Generalization Bounds for Learning Weighted Automata

Borja Balle[a,*], Mehryar Mohri[b,c]

[a] *Department of Mathematics and Statistics, Lancaster University, Lancaster, UK*
[b] *Courant Institute of Mathematical Sciences, New York University, NY, USA*
[c] *Google Research, New York, NY, USA*

Abstract

This paper studies the problem of learning weighted automata from a finite sample of strings with real-valued labels. We consider several hypothesis classes of weighted automata defined in terms of three different measures: the norm of an automaton's weights, the norm of the function computed by an automaton, and the norm of the corresponding Hankel matrix. We present new data-dependent generalization guarantees for learning weighted automata expressed in terms of the Rademacher complexity of these classes. We further present upper bounds on these Rademacher complexities, which reveal key new data-dependent terms related to the complexity of learning weighted automata.

1. Introduction

Weighted finite automata (WFAs) provide a general and highly expressive framework for representing functions mapping strings to real numbers. The mathematical theory behind WFAs, that of rational power series, has been extensively studied in the past (Eilenberg, 1974; Salomaa and Soittola, 1978; Kuich and Salomaa, 1986; Berstel and Reutenauer, 1988) and has been more recently the topic of a dedicated handbook (Droste et al., 2009). WFAs are widely used in modern applications, perhaps most prominently in image processing and speech recognition where the terminology of *weighted automata* seems to have been first introduced and made popular (Culik II and Kari, 1993; Mohri et al., 1996; Pereira and Riley, 1997; Mohri, 1997; Mohri et al., 2008), in several other speech processing applications such as speech synthesis (Sproat, 1995; Allauzen et al., 2004), in phonological and morphological rule compilation (Kaplan and Kay, 1994; Karttunen, 1995; Mohri and Sproat, 1996), in parsing (Roche, 1994; Mohri and Pereira, 1998), machine translation (de Gispert et al., 2010; Allauzen et al., 2014), bioinformatics (Durbin et al., 1998; Allauzen et al., 2008), sequence modeling and prediction (Cortes et al., 2004), formal verification and

*Corresponding author

model checking (Baier et al., 2009; Aminof et al., 2011), in optical character recognition (Breuel, 2008), and in many other areas.

The recent developments in spectral learning (Hsu et al., 2009; Bailly et al., 2009) have triggered a renewed interest in the use of WFAs in machine learning, with several recent successes in natural language processing (Balle et al., 2014a,b) and reinforcement learning (Boots et al., 2009; Hamilton et al., 2013). 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 (Hsu et al., 2009). A refined analysis of such algorithms provides sample bounds which are independent of the size of the Hankel matrices used by the learning algorithm (Denis et al., 2016). 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. See (Balle and Mohri, 2015a) for a recent survey of algorithms for learning WFAs with a discussion of the different assumptions and learning models.

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 (Balle and Mohri, 2012) 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 Balle and Mohri (2012) 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 that 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 (Balle and Mohri, 2012), 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 refined generalization bounds is standard (Koltchinskii and Panchenko, 2000) (see also (Bartlett and Mendelson, 2001) and (Mohri et al., 2012)). It has been successfully used to derive learning guarantees for classification, regression, kernel learning, ranking, and many other machine learning tasks (e.g. see Mohri et al. (2012) 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. As an application of our Rademacher complexity bounds we provide a variety of generalizations bounds for learning with WFAs using a bounded Lipschitz loss function; our bounds include both data-dependent and data-independent bounds.

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 Ishigami and Tani (1997) to be in $O(kn \log n)$. This can be used to show that the Rademacher complexity of this class of DFA is bounded by $O(\sqrt{nk \log n/m})$. For probabilistic finite automata (PFAs), it was shown by Abe and Warmuth (1992) that, in an agnostic setting, a sample of size $\widetilde{O}(kT^2n^2/\varepsilon^2)$ 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 when the error is measured on all strings of length T. New 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 (Kulesza et al., 2014). This has led to interesting upper bounds on the approximation error between minimal WFAs of different sizes (Kulesza et al., 2015). See (Balle et al., 2015) 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; 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 rest of this paper is 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. The bounds are formally stated and proven in Sections 4, 5, and 6. In Section 7 we provide additional bounds required for converting some sample-dependent bounds from Sections 5 and 6 into sample-independent bounds. Finally, the generalizations bounds obtained using the machinery developed in previous sections are given in Section 8.

2. Preliminaries

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 $\Sigma^\star$ the set of all finite strings over the alphabet Σ. The length of $u \in \Sigma^\star$ 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: $\Sigma^{\leq L} = \{x \in \Sigma^\star \colon |x| \leq L\}$. Given two strings $u, v \in \Sigma^\star$ we write uv for their concatenation.

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 $f_A \colon \Sigma^\star \to \mathbb{R}$ defined for all $x = a_1 \cdots a_t \in \Sigma^\star$ by

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

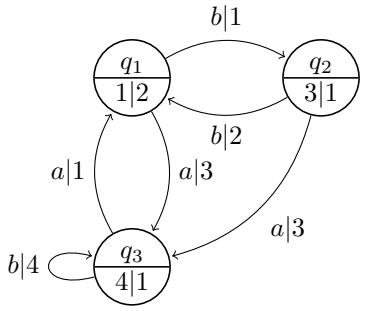
where $\mathbf{A}_x = \mathbf{A}_{a_1} \cdots \mathbf{A}_{a_t}$. This algebraic expression in fact corresponds to summing the weights of all possible paths in the automaton indexed by the symbols in x, where the weight of a single path $(q_0, q_1, \ldots, q_t) \in [n]^{t+1}$ is obtained by multiplying the initial weight of q_0, the weights of all transitions from q_{s-1} to q_s labeled by x_s, and the final weight of state q_t, that is

$$f_A(x) = \sum_{(q_0,\ldots,q_t)\in[n]^{t+1}} \boldsymbol{\alpha}(q_0) \left(\prod_{s=1}^{t} \mathbf{A}_{x_s}(q_{s-1}, q_s) \right) \boldsymbol{\beta}(q_t) \ .$$

See Figure 1 for an example of WFA with 3 states given in terms of its algebraic representation and the equivalent representation as a weighted transition diagram between states.

An arbitrary function $f \colon \Sigma^\star \to \mathbb{R}$ 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 is 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 another minimal WFA computing f. We 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 (Eilenberg, 1974; Salomaa and Soittola, 1978; Berstel and Reutenauer, 2011; Kuich and Salomaa, 1986; Mohri, 2009) 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 (Salomaa and Soittola, 1978; Berstel and Reutenauer, 2011; Kuich and Salomaa, 1986). We will not favor that point of view here, however, since we will not need to make explicit mention 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 $\mathbf{H}_f \in$



(a)

$$\boldsymbol{\alpha} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} \quad \mathbf{A}_a = \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & 3 \\ 1 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} \quad \mathbf{A}_b = \begin{bmatrix} 0 & 1 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

(b)

	ϵ	a	b	aa	ab	ba	bb
ϵ	9	20	29	36	52	53	74
a	20	36	52	60	60	108	208
b	29	53	74	90	100	152	282

(c)

Figure 1: (a) Example of WFA A. Within each circle we present the state name (q_i for the ith state), the initial weight (left number) and the final weight (right number). In particular, $f_A(ab) = 1 \times 3 \times 4 \times 1 + 3 \times 3 \times 4 \times 1 + 4 \times 1 \times 1 \times 1 = 52$. (b) Corresponding initial vector $\boldsymbol{\alpha}$, final vector $\boldsymbol{\beta}$, and transition matrices $\mathbf{A}_a$ and $\mathbf{A}_b$. (c) Finite sub-block of the Hankel matrix $\mathbf{H}_{f_A}$.

$\mathbb{R}^{\Sigma^\star \times \Sigma^\star}$ of a function $f \colon \Sigma^\star \to \mathbb{R}$ is the infinite matrix with rows and columns indexed by all strings with $\mathbf{H}_f(u, v) = f(uv)$ for all $u, v \in \Sigma^\star$. By the theorem of Fliess (1974) (see also (Carlyle and Paz, 1971; Berstel and Reutenauer, 2011) and (Balle and Mohri, 2015a) for a recent proof), $\mathbf{H}_f$ has finite rank n if and only if f is rational and there exists a WFA A with n states computing f, that is, $\operatorname{rank}(f) = \operatorname{rank}(\mathbf{H}_f)$.

2.2. Learning Scenario

Let $\mathcal{Z}$ denote a measurable subset of $\mathbb{R}$. We assume a standard supervised learning scenario where training and test points are drawn i.i.d. according to some unknown distribution D over $\Sigma^\star \times \mathcal{Z}$.

Let $\mathcal{F}$ be a subset of the family of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$, with $\mathcal{Y} \subseteq \mathbb{R}$, and let $\ell \colon \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}_+$ be a loss function measuring the divergence between the prediction $y \in \mathcal{Y}$ made by a function in $\mathcal{F}$ and the target label $z \in \mathcal{Z}$. The learner's objective consists of using a labeled training sample $S = ((x_1, z_1), \ldots, (x_m, z_m))$ of size m to select a function $f \in \mathcal{F}$ with small expected loss, that is

$$\mathfrak{L}_D(f) = \mathop{\mathbb{E}}_{(x,z)\sim D}[\ell(f(x), z)] \ .$$

Our objective 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 of rational functions $f\colon \Sigma^\star \to \mathbb{R}$. Thus, we start with a brief introduction to the main definitions and results regarding the Rademacher complexity of an arbitrary class of functions $\mathcal{F} = \{f\colon \mathcal{X} \to \mathcal{Y}\}$ where $\mathcal{X}$ is the input space and $\mathcal{Y} \subseteq \mathbb{R}$ the output space. Let D be a probability distribution over $\mathcal{X} \times \mathcal{Z}$ for some $\mathcal{Z} \subseteq \mathbb{R}$ and denote by $D_{\mathcal{X}}$ the marginal distribution over $\mathcal{X}$. Suppose $S = (x_1, \ldots, x_m) \overset{\text{iid}}{\sim} D_{\mathcal{X}}^m$ is a sample of m i.i.d. examples 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 $\sigma_i \sim \mathbf{Unif}(\{+1, -1\})$. Note that this is equivalent to considering a random vector $\boldsymbol{\sigma} \sim \mathbf{Unif}(\{+1, -1\}^m)$. 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:

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

The Rademacher complexity of a hypothesis class can be used to derive generalization bounds for a variety of learning tasks (Koltchinskii and Panchenko, 2000; Bartlett and Mendelson, 2001; Mohri et al., 2012). To do so, we need to bound the Rademacher complexity of the associated loss class, for a given loss function $\ell\colon \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}_+$.

