JMLR: Workshop and Conference Proceedings 34:124–138, 2014 Proceedings of the 12th ICGI

# Maximizing a Tree Series in the Representation Space

**Guillaume Rabusseau** GUILLAUME.RABUSSEAU@LIF.UNIV-MRS.FR
**François Denis** FRANCOIS.DENIS@LIF.UNIV-MRS.FR
*Aix Marseille Université, CNRS, LIF, 13288 Marseille Cedex 9, FRANCE*

**Editors:** Alexander Clark, Makoto Kanazawa and Ryo Yoshinaka

## Abstract

This paper investigates the use of linear representations of trees (i.e. mappings from the set of trees into a finite dimensional vector space which are induced by rational series on trees) in the context of structured data learning. We argue that this representation space can be more appealing than the space of trees to handle machine learning problems involving trees. Focusing on a tree series maximization problem, we first analyze its complexity to motivate the use of approximation techniques. We then show how a tree series can be extended to the continuous representation space, we propose an adaptive Metropolis-Hastings algorithm to solve the maximization problem in this space, and we establish convergence guarantees. Finally, we provide some experiments comparing our algorithm with an implementation of the Metropolis-Hastings algorithm in the space of trees.

**Keywords:** Rational tree series, Linear representation, Metropolis-Hastings, Markov chain Monte Carlo

## 1. Introduction

Rational tree series are mappings from the set of trees on a ranked alphabet to the set of real numbers, that can be computed by a weighted tree automaton. An equivalent characterization for a tree series to be rational, is that it admits a finite-dimensional linear representation, which induces a vectorial representation of trees (i.e. a mapping from the set of trees to a finite-dimensional vector space) Berstel and Reutenauer (1982); Denis et al. (2008). Unlike the space of trees which is a discrete space that does not have a natural topological structure, this representation space offers several interesting properties: it is linear, continuous, has a natural metric and may be of small dimension. Our goal is to show on a tree series maximization problem, that working in the representation space rather than directly in the space of trees can be beneficial and lead to better results.

We present a brief motivational example of the tree series maximization problem. In the context of procedural modeling, probabilistic context-free grammars (PCFG) are used as a mean to generate 3D models: in Talton et al. (2011), the authors use parametric conditional PCFGs to generate 3D images of trees (e.g. oaks), buildings or cities. They address the following problem: given such a grammar $G$ and a high-level specification $I$ of the desired image (e.g. a sketch), how can one retrieve a production from the grammar that matches the specification? They formulate this problem in the Bayesian setting by interpreting the distribution $\pi(\cdot)$ on the set $\Delta(G)$ of derivation trees induced by $G$ as the model prior, and a similarity measure $L(I|\cdot)$ between the image generated from a derivation tree and the sketch provided by the user as the likelihood. Finding the best tree that matches the user's

specification then reduces to maximizing the posterior $p(\cdot|I) \propto \pi(\cdot)L(I|\cdot)$. The algorithm they propose is a variant of the Metropolis-Hastings (MH) algorithm, which constructs a Markov chain in the discrete space of trees to explore it. Knowing that the prior distribution $\pi$ is a rational tree series, a natural question arises: could the natural structure of the underlying representation space be exploited to solve this problem? The broader problem we investigate here is the following: given a non-negative tree series $r$, how can we take advantage of a linear representation of trees to find a tree maximizing $r$?

The paper is organized as follows. First we introduce some preliminaries and notations. We analyze the complexity of the maximization problem and of two other problems related to the representation space in Section 3. In Section 4, we first present how the MH algorithm can be implemented in the discrete space of trees. We then show how the series to be maximized can be *extended* to the representation space and we propose an adaptive MH algorithm in this space, for which we establish convergence guarantees. We provide some experiments in Section 5 and we conclude by a discussion in Section 6.

## 2. Preliminaries

**Rational Tree Series**. We refer to Comon et al. (2007) for notions on trees, tree automata and recognizable forests, and to Berstel and Reutenauer (1982); Denis et al. (2008) for notions on rational tree series.

Let $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1 \cup \cdots \cup \mathcal{F}_m$ be a ranked alphabet where symbols in $\mathcal{F}_p$ are of arity $p$. *Trees* on $\mathcal{F}$ are elements of the smallest set $T_\mathcal{F}$ satisfying $\mathcal{F}_0 \subseteq T_\mathcal{F}$ and $f(t_1, \cdots, t_p) \in T_\mathcal{F}$ for all $p > 0$, $f \in \mathcal{F}_p$ and $t_1, \cdots, t_p \in T_\mathcal{F}$. A *context* $c$ of $C_\mathcal{F} \subseteq T_{\mathcal{F}\cup\{\$\}}$ is a tree where the special new symbol \$ (of arity 0) appears exactly once. We denote by $c[t] \in T_\mathcal{F}$ the tree resulting from the substitution of \$ with the tree $t \in T_\mathcal{F}$ in $c \in C_\mathcal{F}$. For any $t \in T_\mathcal{F}$, let $C_\mathcal{F}(t) = \{c \in C_\mathcal{F} \mid \exists t' \in T_\mathcal{F} : t = c[t']\}$ be the set of all suffixes of $t$ (i.e. the set of all contexts resulting from substituting any subtree of $t$ with the special symbol \$).

A *(formal power) tree series* on $T_\mathcal{F}$ is a mapping $r : T_\mathcal{F} \to \mathbb{R}$. Given two tree series $r, s \in \mathbb{R}\langle\langle\mathcal{F}\rangle\rangle$, we define their sum $r + s$ by $(r + s)(t) = r(t) + s(t)$, and their Hadamard product $r \odot s$ by $(r \odot s)(t) = r(t) \cdot s(t)$ for all $t \in T_\mathcal{F}$. We denote by $\mathbb{R}\langle\langle\mathcal{F}\rangle\rangle$ the vector space of tree series on $T_\mathcal{F}$. The *support* of a tree series $r$ is the forest $supp(r) = \{t \in T_\mathcal{F} : r(t) \neq 0\}$. A series $r$ is *recognizable* (or *rational*) if there exists a triple $(V, \mu, \lambda)$, where $V$ is a finite dimensional vector space, $\lambda \in V^*$ is a linear form, and $\mu$ maps each $\mathcal{F}_p$ into the set $\mathcal{L}(V^p; V)$ of $p$-linear mappings from $V^p$ to $V$, such that $r(t) = \lambda(\mu(t))$ for all $t$ in $T_\mathcal{F}$, where $\mu(t) \in V$ is inductively defined by $\mu(f(t_1, \ldots, t_p)) = \mu(f)(\mu(t_1), \ldots, \mu(t_p))$. A tuple $(V, \mu)$ is called a *linear representation* of $T_\mathcal{F}$, the tuple $(V, \mu, \lambda)$ is a *linear representation* of $r$, and the dimension of the vector space $V$ is its *size*.

A *stochastic tree series* is a tree series $r \in \mathbb{R}\langle\langle\mathcal{F}\rangle\rangle$ such that $r(t) \in [0, 1]$ for all $t \in T_\mathcal{F}$ and $\sum_{t\in T_\mathcal{F}} r(t) = 1$. For each context $c \in C_\mathcal{F}$ and each stochastic tree series $r$, we define the probability distribution $c^{-1}r$ on $T_\mathcal{F}$ by $[c^{-1}r](t) = r(c[t]) \Big/ \sum_{t'\in T_\mathcal{F}} r(c[t'])$ for all $t \in T_\mathcal{F}$.

**Theorem 2.1** (Berstel and Reutenauer (1982), Example 4.3, Proposition 5.1)**.** *Given a recognizable forest $L$, the characteristic series $\mathbb{1}_L$ of $L$, defined by $\mathbb{1}_L(t) = 1$ if $t \in L$ and $0$ otherwise, is a rational tree series whose size is polynomial in the number of states of the minimal deterministic tree automaton (DTA) recognizing $L$.*

*The Hadamard product of two rational tree series is a rational tree series, whose size is in $\mathcal{O}(s_1 s_2)$ (where $s_1$ and $s_2$ are the sizes of the two series).*

**MCMC Inference and Metropolis-Hastings.** For an introduction to the Metropolis-Hastings algorithm (MH), see Chib and Greenberg (1995). A *discrete time Markov chain* is a sequence of random variables $X_1, X_2, X_3, \cdots$ taking their values in a (discrete or) measurable state space $\mathcal{X}$, which has the Markov property: $\mathbb{P}(X_{n+1} \in A | X_1 = x_1, \cdots, X_n = x_n) = \mathbb{P}(X_{n+1} \in A | X_n = x_n) := P(x_n, A)$ for all measurable sets $A \subseteq \mathcal{X}$. The function $P$ is called the *transition kernel* of the Markov chain. For $m > 1$, we denote by $P^m$ the $m$-iterated transition kernel defined by

$$P^m(x, A) = \int_{\mathcal{X}} P(x, \mathrm{d}y) P^{m-1}(y, A)$$

A probability distribution $\pi$ with density function $f_\pi$ is called the *stationary distribution* for $P$ if $\int_A f_\pi(x)\,\mathrm{d}x = \int_{\mathcal{X}} f_\pi(x) P(x, A)\,\mathrm{d}x$ for all measurable sets $A$, which we write $\pi = \pi P$. If $\pi$ is the unique stationary distribution of a Markov chain, then the distribution of the samples generated by this chain converges to $\pi$. Suppose that $P$ has a density function $f_P(x, y)$ (i.e. $P(x, A) = \int_A f_P(x, y)\,\mathrm{d}y$), then a sufficient condition for $\pi$ to be the stationary distribution for $P$ is the *detailed balance equation*: $f_\pi(x) f_P(x, y) = f_\pi(y) f_P(y, x)$ for all $x, y \in \mathcal{X}$.

