Journal of Machine Learning Research 7 (2006) 971–981 Submitted 11/05; Revised 5/06; Published 6/06

Lower Bounds and Aggregation in Density Estimation

Guillaume Lecué LECUE@CCR.JUSSIEU.FR
Laboratoire de Probabilités et Modèles Aléatoires
Université Paris 6
4 place Jussieu, BP 188
75252 Paris, France

Editor: Gábor Lugosi

Abstract

In this paper we prove the optimality of an aggregation procedure. We prove lower bounds for aggregation of model selection type of M density estimators for the Kullback-Leibler divergence (KL), the Hellinger's distance and the L_1-distance. The lower bound, with respect to the KL distance, can be achieved by the on-line type estimate suggested, among others, by Yang (2000a). Combining these results, we state that $\log M/n$ is an optimal rate of aggregation in the sense of Tsybakov (2003), where n is the sample size.

Keywords: aggregation, optimal rates, Kullback-Leibler divergence

1. Introduction

Let $(\mathcal{X},\mathcal{A})$ be a measurable space and ν be a σ-finite measure on $(\mathcal{X},\mathcal{A})$. Let $D_n=(X_1,\dots,X_n)$ be a sample of n i.i.d. observations drawn from an unknown probability of density f on $\mathcal{X}$ with respect to ν. Consider the estimation of f from D_n.

Suppose that we have $M\geq 2$ different estimators $\hat{f}_1,\dots,\hat{f}_M$ of f. Catoni (1997), Yang (2000a), Yang (2000b), Nemirovski (2000), Juditsky and Nemirovski (2000), Yang (2001), Tsybakov (2003), Catoni (2004) and Rigollet and Tsybakov (2004) have studied the problem of model selection type aggregation. It consists in construction of a new estimator $\tilde{f}_n$ (called *aggregate*) which is approximatively at least as good as the best among $\hat{f}_1,\dots,\hat{f}_M$. In most of these papers, this problem is solved by using a kind of cross-validation procedure. Namely, the aggregation is based on splitting the sample in two independent subsamples D_m^1 and D_l^2 of sizes m and l respectively, where $m\gg l$ and $m+l=n$. The size of the first subsample has to be greater than the one of the second because it is used for the true estimation, that is for the construction of the M estimators $\hat{f}_1,\dots,\hat{f}_M$. The second subsample is used for the adaptation step of the procedure, that is for the construction of an aggregate $\tilde{f}_n$, which has to mimic, in a certain sense, the behavior of the best among the estimators $\hat{f}_i$. Thus, $\tilde{f}_n$ is measurable w.r.t. the whole sample D_n unlike the first estimators $\hat{f}_1,\dots,\hat{f}_M$. Actually, Nemirovski (2000) and Juditsky and Nemirovski (2000) did not focus on model selection type aggregation. These papers give a bigger picture about the general topic of procedure aggregation and Yang (2004) complemented their results. Tsybakov (2003) improved these results and formulated the three types of aggregation problems (cf. Tsybakov (2003)).

One can suggest different aggregation procedures and the question is how to look for an optimal one. A way to define optimality in aggregation in a minimax sense for a regression problem is suggested in Tsybakov (2003). Based on the same principle we can define optimality for density

aggregation. In this paper we will not consider the sample splitting and concentrate only on the adaptation step, i.e. on the construction of aggregates (following Nemirovski (2000), Juditsky and Nemirovski (2000), Tsybakov (2003)). Thus, the first subsample is fixed and instead of estimators $\hat{f}_1,\dots,\hat{f}_M$, we have fixed functions $f_1,\dots,f_M$. Rather than working with a part of the initial sample we will use, for notational simplicity, the whole sample D_n of size n instead of a subsample D_l^2.

The aim of this paper is to prove the optimality, in the sense of Tsybakov (2003), of the aggregation method proposed by Yang, for the estimation of a density on $(\mathbb{R}^d,\lambda)$ where λ is the Lebesgue measure on $\mathbb{R}^d$. This procedure is a convex aggregation with weights which can be seen in two different ways. Yang's point of view is to express these weights in function of the likelihood of the model, namely

$$\tilde{f}_n(x)=\sum_{j=1}^{M}\tilde{w}_j^{(n)}f_j(x),\quad \forall x\in\mathcal{X}, \tag{1}$$

where the weights are $\tilde{w}_j^{(n)}=(n+1)^{-1}\sum_{k=0}^{n}w_j^{(k)}$ and

$$w_j^{(k)}=\frac{f_j(X_1)\dots f_j(X_k)}{\sum_{l=1}^{M}f_l(X_1)\dots f_l(X_k)},\ \forall k=1,\dots,n \text{ and } w_j^{(0)}=\frac{1}{M}. \tag{2}$$

