Unsupervised statistical learning through label aggregation

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

Institute for Information Transmission Problems

Thursday, november 29, 2012

Natural language analysis : joint work with Thomas Mainguy,

Clustering in RKHS : joint work with Ilaria Giulini.

A new language model

Modeling multiple transmissions :

$$\{X_i^k, i = 1, \ldots, n\} \mapsto \widehat{\theta}(X^k) \mapsto X_i^{k+1} \mapsto \text{ etc.}$$

X^k i.i.d. does not work well : $\widehat{\theta}(X^k)$ will not stay close to $\widehat{\theta}(X^k)$ in the long run.
We assume instead that X^k is exchangeable : this allows for Markov dynamics with more invariant features.

Example : $X_i^k = X_{\sigma(i)}^{k-1}$, where σ is a random permutation.

A “cut and paste” model for language

Let the sample be made of sentences:

$$X_i^k \in D^+ \stackrel{\text{def}}{=} \bigcup_{j=1}^{\infty} D^j,$$

where D is some finite dictionary. We can equivalently represent the sample as

$$X^k = \sum_{i=1}^{n} \delta_{S_i^k}, \text{ where } S_i^k \in D^+.$$

Cutting and pasting with labels, an example :

$[_0$ This is my friend John . $\mapsto$ $[_0$ This is $]_1$ John . $\oplus$ $[_1$ my friend

More formally, let us introduce the set of symbols

$$S = D \cup \{[_i \, i \in \mathbb{N}\} \cup \{]_i, i \in \mathbb{N} \setminus \{0\}\},$$

and the set canonical expressions

$$\mathscr{E}_c = \left\{[_i \, a, i \in \mathbb{N}, a \in \left(D \cup \{]_j, j \in \mathbb{N} \setminus \{0\}\}\right)^+\right\}.$$

Let $\mathfrak{S}_n$ be the set of circular permutations of $\{1, \ldots, n\}$. Given a finite sequence $e = s_1, \ldots, s_\ell \in S^+$, let $\mathfrak{S}(e) = \{e' = s_{\sigma(1)}, \ldots, s_{\sigma(\ell)}, \sigma \in \mathfrak{S}_\ell\}$. Let the set of expressions be $\mathscr{E} = \bigcup_{e \in \mathscr{E}_c} \mathfrak{S}(e)$.

Let us define the set of toric grammars as

$$\mathfrak{G} = \{\mathscr{G} \in \mathscr{M}_+(\mathscr{E}), \mathscr{G}(e) = |\mathfrak{S}(e)|^{-1}\mathscr{G}(\mathfrak{S}(e)), |\mathrm{supp}(\mathscr{G})| < \infty\}.$$

(The set of positive measures on expressions that are invariant under circular permutations and have a finite support.)

Let the set of texts be

$$\mathfrak{T} = \{\mathscr{T} \in \mathfrak{G}, \mathscr{T} = \sum_{i=1}^{n} \sum_{e \in \mathfrak{S}(s_i)} \delta_e, \text{ where } s_i \in [_0\, D^+\}.$$

Introduce the short notations

$$r \otimes e = r \sum_{e' \in \mathfrak{S}(e)} \delta_{e'}, \qquad r \in \mathbb{R}$$

$$e \oplus e' = (1 \otimes e) + (1 \otimes e'),$$

$$\mathscr{G} \oplus e = \mathscr{G} + (1 \otimes e),$$

$$\mathscr{G} \oplus \mathscr{G}' = \mathscr{G} + \mathscr{G}',$$

and define split and merge grammar transformations as

$$\mathscr{G}' = \mathscr{G} \oplus a\,]_i \oplus [_i\, b \ominus ab, \qquad \text{split (or cut)},$$

$$\mathscr{G}' = \mathscr{G} \oplus ab \ominus a\,]_i \ominus [_i\ \ b, \qquad \text{merge (or paste)}.$$

Note that if we do not want to manipulate permutation invariant grammars, we can define them as positive measures on canonical expressions and describe splits and merges as

$$\mathscr{G}' = \mathscr{G} \oplus [_i\, a\,]_j\, c \oplus [_j\, b \ominus [_i\, abc,$$
$$\mathscr{G}' = \mathscr{G} \oplus [_i\, abc \ominus [_i\, a\,]_j\, c \ominus [_j\, b.$$

Generalization through label maps

If we use a new label for each split, sentences will be recombined through the action of merge transformations as they were at the beginning, producing no generalisation. In order produce new sentences, we need to introduce some label identifications. We will do this using label maps $f : \mathbb{N} \to \mathbb{N}$, such that $f(0) = 0$. Let $\mathfrak{F}$ be the set of label maps. They can be used to map non terminal symbols, and consequently also toric grammars, following the obvious rules

$$f([_i) = [_{f(i)}, \quad f(]_i) =]_{f(i)}, \quad f(w) = w, w \in D,$$
$$f(s_1 \cdots s_\ell) = f(s_1) \cdots f(s_\ell), \quad f(\mathscr{G}) = \mathscr{G} \circ f^{-1}.$$

$$\text{Let } \xi_{i,j}(k) = \begin{cases} \min\{i,j\}, & k \in \{i,j\}, \\ k, & \text{otherwise.} \end{cases}$$

Identification rule : $\mathscr{G}' = \xi_{i,j}(\mathscr{G})$, where indices i and j are such that, for some $a \in S^+$, $\mathscr{G}(a]_i)\mathscr{G}(a]_j) > 0$ or $\mathscr{G}([_i\, a)\mathscr{G}([_j\, a) > 0$. We can repeat this transformation as long as possible, and prove that the result does not depend on the order according to which elementary identifications are performed. We will call this result $\chi(\mathscr{G})$.

