Markov substitute models, an alternative to hidden Markov models

Olivier Catoni
Olivier.Catoni@ensae.fr

CREST, CNRS — UMR 9194
Université Paris–Saclay

Sequel seminar, — INRIA Lille

June 9, 2017

Joint work with

Thomas Mainguy

A model for random finite sequences

Finite sequences of words

- Let D be a finite dictionary,
- and $D^+ = \bigcup_{j=1}^{\infty} D^j$ the set of all finite sequences of words of positive length.
- Let ε be the empty sequence (of length zero),
- and $D^* = D^+ \cup \{\varepsilon\}$ the set of finite sequences, including the empty one.

A random sentence

We will present a model

- for random finite sequences,
- that is a family of probability measures included in $\mathscr{M}_+^1(D^+)$, the set of probability measures on D^+.

Markov substitute sets

A string distribution

- Let $\mathscr{D} \subset D^+$ be a domain,
- and $P \in \mathscr{M}_+^1(\mathscr{D})$ a probability distribution on $\mathscr{D}$.
- Let γ be the concatenation operator.

Definition (Markov substitute sets)

A set $B \subset D^+$ is a Markov substitute set of P if and only if

- there exists a function $\beta : B \times B \to \mathbb{R}$, that we will call the substitute exponent, such that
- for any context $(x, z) \in (D^*)^2$,
- for any couple of expressions $(y, y') \in B^2$,
- such that $(\gamma(x, y, z), \gamma(x, y', z)) \in \mathscr{D}^2$,

$$P(\gamma(x, y', z)) = P(\gamma(x, y, z)) \exp(\beta(y, y')).$$

Properties of the substitute exponent

Symmetry and independence from B

- Since $P\big(\gamma(x, y', z)\big) = P\big(\gamma(x, y, z)\big)\exp\big(\beta(y, y')\big)$, the substitute exponent is skew-symmetric:

$$\beta(y', y) = -\beta(y, y').$$

- The substitute exponent does not depend on B : if $\{y, y'\} \subset B \cap B'$, two Markov substitute sets, then we can take the same value of $\beta(y, y')$ to describe the substitute property in B and in B'.

First properties

Proposition (Crossing-over does not change the likelihood)

- *For any Markov substitute set* B *of* $P \in \mathscr{M}_+^1(\mathscr{D})$,
- *for any two contexts* $(x_1, z_1), (x_2, z_2) \in \big(D^*\big)^2$,
- *and any pair* $y_1, y_2 \in B$,
- *such that* $\gamma(x_i, y_j, z_i) \in \mathscr{D}$, $1 \leq i \leq 2, 1 \leq j \leq 2$,

$$
\begin{aligned}
P\big(\gamma(x_1, y_1, z_1)\big) P\big(\gamma(x_2, y_2, z_2)\big) \\
= P\big(\gamma(x_1, y_2, z_1)\big) P\big(\gamma(x_2, y_1, z_2)\big)
\end{aligned}
$$

Elementary properties of substitute sets

Pairs are sufficient

- A subset of a Markov substitute set is itself a Markov substitute set,
- The set $B \subset D^+$ is a Markov substitute set if and only if any pair $\{y, y'\} \subset B$ is a Markov substitute set,
- If B is a Markov subsitute set and $(x, z) \in \left(D^*\right)^2$ is a context, then

$$\gamma(x, B, z) \stackrel{\text{def}}{=} \{\gamma(x, y, z) : y \in B\}$$

is also a Markov substitute set.

Substitute sets as syntax labels

Decomposition of a sentence likelihood

- When B is a Markov substitute set
- and the context $(x,z) \in \left(D^*\right)^2$ is such that $\gamma(x,B,z) \subset \mathscr{D}$,
- the likelihood of a sequence $\gamma(x,y,z)$ decomposes into

$$P\big(\gamma(x,y,z)\big) = P\big(\gamma(x,B,z)\big)\,q_B(y), \qquad y \in B,$$

- where q_B, the substitute measure of B is defined as

$$q_B(y) = \frac{\exp\big(\beta(y',y)\big)}{\displaystyle\sum_{y'' \in B} \exp\big(\beta(y',y'')\big)},$$

this definition being independent of the choice of $y' \in B$.