For a given hypothesis class $\mathcal{F}$, the corresponding loss class $\ell \circ \mathcal{F}$ is given by the set of all functions $\ell \circ f\colon \mathcal{X} \times \mathcal{Z} \to \mathbb{R}_+$ of the form $(x, z) \mapsto \ell(f(x), z)$. Using Talagrand's contraction lemma (Ledoux and Talagrand, 1991, Theorem 4.12), the empirical Rademacher complexity of $\ell \circ \mathcal{F}$ can be bounded in terms of $\widehat{\mathfrak{R}}_S(\mathcal{F})$, when ℓ is μ-Lipschitz with respect to its first argument for some $\mu > 0$, that is when

$$|\ell(y, z) - \ell(y', z)| \leq \mu |y - y'|$$

for all $y, y' \in \mathcal{Y}$ and $z \in \mathcal{Z}$. In that case, the following inequality holds:

$$\widehat{\mathfrak{R}}_{S'}(\ell \circ \mathcal{F}) \leq \mu \widehat{\mathfrak{R}}_S(\mathcal{F}) \ ,$$

where $S' = ((x_1, z_1), \ldots, (x_m, z_m))$ is a sample of size m with $(x_i, z_i) \in \mathcal{X} \times \mathcal{Z}$ and $S = (x_1, \ldots, x_m)$ denotes the sample of elements in $\mathcal{X}$ obtained from S'. When taking expectations over $S' \overset{\text{iid}}{\sim} D^m$ and $S \overset{\text{iid}}{\sim} D_{\mathcal{X}}^m$ we obtain the same bound for the Rademacher complexities $\mathfrak{R}_m(\ell \circ \mathcal{F}) \leq \mu \mathfrak{R}_m(\mathcal{F})$. A typical example of a loss function that is μ-Lipschitz with respect to its first argument is the absolute loss $\ell(y, z) = |y - z|$, which satisfies the condition with $\mu = 1$ for $\mathcal{Y} = \mathcal{Z} = \mathbb{R}$.

We say that a loss function $\ell\colon \mathbb{R} \times \mathcal{Z} \to \mathbb{R}_+$ is M-bounded if $\ell(y, z) \leq M$ for all $y \in \mathbb{R}$ and $z \in \mathcal{Z}$. In order to convert Rademacher complexity bounds into generalization bounds we will use the following standard result. The proof

follows from applying Talagrand's contraction principle to a straightforward extension of (Mohri et al., 2012, Theorem 3.1) for M-bounded loss functions.

Theorem 1. *Let D be a probability distribution over $\mathcal{X} \times \mathcal{Z}$ and let $S = ((x_i, y_i))_{i=1}^m$ be a sample of m i.i.d. examples from D. Suppose $\mathcal{F}$ is a class of functions from $\mathcal{X}$ to $\mathbb{R}$. Assume that the loss $\ell \colon \mathbb{R} \times \mathcal{Z} \to \mathbb{R}_+$ is M-bounded and μ-Lipschitz with respect to its first argument. Then, each of the following statements holds simultaneously for all $f \in \mathcal{F}$ with probability at least $1 - \delta$:*

$$\mathfrak{L}_D(f) \leq \widehat{\mathfrak{L}}_S(f) + 2\mu\mathfrak{R}_m(\mathcal{F}) + M\sqrt{\frac{\log \frac{1}{\delta}}{2m}} \ ,$$

$$\mathfrak{L}_D(f) \leq \widehat{\mathfrak{L}}_S(f) + 2\mu\widehat{\mathfrak{R}}_S(\mathcal{F}) + 3M\sqrt{\frac{\log \frac{2}{\delta}}{2m}} \ .$$

3. Classes of Rational Functions

In this section, we introduce several classes of rational functions. Each of these classes is defined in terms of a different way to measure the complexity of rational functions. The first one is based on the weights of an explicit WFA representation, while the other two are based on intrinsic quantities associated to the function: the norm of the function, and the norm of the corresponding Hankel matrix when viewed as a linear operator on a certain Hilbert space. These three different ways of measuring the complexity of rational functions provide each distinct benefits in the analysis of learning with WFAs. The Rademacher complexity of each of these classes will be analyzed in Sections 4, 5, and 6.

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:

$$A + A' = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_a\} \rangle + \langle \boldsymbol{\alpha}', \boldsymbol{\beta}', \{\mathbf{A}'_a\} \rangle = \langle \boldsymbol{\alpha} + \boldsymbol{\alpha}', \boldsymbol{\beta} + \boldsymbol{\beta}', \{\mathbf{A}_a + \mathbf{A}'_a\} \rangle$$
$$cA = c\langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_a\} \rangle = \langle c\boldsymbol{\alpha}, c\boldsymbol{\beta}, \{c\mathbf{A}_a\} \rangle \ .$$

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 $\mathcal{A}_n$:

$$\|A\|_{p,q} = \|\boldsymbol{\alpha}\|_p + \|\boldsymbol{\beta}\|_q + \max_a \|\mathbf{A}_a\|_q \ ,$$

where $\|\mathbf{A}\|_q$ denotes the matrix norm induced by the corresponding vector norm, that is $\|\mathbf{A}\|_q = \sup_{\|\mathbf{v}\|_q=1} \|\mathbf{A}\mathbf{v}\|_q$.

Now we define the classes of automata we consider. Let $p \in [1, +\infty]$ and $q = 1/(1 - 1/p)$. Then, given the triple of radii $r = (r_\alpha, r_\beta, r_\Sigma) \in \mathbb{R}^3_+$, we denote by $\mathcal{A}_{n,p,r}$ the set of all WFAs A with n states and such that: $\|\boldsymbol{\alpha}\|_p \leq r_\alpha$, $\|\boldsymbol{\beta}\|_q \leq r_\beta$, and $\max_a \|\mathbf{A}_a\|_q \leq r_\Sigma$. Thus, $\mathcal{A}_{n,p,r}$ is contained in the ball of radius $r_\alpha + r_\beta + r_\Sigma$ at the origin in the normed vector space $(\mathcal{A}_n, \|\cdot\|_{p,q})$. We note here that $\mathcal{A}_{n,p,r}$ is a compact subset of $\mathcal{A}_n$ and that for any fixed $x \in \Sigma^\star$ the function $A \mapsto f_A(x)$ is a polynomial in the weights of A and is therefore continuous in the topology induced by $\|\cdot\|_{p,q}$.

3.1.1. Examples

We consider first the class of *deterministic finite automata* (DFA). A DFA can be represented by a WFA where: $\boldsymbol{\alpha}$ is the indicator vector of the initial state; the entries of $\boldsymbol{\beta}$ are values in $\{0, 1\}$ indicating whether a state is accepting or rejecting; and, for any $a \in \Sigma$ and any $i \in [n]$ the ith row of $\mathbf{A}_a$ is either the all-zero vector if there is no transition from the ith state labeled by a, or an indicator vector with a one on the jth position if taking an a-transition from state i leads to state j. Therefore, a DFA $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_a\} \rangle$ satisfies $\|\boldsymbol{\alpha}\|_1 \leq 1$, $\|\boldsymbol{\beta}\|_\infty \leq 1$, and $\max_a \|\mathbf{A}_a\|_\infty \leq 1$ and $\mathcal{A}_{n,1,(1,1,1)}$ contains all DFA with n states.

Another important class of WFA contained in $\mathcal{A}_{n,1,(1,1,1)}$ is that of *probabilistic finite automata* (PFA). To represent a PFA as a WFA, we consider automata where: $\boldsymbol{\alpha}$ is a probability distribution over possible initial states; the vector $\boldsymbol{\beta}$ contains stopping probabilities for every state; and for every $a \in \Sigma$ and $i, j \in [n]$ the entry $\mathbf{A}_a(i, j)$ represents the probability of transitioning from state i to state j while outputting the symbol a. Any WFA satisfying these constraints clearly has $\|\boldsymbol{\alpha}\|_1 = 1$, $\|\boldsymbol{\beta}\|_\infty \leq 1$, and $\|\mathbf{A}_a\|_\infty = \max_i \sum_j |\mathbf{A}_a(i, j)| \leq 1$. The function f_A computed by a PFA A defines a probability distribution over $\Sigma^\star$; i.e. we have $f_A(x) \geq 0$ for all $x \in \Sigma^\star$ and $\sum_{x \in \Sigma^\star} f_A(x) = 1$.

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 $f \colon \Sigma^\star \to \mathbb{R}$ we use $\|f\|_p$ to denote the p-norm of f given by

$$\|f\|_p = \left(\sum_{x \in \Sigma^\star} |f(x)|^p \right)^{1/p} \quad ,$$

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

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:

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

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

3.2.1. Examples and Membership Testing

If A is a PFA, then the function f_A is a probability distribution and we have $f_A \in \mathcal{R}_{1,1}$ and by extension $\mathcal{R}_{p,1}$ for all $p \in [1, +\infty]$. On the other hand, if A is a DFA such that $f_A(x) = 1$ for infinitely many $x \in \Sigma^\star$, then $f_A \in \mathcal{R}_{\infty,1}$, but $f_A \notin \mathcal{R}_p$ for any $p < +\infty$. In fact, it is easy to see that for any $n \geq 0$, $p \in [1, +\infty]$, and $r = (r_\alpha, r_\beta, 1)$ we have $\mathcal{A}_{n,p,r} \subseteq \mathcal{R}_\infty$. The discussion above shows that PFA witness $\mathcal{A}_{n,1,(1,1,1)} \cap \mathcal{R}_1 \neq \emptyset$ and DFA with infinite support witness that $\mathcal{A}_{n,1,(1,1,1)} \cap (\mathcal{R}_\infty \setminus \mathcal{R}_1) \neq \emptyset$. Therefore, the classes $\mathcal{R}_p$ yield a more fine-grained characterization of the complexity of rational functions than what the classes $\mathcal{A}_{n,p,r}$ can provide in general.

On the other hand, testing the membership of a given WFA in $\mathcal{A}_{n,p,r}$ is a straightforward algorithmic task, while testing membership in any of the $\mathcal{R}_p$ can be challenging. Some known results include the following:

- Membership in $\mathcal{R}_{1,r}$ was shown to be semi-decidable in (Bailly and Denis, 2011).
- Membership in $\mathcal{R}_{2,r}$ can be decided in polynomial time (Cortes, Mohri, and Rastogi, 2007).
- Membership in $\mathcal{R}_{\infty,r}$ is in general undecidable (Paz, 1971).