Markov chain Monte Carlo (MCMC) methods are used to solve integration or optimization problems involving a density function $\pi$ on some state space $\mathcal{X}$, which can be unnormalized but satisfies $0 < \int_{\mathcal{X}} \pi < \infty$. MCMC techniques allow one to simulate random variables $X_1, X_2, \cdots, X_N$ drawn from the normalized density $\hat{\pi}(\cdot)$, which can be used to estimate expectation problems of the form $\mathbb{E}_{\hat{\pi}}[f(x)] = \int_{\mathcal{X}} f(x)\hat{\pi}(x)\,\mathrm{d}x \simeq \frac{1}{N}\sum_{i=1}^{N} f(X_i)$ and inference problems of the form $\arg\max_{x \in \mathcal{X}} \hat{\pi}(x) \simeq \arg\max_{X_1, \cdots, X_N} \pi(X_i)$.

A popular MCMC algorithm is the Metropolis-Hastings (MH) algorithm. Let $\pi(x) = p(x)/K$ be a density function on a state space $\mathcal{X}$, where $K = \int_{\mathcal{X}} p(x)\,\mathrm{d}x$ is the unknown normalizing constant. The MH algorithm makes it possible to draw samples from $\pi$, provided that we can evaluate $p$ at any point $x \in \mathcal{X}$. First we choose an arbitrary probability density function $q(\cdot, \cdot)$ on $\mathcal{X} \times \mathcal{X}$ from which samples can easily be drawn. Then, given the current state of the chain $x_n$, we draw a candidate $x^* \sim q(x_n, \cdot)$ and accept it as the next state of the chain with probability

$$\alpha(x_n, x^*) = \min\left\{1, \frac{p(x^*) q(x^*, x_n)}{p(x_n) q(x_n, x^*)}\right\}$$

One can show that the transition kernel of the MH algorithm is

$$P_{MH}(x, A) = \int_A q(x, y)\alpha(x, y)\,\mathrm{d}y + \mathbb{1}_A(x)\left(1 - \int_{\mathcal{X}} q(x, y)\alpha(x, y)\,\mathrm{d}y\right) \tag{1}$$

(where $\mathbb{1}_A$ denotes the characteristic function of the set $A$) and that it satisfies the detailed balance equation for the distribution $\pi$.

**Linear Programming, Convex Sets and Notations.** See Luenberger (2003) for references on linear programming. A *linear program* (LP) is an optimization problem than can be expressed in the following *standard form*:

$$\begin{aligned} &\text{maximize } \mathbf{c}^T \mathbf{x} \\ &\text{subject to } \mathbf{A}\mathbf{x} = \mathbf{b} \text{ and } \mathbf{x} \geq \mathbf{0} \end{aligned} \tag{2}$$

where $\mathbf{x}, \mathbf{c} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$. A vector $\mathbf{x}$ satisfying the constraints in (2) is a *feasible solution*. If at most $m$ of its entries are non-zero it is a *basic feasible solution*, and if it achieves the maximum value of the objective function ($\mathbf{c}^T\mathbf{x}$) it is an *optimal feasible solution*.

**Theorem 2.2** (Fundamental Theorem of Linear Programming)**.** *Given a linear program in standard form (2) where* $\mathbf{A}$ *is an* $m \times n$ *matrix of rank* $m$*,*

*i) if there exists a feasible solution, there exists a basic feasible solution;*

*ii) if there exists an optimal feasible solution, there exists an optimal basic feasible solution.*

Let $V$ be a vector space. For any set $A \subseteq V$, the *convex hull* of $A$ is the smallest convex set of $V$ containing $A$, we denote it $conv(A)$. A *$k$-simplex* is a $k$-dimensional polytope which is the convex hull of its $k+1$ vertices.

For any subset $A$ of a topological space, we denote its closure by $\overline{A}$, its interior by $\mathring{A}$, and its boundary by $\partial A = \overline{A} \setminus \mathring{A}$. For any subset $B = \{\mathbf{v_1}, \cdots, \mathbf{v_n}\}$ of a vector space, we denote by $span(B)$ the vector space spanned by $B$.

## 3. Complexity Study

In this section, we study the complexity of the maximization problem. We want to find a tree in $T_{\mathcal{F}}$ which maximizes a non-negative tree series $\phi : T_F \to \mathbb{R}$. We will show that this problem is undecidable when $\phi$ is rational, and that it is NP-hard even if the support of $\phi$ is finite. We now give a formal definition of the Max-RTS problem and of the other problems we will use to show this result.

**Definition 3.1** (Max-RTS)**.**
*Instance: A non-negative rational tree series $\phi$ and a rational number $\gamma$.*
*Question: Is there a tree $t \in supp(\phi)$ such that $\phi(t) \geq \gamma$ (resp. $>$, $<$, $\leq$)?*

**Definition 3.2** (Max-APA)**.** *An* acceptor probabilistic automaton *(acceptor PA) $A$ of size $n$ over an alphabet $\Sigma$ is a tuple $A = \langle (\mathbf{T}_\sigma)_{\sigma \in \Sigma}, \boldsymbol{\pi}, \boldsymbol{\eta} \rangle$, where for each $\sigma \in \Sigma$, $\mathbf{T}_\sigma \in \mathbb{R}^{n \times n}$ is a row-stochastic transition matrix, and where $\boldsymbol{\pi}, \boldsymbol{\eta} \in \mathbb{R}^n$ are column vectors with only one non-zero entry which is equal to one. An acceptor PA assigns an acceptance probability to each word $w = \sigma_1 \cdots \sigma_m \in \Sigma^*$, given by $P_A(w) = \boldsymbol{\pi}^T \mathbf{T}_w \boldsymbol{\eta}$ where $\mathbf{T}_w = \mathbf{T}_{\sigma_1} \cdots \mathbf{T}_{\sigma_m}$.*

*The Max-APA problem is the following:*
*Instance: An acceptor PA $A$ and a rational number $\gamma$.*
*Question: Is there a word $w \in \Sigma^*$ such that $P_A(w) \geq \gamma$ (resp. $>$, $<$, $\leq$)?*

**Definition 3.3** (3-SAT)**.**
*Instance: A formula $\varphi = \bigwedge_{i=1}^{l} (l_1^i \vee l_2^i \vee l_3^i)$ in conjunctive normal form, such that each clause has 3 literals.*
*Question: Is there a satisfying assignment for $\varphi$?*

It is well known that 3-SAT is NP-complete, and it has been proven in Paz (1971); Blondel and Canterini (2003) that Max-APA is undecidable.

**Theorem 3.1.** *The Max-RTS problem is undecidable*

*Proof.* We reduce the Max-APA problem to Max-RTS. Let $A = \langle (\mathbf{T}_\sigma)_{\sigma\in\Sigma}, \boldsymbol{\pi}, \boldsymbol{\eta}\rangle$ be an acceptor PA of size $n$. Let $\mathcal{F} = \{\diamond\} \cup \{\tilde{\sigma}(\cdot)|\sigma \in \Sigma\}$ be a ranked alphabet. Let $\phi$ be the rational series on $T_\mathcal{F}$ with linear representation $(\mathbb{R}^n, \mu, \lambda)$, where $\lambda$ is the linear form defined by $\lambda(\mathbf{v}) = \boldsymbol{\pi}^T\mathbf{v}$ for all $\mathbf{v} \in \mathbb{R}^n$, and where $\mu$ is defined by $\mu(\diamond) = \boldsymbol{\eta}$ and $\mu(\tilde{\sigma})(\mathbf{v}) = \mathbf{T}_\sigma\mathbf{v}$ for each $\sigma \in \Sigma$.