And the second point of view is to write these weights as exponential ones, as used in Augustin et al. (1997), Catoni (2004), Hartigan (2002), Bunea and Nobel (2005), Juditsky et al. (2005) and Lecué (2005), for different statistical models. Define the empirical Kullback loss $K_n(f)=-(1/n)\sum_{i=1}^{n}\log f(X_i)$ (keeping only the term independent of the underlying density to estimate) for all density f. We can rewrite these weights as exponential weights:

$$w_j^{(k)}=\frac{\exp(-kK_k(f_j))}{\sum_{l=1}^{M}\exp(-kK_k(f_l))},\quad \forall k=0,\dots,n.$$

Most of the results on convergence properties of aggregation methods are obtained for the regression and the gaussian white noise models. Nevertheless, Catoni (1997, 2004), Devroye and Lugosi (2001), Yang (2000a), Zhang (2003) and Rigollet and Tsybakov (2004) have explored the performances of aggregation procedures in the density estimation framework. Most of them have established upper bounds for some procedure and do not deal with the problem of optimality of their procedures. Nemirovski (2000), Juditsky and Nemirovski (2000) and Yang (2004) state lower bounds for aggregation procedure in the regression setup. To our knowledge, lower bounds for the performance of aggregation methods in density estimation are available only in Rigollet and Tsybakov (2004). Their results are obtained with respect to the mean squared risk. Catoni (1997) and Yang (2000a) construct procedures and give convergence rates w.r.t. the KL loss. One aim of this paper is to prove optimality of one of these procedures w.r.t. the KL loss. Lower bounds w.r.t. the Hellinger's distance and L_1-distance (stated in Section 3) and some results of Birgé (2004) and Devroye and Lugosi (2001) (recalled in Section 4) suggest that the rates of convergence obtained in Theorem 2 and 4 are optimal in the sense given in Definition 1. In fact, an approximate bound can be achieved, if we allow the leading term in the RHS of the oracle inequality (i.e. in the upper bound) to be multiplied by a constant greater than one.

The paper is organized as follows. In Section 2 we give a Definition of optimality, for a rate of aggregation and for an aggregation procedure, and our main results. Lower bounds, for different loss functions, are given in Section 3. In Section 4, we recall a result of Yang (2000a) about an exact oracle inequality satisfied by the aggregation procedure introduced in (1).

2. Main Definition and Main Results

To evaluate the accuracy of a density estimator we use the Kullback-Leibler (KL) divergence, the Hellinger's distance and the L_1-distance as loss functions. The *KL divergence* is defined for all densities f, g w.r.t. a $\sigma-$finite measure ν on a space $\mathcal{X}$, by

$$K(f|g) = \begin{cases} \int_{\mathcal{X}} \log\left(\frac{f}{g}\right) f d\nu & \text{if } P_f \ll P_g; \\ +\infty & \text{otherwise,} \end{cases}$$

where P_f (respectively P_g) denotes the probability distribution of density f (respectively g) w.r.t. ν. *Hellinger's distance* is defined for all non-negative measurable functions f and g by

$$H(f,g) = \left\| \sqrt{f} - \sqrt{g} \right\|_2,$$

where the L_2-norm is defined by $\|f\|_2 = \left(\int_{\mathcal{X}} f^2(x) d\nu(x)\right)^{1/2}$ for all functions $f \in L_2(\mathcal{X}, \nu)$. The L_1*-distance* is defined for all measurable functions f and g by

$$v(f,g) = \int_{\mathcal{X}} |f-g| d\nu.$$

The main goal of this paper is to find optimal rate of aggregation in the sense of the definition given below. This definition is an analog, for the density estimation problem, of the one in Tsybakov (2003) for the regression problem.

Definition 1 *Take $M \geq 2$ an integer, $\mathcal{F}$ a set of densities on $(\mathcal{X}, \mathcal{A}, \nu)$ and $\mathcal{F}_0$ a set of functions on $\mathcal{X}$ with values in $\mathbb{R}$ such that $\mathcal{F} \subseteq \mathcal{F}_0$. Let d be a loss function on the set $\mathcal{F}_0$. A sequence of positive numbers $(\psi_n(M))_{n\in\mathbb{N}^*}$ is called* ***optimal rate of aggregation of M functions in*** $(\mathcal{F}_0, \mathcal{F})$ ***w.r.t. the loss*** *d if :*

(i) There exists a constant $C < \infty$, depending only on $\mathcal{F}_0, \mathcal{F}$ and d, such that for all functions $f_1, \ldots, f_M$ in $\mathcal{F}_0$ there exists an estimator $\tilde{f}_n$ (aggregate) of f such that

$$\sup_{f\in\mathcal{F}} \left[\mathbb{E}_f \left[d(f, \tilde{f}_n) \right] - \min_{i=1,\ldots,M} d(f, f_i) \right] \leq C\psi_n(M), \quad \forall n \in \mathbb{N}^*. \tag{3}$$

