
Using PCA for Probabilistic Grammatical Inference on Trees*

Raphaël Bailly

Laboratoire d'Informatique Fondamentale de Marseille
CNRS, Aix-Marseille Université
raphael.bailly@lif.univ-mrs.fr

François Denis

Laboratoire d'Informatique Fondamentale de Marseille
CNRS, Aix-Marseille Université
francois.denis@lif.univ-mrs.fr

Édouard Gilbert

Laboratoire d'Informatique Fondamentale de Lille
INRIA Lille-Nord Europe
edouard.gilbert@inria.fr

Amaury Habrard

Laboratoire d'Informatique Fondamentale de Marseille
CNRS, Aix-Marseille Université
amaury.habrard@lif.univ-mrs.fr

Abstract

We focus on the classical problem in grammatical inference of learning stochastic tree languages from finite samples of trees independently drawn according to a fixed unknown distribution. We consider here the class of stochastic tree languages that can be computed by rational tree series which can be viewed as a strict generalization of probabilistic tree automata. The class of rational stochastic tree languages has an algebraic characterization: All the residuals of a stochastic languages lie in a finite vector subspace. We propose a principle based on Principal Components Analysis to identify this vector subspace. This approach allows us to define a global solution of the problem instead of building an automaton iteratively as done by standard probabilistic grammatical inference algorithm. This is a way to tackle the main drawback of these approaches that is using statistical tests that rely on less and less examples when the structure grows. We provide an algorithm that computes an estimate of the target vector subspace and build a linear representation of a tree series giving an estimation of the target distribution. We notably show that in the case of tree languages, we have to consider the dual vector subspace to build the representation.

1 Introduction

In this article, we focus on the problem of learning probability distributions over trees. Given a sample independently drawn according to a fixed but unknown stochastic tree language p , a classical problem in grammatical inference is to infer an estimate of p in some class of probabilistic models that can be finitely represented [4]. A usual class of representation of stochastic tree language is the class of probabilistic tree automata (PTA) where the parameters stand in $[0, 1]$. This class is the equivalent class of probabilistic automata or hidden Markov models for stochastic strings languages.

Recent approaches have proposed to represent stochastic tree language in a larger class of representation: The class of rational stochastic tree languages (RSTL) that can be represented under a linear form of a tree series [2]. The models of this class can be equivalently represented by weighted tree automata with parameters in \mathbb{R} (hence with weights that can be negative and without any per state

*This work was partially supported by the ANR LAMPADA and SEQUOIA projects and by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.

normalisation condition). This class has two interesting properties: It has a high level of expressiveness since it strictly includes the class of PTA and it admits a canonical form with a minimal number of parameters (see [7] for an illustration in the string case). Based on these properties, linear representations of RSTL are a good candidate from a grammatical inference standpoint. It has notably the characterization that the residuals of the stochastic language lie in a finite dimensional subspace. The goal of an inference algorithm is then to identify this subspace. This was illustrated by the algorithm DEES [9, 8]. This algorithm builds iteratively a linear representation of a tree series by finding a basis of the vector subspace spanned by the stochastic series definable over p . This dimension of this vector subspace is finite and its dimension coincides with the minimal number of states needed by a weighted automaton to compute p . Hence, a first step of a grammatical inference process is to identify this vector subspace. However, this iterative approach suffers from the drawback that the method relies on statistical tests that are done on less and less examples when the structure grows.

In this paper, we study the possibility of using Principal Component Analysis (PCA) techniques to build a rational tree series representing a rational stochastic tree language from a tree sample. PCA has already been used in grammatical inference for another framework in [5] and in [1] for learning rational stochastic string languages. We propose in this paper an algorithm for identifying RSTL. We notably show that the construction of the rational tree series should rely on a projection of the dual vector subspace.

The paper is organised as follows. Section 2 gives the preliminaries on trees and rational tree series. Section 3 is devoted to our algorithm, while the convergence properties are presented in Section 4.

2 Preliminaries

In this section, we introduce the objects that will be used all along in the paper. We mainly follow notations and definitions from [6] about trees. Formal power tree series have been introduced in [2] where the main results appear.

