Journal of Machine Learning Research 17 (2016) 1-32 Submitted 12/14; Revised 6/15; Published 4/16

Dimension-free Concentration Bounds on Hankel Matrices for Spectral Learning

François Denis FRANCOIS.DENIS@LIF.UNIV-MRS.FR
Mattias Gybels MATTIAS.GYBELS@LIF.UNIV-MRS.FR
Aix Marseille Université, CNRS
LIF UMR 7279
13288 Marseille Cedex 9, FRANCE

Amaury Habrard AMAURY.HABRARD@UNIV-ST-ETIENNE.FR
Université de Lyon, UJM-Saint-Etienne, CNRS, UMR 5516
Laboratoire Hubert Curien
F-42023 Saint-Etienne, FRANCE

Editor: Mehryar Mohri

Abstract

Learning probabilistic models over strings is an important issue for many applications. Spectral methods propose elegant solutions to the problem of inferring weighted automata from finite samples of variable-length strings drawn from an unknown target distribution p. These methods rely on a singular value decomposition of a matrix $\mathbf{H}_S$, called the empirical Hankel matrix, that records the frequencies of (some of) the observed strings S. The accuracy of the learned distribution depends both on the quantity of information embedded in $\mathbf{H}_S$ and on the distance between $\mathbf{H}_S$ and its mean $\mathbf{H}_p$. Existing concentration bounds seem to indicate that the concentration over $\mathbf{H}_p$ gets looser with its dimensions, suggesting that it might be necessary to bound the dimensions of $\mathbf{H}_S$ for learning. We prove new *dimension-free* concentration bounds for classical Hankel matrices and several variants, based on prefixes or factors of strings, that are useful for learning. Experiments demonstrate that these bounds are tight and that they significantly improve existing (dimension-dependent) bounds. One consequence of these results is that the spectral learning approach remains consistent even if all the observations are recorded within the empirical matrix.

Keywords: Hankel matrices, Matrix Bernstein bounds, Probabilistic Grammatical Inference, Rational series, Spectral learning

1. Introduction

Many applications in natural language processing, text analysis or computational biology require learning probabilistic models over finite variable-size strings such as probabilistic automata, Hidden Markov Models (HMM), or more generally, weighted automata. Weighted automata exactly model the class of rational series, and their algebraic properties have been widely studied in that context (Droste et al., 2009). In particular, they admit algebraic representations that can be characterized by a set of finite-dimensional linear operators whose ranks are closely linked to the minimum number of states needed to define the automaton. From a machine learning perspective, the objective is then to infer good estimates of these linear operators from finite samples. In this paper, we consider the problem of

learning the linear representation of a weighted automaton, from a finite sample, composed of variable-size strings i.i.d. from an unknown target distribution.

Recently, the seminal papers of (Hsu et al., 2009) for learning HMM and (Bailly et al., 2009) for weighted automata, have defined a new category of approaches - the so-called *spectral methods* - for learning distributions over strings represented by finite state models (Siddiqi et al., 2010; Song et al., 2010; Balle et al., 2012; Balle and Mohri, 2012). Extensions to probabilistic models for tree-structured data (Bailly et al., 2010; Parikh et al., 2011; Cohen et al., 2012), transductions (Balle et al., 2011) or other graphical models (Anandkumar et al., 2012c,b,a; Luque et al., 2012) have also attracted a lot of interest.

Spectral methods suppose that the main parameters of a model can be expressed as the spectrum of a linear operator and estimated from the spectral decomposition of a matrix that sums up the observations. Given a rational series r, the values taken by r can be arranged in a matrix $\mathbf{H}_r$ whose rows and columns are indexed by strings, such that the linear operators defining r can be recovered directly from the right singular vectors of $\mathbf{H}_r$. This matrix is called the Hankel matrix of r.

In a learning context, given a learning sample S drawn from a target distribution p, an empirical estimate $\mathbf{H}_S$ of $\mathbf{H}_p$ is built and then, a rational series $\tilde{p}$ is inferred from the right singular vectors of $\mathbf{H}_S$. However, the size of $\mathbf{H}_S$ increases drastically with the size of S and state of the art approaches consider smaller matrices $\mathbf{H}_S^{U,V}$ indexed by limited subset of strings U and V. It can be shown that the above learning scheme, or slight variants of it, are consistent as soon as the matrix $\mathbf{H}_S^{U,V}$ has full rank (Hsu et al., 2009; Bailly, 2011; Balle et al., 2012) and that the accuracy of the inferred series is directly connected to the concentration distance $\|\mathbf{H}_S^{U,V} - \mathbf{H}_p^{U,V}\|_2$ between the empirical Hankel matrix and its mean (Hsu et al., 2009; Bailly, 2011).

On the one hand, limiting the size of the Hankel matrix avoids prohibitive calculations. Moreover, most existing concentration bounds on sum of random matrices depend on their size and suggest that $\|\mathbf{H}_S^{U,V} - \mathbf{H}_p^{U,V}\|_2$ may become significantly looser with the size of U and V, compromising the accuracy of the inferred model.

On the other hand, limiting the size of the Hankel matrix implies a drastic loss of information: only the strings of S compatible with U and V will be considered. In order to limit the loss of information when dealing with restricted sets U and V, a general trend is to work with other functions than the target p, such as the *prefix* function $\overline{p}(u) \coloneqq \sum_{v\in\Sigma^*} p(uv)$ or the *factor* function $\widehat{p} \coloneqq \sum_{v,w\in\Sigma^*} p(vuw)$ (Balle et al., 2013; Luque et al., 2012). These functions are rational, they have the same rank as p, a representation of p can easily be derived from representations of $\overline{p}$ or $\widehat{p}$ and they allow a better use of the information contained in the learning sample.

A first contribution of this paper is to provide a *dimension free* concentration inequality for $\|\mathbf{H}_S^{U,V} - \mathbf{H}_p^{U,V}\|_2$, by using recent results on tail inequalities for sum of random matrices (Tropp, 2012), and in particular a dimension-free Matrix Bernstein Bound Theorem stated in (Hsu et al., 2011). As a consequence, the spectral learning approach is consistent whatever sets U and V are chosen, and even if they are set to Σ^*, showing that restricting the dimensions of $\mathbf{H}$ is not mandatory.

However, this Matrix Bernstein Bound Theorem cannot be directly applied as such to the prefix and factor series, since the norm of the corresponding random matrices is unbounded. A second contribution of the paper is then to define two classes of parametrized functions,

$\overline{p}_\eta$ and $\widehat{p}_\eta$, that constitute continuous intermediates between p and $\overline{p}$ (resp. p and $\widehat{p}$), and to provide analogous dimension-free concentration bounds for these two classes. Lastly, we adapt a Matrix Bernstein bound theorem for subexponential matrices from (Tropp, 2012) to the dimension free case, using a technique similar as the one used in (Hsu et al., 2011) and we apply it to the prefix Hankel matrices.

These bounds are evaluated on a benchmark made of 11 problems extracted from the PAutomaC challenge (Verwer et al., 2012). These experiments show that the bounds derived from our theoretical results for bounded random matrices are quite tight - compared to the exact values - and that they significantly improve existing bounds, even on matrices of fixed dimensions. By contrast, the bounds obtained in the subexponential case are somewhat loose.

Our theoretical results entail that spectral learning is consistent whatever dimensions of the Hankel matrix are chosen but they give no indication on what should be done in practical cases. We have computed the distance between the spaces spanned by the first right singular vectors of $\mathbf{H}_S^{U,V}$ and $\mathbf{H}_p^{U,V}$ for various sizes of U and V, for each target of our benchmark. These experiments seem to indicate that the best results are obtained by limiting one dimension and taking the other as large as possible but a theoretical justification remains to be provided.

The paper is organized as follows. Section 2 introduces the main notations, definitions and concepts. Section 3 provides some Matrix Berstein bounds that will be used to prove the different results of the paper. Section 4.1 presents a first dimension free-concentration inequality for the standard Hankel matrices. Then, we introduce the prefix and the factor variants and provide analogous concentration results in Sections 4.3 and 4.5 respectively. Section 6 describes some experiments before the conclusion presented in Section 7. The Appendix contains the proof of an original result, which states that the series $u \mapsto p(\Sigma^* u \Sigma^*)$, i.e. the probability that a random string contains a substring u as a factor, may be not rational even if p is rational, explaining why we have considered the less natural series $\widehat{p}$. It also contains two small proofs of known results, in order to keep the paper self-contained.

2. Preliminaries

We first present some preliminary definitions and results about matrices, rational languages, Hankel matrices and spectral learning algorithms for the inference of rational stochastic languages.

2.1 Matrices

The identity matrix of size n is denoted by $\mathbf{I}_n$, or simply by $\mathbf{I}$. Let $\mathbf{M} \in \mathbb{R}^{m \times n}$ be a $m \times n$ real matrix. The *singular values* of $\mathbf{M}$ are the square roots of the eigenvalues of the matrix $\mathbf{M}^\top \mathbf{M}$, where $\mathbf{M}^\top$ denotes the transpose of $\mathbf{M}$: $\sigma_{\max}(\mathbf{M})$ and $\sigma_{\min}(\mathbf{M})$ denote the largest and smallest singular value of $\mathbf{M}$, respectively. The *spectral radius* $\rho(\mathbf{M})$ of a square matrix $\mathbf{M}$ is the supremum among the modulus of the eigenvalues of $\mathbf{M}$. If $\mathbf{M}$ is symmetric, $\sigma_{\max}(\mathbf{M})$ coincides with $\rho(\mathbf{M})$.

Every rank-d matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ admits a factorization of the form $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, called a *reduced singular value decomposition* (SVD), where $\mathbf{U} \in \mathbb{R}^{m \times d}$ and $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_d$, $\mathbf{V} \in \mathbb{R}^{n \times d}$ and $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_d$ and $\mathbf{D}$ is a diagonal matrix whose diagonal elements, listed in descending order,

are the singular values of $\mathbf{M}$. The columns of $\mathbf{U}$ (resp. of $\mathbf{V}$) are called the right-singular vectors of $\mathbf{M}$ (resp. left-singular vectors of $\mathbf{M}$).

The notion of singular values, singular vectors and singular value decomposition can be extended to infinite matrices via the notion of *Hilbert spaces compact operators* (see (Stein and Shakarchi, 2005) for example). Let $(e_i)_{i\in\mathbb{N}}$ be an orthonormal basis of a separable Hilbert space $\mathcal{H}$. A bounded operator $\mathbf{T}$ on $\mathcal{H}$ can be represented by the matrix $(\langle \mathbf{T}(e_i), e_j\rangle_{\mathcal{H}})_{i,j\in\mathbb{N}}$. Compact operators are the closure of finite-rank operators in the uniform operator topology: $\max_{\|x\|=1} \|\mathbf{T}_n(x) - \mathbf{T}(x)\| \to 0$. A sufficient condition for a matrix $\mathbf{M}$ to represent a compact operator is that it has a finite Frobenius norm: $\sum_{i,j\in\mathbb{N}} M[i,j]^2 < \infty$. The matrix of any compact operator admits a reduced SVD. In particular, if $\mathbf{M}$ is the matrix of a finite rank bounded operator, it admits a reduced singular value decomposition $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$.

The *operator norm* $\|\cdot\|_k$ induced by the corresponding vector norm on $\mathbb{R}^n$ is defined by $\|\mathbf{M}\|_k \coloneqq \max_{x\neq 0} \frac{\|\mathbf{M}x\|_k}{\|x\|_k}$. It can be shown that

$$\|\mathbf{M}\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |\mathbf{M}[i,j]|, \|\mathbf{M}\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |\mathbf{M}[i,j]| \text{ and } \|\mathbf{M}\|_2 = \sigma_{\max}(\mathbf{M}).$$

We will mainly use the *spectral norm* $\|\cdot\|_2$ and we will omit the sub index 2 for the sake of simplicity. It can be shown that

$$\|\mathbf{M}\| \leq \sqrt{\|\mathbf{M}\|_1 \|\mathbf{M}\|_\infty} \tag{1}$$

These norms can be extended, under certain conditions, to infinite matrices. For example, the previous inequality remains true (with possibly infinite right-hand side term) if $\mathbf{M}$ represents the matrix of a compact operator in an orthonormal basis of a separable Hilbert space.

A symmetric matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ is *positive semidefinite* if $u^\top \mathbf{M} u \geq 0$ for all vectors $u \in \mathbb{R}^n$. Let $\preccurlyeq$ denotes the *positive semidefinite ordering* (or *Löwner ordering*) on symmetric matrices: $\mathbf{A} \preccurlyeq \mathbf{B}$ means that $\mathbf{B} - \mathbf{A}$ is positive semidefinite. The family of positive semidefinite matrices in $\mathbb{R}^{n\times n}$ forms a convex closed cone.

Any real valued function can be extended to symmetric matrices by the following method: let $\mathbf{A} = \mathbf{U}^\top \mathrm{diag}(\lambda_1, \ldots, \lambda_n)\mathbf{U}$ where $\mathrm{diag}(x_1, \ldots, x_n)$ is the diagonal matrix built over $x_1, \ldots, x_n$ and where $\mathbf{U} \in \mathbb{R}^{n\times n}$ is unitary, i.e. $\mathbf{U}^\top\mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}$; define the matrix $f(\mathbf{A})$ by $f(\mathbf{A}) \coloneqq \mathbf{U}^\top \mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_n))\mathbf{U}$. It can be shown that this definition is independent of the chosen eigenvalue decomposition. The *transfer rule* states that $f \leq g$ implies that $f(\mathbf{A}) \preccurlyeq g(\mathbf{A})$ for any symmetric matrix $\mathbf{A}$. The definition above can be used to define the exponential $e^{\mathbf{A}}$ of a symmetric matrix $\mathbf{A}$ and the logarithm $\log \mathbf{B}$ of a positive semidefinite matrix $\mathbf{B}$. It can be shown that the logarithm preserves the semidefinite order: $\mathbf{0} \preccurlyeq \mathbf{A} \preccurlyeq \mathbf{B}$ implies $\log \mathbf{A} \preccurlyeq \log \mathbf{B}$. See (Tropp, 2012) for a short overview of matrix properties.

2.2 Rational stochastic languages and Hankel matrices

Most classical results on rational series can be found in one of the following references (Berstel and Reutenauer, 1988; Salomaa and Soittola, 1978). Let Σ be a finite alphabet. The set of all finite strings over Σ is denoted by Σ^*, the empty string is denoted by ϵ, the length of

string w is denoted by $|w|$ and Σ^n (resp. $\Sigma^{\leq n}$, resp. $\Sigma^{\geq n}$) denotes the set of all strings of length n (resp. $\leq n$, resp. $\geq n$). For any string w, let $\mathrm{Pref}(w) \coloneqq \{u \in \Sigma^* | \exists v \in \Sigma^* \quad w = uv\}$ and $\mathrm{Suff}(w) \coloneqq \{v \in \Sigma^* | \exists u \in \Sigma^* \quad w = uv\}$.

A *series* is a mapping $r : \Sigma^* \to \mathbb{R}$. The *support* of the series r is the set $\mathrm{supp}(r) = \{u \in \Sigma^* : r(u) \neq 0\}$. A series r is *non negative* if it takes only non negative values. A non negative series r is *convergent* if the sum $\sum_{u \in \Sigma^*} r(u)$ is bounded: for any $A \subseteq \Sigma^*$, let us denote $r(A) \coloneqq \sum_{u \in A} r(u)$. A *stochastic language* p is a probability distribution over Σ^*, i.e. a non negative series p satisfying $p(\Sigma^*) = 1$.

