Extracting Weighted Automata for Approximate Minimization in Language Modelling

Clara Lacroce[*†] CLARA.LACROCE@MAIL.MCGILL.CA
Prakash Panangaden PRAKASH@CS.MCGILL.CA
School of Computer Science, McGill University & Mila , Montréal, Canada

Guillaume Rabusseau GRABUS@IRO.UMONTREAL.CA
DIRO, Université de Montréal & Mila, Montréal, Canada

Editors: Jane Chandlee, Rémi Eyraud, Jeffrey Heinz, Adam Jardine, and Menno van Zaanen

Abstract

In this paper we study the approximate minimization problem for language modelling. We assume we are given some language model as a black box. The objective is to obtain a weighted finite automaton (WFA) that fits within a given size constraint and which mimics the behaviour of the original model while minimizing some notion of distance between the black box and the extracted WFA. We provide an algorithm for the approximate minimization of black boxes trained for language modelling of sequential data over a one-letter alphabet. By reformulating the problem in terms of Hankel matrices, we leverage classical results on the approximation of Hankel operators, namely the celebrated Adamyan-Arov-Krein (AAK) theory. This allows us to use the spectral norm to measure the distance between the black box and the WFA. We provide theoretical guarantees to study the potentially infinite-rank Hankel matrix of the black box, without accessing the training data, and we prove that our method returns an asymptotically-optimal approximation.

Keywords: Approximate minimization, WFA extraction, Hankel matrices, Recurrent Neural Networks, language modelling

1. Introduction

Interpretability and high computational cost are two of the main challenges arising from the use of deep learning models (Doshi-Velez and Kim, 2017). The need to address these issues is at the root of the increasing number of works focusing on knowledge distillation (Hinton et al., 2015). In the case of sequential data, particular attention has been given to the problem of extracting, from a Recurrent Neural Network (RNN) (Hochreiter and Schmidhuber, 1997), a weighted finite automaton (WFA) (Ayache et al., 2018; Rabusseau et al., 2019; Weiss et al., 2019; Okudono et al., 2020; Eyraud and Ayache, 2020; Theertha Suresh et al., 2019; Zhang et al., 2021). In fact, WFAs are a less expensive alternative to RNNs, while still being expressive and suited for sequence modelling and prediction (Denis and Esposito, 2008; Cortes et al., 2004).

The task of knowledge distillation is closely related to the more general *approximate minimization problem*, where the objective is to find a model, smaller than the original one, that imitates its behaviour while minimizing the approximation error. The advantage of

* Corresponding author
† The authors appear in alphabetical order

doing approximate minimization instead of regular extraction is that it allows us to search for the best WFA among those of a predefined size. Since automata benefit from a graphical representation, bounding the number of states can help improve interpretability (Hammerschmidt et al., 2016). In this paper, we tackle the approximate minimization problem for black boxes trained for language modelling over a one-letter alphabet. We remark that, even though this is a very limited setting, it constitutes a first fundamental step towards developing provable approximation algorithms for black box models.

A key point in solving approximation tasks is to decide how to quantify the error. We propose to rewrite the problem in terms of Hankel matrices, mathematical objects related to functions defined on sequential data. In particular, we choose to measure the error in terms of the *spectral norm*, because of some of its desirable features. Indeed, the spectral norm of the Hankel matrix of a WFA can be computed in polynomial time (Balle et al., 2021) and we show that, similarly, minimizing the approximation error between a WFA and a black box model can be (asymptotically) solved optimally in a tractable way. Thus, using our method, we can measure the distance between a given RNN and the extracted WFA. This is particularly valuable, especially in light of the paper of Marzouk and de la Higuera (2020), where the authors show that the general equivalence problem between classes of WFAs and RNNs is at best intractable, if not undecidable. The choice of this norm has the advantage that it allows us to analyze different models through their Hankel matrices, independently of the specific architecture considered. This means that addressing the approximate minimization problem using the spectral norm can facilitate the comparison between different classes of models, and the development of a distance that can be precisely computed and minimized. This is possible because Hankel matrices are at the core of the influential work of Adamyan et al. (1971), which constitutes the main theoretical background on which we build our analysis. This theory has been applied before to the approximate minimization problem for WFAs, but the approach relies on the Hankel matrix considered to have known finite rank, so it cannot be directly generalized (Balle et al., 2021).

Contributions The main contributions of this paper are the following:

- We present a new theoretical framework for WFA extraction from a black box trained for language modelling of sequential data over a one-letter alphabet.
- We use tools from control theory and arguments from random matrix theory to extend the work of Balle et al. (2021) to the case of black boxes having infinite-rank Hankel matrices.
- We propose an algorithm that, given a black box model $\mathcal{M}$ on a one letter alphabet and a target size k, returns a WFA with k states corresponding to an asymptotically-optimal spectral approximation of $\mathcal{M}$. We do not assume any knowledge on the internal structure of the black box, nor on the training data.
- We propose a new way to compute the distance between a black box and the extracted WFA, based on AAK theory. We provide bounds on the approximation error in terms of spectral and ℓ^2 norm, and strategies to improve precision when the rank is infinite.

2. Background

2.1. Notation

Let $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$ be the set of natural, integer and real numbers, respectively. We use bold letters for vectors and matrices; all vectors are column vectors unless otherwise specified. We denote with $\mathbf{v}(i)$, $\mathbf{M}(i,:)$ and $\mathbf{M}(:,j)$ the i-th component of the vector $\mathbf{v}$, and the i-th row and j-th column of $\mathbf{M}$, respectively. A *rank factorization* of $\mathbf{M} \in \mathbb{R}^{p\times q}$ of rank n is a factorization $\mathbf{M} = \mathbf{PQ}$, with $\mathbf{P} \in \mathbb{R}^{p\times n}$, $\mathbf{Q} \in \mathbb{R}^{n\times q}$, with $\mathbf{P}$, $\mathbf{Q}$ of rank n. Let $\mathbf{M} \in \mathbb{R}^{p\times q}$ of rank n, the compact *singular value decomposition* (SVD) of $\mathbf{M}$ is $\mathbf{M} = \mathbf{UDV}^\top$, where $\mathbf{U} \in \mathbb{R}^{p\times n}$, $\mathbf{D} \in \mathbb{R}^{n\times n}$, $\mathbf{V} \in \mathbb{R}^{q\times n}$, with $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{1}$, and $\mathbf{D}$ is diagonal. The columns of $\mathbf{U}$ and $\mathbf{V}$ are called left and right *singular vectors*, while the entries $\sigma_0 \geq \cdots \geq \sigma_{n-1} > 0$ of $\mathbf{D}$ are the *singular values*. The *Moore-Penrose pseudo-inverse* $\mathbf{M}^+$ of $\mathbf{M}$ is the unique matrix such that $\mathbf{MM}^+\mathbf{M} = \mathbf{M}$, $\mathbf{M}^+\mathbf{MM}^+ = \mathbf{M}^+$, with $\mathbf{M}^+\mathbf{M}$ and $\mathbf{MM}^+$ Hermitian.