Production and learning

A production process P_t, $t = 1, \ldots, \sigma$ is a stopped Markov chain on $\mathfrak{G}$, where σ is a stopping time, such that P_t is obtained from P_{t-1} through a random merge operation, the stopping time σ being the time when no more merges are possible.

A learning process S_t, $1, \ldots, \tau$, is a stopped Markov chain on $\mathfrak{G}$ where S_t is obtain from S_{t-1} through a split transformation followed by a label identification made of some sequence of elementary identifications satisfying the above mentioned rule. The stopping time τ is the time when no more moves are possible.

Split and merge process

Once some reference grammar $\mathscr{R}$ has been learnt from an original text $\mathscr{T} \in \mathfrak{T}$, this text, as well as others may be split using the splitting rule

$$\mathscr{G}' = \mathscr{G} \oplus a\,]_i \oplus [_i\, b \ominus ab,$$

where $\mathscr{R}([_i\, b) > 0$. We will call a split process based on this splitting rule a narrow parse process. The stopping time here as usual is the time when no further moves are possible.

Given a reference grammar $\mathscr{R}$ and a starting text $\mathscr{T}$, and considering S, the narrow parse process with reference $\mathscr{R}$, a split and merge process G_t is a Markov chain on $\mathfrak{G}$ with transitions

$$\mathbb{P}\,G_{2t+1}|\,G_{2t} = \mathscr{G} = \mathbb{P}\,S_\tau|\,S_0 = \mathscr{G},$$
$$\mathbb{P}\,G_{2t}|\,G_{2t-1} = \mathscr{G} = \mathbb{P}\,P_\sigma|\,P_0 = \mathscr{G} \text{ and } P_\sigma \in \mathfrak{T}.$$

Theorem

The split and merge process is weakly reversible, in the sense that

$$\mathbb{P}\big(G_1 = \mathscr{G} \big| G_0 = \mathscr{T}\big) > 0$$
$$\iff \mathbb{P}\big(G_2 = \mathscr{T} \big| G_1 = \mathscr{G}\big) > 0 \text{ and } \mathscr{G} \in \bigcup_{t \in \mathbb{N}} \operatorname{supp}(\mathbb{P}_{G_{2t+1}})$$

Consequently

$$\mathbb{P}\big(G_2 = \mathscr{T}' \big| G_0 = \mathscr{T}\big) > 0 \iff \mathbb{P}\big(G_2 = \mathscr{T} \big| G_0 = \mathscr{T}'\big) > 0,$$
$$\mathbb{P}\big(G_3 = \mathscr{G}' \big| G_1 = \mathscr{G}\big) > 0 \iff \mathbb{P}\big(G_3 = \mathscr{G} \big| G_1 = \mathscr{G}'\big) > 0.$$

Invariant features

Along the split and merge process, the following quantities are invariant :

$$\mathscr{G}(wS^*), w \in D, \text{ the words counts,}$$
$$G(\mathscr{E}) - 2\mathscr{G}(\mathscr{E}_c), \text{number of symbols minus two times}$$
$$\text{the number of canonical expressions,}$$
$$\mathscr{G}([_0\, S^*), \text{ the number of global expressions.}$$

Proposition

When $\mathbb{P}_{G_t}(\mathscr{G}) > 0$,

$$\mathscr{G}(\mathscr{E}_c) - \mathscr{G}([_0\, S^*) \leq 2\big[\mathscr{T}(\mathscr{E}) - 2\mathscr{T}(\mathscr{E}_c)\big],$$

so that the reachable sets of the split and merge process are finite, moreover almost surely

$$\sigma \leq 2\big[\mathscr{T}(\mathscr{E}) - 2\mathscr{T}(\mathscr{E}_c)\big],$$
$$\tau \leq 2\big[\mathscr{T}(\mathscr{E}) - 2\mathscr{T}(\mathscr{E}_c)\big],$$

and this is also true for the learning process.

Consequently, the reachable sets of a split and merge process with reference grammar are finite and all states are positive recurrent. To each recurrent class corresponds a language model, given by the corresponding invariant probability measure. So although we use CF grammars with weights to produce texts, we do not define the generated language in the usual way !

Comparison with Markov models

To show the relationships and differences between this language model and the Markov chain model, let us show how one can describe the Markov model in terms of split process and label identifications.
Markov split :

$$\mathscr{G}' = \mathscr{G} \ominus [_0\, w_1 \cdots w_\ell \oplus \bigoplus_{j=1}^{\ell-1} [_{i_{j-1}}\, w_j]_{i_j} \oplus [_{i_{\ell-1}}\, w_\ell .$$

Markov label identification :

$$\mathscr{G}' = \xi_{i,j}(\mathscr{G}), \text{ where there is } w \in D : \mathscr{G}(S^* w]_i)\, \mathscr{G}(S^* w]_j) > 0,$$

so that $]_j$ is a function of the value of w_j.