Let $n \geq 1$ and $\mathbf{M}$ be a morphism defined from Σ^* to $\mathbb{M}_n$, the set of square $n \times n$ matrices with real coefficients. For all $u \in \Sigma^*$, let us denote $\mathbf{M}(u)$ by $\mathbf{M}_u$ and $\Sigma_{x \in \Sigma} \mathbf{M}_x$ by $\mathbf{M}_\Sigma$. A series r over Σ is *rational* if there exists an integer $n \geq 1$, two vectors $I, T \in \mathbb{R}^n$ and a morphism $\mathbf{M} : \Sigma^* \to \mathbb{M}_n$ such that for all $u \in \Sigma^*$, $r(u) = I^\top \mathbf{M}_u T$. The triplet $\langle I, \mathbf{M}, T \rangle$ is called an n-dimensional *linear representation* of r. The vector I can be interpreted as a vector of initial weights, T as a vector of terminal weights and the morphism $\mathbf{M}$ as a set of matrix parameters associated with the letters of Σ. A *rational stochastic language* is thus a stochastic language admitting a linear representation.

Let $U, V \subseteq \Sigma^*$, the *Hankel matrix* $\mathbf{H}_r^{U,V}$, associated with a series r, is the matrix indexed by $U \times V$ and defined by $\mathbf{H}_r^{U,V}[u, v] \coloneqq r(uv)$, for any $(u, v) \in U \times V$. If $U = V = \Sigma^*$, $\mathbf{H}_r^{U,V}$, simply denoted by $\mathbf{H}_r$, is a bi-infinite matrix. In the following, we always assume that $\epsilon \in U \cap V$ and that U and V are ordered in quasi-lexicographic order: strings are first ordered by increasing length and then, according to the lexicographic order. It can be shown that a series r is rational if and only if the rank of the matrix $\mathbf{H}_r$ is finite. The rank of $\mathbf{H}_r$ is equal to the minimal dimension of a linear representation of r: it is called the rank of r. The Hankel matrix $\mathbf{H}_r$ represents a bounded operator if and only if $\sum_{u \in \Sigma^*} r^2(u) < \infty$; in particular, if r is a non negative convergent rational series, then $\mathbf{H}_r$ represents a compact operator, which admits a reduced singular value decomposition.

Let r be a non negative convergent rational series and let $\langle I, \mathbf{M}, T \rangle$ be a minimal d-dimensional linear representation of r. Then, the matrix $\mathbf{I}_d - \mathbf{M}_\Sigma$ is invertible and the sum $\mathbf{I}_d + \mathbf{M}_\Sigma + \ldots + \mathbf{M}_\Sigma^n + \ldots$ converges to $(\mathbf{I}_d - \mathbf{M}_\Sigma)^{-1}$. For any ρ_r such that $\rho(\mathbf{M}_\Sigma) < \rho_r < 1$, there exists a constant $C_r > 0$ such that $r(\Sigma^{\geq n}) \leq C_r \rho_r^n$ for any integer n (we show in Section 6.1 how such constants can be computed in pratical cases). For any integer $k \geq 1$, let us define the moments $S_r^{(k)} \coloneqq \sum_{u_1 u_2 \ldots u_k \in \Sigma^*} r(u_1 u_2 \ldots u_k)$. It can easily be shown that

$$S_r^{(k)} = I^\top (\mathbf{I}_d - \mathbf{M}_\Sigma)^{-k} T. \tag{2}$$

Several rational series can be naturally associated with a rational non negative convergent series r (see (Balle et al., 2014) for example):

- $\overline{r}$, defined by $\overline{r}(u) \coloneqq \sum_{v \in \Sigma^*} r(uv) = r(u\Sigma^*)$, associated with the *prefixes* of the support of r,
- $\widehat{r}$, defined by $\hat{r}(u) \coloneqq \sum_{v,w \in \Sigma^*} r(vuw)$, associated with the *factors* of the support of r.

If p is a stochastic language, it can be noticed that $\overline{p}(u)$ is the probability that a string begins with u and that $\widehat{p}(u) = \mathbb{E}_{v \sim p} |v|_u$, where $|v|_u = \sum_{x,y \in \Sigma^*} \mathbf{1}_{xuy = v}$. We have $\widehat{p}(u) \geq p(\Sigma^* u \Sigma^*)$, the probability that a string contains u as a substring. The function $u \mapsto p(\Sigma^* u \Sigma^*)$ has a

simpler probabilistic interpretation than $\widehat{p}$. However, this function is not rational in general and cannot easily be used in a learning context.

Proposition 1 *There exists a rational stochastic language p of rank one and built on a two-letter alphabet Σ such that the series $u \mapsto p(\Sigma^* u \Sigma^*)$ is not rational.*

Proof See Appendix. ■

If $\langle I, \mathbf{M}, T\rangle$ is a minimal d-dimensional linear representation of r, then $\langle I, \mathbf{M}, (\mathbf{I}_d - \mathbf{M}_\Sigma)^{-1}T\rangle$ (resp. $\langle [I^\top(\mathbf{I}_d - \mathbf{M}_\Sigma)^{-1}]^\top, \mathbf{M}, (I_d - \mathbf{M}_\Sigma)^{-1}T\rangle$) is a minimal linear representation of $\overline{r}$ (resp. of $\widehat{r}$). Conversely, a linear representation of r can be deduced from any linear representation of $\overline{r}$ or of $\widehat{r}$. Clearly,

$$r(\Sigma^*) = S_r^{(1)}, \overline{r}(\Sigma^*) = S_r^{(2)} \text{ and } \widehat{r}(\Sigma^*) = S_r^{(3)}.$$

Let $U, V \subseteq \Sigma^*$. For any string $w \in \Sigma^*$, let us define the matrices $\mathbf{H}_w^{U,V}$, $\overline{\mathbf{H}}_w^{U,V}$ and $\widehat{\mathbf{H}}_w^{U,V}$ by

$$\mathbf{H}_w^{U,V}[u,v] \coloneqq \mathbf{1}_{uv=w}, \overline{\mathbf{H}}_w^{U,V}[u,v] \coloneqq \mathbf{1}_{uv\in \mathrm{Pref}(w)} \text{ and } \widehat{\mathbf{H}}_w^{U,V}[u,v] \coloneqq \sum_{x,y\in\Sigma^*} \mathbf{1}_{xuvy=w}$$

for any $(u,v) \in U \times V$.

For any non empty multiset of strings S, let us define the matrices $\mathbf{H}_S^{U,V}$, $\overline{\mathbf{H}}_S^{U,V}$ and $\widehat{\mathbf{H}}_S^{U,V}$ by

$$\mathbf{H}_S^{U,V} \coloneqq \frac{1}{|S|}\sum_{w\in S} \mathbf{H}_w^{U,V}, \overline{\mathbf{H}}_S^{U,V} \coloneqq \frac{1}{|S|}\sum_{w\in S} \overline{\mathbf{H}}_w^{U,V} \text{ and } \widehat{\mathbf{H}}_S^{U,V} \coloneqq \frac{1}{|S|}\sum_{w\in S} \widehat{\mathbf{H}}_w^{U,V}.$$

Let p_S be the empirical stochastic language associated with S, defined by $p_S(u) \coloneqq \frac{|\{u\in S\}|}{|S|}$. We have

$$\mathbf{H}_{p_S}^{U,V} = \mathbf{H}_S^{U,V}, \mathbf{H}_{\overline{p_S}}^{U,V} = \overline{\mathbf{H}}_S^{U,V} \text{ and } \mathbf{H}_{\widehat{p_S}}^{U,V} = \widehat{\mathbf{H}}_S^{U,V}.$$

For example, let $S = \{a, ab\}$, $U = V = \{\epsilon, a, b\}$. We have

$$\mathbf{H}_S^{U,V} = \begin{pmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 0 & 0 \end{pmatrix}, \quad \overline{\mathbf{H}}_S^{U,V} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1/2 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } \widehat{\mathbf{H}}_S^{U,V} = \begin{pmatrix} 5/2 & 1 & 1/2 \\ 1 & 0 & 1/2 \\ 1/2 & 0 & 0 \end{pmatrix}.$$

2.3 Spectral Algorithm for Learning Rational Stochastic Languages

Rational series admit *canonical linear representations* determined by their Hankel matrix. Let r be a rational series of rank d and $U \subset \Sigma^*$ such that the matrix $\mathbf{H}_r^{U\times\Sigma^*}$ (denoted by $\mathbf{H}$ in the following) has rank d. Moreover, suppose that $\sum_{u\in\Sigma^*} r(u)^2 < \infty$.

- For any string s, let $\mathbf{T}_s$ be the constant matrix whose rows and columns are indexed by Σ^* and defined by $\mathbf{T}_s[u,v] \coloneqq 1$ if $v = us$ and 0 otherwise.
- Let E be a vector indexed by Σ^* whose coordinates are all zero except the first one equals to 1: $E[u] = \mathbf{1}_{u=\epsilon}$ and let P be the vector indexed by Σ^* defined by $P[u] \coloneqq r(u)$.

- Let $\mathbf{H} = \mathbf{L}\mathbf{D}\mathbf{R}^\top$ be a reduced singular value decomposition of $\mathbf{H}$: $\mathbf{R}$ (resp. $\mathbf{L}$) is a matrix whose columns form a set of orthonormal vectors - the right (resp. left) singular vectors of $\mathbf{H}$ - and $\mathbf{D}$ is a $d \times d$ diagonal matrix, composed of the singular values of $\mathbf{H}$.

Then, $\langle \mathbf{R}^\top E, (\mathbf{R}^\top \mathbf{T}_x \mathbf{R})_{x\in\Sigma}, \mathbf{R}^\top P\rangle$ is a linear representation of r (Bailly et al., 2009; Hsu et al., 2009; Bailly, 2011; Balle et al., 2012).

Proposition 2 *$\langle \mathbf{R}^\top E, (\mathbf{R}^\top \mathbf{T}_x \mathbf{R})_{x\in\Sigma}, \mathbf{R}^\top P\rangle$ is a linear representation of r*

Proof See Appendix. ■

The basic spectral algorithm for learning rational stochastic languages aims at identifying the canonical linear representation of the target p determined by its Hankel matrix $\mathbf{H}_p$.

Let S be a sample independently drawn according to p:

- choose sets $U, V \subseteq \Sigma^*$ and build the Hankel matrix $\mathbf{H}_S^{U\times V}$,
- choose a rank d, compute a SVD of $\mathbf{H}_S^{U\times V}$, and consider the d right singular vectors $\mathbf{R}_S$ associated with the d largest singular values,
- build the canonical linear representation $\langle \mathbf{R}_S^\top E, (\mathbf{R}_S^\top \mathbf{T}_x \mathbf{R}_S)_{x\in\Sigma}, \mathbf{R}_S^\top P_S\rangle$ where E and P are the vectors indexed by V s.t. $E[v] = \mathbf{1}_{v=\epsilon}$ and $P[v] \coloneqq p_S(v)$.

Alternative learning strategies consist in learning $\overline{p}$ or $\widehat{p}$, using the same algorithm, and then to compute an estimate of p. In all cases, the accuracy of the learned representation mainly depends on the estimation of $\mathbf{R}$. The Stewart formula (Stewart, 1990) bounds the principle angle θ between the spaces spanned by the right singular vectors of $\mathbf{R}$ and $\mathbf{R}_S$:

$$|\sin(\theta)| \leq \frac{\left\|\mathbf{H}_S^{U\times V} - \mathbf{H}_r^{U\times V}\right\|}{\sigma_{\min}(\mathbf{H}_r^{U\times V})}.$$

According to this formula, the concentration of the Hankel matrix around its mean is critical and the question of limiting the sizes of U and V naturally arises. Note that the Stewart inequality does not give any clear indication on the impact or on the interest of limiting these sets. Indeed, it can be shown that both the numerator and the denominator of the right part of the inequality increase with U and V (see Appendix).

3. Matrix Bernstein bounds

Let p be a rational stochastic language over Σ^*, let ξ be a random variable distributed according to p, let $U, V \subseteq \Sigma^*$ and let $\mathbf{Z}(\xi) \in \mathbb{R}^{|U|\times|V|}$ be a random matrix. For instance, $\mathbf{Z}(\xi)$ may be equal to $\mathbf{H}_\xi^{U,V}$, $\overline{\mathbf{H}}_\xi^{U,V}$ or $\widehat{\mathbf{H}}_\xi^{U,V}$ (ξ will be often omitted for the sake of simplicity). Let S be sample of strings drawn independently according to p.

Concentration bounds for sum of random matrices can be used to estimate the spectral distance between the empirical matrix $\mathbf{Z}_S$ computed on the sample S and its mean.

However, most of classical inequalities depend on the dimensions of the matrices. For eample, the following result describes a simple matrix Bernstein inequality on sum of random matrices (Ahlswede and Winter, 2002; Tropp, 2012).

Suppose that $\mathbb{E}\mathbf{Z}(\xi) = 0$ and let $\nu(\mathbf{Z}) = \max\{\|\mathbb{E}\mathbf{Z}\mathbf{Z}^\top\|, \|\mathbb{E}\mathbf{Z}^\top\mathbf{Z}\|\}$. Then,

$$\Pr\left(\|\mathbf{Z}_S\| \geq \frac{t}{N}\right) \leq (d_1 + d_2)\ \exp\left(-\frac{t^2}{2N\nu(\mathbf{Z}) + 2Mt/3}\right) \tag{3}$$

where N is the size of S, d_1 and d_2 are the dimensions of the matrix $\mathbf{Z}$ and $\|\mathbf{Z}\| \leq M$ almost surely.

We would like to apply this result to $\mathbf{Z} = \mathbf{H}_\xi^{U,V} - \mathbb{E}\mathbf{H}_\xi^{U,V}$, $\mathbf{Z} = \overline{\mathbf{H}}_\xi^{U,V} - \mathbb{E}\overline{\mathbf{H}}_\xi^{U,V}$ and $\mathbf{Z} = \widehat{\mathbf{H}}_\xi^{U,V} - \mathbb{E}\widehat{\mathbf{H}}_\xi^{U,V}$. However, we will see that while $\left\|\mathbf{H}_\xi^{U,V}\right\|$ is bound, $\left\|\overline{\mathbf{H}}_\xi^{U,V}\right\| = \Omega(\max(|U|^{1/2}, |V|^{1/2}))$ in the worst case, and $\left\|\widehat{\mathbf{H}}_\xi^{U,V}\right\|$may be unbounded even for fixed U and V.

These concentration bounds get worse with both sizes of the matrices. Coming back to the discussion at the end of Section 2, they suggest to limit the size of the sets U and V, and therefore, to design strategies to choose optimal sets. However, dimension-free bounds can be obtained.

3.1 A dimension-free Matrix Bernstein bound theorem

We then use recent results from (Tropp, 2012; Hsu et al., 2012) to obtain dimension-free concentration bounds for Hankel matrices.

