About the non-asymptotic behaviour of Bayes estimators

Lucien Birgé
Sorbonne Universités – UPMC Université Paris 06
U.M.R. 7599, L.P.M.A.
F-75005, Paris, France

February 2014

Abstract

This paper investigates the nonasymptotic properties of Bayes procedures for estimating an unknown distribution from \( n \) i.i.d. observations. We assume that the prior is supported by a model \((\mathcal{S}, h)\) (where \( h \) is the Hellinger distance) with suitable metric properties involving the number of small balls that are needed to cover larger ones. We also require that the prior put enough probability on small balls but we do not assume that the true distribution belongs to the model which is therefore only viewed as an approximation of the truth. We only need to have a control of the Kullback-Leibler Information between the true distribution and an approximating one in the model in order to derive our risk bounds.

1 Introduction

The purpose of this paper is to derive in a simple way some non-asymptotic results about posterior distributions and Bayes estimators from a frequentist viewpoint, therefore offering a complementary point of view to the classical results by Ghosal, Ghosh and van der Vaart (2000). It can also be considered as a new presentation of Le Cam (1973 and 1982) with some extensions. In any case, it has been strongly influenced by these three papers.

We shall work here within the following framework: we have at disposal a sample \( \mathbf{X} = (X_1, \ldots, X_n) \) of size \( n \), the \( X_i \) being measurable functions on \((\Omega, \mathcal{A})\) with values in \((\mathcal{X}, \mathcal{X})\) and common unknown distribution \( P \). This distribution is an element of the metric space \((\mathcal{P}, h)\) of all probability measures on \( \mathcal{X} \) endowed with the Hellinger distance \( h \) given by

\[
h^2(R, T) = \frac{1}{2} \int \left( \sqrt{\frac{dR}{d\lambda}} - \sqrt{\frac{dT}{d\lambda}} \right)^2 d\lambda
\]
where $\lambda$ is an arbitrary positive measure which dominates both $R$ and $T$. We then introduce a model for $P$, i.e. a dominated family $\mathcal{F} = \{P_t \mid t \in S\} \subset \mathcal{P}$ of probabilities on $\mathcal{F}$ with densities $f_t = dP_t/d\mu$ with respect to some reference measure $\mu$ on $\mathcal{F}$. We assume that the mapping $t \mapsto P_t$ is one to one which allows us to systematically identify $S$ and $\mathcal{F}$, thus considering $S$ as a metric space with distance $h - h(t,u) = h(P_t,P_u)$ — and the corresponding Borel $\sigma$-algebra. We then introduce a prior distribution $\nu$ on $S$, turning $t$ into a random variable $t$. The prior $\nu$ and the sample $X$ give rise to a posterior distribution $\nu = \nu(X)$ and, given a loss function $w \circ h$ on $S \times S$, to the corresponding Bayes estimator $\tilde{s}$ which minimizes over $S$ the function 

\[ t \mapsto \mathbb{E}_{\nu}[w(h(t,t))]. \]

Our purpose here will be twofold. When $P = P_s$ truly belongs to $\mathcal{F}$ and the metric structure of $(S,h)$ is similar to that of a compact subset of some Euclidean space, we shall study the concentration rate of the posterior distribution $\nu(X)$ of $t$ around $P_s$. When $P$ does not belong to $\mathcal{F}$ or the metric structure of $S$ does not follow the previous requirements, we shall study the performance of the Bayes estimator(s) $P_s$ of $P$ defined via the loss function $w \circ h$.

The main feature of our approach is its non-asymptotic viewpoint, explicit deviation bounds being provided for fixed $n$.

## 2 A toy exemple

We shall consider first, in order to motivate our approach, the very particular situation of a finite or countable parameter set $S \subset \mathbb{N}$ with a prior $\nu$ such that $\nu_j = \nu(\{j\}) > 0$ for all $j \in S$ and assume moreover that the true distribution $P$ belongs to the corresponding model $\mathcal{F}$. Without loss of generality we may take $P = P_0$. Setting $x = (x_1, \ldots, x_n) \in \mathcal{F}^n$ and $f_t(x) = \prod_{i=1}^n f_t(x_i)$, we derive that the posterior probability is given by

\[ \nu(\{k\}) = \frac{\nu_k f_k(X)}{\sum_{j \in S} \nu_j f_j(X)} \quad \text{and} \quad \nu(B) = \frac{\sum_{j \in B} \nu_j f_j(X)}{\sum_{j \in S} \nu_j f_j(X)}. \]

It follows that if $B \ni 0$,

\[ \frac{1}{\nu(B)} = 1 + \frac{\sum_{j \in B^c} \nu_j f_j(X)}{\sum_{j \in B} \nu_j f_j(X)} \quad \text{hence} \quad \nu(B) \geq 1 - \frac{\sum_{j \in B^c} \nu_j f_j(X)}{\nu_0 f_0(X)}. \quad (2.1) \]

In order to control this last quantity we have to control the ratio $f_j(X)/f_0(X)$ which can be done via the following classical lemma:

**Lemma 1** Given $n$ i.i.d. random variables $X_1, \ldots, X_n$ with distribution $P$ and another distribution $Q$,

\[ P\left[ \sum_{i=1}^n \log \left( \frac{dQ}{dP}(X_i) \right) \geq y \right] \leq \exp \left[ -\frac{y}{2} \right] \rho^n(P,Q) \quad \text{with} \quad \rho(P,Q) = \int \sqrt{\frac{dP}{d\lambda} \frac{dQ}{d\lambda}} d\lambda. \]
We recall that \( \rho(P, Q) \) is called the *Hellinger affinity* between \( P \) and \( Q \), the definition being independent of the choice of the dominating measure \( \lambda \). It satisfies

\[
\rho(P, Q) = 1 - h^2(P, Q) \quad \text{and} \quad \rho(P^\otimes_n, Q^\otimes_n) = \rho^n(P, Q) \leq \exp\left[-nh^2(P, Q)\right]. \quad (2.2)
\]