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 on a Hilbert space contained in the free vector space $\mathbb{R}^{\Sigma^\star}$, and consider the Schatten p-norm of $\mathbf{H}_f$ as a measure of complexity of f. To make this more precise we start by noting that the set

$$\mathcal{L}_2 = \{f \colon \Sigma^\star \to \mathbb{R} \mid \|f\|_2 < \infty\}$$

together with the inner product $\langle f, g \rangle = \sum_{x \in \Sigma^\star} f(x)g(x)$ forms a separable Hilbert space. Note that we have the obvious inclusion $\mathcal{R}_2 \subset \mathcal{L}_2$, but not all functions in $\mathcal{L}_2$ are rational. Given an arbitrary function $f \colon \Sigma^\star \to \mathbb{R}$ we identify the Hankel matrix $\mathbf{H}_f$ with a (possibly unbounded) linear operator $\mathbf{H}_f \colon \mathcal{L}_2 \to \mathcal{L}_2$ defined by

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

Recall that an operator $\mathbf{H}_f$ is bounded when its operator norm is finite; i.e. $\|\mathbf{H}_f\| = \sup_{\|g\|_2 \leq 1} \|\mathbf{H}_f g\|_2 < \infty$. Furthermore, a bounded operator is compact if it can be obtained as the limit of a sequence of bounded finite-rank operators under an adequate notion of convergence. In particular, bounded finite-rank operators are compact. Our interest in compact operators on Hilbert spaces stems from the fact that these are precisely the operators for which a notion

equivalent to the SVD for finite matrices can be defined. Thus, if f is a rational function of rank n such that $\mathbf{H}_f$ is bounded (note this implies compactness by Fliess's theorem), then we can use the singular values $\mathfrak{s}_1 \geq \ldots \geq \mathfrak{s}_n$ of $\mathbf{H}_f$ as a measure of the complexity of f. The following result follows from Balle et al. (2015) and gives a useful condition for the boundedness of $\mathbf{H}_f$.

Lemma 2. *Suppose the function $f\colon \Sigma^\star \to \mathbb{R}$ is rational. Then, $\mathbf{H}_f$ is bounded if and only if $\|f\|_2 < \infty$.*

We see that every Hankel matrix $\mathbf{H}_f$ with $f \in \mathcal{R}_2$ has a well-defined SVD. Therefore, for any $f \in \mathcal{R}_2$ it makes sense to define its Schatten–Hankel p-norm as the Schatten p-norm of its Hankel matrix: $\|f\|_{\mathrm{H},p} = \|\mathbf{H}_f\|_{\mathcal{S}_p} = \|(\mathfrak{s}_1, \ldots, \mathfrak{s}_n)\|_p$, where $\mathfrak{s}_i = \mathfrak{s}_i(\mathbf{H}_f)$ is the ith singular value of $\mathbf{H}_f$ and $\mathrm{rank}(\mathbf{H}_f) = n$. 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 $r > 0$, we write $\mathcal{H}_{p,r}$ for class of rational functions with $\|f\|_{\mathrm{H},p} \leq r$.

Note that the discussion above implies $\mathcal{H}_p = \mathcal{R}_2$ for every $p \in [1, +\infty]$, and therefore we can see the classes $\mathcal{H}_{p,r}$ as providing an alternative stratification of $\mathcal{R}_2$ than the classes $\mathcal{R}_{2,r}$. As a consequence of this containment, we also have $\mathcal{R}_1 \subset \mathcal{H}_p$ for every p, and therefore the classes $\mathcal{H}_p$ include all functions computed by probabilistic automata. Since membership in $\mathcal{R}_2$ is efficiently testable (Cortes, Mohri, and Rastogi, 2007), a polynomial time algorithm by Balle, Panangaden, and Precup (2015) can be used to compute $\|f\|_{\mathrm{H},p}$ and thus test membership in $\mathcal{H}_{p,r}$.

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 $S = (x_1, \ldots, x_m) \in (\Sigma^\star)^m$ be a sample of m strings with maximum length $L_S = \max_i |x_i|$. Given $z > 0$ we define $\log_+(z) = \max\{0, \log(z)\}$. The following theorem bounds the empirical Rademacher complexity of $\mathcal{A}_{n,p,r}$ on a sample S for a setting with arbitrary radii $r = (r_\alpha, r_\beta, r_\Sigma)$.

Theorem 3. *Let $r = (r_\alpha, r_\beta, r_\Sigma)$ and $\tilde{r} = \max\{\sqrt{r_\alpha / r_\beta}, \sqrt{r_\beta / r_\alpha}, \sqrt{r_\alpha r_\beta}/r_\Sigma\}$. Define*

$$C = \sqrt{\frac{\log_+(r_\alpha r_\beta)}{kn+2}} + \sqrt{\log_+(r_\Sigma)} + \sqrt{\log_+(\tilde{r})} + 3\sqrt{\log(2)} \ .$$

Then, the following holds for any sample $S \in (\Sigma^\star)^m$:

$$\widehat{\mathfrak{R}}_S(\mathcal{A}_{n,p,r}) \leq 6\sqrt{\frac{n(kn+2) r_\alpha r_\beta r_\Sigma^{L_S}}{m}} \left[C + \sqrt{\log_+\left((L_S+2) r_\Sigma^{L_S/2} \right)} \right] \ .$$

In general, the upper bound of Theorem 3 can grow exponentially with L_S since a WFA in $\mathcal{A}_{n,p,r}$ can assign labels to strings x that grow with $|x|$ whenever $r_\Sigma > 1$. When $r_\Sigma \leq 1$, which still defines a large class of interesting WFAs, the following corollary holds. Using Jensen's inequality, the expected maximum length $L_m = \mathbb{E}_{S\sim D^m}[L_S]$ appear in the bound on the Rademacher complexity $\mathfrak{R}_m(\mathcal{A}_{n,p,r})$.

Corollary 4. *Recall the notation from Theorem 3 and assume that $r_\Sigma \leq 1$. Let $L_m = \mathbb{E}_{S\sim D^m}[L_S]$. Then the following hold for any $m \geq 1$ and any sample $S \in (\Sigma^\star)^m$:*

$$\widehat{\mathfrak{R}}_S(\mathcal{A}_{n,p,r}) \leq 6\sqrt{\frac{n(kn+2)r_\alpha r_\beta}{m}}\left(C + \sqrt{\log(L_S+2)}\right) \ ,$$
$$\mathfrak{R}_m(\mathcal{A}_{n,p,r}) \leq 6\sqrt{\frac{n(kn+2)r_\alpha r_\beta}{m}}\left(C + \sqrt{\log(L_m+2)}\right) \ .$$

Note that the presence of L_m in the bound introduces a dependency on the distribution D, which will lead to 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,[1] 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 $L_m = O(m^{1/s})$. We note that in the first two cases, we obtain a Rademacher complexity bound with rate $O(\sqrt{\log\log m/m})$, while in the power-law case the resulting rate $O(\sqrt{\log m/m})$. In particular, this provides a significant improvement over our previous results in (Balle and Mohri, 2015b) where the rate of $\mathfrak{R}_m(\mathcal{A}_{n,p,r})$ under a power-law distribution with exponent $s+1$ was shown to be $O(\max\{\sqrt{\log m/m}, m^{(1-s)/s}\})$.

4.1. Proof of Theorem 3

We begin the proof by reviewing several well-known facts and definitions related to covering numbers (see e.g. Devroye and Lugosi (2001)). Let $\mathcal{V} \subset \mathbb{R}^m$ be a set of m-dimensional vectors. An (ℓ_2, η)-covering for $\mathcal{V}$ is any set of vectors $\mathcal{C} \subset \mathbb{R}^m$ such that for every vector $\mathbf{v} \in \mathcal{V}$ there exists some vector $\mathbf{w} \in \mathcal{C}$ with ℓ_2-distance from $\mathbf{v}$ at most η; that is,

$$\|\mathbf{v} - \mathbf{w}\|_2 = \sqrt{\sum_{i=1}^{m}(\mathbf{v}_i - \mathbf{w}_i)^2} \leq \eta \ .$$

The ℓ_2-covering number of $\mathcal{V}$ at level η is the cardinality $\mathcal{N}_2(\eta, \mathcal{V})$ of the smallest (ℓ_2, η)-cover for $\mathcal{V}$:

$$\mathcal{N}_2(\eta, \mathcal{V}) = \min\left\{|\mathcal{C}| \colon \mathcal{C} \subset \mathbb{R}^m \text{ is an } (\ell_2, \eta)\text{-cover for } \mathcal{V}\right\} \ .$$

[1] 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 $\mathbb{P}[X > k] \leq O(1/k^{s+1})$.

The connection with Rademacher complexity stems from considering the covering numbers of sets of vectors $\mathcal{V}$ obtained by applying all the possible hypotheses from a class to a given set of examples. Let $S = (x_1, \dots, x_m) \in (\Sigma^\star)^m$ a sample of size m. Given a WFA A, we define $f_A(S) \in \mathbb{R}^m$ by $f_A(S) = (f_A(x_1), \dots, f_A(x_m)) \in \mathbb{R}^m$. Furthermore, we define $\mathcal{A}_{n,p,r}(S) \subset \mathbb{R}^m$ as the set $\mathcal{A}_{n,p,r}(S) = \{f_A(S) \colon A \in \mathcal{A}_{n,p,r}\}$. To prove Theorem 3, we will use the following result which provides a way to convert bounds on the covering numbers of $\mathcal{A}_{n,p,r}(S)$ into bounds on the Rademacher complexity $\widehat{\mathfrak{R}}_S(\mathcal{A}_{n,p,r})$. This result can be obtained using Dudley's chaining technique Dudley (1999).

Lemma 5 (Lemma 27.5 in (Shalev-Shwartz and Ben-David, 2014))**.** *Suppose* $\eta_0 \geq \min_{v \in \mathbb{R}^m} \max_{f_A(S) \in \mathcal{A}_{n,p,r}(S)} \|v - f_A(S)\|_2$. *Let* $\eta_i = 2^{-i}\eta_0$ *for* $i \geq 0$. *If there exist constants* θ_0, θ_1 *such that the* ℓ_2*-covering numbers of* $\mathcal{A}_{n,p,r}(S)$ *satisfy*

$$\sqrt{\log \mathcal{N}_2(\eta_i, \mathcal{A}_{n,p,r}(S))} \leq \theta_0 + i\theta_1$$

for every $i \geq 0$, *then* $\widehat{\mathfrak{R}}_S(\mathcal{A}_{n,p,r}) \leq 6\eta_0(\theta_0 + 2\theta_1)/m$.

In order to derive the required bounds for $\mathcal{N}_2(\eta, \mathcal{A}_{n,p,r}(S))$ we will make use of the following technical results.

