

On Distribution Classes Induced by Probabilistic Automata

(student submission)

Omri Guttman, S. V. N. Vishwanathan, and Robert C. Williamson

Statistical Machine Learning Program
National ICT Australia
RSISE, Australian National University
Canberra, ACT, Australia.

{Omri.Guttman,SVN.Vishwanathan,Bob.Williamson}@nicta.com.au

Abstract. Probabilistic finite automata induce distributions over finite length strings. The two major categories of probabilistic automata are probabilistic finite automata (or *PFA*) and probabilistic deterministic finite automata (or *PDFA*). In PFA models, strings may be generated by a (possibly very large) number of paths through states, while in the PDFA category at most a single generating path exists for each string. In the first part of the paper we characterize the families of distributions inducible by PFA and PDFA models by extending the well known Myhill-Nerode theorem. In the second part we use this characterization to bound the distance between the class of bounded length distributions and the class of distributions inducible by (n -state) PFA and PDFA models.

1 Introduction

Probabilistic finite automata models are pervasive in the machine learning literature and serve a crucial role in numerous real-world engineering applications. In the fields of computational linguistics, speech recognition, computational biology and machine translation, the core technology draws heavily on various forms of PFA models, and the algorithmic and theoretical aspects of the models have been studied extensively. In a recent pair of survey papers (Vidal et al., 2005a), (Vidal et al., 2005b), the state-of-the-art in the PFA field is outlined. We will adhere to the notation used in these papers where relevant.

PFA models can be seen as a means of inducing probability distributions over finite length strings. The PFA family comprises many distinct subfamilies, which in turn induce different classes of probability distributions. However little is known about the relative power of PFAs and PDFAs in approximating arbitrary probability distributions. In Section 3 we formulate and prove characterization results for the families of distributions that can be induced by PFA and PDFA models, extending the well-known Myhill Nerode theorem (Hopcroft and Ullman, 1979).

In Section 4 we use the extended Myhill-Nerode characterization theorems to study the distance between the set of distributions induced by n -state PFA / PDFA models and the class of bounded length distributions. When studying distances between distributions, one is required to adopt a specific distance function. In the PFA literature, two distance functions are typically used: the KL-divergence and the L_1 -distance (formally defined in Section 2). We choose to use the L_1 -distance and motivate our choice in Section 2.3.

2 Preliminaries and Notation

We now formally define the PFA model:

Definition 1. A PFA is a tuple $\mathcal{A} = \langle Q_{\mathcal{A}}, \Sigma, \delta_{\mathcal{A}}, I_{\mathcal{A}}, F_{\mathcal{A}}, P_{\mathcal{A}} \rangle$, where:

- $Q_{\mathcal{A}}$ is a finite set of states,
- Σ is a finite alphabet,
- $\delta_{\mathcal{A}} \subseteq Q_{\mathcal{A}} \times \Sigma \times Q_{\mathcal{A}}$ is a set of transitions,
- $I_{\mathcal{A}} : Q_{\mathcal{A}} \rightarrow \mathbb{R}^+$ are the initial state probabilities,
- $P_{\mathcal{A}} : \delta_{\mathcal{A}} \rightarrow \mathbb{R}^+$ are the transition probabilities,
- $F_{\mathcal{A}} : Q_{\mathcal{A}} \rightarrow \mathbb{R}^+$ are the final state probabilities.

The transition probabilities of non-existing transitions are null, i.e. $P_{\mathcal{A}}(q, \sigma, q') = 0$ for all $(q, \sigma, q') \notin \delta_{\mathcal{A}}, \sigma \in \Sigma$. The initial state probabilities $I_{\mathcal{A}}$ satisfy $\sum_{q \in Q_{\mathcal{A}}} I_{\mathcal{A}}(q) = 1$, while for the transition and final state probabilities we have:

$$F_{\mathcal{A}}(q) + \sum_{\substack{\sigma \in \Sigma \\ q' \in Q_{\mathcal{A}}}} P_{\mathcal{A}}(q, \sigma, q') = 1.$$

The PFA \mathcal{A} induces a distribution over the set of all strings over symbols in Σ (denoted Σ^* , which also includes the empty string ϵ), in the following manner. Let $\theta = (s_0, x_1, s_1, x_2, s_2, \dots, s_{k-1}, x_k, s_k)$ denote a *path* producing the string $x = (x_1 x_2 \dots x_k)$. Equivalently, there is a sequence of transitions $(s_0, x_1, s_1), (s_1, x_2, s_2), \dots, (s_{k-1}, x_k, s_k) \in \delta_{\mathcal{A}}$. The probability of generating such a path is:

$$\Pr_{\mathcal{A}}(\theta) = I_{\mathcal{A}}(s_0) \cdot \left(\prod_{j=1}^k P_{\mathcal{A}}(s_{j-1}, x_j, s_j) \right) \cdot F_{\mathcal{A}}(s_k). \quad (1)$$

Definition 2. A valid path in a PFA \mathcal{A} is a path for some $x \in \Sigma^*$ with probability greater than zero. The set of valid paths in \mathcal{A} will be denoted by $\Theta_{\mathcal{A}}$.

In general, a given string x can be generated by \mathcal{A} through multiple valid paths. Let $\Theta_{\mathcal{A}}(x)$ denote the set of all the valid paths for x in \mathcal{A} . The probability of generating x with \mathcal{A} is:

$$\Pr_{\mathcal{A}}(x) = \sum_{\theta \in \Theta_{\mathcal{A}}(x)} \Pr_{\mathcal{A}}(\theta). \quad (2)$$

The set $\Theta_{\mathcal{A}}(x)$ can potentially have cardinality that is exponential in the length of the string being generated, specifically $\mathcal{O}(|Q_{\mathcal{A}}|^{|x|})$, with $|x|$ denoting

the length of string x . This precludes any attempt at direct calculation. However, a simple dynamic programming recursion known as the *forward* algorithm (discovered by Chang and Hancock (1966) and rediscovered by Baum et al. (1970)) reduces the computational complexity to merely *linear* in $|x|$ and $|\delta|$, where $|x|$ denotes the length of x and $|\delta|$ denotes the number of transitions in \mathcal{A} . We now describe their recursion, which will facilitate the proofs of the Myhill Nerode extension theorems of Section 3.