(ii) There exist some functions $f_1, \ldots, f_M$ in $\mathcal{F}_0$ and $c > 0$ a constant independent of M such that for all estimators $\hat{f}_n$ of f,

$$\sup_{f\in\mathcal{F}} \left[\mathbb{E}_f \left[d(f, \hat{f}_n) \right] - \min_{i=1,\ldots,M} d(f, f_i) \right] \geq c\psi_n(M), \quad \forall n \in \mathbb{N}^*. \tag{4}$$

Moreover, when the inequalities (3) and (4) are satisfied, we say that the procedure $\tilde{f}_n$, appearing in (3), is an **optimal aggregation procedure w.r.t. the loss** *d***.**

Let $A > 1$ be a given number. In this paper we are interested in the estimation of densities lying in

$$\mathcal{F}(A) = \{\text{densities bounded by } A\} \tag{5}$$

and, depending on the used loss function, we aggregate functions in $\mathcal{F}_0$ which can be:

1. $\mathcal{F}_K(A) = \{\text{densities bounded by } A\}$ for KL divergence,
2. $\mathcal{F}_H(A) = \{\text{non-negative measurable functions bounded by } A\}$ for Hellinger's distance,
3. $\mathcal{F}_v(A) = \{\text{measurable functions bounded by } A\}$ for the L_1-distance.

The main result of this paper, obtained by using Theorem 5 and assertion (6) of Theorem 3, is the following Theorem.

Theorem 1 *Let $A > 1$. Let M and n be two integers such that $\log M \leq 16(\min(1, A-1))^2 n$. The sequence*

$$\psi_n(M) = \frac{\log M}{n}$$

is an optimal rate of aggregation of M functions in $(\mathcal{F}_K(A), \mathcal{F}(A))$ (introduced in (5)) w.r.t. the KL divergence loss. Moreover, the aggregation procedure with exponential weights, defined in (1), achieves this rate. So, this procedure is an optimal aggregation procedure w.r.t. the KL-loss.

Moreover, if we allow the leading term "$\min_{i=1,\dots,M} d(f, f_i)$", in the upper bound and the lower bound of Definition 1, to be multiplied by a constant greater than one, then the rate $(\psi_n(M))_{n\in\mathbb{N}^*}$ is said "near optimal rate of aggregation". Observing Theorem 6 and the result of Devroye and Lugosi (2001) (recalled at the end of Section 4), the rates obtained in Theorems 2 and 4:

$$\left(\frac{\log M}{n}\right)^{\frac{q}{2}}$$

are near optimal rates of aggregation for the Hellinger's distance and the L_1-distance to the power q, where $q > 0$.

3. Lower Bounds

To prove lower bounds of type (4) we use the following lemma on minimax lower bounds which can be obtained by combining Theorems 2.2 and 2.5 in Tsybakov (2004). We say that d is a semi-distance on Θ if d is symmetric, satisfies the triangle inequality and $d(\theta, \theta) = 0$.

Lemma 1 *Let d be a semi-distance on the set of all densities on $(\mathcal{X}, \mathcal{A}, \nu)$ and w be a non-decreasing function defined on $\mathbb{R}_+$ which is not identically 0. Let $(\psi_n)_{n\in\mathbb{N}}$ be a sequence of positive numbers. Let $\mathcal{C}$ be a finite set of densities on $(\mathcal{X}, \mathcal{A}, \nu)$ such that $card(\mathcal{C}) = M \geq 2$,*

$$\forall f, g \in \mathcal{C}, f \neq g \Longrightarrow d(f, g) \geq 4\psi_n > 0,$$

and the KL divergences $K(P_f^{\otimes n} | P_g^{\otimes n})$, between the product probability measures corresponding to densities f and g respectively, satisfy, for some $f_0 \in \mathcal{C}$,

$$\forall f \in \mathcal{C}, K(P_f^{\otimes n} | P_{f_0}^{\otimes n}) \leq (1/16)\log(M).$$

Then,

$$\inf_{\hat{f}_n} \sup_{f\in\mathcal{C}} \mathbb{E}_f \left[w(\psi_n^{-1} d(\hat{f}_n, f)) \right] \geq c_1,$$

where $\inf_{\hat{f}_n}$ denotes the infimum over all estimators based on a sample of size n from an unknown distribution with density f and $c_1 > 0$ is an absolute constant.

Now, we give a lower bound of the form (4) for the three different loss functions introduced in the beginning of the section. Lower bounds are given in the problem of estimation of a density on $\mathbb{R}^d$, namely we have $\mathcal{X} = \mathbb{R}^d$ and ν is the Lebesgue measure on $\mathbb{R}^d$.

Theorem 2 *Let M be an integer greater than 2, $A > 1$ and $q > 0$ be two numbers. We have for all integers n such that $\log M \leq 16(\min(1, A-1))^2 n$,*

$$\sup_{f_1,\dots,f_M \in \mathcal{F}_H(A)} \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}(A)} \left[\mathbb{E}_f \left[H(\hat{f}_n, f)^q \right] - \min_{j=1,\dots,M} H(f_j, f)^q \right] \geq c \left(\frac{\log M}{n} \right)^{q/2},$$

where c is a positive constant which depends only on A and q. The sets $\mathcal{F}(A)$ and $\mathcal{F}_H(A)$ are defined in (5) when $\mathcal{X} = \mathbb{R}^d$ and the infimum is taken over all the estimators based on a sample of size n.

