Toric grammars, a new stochastic model

Olivier Catoni

CNRS, INRIA (projet CLASSIC)
Département de Mathématiques et Applications,
ENS, 45 rue d’Ulm, 75 230 Paris Cedex 05,
Olivier.Catoni@ens.fr

Toulouse, Université Paul Sabatier,
Tuesday, March 26, 2013
Toric grammars: joint work with Thomas Mainguy.
Language analysis

Let $S_1, \ldots, S_n$ be $n$ independent copies of the random sentence $S \in D^+ = \bigcup_{k=1}^{\infty} D^k$, where $D$ is a finite dictionary.

The empirical process $\overline{P} = \frac{1}{n} \sum_{i=1}^{n} \delta_{S_i}$ is the starting point to estimate the probability distribution of $S$.

Usual statistical approaches are:

- **kernel estimate:** $\mathbb{E}[f(S)]$ is estimated by
  \[ \int \int f(s')d\overline{P}(s)dk(s,s'), \]
  where $k(s,\cdot)$ is a smoothing kernel.

- **parametric estimate:** $\mathbb{P}_S$ the law of $S$, is estimated by $P_\theta$, where the parameter $\theta$ minimizes $\int \ell(\theta, s) d\overline{P}(s)$, with typically $\ell(\theta, s) = -\log[P_\theta(s)]$. 
Let us consider the space of empirical measures of size $n$

$$\mathcal{E} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \delta_{s_i}, s_i \in D^+ \right\},$$

and $q(P, \cdot) \in \mathcal{M}_1^+(\mathcal{E})$, $P \in \mathcal{E}$, a Markov kernel on this state space. We may estimate $\mathbb{E}[f(S)]$ by

$$\lim_{t \to \infty} \frac{1}{t} \sum_{j=1}^{t} \int_{P \in \mathcal{E}} \int_{s \in D^+} f(s) \ dP(s) \ dq^j(\bar{P}, P).$$
A simple example

Consider a pair of independent random variables \((X, Y)\) and a sample \((X_i, Y_i), 1 \leq i \leq n\), made of \(n\) independent copies of \((X, Y)\). Let \(\sigma \in \mathfrak{S}(\{1, \ldots, n\})\) be a uniform random permutation, and \(\sigma_t\) independent copies of \(\sigma\), independent of everything else. The estimate of \(P(X, Y)\) given by

\[
\lim_{t \to \infty} \frac{1}{t} \sum_{k=1}^{t} \frac{1}{n} \sum_{i=1}^{n} \delta(x_i, y_{\sigma_k(i)}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta(x_i, y_j)
\]

is a sample level kernel estimate with kernel

\[
q = P \left( n^{-1} \sum_{i=1}^{n} \delta_{x_i, y_{\sigma(i)}} \right) \left| n^{-1} \sum_{i=1}^{n} \delta_{x_i, y_i} \right).
\]
Toric grammars

Let \( D_1 = D \cup \{ ]_i, 1 \leq i \leq d_1 \} \), \( D_j = D_{j-1} \cup \{ ]_i, d_{j-1} < i \leq d_j \} \),

Consider any \( x_k \in D_j^+, 1 \leq k \leq n \), let \( \tau_i = \sum_{k=1}^{m} \sum_{t=1}^{\ell(x_k)} 1(x_k, t = ]i) \),

and consider also any \( y_{i,t} \in D_{j-1}^+, d_{j-1} < i \leq d_j, 1 \leq t \leq \tau_i \),

Define
\[
\alpha((x_k, 1 \leq k \leq n), (y_{i,t}, d_{j-1} < i \leq d_j, 1 \leq t \leq \tau_i))
= (\tilde{x}_k, 1 \leq k \leq n)
\]

by replacing each \( ]_i \) by the corresponding \( y_{i,t} \).
Random parsing

Let $X_{0,k} = S_k$, $1 \leq k \leq n$. Let $X_{j,k} \in D_j^+$, $Y_{j,i,t} \in D_{j-1}^+$, $1 \leq j \leq J$, $d_{j-1} < i \leq d_j$, $1 \leq t \leq \tau_{j,i}$, where

$$\tau_{j,i} = \sum_{k=1}^n \sum_{t=1}^{\ell(X_{j,k},t)} 1(X_{j,k},t = [i])$$

be random variables. Let us put $W_j = (X_{j,k}; Y_{j,i,t})$ and let us assume that almost surely

$$\alpha(W_j) = (X_{j-1,k}, 1 \leq k \leq n).$$

Let us assume moreover that

$$\mathbb{P}_{X_{j,k}, Y_{j,i,t}} = \mathbb{P}_{X_{j,1}}^{\otimes n} \prod_{i=d_{j-1}+1}^{d_j} \mathbb{P}_{Y_{j,i,1}}^{\otimes \tau_{j,i}}|_{\tau_{j,i} > 0}, \quad (\mathcal{I}).$$
Let us consider $\tilde{X}_{J,k} = X_{J,k}$ and

$$(\tilde{X}_{j-1,k}, 1 \leq k \leq n) = \alpha[(\tilde{X}_{j,k}, 1 \leq k \leq n), (Y_{j,i}, \sigma_{j,i}(t))],$$

where $\sigma_{j,i}$ are independent uniform random permutations of $\{1, \ldots, \tau_{j,i}\}$. Let us consider the sample level kernel

$$q = \mathbb{P} \left( n^{-1} \sum_{k=1}^{n} \delta_{\tilde{X}_{0,k}} \mid n^{-1} \sum_{k=1}^{n} \delta_{S_k} \right).$$
Proposition

