discussions

This section collects in-detail discussions about two technical aspects of our work. The first one is the relation and consequences of our results with respect to the mathematical theory of low-rank approximation of rational series. The second part makes some remarks about the assumption on the spectral radius of WFA made in our results from Sections III and IV.

A. Low-rank Approximation of Rational Series

We have already discussed how the behavior of the bound (3) matches what intuition suggests. Let us now discuss a little bit more about the quantitative aspects of the bound. In particular, we want to make a few observations about the connection of (3) with low-rank approximation of matrices. We hope these will shed some light on the mathematical theory of low-rank approximation of rational series, and its relations with low-rank approximations of infinite Hankel matrices – a question which certainly deserves further investigation.

Recall that given a rank-n matrix $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ and some $1 \leq \hat{n} < n$, the matrix low-rank approximation problem asks for a matrix $\hat{\mathbf{M}}$ attaining the optimal of the following optimization problem:

$$\min_{\mathrm{rank}(\mathbf{M}') \leq \hat{n}} \|\mathbf{M} - \mathbf{M}'\|_{\mathrm{F}} \enspace .$$

It is well-known the solution to this problem can be computed using the SVD of $\mathbf{M}$ and satisfies the following error bounds in terms of Schatten p-norms:

$$\|\mathbf{M} - \hat{\mathbf{M}}\|_{\mathrm{S},p} = \|(\mathfrak{s}_{\hat{n}+1}, \ldots, \mathfrak{s}_n)\|_p \enspace .$$

Using these results it is straightforward to give lower bounds on the approximation errors achievable by low-rank approximation of rational series in terms of Schatten–Hankel norms (cf. Section III-A). Let $1 \leq p \leq \infty$ and suppose $f \in \ell_{\mathcal{R}}^1$ has rank n and Hankel singular values $\mathfrak{s}_1 \geq \cdots \geq \mathfrak{s}_n$. Then the following holds for every $f' \in \ell_{\mathcal{R}}^1$ with $\mathrm{rank}(f') \leq \hat{n}$:

$$\|f - f'\|_{\mathrm{H},p} \geq \|(\mathfrak{s}_{\hat{n}+1}, \ldots, \mathfrak{s}_n)\|_p \enspace . \tag{5}$$

On the other hand, we define the optimal ℓ^2 approximation error of f with respect to all rational functions of rank at most $\hat{n}$ as

$$\varepsilon_{\hat{n}}^{\mathsf{opt}} = \inf_{\mathrm{rank}(f') \leq \hat{n}} \|f - f'\|_2 \enspace .$$

It is easy to see the infimum will be attained at some $\hat{f}_{\hat{n}}^{\mathsf{opt}} \in \ell_{\mathcal{R}}^2$. If $\hat{f}_{\hat{n}}^{\mathsf{sva}} \in \ell_{\mathcal{R}}^2$ denotes the function realized by the solution obtained from our `SVATruncation` algorithm, then Theorem 12 implies the bound

$$\varepsilon_{\hat{n}}^{\mathsf{opt}} \leq \|f - \hat{f}_{\hat{n}}^{\mathsf{sva}}\|_2 \leq C_f^{1/2} \|(\mathfrak{s}_{\hat{n}+1}, \ldots, \mathfrak{s}_n)\|_1^{1/4} \enspace . \tag{6}$$

Combining the bounds (5) and (6) above, we can conclude that the performance of our approximation $\hat{f}_{\hat{n}}^{\mathsf{sva}}$ with respect to f and $\hat{f}_{\hat{n}}^{\mathsf{opt}}$ can be bounded as follows:

$$\|f - \hat{f}_{\hat{n}}^{\mathsf{opt}}\|_2 \leq \|f - \hat{f}_{\hat{n}}^{\mathsf{sva}}\|_2 \leq C_f^{1/2} \|f - \hat{f}_{\hat{n}}^{\mathsf{opt}}\|_{\mathrm{H},1}^{1/4} \enspace .$$

In future work we plan to investigate the tightness of these bounds and the computational complexity of (approximately) computing $\hat{f}_{\hat{n}}^{\mathsf{opt}}$.