Markov production chain : generates sentences from global expressions through the Markov chain with transitions

$$p\big([_0\, a]_{w_1}, [_0\, a w_2]_{w_2}\big) = \frac{\mathscr{G}\big([_{w_1}\, w_2]_{w_2}\big)}{\mathscr{G}\big([_{w_1}\, S^*\big)},$$

$$p\big([_0\, a]_{w_1}, [_0\, a w_2\big) = \frac{\mathscr{G}\big([_{w_1}\, w_2\big)}{\mathscr{G}\big([_{w_1}\, S^*\big)}.$$

Using successively this learning and this production scheme, we simulate from the Markov chain whose transitions are given by the empirical transition matrix computed on the text we started from.

We see that the toric grammar model is different in many ways :

- the splitting process is random and covers a much broader variety of possible parses,
- the label identification process uses global and forward and backward rules.

Small example

Training text :

```
This is a small example.
More experiments are needed.
This is a small English text.
This example shows that some redundent use of words
is needed.
```

New sentence produced by the model :

```
[0 This English text shows that some redundent use of
words is needed .
```

Reference grammar

```
40 [0 ]2 .
12 [1 ]4 a small ]3
6 [1 ]3 shows that some redundent use of words ]4 ]5
18 [2 This ]1
6 [2 More experiments are ]5
8 [2 ]10 ]4 a small ]3
8 [2 ]6 needed
10 [3 English text
20 [3 example
30 [4 is
12 [5 needed
4 [6 ]10 ]3 shows that some redundent use of words ]4
4 [6 More experiments are
12 [10 This
```

Toric grammars for two word sentences

Let us now focus on the special case where $\mathrm{supp}(\mathscr{T}) \subset [_0 D^2$. This the case of a language where all sentences have two words. There is only one splitting transform available for each sentence, and the identification of labels boils down to

$$[_0\, a\,]_i \oplus [_i\, b \leq \mathscr{G} \text{ and}$$

$$\Big([_0\, c\,]_j \oplus [_j\, b \leq \mathscr{G} \text{ or } [_0\, a\,]_j \oplus [_j\, c \leq \mathscr{G}\Big) \Longrightarrow \mathscr{G}' = \xi_{i,j}(\mathscr{G}) \text{ is allowed.}$$

In this special case it can be related to the following construction.

A model for label aggregation

Consider a couple of random variables $(X, Y) \sim \mathbb{P}_{X,Y}$ and the Markov chain $(\widetilde{X}_t, \widetilde{Y}_t)$ with transitions

$$\mathbb{P}_{\widetilde{X}_t \,|\, \widetilde{X}_j = x_j, \widetilde{Y}_j = y_j, j<t} = \mathbb{P}_{X \,|\, Y = y_{t-1}},$$
$$\mathbb{P}_{\widetilde{Y}_t \,|\, \widetilde{X}_t = x_t, \widetilde{X}_j = x_j, \widetilde{Y}_j = y_j, j<t} = \mathbb{P}_{Y \,|\, X = x_t}.$$

The joint distribution $\mathbb{P}_{\widetilde{X}_1, \widetilde{Y}_t}$ is a non linear functional of $\mathbb{P}_{X,Y}$. If $Y \in \mathscr{Y}$, a finite set of labels, then the Markov chain $\widetilde{Y}_t$ will have positive recurrent states devided into disjoint recurrent classes, inducing disjoint subsets of the pattern space $\mathscr{X}$.

In a more quantitative way, we may define a kernel as

$$K_t(x,x') = \frac{\mathrm{d}\mathbb{P}_{\widetilde{X}_1,\widetilde{X}_t}}{\mathrm{d}\mathbb{P}_X^{\otimes 2}}(x,x').$$

It is symmetric and positive, since it can be written as

$$K_{2t}(x,x') = \int \frac{\mathrm{d}\mathbb{P}_{\widetilde{X}_t|\widetilde{Y}_1=y}}{\mathrm{d}\mathbb{P}_X}(x)\frac{\mathrm{d}\mathbb{P}_{\widetilde{X}_t|\widetilde{Y}_1=y}}{\mathrm{d}\mathbb{P}_X}(x')\,\mathrm{d}\mathbb{P}_Y(y),$$

$$K_{2t+1}(x,x') = \int \frac{\mathrm{d}\mathbb{P}_{\widetilde{X}_t|\widetilde{X}_1=x''}}{\mathrm{d}\mathbb{P}_X}(x)\frac{\mathrm{d}\mathbb{P}_{\widetilde{X}_t|\widetilde{X}_1=x''}}{\mathrm{d}\mathbb{P}_X}\,\mathrm{d}\mathbb{P}_X(x'')$$

If we prefer a normalized kernel, we can also work with

$$\overline{K}_t(x,x') = \frac{K_t(x,x')}{\sqrt{K_t(x,x)K_t(x',x')}}.$$

In the ergodic case, the limit feature map maps all patterns to a single point. In the case when $\mathscr{Y}$ is finite, it is partitioned into a finite number of recurrent classes (because the chain $\widetilde{Y}_t$ is reversible) and the limit feature map maps all the patterns to a finite number of points of the corresponding RKHS.