Proof : For all densities $f_1, \dots, f_M$ bounded by A we have,

$$\sup_{f_1,\dots,f_M \in \mathcal{F}_H(A)} \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}(A)} \left[\mathbb{E}_f \left[H(\hat{f}_n, f)^q \right] - \min_{j=1,\dots,M} H(f_j, f)^q \right] \geq \inf_{\hat{f}_n} \sup_{f \in \{f_1,\dots,f_M\}} \mathbb{E}_f \left[H(\hat{f}_n, f)^q \right].$$

Thus, to prove Theorem 1, it suffices to find M appropriate densities bounded by A and to apply Lemma 1 with a suitable rate.

We consider D the smallest integer such that $2^{D/8} \geq M$ and $\Delta = \{0,1\}^D$. We set $h_j(y) = h(y - (j-1)/D)$ for all $y \in \mathbb{R}$, where $h(y) = (L/D)g(Dy)$ and $g(y) = \mathbb{1}_{[0,1/2]}(y) - \mathbb{1}_{(1/2,1]}(y)$ for all $y \in \mathbb{R}$ and $L > 0$ will be chosen later. We consider

$$f_\delta(x) = \mathbb{1}_{[0,1]^d}(x) \left(1 + \sum_{j=1}^{D} \delta_j h_j(x_1) \right), \quad \forall x = (x_1, \dots, x_d) \in \mathbb{R}^d,$$

for all $\delta = (\delta_1, \dots, \delta_D) \in \Delta$. We take L such that $L \leq D\min(1, A-1)$ thus, for all $\delta \in \Delta$, f_δ is a density bounded by A. We choose our densities $f_1, \dots, f_M$ in $\mathcal{B} = \{f_\delta : \delta \in \Delta\}$, but we do not take all of the densities of $\mathcal{B}$ (because they are too close to each other), but only a subset of $\mathcal{B}$, indexed by a separated set (this is a set where all the points are separated from each other by a given distance) of Δ for the *Hamming distance* defined by $\rho(\delta^1, \delta^2) = \sum_{i=1}^{D} I(\delta_i^1 \neq \delta_i^2)$ for all $\delta^1 = (\delta_1^1, \dots \delta_D^1), \delta^2 = (\delta_1^2, \dots, \delta_D^2) \in \Delta$. Since $\int_{\mathbb{R}} h d\lambda = 0$, we have

$$\begin{aligned} H^2(f_{\delta^1}, f_{\delta^2}) &= \sum_{j=1}^{D} \int_{\frac{j-1}{D}}^{\frac{j}{D}} I(\delta_j^1 \neq \delta_j^2) \left(1 - \sqrt{1 + h_j(x)} \right)^2 dx \\ &= 2\rho(\delta^1, \delta^2) \int_0^{1/D} \left(1 - \sqrt{1 + h(x)} \right) dx, \end{aligned}$$

for all $\delta^1 = (\delta_1^1, \dots, \delta_D^1), \delta^2 = (\delta_1^2, \dots, \delta_D^2) \in \Delta$. On the other hand the function $\varphi(x) = 1 - \alpha x^2 - \sqrt{1+x}$, where $\alpha = 8^{-3/2}$, is convex on $[-1,1]$ and we have $|h(x)| \leq L/D \leq 1$ so, according to Jensen, $\int_0^1 \varphi(h(x))dx \geq \varphi\left(\int_0^1 h(x)dx \right)$. Therefore $\int_0^{1/D} \left(1 - \sqrt{1+h(x)} \right) dx \geq \alpha \int_0^{1/D} h^2(x)dx = (\alpha L^2)/D^3$, and we have

$$H^2(f_{\delta^1}, f_{\delta^2}) \geq \frac{2\alpha L^2}{D^3} \rho(\delta^1, \delta^2),$$

for all $\delta^1, \delta^2 \in \Delta$. According to Varshamov-Gilbert, cf. Tsybakov (2004, p. 89) or Ibragimov and Hasminskii (1980), there exists a $D/8$-separated set, called $N_{D/8}$, on Δ for the Hamming distance such that its cardinal is higher than $2^{D/8}$ and $(0,\dots,0) \in N_{D/8}$. On the separated set $N_{D/8}$ we have,

$$\forall \delta^1, \delta^2 \in N_{D/8}\,, H^2(f_{\delta^1}, f_{\delta^2}) \geq \frac{\alpha L^2}{4D^2}.$$