With any word $w = w_1 \cdots w_m \in \Sigma^*$ we associate the tree $t_w = \widetilde{w}_1(\widetilde{w}_2(\cdots \widetilde{w}_m(\diamond))) \in T_\mathcal{F}$. Check by induction that $\mu(t_w) = \mathbf{T_w}\boldsymbol{\eta}$ for all $w \in \Sigma^*$. It follows that $\lambda(\mu(t_w)) = \boldsymbol{\pi}^T\mu(t_w) = P_A(w)$, and since every tree over $\mathcal{F}$ is of the form $t_w$ for a word $w$, we have $P_A(w) \geq \gamma$ if and only if $\phi(t_w) \geq \gamma$. The proof is similar for the three other inequalities.

□

**Theorem 3.2.** *The Max-RTS problem, with the added constraint that the support of $\phi$ must be finite, is NP-hard.*

*Proof.* We reduce 3-SAT to Max-RTS. Let $\varphi = \bigwedge_{i=1}^{l} C_i = \bigwedge_{i=1}^{l}(l_1^i \vee l_2^i \vee l_3^i)$ be an instance of the 3-SAT problem with variables in $Var(\varphi)$. We consider the ranked alphabet $\mathcal{F} = \{1, 0, \diamond, f(\cdot,\cdot)\} \cup \{C_i(\cdot,\cdot,\cdot)\}_{i=1}^{l}$. For $i = 1 \cdots l$, we define a forest $T_i \subseteq T_\mathcal{F}$ containing the trees of the form $C_i(b_1, b_2, b_3)$ where each $b_i \in \{0, 1\}$ and at least one of them is 1, and

$$T_\varphi := \{f(t_1, f(t_2, f(\cdots, f(t_l, \diamond))))|\ \forall i : t_i \in T_i\}.$$

We interpret the labels of the leaves of a subtree $t_i = C_i(b_1, b_2, b_3)$ as truth values assigned to the corresponding literals. Each tree in $T_\varphi$ puts at least one literal in every clause to true, but not every tree defines a valid assignment (since assignments can be contradictory); and for each valid assignment, there exists a tree in $T_\varphi$ from which it can be deduced.

We define the rational series $\phi$ by its linear representation $(V, \mu, \lambda)$:

- $V$ is a vector space with basis $\{\mathbf{e}_\top, \mathbf{e}_\bot, \mathbf{e}_\diamond\} \cup \{\mathbf{e_x}, \mathbf{e}_{\neg\mathbf{x}} \mid x \in Var(\varphi)\}$
- $\mu$ is defined by $\mu(1) = \mathbf{e}_\top$, $\mu(0) = \mathbf{e}_\bot$, $\mu(\diamond) = \mathbf{e}_\diamond$, and for any basis vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$

$$\mu(f)(\mathbf{e_1}, \mathbf{e_2}) = \begin{cases} 2\mathbf{e_1} & \text{if } \mathbf{e_2} = \mathbf{e}_\diamond \\ 2\mathbf{e}_\bot & \text{if } \mathbf{e_1} = \mathbf{e_l} \text{ and } \mathbf{e_2} = \mathbf{e}_{\neg\mathbf{l}} \text{ for a literal } l \text{ , and} \\ \mathbf{e_1} + \mathbf{e_2} & \text{otherwise} \end{cases}$$

$$\mu(C)(\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3}) = \begin{cases} \displaystyle\sum_{i:\mathbf{e_i}=\mathbf{e}_\top} \mathbf{e_{l_i}} + \sum_{i:\mathbf{e_i}=\mathbf{e}_\bot} \mathbf{e}_{\neg\mathbf{l_i}} & \text{if } \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\} \cap \{\mathbf{e}_\top, \mathbf{e}_\bot\} \neq \emptyset \\ \mathbf{0} & \text{otherwise} \end{cases}$$

  for each clause $C = l_1 \vee l_2 \vee l_3$ (where it is understood that $\mathbf{e}_{\neg\neg\mathbf{x}} = \mathbf{e_x}$ for all $x \in Var(\varphi)$).
- $\lambda$ is defined on the basis vectors by $\lambda(\mathbf{e}) = 0$ if $\mathbf{e} = \mathbf{e}_\bot$ and 1 otherwise

For a tree $t \in T_\varphi$, each $\mathbf{v_i} = \mu(C_i(b_1, b_2, b_3))$ is the sum of three basis vectors, each one of them corresponding to a literal which should be set to true by the assignment (e.g. if $C = x \vee \neg y \vee z$, then $\mu(C(0,0,1)) = \mathbf{e_{\neg x}} + \mathbf{e_y} + \mathbf{e_z}$). Thus, developing $\mu(t) = \mu(f)(\mathbf{v_1}, \mu(f)(\cdots, \mu(f)(\mathbf{v_l}, \mathbf{e}_\diamond)))$ by multilinearity, we have that $\mu(t)$ is the sum of $6^l$ basis vectors, all of them being different from $\mathbf{e}_\perp$ if and only if there is no contradiction in the assignment induced by this tree.

Consider the tree series $\phi' = \phi \odot \mathbb{1}_{T_\varphi}$ (which is rational by Theorem 2.1, with finite support). It follows from the construction of $\phi$ that there exists a tree $t \in supp(\phi')$ such that $\phi'(t) \geq 6^l$ (resp. $\phi'(t) > 6^l - 1$) if and only if there exists an assignment satisfying $\varphi$. Since the forest $T_\varphi$ is recognizable by a DTA with $3(l+1)$ states, the size of the representation of $\phi'$ is polynomial in the size of the 3-SAT problem (and so is the encoding of $\gamma$). This construction can thus be carried out in polynomial time, hence the NP-hardness of the Max-RTS problem with finite support for the first two inequalities.

To prove the result for the remaining inequalities, define $\lambda$ by $\lambda(\mathbf{e}) = 1$ if $\mathbf{e} = \mathbf{e}_\perp$ and 0 otherwise, and the series $\phi_{+1} : t \mapsto \phi(t)+1$. Consider the rational tree series $\phi' = \phi_{+1} \odot \mathbb{1}_{T_\varphi}$, and check that there exists a tree $t \in supp(\phi')$ such that $\phi'(t) \leq 1$ (resp. $\phi'(t) < 2$) if and only if there exists an assignment satisfying $\varphi$. □

We now state two other complexity results of problems related to the representation space (we omit the proofs of these results due to lack of space):

- Let $(V, \mu, \lambda)$ be a representation of a rational tree series $\phi$. Given a point $\mathbf{x}$ in $V$, can we find a tree $t \in supp(\phi)$ such that its projection $\mu(t)$ in $V$ is in a small ball around $\mathbf{x}$? One can show that Max-RTS is Turing-reducible to this problem in polynomial time, which implies that it is undecidable, and that it is NP-hard when the support of $\phi$ is finite.

- Given a linear representation $(V, \mu)$ of $T_\mathcal{F}$, can we decide whether the application $\mu : T_\mathcal{F} \to V$ is injective? A straightforward reduction from the problem of the freeness of matrix semi-groups for $3 \times 3$ matrices with non-negative integer Klarner et al. (1991) shows that this problem is undecidable.

## 4. MCMC Inference in the Representation Space

Let $\phi$ be the non-negative tree series to maximize. We know from the previous section that finding a tree maximizing $\phi$ is a difficult problem. In this section, we present two methods to get an estimate of such a tree. First, we implement the Metropolis-Hastings algorithm directly in the space of trees (this is in some way a reformulation of the algorithm proposed in Talton et al. (2011) in our setting). Then, we propose a method to solve this problem in a representation space: given a linear representation $(V, \mu)$ of $T_\mathcal{F}$, we propose a continuous extension $\widetilde{\phi}$ of $\phi$ to $V$ and an implementation of the MH algorithm targeting $\widetilde{\phi}$ in the representation space $V$. We end this section by showing the convergence of the proposed algorithm.

### 4.1. Metropolis-Hastings in $T_\mathcal{F}$

A first method to get an estimate of a tree maximizing the series $\phi$ is to implement the MH algorithm directly in the space of trees, thus constructing a Markov chain in $T_\mathcal{F}$ whose

stationary distribution is proportional to $\phi$. We only need to define the jump probability from the current state of the chain $t \in T_{\mathcal{F}}$.

Let $\pi$ be a stochastic tree series, and for each $t \in T_{\mathcal{F}}$, let $q_t$ be a probability distribution on the contexts $C_{\mathcal{F}}(t)$ such that $q_t(c) > 0$ for all $c \in C_{\mathcal{F}}(t)$. Let $t$ be the current state of the chain. To generate a new candidate, we first draw a context $c \in C_{\mathcal{F}}(t)$ from $q_t(\cdot)$. We then draw a new subtree $\tau'$ from $c^{-1}\pi(\cdot)$ and set $t^* = c[\tau']$. The jump probability from the tree $t$ to $t^*$ is $q_t(c) \cdot [c^{-1}\pi](\tau') \propto q_t(c)\pi(t^*)$, which leads to the following acceptance probability

$$\alpha(t, t^*) = \min\left\{1, \frac{\phi(t^*)q_{t^*}(c)\pi(t)}{\phi(t)q_t(c)\pi(t^*)}\right\}$$

We can then run the chain in $T_{\mathcal{F}}$ with the usual acceptance-rejection method while keeping track of the tree maximizing $\phi$.

