Relevant Representations for the Inference of Rational Stochastic Tree Languages

François Denis[1] Edouard Gilbert[2] Amaury Habrard[1]
Faïssal Ouardi[1] Marc Tommasi[2]

[1]Laboratoire d'Informatique Fondamentale de Marseille (LIF)
CNRS, Aix-Marseille Université, France

[2]Laboratoire d'Informatique Fondamentale de Lille (L.I.F.L.), INRIA
and É.N.S. Cachan, France

ICGI 2008

Outline

Outline

Trees

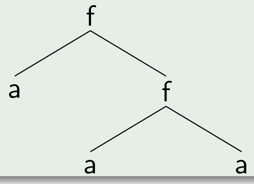
$\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1 \cup \cdots \cup \mathcal{F}_p$: a ranked alphabet

$\mathcal{F}_m$: function symbols of *arity* m.

$T(\mathcal{F})$: all the *trees* constructed from $\mathcal{F}$.

Example:

$\mathcal{F} = \{f(\cdot,\cdot), a\}$; $f(a, f(a,a)) \in T(\mathcal{F})$.



Stochastic Tree Languages

Stochastic tree language: Probability distribution over $T(\mathcal{F})$

$$p : T(\mathcal{F}) \to \mathbb{R}$$

- for any $t \in T(\mathcal{F})$, $0 \leq p(t) \leq 1$ and
- $\sum_{t \in T(\mathcal{F})} p(t) = 1$.

Formal power tree series over $T(\mathcal{F})$

$$r : T(\mathcal{F}) \to \mathbb{R}.$$

Notation: $\mathbb{R}\langle\langle T(\mathcal{F})\rangle\rangle$ (vector space).

A Basic Problem in Probabilistic Grammatical Inference

The Problem

Data $t_1, \ldots, t_n \in T(\mathcal{F})$ independently drawn according to a fixed unknown stochastic tree language p.

Goal Infer an estimate of p in some class of probabilistic models.

Probabilistic models

- Probabilistic tree automata
- Linear representations of rational tree series

Probabilistic Tree Automata

A distribution over $T(\mathcal{F})$ according to a PA with one state

$$\mathcal{A}_\alpha : \ \Delta_\alpha = \{q \xrightarrow{\alpha} a, \ \ q \xrightarrow{1-\alpha} f(q,q)\}, \ \ \tau(q) = 1, \ \ 0 \leq \alpha \leq 1$$

$$p_\alpha(f(a, f(a,a))) = \alpha^3(1-\alpha)^2$$

Less simple than in the word case

- p_α is a stochastic language iff $\alpha \geq 1/2$.
- Is it decidable whether a PA defines a stochastic language?
- The average tree size: $1/(2\alpha - 1)$. Unbounded if $\alpha = 1/2$.
- It is polynomially decidable whether a PA defines a stochastic language with bounded average size.

Linear Representations of Rational Tree Languages

A series $r \in \mathbb{R}\langle\langle T(\mathcal{F})\rangle\rangle$ is rational iff there exists a triple (V, μ, λ):

- V is a finite dimensional vector space over $\mathbb{R}$,
- μ maps any $f \in \mathcal{F}_p$ to a p-linear mapping $\mu(f) \in \mathcal{L}(V^p; V)$,
- λ is a linear form $V \to \mathbb{R}$,
- $r(t) = \lambda\mu(t)$, where $\mu(f(t_1, \ldots, t_p)) = \mu(f)(\mu(t_1), \ldots, \mu(t_p))$.

Example

- $V = \mathbb{R}$ and let $e_1 \neq 0$ a basis of $\mathbb{R}$,
- $\mu(a) = \alpha e_1$, $\mu(f)(e_1, e_1) = (1 - \alpha)e_1$,
- $\lambda(e_1) = 1$.

$$\lambda\mu(f(a, f(a, a))) = \alpha^3(1 - \alpha)^2$$

Rational Stochastic Tree Languages

Stochastic languages

A *rational stochastic tree language (RSTL)* is a stochastic language that has a linear representation.

- Every stochastic language computed by a probabilistic automaton is rational.
- Some RSTL cannot be computed by a probabilistic automaton.
- It is undecidable whether a linear representation represents a stochastic language.
- A RSTL can be equivalently represented by a weighted tree automaton, minimal in the number of states (vector space).