2.1 The Forward Algorithm

In describing the forward algorithm we follow the notation of Vidal et al. (2005a) and duplicate the relevant part of their exposition. Define $\alpha_x(i, q)$, $\forall q \in Q$ and $0 \leq i \leq |x|$, as the probability of generating the prefix $x_1 \dots x_i$ and reaching state q :

$$\alpha_x(i, q) = \sum_{(s_0, s_1, \dots, s_i) \in \Theta_{\mathcal{A}}(x_1 \dots x_i)} I_{\mathcal{A}}(s_0) \cdot \prod_{j=1}^i P_{\mathcal{A}}(s_{j-1}, x_j, s_j) \cdot 1(q, s_i), \quad (3)$$

where $1(q, q') = 1$ if $q = q'$ and 0 if $q \neq q'$. Equation (3) can be calculated using the following recursion:

$$\alpha_x(0, q) = I_{\mathcal{A}}(q), \quad (4a)$$

$$\alpha_x(i, q) = \sum_{q' \in Q} \alpha_x(i-1, q') \cdot P_{\mathcal{A}}(q', x_i, q), \quad 1 \leq i \leq |x|. \quad (4b)$$

We note that the forward densities $\alpha_x(i, q)$ distribute over $Q_{\mathcal{A}}$ (for all x and i). For a string $x \in \Sigma^*$, the following proposition is straightforward:

$$\Pr_{\mathcal{A}}(x) = \sum_{q \in Q} \alpha_x(|x|, q) \cdot F_{\mathcal{A}}(q). \quad (5)$$

The evaluation of α can thus be performed in $\mathcal{O}(|x| \cdot |\delta|)$ calculations.

2.2 Probabilistic Deterministic Finite Automata

An interesting special case of the PFA occurs when the following restrictions are placed:

- $\exists q_0 \in Q$ (initial state), such that $I_{\mathcal{A}}(q_0) = 1$,
- $\forall q \in Q, \forall \sigma \in \Sigma, |\{q' : (q, \sigma, q') \in \delta_{\mathcal{A}}\}| \leq 1$.

The resulting model is termed a probabilistic deterministic finite automaton, or *PDFA* (also referred to as the deterministic probabilistic finite automaton (*DPFA*) or the stochastic deterministic finite automaton (*SDFA*)¹). In the PDFA model, the *single* path through states can be deterministically recovered, given the generated string.

¹ The corresponding term in the information theory literature is the *unifilar finite-state source* or *unifilar source*.

2.3 Distance Functions Between Distributions

Given two distributions D_1 and D_2 over Σ^* , we will use the following notions of distance: For $1 \leq p < \infty$, the L_p distance between D_1 and D_2 is defined as:

$$\|D_1 - D_2\|_p = \left[\sum_{s \in \Sigma^*} |D_1(s) - D_2(s)|^p \right]^{1/p}.$$

The KL-divergence between D_1 and D_2 is defined as:

$$\text{KL}(D_1 \| D_2) = \sum_{s \in \Sigma^*} D_1(s) \log \left(\frac{D_1(s)}{D_2(s)} \right).$$

Note that the KL divergence is *not* a metric. As shown in (Cover and Thomas, 1991), for any pair of distributions D_1 and D_2 , it holds that $\text{KL}(D_1 \| D_2) \geq \frac{1}{2 \ln 2} \|D_1 - D_2\|_1^2$, which in turn upper-bounds all L_p -distances, making this notion of distance stronger. Learnability of PDFA models (as defined by Clark and Thollard (2004)) has been widely investigated. There are indications that the general PDFA learning problem is hard. For instance, Kearns et al. (1994) showed that KL-PAC learnability of PDFA implies the computability of the *noisy parity function*, thus violating the *noisy parity assumption*, widely believed to be true in the cryptography community (see e.g. Kearns, 1993). This is demonstrated by showing how by KL-PAC learning a specific family of (acyclic) PDFA, one can evaluate the noisy parity function. We have shown (and will publish in the full version of this paper) that even in the (weaker) L_1 -PAC learning model, general PDFA learnability still violates the noisy parity assumption. This implies that the difficulty is inherent to PDFA learning, and not merely an artifact of the KL-divergence. We also mention that this specific reduction does not extend to the L_p -PAC, $p > 1$ learning framework. Thus motivated, we adopt the L_1 -distance as the distance function to be used throughout the rest of the paper.

3 An Extension of the Myhill-Nerode Theorem

In this section we formulate and prove two extensions of the Myhill-Nerode theorem for PDFA and PFA models. We begin by formally defining the notions of a *suffix distribution* and the *suffix set*:

Definition 3. Given a distribution D over Σ^* and a string $s \in \Sigma^*$ such that $D(s) > 0$, the *suffix distribution* $D_{[s]}$ is defined by:

$$D_{[s]}(x) = \frac{D(sx)}{D(s)} \quad \forall x \in \Sigma^*. \quad (6)$$

Definition 4. The set of all suffix distributions of a given distribution D is denoted by $\text{suff}(D)$:

$$\text{suff}(D) = \{D_{[s]}\}_{s \in \Sigma^*}.$$

Note that Definition 3 implies $D \in \text{succ}(D)$. For the class of PDFA models, a Myhill-Nerode type result was formulated and proved in (Carrasco and Oncina, 1999), where the authors defined a *canonical generator*, or a minimal PDFA inducing a specific distribution. We present the following *characterization*, completing the extension:

Theorem 1 (Myhill-Nerode Extension for PDFA Models). *Let Σ be a finite alphabet. The following two statements are equivalent:*

1. *A distribution D over Σ^* can be induced by an n -state PDFA.*
2. *All suffix distributions of D are members of a set of at most n fixed distributions over Σ^* .*

A further extension applies for the general class of PFA models:

Theorem 2 (Myhill-Nerode Extension for PFA Models). *The following two statements are equivalent:*

1. *A distribution D over Σ^* can be induced by an n -state PFA.*
2. *All suffix distributions of D are in the convex hull of at most n fixed distributions over Σ^* .*

We proceed to prove Theorem 2; the proof of Theorem 1 follows as a special case.

