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 $S_1, \dots, S_n$ 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 $\overline{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^{n} \delta_{S_i}$ is the starting point to estimate the probability distribution of S.

Usual statistical approaches are :

- kernel estimate: $\mathbb{E}\big[f(S)\big]$ is estimated by $\int\int f(s')\mathrm{d}\overline{\mathbb{P}}(s)\mathrm{d}k(s,s')$, where $k(s,\cdot)$ is a smoothing kernel.
- parametric estimate: $\mathbb{P}_S$ the law of S, is estimated by P_θ, where the parameter θ minimizes $\int \ell(\theta, s)\,\mathrm{d}\overline{\mathbb{P}}(s)$, with typically $\ell(\theta, s) = -\log\big[P_\theta(s)\big]$.

Sample level kernel estimate

Let us consider the space of empirical measures of size n

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

and $q(P, \cdot) \in \mathscr{M}_+^1(\mathscr{E})$, $P \in \mathscr{E}$, a Markov kernel on this state space. We may estimate $\mathbb{E}[f(S)]$ by

$$\lim_{t \to \infty} \frac{1}{t} \sum_{j=1}^{t} \int_{P \in \mathscr{E}} \int_{s \in D^+} f(s) \, \mathrm{d}P(s) \, \mathrm{d}q^j(\overline{\mathbb{P}}, P).$$

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 $\sigma \in \mathfrak{S}(\{1, \dots, n\})$ be a uniform random permutation, and σ_t independent copies of σ, independent of everything else. The estimate of $\mathbb{P}_{(X,Y)}$ 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

$$q = \mathbb{P}_{n^{-1} \sum_{i=1}^{n} \delta_{x_i, y_{\sigma(i)}} \Big| n^{-1} \sum_{i=1}^{n} \delta_{x_i, y_i}}.$$

Toric grammars

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

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

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

Define

$$\alpha\big((x_k, 1 \leq k \leq n), (y_{i,t}, d_{j-1} < i \leq d_j, 1 \leq t \leq \tau_i)\big)$$
$$= (\widetilde{x}_k, 1 \leq k \leq n)$$

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

$$\mathbb{P}_{X_{j,k},\, Y_{j,i,t}} = \mathbb{P}_{X_{j,1}}^{\otimes n} \prod_{i=d_{j-1}+1}^{d_j} \mathbb{P}_{Y_{j,i,1} | \tau_{j,i} > 0}^{\otimes \tau_{j,i}}, \qquad (\mathscr{I}).$$

Sample level kernel

Let us consider $\widetilde{X}_{J,k} = X_{J,k}$ and

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

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

$$q = \mathbb{P}_{n^{-1}\sum_{k=1}^{n} \delta_{\widetilde{X}_{0,k}} | n^{-1}\sum_{k=1}^{n} \delta_{S_k}}.$$

Proposition

The sample level kernel q is reversible, with invariant measure $\mathbb{P}_{\overline{\mathbb{P}}} : \mathbb{P}_{\overline{\mathbb{P}}}(P)q(P,Q) = \mathbb{P}_{\overline{\mathbb{P}}}(Q)q(Q,P)$. The sample level kernel estimate

$$\widehat{\mathbb{P}} = \frac{1}{T}\sum_{t=1}^{T}\int P\,\mathrm{d}q^t(\overline{\mathbb{P}},P)$$

is unbiased in the sense that

$$\mathbb{E}(\widehat{\mathbb{P}}) = \mathbb{P}_S.$$

A small recursive example

Here $J = d_J = 1$.

$$\begin{aligned}
\mathbb{P}_{X_{1,1}}(a^m]_1) &= 2^{-m}, && m \geq 1,\\
\mathbb{P}_{Y_{1,1,1}}(b) &= 1/2,\\
\mathbb{P}_{Y_{1,1,1}}(ab) &= 1/2,\\
\mathbb{P}_{S}(ab) &= 1/4,\\
\mathbb{P}_{S}(a^m b) &= 3 \times 2^{-(m+1)}, && m \geq 2,\\
\mathbb{P}_{(W_1|S=ab)}(a]_1, b) &= 1,\\
\mathbb{P}_{(W_1|S=a^m b)}(a^m]_1, b) &= 1/3, && m \geq 2,\\
\mathbb{P}_{(W_1|S=a^m b)}(a^{m-1}]_1, ab) &= 2/3, && m \geq 2.
\end{aligned}$$