Lemma 6 (Corollary 4.3 in Vershynin (2009))**.** *A ball* $\mathcal{B}$ *of radius* $R > 0$ *in a real* d*-dimensional Banach space can be covered by* $R^d(2 + 1/\rho)^d$ *balls of radius* $\rho > 0$ *with centres inside* $\mathcal{B}$.

Lemma 7. *Let* $r = (r_\alpha, r_\beta, r_\Sigma)$ *and* $\bar{r} = \max\{r_\alpha, r_\beta, r_\alpha r_\beta / r_\Sigma\}$. *If* $A, A' \in \mathcal{A}_{n,p,r}$ *then the following holds for any* $x \in \Sigma^\star$*:*

1. $|f_A(x)| \leq r_\alpha r_\beta r_\Sigma^{|x|}$,
2. $|f_A(x) - f_{A'}(x)| \leq \bar{r} r_\Sigma^{|x|}(\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_p + \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_q + |x| \max_a \|\mathbf{A}_a - \mathbf{A}'_a\|_q)$.

Proof. The first bound follows from applying Hölder's inequality and the sub-multiplicativity of the norms used in the definition of $\|A\|_{p,q}$ to (1). The second bound was proven in (Balle and Mohri, 2012) (see also (Balle, 2013, Lemma 5.4.2)). □

The two lemmas above can be combined to obtain the following bound on the covering numbers of $\mathcal{A}_{n,p,r}(S)$.

Lemma 8. *Let* $r = (r_\alpha, r_\beta, r_\Sigma)$ *and* $\bar{r} = \max\{r_\alpha, r_\beta, r_\alpha r_\beta / r_\Sigma\}$. *The* ℓ_2*-covering numbers of* $\mathcal{A}_{n,p,r}$ *can be bounded as follows:*

$$\mathcal{N}_2(\eta, \mathcal{A}_{n,p,r}(S)) \leq r_\alpha^n r_\beta^n r_\Sigma^{kn^2} \left(2 + \frac{\sqrt{m}(L_S + 2)\bar{r} r_\Sigma^{L_S}}{\eta}\right)^{n(kn+2)} .$$

Proof. Let $\rho_\alpha, \rho_\beta, \rho_\Sigma > 0$ be some parameters to be chosen later. Using Lemma 6, we can find the following coverings:

1. $\mathcal{C}_\alpha \subset \mathcal{B}_\alpha = \{\boldsymbol{\alpha} \in \mathbb{R}^n \colon \|\boldsymbol{\alpha}\|_p \leq r_\alpha\}$ of size at most $r_\alpha^n(2+1/\rho_\alpha)^n$ containing the centres of a covering of $\mathcal{B}_\alpha$ by balls of radius ρ_α.
2. $\mathcal{C}_\beta \subset \mathcal{B}_\beta = \{\boldsymbol{\beta} \in \mathbb{R}^n \colon \|\boldsymbol{\beta}\|_a \leq r_\beta\}$ of size at most $r_\beta^n(2+1/\rho_\beta)^n$ containing the centres of a covering of $\mathcal{B}_\beta$ by balls of radius ρ_β .
3. $\mathcal{C}_\Sigma \subset \mathcal{B}_\Sigma = \{(\mathbf{A}_{a_1}, \ldots, \mathbf{A}_{a_k}) \in (\mathbb{R}^{n\times n})^k \colon \max_a \|\mathbf{A}_a\|_q \leq r_\Sigma\}$ of size at most $r_\Sigma^{kn^2}(2+1/\rho_\Sigma)^{kn^2}$ containing the centres of a covering of $\mathcal{B}_\Sigma$ by balls of radius ρ_Σ with respect to the norm $\|(\mathbf{A}_{a_1}, \ldots, \mathbf{A}_{a_k})\| = \max_a \|\mathbf{A}_a\|_q$.

Now, given any automaton $A \in \mathcal{A}_{n,p,r}$ with $A = \langle \boldsymbol{\alpha}, \boldsymbol{\beta}, \{\mathbf{A}_a\}\rangle$ we can define another automaton $A' \in \mathcal{A}_{n,p,r}$ with $A = \langle \boldsymbol{\alpha}', \boldsymbol{\beta}', \{\mathbf{A}'_a\}\rangle$ such that $\boldsymbol{\alpha}' \in \mathcal{C}_\alpha$ with $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_p \leq \rho_\alpha$, $\boldsymbol{\beta}' \in \mathcal{C}_\beta$ with $\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_q \leq \rho_\beta$, and $(\mathbf{A}'_{a_1}, \ldots, \mathbf{A}'_{a_k}) \in \mathcal{C}_\Sigma$ with $\max_a \|\mathbf{A}_a - \mathbf{A}'_a\|_q \leq \rho_\Sigma$. Note that this is possible because of how the coverings were defined. Now using Lemma 7 and $L_S = \max_{x\in S} |x|$, we get that the following holds for any $x \in S$:

$$|f_A(x) - f'_A(x)| \leq \bar{r} r_\Sigma^{L_S}(\rho_\alpha + \rho_\beta + L_S\rho_\Sigma) \ .$$

Let $\mathcal{C} = \mathcal{C}_\alpha \times \mathcal{C}_\beta \times \mathcal{C}_\Sigma$ be the set of WFAs obtained by taking initial weights in $\mathcal{C}_\alpha$, final weights in $\mathcal{C}_\beta$ and transition weights in $\mathcal{C}_\Sigma$. By definition of the coverings we have $\mathcal{C} \subset \mathcal{A}_{n,p,r}$ and $\mathcal{C}(S) \subset \mathcal{A}_{n,p,r}(S)$. Therefore, if $\bar{r} r_\Sigma^{L_S}(\rho_\alpha + \rho_\beta + L_S\rho_\Sigma) \leq \eta/\sqrt{m}$, then $\mathcal{C}(S)$ is an (ℓ_2, η)-covering for $\mathcal{A}_{n,p,r}(S)$ of size

$$|\mathcal{C}(S)| \leq r_\alpha^n r_\beta^n r_\Sigma^{kn^2} \left(2 + \frac{1}{\rho_\alpha}\right)^n \left(2 + \frac{1}{\rho_\beta}\right)^n \left(2 + \frac{1}{\rho_\Sigma}\right)^{kn^2} \ .$$

Finally, by taking $\rho_\alpha = \rho_\beta = \rho_\Sigma = \eta/\sqrt{m}(L_S + 2)\bar{r} r_\Sigma^{L_S}$ we obtain a covering of the required size. □

The proof now follows by applying the bound in the previous lemma to the chaining result provided by Lemma 5. Start by defining the following quantities:

$$\theta'_0 = \log_+\left(r_\alpha^n r_\beta^n (2r_\Sigma)^{n(kn+2)}\right) + n(kn+2)\log_+\left((L_S+2)\tilde{r} r_\Sigma^{L_S/2}\right) \ ,$$
$$\theta'_1 = n(kn+2)\log(2) \ ,$$

where Let $\tilde{r} = \bar{r}/\sqrt{r_\alpha r_\beta} = \max\{\sqrt{r_\alpha/r_\beta}, \sqrt{r_\beta/r_\alpha}, \sqrt{r_\alpha r_\beta}/r_\Sigma\}$ as in the statement of Theorem 3. We note that both θ'_0 and θ'_1 are non-negative. In order to apply the chaining lemma we note that by setting $\mathbf{v} = 0$ in the definition of η_0 we can apply Lemma 7 and take $\eta_0 = (m r_\alpha r_\beta r_\Sigma^{L_S})^{1/2}$. The bound on $\mathcal{N}_2(\eta, \mathcal{A}_{n,p,r}(S))$ for $\eta = \eta_i = 2^{-i}\eta_0$ now yields:

$$\begin{aligned}\log \mathcal{N}_2\left(\eta_i, \mathcal{A}_{n,p,r}(S)\right) &\leq \log\left(r_\alpha^n r_\beta^n r_\Sigma^{n(kn+2)}\right) + n(kn+2)\log\left(2 + \frac{2^i(L_S+2)\bar{r} r_\Sigma^{L_S}}{\sqrt{r_\alpha r_\beta r_\Sigma^{L_S}}}\right) \\ &= \log\left(r_\alpha^n r_\beta^n r_\Sigma^{n(kn+2)}\right) + n(kn+2)\log\left(2 + 2^i(L_S+2)\tilde{r} r_\Sigma^{L_S/2}\right) \ .\end{aligned}$$

Furthermore, using the definition $\log_+(z) = \max\{0, \log(z)\} = \log(\max\{1, z\})$, we obtain for any $i \geq 1$:

$$\begin{aligned}
\log\left(2 + 2^i(L_S+2)\tilde{r}r_\Sigma^{L_S/2}\right) &\leq \log(4) + \log\left(\max\left\{1, 2^{i-1}(L_S+2)\tilde{r}r_\Sigma^{L_S/2}\right\}\right) \\
&= \log(4) + \log_+\left(2^{i-1}(L_S+2)\tilde{r}r_\Sigma^{L_S/2}\right) \\
&\leq \log(4) + \log_+\left((L_S+2)\tilde{r}r_\Sigma^{L_S/2}\right) + \log_+(2^{i-1}) \\
&= \log(2) + \log_+\left((L_S+2)\tilde{r}r_\Sigma^{L_S/2}\right) + i\log(2) \ ,
\end{aligned}$$

where the second inequality uses $\log_+(z_1 z_2) \leq \log_+(z_1) + \log_+(z_2)$. Therefore we obtain the bound $\log \mathcal{N}_2(\eta_i, \mathcal{A}_{n,p,r}(S)) \leq \theta_0' + i\theta_1'$, which implies a bound of the form required by Lemma 5 with $\theta_0 = \sqrt{\theta_0'}$ and $\theta_1 = \sqrt{\theta_1'}$. An application of the Lemma 5 now gives:

$$\widehat{\mathfrak{R}}_S(\mathcal{A}_{n,p,r}) \leq 6\sqrt{\frac{r_\alpha r_\beta r_\Sigma^{L_S}}{m}}\left(\theta_0 + 2\theta_1\right) \ .$$

Finally, we obtain the bound in Theorem 3 by noting that $(\theta_0 + 2\theta_1)/\sqrt{n(kn+2)}$ can be further upper bounded by

$$\sqrt{\frac{\log_+(r_\alpha r_\beta)}{kn+2}} + \sqrt{\log_+(r_\Sigma)} + \sqrt{\log_+(\tilde{r})} + 3\sqrt{\log(2)} + \sqrt{\log_+\left((L_S+2)r_\Sigma^{L_S/2}\right)} \ .$$

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 $\mathbf{f} \in \mathbb{R}^{\Sigma^\star}$ with $\|\mathbf{f}\|_p \leq r$. That is, $\mathbf{f}$ is an infinite vector indexed by strings in $\Sigma^\star$ whose xth entry is $\mathbf{f}_x = f(x)$. An important observation is that using this notation, for any given $x \in \Sigma^\star$, we can write $f(x)$ as the inner product $\langle \mathbf{f}, \mathbf{e}_x \rangle$, where $\mathbf{e}_x \in \mathbb{R}^{\Sigma^\star}$ is the indicator vector corresponding to string x.