In the particular case were $\phi(\cdot) = p(\cdot)L(\cdot)$ is an unnormalized posterior distribution ($p$ is the prior and $L$ the likelihood), a possible choice for $\pi$ is the prior $p$, which simplifies the acceptance probability to $\alpha(t, t^*) = \min\left\{1, \frac{q_{t^*}(c)L(t)}{q_t(c)L(t^*)}\right\}$. In this context, the algorithm is similar to the one proposed in Talton et al. (2011).

### 4.2. Extending $\phi$ to the Representation Space

Let $\phi$ be the non-negative tree series to maximize over $T_{\mathcal{F}}$, and $(V, \mu)$ be a linear representation of $T_{\mathcal{F}}$. We assume that $\phi$ is *bounded*, that its sum over $T_{\mathcal{F}}$ is *finite* ($\sum_{t \in T_{\mathcal{F}}} \phi(t) < \infty$) and that $\mu(T_{\mathcal{F}})$ is *bounded in* $V$ (the first three assumptions are satisfied by any stochastic tree series, and the linear representation induced by a rational stochastic tree series satisfies the last one)[1].

We want to define a non-negative function $\widetilde{\phi} : \mathcal{X} \to \mathbb{R}$ where $\mu(T_{\mathcal{F}}) \subseteq \mathcal{X} \subseteq V$, which extends $\phi$ to $V$. This function $\widetilde{\phi}$ should be such that one of its maximizing points is $\mu(\hat{t})$ where $\hat{t}$ is a maximizing tree of $\phi$.

The function $\phi$ naturally suggests a value for points in $\mu(T_{\mathcal{F}})$, but since we cannot assume that $\mu$ is injective, we can only require that $\widetilde{\phi}(\mu(t)) = \max\{\phi(t') : \mu(t) = \mu(t'), t' \in T_{\mathcal{F}}\}$ for all $t \in T_{\mathcal{F}}$. We now need to extend $\widetilde{\phi}$ to the rest of the representation space, or at least a subset of it containing $\mu(T_{\mathcal{F}})$. A first idea is to define $\widetilde{\phi}$ on $conv(\mu(T_{\mathcal{F}}))$ as follows: for a point $\mathbf{x} \in conv(\mu(T_{\mathcal{F}}))$, consider all the convex combinations $\sum \alpha_i \mu(t_i)$ that are equal to $\mathbf{x}$ and their corresponding scores $\sum_i \alpha_i \phi(t_i)$, and set $\widetilde{\phi}(\mathbf{x})$ equal to the best score. However, in order to use approximation techniques like the Metropolis-Hastings algorithm in the representation space, we need $\widetilde{\phi}$ to be continuous on its domain $\mathcal{X}$ (which we would like to be compact and closed), and we cannot ensure these properties with this definition. In this section, we first present an alternative definition of $\widetilde{\phi}$, and we then show that this function has the desired properties.

Let $D = \dim(V) + 1$ and let $\mathcal{S} = \{\mathbf{s_1}, \cdots, \mathbf{s_D}\} \subseteq V$ be the set of vertices of a $(D-1)$-simplex $\mathcal{X} = conv(\mathcal{S})$ whose interior $\mathring{\mathcal{X}}$ contains $\mu(T_{\mathcal{F}})$. Let $T_{\mathcal{S}} = \{s_1, \cdots, s_D\}$ be a set of new symbols, we extend $\phi$ and $\mu$ to $T_{\mathcal{F}} \cup T_{\mathcal{S}}$ by setting $\phi(s_i) = 0$ and $\mu(s_i) = \mathbf{s_i}$ for $1 \leq i \leq D$. We denote by $\mathcal{C}^n$ the set of all positive vectors of $\mathbb{R}^n$ with unit $\ell_1$-norm: $\mathcal{C}^n = \{\boldsymbol{\alpha} \in [0,1]^n : \sum \alpha_i = 1\}$.

1. Note that $\phi$ does not need to be rational, all the results of this section hold for any non-negative series over $T_{\mathcal{F}}$ and any linear representation of $T_{\mathcal{F}}$ that satisfy these assumptions.

**Definition 4.1.** *For all non-empty sets of trees $T \subseteq T_{\mathcal{F}}$, we define the function $\widetilde{\phi}_T : \mathcal{X} \to \mathbb{R}$ by*

$$\widetilde{\phi}_T(\mathbf{x}) = \sup_{\substack{n>0,\ \boldsymbol{\alpha}\in\mathcal{C}^n \\ t_1\cdots t_n \in T\cup T_{\mathcal{S}}}} \left\{ \sum_{i=1}^{n} \alpha_i \phi(t_i) : \mathbf{x} = \sum_{i=1}^{n} \alpha_i \mu(t_i) \right\} \tag{3}$$

*We say that a tuple $(\boldsymbol{\alpha}, \{t_1, \cdots, t_n\})$, where $\boldsymbol{\alpha} \in \mathcal{C}^n$ and $\{t_1, \cdots, t_n\} \subseteq T_{\mathcal{F}} \cup T_{\mathcal{S}}$, is* a solution (of length $n$) for $\widetilde{\phi}_T(\mathbf{x})$ *if and only if $\mathbf{x} = \sum_{i=1}^{n} \alpha_i \mu(t_i)$ and $\widetilde{\phi}_T(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \phi(t_i)$.*

Note that if $T = \{t_1, \cdots, t_n\}$ is finite, then the function $\widetilde{\phi}_T$ coincides with the objective function of the following LP problem:

$$\begin{aligned} &\text{maximize } \boldsymbol{\Phi}^T \boldsymbol{\alpha} \\ &\text{subject to } \begin{bmatrix} 1 & \dots & 1 & 1 & \cdots & 1 \\ | & \dots & | & | & \dots & | \\ \mu(s_1) & \dots & \mu(s_D) & \mu(t_1) & \dots & \mu(t_n) \\ | & \dots & | & | & \dots & | \end{bmatrix} \boldsymbol{\alpha} = \begin{bmatrix} 1 \\ | \\ | \\ \mathbf{x} \\ | \\ | \end{bmatrix} \\ &\text{and } \boldsymbol{\alpha} \geq \mathbf{0} \end{aligned} \tag{4}$$

where $\boldsymbol{\Phi} = (\phi(s_1), \cdots, \phi(s_D), \phi(t_1), \cdots, \phi(t_n))$ and the unknown $\boldsymbol{\alpha}$ are vectors in $\mathbb{R}^{D+n}$.

The extension of $\phi$ we propose is the function $\widetilde{\phi}_{T_{\mathcal{F}}}$ defined on the set $\mathcal{X}$. Intuitively, $\widetilde{\phi}_{T_{\mathcal{F}}}$ is the smallest concave function such that $\widetilde{\phi}_{T_{\mathcal{F}}}(\mu(t)) \geq \phi(t)$ for any $t \in T_{\mathcal{F}}$. We now prove that the function $\widetilde{\phi}_T$ is a well-defined function for any non-empty $T \subseteq T_{\mathcal{F}}$, continuous on $\mathcal{X}$, and that one of its maximizing points coincides with $\mu(\hat{t})$ for some tree $\hat{t}$ such that $\phi(\hat{t}) = \max_{t\in T} \phi(t)$. To make the notations less cluttered, we denote in this section by $\widetilde{\phi}$ the function $\widetilde{\phi}_T$ for an arbitrary non-empty set of trees $T \subseteq T_{\mathcal{F}}$.

**Proposition 4.1.** *The function $\widetilde{\phi}$ is a well-defined function, and $\widetilde{\phi}(\mu(t)) \geq \phi(t)$ for all $t \in T$.*

*Proof.* The set in (3) is non-empty (any $\mathbf{x} \in \mathcal{X}$ can be expressed as a convex combination of points in $\mathcal{S}$) and has a supremum (it is bounded above by $\max_{t\in T_{\mathcal{F}}} \phi(t)$), thus $\widetilde{\phi}$ is a well-defined function. The second point is a direct consequence of the definition of $\widetilde{\phi}$. □

We now show that the supremum in (3) is always reached, and that for all $\mathbf{x} \in \mathcal{X}$ there exists a solution of length $D$ for $\widetilde{\phi}(\mathbf{x})$.