Proof. 1 \Rightarrow 2: Assuming the distribution D is induced by some PFA $\mathcal{A} = \langle Q_{\mathcal{A}}, \Sigma, \delta_{\mathcal{A}}, I_{\mathcal{A}}, F_{\mathcal{A}}, P_{\mathcal{A}} \rangle$ with $|Q_{\mathcal{A}}| \leq n$, we show that the set of its suffix distributions is contained in the convex hull of at most n distributions. For simplicity we assume $|Q_{\mathcal{A}}| = n$.

Let e_k be a vector with the k^{th} entry equal to 1 and all other entries 0. Define the PFA \mathcal{A}_k ($k = 1, \dots, n$) by $\mathcal{A}_k = \langle Q_{\mathcal{A}}, \Sigma, \delta_{\mathcal{A}}, e_k, F_{\mathcal{A}}, P_{\mathcal{A}} \rangle$ (i.e. a PFA identical to \mathcal{A} , except for the initial state distribution which is concentrated on the k^{th} state). Let $s \in \Sigma^*$ be a prefix, $x \in \Sigma^*$ a suffix, and let D_k , $k = 1, \dots, n$, denote the distribution induced by \mathcal{A}_k . In analogy to Equation (3), the forward density for \mathcal{A}_k is given by:

$$\alpha_x^k(i, q) = I_{\mathcal{A}}(s_k) \cdot \sum_{(s_1, \dots, s_i) \in \Theta_{\mathcal{A}}(x_1 \dots x_i)} \prod_{j=1}^i P_{\mathcal{A}}(s_{j-1}, x_j, s_j) \cdot 1(q, s_i),$$

or in other words, $\alpha_x^k(i, q)$ denotes the probability of generating the prefix $x_1 \dots x_i$ and reaching state q after *starting* from state s_k . As there exist only n possible states to reach, we have the following equalities:

$$D_k(s) = \sum_{\ell=1}^n \alpha_s^k(|s|, q_{\ell}); \quad D_k(sx) = \sum_{\ell=1}^n \alpha_s^k(|s|, q_{\ell}) D_{\ell}(x). \quad (7)$$

$D_k(x)$ denotes the probability of generating a string $x \in \Sigma^*$ conditioned on the fact that we started from state q_k . Since the probability of starting in

the state q_k is given by $I_{\mathcal{A}}(q_k)$, the overall probability of generating x becomes $D(x) = \sum_{k=1}^n I_{\mathcal{A}}(q_k)D_k(x)$. Using Definition 3 and plugging in (7), we get:

$$\begin{aligned} D_{[s]}(x) &= \frac{D(sx)}{D(s)} = \frac{\sum_{k=1}^n I_{\mathcal{A}}(q_k)D_k(sx)}{\sum_{k=1}^n I_{\mathcal{A}}(q_k)D_k(s)} \\ &= \frac{\sum_{k=1}^n I_{\mathcal{A}}(q_k) \cdot \sum_{\ell=1}^n \alpha_s^k(|s|, q_\ell) D_\ell(x)}{\sum_{k=1}^n I_{\mathcal{A}}(q_k) \cdot \sum_{\ell=1}^n \alpha_s^k(|s|, q_\ell)} \\ &= \sum_{\ell=1}^n \left[\frac{\sum_{k=1}^n I_{\mathcal{A}}(q_k) \cdot \alpha_s^k(|s|, q_\ell)}{\sum_{k=1}^n I_{\mathcal{A}}(q_k) \cdot \sum_{\ell=1}^n \alpha_s^k(|s|, q_\ell)} \right] D_\ell(x). \end{aligned} \quad (8)$$

The bracketed coefficients are nonnegative, sum to 1, and do not depend on x , so the resulting expression is a convex combination of the distributions $\{D_\ell(\cdot)\}_{\ell=1}^n$, proving the claim.

2 \Rightarrow 1: The distribution D is in $\text{stuff}(D)$ and therefore in $\text{co}(D_1, \dots, D_n)$ (the convex hull of (D_1, \dots, D_n)). Thus, there exist nonnegative, sum-to-one (*i.e.*, convex) coefficients $\{\pi_1, \dots, \pi_n\}$ such that $D = \sum_{i=1}^n \pi_i D_i$.

