HAL open science

Suboptimality of Penalized Empirical Risk Minimization in Classification

Guillaume Lecué

To cite this version:

Guillaume Lecué. Suboptimality of Penalized Empirical Risk Minimization in Classification. 2007. hal-00138840

HAL Id: hal-00138840

https://hal.science/hal-00138840v1

Preprint submitted on 27 Mar 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

hal-00138840, version 1 - 27 Mar 2007

Suboptimality of Penalized Empirical Risk Minimization in Classification.

Guillaume Lecué **

Laboratoire de Probabilités et Modèles Aléatoires (UMR CNRS 7599)
Université Paris VI
4 pl.Jussieu, BP 188, 75252 Paris, France ,
lecue@ccr.jussieu.fr

Abstract Let $\mathcal{F}$ be a set of M classification procedures with values in $[-1,1]$. Given a loss function, we want to construct a procedure which mimics at the best possible rate the best procedure in $\mathcal{F}$. This fastest rate is called optimal rate of aggregation. Considering a continuous scale of loss functions with various types of convexity, we prove that optimal rates of aggregation can be either $((\log M)/n)^{1/2}$ or $(\log M)/n$. We prove that, if all the M classifiers are binary, the (penalized) Empirical Risk Minimization procedures are suboptimal (even under the margin/low noise condition) when the loss function is somewhat more than convex, whereas, in that case, aggregation procedures with exponential weights achieve the optimal rate of aggregation.

1 Introduction

Consider the problem of binary classification. Let $(\mathcal{X},\mathcal{A})$ be a measurable space. Let (X,Y) be a couple of random variables, where X takes its values in $\mathcal{X}$ and Y is a random label taking values in $\{-1,1\}$. We denote by π the probability distribution of (X,Y). For any function $\phi:\mathbb{R}\longmapsto\mathbb{R}$, define the $\phi-$risk of a real valued classifier $f:\mathcal{X}\longmapsto\mathbb{R}$ by

$$A^{\phi}(f)=\mathbb{E}[\phi(Yf(X))].$$

Many different losses have been discussed in the literature along the last decade (cf. [10,13,26,14,6]), for instance:

$\phi_0(x)=\mathbb{1}_{(x\leq 0)}$	classical loss or $0-1$ loss
$\phi_1(x)=\max(0,1-x)$	hinge loss (SVM loss)
$x\longmapsto\log_2(1+\exp(-x))$	logit-boosting loss
$x\longmapsto\exp(-x)$	exponential boosting loss
$x\longmapsto(1-x)^2$	squared loss
$x\longmapsto\max(0,1-x)^2$	2-norm soft margin loss

We will be especially interested in losses having convex properties as it is considered in the following definition (cf. [17]).

** Paper to be considered for the Mark Fulk Award for the "best student paper".

Definition 1. *Let $\phi : \mathbb{R} \longmapsto \mathbb{R}$ be a function and β be a positive number. We say that ϕ is $\beta-$***convex on** $[-1,1]$ *when*

$$[\phi'(x)]^2 \leq \beta\phi''(x), \quad \forall |x| \leq 1.$$

For example, logit-boosting loss is $(e/\log 2)-$convex, exponential boosting loss is $e-$convex, squared and $2-$norm soft margin losses are $2-$convex.

We denote by f_ϕ^* a function from $\mathcal{X}$ to $\mathbb{R}$ which minimizes A^ϕ over all real-valued functions and by $A_*^\phi \stackrel{\text{def}}{=} A^\phi(f_\phi^*)$ the minimal $\phi-$risk. In most of the cases studied f_ϕ^* or its sign is equal to the Bayes classifier

$$f^*(x) = \text{sign}(2\eta(x) - 1),$$

where η is the conditional probability function $x \longmapsto \mathbb{P}(Y = 1|X = x)$ defined on $\mathcal{X}$ (cf. [3,26,34]). The Bayes classifier f^* is a minimizer of the ϕ_0-risk (cf. [11]).