Substitute sets as syntax labels

Syntax labels

- Due to the decomposition

$$P(\gamma(x,y,z)) = P(\gamma(x,B,z))\,q_B(y), \quad y \in B,$$

- the substitute set B behaves as a syntax label ℓ_B,
- the likelihood of $\gamma(x,y,z)$ being deduced from
- the likelihood of the syntactic construction $\gamma(x,\ell_B,z)$
- and the likelihood $q_B(y)$
- of the rewriting rule $\ell_B \to y$,

Multiple parsings

- but the parsing of $\gamma(x,y,x')$ into $\gamma(x,\ell_B,x')$ may not be unique.

The set of $\mathscr{B}$-Markov probability measures

Model definition

- For any given finite family $\mathscr{B}$ of finite subsets of D^+,
- for any domain $\mathscr{D} \subset D^+$,
- the probability measure $P \in \mathscr{M}_+^1(\mathscr{D})$
- is said to be a $\mathscr{B}$-Markov probability measure on $\mathscr{D}$,
- if and only if all sets $B \in \mathscr{B}$ are Markov substitute sets of P.
- The notation $\mathfrak{M}(\mathscr{D}, \mathscr{B})$ will stand for the set of $\mathscr{B}$-Markov probability measures on the domain $\mathscr{D}$.

Substitute Markov models in action

A training sample

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

Syntax rules, inferred from the training sample

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

Syntax rules, inferred from the training sample

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

New sentences discovered

```
[0 Paul is driving a car too .
[0 Paul is doing some shopping .
[0 Paul is laughing .
[0 Paul is riding a horse too .
[0 Paul is running too .
[0 Paul is running .
[0 Paul is not interested in sports too .
[0 Paul is not interested in sports .
[0 Paul is a clever guy too .
[0 Paul is a clever guy .
[0 Paul is walking too .
[0 Peter is driving a car too .
[0 Peter is driving a car .
[0 Peter is doing some shopping .
[0 Peter is laughing .
[0 Peter is riding a horse too .
[0 Peter is riding a horse .
[0 Peter is running too .
[0 Peter is running .
[0 Peter is not interested in sports .
```

New sentences discovered

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

New sentences discovered

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

New sentences discovered

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

$\mathscr{B}$-Markov models form exponential families

also known as Gibbs measures

The substitute graph on $\mathscr{D}$

$$\mathscr{G}(\mathscr{D},\mathscr{B}) = \Big\{\big(\gamma(x,y,z),\gamma(x,y',z)\big), \\ (x,z)\in\big(D^*\big)^2,(y,y')\in B^2, B\in\mathscr{B}\Big\}\cap\big(\mathscr{D}\times\mathscr{D}\big)$$

defines an equivalence relation $\sim_{\mathscr{B}}$ on the domain $\mathscr{D}$.

- The components $\mathscr{D}/\sim_{\mathscr{B}}$ are the connected components of the graph.
- The support of any $P\in\mathfrak{M}(\mathscr{D},\mathscr{B})$ is necessarily a union of components: for some $\mathscr{C}_P\subset\mathscr{D}/\sim_{\mathscr{B}}$

$$\mathrm{supp}(P) = \bigcup_{C\in\mathscr{C}_P} C$$

$\mathscr{B}$-Markov models form exponential families

also known as Gibbs measures

$\mathscr{B}$-Markov models with a given support

Conversely, for any $\mathscr{C} \subset \mathscr{D}/{\sim_{\mathscr{B}}}$, the set $\mathfrak{M}_{\mathscr{C}}(\mathscr{D}, \mathscr{B})$ of $\mathscr{B}$-Markov probability measures with support $\bigcup_{C \in \mathscr{C}} C$ is non-empty.

$\mathscr{B}$-Markov models form exponential families

also known as Gibbs measures

Independent $\mathscr{B}$-Markov processes

- Consider $\xi \in \mathscr{M}_+(D)$, such that $r = 1 - \xi(D) > 0$,
- and let $\widetilde{P}(w) = \dfrac{r}{1-r} \displaystyle\prod_{j=1}^{k} \xi(w_j), \quad w \in D^k, k \in \mathbb{N} \setminus \{0\}$.
- Remark that $\widetilde{P} \in \mathfrak{M}(D^+, \{\mathrm{supp}(\xi)^+\})$.
- For any family $\mathscr{B}$ of subsets of D^+, any domain $\mathscr{D} \subset D^+$, any $\mathscr{C} \subset \mathscr{D}/\sim_{\mathscr{B}}$, any probability measure $\mu \in \mathscr{M}_+^1(\mathscr{C})$, the probability P defined as $P(s) = \displaystyle\sum_{C \in \mathscr{C}} \mathbb{1}(s \in C) \mu(C) \widetilde{P}(s)/\widetilde{P}(C)$, $s \in \mathscr{D}$ belongs to $\mathfrak{M}_{\mathscr{C}}(\mathscr{D}, \mathscr{B})$.

