Learning rational stochastic tree languages

Francois Denis, Amaury Habrard

Learning Rational Stochastic Tree Languages*

François Denis and Amaury Habrard

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

Abstract. We consider the problem of learning stochastic tree languages, i.e. probability distributions over a set of trees $T(\mathcal{F})$, from a sample of trees independently drawn according to an unknown target P. We consider the case where the target is a rational stochastic tree language, i.e. it can be computed by a rational tree series or, equivalently, by a multiplicity tree automaton. In this paper, we provide two contributions. First, we show that rational tree series admit a canonical representation with parameters that can be efficiently estimated from samples. Then, we give an inference algorithm that identifies the class of rational stochastic tree languages in the limit with probability one.

1 Introduction

In this paper, we stand in the field of probabilistic grammatical inference and we focus on the learning of stochastic tree languages. A *stochastic tree language* is a probability distribution over the set of trees $T(\mathcal{F})$ built on a ranked finite alphabet $\mathcal{F}$. Given a sample of trees independently drawn according to an unknown stochastic language P, we aim at finding an estimate of P in a given class of models such as *probabilistic tree automata*. Carrasco *et al.* have proposed to learn *deterministic* stochastic tree automata [1]. Specific work for *probabilistic k-testable tree languages* was presented in [2] and for learning stochastic grammars in [3]. However, to our knowledge, no efficient inference algorithm capable of identifying the whole class of probabilistic tree automata is known.

Here, we can make a parallel with results on stochastic languages on strings. Indeed, there exists no efficient algorithm capable of identifying the whole class of probabilistic automata on strings either and the main reason is that we cannot define a canonical structure for these models. Most former results deal with specific subclasses of the class of probabilistic automata. Recently, it has been proposed to consider a larger class of models: the class $\mathcal{S}_{\mathbb{R}}^{rat}$ of rational stochastic languages [4]. In the field of strings, a *rational stochastic language* is a stochastic language that can be computed by a *multiplicity automaton*, whose parameters may be positive or negative. Rational stochastic languages have a minimal canonical representation while such canonical representations do not exist for

* This work was partially supported by the Marmota project ANR-05-MMSA-0016

probabilistic automata. And it has been shown that the class of rational stochastic languages can be inferred in the limit with probability 1 [5,6]. The aim of this paper is to study an extension of these results to the case of trees.

A tree series is a mapping from $T(\mathcal{F})$ to $\mathbb{R}$. Rational tree series have been studied in [7,8]. As far as we know, very few approaches have focused on the learning of tree series but we can mention two papers that stand in a variant of the MAT learning model of Angluin: [9] in a general case and [10] in a deterministic case. But, to the best of our knowledge, this is the first attempt for learning rational stochastic tree languages. Note that the adaptation to trees is not trivial. Prefixes and suffixes of a string are also strings. The equivalent notions for trees are *subtrees* and *contexts* (a context c is a tree one leaf of which acts as a variable, so that substituting a tree t to the variable yields a new tree $c[t]$), which are not similar objects. In the case of words, it can be shown that any rational series r has a canonical representation that can be built on derived rational series of the form $\dot{u}r$ such that $\dot{u}r(v) = r(uv)$ for any string v. The corresponding notion for trees could be rational series of the form $\dot{c}r$ where c is a context, which associates $r(c[t])$ with each tree t. However, it seems impossible to build a canonical representation on them and we need to consider much more sophisticated objects. Let $\mathbb{R}\langle\langle T(\mathcal{F})\rangle\rangle$ be the vector space composed of all series defined on $T(\mathcal{F})$, let $r \in T(\mathcal{F})$ be a tree rational series, let W be the subspace of $\mathbb{R}\langle\langle T(\mathcal{F})\rangle\rangle$ spanned by all the series of the form $\dot{c}r$.

The first result of this paper shows that a canonical representation of r can be defined on the dual vector space W^* composed of all the linear forms defined on W. We show that given an order on $T(\mathcal{F})$, a canonical basis $\{\overline{t_1}, \ldots, \overline{t_n}\}$ - whose elements naturally correspond to trees - can be defined for W^*. This point is important from a machine learning perspective. We show that such a basis can be extracted from any sufficiently large sample of trees drawn according to the target. This leads us to the inference part of our paper.

Our second contribution consists in proposing an inference algorithm which identifies in the limit any rational stochastic tree language with probability one. We show that there exists a sample size above which, the structure of the canonical representation is identified with probability one. Moreover, we show that the parameters output by the algorithm converge to the true parameters at a convergence rate equal to $O(|S|^{\gamma})$ for any $\gamma \in]-1/2, 0[$.

The paper is organized as follows. In Section 2, we introduce some preliminaries on tree series. The canonical linear representation for rational tree series is presented in Section 3. We propose our inference algorithm in Section 4. We conclude by a discussion and a description of future work in Section 5.

2 Preliminaries

2.1 Formal Power Series on Trees

See [11] for references on trees. Let $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1 \cup \cdots \cup \mathcal{F}_p$ be a ranked alphabet where the elements in $\mathcal{F}_m$ are the function symbols of *arity* m. Let $T(\mathcal{F})$ be the

set of all the *trees* that can be constructed from $\mathcal{F}$. Let us define the *height* of a tree t by: $height(t) = 0$ if $t \in \mathcal{F}_0$ and $height(t) = 1 + Max\{height(t_i)|i = 1..m\}$ if $t = f(t_1, \ldots, t_m)$. For any integer n, let us define $T^n(\mathcal{F})$ (resp. $T^{\leq n}(\mathcal{F})$) the set of trees whose height is equal to n (resp. $\leq n$).

Let \$ be a zero arity function symbol not in $\mathcal{F}_0$. A *context* is an element of $T(\mathcal{F} \cup \{\$\})$ such that the symbol \$ appears exactly once. We denote by $C(\mathcal{F})$ the set of all the contexts that can be defined over $\mathcal{F}$. Let t be a tree and let c be a context, $c[t]$ denotes the tree obtained by substituting the symbol \$ in the context c by the tree t. A subset A of $T(\mathcal{F})$ is *prefixial* if for any $c \in C(\mathcal{F})$ and any $t \in T(\mathcal{F})$, $c[t] \in A \Rightarrow t \in A$.

A *formal power tree series* on $T(\mathcal{F})$ is a mapping $r : T(\mathcal{F}) \to \mathbb{R}$. The set of all formal power series on $T(\mathcal{F})$ is denoted by $\mathbb{R}\langle\langle T(\mathcal{F})\rangle\rangle$. It is a vector space, when provided with addition and multiplication by a scalar.

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

Definition 1. 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* $\mu : \mathcal{F} \to \mathcal{L}$ *maps* $\mathcal{F}_p$ *into* $\mathcal{L}(V^p; V)$ *for each* $p \geq 0$*.*

Thus for each $f \in \mathcal{F}_p$, $\mu(f) : V^p \to V$ is p-linear. It can easily be shown that μ extends uniquely to a morphism $\mu : T(\mathcal{F}) \to V$ by the formula

$$\mu(f(t_1, \ldots, t_p)) = \mu(f)(\mu(t_1), \ldots, \mu(t_p)). \tag{1}$$

The μ function can be extended to work over contexts. Let $\overline{\mu} : C(\mathcal{F}) \to \mathcal{L}(V; V)$ be inductively by $\overline{\mu}(\$)(v) = v$ and $\overline{\mu}(f(t_1, \ldots, t_{i-1}, c, t_{i+1}, \ldots, t_n))(v) = \mu(f)(\mu(t_1), \ldots, \mu(t_{i-1}), \overline{\mu}(c)(v), \mu(t_{i+1}), \ldots, \mu(t_n))$.