Our framework is the same as the one considered, among others, by [27,33,7] and [29,17]. We have a family $\mathcal{F}$ of M classifiers $f_1, \ldots, f_M$ and a loss function ϕ. Our goal is to mimic the oracle $\min_{f\in\mathcal{F}}(A^\phi(f) - A_*^\phi)$ based on a sample D_n of n i.i.d. observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y). These classifiers may have been constructed from a previous sample or they can belong to a dictionary of simple prediction rules like decision stumps. The problem is to find a strategy which mimics as fast as possible the best classifier in $\mathcal{F}$. Such strategies can then be used to construct efficient adaptive estimators (cf. [27,22,23,9]). We consider the following definition, which is inspired by the one given in [29] for the regression model.

Definition 2. *Let ϕ be a loss function. The remainder term $\gamma(n, M)$ is called* **optimal rate of aggregation for the $\phi-$risk**, *if the following two inequalities hold.*

i) For any finite set $\mathcal{F}$ of M functions from $\mathcal{X}$ to $[-1,1]$, there exists a statistic $\tilde{f}_n$ such that for any underlying probability measure π and any integer $n \geq 1$,

$$\mathbb{E}[A^\phi(\tilde{f}_n) - A_*^\phi] \leq \min_{f\in\mathcal{F}} \left(A^\phi(f) - A_*^\phi\right) + C_1\gamma(n, M). \tag{1}$$

ii) There exists a finite set $\mathcal{F}$ of M functions from $\mathcal{X}$ to $[-1,1]$ such that for any statistic $\bar{f}_n$ there exists a probability distribution π such that for all $n \geq 1$

$$\mathbb{E}\left[A^\phi(\bar{f}_n) - A_*^\phi\right] \geq \min_{f\in\mathcal{F}} \left(A^\phi(f) - A_*^\phi\right) + C_2\gamma(n, M). \tag{2}$$

Here C_1 and C_2 are absolute positive constants which may depend on ϕ. Moreover, when the above two properties i) and ii) are satisfied, we say that the procedure $\tilde{f}_n$, appearing in (1), is an **optimal aggregation procedure for the $\phi-$risk**.

The paper is organized as follows. In the next Section we present three aggregation strategies that will be shown to attain the optimal rates of aggregation. Section 3 presents performance of these procedures. In Section 4 we give some proofs of the optimality of these procedures depending on the loss function. In Section 5 we state a result on suboptimality of the penalized Empirical Risk Minimization procedures and of procedures called selectors. In Section 6 we give some remarks. All the proofs are postponed to the last Section.

2 Aggregation Procedures

We introduce procedures that will be shown to achieve optimal rates of aggregation depending on the loss function $\phi : \mathbb{R} \longmapsto \mathbb{R}$. All these procedures are constructed with the empirical version of the $\phi-$risk and the main idea is that a classifier f_j with a small empirical $\phi-$risk is likely to have a small $\phi-$risk. We denote by

$$A_n^{\phi}(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i f(X_i))$$

the empirical $\phi-$risk of a real-valued classifier f.

The Empirical Risk Minimization (**ERM**) procedure, is defined by

$$\tilde{f}_n^{ERM} \in \operatorname{Arg}\min_{f\in\mathcal{F}} A_n^{\phi}(f). \tag{3}$$

This is an example of what we call a **selector** which is an aggregate with values in the family $\mathcal{F}$. Penalized ERM procedures are also examples of selectors.

The Aggregation with Exponential Weights (**AEW**) procedure is given by

$$\tilde{f}_n^{AEW} = \sum_{f\in\mathcal{F}} w^{(n)}(f) f, \tag{4}$$

where the weights $w^{(n)}(f)$ are defined by

$$w^{(n)}(f) = \frac{\exp\left(-nA_n^{\phi}(f)\right)}{\sum_{g\in\mathcal{F}} \exp\left(-nA_n^{\phi}(g)\right)}, \quad \forall f \in \mathcal{F}. \tag{5}$$