B. Spectral Radius Assumptions

The algorithms presented in Sections III and IV assume their input is a WFA $A = \langle \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\infty, \{\mathbf{A}_\sigma\} \rangle$ such that $\mathbf{A}^{\otimes} = \sum_\sigma \mathbf{A}_\sigma \otimes \mathbf{A}_\sigma$ has spectral radius $\rho(\mathbf{A}^{\otimes}) < 1$. This condition is used in order to guarantee the existence of a closed-form expression for the summation of the series $\sum_{t \geq 0} (\mathbf{A}^{\otimes})^t$. Algorithm `ComputeSVA` uses this expression for computing the Gram matrices $\mathbf{G}_p = \mathbf{P}_A^\top \mathbf{P}_A$ and $\mathbf{G}_s = \mathbf{S}_A^\top \mathbf{S}_A$ associated with the FB rank factorization $\mathbf{H}_{f_A} = \mathbf{P}_A \mathbf{S}_A^\top$ induced by A. A first important remark is that since $\rho(\mathbf{A}^{\otimes})$ is defined in terms of the eigenvalues of $\mathbf{A}^{\otimes}$, the assumption can be tested efficiently. The rest of this sections discusses the following two questions: (1) is the assumption always true in general? and, (2) if not, is there an alternative way to compute the Gram matrices needed by `ComputeSVA`?

Regarding the first question, let us start by pointing out a natural way in which one could try to prove that the assumption always hold. This approach is based on the following result due to F. Denis [36].

Proposition 17. *Let $A = \langle \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\infty, \{\mathbf{A}_\sigma\} \rangle$ be a minimal WFA realizing $f_A \in \ell_{\mathcal{R}}^1$. Then the spectral radius of $\mathbf{A} = \sum_\sigma \mathbf{A}_\sigma$ satisfies $\rho(\mathbf{A}) < 1$.*

In view of this, a natural question to ask is whether the fact $\rho(\mathbf{A}) < 1$ implies $\rho(\mathbf{A}^{\otimes}) < 1$. While this follows from the equation $\rho(\mathbf{M} \otimes \mathbf{M}) = \rho(\mathbf{M})^2$ in the case with $|\Sigma| = 1$, the result is not true in general for arbitrary matrices. In fact, obtaining interesting bounds on the spectral radius of matrices of the form $\mathbf{M}_1 \otimes \mathbf{M}_1 + \mathbf{M}_2 \otimes \mathbf{M}_2$ is an area of active research in linear algebra [37]. Following this approach would require proving new bounds along these lines that apply to matrices defining WFA for absolutely convergent rational series. An alternative approach based on Proposition 17 could be to show that the automaton computing f_A^2 obtained in Lemma 8 is minimal. However, this is not true in general as witnessed by the following example. Let A be the WFA over $\Sigma = \{a, b\}$ with 2 states given by: $\boldsymbol{\alpha}_0^\top = [1 \;\; 0]$, $\boldsymbol{\alpha}_\infty^\top = [1/3 \;\; 1/3]$,

$$\mathbf{A}_a = \begin{bmatrix} 0 & 1/3 \\ 1/3 & 0 \end{bmatrix} ,$$

$$\mathbf{A}_b = \begin{bmatrix} -1/3 & 0 \\ 0 & 1/3 \end{bmatrix} .$$

Note that $\|f_A\|_1 = 1$ and therefore we have $f_A \in \ell^1_{\mathcal{R}}$. It is easy to see, by looking at the rows of $\mathbf{H}_{f_A}$ corresponding to prefixes λ and a, that $\operatorname{rank}(f_A) = 2$; thus, A is minimal. On the other hand, one can check that $f_A^2(x) = 3^{-2(|x|+1)}$ for every $x \in \Sigma^\star$. Thus, f_A^2 has rank 1 and the 4-state WFA for f_A^2 constructed in Lemma 8 is not minimal. In conclusion, though we have not been able to provide a counter-example to the fact that $\rho(\mathbf{A}^{\otimes}) < 1$ when A is a minimal WFA realizing a function $f_A \in \ell^1_{\mathcal{R}}$, we suspect that making progress in either proving or disproving this assumption will require a deeper understanding of the structure of absolutely convergent rational series.