2.1 Trees and contexts

Let $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1 \cup \dots \cup \mathcal{F}_p$ be a ranked alphabet where the elements in \mathcal{F}_m are the function symbols of rank m . We make the assumption that $\mathcal{F}_0 \neq \emptyset$. The set of *trees* over a ranked alphabet \mathcal{F} is the smallest set $T_{\mathcal{F}}$ satisfying $\mathcal{F}_0 \subseteq T_{\mathcal{F}}$, for $f \in \mathcal{F}_m$, $m \geq 1$, and $t_1, \dots, t_m \in T_{\mathcal{F}}$, $f(t_1, \dots, t_m) \in T_{\mathcal{F}}$.

Contexts are element c of $C_{\mathcal{F}} \subset T_{\mathcal{F} \cup \{\$\}}$ where $\$$ is variable that appears exactly once as a leaf in c ($\$$ has a rank 0 and $\$ \notin \mathcal{F}$). Given a context $c \in C_{\mathcal{F}}$ and a tree $t \in T_{\mathcal{F}}$, one can build a tree $c[t] \in T_{\mathcal{F}}$ by replacing the (unique) occurrence of $\$$ in c by the tree t .

Example 1. Let $\mathcal{F}_0 = \{a, b\}$, $\mathcal{F}_1 = \{g(\cdot)\}$ and $\mathcal{F}_2 = \{f(\cdot, \cdot)\}$. Then $t = f(a, g(b)) \in T_{\mathcal{F}}$ (Figure 1(a)), $c = f(a, \$) \in C_{\mathcal{F}}$ (Figure 1(b)) and $c[t] = f(f(a, g(b)), a)$ (Figure 1(c)).

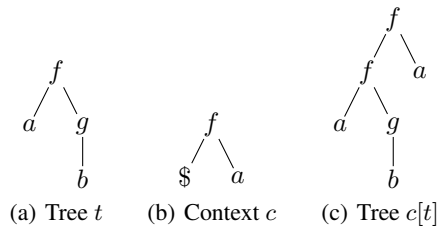


Figure 1: An example of tree $t = f(a, g(b))$, context $c = f(\$, a)$ and their composition $c[t] = f(f(a, g(b)), a)$, as defined in Example 1.

2.2 Tree series

A (*formal power*) *tree series* on $T(\mathcal{F})$ is a mapping $r : T(\mathcal{F}) \rightarrow \mathbb{R}$. The vector space of all tree series on $T(\mathcal{F})$ is denoted by $\mathbb{R}\langle\langle T_{\mathcal{F}} \rangle\rangle$.

Let V be a finite dimensional vector space over \mathbb{R} . We denote by $\mathcal{L}(V^m; V)$ the set of m -linear mappings from V^m to V . Let $\mathcal{L} = \cup_{m \geq 0} \mathcal{L}(V^m; V)$. We denote by V^* the dual space of V , i.e. the vector space composed of all the linear forms defined on V .

A *linear representation* of $T(\mathcal{F})$ is a couple (V, μ) , where V is a finite dimensional vector space over \mathbb{R} , and where $\mu : \mathcal{F} \rightarrow \mathcal{L}$ maps \mathcal{F}_m into $\mathcal{L}(V^m; V)$ for each $m \geq 0$. Thus for each $f \in \mathcal{F}_m$, $\mu(f) : V^m \rightarrow V$ is m -linear. Function μ extends uniquely to a morphism $\mu : T(\mathcal{F}) \rightarrow V$ by: $\mu(f(t_1, \dots, t_m)) = \mu(f)(\mu(t_1), \dots, \mu(t_m))$. Let $V_{T(\mathcal{F})}$ be the vector subspace of V spanned by $\mu(T(\mathcal{F}))$: $(V_{T(\mathcal{F})}, \mu)$ is a linear representation of $T(\mathcal{F})$.

Let r be a tree series over $T(\mathcal{F})$, r is said to be *recognizable* if there exists a triple (V, μ, τ) , where (V, μ) is a linear representation of $T(\mathcal{F})$, and $\lambda : V \rightarrow \mathbb{R}$ is a linear form, such that $r(t) = \tau(\mu(t))$ for all $t \in T(\mathcal{F})$. The triple (V, μ, τ) is called a *linear representation for r* . In [2], it has been shown that the notions of recognizable tree series and rational tree series coincide. From now on, we shall refer to them by using the term of *rational tree series*. Rational tree series can be equivalently represented by weighted tree automata [3].

Definition 1. A *stochastic tree language* over $T(\mathcal{F})$ is a tree series $r \in \mathbb{R}\langle\langle T(\mathcal{F}) \rangle\rangle$ such that for any $t \in T(\mathcal{F})$, $0 \leq r(t) \leq 1$ and $\sum_{t \in T(\mathcal{F})} r(t) = 1$. The sum is well defined since r is non negative. The set of stochastic tree languages is denoted by $\mathcal{S}(T(\mathcal{F}))$.