Theorem 9. *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 $r > 0$:*

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

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, we can write

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

where the last equality 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 10. *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 9 and Jensen's inequality:

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

The lower bound follows directly from Khintchine–Kahane's inequality (see (Mohri et al., 2012, Theorem D.4)):

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

which completes the proof. □

The following definition 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 $C_S = \max_{x\in\Sigma^\star} s_x$ and note we have the straightforward bounds $1 \leq C_S \leq m$.

Corollary 11. *For any $m \geq 1$, any $S \in (\Sigma^\star)^m$, and any $r > 0$, the following upper bound holds:*

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

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 $\max_{\mathbf{v}_x\in V} \|\mathbf{v}_x\|_2 = \sqrt{C_S}$. Now we can apply Theorem 9 with $q = \infty$ an rewrite the supremum norm over $\mathbb{R}^{\Sigma^\star}$ as a maximum over the entries with non-zero entries and get

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

where we used $V \cup (-V)$ to mimic the absolute value in the definition of $\|\cdot\|_\infty$. Therefore, using Massart's Lemma we obtain

$$\widehat{\mathfrak{R}}_S(\mathcal{R}_{1,r}) \leq \frac{r\sqrt{2C_S \log(2m)}}{m} \quad . \qquad \square$$

Note in this case we cannot rely on the Khintchine–Kahane inequality to obtain lower bounds because there is no version of this inequality for the case $q = \infty$.

We can easily convert the above empirical bound into a standard Rademacher complexity bound by defining the expectation $C_m = \mathbb{E}_{S\sim D^m}[C_S]$ over a distribution D on $\Sigma^\star$. Note that C_m is the expected maximum number of collisions (repeated strings) in a sample of size m drawn from D. We provide a bound for C_m in terms of m in Section 7.

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 $f\colon \Sigma^\star \to \mathbb{R}$ is the bi-infinite matrix $\mathbf{H}_f \in \mathbb{R}^{\Sigma^\star\times\Sigma^\star}$ whose entries are defined by $\mathbf{H}_f(u,v) = f(uv)$. Note that any string $x \in \Sigma^\star$ admits $|x|+1$ decompositions $x = uv$ into a prefix $u \in \Sigma^\star$ and a suffix $v \in \Sigma^\star$. Thus, $\mathbf{H}_f$ contains a high degree of redundancy: for any $x \in \Sigma^\star$, $f(x)$ 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 $\|\mathbf{M}\|_{\mathcal{S}_p}$ denote the p-Schatten norm of $\mathbf{M}$ defined as the ℓ_p norm of the singular values of $\mathbf{M}$, i.e. $\|\mathbf{M}\|_{\mathcal{S}_p} = (\sum_{i\geq 1} \mathfrak{s}_i(\mathbf{M})^p)^{1/p}$. We also recall that given two matrices $\mathbf{M}_1, \mathbf{M}_2$ of the same dimensions, the matrix inner product is defined as $\langle \mathbf{M}_1, \mathbf{M}_2\rangle = \mathrm{Tr}(\mathbf{M}_1^\top \mathbf{M}_2)$. Von Neumann's trace inequality Mirsky (1975) provides the following bound for the matrix inner product in terms of the singular values of the matrices: $|\langle \mathbf{M}_1, \mathbf{M}_2\rangle| \leq \sum_{i\geq 1} \mathfrak{s}_i(\mathbf{M}_1)\mathfrak{s}_i(\mathbf{M}_2)$.

Theorem 12. *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 $\Sigma^\star$. For any decomposition $x_i = u_i v_i$ of the strings in S and any $r > 0$, the following inequality holds:*

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

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

$\mathrm{Tr}(\mathbf{e}_{v_i}\mathbf{e}_{u_i}^\top \mathbf{H}_f)$, we can use the linearity of the trace to write

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

Finally, by applying von Neumann's trace inequality to this matrix inner product, and then using Hölder's inequality to the inner product between the singular values of both matrices, the following holds:

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

Dividing the above by m yields the desired result. □

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 12 for the cases $p = 1$ and $p = 2$. First, note that the corresponding q-Schatten norms have given names: $\|\mathbf{R}\|_{\mathcal{S}_2} = \|\mathbf{R}\|_{\mathrm{F}}$ is the Frobenius norm, and $\|\mathbf{R}\|_{\mathcal{S}_\infty} = \|\mathbf{R}\|_{\mathrm{op}}$ is the operator norm.

Corollary 13. *For any $m \ge 1$ and any $r > 0$, the Rademacher complexity of $\mathcal{H}_{2,r}$ can be bounded as follows:*

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

Proof. In view of Theorem 12 and using Jensen's inequality, we can write

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

which concludes the proof. □

To bound the Rademacher complexity of $\mathcal{H}_{p,r}$ in the case $p = 1$ we will need the following moment bound for the operator norm of a random matrix from Tropp (2015).

Theorem 14 (Corollary 7.3.2 Tropp (2015)). *Let $c_1 = (2 + 8/\log(2))/3$ and $c_2 = \sqrt{2} + 4/\sqrt{\log(2)}$. Suppose $\mathbf{M} = \sum_i \mathbf{M}_i$ is a sum of i.i.d. random matrices with $\mathbb{E}[\mathbf{M}_i] = \mathbf{0}$ and $\|\mathbf{M}_i\|_{\mathrm{op}} \leq M$. Let $\sum_i \mathbb{E}[\mathbf{M}_i \mathbf{M}_i^\top] \preccurlyeq \mathbf{V}_1$, $\sum_i \mathbb{E}[\mathbf{M}_i^\top \mathbf{M}_i] \preccurlyeq \mathbf{V}_2$, and $\mathbf{V} = \mathrm{diag}(\mathbf{V}_1, \mathbf{V}_2)$. If $d = \mathrm{Tr}(\mathbf{V})/\|\mathbf{V}\|_{\mathrm{op}}$ and $\nu = \|\mathbf{V}\|_{\mathrm{op}}$, then we have*

$$\mathbb{E}[\|\mathbf{M}\|_{\mathrm{op}}] \leq c_1 M \log(d+1) + c_2 \sqrt{\nu \log(d+1)} \ .$$

We now introduce a combinatorial number depending on S and the decomposition selected for each string x_i. Let $U_S = \max_{u \in \Sigma^\star} |\{i \colon u_i = u\}|$ and $V_S = \max_{v \in \Sigma^\star} |\{i \colon v_i = v\}|$. 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. 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 15. *Let $c_1 = (2 + 8/\log(2))/3$ and $c_2 = \sqrt{2} + 4/\sqrt{\log(2)}$. For any $m \geq 1$, any $S \in (\Sigma^\star)^m$, and any $r > 0$, the following upper bound holds:*

$$\widehat{\mathfrak{R}}_S(\mathcal{H}_{1,r}) \leq \frac{r}{m} \left[c_1 \log(2m+1) + c_2 \sqrt{W_S \log(2m+1)} \right] \ .$$

Proof. First note that we can apply Theorem 14 to the random matrix $\mathbf{R}$ by letting $\mathbf{V}_1 = \sum_i \mathbf{e}_{u_i} \mathbf{e}_{u_i}^\top$ and $\mathbf{V}_2 = \sum_i \mathbf{e}_{v_i} \mathbf{e}_{v_i}^\top$. In this case we have $d = 2m$, $\nu = \max\{\|\sum_i \mathbf{e}_{u_i} \mathbf{e}_{u_i}^\top\|_{\mathrm{op}}, \|\sum_i \mathbf{e}_{v_i} \mathbf{e}_{v_i}^\top\|_{\mathrm{op}}\}$, and we get:

$$\mathbb{E}[\|\mathbf{R}\|_{\mathrm{op}}] \leq c_1 \log(2m+1) + c_2 \sqrt{\nu \log(2m+1)} \ .$$

Next, observe that $\mathbf{V}_1 = \sum_i \mathbf{e}_{u_i} \mathbf{e}_{u_i}^\top \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$ is a diagonal matrix with $\mathbf{V}_1(u,u) = \sum_i \mathbb{I}_{u=u_i}$. Thus, $\|\mathbf{V}_1\|_{\mathrm{op}} = \max_u \mathbf{V}_1(u,u) = \max_{u \in \Sigma^\star} |\{i \colon u_i = u\}| = U_S$. Similarly, we have $\|\mathbf{V}_2\|_{\mathrm{op}} = V_S$. Thus, since the decomposition of the strings in S is arbitrary, we can choose it such that $\nu = W_S$. Applying Theorem 12 now yields the desired bound. □

We can again convert the above empirical bound into a standard Rademacher complexity bound by defining the expectation $W_m = \mathbb{E}_{S \sim D^m}[W_S]$ over a distribution D on $\Sigma^\star$. We provide a bound for W_m in terms of m in next section.

7. Distribution-Dependent Rademacher Complexity Bounds

The bounds on the Rademacher complexity of $\mathcal{R}_{1,r}$ and $\mathcal{H}_{1,r}$ we presented in the previous section identify two important distribution-dependent parameters, $C_m = \mathbb{E}_S[C_S]$ and $W_m = \mathbb{E}_S[W_S]$, that reflect the impact of the distribution D on the complexity of learning these classes of rational functions. We now derive upper bounds on C_m and W_m in terms of m to give more explicit bounds on the Rademacher complexities $\mathfrak{R}_m(\mathcal{R}_{1,r})$ and $\mathfrak{R}_m(\mathcal{H}_{1,r})$.

7.1. Distribution-Dependent Bounds for $\mathfrak{R}_m(\mathcal{R}_{1,r})$

We start by rewriting C_S in a convenient way. Let $\mathcal{E} = \{e_x \colon \Sigma^\star \to \mathbb{R} \,|\, x \in \Sigma^\star\}$ be the class of all indicators on $\Sigma^\star$ given by $e_x(y) = \mathbb{I}_{x=y}$. Recall that given $S = (x_1, \ldots, x_m)$ we defined $s_x = |\{i \colon x_i = x\}|$ and $C_S = \sup_{x \in \Sigma^\star} s_x$. Using $\mathcal{E}$ we can rewrite these as $s_x = \sum_{i=1}^m e_x(x_i)$ and

$$C_S = \sup_{e_x \in \mathcal{E}} \sum_{i=1}^m e_x(x_i) \ .$$

Let $D_{\max} = \max_{x \in \Sigma^\star} \mathbb{P}_D[x]$ be the maximum probability of any strings with respect to the distribution D.