The second question is whether it is possible to compute an SVA efficiently for a WFA such that $\rho(\mathbf{A}^{\otimes}) \geq 1$. The key ingredient here is to provide an alternative way of computing the Gram matrices required in `ComputeSVA`. A first remark is that such Gram matrices are guaranteed to exist regardless of whether the assumption on the spectral radius of $\mathbf{A}^{\otimes}$ holds or not; this follows from the proof of Lemma 9. It also follows from the same proof that each entry $\mathbf{G}_p(i, j)$ corresponds to the inner product $\langle p_i, p_j \rangle$ between two rational functions in $\ell^2_{\mathcal{R}}$; the same holds for the entries of $\mathbf{G}_s$. This observation suggests that, instead of computing the Gram matrices in "one shot" as in Lemma 9, it might be possible to compute each entry $\mathbf{G}_p(i, j)$, $1 \leq i \leq j \leq n$, separately – note the constraint on the indices exploits the fact that $\mathbf{G}_p$ is symmetric. This can be done as follows. Recall that $A_{p,i} = \langle \boldsymbol{\alpha}_0, \mathbf{e}_i, \{\mathbf{A}_\sigma\} \rangle$ computes p_i for all $i \in [n]$. Then the function $f_{p,i,j} = p_i \cdot p_j$ is computed by the WFA $A_{p,i,j} = \langle \boldsymbol{\alpha}_0^{\otimes 2}, \mathbf{e}_i \otimes \mathbf{e}_j, \{\mathbf{A}_\sigma^{\otimes 2}\} \rangle$ with n^2 states. Now observe that by Hölder's inequality we have $f_{p,i,j} \in \ell^1_{\mathcal{R}}$. Therefore, if $\tilde{\mathbf{A}}_{p,i,j} = \langle \tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\alpha}}_\infty, \{\tilde{\mathbf{A}}_\sigma\} \rangle$ is a minimization of $A_{p,i,j}$ with $\operatorname{rank}(f_{p,i,j})$ states, then we must have $\rho(\tilde{\mathbf{A}}) < 1$ by Proposition 17, where $\tilde{\mathbf{A}} = \sum_\sigma \tilde{\mathbf{A}}_\sigma$. Using the same argument as in Lemma 8 we can conclude that

$$\mathbf{G}_p(i, j) = \langle p_i, p_j \rangle = \sum_{x \in \Sigma^\star} f_{p,i,j}(x) = \tilde{\boldsymbol{\alpha}}_0^\top (\mathbf{I} - \tilde{\mathbf{A}})^{-1} \tilde{\boldsymbol{\alpha}}_\infty .$$

This gives an alternate procedure for computing $\mathbf{G}_p$ and $\mathbf{G}_s$ which involves $\Theta(n^2)$ WFA minimizations of automata with n^2 states, each of them taking time $O(n^6)$ (cf. [27]). Hence, the cost of this alternate procedure is of order $O(n^8)$, and should only be used when it is not possible to use the $O(n^6)$ procedure given in Section III-B.

VI. Conclusions and Future Work

With this paper we initiate a systematic study of approximate minimization problems of quantitative systems. Here we have focused our attention on weighted finite automata realizing absolutely convergent rational series. These are, of course, not all rational series but include many situations of interest, for example, all fully probabilistic automata. We have shown how the connection between rational series and infinite Hankel matrices yields powerful tools for analysing approximate minimization problems for WFA: the singular value decomposition of infinite Hankel matrices and the singular values themselves. Our first contribution: an algorithm for computing the SVD of an infinite Hankel matrix by operating on its "compressed" representation as a WFA uses these tools in a crucial way. Such a decomposition leads us to our second contribution: the definition of the singular value automaton (SVA) associated with a rational function f. SVA provide a new canonical form for WFA which is unique under the same conditions guaranteeing uniqueness of the SVD decomposition for Hankel matrices. We were also able to give an efficient algorithm for computing the SVA of a rational function f from any WFA realizing f.

Our second set of contributions are related to the application of SVA canonical forms to the approximate minimization of WFA. The algorithm `SVATruncation` and the corresponding analysis presented in Theorem 12 provide novel and rigorous bounds on the quality of our approximations measured in terms of $\|f - \hat{f}\|_2$, the ℓ^2 norm between the original and minimized functions. The importance of such bounds lies in the fact that they depend only on intrinsic quantities associated with f.