A *rational stochastic tree language* (RSTL) is a stochastic tree language which admits a linear representation. The set of rational stochastic tree languages is denoted by $\mathcal{S}^{\text{rat}}\langle\langle T(\mathcal{F}) \rangle\rangle$. Note that it can be shown that there exists some rational stochastic tree language that cannot be computed by any probabilistic tree automaton (see [7] for an example in the case of stochastic word languages).

2.3 Canonical form

We now define the canonical representation of a rational tree series [9].

Let $c \in C_{\mathcal{F}}$. We define the linear mapping $\dot{c} : \mathbb{R}\langle\langle T(\mathcal{F}) \rangle\rangle \rightarrow \mathbb{R}\langle\langle T(\mathcal{F}) \rangle\rangle$ by

$$\dot{c}(r)(t) = r(c[t]).$$

Let $r \in \mathbb{R}\langle\langle T(\mathcal{F}) \rangle\rangle$. Let us denote by W_r the vector subspace of $\mathbb{R}\langle\langle T(\mathcal{F}) \rangle\rangle$ spanned by $\{\dot{c}r | c \in C_{\mathcal{F}}\}$. The elements of $\{\dot{c}r | c \in C_{\mathcal{F}}\}$ are called the residuals of r . It can be shown that r is rational if and only if the dimension of W_r is finite [9]. Let W_r^* be the dual space of W_r , i.e. the set of all linear forms on W_r . For any $t \in T(\mathcal{F})$, let $\bar{t} \in W_r^*$ be defined by: $\forall s \in W_r$, $\bar{t}(s) = s(t)$. It can be shown that there exist trees t_1, \dots, t_n such that $(\bar{t}_1, \dots, \bar{t}_n)$ forms a basis of W_r^* . Let us define the linear representation (W_r^*, μ, τ) as follows:

- for any $f \in \mathcal{F}_m$, define $\mu(f)(\bar{t}_{i_1}, \dots, \bar{t}_{i_m}) = \overline{f(t_{i_1}, \dots, t_{i_m})}$.
- $\tau \in (W_r^*)^* = W_r$ by $\tau(\bar{t}) = r(t)$.

Theorem 1. [9] (W_r^*, μ, τ) is a linear representation of r which is called the canonical linear representation of r . It can be embedded into any linear representation of r ; in particular, its dimension is minimal.

Example 2. Let $\mathcal{F} = \{a, f(\cdot, \cdot)\}$ a ranked alphabet, $|t|_f$ denotes the number of nodes f in the tree t . Consider the rational stochastic tree language $p = 2p_1 - p_2$ such that $p_1(t) = \frac{2^{|t|_f+1}}{3^{2|t|_f+1}}$, $p_2(t) = \frac{3^{|t|_f+1}}{4^{2|t|_f+1}}$ then $p(t) = \frac{2^{5|t|_f+4} - 3^{3|t|_f+2}}{3^{2|t|_f+1} \times 4^{2|t|_f+1}}$. It can be shown that p_1 , p_2 and p define stochastic languages and that p has a 2-dimensional linear representation [8]. Now, for any context c and any tree t :

$$\bar{t}(\dot{c}p) = p(c[t]) = \frac{2^{5|t|_f+5|c|_f+4} - 3^{3|t|_f+3|c|_f+2}}{3^{2|t|_f+2|c|_f+1} \times 4^{2|t|_f+2|c|_f+1}}.$$

Since p has a 2-dimensional linear representation, the dimension of W_r^* is ≤ 2 . Let $c_0 = \$$ and $c_1 = f(a, \$)$, we have:

$$\bar{a}(\dot{c}_0 p) = \frac{7}{3 \times 2^2}, \bar{a}(\dot{c}_1 p) = \overline{f(a, a)}(c_0) = \frac{269}{3^3 \times 2^6}, \text{ and } \overline{f(a, a)}(\dot{c}_1 p) = \frac{9823}{3^5 \times 2^{10}}.$$