**Proposition 4.2.** *Let $\mathbf{x} \in \mathcal{X}$. If there exists a solution of length $n > D$ for $\widetilde{\phi}(\mathbf{x})$, then there exists a solution of length $D$ for $\widetilde{\phi}(\mathbf{x})$.*

*Proof.* Let $(\boldsymbol{\alpha}, A)$ be a solution of length $n > D$ for $\widetilde{\phi}(\mathbf{x})$. The vector $\boldsymbol{\alpha}$ is an optimal feasible solution of the LP problem described in (4) with $T = A = \{t_1, \cdots, t_n\}$, and the $D \times (D+n)$ matrix in this LP problem has rank $D$. It follows from Theorem 2.2 that there exists an optimal basic feasible solution from which we can extract a vector $\boldsymbol{\beta} \in \mathcal{C}^D$ and a subset $B \subseteq A$ of cardinality $D$ such that $(\boldsymbol{\beta}, B)$ is a solution of length $D$ for $\widetilde{\phi}(\mathbf{x})$. □

**Lemma 4.1.** *Let $(t_n)_n$ be a sequence in $T_{\mathcal{F}}$. If the sequence $(\mu(t_n))_n$ converges to $\mathbf{x} \notin \mu(T_{\mathcal{F}})$, then $\lim_n \phi(t_n) = 0$.*

*Proof.* Let $\varepsilon > 0$. Since $\sum_{t \in T_{\mathcal{F}}} \phi(t) < \infty$, the set $T_\varepsilon = \{t \in T_{\mathcal{F}} : \phi(t) \geq \varepsilon\}$ is finite. Moreover, since $\lim_n \mu(t_n) \notin \mu(T_{\mathcal{F}})$, each tree $t_m$ appears only a finite number of times in $(t_n)_n$. Consequently, there exists an integer $N$ such that for all $n \geq N$ we have $t_n \notin T_\varepsilon$, hence $\lim_n \phi(t_n) = 0$. □

**Theorem 4.1.** *For all $\mathbf{x} \in \mathcal{X}$, there exists a solution of length $D$ for $\widetilde{\phi}(\mathbf{x})$.*

*Proof.* This result is straightforward if the subset of trees $T$ is finite. Let $\mathbf{x} \in \mathcal{X}$ and suppose that the supremum in (3) is not reached, this implies that for each $(\boldsymbol{\alpha}, \{t_1, \cdots, t_n\}) \subseteq \mathcal{C}^n \times T$ such that $\mathbf{x} = \sum_i \alpha_i \mu(t_i)$, we can find $(\boldsymbol{\alpha'}, \{t'_1, \cdots, t'_{n'}\}) \subseteq \mathcal{C}^{n'} \times T$ such that $\mathbf{x} = \sum_i \alpha'_i \mu(t'_i)$ and $\sum_i \alpha_i \phi(t_i) < \sum_i \alpha'_i \phi(t'_i)$. Thus there exists a sequence $(\boldsymbol{\alpha^n})_n$ in $\mathcal{C}^D$ and sequences $(t_i^n)_n$ in $T \cup T_{\mathcal{S}}$ for $i = 1 \cdots D$ such that

$$\sum_{i=1}^{D} \alpha_i^n \mu(t_i^n) = \mathbf{x} \text{ for all } n, \text{ and } \lim_n \sum_{i=1}^{D} \alpha_i^n \phi(t_i^n) = \widetilde{\phi}(\mathbf{x}).$$

Since $\mathcal{C}^D$ and $\mathcal{X}$ are compact, we can extract a subsequence $(\boldsymbol{\alpha^{\sigma(n)}})_n$ converging to $\boldsymbol{\beta} \in \mathcal{C}^D$, and subsequences $(\mu(t_i^{\sigma(n)}))_n$ converging to $\mathbf{x_i} \in \mathcal{X}$ for $i = 1 \cdots D$. Let $I$ be the set of indices $i$ such that $\mathbf{x_i} = \mu(t_i)$ for a tree $t_i \in T \cup T_{\mathcal{S}}$, and $J$ be the set of remaining indices. For all $j \in J$, it follows from Lemma 4.1 that $\lim_n \phi(t_j^n) = 0$, hence $\widetilde{\phi}(\mathbf{x}) = \lim_n \sum_{i=1}^D \alpha_i^n \phi(t_i^n) = \sum_{i \in I} \beta_i \phi(t_i)$. For each $j \in J$ we have $\mathbf{x_j} \in conv(\mu(T_{\mathcal{S}}))$, thus there exists $\boldsymbol{\gamma^j} \in \mathcal{C}^D$ such that $\mathbf{x_j} = \sum_{k=1}^D \gamma_k^j \mu(s_k)$. We then have $\mathbf{x} = \sum_{i \in I} \beta_i \mu(t_i) + \sum_{j \in J} \beta_j \sum_{k=1}^D \gamma_k^j \mu(s_k)$, and since $\phi(s_k) = 0$ for all $s_k \in T_{\mathcal{S}}$, $\widetilde{\phi}(\mathbf{x}) = \sum_{i \in I} \beta_i \phi(t_i) + \sum_{j \in J} \beta_j \sum_{k=1}^D \gamma_k^j \phi(s_k)$. We can then reduce this solution of length $(|I| + D)$ for $\widetilde{\phi}(\mathbf{x})$ to one of length $D$ using Proposition 4.2. □

**Corollary 4.1.** *The function $\widetilde{\phi}$ is concave.*

*Proof.* Let $n > 0$, $\boldsymbol{\alpha} \in \mathcal{C}^n$ and $\mathbf{x}, \mathbf{x_1}, \cdots, \mathbf{x_n} \in \mathcal{X}$ such that $\mathbf{x} = \sum_{i=1}^n \alpha_i \mathbf{x_i}$. For each $1 \leq i \leq n$, let $(\boldsymbol{\gamma^i}, \{t_1^i, \cdots, t_D^i\})$ be a solution of length $D$ for $\widetilde{\phi}(\mathbf{x_i})$. We have $\mathbf{x} = \sum_{i=1}^n \alpha_i \mathbf{x_i} = \sum_{i=1}^n \alpha_i \sum_{j=1}^D \gamma_j^i \mu(t_j^i)$, hence $\widetilde{\phi}(\mathbf{x}) \geq \sum_{i=1}^n \alpha_i \sum_{j=1}^D \gamma_j^i \phi(t_j^i) = \sum_{i=1}^n \alpha_i \widetilde{\phi}(\mathbf{x_i})$. □

**Corollary 4.2.** *Let $T \subseteq T_{\mathcal{F}}$ and $\hat{t}$ be a tree in $T$. If $\hat{t} \in \arg\max_{t \in T} \phi(t)$ then $\mu(\hat{t}) \in \arg\max_{\mathbf{x} \in \mathcal{X}} \widetilde{\phi}(\mathbf{x})$.*

*Proof.* Let $\hat{t} \in \arg\max_{t \in T} \phi(t)$. For any $\mathbf{x} \in \mathcal{X}$, let $(\boldsymbol{\alpha}, \{t_1, \cdots, t_D\})$ be a solution of length $D$ for $\widetilde{\phi}(\mathbf{x})$, we have $\widetilde{\phi}(\mathbf{x}) = \sum_{i=1}^D \alpha_i \phi(t_i) \leq \phi(\hat{t}) \leq \widetilde{\phi}(\mu(\hat{t}))$, thus $\mu(\hat{t}) \in \arg\max_{\mathbf{x} \in \mathcal{X}} \widetilde{\phi}(\mathbf{x})$. □

We now study the continuity of $\widetilde{\phi}$ on $\mathcal{X}$.