Intuitively, for every $\sigma \in \Sigma$ the distribution $(D_i)_{[\sigma]}$ is in $\text{co}(D_1, \dots, D_n)$, and hence can be written as $(D_i)_{[\sigma]}(x) = \sum_j \lambda_{\sigma,i}^j D_j(x)$ for some set of convex coefficients $\lambda_{\sigma,i}^j$. When generating a string $s = s_1 s_2 \dots s_k$, we first pick D_i with probability π_i . Then we pick D_j with probability $\lambda_{s_1,i}^j$, output s_1 , and proceed to generate $s_2 \dots s_k$. But since $(D_j)_{[s_2]}$ is itself in $\text{co}(D_1, \dots, D_n)$, the process is recursively repeated. Formally:

$$\begin{aligned} D(s) &= \sum_{i=1}^n \pi_i D_i(s_1 s_2 \dots s_k) = \sum_{i=1}^n \pi_i (D_i)_{[s_1]}(s_2 \dots s_k) \\ &= \sum_{i=1}^n \pi_i \sum_{j_1=1}^n \lambda_{s_1,i}^{j_1} D_{j_1}(s_2 \dots s_k) = \sum_{i=1}^n \pi_i \sum_{j_1=1}^n \lambda_{s_1,i}^{j_1} \sum_{j_2=1}^n \lambda_{s_2,j_1}^{j_2} D_{j_2}(s_3 \dots s_k) \\ &= \dots = \sum_{i=1}^n \pi_i \sum_{j_1=1}^n \lambda_{s_1,i}^{j_1} \sum_{j_2=1}^n \lambda_{s_2,j_1}^{j_2} \dots \sum_{j_k=1}^n \lambda_{s_k,j_{(k-1)}}^{j_k} D_{j_k}(\epsilon). \end{aligned}$$

In order to construct an n -state PFA $\mathcal{A} = \langle Q_{\mathcal{A}}, \Sigma, \delta_{\mathcal{A}}, I_{\mathcal{A}}, F_{\mathcal{A}}, P_{\mathcal{A}} \rangle$ which induces the above distribution, we proceed as follows: identify each D_i with a state q_i of \mathcal{A} and set $I_{\mathcal{A}} = (\pi_1, \dots, \pi_n)$. The transition probabilities $P_{\mathcal{A}}(q_i, \sigma, q_j)$ are set to $\lambda_{\sigma,i}^j$, and the final state probabilities are set to $F_{\mathcal{A}}(q_i) = D_i(\epsilon)$. Using Equation (5) and writing out the recursion in (4), we have $\text{Pr}_{\mathcal{A}}(s) = D(s)$ for any string $s \in \Sigma^*$. ■

The proof of Theorem 1 follows as a simple special case. In the first direction, the forward densities and initial state probabilities of (8) reduce to delta distributions, implying the suffix probabilities are in the set $\{D_1, \dots, D_n\}$. In the second direction, the PDFAs case has a delta initial distribution, setting the initial state. The parameters $\{\lambda_{\sigma,i}^j\}_{i,j=1}^n$ adhere to the PDFAs constraints, and the constructed automaton \mathcal{A} reduces to a PDFAs.

4 Bounds on Distances Between Distribution Classes

In this section we seek to answer the following question: how well can PFA and PDFA models approximate general distributions of bounded length. To define the problem rigorously, we need to choose a suitable distance metric and impose a number of constraints. In the sections below we formulate and prove lower bounds on the ability of (n -state) PDFA and PFA models to approximate the class of bounded length distributions. But first we present an immediate upper bound:

Lemma 1. *Let D be a distribution with length bounded by L (i.e. $w \sim D$ implies $|w| \leq L$). Then A PDFA \mathcal{A} with $|\Sigma|^{L+1} - 1$ states can be constructed which induces a distribution exactly matching D .*

Proof. We construct a $|\Sigma|$ -ary tree of depth (at most) $L + 1$ and set \mathcal{A} 's initial state q_0 to be the tree's root. For q_i denoting one of the tree's internal nodes, we let $w(q_i)$ denote the sequence of letters that had been traversed while reaching q_i from q_0 . For each alphabet letter $\sigma \in \Sigma$ and internal node q_i we label one outgoing edge from q_i with σ , thus defining the state transition function $\delta_{\mathcal{A}}$. For a pair of nodes (q_i, q_j) such that $(q_i, \sigma, q_j) \in \delta_{\mathcal{A}}$, we define the transition probability $P_{\mathcal{A}}(q_i, \sigma, q_j)$ by:

$$P_{\mathcal{A}}(q_i, \sigma, q_j) = \frac{\sum_{x \in \Sigma^*} D(w(q_j)x)}{\sum_{x \in \Sigma^*} D(w(q_i)x)}.$$

The final state probability for state q_i is set to:

$$F_{\mathcal{A}} = \frac{D(w(q_i))}{\sum_{x \in \Sigma^+} D(w(q_i)x)},$$

where $\Sigma^+ = \Sigma^* \setminus \{\epsilon\}$. It is readily shown that the PDFA \mathcal{A} thus constructed induces the distribution D . ■

A natural question regards the PFA / PDFA models' ability to approximate the class of bounded *expected* length distributions. The difficulty of such PFA approximation is shown in the following lemma, for which a proof sketch is presented in Appendix A. The proof relies on the lower bound for PFA approximation presented in Section 4.1 below. An analogous result for PDFA-based approximation can also be formulated, based on the lower bound shown in Section 4.2.

Lemma 2. *For any $\epsilon > 0$ there exists a distribution D^{**} of expected length bounded by L (i.e. $\mathbb{E}_{w \sim D} |w| \leq L$), such that for any PFA \mathcal{A} with no more than $|\Sigma|^{\lfloor \frac{L-1}{8\epsilon} \rfloor}$ states, the following will hold:*