$\mathscr{B}$-Markov models form exponential families

also known as Gibbs measures

Active pairs

- Consider any domain $\mathscr{D} \subset D^+$ and any family $\mathscr{B} \subset 2^{D^+}$
- let $\mathscr{P}$ be a minimal set of pairs such that $\mathfrak{M}(\mathscr{D}, \mathscr{B}) = \mathfrak{M}(\mathscr{D}, \mathscr{P})$, implying that $\mathscr{D}/\sim_{\mathscr{B}} = \mathscr{D}/\sim_{\mathscr{P}}$.
- Let $\mathscr{C} \subset \mathscr{D}/\sim_{\mathscr{B}}$.
- Define the set of active pairs
$$\mathscr{A} = \Big\{ \{y, y'\} \in \mathscr{P}, \text{ for some } x, z \in D^*,\ C \in \mathscr{C}, \\ \gamma(x, \{y, y'\}, z) \subset C \Big\}.$$

$\mathscr{B}$-Markov models form exponential families

also known as Gibbs measures

Free pairs and Gibbs measures

- There is a nonempty subset $\mathscr{F} \subset \mathscr{A}$ of free pairs,
- and energy functions $U_i : \bigcup_{C \in \mathscr{C}} C \to \mathbb{R}$, where $i \in \mathscr{I} \stackrel{\text{def}}{=} \mathscr{F} \cup \mathscr{C}$, such that,
- defining $Z_\beta = \displaystyle\sum_{C \in \mathscr{C}} \sum_{s \in C} \exp\left(- \sum_{i \in \mathscr{I}} \beta_i U_i(s) \right)$.
- and $P_\beta(s) = Z_\beta^{-1} \exp\left(- \displaystyle\sum_{i \in \mathscr{I}} \beta_i U_i(s) \right), \quad s \in \displaystyle\bigcup_{C \in \mathscr{C}} C$
- $\mathfrak{M}_{\mathscr{C}}(\mathscr{D}, \mathscr{B}) = \left\{ P_\beta : \beta \in \mathbb{R}^{\mathscr{I}}, Z_\beta < \infty \right\}$,
- and such that moreover

$$\left(P_\beta = P_{\beta'} \text{ and } Z_\beta = Z_{\beta'} \right) \iff \beta = \beta'.$$

$\mathscr{B}$-Markov models form exponential families

also known as Gibbs measures

Substitute exponents from temperature parameters

- For any $i = \{y, y'\} \in \mathscr{F}$, where $y < y'$, $\beta(y, y') = \beta_i$,
- and for any $j = \{z, z'\} \in \mathscr{A} \setminus \mathscr{F}$, where $z < z'$,

$$\beta(z, z') = \sum_{i \in \mathscr{F}} \beta_i \, e_{i,j},$$

- for some matrix $\left(e_{i,j},\, i \in \mathscr{F},\, j \in \mathscr{A} \setminus \mathscr{F}\right)$,
- while the substitute exponents for non active pairs in $\mathscr{P} \setminus \mathscr{A}$ can be set arbitrarily.

Some ideas from the proof: the loop constraint

- For any path $(x_0, \ldots, x_k)$ in the substitute graph $\mathscr{G}(\mathscr{D}, \mathscr{A})$, there are pairs $\{y_j, y'_j\} \in \mathscr{A}$ such that one goes from x_{j-1} to x_j by changing y_j into y'_j.
- Therefore, if $P \in \mathfrak{M}_{\mathscr{C}}(\mathscr{D}, \mathscr{B})$,

$$P(x_k) = P(x_1) \exp\left(\sum_{j=1}^{k} \beta(y_j, y'_j)\right)$$

$$= P(x_1) \exp\left(\sum_{p \in \mathscr{A}} -\beta(p)\, V_p(x_0, \ldots, x_k)\right),$$

where

$$V_p(x_1, \ldots, x_k) = \sum_{j=1}^{k} \Big[\mathbb{1}(y_j > y'_j) - \mathbb{1}(y_j < y'_j)\Big] \mathbb{1}\Big(p = \{y_j, y'_j\}\Big).$$

- We have to meet the constraint $\sum_{p \in \mathscr{A}} \beta(p)\, V_p(\ell) = 0$ for all $\ell \in \mathfrak{L}(\mathscr{C})$ the set of loops of $\mathscr{G}$ included in the support of P.

