On the Rademacher Complexity of Weighted Automata

Borja Balle[1] and Mehryar Mohri[2,3]

[1] School of Computer Science, McGill University, Montréal, Canada
[2] Courant Institute of Mathematical Sciences, New York, NY
[3] Google Research, New York, NY

Abstract. Weighted automata (WFAs) provide a general framework for the representation of functions mapping strings to real numbers. They include as special instances deterministic finite automata (DFAs), hidden Markov models (HMMs), and predictive states representations (PSRs). In recent years, there has been a renewed interest in weighted automata in machine learning due to the development of efficient and provably correct spectral algorithms for learning weighted automata. Despite the effectiveness reported for spectral techniques in real-world problems, almost all existing statistical guarantees for spectral learning of weighted automata rely on a strong realizability assumption. In this paper, we initiate a systematic study of the learning guarantees for broad classes of weighted automata in an agnostic setting. Our results include bounds on the Rademacher complexity of three general classes of weighted automata, each described in terms of different natural quantities. Interestingly, these bounds underline the key role of different data-dependent parameters in the convergence rates.

1 Introduction

Weighted finite automata (WFAs) provide a general and highly expressive framework for representing functions mapping strings to real numbers. The properties of WFAs or their mathematical counterparts, rational power series, have been extensively studied in the past [17, 33, 12, 25, 30]. WFAs have also been used in a variety of applications, including speech recognition [31], image compression [2], natural language processing [23], model checking [3], and machine translation [19]. See also [9] for a recent survey of algorithms for learning WFAs.

The recent developments in spectral learning [21, 4] have triggered a renewed interest in the use of WFAs in machine learning, with several recent successes in natural language processing [6, 7] and reinforcement learning [13, 20]. The interest in spectral learning algorithms for WFAs is driven by the many appealing theoretical properties of such algorithms, which include their polynomial-time complexity, the absence of local minima, statistical consistency, and finite sample bounds *à la* PAC [21]. However, the typical statistical guarantees given for the hypotheses used in spectral learning only hold in the realizable case. That is, these analyses assume that the labeled data received by the algorithm is sampled from some unknown WFA. While this assumption is a reasonable starting point for theoretical analyses, the results obtained in this setting fail to explain the good performance of spectral algorithms in many practical applications where the data is typically not generated by a WFA.

There exists of course a vast literature in statistical learning theory providing tools to analyze generalization guarantees for different hypothesis classes in classification, regression, and other learning tasks. These guarantees typically hold in an agnostic setting where the data is drawn i.i.d. from an arbitrary distribution. For spectral learning of WFAs, an algorithm-dependent agnostic generalization bound was proven in [8] using a stability argument. This seems to have been the first analysis to provide statistical guarantees for learning WFAs in an agnostic setting. However, while [8] proposed a broad family of algorithms for learning WFAs parametrized by several choices of loss functions and regularizations, their bounds hold only for one particular algorithm within this family.

In this paper, we start the systematic development of algorithm-independent generalization bounds for learning with WFAs, which apply to all the algorithms proposed in [8], as well as to others using WFAs as their hypothesis class. Our approach consists of providing upper bounds on the Rademacher complexity of general classes of WFAs. The use of Rademacher complexity to derive generalization bounds is standard [24] (see also [11] and [32]). It has been successfully used to derive statistical guarantees for classification, regression, kernel learning, ranking, and many other machine learning tasks (e.g. see [32] and references therein). A key benefit of Rademacher complexity analyses is that the resulting generalization bounds are data-dependent.

Our main results consist of upper bounds on the Rademacher complexity of three broad classes of WFAs. The main difference between these classes is the quantities used for their definition: the norm of the transition weight matrix or initial and final weight vectors of a WFA; the norm of the function computed by a WFA; and, the norm of the Hankel matrix associated to the function computed by a WFA. The formal definitions of these classes is given in Section 3. Let us point out that our analysis of the Rademacher complexity of the class of WFAs described in terms of Hankel matrices directly yields theoretical guarantees for a variety of spectral learning algorithms. We will return to this point when discussing the application of our results.

Related Work. To the best of our knowledge, this paper is the first to provide general tools for deriving learning guarantees for broad classes of WFAs. However, there exists some related work providing complexity bounds for some sub-classes of WFAs in agnostic settings. The VC-dimension of deterministic finite automata (DFAs) with n states over an alphabet of size k was shown by [22] to be in $O(kn \log n)$. For probabilistic finite automata (PFAs), it was shown by [1] that, in an agnostic setting, a sample of size $\widetilde{O}(kn^2/\varepsilon^2)$ is sufficient to learn a PFA with n states and k symbols whose log-loss error is at most ε away from the optimal one in the class. Learning bounds on the Rademacher complexity of DFAs and PFAs follow as straightforward corollaries of the general results we present in this paper.

Another recent line of work, which aims to provide guarantees for spectral learning of WFAs in the non-realizable setting, is the so-called low-rank spectral learning approach [27]. This has led to interesting upper bounds on the approximation error between minimal WFAs of different sizes [26]. See [10] for a polynomial-time algorithm for computing these approximations. This approach, however, is more limited than ours for two reasons. First, because it is algorithm-dependent. And second, because it assumes that the data is actually drawn from some (probabilistic) WFA, albeit one that is larger than any of the WFAs in the hypothesis class considered by the algorithm.

The following sections of this paper are organized as follows. Section 2 introduces the notation and technical concepts used throughout. Section 3 describes the three classes of WFAs for which we provide Rademacher complexity bounds, and gives a brief overview of our results. Our learning bounds are formally stated and proven in Sections 4, 5, and 6.

2 Preliminaries and Notation

2.1 Weighted Automata, Rational Functions, and Hankel Matrices

Let Σ be a finite alphabet of size k. Let ϵ denote the empty string and Σ^* the set of all finite strings over the alphabet Σ. The length of $u \in \Sigma^*$ is denoted by $|u|$. Given an integer $L \geq 0$, we denote by $\Sigma^{\leq L}$ the set of all strings with length at most L: $\Sigma^{\leq L} = \{x \in \Sigma^* \colon |x| \leq L\}$.