The Cumulative Aggregation with Exponential Weights (**CAEW**) procedure, is defined by

$$\tilde{f}_{n,\beta}^{CAEW} = \frac{1}{n}\sum_{k=1}^{n} \tilde{f}_{k,\beta}^{AEW}, \tag{6}$$

where $\tilde{f}_{k,\beta}^{AEW}$ is constructed as in (4) based on the sample $(X_1, Y_1), \ldots, (X_k, Y_k)$ of size k and with the 'temperature' parameter $\beta > 0$. Namely,

$$\tilde{f}_{k,\beta}^{AEW} = \sum_{f\in\mathcal{F}} w_{\beta}^{(k)}(f) f, \text{ where } w_{\beta}^{(k)}(f) = \frac{\exp\left(-\beta^{-1}kA_k^{\phi}(f)\right)}{\sum_{g\in\mathcal{F}} \exp\left(-\beta^{-1}kA_k^{\phi}(g)\right)}, \quad \forall f \in \mathcal{F}.$$

The idea of the ERM procedure goes to Le Cam and Vapnik. Exponential weights have been discussed, for example, in [2,15,19,33,7,25,35,1] or in [32,8] in the on-line prediction setup.

3 Exact Oracle Inequalities.

We now recall some known upper bounds on the excess risk. The first point of the following Theorem goes to [31], the second point can be found in [18] or [9] and the last point, dealing with the case of a $\beta-$convex loss function, is Corollary 4.4 of [17].

Theorem 1. *Let $\phi : \mathbb{R} \longmapsto \mathbb{R}$ be a bounded loss function. Let $\mathcal{F}$ be a family of M functions $f_1, \ldots, f_M$ with values in $[-1, 1]$, where $M \geq 2$ is an integer.*

i) The Empirical Risk Minimization procedure $\tilde{f}_n = \tilde{f}_n^{ERM}$ satisfies

$$\mathbb{E}[A^\phi(\tilde{f}_n) - A_*^\phi] \leq \min_{f \in \mathcal{F}}(A^\phi(f) - A_*^\phi) + C\sqrt{\frac{\log M}{n}}, \tag{7}$$

where $C > 0$ is a constant depending only on ϕ.

ii) If ϕ is convex, then the CAEW procedure $\tilde{f}_n = \tilde{f}_n^{CAEW}$ with "temperature parameter" $\beta = 1$ and the AEW procedure $\tilde{f}_n = \tilde{f}_n^{AEW}$ satisfy (7).

iii) If ϕ is $\beta-$convex for a positive number β, then the CAEW procedure with "temperature parameter" β, satisfies

$$\mathbb{E}[A^\phi(\tilde{f}_{n,\beta}^{CAEW}) - A_*^\phi] \leq \min_{f \in \mathcal{F}}(A^\phi(f) - A_*^\phi) + \beta\frac{\log M}{n}.$$

4 Optimal Rates of Aggregation.

To understand how behaves the optimal rate of aggregation depending on the loss we introduce a "continuous scale" of loss functions indexed by a non negative number h,

$$\phi_h(x) = \begin{cases} h\phi_1(x) + (1-h)\phi_0(x) & \text{if } 0 \leq h \leq 1 \\ (h-1)x^2 - x + 1 & \text{if } h > 1, \end{cases}$$

defined for any $x \in \mathbb{R}$, where ϕ_0 is the $0-1$ loss and ϕ_1 is the hinge loss.

This set of losses is representative enough since it describes different type of convexity: for any $h > 1$, ϕ_h is $\beta-$convex on $[-1, 1]$ with $\beta \geq \beta_h \stackrel{\text{def}}{=} (2h-1)^2/(2(h-1)) \geq 2$, for $h = 1$ the loss is linear and for $h < 1$, ϕ_h is non-convex. For $h \geq 0$, we consider

$$A_h(f) \stackrel{\text{def}}{=} A^{\phi_h}(f), f_h^* \stackrel{\text{def}}{=} f_{\phi_h}^* \text{ and } A_h^* \stackrel{\text{def}}{=} A_*^{\phi_h} = A^{\phi_h}(f_h^*).$$