It can be shown that for any context c and any term t, $\overline{\mu}(c)(\mu(t)) = \mu(c[t])$.

Let (V, μ) be a linear representation of $T(\mathcal{F})$ and let $V_{T(\mathcal{F})}$ be the vector subspace of V spanned by $\mu(T(\mathcal{F}))$. It can be shown that $(V_{T(\mathcal{F})}, \mu)$ is also a linear representation of $T(\mathcal{F})$. Let A be a prefixial subset of $T(\mathcal{F})$ and let V_A be the subspace of V spanned by $\mu(A)$. Suppose that for any integer m, any $f \in \mathcal{F}_m$ and any $t_1, \ldots, t_m \in A$, $\mu(f(t_1, \ldots, t_m)) \in V_A$. Then, $V_A = V_{T(\mathcal{F})}$. As a consequence, a basis of $V_{T(\mathcal{F})}$ can be extracted from $\mu(A)$. Therefore, given a linear representation (V, μ) of $T(\mathcal{F})$, a basis of $V_{T(\mathcal{F})}$ can be computed within polynomial time.

Definition 2. *Let* r *be a formal series over* $T(\mathcal{F})$*,* r *is a* recognizable tree series *if there exists a triple* (V, μ, λ)*, where* (V, μ) *is a linear representation of* $T(\mathcal{F})$*, and* $\lambda : V \to \mathbb{R}$ *is a linear form, such that* $r(t) = \lambda(\mu(t))$ *for all* t *in* $T(\mathcal{F})$*.*

We say that (V, μ, λ) is *trimmed* if (i) $V = V_{T(\mathcal{F})}$ and (ii) $\forall v \in V \setminus \{0\}, \exists c \in C(\mathcal{F}), \lambda\overline{\mu}(c)(v) \neq 0$.

Rational tree series have been studied in [7]. 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. Note also that rational series on strings can be seen as particular cases of rational series on trees and hence, counterexamples designed in the first field can be directly exported in the second one.

Example 1. Let $\mathcal{F} = \{a, b, g(\cdot), f(\cdot,\cdot)\}$, let $V = \mathbb{R}^2$ and let (e_1, e_2) be a basis of V. We define μ, λ and r by:

$$\mu(a) = 2e_1/3, \mu(b) = e_2/2, \mu(g)(e_1) = e_2/2, \mu(g)(e_2) = 0,$$

$$\mu(f)(e_i, e_j) = \begin{cases} e_1/3 \text{ if } i = 1 \text{ and } j = 2 \\ 0 \text{ otherwise} \end{cases}$$

and

$$\lambda(e_1) = 1, \lambda(e_2) = 0 \text{ and } r(t) = \lambda\mu(t) \text{ for any term } t.$$

We have $\mu(f(a,b)) = \mu(f)(\mu(a), \mu(b)) = e_1/9$ and
$\mu(f(a, g(a))) = \mu(f)(\mu(a), \mu(g)(\mu(a))) = 2e_1/27$.
Hence, $r(a) = 2/3,\ r(b) = 0,\ r(f(a,b)) = 1/9,\ r(f(a, g(a))) = 2/27$.

Definition 3. *A* multiplicity tree automaton *(MA) over $\mathcal{F}$ is a tuple $\mathcal{A} = (Q, \mathcal{F}, \tau, \delta)$ where Q is a set of states, τ is a mapping from Q to $\mathbb{R}$ and δ is a mapping from $\cup_{m\geq 0}\mathcal{F}_m \times Q^m \times Q$ to $\mathbb{R}$.*

A multiplicity automaton is a device that can be used to compute tree series. They can be interpreted in a bottom-up or a top-down way, since $\delta(f, q_1, \ldots, q_m, q) = w$ can be rewritten as a bottom-up rule or a top-down rule:

$$f(q_1, \ldots, q_m) \xrightarrow{w} q \text{ or } q \xrightarrow{w} f(q_1, \ldots, q_m).$$

A *probabilistic tree automaton* (PA) is an MA $\mathcal{A} = (Q, \mathcal{F}, \tau, \delta)$ which satisfies the following conditions:

- δ and τ take their values in $[0, 1]$,
- $\sum_{q\in Q} \tau(q) = 1$,
- for any $q \in Q$, $\sum_{f(q_1,\ldots,q_m)\xrightarrow{w}q} w = 1$.

Multiplicity automata and linear representations are two equivalent ways to represent rational series. For example, let (V, μ, λ) be a linear representation of the formal series r defined on $T(\mathcal{F})$ and let $B = (e_1, \ldots, e_n)$ be a basis of V. A multiplicity automaton $\mathcal{A} = (Q, \mathcal{F}, \lambda, \delta)$ can be associated with (V, μ, λ, B) as follows:

- $Q = \{e_1, \ldots, e_n\}$,
- $\delta(f, e_{i_1}, \ldots, e_{i_m}, e_j) = w_j$ for any $f \in \mathcal{F}_m$ where $\mu(f)(e_{i_1}, \ldots, e_{i_m}) = \sum_j w_j e_j$.

Conversely, an equivalent linear representation can be associated with any multiplicity automaton.

Example 2. It can easily be shown that the linear representation described in Example 1 is equivalent to the probabilistic automaton defined by: $Q = \{e_1, e_2\}$, $\tau(e_1) = 1$, $\tau(e_2) = 0$ and

$$\delta = \{e_1 \xrightarrow{2/3} a, e_1 \xrightarrow{1/3} f(e_1, e_2), e_2 \xrightarrow{1/2} b, e_2 \xrightarrow{1/2} g(e_1)\}.$$

2.2 Rational Stochastic Tree Languages

Definition 4. *A* stochastic tree language *over* $T(\mathcal{F})$ *is a tree series* $r \in K\langle\langle T(\mathcal{F})\rangle\rangle$ *such that for any* $t \in T(\mathcal{F})$, $0 \leq r(t) \leq 1$ *and* $\sum_{t \in T(\mathcal{F})} r(t) = 1$.

Therefore, a *rational stochastic tree language* is a stochastic tree language which admits a linear representation. Stochastic languages that can be computed by a probabilistic automaton are rational. However, the converse is false: there exists a rational stochastic tree language that cannot be computed by a probabilistic automaton [4]. Moreover, it can be shown that the rational series computed by a PA is not always a stochastic language. For example, it can easily be shown that the PA defined by $Q = \{q\}$, $\tau(q) = 1$, $\delta = \{q \xrightarrow{\alpha} a, q \xrightarrow{1-\alpha} f(q, q)\}$ defines a stochastic language iff $\alpha \geq 1/2$. When $\alpha < 1/2$, $\sum_{t \in T(\mathcal{F})} r(t) < 1$ [12].