In order to apply Lemma 1, we need to control the KL divergences too. Since we have taken $N_{D/8}$ such that $(0,\dots,0) \in N_{D/8}$, we can control the KL divergences w.r.t. P_0, the Lebesgue measure on $[0,1]^d$. We denote by P_δ the probability of density f_δ w.r.t. the Lebesgue's measure on $\mathbb{R}^d$, for all $\delta \in \Delta$. We have,

$$\begin{aligned} K(P_\delta^{\otimes n}|P_0^{\otimes n}) &= n\int_{[0,1]^d} \log\left(f_\delta(x)\right) f_\delta(x)dx \\ &= n\sum_{j=1}^{D} \int_{\frac{j-1}{D}}^{j/D} \log\left(1+\delta_j h_j(x)\right)\left(1+\delta_j h_j(x)\right)dx \\ &= n\left(\sum_{j=1}^{D}\delta_j\right)\int_0^{1/D} \log(1+h(x))(1+h(x))dx, \end{aligned}$$

for all $\delta = (\delta_1,\dots,\delta_D) \in N_{D/8}$. Since $\forall u > -1, \log(1+u) \leq u$, we have,

$$K(P_\delta^{\otimes n}|P_0^{\otimes n}) \leq n\left(\sum_{j=1}^{D}\delta_j\right)\int_0^{1/D}(1+h(x))h(x)dx \leq nD\int_0^{1/D} h^2(x)dx = \frac{nL^2}{D^2}.$$

Since $\log M \leq 16(\min(1,A-1))^2 n$, we can take L such that $(nL^2)/D^2 = \log(M)/16$ and still having $L \leq D\min(1,A-1)$. Thus, for $L = (D/4)\sqrt{\log(M)/n}$, we have for all elements δ^1, δ^2 in $N_{D/8}$, $H^2(f_{\delta^1}, f_{\delta^2}) \geq (\alpha/64)(\log(M)/n)$ and $\forall \delta \in N_{D/8}\,, K(P_\delta^{\otimes n}|P_0^{\otimes n}) \leq (1/16)\log(M)$.

Applying Lemma 1 when d is H, the Hellinger's distance, with M densities $f_1,\dots,f_M$ in $\left\{f_\delta : \delta \in N_{D/8}\right\}$ where $f_1 = \mathbb{1}_{[0,1]^d}$ and the increasing function $w(u) = u^q$, we get the result.

■

Remark 1 *The construction of the family of densities $\left\{f_\delta : \delta \in N_{D/8}\right\}$ is in the same spirit as the lower bound of Tsybakov (2003), Rigollet and Tsybakov (2004). But, as compared to Rigollet and Tsybakov (2004), we consider a different problem (model selection aggregation) and as compared to Tsybakov (2003), we study in a different context (density estimation). Also, our risk function is different from those considered in these papers.*

Now, we give a lower bound for KL divergence. We have the same result as for square of Hellinger's distance.

Theorem 3 *Let $M \geq 2$ be an integer, $A > 1$ and $q > 0$. We have, for any integer n such that $\log M \leq 16(\min(1,A-1))^2 n$,*

$$\sup_{f_1,\dots,f_M \in \mathcal{F}_K(A)} \inf_{\hat{f}_n} \sup_{f\in\mathcal{F}(A)} \left[\mathbb{E}_f\left[(K(f|\hat{f}_n))^q\right] - \min_{j=1,\dots,M}(K(f|f_j))^q\right] \geq c\left(\frac{\log M}{n}\right)^q, \qquad (6)$$

and

$$\sup_{f_1,\dots,f_M\in\mathcal{F}_K(A)}\inf_{\hat{f}_n}\sup_{f\in\mathcal{F}(A)}\left[\mathbb{E}_f\left[(K(\hat{f}_n|f))^q\right]-\min_{j=1,\dots,M}(K(f_j|f))^q\right]\geq c\left(\frac{\log M}{n}\right)^q, \tag{7}$$

where c is a positive constant which depends only on A. The sets $\mathcal{F}(A)$ and $\mathcal{F}_K(A)$ are defined in (5) for $\mathcal{X}=\mathbb{R}^d$.

Proof : Proof of the inequality (7) of Theorem 3 is similar to the one for (6). Since we have for all densities f and g,

$$K(f|g)\geq H^2(f,g),$$

(a proof is given in Tsybakov, 2004, p. 73), it suffices to note that, if $f_1,\dots,f_M$ are densities bounded by A then,

$$\sup_{f_1,\dots,f_M\in\mathcal{F}_K(A)}\inf_{\hat{f}_n}\sup_{f\in\mathcal{F}(A)}\left[\mathbb{E}_f\left[(K(f|\hat{f}_n))^q\right]-\min_{j=1,\dots,M}(K(f|f_i))^q\right]$$
$$\geq\inf_{\hat{f}_n}\sup_{f\in\{f_1,\dots,f_M\}}\left[\mathbb{E}_f\left[(K(f|\hat{f}_n))^q\right]\right]\geq\inf_{\hat{f}_n}\sup_{f\in\{f_1,\dots,f_M\}}\left[\mathbb{E}_f\left[H^{2q}(f,\hat{f}_n)\right]\right],$$

to get the result by applying Theorem 2.