The sample level kernel $q$ is reversible, with invariant measure

\[ \mathbb{P} = \mathbb{P}(P)q(P, Q) = \mathbb{P}(Q)q(Q, P). \]

The sample level kernel estimate

\[ \hat{\mathbb{P}} = \frac{1}{T} \sum_{t=1}^{T} \int P \, dq^{t}(\overline{\overline{P}}, P) \]

is unbiased in the sense that

\[ \mathbb{E}(\hat{\mathbb{P}}) = \mathbb{P} \, S. \]
A small recursive example

Here \( J = d_J = 1 \).

\[
\begin{align*}
\mathbb{P}_{X_{1,1}}(a^m)_1 &= 2^{-m}, \quad m \geq 1, \\
\mathbb{P}_{Y_{1,1}}(b) &= 1/2, \\
\mathbb{P}_{Y_{1,1}}(ab) &= 1/2, \\
\mathbb{P}_{S}(ab) &= 1/4, \\
\mathbb{P}_{S}(a^m b) &= 3 \times 2^{-(m+1)}, \quad m \geq 2, \\
\mathbb{P}(W_1|S = ab)(a)_1, b) &= 1, \\
\mathbb{P}(W_1|S = a^m b)(a^m)_1, b) &= 1/3, \quad m \geq 2, \\
\mathbb{P}(W_1|S = a^m b)(a^{m-1})_1, ab) &= 2/3, \quad m \geq 2.
\end{align*}
\]
A small natural language example

1. He is a clever guy.
2. He is doing some shopping.
3. He is laughing.
4. He is not interested in sports.
5. He is walking.
6. He likes to walk in the streets.
7. I am driving a car.
8. I am riding a horse too.
9. I am running.
10. Paul is crossing the street.
11. Paul is driving a car.
12. Paul is riding a horse.
13. Paul is walking.
14. Peter is walking.
15. While I was walking, I saw Paul crossing the street.
1 [0 Paul is driving a car too.
1 [0 Paul is doing some shopping.
1 [0 Paul is laughing.
1 [0 Paul is riding a horse too.
1 [0 Paul is running too.
1 [0 Paul is running.
1 [0 Paul is not interested in sports too.
1 [0 Paul is not interested in sports.
1 [0 Paul is a clever guy too.
1 [0 Paul is a clever guy.
1 [0 Paul is walking too.
1 [0 Peter is driving a car too.
1 [0 Peter is driving a car.
1 [0 Peter is doing some shopping.
1 [0 Peter is laughing.
1 [0 Peter is riding a horse too.
1 [0 Peter is riding a horse.
1 [0 Peter is running too.
1 [0 Peter is running.
1 [0 Peter is not interested in sports.
Peter is a clever guy.
Peter is crossing the street.
He is driving a car too.
He is driving a car.
He is riding a horse too.
He is riding a horse.
He is running too.
He is running.
He is not interested in sports too.
He is crossing the street too.
He is crossing the street.
He is walking too.
I am driving a car too.
I am doing some shopping.
I am laughing too.
I am laughing.
I am riding a horse.
I am not interested in sports.
I am a clever guy.
I am crossing the street too.
I am crossing the street.
I am walking too.
I am walking.
While I was driving a car, I saw Paul doing some shopping.
While I was driving a car, I saw Paul riding a horse.
While I was doing some shopping, I saw Paul walking.
While I was riding a horse, I saw Paul driving a car.
While I was riding a horse, I saw Paul running.
While I was riding a horse, I saw Paul walking.
While I was riding a horse, I saw Peter not interested in sports.
While I was running, I saw Paul laughing.
While I was running, I saw Paul not interested in sports.
While I was running, I saw Paul a clever guy.
While I was running, I saw Paul walking.
While I was not interested in sports, I saw Paul driving a car.
While I was not interested in sports, I saw Paul riding a horse.
While I was a clever guy, I saw Paul running.
While I was a clever guy, I saw Paul crossing the street.
While I was a clever guy, I saw Paul walking.
While I was crossing the street, I saw Paul riding a horse.
While I was crossing the street, I saw Paul running.
While I was crossing the street, I saw Paul crossing the street.
While I was crossing the street, I saw Paul walking.
While I was crossing the street, I saw Peter walking.
While I was walking, I saw Paul driving a car.
While I was walking, I saw Paul laughing.
While I was walking, I saw Paul riding a horse.
While I was walking, I saw Paul running.
While I was walking, I saw Paul not interested in sports.
While I was walking, I saw Paul crossing the street too.
While I was walking, I saw Paul walking.
While I was walking, I saw Peter not interested in sports.
While I was walking, I saw Peter walking.
He likes to walk 3 streets.

He is a clever guy.

He is doing some shopping.

He is laughing.

He is not interested in sports.

He is riding a horse.

He is riding a horse.

He is running.

7 am.

Paul is.

He is crossing a street.

He is driving a car.

4 is.

He is walking.

Peter is.

While he was, he saw.

He is.

Peter is.
While 7 was 5, 7 saw 4.

7 am

Paul is too

the

Peter

Paul crossing street

Paul driving car

Paul riding horse

Paul walking

5 too

clever guy

Paul doing some shopping

Paul laughing

Paul not interested sports

Paul running

in

I

a