It follows from this lemma that \( \mathbb{P}_0 \left[f_j(X) \geq e^{-2z} \nu_0 f_0(X)\right] \leq e^z \nu_0^{-1/2} \rho^n(P_0, P_j) \) and

\[
\mathbb{P}_0[X \in \Gamma] \leq e^z \nu_0^{-1/2} \sum_{j \in B^c} \rho^n(P_0, P_j) \quad \text{with} \quad \Gamma = \left\{ x \left| \sup_{j \in B^c} f_j(x) \geq e^{-2z} f_0(x) \right. \right\}.
\]

When \( X \not\in \Gamma \), (2.1) implies that \( \sigma(B) \geq 1 - e^{-2z} \nu(B^c) \geq 1 - e^{-2z} \) hence for \( z \) large enough the posterior will be concentrated on the set \( B \) with a probability at least

\[
\mathbb{P}_0[X \in B^c] \geq 1 - e^z \nu_0^{-1/2} \sum_{j \in B^c} \exp\left[-nh^2(P_0, P_j)\right].
\]

If we choose for \( B \) the ball centered at \( P_0 \) with radius \( k/\sqrt{n} \), \( k \geq 1 \) and if \( N_l \) denotes the number of points \( j \) in \( S \) such that \( l/\sqrt{n} \leq h(P_0, P_j) < (l + 1)/\sqrt{n} \), then

\[
\mathbb{P}_0[X \in B^c] \geq 1 - e^z \nu_0^{-1/2} \sum_{l \geq 1} N_l \exp\left[-l^2\right].
\]

If the series converge, the sum can be made arbitrarily small by taking \( k \) large enough but, in order that \( \mathbb{P}_0[X \in B^c] \) be close to one, \( k \) should also depend on \( z \) and \( \nu_0 \). We see that a very small value of \( \nu_0 \) leads to worse concentration performance of the posterior distribution.

3 Our assumptions

3.1 Definitions and notations

In the sequel, we shall always distinguish a special point \( s \in S \), whatever the situation of \( P \) with respect to the model \( \mathscr{X} \). We shall denote by \( \mathscr{B}(r) \), with \( r > 0 \), the Hellinger ball of center \( s \) and radius \( r \) in \( S \), by \(|N|\) the cardinality of the set \( N \) and by \( \mathbb{N}^* \) the set of positive integers \( \mathbb{N} \setminus \{0\} \). For any measurable subset \( B \) of \( S \) such that \( \nu(B) > 0 \), we define the density \( g_B \) with respect to \( \mu^\otimes_n \) and the probability \( P_B \) on \( \mathcal{X}^n \) by

\[
g_B(x) = \frac{1}{\nu(B)} \int_B f_t(x) \, d\nu(t) \quad \text{and} \quad P_B = g_B \cdot \mu^\otimes_n. \quad (3.1)
\]

We denote by \( P_B \) the probability on \( \Omega \) that gives \( X \) the distribution \( P_B \).

We also want to recall some well-known facts about metric entropies and related notions that can, for instance, be found in Kolmogorov and Tikhomirov (1961) or Lorentz (1966). Let \((M, d)\) be a metric space; a subset \( N \) of \( M \) is called \( x \)-separated (with \( x > 0 \)) if any two distinct points in \( N \) are at a distance larger than \( x \). Given a subset \( A \) of \( M \) we denote by \( N_A(x) \) the smallest number such that \( A \) can be covered by \( N_A(x) \) balls in \( M \) of radius \( x > 0 \).

**Lemma 2** Let \( A \) be a subset of a metric space \((M, d)\) and \( N \) be a maximal \( x \)-separated subset \( N \) of \( A \). Then \( N_A(x) \leq |N| \leq N_A(x/2) \).

3
3.2 Two assumptions

In our toy example, the performance of Bayes procedures depends on the amount of mass that the prior puts on the true distribution $P_0$ and on the number of points that are contained in Hellinger balls centered on it. In the general situation, we have to make similar assumptions on both the metric structure of $(S,h)$ and the way the measure $\nu$ behaves in vicinities of $s$.

**Assumption 1**

There exists a non-increasing function $D$ from $(0,1/4]$ to $[1, +\infty)$ such that, for all $x \in (0,1/4]$, any ball $B_{4x}$ with radius $4x$ in $S$ and any $x$-separated subset $N$ of $B_{4x}$, $|N| \leq \exp[D(x)]$.

This assumption holds for Euclidean spaces $\mathbb{R}^k$ with a function $D$ bounded by $k \log 9$ as can be checked in the following way: all balls of radius $x/2$ with centers in $N$ are disjoint and included in a ball of radius $9x/2$ so that their number is bounded by $9^k$. This result can be used when $S \subset \mathbb{R}^k$, which means that $\mathcal{S}$ is a parametric model, and there exists a polynomial relationship in $S$ between the Hellinger and Euclidean distances of the form

$$a\|t-u\|^\alpha \leq h(t,u) \leq A\|t-u\|^\alpha, \quad \text{for some } A > a > 0, \alpha > 0 \text{ and all } t, u \in S. \quad (3.2)$$

More generally, checking this assumption amounts to make entropy computations as shown by Lemma 2. Indeed, since

$$h(t,u) = \frac{1}{\sqrt{2}} \left\| \sqrt{f_t} - \sqrt{f_u} \right\|_2 \quad \text{with} \quad \|f-g\|_2^2 = \int (f-g)^2 d\mu,$$

one can use the mapping $t \mapsto \psi(t) = \sqrt{f_t}$ from $S$ into $L_2(\mu)$ to derive the entropy of $S$ from the entropy of $\psi(S)$ in $L_2(\mu)$. For suitable models $S$, the images $\psi(S)$ are compact subsets of function spaces (like Hölder and Besov spaces) for which these entropy counts are well-known, typically resulting in functions of the type $D(x) = cx^{-\alpha}$.