The free pairs

- Let $\{ V_p, p \in \mathscr{A} \setminus \mathscr{F} \}$ be a vector basis of
$$\mathbf{span}\{ V_p \in \mathbb{R}^{\mathfrak{L}(\mathscr{C})}, p \in \mathscr{A} \}.$$
- For any $p \in \mathscr{F}$, $V_p = -\sum_{q \in \mathscr{A} \setminus \mathscr{F}} e_{p,q} V_q$, for some matrix $e_{p,q}$, $p \in \mathscr{F}, q \in \mathscr{A} \setminus \mathscr{F}$.
- The constraint writes as $\displaystyle\sum_{q \in \mathscr{A} \setminus \mathscr{F}} \Big(\beta_q - \sum_{p \in \mathscr{F}} \beta_p e_{p,q} \Big) V_q = 0$
- and is equivalent to $\beta_q = \sum_{p \in \mathscr{F}} \beta_p e_{p,q}$, $q \in \mathscr{A} \setminus \mathscr{F}$.
- For any path $\pi_{x_C,x} \in \mathscr{G}(\mathscr{D}, \mathscr{A})$, joining $x_C \in C \in \mathscr{C}$ to x (so that $x \in C$), the energy function
$$U_p(\pi_{x_C,x}) = V_p(\pi_{x_C,x}) + \sum_{q \in \mathscr{A} \setminus \mathscr{F}} e_{p,q} V_q(\pi_{x_C,x}) = U_p(x)$$
depends only on x, because $U_p(\ell) = 0$ on loops $\ell \in \mathfrak{L}(\mathscr{C})$.

The Gibbs measure

- Therefore $P(x) = P(x_C) \exp\left(- \sum_{p \in \mathscr{F}} \beta_p U_p(x) \right)$

$$= \exp\left(- \sum_{C \in \mathscr{C}} \underbrace{\mathbb{1}(x \in C)}_{= U_C(x)} \underbrace{\log(1/P(x_C))}_{= \beta_C} - \sum_{p \in \mathscr{F}} \beta_p U_p(x) \right)$$

$$= \exp\left(- \sum_{i \in \mathscr{C} \cup \mathscr{F}} \beta_i U_i(x) \right).$$

- In this construction we get $Z_\beta = 1$.
- One can check that the converse is true:

  if $P(x) = Z_\beta^{-1} \exp\left(- \sum_{i \in \mathscr{C} \cup \mathscr{F}} \beta_i U_i(x) \right)$, where $Z_\beta < \infty$,
- then $P \in \mathfrak{M}_{\mathscr{C}}(\mathscr{D}, \mathscr{B})$. $\square$

A toy example

Recursive structures are possible

- Let $D = \{a, b, c\}$, $\mathscr{D} = D^+$,
- and $\mathscr{B} = \{\{a, ab\}, \{c, bc\}\}$.
- Consider $C_1 = \{ab^n c, n \in \mathbb{N}\}$,

  $C_2 = \{b^m cab^n, (m, n) \in \mathbb{N}^2\}$,

  $C_3 = \{b^k cab^m cab^n, (k, m, n) \in \mathbb{N}^3\}$.
- Remark that $C_j \in D^+/\!\sim_{\mathscr{B}}$, $1 \leq j \leq 3$.

The support may change the number of free pairs

- In C_1, the loop $ac \xrightarrow{(a,ab)} abc \xrightarrow{(bc,c)} ac$ is the only constraint,

$$\mathfrak{M}_{\{C_1\}}(D^+,\mathscr{B}) = \Big\{ P_r \in \mathscr{M}^1_+(C_1) : \\ P_r(ab^n c) = r(1-r)^n, \quad n \in \mathbb{N}, r \in]0,1[\Big\}.$$

- In C_2, there is no loop constraint, so that

$$\mathfrak{M}_{\{C_2\}}(D^+,\mathscr{B}) = \Big\{ P_{r,t} \in \mathscr{M}^1_+(C_2) : \\ P_{r,t}(b^m cab^n) = rt(1-r)^m(1-t)^n, \\ (m,n) \in \mathbb{N}^2, (r,t) \in]0,1[^2 \Big\}.$$