Theorem 3 *(Hsu et al., 2012). Let $\xi_1, \ldots, \xi_N$ be random variables, and for each $i = 1, \ldots, N$, let $\mathbf{X}_i(\xi_i)$ be a symmetric matrix-valued functional of ξ_i[1]. For any $\eta \in \mathbb{R}$ and any $t > 0$,*

$$Pr\left[\left\|\eta \sum_{i=1}^{N} \mathbf{X}_i - \sum_{i=1}^{N} \log \mathbb{E}[\exp(\eta \mathbf{X}_i)]\right\| > t\right] \leq \operatorname{Tr}\left(\mathbb{E}\left[-\eta \sum_{i=1}^{N} \mathbf{X}_i + \sum_{i=1}^{N} \log \mathbb{E}[\exp(\eta \mathbf{X}_i)]\right]\right) \cdot (e^t - t - 1)^{-1}.$$

A matrix Bernstein bound can be derived from previous Theorem.

Theorem 4 *(Hsu et al., 2012). If there exists $b > 0, \sigma > 0, k > 0$ s.t. for all $i = 1, \ldots, N$, $\mathbb{E}_i[\mathbf{X}_i] = 0$, $\|\mathbf{X}_i\| \leq b$, $\left\|\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_i(\mathbf{X}_i^2)\right\| \leq \sigma^2$ and $\mathbb{E}\left[\operatorname{Tr}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_i(\mathbf{X}_i^2)\right)\right] \leq \sigma^2 k$ almost surely, then for all $t > 0$,*

$$Pr\left[\left\|\frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i\right\| > \sqrt{\frac{2\sigma^2 t}{N}} + \frac{bt}{3N}\right] \leq k \cdot t(e^t - t - 1)^{-1}.$$

1. (Hsu et al., 2011) consider the more general case, where $\mathbf{X}_i$ be a symmetric matrix-valued functional of $\xi_1, \ldots, \xi_i$.

Previous theorem is valid for symmetric matrices, but it can be extended to general real-valued matrices thanks to the principle of dilation.

Let $\mathbf{Z}$ be a matrix, the *dilation* of $\mathbf{Z}$ is the symmetric matrix $\mathbf{X}$ defined by

$$\mathbf{X} = \begin{bmatrix} 0 & \mathbf{Z} \\ \mathbf{Z}^T & 0 \end{bmatrix}. \text{ Then } \mathbf{X}^2 = \begin{bmatrix} \mathbf{Z}\mathbf{Z}^T & 0 \\ 0 & \mathbf{Z}^T\mathbf{Z} \end{bmatrix}$$

and $\|\mathbf{X}\| = \|\mathbf{Z}\|$, $\mathrm{Tr}(\mathbf{X}^2) = \mathrm{Tr}(\mathbf{Z}\mathbf{Z}^T) + \mathrm{Tr}(\mathbf{Z}^T\mathbf{Z})$ and $\left\|\mathbf{X}^2\right\| \leq \max(\left\|\mathbf{Z}\mathbf{Z}^T\right\|, \left\|\mathbf{Z}^T\mathbf{Z}\right\|)$.

We can then reformulate previous theorem as follows.

Theorem 5 *Let $\xi_1, \ldots, \xi_N$ be i.i.d. random variables, and for $i = 1, \ldots, N$, let $\mathbf{Z}_i = \mathbf{Z}(\xi_i)$ be i.i.d. matrices and $\mathbf{X}_i$ the dilation of $\mathbf{Z}_i$. If there exists $b > 0, \sigma > 0$, and $k > 0$ such that $\mathbb{E}[\mathbf{X}_1] = 0$, $\|\mathbf{X}_1\| \leq b, \left\|\mathbb{E}(\mathbf{X}_1^2)\right\| \leq \sigma^2$ and $\mathrm{Tr}(\mathbb{E}(\mathbf{X}_1^2)) \leq \sigma^2 k$ almost surely, then for all $t > 0$,*

$$Pr\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\right\| > \sqrt{\frac{2\sigma^2 t}{N}} + \frac{bt}{3N}\right] \leq k \cdot t(e^t - t - 1)^{-1}.$$

We will then make use of this theorem to derive our new concentration bounds. Section 4.1 deals with the standard case, Section 4.3 with the prefix case and Section 4.5 with the factor case.

3.2 The subexponential case

Theorem 5 needs that the random matrices are bounded. However, the norm of $\overline{\mathbf{H}}_{\xi}^{U,V}$ depends on the size of U and V and $\widehat{\mathbf{H}}_{\xi}^{U,V}$ may be unbounded even in U and V are finite. Fortunately, Bernstein inequalities for subexponential random variables have been extended to unbounded random matrices whose moments grow at a limited rate [Th. 6.2 in (Tropp, 2012)]. We adapt this result to the dimension-free case in a similar way as what has been done in (Hsu et al., 2012) for Theorem 4.

Theorem 6 *[Matrix Bernstein bound: subexponential case.] If there exist $k > 0$, $R > 0$, and a symmetric matrix $\mathbf{A}$ such that for all $i = 1, \ldots, N$, $\mathbb{E}\mathbf{X}_i = 0$, $\mathbb{E}\mathbf{X}_i^p \preccurlyeq \frac{p!}{2}R^{p-2}\mathbf{A}^2$ for any integer $p \geq 2$, $\mathrm{Tr}(\mathbf{A}^2) \leq k\sigma^2$ where $\sigma^2 = \left\|\mathbf{A}^2\right\|$, then for any $t > 0$,*

$$Pr\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\right\| > \frac{Rt}{N} + \sqrt{\frac{2\sigma^2 t}{N}}\right] \leq k \cdot t(e^t - t - 1)^{-1}.$$

Proof Let $0 < \eta < 1/R$. We have

$$\mathbb{E}e^{\eta\mathbf{X}_i} = \mathbb{E}\sum_{p\geq 0}\frac{1}{p!}\eta^p\mathbf{X}_i^p = \mathbf{I} + \sum_{p\geq 2}\frac{1}{p!}\eta^p\mathbb{E}\mathbf{X}_i^p \preccurlyeq \mathbf{I} + \frac{\eta^2}{2(1-\eta R)}\mathbf{A}^2.$$

Hence, by using the monotonicity of the logarithm function and the transfer rule applied to the inequality $\log(1+x) \leq x$,

$$\log \mathbb{E}e^{\eta\mathbf{X}_i} \preccurlyeq \frac{\eta^2}{2(1-\eta R)}\mathbf{A}^2.$$

Now, let η and t such that

$$\left\|\eta \sum_{i=1}^{N} \mathbf{X}_i - \sum_{i=1}^{N} \log \mathbb{E} e^{\eta \mathbf{X}_i}\right\| \leq t.$$

Then,

$$\left\|\eta \sum_{i=1}^{N} \mathbf{X}_i\right\| \leq t + \left\| \sum_{i=1}^{N} \log \mathbb{E} e^{\eta \mathbf{X}_i}\right\| \leq t + N \frac{\eta^2}{2(1-\eta R)} \left\|\mathbf{A}^2\right\|$$

and

$$\left\|\frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i\right\| \leq \frac{t}{N\eta} + \frac{\eta}{2(1-\eta R)} \sigma^2.$$

Moreover,

$$\mathrm{Tr}\left(\mathbb{E}\left[-\eta \sum_{i=1}^{N} \mathbf{X}_i + \sum_{i=1}^{N} \log \mathbb{E} e^{\eta \mathbf{X}_i}\right]\right) \leq \frac{N\eta^2}{2(1-\eta R)} k\sigma^2.$$

Hence,

$$\begin{aligned}\Pr\left[\left\|\frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i\right\| > \frac{t}{N\eta} + \frac{\eta}{2(1-\eta R)} \sigma^2\right] &\leq \Pr\left[\left\|\eta \sum_{i=1}^{N} \mathbf{X}_i - \sum_{i=1}^{N} \log \mathbb{E} e^{\eta \mathbf{X}_i}\right\| > t\right] \\ &\leq \frac{N\eta^2}{2(1-\eta R)} k\sigma^2 (e^t - t - 1)^{-1} \text{from Theorem 3.}\end{aligned}$$

It can easily be checked that $\frac{t}{N\eta} + \frac{\eta}{2(1-\eta R)}\sigma^2$ takes its minimal positive value m at

$$\eta_{\min} = \frac{\sqrt{\frac{2t}{N\sigma^2}}}{1 + R\sqrt{\frac{2t}{N\sigma^2}}} \text{ and that } m = \frac{Rt}{N} + \sqrt{\frac{2\sigma^2 t}{N}}.$$

We have also

$$\frac{\eta_{\min}^2}{1 - R\eta_{\min}} \leq \left(\frac{\eta_{\min}}{1 - R\eta_{\min}}\right)^2 = \frac{2t}{N\sigma^2}$$

which entails the theorem. ■

4. Concentration bounds for Hankel matrices: main results

In this section, we present the main results of the paper. All proofs are reported in Section 5.

Let p be a rational stochastic language over Σ^*, let S be a sample independently drawn according to p, and let $U, V \subseteq \Sigma^*$ be finite set of strings.

4.1 Concentration Bound for the Hankel Matrix $\mathbf{H}_p^{U,V}$

We first describe a bound on $\|\mathbf{H}_S^{U,V} - \mathbf{H}_p^{U,V}\|$, independent from the sizes of U and V.

Let ξ be a random variable distributed according to p, let $\mathbf{Z}(\xi) = \mathbf{H}_\xi^{U,V} - \mathbf{H}_p^{U,V}$ be the random matrix defined by $\mathbf{Z}[u,v] = \mathbf{1}_{\xi=uv} - p(uv)$ and let $\mathbf{X}$ be the dilation of $\mathbf{Z}$.

Clearly, $\mathbb{E}\,\mathbf{X} = 0$. In order to apply Theorem 5, it is necessary to compute the parameters b, σ and k. We show in Lemma 14 that $\|\mathbf{X}\| \leq 2$, $\mathbb{E}\,Tr(\mathbf{X}^2) \leq 2S_p^{(2)}$ and $\left\|\mathbb{E}\,\mathbf{X}^2\right\| \leq S_p^{(2)}$ which entails that 4 conditions of Theorem 5 are fulfilled with $b = 2, \sigma^2 = S_p^{(2)}$ and $k = 2$.

Theorem 7 *Let p be a rational stochastic language and let S be a sample of N strings drawn i.i.d. from p. For all $t > 0$,*

$$Pr\left[\left\|\mathbf{H}_S^{U,V} - \mathbf{H}_p^{U,V}\right\| > \sqrt{\frac{2S_p^{(2)}t}{N} + \frac{2t}{3N}}\right] \leq 2t(e^t - t - 1)^{-1}.$$

This bound is independent from U and V. It can be noticed that the proof of Lemma 14 also provides a dimension dependent bound by replacing $S_p^{(2)}$ with $\sum_{(u,v)\in U\times V} p(uv)$, which may result in a significative improvement if U or V are small.

The moment $S_p^{(2)}$ is generally unknown. However, it can be estimated from S. Indeed, $S_p^{(2)} = \sum_{u,v\in\Sigma^*} p(uv) = \sum_{w\in\Sigma^*}(|w|+1)p(w) = \mathbb{E}|\xi|+1$ and $\frac{1}{N}\sum_{w\in S}|w|+1$ is an natural estimate for $S_p^{(2)}$.The random variable $|\xi|$ is sub-exponential and its concentration around its mean can be estimated using Bernstein-type inequalities (see (Vershynin, 2012) for example). Thus, Theorem 7 can easily be reformulated replacing $S_p^{(2)}$ by its estimate.

4.2 Concentration Bound for the smoothed prefix Hankel Matrix $H_{\overline{p}_\eta}^{U,V}$

The random matrix $\overline{\mathbf{Z}}(\xi) = \overline{\mathbf{H}}_\xi^{U,V} - \mathbf{H}_{\overline{p}}^{U,V}$ is defined by $\overline{\mathbf{Z}}[u,v] = \mathbf{1}_{uv\in Pref(\xi)} - \overline{p}(uv)$, where references to U and V are omitted for the sake of readability. It can easily be shown that $\left\|\overline{\mathbf{Z}}\right\|$ may be unbounded if U or V are unbounded. For example, consider the stochastic language defined on a one-letter alphabet $\Sigma = \{a\}$ by $p(a^n) = (1-\rho)\rho^n$. If $U = V = \Sigma^{\leq n}$, $\overline{\mathbf{Z}}$ may be equal to the $(n+1)\times(n+1)$ upper triangular all-ones matrix whose norm is $\Theta(n)$. Hence, Theorem 5 cannot be directly applied to obtain dimension-free bounds. This suggests that the concentration of $\overline{\mathbf{Z}}$ around its mean could be far weaker than the concentration of $\mathbf{Z}$.

For any $\eta \in [0,1]$, we define a smoothed variant[2] of $\overline{p}$ by

$$\overline{p}_\eta(u) \coloneqq \sum_{x\in\Sigma^*} \eta^{|x|} p(ux) = \sum_{n\geq 0} \eta^n p(u\Sigma^n). \tag{4}$$

Note that $\overline{p}_1 = \overline{p}$, $\overline{p}_0 = p$ and that $p(u) \leq \overline{p}_\eta(u) \leq \overline{p}(u)$ for every string u: the functions $\overline{p}_\eta$ are natural intermediates between p and $\overline{p}$.

Any function $\overline{p}_\eta$ can be used to compute p:

$$p(u) = \overline{p}_\eta(u) - \eta\overline{p}_\eta(u\Sigma). \tag{5}$$

Moreover, when p is rational, each $\overline{p}_\eta$ is also rational and a linear representation of p can be derived from any linear representation of $\overline{p}_\eta$. More precisely,

2. Note that our *smoothed variant* can also be interpreted as a *discounted variant* since the parameter η helps to reduce the impact of long strings.

Proposition 8 *Let p be a rational stochastic language, let $\langle I, (\mathbf{M}_x)_{x\in\Sigma}, T\rangle$ be a minimal linear representation of p and let $\overline{T}_\eta = (\mathbf{I}_d - \eta\mathbf{M}_\Sigma)^{-1}T$. Then, $\langle I, (\mathbf{M}_x)_{x\in\Sigma}, \overline{T}_\eta\rangle$ is a linear representation of $\overline{p}_\eta$. Hence, $S^{(k)}_{\overline{p}_\eta} = I^T(\mathbf{I}_d - \mathbf{M}_\Sigma)^{-k}(\mathbf{I}_d - \eta\mathbf{M}_\Sigma)^{-1}T$. In particular, $S^{(k)}_{\overline{p}} = S^{(k+1)}_p$.*

Therefore, it is a consistent learning strategy to learn $\overline{p}_\eta$ from the data, for some η, and next, to derive p. A theoretical study that would guide the choice of the parameter η remains to be done. In the absence of such indications, its value can be set by cross validation.

For any $0 \leq \eta \leq 1$, let $\overline{\mathbf{Z}}_\eta(\xi)$ be the random matrix defined by

$$\overline{\mathbf{Z}}_\eta[u,v] \coloneqq \sum_{x\in\Sigma^*} \eta^{|x|}\mathbf{1}_{\xi=uvx} - \overline{p}_\eta(uv) = \sum_{x\in\Sigma^*} \eta^{|x|}(\mathbf{1}_{\xi=uvx} - p(uvx)).$$

for any $(u,v) \in U\times V$. It is clear that $\mathbb{E}\,\overline{\mathbf{Z}}_\eta = 0$. We show that $\left\|\overline{\mathbf{Z}}_\eta\right\|$ is bounded by $\frac{1}{1-\eta} + S^{(1)}_{\overline{p}_\eta}$ if $\eta < 1$ (Lemma 15), that $\left\|\mathbb{E}\,\overline{\mathbf{X}}^2_\eta\right\| \leq S^{(2)}_{\overline{p}_\eta}$ and $\mathbb{E}\,Tr(\overline{\mathbf{X}}^2_\eta) \leq 2S^{(2)}_{\overline{p}_\eta}$. (Lemma 17).