It also follows from Lemma 2 that Assumption 1 implies that $N_{B_{4x}}(x) \leq \exp[D(x)]$. More generally, iterating the procedure and taking into account the fact that $D$ is non-increasing so that $D(4x) \leq D(x)$, we see that $N_{S}(4^{-j}) \leq \exp[jD(4^{-j})]$ for all $j \geq 1$.

**Assumption 2**

Let $B(s)$ be the closed Hellinger ball with center $s$ and radius $r$ for a given point $s \in S$. There exists a non-increasing function $\beta$ from $(0,1/8]$ to $[1, +\infty)$ such that the following property holds for $\nu$ and some $\gamma \in [0, 2]$:

$$\nu(B(xr)) \leq \exp[x^\gamma \beta(r)] \nu(B(r)) \quad \text{for all } x \geq 8 \quad \text{and} \quad 0 < r \leq 1/8. \quad (3.3)$$

Here are some comments in order. First of all, if the assumption holds for some $\gamma < 2$ it also holds with $\gamma = 2$. On the one hand, since $x \geq 8$ and $\gamma \geq 0$ the right-hand side of (3.3) is not smaller than $\exp[8^\gamma \beta(r)] \nu(B(r))$ and the assumption is satisfied as soon as $\nu(B(r)) \geq \exp[-8^\gamma \beta(r)]$. On the other hand, setting $x = r^{-1}$, we see that (3.3) requires that $\nu(B(r)) \geq \exp[-r^{-\gamma} \beta(r)]$.

The Lebesgue measure on $\mathbb{R}^k$ with its Euclidean distance actually satisfies (3.3) with $\gamma = 1$ and $\beta(r) = (k/8) \log 8$ (among other possible choices) and this remains true for
the uniform probability on a convex subset \( S \) of \( \mathbb{R}^k \) with positive Lebesgue measure. It follows that if \( S \) is such a subset of \( \mathbb{R}^k \) and (3.2) holds, Assumption 2 also holds for \((S,h)\) and the uniform prior on \( S \) with a bounded function \( \beta \).

4 Concentration of the posterior distribution

We start with an important auxiliary result to be proved in Section 6 below.

**Theorem 1** Let \( \pi(X) \) be the posterior distribution of \( t \) and \( s \) be a given element of \( S \). Let Assumptions 1 and 2 hold and \( r > 0 \), \( J_0, n \in \mathbb{N}^* \) satisfy

\[
4^{J_0} nr^2 \geq \left[ 2^{(J_0+3)\gamma} \beta(r) \right] \sqrt{2D(2^{J_0+1}r)}. \tag{4.1}
\]

For all \( J \geq J_0 \), there exists some \( \Gamma_J \subset \mathcal{X}^n \) with

\[
P_{\mathcal{B}(r)}[\Gamma_J] = \int_{\Gamma_J} g_{\mathcal{B}(r)}(x) d\mu^{\otimes n}(x) \leq 1.05 \exp\left[ -4^J nr^2 / 2 \right],
\]
such that if \( X \notin \Gamma_J \), the posterior distribution \( \pi \) satisfies

\[
\pi\left[ \mathcal{B}(2^{J+1}r) \right] \geq 1 - 0.01 \exp\left[ -4^J nr^2 \right] \quad \text{for all } l \in \mathbb{N}.
\]

Let us now see how this applies to our problem for which the true distribution of \( X \) is not \( P_{\mathcal{B}(r)} \) but \( P^{\otimes n} \).

4.1 The finite dimensional case

We assume here that \( P = P_s \in \mathcal{P} \), that \( \gamma < 2 \) and that both \( D \) and \( \beta \) are bounded functions so that \( D(z) \leq \overline{D} \) and \( \beta(z) \leq \overline{\beta} \) for all \( z \). Let \( r = \exp(-z)/\sqrt{n} \) with \( z \geq 1 \). We choose for \( J_0 \) the smallest integer that satisfies (4.1) together with \( 4^{J_0} \geq 2ze^{2z} \), which implies that

\[
4^{J_0} \exp(-2z) \geq \left( 2^{(J_0+3)\gamma} \overline{\beta} \right) \sqrt{2\overline{D} \vee z}. \tag{4.2}
\]

This is possible for \( J_0 \) large enough since \( \gamma < 2 \). Since \( h(t,s) \leq \exp(-z)/\sqrt{n} \) for \( t \in \mathcal{B}(r) \),

\[
h^2(P_s^{\otimes n}, P_\varnothing^{\otimes n}) \leq \exp(-2z) \quad \text{by (2.2)}.
\]

We then use the following lemma to be proved in Section 6.

**Lemma 3** Let \( \lambda \) be a probability on \( S \) such that for some \( P \in \mathcal{P} \), \( \lambda\{t \mid h(P,P_t) \leq r \} = 1 \) and let \( P_\lambda \) be the distribution with density \( f_\lambda(x) = \int_S f_t(x) d\lambda(t) \) with respect to \( \mu \). Then, \( \rho(P,P_\lambda) \geq 1 - r^2 \) or, equivalently, \( h(P,P_\lambda) \leq r \).