A WFA over the alphabet Σ with $n \geq 1$ states is a tuple $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_a\}_{a \in \Sigma} \rangle$ where $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$ are the initial and final weights, and $\mathbf{A}_a \in \mathbb{R}^{n \times n}$ the transition matrix whose entries give the weights of the transitions labeled with a. Every WFA A defines a function $f_A \colon \Sigma^* \to \mathbb{R}$ defined for all $x = a_1 \cdots a_t \in \Sigma^*$ by

$$f_A(x) = f_A(a_1 \cdots a_t) = \boldsymbol{\alpha}^\top \mathbf{A}_{a_1} \cdots \mathbf{A}_{a_t} \boldsymbol{\beta} = \boldsymbol{\alpha}^\top \mathbf{A}_x \boldsymbol{\beta} \enspace , \tag{1}$$

where $\mathbf{A}_x = \mathbf{A}_{a_1} \cdots \mathbf{A}_{a_t}$. A function $f \colon \Sigma^* \to \mathbb{R}$ is said to be *rational* if there exists a WFA A such that $f = f_A$. The rank of f is denoted by $\mathrm{rank}(f)$ and defined as the minimal number of states of a WFA A such that $f = f_A$. Note that minimal WFAs are not unique. In fact, it is not hard to see that, for any minimal WFA $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_a\} \rangle$ with $f = f_A$ and any invertible matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$, $A^{\mathbf{Q}} = \langle \mathbf{Q}^\top \boldsymbol{\alpha}, \mathbf{Q}^{-1} \boldsymbol{\beta}, \{\mathbf{Q}^{-1} \mathbf{A}_a \mathbf{Q}\} \rangle$ is also a minimal WFA computing f. We will sometimes write $A(x)$ instead of $f_A(x)$ to emphasize the fact that we are considering a specific parametrization of f_A. Note that for the purpose of this paper we only consider weighted automata over the familiar field of real numbers with standard addition and multiplication (see [17, 33, 12, 25, 30] for more general definitions of WFAs over arbitrary semirings). Functions mapping strings to real numbers can also be viewed as non-commutative formal power series, which often helps deriving rigorous proofs in formal language theory [33, 12, 25]. We will not favor that point of view here, however, since we will not make use of the algebraic properties offered by that perspective.

An alternative method to represent rational functions independently of any WFA parametrization is via their *Hankel matrices.* The Hankel matrix $\mathbf{H}_f \in \mathbb{R}^{\Sigma^* \times \Sigma^*}$ of a function $f \colon \Sigma^* \to \mathbb{R}$ is the infinite matrix with rows and columns indexed by all strings with $\mathbf{H}_f(u, v) = f(uv)$ for all $u, v \in \Sigma^*$. By the theorem of Fliess [18] (see also [14] and [12]), $\mathbf{H}_f$ has finite rank n if and only if f is rational and there exists a WFA A with n states computing f, that is, $\text{rank}(f) = \text{rank}(\mathbf{H}_f)$.

2.2 Rademacher Complexity

Our objetive is to derive learning guarantees for broad families of weighted automata or rational functions used as hypothesis sets in learning algorithms. To do so, we will derive upper bounds on the Rademacher complexity of different classes $\mathcal{F}$ of rational functions $f \colon \Sigma^* \to \mathbb{R}$. Thus, we first briefly introduce the definition of the Rademacher complexity of an arbitrary class of functions $\mathcal{F}$. Let D be a probability distribution over Σ^*. Suppose $S = (x_1, \ldots, x_m) \overset{\text{iid}}{\sim} D^m$ is a sample of m i.i.d. strings drawn from D. The *empirical Rademacher complexity* of $\mathcal{F}$ on S is defined as follows:

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i)\right] \ ,$$

where the expectation is taken over the m independent Rademacher random variables $\sigma_i \sim \mathbf{Unif}(\{+1, -1\})$. The *Rademacher complexity* of $\mathcal{F}$ is defined as the expectation of $\widehat{\mathfrak{R}}_S(\mathcal{F})$ over the draw of a sample S of size m:

$$\mathfrak{R}_m(\mathcal{F}) = \mathop{\mathbb{E}}_{S \sim D^m}\left[\widehat{\mathfrak{R}}_S(\mathcal{F})\right] \ .$$

Rademacher complexity bounds can be directly used to derive data-dependent generalization bounds for a variety of learning tasks [24, 11, 32]. Since the derivation of these learning bounds from Rademacher complexity bounds is now standard and depends on the learning task, we will not provide them here explicitly. Instead, we will discuss multiple applications of our techniques in an extended version of this paper, which will also contain explicit generalization bounds for several set-ups relevant to practical applications.

3 Classes of Rational Functions

In this section, we introduce three different classes of rational functions described in terms of distinct quantities. These quantities, such as the number of states of a WFA representation, the norm of the rational function, or that of its Hankel matrix, control the complexity of the classes of rational functions in distinct ways and each class admits distinct benefits in the analysis of learning with WFAs.

3.1 The Class $\mathcal{A}_{n,p,r}$

We start by considering the case where each rational function is given by a fixed WFA representation. Our learning bounds would then naturally depend on the number of states and the weights of the WFA representations.

Fix an integer $n > 0$ and let $\mathcal{A}_n$ denote the set of all WFAs with n states. Note that any $A \in \mathcal{A}_n$ is identified by the $d = n(kn+2)$ parameters required to specify its initial, final, and transition weights. Thus, we can identify $\mathcal{A}_n$ with the vector space $\mathbb{R}^d$ by suitably defining addition and scalar multiplication. In particular, given $A, A' \in \mathcal{A}_n$ and $c \in \mathbb{R}$, we define:

$$\begin{aligned} A + A' &= \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_a\}\rangle + \langle \boldsymbol{\alpha}', \boldsymbol{\beta}', \{\mathbf{A}'_a\}\rangle = \langle \boldsymbol{\alpha} + \boldsymbol{\alpha}', \boldsymbol{\beta} + \boldsymbol{\beta}', \{\mathbf{A}_a + \mathbf{A}'_a\}\rangle \\ cA &= c\langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_a\}\rangle = \langle c\boldsymbol{\alpha}, c\boldsymbol{\beta}, \{c\mathbf{A}_a\}\rangle \enspace . \end{aligned}$$