Since $\bar{a}(\dot{c}_0 p) \times \overline{f(a, a)}(\dot{c}_1 p) \neq \bar{a}(\dot{c}_1 p) \times \overline{f(a, a)}(\dot{c}_0 p)$, \bar{a} and $\overline{f(a, a)}$ are linearly independent. Then, $(\bar{a}, \overline{f(a, a)})$ is a basis of W_r^* . We define τ and μ by:

$$\begin{aligned} \tau(\bar{a}) &= p(a) = \frac{7}{3 \times 2^2} \text{ and } \tau(\overline{f(a, a)}) = p(f(a, a)) = \frac{269}{3^3 \times 2^6} \text{ and} \\ \mu(a) &= \bar{a}, \mu(f)(\bar{a}, \bar{a}) = \overline{f(a, a)}, \\ \mu(f)(\bar{a}, \overline{f(a, a)}) &= \mu(f)(\overline{f(a, a)}, \bar{a}) = \frac{-54}{2^4 \times 3^4} \bar{a} + \frac{59}{2^4 \times 3^2} \overline{f(a, a)}, \\ \mu(f)(\overline{f(a, a)}, \overline{f(a, a)}) &= \frac{-3186}{2^8 \times 3^6} \bar{a} + \frac{2617}{2^8 \times 3^4} \overline{f(a, a)}. \end{aligned}$$

Note, it is possible to change this representation in order to ensure that each element of the basis in this new representation defines a stochastic language and then use this representation as a generative model [8].

We consider in this paper the Hilbert space $\ell_2(T_{\mathcal{F}})$ composed of the rational tree series $r \in \mathbb{R}\langle\langle T_{\mathcal{F}} \rangle\rangle$ such that $\sum_{t \in T_{\mathcal{F}}} r(t)^2 < \infty$ and where the dot product is defined by $(r, s) = \sum_{t \in T_{\mathcal{F}}} r(t)s(t)$. Hence, $\|r\| = (\sum_{t \in T_{\mathcal{F}}} r(t)^2)^{1/2}$. Clearly, $S^{\text{rat}}\langle\langle T_{\mathcal{F}} \rangle\rangle \in \ell^2(T_{\mathcal{F}})$ and for any $r \in \ell_2(T_{\mathcal{F}})$, $\dot{c}r \in \ell_2(T_{\mathcal{F}})$.

3 Principle of the Algorithm

Let $p \in S^{\text{rat}}\langle\langle T_{\mathcal{F}} \rangle\rangle$ be a rational stochastic tree language and let $C = \{c_1, \dots, c_m\}$ be a finite set of contexts such that the empty context $\$ \in C$. We consider the finite dimensional subspace W of $\ell_2(T_{\mathcal{F}})$ spanned by the set $\{\dot{c}p \mid c \in C\}$, i.e. the residuals of p restricted to the contexts defined by the set C . W^* denotes the set of all linear forms over W . We define the following dot product in W^* by: $(f, g) = \sum_{c \in C} f(\dot{c}p)g(\dot{c}p)$. We denote by $\|\cdot\|_C$ the norm induced by this dot product, i.e. $\|f\|_C = (\sum_{c \in C} (f(\dot{c}p))^2)^{1/2}$. Π_{W^*} denotes the orthogonal projection over W^* .

Let S be a sample of trees independently and identically drawn according to p and let p_S be the empirical distribution defined from S . W_S denotes the vector subspace of $\ell_2(T_{\mathcal{F}})$ spanned by $\{\dot{c}p_S \mid c \in C\}$. We first build from S an estimate \hat{W}_S^* of W^* and then we show that \hat{W}_S^* can be used to build a linear representation such that its associated rational series approximate the target p . In this section, we implicitly suppose that the dimension d of \hat{W}_S^* is known. We will show in the next section how it may be estimated from the data. In the following, $Subtrees(S)$ denotes the set of subtrees that can be extracted from S .

3.1 Estimating the Target Space

Given the set of contexts C , let us define for any $t \in T_{\mathcal{F}}$, $\bar{t}_{p_S} \in W_S^* : W_S \mapsto \mathbb{R}$ such that $\bar{t}_{p_S} = \bar{t}|_{W_S}$ i.e. the restriction of $\bar{t} \in \ell_2(T_{\mathcal{F}})^*$ to W_S : $\bar{t}_{p_S}(\dot{c}p_S) = \bar{t}(\dot{c}p_S) = p_S(c[t])$.