Non supervised clustering on the sphere of a RKSH

Let us now start with a pattern sample $(X_i)_{i=1,\dots,n} \sim \mathbb{Q}^{\otimes n}$, $X_i \in \mathscr{X}$, some pattern space and let $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}_+$ be a normalized positive symmetric kernel, and $\Phi : \mathscr{X} \to \mathscr{H}$ the corresponding feature map.
We can compute the Principal Component Analysis of $\Phi(X)$ considering the problem

$$\sup_{\substack{\Lambda' \text{ orthogonal} \\ \text{normed sequence} \\ \text{of } \mathbb{L}^2(\mathscr{X},\mathbb{Q}), \\ |\Lambda'| \leq p}} \sum_{\mu \in \Lambda'} \int \mu(x) K(x,y) \mu(y) \, \mathrm{d}\mathbb{Q}(x) \mathrm{d}\mathbb{Q}(y) = \sum_{\mu \in \Lambda} s(\mu)^2.$$

We will talk later about the estimation problem due to the fact that we know $\overline{P} = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ instead of $\mathbb{Q}$. The principal normed vectors in $\mathscr{H}$ are the images of $v_\mu = s(\mu)^{-1} \mu$.

Let us put

$$K(y, v_\mu) = \langle \Phi(y), \Phi(v_\mu) \rangle = s(\mu)^{-1} \int \mu(x) K(x, y) \, \mathrm{d}\mathbb{Q}(x) \stackrel{\mathrm{def}}{=} y_\mu.$$

We can define a joint distribution on $\mathscr{X} \times (\Lambda \cup \{\Delta\})$, where Δ is some additional label, setting

$$\mathbb{P}_X = \mathbb{Q}, \qquad \mathbb{P}_{\mu|X} = x_\mu^2, \qquad \mathbb{P}_{\Delta|X} = 1 - \sum_{\mu \in \Lambda} x_\mu^2.$$

This gives

$$\mathbb{P}(\mu) = s(\mu)^2, \qquad \mathbb{P}(\Delta) = s(\Delta)^2 \stackrel{\mathrm{def}}{=} 1 - \sum_{\mu \in \Lambda} s(\mu)^2,$$

$$\frac{\mathrm{d}\mathbb{P}_{X|\mu}}{\mathrm{d}\mathbb{Q}}(x) = \frac{x_\mu^2}{s(\mu)^2}, \qquad \mu \in \overline{\Lambda} = \Lambda \cup \{\Delta\}, x_\Delta \stackrel{\mathrm{def}}{=} \sqrt{1 - \sum_{\mu \in \Lambda} x_\mu^2}.$$

Let us introduce $q \in \mathscr{M}_+^1(\overline{\Lambda} \times \overline{\Lambda})$, the joint distribution of pairs of labels

$$q(\mu, \nu) = \int x_\mu^2 x_\nu^2 \mathrm{d}\mathbb{Q}(x), \mu, \nu \in \overline{\Lambda}.$$

Let us put more generally $q^t(\mu, \nu) = \sum_{\xi \in \overline{\Lambda}} q^{t-1}(\mu, \xi) q(\nu | \xi)$, with the convention that $q^0(\mu, \nu) = s(\mu)^2 \delta_\mu(\nu)$. This leads to considering the following kernel that aggregates principal components, following the label aggregation scheme of the previous section:

$$\overline{K}(x, y) = \sum_{\mu, \nu} \frac{x_\mu^2}{s(\mu)^2} \frac{y_\nu^2}{s(\nu)^2} q^{t-1}(\mu, \nu).$$

Principal component smoothing

$$\overline{K}_{2t+1}(x,y) = \sum_{\mu,\nu,\xi\in\overline{\Lambda}} \frac{x_\mu^2}{s(\mu)^2} q^t(\mu|\xi) \frac{y_\nu^2}{s(\nu)^2} q^t(\nu|\xi) s(\xi)^2$$

$$= \sum_{\xi\in\overline{\Lambda}} \frac{\overline{x}_\xi^2 \overline{y}_\xi^2}{s(\xi)^2},$$

$$\text{where } \overline{x}_\xi \stackrel{\text{def}}{=} \left(\sum_{\mu\in\overline{\Lambda}} \frac{x_\mu^2}{s(\mu)^2} q^t(\mu|\xi) s(\xi)^2 \right)^{1/2}.$$