Let P be a stochastic tree language over $T(\mathcal{F})$. We consider infinite samples S composed of trees independently drawn according to P. For any integer m, let S_m be the sample composed of the m first elements of S. We denote by P_{S_m} the empirical distribution on $T(\mathcal{F})$ associated with S_m. Let $\mathcal{A} = (A_i)_{i \in I}$ be a family of subsets of $T(\mathcal{F})$. It can be shown [13,14] that for any confidence parameter δ and any integer m, with a probability greater than $1 - \delta$, for any $i \in I$,

$$|P_{S_m}(A_i) - P(A_i)| \leq C\sqrt{\frac{d - \log\frac{\delta}{4}}{m}}. \tag{2}$$

where d is the Vapnik-Chervonenkis dimension of $\mathcal{A}$ and C is a universal constant. In particular, with a probability greater than $1 - \delta$, for any $t \in T(\mathcal{F})$,

$$|P_{S_m}(t) - P(t)| \leq C\sqrt{\frac{1 - \log\frac{\delta}{4}}{m}}. \tag{3}$$

Let $\Psi(d, \epsilon, \delta) = \frac{C^2}{\epsilon^2}(d - \log\frac{\delta}{4})$. One can easily verify that if $m \geq \Psi(d, \epsilon, \delta)$, with a probability greater than $1 - \delta$, $|P_{S_m}(A_i) - P(A_i)| \leq \epsilon$ for any index i.

Borel-Cantelli Lemma states that if $(A_k)_{k \in \mathbb{N}}$ is a family of events such that $\sum_k P(A_k) < \infty$, the probability that a finite number of events A_k occur is equal to 1.

Check that for any α such that $-1/2 < \alpha < 0$ and any $\beta < -1$, if we define $\epsilon_k = k^\alpha$ and $\delta_k = k^\beta$, then there exists K such that for all $k \geq K$, we have $k \geq \Psi(1, \epsilon_k, \delta_k)$. For such choices of α and β, we have $\lim_{k \to \infty} \epsilon_k = 0$ and $\sum_{k \geq 1} \delta_k < \infty$. Therefore, from Borel-Cantelli Lemma, it can easily be shown that with probability 1, there exists K such that for any $k \geq K$, for any $t \in T(\mathcal{F})$,

$$|P_{S_k}(t) - P(t)| \leq \epsilon_k. \tag{4}$$

3 A Canonical Linear Representation for Rational Tree Series

The main goal of the paper is to show that any rational stochastic tree language P can be inferred in the limit from an infinite sample drawn according to P with probability 1. The first step is to define the *canonical linear representation* of a rational tree series r, whose components only depend on r.

3.1 Defining the Canonical Representation

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

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

Lemma 1. *Let (V, μ, λ) be a linear representation of the rational series r. For any context c, $\dot{c}r$ is rational and $(V, \overline{\mu}(c) \circ \mu, \lambda)$ is a linear representation of $\dot{c}r$.*

Proof. Indeed, for any term t, $\dot{c}r(t) = r(c[t]) = \lambda\mu(c[t]) = \lambda(\overline{\mu}(c) \circ \mu)(t)$. □

Let r be a formal power series on $T(\mathcal{F})$. Let us denote by W_r the vector subspace of $\mathbb{R}\langle\langle T(\mathcal{F})\rangle\rangle$ spanned by $\{\dot{c}r | c \in C(\mathcal{F})\}$.

Lemma 2. *If r is rational, then the dimension of W_r is finite.*

Proof. Let (V, μ, λ) be a linear representation of r. For any context c, $\lambda\overline{\mu}(c) \in V^*$. Since the dimension of V^* is finite, there exist $c_1, \ldots c_n$ s. t. for any $c \in C(\mathcal{F})$, there exists $\alpha_1, \ldots, \alpha_n$ s.t. $\lambda\overline{\mu}(c) = \sum_i \alpha_i \lambda\overline{\mu}(c_i)$. Check that $\{\dot{c_1}r, \ldots, \dot{c_n}r\}$ spans W_r. □

Let W_r^* be the dual space of W_r, i.e. the set of all linear forms defined over W_r. For any $t \in T(\mathcal{F})$, let $\overline{t} \in W_r^*$ be defined by: $\forall s \in W_r, \overline{t}(s) = s(t)$.

Lemma 3. *Let $f(u_1, \ldots, u_i, \ldots, u_p), t_1, \ldots, t_n \in T(\mathcal{F})$ and suppose that $\overline{u_i} = \sum_{j=1}^n \alpha_j \overline{t_j}$ for some index i. Then,*
$\overline{f(u_1, \ldots, u_i, \ldots, u_p)} = \sum_{j=1}^n \alpha_j \overline{f(u_1, \ldots, t_j, \ldots, u_p)}$.

Proof. Let c_i be the context $f(u_1, \ldots, \$, \ldots, u_n)$ where \$ is at the i-th position. For any $s \in W_r$,

$$\overline{f(u_1, \ldots, u_i, \ldots, u_p)}(s) = \overline{u_i}(\dot{c_i}s) = \sum_{j=1}^{n} \alpha_j \overline{t_j}(\dot{c_i}s) = \sum_{j=1}^{n} \alpha_j \overline{f(u_1, \ldots, t_j, \ldots, u_p)}(s). \square$$

Suppose that the dimension of W_r is finite and let $\{c_1^{-1}r, \ldots, c_n^{-1}r\}$ be a basis of W_r. One can show that there exists n terms $t_1, \ldots, t_n$ such that the rank of the matrix $(c_i^{-1}r(t_j))_{1 \leq i,j \leq n}$ is n. Therefore, $(\overline{t_1}, \ldots, \overline{t_n})$ is a basis of W_r^*.

Let r be a rational series. We know that the dimension of W_r is finite. Let $t_1, \ldots, t_n$ be n terms such that $(\overline{t_1}, \ldots, \overline{t_n})$ is a basis of W_r^*. We define a linear representation (W_r^*, ν, τ) of r as follows:

– for any $f \in \mathcal{F}_p$, define $\nu(f) \in \mathcal{L}((W_r^*)^p; W_r^*)$ by $\nu(f)(\overline{t_{i_1}}, \ldots, \overline{t_{i_p}}) = \overline{f(t_{i_1}, \ldots, t_{i_p})}$.
– $\tau \in (W_r^*)^* = W_r$ by $\tau(\overline{t}) = r(t)$.

Lemma 4. *For any term $t \in T(\mathcal{F})$, $\nu(t) = \overline{t}$.*

Proof. Let $t = f(s_1, \ldots, s_p) \in T(\mathcal{F})$ and let $\overline{s_i} = \sum_{j=1}^{n} \alpha_i^j \overline{t_i}$. Using the previous lemma, we have

$$\nu(f)(\overline{s_1}, \ldots, \overline{s_p}) = \sum_{j_1, \ldots, j_p} \alpha_1^{j_1} \ldots \alpha_p^{j_p} \overline{f(t_{j_1}, \ldots, t_{j_p})} = \overline{f(s_1, \ldots, s_p)}$$

Remark that ν and τ do not depend on any basis chosen for W_r^*.

Theorem 1. *(W_r^*, ν, τ) is a trimmed linear representation of r which is called the* canonical linear representation *of r.*

Proof. For any term t, $\tau(\nu(t)) = \tau(\overline{t}) = r(t)$. Therefore, (W_r^*, ν, τ) is a linear representation of r. By construction, $\nu(T(\mathcal{F}))$ spans W_r^*. Now, let $w \in W_r^* \setminus \{0\}$ and let $\{\overline{t_1}, \ldots, \overline{t_n}\}$ be a basis of W_r^*. There exist $\alpha_1, \ldots, \alpha_n$ not all zero s.t. $w = \sum \alpha_i \overline{t_i}$. Since $\{\overline{t_1}, \ldots, \overline{t_n}\}$ is linearly independent, there exists a context c such that $\sum \alpha_i \overline{t_i}(c) = \tau\overline{\nu}(c)(w) \neq 0$. Therefore, (W_r^*, ν, τ) is trimmed. □

Given a total order $\leq$ on $T(\mathcal{F})$, there exists a unique subset B of $\nu(T(\mathcal{F}))$ which is a basis of W_r^* and such that for any $s \in T(\mathcal{F})$, $\overline{s} \in B$ or $\{\overline{s}\} \cup \{\overline{t} \in B | t \leq s\}$ is linearly dependent. We say that B is the canonical basis of W_r^* (wrt $\leq$).