Theorem 2. *Let $M \geq 2$ be an integer. Assume that the space $\mathcal{X}$ is infinite.*

If $0 \leq h < 1$, then the optimal rate of aggregation for the ϕ_h-risk is achieved by the ERM procedure and is equal to

$$\sqrt{\frac{\log M}{n}}.$$

For $h = 1$, the optimal rate of aggregation for the ϕ_1-risk is achieved by the ERM, the AEW and the CAEW (with 'temperature' parameter $\beta = 1$) procedures and is equal to

$$\sqrt{\frac{\log M}{n}}.$$

If $h > 1$ then, the optimal rate of aggregation for the ϕ_h-risk is achieved by the CAEW, with 'temperature' parameter β_h and is equal to

$$\frac{\log M}{n}.$$

5 Suboptimality of Penalized ERM Procedures.

In this Section we prove a lower bound under the margin assumption for any selector and we give a more precise lower bound for penalized ERM procedures. First, we recall the definition of the margin assumption introduced in [30].

Margin Assumption(MA): *The probability measure π satisfies the margin assumption MA(κ), where $\kappa \geq 1$ if we have*

$$\mathbb{E}[|f(X) - f^*(X)|] \leq c(A_0(f) - A_0^*)^{1/\kappa}, \tag{8}$$

for any measurable function f with values in $\{-1, 1\}$

We denote by $\mathcal{P}_\kappa$ the set of all probability distribution π satisfying MA(κ).

Theorem 3. *Let $M \geq 2$ be an integer, $\kappa \geq 1$ be a real number, $\mathcal{X}$ be infinite and $\phi : \mathbb{R} \longmapsto \mathbb{R}$ be a loss function such that $a_\phi \stackrel{\text{def}}{=} \phi(-1) - \phi(1) > 0$. There exists a family $\mathcal{F}$ of M classifiers with values in $\{-1, 1\}$ satisfying the following.*

Let $\tilde{f}_n$ be a selector with values in $\mathcal{F}$. Assume that $\sqrt{(\log M)/n} \leq 1/2$. There exists a probability measure $\pi \in \mathcal{P}_\kappa$ and an absolute constant $C_3 > 0$ such that $\tilde{f}_n$ satisfies

$$\mathbb{E}\left[A^\phi(\tilde{f}_n) - A_*^\phi\right] \geq \min_{f \in \mathcal{F}} \left(A^\phi(f) - A_*^\phi\right) + C_3 \Big(\frac{\log M}{n}\Big)^{\frac{\kappa}{2\kappa - 1}}. \tag{9}$$

Consider the penalized ERM procedure $\tilde{f}_n^{pERM}$ associated with $\mathcal{F}$, defined by

$$\tilde{f}_n^{pERM} \in \operatorname{Arg}\min_{f \in \mathcal{F}} (A_n^\phi(f) + \operatorname{pen}(f))$$

where the penalty function $\operatorname{pen}(\cdot)$ *satisfies* $|\operatorname{pen}(f)| \leq C\sqrt{(\log M)/n}, \forall f \in \mathcal{F}$, *with* $0 \leq C < \sqrt{2}/3$. *Assume that* $1188\pi C^2 M^{9C^2} \log M \leq n$. *If $\kappa > 1$ then, there*

exists a probability measure $\pi \in \mathcal{P}_\kappa$ and an absolute constant $C_4 > 0$ such that the penalized ERM procedure $\tilde{f}_n^{pERM}$ satisfies

$$\mathbb{E}\left[A^\phi(\tilde{f}_n^{pERM}) - A_*^\phi\right] \geq \min_{f\in\mathcal{F}}\left(A^\phi(f) - A_*^\phi\right) + C_4\sqrt{\frac{\log M}{n}}.$$

Remark 1 *Inspection of the proof shows that Theorem 3 is valid for any family $\mathcal{F}$ of classifiers $f_1, \ldots, f_M$, with values in $\{-1, 1\}$, such that there exist points $x_1, \ldots, x_{2^M}$ in $\mathcal{X}$ satisfying $\left\{(f_1(x_j), \ldots, f_M(x_j)) : j = 1, \ldots, 2^M\right\} = \{-1, 1\}^M$.*