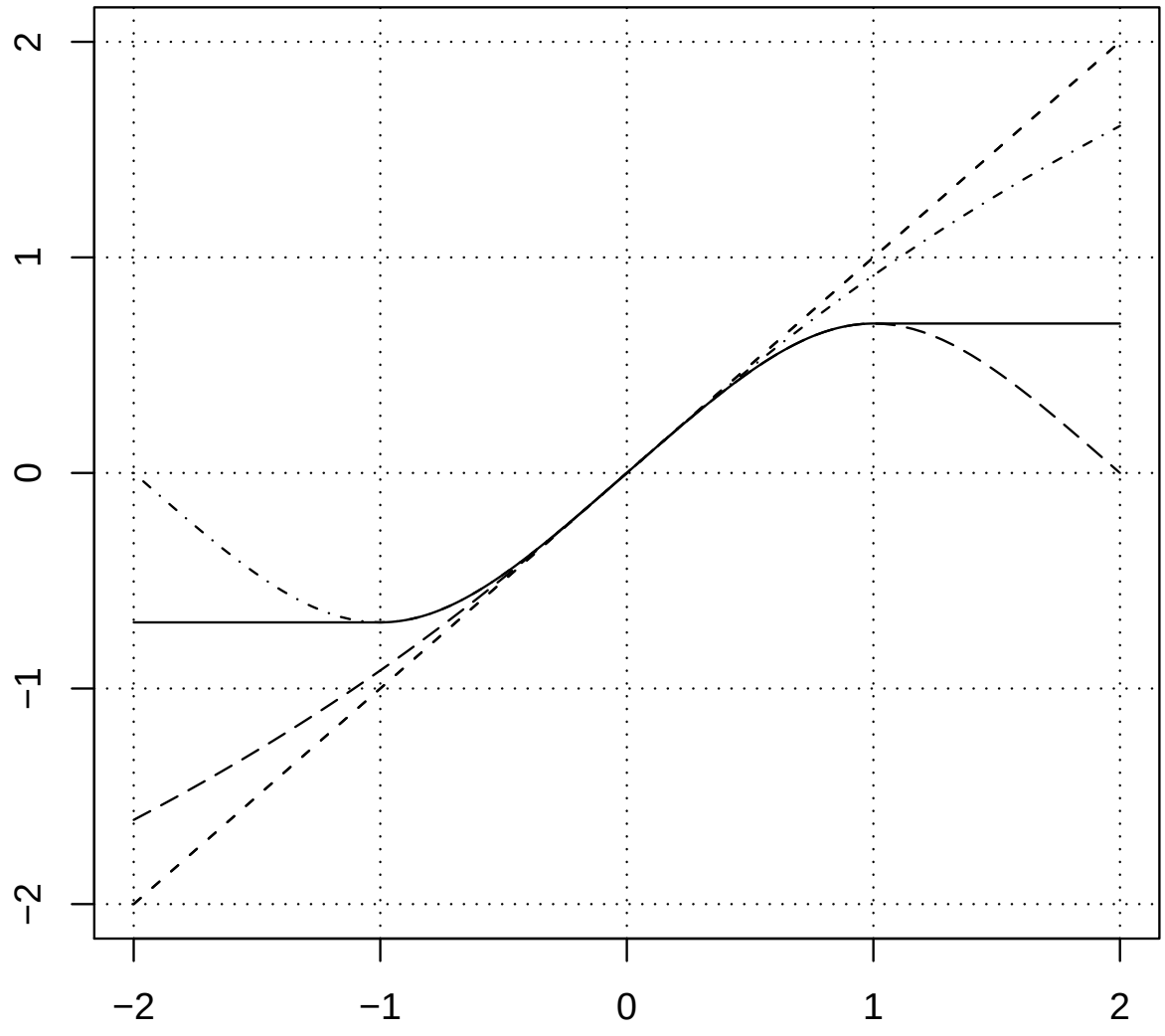
This bears an analogy to scattering.
In order to implement this method in a stable way, one has to solve the problem of estimating the first principal components in large or even infinite dimension.

Principal component analysis in large dimension

Let us assume that $X_i \in \mathbb{R}^d$, where d may be large and $X_i \sim \mathbb{Q}$.
Question : Estimate $\int \langle x, \theta \rangle^2 \, \mathrm{d}\mathbb{Q}(x)$ for all $\theta \in \mathbb{R}^d$, or equivalently all $\theta \in S_d$, the sphere of $\mathbb{R}^d$.
We will give both dimension dependent and dimension independent PAC-Bayes bounds.

Let us introduce the influence function

$$\psi(z) = \begin{cases} \log(2), & z \geq 1, \\ -\log(1 - z + z^2/2), & 0 \leq z \leq 1, \\ -\psi(-z), & z \leq 0. \end{cases}$$



$z \mapsto \psi(z)$, compared with $z \mapsto z$
$z \mapsto \log(1+z+z^2/2)$, and $z \mapsto -\log(1-z+z^2/2)$

It is symmetric, non decreasing, bounded and satisfies

$$-\log\bigl(\min\{\log(2), 1 - z + z^2/2\}\bigr) \le \psi(z) \le \log\bigl(\min\{\log(2), 1 + z + z^2/2\}\bigr), \qquad z \in \mathbb{R}.$$

Dimension dependent bounds

Let $\overline{\mathbb{P}} = \frac{1}{n}\delta_{X_i}$ and

$$r_\lambda(\theta) = \lambda^{-1} \int \psi\big[\lambda\big[\langle\theta, x\rangle^2 - 1)\big]\,\mathrm{d}\overline{\mathbb{P}}(x),$$

where $\lambda > 0$ will be chosen later. Remark that

$$\int \lim_{\lambda\to 0} r_\lambda(\theta)\,\mathrm{d}\mathbb{P}^{\otimes n} = N(\theta) - 1.$$