Lemma 16. *The following holds for any distribution D over $\Sigma^\star$ and any $m \geq 1$:*

$$mD_{\max} \leq C_m \leq mD_{\max} + 2m\mathfrak{R}_m(\mathcal{E}) \ .$$

Proof. We start by noting that using the expression for C_S given above we can bound $C_m = \mathbb{E}_S[C_S]$ as follows:

$$\begin{aligned}
C_m &= \mathop{\mathbb{E}}_{S \sim D^m}\left[\sup_{e_x \in \mathcal{E}} \sum_{i=1}^m e_x(x_i)\right] \\
&= \mathop{\mathbb{E}}_{S \sim D^m}\left[\sup_{e_x \in \mathcal{E}} \sum_{i=1}^m \left(e_x(x_i) + \mathop{\mathbb{E}}_{x_i' \sim D}[e_x(x_i')] - \mathop{\mathbb{E}}_{x_i' \sim D}[e_x(x_i')]\right)\right] \\
&\leq \mathop{\mathbb{E}}_{S \sim D^m}\left[\sup_{e_x \in \mathcal{E}} \sum_{i=1}^m \mathop{\mathbb{E}}_{x_i' \sim D}[e_x(x_i')]\right] + \mathop{\mathbb{E}}_{S \sim D^m}\left[\sup_{e_x \in \mathcal{E}} \sum_{i=1}^m \left(e_x(x_i) - \mathop{\mathbb{E}}_{x_i' \sim D}[e_x(x_i')]\right)\right] \\
&= m \sup_{e_x \in \mathcal{E}} \mathop{\mathbb{E}}_{x' \sim D}[e_x(x')] + \mathop{\mathbb{E}}_{S \sim D^m}\left[\sup_{e_x \in \mathcal{E}} \sum_{i=1}^m \left(e_x(x_i) - \mathop{\mathbb{E}}_{x_i' \sim D}[e_x(x_i')]\right)\right] \\
&\leq m \sup_{e_x \in \mathcal{E}} \mathop{\mathbb{E}}_{x' \sim D}[e_x(x')] + \mathop{\mathbb{E}}_{S \sim D^m}\left[\sup_{e_x \in \mathcal{E}} \left|\sum_{i=1}^m \left(e_x(x_i) - \mathop{\mathbb{E}}_{x_i' \sim D}[e_x(x_i')]\right)\right|\right] \ ,
\end{aligned}$$

where the second line introduces fresh samples $x_i' \sim D$ independent from S and the third line uses the sub-additivity of the supremum. Now, note that using the definition $e_x(x') = \mathbb{I}_{x=x'}$ we get

$$\sup_{e_x \in \mathcal{E}} \mathop{\mathbb{E}}_{x' \sim D}[e_x(x')] = \sup_{x \in \Sigma^\star} \mathbb{P}_{x' \sim D}[x' = x] = D_{\max} \ .$$

On the other hand, a standard symmetrization argument yields:

$$\mathop{\mathbb{E}}_{S \sim D^m}\left[\sup_{e_x \in \mathcal{E}} \left|\sum_{i=1}^m \left(e_x(x_i) - \mathop{\mathbb{E}}_{x_i' \sim D}[e_x(x_i')]\right)\right|\right] \leq 2m\mathfrak{R}_m(\mathcal{E}) \ .$$

To get the lower bound, note that by Jensen's inequality we also have

$$\begin{aligned} mD_{\max} &= m \sup_{e_x \in \mathcal{E}} \mathop{\mathbb{E}}_{x' \sim D}[e_x(x')] = \sup_{e_x \in \mathcal{E}} \mathop{\mathbb{E}}_{S \sim D^m} \left[\sum_{i=1}^{m} e_x(x_i) \right] \\ &\leq \mathop{\mathbb{E}}_{S \sim D^m} \left[\sup_{e_x \in \mathcal{E}} \sum_{i=1}^{m} e_x(x_i) \right] = C_m \ . \end{aligned}$$

□

To bound the Rademacher complexity $\mathfrak{R}_m(\mathcal{E})$ we will use the following lemma.

Lemma 17. *For any distribution D over $\Sigma^\star$ and any sample size $m \geq 1$ the following inequality holds for the Rademacher complexity of $\mathcal{E}$:*

$$\mathfrak{R}_m(\mathcal{E}) \leq \sqrt{\frac{\log(2)}{2m}}.$$

Proof. Let S be a sample of size m with $p \leq m$ distinct strings and with $n_1, \ldots, n_p$ occurrences for each of these strings, thus $\sum_{k=1}^{p} n_k = m$. The empirical Rademacher complexity of $\mathcal{E}$ can for that sample be expressed as follows:

$$\begin{aligned} \widehat{\mathfrak{R}}_S(\mathcal{E}) = \frac{1}{m} \mathbb{E} \left[\sup_x \sum_{i=1}^{m} \sigma_i 1_{x_i = x} \right] &= \frac{1}{m} \mathbb{E} \left[\max \left(\max_{k \in [p]} \sum_{j=1}^{n_k} \sigma_{k,j}, 0 \right) \right] \\ &\leq \frac{1}{2m} \mathbb{E} \left[\max_{k \in [p]} \left| \sum_{j=1}^{n_k} \sigma_{k,j} \right| \right] \\ &= \frac{1}{2m} \mathbb{E} \left[\max_{\substack{k \in [p] \\ s \in \{\pm 1\}}} s \sum_{j=1}^{n_k} \sigma_{k,j} \right] \ , \end{aligned}$$

where we introduced m independent Rademacher variables $\sigma_{k,j}$ indexed per string k and occurrence j instead of the original random variables σ_i. By the convexity of the exponential function and Jensen's inequality, we can write for any $t > 0$,

$$\begin{aligned} \exp \left(t \, \mathbb{E} \left[\max_{\substack{k \in [p] \\ s \in \{\pm 1\}}} s \sum_{j=1}^{n_k} \sigma_{k,j} \right] \right) &\leq \mathbb{E} \left[\exp \left(t \max_{\substack{k \in [p] \\ s \in \{\pm 1\}}} s \sum_{j=1}^{n_k} \sigma_{k,j} \right) \right] \\ &= \mathbb{E} \left[\max_{\substack{k \in [p] \\ s \in \{\pm 1\}}} e^{ts \sum_{j=1}^{n_k} \sigma_{k,j}} \right] \\ &\leq \sum_{\substack{k \in [p] \\ s \in \{\pm 1\}}} \mathbb{E} \left[e^{ts \sum_{j=1}^{n_k} \sigma_{k,j}} \right] \leq 2 \sum_{k=1}^{p} e^{\frac{t^2 n_k}{2}} \ , \end{aligned}$$

where the last step holds by Hoeffding's inequality. Let n_k^* denote the largest n_k, $k \in [p]$. Then, taking the log of both sides of the inequality and choosing t to optimize the upper bound ($t = \sqrt{2\log(2p)/n_k^*}$) yields

$$\mathbb{E}\left[\max_{\substack{k\in[p]\\ s\in\{\pm 1\}}} s\sum_{j=1}^{n_k}\sigma_{k,j}\right] \leq \frac{1}{t}\log\left[2\sum_{k=1}^{p} e^{\frac{t^2 n_k}{2}}\right]$$
$$\leq \frac{1}{t}\log\left(2pe^{\frac{t^2 n_k^*}{2}}\right) = \frac{\log(2p)}{t} + \frac{tn_k^*}{2} \leq \sqrt{2n_k^*\log(2p)} \ .$$

Thus, the following inequality holds:

$$\widehat{\mathfrak{R}}_S(\mathcal{E}) \leq \frac{\sqrt{2n_k^*\log(2p)}}{2m} \ .$$

It is straightforward to verify that the right-hand side is maximized for $p = 1$ and $n_k^* = m$, that is for a sample made of a single string repeated m times. This implies that the inequality $\widehat{\mathfrak{R}}_S(\mathcal{E}) \leq \sqrt{\log(2)/2m}$ holds for all samples S of size m. □

A straightforward application of Jensen's inequality now yields $\mathbb{E}_S[\sqrt{C_S}] \leq \sqrt{mD_{\max} + \sqrt{2\log(2)m}}$. Plugging this bound into Corollary 11 we get the following.

Corollary 18. *For any $m \geq 1$ and any $r > 0$ we have:*

$$\mathfrak{R}_m(\mathcal{R}_{1,r}) \leq \frac{r}{\sqrt{m}}\sqrt{2\log(2m)\left(D_{\max} + \sqrt{\frac{2\log(2)}{m}}\right)}.$$

7.2. Distribution-Dependent Bounds for $\mathfrak{R}_m(\mathcal{H}_{1,r})$

Next we provide bounds for W_m. Given a sample $S = (x_1, \ldots, x_m)$ we will say that the tuples of pairs of strings $S' = ((u_1, v_1), \ldots, (u_m, v_m)) \in (\Sigma^\star \times \Sigma^\star)^m$ form a *split* of S if $x_i = u_i v_i$ for all $1 \leq i \leq m$. We denote by $S^\vee$ the set of all possible splits of a sample S. We also define coordinate projections $\pi_j \colon \Sigma^\star \times \Sigma^\star \to \Sigma^\star$ given by $\pi_1(u, v) = u$ and $\pi_2(u, v) = v$. Now recall that $W_m = \mathbb{E}_S[W_S]$ and note we can rewrite the definition of W_S as

$$W_S = \min_{S'\in S^\vee}\max_{j=1,2}\sup_{e_x\in\mathcal{E}}\sum_{i=1}^{m} e_x(\pi_j(u_i, v_i))$$
$$= \min_{S'\in S^\vee}\sup_{e\in\mathcal{E}^\vee}\sum_{i=1}^{m} e(u_i, v_i) \ ,$$

where $\mathcal{E}^\vee = (\mathcal{E}\circ\pi_1)\cup(\mathcal{E}\circ\pi_2)$ and $\mathcal{E}\circ\pi_j$ is the set of functions of the form $e_x(\pi_j(u, v))$. Finally, given a distribution D over $\Sigma^\star$ we define the parameter

$$D_{\max}^\vee = \sup_{x\in\Sigma^\star}\max\left\{\sum_{v\in\Sigma^\star}\frac{1}{|x|+|v|+1}\mathbb{P}_D[xv], \sum_{u\in\Sigma^\star}\frac{1}{|x|+|u|+1}\mathbb{P}_D[ux]\right\} \ .$$

Note that the first term in the maximum above is the probability of obtaining x by first sampling a random string from D and then sampling a prefix from that string uniformly at random. Similarly, the second term is the probability of obtaining x as a random suffix from a string sampled from D. With these definitions we have the following result.