**Theorem 4.2.** *The function $\widetilde{\phi}$ is continuous on $\mathcal{X}$.*

*Proof.* We first prove the continuity at a point $\mathbf{x_0} \in \mathring{\mathcal{X}}$ in the interior of $\mathcal{X}$. Let $(\mathbf{x_n})_n$ be a sequence in $\mathring{\mathcal{X}}$ converging to $\mathbf{x_0}$. For each $n$, we can express $\mathbf{x_n}$ (resp. $\mathbf{x_0}$) as a convex combination of a point on the boundary of $\mathcal{X}$ and $\mathbf{x_0}$ (resp. $\mathbf{x_n}$), i.e. there exist $\alpha_n, \beta_n \in ]0,1]$ and $\mathbf{v_n}, \mathbf{w_n} \in \partial\mathcal{X}$ such that

$$\mathbf{x_n} = \alpha_n \mathbf{x_0} + (1-\alpha_n)\mathbf{v_n} \quad \text{and} \quad \mathbf{x_0} = \beta_n \mathbf{x_n} + (1-\beta_n)\mathbf{w_n} \tag{5}$$

By concavity of $\widetilde{\phi}$ and since $\widetilde{\phi}(\mathbf{v_n}) = \widetilde{\phi}(\mathbf{w_n}) = 0$ for all $n$, we can deduce

$$\alpha_n \widetilde{\phi}(\mathbf{x_0}) \leq \widetilde{\phi}(\mathbf{x_n}) \leq \frac{\widetilde{\phi}(\mathbf{x_0})}{\beta_n} \tag{6}$$

On the other hand, by going to the limit in (5) we have $(1 - \lim_n \alpha_n)(\mathbf{x_0} - \lim_n \mathbf{v_n}) = (1-\lim_n \beta_n)(\mathbf{x_0} - \lim_n \mathbf{w_n}) = \mathbf{0}$ and since $\lim_n \mathbf{v_n} \in \partial\mathcal{X}$ (resp. $\lim_n \mathbf{w_n} \in \partial\mathcal{X}$) and $\mathbf{x_0} \in \mathring{\mathcal{X}}$, we have $\mathbf{x_0} - \lim_n \mathbf{v_n} \neq \mathbf{0}$ (resp. $\mathbf{x_0} - \lim_n \mathbf{w_n} \neq \mathbf{0}$), hence $\lim_n \alpha_n = \lim_n \beta_n = 1$ and it follows from (6) that $\lim_n \widetilde{\phi}(\mathbf{x_n}) = \widetilde{\phi}(\mathbf{x_0})$.

We now consider the case where $\mathbf{x_0} \in \partial\mathcal{X}$ is on the boundary of $\mathcal{X}$. Assume for convenience that $\mathbf{x_0} \in conv(\mu(s_1, \cdots, s_{D-1})) = S$ and let $\mathcal{H} = span(\mu(s_1) - \mu(s_2), \cdots, \mu(s_1) - \mu(s_{D-1}))$ be the hyperplane parallel to $S$. Let $(\mathbf{x_n})_n$ be a sequence in $\mathring{\mathcal{X}}$ converging to $\mathbf{x_0}$, and assume that the solution for each $\widetilde{\phi}(\mathbf{x_n})$ can be written as

$$\mathbf{x_n} = \alpha_n \mathbf{v_n} + (1-\alpha_n)\mathbf{w_n} \quad \text{and} \quad \widetilde{\phi}(\mathbf{x_n}) = \alpha_n \widetilde{\phi}(\mathbf{v_n}) + (1-\alpha_n)\widetilde{\phi}(\mathbf{w_n}) \tag{7}$$

where $\mathbf{v_n} \in conv(\mu(T_{\mathcal{F}}))$ and $\mathbf{w_n} \in S$ (this is true as soon as $\mathbf{x_n}$ gets close enough to $\mathbf{x_0}$). We can then decompose each $\mathbf{v_n}$ as $\mathbf{v_n} = \mathbf{v_n^S} + \mathbf{v_n^\perp}$ where $\mathbf{v_n^S} \in S$ and $\mathbf{v_n^\perp} \in \mathcal{H}^\perp$, and it follows from (7) that

$$\mathbf{x_n} - \mathbf{w_n} = \alpha_n(\mathbf{v_n^S} - \mathbf{w_n}) + \alpha_n \mathbf{v_n^\perp} \tag{8}$$

Since the closed set $conv(\mu(T_{\mathcal{F}}))$ is a subset of the open set $\mathring{\mathcal{X}}$, we have

$$\|\mathbf{v_n^\perp}\| \geq \min\{\|\mathbf{x_o} - \mathbf{x_\partial}\| : \mathbf{x_o} \in conv(\mu(T_{\mathcal{F}})), \mathbf{x_\partial} \in \partial\mathcal{X}\} > 0$$

for all $n$, hence $\lim_n \|\mathbf{v_n^\perp}\| > 0$. Since $\mathbf{v_n^S} - \mathbf{w_n} \in \mathcal{H}$, by taking the scalar product with $\mathbf{v_n^\perp}$ in (8), we obtain $\langle \mathbf{x_n} - \mathbf{w_n}, \mathbf{v_n^\perp} \rangle = \alpha_n \|\mathbf{v_n^\perp}\|^2$. The left-hand side of this last equality converges to 0 as $n$ grows to infinity (because $\mathbf{x_0} - \mathbf{w_n} \in \mathcal{H}$), which implies $\lim_n \alpha_n = 0$ (because $\lim_n \|\mathbf{v_n^\perp}\| > 0$), hence $\lim_n \widetilde{\phi}(\mathbf{x_n}) = \lim_n \widetilde{\phi}(\mathbf{w_n}) = 0 = \widetilde{\phi}(\mathbf{x_0})$.

□

### 4.3. Metropolis-Hastings in the Representation Space

Let $\pi$ be a stochastic tree series whose support is the whole set of trees $T_{\mathcal{F}}$ (i.e. $\pi(t) > 0$ for all $t \in T_{\mathcal{F}}$). For each $t \in T_{\mathcal{F}}$, let $q_t$ be a probability distribution on the contexts $C_{\mathcal{F}}(t)$ such that $q_t(c) > 0$ for all $c \in C_{\mathcal{F}}(t)$. Let $(V, \mu)$ be a linear representation of $T_{\mathcal{F}}$.

Instead of using an MCMC algorithm directly in the space of trees $T_{\mathcal{F}}$, we will use $\widetilde{\phi}$ to work in the underlying continuous space $V$. Since the function $\widetilde{\phi}_T$ is continuous and non-negative on the compact space $\mathcal{X}$ for any $T \subseteq T_{\mathcal{F}}$, we can define its normalized counterpart $\hat{\phi}_T(\cdot) = \widetilde{\phi}_T(\cdot) / \int_{\mathcal{X}} \widetilde{\phi}_T(x)\,\mathrm{d}x$. Our algorithm develops a Markov chain in $V$ targeting the distribution $\hat{\phi}_{T_{\mathcal{F}}}$: while evolving in this space, we construct successive sets

of trees $T_1 \subseteq T_2 \subseteq \cdots$; at each step $n$, we use the traditional MH acceptance probability targeting the distribution $\hat{\phi}_{T_n}$, where the jump distribution $q(\cdot,\cdot)$ can be any symmetric density function satisfying $\inf_{\mathbf{x},\mathbf{y}\in\mathcal{X}} q(\mathbf{x},\mathbf{y}) > 0$ (e.g. $q(\mathbf{x},\cdot)$ is the multivariate normal distribution truncated to $\mathcal{X}$, with mean $\mathbf{x}$ and variance $\sigma^2\mathbf{I}$).

Let $\mathbf{x} \in \mathcal{X}$ and $T \subseteq T_{\mathcal{F}}$, for each solution $(\boldsymbol{\alpha}, R)$ for $\widetilde{\phi}_T(\mathbf{x})$ we define the probability distribution $p_{\boldsymbol{\alpha}}$ on $R = \{t_1, \cdots, t_n\}$ by $p_{\boldsymbol{\alpha}}(t_i) = 0$ if $t_i \in T_{\mathcal{S}}$ and $p_{\boldsymbol{\alpha}}(t_i) = \alpha_i / \sum_{t_i \in R \cap T_{\mathcal{F}}} \alpha_i$ otherwise.

**Algorithm 1** Adaptive Metropolis-Hastings in $V$

**Input:** $\mathbf{x_n} \in \mathcal{X}$, $T_n = \{t_1, \cdots, t_n\} \subseteq T_{\mathcal{F}}$
**Output:** $\mathbf{x_{n+1}} \in \mathcal{X}$, $T_{n+1} \subseteq T_{\mathcal{F}}$
1: Draw $\mathbf{x}^* \sim q(\mathbf{x_n}, \cdot)$
2: Let $(\boldsymbol{\alpha}, R)$ be a solution for $\widetilde{\phi}_{T_n}(\mathbf{x}^*)$
3: Draw $t \in R \sim p_{\boldsymbol{\alpha}}(\cdot)$, a context $c \in C_{\mathcal{F}}(t) \sim q_t(\cdot)$ and $\tau \in T_{\mathcal{F}} \sim c^{-1}\pi(\cdot)$
4: $t_{n+1} \leftarrow c[\tau]$, $T_{n+1} \leftarrow T_n \cup \{t_{n+1}\}$
5: Accept $\mathbf{x}^*$ (i.e. $\mathbf{x_{n+1}} \leftarrow \mathbf{x}^*$, otherwise $\mathbf{x_{n+1}} \leftarrow \mathbf{x_n}$) with probability

$$\alpha(\mathbf{x_n}, \mathbf{x}^*) = \min\left\{1, \frac{\hat{\phi}_{T_n}(\mathbf{x}^*) q(\mathbf{x}^*, \mathbf{x_n})}{\hat{\phi}_{T_n}(\mathbf{x_n}) q(\mathbf{x_n}, \mathbf{x}^*)}\right\} = \min\left\{1, \frac{\widetilde{\phi}_{T_n}(\mathbf{x}^*)}{\widetilde{\phi}_{T_n}(\mathbf{x_n})}\right\}$$

Algorithm 1 shows how to get the next state of the chain given the current one; to get an estimate of the tree maximizing $\phi$, we start with $\mathbf{x_1}$ and $T_1 = \{t_1\}$ chosen randomly, and evolve the chain in $\mathcal{X}$ while keeping track of the tree in $T_n$ maximizing $\phi$. This algorithm is close to the traditional MH algorithm, but at each step the target distribution slightly changes $(\hat{\phi}_{T_1}, \hat{\phi}_{T_2}, \cdots)$. Such Monte-Carlo algorithms are called *adaptive*, and their convergence is more tedious to assess than in the traditional case (see Roberts and Rosenthal (2007)).