We may assume that $G \stackrel{\text{def}}{=} \int xx^t\,\mathrm{d}\mathbb{P}(x)$ is invertible, since $\mathbb{P}\big(X \in \mathbf{Im}(G)\big) = 1$ and we can replace $\mathbb{R}^d$ with $\mathbf{Im}(G)$. Let

$$\widehat{\alpha}(\theta) = \sup\big\{\alpha \in \mathbb{R}_+ : r_\lambda(\alpha\theta) \le 0\big\}, \qquad \widehat{N}(\theta) = \widehat{\alpha}^{-2}.$$

$$\text{Let } s_4 = \left(\int \| G^{-1/2} x \|^4 \, \mathrm{d}\mathbb{P}(x) \right)^{1/4},$$

$$\kappa = \sup \Big\{ \int \langle \theta, x \rangle^4 \, \mathrm{d}\mathbb{P}(x), \theta \in \mathbb{R}^d, \int \langle \theta, x \rangle^2 \, \mathrm{d}\mathbb{P}(x) = 1 \Big\},$$

$$\lambda = \sqrt{\frac{2}{(\kappa - 1)n} \left[\log(\epsilon^{-1}) + \frac{(1 + 18c) s_4^2}{4\sqrt{\kappa}(1 + 4c)} \right]},$$

$$\eta = (\kappa - 1)\lambda = \sqrt{\frac{2(\kappa - 1)}{n} \left[\log(\epsilon^{-1}) + \frac{(1 + 18c) s_4^2}{4\sqrt{\kappa}(1 + 4c)} \right]}$$

$$\gamma = 2\sqrt{\frac{(1 + 4c) s_4^2 \sqrt{\kappa}}{n}},$$

$$\mu = \gamma + \eta,$$

$$\xi = \frac{\kappa \eta}{2(\kappa - 1)}.$$

Proposition

If $2\mu + \xi < 1$, *which is the case when*

$$n > \left(4\kappa^{1/4} s_4 \sqrt{1+4c} + \left(2 + \frac{\kappa}{2(\kappa-1)} \right) \times \sqrt{2(\kappa-1)\left[\log(\epsilon^{-1}) + \frac{(1+18c)s_4^2}{4\sqrt{\kappa}(1+4c)} \right]} \right).$$

With probability at least $1-2\epsilon$, *for any* $\theta \in \mathbb{R}^d$,

$$\left| N(\theta)/\widehat{N}(\theta) - 1 \right| \leq \frac{\mu}{1-2\mu}.$$

Dimension free bounds

With probability $1-2\epsilon$, for any $\theta \in S_d$, and some estimator $\widehat{N}$ to be described in the proof,

$$\mathbb{1}\left(4\mu < 1\right)\left|N(\theta) - \widehat{N}(\theta)\right| \leq N(\theta)\frac{\mu}{1-4\mu},$$

where

$$\mu = \left(2\sqrt{\frac{(1+4c)s_4^2\sqrt{\kappa}}{nN(\theta)}} + \sqrt{\frac{2(\kappa-1)}{n}\left[\log\left(\frac{g+1}{\epsilon}\right) + \frac{(1+18c)s_4^2}{4(1+4c)N(\theta)\sqrt{\kappa}}\exp\left(\frac{\log(n)}{2g}\right)\right]}\right) \times \cosh\left(\frac{\log(n)}{2g}\right),$$

where $g \in \mathbb{N}$ is a grid parameter (you can take $g = n$ for instance) and this time $s_4 = \left(\int \|x\|^4 \mathrm{d}\mathbb{P}(x)\right)^{1/4}$.

Proof

Let this time $r_\lambda(\theta) = \int \psi(\langle\theta, x\rangle^2 - \lambda)\, \mathrm{d}\overline{\mathbb{P}}(x)$. Consider the Gaussian parameter perturbations $\pi_\theta = \mathcal{N}(\theta, \beta^{-1}\mathbb{I}_d)$, where $\mathbb{I}_d$ is the identity matrix of size $d \times d$. Let

$$\widehat{\alpha}(\theta) = \sup\{\alpha \in \mathbb{R}_+ : r_\lambda(\alpha\theta) \leq 0\}.$$

Proposition

Let $c = \frac{15}{\log(4)} \leq 11$.

$$\psi(\langle\theta, x\rangle^2 - \lambda) \leq \int \log\Bigg[1 + \langle\theta', x\rangle^2 - \lambda - \frac{\|x\|^2}{\beta} + \frac{1}{2}\left(\langle\theta', x\rangle^2 - \lambda - \frac{\|x\|^2}{\beta}\right)^2 + \frac{2c\|x\|^2}{\beta}\left(4\langle\theta', x\rangle^2 + \frac{5\|x\|^2}{\beta}\right)\Bigg] \mathrm{d}\pi_\theta(\theta').$$