We can view $\mathcal{A}_n$ as a normed vector space by endowing it with any norm from the following family. Let $p, q \in [1, +\infty]$ be Hölder conjugates, i.e. $p^{-1} + q^{-1} = 1$. It is easy to check that the following defines a norm on $\mathcal{A}_n$:

$$\|A\|_{p,q} = \max\left\{ \|\boldsymbol{\alpha}\|_p, \|\boldsymbol{\beta}\|_q, \max_a \|\mathbf{A}_a\|_q \right\} \enspace ,$$

where $\|\mathbf{A}\|_q$ denotes the matrix norm induced by the corresponding vector norm, that is $\|\mathbf{A}\|_q = \sup_{\|\mathbf{v}\|_q = 1} \|\mathbf{A}\mathbf{v}\|_q$. Given $p \in [1, +\infty]$ and $q = 1/(1 - 1/p)$, we denote by $\mathcal{A}_{n,p,r}$ the set of all WFAs A with n states and $\|A\|_{p,q} \leq r$. Thus, $\mathcal{A}_{n,p,r}$ is the ball of radius r at the origin in the normed vector space $(\mathcal{A}_n, \|\cdot\|_{p,q})$.

3.2 The Class $\mathcal{R}_{p,r}$

Next, we consider an alternative quantity measuring the complexity of rational functions that is *independent* of any WFA representation: their norm.

Given $p \in [1, \infty]$ and $f \colon \Sigma^* \to \mathbb{R}$ we use $\|f\|_p$ to denote the p-norm of f given by

$$\|f\|_p = \left[\sum_{x \in \Sigma^*} |f(x)|^p \right]^{\frac{1}{p}} \enspace ,$$

which in the case $p = \infty$ amounts to $\|f\|_\infty = \sup_{x \in \Sigma^*} |f(x)|$.

Let $\mathcal{R}_p$ denote the class of rational functions with finite p-norm: $f \in \mathcal{R}_p$ if and only if f is rational and $\|f\|_p < +\infty$. Given some $r > 0$ we also define $\mathcal{R}_{p,r}$, the class of functions with p-norm bounded by r:

$$\mathcal{R}_{p,r} = \{ f \colon \Sigma^* \to \mathbb{R} \mid f \text{ rational and } \|f\|_p \leq r \} \enspace .$$

Note that this definition is independent of the WFA used to represent f.

3.3 The Class $\mathcal{H}_{p,r}$

Here, we introduce a third class of rational functions described via their Hankel matrices, a quantity that is also independent of their WFA representations. To do so, we represent a function f using its Hankel matrix $\mathbf{H}_f$, interpret this matrix as a linear operator $\mathbf{H}_f : \mathbb{R}^{\Sigma^*} \to \mathbb{R}^{\Sigma^*}$ on the free vector space $\mathbb{R}^{\Sigma^*}$, and consider the Schatten p-norm of $\mathbf{H}_f$ as a measure of the complexity of f.

We now proceed to make this more precise. We identify a function $g \colon \Sigma^* \to \mathbb{R}$ with an infinite vector $g \in \mathbb{R}^{\Sigma^*}$. It follows from the definition of a Hankel matrix that we can interpret $\mathbf{H}_f$ as an operator given by

$$(\mathbf{H}_f g)(x) = \sum_{y \in \Sigma^*} f(xy) g(y) \ .$$

Note the similarity of the operation $g \mapsto \mathbf{H}_f g$ with a convolution between f and g. The following result of [10] shows that $\|f\|_1 < \infty$ is a sufficient condition for this operation to be defined.

Lemma 1. *Let $p \in [1, +\infty]$. Assume that $f \colon \Sigma^* \to \mathbb{R}$ satisfies the condition $\|f\|_1 < \infty$. Then, $\|g\|_p < \infty$ implies $\|\mathbf{H}_f g\|_p < \infty$.*

This shows that for $f \in \mathcal{R}_1$ the operator $\mathbf{H}_f : \mathcal{R}_p \to \mathcal{R}_p$ is *bounded* for every $p \in [1, +\infty]$. By the Theorem of Fliess, the matrix $\mathbf{H}_f$ has finite rank when f is rational. Thus, this implies (by considering the case $p = 2$) that the bi-infinite matrix $\mathbf{H}_f$ admits a singular value decomposition whenever $f \in \mathcal{R}_1$. In that case, it makes sense to define the Schatten–Hankel p-norm of $f \in \mathcal{R}_1$ as $\|f\|_{\mathrm{H},p} = \|(\mathfrak{s}_1, \ldots, \mathfrak{s}_n)\|_p$, where $\mathfrak{s}_i = \mathfrak{s}_i(\mathbf{H}_f)$ is the ith singular value of $\mathbf{H}_f$ and $\operatorname{rank}(\mathbf{H}_f) = n$. That is, the Schatten–Hankel p-norm of f is exactly the Schatten p-norm of $\mathbf{H}_f$.

Using this notation, we can define several classes of rational functions. For a given $p \in [1, +\infty]$, we denote by $\mathcal{H}_p$ the class of rational functions with $\|f\|_{\mathrm{H},p} < \infty$ and, for any $r > 0$, by $\mathcal{H}_{p,r}$ the class of rational functions with $\|f\|_{\mathrm{H},p} \leq r$.

3.4 Overview of Results

In addition to proving general bounds on the Rademacher complexity of the three classes just described, we will also highlight their application in some important special cases.

Here, we briefly discuss these special cases, stress different properties of the classes of WFAs to which these results apply, and mention several well-known sub-families within each class. We also briefly touch upon the problem of deciding the membership of a given WFA in any of the particular classes defined above.

– $\mathcal{A}_{n,p,r}$ in the case $r = 1$ (Corollary 1): note that for $r = 1$ and $p = 1$, $\mathcal{A}_{n,p,r}$ includes all DFAs and PFAs since for these classes of automata $\boldsymbol{\alpha}$ is either an indicator vector or a probability distribution over states, hence $\|\boldsymbol{\alpha}\|_1 = 1$; $\boldsymbol{\beta}$ has all its entries in $[0, 1]$ since it consists of accept/reject labels or stopping probabilities, hence $\|\boldsymbol{\beta}\|_\infty \leq 1$; and, for any $a \in \Sigma$ and any $i \in [1, n]$, the inequality $\sum_j |\mathbf{A}_a(i,j)| \leq 1$ holds since the transitions can reach at most one state per symbol, or represent a probability distribution over next states, hence $\|\mathbf{A}_a\|_\infty \leq 1$.