$$\|D^{**} - D_{\mathcal{A}}\|_1 \geq \epsilon.$$

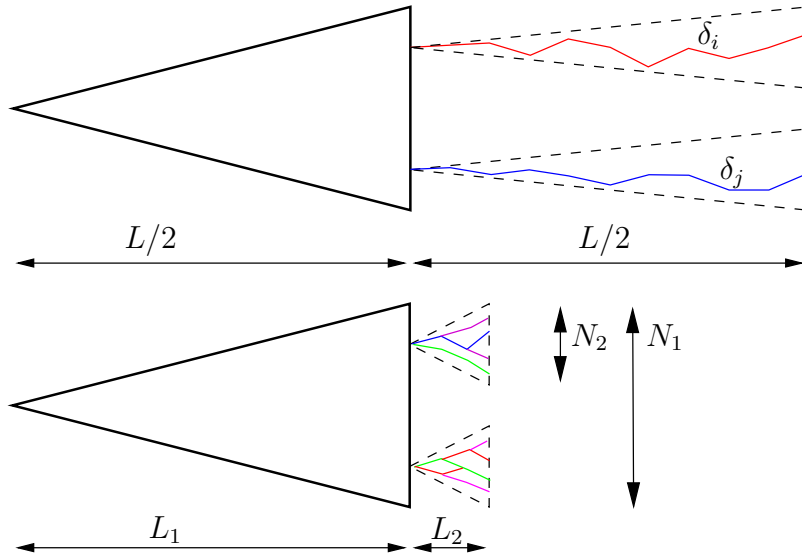


Fig. 1. An illustration of the techniques used for proving lower bounds on the approximation ability of PFA (top) and PDFA (bottom) models. The triangles on the left illustrate uniform distributions, while the suffix distributions on the right hand side serve as “symmetry-breakers”.

We proceed to formulate and prove lower bounds for approximation of bounded length distributions by PDFA / PFA models. The techniques used for obtaining the lower bounds are illustrated in Figure 1. In both cases, the target distribution is composed of a uniform prefix distribution followed by a “symmetry-breaking” set of suffixes.

In the case of the PFA, we require each suffix to be poorly approximable by the convex hull of all other suffixes, so all the chosen suffixes are distinct delta distributions. Thus, the L_1 -distance between each suffix distribution and the convex hull of all other suffix distributions equals 2. In the PDFA case, we require a weaker condition, namely that the L_1 -distance between each suffix pair is at least $1/2$. Therefore, (exponentially) shorter suffix lengths suffice.

4.1 Lower bound for PFA Approximation of Bounded Length Distributions

In this section we apply the Myhill-Nerode theorem for PFA to derive a lower bound on the models’ approximation ability.

Theorem 3. *There exists a distribution D^* of bounded length L such that for any distribution \hat{D} induced from a PFA with no more than $|\Sigma|^{(\frac{L}{2}-1)}$ states the following will hold: $\|D^* - \hat{D}\|_1 \geq 1/2$.*

Before proving the theorem, we define the notion of *suffix mass*:

Definition 5. The suffix mass of a family of probability distributions \mathcal{D} is defined by:

$$SM(\mathcal{D}) = \max_{D \in \mathcal{D}} \sum_{w \in \Sigma^*} \max_{S \in \text{suffix}(D)} S(w).$$

For the family of distributions induced by n -state PFAs we now show:

Lemma 3. Let $\mathcal{D}_n = \{D : D \text{ is induced by an } n\text{-state PFA}\}$. Then $SM(\mathcal{D}_n) \leq n$.

Proof. Denote the n -dimensional probability simplex by $\Delta^n = \{\alpha \in \mathbb{R}^n : \alpha_i \geq 0, i = 1, \dots, n, \sum_{i=1}^n \alpha_i = 1\}$. By the Myhill-Nerode extension for PFA (Theorem 2), all suffix distributions $S \in \text{suffix}(D)$ in Definition 5 reside in the convex hull of at most n distributions, which we denote by $\{D_1, \dots, D_n\}$. We thus have:

$$\sum_{w \in \Sigma^*} \max_{S \in \text{suffix}(D)} S(w) = \sum_{w \in \Sigma^*} \max_{\alpha \in \Delta^n} \sum_{i=1}^n \alpha_i D_i(w).$$

However, for each $w \in \Sigma^*$ there exists an index $i(w) \in \{1, \dots, n\}$ such that $D_{i(w)}(w) = \max_{\alpha \in \Delta^n} \sum_{i=1}^n \alpha_i D_i(w)$. Observing that $\{i(w) : w \in \Sigma^*\} \subseteq \{1, \dots, n\}$, we have:

$$\sum_{w \in \Sigma^*} \max_{\alpha \in \Delta^n} \sum_{i=1}^n \alpha_i D_i(w) = \sum_{w \in \Sigma^*} D_{i(w)}(w) \leq \sum_{i=1}^n 1 = n,$$

and the proof immediately follows. ■

We will also require the following technical lemma, for which a proof is supplied in Appendix B:

Lemma 4. Given any discrete distribution D represented by d_1, \dots, d_N with $d_i \geq 0$, $\sum_{i=1}^N d_i = 1$ and any sequence of numbers $S = (s_1, \dots, s_N)$ such that $0 \leq s_i \leq 1$, the following inequality holds:

$$\sum_{i=1}^N |1/N - d_i s_i| \geq 1 - \frac{1}{N} \sum_{i=1}^N s_i.$$

We are now in a position to prove the negative result mentioned earlier:

Proof. (**Theorem 3**) Define the target distribution D^* as follows:

$$D^*(w) = \begin{cases} |\Sigma|^{-L/2} & w = w'w', \quad w' \in \Sigma^{L/2} \\ 0 & \text{otherwise.} \end{cases}$$

Let \mathcal{D} be some family of distributions with $SM(\mathcal{D}) \leq |\Sigma|^{(\frac{L}{2}-1)}$, let $\widehat{D} \in \mathcal{D}$ be an approximating distribution and enumerate the set $\Sigma^{L/2}$ as w_1, \dots, w_N with

$N = |\Sigma|^{L/2}$. We proceed to lower-bound the L_1 -distance:

$$\begin{aligned} \|D^* - \widehat{D}\|_1 &= \sum_{w \in \Sigma^*} |D^*(w) - \widehat{D}(w)| \geq \sum_{w \in \Sigma^L} |D^*(w) - \widehat{D}(w)| \\ &= \sum_{i=1}^N |D^*(w_i w_i) - \widehat{D}(w_i w_i)| + \sum_{i=1}^N \sum_{\substack{w_j \in \Sigma^L \\ w_j \neq w_i}} |D^*(w_i w_j) - \widehat{D}(w_i w_j)| \\ &\geq \sum_{i=1}^N \left| \frac{1}{N} - \widehat{D}(w_i w_i) \right|. \end{aligned}$$

Plugging Lemma 3 above into Lemma 4, we have shown $\|D^* - \widehat{D}\|_1 \geq 1/2$. ■

The preceding analysis is loose in the sense that (asymptotically) a square factor in the number of states separates between the upper and the lower bounds of Lemma 1 and Theorem 3 respectively. A tighter analysis for PDFA models follows.