3.2 Building the Canonical Representation

Given an n-dimensional trimmed linear representation (V, μ, λ) for r, it is possible to build the canonical representation of r in time polynomial in n^p where p is the maximal arity of symbols in $\mathcal{F}$. The proof of this result relies on the following lemma:

Lemma 5. *Let (V, μ, λ) be an n-dimensional trimmed linear representation (V, μ, λ) for r and let $t_1, \ldots, t_m \in T(\mathcal{F})$. Then, $\{\overline{t_1}, \ldots, \overline{t_m}\}$ is linearly independent iff $\{\mu(t_1), \ldots, \mu(t_m)\}$ is linearly independent.*

Proof. Suppose that $\{\mu(t_1), \ldots, \mu(t_m)\}$ is linearly independent in V and let $\alpha_1, \ldots, \alpha_m$ be such that $\sum \alpha_{i=1}^m \overline{t_i} = 0$. For any context c, $\sum_i \alpha_i \overline{t_i}(c) = \sum_i r(c[t_i]) = \lambda\overline{\mu}(c)(\sum_i \alpha_i \mu(t_i)) = 0$. Therefore, $\sum_i \alpha_i \mu(t_i) = 0$ since (V, μ, λ) is trimmed and $\alpha_i = 0$ for $i = 1, \ldots, n$ since $\{\mu(t_1), \ldots, \mu(t_m)\}$ is linearly independent: hence, $\{\overline{t_1}, \ldots, \overline{t_m}\}$ is linearly independent.

Suppose that $\{\mu(t_1), \ldots, \mu(t_m)\}$ is linearly dependent and let $\sum_i \alpha_{i=1}^m \mu(t_i) = 0$ where the α_i are not all zero. For any context c,

$$\sum_i \alpha_i \overline{t_i}(c) = \sum_i \alpha_i r(c[t_i]) = \sum_i \alpha_i \lambda\mu(c[t_i]) = \lambda\overline{\mu}(c)(\sum_i \alpha_i \mu(t_i)) = 0.$$

Therefore, $\sum_{i=1}^m \alpha_i \overline{t_i} = 0$. □

```
Data      : A trimmed linear representation (V, μ, λ) for r
Result    : A basis B of W_r^*
begin
    B ← ∅;    is_a_basis ← False;
    while not is_a_basis do
        is_a_basis ← True;
        for every f ∈ F do
            let p = arity(f);
            for t_1, ..., t_p ∈ B do
                if B ∪ f(t_1, ..., t_p)‾ is linearly independent then
                    B = B ∪ f(t_1, ..., t_p)‾;    is_a_basis ← False;
end
```

Algorithm 1: Building a canonical linear representation of r

Proposition 1. *Given an n-dimensional trimmed linear representation (V, μ, λ) for the rational series r, a basis for W_r^* can be computed in time polynomial in n^p.*

Proof. One can verify that Algorithm 1 computes a basis of W_r^*. □

One can remark that the linear representation is only used to check whether $B \cup \overline{f(t_1, \ldots, t_p)}$ is linearly independent. Therefore, the linear representation can be replaced by an oracle that says whether $B \cup \overline{f(t_1, \ldots, t_p)}$ is linearly independent. Such an oracle could be achieved, in a variant of the MAT learning model of Angluin, by using a *membership oracle* which would compute $r(t)$ for any tree t and an *equivalence oracle* which would say whether the current representation computes r, and would provide a counterexample $(t, r(t))$ otherwise. See [9,10] for related work.

Example 3. Let us consider the previous example.

- $\overline{a} \neq 0$ since $\overline{a}(\$) = 2/3$.
- $\{\overline{a}, \overline{b}\}$ is linearly independent since $\overline{a}(f(a, \$)) = 0$ and $\overline{b}(f(a, \$)) = 1/9$.
- We have $\overline{f(a, a)} = \overline{g(b)} = \overline{f(b, a)} = \overline{f(b, b)} = 0$.
- We have also $\overline{g(a)} = 2\overline{b}/3$ and $\overline{f(a, b)} = \overline{a}/6$.

Therefore, $\{\overline{a}, \overline{b}\}$ is a basis of the canonical linear representation of r.

4 Inference of Rational Tree Series in the Limit

In this section, we show how to identify in the limit a canonical linear representation of a rational stochastic tree language P from an infinite sample S of trees independently drawn according to P.

Let (W^*, ν, τ) be the canonical linear representation of the target. Given a total order $\leq$ on $T(\mathcal{F})$ satisfying $height(t) < height(t') \Rightarrow t \leq t'$, the aim of the algorithm is to identify the canonical basis $B = \{\overline{t_1}, \ldots, \overline{t_n}\}$ of W^* associated with $\leq$. Let t_{max} be the maximal element of $\{t_1, \ldots, t_n\}$. Let S be an infinite sample independently drawn according to P and let S_m be the sample composed of the m first elements of S. We have to show that with probability one, there exists an integer N such that for any $m \geq N$, the following properties can be identified from S_m:

- $B = \{\overline{t_1}, \ldots, \overline{t_n}\}$ is linearly independent,
- for any $t \leq t_{max}$, $B \cup \{\overline{t}\}$ is linearly dependent,
- for any $f \in \mathcal{F}$ and any $1 \leq i_1, \ldots, i_p \leq n$, $B \cup \{\overline{f(t_{i_1}, \ldots, t_{i_p})}\}$ is linearly dependent, where p is the arity of f.

Given these relations, a linear representation (W^*, ν_m, τ_m) can be computed. Then, we have to show that the (multi-) linear mappings $\nu_m(f)$ for any $f \in \mathcal{F}$ and τ_m converge to the correct ones.

Since we are working on finite samples S_m, we cannot consider exact linear dependencies. Let T be a finite subset of $T(\mathcal{F})$, let S_m be a finite sample composed of m trees independently drawn from the target, let $t \in T(\mathcal{F})$, let $\{x_s | s \in T\}$ be a set of variables and let $\epsilon > 0$. We denote by $I(T, t, S_m, \epsilon)$ the following set of inequalities :

$$I(T, t, S_m, \epsilon) = \{|\overline{t}(\dot{c}P_S) - \sum_{s \in T} x_s \overline{s}(\dot{c}P_S)| \leq \epsilon | c \in C(S_m)\}$$

where P_S is the empirical distribution on S_m and where $C(S) = \{c \in C(\mathcal{F}) | \exists t \in T(\mathcal{F}) \text{ s.t. } c[t] \in S_m\}$.

Let S be an infinite sample of the target P. Suppose that $\{\overline{t}\} \cup \{\overline{s} | s \in T\}$ is linearly independent. We show that, with probability 1, there exists $\epsilon > 0$ and a sample size from which $I(T, t, S_m, \epsilon)$ has no solution.