### 4.4. Proof of Convergence

In this section, we prove the following theorem:

**Theorem 4.3.** *The distribution of the $X_n$'s generated by Algorithm 1 converges to $\hat{\phi}_{T_{\mathcal{F}}}$.*

For a set of trees $T \subseteq T_{\mathcal{F}}$, let $P_T$ denote the transition kernel of the Metropolis-Hastings algorithm targeting the distribution $\hat{\phi}_T$ (cf. Eq. 1) and $P_T^n(\mathbf{x},\cdot)$ denote this transition kernel starting from $\mathbf{x}$ after $n$ steps; for any $\varepsilon > 0$ and $\mathbf{x} \in \mathcal{X}$, we define the quantity

$$M_\varepsilon(\mathbf{x}, T) = \inf\{n \geq 0 : \|P_T^n(\mathbf{x},\cdot) - \hat{\phi}_T(\cdot)\|_{TV} \leq \varepsilon\}$$

where $\|P_T^n(\mathbf{x},\cdot) - \hat{\phi}_T(\cdot)\|_{TV} = \sup_{A\subseteq\mathcal{X}} |P_T^n(\mathbf{x}, A) - \hat{\phi}_T(A)|$ is the total variation distance.

Theorem 2.1 and Corollary 2.2 in Fort et al. (2012) can be stated as follows: if the conditions (A1) to (A4) below hold, then $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \hat{\phi}_{T_{\mathcal{F}}}(f)$ for any bounded continuous function $f$ (i.e. the distribution of the $X_n$'s converges to $\hat{\phi}_{T_{\mathcal{F}}}$)

(A1) For all $T \subseteq T_{\mathcal{F}}$, the probability distribution $\hat{\phi}_T$ is such that $\hat{\phi}_T P_T = \hat{\phi}_T$ (i.e. $\hat{\phi}_T$ is the stationary distribution of the Markov chain with transition kernel $P_T$).

(A2) The sequence $(\hat{\phi}_{T_n})_{n\geq 1}$ converges weakly to $\hat{\phi}_{T_{\mathcal{F}}}$ $\mathbb{P}$-a.s.

(A3) The sequence $(\sup_{x\in V} \|P_{T_n}(x,\cdot) - P_{T_{n-1}}(x,\cdot)\|_{TV})_n$ converges to zero in probability

(A4) For any $\varepsilon > 0$, the sequence $(M_\varepsilon(X_n, T_n))_n$ is bounded in probability

For any $T \subseteq T_{\mathcal{F}}$, the transition kernel $P_T$ is by construction the MH kernel targeting $\hat{\phi}_T$, so $\hat{\phi}_T$ is the stationary distribution for the Markov chain with transition kernel $P_T$ and condition (A1) is satisfied.

To prove that (A2) is satisfied, first remark that for any $\varepsilon > 0$ we can find a finite set of trees $R = \{t_1 \cdots, t_k\}$ in $T_{\mathcal{F}}$ such that $\|\hat{\phi}_R(\cdot) - \hat{\phi}_{T_{\mathcal{F}}}(\cdot)\|_{TV} \leq \varepsilon$. Since the empty context can be drawn at line 3 of Algorithm 1 with non-zero probability, any tree $t \in T_{\mathcal{F}}$ can be drawn and added to the current set of trees with non-zero probability; it follows that the probability that there exists a step $m$ where $T_m \supseteq R$ is strictly positive, hence $\mathbb{P}(\lim_{n\to\infty} \|\hat{\phi}_{T_n}(\cdot) - \hat{\phi}_{T_{\mathcal{F}}}(\cdot)\|_{TV} = 0) = 1$.

The only difference between the definition of two successive transition kernels $P_{T_{m-1}}$ and $P_{T_m}$ are the functions $\hat{\phi}_{T_{m-1}}$ and $\hat{\phi}_{T_m}$, and we have established the weak convergence of the sequence $(\hat{\phi}_{T_n})_n$. It follows that condition (A3) is a direct consequence of condition (A2).

To prove that the last condition is satisfied, we use the following result.

**Theorem 4.4** (Roberts and Rosenthal (2004), Theorem 8)**.** *Consider a Markov chain with transition kernel $P$ and stationary distribution $\pi$. Suppose there exist a positive integer $n_0$, $\varepsilon > 0$, and a probability measure $\nu(\cdot)$ on $\mathcal{X}$ such that $P^{n_0}(x, A) \geq \varepsilon \cdot \nu(A)$ for all $x \in \mathcal{X}$ and all measurable set $A \subseteq \mathcal{X}$. Then the chain is uniformly ergodic and $\|P^n(x,\cdot) - \pi(\cdot)\|_{TV} \leq (1-\varepsilon)^{\lfloor n/n_0 \rfloor}$ for all $x \in \mathcal{X}$.*

We first prove the following lemma.

**Lemma 4.2.** *For any set of trees $T$ such that $T_1 \subseteq T \subseteq T_{\mathcal{F}}$, there exist $\varepsilon > 0$ and a probability measure $\nu_T(\cdot)$ on $\mathcal{X}$ such that $P_T(\mathbf{x},\cdot) \geq \varepsilon \cdot \nu_T(\cdot)$ for all $\mathbf{x} \in \mathcal{X}$.*

*Proof.* Let $m = \min_{\mathbf{x},\mathbf{y}\in\mathcal{X}} q(\mathbf{x},\mathbf{y})$ (recall that we have $m > 0$ by the choice of $q$) and let $T$ be a set of trees such that $T_1 \subseteq T \subseteq T_{\mathcal{F}}$, we know that $\max \widetilde{\phi}_T := \max_{\mathbf{x}\in\mathcal{X}} \widetilde{\phi}_T(\mathbf{x}) < \infty$. It follows from the definition of the MH transition kernel (cf. Eq. 1) that for any measurable set $A \subseteq \mathcal{X}$ and $\mathbf{x} \in \mathcal{X}$

$$P_T(\mathbf{x}, A) \geq \int_A q(\mathbf{x},\mathbf{y}) \min\left\{1, \frac{\widetilde{\phi}_T(\mathbf{y})}{\widetilde{\phi}_T(\mathbf{x})}\right\} \mathrm{d}\mathbf{y} \geq m \int_A \frac{\widetilde{\phi}_T(\mathbf{y})}{\max \widetilde{\phi}_T} \mathrm{d}\mathbf{y}$$

We define the probability measure $\nu_T(\cdot)$ by $\nu_T(B) = \frac{1}{Z}\int_B \frac{\widetilde{\phi}_T(\mathbf{y})}{\max \widetilde{\phi}_T} \mathrm{d}\mathbf{y}$ for all measurable sets $B \subseteq \mathcal{X}$, where $Z = \int_{\mathcal{X}} \frac{\widetilde{\phi}_T(\mathbf{y})}{\max \widetilde{\phi}_T} \mathrm{d}\mathbf{y}$. We then have $P_T(\mathbf{x},\cdot) \geq \varepsilon_T \cdot \nu_T(\cdot)$, where $\varepsilon_T = m \int_{\mathcal{X}} \frac{\widetilde{\phi}_T(\mathbf{y})}{\max \widetilde{\phi}_T} \mathrm{d}\mathbf{y} > 0$. As a direct consequence of $T_1 \subseteq T \subseteq T_{\mathcal{F}}$ we have $\widetilde{\phi}_{T_1} \leq \widetilde{\phi}_T \leq$

$\widetilde{\phi}_{T_{\mathcal{F}}}$, which implies $\int_{\mathcal{X}} \widetilde{\phi}_T \geq \int_{\mathcal{X}} \widetilde{\phi}_{T_1}$ and $\max \widetilde{\phi}_T \leq \max \widetilde{\phi}_{T_{\mathcal{F}}}$. We can then deduce $\varepsilon_T \geq m \int_{\mathcal{X}} \frac{\widetilde{\phi}_{T_1}(\mathbf{y})}{\max \widetilde{\phi}_{T_{\mathcal{F}}}} \, \mathrm{d}\mathbf{y} = \varepsilon^* > 0$ hence $P_T(\mathbf{x}, \cdot) \geq \varepsilon^* \cdot \nu_T(\cdot)$ for all $\mathbf{x} \in \mathcal{X}$ and $T_1 \subseteq T \subseteq T_{\mathcal{F}}$. □

It then follows from Theorem 4.4 that $\|P_T^n(\mathbf{x}, \cdot) - \hat{\phi}_T(\cdot)\|_{TV} \leq (1 - \varepsilon^*)^n$ for all $T_1 \subseteq T \subseteq T_{\mathcal{F}}$ and $n > 1$. Hence $M_\varepsilon(\mathbf{x}, T) \leq \frac{\ln \varepsilon}{\ln(1-\varepsilon^*)}$ for all $\varepsilon > 0$, which shows that condition (A4) is satisfied and ends the proof of Theorem 4.3.