4.2 Lower bound for PDFA Approximation of Bounded Length Distributions

In this section we present a sharper lower bound when the approximating class is restricted to n -state PDFA models. We assume for simplicity that the alphabet's cardinality is 2, and therefore all logarithms used below are base 2. Before presenting the bound we quote the following result (Pisier, 1989) regarding the packing number on the probability simplex:

Lemma 5. *Let Δ^d denote the d -dimensional probability simplex. Then the number of distributions on Δ^d which are at least ε -separated in the L_1 -norm (denoted as $D(\varepsilon, \Delta^d, \|\cdot\|_1)$) is lower bounded by:*

$$D(\varepsilon, \Delta^d, \|\cdot\|_1) \geq \left(\frac{C_0}{\varepsilon} \right)^{d-1},$$

where C_0 is an absolute constant greater than 1.

Theorem 4. *Suppose $\Sigma = \{0, 1\}$. For any positive integer L , there exists a distribution D^* over $\Sigma^{L+\log(L+1)}$ such that for any distribution \widehat{D} induced from a PDFA with no more than $2^{(L-1)}$ states, $\|D^* - \widehat{D}\|_1 \geq 1/8$.*

Proof. We set all L -length prefixes of D^* to be equiprobable (i.e. of probability 2^{-L} each) and denote $N_1 = 2^L$. For $i \in \{1, \dots, N_1\}$, we denote the target suffix distribution following prefix i by D_i^* . For the same prefix, we denote the approximating distribution's prefix probability by \widehat{d}_i and its corresponding suffix distribution by \widehat{D}_i .

We construct D^* such that each suffix distribution D_i^* is composed solely of L_2 -length strings. We wish to construct a set of suffixes of cardinality N_1 ,

such that each pair is at least $1/2$ -separated (i.e. $\|D_{i_1}^* - D_{i_2}^*\|_1 \geq 1/2$, $i_1 \neq i_2$). Denoting $N_2 = 2^{L_2}$ (the number of possible length- L_2 strings), we appeal to Lemma 5. In this case, the dimensionality d of the simplex is N_2 . Setting $L_2 = \log(L+1)$ and using Lemma 5, we see that the number of possible distributions conforming to the demands is at least N_1 , as desired.

We now proceed to lower-bound $\|D^* - \widehat{D}\|_1$. Given that the approximating PDFA has (at most) $N_1/2$ states, there exist (at least) $N_1/2$ ‘‘coupled’’ pairs (i_1, i_2) such that $\widehat{D}_{i_1} = \widehat{D}_{i_2}$. Writing out $\|D^* - \widehat{D}\|_1$, we get:

$$\begin{aligned} \|D^* - \widehat{D}\|_1 &= \sum_{w \in \Sigma^*} |D^*(w) - \widehat{D}(w)| \geq \sum_{w \in \Sigma^{(L+L_2)}} |D^*(w) - \widehat{D}(w)| \\ &\geq \sum_{i=1}^{N_1} \left\| \frac{1}{N_1} D_i^* - \widehat{d}_i \widehat{D}_i \right\|_1. \end{aligned}$$

Using the standard norm inequality $\|x - y\| \geq \left| \|x\| - \|y\| \right|$, we find that each term in the sum is lower-bounded by $\left| \frac{1}{N_1} - \widehat{d}_i \right|$. For the coupled pair (i_1, i_2) , we therefore have:

$$\begin{aligned} t &:= \left\| \frac{1}{N_1} D_{i_1}^* - \widehat{d}_{i_1} \widehat{D}_{i_1} \right\|_1 + \left\| \frac{1}{N_1} D_{i_2}^* - \widehat{d}_{i_2} \widehat{D}_{i_2} \right\|_1 \\ &\geq \left| \frac{1}{N_1} - \widehat{d}_{i_1} \right| + \left| \frac{1}{N_1} - \widehat{d}_{i_2} \right| \geq \left| \widehat{d}_{i_1} - \widehat{d}_{i_2} \right|. \end{aligned}$$

However, as $\widehat{D}_{i_1} = \widehat{D}_{i_2}$, we also have that:

$$\begin{aligned} t &= \left\| \frac{1}{N_1} D_{i_1}^* - \widehat{d}_{i_1} \widehat{D}_{i_1} \right\|_1 + \left\| \frac{1}{N_1} D_{i_2}^* - \widehat{d}_{i_2} \widehat{D}_{i_2} \right\|_1 \\ &\geq \left\| \frac{1}{N_1} (D_{i_1}^* - D_{i_2}^*) - (\widehat{d}_{i_1} - \widehat{d}_{i_2}) \widehat{D}_{i_1} \right\|_1 \\ &\geq \left\| \frac{1}{N_1} (D_{i_1}^* - D_{i_2}^*) \right\|_1 - \left| \widehat{d}_{i_1} - \widehat{d}_{i_2} \right| \\ &\geq \frac{1}{2N_1} - \left| \widehat{d}_{i_1} - \widehat{d}_{i_2} \right|. \end{aligned}$$