Lemma 19. *The following holds for any distribution D over $\Sigma^\star$ and any $m \geq 1$:*

$$W_m \leq m D^\vee_{\max} + 2m\mathfrak{R}_m(\mathcal{E}^\vee) \ .$$

Proof. We start by upper bounding the $\min_{S' \in S^\vee}$ by the expectation $\mathbb{E}_{S' \sim \mathbf{Unif}(S^\vee)}$ over a split chosen uniformly at random:

$$\begin{aligned}
W_m &= \mathop{\mathbb{E}}_{S \sim D^m}\left[\min_{S' \in S^\vee} \sup_{e \in \mathcal{E}^\vee} \sum_{i=1}^m e(u_i, v_i)\right] \\
&\leq \mathop{\mathbb{E}}_{S \sim D^m} \mathop{\mathbb{E}}_{S' \sim \mathbf{Unif}(S^\vee)}\left[\sup_{e \in \mathcal{E}^\vee} \sum_{i=1}^m e(u_i, v_i)\right] \\
&\leq \sup_{e \in \mathcal{E}^\vee} \mathop{\mathbb{E}}_{S \sim D^m} \mathop{\mathbb{E}}_{S' \sim \mathbf{Unif}(S^\vee)}\left[\sum_{i=1}^m e(u_i, v_i)\right] \\
&+ \mathop{\mathbb{E}}_{S \sim D^m} \mathop{\mathbb{E}}_{S' \sim \mathbf{Unif}(S^\vee)}\left[\sup_{e \in \mathcal{E}^\vee}\left|\sum_{i=1}^m \left(e(u_i, v_i) - \mathop{\mathbb{E}}_{x'_i \sim D} \mathop{\mathbb{E}}_{(u'_i, v'_i) \sim \mathbf{Unif}(\{x'_i\}^\vee)}[e(u'_i, v'_i)]\right)\right|\right] \ .
\end{aligned}$$

The same argument we used in Lemma 16 shows that the second term in the last sum above can be bounded by $2m\mathfrak{R}_m(\mathcal{E}^\vee)$. To compute the first term in the sum note that given a string y and a random split $(u, v) \sim \mathbf{Unif}(\{y\}^\vee)$, the probability that $u = x$ for some fixed $x \in \Sigma^\star$ is $1/(|y|+1)$ if x is a prefix of y and 0 otherwise. Thus, we let $e = e_x \circ \pi_1 \in \mathcal{E}^\vee$ and write

$$\begin{aligned}
\mathop{\mathbb{E}}_{S \sim D^m} \mathop{\mathbb{E}}_{S' \sim \mathbf{Unif}(S^\vee)}\left[\sum_{i=1}^m e(u_i, v_i)\right] &= m \mathop{\mathbb{E}}_{x' \sim D} \mathop{\mathbb{E}}_{(u,v) \sim \mathbf{Unif}(\{x'\}^\vee)} e_x(u) \\
&= m \mathbb{P}_{x' \sim D, (u,v) \sim \mathbf{Unif}(\{x'\}^\vee)} \mathbb{I}_{u=x} \\
&= m \sum_{x' \in x\Sigma^\star} \frac{1}{|x'|+1} \mathbb{P}_D[x'] \\
&= m \sum_{v \in \Sigma^\star} \frac{1}{|x|+|v|+1} \mathbb{P}_D[xv] \ .
\end{aligned}$$

Similarly, if we have $e = e_x \circ \pi_2 \in \mathcal{E}^\vee$ then

$$\mathop{\mathbb{E}}_{S \sim D^m} \mathop{\mathbb{E}}_{S' \sim \mathbf{Unif}(S^\vee)}\left[\sum_{i=1}^m e(u_i, v_i)\right] = m \sum_{u \in \Sigma^\star} \frac{1}{|x|+|u|+1} \mathbb{P}_D[ux] \ .$$

Thus, we can combine these equations to show that $W_m \leq m D^\vee_{\max} + 2m\mathfrak{R}_m(\mathcal{E}^\vee)$. □

Next lemma shows how to use Lemma 17 in order to bound the Rademacher complexity $\mathfrak{R}_m(\mathcal{E}^\vee)$.

Lemma 20. *For any distribution D over $\Sigma^\star$ and any sample size $m \geq 1$ the following inequality holds for the Rademacher complexity of $\mathcal{E}^\vee$:*

$$\mathfrak{R}_m(\mathcal{E}^\vee) \leq \sqrt{\frac{2\log(2)}{m}}.$$

Proof. Let $S^\vee = ((u_1, v_1), \ldots, (u_m, v_m))$ be a sample of size m. Then, by definition of $\mathcal{E}^\vee$, we can write

$$\begin{aligned}
\widehat{\mathfrak{R}}_{S^\vee}(\mathcal{E}^\vee) &= \frac{1}{m}\,\mathbb{E}\left[\sup_x \sum_{i=1}^m \sigma_i 1_{u_i=x} + \sup_x \sum_{i=1}^m \sigma_i 1_{v_i=x}\right] \\
&= \frac{1}{m}\,\mathbb{E}\left[\sup_x \sum_{i=1}^m \sigma_i 1_{u_i=x}\right] + \frac{1}{m}\,\mathbb{E}\left[\sup_x \sum_{i=1}^m \sigma_i 1_{v_i=x}\right] \\
&= \widehat{\mathfrak{R}}_{S^1}(\mathcal{E}) + \widehat{\mathfrak{R}}_{S^2}(\mathcal{E}),
\end{aligned}$$

where $S^1 = (u_1, \ldots, u_m)$ and $S^2 = (v_1, \ldots, v_m)$. This implies $\mathfrak{R}_m(\mathcal{E}^\vee) \leq 2\mathfrak{R}_m(\mathcal{E})$. The result then follows by the bound on $\mathfrak{R}_m(\mathcal{E})$ of Lemma 17. □

Finally, using Jensen's inequality on the bound from Corollary 15 we obtain the following.

Corollary 21. *Let $c_1 = (2 + 8/\log(2))/3$ and $c_2 = \sqrt{2} + 4/\sqrt{\log(2)}$. For any $m \geq 1$ and any $r > 0$ we have:*

$$\mathfrak{R}_m(\mathcal{H}_{1,r}) \leq \frac{c_1 r \log(2m+1)}{m} + \frac{c_2 r}{\sqrt{m}}\sqrt{\log(2m+1)\left(D^\vee_{\max} + \sqrt{\frac{8\log(2)}{m}}\right)}\,.$$

8. Learning and Sample Complexity Bounds

We now have all the ingredients to derive generalization bounds for learning with weighted automata for all the classes of weighted automata and rational functions introduced in the previous sections. Our learning bounds hold for loss functions that are bounded and Lipschitz. In cases where we have different bounds for the empirical and expected Rademacher complexities we also give two versions of the generalization bounds. All these bounds can be used to derive learning algorithms for weighted automata provided the right-hand side can be optimized over the corresponding hypothesis class. We will discuss in the next section some open problems related to devising efficient algorithms to solve these optimization problems. The proofs of these theorems are straightforward combinations of the bounds on the Rademacher complexity proven in the previous sections with the generalization bounds of Theorem 1.

Theorem 22. *Let D be a probability distribution over $\Sigma^\star \times \mathcal{Z}$ and let $S = ((x_i, y_i))_{i=1}^m$ be a sample of m i.i.d. examples from D. Assume that the loss $\ell\colon \mathbb{R} \times \mathcal{Z} \to \mathbb{R}_+$ is M-bounded and μ-Lipschitz with respect to its first argument. Fix $\delta > 0$. Then, the following statements hold:*

1. *For all $n \geq 1$ and $p \in [1, +\infty]$, with probability at least $1-\delta$, the following holds simultaneously for all $A \in \mathcal{A}_{n,p,r}$ with $r = (r_\alpha, r_\beta, r_\Sigma)$ and $r_\Sigma \leq 1$:*

$$\mathfrak{L}_D(A) \leq \widehat{\mathfrak{L}}_S(A) + 12\mu c_0 \sqrt{\frac{n(kn+2)r_\alpha r_\beta}{m}} + M\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \ ,$$

   where

$$c_0 = \sqrt{\frac{\log_+(r_\alpha r_\beta)}{kn+2}} + \sqrt{\log_+(r_\Sigma)} + \sqrt{\log_+(\tilde{r})} + 3\sqrt{\log(2)} + \sqrt{\log(L_m+2)} \ .$$

2. *For all $r > 0$, with probability at least $1-\delta$, the following holds simultaneously for all $f \in \mathcal{R}_{2,r}$:*

$$\mathfrak{L}_D(f) \leq \widehat{\mathfrak{L}}_S(f) + \frac{2\mu r}{\sqrt{m}} + M\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \ .$$

3. *For all $r > 0$, with probability at least $1-\delta$, the following holds simultaneously for all $f \in \mathcal{R}_{1,r}$:*

$$\mathfrak{L}_D(f) \leq \widehat{\mathfrak{L}}_S(f) + \frac{2\mu r}{\sqrt{m}} \sqrt{2\log(2m)\left(D_{\max} + \sqrt{\frac{2\log(2)}{m}}\right)} + M\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \ .$$

4. *For all $r > 0$, with probability at least $1-\delta$, the following holds simultaneously for all $f \in \mathcal{H}_{2,r}$:*

$$\mathfrak{L}_D(f) \leq \widehat{\mathfrak{L}}_S(f) + \frac{2\mu r}{\sqrt{m}} + M\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \ .$$

5. *For all $r > 0$, with probability at least $1-\delta$, the following holds simultaneously for all $f \in \mathcal{H}_{1,r}$:*

$$\mathfrak{L}_D(f) \leq \widehat{\mathfrak{L}}_S(f) + \frac{2c_1\mu r \log(2m+1)}{m} + \frac{2\mu c_2 r}{\sqrt{m}} \sqrt{\log(2m+1)\left(D^\vee_{\max} + \sqrt{\frac{8\log(2)}{m}}\right)} + M\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \ ,$$

   where $c_1 = (2 + 8/\log(2))/3$ and $c_2 = \sqrt{2} + 4/\sqrt{\log(2)}$.