- In C_3, the loop constraint is the same as in C_1, so that

$$\mathfrak{M}_{\{C_3\}}(D^+,\mathscr{B}) = \Big\{ P_r \in \mathscr{M}^1_+(C_3) : \\ P_r\big(b^k cab^m cab^n\big) = r(1-r)^{k+m+n} \\ (k,m,n) \in \mathbb{N}^3, r \in]0,1[\Big\}.$$

The support may change the number of minimal pairs

- In $\mathfrak{M}_{\{C_3\}}(D^+, \mathscr{B})$, the set of substitute pairs $\mathscr{B}$ is minimal,
- whereas it is not in $\mathfrak{M}_{\{C_1\}}(D^+, \mathscr{B})$, indeed

$$\begin{aligned}\mathfrak{M}_{\{C_1\}}(D^+, \mathscr{B}) = \mathfrak{M}_{\{C_1\}}\Big(D^+, \{\{a, ab\}\}\Big) \\ = \mathfrak{M}_{\{C_1\}}\Big(D^+, \{\{c, bc\}\}\Big).\end{aligned}$$

Relation with Markov chains

or more accurately with Markov random fields

Markov chains are $\mathscr{B}$-Markov processes

- Consider a finite dictionary D, the domain $\mathscr{D} = D^L$
- and the substitute sets $\mathscr{B} = \left\{ \gamma(a, D, b), (a, b) \in D^2 \right\}$.
- The components of the state space are $D^L/\!\sim_{\mathscr{B}} = \{\gamma(a, D^{L-2}, b) : (a, b) \in D^2\}$.
- The model $\mathfrak{M}(D^L, \mathscr{B})$ contains the law of all time homogeneous Markov chains $(S_1, \ldots, S_L)$ with positive transition matrix M.

Relation with Markov chains

or more accurately with Markov random fields

Some $\mathscr{B}$-Markov models are Markov random fields

- Conversely for any process $S \sim P \in \mathfrak{M}(D^L, \mathscr{B})$,
- there is a time-homogeneous Markov chain $(X_1, \ldots, X_L)$ such that
- for any boundary conditions $(a, b)^2 \in D^2$ such that $\mathbb{P}(S_1 = a, S_L = b) > 0$,
- $\mathbb{P}_{S_2, \ldots S_{L-1} \,|\, S_1 = a, S_L = b} = \mathbb{P}_{X_2, \ldots X_{L-1} \,|\, X_1 = a, X_L = b}$.
- Moreover, the marginal distribution of the pair (S_1, S_L) can be arbitrary, while this is not the case for the distribution of (X_1, X_L).
- In other words, S is a one-dimensional Markov random field.

Simulating a $\mathscr{B}$-Markov process

Some Metropolis algorithm

- To simulate $P \in \mathfrak{M}_{\mathscr{C}}(\mathscr{D}, \mathscr{B})$, we need to know $P(C), C \in \mathscr{C}$
- and the substitute exponents, or equivalently $P(y)/P(x)$ for each $(x, y) \in \mathscr{G}(\mathscr{D}, \mathscr{B})$.
- Let $q(x, y)$ be a Markov kernel on $\mathscr{D} \times \mathscr{D}$ such that $\left\{(x, y) \in \mathscr{D}^2 : q(x, y) > 0\right\} = \mathscr{G}(\mathscr{D}, \mathscr{B}) \cup \left\{(x, x) : x \in \mathscr{D}\right\}$.
- Choose $x_C \in C$, $C \in \mathscr{C}$, and define the Markov kernel

$$\begin{cases} M(x, y) = q(x, y) \underbrace{\left(1 \wedge \frac{P(y) q(y, x)}{P(x) q(x, y)}\right)}_{\text{acceptance probability}}, & x \neq y \in \mathscr{D}, \\ M(x, x) = 1 - \sum_{y, y \neq x} M(x, y). & \end{cases}$$

- For any y in $\mathscr{D}$, $P(y) = \lim_{n \to \infty} \sum_{C \in \mathscr{C}} P(C) M^n(x_C, y)$