Outline

1. The Basic Problem
2. A Canonical Linear Representation for Rational Tree Series
3. Contributions
   - Normalization of the Model as a Generative Model
   - Strongly Consistent Model
   - Unranked Trees

Word Languages: The Notion of Residual Languages

Languages: $L \subseteq \Sigma^*, u \in \Sigma^*$

$$u^{-1}L = \{v \in \Sigma^* | uv \in L\}$$

Series: $r \in \mathbb{R}\langle\langle T(\mathcal{F})\rangle\rangle, u \in \Sigma^*$

$$\dot{u}r(v) = r(uv)$$

Residual language is a key notion for inference because:

- residual languages are intrinsic components
- they are observable on samples
- they yield canonical representations.

Contexts

$\$$: a zero arity function symbol not in $\mathcal{F}_0$.

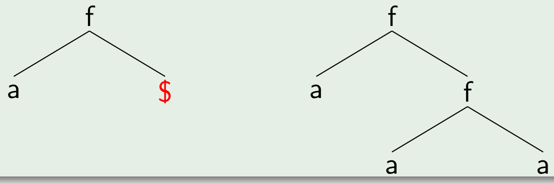
A *context* is an element of $T(\mathcal{F} \cup \{\$\})$ s.t. $\$$ appears exactly once.

$C(\mathcal{F})$: all contexts over $\mathcal{F}$.

$c[t]$: the tree obtained by substituting $\$$ by t.

Example:

$c = f(a, \$)$ $\quad\quad$ $c[f(a,a)] = f(a, f(a,a))$



An Algebraic Characterization of Rational Series

Contexts operate on tree series

Let $c \in C(\mathcal{F})$. Define $\dot{c} : \mathbb{R}\langle\langle T(\mathcal{F})\rangle\rangle \to \mathbb{R}\langle\langle T(\mathcal{F})\rangle\rangle$ by

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

Example

$c = f(a, \$), t = f(a, a), \dot{c}r(t) = r(f(a, f(a, a)))$.

Let $r \in T(\mathcal{F})$, consider $W_r = [\{\dot{c}r | c \in C(\mathcal{F})\}] \subseteq \mathbb{R}\langle\langle T(\mathcal{F})\rangle\rangle$

the vector subspace of $\mathbb{R}\langle\langle T(\mathcal{F})\rangle\rangle$ spanned by the series $\dot{c}r$.

Theorem: r is rational iff the dimension of W_r is finite.

The Canonical Linear Representation of Rational Series

$$W_r = [\{\dot{c}r | c \in C(\mathcal{F})\}] \ ; \ W_r^* \text{ dual space of } W_r$$

- No natural linear representation of r on W_r

The Canonical Linear Representation of Rational Series

$$W_r = [\{\dot{c}r | c \in C(\mathcal{F})\}] \ ; \ W_r^* \text{ dual space of } W_r$$

- No natural linear representation of r on W_r
- $T(\mathcal{F})$ is naturally embedded in W_r^*:

$$t \rightarrow \overline{t} \text{ s.t. } \overline{t}(\dot{c}r) = r(c[t])$$

The Canonical Linear Representation of Rational Series

$$W_r = [\{\dot{c}r | c \in C(\mathcal{F})\}] \; ; \; W_r^* \text{ dual space of } W_r$$

- No natural linear representation of r on W_r
- $T(\mathcal{F})$ is naturally embedded in W_r^*:

$$t \rightarrow \overline{t} \text{ s.t. } \overline{t}(\dot{c}r) = r(c[t])$$

- $\{\overline{t} | t \in T(\mathcal{F})\}$ spans W_r^*

The Canonical Linear Representation of Rational Series

$$W_r = [\{\dot{c}r | c \in C(\mathcal{F})\}] \ ; \ W_r^* \text{ dual space of } W_r$$

- No natural linear representation of r on W_r
- $T(\mathcal{F})$ is naturally embedded in W_r^*:

$$t \rightarrow \overline{t} \text{ s.t. } \overline{t}(\dot{c}r) = r(c[t])$$

- $\{\overline{t} | t \in T(\mathcal{F})\}$ spans W_r^*
- the canonical linear representation of r:
  (W_r^*, μ, λ) where $\mu(t) = \overline{t}$ and $\lambda = r$ $(W_r^{**} = W_r)$

Building the Canonical Linear Representation

$$\mathcal{F} = \{f(,), a\}, \tau(q) = 1, p_\alpha : q \xrightarrow{\alpha} a, q \xrightarrow{1-\alpha} f(q, q)$$

Building the Canonical Linear Representation

$$\mathcal{F} = \{f(,), a\}, \tau(q) = 1, p_\alpha : q \xrightarrow{\alpha} a, q \xrightarrow{1-\alpha} f(q, q)$$

Let $p = 2p_{2/3} - p_{3/4}$: $\sum_t p(t) = 1$ and $\forall t, p(t) \geq 0$.