Therefore, we can apply Theorem 5 with $b = \frac{1}{1-\eta} + S^{(1)}_{\overline{p}_\eta}, \sigma^2 = S^{(2)}_{\overline{p}_\eta}$ and $k = 2$.

Theorem 9 *Let p be a rational stochastic language, let S be a sample of N strings drawn i.i.d. from p and let $0 \leq \eta < 1$. For all $t > 0$,*

$$Pr\left[\left\|\overline{\mathbf{H}}^{U,V}_{\eta,S} - \mathbf{H}^{U,V}_{\overline{p}_\eta}\right\|_2 > \sqrt{\frac{2S^{(2)}_{\overline{p}_\eta}t}{N}} + \frac{t}{3N}\left[\frac{1}{1-\eta} + S^{(1)}_{\overline{p}_\eta}\right]\right] \leq 2t(e^t - t - 1)^{-1}.$$

Remark that when $\eta = 0$ we find back the concentration bound of Theorem 7. When U and V are finite, a careful examination of the proof of Lemma 15 shows that $S^{(2)}_{\overline{p}_\eta}$ can be replaced with $\sum_{(u,v)\in U\times V}\overline{p}_\eta(uv)$, which may provide a significant better bound if U and V are small. Moreover, Inequality 8 can be used to provide a finite bound depending on the sizes of U and V, when $\eta = 1$.

As for Theorem 7, the moment $S^{(2)}_{\overline{p}_\eta}$ is generally unknown but it can be estimated from S, with controled accuracy, providing a reformulation of the theorem that would not depend on this parameter.

4.3 Concentration Bound for the prefix Hankel Matrix $H^{U,V}_{\overline{p}}$

Theorem 9 does not provide a dimension-free bound for the prefix Hankel matrices $\overline{\mathbf{H}}^{U,V}_S$. However, we show that $\overline{\mathbf{Z}}(\xi)$ is a subexponential random matrix (Lemma 21), and that Theorem 6 can be used to provide a bound in this case.

More precisely, let $\mathbf{X}(\xi)$ be the dilation of the matrix $\overline{\mathbf{Z}}(\xi)$, let $C_p > 0$ and $0 < \rho_p < 1$ be such that $p(\Sigma^n) \leq C_p\rho_p^n$ for any integer n.

For any $0 < \beta < -\ln\rho_p$, we show in Lemma 21 the existence of a symmetric matrix $\mathbf{A}^2$ satisfying $\left\|\mathbf{A}^2\right\| \leq K(1-\rho_p e^\beta)^{-1}, \mathrm{Tr}(\mathbf{A}^2) \leq 2K(1-\rho_p e^\beta)^{-2}$and such that for any $k \geq 0$,

$\mathbb{E}\mathbf{X}^k \preccurlyeq \frac{k!}{2} R^{k-2}\mathbf{A}^2$ where $R = e^{1/e}\beta^{-1}$ and $K = 2e^{3/e}\beta^{-3}C_p S_p^{(3)} e^{\beta S_p^{(2)}}$ is a constant that only depends on p and β.

Hence, we can apply Theorem 6 to obtain the following dimension free bound:

Theorem 10 *Let p be a rational stochastic language and let S be a sample of N strings drawn i.i.d. from p. For all $t > 0$,*

$$Pr\left[\left\|\overline{\mathbf{H}}_S^{U,V} - \mathbf{H}_{\overline{p}}^{U,V}\right\|_2 > \frac{e^{1/e}t}{N\beta} + \sqrt{\frac{2\sigma^2 t}{N}}\right] \leq \frac{2}{1-\rho_p e^{\beta}} \cdot t(e^t - t - 1)^{-1}$$

where $\sigma^2 = K(1-\rho_p e^{\beta})^{-2}$, $K = \beta^{-2}C_p S_p^{(3)} e^{\beta S_p^{(2)}}$ and $0 < \beta < -\ln \rho_p$.

Note that β can be set to $1-\rho_p$ but depending on the particular values of the terms, β can be adjusted to obtain better bounds.

Thus, Theorem 10 describes a dimension free concentration bound for the prefix Hankel matrix. However, the constants K and σ that occur in this bound can be very large, which makes it impossible to use it in practical cases, such as those we consider in Section 6.

4.4 Concentration Bound for the smoothed factor Hankel Matrix $H_{\widehat{p}_\eta}^{U,V}$

The random matrix $\widehat{\mathbf{Z}}(\xi) = \widehat{\mathbf{H}}_{\xi}^{U,V} - \mathbf{H}_{\widehat{p}^{U,V}}$ is defined by

$$\widehat{\mathbf{Z}}[u,v] = \sum_{x,y\in\Sigma^*} \mathbf{1}_{\xi=xuvy} - \widehat{p}(uv),$$

where references to U and V are omitted for the sake of readability. $\left\|\widehat{\mathbf{Z}}\right\|$ is unbounded if the support of p is unbounded. Indeed, $\widehat{\mathbf{Z}}[\epsilon,\epsilon] = |\xi| + 1 - \widehat{p}(\epsilon)$. Hence, Theorem 5 cannot be directly applied either.

We can also define smoothed variants $\widehat{p}_\eta$ of $\widehat{p}$, for any $\eta \in [0,1]$ by

$$\widehat{p}_\eta(u) = \sum_{x,y\in\Sigma^*} \eta^{|xy|} p(xuy) = \sum_{m,n\geq 0} \eta^{m+n} p(\Sigma^m u \Sigma^n)$$

which have properties similar to functions $\overline{p}_\eta$:

- $p \leq \widehat{p}_\eta \leq \widehat{p}$, $\widehat{p}_1 = \widehat{p}$ and $\widehat{p}_0 = p$,
- when p is rational, each $\widehat{p}_\eta$ is also rational,
- if $\langle I, (\mathbf{M}_x)_{x\in\Sigma}, T\rangle$ be a minimal linear representation of p then $\langle \widehat{I}_\eta, (\mathbf{M}_x)_{x\in\Sigma}, \overline{T}_\eta\rangle$ is a linear representation of $\hat{p}_\eta$, where $\widehat{I}_\eta = (I_d - \eta\mathbf{M}_\Sigma^\intercal)^{-1} I$,
- I and T can be computed from $\widehat{I}_\eta$ and $\widehat{T}_\eta$ when η and $\mathbf{M}_\Sigma$ are known:

$$T = (\mathbf{I}_d - \eta\mathbf{M}_\Sigma)\widehat{T}_\eta, I = (\mathbf{I}_d - \eta\mathbf{M}_\Sigma)\widehat{I}_\eta \text{ and} \tag{6}$$

$$p(u) = \widehat{p}_\eta(u) - \eta\widehat{p}_\eta(u\Sigma) - \eta\widehat{p}_\eta(\Sigma u) + \eta^2\widehat{p}_\eta(\Sigma u\Sigma) \tag{7}$$

- therefore, it is a consistent learning strategy to learn $\widehat{p}_\eta$ from the data, for some η, and next, to derive p.

For $\eta \in [0,1]$, let $\widehat{\mathbf{Z}}_\eta(\xi)$ be the random matrix defined by

$$\widehat{\mathbf{Z}}_\eta[u,v] = \sum_{x,y\in\Sigma^*} \eta^{|xy|}\mathbf{1}_{\xi=xuvy} - \widehat{p}_\eta(uv) = \sum_{x,y\in\Sigma^*} \eta^{|xy|}(\mathbf{1}_{\xi=xuvy} - p(xuvy))$$

for any $(u,v) \in U \times V$. Clearly, $\mathbb{E}\,\widehat{\mathbf{Z}}_\eta = 0$. We show that $\left\|\widehat{\mathbf{Z}}_\eta\right\|$ is bounded by $(1-\eta)^{-2} + S^{(1)}_{\widehat{p}_\eta}$ if $\eta < 1$ (lemma 24).

While $\overline{p}$ is bounded by 1, a property which is often used in the proofs, $\widehat{p}$ is unbounded when η converges to 1. Let us introduce a new constant K_η defined by

$$K_\eta = \begin{cases} 1 & \text{if } \eta \le e^{-1} \\ (-e\eta\ln\eta)^{-1} & \text{otherwise.} \end{cases}$$

We show in Lemma 26 that

$$\left\|\mathbb{E}\widehat{\mathbf{X}}^2_\eta\right\| \le K_\eta S^{(2)}_{\widehat{p}_\eta} \text{ and } Tr(\mathbb{E}(\widehat{\mathbf{X}}^2_\eta)) \le 2K_\eta S^{(2)}_{\widehat{p}_\eta}.$$

Eventually, we can apply Theorem 5 with $b = (1-\eta)^{-2} + S^{(1)}_{\widehat{p}_\eta}, \sigma^2 = K_\eta S^{(2)}_{\widehat{p}_\eta}$ and $k = 2$.

Theorem 11 *Let p be a rational stochastic language, let S be a sample of N strings drawn i.i.d. from p and let $0 \le \eta < 1$. For all $t > 0$,*

$$Pr\left[\|\widehat{\mathbf{H}}^{U,V}_{\eta,S} - \mathbf{H}^{U,V}_{\widehat{p}_\eta}\|_2 > \sqrt{\frac{2K_\eta S^{(2)}_{\widehat{p}_\eta} t}{N}} + \frac{t}{3N}\left[\frac{1}{(1-\eta)^2} + S^{(1)}_{\widehat{p}_\eta}\right]\right] \le 2t(e^t - t - 1)^{-1}.$$

Remark that when $\eta = 0$ we find back the concentration bound of Theorem 7. As in the previous cases, $S^{(2)}_{\widehat{p}_\eta}$ can be replaced with $\sum_{(u,v)\in U\times V}\widehat{p}_\eta(uv)$, which may provide a significant better bound if U and V are small. However, it not possible to use these results to obtain a bound, even depending on U and V, when $\eta = 1$.

As for the previous theorems, the moment $S^{(2)}_{\widehat{p}_\eta}$ can be estimated from S in order to provide a reformulation of the theorem that would not depend on this parameter.

4.5 Concentration Bound for the factor Hankel Matrix $H^{U,V}_{\widehat{p}}$

$\widehat{\mathbf{Z}}(\xi)$ is not a subexponential random matrix, and therefore, Theorem 6 cannot be used to provide a bound in this case. This suggests that the concentration of $\widehat{\mathbf{Z}}(\xi)$ around its mean is quite loose.

5. Proofs of all concentration bounds results

In this section, we detail the proofs of all the results stated in Section 4, keeping the titles of all subsections and the notations that have been introduced in the corresponding subsection.

5.1 Concentration Bound for the Hankel Matrix $\mathbf{H}_p^{U,V}$: proofs

Recall that $\mathbf{X}$ is the dilation of the random matrix $\mathbf{Z}(\xi) = \mathbf{H}_\xi^{U,V} - \mathbf{H}_p^{U,V}$.

Clearly, $\mathbb{E}\,\mathbf{X} = 0$. We need technical lemmas in order to obtains bound on $\mathbb{E}\,\mathbf{X}^2$ and $\mathbb{E}\,Tr(\mathbf{X}^2)$ and apply Theorem 5.

Lemma 12 *Let X and Y be two random variables such that $0 \leq X, Y \leq M$. Then,*

$$|\mathbb{E}\,(X - \mathbb{E}\,X)(Y - \mathbb{E}\,Y)| \leq M \min\{\mathbb{E}\,X, \mathbb{E}\,Y\}.$$

Proof Indeed, $0 \leq \mathbb{E}\,XY \leq M \min\{\mathbb{E}\,X, \mathbb{E}\,Y\}$ and $0 \leq \mathbb{E}\,X\mathbb{E}\,Y \leq M \min\{\mathbb{E}\,X, \mathbb{E}\,Y\}$, which entails the lemma. ■

Lemma 13 *For any $u, u' \in U$, $v, v' \in V$,*

$$|\mathbb{E}\,\mathbf{Z}[u, v]\mathbf{Z}[u', v']| \leq \min\{p(uv), p(u'v')\}.$$

Proof This a corollary of Lemma 12 with $X \coloneqq \mathbf{1}_{\xi=uv}$, $Y \coloneqq \mathbf{1}_{\xi=u'v'}$ and $M = 1$. ■

Lemma 14 $\|\mathbf{X}\| \leq 2$, $\mathbb{E}\,Tr(\mathbf{X}^2) \leq 2S_p^{(2)}$ *and* $\left\|\mathbb{E}\,\mathbf{X}^2\right\| \leq S_p^{(2)}$.

Proof

1. $\forall u \in U$, $\sum_{v\in V} |\mathbf{Z}[u, v]| = \sum_{v\in V} |\mathbf{1}_{\xi=uv} - p(uv)| \leq 1 + p(u\Sigma^*) \leq 2$. Therefore, $\|\mathbf{Z}\|_\infty \leq 2$. In a similar way, it can be shown that $\|\mathbf{Z}\|_1 \leq 2$. Hence,

$$\|\mathbf{X}\| = \|\mathbf{Z}\| \leq \sqrt{\|\mathbf{Z}\|_\infty \|\mathbf{Z}\|_1} \leq 2.$$

2. For all $(u, u') \in U^2$: $\mathbf{Z}\mathbf{Z}^T[u, u'] = \sum_{v\in V} \mathbf{Z}[u, v]\mathbf{Z}_{u',v}$. Therefore,

$$\mathbb{E}\,Tr(\mathbf{Z}\mathbf{Z}^T) = \mathbb{E} \sum_{u\in U} \mathbf{Z}\mathbf{Z}^T[u, u] = \mathbb{E} \sum_{u\in U, v\in V} \mathbf{Z}[u, v]\mathbf{Z}[u, v] \leq \sum_{u,v\in\Sigma^*} \mathbb{E}\,\mathbf{Z}[u, v]^2 \leq \sum_{u,v\in\Sigma^*} p(uv) \leq S_p^{(2)}.$$

In a similar way, it can be proved that $\mathbb{E}\,Tr(\mathbf{Z}^T\mathbf{Z}) \leq S_p^{(2)}$. Therefore, $\mathbb{E}\,Tr(\mathbf{X}^2) \leq 2S_p^{(2)}$.

3. For any $u \in U$,

$$\sum_{u'\in U} |\mathbb{E}\,\mathbf{Z}\mathbf{Z}^T[u, u']| \leq \sum_{u'\in U, v\in V} |\mathbb{E}\,\mathbf{Z}[u, v]\mathbf{Z}[u', v]| \leq \sum_{u'\in U, v\in V} p(u'v) \leq S_p^{(2)}.$$

Hence, $\left\|\mathbf{Z}\mathbf{Z}^T\right\|_\infty \leq S_p^{(2)}$. It can be proved, in a similar way, that $\left\|\mathbf{Z}^T\mathbf{Z}\right\|_\infty \leq S_p^{(2)}$, $\left\|\mathbf{Z}\mathbf{Z}^T\right\|_1 \leq S_p^{(2)}$ and $\left\|\mathbf{Z}^T\mathbf{Z}\right\|_1 \leq S_p^{(2)}$. Therefore, $\left\|\mathbf{X}^2\right\| \leq S_p^{(2)}$. ■

5.2 Concentration Bound for the smoothed prefix Hankel Matrix $H_{\overline{p}_\eta}^{U,V}$: proofs

Proof (Proposition 8.)