The lemma shows that \( \rho(P_{\mathcal{B}(r)}, P_s^{\otimes n}) \geq 1 - \exp(-2z) \), hence by a classical result of Le Cam relating the variation distance to the Hellinger affinity,

\[
\sup_A \left[ P_{\mathcal{B}(r)}(A) - P_s^{\otimes n}(A) \right]^2 \leq 1 - r^2 \left( P_{\mathcal{B}(r)}, P_s^{\otimes n} \right) \leq 1 - (1 - \exp(-2z))^2 \leq 2 \exp(-2z). \tag{4.3}
\]

Then, by Theorem 1 and (4.2), apart when \( X \in \Gamma_0 \) with

\[
\mathbb{P}_s[X \in \Gamma_0] \leq 1.05 \exp\left[ -4^{J_0} e^{-2z} / 2 \right] + \sqrt{2} \exp(-z) < (3/2) \exp(-z),
\]

5
the posterior distribution $\nu(X)$ concentrates in the following way:

$$
\nu \left( 2^{J_0 + l + 2} e^{-z} / \sqrt{n} \right) \geq 1 - 1.01 \exp \left[ -4^{J_0 + l} e^{-2z} \right] \geq 1 - 1.01 \exp \left[ -4^{l+1/2} z \right]
$$

(4.4)

for all $l \in \mathbb{N}$. Fixing a value of $z$ which provides a suitable control of $P_s[X \in \Gamma_0]$ amounts to fixing the value of $J_0$ as the smallest integer satisfying (4.2) which does not depend on $n$. Then (4.4) provides concentration properties of the posterior distribution for balls with center $s$ and radius at least $2^{J_0 + 2} e^{-z} / \sqrt{n}$. This confirms that, in the finite dimensional case, the posterior concentrates around the true value $s$ at rate $n^{-1/2}$ (with respect to the Hellinger distance) as shown by Le Cam (1973) and Ibragimov and Has’minskii (1981). For parametric models with $S \in \mathbb{R}^k$ such that (3.2) holds, we can also derive concentration rates for the posterior with respect to the Euclidean distance.

### 4.2 The general case

If the requirements of Section 4.1 are not satisfied because $P \not\in \mathcal{S}$ or the function $D \vee \beta$ is unbounded in a vicinity of 0, we set $J_0 = 1$ and fix $r = r_n$ satisfying

$$
4n r_n^2 \geq \left[ 2^{l_J} \beta(r_n) \right] \vee \left[ 2D(4r_n) \right].
$$

(4.5)

If $D \vee \beta$ is unbounded $nr_n^2$ goes to infinity with $n$. Once $r_n$ has been fixed, we choose an integer $J \geq 1$ and get from Theorem 1 that, except on a set $\Gamma_J$ with $P_{\mathcal{S}(r_n)}[\Gamma_J] \leq 1.05 \exp \left[ -4^l r_n^2 / 2 \right]$, the posterior distribution $\nu(X)$ satisfies

$$
\nu \left( 2^{J_0 + l + 2} r_n \right) \geq 1 - 1.01 \exp \left[ -4^{J_0 + l} r_n^2 \right]
$$

for all $l \in \mathbb{N}$. Slightly increasing $r_n$ if necessary, we may actually choose $2^J \sqrt{n} r_n = k_n$ as an integer and, since by (4.5) $nr_n^2 \geq 1/2$, $k_n \geq 2$. Then, except on the set $\Gamma_J$ with

$$
P_{\mathcal{S}(r_n)}[\Gamma_J] \leq 1.05 \exp \left[ -k_n^2 / 2 \right],
$$

the posterior distribution $\nu(X)$ satisfies

$$
\nu \left( 2^{J_0 + l + 2} k_n / \sqrt{n} \right) \geq 1 - 1.01 \exp \left[ -4^l k_n^2 \right]
$$

for all $l \in \mathbb{N}$. Unfortunately, in such a case, we cannot use the previous recipe in order to bound $|P_{\mathcal{S}(r_n)}[\Gamma_J] - P_{\mathcal{S}_n}(\Gamma_J)|$ as we did in (4.3) and need to use another trick.

### 5 Convergence of Bayes estimators

#### 5.1 A preliminary result

Let us consider the following situation. We are given a probability $Q$ on some metric space $(\mathcal{S}, d)$ (with its Borel $\sigma$-algebra) with diameter $\sup_{t,u \in \mathcal{S} \times \mathcal{S}} d(t, u) \leq 1$. We assume
that \(Q\) satisfies the following concentration property around a special point \(s \in S\) with \(\eta = 2^{-J}\) for some \(J \in \mathbb{N}^*\):

\[
Q[\mathcal{B}(s,2^j\eta)] \geq 1 - a \exp \left( -B4^j \right) \quad \text{for } 0 \leq j \leq J.
\]  

(5.1)

with

\[
a > 0, \quad B \geq 2, \quad \text{and} \quad a e^{-B} \leq 1/5.
\]

Then \(Q[\mathcal{B}(s,\eta)] \geq 4/5\) and the following result holds.

**Proposition 1** Let \(w\) be some nondecreasing function on \([0,1]\) with the following properties:

\[
w(0) = 0 \quad \text{and} \quad x^\delta w(\eta) \leq w(x\eta) \leq a' \exp \left( B' x^2 \right) w(\eta) \quad \text{for } 2 \leq x \leq \eta^{-1},
\]

(5.2)

with \(\delta, \ a', \ B' > 0\). If \(t\) is distributed according to some \(Q\) which satisfies (5.1) and \(B' \leq (B-1)/4\), any minimizer \(\bar{t}\) of the function \(t \mapsto \mathbb{E}[w(d(t,t))]\) satisfies

\[
d(\bar{t},s) \leq \left( \left( \frac{5}{4} \right) (1 + aa'/2) \right)^{1/\delta} + 1 \eta.
\]