Let $k \geq 0$ be an integer. The first step consists of finding the k -dimensional vector subspace of W_S^* : $W_{S,k}^*$ which minimizes the distance to $\{\bar{t}_{p_S} \mid t \in T_{\mathcal{F}}\}$:

$$W_{S,k}^* = \underset{\dim(V^*)=k, V^* \subseteq W_S^*}{\text{Argmin}} \sum_{t \in T_{\mathcal{F}}} \|\bar{t}_{p_S} - \Pi_{V^*} \bar{t}_{p_S}\|_C^2.$$

Since the support of p_S is finite, $W_{S,k}^*$ can be computed using PCA. Let t_1, \dots, t_n be an enumeration of $Subtrees(S)$. We define for every $c \in C$:

$$mean_S(c) = \frac{1}{|Subtrees(S)|} \sum_{t \in Subtrees(S)} \bar{t}_{p_S}(\dot{c}p_S) = \frac{1}{|Subtrees(S)|} \sum_{t \in Subtrees(S)} p_S(c[t]).$$

Let x_t be the vector of $\mathbb{R}^{|C|}$ defined by:

$$x_t[j] = \bar{t}_{p_S}(\dot{c}_j p_S) - \text{mean}_S(c_j) = p_S(c_j[t]) - \text{mean}_S(c_j).$$

Let X_S the matrix containing the vectors x_t as rows. Intuitively, this matrix is defined according the \bar{t}_{p_S} operators of the dual space on its rows and the series \dot{c}_p on its columns.

The matrix $M_S = X_S X_S^T$, with X_S^T the transpose of X_S , is positive semi-definite and the eigenvectors (w_1, \dots, w_k) corresponding to the k largest (positive) eigenvalues form an orthogonal basis of $W_{S,k}^*$.

3.2 Building the Linear Representation From the Dual Space

Now, we show how to build a linear representation $(W_{S,k}^*, \mu_k, \tau)$ from $W_{S,k}^*$.

First, note that for any $a \in \mathcal{F}_0$, $\mu_k(a)$ is simply defined by the orthogonal projection of \bar{a} over $W_{S,k}^*$.

Now, we have to define the other μ_k operators. First, we fix a basis of W_S^* , let $B_S^* = \{\bar{t}_1, \dots, \bar{t}_l\}$ this basis. For any $f \in \mathcal{F}_m$, $m > 0$, and $\bar{t}_{i_1}, \dots, \bar{t}_{i_m}$ elements of B_S^* , then we define the multi-linear operator μ such that:

$$\mu(f)(\bar{t}_{i_1}, \dots, \bar{t}_{i_m}) = \overline{f(t_{i_1}, \dots, t_{i_m})}.$$

This definition determines a unique multi-linear form over W_S^* . Now, recall that $W_{S,k}^* \subseteq W_S^*$, we define $\mu_k(f)$ over $W_{S,k}^*$ such that for every $v_1, \dots, v_m \in W_{S,k}^*$:

$$\mu_k(f)(v_1, \dots, v_m) = \Pi_{W_{S,k}^*} \mu(f)(v_1, \dots, v_m).$$

The linear representation $(W_{S,k}^*, \mu_k, \tau)$ is then built such that:

- For $a \in \mathcal{F}_0$, $\mu_k(a) = \Pi_{W_{S,k}^*} \bar{a}$.
- For $f \in \mathcal{F}_m$, $m > 0$, for any $(v_1, \dots, v_m) \in (W_{S,k}^*)^m$:
 $\mu_k(f)(v_1, \dots, v_m) = \Pi_{W_{S,k}^*} \mu(f)(v_1, \dots, v_m)$
- τ is defined such that $\tau(v_i) = v_i(\dot{\$}p_S) = v_i(p_S)$ for any $v_i \in W_{S,k}^*$.

The different steps of the algorithm are described in Algorithm 1.

Data: A sample $S = \{t_i \in T_{\mathcal{F}}, 1 \leq i \leq |S|\}$ i.i.d. according to a distribution p , a dimension d and a finite set of contexts C .