Hence for all coupled pairs (i_1, i_2) , $t \geq \max(\frac{1}{2N_1} - b, b)$ where $b = |\widehat{d}_{i_1} - \widehat{d}_{i_2}|$. Since $1/(2N_1) - b > b$ for $b < 1/(4N_1)$, we have $t > 1/(4N_1)$ for any possible value of b . In other words, for all possible values of $(\widehat{d}_{i_1}, \widehat{d}_{i_2})$ we have:

$$\left\| \frac{1}{N_1} D_{i_1}^* - \widehat{d}_{i_1} \widehat{D}_{i_1} \right\|_1 + \left\| \frac{1}{N_1} D_{i_2}^* - \widehat{d}_{i_2} \widehat{D}_{i_2} \right\|_1 \geq \frac{1}{4N_1}.$$

Summing over (at least) $N_1/2$ coupled pairs, we obtain $\|D^* - \widehat{D}\|_1 \geq \frac{1}{8}$. \blacksquare

5 Discussion and Conclusion

The lower bounds presented in Sections 4.1 and 4.2 are conceptually loose in the following sense: we made only *partial* use of the Myhill-Nerode extension theorems. Namely, when constructing the PFA lower bound, we only used Theorem 2 indirectly via Lemma 3. The lemma does not utilize the fact that all PFA suffix distributions must *recursively* and exclusively contain suffix distributions which also conform to the conditions of the theorem. A proof technique utilizing this additional information could potentially provide a tighter result. In the PDFFA case, a similar criticism holds true, but the result obtained is tight to a logarithmic factor, leaving little room for improvement.

Our main goal for extending the research presented in this paper is to attain a complete understanding of the relationship between distributions induced by PFA and PDFFA families. Specifically, we seek to understand how well (and under which circumstances) PDFFA models can approximate distributions induced by PFA models. This problem was addressed in (Zeitouni et al., 1992), where assuming a certain condition (a lower bound on *all* the approximated PFA's transition probabilities), an L_∞ -approximation result was shown. In this result, however, the number of states required grows exponentially with the (inverse of the) accuracy parameter; the approximation is in the (weak) L_∞ sense, and the condition assumed may be unnecessarily strong. A more complete understanding would be theoretically desirable, and could potentially have practical implications.

Acknowledgements

We wish to thank Tim Sears for helpful discussions. National ICT Australia is funded by the Australian Government's Department of Communications, Information Technology and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Center of Excellence program. This work is also supported by the IST Program of the European Community, under the Pascal Network of Excellence, IST-2002-506778.

References

- L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.
- R. C. Carrasco and J. Oncina. Learning deterministic regular grammars from stochastic samples in polynomial time. *Theoret. Inform. and Appl.*, 33(1): 1–20, 1999.
- R. W. Chang and J. C. Hancock. On receiver structures for channels having memory. *IEEE Transactions on Information Theory*, IT-12:463–468, October 1966.

- A. Clark and F. Thollard. PAC-learnability of probabilistic deterministic finite state automata. *Journal of Machine Learning Research*, 5:473–497, 2004.
- T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.
- J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, Reading, Massachusetts, first edition, 1979.
- M. Kearns. Efficient noise-tolerant learning from statistical queries. In *Proc. 25th Annu. ACM Sympos. Theory Comput. (STOC)*, pages 392–401. ACM Press, New York, NY, 1993.
- M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie. On the learnability of discrete distributions. In *Proc. 26th Annu. ACM Sympos. Theory Comput. (STOC)*, pages 273–282, 1994.
- Gilles Pisier. *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge, 1989.
- E. Vidal, F. Thollard, C. de la Higuera, F. Casacuberta, and R. C. Carrasco. Probabilistic finite-state machines – Part I. *IEEE Trans. on Pattern analysis and Machine Intelligence*, 27(7):1013–1025, July 2005a.
- E. Vidal, F. Thollard, C. de la Higuera, F. Casacuberta, and R. C. Carrasco. Probabilistic finite-state machines – Part II. *IEEE Trans. on Pattern analysis and Machine Intelligence*, 27(7):1026–1039, July 2005b.
- Ofer Zeitouni, Jacob Ziv, and Neri Merhav. When is the generalized likelihood ratio test optimal? *IEEE Transactions on Information Theory*, 38(5):1597–1602, 1992.

A Proof of Lemma 2

We present a proof sketch, building on the lower bound of Section 4.1. Let the distribution D^* be as defined in Section 4.1 with L replaced by L_0 . We define D^{**} as follows:

$$D^{**}(w) = \begin{cases} 1 - 4\varepsilon & w = 0. \\ 4\varepsilon \cdot D^*(w') & w = 1w', \forall w' \in \Sigma^*. \end{cases}$$

It follows from the definition that $\mathbb{E}_{w \sim D^{**}} |w| = 1 - 4\varepsilon + 4\varepsilon(1 + L_0) = 1 + 4\varepsilon L_0$. Selecting $L_0 \leq \frac{L-1}{4\varepsilon}$ guarantees $\mathbb{E}_{w \sim D^{**}} |w| \leq L$. An ε -approximation of D^{**} can only be achieved if D^* is approximated to accuracy $1/2$, which by Lemma 3 cannot be achieved using a PFA of less than $|\Sigma|^{\binom{L_0}{2}-1}$ states. Substituting $L_0 = \frac{L-1}{4\varepsilon}$ concludes the proof. ■

B Proof of Lemma 4

We follow a two step approach. The first step is a reduction to an equivalent problem. The second step shows a bound on the equivalent problem, proving the inequality.