Replicated sample

- Consider some (deterministic) sample $(x_1, \ldots, x_n) \in \left(D^+\right)^n$.
- Take m copies $x_1, \ldots, x_N$, where $N = nm$.
- Let $\mu_N = \dfrac{1}{|\mathfrak{S}_N|} \displaystyle\sum_{\sigma \in \mathfrak{S}_N} \delta_{x \circ \sigma} \in \mathscr{M}_+^1\left(\left(D^+\right)^N\right)$ be the uniform measure on the permutations of the replicated sample.
- Let $p = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \delta_{x_i} \in \mathscr{M}_+^1\left(D^+\right)$ be the empirical measure of the original sample.
- Remark that μ_N is symmetric and consequently p-chaotic:
$$\lim_{N \to \infty} \int \varphi_1(x_1)\varphi_2(x_2)\mathrm{d}\mu_N(x_1, \ldots x_N) = \int \varphi_1(x_1)\mathrm{d}p(x_1) \int \varphi_2(x_2)\mathrm{d}p(x_2).$$

Conditions on the model

- Consider a substitute model $\mathfrak{M}_{\mathscr{C}}(\mathscr{D}, \mathscr{B})$ such that
- the domain contains the sample: $\{x_i, 1 \leq i \leq n\} \subset \mathscr{D}$,
- all members of substitute sets are present in the sample: $\displaystyle\sum_{i=1}^{n} \mathbb{1}\left(y \prec x_i\right) > 0$, for any $y \in B \in \mathscr{B}$, where $y \prec x$ means that y is a subsequence of x, or in other words that for some $(a, b) \in (D^*)^2$, $x = \gamma(a, y, b)$,
- all components of the support are present in the sample: $\mathscr{C} = \left\{C \in \mathscr{D}/{\sim_{\mathscr{B}}} : C \cap \{x_1, \ldots, x_n\} \neq \varnothing\right\}$.

Conditions on crossing-over dynamics

- Consider a Markov transition kernel $Q_N(x,y), x,y \in \mathscr{D}^N$, such that
- $$Q_N\Big[(\gamma(a,b,c),\gamma(a',b',c'),x_3,\dots,x_N);\\ (\gamma(a,b',c),\gamma(a',b,c'),x_3,\dots,x_N)\Big] > 0$$
  for any $(a,c) \in (D^*)^2$, $\{b,b'\} \subset B \in \mathscr{B}$ and $(x_3,\dots,x_N) \in \mathscr{D}$,
- Q_N is permutation invariant and symmetric: $Q_N(x \circ \sigma, y \circ \sigma') = Q_N(x,y) = Q_N(y,x)$, for any $x,y \in (D^+)^N$, and any $\sigma, \sigma' \in \mathfrak{S}_N$,
- Q_N is aperiodic, that will be the case for instance if $Q_N(x,x) > 0$, for any $x \in \mathscr{D}$.
- $$Q_N(x,y) > 0 \implies \sum_{j=1}^N U_i(x_j) = \sum_{j=1}^N U_i(y_j),\ i \in \mathscr{I}.$$

Crossing-over dynamics
and the maximum likehood estimator

Propagation of chaos

- Consider the empirical measure
$$M_N : x \in \mathscr{D}^N \mapsto M_N(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i} \in \mathscr{M}_+^1(\mathscr{D}).$$
- Let $\nu_{N,k} = \mu_N Q_N^k$ be the marginal of the crossing-over dynamics after k iterations,
- Let $\nu_N = \lim_{k\to\infty} \nu_{N,k}$. As Q_N is symmetric, ν_N is the uniform measure on its support.
- The law of the empirical measure $m_N = \nu_N \circ M_N^{-1}$ converges towards the likelihood estimator: $\lim_{N\to\infty} m_N = \delta_m$, where $m = \arg \max_{P \in \mathfrak{M}_{\mathscr{C}}\left(\mathscr{D},\mathscr{B}\right)} \prod_{i=1}^{n} P(x_i)$.
- Moreover ν_N is m-chaotic.

Elements of proof

Some combinatorics

- Since ν_N is uniform on its support and $m_N = \nu_N \circ M_N^{-1}$,

$$m_N(\rho) = Z_N^{-1} \frac{N!}{\prod_{x \in \mathscr{D}} (N\rho(x))!}$$

$$\asymp \exp\Big\{ N \Big[H(\rho) - \sup_{\rho' \in \mathrm{supp}(m_N)} H(\rho') \Big] \pm c \log(N) \Big\},$$

from Stirling's formula, where $H(\rho) = -\sum_{x \in \mathscr{D}} \rho(x) \log(\rho(x))$

is Shannon's entropy.