- $\mathcal{R}_{p,r}$ in the cases $p = 1$ and $p = 2$ (Corollaries 3 and 2): we note here that PFAs with stopping probabilities are contained in $\mathcal{R}_1$, while there are PFAs without stopping probabilities in $\mathcal{R}_2 \backslash \mathcal{R}_1$. In general, given a WFA, membership in $\mathcal{R}_{1,r}$ is semi-decidable [5], while membership in $\mathcal{R}_{2,r}$ can be decided in polynomial time [15].
- $\mathcal{H}_{p,r}$ in the cases $p = 1$ and $p = 2$ (Corollaries 5 and 4): as mentioned above, membership in $\mathcal{R}_1$ is sufficient to show membership in $\mathcal{H}_p$ for all $1 \leq p \leq \infty$. Assuming membership in $\mathcal{H}_\infty$, it is possible to decide membership in $\mathcal{H}_{p,r}$ in polynomial time [10].

4 Rademacher Complexity of $\mathcal{A}_{n,p,r}$

In this section, we present an upper bound on the Rademacher complexity of the class of WFAs $\mathcal{A}_{n,p,r}$. To bound $\mathfrak{R}_m(\mathcal{A}_{n,p,r})$, we will use an argument based on covering numbers. We first introduce some notation, then state our general bound and related corollaries, and finally prove the main result of this section.

Let $S = (x_1, \ldots, x_m) \in (\Sigma^*)^m$ be a sample of m strings with maximum length $L_S = \max_i |x_i|$. The expectation of this quantity over a sample of m strings drawn i.i.d. from some fixed distribution D will be denoted by $L_m = \mathbb{E}_{S \sim D^m}[L_S]$. It is interesting at this point to note that L_m appears in our bound and introduces a dependency on the distribution D which will exhibit different growth rates depending on the behavior of the tails of D. For example, it is well known that if the random variable $|x|$ for $x \sim D$ is sub-Gaussian,[4] then $L_m = O(\sqrt{\log m})$. Similarly, if the tail of D is sub-exponential, then $L_m = O(\log m)$ and if the tail is a power-law with exponent $s+1$, $s > 0$, then $L_m = O(m^{1/s})$. Note that in the latter case the distribution of $|x|$ has finite variance if and only if $s > 1$.

Theorem 1. *The following inequality holds for every sample $S \in (\Sigma^*)^m$:*

$$\widehat{\mathfrak{R}}_S(\mathcal{A}_{n,p,r}) \leq \inf_{\eta > 0} \left(\eta + r^{L_S+2} \sqrt{\frac{2n(kn+2)\log\left(2r + \frac{r^{L_S+2}(L_S+2)}{\eta}\right)}{m}} \right) .$$

By considering the case $r = 1$ and choosing $\eta = (L_S + 2)/m$ we obtain the following corollary.

[4] Recall that a non-negative random variable X is sub-Gaussian if $\mathbb{P}[X > k] \leq \exp(-\Omega(k^2))$, sub-exponential if $\mathbb{P}[X > k] \leq \exp(-\Omega(k))$, and follows a power-law with exponent $(s+1)$ if $\mathbb{P}[X > k] \leq O(1/k^{s+1})$.

Corollary 1. *For any $m \geq 1$ and $n \geq 1$ the following inequality holds:*

$$\mathfrak{R}_m(\mathcal{A}_{n,p,1}) \leq \sqrt{\frac{2n(kn+2)\log(m+2)}{m}} + \frac{L_m+2}{m} \ .$$

4.1 Proof of Theorem 1

We begin the proof by recalling several well-known facts and definitions related to covering numbers (see e.g. [16]). Let $V \subset \mathbb{R}^m$ be a set of vectors and $S = (x_1, \dots, x_m) \in (\Sigma^*)^m$ a sample of size m. Given a WFA A, we define $A(S) \in \mathbb{R}^m$ by $A(S) = (A(x_1), \dots, A(x_m)) \in \mathbb{R}^m$. We say that V is an (ℓ_1, η)-cover for S with respect to $\mathcal{A}_{n,p,r}$ if for every $A \in \mathcal{A}_{n,p,r}$ there exists some $\mathbf{v} \in V$ such that

$$\frac{1}{m}\|\mathbf{v} - A(S)\|_1 = \frac{1}{m}\sum_{i=1}^{m} |\mathbf{v}_i - A(x_i)| \leq \eta \ .$$

The ℓ_1-covering number of S at level η with respect to $\mathcal{A}_{n,p,r}$ is defined as follows:

$$\mathcal{N}_1(\eta, \mathcal{A}_{n,p,r}, S) = \min\{|V| : \ V \subset \mathbb{R}^m \text{ is an } (\ell_1,\eta)\text{-cover for } S \text{ w.r.t. } \mathcal{A}_{n,p,r} \ \} \ .$$

A typical analysis based on covering numbers would now proceed to obtain a bound on the growth of $\mathcal{N}_1(\eta, \mathcal{A}_{n,p,r}, S)$ in terms of the number of strings m in S. Our analysis requires a slightly finer approach where the size of S is characterized by m and L_S. Thus, we also define for every integer $L \geq 0$ the following covering number

$$\mathcal{N}_1(\eta, \mathcal{A}_{n,p,r}, m, L) = \max_{S \in (\Sigma^{\leq L})^m} \mathcal{N}_1(\eta, \mathcal{A}_{n,p,r}, S) \ .$$

The first step in the proof of Theorem 1 is to bound $\mathcal{N}_1(\eta, \mathcal{A}_{n,p,r}, m, L)$. In order to derive such a bound, we will make use of the following technical results.

Lemma 2 (Corollary 4.3 in [35]). *A ball of radius $R > 0$ in a real d-dimensional Banach space can be covered by $R^d(2+1/\rho)^d$ balls of radius $\rho > 0$.*

Lemma 3. *Let $A, B \in \mathcal{A}_{n,p,r}$. Then the following hold for any $x \in \Sigma^*$:*

1. $|A(x)| \leq r^{|x|+2}$,
2. $|A(x) - B(x)| \leq r^{|x|+1}(|x|+2)\|A - B\|_{p,q}$.