A *Hilbert space* is a complete normed vector space where the norm arises from an inner product. Let X, Y be Hilbert spaces. A linear operator $T : X \to Y$ is *bounded* if it has finite *operator norm*, *i.e.* $\|T\|_{op} = \sup_{\|g\|_X \leq 1} \|Tg\|_Y < \infty$, while is *compact* if it is the limit of finite rank operators in the operator norm. We write $T^i \to T$ if T is the limit of the sequence of operators $\{T^i\}_{i\geq 0}$. Let $T : X \to Y$ compact, the *adjoint* T^* is the linear operator $T^* : Y \to X$ such that $\langle Tx, y\rangle_Y = \langle x, T^*y\rangle_X$, where $\langle\cdot,\cdot\rangle$ is the inner product of the corresponding Hilbert space, $x \in X$, $y \in Y$. The *singular numbers* $\{\sigma_n\}_{n\geq 0}$ of T are the square roots of the eigenvalues of T^*T, arranged in decreasing order. A singular number is *simple* if it is not repeated. Let $\mathbf{T}$ be the infinite matrix associated with T by some canonical orthonormal basis. The Hilbert-Schmidt decomposition generalizes the compact SVD for the matrix of a compact operator T: $\mathbf{Tx} = \sum_{n\geq 0} \sigma_n \langle \mathbf{x}, \boldsymbol{\xi}_n\rangle \boldsymbol{\eta}_n$ (see Zhu (1990)). The *spectral norm* $\|\mathbf{T}\|$ of the matrix of the operator T is the largest singular number, and corresponds to the operator norm of T. Let $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$ and $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$ be the complex unit circle and disc, respectively. Let $p > 1$, $\mathcal{L}^p(\mathbb{T})$ is the space of measurable functions on $\mathbb{T}$ with the p-th power of their absolute value Lebesgue integrable.

2.2. Hankel Matrix and WFAs

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

Weighted finite automata are a class of models defined over sequential data. A *weighted finite automaton* (WFA) of n states over Σ is a tuple $A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}, \boldsymbol{\beta}\rangle$, 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 matrix containing the transition weights associated with each symbol $a \in \Sigma$. While WFAs can in general be defined over semirings, we will 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

$f_A(x) = \boldsymbol{\alpha}^\top \mathbf{A}_{x_1} \cdots \mathbf{A}_{x_t} \boldsymbol{\beta} = \boldsymbol{\alpha}^\top \mathbf{A}_x \boldsymbol{\beta}$. We say that f is *rational* if there exists a WFA A with $f = f_A$, and the *rank* of f is the size of the smallest WFA realizing f. We can use the Hankel matrix $\mathbf{H}_f$ to recover information about the WFA.

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

Given a Hankel matrix $\mathbf{H}_f$ of rank n, we can recover the minimal WFA A realizing f by using the method proposed in Balle et al. (2014), an efficient *spectral algorithm* which is robust to noise. In particular, we can consider a basis $\mathcal{B} = (\mathcal{P}, \mathcal{S})$, with $\mathcal{P}, \mathcal{S} \subset \Sigma^*$, and a sub-block $\mathbf{H}_\mathcal{B}$ of $\mathbf{H}_f$ defined over $\mathcal{B}$. The method can be applied whenever $\mathcal{B}$ is *prefix-closed* and *complete*, *i.e.*, when $\mathcal{P} = \mathcal{P}' \cdot \Sigma'$ for some $\mathcal{P}'$, and $\mathbf{H}_\mathcal{B}$ has rank n. In this case, we can consider the sub-block $\mathbf{H}_a$ defined over $\mathcal{B}$ by $\mathbf{H}_a(u, v) = \mathbf{H}(u \cdot a, v)$ for each $a \in \Sigma'$, and the vectors $\mathbf{h}_{\mathcal{P},\varepsilon}$, $\mathbf{h}_{\varepsilon,\mathcal{S}}$ having coordinates $\mathbf{h}_{\mathcal{P},\varepsilon}(u) = \mathbf{H}(u, \varepsilon)$ and $\mathbf{h}_{\varepsilon,\mathcal{S}}(v) = \mathbf{H}(\varepsilon, v)$. Then, from the rank factorization $\mathbf{H}_\varepsilon = \mathbf{P}\mathbf{S}$ we can compute a minimal WFA $A = \langle \boldsymbol{\alpha}, \{\mathbf{A}_a\}, \boldsymbol{\beta} \rangle$ for f:

$$\boldsymbol{\alpha}^\top = \mathbf{h}_{\varepsilon,\mathcal{S}}^\top \mathbf{S}^+, \quad \boldsymbol{\beta} = \mathbf{P}^+ \mathbf{h}_{\mathcal{P},\varepsilon}, \quad \mathbf{A}_a = \mathbf{P}^+ \mathbf{H}_a \mathbf{S}^+. \tag{1}$$

2.3. Recurrent Neural Networks

Recurrent Neural Networks (Hochreiter and Schmidhuber, 1997), or RNNs, are a class of neural networks designed to process sequential data. Unlike feedforward neural networks, RNNs maintain an internal memory based on history information through the hidden states. At each timestep, a RNN receives an input and returns a new state vector, depending on the input and on the sequence received so far. There exists several types of architectures for these models, which makes them well suited for a variety of tasks (Weiss et al., 2018; Merrill et al., 2020). Analogously to Ayache et al. (2018) and Weiss et al. (2019), we focus on LM-RNNs, where the RNN is trained for *language modelling*, and the task is to predict the next element in a sequence. Thus, a LM-RNN can be seen as computing the probability associated to a string, and can then be represented by a Hankel matrix.

2.4. AAK Theory

The key idea behind our method is that, since a model computing $f : \Sigma^* \to \mathbb{R}$ corresponds to a Hankel matrix $\mathbf{H} = \mathbf{H}_f$, the minimization problem can be reformulated using Hankel matrices. The objective becomes to find a Hankel matrix $\mathbf{G}$ that approximates $\mathbf{H}$ optimally in the spectral norm, and then extract a WFA from it. This approach has been explored before by Balle et al. (2021), but their method does not generalize to infinite-rank matrices. We recall a well known result in low-rank matrix approximation.

Theorem 3 (Eckart and Young (1936)) *Let $\mathbf{H}$ be a Hankel matrix of rank n, and let $\sigma_0 \geq \cdots \geq \sigma_{n-1} > 0$ be its singular numbers. Then, if $\mathbf{R}$ is a matrix of rank k, we have $\|\mathbf{H} - \mathbf{R}\| \geq \sigma_k$, and the minimum is attained when $\mathbf{R}$ is the truncated SVD of $\mathbf{H}$.*

Unfortunately this result does not solve our problem, since truncating the SVD does not necessarily produce a Hankel matrix, which is required to recover a WFA. When $|\Sigma| = 1$, the issue can be solved using a theory of optimal approximation called Adamyan-Arov-Krein

(AAK) theory (Adamyan et al., 1971), which allows us to search for the best approximation directly in the set of finite-rank Hankel matrices. In order to introduce AAK theory, for the rest of this section we will assume $|\Sigma| = 1$. The same assumption will be required in the contribution (for more details we refer the reader to Section 3.1). When the alphabet only has one letter, we can denote a string with the number corresponding to how many times the single character is repeated (*e.g.* $'aaa' = 3$), and we can identify Σ^* with $\mathbb{N}$. Let ℓ^2 be the Hilbert space of square-summable sequences over $\mathbb{N}$. We interpret the Hankel matrix $\mathbf{H}_f$ associated to $f : \mathbb{N} \to \mathbb{R}$ as the expression, in terms of the canonical basis, of a linear Hankel operator $H_f : \ell^2 \to \ell^2$. To reformulate the problem in the setting of AAK theory, we embed ℓ^2 into $\ell^2(\mathbb{Z})$, and apply the Fourier isomorphism to associate a complex function to each sequence in $\ell^2(\mathbb{Z})$. In fact, a function $\phi(z) \in \mathcal{L}^2(\mathbb{T})$ in the complex variable z can be represented by its Fourier expansion $\phi(z) = \sum_{n\in\mathbb{Z}} \widehat{\phi}(n) z^n$, and identified with the sequence of its Fourier coefficients $\widehat{\phi}(n) = \int_{\mathbb{T}} \phi(z)\bar{z}^n dz$, $n \in \mathbb{Z}$ using the orthonormal basis $\{z^n\}_{n\in\mathbb{Z}}$. Then we partition the function space $\mathcal{L}^2(\mathbb{T})$ into two subspaces.