**Proof:** Let us first evaluate \(\mathbb{E}[w(d(t,s))]\). Since \(B \geq 4B' + 1\),

\[
\mathbb{E}[w(d(t,s))] \leq w(\eta) + \sum_{j=0}^{J-1} w(2^j+1\eta) \mathbb{P}[2^j \eta < d(t,s)] \leq 2^{j+1} \eta
\]

\[
\leq w(\eta) + a \sum_{j=0}^{J-1} w(2^{j+1}\eta) \exp \left( -B4^j \right)
\]

\[
\leq w(\eta) \left[ 1 + aa' \sum_{j=0}^{J-1} \exp \left( -B4^j + B'4^{j+2} \right) \right]
\]

\[
\leq w(\eta) \left[ 1 + aa' \sum_{j=0}^{J-1} \exp \left( -4^j \right) \right] \leq (1 + aa'/2)w(\eta).
\]

Now, if \(t_1 \in S\) with \(d(t_1,s) \geq [2^j + 1] \eta\) and \(d(t,s) \leq \eta\), then \(d(t,t_1) \geq 2^j \eta\) hence,

\[
\mathbb{E}[w(d(t,t_1))] \geq w(2^j \eta) Q[\eta \leq t] > (4/5)2^{j\delta}w(\eta) \quad \text{if } d(t_1,s) \geq [2^j + 1] \eta.
\]

It follows that \(t_1\) cannot be a minimizer of the function \(t \mapsto \mathbb{E}[w(d(t,t))]\) as soon as

\[
(4/5)2^{j\delta} > (1 + aa'/2) \quad \text{or} \quad 2^j > \left( \frac{5}{4} \right) (1 + aa'/2) \right)^{1/\delta}
\]

and the conclusion follows. \(\square\)

It is immediate to check that functions \(w(z)\) of the form \(z^{\delta}\) satisfy (5.2) with \(B' = 1/4 \leq (B-1)/4\) (since \(B \geq 2\)) if \(\delta \log x \leq \log a' + x^2/4\) for \(x \geq 2\), which holds when

\[
a' = \sup_{x \geq 2} \exp \left( \delta \log x - x^2/4 \right).
\]

But there are other cases.
Proposition 2 The function \( w(z) = \exp[\theta z^\delta] - 1 \) with \( \theta > 0 \) and \( 0 < \delta \leq 2 \) satisfies (5.2) with \( a' = B' = \theta \eta^\delta \).

Proof: On the one hand,
\[
w(\theta \eta) = \exp \left[ \theta x^\delta \eta^\delta \right] - 1 \geq x^\delta \left( \exp \left[ \theta \eta^\delta \right] - 1 \right) = x^\delta w(\eta),
\]
since \( y(z^2 - 1) \leq e^{yz} - 1 \) for \( z \geq 0 \) and \( y \geq 1 \). On the other hand, since \( w(\eta) > \theta \eta^\delta \), it is enough to prove that \( \exp[\theta(x^\delta \eta^\delta)] - 1 \leq a' \theta \eta^\delta \exp[B'x^2] \). Expanding both sides in series we see that the inequality is satisfied provided that \( (\theta(x^\delta \eta^\delta))^k \leq a' \theta \eta^\delta (B'x^2)^k \)
for all \( k \geq 1 \) or equivalently \( 1 \leq a'B'x^{(2-\delta)k} (B'\theta^{-1}\eta^{-\delta})^{k-1} \) which holds in view of our assumptions on \( \delta, \theta, a' \) and \( B' \). \[\square\]

5.2 Application to Bayesian estimators

The previous arguments can in particular be applied to Bayesian estimators in the context of Section 4.2. We know that, in this case, when \( X \) is distributed according to \( P_{\theta(\eta)} \), with probability at least \( 1 - 1.05 \exp[-k_n^2/2] \), the posterior distribution \( \pi(X) \) satisfies the assumptions (5.1) with \( \eta = 4k_n/\sqrt{n} \), \( a = 1.01 \) and \( B = k_n^2 \). Therefore, according to the previous section, for any loss function \( w \) that satisfies (5.2) with \( B' \leq (k_n^2 - 1)/4 \), the corresponding Bayes estimator \( \tilde{s}(X) \) satisfies, since \( k_n = 2J \sqrt{n}r_n \),
\[
h \left( \tilde{s}(X), s \right) \leq 4 \left( [(5/4)(1 + aa'/2)]^{1/\delta} + 1 \right) k_n/\sqrt{n} = \Delta 2^J r_n,
\]
with \( \Delta = 4 \left( [(5/4)(1 + aa'/2)]^{1/\delta} + 1 \right) \), hence
\[
P_{\theta(\eta)} \left[ h \left( \tilde{s}(X), s \right) > \Delta 2^J r_n \right] \leq 1.05 \exp \left[ -4^J nr_n^2/2 \right] \quad \text{for all} \quad J \geq 0
\]
and
\[
P_{\theta(r_n)} \left[ h^2 \left( \tilde{s}(X), s \right) > z \right] \leq 1.05 \exp \left[ -nz/(8\Delta^2) \right] \quad \text{for all} \quad z \geq (\Delta r_n)^2.
\]

This applies when \( w(z) = z^\delta \) since then one can take \( B' = 1/4 \) but also for losses of the type considered in Proposition 2 provided that \( B' = \theta(4k_n)^{-n^{-\delta}/2} \leq (k_n^2 - 1)/4 \). This always holds for \( n \) large enough (depending on \( \theta \) and \( \delta \)) since \( 0 < \delta \leq 2 \).