Building the Canonical Linear Representation

Let $p = 2p_{2/3} - p_{3/4}$: $\sum_t p(t) = 1$ and $\forall t, p(t) \geq 0$.

$$p(a) = \frac{7}{12}, p(f(a,a)) = \frac{269}{1728}, p(f(a, f(a,a))) = p(f(f(a,a),a)) = \frac{9823}{248832}, \ldots$$

Building the Canonical Linear Representation

Let $p = 2p_{2/3} - p_{3/4}$: $\sum_t p(t) = 1$ and $\forall t, p(t) \geq 0$.

$$p(a) = \frac{7}{12}, p(f(a,a)) = \frac{269}{1728}, p(f(a,f(a,a))) = p(f(f(a,a),a)) = \frac{9823}{248832}, \ldots$$

Oracle: Is $\overline{a} = 0$? i.e. for every context c, $p(c[a]) = 0$?

Building the Canonical Linear Representation

Let $p = 2p_{2/3} - p_{3/4}$: $\sum_t p(t) = 1$ and $\forall t, p(t) \geq 0$.

$$p(a) = \frac{7}{12}, p(f(a,a)) = \frac{269}{1728}, p(f(a,f(a,a))) = p(f(f(a,a),a)) = \frac{9823}{248832}, \dots$$

Oracle: Is $\overline{a} = 0$? i.e. for every context c, $p(c[a]) = 0$?

Answer: NO, consider $c = \$$.

Let $B = \{\overline{a}\}$.

Building the Canonical Linear Representation

Let $p = 2p_{2/3} - p_{3/4}$: $\sum_t p(t) = 1$ and $\forall t, p(t) \geq 0$.

$$p(a) = \frac{7}{12}, p(f(a,a)) = \frac{269}{1728}, p(f(a,f(a,a))) = p(f(f(a,a),a)) = \frac{9823}{248832}, \ldots$$

Oracle: Is $\overline{f(a,a)}$ colinear to $\overline{a}$?
i.e. $\exists \alpha$, for every context c, $p(c[f(a,a)]) = \alpha p(c[a])$?

Let $B = \{\overline{a}\}$.

Building the Canonical Linear Representation

Let $p = 2p_{2/3} - p_{3/4}$: $\sum_t p(t) = 1$ and $\forall t, p(t) \geq 0$.

$$p(a) = \frac{7}{12}, p(f(a,a)) = \frac{269}{1728}, p(f(a,f(a,a))) = p(f(f(a,a),a)) = \frac{9823}{248832}, \ldots$$

Oracle: Is $\overline{f(a,a)}$ colinear to $\overline{a}$?
i.e. $\exists \alpha$, for every context c, $p(c[f(a,a)]) = \alpha p(c[a])$?

Answer: NO, consider $c_1 = \$$ and $c_2 = f(a, \$)$.

Let $B = \{\overline{a}, \overline{f(a,a)}\}$.

Building the Canonical Linear Representation

Let $p = 2p_{2/3} - p_{3/4}$: $\sum_t p(t) = 1$ and $\forall t, p(t) \geq 0$.

$$p(a) = \frac{7}{12}, p(f(a,a)) = \frac{269}{1728}, p(f(a,f(a,a))) = p(f(f(a,a),a)) = \frac{9823}{248832}, \dots$$

Oracle: Is $\overline{f(a,f(a,a))}$ colinear to $\overline{a}, \overline{f(a,a)}$?

Let $B = \{\overline{a}, \overline{f(a,a)}\}$.

Building the Canonical Linear Representation

Let $p = 2p_{2/3} - p_{3/4}$: $\sum_t p(t) = 1$ and $\forall t, p(t) \geq 0$.

$$p(a) = \frac{7}{12}, p(f(a,a)) = \frac{269}{1728}, p(f(a,f(a,a))) = p(f(f(a,a),a)) = \frac{9823}{248832}, \ldots$$

Oracle: Is $\overline{f(a,f(a,a))}$ colinear to $\overline{a}, \overline{f(a,a)}$?

Answer: YES,

$$\overline{f(a,f(a,a))} = \frac{-54}{2^4 \times 3^4}\overline{a} + \frac{59}{2^4 \times 3^2}\overline{f(a,a)}.$$

Let $B = \{\overline{a}, \overline{f(a,a)}\}$.

Building the Canonical Linear Representation

Let $p = 2p_{2/3} - p_{3/4}$: $\sum_t p(t) = 1$ and $\forall t, p(t) \geq 0$.

$$p(a) = \frac{7}{12}, p(f(a,a)) = \frac{269}{1728}, p(f(a,f(a,a))) = p(f(f(a,a),a)) = \frac{9823}{248832}, \ldots$$

Oracle: Is $\overline{f(f(a,a),a)}$ colinear to $\overline{a}, \overline{f(a,a)}$?