Result: A linear representation A of a tree series $(W_{S,d}^*, \mu_d, \tau)$

$T = \text{Subtrees}(S)$;
Let X a $|T| \times |C|$ matrix; $X[i, j] = p_S(c_j[t_i]) - \frac{1}{|T|} \sum_{t \in \text{Subtrees}(S)} p_S(c_j[t])$;
 $M = X X^T$ /* variance-covariance matrix */;
 $(\lambda_i, w_i) \leftarrow$ eigenvalues of M in decreasing order and corresponding eigenvectors;
Let $B_{S,d}^* = \{w_1, \dots, w_d\}$ /* $\Pi_{B_{S,d}^*}$ orthogonal projection on the vector subspace $W_{S,d}^*$ spanned by $B_{S,d}^*$ */;
Let μ an operator defined over W_S^* and $v_1, \dots, v_m \in W_{S,d}^* \subseteq W_S^*$;
foreach $f \in \mathcal{F}$ **do**
 if $f \in \mathcal{F}_0$ **then** $\mu_d(f) = \Pi_{B_{S,d}^*} \bar{f}$;
 if $f \in \mathcal{F}_m$, $m > 0$ **then** $\mu_d(f)(v_1, \dots, v_m) = \Pi_{B_{S,d}^*} \mu(f)(v_1, \dots, v_m)$;
end
 $\tau(v_i) = v_i(\dot{\$}p_S) = v_i(p_S)$ for any $v_i \in W_{S,d}^*$;
return $A = (W_{S,d}^*, \mu_d, \tau)$;

Algorithm 1: Building a linear representation corresponding to a sample S and a dimension d .

4 Consistency

4.1 Consistency when the rank of target is known

We show that the solution computed by the algorithm converges to the target as the size of the sample S increases.

Proposition 1. *Let S be a sample i.i.d. according to a rational stochastic language p with rank d , and p_S the empirical distribution deduced from S . Let W the vector space spanned by $\{\dot{c}p \mid c \in C\}$ with $C \subset C_{\mathcal{F}}$, $|C| = m < \infty$, W^* being the linear forms over W and Π_{W^*} is the orthogonal projection on W^* . Every \bar{t}_p is a mapping from W to \mathbb{R} such that for any $c \in C$, $\bar{t}_p(\dot{c}p) = p(c|t)$.*

W_S is defined in a symmetrical way as the vector space spanned by $\{\dot{c}p_S \mid c \in C\}$.

Let $W_{S,d}^*$ be the d -dimension dual vector subspace which minimizes $\sum_{t \in T_{\mathcal{F}}} \|\bar{t}_{p_S} - \Pi_{W_{S,d}^*}(\bar{t}_{p_S})\|_C^2$. Then $\mathbb{E}[\|\Pi_{W_{S,d}^*}(\bar{t}_{p_S}) - \bar{t}_p\|_C] \rightarrow 0$ uniformly wrt t as the size of S increases.

Proof. Let $t \in T_{\mathcal{F}}$. As $p_S(t)$ is a binomial distribution, one can compute the variance $\mathbb{V}[p_S(t)] = \mathbb{E}[(p_S(t) - p(t))^2] = \frac{p(t)(1-p(t))}{|S|} \leq \frac{p(t)}{|S|}$. Thus,

$$\mathbb{E}\left[\sum_t \|\bar{t}_{p_S} - \bar{t}_p\|_C^2\right] \leq \sum_{c \in C} \frac{p(c|T_{\mathcal{F}})}{|S|} \quad (1)$$

and tends to 0 as the size of S increases. In particular,

$$\mathbb{E}[\|\bar{t}_{p_S} - \bar{t}_p\|_C] \leq \sqrt{\frac{m}{|S|}} \quad (2)$$

and converges uniformly towards 0 as $|S|$ increases.

Note that one also has:

$$\sum_{c \in C} \mathbb{E}[\|\dot{c}p_S - \dot{c}p\|^2] \leq \frac{m}{|S|} \quad (3)$$

$$\forall c \in C, \mathbb{E}[\|\dot{c}p_S - \dot{c}p\|] \leq \sqrt{\frac{m}{|S|}} \quad (4)$$

Now, $\mathbb{E}[\sum_t \|\bar{t}_{p_S} - \Pi_{W^*}(\bar{t}_{p_S})\|_C^2] \leq \mathbb{E}[\sum_t \|\bar{t}_{p_S} - \bar{t}_p\|_C^2]$ as $\bar{t}_p \in W^*$. Then

$$\mathbb{E}\left[\sum_t \|\bar{t}_{p_S} - \Pi_{W_{S,d}^*}(\bar{t}_{p_S})\|_C^2\right] \leq \frac{m}{|S|} \quad (5)$$

as $|S| \rightarrow \infty$ by minimality of $W_{S,d}^*$, and we obtain that $\mathbb{E}[\|\bar{t}_{p_S} - \Pi_{W_{S,d}^*}(\bar{t}_{p_S})\|_C] \leq \sqrt{\frac{m}{|S|}}$ and converges uniformly towards 0 as $|S|$ increases.