Definition 4 *For $0 < p \leq \infty$, the **Hardy space** $\mathcal{H}^p$ and the **negative Hardy space** $\mathcal{H}^p_-$ on $\mathbb{T}$ are the subspaces of $\mathcal{L}^p(\mathbb{T})$ defined as:*

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

Since the elements of $\mathcal{H}^p$ can be canonically identified with the set of p-integrable functions analytic in $\mathbb{D}$, we will make no difference between these functions in the complex unit disc and their boundary value on the complex unit circle (Nikol'Skii, 2002).

It is possible to characterize Hankel operators using Hardy spaces (more details can be found in Nikol'Skii (2002)). Let $\mathbb{P}_- : \mathcal{L}^2(\mathbb{T}) \to \mathcal{H}^2_-$ be the orthogonal projection on the negative Hardy space.

Definition 5 *Let $\phi(z)$ be a function in $\mathcal{L}^2(\mathbb{T})$. A **Hankel operator** is an operator $H_\phi : \mathcal{H}^2 \to \mathcal{H}^2_-$ defined by $H_\phi f(z) = \mathbb{P}_- \phi f(z)$. The function $\phi(z)$ is a **symbol** for H_ϕ.*

From now on, Hankel operators will always be interpreted in Hardy spaces. We recall that a complex function $\phi(z)$ is *rational* if $\phi(z) = p(z)/q(z)$, with $p(z)$ and $q(z)$ polynomials, and it is *strictly proper* if the degree of $p(z)$ is strictly smaller than that of $q(z)$. Finite rank Hankel operators are closely related to the theory of rational functions.

Theorem 6 (Kronecker (1881)) *Let H_ϕ be a bounded Hankel operator with matrix $\mathbf{H}$. Then $\mathbf{H}$ has finite rank if and only if $\mathbb{P}_-\phi$ is a strictly proper rational function. Moreover the rank of $\mathbf{H}$ is equal to the number of poles in $\mathbb{D}$ (counted with multiplicities) of $\mathbb{P}_-\phi$.*

We remark an important property of Hankel matrices (see Appendix A for an example).

Remark 7 *On the one hand we can consider the matrix $\mathbf{H}$ with respect to the basis of ℓ^2, and associate $\mathbf{H}$ with the function $f : \mathbb{N} \to \mathbb{R}$. In this case $\mathbf{H}(i, j) = f(i + j)$ for $i, j \geq 0$. On the other hand, we can look at $\mathbf{H}$ with respect to the standard orthonormal bases of $\mathcal{H}^2$ and $\mathcal{H}^2_-$. Now, $\mathbf{H}$ is associated with $\phi(z) \in \mathcal{L}^2(\mathbb{T})$, and we have $\mathbf{H}(j, k) = \widehat{\phi}(-j - k - 1)$. Note that f and ϕ are related through the Fourier isomorphism, with $f(n) = \widehat{\phi}(-n - 1)$.*

The core of AAK theory is that, when minimizing a compact Hankel operator, the constraint of preserving the Hankel property does not affect the quality of the approximation.

Theorem 8 (Adamyan et al. (1971)) *Let H be a compact Hankel operator, with matrix $\mathbf{H}$ of rank n and singular numbers $\sigma_0 \geq \cdots \geq \sigma_{n-1} > 0$. Then, there exists a unique Hankel operator G_k of rank $k < n$ such that $\|\mathbf{H} - \mathbf{G}_k\| = \sigma_k$,* i.e. *G_k is the optimal approximation.*

Using the following theorem, based on the proof of Theorem 8, we can find a symbol for the best approximation. We recall that a *σ-Schmidt pair* $\{\boldsymbol{\xi}, \boldsymbol{\eta}\}$ for H is a couple of vectors such that: $\mathbf{H}\boldsymbol{\xi} = \sigma\boldsymbol{\eta}$ and $\mathbf{H}^*\boldsymbol{\eta} = \sigma\boldsymbol{\xi}$.

Theorem 9 (Chui and Chen (1997)) *Let $\{\boldsymbol{\xi}_k, \boldsymbol{\eta}_k\}$ be any σ_k-Schmidt pair for H. We consider a bi-infinite upper triangular matrix $\mathbf{T}$, having zeros on the main diagonal, first row $\mathbf{T}(0,k) = \mathbf{H}(0,k-1)$ for $k > 0$, and remaining entries defined by $\mathbf{T}(j,k) = \mathbf{T}(j+1,k+1)$. Let $\mathbf{z} = \begin{pmatrix} 1 & z & z^2 & \dots \end{pmatrix}^\top$ where z is the complex variable. Then, the rational function $r(z)$ corresponding to the symbol of the best approximation of rank k is:*

$$r(z) = \mathbb{P}_- \left(\frac{\mathbf{z}^\top \mathbf{T} \boldsymbol{\xi}}{\mathbf{z}^\top \boldsymbol{\xi}} \right). \tag{2}$$

For an example of the matrix $\mathbf{T}$, we refer the reader to Equation 6.

We conclude emphasizing the important relation between matrix and operator.

Remark 10 *A Hankel matrix $\mathbf{H}$ can be seen as the representation of a Hankel operator H by means of a canonical basis. As noted in Remark 7, H can be viewed as acting between sequences or between Hardy spaces, depending on the basis used. While we are interested in matrices, most of the results are stated for operators. Working with the basis of the Hardy space let us alternate between matrix and operator, transferring results from one interpretation to the other. Moreover, we recall that $\|H - G\| = \|\mathbf{H} - \mathbf{G}\|$, where on the left we have operators and operator norm, and on the right matrices and spectral norm. While we keep the notation distinct to remain faithful to definitions (*e.g.*, compactness is defined for H, not for $\mathbf{H}$), for an intuition of the results one can think in terms of Hankel matrices.*

3. Asymptotically-Optimal Approximate Minimization

We are now ready to introduce the main contribution of this paper.

3.1. Problem Formulation

We recall that a bounded operator H is compact if and only if there exists a sequence of finite rank operators $\{H^i\}_{i\geq 0}$ converging to it, *i.e.* if $H^i \to H$. Let G_k and G_k^i be the rank k optimal approximations to H and H^i, respectively, according to Theorem 8. We say that the sequence of matrices $\{\mathbf{G}_k^i\}_{i\geq 0}$ is an *asymptotic sequence* for $\mathbf{G}_k$, if the corresponding sequence of operators $\{G_k^i\}_{i\geq 0}$ converges to the operator G_k, *i.e.*, if $G_k^i \to G_k$. Note that, if $\{\sigma_j\}_{j\geq 0}$ are the singular numbers of H, for an asymptotic sequence we have:

$$\lim_{i\to\infty} \|H - G_k^i\| = \sigma_k. \tag{3}$$

We can now formally define the approximation problem. Let $|\Sigma| = 1$, $\Sigma^* = \mathbb{N}$. We consider a LM-RNN computing a function $f : \mathbb{N} \to \mathbb{R}$, with Hankel matrix $\mathbf{H}$ corresponding

to the operator H. Let k be the target size of the approximation, and $n > k$. We denote with G_k the optimal rank k approximation of H. We say that a WFA $\widehat{A}_k^n$ with k states is an *asymptotically-optimal* (n,k)*-approximation* for the LM-RNN if the Hankel matrix $\mathbf{G}_k^n$ of $\widehat{A}_k^n$ belongs to an asymptotic sequence for $\mathbf{G}_k$.