For every string u, $\overline{p}_\eta(u) = \sum_{n\geq 0} I^T \mathbf{M}_u \eta^n \mathbf{M}_\Sigma^n T = I^T \mathbf{M}_u (\sum_{n\geq 0} \eta^n \mathbf{M}_\Sigma^n) T = I^T \mathbf{M}_u \overline{T}_\eta$. The expression of $S_{\overline{p}_\eta}^{(k)}$ comes directly from Equation 2. ■

Lemma 15 *For any $U, V \subseteq \Sigma^*$,*

$$\left\|\overline{\mathbf{Z}}_\eta\right\| \leq \frac{1}{1-\eta} + S_{\overline{p}_\eta}^{(1)}.$$

Proof Indeed, let $u \in U$.

$$\begin{aligned}\sum_{v\in V} |\overline{\mathbf{Z}}_\eta[u,v]| &\leq \sum_{v,x\in\Sigma^*} \eta^{|x|} \mathbf{1}_{\xi=uvx} + \sum_{v\in\Sigma^*} \overline{p}_\eta(uv) \\ &\leq (1+\eta+\ldots+\eta^{|\xi|-|u|}) + S_{\overline{p}_\eta}^{(1)} \\ &\leq \frac{1}{1-\eta} + S_{\overline{p}_\eta}^{(1)}.\end{aligned}$$

Hence, $\left\|\overline{\mathbf{Z}}_\eta\right\|_\infty \leq \frac{1}{1-\eta} + S_{\overline{p}_\eta}^{(1)}$. Similarly, $\left\|\overline{\mathbf{Z}}_\eta\right\|_1 \leq \frac{1}{1-\eta} + S_{\overline{p}_\eta}^{(1)}$, which completes the proof. ■

When U and V are bounded, let l be the maximal length of a string in $U \cup V$. It can easily be shown that $\left\|\overline{\mathbf{Z}}_\eta\right\| \leq l + 1 + S_{\overline{p}_\eta}^{(1)}$ and therefore, in that case,

$$\left\|\overline{\mathbf{Z}}_\eta\right\| \leq Min(l+1, \frac{1}{1-\eta}) + S_{\overline{p}_\eta}^{(1)} \tag{8}$$

which holds even if $\eta = 1$.

Lemma 16 $|\mathbb{E}(\overline{\mathbf{Z}}_\eta[u,v]\overline{\mathbf{Z}}_\eta[u',v'])| \leq \min\{\overline{p}_\eta(uv), \overline{p}_\eta(u'v')\}$, *for any $u, u' \in U$ and $v, v' \in V$.*

Proof This a corollary of Lemma 12 with $X \coloneqq \sum_{x\in\Sigma^*} \eta^{|x|} \mathbf{1}_{\xi=uvx}$, $Y \coloneqq \sum_{x\in\Sigma^*} \eta^{|x|} \mathbf{1}_{\xi=u'v'x}$ and $M = 1$. ■

Let $\overline{\mathbf{X}}_\eta$ be the dilation of $\overline{\mathbf{Z}}_\eta$. We can now propose bounds for $\mathbb{E}\,\overline{\mathbf{X}}_\eta^2$ and $\mathbb{E}\,Tr(\overline{\mathbf{X}}_\eta^2)$.

Lemma 17

$$\left\|\mathbb{E}\,\overline{\mathbf{X}}_\eta^2\right\| \leq S_{\overline{p}_\eta}^{(2)} \text{ and } \mathbb{E}\,Tr(\overline{\mathbf{X}}_\eta^2) \leq 2S_{\overline{p}_\eta}^{(2)}.$$

Proof Indeed,

$$\left\|\mathbb{E}\,\overline{\mathbf{Z}}_\eta \overline{\mathbf{Z}}_\eta^T\right\|_\infty \leq \max_{u\in\Sigma^*} \sum_{u',v\in\Sigma^*} |\mathbb{E}\,\overline{\mathbf{Z}}_\eta[u,v]\overline{\mathbf{Z}}_\eta[u',v]| \leq \sum_{u',v} \overline{p}_\eta(u'v) \leq S_{\overline{p}_\eta}^{(2)}.$$

We have also

$$\left\|\mathbb{E}\,\overline{\mathbf{Z}}_\eta \overline{\mathbf{Z}}_\eta^T\right\|_1 \leq S_{\overline{p}_\eta}^{(2)} \text{ and therefore } \left\|\mathbb{E}\,\overline{\mathbf{Z}}_\eta\ \overline{\mathbf{Z}}_\eta^T\right\| \leq S_{\overline{p}_\eta}^{(2)}.$$

Similar computations provide all the inequalities. ■

5.3 Concentration Bound for the prefix Hankel Matrix $H_{\overline{p}}^{U,V}$: proofs

Let $\mathbf{X}(\xi)$ be the dilation of the matrix $\overline{\mathbf{Z}}(\xi)$. It can easily be shown that

$$\mathbf{X}^{2k+1} = \begin{bmatrix} 0 & \left(\overline{\mathbf{Z}} \cdot \overline{\mathbf{Z}}^\top\right)^k \overline{\mathbf{Z}} \\ \overline{\mathbf{Z}}^\top \left(\overline{\mathbf{Z}} \cdot \overline{\mathbf{Z}}^\top\right)^k & 0 \end{bmatrix} \text{ and } \mathbf{X}^{2k} = \begin{bmatrix} \left(\overline{\mathbf{Z}} \cdot \overline{\mathbf{Z}}^\top\right)^k & 0 \\ 0 & \left(\overline{\mathbf{Z}}^\top \cdot \overline{\mathbf{Z}}\right)^k \end{bmatrix}.$$

Let $t \in \Sigma^*$ be a realization of ξ.

Lemma 18 *For any strings $u, v, t \in \Sigma^*$ and any stochastic language p,*

$$\sum_{v \in V} |\overline{\mathbf{Z}}[u,v]| \leq |t| + \overline{p}(u\Sigma^*), \sum_{u \in U} |\overline{\mathbf{Z}}[u,v]| \leq |t| + \overline{p}(\Sigma^* u)$$

and

$$\sum_{u \in U, v \in V} |\overline{\mathbf{Z}}[u,v]| \leq \frac{|t|(|t|+3)}{2} + S_p^{(3)}.$$

Proof We have $\sum_{v \in V} |\overline{\mathbf{Z}}[u,v]| \leq \sum_{w \in \Sigma^*, uw \in \mathrm{Pref}(t)} |1 - \overline{p}(uw)| + \overline{p}(u\Sigma^*)$. Moreover, if $u = w = \epsilon$, $\overline{p}(uw) = 1$. Hence, there are at most $|t|$ strings w such that $uw \in \mathrm{Pref}(t)$ and $|1 - \overline{p}(uw)| \neq 0$, which proves the first inequality. The second one is proved in a similar way. A similar argument proves that there are at most $(|t|+1)(|t|+2)/2 - 1 = |t|(|t|+3)/2$ pairs of words u, w such that $uw \in \mathrm{Pref}(t)$ and $|1 - \overline{p}(uw)| \neq 0$, which entails the third inequality. ■

Lemma 19 *Let $\mathbf{M}$ be a matrix of the form $(\overline{\mathbf{Z}}^\top)^e (\overline{\mathbf{Z}} \cdot \overline{\mathbf{Z}}^\top)^k \overline{\mathbf{Z}}^f$, where $k \in \mathbb{N}$ and $e, f \in \{0,1\}$. Then, for any strings $u, v \in \Sigma^*$, $|\mathbf{M}[u,v]| \leq (|t| + S_p^{(2)})^h$ where $h = e + 2k + f$. Moreover, $\sum_{u \in U} |\mathbf{M}[u,v]|$ and $\sum_{v \in V} |\mathbf{M}[u,v]|$ are bounded by $S_p^{(3)}(|t|+1)(|t| + S_p^{(2)})^h$.*

Proof By induction on $h = e + 2k + f$. The inequality is obvious if $h = 0$. Let $\mathbf{M} = \overline{\mathbf{Z}}\mathbf{N}$.

$$|\mathbf{M}[u,v]| \leq \sum_{w \in V} |\overline{\mathbf{Z}}[u,w]\mathbf{N}[w,v]| \leq (|t| + S_p^{(2)})^{h-1} \sum_{w \in V} |\overline{\mathbf{Z}}[u,w]| \text{ by induction hypothesis.}$$

By Lemma 18, we have

$$|\mathbf{M}[u,v]| \leq (|t| + S_p^{(2)})^h \text{ since } \overline{p}(u\Sigma^*) \leq S_p^{(2)}$$

and

$$\sum_{u \in U} |\mathbf{M}[u,v]| \leq (|t| + S_p^{(2)})^{h-1} \left(\frac{|t|(|t|+3)}{2} + S_p^{(3)} \right) \leq S_p^{(3)}(|t|+1)(|t| + S_p^{(2)})^h$$

since $1 \leq S_p^{(3)}$ and $(\frac{|t|(|t|+3)}{2} + 1) \leq (|t|+1)(|t| + S_p^{(2)})$.

The other cases are proved in a similar way. ■

Corollary 20 *For any integer k, $\left\|\mathbf{X}^k\right\| \leq S_p^{(3)}(|t|+1)(|t| + S_p^{(2)})^k$.*

Proof Indeed, it can easily be shown that for any $k \in \mathbb{N}$ and $e \in \{0,1\}$, $\left\|\mathbf{X}^{2k+e}\right\| = \left\|\left(\overline{\mathbf{Z}} \cdot \overline{\mathbf{Z}}^\top\right)^k \overline{\mathbf{Z}}^e\right\|$. The result is a consequence of Lemma 19. ■

Let $C_p > 0$ and $0 < \rho_p < 1$ be such that $p(\Sigma^n) \leq C_p \rho_p^n$.

Let $\mathbf{X}^k(t) = \mathbf{U}_t^\top diag(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)\mathbf{U}_t$ be an eigenvalue decomposition of $\mathbf{X}^k$, where $\lambda_1, \lambda_2, \ldots, \lambda_r$ are the non zero eigenvalues of $\mathbf{X}^k(t)$, and let $\mathbf{J}_t = diag(1, \ldots, 1, 0, \ldots, 0)$ the matrix whose coefficients are all equal to 0 but the r upper diagonal elements which are equal to 1. We have

$$\mathbf{X}^k \preccurlyeq \mathbf{U}^\top diag(|\lambda_1|, \ldots, |\lambda_r|, 0, \ldots, 0)\mathbf{U} \preccurlyeq \left\|\mathbf{X}_t^k\right\| \mathbf{U}_t^\top \mathbf{J}_t \mathbf{U}_t. \tag{9}$$

For any $n \in \mathbb{N}$, let

$$\mathbf{M}_n = \sum_{t \in \Sigma^n} \frac{p(t)}{p(\Sigma^n)} \mathbf{U}_t^\top \mathbf{J}_t \mathbf{U}_t \text{ if } p(\Sigma^n) \neq 0 \text{ and } \mathbf{M}_n = 0 \text{ otherwise.}$$

We can remark that $\|\mathbf{M}_n\| \leq 1$ and $0 \preccurlyeq \mathbf{M}_n$.

Let $\mathbf{A}$ be the symmetric matrix such that

$$\mathbf{A}^2 = K \sum_{n \geq 0} \rho_p^n e^{\beta n} \mathbf{M}_n$$

where $K = 2e^{3/e}\beta^{-3} C_p S_p^{(3)} e^{\beta S_p^{(2)}}$ and $0 < \beta < -\ln \rho_p$. For example, β can be taken equal to $1 - \rho_p$.

Since $\rho_p e^\beta < 1$ and $\|\mathbf{M}_n\| \leq 1$, $\mathbf{A}$ is well defined. Moreover, $0 \preccurlyeq \mathbf{A}^2$.

Lemma 21 *We have* $\left\|\mathbf{A}^2\right\| \leq K(1 - \rho_p e^\beta)^{-1}, \mathrm{Tr}(\mathbf{A}^2) \leq 2K(1 - \rho_p e^\beta)^{-2}$ *and for any* $k \geq 0$, $\mathbb{E}\mathbf{X}^k \preccurlyeq \frac{k!}{2} R^{k-2} \mathbf{A}^2$ *where* $R = e^{1/e}\beta^{-1}$.

Proof The bound on $\|\mathbf{A}\|$ is straightforward.

The rank of $\mathbf{X}^k$, equal to the rank of $\mathbf{J}_t$, is bounded by $2(|t| + 1)$ and hence, $\mathrm{Tr}(\mathbf{M}_n) \leq 2(n+1)$. The bound on $\mathrm{Tr}(\mathbf{A}^2)$ comes from the following classical equality: if $|x| < 1$ then, $\sum_{n \geq 0}(n+1)x^n = (1-x)^{-2}$.

We have

$$\begin{aligned}
\mathbb{E}\mathbf{X}^k &\preccurlyeq \sum_t p(t) S_p^{(3)} (|t|+1)(|t|+S_p^{(2)})^k \mathbf{U}_t^\top \mathbf{J}_t \mathbf{U}_t && \text{from } Eq\ 9 \text{ and } Cor.\ 20 \\
&\preccurlyeq \sum_t p(t) S_p^{(3)} (|t|+S_p^{(2)})^{k+1} \mathbf{U}_t^\top \mathbf{J}_t \mathbf{U}_t && \text{since } S_p^{(2)} \geq 1 \\
&= S_p^{(3)} \sum_{n\geq 0} p(\Sigma^n)(n+S_p^{(2)})^{k+1} \mathbf{M}_n \\
&\preccurlyeq C_p S_p^{(3)} \sum_{n\geq 0} \rho_p^n \frac{[(n+S_p^{(2)})\beta]^{k+1}}{(k+1)!} \frac{(k+1)!}{\beta^{k+1}} \mathbf{M}_n && \text{since } p(\Sigma^n) \leq C_p \rho_p^n \\
&\preccurlyeq C_p S_p^{(3)} \sum_{n\geq 0} \rho_p^n e^{(n+S_p^{(2)})\beta} \frac{(k+1)!}{\beta^{k+1}} \mathbf{M}_n && \text{since } x^k/k! \leq e^x \\
&\preccurlyeq k! \left(\frac{e^{1/e}}{\beta}\right)^k \frac{e^{1/e}}{\beta} C_p S_p^{(3)} e^{S_p^{(2)}\beta} \sum_{n\geq 0} (\rho_p e^\beta)^n \mathbf{M}_n && \text{since } k+1 \leq e^{(k+1)/e} \\
&= \frac{k!}{2} R^{k-2} \mathbf{A}^2.
\end{aligned}$$

■

5.4 Concentration Bound for the smoothed factor Hankel Matrix $H_{\widehat{p}_\eta}^{U,V}$: proofs

Lemma 22 *Let $0 < \eta \leq 1$. For any integer n, $(n+1)\eta^n \leq K_\eta$.*

Proof Let $f(x) = (x+1)\eta^x$. We have $f'(x) = \eta^x(1 + (x+1)\ln\eta)$ and f takes its maximum for $x_M = -1 - 1/\ln\eta$, which is positive if and only if $\eta > 1/e$. We have $f(x_M) = (-e\eta\ln\eta)^{-1}$. ■

Lemma 23 *Let $w, u \in \Sigma^*$. Then,*

$$\sum_{x,y\in\Sigma^*} \eta^{|xy|} \mathbf{1}_{w=xuy} \leq K_\eta \text{ and } \widehat{p}(u) \leq K_\eta p(\Sigma^* u \Sigma^*).$$

Proof Indeed, if $w = xuy$, then $|xy| = |w| - |u|$ and u appears at most $|w| - |u| + 1$ times as a factor of w. Therefore, $\sum_{x,y\in\Sigma^*} \eta^{|xy|}\mathbf{1}_{w=xuy} \leq (|w| - |u| + 1)\eta^{|w|-|u|} \leq K_\eta$. Moreover,

$$\widehat{p}(u) = \sum_{x,y\in\Sigma^*} \eta^{|xy|} p(xuy) = \sum_{w\in\Sigma^* u \Sigma^*} p(w) \sum_{x,y\in\Sigma^*} \eta^{|xy|} \mathbf{1}_{w=xuvy} \leq K_\eta p(\Sigma^* u \Sigma^*).$$

■

For $\eta \in [0,1]$, let $\widehat{\mathbf{Z}}_\eta(\xi)$ be the random matrix defined by

$$\widehat{\mathbf{Z}}_\eta[u,v] = \sum_{x,y\in\Sigma^*} \eta^{|xy|} \mathbf{1}_{\xi=xuvy} - \widehat{p}_\eta(uv) = \sum_{x,y\in\Sigma^*} \eta^{|xy|} (\mathbf{1}_{\xi=xuvy} - p(xuvy))$$

for any $(u,v) \in U \times V$. Clearly, $\mathbb{E}\,\widehat{\mathbf{Z}}_\eta = 0$. We show below that $\|\widehat{\mathbf{Z}}_\eta\|$ is bounded if $\eta < 1$.