Remark 2 *If we use a penalty function such that $|\mathrm{pen}(f)| \leq \gamma n^{-1/2}, \forall f \in \mathcal{F}$, where $\gamma > 0$ is an absolute constant (i.e. $0 \leq C \leq \gamma(\log M)^{-1/2}$), then the condition "$1188\pi C^2 M^{9C^2} \log M \leq n$" of Theorem 3 is equivalent to "n greater than a constant".*

Theorem 3 states that the ERM procedure (and even penalized ERM procedures) cannot mimic the best classifier in $\mathcal{F}$ with rates faster than $((\log M)/n)^{1/2}$ if the basis classifiers in $\mathcal{F}$ are different enough, under a very mild condition on the loss ϕ. If there is no margin assumption (which corresponds to the case $\kappa = +\infty$), the result of Theorem 3 can be easily deduced from the lower bound in Chapter 7 of [11]. The main message of Theorem 3 is that such a negative statement remains true even under the margin assumption MA(κ). Selectors aggregate cannot mimic the oracle faster than $((\log M)/n)^{1/2}$ in general. Under MA(κ), they cannot mimic the best classifier in $\mathcal{F}$ with rates faster than $((\log M)/n)^{\kappa/(2\kappa-1)}$ (which is greater than $(\log M)/n$ when $\kappa > 1$). We know, according to Theorem 1, that the CAEW procedure mimics the best classifier in $\mathcal{F}$ at the rate $(\log M)/n$ if the loss is $\beta-$convex. Thus, penalized ERM procedures (and more generally, selectors) are suboptimal aggregation procedures when the loss function is $\beta-$convex even if we add the constraint that π satisfies MA(κ).

We can extend Theorem 3 to a more general framework [24] and we obtain that, if the loss function associated with a risk is somewhat more than convex then it is better to use aggregation procedures with exponential weights instead of selectors (in particular penalized ERM or pure ERM). We do not know whether the lower bound (9) is sharp, i.e., whether there exists a selector attaining the reverse inequality with the same rate.

6 Discussion.

We proved in Theorem 2 that the ERM procedure is optimal only for non-convex losses and for the borderline case of the hinge loss. But, for non-convex losses, the implementation of the ERM procedure requires minimization of a function which is not convex. This is hard to implement and not efficient from a practical point of view. In conclusion, the ERM procedure is theoretically optimal only for non-convex losses but in that case it is practically inefficient and it is practically efficient only for the cases where ERM is theoretically suboptimal.

For any convex loss ϕ, we have $\frac{1}{n}\sum_{k=1}^{n} A^{\phi}(\tilde{f}_{k,\beta}^{AEW}) \leq A^{\phi}(\tilde{f}_{\beta}^{CAEW})$. Next, less observations are used for the construction of $\tilde{f}_{k,\beta}^{AEW}, 1 \leq k \leq n-1$, than for the construction of $\tilde{f}_{n,\beta}^{AEW}$. We can therefore expect the $\phi-$risk of $\tilde{f}_{n,\beta}^{AEW}$ to be smaller than the $\phi-$risk of $\tilde{f}_{k,\beta}^{AEW}$ for all $1 \leq k \leq n-1$ and hence smaller than the $\phi-$risk of $\tilde{f}_{n,\beta}^{CAEW}$. Thus, the AEW procedure is likely to be an optimal aggregation procedure for the convex loss functions.

The hinge loss happens to be really hinge for different reasons. For losses "between" the $0-1$ loss and the hinge loss ($0 \leq h \leq 1$), the ERM is an optimal aggregation procedure and the optimal rate of aggregation is $\sqrt{(\log M)/n}$. For losses "over" the hinge loss ($h > 1$), the ERM procedure is suboptimal and $(\log M)/n$ is the optimal rate of aggregation. Thus, there is a breakdown point in the optimal rate of aggregation just after the hinge loss. This breakdown can be explained by the concept of margin : this argument has not been introduced here by the lack of space, but can be found in [24]. Moreover for the hinge loss we get, by linearity

$$\min_{f\in\mathcal{C}} A_1(f) - A_1^* = \min_{f\in\mathcal{F}} A_1(f) - A_1^*,$$

where $\mathcal{C}$ is the convex hull of $\mathcal{F}$. Thus, for the particular case of the hinge loss, "model selection" aggregation and "convex" aggregation are identical problems (cf. [21] for more details).