A small natural language example

```
1 [0 He is a clever guy .
1 [0 He is doing some shopping .
1 [0 He is laughing .
1 [0 He is not interested in sports .
1 [0 He is walking .
1 [0 He likes to walk in the streets .
1 [0 I am driving a car .
1 [0 I am riding a horse too .
1 [0 I am running .
1 [0 Paul is crossing the street .
1 [0 Paul is driving a car .
1 [0 Paul is riding a horse .
1 [0 Paul is walking .
1 [0 Peter is walking .
1 [0 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 .
```

```
1 [0 Peter is a clever guy .
1 [0 Peter is crossing the street .
1 [0 He is driving a car too .
1 [0 He is driving a car .
1 [0 He is riding a horse too .
1 [0 He is riding a horse .
1 [0 He is running too .
1 [0 He is running .
1 [0 He is not interested in sports too .
1 [0 He is crossing the street too .
1 [0 He is crossing the street .
1 [0 He is walking too .
1 [0 I am driving a car too .
1 [0 I am doing some shopping .
1 [0 I am laughing too .
1 [0 I am laughing .
1 [0 I am riding a horse .
1 [0 I am not interested in sports .
1 [0 I am a clever guy .
1 [0 I am crossing the street too .
1 [0 I am crossing the street .
1 [0 I am walking too .
1 [0 I am walking .
```

```
1 [0 While I was driving a car , I saw Paul doing some shopping too .
1 [0 While I was driving a car , I saw Paul doing some shopping .
1 [0 While I was driving a car , I saw Paul riding a horse .
1 [0 While I was driving a car , I saw Paul crossing the street .
1 [0 While I was driving a car , I saw Paul walking .
1 [0 While I was driving a car , I saw Peter riding a horse .
1 [0 While I was doing some shopping , I saw Paul riding a horse .
1 [0 While I was doing some shopping , I saw Paul walking .
1 [0 While I was laughing too , I saw Peter crossing the street .
1 [0 While I was laughing , I saw Peter riding a horse .
1 [0 While I was riding a horse , I saw Paul driving a car too .
1 [0 While I was riding a horse , I saw Paul driving a car .
1 [0 While I was riding a horse , I saw Paul laughing .
```

```
1 [0 While I was riding a horse , I saw Paul running .
1 [0 While I was riding a horse , I saw Paul walking .
1 [0 While I was riding a horse , I saw Peter not interested in sports .
1 [0 While I was running , I saw Paul laughing .
1 [0 While I was running , I saw Paul not interested in sports .
1 [0 While I was running , I saw Paul a clever guy .
1 [0 While I was running , I saw Paul walking .
1 [0 While I was not interested in sports , I saw Paul driving a car .
1 [0 While I was not interested in sports , I saw Paul riding a horse .
1 [0 While I was a clever guy , I saw Paul running .
1 [0 While I was a clever guy , I saw Paul crossing the street .
1 [0 While I was a clever guy , I saw Paul walking .
1 [0 While I was crossing the street , I saw Paul riding a horse .
1 [0 While I was crossing the street , I saw Paul running .
1 [0 While I was crossing the street , I saw Paul crossing the street .
1 [0 While I was crossing the street , I saw Paul walking .
1 [0 While I was crossing the street , I saw Peter walking .
1 [0 While I was walking , I saw Paul driving a car .
1 [0 While I was walking , I saw Paul laughing .
1 [0 While I was walking , I saw Paul riding a horse .
1 [0 While I was walking , I saw Paul running .
1 [0 While I was walking , I saw Paul not interested in sports .
1 [0 While I was walking , I saw Paul crossing the street too .
1 [0 While I was walking , I saw Paul walking .
1 [0 While I was walking , I saw Peter not interested in sports .
1 [0 While I was walking , I saw Peter walking .
```

```
10 [0 He likes to walk ]6 ]3 streets .
2 [0 ]1 ]8 clever guy .
2 [0 ]1 doing some shopping .
2 [0 ]1 laughing .
2 [0 ]1 not interested ]6 sports .
2 [0 ]1 riding ]8 horse .
2 [0 ]1 riding ]8 horse ]2 .
2 [0 ]1 running .
24 [0 ]7 am ]5 .
28 [0 Paul is ]5 .
40 [0 He is ]5 .
4 [0 ]1 crossing ]3 street .
4 [0 ]1 driving ]8 car .
5 [0 ]4 is ]5 .
6 [0 ]1 walking .
7 [0 Peter is ]5 .
8 [0 While ]7 was ]5 , ]7 saw ]4 ]5 .
10 [1 He is
2 [1 Peter is
```

2 [1 While]7 was]5 ,]7 saw]4
6 [1]7 am
8 [1 Paul is
2 [2 too
30 [3 the
14 [4 Paul
1 [4 Peter
16 [5 crossing]3 street
16 [5 driving]8 car
16 [5 riding]8 horse
34 [5 walking
8 [5]5 too
8 [5]8 clever guy
8 [5 doing some shopping
8 [5 laughing
8 [5 not interested]6 sports
8 [5 running
20 [6 in
50 [7 I
50 [8 a