Recalling that the Kullback-Leibler divergence between two probabilities \( P, Q \) is given by
\[
K(P, Q) = \int \log \left( \frac{dP}{dQ} \right) dP \in [0, +\infty] \quad \text{if} \quad P \ll Q \quad \text{and} \quad K(P, Q) = +\infty \quad \text{otherwise},
\]
we may now use the following result which is a suitable version due to Yannick Baraud of an old result of Andrew Barron:

Proposition 3 For any pair \( Q, R \) of probabilities such that \( K(R, Q) < +\infty \) and any random variable \( T \) such that \( P_Q[T \geq z] \leq ae^{-bz} \) for all \( z \geq z_0 \geq 0 \) with \( a, b > 0 \), we have
\[
E_R[T] \leq z_0 + b^{-1} \left( 1 + A + \sqrt{2A} \right) \quad \text{with} \quad A = \log \left( 1 + ae^{-b z_0} \right) + K(R, Q).
\]
We shall apply it with \( Q = P_{\mathcal{B}(r_n)} \) and \( R = P^\otimes n \), \( T = h^2(\tilde{s}(X), s) \), \( z_0 = (\Delta r_n)^2 \), \( a = 1.05 \) and \( b = n/(8\Delta^2) \). This leads to

\[
\mathbb{E} \left[ h^2(\tilde{s}(X), s) \right] \leq (\Delta r_n)^2 + (8\Delta^2) n^{-1} \left( 1 + A + \sqrt{2A} \right)
\]

with

\[
A = \log \left( 1 + 1.05 \exp \left[ -nr_n^2/8 \right] \right) + K \left( P^\otimes n, P_{\mathcal{B}(r_n)} \right)
\]

\[
< 1.05 \exp \left[ -nr_n^2/8 \right] + K \left( P^\otimes n, P_{\mathcal{B}(r_n)} \right).
\]

Since \( K(P, Q) = \int \log(dP/dQ)(t) dP \) and \( -\log \) is a convex function,

\[
K \left( P, \sum_{i=1}^m \alpha_i Q_i \right) \leq \sum_{i=1}^m \alpha_i K(P, Q_i) \quad \text{if} \quad \sum_{i=1}^m \alpha_i = 1 \quad \text{and} \quad \alpha_i \geq 0 \quad \text{for all} \ i.
\]

It follows that

\[
K \left( P^\otimes n, P_{\mathcal{B}(r_n)} \right) \leq \frac{1}{\nu(\mathcal{B}(r_n))} \int_{\mathcal{B}(r_n)} K \left( P^\otimes n, P_t^n \right) d\nu(t)
\]

\[
= \frac{n}{\nu(\mathcal{B}(r_n))} \int_{\mathcal{B}(r_n)} K(P, P_t) d\nu(t).
\]

Therefore, if \( \sup_{t \in \mathcal{B}(r_n)} K(P, P_t) \leq \kappa_n^2 \), we derive that \( K \left( P^\otimes n, P_{\mathcal{B}(r_n)} \right) \leq n\kappa_n^2 \) and by \( (5.3) \),

\[
\mathbb{E} \left[ h^2(\tilde{s}(X), s) \right] \leq C \left( r_n^2 + \kappa_n^2 \right),
\]

where the constant \( C \) depends on all other constants involved in the assumptions. It should be noticed that, due to the regularization properties of the averaging operation,

\[
\frac{1}{\nu(\mathcal{B}(r_n))} \int_{\mathcal{B}(r_n)} K(P, P_t) d\nu(t)
\]

can be substantially smaller than \( \sup_{t \in \mathcal{B}(r_n)} K(P, P_t) \).

The control of \( K(P, P_t) \) can derive, in the case of bounded likelihood ratios, from the following result that we include for the sake of completeness.

**Lemma 4** For any two probabilities \( P \) and \( Q \),

\[
K(P, Q) \geq -2 \log[r(P, Q)] \geq 2h^2(P, Q).
\]

Moreover, if \( \sup_x (dP/dQ)(x) = M < +\infty \),

\[
2 \leq \frac{K(P, Q)}{h^2(P, Q)} \leq K(M) \quad \text{with} \quad K(x) = \frac{x(x \log x - 1) + 1}{(x + 1)/2 - \sqrt{x}}.
\]

In particular, \( K(x) \leq 4 + 2\log x \) for \( x \geq 1 \).
Proof: To prove (5.4) we use Jensen’s Inequality,

\[-\frac{1}{2} K(P, Q) = \int \log \left( \sqrt{\frac{dQ}{dP}} \right) dP \leq \log \left( \int \sqrt{dQ dP} \right) = \log[\rho(P, Q)],\]

and \( \log (1 - h^2(P, Q)) \leq -h^2(P, Q) \). Let \( \eta = dP/dQ \) so that \( \int \eta dQ = 1 \). Then

\[
K(P, Q) = \int \log(\eta) dP = \int \eta \log(\eta) dQ = \int (\eta [\log(\eta) - 1] + 1) dQ = \int K(\eta) \left[ (\eta + 1)/2 - \sqrt{\eta} \right] dQ,
\]

while

\[
h^2(P, Q) = 1 - \int \sqrt{\eta} dQ = \int \left[ (\eta + 1)/2 - \sqrt{\eta} \right] dQ.
\]

Since the function \( K \) is continuous and increasing on \( \mathbb{R}_+ \) with \( K(0) = 2 \), it follows that \( 2 \leq K(\eta) \leq K(M) \) hence (5.5). The bound on \( K \) follows from calculus. \( \square \)

6 Proofs

6.1 Proof of Lemma 3

We may assume, without loss of generality, that \( \mu \) dominates \( P \) with \( dP/d\mu = g \). Since \( \lambda \) is a probability, Jensen’s Inequality implies that

\[
\sqrt{f_{\lambda}(x)} \geq \int_{S} \sqrt{f_{t}(x)} d\lambda(t).
\]

If \( h(P, P_s) \leq r \), then \( \rho(P, P_s) \geq 1 - r^2 \) and, according to Fubini’s Theorem,

\[
\rho(P, P_\lambda) = \int_{X} \sqrt{g(x)f_{\lambda}(x)} d\mu(x) \geq \int_{X} \sqrt{g(x)} \left( \int_{S} \sqrt{f_{t}(x)} d\lambda(t) \right) d\mu(x) = \int_{S} d\lambda(t) \int_{X} \sqrt{g(x)f_{t}(x)} d\mu(x) = \int_{S} \rho(P, P_t) d\lambda(t) \geq 1 - r^2. \quad \square
\]

6.2 Likelihood ratio tests

Our results are based on the following fundamental result which is due to Le Cam and can be found in his book, Le Cam (1986, Lemma 2 page 477).