■

With the same method as Theorem 1, we get the result below for the L_1-distance.

Theorem 4 *Let $M\geq 2$ be an integer, $A>1$ and $q>0$. We have for any integers n such that $\log M\leq 16(\min(1,A-1))^2 n$,*

$$\sup_{f_1,\dots,f_M\in\mathcal{F}_v(A)}\inf_{\hat{f}_n}\sup_{f\in\mathcal{F}(A)}\left[\mathbb{E}_f\left[v(f,\hat{f}_n)^q\right]-\min_{j=1,\dots,M}v(f,f_i)^q\right]\geq c\left(\frac{\log M}{n}\right)^{q/2}$$

where c is a positive constant which depends only on A. The sets $\mathcal{F}(A)$ and $\mathcal{F}_v(A)$ are defined in (5) for $\mathcal{X}=\mathbb{R}^d$.

Proof : The only difference with Theorem 2 is in the control of the distances. With the same notations as the proof of Theorem 2, we have,

$$v(f_{\delta^1},f_{\delta^2})=\int_{[0,1]^d}|f_{\delta^1}(x)-f_{\delta^2}(x)|dx=\rho(\delta^1,\delta^2)\int_0^{1/D}|h(x)|dx=\frac{L}{D^2}\rho(\delta^1,\delta^2),$$

for all $\delta^1,\delta^2\in\Delta$. Thus, for $L=(D/4)\sqrt{\log(M)/n}$ and $N_{D/8}$, the $D/8$-separated set of Δ introduced in the proof of Theorem 2, we have,

$$v(f_{\delta^1},f_{\delta^2})\geq\frac{1}{32}\sqrt{\frac{\log(M)}{n}},\quad\forall\delta^1,\delta^2\in N_{D/8}\text{ and }K(P_\delta^{\otimes n}|P_0^{\otimes n})\leq\frac{1}{16}\log(M),\quad\forall\delta\in\Delta.$$

Therefore, by applying Lemma 1 to the L_1-distance with M densities $f_1,\dots,f_M$ in $\left\{f_\delta:\delta\in N_{D/8}\right\}$ where $f_1=\mathbb{1}_{[0,1]^d}$ and the increasing function $w(u)=u^q$, we get the result.

■

4. Upper Bounds

In this section we use an argument in Yang (2000a) (see also Catoni, 2004) to show that the rate of the lower bound of Theorem 3 is an optimal rate of aggregation with respect to the KL loss. We use an aggregate constructed by Yang (defined in (1)) to attain this rate. An upper bound of the type (3) is stated in the following Theorem. Remark that Theorem 5 holds in a general framework of a measurable space $(\mathcal{X},\mathcal{A})$ endowed with a σ-finite measure ν.

Theorem 5 (Yang) *Let $X_1,\dots,X_n$ be n observations of a probability measure on $(\mathcal{X},\mathcal{A})$ of density f with respect to ν. Let $f_1,\dots,f_M$ be M densities on $(\mathcal{X},\mathcal{A},\nu)$. The aggregate $\tilde{f}_n$, introduced in (1), satisfies, for any underlying density f,*

$$\mathbb{E}_f\left[K(f|\tilde{f}_n)\right] \leq \min_{j=1,\dots,M} K(f|f_j) + \frac{\log(M)}{n+1}. \tag{8}$$

Proof : Proof follows the line of Yang (2000a), although he does not state the result in the form (3), for convenience we reproduce the argument here. We define $\hat{f}_k(x;X^{(k)}) = \sum_{j=1}^M w_j^{(k)} f_j(x)$, $\forall k = 1,\dots,n$ (where $w_j^{(k)}$ is defined in (2) and $x^{(k)} = (x_1,\dots,x_k)$ for all $k\in\mathbb{N}$ and $x_1,\dots,x_k\in\mathcal{X}$) and $\hat{f}_0(x;X^{(0)}) = (1/M)\sum_{j=1}^M f_j(x)$ for all $x\in\mathcal{X}$. Thus, we have

$$\tilde{f}_n(x;X^{(n)}) = \frac{1}{n+1}\sum_{k=0}^n \hat{f}_k(x;X^{(k)}).$$

Let f be a density on $(\mathcal{X},\mathcal{A},\nu)$. We have

$$\begin{aligned}
\sum_{k=0}^n \mathbb{E}_f\left[K(f|\hat{f}_k)\right] &= \sum_{k=0}^n \int_{\mathcal{X}^{k+1}} \log\left(\frac{f(x_{k+1})}{\hat{f}_k(x_{k+1};x^{(k)})}\right)\prod_{i=1}^{k+1} f(x_i)d\nu^{\otimes(k+1)}(x_1,\dots,x_{k+1}) \\
&= \int_{\mathcal{X}^{n+1}} \left(\sum_{k=0}^n \log\left(\frac{f(x_{k+1})}{\hat{f}_k(x_{k+1};x^{(k)})}\right)\right)\prod_{i=1}^{n+1} f(x_i)d\nu^{\otimes(n+1)}(x_1,\dots,x_{n+1}) \\
&= \int_{\mathcal{X}^{n+1}} \log\left(\frac{f(x_1)\dots f(x_{n+1})}{\prod_{k=0}^n \hat{f}_k(x_{k+1};x^{(k)})}\right)\prod_{i=1}^{n+1} f(x_i)d\nu^{\otimes(n+1)}(x_1,\dots,x_{n+1}),
\end{aligned}$$