The moments $S^{(k)}_{\widehat{p}_\eta}$ satisfy $S^{(k)}_{\widehat{p}_\eta} = I^\top (I_d - \eta \mathbf{M}_\Sigma)^{-1}(I_d - \mathbf{M}_\Sigma)^{-k}(I_d - \eta \mathbf{M}_\Sigma)^{-1} T$, $S^{(k)}_{\widehat{p}_0} = S^{(k)}_p$ and $S^{(k)}_{\widehat{p}_1} = S^{(k+2)}_p$.

Lemma 24

$$\left\|\widehat{\mathbf{Z}}_\eta\right\| \leq (1-\eta)^{-2} + S^{(1)}_{\widehat{p}_\eta}.$$

Proof Indeed, for all u,

$$\sum_{v \in V} |\widehat{\mathbf{Z}}_\eta[u,v]| \leq \sum_{v,x,y \in \Sigma^*} [\eta^{|xy|} \mathbf{1}_{\xi = xuvy} + \hat{p}_\eta(uv)] \leq \sum_{x,y \in \Sigma^*} \eta^{|xy|} \mathbf{1}_{\xi \in x\Sigma^* y} + S^{(1)}_{\hat{p}_\eta} \leq (1-\eta)^{-2} + S^{(1)}_{\widehat{p}_\eta}.$$

Hence, $\left\|\widehat{\mathbf{Z}}_\eta\right\|_\infty \leq (1-\eta)^{-2} + S^{(1)}_{\hat{p}_\eta}$. Similarly, $\left\|\overline{\mathbf{Z}}_\eta\right\|_1 \leq (1-\eta)^{-2} + S^{(1)}_{\hat{p}_\eta}$, which completes the proof. ■

Lemma 25 *For any $u, u', v, v' \in \Sigma^*$, $|\mathbb{E}(\widehat{\mathbf{Z}}_\eta[u,v]\widehat{\mathbf{Z}}_\eta[u',v'])| \leq K_\eta \min\{\widehat{p}(uv), \widehat{p}(u'v')\}$.*

Proof This a corollary of Lemmas 12 and 23 with $X \coloneqq \sum_{x,y \in \Sigma^*} \eta^{|xy|} \mathbf{1}_{\xi = xuvy}$, $Y \coloneqq \sum_{x,y \in \Sigma^*} \eta^{|xy|} \mathbf{1}_{\xi = xu'v'y}$ and $M = K_\eta$. ■

Let $\widehat{\mathbf{X}}_\eta$ be the dilation of $\widehat{\mathbf{Z}}_\eta$. We can now propose bounds for $\mathbb{E}\,\widehat{\mathbf{X}}^2_\eta$ and $\mathbb{E}\,Tr(\widehat{\mathbf{X}}^2_\eta)$.

Lemma 26

$$\left\|\mathbb{E}\widehat{\mathbf{X}}^2_\eta\right\| \leq K_\eta S^{(2)}_{\widehat{p}_\eta} \text{ and } Tr(\mathbb{E}(\widehat{\mathbf{X}}^2_\eta)) \leq 2K_\eta S^{(2)}_{\widehat{p}_\eta}.$$

Proof Indeed,

$$||\mathbb{E}(\widehat{\mathbf{Z}}\widehat{\mathbf{Z}}^\top)||_\infty \leq \max_u \sum_{u',v} |\mathbb{E}(\widehat{\mathbf{Z}}_\eta[u,v]\widehat{\mathbf{Z}}_\eta[u',v])| \leq K_\eta \sum_{u',v} \widehat{p}(u'v) \leq K_\eta S^{(2)}_{\widehat{p}_\eta}.$$

We have also

$$||\mathbb{E}(\widehat{\mathbf{Z}}\widehat{\mathbf{Z}}^\top)||_1 \leq K_\eta S^{(2)}_{\widehat{p}_\eta} \text{ and therefore } ||\mathbb{E}(\widehat{\mathbf{Z}}\widehat{\mathbf{Z}}^\top)|| \leq K_\eta S^{(2)}_{\widehat{p}_\eta}.$$

Similar computations provide all inequalities. ■

5.5 The factor Hankel Matrix $H^{U,V}_{\widehat{p}}$ is not subexponential: proof

$\widehat{\mathbf{Z}}(\xi)$ can be not a subexponential random matrix. Let $\Sigma = \{a\}$ be a one-letter alphabet and let p be the rational stochastic language defined by $p(a^n) = 2^{-(n+1)}$. When $\xi = a^n$, $\widehat{\mathbf{H}}_\xi$ is the matrix defined by $\widehat{\mathbf{H}}_\xi[i,j] = n+1-(i+j)$ if $i+j \leq n$ and 0 otherwise. Let $\widehat{\mathbf{H}}_n \in \mathbb{R}^{n \times n}$ be the nonnegative symmetric matrix defined by

$$\widehat{\mathbf{H}}_n = \begin{pmatrix} n & n-1 & \cdots & \cdots & 1 \\ n-1 & n-2 & \cdots & 1 & 0 \\ \vdots & \vdots & \iddots & & \vdots \\ \vdots & 1 & & & \\ 1 & 0 & \cdots & & 0 \end{pmatrix}$$

It can easily be deduced from the definition of a subexponential random matrix that if $\widehat{\mathbf{Z}}(\xi)$ were subexponential then, there would exist constants $C, R > 0$ such that for every integer k,

$$\max_{n\geq 0} 2^{-n}\|\widehat{\mathbf{H}}_n^k\| \leq \|\sum_{n\geq 0} 2^{-n}\widehat{\mathbf{H}}_n^k\| \leq Ck!R^k.$$

Proposition 27 *$\widehat{\mathbf{Z}}(\xi)$ is not subexponential.*

We need the following Lemma.

Lemma 28 *Bo (2000) Let A be a $n \times n$ nonnegative symmetric matrix with positive row sums $d_1, \ldots, d_n$. Then, the spectral radius of A satisfies $\rho(A) \geq n^{-1/2}\sqrt{\sum_{i=1}^n d_i^2}$.*

Proof (of Proposition 27) Lemma 28 applied to the matrix $\widehat{\mathbf{H}}_n$ gives that $\rho(\widehat{\mathbf{H}}_n) = \Omega(n^2)$. Indeed, we have the row sums d_i satisfy $d_i \geq (n+1-i)^2/2$ and $\sum_{i=1}^n d_i^2 = \Theta(n^5)$. Hence, for every integer k, $\|\widehat{\mathbf{H}}_n^k\| \geq n^{2k}$. Now, taking $n = k$, we should have $2^{-k}k^{2k} \leq CR^k k! \leq CR^k k^k$ for every integer k, which is false. ■

6. Experiments

The theoretical bounds described in the previous Sections have been evaluated on the benchmark of PAutomaC (Verwer et al., 2012).

6.1 Presentation of the benchmark

The benchmark of PAutomaC provides samples of strings generated from probabilistic automata and designed to evaluate probabilistic automata learning. We have selected eleven problems from that benchmark, for which the sparsity of the Hankel matrices makes the use of standard SVD algorithms available from `NumPy` or `SciPy` possible. Table 1 provides some information about the selected problems.

- Target models are of different types: non deterministic probabilistic automata (PA), deterministic probabilistic automata (DPA) and hidden Markov models (HMM). Each target is a rational stochastic language. We display the size of the corresponding alphabet, its 2nd and 3rd moments and the spectral radius ρ of $\mathbf{M}_\Sigma$[3], for a minimal representation $\langle I, \mathbf{M}, T\rangle$ of p. We display the number of states of the target automaton and the rank of the corresponding Hankel matrix computed using `NumPy`: the true rank of the target lies between these two values. We also provide constants C_p and ρ_p satisfying $p(\Sigma^n) \leq C_p\rho_p^n$ for any integer n[4].

3. Since the matrices $\mathbf{M}_\Sigma$ corresponding to two minimal representations are similar, the spectral radius ρ only depends on the underlying rational series.
4. From Gelfand's formula, $\|\mathbf{M}_\Sigma^k\|^{1/k}$ converges to ρ when $k \to \infty$. For any k satisfying $\|\mathbf{M}_\Sigma^k\|^{1/k} < 1$, we can take $\rho_p = \|\mathbf{M}_\Sigma^k\|^{1/k}$ and $C_p = \max_{0\leq r<k} \min_{0\leq s\leq r} \|I^\top \mathbf{M}^s\| \cdot \|\mathbf{M}^{r-s}T\|$. We have noted in practice that when k increases, ρ_p decreases to ρ while C_p increases very slowly. We have uniformly taken the values computed for $k = 100$.

- Each problem comprises a sample S of strings independently drawn from the target. We provide the cardinal of S, the maximal length and the average length of strings in S.

- The empirical Hankel matrices are built on the prefixes, suffixes or factors of elements of S. We provide their size computed as the product of the number of non null rows by the number of non null columns. Almost all their cells are null: we provide the sparsity ratio.

Table 1: The 11 selected problems. The size of the Hankel matrices matrices is expressed in billions, where 1 g stands for 10^9. Sparsity indicates the ratio of non zero entries in the matrix: for example, there are $5.3 \times 1.9\ 10^4$ non empty cells in $\mathbf{H}_S$ for pb 3.

Pautomac ID	**3**	**4**	**7**	**15**	**25**	**29**	**31**	**38**	**39**	**40**	**42**
Target											
Type	PA	PA	DPA	PA	HMM	PA	PA	HMM	PA	DPA	DPA
$\|\Sigma\|$	4	4	13	14	10	6	5	10	14	14	9
NbStates	25	12	12	26	40	36	12	14	6	65	6
Rank	25	10	12	26	28	36	12	13	6	65	6
$S_p^{(2)}$	8.23	6.25	6.52	13.40	10.65	6.35	6.97	8.09	8.82	9.74	7.39
$S_p^{(3)}$	57.84	31.06	29.61	160.92	93.34	38.11	43.53	65.87	90.81	111.84	62.11
$\rho(M_\Sigma)$	0.85	0.77	0.72	0.92	0.88	0.83	0.84	0.88	0.90	0.91	0.88
ρ_p	0.87	0.79	0.73	0.95	0.92	0.87	0.86	0.91	0.92	0.96	0.90
C_p	0.24	0.42	0.80	0.09	0.21	0.37	0.23	0.29	0.29	0.26	0.25
Sample											
$\|S\|$	20 **k**	100 **k**	20 **k**	20 **k**	20 **k**	20 **k**	20 **k**	20 **k**	20 **k**	20 **k**	20 **k**
Avg $\|w\|$	7.22	5.26	5.52	12.46	9.72	5.29	6.00	7.18	7.74	8.72	6.36
max $\|w\|$	67	55	36	110	90	59	59	84	106	106	70
Hankel mat.											
$\|Pref\| \times \|Suff\|$	1.9g	0.6g	0.2g	28g	13g	0.5g	1.4g	8.0g	7.7g	15g	3.4g
H_S *sparsity*	5.3e-5	1.9e-4	2.1e-4	9.0e-6	1.5e-5	1.2e-4	6.1e-5	1.8e-5	1.9e-5	1.1e-5	3.3e-5
$\|Pref\| \times \|Fact\|$	11g	1.8g	0.7g	291g	99g	2.4g	7.6g	60g	76g	165g	25g
$\overline{H}_S$ *sparsity*	5.8e-5	1.9e-4	2.1e-4	1.0e-6	1.6e-5	1.2e-4	6.6e-5	1.9e-5	2.0e-5	1.2e-5	3.5e-5
$\|Fact\| \times \|Fact\|$	73g	6.4g	3g	3363g	797g	15.7g	44g	460g	761g	1925g	202g
$\widehat{H}_S$ *sparsity*	5.8e-5	2.0e-4	2.0e-4	1.0e-6	1.6e-5	1.2e-4	6.9e-5	2.0e-5	2.0e-5	1.2e-5	3.6e-5

Figure 1 shows the typical behavior of $S^{(1)}_{\overline{p}_\eta}$ and $S^{(1)}_{\widehat{p}_\eta}$, similar for all the problems.

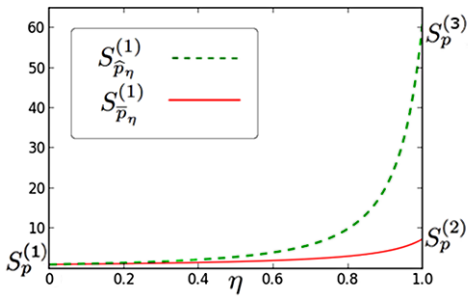

Figure 1: Behavior of $S^{(1)}_{\overline{p}_\eta}$ and $S^{(1)}_{\widehat{p}_\eta}$ for $\eta \in [0;1]$.

6.2 Accuracy of the bounds

For each problem, the exact value of $\|\mathbf{H}_S^{U,V} - \mathbf{H}_p^{U,V}\|_2$ is computed for sets U and V of the form $\Sigma^{\leq l}$, where we have maximized l according to our computing resources. This value is compared to the bounds provided by Theorem 7 and Equation (3), with $\delta = 0.05$ (Table 2). The optimized bound ("opt."), refers to the case where σ^2 has been calculated over $U \times V$ rather than $\Sigma^* \times \Sigma^*$ (see the remark at the end of Section 4.1). Tables 3 and 4 show analog comparisons for the prefix and the factor cases with different values of η. We can remark that our dimension-free bounds are significantly more accurate than the one provided by Equation (3). Similar results have been obtained for all the problems of PautomaC.