Let $B = \{\overline{a}, \overline{f(a,a)}\}$.

Building the Canonical Linear Representation

Let $p = 2p_{2/3} - p_{3/4}$: $\sum_t p(t) = 1$ and $\forall t, p(t) \geq 0$.

$$p(a) = \frac{7}{12}, p(f(a,a)) = \frac{269}{1728}, p(f(a,f(a,a))) = p(f(f(a,a),a)) = \frac{9823}{248832}, \ldots$$

Oracle: Is $\overline{f(f(a,a),a)}$ colinear to $\overline{a}, \overline{f(a,a)}$?

Answer: YES,

$$\overline{f(a,f(a,a))} = \frac{-54}{2^4 \times 3^4}\overline{a} + \frac{59}{2^4 \times 3^2}\overline{f(a,a)}.$$

Let $B = \{\overline{a}, \overline{f(a,a)}\}$.

Building the Canonical Linear Representation

Let $p = 2p_{2/3} - p_{3/4}$: $\sum_t p(t) = 1$ and $\forall t, p(t) \geq 0$.

$$p(a) = \frac{7}{12}, p(f(a,a)) = \frac{269}{1728}, p(f(a,f(a,a))) = p(f(f(a,a),a)) = \frac{9823}{248832}, \ldots$$

Oracle: Is $\overline{f(f(a,a),f(a,a))}$ colinear to $\overline{a}, \overline{f(a,a)}$?

Let $B = \{\overline{a}, \overline{f(a,a)}\}$.

Building the Canonical Linear Representation

Let $p = 2p_{2/3} - p_{3/4}$: $\sum_t p(t) = 1$ and $\forall t, p(t) \geq 0$.

$$p(a) = \frac{7}{12}, p(f(a,a)) = \frac{269}{1728}, p(f(a,f(a,a))) = p(f(f(a,a),a)) = \frac{9823}{248832}, \ldots$$

Oracle: Is $\overline{f(f(a,a),f(a,a))}$ colinear to $\overline{a}, \overline{f(a,a)}$?

Answer: YES,

$$\overline{f(f(a,a),f(a,a))} = \frac{-3186}{2^8 \times 3^6}\overline{a} + \frac{2617}{2^8 \times 3^4}\overline{f(a,a)}.$$

Let $B = \{\overline{a}, \overline{f(a,a)}\}$.

Building the Canonical Linear Representation

$$p = 2p_{2/3} - p_{3/4}$$

$B = \{\overline{a}, \overline{f(a,a)}\}$.

$$\begin{aligned}
\mu(a) &= \overline{a} \\
\mu(f)(\overline{a}, \overline{a}) &= \overline{f(a,a)} \\
\mu(f)(\overline{a}, \overline{f(a,a)}) &= \frac{-54}{2^4 \times 3^4}\overline{a} + \frac{59}{2^4 \times 3^2}\overline{f(a,a)} \\
\mu(f)(\overline{f(a,a)}, \overline{a}) &= \frac{-54}{2^4 \times 3^4}\overline{a} + \frac{59}{2^4 \times 3^2}\overline{f(a,a)} \\
\mu(f)(\overline{f(a,a)}, \overline{f(a,a)}) &= \frac{-3186}{2^8 \times 3^6}\overline{a} + \frac{2617}{2^8 \times 3^4}\overline{f(a,a)}
\end{aligned}$$

$\lambda(\overline{a}) = p(a) = \frac{7}{12}; \lambda(\overline{f(a,a)}) = p(f(a,a)) = \frac{269}{1728}$

Algorithm DEES; Independence Test

S a finite sample i.i.d. from p; B current basis; s vector candidate

$$\forall \alpha_t \in \mathbb{R}, \overline{s} \neq \sum_{t \in B} \alpha_t \overline{t}$$

$$\simeq$$

$$\bigwedge_{c : \exists t \;\; c[t] \in S} \left\{ |p_S(c[s]) - \sum_{t \in B} \alpha_t p_S(c[t])| \leq \epsilon \right\} \text{ has no solution.}$$

Take $\epsilon = |S|^{-\gamma}$ where $\gamma \in]0, 1/2[$ (VC bounds).

Properties of DEES

Theorem [F. Denis and A. Habrard, ALT'07]

DEES identifies the correct basis in the limit with probability one and the parameters converge to the correct ones in $O(|S|^{-1/2})$.

But ...