Proof. The first bound follows from applying Hölder's inequality and the submultiplicativity of the norms in the definition of $\|A\|_{p,q}$ to (1). The second bound was proven in [8]. □

Combining these lemmas yields the following bound on the covering number $\mathcal{N}_1(\eta, \mathcal{A}_{n,p,r}, m, L)$.

Lemma 4.

$$\mathcal{N}_1(\eta, \mathcal{A}_{n,p,r}, m, L) \leq r^{n(kn+2)}\left(2 + \frac{r^{L+1}(L+2)}{\eta}\right)^{n(kn+2)} \ .$$

Proof. Let $d = n(kn+2)$. By Lemma 2 and Lemma 3, for any $\rho > 0$, there exists a finite set $\mathcal{C}_\rho \subset \mathcal{A}_{n,p,r}$ with $|\mathcal{C}_\rho| \leq r^d(2 + 1/\rho)^d$ such that: for every $A \in \mathcal{A}_{n,p,r}$ there exists $B \in \mathcal{C}_\rho$ satisfying $|A(x) - B(x)| \leq r^{|x|+1}(|x| + 2)\rho$ for every $x \in \Sigma^*$. Thus, taking $\rho = \eta/(r^{L+1}(L + 2))$ we see that for every $S \in (\Sigma^{\leq L})^m$ the set $V = \{B(S)\colon B \in \mathcal{C}_\rho\} \subset \mathbb{R}^m$ is an η-cover for S with respect to $\mathcal{A}_{n,p,r}$. □

The last step of the proof relies on the following well-known result due to Massart.

Lemma 5 (Massart [28]). *Given a finite set of vectors $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\} \subset \mathbb{R}^m$, the following holds*

$$\frac{1}{m}\,\mathbb{E}\left[\max_{\mathbf{v}\in V}\langle \boldsymbol{\sigma}, \mathbf{v}\rangle\right] \leq \left(\max_{\mathbf{v}\in V}\|\mathbf{v}\|_2\right)\frac{\sqrt{2\log(N)}}{m} \quad ,$$

where the expectation is over the vector $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)$ whose entries are independent Rademacher random variables $\sigma_i \sim \mathbf{Unif}(\{+1, -1\})$.

Fix $\eta > 0$ and let $V_{S,\eta}$ be an (ℓ_1, η)-cover for S with respect to $\mathcal{A}_{n,p,r}$. By Massart's lemma, we can write

$$\widehat{\mathfrak{R}}_S(\mathcal{A}_{n,p,r}) \leq \eta + \left(\max_{\mathbf{v}\in V_{S,\eta}}\|\mathbf{v}\|_2\right)\frac{\sqrt{2\log|V_{S,\eta}|}}{m} \quad . \tag{2}$$

Since $|A(x_i)| \leq r^{L_S+2}$ by Lemma 3, we can restrict the search for (ℓ_1, η)-covers for S to sets $V_{S,\eta} \subset \mathbb{R}^m$ where all $\mathbf{v} \in V_{S,\eta}$ must satisfy $\|\mathbf{v}\|_\infty \leq r^{L_S+2}$. By construction, such a covering satisfies $\max_{\mathbf{v}\in V_{S,\eta}}\|\mathbf{v}\|_2 \leq r^{L_S+2}\sqrt{m}$. Finally, plugging in the bound for $|V_{S,\eta}|$ given by Lemma 4 into (2) and taking the infimum over all $\eta > 0$ yields the desired result. □

5 Rademacher Complexity of $\mathcal{R}_{p,r}$

In this section, we study the complexity of rational functions from a different perspective. Instead of analyzing their complexity in terms of the parameters of WFAs computing them, we consider an intrinsic associated quantity: their norm. We present upper bounds on the Rademacher complexity of the classes of rational functions $\mathcal{R}_{p,r}$ for any $p \in [1, +\infty]$ and $r > 0$.

It will be convenient for our analysis to identify a rational function $f \in \mathcal{R}_{p,r}$ with an infinite-dimensional vector $\mathbf{f} \in \mathbb{R}^{\Sigma^*}$ with $\|\mathbf{f}\|_p \leq r$. That is, $\mathbf{f}$ is an infinite vector indexed by strings in Σ^* whose xth entry is $\mathbf{f}_x = f(x)$. An important observation is that using this notation, for any given $x \in \Sigma^*$, we can write $f(x)$ as the inner product $\langle \mathbf{f}, \mathbf{e}_x\rangle$, where $\mathbf{e}_x \in \mathbb{R}^{\Sigma^*}$ is the indicator vector corresponding to string x.

Theorem 2. *Let $p^{-1} + q^{-1} = 1$. Let $S = (x_1, \ldots, x_m)$ be a sample of m strings. Then, the following holds for any $r > 0$:*

$$\widehat{\mathfrak{R}}_S(\mathcal{R}_{p,r}) = \frac{r}{m}\,\mathbb{E}\left[\left\|\sum_{i=1}^m \sigma_i \mathbf{e}_{x_i}\right\|_q\right] \quad ,$$

where the expectation is over the m independent Rademacher random variables $\sigma_i \sim \mathbf{Unif}(\{+1,-1\})$.

Proof. In view of the notation just introduced described, we can write

$$\begin{aligned}\widehat{\mathfrak{R}}_S(\mathcal{R}_{p,r}) = \mathbb{E}\left[\sup_{f\in\mathcal{R}_{p,r}} \frac{1}{m}\sum_{i=1}^{m}\langle \mathbf{f}, \sigma_i \mathbf{e}_{x_i}\rangle\right] &= \frac{1}{m}\,\mathbb{E}\left[\sup_{f\in\mathcal{R}_{p,r}} \left\langle \mathbf{f}, \sum_{i=1}^{m}\sigma_i \mathbf{e}_{x_i}\right\rangle\right] \\ &= \frac{r}{m}\,\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sigma_i \mathbf{e}_{x_i}\right\|_q\right] \quad ,\end{aligned}$$

where the last inequality holds by definition of the dual norm. □

The next corollaries give non-trivial bounds on the Rademacher complexity in the case $p = 1$ and the case $p = 2$.