- Moreover $|\mathrm{supp}(m_N)| \leq N^{|\mathscr{D}|}$,
- implying that $\lim_{N \to \infty} m_N \Big(\arg \max_{\rho \in \mathrm{supp}(m_N)} H(\rho) \Big) = 1$.

The limit support

- Consider
$$\mathscr{Q} = \Big\{ \delta_{\gamma(a,b',c)} + \delta_{\gamma(a',b,c')} - \delta_{\gamma(a,b,c)} - \delta_{\gamma(a',b',c')}, \\ a, c, a', c' \in D^*, \{b, b'\} \subset B \in \mathscr{B}, \\ \gamma(a,b,c), \gamma(a,b',c), \gamma(a',b,c'), \gamma(a',b',c') \in \bigcup \mathscr{C} \Big\}.$$
- Remark that $\lim_{N\to\infty} \mathrm{supp}(m_N) = A = \Big\{ p + \sum_{\xi \in \mathscr{Q}} \alpha(\xi)\xi, \alpha \in \mathbb{R}^{\mathscr{Q}} \Big\} \cap \mathscr{M}_+^1(\mathscr{D})$ is a convex set.
- Let $m = \arg\max_{\rho \in A} H(\rho)$. One can prove that $\mathrm{supp}(m) = \bigcup \mathscr{C}$, and that $\dfrac{m(\gamma(a,b',c))}{m(\gamma(a,b,c))} = \dfrac{m(\gamma(a',b',c'))}{m(\gamma(a',b,c'))}$, under the same conditions as in the definition of $\mathscr{Q}$. This is a consequence of $\dfrac{\partial}{\partial\alpha}_{|\alpha=0} H(m + \alpha\xi) = 0$, and implies that $m \in \mathfrak{M}_{\mathscr{C}}(\mathscr{D}, \mathscr{B})$.

The maximum likelihood estimator

- Remark that $\int U_i(x)\,\mathrm{d}m(x) = \frac{1}{n}\sum_{j=1}^{n} U_i(x_j), i \in \mathscr{I}$, since for any $\xi \in \mathscr{Q}$ and any $i \in \mathscr{I}$, $\int U_i(x)\,\mathrm{d}\xi(x) = 0$.
- As we have seen that $m \in \mathfrak{M}_{\mathscr{C}}(\mathscr{D},\mathscr{B})$
- we decuce that m is the maximum likelihood estimator of the original sample $(x_1,\ldots,x_n)$,

$$m = \arg\max_{P\in\mathfrak{M}_{\mathscr{C}}(\mathscr{D},\mathscr{B})} \prod_{i=1}^{n} P(x_i).$$

Convergence of the empirical measure

- Since $\lim_{N\to\infty} m_N\Big(\arg\max_{\rho\in \mathrm{supp}(m_N)} H(\rho)\Big) = 1$, $\lim_{N\to\infty} \mathrm{supp}(m_N) = A$ and $m = \arg\max_{\rho\in A} H(\rho)$,
- $\lim_{N\to\infty} \int \mathbb{1}\Big(H(\rho) \leq H(m) - \eta\Big) \mathrm{d}m_N(\rho) = 0$, and consequently, H being strictly concave on A, a finite dimensional convex set,
- $\lim_{N\to\infty} \int \mathbb{1}\Big(|\rho - m| \geq \eta\Big) \mathrm{d}m_N(\rho) = 0$, $\eta > 0$, implying that $\lim_{N\to\infty} m_N = \delta_m$ and consequently that ν_N is m-chaotic.

$$\lim_{N\to\infty} \int \varphi_1(x_1)\varphi_2(x_2)\mathrm{d}\nu_N(x_1,\dots x_N) = \int \varphi_1(x_1)\mathrm{d}m(x_1) \int \varphi_2(x_2)\mathrm{d}m(x_2).$$

Conclusion

Summary

- We have a parametric model for some probability ratios
  $$\frac{P(\gamma(x,y,z))}{P(\gamma(x,y',z))} = \exp(\beta(y,y'))$$
- We get exponential families for any given support.
- The number of parameters is related to linear loop constraints.
- Crossing-over dynamics compute the maximum likelihood estimator "automatically", without requiring any explicit estimate of the substitute exponents.

Conclusion

Further questions

- We can use Context Free Grammars to describe substitute sets more efficiently.
- How can we compute an estimate of $P(x)$?
- How to select the model, that is how to choose the family $\mathscr{B}$ of substitute sets ?