The present paper opens the door to many possible extensions. First and foremost, we will seek further applications and properties of the SVA canonical form for WFA. For example, a simple question that remains unanswered is to what extent the equations in Lemma 11 are enough to characterize the weights of an SVA. In the near future we are also interested in conducting a thorough empirical study by implementing the algorithms presented here. This should serve to validate our ideas and explore their possible applications to machine learning and other applied fields where WFA are used frequently. We will also set out to study the tightness of the bound in Theorem 12 in practical situations, and conjecture further refinements if necessary. It should also be possible to extend our results to other classes of systems closely related to weighted automata. In particular, we want to study approximate minimization problems for weighted tree automata and weighted context-free grammars, for which the notions of Hankel matrix can be naturally extended. Along these lines, it will be interesting to compare our approach to recent works that improve the running time of parsing algorithms by reducing the size of probabilistic context-free grammars using low-rank tensor approximations [38], [39].

Though we have not emphasized it in the present paper, this work is inspired, in part, from the general co-algebraic view of Brzozowski-style minimization [2]. We have expressed everything in very concrete matrix algebra terms because we are using the singular value decomposition in a crucial way. However, there are other minimization schemes for other types of automata coming from other dualities [1] for which we think similar approximate minimization schemes can be developed. A general abstract notion of approximate minimization is, of course, a very tempting subject to pursue and, after we have more examples, it would be certainly high on our agenda. For the moment, however, we will concentrate on concrete instances.

ACKNOWLEDGMENT

We thank F. Denis and anonymous referees for their useful comments on an earlier version of this manuscript. The authors gratefully acknowledge funding received from NSERC.

REFERENCES

[1] N. Bezhanishvili, C. Kupke, and P. Panangaden, "Minimization via duality," in *Logic, Language, Information and Computation - 19th International Workshop, WoLLIC 2012, Buenos Aires, Argentina, September 3-6, 2012. Proceedings*, ser. Lecture Notes in Computer Science, vol. 7456. Springer, 2012, pp. 191–205.

[2] F. Bonchi, M. M. Bonsangue, H. H. Hansen, P. Panangaden, J. Rutten, and A. Silva, "Algebra-coalgebra duality in Brzozowski's minimization algorithm," *ACM Transactions on Computational Logic*, 2014.

[3] J. A. Brzozowski, "Canonical regular expressions and minimal state graphs for definite events," in *Proceedings of the Symposium on Mathematical Theory of Automata*, ser. MRI Symposia Series, J. Fox, Ed. Polytechnic Press of the Polytechnic Institute of Brooklyn, April 1962, no. 12, pp. 529–561, book appeared in 1963.

[4] M. Boreale, "Weighted bisimulation in linear algebraic form," in *CONCUR 2009-Concurrency Theory*. Springer, 2009, pp. 163–177.

[5] F. Bonchi, M. Bonsangue, M. Boreale, J. Rutten, and A. Silva, "A coalgebraic perspective in linear weighted automata," *Information and Computation*, vol. 211, pp. 77–105, 2012.

[6] D. Hsu, S. M. Kakade, and T. Zhang, "A spectral algorithm for learning hidden Markov models," in *COLT*, 2009.

[7] R. Bailly, F. Denis, and L. Ralaivola, "Grammatical inference as a principal component analysis problem," in *ICML*, 2009.

[8] S. M. Siddiqi, B. Boots, and G. Gordon, "Reduced-rank hidden Markov models," in *AISTATS*, 2010.

[9] B. Boots, S. Siddiqi, and G. Gordon, "Closing the learning-planning loop with predictive state representations," in *Proceedings of Robotics: Science and Systems VI*, 2009.

[10] B. Balle, A. Quattoni, and X. Carreras, "A spectral learning algorithm for finite state transducers," in *ECML-PKDD*, 2011.

[11] R. Bailly, X. Carreras, and A. Quattoni, "Unsupervised spectral learning of finite state transducers," in *NIPS*, 2013.