7 Proofs.

Proof of Theorem 2: The optimal rates of aggregation of Theorem 2 are achieved by the procedures introduced in Section 2. Depending on the value of h, Theorem 1 provides the exact oracle inequalities required by the point (1) of Definition 2. To show optimality of these rates of aggregation, we need only to prove the corresponding lower bounds. We consider two cases: $0 \leq h \leq 1$ and $h > 1$. Denote by $\mathcal{P}$ the set of all probability distributions on $\mathcal{X} \times \{-1, 1\}$.

Let $0 \leq h \leq 1$. It is easy to check that the Bayes rule f^* is a minimizer of the ϕ_h-risk. Moreover, using the inequality $A_1(f) - A_1^* \geq A_0(f) - A_0^*$, which holds for any real-valued function f (cf. [34]), we have for any prediction rules $f_1, \ldots, f_M$ (with values in $\{-1, 1\}$) and for any finite set $\mathcal{F}$ of M real valued functions,

$$\inf_{\hat{f}_n} \sup_{\pi\in\mathcal{P}} \left(\mathbb{E}\left[A_h(\hat{f}_n) - A_h^*\right] - \min_{f\in\mathcal{F}}(A_h(f) - A_h^*) \right) \tag{10}$$
$$\geq \inf_{\hat{f}_n} \sup_{\substack{\pi\in\mathcal{P} \\ f^*\in\{f_1,\ldots,f_M\}}} \left(\mathbb{E}\left[A_h(\hat{f}_n) - A_h^*\right] \right) \geq \inf_{\hat{f}_n} \sup_{\substack{\pi\in\mathcal{P} \\ f^*\in\{f_1,\ldots,f_M\}}} \left(\mathbb{E}\left[A_0(\hat{f}_n) - A_0^*\right] \right).$$

Let N be an integer such that $2^{N-1} \leq M$, $x_1, \ldots, x_N$ be N distinct points of $\mathcal{X}$ and w be a positive number satisfying $(N-1)w \leq 1$. Denote by P^X the probability measure on $\mathcal{X}$ such that $P^X(\{x_j\}) = w$, for $j = 1, \ldots, N-1$

and $P^X(\{x_N\}) = 1-(N-1)w$. We consider the cube $\Omega = \{-1,1\}^{N-1}$. Let $0 < \mathfrak{h} < 1$. For all $\sigma = (\sigma_1, \ldots, \sigma_{N-1}) \in \Omega$ we consider

$$\eta_\sigma(x) = \begin{cases} (1+\sigma_j \mathfrak{h})/2 & \text{if } x = x_1, \ldots, x_{N-1}, \\ 1 & \text{if } x = x_N. \end{cases}$$

For all $\sigma \in \Omega$ we denote by π_σ the probability measure on $\mathcal{X} \times \{-1,1\}$ defined by its marginal P^X on $\mathcal{X}$ and its conditional probability function η_σ.

We denote by ρ the Hamming distance on Ω. Let $\sigma, \sigma' \in \Omega$ such that $\rho(\sigma,\sigma') = 1$. Denote by H the Hellinger's distance. Since $H^2\left(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}\right) = 2\Big(1 - \Big(1 - H^2(\pi_\sigma, \pi_{\sigma'})/2\Big)^n\Big)$ and $H^2(\pi_\sigma, \pi_{\sigma'}) = 2w(1-\sqrt{1-\mathfrak{h}^2})$, then, the Hellinger's distance between the measures $\pi_\sigma^{\otimes n}$ and $\pi_{\sigma'}^{\otimes n}$ satisfies

$$H^2\left(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}\right) = 2\left(1 - (1 - w(1-\sqrt{1-\mathfrak{h}^2}))^n\right).$$

Take w and $\mathfrak{h}$ such that $w(1-\sqrt{1-\mathfrak{h}^2}) \le n^{-1}$. Then, $H^2\left(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}\right) \le 2(1-e^{-1}) < 2$ for any integer n.