Corollary 2. *For any $m \geq 1$ and any $r > 0$, the following inequalities hold:*

$$\frac{r}{\sqrt{2m}} \leq \mathfrak{R}_m(\mathcal{R}_{2,r}) \leq \frac{r}{\sqrt{m}}.$$

Proof. The upper bound follows directly from Theorem 2 and Jensen's inequality:

$$\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sigma_i \mathbf{e}_{x_i}\right\|_2\right] \leq \sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sigma_i \mathbf{e}_{x_i}\right\|_2^2\right]} = \sqrt{m} \ .$$

The lower bound is obtained using Khintchine–Kahane's inequality (see appendix of [32]):

$$\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sigma_i \mathbf{e}_{x_i}\right\|_2\right]^2 \geq \frac{1}{2}\,\mathbb{E}\left[\left\|\sum_{i=1}^{m}\sigma_i \mathbf{e}_{x_i}\right\|_2^2\right] = \frac{m}{2},$$

which completes the proof. □

The following definitions will be needed to present our next corollary. Given a sample $S = (x_1, \ldots, x_m)$ and a string $x \in \Sigma^\star$ we denote by $s_x = |\{i \colon x_i = x\}|$ the number of times x appears in S. Let $M_S = \max_{s\in\Sigma^\star} s_x$. Given a probability distribution D over $\Sigma^\star$ we also define $M_m = \mathbb{E}_{S\sim D^m}[M_S]$. Note that M_m is the expected maximum number of collisions (repeated strings) in a sample of size m drawn from D, and that we have the straightforward bounds $1 \leq M_S \leq m$.

Corollary 3. *For any $m \geq 1$ and any $r > 0$, the following upper bound holds:*

$$\mathfrak{R}_m(\mathcal{R}_{1,r}) \leq \frac{r\sqrt{2M_m \log(2m)}}{m}.$$

Proof. Let $S = (x_1, \ldots, x_m)$ be a sample with m strings. For any $x \in \Sigma^\star$ define the vector $\mathbf{v}_x \in \mathbb{R}^m$ given by $\mathbf{v}_x(i) = \mathbb{I}_{x_i = x}$. Let V be the set of vectors $\mathbf{v}_x$ which are not identically zero, and note we have $|V| \leq m$. Also note that by construction we have $\max_{\mathbf{v}_x \in V} \|\mathbf{v}_x\|_2 = \sqrt{M_S}$. Now, by Theorem 2 we have

$$\widehat{\mathfrak{R}}_S(\mathcal{R}_{1,r}) = \frac{r}{m} \mathbb{E}\left[\left\| \sum_{i=1}^{m} \sigma_i \mathbf{e}_{x_i} \right\|_\infty \right] = \frac{r}{m} \mathbb{E}\left[\max_{\mathbf{v}_x \in V \cup (-V)} \langle \boldsymbol{\sigma}, \mathbf{v}_x \rangle \right] \enspace .$$

Therefore, using Massart's Lemma we get

$$\widehat{\mathfrak{R}}_S(\mathcal{R}_{1,r}) \leq \frac{r\sqrt{2 M_S \log(2m)}}{m} \enspace .$$

The result now follows from taking the expectation over S and using Jensen's inequality to see that $\mathbb{E}[\sqrt{M_S}] \leq \sqrt{M_m}$. □

Note in this case we cannot rely on the Khintchine–Kahane inequality to obtain lower bounds on $\mathfrak{R}_m(\mathcal{R}_{1,r})$ because there is no version of this inequality for the case $q = \infty$.

6 Rademacher Complexity of $\mathcal{H}_{p,r}$

In this section, we present our last set of upper bounds on the Rademacher complexity of WFAs. Here, we characterize the complexity of WFAs in terms of the spectral properties of their Hankel matrix.

The Hankel matrix of a function $f \colon \Sigma^* \to \mathbb{R}$ is the bi-infinite matrix $\mathbf{H}_f \in \mathbb{R}^{\Sigma^* \times \Sigma^*}$ whose entries are defined by $\mathbf{H}_f(u, v) = f(uv)$. Note that any string $x \in \Sigma^*$ admits $|x| + 1$ decompositions $x = uv$ into a prefix $u \in \Sigma^*$ and a suffix $v \in \Sigma^*$. Thus, $\mathbf{H}_f$ contains a high degree of redundancy: for any $x \in \Sigma^*$, $f(x)$ is the value of at least $|x| + 1$ entries of $\mathbf{H}_f$ and we can write $f(x) = \mathbf{e}_u^\top \mathbf{H}_f \mathbf{e}_v$ for any decomposition $x = uv$.

Let $\mathfrak{s}_i(\mathbf{M})$ denote the ith singular value of a matrix $\mathbf{M}$. For $1 \leq p \leq \infty$, let $\|\mathbf{M}\|_{\mathrm{S},p}$ denote the p-Schatten norm of $\mathbf{M}$ defined by $\|\mathbf{M}\|_{\mathrm{S},p} = \left[\sum_{i \geq 1} \mathfrak{s}_i(\mathbf{M})^p\right]^{\frac{1}{p}}$.

Theorem 3. *Let $p, q \geq 1$ with $p^{-1} + q^{-1} = 1$ and let $S = (x_1, \ldots, x_m)$ be a sample of m strings in Σ^*. For any decomposition $x_i = u_i v_i$ of the strings in S and any $r > 0$, the following inequality holds:*

$$\widehat{\mathfrak{R}}_S(\mathcal{H}_{p,r}) \leq \frac{r}{m} \mathbb{E}\left[\left\| \sum_{i=1}^{m} \sigma_i \mathbf{e}_{u_i} \mathbf{e}_{v_i}^\top \right\|_{\mathrm{S},q} \right] \enspace .$$