[12] R. Bailly, "Quadratic weighted automata: Spectral algorithm and likelihood maximization," in *ACML*, 2011.

[13] B. Balle and M. Mohri, "Spectral learning of general weighted automata via constrained matrix completion," in *NIPS*, 2012.

[14] A. Recasens and A. Quattoni, "Spectral learning of sequence taggers over continuous sequences," in *ECML-PKDD*, 2013.

[15] B. Balle, X. Carreras, F. Luque, and A. Quattoni, "Spectral learning of weighted automata: A forward-backward perspective," *Machine Learning*, 2014.

[16] M. Mohri, F. C. N. Pereira, and M. Riley, "Speech recognition with weighted finite-state transducers," in *Handbook on Speech Processing and Speech Communication*, 2008.

[17] J. Albert and J. Kari, "Digital image compression," in *Handbook of Weighted Automata*, 2009.

[18] K. Knight and J. May, "Applications of weighted automata in natural language processing," in *Handbook of Weighted Automata*, 2009.

[19] C. Baier, M. Größer, and F. Ciesinski, "Model checking linear-time properties of probabilistic systems," in *Handbook of Weighted automata*, 2009.

[20] A. de Gispert, G. Iglesias, G. Blackwood, E. Banga, and W. Byrne, "Hierarchical phrase-based translation with weighted finite-state transducers and shallow-n grammars," *Computational Linguistics*, 2010.

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, 1977.

[22] B. Balle, W. Hamilton, and J. Pineau, "Methods of moments for learning stochastic languages: Unified presentation and empirical comparison," in *ICML*, 2014.

[23] A. Kulesza, N. R. Rao, and S. Singh, "Low-Rank Spectral Learning," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, 2014, pp. 522–530.

[24] A. Kulesza, N. Jiang, and S. Singh, "Low-rank spectral learning with weighted loss functions," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 2015.

[25] S. Kiefer and B. Wachter, "Stability and complexity of minimising probabilistic automata," in *Proceedings of the 41st International Colloquium on Automata, Languages and Programming (ICALP), part II*, ser. LNCS, J. E. et al., Ed., vol. 8573. Copenhagen, Denmark: Springer, 2014, pp. 268–279.

[26] K. Zhu, *Operator Theory in Function Spaces*. American Mathematical Society, 1990, vol. 138.

[27] J. Berstel and C. Reutenauer, *Noncommutative Rational Series with Applications*. Cambridge University Press, 2011.

[28] P. Dupont, F. Denis, and Y. Esposito, "Links between probabilistic automata and hidden Markov models: probability distributions, learning models and induction algorithms," *Pattern Recognition*, 2005.

[29] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Readings in speech recognition*, A. Waibel and K. Lee, Eds., 1990, pp. 267–296.

[30] S. Singh, M. James, and M. Rudary, "Predictive state representations: a new theory for modeling dynamical systems," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 2004.

[31] R. Bailly and F. Denis, "Absolute convergence of rational series is semi-decidable," *Inf. Comput.*, vol. 209, no. 3, pp. 280–295, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.ic.2010.11.004

[32] L. N. Trefethen and D. Bau III, *Numerical linear algebra*. Siam, 1997.

[33] B. Balle, P. Panangaden, and D. Precup, "A canonical form for weighted automata and applications to approximate minimization," *CoRR*, vol. abs/1501.06841, 2015. [Online]. Available: http://arxiv.org/abs/1501.06841

[34] R. Bhatia, *Matrix analysis*. Springer, 1997, vol. 169.

[35] R. Bailly, "Méthodes spectrales pour l'inférence grammaticale probabiliste de langages stochastiques rationnels," Ph.D. dissertation, Aix-Marseille Université, 2011.

[36] F. Denis, Personal communication, 2015.

[37] S. V. Lototsky, "Simple spectral bounds for sums of certain kronecker products," *Linear Algebra and its Applications*, vol. 469, pp. 114–129, 2015.

[38] M. Collins and S. B. Cohen, "Tensor decomposition for fast parsing with latent-variable PCFGs," in *NIPS*, 2012.

[39] S. B. Cohen, G. Satta, and M. Collins, "Approximate PCFG parsing using tensor decomposition," in *NAACL*, 2013.