Lemma 6. *Let P be a stochastic language and let $\{t_0, t_1, \ldots, t_n\}$ be a set of trees such that $\{\overline{t_0}, \overline{t_1}, \ldots, \overline{t_n}\}$ is linearly independent. Then, with probability one, for any infinite sample S of P, there exists a positive number ϵ and an integer M such that for every $m \geq M$, $I(\{t_1, \ldots, t_n\}, t_0, S_m, \epsilon)$ has no solution.*

Proof. Let S be an infinite sample of P. Suppose that for every $\epsilon > 0$ and every integer M, there exists $m \geq M$ such that $I(\{t_1, \ldots, t_n\}, t_0, S_m, \epsilon)$ has a solution. Then, for any integer k, there exists $m_k \geq k$ such that $I(\{t_1, \ldots, t_n\}, t_0, S_{m_k}, 1/k)$ has a solution $(\alpha_{1,k}, \ldots, \alpha_{n,k})$.

Let $\rho_k = Max\{1, |\alpha_{1,k}|, \ldots, |\alpha_{n,k}|\}$, $\gamma_{0,k} = 1/\rho_k$ and $\gamma_{i,k} = -\alpha_{i,k}/\rho_k$ for $1 \leq i \leq n$. For every k, $Max\{|\gamma_{i,k}| : 0 \leq i \leq n\} = 1$. Check that for any context c: $\forall k \geq 0$, $\left|\sum_{i=0}^{n} \gamma_{i,k} \overline{t_i}(\dot{c}P_{S_{m_k}})\right| \leq \frac{1}{\rho_k k} \leq \frac{1}{k}$.

There exists a subsequence $(\alpha_{1,\phi(k)}, \ldots, \alpha_{n,\phi(k)})$ of $(\alpha_{1,k}, \ldots, \alpha_{n,k})$ such that $(\gamma_{0,\phi(k)}, \ldots, \gamma_{n,\phi(k)})$ converges to $(\gamma_0, \ldots, \gamma_n)$. We show below that we should

have $\sum_{i=0}^{n} \gamma_i \overline{t_i}(\dot{c}P) = 0$ for every context c, which is contradictory with the independence assumption since $Max\{\gamma_i : 0 \leq i \leq n\} = 1$ and hence, some γ_i is not zero.

Let $c \in C(\mathcal{F})$. With probability 1, there exists an integer k_0 such that $c \in C(S_{m_k})$ for any $k \geq k_0$. For such a k, we can write

$\gamma_i \overline{t_i}(\dot{c}P) = (\gamma_i \overline{t_i}(\dot{c}P) - \gamma_i \overline{t_i}(\dot{c}P_{S_{m_k}})) + (\gamma_i - \gamma_{i,\phi(k)})\overline{t_i}(\dot{c}P_{S_{m_k}}) + \gamma_{i,\phi(k)}\overline{t_i}(\dot{c}P_{S_{m_k}})$

and therefore $\left|\sum_{i=0}^{n} \gamma_i \overline{t_i}(\dot{c}P)\right| \leq \sum_{i=0}^{n} |\overline{t_i}(\dot{c}P - \dot{c}P_{S_{m_k}})| + \sum_{i=0}^{n} |\gamma_i - \gamma_{i,\phi(k)}| + \frac{1}{k}$ which converges to 0 when k tends to infinity.

□

Let S be an infinite sample of the target P. Suppose that $\overline{t} = \sum_{s \in T} \alpha_s \overline{s}$. We show that, with probability 1, for any $\gamma \in]-1/2, 0[$, there exists a sample size M from which, $I(T, t, S_m, m^\gamma)$ has a solution for any $m \geq M$.

Lemma 7. *Let P be a stochastic language and let $t_0, t_1, \ldots, t_n$ be a set of trees such that there exist $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that $\overline{t_0} = \sum_{i=1}^{n} \alpha_i \overline{t_i}$. Then, for any $\gamma \in]-1/2, 0[$, with probability one, for any infinite sample S of P, there exists K s.t. $I(\{t_1, \ldots, t_n\}, t_0, S_k, k^\gamma)$ has a solution for every $k \geq K$.*

Proof. Let S an infinite sample of P. Let $\alpha_0 = 1$ and let $R = Max\{|\alpha_i| : 0 \leq i \leq n\}$. With probability one, there exists K_1 s.t. $\forall k \geq K_1$, $k \geq \Psi(1, [k^\gamma (n+1)R]^{-1}, [(n+1)k^2]^{-1})$ (see definition of Ψ in Section 2). Let $k \geq K_1$, for any $c \in C(\mathcal{F})$,

$$|\overline{t_0}(\dot{c}P_{S_k}) - \sum_{i=1}^{n} \alpha_i \overline{t_i}(\dot{c}P_{S_k})| \leq |\overline{t_0}(\dot{c}P_{S_k}) - \overline{t_0}(\dot{c}P)| + \sum_{i=1}^{n} |\alpha_i| |\overline{t_i}(\dot{c}P_{S_k}) - \overline{t_i}(\dot{c}P)|.$$

From the definition of Ψ, with probability greater than $1 - \frac{1}{k^2}$, for any $i = 0, \ldots, n$ and any context c, $|\overline{t_i}(\dot{c}P_{S_k}) - \overline{t_i}(\dot{c}P)| \leq [k^{-\gamma}(n+1)R]^{-1}$ and therefore $|\overline{t_0}(\dot{c}P_{S_k}) - \sum_{i=1}^{n} \alpha_i \overline{t_i}(\dot{c}P_{S_k})| \leq k^\gamma$. For any integer $k \geq K_1$, let E_k be the event: $|\overline{t_0}(\dot{c}P_{S_k}) - \sum_{i=1}^{n} \alpha_i \overline{t_i}(\dot{c}P_{S_k})| > k^\gamma$. Since $Pr(E_k) < 1/k^2$, from the Borel-Cantelli Lemma, the probability that a finite number of E_k occurs is 1.

Therefore, with probability 1, there exists an integer K such that for any $k \geq K$, $I(\{t_1, \ldots, t_n\}, t_0, S_k, k^\gamma)$ has a solution. □

In the next lemma, we focus on the convergence of the parameters found when resolving an inequation system.

Lemma 8. *Let $P \in \mathcal{S}(T(\mathcal{F}))$, let $t_0, t_1, \ldots, t_n$ such that $\{\overline{t_1}, \ldots, \overline{t_n}\}$ is linearly independent and let $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ be such that $\overline{t_0} = \sum_{i=1}^{n} \alpha_i \overline{t_i}$. Then, for any $\gamma \in]-1/2, 0[$, with probability one, for any infinite sample S of P, there exists an integer K such that $\forall k \geq K$, any solution $\widehat{\alpha_1}, \ldots, \widehat{\alpha_n}$ of $I(\{t_1, \ldots, t_n\}, t_0, S_k, k^\gamma)$ satisfies $|\widehat{\alpha_i} - \alpha_i| < O(k^\gamma)$ for $1 \leq i \leq n$.*

Proof. Let $c_1, \ldots, c_n \in C(\mathcal{F})$ be such that the square matrix M defined by $M[i,j] = \overline{t_j}(\dot{c_i}P)$ for $1 \leq i, j \leq n$ is invertible. Let $A = (\alpha_1, \ldots, \alpha_n)^t$, $U = (\overline{t_0}(\dot{c_1}P), \ldots, \overline{t_0}(\dot{c_n}P))^t$. We have $M \times A = U$. Let S be an infinite sample of