Let $\sigma \in \Omega$ and $\hat{f}_n$ be an estimator with values in $\{-1,1\}$ (only the sign of a statistic is used when we work with the $0-1$ loss). For $\pi = \pi_\sigma$, we have

$$\mathbb{E}_{\pi_\sigma}[A_0(\hat{f}_n) - A_0^*] \ge \mathfrak{h} w \mathbb{E}_{\pi_\sigma}\Big[\sum_{j=1}^{N-1} |\hat{f}_n(x_j) - \sigma_j|\Big].$$

Using Assouad's Lemma (cf. Lemma 1), we obtain

$$\inf_{\hat{f}_n} \sup_{\sigma \in \Omega} \Big(\mathbb{E}_{\pi_\sigma}\Big[A_0(\hat{f}_n) - A_0^*\Big]\Big) \ge \mathfrak{h} w \frac{N-1}{4e^2}. \tag{11}$$

Take now $w = (n\mathfrak{h}^2)^{-1}$, $N = \lceil \log M/\log 2 \rceil$, $\mathfrak{h} = \big(n^{-1}\lceil \log M/\log 2\rceil\big)^{1/2}$. We complete the proof by replacing w, $\mathfrak{h}$ and N in (11) and (10) by their values.

For the case $h > 1$, we consider an integer N such that $2^{N-1} \le M$, $N-1$ different points $x_1, \ldots, x_N$ of $\mathcal{X}$ and a positive number w such that $(N-1)w \le 1$. We denote by P^X the probability measure on $\mathcal{X}$ such that $P^X(\{x_j\}) = w$ for $j = 1, \ldots, N-1$ and $P^X(\{x_N\}) = 1-(N-1)w$. Denote by Ω the cube $\{-1,1\}^{N-1}$. For any $\sigma \in \Omega$ and $h > 1$, we consider the conditional probability function η_σ in two different cases. If $2(h-1) \le 1$ we take

$$\eta_\sigma(x) = \begin{cases} (1+2\sigma_j(h-1))/2 & \text{if } x = x_1, \ldots, x_{N-1} \\ 2(h-1) & \text{if } x = x_N, \end{cases}$$

and if $2(h-1) > 1$ we take

$$\eta_\sigma(x) = \begin{cases} (1+\sigma_j)/2 & \text{if } x = x_1, \ldots, x_{N-1} \\ 1 & \text{if } x = x_N. \end{cases}$$

For all $\sigma \in \Omega$ we denote by π_σ the probability measure on $\mathcal{X} \times \{-1,1\}$ with the marginal P^X on $\mathcal{X}$ and the conditional probability function η_σ of Y knowing X.

Consider

$$\rho(h) = \begin{cases} 1 & \text{if } 2(h-1) \le 1 \\ (4(h-1))^{-1} & \text{if } 2(h-1) > 1 \end{cases} \text{ and } g^*_\sigma(x) = \begin{cases} \sigma_j \text{ if } x = x_1, \dots, x_{N-1} \\ 1 \text{ if } x = x_N. \end{cases}$$

A minimizer of the ϕ_h-risk when the underlying distribution is π_σ is given by

$$f^*_{h,\sigma} \stackrel{\text{def}}{=} \frac{2\eta_\sigma(x) - 1}{2(h-1)} = \rho(h) g^*_\sigma(x), \quad \forall x \in \mathcal{X},$$

for any $h > 1$ and $\sigma \in \Omega$.