## 5. Experiments

Let $\mathcal{F} = \{f(\cdot, \cdot), a\}$ be a ranked alphabet. We consider a simple *mismatch* distance $d$ on $T_{\mathcal{F}}$ which counts the number of differences between two trees (i.e. nodes in the same position with different labels, and positions for which only one of the two trees has a node; for example $d(f(a, a), f(a, f(a, a))) = 3$). Formally, for any trees $t_1, t_2, t_3, t_4 \in T_{\mathcal{F}}$, $d(a, a) = 0$, $d(f(t_1, t_2), f(t_3, t_4)) = d(t_1, t_3) + d(t_2, t_4)$ and $d(f(t_1, t_2), a) = 1 + |t_1| + |t_2|$, where $|t|$ is the number of nodes in the tree $t$.

Given a tree $\hat{t} \in T_{\mathcal{F}}$, we define the tree series $\phi_{\hat{t}}(\cdot) = \exp\{-d(\hat{t}, \cdot)\}$. We compare the MH algorithm in the space of trees and Algorithm 1 for the task of retrieving the tree $\hat{t}$, which is equivalent to finding the tree maximizing the series $\phi_{\hat{t}}$.

Let $\mathcal{A}$ be the weighted tree automaton with two states $q_1$ and $q_2$, initial weights $\iota(q_1) = \iota(q_2) = 0.5$, and the set of rules

$$\{q_1 \xrightarrow{0.9} f(q_1, q_2), q_1 \xrightarrow{0.1} a, q_2 \xrightarrow{0.4} f(q_2, q_2), q_2 \xrightarrow{0.6} a\}.$$

For any tree $t \in T_{\mathcal{F}}$, let $q_t$ be the distribution on $C_{\mathcal{F}}(t)$ defined by the following process: (i) randomly choose an integer $h$ between 1 and the height of $t$, (ii) randomly choose a node $n$ within the nodes of depth $h$ in $t$, and (iii) replace the subtree rooted in $n$ with \$.

For three different target trees $\hat{t}$ of different sizes (generated by the automaton $\mathcal{A}$), we run the MH algorithm in $T_{\mathcal{F}}$ and Algorithm 1 to maximize the series $\phi_{\hat{t}}$ until recovery of the tree $\hat{t}$. For both algorithms, we use the distribution induced by the automaton $\mathcal{A}$ as the stochastic tree series $\pi$, and the distribution $q_t$ to draw a context in $C_{\mathcal{F}}(t)$. For Algorithm 1, we use the representation space induced by $\mathcal{A}$ and a truncated normal for the distribution $q(\cdot, \cdot)$ on this space.

The results of this experiment are shown in Figure 1, where we plot the average of $\phi_{\hat{t}}(t)$ over 500 runs for the best tree found so far as a function of the number of iterations. This experiment shows that as the task of retrieving $\hat{t}$ gets more difficult, working in the representation space leads to better performances.

## 6. Conclusion

We proposed an algorithm to solve a tree inference problem in a representation space of $T_{\mathcal{F}}$ and we showed its convergence. In this algorithm, the tree generation process is parallel to the construction of the Markov chain in the representation space. Unlike the Metropolis-Hastings algorithm in $T_{\mathcal{F}}$ which *forgets* each generated tree, this algorithm uses each one of them to learn more about the function $\widetilde{\phi}_{T_{\mathcal{F}}}$, thus indirectly learning about $\phi$. In doing so, we take advantage of the representation space by using it as a foundation on which

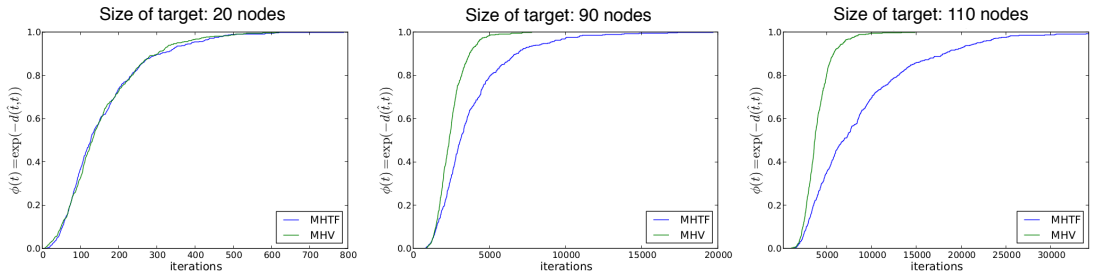


Figure 1: Comparison of the MH algorithm in the space of trees (MHTF) and the adaptive MH algorithm in the representation space (MHV) for different sizes of $\hat{t}$.

we progressively build a map, focusing around interesting regions of the space for the tree generation process.

This work could be extended in several ways. First, the choice of the linear representation is a key step of this method and hand-crafting it could be a tedious task, we intend to investigate how this linear representation could be learnt progressively: each generated tree gives us information on the discriminative power of the representation space, and we could use this information to modify it while exploring the space. Then, the canonical representation of rational tree series introduced in Denis and Habrard (2007) induces a representation space which is tightly linked to the space of contexts, we want to investigate how this space could be used in a similar fashion. Finally, we want to explore how we could work in the representation space to solve other learning problems involving trees (e.g. tree classification). We strongly believe that we can use the representation space as a powerful tool in this context.

## Acknowledgments

This work has been carried out thanks to the support of the ARCHIMEDE Labex (ANR-11-LABX- 0033) and the A*MIDEX project (ANR-11-IDEX-0001-02) funded by the "Investissements d'Avenir" French government program managed by the ANR.

## References

Jean Berstel and Christophe Reutenauer. Recognizable formal power series on trees. *Theoret. Comput. Sci.*, 18(2):115–148, 1982.

Vincent Blondel and Vincent Canterini. Undecidable problems for probabilistic automata of fixed dimension. *Theory Comput. Syst.*, 36(3):231–245, 2003.

Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings Algorithm. *American Statistician*, 49:327–335, 1995. doi: 10.1080/00031305.1995.10476177.

H. Comon, M. Dauchet, R. Gilleron, C. Löding, F. Jacquemard, D. Lugiez, S. Tison, and M. Tommasi. Tree automata techniques and applications, 2007.

François Denis and Amaury Habrard. Learning rational stochastic tree languages. In *Proceedings of the 18th international conference on Algorithmic Learning Theory*, ALT '07, 2007.

François Denis, Édouard Gilbert, Amaury Habrard, Faïssal Ouardi, and Marc Tommasi. Relevant representations for the inference of rational stochastic tree languages. In *Proceedings of the 9th international colloquium on Grammatical Inference: Algorithms and Applications*, ICGI '08, 2008.

G. Fort, E. Moulines, and P. Priouret. Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *The Annals of Statistics*, 39(6):3262–3289, 2012.

David A. Klarner, Jean-Camille Birget, and Wade Satterfield. On the undecidability of the freeness of integer matrix semigroups. *Internat. J. Algebra Comput.*, 1(2):223–226, 1991.

D.G. Luenberger. *Linear and Nonlinear Programming: Second Edition*. Springer, 2003. ISBN 9781402075933. URL http://books.google.fr/books?id=QY9BjisUT1gC.

Azaria Paz. *Introduction to Probabilistic Automata (Computer Science and Applied Mathematics)*. Academic Press, Inc., Orlando, FL, USA, 1971. ISBN 0125476507.

G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

Gareth O. Roberts and Jeffrey S. Rosenthal. Coupling and ergodicity of adaptive MCMC. *Journal of Applied Probabilities*, 44:458–475, 2007.

Jerry O. Talton, Yu Lou, Steve Lesser, Jared Duke, Radomír Mech, and Vladlen Koltun. Metropolis procedural modeling. *ACM Trans. Graph.*, 30(2), 2011.