This implies that $\mathbb{E}[\|\Pi_{W_{S,d}^*}(\bar{t}_{p_S}) - \bar{t}_p\|_C] \leq \mathbb{E}[\|\bar{t}_{p_S} - \Pi_{W_{S,d}^*}(\bar{t}_{p_S})\|_C] + \mathbb{E}[\|\bar{t}_{p_S} - \bar{t}_p\|_C] \leq 2\sqrt{\frac{m}{|S|}}$ and converges uniformly towards 0 as $|S|$ increases. \square

We can deduce from this proposition that if $\{w_1, \dots, w_d\}$ forms a basis of W^* , if for any $w \in W^*$ $w = \sum_{i=1}^d \alpha_i w_i$ and if $\Pi_{W_{S,d}^*} w = \sum_{i=1}^d \hat{\alpha}_i \Pi_{W_{S,d}^*} w_i$, then $\mathbb{E}[\text{Max}(\|\alpha_i - \hat{\alpha}_i\|)] = O(|S|^{-1/2})$.

4.2 Convergence of eigenvalues

In this section we prove some results about the eigenvalues provided by the PCA. We will consider the PCA on the columns of the matrix M rather than on the lines as done in Algorithm 1 since both of two approaches provide the same eigenvalues. As before, a finite set C of contexts is considered.

We will use two results concerning eigenvalues:

Lemma 1. [10] Let M be the variance-covariance matrix of the set $\{\dot{c}p, c \in C\}$, with $|C| = m$. Let $\{\lambda_i, 1 \leq i \leq m\}$ be the set of eigenvalues of M . Let W_d be the d -dimension vector subspace which minimizes $\sum_{c \in C} \|\dot{c}p - \Pi_{W_d}(\dot{c}p)\|^2$. Then

- $\forall d$ s.t. $1 \leq d \leq m$,
 $\lambda_d = \max_{\dim(W)=d} \min_{w \in W \setminus \{0\}} \sum_{c \in C} \|\Pi_w(\dot{c}p)\|^2$.
- $\sum_{d+1 \leq i \leq m} \lambda_i = \sum_{c \in C} \|\dot{c}p - \Pi_{W_d}(\dot{c}p)\|^2$.

Proposition 2. Let S be a sample i.i.d. according to a rational stochastic language p with rank d , and p_S the empirical distribution deduced from S . Let Π_W be the orthogonal projection on W . Let $W_{S,d}$ be the d -dimension vector subspace which minimizes $\sum_{c \in C} \|\dot{c}p_S - \Pi_{W_{S,d}}(\dot{c}p_S)\|^2$. Let $\{\lambda_i, 1 \leq i \leq m\}$ be the set of eigenvalues of the variance-covariance matrix of $\{\dot{c}p_S, c \in C\}$, with $m = |C|$.

Then $\mathbb{E}[\sum_{d+1 \leq i \leq m} \lambda_i]$ tends to 0 as the size of S increases.

Proof. $\mathbb{E}[\sum_{d+1 \leq i \leq |C|} \lambda_i] = \sum_{c \in C} \|\dot{c}p_S - \Pi_{W_{S,d}}(\dot{c}p_S)\|^2 \leq \frac{m}{|S|}$. This tends to 0 as $|S|$ tends to infinity. □

Proposition 3. Let S be a infinite sample i.i.d. according to a rational stochastic language p with rank d , and S_n the n first elements. Let p_S the empirical distribution deduced from S , and p_{S_n} the one deduced from S_n . Let

$$\lambda_k = \max_{\dim(W)=k} \min_{w \in W \setminus \{0\}} \sum_{c \in C_F} \|\Pi_w(\dot{c}p)\|^2$$

$$\lambda_{k,n} = \mathbb{E}[\max_{\dim(W_{S,k})=k} \min_{w \in W_{S,k} \setminus \{0\}} \sum_{c \in C_F} \|\Pi_w(\dot{c}p_{S_n})\|^2].$$

Then $\lim_{n \rightarrow \infty} \lambda_{k,n} = \lambda_k$.

Proof. Then, $\|\Pi_w(\dot{c}p)\|^2 - \|\Pi_w(\dot{c}p_{S_n})\|^2 = (\|\Pi_w(\dot{c}p)\| - \|\Pi_w(\dot{c}p_{S_n})\|)(\|\Pi_w(\dot{c}p)\| + \|\Pi_w(\dot{c}p_{S_n})\|)$ with $(\|\Pi_w(\dot{c}p)\| - \|\Pi_w(\dot{c}p_{S_n})\|) \leq (\|\Pi_w(\dot{c}p) - \Pi_w(\dot{c}p_{S_n})\|) \leq \|\dot{c}p - \dot{c}p_{S_n}\|$ because Π_w is 1-lipschitz. For the same reason, $\|\Pi_w(\dot{c}p)\| + \|\Pi_w(\dot{c}p_{S_n})\| \leq \|\dot{c}p\| + \|\dot{c}p_{S_n}\| \leq 2$.