Indeed,

$$\psi\Big(\int h\,\mathrm{d}\rho\Big) \leq \int \psi(h)\,\mathrm{d}\rho + \min\{\log(4), \mathbf{Var}(h\mathrm{d}\rho)\},$$

because $y \mapsto \psi(y) + \big(y - \int h\,\mathrm{d}\rho\big)^2$ is convex, since $\psi''(y) \geq -2$. As moreover

$$\langle\theta, x\rangle^2 - \lambda = \int \langle\theta', x\rangle^2\,\mathrm{d}\pi_\theta(\theta') - \lambda - \frac{\|x\|^2}{\beta},$$

we get

$$\psi(\langle\theta, x\rangle^2 - \lambda \leq \int \psi\Big(\langle\theta', x\rangle^2 - \frac{\|x\|^2}{\beta} - \lambda\Big)\,\mathrm{d}\pi_\theta(\theta') + \min\Big\{\log(4), \frac{4\|x\|^2\langle\theta, x\rangle^2}{\beta} + \frac{2\|x\|^4}{\beta^2}\Big\}$$

Lemma

If $W \sim \mathcal{N}(0, \sigma^2)$,

$$\min\{a, bm^2 + c\} \leq \mathbb{E}\Big(\min\{2a, 2b(m + W)^2 + 2b\sigma^2 + c\}\Big),$$

$$a, b, c \in \mathbb{R}_+, m \in \mathbb{R}.$$

The proof of this lemma is based on the inequalities $m^2 \leq 2(m + W)^2 + 2\,W^2$ and

$$\min\{a, y + z\} \leq \min\{a, y\} + \min\{a, z\}$$

$$\leq \min\{2a, y + z\}, \quad a, y, z \in \mathbb{R}_+.$$

Accordingly

$$\psi(\langle\theta,x\rangle^2-\lambda)\le\int\psi\bigg(\langle\theta',x\rangle^2-\frac{\|x\|^2}{\beta}-\lambda\bigg)\,\mathrm{d}\pi_\theta(\theta')$$
$$+\int\min\bigg\{4\log(2),\frac{8\|x\|^2\langle\theta',x\rangle^2}{\beta}+\frac{10\|x\|^4}{\beta^2}\bigg\}\,\mathrm{d}\pi_\theta(\theta').$$

We will now use

Lemma

For any $a, b, y, \in \mathbb{R}_+$ *and* $c = \frac{a}{b}\left[\exp(b) - 1\right]$,

$$\log(a) + \min\{b, y\} \leq \log(a + cy).$$

Applying this lemma to $a \leq 2$, $b = 4\log(2)$, and the corresponding $c = \frac{15}{\log(4)}$ ends the proof of the previous proposition.

PAC-Bayes bound

Proposition

For any measure $\nu \in \mathscr{M}_+^1(\Theta)$*, real number* $a > -1$ *and any measurable function* $f : \mathscr{X} \times \Theta \to [a, +\infty[$*, with probability at least* $1-\epsilon$*, for any* $\rho \in \mathscr{M}_+^1(\Theta)$ *such that* $\mathscr{K}(\rho, \nu) < \infty$*,*

$$\int \log\left[1+f(x,\theta)\right] \mathrm{d}\rho(\theta)\, \mathrm{d}\overline{\mathbb{P}}(x) \leq \int f(x,\theta)\, \mathrm{d}\rho(\theta)\, \mathrm{d}\mathbb{P}(x) + \frac{\mathscr{K}(\rho,\nu) - \log(\epsilon)}{n}.$$

The proof is based on the fact that

$$\int h \mathrm{d}\rho - \mathscr{K}(\rho,\nu) \leq \log\left(\int \exp(h) \mathrm{d}\nu\right),$$

for any upper-bounded measurable function $h : \Theta \to \mathbb{R}$.

We get with probability $1-\epsilon$, for any $\theta \in \mathbb{R}^d$,

$$\int \psi(\langle \theta, x\rangle^2 - \lambda)\, \mathrm{d}\overline{\mathbb{P}}(x) \leq \int \Bigg[\langle \theta, x\rangle^2 - \lambda + \frac{1}{2}\left(\left(\langle \theta, x\rangle^2 - \lambda\right)^2 + \frac{4}{\beta}\langle \theta, x\rangle^2 \|x\|^2 + \frac{2}{\beta^2}\|x\|^4 \right) + \frac{2c\|x\|^2}{\beta}\left(\frac{9\|x\|^2}{\beta} + 4\langle \theta, x\rangle^2 \right) \Bigg] \mathrm{d}\mathbb{P}(x) + \frac{\beta\|\theta\|^2}{2n} + \frac{\log(\epsilon^{-1})}{n}.$$

Using the Cauchy-Schwartz inequality will make the following quantities appear

$$
\begin{aligned}
s_4 &= \left(\int \|x\|^4 \, \mathrm{d}\mathbb{P}(x) \right)^{1/4}, \\
\kappa &= \sup \left\{ \int \langle \theta, x \rangle^4 \, \mathrm{d}\mathbb{P}(x), \theta \in \mathbb{R}^d \text{ s. t. } \int \langle \theta, x \rangle^2 \, \mathrm{d}\mathbb{P}(x) = 1 \right\}, \\
\xi &= \frac{\kappa \lambda}{2}, \\
\gamma &= \lambda(\kappa - 1) + \frac{2}{\beta}(1 + 4c) s_4^2 \sqrt{\kappa}, \\
\eta &= \frac{\lambda}{2}(\kappa - 1) + \frac{2}{\beta}(1 + 4c) s_4^2 \sqrt{\kappa} + \frac{(1 + 18c) s_4^4}{\beta^2 \lambda} - \frac{\log\left[\nu(\lambda, \beta)\epsilon\right]}{n\lambda}, \\
\delta &= \frac{\beta}{2n\lambda}.
\end{aligned}
$$