Intuitively, we can consider a sequence of finite rank matrices $\{\mathbf{H}^i\}_{i\geq 0}$ converging to $\mathbf{H}$, and associate to each of them a WFA (Theorem 6). This means that we have a sequence of WFAs of increasing size that "converges" to the LM-RNN. The matrix $\mathbf{G}_k$ of rank k corresponds to the optimal approximation for $\mathbf{H}$, *i.e.*, it is the WFA $\widehat{A}_k$ with k states that best approximate the LM-RNN. Now, from the sequence of matrices $\mathbf{G}_k^i$ of optimal rank k approximations, we obtain a second sequence of WFAs $\widehat{A}_k^i$, all having size k. When $\{\mathbf{G}_k^i\}_{i\geq 0}$ is an asymptotic sequence for $\mathbf{G}_k$, the corresponding sequence of WFAs "converges" to $\widehat{A}_k$.

We will study the convergence of asymptotic sequences in the next section. In particular, we will prove that a solution for the asymptotically-optimal problem can be obtained from Theorem 8, but it is not unique, since different sequences $\{\mathbf{H}^i\}_{i\geq 0}$ lead to different approximations. Nonetheless, we will show that we can get arbitrarily close to the optimum.

We briefly remark that it is possible to consider an alternative formulation of the approximate minimization problem (Kung and Lin, 1981). In this case, instead of fixing the size of the approximation, we set the tolerance allowed for the approximation error. Thus, the objective becomes to find the smallest possible WFA such that the spectral norm of the approximation error is smaller than a fixed constant ρ of choice. In this case, if $\rho \in (\sigma_k, \sigma_{k-1})$, then the best approximation has size at least $k-1$, and can be found following the same solution we will present for the standard approximation problem.

Assumptions The main limitation of this approach is that the results outlined in Section 2.4 can be applied only if $|\Sigma| = 1$. In this case, Σ^* can be identified with $\mathbb{N}$, and canonically embedded into $\mathbb{Z}$. This fundamental step allows us to use the Fourier isomorphism to reformulate the problem in the Hardy space, where it can be solved using Theorem 8. If $|\Sigma| > 1$, Σ^* is a free non-abelian monoid, therefore it cannot be embedded into $\mathbb{Z}$. Therefore, for the rest of the paper we will assume $|\Sigma| = 1$, and identify $\Sigma^* = \mathbb{N}$.

We remark that the proof of Theorem 8 is constructive only for compact operators. We will show that compactness is automatically respected by LM-RNNs (Theorem 12) and that the necessary condition is actually less restrictive. In fact, if f is the function computed by the black box considered, it is enough that $f \in \ell^1$. Thus, even though we mainly refer to LM-RNNs, the proposed algorithm can be applied to any black box for language modelling on a one-letter alphabet, for example transformers (Vaswani et al., 2017).

3.2. Compactness of the Hankel Matrix

To apply the results of Section 2.4, we need to find a way to test for compactness. This is the main theoretical challenge addressed by the paper. In fact, with matrices of known finite rank, like in the case of WFAs, compactness is achieved by requiring $f \in \ell^2$ (Balle et al., 2019). Then, the problem can be rewritten in terms of finite matrices, the Gramians, and it is possible to find an algorithm returning the parameters of the unique best approximating WFA (Balle et al., 2021). Instead, in the case of LM-RNNs we don't have access to the full bi-infinite Hankel matrix $\mathbf{H}$, and the unknown rank might not be finite. Therefore, there is no guarantee that the problem can be solved algorithmically.

As noted before, the operator H is compact if and only if there is a sequence of finite rank operators $\{H^i\}_{i\geq 0}$, with $H^i \to H$. Given such converging sequence, the key idea is to find the best approximation of rank k for each of its element. Note that every H^i has known finite rank, so the approximation problem can be solved algorithmically. It remains to ensure the continuity of the approximation: if G_k and G_k^i are the optimal approximations of H and of H^i, respectively, we want $\{\mathbf{G}_k^i\}_{i\geq 0}$ to be an asymptotic sequence for $\mathbf{G}_k$, so that $G_k^i \to G_k$. This problem has been analyzed, for signal processing, in the fundamental work of Chui et al. (1991) and Chui and Li (1994). We recall the following result.

Theorem 11 (Chui and Li (1994)) *Let H be a bounded Hankel operator, $\{\sigma_i\}_{i\geq 0}$, its singular numbers. Suppose to have a sequence $\{H^i\}_{i\geq 0}$ of bounded Hankel operators such that $H^i \to H$. Let G_k and G_k^i be the unique optimal approximations of rank k of H and of H^i for any i, respectively. If $\sigma_{k-1} \neq \sigma_k$, then the sequence $\{G_k^i\}_i \geq 0$ converges to G_k.*

This theorem gives us the conditions under which we can solve the approximation problem (at least asymptotically) for the matrix $\mathbf{H}(i,j) = f(i+j)$ of the LM-RNN.

The first step is to find a converging sequence of operators (matrices). We can define one by truncation: let $t \geq 0$, we consider the sequence of matrices defined as:

$$\mathbf{H}^t(i,j) = \begin{cases} f(i+j) & \text{if } i+j \leq t \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

(see Equation 6 for an example). We have the following theorem:

Theorem 12 *Let $|\Sigma| = 1$. Let $f : \mathbb{N} \to \mathbb{R}$ be the function computed by a black box for language modelling, $\mathbf{H}$ the Hankel matrix. Let $\{\mathbf{H}^t\}_{t\geq 0}$ as in Equation 4. Then, since $f \in \ell^1$, we have that $H^t \to H$.*

We remark that the proof, that can be found in Appendix B, relies only on $f \in \ell^1$. Note that we have found a sequence of finite rank operators converging to H, thus $\mathbf{H}$ is compact.

The second step is to ensure that the property $\sigma_{k-1} \neq \sigma_k$ on the singular numbers of H holds when k is the size of the best approximation. This condition cannot be tested experimentally, since we don't have access to the infinite Hankel matrix $\mathbf{H}$. Instead, we can address the problem by using arguments from random matrix theory. In fact, up to at worst a small perturbation, we can view any $\mathbf{H}^t$ for $t > 0$ as a random matrix having only simple singular values with probability one, and this property holds (in limit) also for $\mathbf{H}$ (von Neumann and Wigner, 1993; Tao and Vu, 2014). Note that compact operators have simple spectrum after arbitrarily small perturbations, which do not have a big effect on the quality of the result since the spectrum of symmetric matrices is very stable (Hörmander and Melin, 1994; Kato, 2013; Tao, 2012). In practice, for most settings the Hankel matrix $\mathbf{H}$ will satisfy the condition of Theorem 11 with probability one. This is the case, for example, of RNNs trained using a gradient based method with a random initialization. On the other hand, to keep our analysis general, we also need to consider an adversarial setting, in which the black box to approximate is specifically chosen to have $\sigma_{k-1} = \sigma_k$. To avoid this kind of situation we can add some random noise to $\mathbf{H}$ post training. To preserve compactness, it is important to choose the Hankel matrix of noise $\mathbf{N}$ appropriately. For instance, $\mathbf{N}$ can be a Hankel matrix, with first row $\mathbf{N}(0,j)$ sampled uniformly in the interval

$[-(j+2)^{-p}, (j+2)^{-p}]$, with $p \geq 2$ fixed, so that the operator N is compact. Moreover, for every $\varepsilon > 0$, we can find an exponent $p \geq 2$ such that $\|\mathbf{N}\| \leq \varepsilon$, so the perturbation can be chosen to be arbitrarily small. Note that $\mathbf{H} + \mathbf{N}$ is then a random matrix corresponding to a compact Hankel operator, and satisfies the conditions of Theorem 11 with probability one. We will address the additional error due to small perturbations in Section 4.

We are finally ready to show that if $\mathbf{H}^n$ belongs to the sequence of bi-infinite truncation matrices $\{\mathbf{H}^t\}_{t \geq 0}$ introduced in Equation 4, an asymptotically-optimal (n, k)-approximation can be found by solving the problem described by Theorem 8 for $\mathbf{H}^n$ (proof in Appendix B).