Proposition 4 Let \( \{P_t, t \in S\} \) be a dominated family of probability measures on \( \mathcal{X} \) with densities \( f_t \) with respect to some measure \( \mu \). We consider two distributions \( Q_1, Q_2 \) on \( \mathcal{X} \), two positive numbers \( r_1, r_2 \) such that \( h(Q_1, Q_2) > r_1 + r_2 \) and two probability
distributions $\lambda_1, \lambda_2$ on $S$ such that $\lambda_j \{t \mid h(P_t, Q_j) \leq r_j\} = 1$ for $j = 1, 2$. We define the distributions $R_1, R_2$ on $\mathcal{X}^n$ by their densities with respect to $\mu^\otimes n$,

$$\frac{dR_j}{d\mu^\otimes n}(x) = \int_S f_t(x) d\lambda_j(t) \quad \text{for all } x \in \mathcal{X}^n \text{ and } j = 1, 2 \quad \text{with } f_t(x) = \prod_{i=1}^n f_i(x_i).$$

Then

$$\rho(R_1, R_2) \leq \left[ 1 - (h(Q_1, Q_2) - r_1 - r_2)^2 \right]^n.$$

**Corollary 1** Let $\{P_t, t \in S\}$ be a dominated family of probability measures on $\mathcal{X}$ with densities $f_t$ with respect to some measure $\mu$ and let $\nu$ be a probability on $S$. Let $B_1$ and $B_2$ be two measurable subsets of $S$ such that $\nu(B_1)\nu(B_2) > 0$ and $B_j \subset B_j$ for $j = 1, 2$ where $\mathcal{B}_1$ and $\mathcal{B}_2$ are two Hellinger balls in $S$ such that $h(\mathcal{B}_1, \mathcal{B}_2) > 0$. Then, given a random element $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$ with density $g_{B_1}$ with respect to $\mu^\otimes n$ given by (3.1),

$$\mathbb{P}_{B_1} \left[ \log \left( \frac{g_{B_2}(X)}{g_{B_1}(X)} \right) \geq y \right] \leq e^{-y/2} \rho(P_{B_1}, P_{B_2}) \leq \exp \left[ -\frac{y}{2} - nh^2(\mathcal{B}_1, \mathcal{B}_2) \right] \quad \text{for all } y \in \mathbb{R}. \quad (6.1)$$

**Proof:** The first inequality is a classical application of the exponential inequality:

$$\mathbb{P}_R \left[ \log \left( \frac{dQ}{dR}(X) \right) \geq y \right] \leq \mathbb{E}_R \left[ e^{-y/2} \sqrt{\frac{dQ}{dR}(X)} \right]$$

and the second follows from Proposition 4 applied with $\lambda_j = P_{B_j}$ and $\mathcal{B}_j = \{t \mid h(P_t, Q_j) \leq r_j\}$ so that $h(\mathcal{B}_1, \mathcal{B}_2) \geq h(Q_1, Q_2) - r_1 - r_2$. \hfill \Box

### 6.3 Proof of Theorem 1

The posterior distribution $\mathcal{P}(A)$ of a subset $A \supseteq \mathcal{B}(r)$ of $S$ is given by

$$\frac{1}{\mathcal{P}(A)} = \int_S f_t(X) d\nu(t) \left[ \int_A f_t(X) d\nu(t) \right]^{-1}$$

$$\leq 1 + \int_{A^c} f_t(X) d\nu(t) \left[ \int_{\mathcal{B}(r)} f_t(X) d\nu(t) \right]^{-1},$$

so that

$$\mathcal{P}(A) \geq 1 - \int_{A^c} f_t(X) d\nu(t) \left[ \nu(\mathcal{B}(r)) \right]^{-1} g_{\mathcal{B}(r)}(X)^{-1}.$$ 

Setting $R = 2^{j+2} r$ and $A = \mathcal{B}(R)$, we can partition $A^c$ into sets $F_{j,k}$, $j \geq 1, 1 \leq k \leq K_j$ in the following way. We set $F_j = \mathcal{B}(2^j r) \setminus \mathcal{B}(2^{j-1} r)$ and consider a maximal $(2^{j-2} r)$-separated subset $N_j$ of $F_j$, hence of $\mathcal{B}(2^j r)$. By Assumption 1,

$$K_j = |N_j| \leq \exp \left[ D(2^{j-2} r) \right] = \exp \left[ D(2^{j+1} r) \right]. \quad (6.2)$$
This induces a covering of \( F_j \) by \( K_j \) balls \( B_{j,k}, 1 \leq k \leq K_j \) of radius \( 2^{j-2}R \) and centers in \( N_j \) from which one can deduce a partition of \( F_j \) into \( K_j \) subsets \( F_{j,k} \), each one being included in the corresponding ball \( B_{j,k} \) with center in \( F_j \) so that \( h(s, B_{j,k}) \geq 2^{j-2}R \) for \( 1 \leq k \leq K_j \). Then

\[
\int_{[\mathcal{B}(R)]^c} f_t(X) \, d\nu(t) = \sum_{j \geq 1} \sum_{k=1}^{K_j} \int_{F_{j,k}} f_t(X) \, d\nu(t) = \sum_{j \geq 1} \sum_{k=1}^{K_j} \nu(F_{j,k}) g_{F_{j,k}}(X).
\]