Proposition

With probability at least $1-2\epsilon$, *for any* $\theta \in \mathbb{R}^d$,

$$\frac{r_\lambda(\alpha\theta)}{\lambda} \leq \xi\left(\frac{N(\theta)}{\lambda}\alpha^2 - 1\right)^2 + (1+\gamma)\left(\frac{N(\theta)}{\lambda}\alpha^2 - 1\right) + \eta + \delta\|\theta\|^2\alpha^2,$$

$$\frac{r_\lambda(\alpha\theta)}{\lambda} \geq -\xi\left(\frac{\alpha^2 N(\theta)}{\lambda} - 1\right)^2 + \left(1-\gamma-\frac{\lambda\|\theta\|^2\delta}{N(\theta)}\right)\left(\frac{N(\theta)}{\lambda}\alpha^2 - 1\right) - \eta - \frac{\lambda\|\theta\|^2\delta}{N(\theta)}.$$

Let

$$\Phi_-(z) = z\left(1 - \frac{\eta + \frac{\delta\lambda}{z}}{1+\gamma-\eta-\frac{\delta\lambda}{z}}\right)\mathbb{1}\left(\xi - \gamma + \eta + \frac{\delta\lambda}{z} < 1\right),$$

$$\Phi_+(z) = z\left(1 + \frac{\eta + \frac{\delta\lambda}{z}}{1-\gamma-\eta-\frac{2\delta\lambda}{z}}\right)^{-1}\mathbb{1}\left(\xi + \gamma + \eta + \frac{2\delta\lambda}{z} < 1\right).$$

Replacing $N(\theta)$ with $\Phi_-\left(\lambda/\widehat{\alpha}^2\right)$ in the first inequality of the previous proposition and $\lambda/\widehat{\alpha}^2$ with $\Phi_+\left[N(\theta)\right]$ in the second inequality, we can prove that

$$\Phi_-\left(\frac{\lambda}{\widehat{\alpha}^2}\right) \leq N(\theta),$$

$$\Phi_+\left[N(\theta)\right] \leq \frac{\lambda}{\widehat{\alpha}^2}.$$

Proposition

With probability at least $1-2\epsilon$, *for any* $\theta \in \mathbb{R}^d$,

$$B_- = \sup_{\lambda,\beta} \Phi_-\left(\frac{\lambda}{\widehat{\alpha}^2}\right) \leq N(\theta) \leq \inf_{\lambda,\beta} \Phi_+\left(\frac{\lambda}{\widehat{\alpha}^2}\right) = B_+.$$

Let us put $\widehat{N}(\theta) = \dfrac{B_- + B_+}{2}$, we get with probability at least $1-2\epsilon$, for any $\lambda, \beta \in \mathbb{R}_+$,

$$N(\theta) - \widehat{N}(\theta) \leq \frac{N - B_-}{2} = \frac{N - \Phi_-\left(\lambda/\widehat{\alpha^2}\right)}{2} \leq \frac{N(\theta) - \Phi_- \circ \Phi_+\left[N(\theta)\right]}{2},$$

$$\widehat{N}(\theta) - N(\theta) \leq \frac{\Phi_+^{-1}\left(\lambda/\widehat{\alpha}^2\right) - N(\theta)}{2} \leq \frac{\Phi_+^{-1} \circ \Phi_-^{-1}\left[N(\theta)\right] - N(\theta)}{2}.$$

Putting $r(z) = \dfrac{\eta + \frac{\delta\lambda}{z}}{1-\gamma-\eta-\frac{2\delta\lambda}{z}}$, and $c(z) = \eta + \dfrac{\delta\lambda}{z}$, we can prove that

$$\Phi_-(z) \geq z\big[1-r(z)\big]\mathbb{1}\big[4c(z)<1\big],$$
$$\Phi_+(z) \geq z\big[1+r(z)\big]^{-1}\mathbb{1}\big[4c(z)<1\big]$$
$$\Phi_-^{-1}(z)\mathbb{1}\big[4c(z)<1\big] \leq z\big[1-r(z)\big]^{-1},$$
$$\Phi_+^{-1}(z)\mathbb{1}\big[4c(z)<1\big] \leq z\big[1+r(z)\big].$$

From this we can deduce

Proposition

With probablity at least $1-2\epsilon$, *for any* $\theta \in \mathbb{R}^d$,

$$\mathbb{1}\big[4c(N(\theta)) < 1\big]\Big|N(\theta) - \widehat{N}(\theta)\Big| \leq N(\theta)\frac{c\big(N(\theta)\big)}{1-4c\big(N(\theta)\big)}.$$

The announced result is then obtained by optimizing the value of $c\big(N(\theta)\big)$ by appropriate choices of λ and β.