P, let $k \in \mathbb{N}$ and let $\widehat{\alpha_1}, \ldots, \widehat{\alpha_n}$ be a solution of $I(\{t_1, \ldots, t_n\}, t_0, S_k, k^\gamma)$. Let M_k be the square matrix defined by $M_k[i,j] = \overline{t_j}(\dot{c_i} P_{S_k})$ for $1 \leq i, j \leq n$, let $A_k = (\widehat{\alpha_1}, \ldots, \widehat{\alpha_n})^t$ and $U_k = (\overline{t_0}(\dot{c_1} P_{S_k}), \ldots, \overline{t_0}(\dot{c_n} P_{S_k}))^t$. We have

$$\|M_k A_k - U_k\|^2 = \sum_{i=1}^{n} [\overline{t_0}(\dot{c_i} P_{S_k}) - \sum_{j=1}^{n} \widehat{\alpha_j} \overline{t_j}(\dot{c_i} P_{S_k})]^2 \leq n k^{2\gamma}.$$

Check that $A - A_k = M^{-1}(MA - U + U - U_k + U_k - M_k A_k + M_k A_k - M A_k)$ and therefore, for any $1 \leq i \leq n$

$$|\alpha_i - \widehat{\alpha_i}| \leq \|A - A_k\| \leq \|M^{-1}\|(\|U_0 - U_k\| + n^{1/2} k^\gamma + \|M_k - M\| \|A_k\|).$$

Now, by using Equation 4 and Borel-Cantelli Lemma as in the proof of Lemma 7, with probability 1, there exists K such that for all $k \geq K$, $\|U_0 - U_k\| < O(k^\gamma)$ and $\|M_k - M\| < O(k^\gamma)$. Therefore, for all $k \geq K$, any solution $\widehat{\alpha_1}, \ldots, \widehat{\alpha_n}$ of $I(\{t_1, \ldots, t_n\}, t_0, S_k, k^\gamma)$ satisfies $|\widehat{\alpha_i} - \alpha_i| < O(k^\gamma)$ for $1 \leq i \leq n$. □

The learning algorithm is presented in Algorithm 2 and works as follows. We suppose that a total order is defined over $T(\mathcal{F})$ such that $height(t) < height(t') \Rightarrow t \leq t'$. To begin with, we extract the first constant symbol a_0 of the learning sample and we put it in the basis set B. We define the frontier set (FS) to be composed of all the trees of the form $f(a_0, \ldots, a_0)$. Note that FS contains all the constant symbols different from a_0. Then, the algorithm processes the frontier set while it is not empty. For each tree t in this set, we check if it can approximately be expressed according to a linear combination of the elements of the current basis. If the answer is no, we add t to the basis and we enlarge the frontier set by adding all the trees of the form $f(t_1, \ldots, t_m)$ where every $t_i \in B$. Otherwise, we use the linear relation obtained from the inequation system to complete the definition of μ.

We can now present the theorem of convergence in the limit.

Theorem 2. *Let P be a rational stochastic tree language defined on $T(\mathcal{F})$, let (V, μ, λ) be the canonical linear representation of P, let $B = \{\overline{t_1}, \ldots, \overline{t_n}\}$ the canonical basis of V (associated with some known total order on $T(\mathcal{F})$) and let $\gamma \in]-1/2, 0[$. Then, with probability one, for any infinite sample S of P, there exists an integer K such that for any $k \geq K$, $Algo(S_k, \gamma)$ identifies B. Moreover, let (V, μ_k, λ_k) be the linear representation output by the algorithm. There exists a constant C such that $|\mu_k(f)(t_{i_1}, \ldots, t_{i_n}) - \mu(f)(t_{i_1}, \ldots, t_{i_n})| \leq C k^\gamma$ and $|\lambda_k(t_i) - \lambda(t_i)| \leq C k^\gamma$ for any $f \in \mathcal{F}$ and any elements t_i, t_{i_j} of B.*

Proof. Lemmas 6 and 7 prove that the basis B will be identified from some step with probability one. Lemma 8 can then be used to prove the last part of the theorem. □

When P is a rational stochastic tree language which takes its values in the set of rational numbers $\mathbb{Q}$, the algorithm can be completed to exactly identify it. The proof is based on the representation of real numbers by continuous fractions. See [15] for a survey on continuous fractions and [16] for a similar application.

Data : S a finite sample of k trees, $\gamma \in]-1/2, 0[$
Result : a linear representation (V, λ, μ)
begin
  $a_0 \leftarrow min(\mathcal{F}_0 \cap Subtrees(S))$;
  $B \leftarrow \{\overline{a_0}\}$; $\mu(a_0) \leftarrow \overline{a_0}$; $\lambda_{\overline{a_0}} \leftarrow P_s(a_0)$;
  $FS \leftarrow \bigcup_{f \in \mathcal{F}_p, p \geq 0} \{f(t_{j_1}, \ldots, t_{j_p}) | \overline{t_{i_j}} \in B\}$; $FS \leftarrow FS \backslash \{a_0\}$;
  while $FS \neq \emptyset$ **do**
    $t \leftarrow min(FS)$; $FS \leftarrow FS \backslash \{t\}$;
    if $I(B, t, S, k^\gamma)$ *has no solution* **then**
      $B \leftarrow B \cup \{\overline{t}\}$; $\mu(t) \leftarrow \overline{t}$; $\lambda_{\overline{t}} \leftarrow P_S(t)$;
      $FS \leftarrow FS \bigcup_{f \in \mathcal{F}_p, p \geq 1} \{f(t_{j_1}, \ldots, t_{j_p}) | \overline{t_{j_i}} \in B\}$;
    else
      Let $(\alpha_{t_i})_{t_i \in B}$ a solution of I; $\mu(t) \leftarrow \sum_{t_i \in B} \alpha_{t_i} \overline{t_i}$;
end

Algorithm 2: Learning algorithm Algo(S,γ)

Let (ϵ_n) be a sequence of non negative real numbers which converges to 0, let $x \in \mathbb{Q}$, let (y_n) be a sequence of elements of $\mathbb{Q}$ such that $|x - y_n| \leq \epsilon_n$ for all but finitely many n. It can be shown that there exists an integer N such that, for any $n \geq N$, x is the unique rational number $\frac{p}{q}$ which satisfies $\left|y_n - \frac{p}{q}\right| \leq \epsilon_n \leq \frac{1}{q^2}$. Moreover, the unique solution of these inequalities can be computed from y_n.

Let P be a rational stochastic tree language which takes its values in $\mathbb{Q}$, let $\gamma \in]-1/2, 0[$, let S be an infinite sample of P and let (V, μ_k, λ_k) the linear representation output by the algorithm on input (S_k, γ). Let (V, μ'_k, λ'_k) be the representation derived from (V, μ_k, λ_k) by replacing every parameter $\alpha_k = \mu_k(\overline{f(t_{i_1}, \ldots, t_{i_n})})$ or $\alpha_k = \lambda_k(t_i)$ with a solution $\frac{p}{q}$ of $\left|\alpha_k - \frac{p}{q}\right| \leq k^{\gamma/2} \leq \frac{1}{q^2}$ and let $Algo'$ be the corresponding algorithm.