Since \( J + j \geq J_0 + 1 \geq 2 \),
\[
h(\mathcal{B}(r), B_{j,k}) \geq 2^{j-2}R - r = (2^{J+j} - 1) r = 2^{J+j} r \left( 1 - 2^{-(J+j)} \right) \geq (3/4) 2^{J+j} r,
\]
and, since \( F_{j,k} \subset B_{j,k} \), it follows from (6.1) that
\[
\mathbb{P}_{\mathcal{B}(r)} \left[ \log \left( \frac{g_{F_{j,k}}(X)}{g_{\mathcal{B}(r)}(X)} \right) \geq y \right] \leq \exp \left[ -\frac{1}{2} (y + 4^{J+j}nr^2) \right] \quad \text{for all } y \in \mathbb{R}.
\]

Therefore
\[
\mathbb{P}_{\mathcal{B}(r)} \left[ \log \left( \frac{g_{F_{j,k}}(X)}{g_{\mathcal{B}(r)}(X)} \right) \geq -4^{J+j}nr^2/2 \right] \leq \exp \left[ -4^{J+j}nr^2 \right].
\]

This means that
\[
\frac{g_{F_{j,k}}(X)}{g_{\mathcal{B}(r)}(X)} < \exp \left[ -4^{J+j}nr^2/2 \right] \quad \text{for all } j \geq 1 \text{ and } 1 \leq k \leq K_j,
\]
except if \( X \in \Gamma \) with
\[
P_{\mathcal{B}(r)}[\Gamma] \leq \sum_{j \geq 1} K_j \exp \left[ -4^{J+j}nr^2 \right] \leq \sum_{j \geq 1} \exp \left[ -4^{J+j}nr^2 + D(2^{J+j}r) \right],
\]
by (6.2). Since \( D \) is non-increasing, \( J \geq J_0 \geq 1 \) and \( j \geq 1 \), it follows from (4.1) that \( D(2^{J+j}r) \leq 4^{J+j}nr^2/8 \) and \( 4^Jnr^2 \geq 2 \), hence
\[
P_{\mathcal{B}(r)}[\Gamma] \leq \sum_{j \geq 1} \exp \left[ -4^{J+j}nr^2/8 \right] \leq \exp \left[ -4^Jnr^2/2 \right] \sum_{j \geq 0} \exp \left[ -4^Jnr^2 \left( 4^k - 1 \right)/2 \right]
\]
\[
< 1.05 \exp \left[ -4^Jnr^2/2 \right].
\]

Putting everything together, we derive that, when \( X \in \Gamma^c \), for \( l \geq 0 \),
\[
\mathbb{P} \left[ (\mathcal{B}(2^lR))^c \right] \leq \frac{1}{\nu(\mathcal{B}(r))} \sum_{j \geq l+1} \sum_{k=1}^{K_j} \nu(F_{j,k}) \frac{g_{F_{j,k}}(X)}{g_{\mathcal{B}(r)}(X)}
\]
\[
\leq \frac{1}{\nu(\mathcal{B}(r))} \sum_{j \geq l+1} \sum_{k=1}^{K_j} \nu(F_{j,k}) \exp \left[ -4^{J+j}nr^2/2 \right]
\]
\[
\leq \sum_{j \geq l+1} \frac{\nu(F_{j,k})}{\nu(\mathcal{B}(r))} \exp \left[ -4^{J+j}nr^2/2 \right]
\]
\[
\leq \sum_{j \geq l+1} \frac{\nu(\mathcal{B}(2^{J+j+2}r))}{\nu(\mathcal{B}(r))} \exp \left[ -4^{J+j}nr^2/2 \right]
\]
\[
\leq \sum_{j \geq l+1} \exp \left[ 2^{(J+j+2)\gamma} \beta(r) - (4^{J+j}nr^2/2) \right]
\]
\[
12
\]
by Assumption 2. Since $2^{(J_0 + 3)\gamma \beta(r)} \leq 4J_0nr^2$ and $\gamma \leq 2$, $2^{(J + j + 2)\gamma \beta(r)} \leq 4^{J+j}nr^2/4$

and

$$\mathcal{N} \left[ \left( \mathcal{B}(2^lR) \right)^c \right] \leq \sum_{j \geq l+1} \exp \left[ -4^{J+l-1}nr^2 \right] = \exp \left[ -4^{J+l}nr^2 \right] \sum_{k \geq 0} \exp \left[ - \left( 4^k - 1 \right) 4^{l+k}nr^2 \right] < 1.01 \exp \left[ -4^{J+l}nr^2 \right]$$

which concludes the proof ot Theorem 1.

**Acknowledgements** I would like to thank Judith Rousseau for her invitation to participate to the Bayesian Nonparametrics Conference 2013, which strongly pushed me to write down some thoughts about Bayesian procedures.

**References**

BARRON, A.R. (1991). Complexity regularization with applications to artificial neural networks. In *Nonparametric Functional Estimation* (G. Roussas, ed.). Kluwer, Dordrecht, 561-576.

GHOSAL, S., GHOSH, J.K. and van der VAART, A.W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28**, 500-531.

IBRAGIMOV, I.A. and HAS’MINSKII, R.Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York.

KOLMOGOROV, A.N. and TIKHOMIROV, V.M. (1961). $\varepsilon$-entropy and $\varepsilon$-capacity of sets in function spaces. *Amer. Math. Soc. Transl. (2)* **17**, 277-364.

Le CAM, L.M. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1**, 38-53.

Le CAM, L.M. (1982). On the risk of Bayes estimates. *Statistical Decision Theory and Related Topics III, Vol. 2* (S.S. Gupta and J.O. Berger, eds.), 121-137. Academic Press, New York.

Le CAM, L.M. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.

LORENTZ, G.G. (1966). *Approximation of Functions*. Holt, Rinehart, Winston, New York.

Lucien BIRGÉ
UMR 7599 “Probabilités et modèles aléatoires”
Laboratoire de Probabilités, boîte 188
Université Paris 06, 4 Place Jussieu
F-75252 Paris Cedex 05
France
e-mail: lucien.birge@upmc.fr