Theorem 13 *Let $\mathbf{H}$ and $\mathbf{H}^n$ be as above, and assume $\sigma_k \neq \sigma_{k-1}$. If $\mathbf{G}_k^n$ is the optimal approximation of $\mathbf{H}^n$ according to Theorem 8, then a WFA having Hankel matrix $\mathbf{G}_k^n$ is an asymptotically-optimal (n, k)-approximation, and we have:*

$$\sigma_k \leq \|\mathbf{H} - \mathbf{G}_k^n\| \leq \sigma_k + 2\left(1 - \sum_{i=0}^{n} f(i)\right). \tag{5}$$

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

3.3. Algorithm

When the rank r of $\mathbf{H}$ is finite and known, it is possible to find directly the optimal approximation. This can be done by first extracting a WFA of size r from the LM-RNN (Ayache et al., 2018), and then applying the algorithm of Balle et al. (2021) to obtain the unique optimal WFA. Therefore, in our algorithm we focus on the case in which the rank r is unknown, and look for an asymptotically optimal approximation. This entails assuming that the truncation $\mathbf{H}^n$ has full rank: if this was not the case, since $\mathbf{H}^n$ is the leading principal submatrix of $\mathbf{H}$, we would have $r = \operatorname{rank}(\mathbf{H}^n)$ (Al'pin, 2017).

To simplify the notation across this section, we set $f_i = f(i)$. We recall the two bi-infinite matrices necessary to find the best approximation:

$$\mathbf{H}^n = \begin{pmatrix} f_0 & f_1 & \dots & f_{n-1} & 0 & \dots \\ f_1 & & \iddots & \iddots & \vdots & \\ \vdots & \iddots & \iddots & & \vdots & \\ f_{n-1} & \iddots & & & \vdots & \\ 0 & \dots & \dots & \dots & 0 & \\ \vdots & & & & & \ddots \end{pmatrix}, \qquad \mathbf{T} = \begin{pmatrix} 0 & f_0 & \dots & f_{n-1} & 0 & \dots \\ \vdots & \ddots & \ddots & & f_{n-1} & \\ \vdots & & \ddots & \ddots & \vdots & \\ \vdots & & & \ddots & f_0 & \\ 0 & \dots & \dots & \dots & 0 & \\ \vdots & & & & & \ddots \end{pmatrix}. \tag{6}$$

The key to successfully implement Theorem 8, which applies only to infinite matrices, is in the definition of the truncation. In fact, this allows us to discard the zero-part and work only with the $n \times n$ sub-block of $\mathbf{H}^n$, which we will still denote with $\mathbf{H}^n$ for the sake of simplicity. Analogously, if $\mathbf{z}$ and $\mathbf{T}$ are the infinite vector and matrix defined in Theorem 9 for $\mathbf{H}$, in the algorithm we will consider the truncations associated to $\mathbf{H}^n$:

$$\mathbf{z}^n(i) = \mathbf{z}(i), \quad \mathbf{T}^n(i,j) = \mathbf{T}(i,j) \quad \text{for } i,j < n, \qquad \mathbf{z}^n \in \mathbb{R}^{n \times 1}, \mathbf{T}^n \in \mathbb{R}^{n \times n} \tag{7}$$

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

We can finally analyze the building blocks of Algorithm 1.

Algorithm 1: `AAKmethod`

input : A trained LM-RNN $\mathcal{M}$ of unknown rank, a target number of states k the size of the truncation $n > k$, a perturbation matrix $\mathbf{N}^n$ as in Section 3.2

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

Let $\widetilde{\mathbf{H}}^n \leftarrow$ `GetHankel(`$\mathcal{M}, n, \mathbf{N}^n$`)`

Let σ_k^n, $\boldsymbol{\xi}^n \leftarrow$ `ComputeEigenpair(`$\widetilde{\mathbf{H}}^n$`)`

Let $\mathbf{T}^n$, $\mathbf{z}^n$ defined as in Equation 7

Let $\psi(z) = \frac{(\mathbf{z}^n)^\top \mathbf{T}^n \boldsymbol{\xi}^n}{(\mathbf{z}^n)^\top \boldsymbol{\xi}^n}$

Let $r(z) \leftarrow$ `ExtractRational(`$\psi(z)$`)`

Let $\mathbf{G}_k^n \leftarrow$ `RecoverMatrix(`$r(z), k+1$`)`

Let $\widehat{A}_k^n \leftarrow$ `SpectralMethod(`$\mathbf{G}_k^n, \mathcal{B}$`)`

return $\widehat{A}_k^n$

Filling the Matrix Following Ayache et al. (2018), we consider a trained LM-RNN, and use it to fill the entries of a Hankel matrix $\mathbf{H}^n$. We obtain a $n \times n$ Hankel matrix $\mathbf{H}^n$, having entries f_n on the first n anti-diagonals, and zeroes everywhere else. As mentioned in Section 3.2, we add a perturbation to $\mathbf{H}^n$, *i.e.* a random Hankel matrix of noise $\mathbf{N}^n$, which can be set to zero when the singular numbers σ_k and σ_{k-1} of $\mathbf{H}$ are known to be distinct. The output of `GetHankel` is the perturbed matrix $\widetilde{\mathbf{H}}^n = \mathbf{H}^n + \mathbf{N}^n$.

Computing a Schmidt Pair The function `ComputeEigenpair` returns the singular number σ_k^n of $\widetilde{\mathbf{H}}^n$, and a corresponding singular vector. Since $\widetilde{\mathbf{H}}^n$ has finite rank and is symmetric, its singular numbers are the absolute values of the corresponding eigenvalues, *i.e.* $\sigma_k^n = |\lambda_k|$. Analogously, given the eigenvalue λ_k and a corresponding eigenvector $\mathbf{v}_k^n$, a Schmidt pair is given by $(\boldsymbol{\xi}^n, \boldsymbol{\eta}^n)$, with $\boldsymbol{\xi}^n = \mathbf{v}_k^n$, $\boldsymbol{\eta}^n = \mathrm{sgn}(\lambda_k)\mathbf{v}_k^n$, and $\mathrm{sgn}(\lambda_k) = \lambda_k/|\lambda_k|$.

Rational function From Theorem 6 we know that finite rank Hankel matrices correspond to strictly proper rational functions, with all the poles inside the complex unit disc. In order to find the best approximation, we apply Equation 2 from Theorem 9, and obtain a function $\psi(z) = \frac{a(z)}{b(z)}$. Note that we are interested in keeping only $r(z) = \mathbb{P}_- \psi(z)$, as $\psi(z)$ might contain poles outside the unit disc. Since the poles of $\psi(z)$ correspond to the zeros of $b(z)$, we can isolate the part of the function with poles inside the unit disc using partial fraction decomposition. This method allows us to rewrite the rational function $\psi(z) = \frac{a(z)}{b(z)}$ as:

$$\psi(z) = \frac{a(z)}{b(z)} = c(z) + \sum_i \frac{a_i(z)}{b_i(z)} \tag{8}$$

where each $\frac{a_i(z)}{b_i(z)}$ is a strictly proper rational function, and each factor b_i of the denominator is a power of an irreducible polynomial. Now, we analyze the zero of each b_i: if it is outside or on the complex unit disc, then we discard the term $\frac{a_i(z)}{b_i(z)}$. The output of `ExtractRational` is the sum of the remaining terms, corresponding to the component in $\mathcal{H}_-^2$ of $\psi(z)$. We remark that the partial fraction decomposition can be computed efficiently, with the naive implementation having complexity $O(n^3)$ for a fraction with n poles (Kung and Tong, 1977).