Proof. For any $1 \leq i \leq m$, let $x_i = u_i v_i$ be an arbitrary decomposition and let $\mathbf{R}$ denote $\mathbf{R} = \sum_{i=1}^{m} \sigma_i \mathbf{e}_{u_i} \mathbf{e}_{v_i}^\top$. Then, in view of the identity $f(x_i) = \mathbf{e}_{u_i}^\top \mathbf{H}_f \mathbf{e}_{v_i}$, we can write

$$\begin{aligned}\widehat{\mathfrak{R}}_S(\mathcal{H}_{p,r}) &= \mathbb{E}\left[\sup_{f\in\mathcal{H}_{p,r}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i \mathbf{e}_{u_i}^\top \mathbf{H}_f \mathbf{e}_{v_i}\right] \\ &= \frac{1}{m}\,\mathbb{E}\left[\sup_{f\in\mathcal{H}_{p,r}} \sum_{i=1}^{m} \operatorname{Tr}\left(\sigma_i \mathbf{e}_{v_i}\mathbf{e}_{u_i}^\top \mathbf{H}_f\right)\right] = \frac{1}{m}\,\mathbb{E}\left[\sup_{f\in\mathcal{H}_{p,r}} \langle \mathbf{R}, \mathbf{H}_f\rangle\right] \ .\end{aligned}$$

Then, by von Neumann's trace inequality [29] and Hölder's inequality, the following holds:

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{H}_{p,r}} \langle \mathbf{R}, \mathbf{H}_f\rangle\right] &\leq \mathbb{E}\left[\sup_{f\in\mathcal{H}_{p,r}} \sum_{j\geq 1} \mathfrak{s}_j(\mathbf{R})\cdot\mathfrak{s}_j(\mathbf{H}_f)\right] \\ &\leq \mathbb{E}\left[\sup_{f\in\mathcal{H}_{p,r}} \|\mathbf{R}\|_{\mathrm{S},q}\|\mathbf{H}_f\|_{\mathrm{S},p}\right] = r\,\mathbb{E}\left[\|\mathbf{R}\|_{\mathrm{S},q}\right] \ ,\end{aligned}$$

which completes the proof. □

Note that, in this last result, the equality condition for von Neumann's inequality cannot be used to obtain a lower bound on $\widehat{\mathfrak{R}}_S(\mathcal{H}_{p,r})$ since it requires the simultaneous diagonalizability of the two matrices involved, which is difficult to control in the case of Hankel matrices.

As in the previous sections, we now proceed to derive specialized versions of the bound of Theorem 3 for the cases $p = 1$ and $p = 2$. First, note that the corresponding q-Schatten norms have given names: $\|\mathbf{R}\|_{\mathrm{S},2} = \|\mathbf{R}\|_{\mathrm{F}}$ is the Frobenius norm, and $\|\mathbf{R}\|_{\mathrm{S},\infty} = \|\mathbf{R}\|_{\mathrm{op}}$ is the operator norm.

Corollary 4. *For any $m \geq 1$ and any $r > 0$, the Rademacher complexity of $\mathcal{H}_{2,r}$ can be bounded as follows:*

$$\mathfrak{R}_m(\mathcal{H}_{2,r}) \leq \frac{r}{\sqrt{m}} .$$

Proof. In view of Theorem 3 and using Jensen's inequality, we can write

$$\begin{aligned}\mathfrak{R}_m(\mathcal{H}_{2,r}) \leq \frac{r}{m}\,\mathbb{E}\left[\|\mathbf{R}\|_{\mathrm{F}}\right] &\leq \frac{r}{m}\sqrt{\mathbb{E}\left[\|\mathbf{R}\|_F^2\right]} \\ &= \frac{r}{m}\sqrt{\mathbb{E}\left[\sum_{i,j=1}^{m} \sigma_i\sigma_j \langle \mathbf{e}_{u_i}\mathbf{e}_{v_i}^\top, \mathbf{e}_{u_j}\mathbf{e}_{v_j}^\top\rangle\right]} \\ &= \frac{r}{m}\sqrt{\mathbb{E}\left[\sum_{i=1}^{m} \langle \mathbf{e}_{u_i}\mathbf{e}_{v_i}^\top, \mathbf{e}_{u_i}\mathbf{e}_{v_i}^\top\rangle\right]} = \frac{r}{\sqrt{m}} \ ,\end{aligned}$$

which concludes the proof. □

We now introduce a combinatorial number depending on S and the decomposition selected for each string x_i. Let $U_S = \max_{u \in \Sigma^*} |\{i \colon u_i = u\}|$ and $V_S = \max_{v \in \Sigma^*} |\{i \colon v_i = v\}|$. Then, we define $W_S = \min \max\{U_S, V_S\}$, where then minimum is taken over all possible decompositions of the strings in S. If S is sampled from a distribution D, we also define $W_m = \mathbb{E}_{S \sim D^m}[W_S]$. It is easy to show that we have the bounds $1 \leq W_S \leq m$. Indeed, for the case $W_S = m$ consider a sample with m copies of the empty string, and for the case $W_S = 1$ consider a sample with m different strings of length m. The following result can be stated using this definition.

Corollary 5. *There exists a universal constant $C > 0$ such that for any $m \geq 1$ and any $r > 0$, the following inequality holds:*

$$\mathfrak{R}_m(\mathcal{H}_{1,r}) \leq \frac{Cr\left(\log(m+1) + \sqrt{W_m \log(m+1)}\right)}{m} \ .$$

Proof. First, note that by Corollary 7.3.2 of [34] applied to the random matrix $\mathbf{R}$, the following inequality holds:

$$\mathbb{E}[\|\mathbf{R}\|_{\mathrm{op}}] \leq C\left(\log(m+1) + \sqrt{\mu \log(m+1)}\right) \ ,$$

where $\mu = \max\{\| \sum_i \mathbf{e}_{u_i} \mathbf{e}_{u_i}^\top \|_{\mathrm{op}}, \| \sum_i \mathbf{e}_{v_i} \mathbf{e}_{v_i}^\top \|_{\mathrm{op}}\}$ and $C > 0$ is a constant. Next, observe that $\mathbf{D} = \sum_i \mathbf{e}_{u_i} \mathbf{e}_{u_i}^\top \in \mathbb{R}^{\Sigma^* \times \Sigma^*}$ is a diagonal matrix with $\mathbf{D}(u,u) = \sum_i \mathbb{I}_{u=u_i}$. Thus, $\|\mathbf{D}\|_{\mathrm{op}} = \max_u \mathbf{D}(u,u) = \max_{u \in \Sigma^*} |\{i \colon u_i = u\}| = U_S$. Similarly, we have $\| \sum_i \mathbf{e}_{v_i} \mathbf{e}_{v_i}^\top \|_{\mathrm{op}} = V_S$. Thus, since the decomposition of the strings in S is arbitrary, we can choose it such that $\mu = W_S$. In addition, Jensen's inequality implies $\mathbb{E}_S[\sqrt{W_S}] \leq \sqrt{W_m}$. Applying Theorem 3 now yields the desired bound. □

7 Conclusion

We introduced three general classes of WFAs described via different natural quantities and for each, proved upper bounds on their Rademacher complexity. An interesting property of these bounds is the appearance of different combinatorial parameters tying the sample to the convergence rate, whose nature depends on the way chosen to measure the complexity of the hypotheses: the length of the longest string L_S for $\mathcal{A}_{n,p,r}$; the maximum number of collisions M_S for $\mathcal{R}_{p,r}$; and, the minimum number of prefix or suffix collisions over all possible splits W_S for $\mathcal{H}_{p,r}$.