but $\prod_{k=0}^n \hat{f}_k(x_{k+1};x^{(k)}) = (1/M)\sum_{j=1}^M f_j(x_1)\dots f_j(x_{n+1}), \forall x_1,\dots,x_{n+1}\in\mathcal{X}$ thus,

$$\sum_{k=0}^n \mathbb{E}_f\left[K(f|\hat{f}_k)\right] = \int_{\mathcal{X}^{n+1}} \log\left(\frac{f(x_1)\dots f(x_{n+1})}{\frac{1}{M}\sum_{j=1}^M f_j(x_1)\dots f_j(x_{n+1})}\right)\prod_{i=1}^{n+1} f(x_i)d\nu^{\otimes(n+1)}(x_1,\dots,x_{n+1}),$$

moreover $x\longmapsto \log(1/x)$ is a decreasing function so,

$$\begin{aligned}
&\sum_{k=0}^n \mathbb{E}_f\left[K(f|\hat{f}_k)\right] \\
&\leq \min_{j=1,\dots,M}\left\{\int_{\mathcal{X}^{n+1}} \log\left(\frac{f(x_1)\dots f(x_{n+1})}{\frac{1}{M}f_j(x_1)\dots f_j(x_{n+1})}\right)\prod_{i=1}^{n+1} f(x_i)d\nu^{\otimes(n+1)}(x_1,\dots,x_{n+1})\right\} \\
&\leq \log M + \min_{j=1,\dots,M}\left\{\int_{\mathcal{X}^{n+1}} \log\left(\frac{f(x_1)\dots f(x_{n+1})}{f_j(x_1)\dots f_j(x_{n+1})}\right)\prod_{i=1}^{n+1} f(x_i)d\nu^{\otimes(n+1)}(x_1,\dots,x_{n+1})\right\},
\end{aligned}$$

finally we have,

$$\sum_{k=0}^{n} \mathbb{E}_f\left[K(f|\hat{f}_k)\right] \leq \log M + (n+1) \inf_{j=1,\ldots,M} K(f|f_j). \tag{9}$$

On the other hand we have,

$$\mathbb{E}_f\left[K(f|\tilde{f}_n)\right] = \int_{\mathcal{X}^{n+1}} \log\left(\frac{f(x_{n+1})}{\frac{1}{n+1}\sum_{k=0}^{n}\hat{f}_k(x_{n+1};x^{(k)})}\right) \prod_{i=1}^{n+1} f(x_i) d\nu^{\otimes(n+1)}(x_1,\ldots,x_{n+1}),$$

and $x \longmapsto \log(1/x)$ is convex, thus,

$$\mathbb{E}_f\left[K(f|\tilde{f}_n)\right] \leq \frac{1}{n+1}\sum_{k=0}^{n} \mathbb{E}_f\left[K(f|\hat{f}_k)\right]. \tag{10}$$

Theorem 5 follows by combining (9) and (10).

■

Birgé constructs estimators, called *T-estimators* (the "T" is for "test"), which are adaptive in aggregation selection model of M estimators with a residual proportional at $(\log M/n)^{q/2}$ when Hellinger and L_1-distances are used to evaluate the quality of estimation (cf. Birgé (2004)). But it does not give an optimal result as Yang, because there is a constant greater than 1 in front of the main term $\min_{i=1,\ldots,M} d^q(f,f_i)$ where d is the Hellinger distance or the L_1 distance. Nevertheless, observing the proof of Theorem 2 and 4, we can obtain

$$\sup_{f_1,\ldots,f_M\in\mathcal{F}(A)} \inf_{\hat{f}_n} \sup_{f\in\mathcal{F}(A)} \left[\mathbb{E}_f\left[d(f,\hat{f}_n)^q\right] - C(q) \min_{i=1,\ldots,M} d(f,f_i)^q\right] \geq c\left(\frac{\log M}{n}\right)^{q/2},$$

where d is the Hellinger or L_1-distance, $q>0$ and $A>1$. The constant $C(q)$ can be chosen equal to the one appearing in the following Theorem. The same residual appears in this lower bound and in the upper bounds of Theorem 6, so we can say that

$$\left(\frac{\log M}{n}\right)^{q/2}$$

is near optimal rate of aggregation w.r.t. the Hellinger distance or the L_1-distance to the power q, in the sense given at the end of Section 2. We recall Birgé's results in the following Theorem.