Theorem 3. *Let P be a rational stochastic tree language which takes its values in $\mathbb{Q}$, let $\gamma \in]-1/2, 0[$, and let (V, μ, λ) be its canonical linear representation. Then, with probability one, for any infinite sample S of P, there exists an integer K such that $\forall k \geq K$, $Algo'(S_k, \gamma)$ returns (V, μ, λ).*

Proof. From the previous theorem, for every parameter α of (V, μ, λ), the corresponding parameter α_k in (V, μ_k, λ_k) satisfies $|\alpha - \alpha_k| \leq Ck^\gamma$ for some constant C, from some step k, with probability one. Therefore, if k is sufficiently large, we have $|\alpha - \alpha_k| \leq k^{\gamma/2}$ and there exists an integer K such that $\alpha = p/q$ is the unique solution of $\left|\alpha - \frac{p}{q}\right| \leq k^{\gamma/2} \leq \frac{1}{q^2}$. Therefore, the parameter corresponding to α in the linear representation output by $Algo'(S_k, \gamma)$ is α itself. □

Example 4. To illustrate the principle of our algorithm. Consider the following learning sample made up of 20 trees (the number of occurrences of each term is indicated inside brackets):

$\{a[13], f(a,b)[4], f(a,g(a))[1], f(a,g(f(a,g(a))))[1], f(f(f(a,g(a)),b),b)[1]\}$.

In a first step the algorithm puts $\overline{a}$ in the basis and sets $\mu(a) = \overline{a}$.

Next, the algorithm considers the constant symbol b. To check if $\overline{b}$ should belong to the basis, the algorithm constructs a set of inequations with the contexts definable in the learning set. For sake of simplicity, we will not consider all the contexts, but only 3 of them $c_0 = \$$, $c_1 = f(\$, b)$, $c_2 = f(a, \$)$. We obtain the following inequation system:

$$|\overline{b}(\dot{c}_0 p_S) - X_{\overline{a}}\overline{a}(\dot{c}_0 p_S)| = |p_S(c_0[b]) - X_{\overline{a}} p_S(c_0[a])| = |0 - X_{\overline{a}}\tfrac{13}{20}| \leq \epsilon$$
$$|\overline{b}(\dot{c}_1 p_S) - X_{\overline{a}}\overline{a}(\dot{c}_1 p_S)| = |p_S(c_1[b]) - X_{\overline{a}} p_S(c_1[a])| = |\tfrac{4}{20} - X_{\overline{a}}0| \leq \epsilon$$
$$|\overline{b}(\dot{c}_2 p_S) - X_{\overline{a}}\overline{a}(\dot{c}_2 p_S)| = |p_S(c_2[b]) - X_{\overline{a}} p_S(c_2[a])| = |0 - X_{\overline{a}}\tfrac{4}{20}| \leq \epsilon$$

If we set ϵ to 0.1, the systems admits no solution and then $\overline{b}$ is added to the basis with $\lambda_{\overline{b}} = 0$.

The algorithm examines the terms $f(a,a)$, $g(a)$, $f(a,b)$, $f(b,a)$, $f(b,b)$, $g(b)$. Since, the values of p_S according to the 3 contexts is null for $f(a,a)$ $f(b,a)$, $f(b,b)$ and $g(b)$ we do not show the inequation systems.

For $g(a)$ the system obtained is:

$$|\overline{g(a)}(\dot{c}_0 p_S) - X_{\overline{a}}\overline{a}(\dot{c}_0 p_S) - X_{\overline{b}}\overline{b}(\dot{c}_0 p_S)| = |0 - X_{\overline{a}}\tfrac{13}{20} - X_{\overline{a}}0| \leq \epsilon$$
$$|\overline{g(a)}(\dot{c}_1 p_S) - X_{\overline{a}}\overline{a}(\dot{c}_1 p_S) - X_{\overline{b}}\overline{b}(\dot{c}_1 p_S)| = |0 - X_{\overline{a}}\tfrac{4}{20} - X_{\overline{b}}| \leq \epsilon$$
$$|\overline{g(a)}(\dot{c}_2 p_S) - X_{\overline{a}}\overline{a}(\dot{c}_2 p_S) - X_{\overline{b}}\overline{b}(\dot{c}_2 p_S)| = |\tfrac{1}{20} - X_{\overline{a}}0 - X_{\overline{b}}\tfrac{4}{20}| \leq \epsilon$$

$X_{\overline{a}} = 0$ and $X_{\overline{b}} = \frac{1}{4}$ is a solution of the system, then the algorithm sets $\mu(g)(\overline{a}) = \frac{1}{4}\overline{b}$.

For $f(a,b)$, the inequation system is:

$$|\overline{f(a,b)}(\dot{c}_0 p_S) - X_{\overline{a}}\overline{a}(\dot{c}_0 p_S) - X_{\overline{b}}\overline{b}(\dot{c}_0 p_S)| = |\tfrac{4}{20} - X_{\overline{a}}\tfrac{13}{20} - X_{\overline{a}}0| \leq \epsilon$$
$$|\overline{f(a,b)}(\dot{c}_1 p_S) - X_{\overline{a}}\overline{a}(\dot{c}_1 p_S) - X_{\overline{b}}\overline{b}(\dot{c}_1 p_S)| = |0 - X_{\overline{a}}\tfrac{4}{20} - X_{\overline{b}}0| \leq \epsilon$$
$$|\overline{f(a,b)}(\dot{c}_2 p_S) - X_{\overline{a}}\overline{a}(\dot{c}_2 p_S) - X_{\overline{b}}\overline{b}(\dot{c}_2 p_S)| = |0 - X_{\overline{a}}0 - X_{\overline{b}}\tfrac{4}{20}| \leq \epsilon$$

$X_{\overline{a}} = \frac{4}{13}$ and $X_{\overline{b}} = 0$ is a solution of the system, then the algorithm sets $\mu(f)(\overline{a}, \overline{b}) = \frac{4}{13}\overline{a}$. The representation obtained is finally:

$\mu(a) = \overline{a}$, $\mu(b) = \overline{b}$, $\mu(g)(\overline{a}) = \frac{1}{4}\overline{b}$, $\mu(f)(\overline{a}, \overline{b}) = \frac{4}{13}\overline{a}$, $\lambda_{\overline{a}} = \frac{13}{20}$, $\lambda_{\overline{b}} = 0$.

5 Discussion, Future Work and Conclusion

We have proved a theoretical result: rational stochastic tree languages are identifiable in the limit with probability one. The inference algorithm we use runs within polynomial time and approximates the parameters of the model with usual statistical rates of convergence. How can it be used in practical cases? Can it be improved?

First of all, the algorithm highly relies on an inequation system which aims at detecting linear combinations

$$I(T, t, S_n, \epsilon) = \{|\overline{t}(\dot{c}P_{S_n}) - \sum_{s \in T} x_s \overline{s}(\dot{c}P_{S_n})| \leq \epsilon | c \in C(S_n)\}.$$

However, this system uses contexts which can be poorly represented in current samples. We can overcome this drawback by using *generalized contexts*, i.e. contexts containing several variables.