- In the model output, the states may not define stochastic languages.
- The parameters are not normalized.
- Before convergence, the model output may not define a stochastic language.

Outline

1. The Basic Problem
2. A Canonical Linear Representation for Rational Tree Series
3. Contributions
   - Normalization of the Model as a Generative Model
   - Strongly Consistent Model
   - Unranked Trees

The Normalization of the Model

$$q \rightarrow q_0, 7/12 + q_1, 269/1728$$
$$q_0 \rightarrow a, 1 + f(q_0, q_1), \frac{-54}{2^4 3^4} + f(q_1, q_0), \frac{-54}{2^4 3^4} + f(q_1, q_1), \frac{-3186}{2^8 3^6}$$
$$q_1 \rightarrow f(q_0, q_0), 1 + f(q_0, q_1), \frac{59}{2^4 3^2} + f(q_1, q_0), \frac{59}{2^4 3^2} + f(q_1, q_1), \frac{2617}{2^8 3^4}$$

The Normalization of the Model

$$q \rightarrow q_0, 7/12 + q_1, 269/1728$$

$$q_0 \rightarrow a, 1 + f(q_0, q_1), \frac{-54}{2^4 3^4} + f(q_1, q_0), \frac{-54}{2^4 3^4} + f(q_1, q_1), \frac{-3186}{2^8 3^6}$$

$$q_1 \rightarrow f(q_0, q_0), 1 + f(q_0, q_1), \frac{59}{2^4 3^2} + f(q_1, q_0), \frac{59}{2^4 3^2} + f(q_1, q_1), \frac{2617}{2^8 3^4}$$

Theorem

For any rational stochastic language, there exists a normalized representation with a basis chosen to ensure that:

- Each state defines a stochastic language.
- The weights of the transitions are normalized.

After Renormalization

- $\forall$ state lhs: Sum of the transition weights is one.
- $\forall$ pair (state-lhs,symbol): Sum of the transition weights ≥ 0.

$$
\begin{aligned}
q &\rightarrow q_0, 1 \\
q_0 &\rightarrow a, \frac{7}{12} + f(q_0, q_0), \frac{-269}{50} + f(q_0, q_1), \frac{259}{50} + f(q_1, q_0), \frac{259}{50}, \\
&\quad + f(q_1, q_1), \frac{-1369}{300} \\
q_1 &\rightarrow a, \frac{269}{444} + f(q_0, q_0), \frac{-3024}{925} + f(q_0, q_1), \frac{2664}{925} + f(q_1, q_0), \frac{2664}{925} \\
&\quad + f(q_1, q_1), \frac{-23273}{11100}
\end{aligned}
$$

- Efficient propagative method for computing the normalization.
- Still negative weights $\longrightarrow$ specific generation algorithm.

Notion of Strong Consistency

Rational Stochastic Tree Language Strongly Consistent

Bounded average tree size: $\sum_t p(t)|t| < \infty$

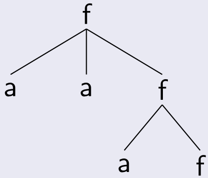
Theorem

For a strongly consistent RSTL, the spectral radius of the *"expectation matrix"* A taken from the normalized representation is strictly less than 1 ($\rho(A) < 1$).

Errata: Some hypotheses are missing in Proposition 1 see
`http://hal.archives-ouvertes.fr/hal-00293511/en`
(the series $\sum_{t \in T(\mathcal{F})} p_i(t)$ and $\sum_{t \in T(\mathcal{F})} p_i(t)|t|$ have to be absolutely convergent)

Adapting the Framework to Unranked Trees

Unranked Trees



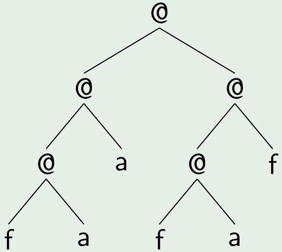
- Unranked tree series

$\Leftrightarrow$ **Bijection**

$\Leftrightarrow$ **Equivalence**

Ranked Trees

$\mathcal{F}_0 = \{a, f\}$ $\mathcal{F}_2 = \{@\}$



- Ranked tree series

All the inference results apply: Convert the data and use DEES

Conclusion: Learning RSTL from *i.i.d.* samples

- DEES may output irrelevant representations.
- Our contributions:

  - Existence and construction of a normalized representation.
  - Algorithm for generating trees from the distribution.
  - Strong consistency.
  - Application to unranked trees.

- $\Rightarrow$ When the models do not define stochastic languages, a distribution can be extracted and controlled if $\rho(A) < 1$.
- $\Rightarrow$ A prototype software is being developed (Piccata).