Theorem 6 (Birgé) *If we have n observations of a probability measure of density f w.r.t. ν and $f_1,\ldots,f_M$ densities on $(\mathcal{X},\mathcal{A},\nu)$, then there exists an estimator $\tilde{f}_n$ (T-estimator) such that for any underlying density f and $q>0$, we have*

$$\mathbb{E}_f\left[H(f,\tilde{f}_n)^q\right] \leq C(q)\left(\min_{j=1,\ldots,M} H(f,f_j)^q + \left(\frac{\log M}{n}\right)^{q/2}\right),$$

and for the L_1-distance we can construct an estimator $\tilde{f}_n$ which satisfies :

$$\mathbb{E}_f\left[v(f,\tilde{f}_n)^q\right] \leq C(q)\left(\min_{j=1,\ldots,M} v(f,f_j)^q + \left(\frac{\log M}{n}\right)^{q/2}\right),$$

where $C(q)>0$ is a constant depending only on q.

Another result, which can be found in Devroye and Lugosi (2001), states that the minimum distance estimate proposed by Yatracos (1985) (cf. Devroye and Lugosi (2001, p. 59)) achieves the same aggregation rate as in Theorem 6 for the L_1-distance with $q=1$. Namely, for all $f, f_1, \dots, f_M \in \mathcal{F}(A)$,

$$\mathbb{E}_f\left[v(f,\check{f}_n)\right] \leq 3 \min_{j=1,\dots,M} v(f,f_j) + \sqrt{\frac{\log M}{n}},$$

where $\check{f}_n$ is the estimator of Yatracos defined by

$$\check{f}_n = \arg \min_{f\in\{f_1,\dots,f_M\}} \sup_{A\in\mathcal{A}} \left| \int_A f - \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i\in A\}} \right|,$$

and $\mathcal{A} = \left\{ \{x : f_i(x) > f_j(x)\} : 1 \leq i,j \leq M \right\}$.

References

N. H. Augustin, S. T. Buckland, and K. P. Burnham. Model selection: An integral part of inference. *Biometrics*, 53:603–618, 1997.

A. Barron and G. Leung. Information theory and mixing least-square regressions. 2004. manuscript.

L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *To appear in Annales of IHP*, 2004. Available at http://www.proba.jussieu.fr/mathdoc/textes/PMA-862.pdf.

F. Bunea and A. Nobel. Online prediction algorithms for aggregation of arbitrary estimators of a conditional mean. 2005. Submitted to IEEE Transactions in Information Theory.

O. Catoni. A mixture approach to universal model selection. 1997. preprint LMENS-97-30, available at http://www.dma.ens.fr/EDITION/preprints/.

O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. Ecole d'été de Probabilités de Saint-Flour 2001, Lecture Notes in Mathematics. Springer, N.Y., 2004.

L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. 2001. Springer, New-York.

J.A. Hartigan. Bayesian regression using akaike priors. 2002. Yale University, New Haven, Preprint.

I.A. Ibragimov and R.Z. Hasminskii. An estimate of density of a distribution. Studies in mathematical stat. IV. Zap. Nauchn. Semin., LOMI, 98(1980),61–85.

A. Juditsky, A. Nazin, A.B. Tsybakov and N. Vayatis. Online aggregation with mirror-descent algorithm. 2005. Preprint n.987, Laboratoire de Probabilités et Modèle aléatoires, Universités Paris 6 and Paris 7 (available at http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2005).

A. Juditsky and A. Nemirovski. Functionnal aggregation for nonparametric estimation. *Ann. of Statist.*, 28:681–712, 2000.

G. Lecué. Simultaneous adaptation to the marge and to complexity in classification. 2005. Available at http://hal.ccsd.cnrs.fr/ccsd-00009241/en/.

A. Nemirovski. *Topics in Non-parametric Statistics*. Springer, N.Y., 2000.

P. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. 2004. Manuscript.

A.B. Tsybakov. Optimal rates of aggregation. *Computational Learning Theory and Kernel Machines. B.Schölkopf and M.Warmuth, eds. Lecture Notes in Artificial Intelligence*, 2777:303–313, 2003. Springer, Heidelberg.

A.B. Tsybakov. *Introduction à l'estimation non-paramétrique*. Springer, 2004.

Y. Yang. Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87, 2000a.

Y. Yang. Combining Different Procedures for Adaptive Regression. *Journal of Multivariate Analysis*, 74:135–161, 2000b.

Y. Yang. Adaptive regression by mixing. *Journal of American Statistical Association*, 96:574–588, 2001.

Y. Yang. Aggregating regression procedures to impose performance. *Bernoulli*, 10(1):25–47, 2004.

T. Zhang. From epsilon-entropy to KL-complexity: analysis of minimum information complexity density estimation. 2003. Tech. Report RC22980, IBM T.J.Watson Research Center.