Theorem 23. *Let D be a probability distribution over $\Sigma^\star \times \mathbb{R}$ and let $S = ((x_i, y_i))_{i=1}^m$ be an i.i.d. sample of size m drawn from D. Assume that the loss $\ell\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is M-bounded and μ-Lipschitz with respect to its first argument. Fix $\delta > 0$. Then, the following statements hold:*

1. *For all $n \geq 1$ and $p \in [1, +\infty]$, with probability at least $1 - \delta$, the following holds simultaneously for all $A \in \mathcal{A}_{n,p,r}$ with $r = (r_\alpha, r_\beta, r_\Sigma)$ and $r_\Sigma \leq 1$:*

$$\mathfrak{L}_D(A) \leq \widehat{\mathfrak{L}}_S(A) + 12\mu c_0 \sqrt{\frac{n(kn+2)r_\alpha r_\beta}{m}} + 3M\sqrt{\frac{\log \frac{1}{\delta}}{2m}} \ ,$$

   where

$$c_0 = \sqrt{\frac{\log_+(r_\alpha r_\beta)}{kn+2}} + \sqrt{\log_+(r_\Sigma)} + \sqrt{\log_+(\tilde{r})} + 3\sqrt{\log(2)} + \sqrt{\log(L_S + 2)} \ .$$

2. *For all $r > 0$, with probability at least $1 - \delta$, the following holds simultaneously for all $f \in \mathcal{R}_{1,r}$:*

$$\mathfrak{L}_D(f) \leq \widehat{\mathfrak{L}}_S(f) + \frac{2\mu r\sqrt{2C_S \log(2m)}}{m} + 3M\sqrt{\frac{\log \frac{2}{\delta}}{2m}} \ .$$

3. *For all $r > 0$, with probability at least $1 - \delta$, the following holds simultaneously for all $f \in \mathcal{H}_{1,r}$:*

$$\mathfrak{L}_D(f) \leq \widehat{\mathfrak{L}}_S(f) + \frac{2\mu c_1 r \log(2m+1)}{m} + \frac{2\mu c_2 r\sqrt{W_S \log(2m+1)}}{m} + 3M\sqrt{\frac{\log \frac{2}{\delta}}{2m}} \ ,$$

   where $c_1 = (2 + 8/\log(2))/3$ and $c_2 = \sqrt{2} + 4/\sqrt{\log(2)}$.

9. Conclusion

We presented the first algorithm-independent generalization bounds for learning with wide classes of WFAs. We introduced three ways to parametrize the complexity of WFAs and rational functions, each described by a different natural quantity associated with the automaton or function. We pointed out the merits of each description in the analysis of the problem of learning with WFAs, and proved upper bounds on the Rademacher complexity of several classes defined in terms of these parameters. An interesting property of these bounds is the appearance of different combinatorial parameters that tie the sample to the convergence rate: the length of the longest string L_S for $\mathcal{A}_{n,p,r}$; the maximum number of collisions C_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 (Hsu et al., 2009; Balle and Mohri, 2012). 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 of Theorem 12 for deriving strategies to help decide which prefixes and suffixes to select 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 (Balle et al., 2014a).

In (Balle and Mohri, 2012), we proposed an efficient algorithm for learning WFAs that works in two steps: a matrix completion procedure applied to Hankel matrices followed by a spectral method to obtain a WFA from such Hankel matrix. Although each of these two steps solves an optimization problem without local minima, it is not clear from the analysis that the solution of the combined procedure is close to the empirical risk minimizer of any of the classes introduced in this paper. Nonetheless, we expect that the tools developed in this paper will prove useful in analyzing variants of this algorithm and will further help design new algorithms for efficiently learning interesting classes of WFA.

Acknowledgments

MM's work was partly funded by NSF CCF-1535987 and IIS-1618662.

References

Abe, N., Warmuth, M. K., 1992. On the computational complexity of approximating distributions by probabilistic automata. Machine Learning.

Allauzen, C., Byrne, B., de Gispert, A., Iglesias, G., Riley, M., 2014. Pushdown automata in statistical machine translation. Computational Linguistics 40 (3), 687–723.

Allauzen, C., Mohri, M., Riley, M., 2004. Statistical modeling for unit selection in speech synthesis. In: Proceedings of ACL.

Allauzen, C., Mohri, M., Talwalkar, A., 2008. Sequence kernels for predicting protein essentiality. In: Proceedings of ICML.

Aminof, B., Kupferman, O., Lampert, R., 2011. Formal analysis of online algorithms. In: Proceedings of ATVA.

Baier, C., Größer, M., Ciesinski, F., 2009. Model checking linear-time properties of probabilistic systems. In: Handbook of Weighted automata. Springer.

Bailly, R., Denis, F., 2011. Absolute convergence of rational series is semi-decidable. Inf. Comput.

Bailly, R., Denis, F., Ralaivola, L., 2009. Grammatical inference as a principal component analysis problem. In: ICML.

Balle, B., 2013. Learning finite-state machines: Algorithmic and statistical aspects. Ph.D. thesis, Universitat Politècnica de Catalunya.

Balle, B., Carreras, X., Luque, F., Quattoni, A., 2014a. Spectral learning of weighted automata: A forward-backward perspective. Machine Learning.

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

Balle, B., Mohri, M., 2012. Spectral learning of general weighted automata via constrained matrix completion. In: NIPS.

Balle, B., Mohri, M., 2015a. Learning weighted automata. In: CAI.

Balle, B., Mohri, M., 2015b. On the Rademacher complexity of weighted automata. In: Proceedings of ALT.

Balle, B., Panangaden, P., Precup, D., 2015. A canonical form for weighted automata and applications to approximate minimization. In: Logic in Computer Science (LICS).

Bartlett, P. L., Mendelson, S., 2001. Rademacher and gaussian complexities: Risk bounds and structural results. In: COLT.

Berstel, J., Reutenauer, C., 1988. Rational Series and Their Languages. Springer.

Berstel, J., Reutenauer, C., 2011. Noncommutative rational series with applications. Cambridge University Press.

Boots, B., Siddiqi, S., Gordon, G., 2009. Closing the learning-planning loop with predictive state representations. In: RSS.

Breuel, T. M., 2008. The OCRopus open source OCR system. In: Proceedings of IS&T/SPIE.

Carlyle, J. W., Paz, A., 1971. Realizations by stochastic finite automata. J. Comput. Syst. Sci. 5 (1).

Cortes, C., Haffner, P., Mohri, M., 2004. Rational kernels: Theory and algorithms. Journal of Machine Learning Research 5.

Cortes, C., Mohri, M., Rastogi, A., 2007. l_p distance and equivalence of probabilistic automata. International Journal of Foundations of Computer Science.

Culik II, K., Kari, J., 1993. Image compression using weighted finite automata. Computers & Graphics 17 (3).

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

Denis, F., Gybels, M., Habrard, A., 2016. Dimension-free concentration bounds on hankel matrices for spectral learning. Journal of Machine Learning Research 17 (31), 1–32.

Devroye, L., Lugosi, G., 2001. Combinatorial methods in density estimation. Springer.

Droste, M., Kuich, W., Vogler, H. (Eds.), 2009. Handbook of weighted automata. EATCS Monographs on Theoretical Computer Science. Springer.

Dudley, R. M., 1999. Uniform central limit theorems. Vol. 23. Cambridge Univ Press.

Durbin, R., Eddy, S. R., Krogh, A., Mitchison, G. J., 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.

Eilenberg, S., 1974. Automata, Languages and Machines. Academic Press.

Fliess, M., 1974. Matrices de Hankel. Journal de Mathématiques Pures et Appliquées 53.

Hamilton, W. L., Fard, M. M., Pineau, J., 2013. Modelling sparse dynamical systems with compressed predictive state representations. In: ICML.

Hsu, D., Kakade, S. M., Zhang, T., 2009. A spectral algorithm for learning hidden Markov models. In: COLT.

Ishigami, Y., Tani, S., 1997. Vc-dimensions of finite automata and commutative finite automata with k letters and n states. Discrete Applied Mathematics.

Kaplan, R. M., Kay, M., 1994. Regular models of phonological rule systems. Computational Linguistics 20 (3).

Karttunen, L., 1995. The replace operator. In: Proceedings of ACL.

Koltchinskii, V., Panchenko, D., 2000. Rademacher processes and bounding the risk of function learning. In: High Dimensional Probability II. Birkhäuser, pp. 443–459.

Kuich, W., Salomaa, A., 1986. Semirings, Automata, Languages. No. 5 in EATCS Monographs on Theoretical Computer Science. Springer-Verlag, Berlin-New York.

Kulesza, A., Jiang, N., Singh, S., 2015. Low-rank spectral learning with weighted loss functions. In: AISTATS.

Kulesza, A., Rao, N. R., Singh, S., 2014. Low-Rank Spectral Learning. In: AISTATS.

Ledoux, M., Talagrand, M., 1991. Probability in Banach spaces. Springer-Verlag.

Mirsky, L., 1975. A trace inequality of John von Neumann. Monatshefte für Mathematik.

Mohri, M., 1997. Finite-state transducers in language and speech processing. Computational Linguistics 23 (2).

Mohri, M., 2009. Weighted automata algorithms. In: Handbook of Weighted Automata. Monographs in Theoretical Computer Science. Springer, pp. 213–254.

Mohri, M., Pereira, F., Riley, M., 1996. Weighted automata in text and speech processing. In: Proceedings of ECAI-96 Workshop on Extended finite state models of language.

Mohri, M., Pereira, F. C. N., 1998. Dynamic compilation of weighted context-free grammars. In: Proceedings of COLING-ACL.

Mohri, M., Pereira, F. C. N., Riley, M., 2008. Speech recognition with weighted finite-state transducers. In: Handbook on Speech Processing and Speech Comm. Springer.

Mohri, M., Rostamizadeh, A., Talwalkar, A., 2012. Foundations of machine learning. MIT press.

Mohri, M., Sproat, R., 1996. An efficient compiler for weighted rewrite rules. In: Proceedings of ACL.

Paz, A., 1971. Introduction to Probabilistic Automata. Academic Press, Inc.

Pereira, F., Riley, M., 1997. Speech recognition by composition of weighted finite automata. In: Finite-State Language Processing. MIT Press.

Roche, E., 1994. Two parsing algorithms by means of finite state transducers. In: 15th International Conference on Computational Linguistics, COLING 1994, Kyoto, Japan, August 5-9, 1994. pp. 431–435.

Salomaa, A., Soittola, M., 1978. Automata-Theoretic Aspects of Formal Power Series. Springer-Verlag: New York.

Shalev-Shwartz, S., Ben-David, S., 2014. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.

Sproat, R., 1995. A finite-state architecture for tokenization and grapheme-to-phoneme conversion in multilingual text analysis. In: Proceedings of the ACL SIGDAT Workshop. ACL.

Tropp, J. A., 2015. An introduction to matrix concentration inequalities. Foundations and Trends® in Machine Learning 8 (1-2), 1–230.

Vershynin, R., 2009. Lectures in Geometrical Functional Analysis. Preprint.