We have not reported experimental results based on Theorem 10, as for all the problems we consider, the constant σ is extremely large. For example, on Problem 3, with $\beta = 1 - \rho_p$, and using the parameters of Table 1, we have $\sigma \simeq 5308$, which would provide non significant accuracy values.

Table 2: Concentration bounds for $\|\mathbf{H}_S^{U,V} - \mathbf{H}_p^{U,V}\|$ where $U = V = \Sigma^{\leq l}$.

Problem number	3	4	7	15	25	29	31	38	39	40	42
l	8	9	8	5	5	9	7	4	6	4	7
$\|\mathbf{H}_S^{U,V} - \mathbf{H}_p^{U,V}\|$	0.005	0.003	0.006	0.004	0.003	0.005	0.005	0.006	0.005	0.004	0.005
Eq. (3)	0.100	0.039	0.088	0.127	0.115	0.088	0.092	0.097	0.103	0.105	0.095
Th. 7	0.067	0.026	0.060	0.085	0.076	0.059	0.062	0.066	0.069	0.073	0.063
Th. 7 (opt.)	0.048	0.023	0.053	0.028	0.032	0.047	0.044	0.028	0.033	0.024	0.038

Table 3: Concentration bounds for $\|\overline{\mathbf{H}}_S^{U,V} - \mathbf{H}_{\overline{p},\eta}^{U,V}\|$ (prefix case) where $U = V = \Sigma^{\leq l}$. The first part of the array is computed for $\eta = 1/2$, the second part for $\eta = 1$. The limiting case $\eta = 1$ (Th. 9 (opt.)) uses the remark at the end of Section 4.3

Problem number	3	4	7	15	25	29	31	38	39	40	42
l	8	9	8	5	5	9	7	4	6	4	7
$\|\overline{\mathbf{H}}_S^{U,V} - \mathbf{H}_{\overline{p},1/2}^{U,V}\|$	0.007	0.004	0.009	0.004	0.004	0.006	0.007	0.006	0.006	0.004	0.006
Eq. (3)	0.140	0.052	0.121	0.186	0.164	0.121	0.126	0.136	0.148	0.154	0.134
Th. 9	0.089	0.034	0.078	0.116	0.103	0.077	0.081	0.088	0.093	0.098	0.084
Th. 9 (opt.)	0.064	0.030	0.070	0.040	0.046	0.062	0.058	0.037	0.043	0.032	0.050
$\|\overline{\mathbf{H}}_S^{U,V} - \mathbf{H}_{\overline{p},1}^{U,V}\|$	0.014	0.006	0.022	0.012	0.015	0.012	0.018	0.013	0.014	0.009	0.013
Eq. (3)	0.281	0.089	0.200	0.476	0.361	0.229	0.241	0.291	0.355	0.387	0.293
Th. 9 (opt.)	0.128	0.052	0.117	0.106	0.106	0.118	0.110	0.078	0.102	0.0761	0.108

6.3 Implication for learning

The theoretical results of the last sections show that $\left\|\mathbf{H}_S^{U,V} - \mathbf{H}^{U,V}\right\|$, and similar expressions for other variants of the Hankel matrices, are bounded by a term that converges to 0 as the size of S increases, and is independent from U and V. This entails that the spectral learning

Table 4: Concentration bounds for $\|\widehat{\mathbf{H}}_S^{U,V} - \mathbf{H}_{\hat{p},\eta}^{U,V}\|$ (factor case) where $U = V = \Sigma^{\leq l}$ and $\eta = 1/e$.

Problem number	3	4	7	15	25	29	31	38	39	40	42
l	6	7	5	4	4	6	6	4	4	4	5
$\|\widehat{\mathbf{H}}_S^{U,V} - \mathbf{H}_{\widehat{p},1/e}^{U,V}\|$	0.007	0.003	0.007	0.004	0.003	0.005	0.007	0.006	0.007	0.005	0.006
Eq. (3)	0.148	0.056	0.127	0.206	0.177	0.130	0.138	0.152	0.155	0.177	0.142
Th. 11	0.099	0.037	0.086	0.129	0.114	0.0845	0.090	0.098	0.103	0.109	0.093
Th. 11 (opt.)	0.060	0.030	0.062	0.036	0.041	0.056	0.059	0.0401	0.036	0.035	0.044

algorithm is consistent, whatever sets U and V are chosen, as soon as the rank of $\mathbf{H}^{U,V}$ is equal to the rank of the target, and even if we set $U = V = \Sigma^*$. But these concentration bounds give no indication of what should be done in pratical cases.

The spectral learning algorithm first computes the r-first right singular vectors $R_S^{U,V}$ of $\mathbf{H}_S^{U,V}$ and then build a linear representation from $R_S^{U,V}$. Since an exact linear representation of the target can be computed from the r-first right singular vectors $R^{U,V}$ of $\mathbf{H}^{U,V}$, where r is the rank of the target, the distance between the linear spaces spanned by $R_S^{U,V}$ and $R^{U,V}$ seems to be a relevant measure to evaluate the impact on learning of the choice of U and V.

There are several ways to measure the distance between two linear spaces. Most of them are based on the principal angles $\theta_1 \geq \theta_2 \geq \ldots \geq \theta_r$ between them. The largest principal angle θ_1 is a harsh measure since, even if the two spaces coincide along the last $r-1$ principal angles, the distance between the two spaces can be large. We have considered the following measure

$$d(span(R^{U,V}), span(R_S^{U,V})) = 1 - \frac{1}{r}\sum_{i=1}^{r} \cos\theta_i \tag{10}$$

which is equal to 0 if the spaces coincide and 1 if they are completely orthogonal, and which takes into account all the principal angles.

The tables 5 to 9 show the distance between $span(R^{U,V})$ and $span(R_S^{U,V})$ for p, $\overline{p}_{1/2}$, $\overline{p}$, $\widehat{p}_{1/2}$ and $\widehat{p}$, and for matrices having 100 columns and a variable number of rows, from 100 to 20,000 (i.e. $|V| = 100$ and $100 \leq |U| \leq 20,000$).

These tables show that

- the distance between the empirical and true singular vector spaces is smaller for the factor variant than for the prefix variant, and smaller for the prefix variant than for the classical Hankel matrices
- for both the prefix and factor variants, the distance is smaller for $\eta = 1$ than for $\eta = 1/2$ (and $\eta = 0$)
- in most cases, the distance computed for $|U| = 20,000$ is either minimal or not very far from the minimum.

We have run similar experiences for increasing values of $|V|$, from 100 to 20,000. The tables are very similar but the distances systematically increase with V. Table 10 shows the

Table 5: Distance between the spaces spanned by the r first right singular vectors of $\mathbf{H}^{U,V}$ and $\mathbf{H}_S^{U,V}$ for $|V| = 100$ and $100 \leq |U| \leq 20,000$. Entries must be scaled by 10^{-1}.

	3	4	7	15	25	29	31	38	39	40	42
100	2.096	0.011	0.841	1.643	4.171	1.495	0.985	3.375	0.132	1.789	0.031
200	2.079	0.011	0.005	1.466	3.988	1.512	0.902	3.304	0.091	1.609	0.031
500	1.934	0.011	0.004	1.417	3.901	1.457	0.917	2.915	0.104	1.593	0.029
1000	1.883	0.010	0.004	1.421	3.530	1.363	0.908	2.578	0.108	1.501	0.029
2000	1.813	0.010	0.004	1.382	3.529	1.358	0.919	2.511	0.125	1.512	0.029
5000	1.766	0.010	0.004	1.423	3.442	1.134	0.940	2.380	0.172	1.485	0.029
10000	1.755	0.010	0.004	1.424	3.431	1.136	0.961	2.284	0.269	1.390	0.029
20000	1.739	0.011	0.004	1.401	3.257	1.283	0.986	2.051	0.728	1.457	0.029

Table 6: Distance between the spaces spanned by the r first right singular vectors of $\mathbf{H}_{\overline{p}_{0.5}}^{U,V}$ and $\overline{\mathbf{H}}_{0.5,S}^{U,V}$ for $|V| = 100$ and $100 \leq |U| \leq 20,000$. Entries must be scaled by 10^{-1}.

	3	4	7	15	25	29	31	38	39	40	42
100	1.880	0.010	0.535	1.683	4.211	1.342	0.304	2.931	0.077	1.782	0.015
200	1.717	0.009	0.004	1.631	3.928	1.281	0.251	2.867	0.049	1.730	0.015
500	1.646	0.009	0.004	1.500	3.847	1.161	0.284	2.590	0.047	1.578	0.013
1000	1.623	0.009	0.003	1.558	3.537	1.137	0.297	2.443	0.047	1.529	0.013
2000	1.549	0.009	0.003	1.504	3.514	1.107	0.329	2.391	0.047	1.524	0.013
5000	1.491	0.009	0.003	1.501	3.317	1.098	0.439	2.129	0.048	1.486	0.013
10000	1.459	0.009	0.003	1.495	3.284	1.059	0.783	1.785	0.049	1.450	0.013
20000	1.425	0.009	0.003	1.548	3.203	0.950	0.955	1.690	0.052	1.424	0.013

Table 7: Distance between the spaces spanned by the r first right singular vectors of $\mathbf{H}_{\overline{p}}^{U,V}$ and $\overline{\mathbf{H}}_S^{U,V}$ for $|V| = 100$ and $100 \leq |U| \leq 20,000$. Entries must be scaled by 10^{-1}.

	3	4	7	15	25	29	31	38	39	40	42
100	1.281	0.010	0.554	1.729	4.095	1.221	0.224	2.778	0.009	1.662	0.005
200	1.185	0.008	0.007	1.472	4.016	1.237	0.204	2.732	0.006	1.497	0.005
500	1.074	0.007	0.006	1.388	3.886	1.194	0.238	2.619	0.006	1.352	0.005
1000	1.060	0.007	0.006	1.358	3.686	1.227	0.255	2.391	0.006	1.313	0.005
2000	1.054	0.007	0.006	1.179	3.682	1.174	0.288	2.349	0.006	1.341	0.005
5000	1.063	0.007	0.006	1.169	3.347	1.164	0.313	2.263	0.006	1.289	0.005
10000	1.073	0.007	0.006	1.181	3.244	1.165	0.332	2.156	0.006	1.347	0.005
20000	1.088	0.007	0.006	1.213	3.100	1.165	0.357	1.825	0.006	1.356	0.005

Table 8: Distance between the spaces spanned by the r first right singular vectors of $\mathbf{H}^{U,V}_{\widehat{p}_{0.5}}$ and $\widehat{\mathbf{H}}^{U,V}_{0.5,S}$ for $|V| = 100$ and $100 \leq |U| \leq 20,000$. Entries must be scaled by 10^{-1}.

	3	4	7	15	25	29	31	38	39	40	42
100	1.917	0.009	0.004	1.317	3.507	1.229	0.273	2.814	0.082	1.780	0.013
200	1.832	0.008	0.005	1.328	3.445	1.234	0.266	2.740	0.072	1.683	0.013
500	1.796	0.008	0.004	1.344	3.410	1.231	0.284	2.515	0.069	1.600	0.009
1000	1.781	0.008	0.003	1.363	3.312	1.164	0.288	2.344	0.048	1.585	0.009
2000	1.738	0.008	0.003	1.324	3.281	1.221	0.311	2.319	0.048	1.596	0.009
5000	1.727	0.007	0.003	1.323	3.262	1.137	0.315	2.247	0.047	1.475	0.009
10000	1.718	0.007	0.003	1.321	3.156	1.094	0.319	2.164	0.047	1.480	0.009
20000	1.652	0.007	0.003	1.351	3.110	1.065	0.324	2.119	0.047	1.452	0.009

Table 9: Distance between the spaces spanned by the r first right singular vectors of $\mathbf{H}^{U,V}_{\widehat{p}}$ and $\widehat{\mathbf{H}}^{U,V}_{S}$ for $|V| = 100$ and $100 \leq |U| \leq 20,000$. Entries must be scaled by 10^{-1}.

	3	4	7	15	25	29	31	38	39	40	42
100	1.185	0.004	0.005	0.557	2.264	0.901	0.237	1.977	0.004	1.512	0.002
200	1.065	0.004	0.005	0.480	2.153	0.893	0.244	1.922	0.003	1.307	0.002
500	1.064	0.004	0.004	0.524	2.162	0.797	0.254	1.783	0.003	1.290	0.002
1000	1.048	0.004	0.004	0.529	2.098	0.784	0.260	1.613	0.002	1.294	0.002
2000	1.029	0.004	0.004	0.537	2.075	0.803	0.267	1.597	0.002	1.268	0.002
5000	1.011	0.004	0.004	0.570	2.085	0.822	0.280	1.560	0.002	1.278	0.002
10000	1.012	0.005	0.004	0.584	2.072	0.882	0.287	1.513	0.002	1.278	0.002
20000	1.011	0.005	0.004	0.620	2.072	0.914	0.297	1.499	0.002	1.279	0.002

Table 10: Distance between the spaces spanned by the r first right singular vectors of $\mathbf{H}^{U,V}$ and $\mathbf{H}^{U,V}_{S}$ for $100 \leq |U| \leq 20,000$ and $100 \leq |V| \leq 20,000$ for problem 3. Entries must be scaled by 10^{-1}.

	100	300	1,000	2,000	5,000	20,000
100	2.096	2.399	2.660	2.747	2.887	3.116
200	2.079	2.292	2.512	2.608	2.752	2.974
500	1.934	2.196	2.350	2.431	2.575	2.808
1000	1.883	2.134	2.306	2.401	2.543	2.767
2,000	1.813	2.052	2.239	2.331	2.475	2.711
5,000	1.766	1.981	2.183	2.274	2.410	2.664
10,000	1.755	1.924	2.132	2.221	2.349	2.615
20,000	1.739	1.876	2.077	2.159	2.276	2.543

results for $d(span(\mathbf{H}^{U,V}), span(\mathbf{H}_S^{U,V}))$ computed on problem 3 - the behavior is similar for all the problems and all other variants.

These experiments call for the following recommendations that remain to be confirmed by further theoretical studies.

- use the data to infer first the factor variant of p, rather than the prefix variant or p itself,
- use a small number of columns,
- use as many rows as available, unless a specific information on the domain indicate to bound their number.

7. Conclusion

We have provided dimension-free concentration inequalities for Hankel matrices in the context of spectral learning of rational stochastic languages. These bounds cover 3 cases, each one corresponding to a specific way to exploit the strings under observation, paying attention to the strings themselves, to their prefixes or to their factors. For the last two cases, we introduced parametrized variants which allow a trade-off between the rate of the concentration and the exploitation of the information contained in data.

Experiments on the PAutomaC benchmark show that our dimension-free bounds are quite tight (except the subexponential bound for the prefix variant) and significantly more accurate than the bounds provided by classically used dimension-dependent bounds.

A consequence of these dimension-free inequalities is that the spectral learning algorithm is consistent, even if the whole empirical Hankel matrix is used, suggesting that the choice of relevant sets of rows and columns is maybe not critical. However, they do not provide any indication on what should be done in pratical cases. Experiments indicate that the singular vector spaces computed from the empirical Hankel matrices converge more quickly to the true singular vector spaces for the factor variant of the Hankel matrix - which is consistent with the experiments in (Balle et al., 2014), a small number of columns and a large number of rows. It would be interesting to obtain concentration results which confirm these practice. Another research direction would be to link up the prefix and factor cases to concentration bounds for sum of random tensors and to generalize the results to the case where a fixed number ≥ 1 of factors is considered for each string.