Recovering the Matrix In the previous step we have obtained a strictly proper rational function $r(z) = \frac{p(z)}{q(z)}$, where $p(z) = \sum_{i=1}^{k} p_i z^{k-i}$ and $q(z) = z^k + \sum_{i=1}^{k} q_i z^{k-i}$ are relatively prime, and $q(z)$ has degree k. As seen in subsection 2.4, if $r(z) = \sum_{n\geq 0} g_n z^{-n-1}$, then $\mathbf{G}_k^n(j,k) = g_{j+k}$. The coefficients g_i of the Hankel matrix can be recovered from the following set of equations, obtained from the constructive proof of Theorem 6 (Chui and Chen, 1997):

$$\begin{cases} g_0 = p_1 \\ \dots \\ g_{k-1} = p_k - g_{k-2}q_1 - \dots - g_0 q_{k-1} \end{cases} \qquad \begin{cases} g_k + \sum_{i=1}^{k} q_i g_{k-i} = 0 \\ g_{k+1} + \sum_{i=1}^{k} q_i g_{k+1-i} = 0 \\ \dots \end{cases} . \tag{9}$$

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

Extracting the WFA We can finally recover the minimal WFA $\widehat{A}_k^n = \langle \boldsymbol{\alpha}, \mathbf{A}, \boldsymbol{\beta} \rangle$ with k states realizing the function $g : \Sigma^* \to \mathbb{R}$ such that $\mathbf{G}_k^n(i,j) = g_{i+j}$. We use the spectral method outlined in subsection 2.2. The key point of the algorithm is to select a prefix-closed and complete basis $\mathcal{B}$. As noted before, since we are working with a one-letter alphabet, the Hankel matrix $\mathbf{G}$ is symmetric. In this case, if $\mathbf{G}_k^n$ has rank k, then the size of the biggest leading principal submatrix is $k \times k$ (Al'pin, 2017). Consequently, the natural choice for $\mathcal{B} = (\mathcal{P}, \mathcal{S})$ is to have $\mathcal{P} = \mathcal{S}$, with $\mathcal{P}$ containing all the strings having size strictly smaller than k. Following the notation of Section 2.2, $\mathbf{H}_\varepsilon$ corresponds to the $k \times k$ leading principal submatrix of $\mathbf{G}_k^n$, and $\mathbf{h}_{\mathcal{P},\varepsilon}$, $\mathbf{h}_{\varepsilon,\mathcal{S}}$ are its first column and row, respectively. Finally, $\mathbf{H}_a$ is the sub-block of $\mathbf{G}_k^n$ having the same rows as $\mathbf{H}_\varepsilon$, and the columns obtained by shifting each individual column of $\mathbf{H}_\varepsilon$ by one column. Using Equation 1 we obtain the WFA $\widehat{A}_k^n$.

4. Error and Convergence

If the matrix of the LM-RNN has finite rank, the unique optimal approximation of size k can be recovered, and the error, which can be computed using Gramian matrices, is given by σ_k (Balle et al., 2021). Moreover, due to the ordering of the singular numbers, the error is guaranteed to decrease when the size of the approximation gets closer to the actual rank of the matrix. On the other hand, if the rank is not finite we can only recover an asymptotically-optimal solution, and a bound for the error. As seen in Theorem 13, we can estimate how far we are from the optimal error σ_k:

$$\|\mathbf{H} - \mathbf{G}_k^n\| \leq \sigma_k + 2\left(1 - \sum_{i=0}^{n} f(i)\right). \tag{10}$$

We know that $f \in \ell^1$, so $f(n) \to 0$, meaning that "little" probability is allocated to very long strings. Thus, a direct way to reduce the error is to select the biggest possible n. An estimate for σ_k in terms of σ_k^n can be obtained using Lemma 17 in Appendix B:

$$|\sigma_k - \sigma_k^n| \leq 1 - \sum_{i=0}^{n} f(i). \tag{11}$$

An alternative way to reduce the error when additional information is available is to explore other types of truncations, to try to improve the convergence rate (Chui and Li, 1994).

If a matrix of noise $\mathbf{N}$ is added to $\mathbf{H}$ (see Section 3.2), we need to consider its effect on the error. Given the infinite matrix $\mathbf{N}$, we consider the matrix $\mathbf{N}^n$ obtained by truncation in a way analogous to Equation 4. We obtain the following bound (proof in Appendix B).

Theorem 14 *Let $\mathbf{N}^n$ be defined as above, and let $\widetilde{\mathbf{G}}_k^n$ be an asymptotically-optimal (n,k)-approximation of $\widetilde{\mathbf{H}}^n = \mathbf{H}^n + \mathbf{N}^n$. Then the error is bounded by:*

$$\|\mathbf{H} - \widetilde{\mathbf{G}}_k^n\| \leq \|\mathbf{H} - \mathbf{G}_k^n\| + 2\|\mathbf{N}^n\|. \tag{12}$$

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

Finally, as noted by Balle et al. (2021), the ℓ^2-norm is bounded by the spectral norm.

Theorem 15 *Let $f : \mathbb{N} \to \mathbb{R}$, $\mathbf{H}$ and $\mathbf{H}^n$ as before. Let $\widehat{A}_k^n$ be an asymptotically-optimal (n,k)-approximation computing $g : \mathbb{N} \to \mathbb{R}$, with matrix $\mathbf{G}_k^n$. Then: $\|f - g\|_{\ell^2} \leq \|\mathbf{H} - \mathbf{G}_k^n\|$.*

A point deserving further investigation is to understand how our approximation method performs with respect to other metrics (such as word error rate or normalized discounted cumulative gain). This could help evaluate how meaningful it is to use the spectral norm in an experimental setting, but the comparison is possible only for multi-letter alphabets.

5. Related Work

Several works in the literature analyze the relation between RNNs and WFAs. The work of Rabusseau et al. (2019) highlight a structural correspondence between WFAs and second order RNNs with linear activation function, showing that they are expressively equivalent. Weiss et al. (2019) propose a method to extract probabilistic deterministic finite automata from RNNs, based on conditional probabilities and on a local tolerance to compare observations. Analogously, Okudono et al. (2020) use spectral learning and regression methods to extract a WFA from a RNN trained on rational languages. Ayache et al. (2018) and Eyraud and Ayache (2020) propose a spectral algorithm to extract a WFA from a black box model for language modelling, without accessing the training samples.

The approximate minimization problem has been studied also for other types of models. For finite state machines, Balle et al. (2015, 2019) and Balle and Rabusseau (2020) present a technique based on the canonical expressions of weighted and weighted tree automata, respectively. Balle et al. (2021) use AAK theory to address the optimal spectral-norm approximate minimization problem for a large class of WFAs over a one-letter alphabet. The control theory community has studied this problem in the context of linear time-invariant systems (Antoulas, 2005). A first approximation algorithm is due to Kung (1980); Kung and Lin (1981), followed by state-space solutions from Glover (1984) and from Gu (2005); Ball and Ran (1987); Chui and Chen (1997) for optimal continuous and sub-optimal discrete case, respectively. The fundamental work of Chui et al. (1991, 1992); Chui and Li (1994), that provides some of the theoretical results we used, analyzes the continuity of approximation and truncation methods in signal processing.

6. Conclusion

In this paper we studied the approximate minimization problem for black boxes trained for language modelling of sequential data over a one-letter alphabet. To solve this problem, we applied the AAK theory for Hankel operators (Adamyan et al., 1971) and continuity results from the control theory literature (Chui and Li, 1994; Chui et al., 1991). This allowed us to extended the contribution of Balle et al. (2021) to the case of infinite-rank Hankel matrices. Given a language model and a target size as input, we provided an algorithm to extract a WFA corresponding to an asymptotically-optimal approximation in the spectral norm. The algorithm can be applied to black box models like RNNs or transformers.