Thus, for all $\epsilon > 0$ there exists N_ϵ such that $\forall c \in C, \forall w \in \ell_2(T_{\mathcal{F}}), \forall n > N_\epsilon, \mathbb{E}[\|\Pi_w(\dot{c}p)\|^2 - \|\Pi_w(\dot{c}p_{S_n})\|^2] \leq 2\mathbb{E}[\|\dot{c}p - \dot{c}p_{S_n}\|] \leq \frac{\epsilon}{|C|}$.

One then has:

$$\lambda_k - \epsilon \geq \lambda_{k,N_\epsilon} \geq \lambda_k - \epsilon$$

and we get the conclusion. □

From this proposition, it follows that in the limit, for any target space of dimension d , there will be a gap between the d^{th} and the $(d+1)^{th}$ eigenvalues. Thus, the correct dimension can be assessed by an appropriate statistical test (see [1] for an example of a heuristic test).

5 Conclusion and Discussion

We have studied the problem of learning stochastic tree languages from finite samples drawn i.i.d. from an unknown distribution p and proposed a new approach for identifying rational stochastic tree languages, i.e. stochastic languages that can be represented by rational tree series. Indeed, most classical inference algorithms in probabilistic grammatical inference build an automaton or a grammar iteratively from a sample S ; starting from an automaton composed of only one state, then they have to decide whether a new state must be added to the structure. This iterative decision relies on a statistical test with a known drawback: as the structure grows, the test relies on less and less examples. Instead of this iterative approach, we tackle the problem globally and our algorithm computes

in one step the space needed to build the output automaton. That is, we have reduced the problem set in the classical probabilistic grammatical inference framework into a classical optimization problem. We now need to experimentally study and compare our approach to existing ones on real data: this is a work in progress. A further consequence of our approach is that it will be possible to introduce non linearity via the kernel PCA technique developed in [10] and by the Hilbert space embedding of distributions proposed in [11].

References

- [1] Raphaël Bailly, François Denis, and Liva Ralavaiola. Grammatical inference as a principal component analysis problem. In *Proceedings of the 26th International Conference on Machine Learning*, pages 33–40, Montréal, Canada, June 2009. Omnipress.
- [2] Jean Berstel and Christophe Reutenauer. Recognizable formal power series on trees. *Theoretical computer science*, 18:115–148, 1982.
- [3] Björn Borchardt. *The Theory of Recognizable Tree Series*. PhD thesis, TU Dresden, 2004.
- [4] R.C. Carrasco, J. Oncina, and J. Calera-Rubio. Stochastic inference of regular tree languages. *Machine Learning*, 44(1/2):185–197, 2001.
- [5] A. Clark, C. Costa Florêncio, and C. Watkins. Languages as hyperplanes: grammatical inference with string kernels. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 90–101, 2006.
- [6] Hubert Comon, Max Dauchet, Rémi Gilleron, Florent Jacquemard, Denis Lugiez, Christof Löding, Sophie Tison, and Marc Tommasi. Tree automata techniques and applications. Available on: <http://tata.gforge.inria.fr/>, 2007. release October, 12th 2007.
- [7] F. Denis and Y. Esposito. On rational stochastic languages. *Fundamenta Informaticae*, 86:41–77, 2008.
- [8] F. Denis, E. Gilbert, A. Habrard, F. Ouardi, and M. Tommasi. Relevant representations for the inference of rational stochastic tree languages. In *Grammatical Inference: Algorithms and Applications, 9th International Colloquium*, pages 57–70. Springer, 2008.
- [9] François Denis and Amaury Habrard. Learning rational stochastic tree languages. In *Algorithmic learning theory*, volume 4754 of *Lecture Notes in Artificial Intelligence*, pages 242–256. 18th International Conference, ALT 2007, Springer-Verlag, Octobre 2007.
- [10] J. Shawe-Taylor, N. Cristianini, and J.S. Kandola. On the concentration of spectral properties. In *Advances in Neural Information Processing Systems*, volume 14, pages 511–517. MIT Press, 2001.
- [11] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *18th International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.