Another important feature of our bounds for the classes $\mathcal{H}_{p,r}$ is that they depend on spectral properties of Hankel matrices, which are commonly used in spectral learning algorithms for WFAs [21, 8]. We hope to exploit this connection in the future to provide more refined analyses of these learning algorithms. Our results can also be used to improve some aspects of existing spectral learning algorithms. For example, it might be possible to use the analysis in Theorem 3 for deriving strategies to help choose which prefixes and suffixes to consider in algorithms working with finite sub-blocks of an infinite Hankel matrix. This is a problem of practical relevance when working with large amounts of data which require balancing trade-offs between computation and accuracy [6].

Acknowledgments

This work was partly funded by the NSF award IIS-1117591 and NSERC.

References

1. Abe, N., Warmuth, M.K.: On the computational complexity of approximating distributions by probabilistic automata. Machine Learning (1992)
2. Albert, J., Kari, J.: Digital image compression. In: Handbook of weighted automata. Springer (2009)
3. Baier, C., Größer, M., Ciesinski, F.: Model checking linear-time properties of probabilistic systems. In: Handbook of Weighted automata. Springer (2009)
4. Bailly, R., Denis, F., Ralaivola, L.: Grammatical inference as a principal component analysis problem. In: ICML (2009)
5. Bailly, R., Denis, F.: Absolute convergence of rational series is semi-decidable. Inf. Comput. (2011)
6. Balle, B., Carreras, X., Luque, F., Quattoni, A.: Spectral learning of weighted automata: A forward-backward perspective. Machine Learning (2014)
7. Balle, B., Hamilton, W., Pineau, J.: Methods of moments for learning stochastic languages: Unified presentation and empirical comparison. In: ICML (2014)
8. Balle, B., Mohri, M.: Spectral learning of general weighted automata via constrained matrix completion. In: NIPS (2012)
9. Balle, B., Mohri, M.: Learning weighted automata. In: CAI (2015)
10. Balle, B., Panangaden, P., Precup, D.: A canonical form for weighted automata and applications to approximate minimization. In: Logic in Computer Science (LICS) (2015)
11. Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: Risk bounds and structural results. In: COLT (2001)
12. Berstel, J., Reutenauer, C.: Noncommutative rational series with applications. Cambridge University Press (2011)
13. Boots, B., Siddiqi, S., Gordon, G.: Closing the learning-planning loop with predictive state representations. In: RSS (2009)
14. Carlyle, J.W., Paz, A.: Realizations by stochastic finite automata. J. Comput. Syst. Sci. 5(1) (1971)
15. Cortes, C., Mohri, M., Rastogi, A.: Lp distance and equivalence of probabilistic automata. International Journal of Foundations of Computer Science (2007)
16. Devroye, L., Lugosi, G.: Combinatorial methods in density estimation. Springer (2001)
17. Eilenberg, S.: Automata, Languages and Machines, vol. A. Academic Press (1974)
18. Fliess, M.: Matrices de Hankel. Journal de Mathématiques Pures et Appliquées 53 (1974)
19. de Gispert, A., Iglesias, G., Blackwood, G., Banga, E., Byrne, W.: Hierarchical phrase-based translation with weighted finite-state transducers and shallow-n grammars. Computational Linguistics (2010)

20. Hamilton, W.L., Fard, M.M., Pineau, J.: Modelling sparse dynamical systems with compressed predictive state representations. In: ICML (2013)
21. Hsu, D., Kakade, S.M., Zhang, T.: A spectral algorithm for learning hidden Markov models. In: COLT (2009)
22. Ishigami, Y., Tani, S.: Vc-dimensions of finite automata and commutative finite automata with k letters and n states. Discrete Applied Mathematics (1997)
23. Knight, K., May, J.: Applications of weighted automata in natural language processing. In: Handbook of Weighted Automata. Springer (2009)
24. Koltchinskii, V., Panchenko, D.: Rademacher processes and bounding the risk of function learning. In: High Dimensional Probability II. pp. 443–459. Birkhäuser (2000)
25. Kuich, W., Salomaa, A.: Semirings, Automata, Languages. No. 5 in EATCS Monographs on Theoretical Computer Science, Springer-Verlag, Berlin-New York (1986)
26. Kulesza, A., Jiang, N., Singh, S.: Low-rank spectral learning with weighted loss functions. In: AISTATS (2015)
27. Kulesza, A., Rao, N.R., Singh, S.: Low-Rank Spectral Learning. In: AISTATS (2014)
28. Massart, P.: Some applications of concentration inequalities to statistics. Annales de la Faculté des Sciences de Toulouse (2000)
29. Mirsky, L.: A trace inequality of John von Neumann. Monatshefte für Mathematik (1975)
30. Mohri, M.: Weighted automata algorithms. In: Handbook of Weighted Automata, pp. 213–254. Monographs in Theoretical Computer Science, Springer (2009)
31. Mohri, M., Pereira, F.C.N., Riley, M.: Speech recognition with weighted finite-state transducers. In: Handbook on Speech Processing and Speech Comm. Springer (2008)
32. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of machine learning. MIT press (2012)
33. Salomaa, A., Soittola, M.: Automata-Theoretic Aspects of Formal Power Series. Springer-Verlag: New York (1978)
34. Tropp, J.A.: An Introduction to Matrix Concentration Inequalities. ArXiv abs/1501.01571 (2015)
35. Vershynin, R.: Lectures in Geometrical Functional Analysis. Preprint (2009)