The use of approximate minimization over regular extraction has the advantage that it allows us to choose the size of the approximation and search the optimal WFA within this constraint. This is particularly useful when the extracted WFA is used for interpretability. In fact, every WFA has a graphical representation, but this is helpful only when the number of states is small enough to actually make it readable. Moreover, approximate minimization can be used to reduce the computational cost of the task considered, as the new model is smaller and easier to compute than the original one.

While the choice of the spectral norm to evaluate the approximation deserves further investigation, we think that it constitutes an interesting way to approach the problem of approximating black boxes with WFAs. In particular, it allows us to precisely compute the distance between different classes of models, for example RNNs and WFAs, and to (asymptotically) find the optimal approximation of a given size.

The one-letter setting is certainly restrictive, but it is a first step towards developing provable approximation algorithms for black box models. In fact, it allows us to introduce AAK techniques in the context of black boxes for language modelling. The application of this rich mathematical theory has shown to be very effective in areas like control theory or signal processing, and our work highlights fruitful connections with these fields. Moreover, one-letter alphabets have proven to be of independent interest when dealing with automata, as in this case the classes of regular and of context-free languages collapse (Pighizzini, 2015).

The natural next step for future work is to extend our result to larger alphabets. This cannot be done directly, since the correspondence with Hardy spaces holds only in the one-letter case. Even though a non-commutative version of AAK theory has been recently studied (Popescu, 2003), adapting this extension to functions on sequential data remains challenging. Nonetheless, the strong theoretical foundations of this work, together with a provable algorithm for the approximate minimization problem and the possibility to compute the distance between different models, make this direction worth pursuing.

Acknowledgments

The authors would like to thank Doina Precup for supporting this work, Borja Balle for fruitful discussions on the approximate minimization problem, and Maxime Wabartha for help with the problem formulation and for a detailed feedback.

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.

Yu.A. Al'pin. The Hankel Matrix Rank Theorem Revisited. *Linear Algebra and its Applications*, 534:97–101, 2017. ISSN 0024–3795. doi: https://doi.org/10.1016/j.laa.2017.08.010. URL https://www.sciencedirect.com/science/article/pii/S002437951730486X.

Athanasios C. Antoulas. *Approximation of Large-Scale Dynamical Systems*. SIAM, 2005.

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 Andre CM. Ran. Optimal Hankel norm model reductions and Wiener–Hopf factorization I: The canonical case. *SIAM Journal on Control and Optimization*, 25 (2):362–382, 1987.

Borja Balle and Guillaume Rabusseau. Approximate minimization of weighted tree automata. *Information and Computation*, page 104654, 2020.

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. A Canonical Form for Weighted Automata and Applications to Approximate Minimization. In *30th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2015, Kyoto, Japan, July 6-10, 2015*, pages 701–712. IEEE Computer Society, 2015. doi: 10.1109/LICS.2015.70. URL https://doi.org/10.1109/LICS.2015.70.

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. *arXiv preprint arXiv:2102.06860*, 2021.

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

Charles K. Chui and Guanrong Chen. *Discrete H^∞ Optimization With Applications in Signal Processing and Control Systems*. Springer-Verlag, 1997.

Charles K. Chui and Xin Li. Continuity of Best Hankel Approximation and Convergence of Near–Best Approximants. *SIAM J. Control Optim.*, 32(6):1769—-1781, November 1994. ISSN 0363-0129. doi: 10.1137/S0363012992232245. URL https://doi.org/10.1137/S0363012992232245.

Charles K Chui, Xin Li, and Joseph D Ward. System Reduction Via Truncated Hankel Matrices. *Mathematics of Control, Signals and Systems*, 4(2):161–175, 1991.

Charles K. Chui, Xin Li, and Joseph D. Ward. Rate of Convergence of Schmidt Pairs and Rational Functions Corresponding to Best Approximants of Truncated Hankel Operators. *Math. Control. Signals Syst.*, 5(1):67–79, 1992. doi: 10.1007/BF01211976. URL https://doi.org/10.1007/BF01211976.

Corinna Cortes, Patrick Haffner, and Mehryar Mohri. Rational Kernels: Theory and Algorithms. *Journal of Machine Learning Research (JMLR)*, 5:1035–1062, 2004. URL http://www.cs.nyu.edu/~mohri/postscript/jmlr.pdf.

François Denis and Yann Esposito. On Rational Stochastic Languages. *Fundamenta Informaticae*, 86(1, 2):41–77, 2008.

Finale Doshi-Velez and Been Kim. Towards a Rigorous Science of Interpretable Machine Learning, 2017.

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.

Keith Glover. All Optimal Hankel–Nnorm Approximations of Linear Multivariable Systems and their $\mathcal{L}^{\infty}$–Error Bounds. *International Journal of Control*, 39(6):1115–1193, 1984. doi: 10.1080/00207178408933239. URL https://doi.org/10.1080/00207178408933239.

Guoxiang Gu. All Optimal Hankel–Norm Approximations and their Eerror Bounds in Discrete–Time. *International Journal of Control*, 78(6):408–423, 2005. doi: 10.1080/00207170500110988.

Christian Albert Hammerschmidt, Sicco Verwer, Qin Lin, and Radu State. Interpreting Finite Automata for Sequential Data, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

Lars Hörmander and Anders Melin. A Remark on Perturbations of Compact Operators. *Mathematica Scandinavica*, 75(2):255–262, 1994. ISSN 00255521, 19031807. URL http://www.jstor.org/stable/24491887.

Suk-Geun Hwang. Cauchy's Interlace Theorem for Eigenvalues of Hermitian Matrices. *The American Mathematical Monthly*, 111(2):157–159, 2004. doi: 10.1080/00029890.2004.11920060. URL https://doi.org/10.1080/00029890.2004.11920060.

Tosio Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.

M. Kreĭn and I. Gohberg. *Introduction to the Theory of Linear Nonselfadjoint Operators in Hilbert Space*, volume 18 of *Translations of Mathematical Monographs*. American Mathematical Society, 1969.

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

H. T. Kung and D. M. Tong. Fast Algorithms for Partial Fraction Decomposition. *SIAM J. Comput.*, 6(3):582–593, 1977. doi: 10.1137/0206042. URL https://doi.org/10.1137/0206042.

Sun-Yuan Kung. Optimal Hankel-Nnorm Model Reductions: Scalar Systems. In *Proceedings of the 1980 Joint Automation Control Conference,San Francisco, CA*, page Paper FA8.A, 1980.

Sun-Yuan Kung and David W. Lin. Optimal Hankel-Norm Model Reductions: Multivariable Systems. *IEEE Transactions Automation Control*, 26:832–852, 1981.

Reda Marzouk and Colin de la Higuera. Distance and equivalence between finite state machines and recurrent neural networks: Computational results, 2020.

William Merrill, Gail Weiss, Yoav Goldberg, Roy Schwartz, Noah A. Smith, and Eran Yahav. A formal hierarchy of RNN architectures. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 443–459. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.43. URL https://doi.org/10.18653/v1/2020.acl-main.43.

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.

G. Pighizzini. Investigations on Automata and Languages Over a Unary Alphabet. *Int. J. Found. Comput. Sci.*, 26:827–850, 2015.

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.

Frigyes Riesz and Béla Szökefalvi-Nagy. *Functional Analysis [by] Frigyes Riesz and Béla Sz.-Nagy. Translated from the 2nd French ed. by Leo F. Boron.* F. Ungar Pub. Co., New York, 1955.

Terence Tao. *Topics in Random Matrix Theory*, volume 132. American Mathematical Soc., 2012.

Terence Tao and Van Vu. Random Matrices Have Simple Spectrum, 2014.