Let $\$_0, \$_1, \ldots, \$_k$ be zero arity function symbols not in $\mathcal{F}_0$. A generalized context is an element of $T(\mathcal{F} \cup \{\$_0, \$_1, \ldots, \$_k\})$ such that $\$_0$ appears exactly once and each other new symbol appears at most once. Now, for any stochastic languages P and any generalized context c, we define

$$\overline{t}(\dot{c}P) = \dot{c}P(t) = \sum_{t_1,\ldots,t_k \in T(\mathcal{F})} P(c[\$_0 \leftarrow t, \$_1 \leftarrow t_1, \ldots, \$_k \leftarrow t_k]).$$

We can then replace the inequation system $I(T, t, S_n, \epsilon)$ with

$$I(T, t, S_n, \epsilon) = \{|\overline{t}(\dot{c}P_{S_n}) - \sum_{s \in T} x_s \overline{s}(\dot{c}P_{S_n})| \leq \epsilon | c \in \mathcal{C}_k^g(S_n)\}$$

where $C_k^g(S_n)$ is the set of generalized context with k variables occurring in S_n.

If the number of new variables in not bounded, the VC-dimension of the set of generalized contexts is unbounded. However, it can easily be shown that the VC-dimension of the set of generalized contexts with k variables is bounded by $2k+1$. Therefore, we can adjust the number of variables to the size of the current learning sample in the inference algorithm in order to avoid overfitting.

Next, the rational series r output by the inference algorithm is not a stochastic language. Moreover, it may happen that the sum $\sum_{t \in T(\mathcal{F})} r(t)$ diverges. We conjecture that as soon as the size of the learning sample is large enough, with a high probability, the sum $\sum_{t \in T(\mathcal{F})} r(t)$ is absolutely convergent, i.e. $\sum_{t \in T(\mathcal{F})} |r(t)|$ converges. Moreover, let (V, μ, λ) be the canonical linear representation of a rational tree series r and let $B = \{\overline{t_1}, \ldots, \overline{t_n}\}$ be a basis of V. For any tree t and any index i, let α_i^t be such that $\overline{t} = \sum_{i=1}^n \alpha_i^t \overline{t_i}$. We have $r(t) = \sum_{i=1}^n \alpha_i^t r(t_i)$. We also conjecture that $\sum_{t \in T(\mathcal{F})} \alpha_i^t$ is absolutely convergent for any index i so that, $s_i = \sum_{t \in T(\mathcal{F})} \alpha_i^t$ is defined without ambiguity. One can show that s_i can be efficiently estimated.

Given these properties, it is possible to normalize the linear representation output by the algorithm in such a way that it computes a series $\overline{r}$ satisfying $\sum_{t \in T(\mathcal{F})} |\overline{r}(t)| < \infty$ and $\sum_{t \in T(\mathcal{F})} \overline{r}(t) = 1$. Let (V, μ_N, λ_N) be defined by

- $\forall f \in \mathcal{F}$, $[\mu_N(f)(\overline{t_{j_1}}, \ldots, \overline{t_{j_p}})]_i = [\mu(f)(\overline{t_{j_1}}, \ldots, \overline{t_{j_p}})]_i \cdot \pi_{k=1}^p s_{j_k}/s_i$.
- $\lambda_N(\overline{t_i}) = \lambda(\overline{t_i}) \times s_i$ for any element of λ_N.

It can easily be shown that (V, μ_N, λ_N) computes r and that $\sum_{\overline{t_{j_1}}, \ldots, \overline{t_{j_p}} \in B} [\mu_N(f(\overline{t_{j_1}}, \ldots, \overline{t_{j_p}}))]_i = 1$.
We can then adjust the linear form λ by multiplying each of its coordinates by a constant in order to get a series $\overline{r}$ which sums to 1.

However, it may happen that the series $\overline{r}$ takes negative values. We call such a series, a *pseudo-stochastic language*. From these languages, we can extract a probability distribution $P_{\overline{r}}$ such that $P_{\overline{r}}(t) = 0$ if $\overline{r}(t) < 0$ and otherwise

$P_{\overline{r}}(t) = b_t\overline{r}(t)$ with a normalization that compensates the loss of the negative values. We may compute this distribution iteratively when developing a tree. Suppose that at a given step, we are building a tree with some leaves labeled by states. We choose to develop a new branch from any of these states. We consider all the transitions leaving from the considered state grouped by symbols. If all the possible expansions with a given symbol lead to a negative value, then we omit this symbol and we renormalized the probabilities of the other expansions. Note that when r defines a stochastic language, $P_{\overline{r}} = r$ since there will be no negative values. See [6] for a more detailed description of this point, in the case of pseudo-stochastic languages defined on strings.

To conclude, we have studied in this paper the inference of a stochastic tree language P from a sample of trees independently drawn according to P. We have proposed to work in the class of rational stochastic tree languages that are stochastic languages computed by rational tree series. We have presented two contributions. First, we have shown that rational tree series admit a canonical linear representation. Then, we have proposed an inference algorithm which identifies in the limit the class of rational stochastic tree languages. Our future work will concern improvements of our approach in practical cases as evoked in the previous discussion.

References

1. Carrasco, R., Oncina, J., Calera-Rubio, J.: Stochastic inference of regular tree languages. Machine Learning **44**(1/2) (2001) 185–197
2. Rico-Juan, J., Calera, J., Carrasco, R.: Probabilistic k-testable tree-languages. In: Proceedings of ICGI 2000. Volume 1891 of LNCS., Springer (2000) 221–228
3. Abe, N., Mamitsuka, H.: Predicting protein secondary structure using stochastic tree grammars. Machine Learning Journal **29**(2-3) (1997) 275–301
4. Denis, F., Esposito, Y.: Rational stochastic languages. Technical report, LIF - Université de Provence, http://hal.ccsd.cnrs.fr/ccsd-00019728 (2006)
5. Denis, F., Esposito, Y., Habrard, A.: Learning rational stochastic languages. In: Proceedings of COLT'06. Volume 4005 of LNCS. (2006) Springer.
6. Habrard, A., Denis, F., Esposito, Y.: Using pseudo-stochastic rational languages in probabilistic grammatical inference. In: Proceedings of the 8th International Colloquium on Grammatical Inference (ICGI'06). Volume 4201 of LNCS., Springer (2006) 112–124
7. Berstel, J., Reutenauer, C.: Recognizable formal power series on trees. Theoretical Computer Science **18** (1982) 115–148
8. Ésik, Z., Kuich, W.: Formal tree series. Journal of Automata, Languages and Combinatorics **8**(2) (2003) 219–285
9. Habrard, A., Oncina, J.: Learning multiplicity tree automata. In: Proceedings of the 8th International Colloquium on Grammatical Inference (ICGI'06). Volume 4201 of LNCS., Springer (2006)
10. Drewes, F., Vogler, H.: Learning deterministically recognizable tree serie. Journal of Automata, Languages and Combinatorics (2007)

11. Comon, H., Dauchet, M., Gilleron, R., Jacquemard, F., Lugiez, D., Tison, S., Tommasi, M.: Tree Automata Techniques and Applications . Available from: http://www.grappa.univ-lille3.fr/tata (1997)
12. Wetherell, C.S.: Probabilistic languages: A review and some open questions. ACM Comput. Surv. **12**(4) (1980) 361–379
13. Vapnik, V.: Statistical Learning Theory. John Wiley (1998)
14. Lugosi, G.: Pattern classification and learning theory. In: Principles of Nonparametric Learning. Springer (2002)
15. Hardy, G., Wright, M.: An introduction to the theory of numbers. Oxford University Press (1979)
16. Denis, F., Esposito, Y.: Learning classes of probabilistic automata. In: COLT 2004. Volume 3120 of LNAI. (2004) 124–139