We start by partitioning the summands $d_1 s_1, \dots, d_N s_N$ into two mutually exclusive sets I_1, I_2 according to the following rule:

$$\begin{aligned} i \in I_1 & \quad \text{if } d_i s_i \leq \frac{1}{N}, \\ i \in I_2 & \quad \text{if } d_i s_i > \frac{1}{N}. \end{aligned}$$

Adding and subtracting $\sum_{i \in I_2} (1/N - d_i s_i)$, we get:

$$\sum_{i=1}^N |1/N - d_i s_i| = 1 - \sum_{i=1}^N d_i s_i + 2 \sum_{i \in I_2} \left(d_i s_i - \frac{1}{N} \right).$$

On the right hand side of the equality above, any term with $d_i s_i > 1/N$ is “doubly penalized” by the last sum, raising the suspicion that such terms will not participate in an optimal solution. Indeed, we now show that given a specific pair $\{D, S\}$ we can construct an alternative pair $\{\widehat{D}, \widehat{S}\}$ such that $\widehat{d}_i \widehat{s}_i \leq 1/N$ for all i , and $\sum_{i=1}^N |1/N - \widehat{d}_i \widehat{s}_i| \leq \sum_{i=1}^N |1/N - d_i s_i|$. The constructive proof is detailed in algorithm 1.

Algorithm 1: Reduction to a Uniformly Bounded Distribution \widehat{D}

Input: A distribution $D = \{d_1, \dots, d_N\}$ and a set $S = \{s_1, \dots, s_N\}$ with $0 \leq s_i \leq 1$.

Output: A distribution \widehat{D} and a set \widehat{S} with $0 \leq \widehat{s}_i \leq 1$, $\sum_{i=1}^N \widehat{s}_i \leq \sum_{i=1}^N s_i$, such that $\widehat{d}_i \widehat{s}_i \leq 1/N$ for all i , and $\sum_{i=1}^N |1/N - \widehat{d}_i \widehat{s}_i| \leq \sum_{i=1}^N |1/N - d_i s_i|$.

while $\exists i \in \{1, \dots, N\}$ such that $d_i s_i > 1/N$ **do**

Find $j \in \{1, \dots, N\}$ such that $d_j < 1/N$

if $d_i s_i - 1/N \geq 1/N - d_j$ **then**

Set $d_i = d_i - (1/N - d_j)$

Set $d_j = 1/N$

else

Set $d_i = 1/N$

Set $d_j = d_j - (1/N - d_i s_i)$

end

end

Set $\widehat{D} = D, \widehat{S} = S$.

To prove the algorithm’s correctness we note that no operation performed over the algorithm’s run will increase $\sum_{i=1}^N |1/N - d_i s_i|$, and that the algorithm terminates after a finite number of steps.

If at any stage of the while loop of line 1 a suitable i is found, the existence of a suitable j in line 1 is assured ($d_i s_i > 1/N \Rightarrow d_i > 1/N \Rightarrow \exists d_j < 1/N$). The number of steps performed is bounded by the number of pairs (i, j) and is therefore finite. In all reassigment operations, the values of both $|1/N - d_i s_i|$

and $|1/N - d_j s_j|$ are not increased, assuring that upon termination we have $\sum_{i=1}^N |1/N - \widehat{d}_i \widehat{s}_i| \leq \sum_{i=1}^N |1/N - d_i s_i|$. We have thus shown that:

$$\min_{\{D,S\}} \sum_{i=1}^N |1/N - d_i s_i| = \min_{\{D,S\}} \sum_{i=1}^N (1/N - d_i s_i) \quad \text{s.t.} \quad d_i s_i \leq 1/N \quad \forall i.$$

We now examine how the selections of d_1 and s_1 affect both $(1/N - d_1 s_1)$ and $\sum_{i=2}^N (1/N - d_i s_i)$, under the condition $d_i s_i \leq 1/N$. If $d_1 = 1/N - \varepsilon$, we immediately have $1/N - d_1 s_1 \geq \varepsilon$, due to $s_1 \leq 1$. If $d_1 = 1/N + \varepsilon$ we have:

$$\sum_{i=2}^N (1/N - d_i s_i) \geq \sum_{i=2}^N (1/N - d_i) \geq \sum_{i=2}^N 1/N - [(N-1)/N - \varepsilon] = \varepsilon.$$

Writing s_1 as $1 - \delta_1$, $0 \leq \delta_1 \leq 1$, we again examine the effect of the choice on $IP_1 = (1/N - d_1 s_1)$ (denoting ‘‘immediate penalty’’) and on $SP_1 = \sum_{i=2}^N (1/N - d_i s_i)$ (denoting ‘‘subsequent penalty’’):

$$\begin{aligned} IP_1 &= 1/N - d_1 s_1 = 1/N - d_1(1 - \delta_1), \\ SP_1 &= \sum_{i=2}^N (1/N - d_i s_i) \geq \max\{d_1 - 1/N, 0\}. \end{aligned}$$

We can now deduce $IP_1 + SP_1 \geq \delta_1/N$ by considering the following three cases:

$$\begin{aligned} (i) \quad d_1 < 1/N & \quad IP_1 > \delta_1/N. \\ (ii) \quad 1/N \leq d_1 \leq \frac{1}{(1-\delta_1)N} & \quad SP_1 \geq d_1 - 1/N \\ & \Rightarrow IP_1 + SP_1 \geq 1/N - d_1 s_1 + d_1 - 1/N \\ & \quad = d_1(1 - s_1) \geq \delta_1/N. \\ (iii) \quad d_1 > \frac{1}{(1-\delta_1)N} & \quad \text{Violates the constraint } s_1 d_1 \leq 1/N. \end{aligned}$$

Summing over $i = 1, \dots, N$ we get:

$$\sum_{i=1}^N (1/N - d_i s_i) = \sum_{i=1}^N IP_i + SP_i \geq \frac{1}{N} \sum_{i=1}^N \delta_i = \frac{1}{N} \sum_{i=1}^N (1 - s_i) = 1 - \frac{1}{N} \sum_{i=1}^N s_i,$$

proving the lemma. \blacksquare