Ananda Theertha Suresh, Brian Roark, Michael Riley, and Vlad Schogol. Approximating Probabilistic Models as Weighted Finite Automata. *arXiv e-prints*, pages arXiv–1905, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All You Need. *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

J. von Neumann and E. P. Wigner. *Über das Verhalten von Eigenwerten bei adiabatischen Prozessen*, pages 294–297. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993. ISBN 978-3-662-02781-3. doi: 10.1007/978-3-662-02781-3_20. URL https://doi.org/10.1007/978-3-662-02781-3_20.

Gail Weiss, Yoav Goldberg, and Eran Yahav. On the Practical Computational Power of Finite Precision RNNs for Language Recognition. *CoRR*, 2018. URL http://arxiv.org/abs/1805.04908.

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.

Xiyue Zhang, Xiaoning Du, Xiaofei Xie, Lei Ma, Yang Liu, and Meng Sun. Decision-Guided Weighted Automata Extraction from Recurrent Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11699–11707, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17391.

Kehe Zhu. *Operator Theory in Function Spaces*, volume 138. American Mathematical Society, 1990.

Appendix A. Example

In this section we show an illustrative example, analogous to the one presented by Balle et al. (2021).

We consider the function $f : \mathbb{N} \to \mathbb{R}$, computing a probability, defined as:

$$f(k) = \begin{cases} 0 & \text{if } k \text{ is odd} \\ \frac{8}{9}3^{-k} & \text{if } k \text{ is even} \end{cases}$$

The corresponding Hankel matrix is:

$$\mathbf{H} = \begin{pmatrix} f(0) & f(1) & f(2) & \dots \\ f(1) & f(2) & f(3) & \dots \\ f(2) & f(3) & f(4) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} \frac{8}{9} & 0 & \frac{8}{81} & \dots \\ 0 & \frac{8}{81} & 0 & \dots \\ \frac{8}{81} & 0 & \frac{8}{729} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{13}$$

If we consider the Hankel matrix with respect to the basis of the Hardy space, since $\mathbf{H}(j,k) = \widehat{\phi}(-j-k-1)$, we have:

$$\mathbf{H} = \begin{pmatrix} \frac{8}{9} & 0 & \frac{8}{81} & \dots \\ 0 & \frac{8}{81} & 0 & \dots \\ \frac{8}{81} & 0 & \frac{8}{729} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} \widehat{\phi}(-1) & \widehat{\phi}(-2) & \widehat{\phi}(-3) & \dots \\ \widehat{\phi}(-2) & \widehat{\phi}(-3) & \widehat{\phi}(-4) & \dots \\ \widehat{\phi}(-3) & \widehat{\phi}(-4) & \widehat{\phi}(-5) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

and the rational component of a symbol for $\mathbf{H}$ is:

$$\mathbb{P}_- \phi = \sum_{n \geq 0} \widehat{\phi}(-n-1) z^{-n-1} = \sum_{n \geq 0} \frac{8}{9} 9^{-n} z^{-2n-1} = \frac{8z}{9z^2 - 1}.$$

Appendix B. Proofs

Proof of Theorem 12

We briefly recall the following theorem, due to Nehari (1957), which will be used in the proof of Theorem 12.

Theorem 16 (Nehari (1957)) *Let $\phi \in \mathcal{L}^2(\mathbb{T})$ be a symbol of the Hankel operator on Hardy spaces $H_\phi : \mathcal{H}^2 \to \mathcal{H}^2_-$. Then, H_ϕ is bounded on $\mathcal{H}^2$ if and only if there exists $\psi \in \mathcal{L}^\infty(\mathbb{T})$ such that $\widehat{\psi}(m) = \widehat{\phi}(m)$ for all $m < 0$. If the conditions above are satisfied, then:*

$$\|H_\phi\| = \inf\{\|\psi\|_\infty : \widehat{\psi}(m) = \widehat{\phi}(m),\, m < 0\}. \tag{14}$$

We can now prove Theorem 12.

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

$$\|H - H^t\| \leq \left\| \sum_{i=0}^{\infty} f(i) z^{-i-1} - \sum_{i=0}^{t} f(i) z^{-i-1} \right\|_\infty \leq \left\| \sum_{i=t+1}^{\infty} f(i) z^{-i-1} \right\|_\infty \leq \sum_{i=t+1}^{\infty} |f(i)| \tag{15}$$

where the first inequality follows from Theorem 16. Since the black box is trained for language modelling, we have that $\sum_{k\geq 0}|f(k)| = 1$. Thus, $f \in \ell^1$, and it follows directly that $H^t \to H$. ■

Proof of Theorem 13

We recall the following result from Riesz and Szökefalvi-Nagy (1955) (see Kreĭn and Gohberg (1969) for the proof and for a more general version of this theorem), as it constitutes a fundamental step in the proof of Theorem 13.

Lemma 17 (Riesz and Szökefalvi-Nagy (1955)) *Let T, S be two self-adjoint compact operators, and let σ_k^T, σ_k^S for $k \geq 0$ be their singular numbers. Then:*

$$|\sigma_k^T - \sigma_k^S| \leq \|\mathbf{S} - \mathbf{T}\|. \tag{16}$$

We can now prove Theorem 13.

Proof Let σ_k^n be the singular number $k+1$ of the operator H^n, and let $\mathbf{G}_k^n$ be the optimal approximation described by Theorem 8, *i.e.* $\|\mathbf{H}^n - \mathbf{G}_k^n\| = \sigma_k^n$. We have:

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

From Theorem 3 we know that $\|\mathbf{H} - \mathbf{G}_k^n\| \geq \sigma_k^n$. On the other hand, using Lemma 17 and Cauchy's interlace theorem (Hwang, 2004), we obtain $\sigma_k^n \leq \sigma_k + \|\mathbf{H} - \mathbf{H}^n\|$. It follows that:

$$\sigma_k \leq \|\mathbf{H} - \mathbf{G}_k^n\| \leq \sigma_k + 2\|\mathbf{H} - \mathbf{H}^n\|. \tag{17}$$

Now, $\mathbf{H}^n$ belongs to the sequence of truncating matrices $\{\mathbf{H}^t\}_{t\geq 0}$, and $H^t \to H$ (Theorem 12). Since $\sigma_k \neq \sigma_{k-1}$, the conditions of Theorem 11 hold. Therefore, the sequence of matrices of best approximations $\{\mathbf{G}_k^t\}_{t\geq 0}$ is an asymptotic sequence for $\mathbf{G}_k$, and $\mathbf{G}_n^k$ belongs to it. Thus, the WFA having matrix $\mathbf{G}_n^k$ is an asymptotically-optimal (n,k)-approximation, and Equation 3 holds. Moreover, from Equation 15, we have:

$$\|\mathbf{H} - \mathbf{G}_k^n\| \leq \sigma_k + 2\|\mathbf{H} - \mathbf{H}^n\| \leq \sigma_k + 2\sum_{i=n+1}^{\infty} f(i) = \sigma_k + 2\left(1 - \sum_{i=0}^{n} f(i)\right). \tag{18}$$

■

Proof of Theorem 14

Proof Let $\widetilde{\mathbf{G}}_k^n$ and $\widetilde{\sigma}_k^n$ be the optimal approximation and the $(k+1)$-th singular number of $\mathbf{H}^n + \mathbf{N}^n$, respectively. From Theorem 8 we have:

$$\|\mathbf{H}^n + \mathbf{N}^n - \widetilde{\mathbf{G}}_k^n\| = \widetilde{\sigma}_k^n. \tag{19}$$

Then:

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

where we used Equation 19 for the second step, and we used Lemma 17 for the last two (first with σ_k^n and $\widetilde{\sigma}_k^n$, then with σ_k^n and σ_k). ■

Proof of Theorem 15

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

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

The second equation follows by definition and by observing that matrix difference is always computed entry-wise, while the last inequality is a consequence of Equation 5. ■