Acknowledgments

This work has been carried out thanks to the support of the ARCHIMEDE Labex (ANR-11-LABX- 0033) and the A*MIDEX project (ANR-11-IDEX-0001-02) funded by the "Investissements d'Avenir" French government program managed by the ANR.

References

R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

A. Anandkumar, D.P. Foster, D. Hsu, S. Kakade, and Y.-K. Liu. A spectral algorithm for latent dirichlet allocation. In *Proceedings of NIPS*, pages 926–934, 2012a.

A. Anandkumar, D. Hsu, F. Huang, and S. Kakade. Learning mixtures of tree graphical models. In *Proceedings of NIPS*, pages 1061–1069, 2012b.

A. Anandkumar, D. Hsu, and S.M. Kakade. A method of moments for mixture models and hidden markov models. *Proceedings of COLT - Journal of Machine Learning Research - Proceedings Track*, 23:33.1–33.34, 2012c.

R. Bailly. *Méthodes spectrales pour l'inférence grammaticale probabiliste de langages stochastiques rationnels*. PhD thesis, Aix-Marseille Université, 2011.

R. Bailly, F. Denis, and L. Ralaivola. Grammatical inference as a principal component analysis problem. In *Proceedings of ICML*, page 5, 2009.

R. Bailly, A. Habrard, and F. Denis. A spectral approach for probabilistic grammatical inference on trees. In *Proceedings of ALT*, pages 74–88, 2010.

B. Balle and M. Mohri. Spectral learning of general weighted automata via constrained matrix completion. In *Proceedings of NIPS*, pages 2168–2176, 2012.

B. Balle, A. Quattoni, and X. Carreras. A spectral learning algorithm for finite state transducers. In *Proceedings of ECML/PKDD (1)*, pages 156–171, 2011.

B. Balle, A. Quattoni, and X. Carreras. Local loss optimization in operator models: A new insight into spectral learning. In *Proceedings of ICML*, 2012.

B. Balle, X. Carreras, F. M. Luque, and A. Quattoni. Spectral learning of weighted automata: A forward-backward perspective. To appear in Machine Learning, 2013.

B. Balle, W. L. Hamilton, and J. Pineau. Methods of moments for learning stochastic languages: Unified presentation and empirical comparison. In *Proceedings of ICML*, 2014.

J. Berstel and C. Reutenauer. *Rational series and their languages*. EATCS monographs on theoretical computer science. Springer-Verlag, Berlin, New York, 1988. ISBN 0-387-18626-3. URL `http://opac.inria.fr/record=b1086956`. Translation of: Les sries rationnelles et leurs langages.

Z. Bo. On the spectral radius of nonnegative matrices. *Australasian Journal of Combinatorics*, 22:301–306, 2000.

S. B. Cohen, K. Stratos, M. Collins, D. P. Foster, and L. H. Ungar. Spectral learning of Latent-Variable PCFGs. In *ACL (1)*, pages 223–231. The Association for Computer Linguistics, 2012. ISBN 978-1-937284-24-4.

M. Droste, W. Kuich, and H. Vogler, editors. *Handbook of Weighted Automata*. Springer, 2009.

D. Hsu, S.M. Kakade, and T. Zhang. A spectral algorithm for learning hidden markov models. In *Proceedings of COLT*, 2009.

D. Hsu, S. M. Kakade, and T. Zhang. Dimension-free tail inequalities for sums of random matrices. *ArXiv e-prints*, 2011.

D. Hsu, S. Kakade, and T. Zhang. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electron. Commun. Probab.*, 17:no. 14, 1–13, 2012.

F.M. Luque, A. Quattoni, B. Balle, and X. Carreras. Spectral learning for non-deterministic dependency parsing. In *Proceedings of EACL*, pages 409–419, 2012.

A.P. Parikh, L. Song, and E.P. Xing. A spectral algorithm for latent tree graphical models. In *Proceedings of ICML*, pages 1065–1072, 2011.

A. Salomaa and M. Soittola. *Automata-theoretic aspects of formal power series.* Texts and monographs in computer science. Springer, 1978. ISBN 978-0-387-90282-1.

S. Siddiqi, B. Boots, and G.J. Gordon. Reduced-rank hidden Markov models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-2010)*, 2010.

L. Song, B. Boots, S.M. Siddiqi, G.J. Gordon, and A.J. Smola. Hilbert space embeddings of hidden markov models. In *Proceedings of ICML*, pages 991–998, 2010.

E. M. Stein and R. Shakarchi. *Real analysis : measure theory, integration, and Hilbert spaces*. Princeton lectures in analysis. Princeton University press, Princeton (N.J.), Oxford, 2005. ISBN 0-691-11386-6. URL `http://opac.inria.fr/record=b1133853`.

G. W. Stewart. Perturbation theory for the singular value decomposition. In *SVD and Signal Processing II: Algorithms, Analysis and Applications*, pages 99–109. Elsevier, 1990.

J.A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

R. Vershynin. *Compressed Sensing*, chapter 5. Introduction to the non-asymptotic analysis of random matrices, pages 210–268. Cambridge University Press, 2012.

S. Verwer, R. Eyraud, and C. de la Higuera. Results of the PAutomaC probabilistic automaton learning competition. *Journal of Machine Learning Research - Proceedings Track*, 21: 243–248, 2012.

8. Appendix

8.1 Proof of Proposition 1

On a one-letter alphabet, for any non negative rational convergent series r, the series $u \mapsto r(\Sigma^* u \Sigma^*)$ is rational. Indeed, $r(\Sigma^* u \Sigma^*) = r(u\Sigma^*) = \overline{r}(u)$ and $\overline{r}$ is rational. On the other hand, this property may be false as soon as the alphabet contains at least two letters.

Proposition 1 *There exists a rational stochastic language p of rank 1 and built over a two-letter alphabet such that the series $u \mapsto p(\Sigma^* u \Sigma^*)$ is not rational.*

Proof Let $\Sigma = \{a, b\}$ and p be the rational stochastic language defined by $p(u) \coloneqq \alpha^{|u|_a} \beta^{|u|_b} \gamma$ where $\alpha, \beta, \gamma > 0$, $\alpha + \beta + \gamma = 1$ and where $|u|_x$ denotes the number of occurrences of the letter $x \in \Sigma$ in u. We have

$$p(\Sigma^*) = \sum_{u \in \Sigma^*} \alpha^{|u|_a} \beta^{|u|_b} \gamma = \gamma \sum_{n=0}^{\infty} \sum_{m=0}^{n} \binom{n}{m} \alpha^m \beta^{n-m} = \gamma \sum_{n=0}^{\infty} (\alpha + \beta)^n = \frac{\gamma}{1 - \alpha - \beta} = 1.$$

Let f be the series defined by $f(u) \coloneqq p(\Sigma^* u \Sigma^*)$. Let us compute $f(a^n)$ for any integer n. Clearly, $f(\varepsilon) = 1$. Let $n \geq 1$. We can write

$$\Sigma^* = \bigcup_{m=0}^{n-1} \{a^m\} \cup a^n \Sigma^* \cup \bigcup_{m=0}^{n-1} a^m b \Sigma^*$$

and

$$\begin{aligned} f(a^n) &= p(a^n \Sigma^*) + \sum_{m=0}^{n-1} p(a^m b \Sigma^* a^n \Sigma^*) \\ &= \alpha^n + \sum_{m=0}^{n-1} \alpha^m \beta p(\Sigma^* a^n \Sigma^*) \\ &= \alpha^n + \frac{1 - \alpha^n}{1 - \alpha} \beta f(a^n) \end{aligned}$$

and therefore,

$$f(a^n) = (1 - \alpha) \frac{\alpha^n}{\gamma + \beta \alpha^n}.$$

Suppose that f is rational. Then, every submatrix of the Hankel matrix of f is of finite rank. In particular, there exists an index k and real coefficients $\lambda_0, \lambda_1, \ldots, \lambda_{k-1}$ such that for any integer p,

$$f(a^{k+p}) = \sum_{i=0}^{k-1} \lambda_i f(a^{i+p})$$

which is equivalent to

$$\sum_{i=0}^{k-1} \lambda_i \alpha^{i-k} \frac{\gamma + \beta \alpha^{k+p}}{\gamma + \beta \alpha^{i+p}} = 1.$$

Let $g(z)$ be the complex function defined by

$$g(z) \coloneqq \left(\sum_{i=0}^{k-1} \mu_i \frac{\delta + \alpha^k z}{\delta + \alpha^i z} \right) - 1$$

where $\delta = \gamma/\beta$ and $\mu_i = \lambda_i \alpha^{i-k}$ for $0 \leq i \leq k-1$.

The function g as poles at $-\delta\alpha^{-i}$ and hence, is analytic on a neighborhood V of 0. Since $g(\alpha^p) = 0$ for any integer p, the principle of permanence shows that g is uniformly equal to 0 on V, i.e.

$$\sum_{i=0}^{k-1} \mu_i (\delta + \alpha^i z)^{-1} = (\delta + \alpha^k z)^{-1}, \forall z \in V.$$

In particular, these two functions and all their derivatives are equal for $z = 0$: we obtain the system

$$\sum_{i=0}^{k-1} \mu_i \alpha^{ih} = \alpha^{kh} \text{ for every } h \geq 0. \tag{11}$$

The Vandermonde matrix

$$\begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \alpha & \alpha^2 & \dots & \alpha^k \\ 1 & \alpha^2 & \alpha^4 & \dots & \alpha^{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha^k & \alpha^{2k} & \dots & \alpha^{k^2} \end{pmatrix}$$

has a non zero determinant since $\alpha^i \neq \alpha^j$ for $i \neq j$ and therefore, the unique solution of the system $\sum_{i=0}^{k} \mu_i \alpha^{ih} = 0$ for $0 \leq h \leq k$ is $\mu_0 = \mu_1 = \dots = \mu_k = 0$ and the system (11) has no solution. ■

8.2 Proof of Proposition 2

From the definition of $\mathbf{T}_s$, it can easily be shown that the mapping $s \mapsto \mathbf{T}_s$ is a morphism: $\mathbf{T}_{s_1}\mathbf{T}_{s_2}[u,v] = \sum_{w \in \Sigma^*} \mathbf{T}_{s_1}[u,w]\mathbf{T}_{s_2}[w,v] = 1$ iff $v = us_1s_2$ and 0 otherwise.
If $\mathbf{X}$ is a matrix whose rows are indexed by Σ^*, we have $\mathbf{T}_s\mathbf{X}[u,v] = \sum_w \mathbf{T}_s[u,w]\mathbf{X}[w,v] = \mathbf{X}[us,v]$: i.e. the rows of $\mathbf{T}_S\mathbf{X}$ are included in the set of rows of $\mathbf{X}$. Then, it follows from the definition of E that $E^\top\mathbf{T}_s$ is equal to the first row of $\mathbf{T}_s$ (indexed by ϵ) with all coordinates equal to zero except the one indexed by s which equal 1.

Now, from the reduced singular value decomposition of $\mathbf{H} = \mathbf{L}\mathbf{D}\mathbf{R}^\top$ at rank d, $\mathbf{R}$ is a matrix of dimension $\infty \times d$ whose columns form a set of orthonormal vectors - the right singular vectors of $\mathbf{H}$ - such that $\mathbf{R}^\top\mathbf{R} = \mathbf{I}_d$ and $\mathbf{R}\mathbf{R}^\top\mathbf{H}^\top = \mathbf{H}^\top$ ($\mathbf{R}\mathbf{R}^\top$ is the orthogonal projection on the subspace spanned by the rows of $\mathbf{H}$).
One can easily deduce, by a recurrence over n, that for every string $u = x_1 \dots x_n$,

$$(\mathbf{R}^\top\mathbf{T}_{x_1}\mathbf{R}) \circ \dots \circ (\mathbf{R}^\top\mathbf{T}_{x_n}\mathbf{R})\mathbf{R}^\top\mathbf{H}^\top = \mathbf{R}^\top\mathbf{T}_u\mathbf{H}^\top.$$

Indeed, the inequality is trivially true for $n = 0$ since $\mathbf{T}_\epsilon = \mathbf{I}_d$. Then, we have that $\mathbf{R}^\top\mathbf{T}_x\mathbf{R}\mathbf{R}^\top\mathbf{T}_u\mathbf{H}^\top = \mathbf{R}^\top\mathbf{T}_x\mathbf{T}_u\mathbf{H}^\top = \mathbf{R}^\top\mathbf{T}_{xu}\mathbf{H}^\top$ since the columns of $\mathbf{T}_u\mathbf{H}^\top$ are rows of $\mathbf{H}$ and $\mathbf{T}$ is a morphism.

If $P^\top$ is the first row of $\mathbf{H}$ then: $E^\top\mathbf{R}(\mathbf{R}^\top\mathbf{T}_{x_1}\mathbf{R})\circ\ldots\circ(\mathbf{R}^\top\mathbf{T}_{x_n}\mathbf{R})\mathbf{R}^\top P = E^\top\mathbf{T}_u P = r(u)$. Thus, $\langle\mathbf{R}^\top E, (\mathbf{R}^\top\mathbf{T}_x\mathbf{R})_{x\in\Sigma}, \mathbf{R}^\top P\rangle$ is a linear representation of r of dimension d.

8.3 Monotonicity in Stewart formula.

Let us first recall the min – max characterization of singular values derived from the Courant-Fischer Theorem: for any matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$,

$$\sigma_k(\mathbf{A}) = \min_{u_1,\ldots,u_{k-1}\in\mathbb{R}^n}\ \max_{x\in\mathbb{R}^n,\|x\|=1,x\in[u_1,\ldots,u_{k-1}]^\perp} \|\mathbf{A}x\|\,.$$

Let $\mathbf{B}$ be the result obtained by replacing all elements in the last column of $\mathbf{A}$ with 0 and let $u_k\in\mathbb{R}^n$ be defined by $u_k[i] = \mathbf{1}_{i=n}$.

$$\begin{aligned}\sigma_k(\mathbf{A}) &\geq \min_{u_1,\ldots,u_{k-1}\in\mathbb{R}^n}\ \max_{x\in\mathbb{R}^n,\|x\|=1,x\in[u_1,\ldots,u_k]^\perp} \|\mathbf{A}x\| \\ &= \min_{u_1,\ldots,u_{k-1}\in\mathbb{R}^n}\ \max_{x\in\mathbb{R}^n,\|x\|=1,x\in[u_1,\ldots,u_{k-1}]^\perp} \|\mathbf{B}x\| = \sigma_k(\mathbf{B}).\end{aligned}$$

A similar argument holds if we delete a row of $\mathbf{A}$. Then, it can easily be shown by induction that if $\mathbf{B}$ is obtained by deleting some rows and columns in $\mathbf{A}$, then $\sigma_k(\mathbf{A})\geq\sigma_k(\mathbf{B})$ (as far as $\sigma_k(\mathbf{B})$ is defined).

Therefore, if $U\subseteq U'$ and $V\subseteq V'$, then $\left\|\mathbf{H}_S^{U\times V} - \mathbf{H}_r^{U\times V}\right\| \leq \left\|\mathbf{H}_S^{U'\times V'} - \mathbf{H}_r^{U'\times V'}\right\|$ and $\sigma_{\min}(\mathbf{H}_r^{U\times V}) \leq \sigma_{\min}(\mathbf{H}_r^{U'\times V'})$.