When we choose $\{f^*_{h,\sigma} : \sigma \in \Omega\}$ for the set $\mathcal{F} = \{f_1, \dots, f_M\}$ of basis functions, we obtain

$$\sup_{\{f_1,\dots,f_M\}} \inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}} \left(\mathbb{E}\left[A_h(\hat{f}_n) - A^*_h\right] - \min_{j=1,\dots,M} (A_h(f_j) - A^*_h) \right)$$
$$\ge \inf_{\hat{f}_n} \sup_{\substack{\pi \in \mathcal{P}: \\ f^*_h \in \{f^*_{h,\sigma} : \sigma \in \Omega\}}} \left(\mathbb{E}\left[A_h(\hat{f}_n) - A^*_h\right] \right).$$

Let σ be an element of Ω. Under the probability distribution π_σ, we have $A_h(f) - A^*_h = (h-1)\mathbb{E}[(f(X) - f^*_{h,\sigma}(X))^2]$, for any real-valued function f on $\mathcal{X}$. Thus, for a real valued estimator $\hat{f}_n$ based on D_n, we have

$$A_h(\hat{f}_n) - A^*_h \ge (h-1) w \sum_{j=1}^{N-1} (\hat{f}_n(x_j) - \rho(h)\sigma_j)^2.$$

We consider the projection function $\psi_h(x) = \psi(x/\rho(h))$ for any $x \in \mathcal{X}$, where $\psi(y) = \max(-1, \min(1, y)), \forall y \in \mathbb{R}$. We have

$$\begin{aligned} \mathbb{E}_\sigma[A_h(\hat{f}_n) - A^*_h] &\ge w(h-1) \sum_{j=1}^{N-1} \mathbb{E}_\sigma(\psi_h(\hat{f}_n(x_j)) - \rho(h)\sigma_j)^2 \\ &\ge w(h-1)(\rho(h))^2 \sum_{j=1}^{N-1} \mathbb{E}_\sigma(\psi(\hat{f}_n(x_j)) - \sigma_j)^2 \\ &\ge 4w(h-1)(\rho(h))^2 \inf_{\hat{\sigma} \in [0,1]^{N-1}} \max_{\sigma \in \Omega} \mathbb{E}_\sigma \left[\sum_{j=1}^{N-1} |\hat{\sigma}_j - \sigma_j|^2 \right], \end{aligned}$$

where the infimum $\inf_{\hat{\sigma} \in [0,1]^{N-1}}$ is taken over all estimators $\hat{\sigma}$ based on one observation from the statistical experience $\{\pi_\sigma^{\otimes n} | \sigma \in \Omega\}$ and with values in $[0,1]^{N-1}$.

For any $\sigma, \sigma' \in \Omega$ such that $\rho(\sigma, \sigma') = 1$, the Hellinger's distance between the measures $\pi_\sigma^{\otimes n}$ and $\pi_{\sigma'}^{\otimes n}$ satisfies

$$H^2\left(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}\right) = \begin{cases} 2\left(1 - (1 - 2w(1 - \sqrt{1-h^2}))^n\right) & \text{if } 2(h-1) < 1 \\ 2\left(1 - (1 - 2w(1 - \sqrt{3/4}))^n\right) & \text{if } 2(h-1) \ge 1 \end{cases}.$$

We take

$$w = \begin{cases} (2n(h-1)^2) & \text{if } 2(h-1) < 1 \\ 8n^{-1} & \text{if } 2(h-1) \geq 1. \end{cases}$$

Thus, we have for any $\sigma, \sigma' \in \Omega$ such that $\rho(\sigma, \sigma') = 1$,

$$H^2\left(\pi_\sigma^{\otimes n}, \pi_{\sigma'}^{\otimes n}\right) \leq 2(1 - e^{-1}).$$

To complete the proof we apply Lemma 1 with $N = \lceil (\log M)/n \rceil$.

Proof of Theorem 3: Consider $\mathcal{F}$ a family of classifiers $f_1, \ldots, f_M$, with values in $\{-1, 1\}$, such that there exist 2^M points $x_1, \ldots, x_{2^M}$ in $\mathcal{X}$ satisfying $\left\{(f_1(x_j), \ldots, f_M(x_j)) : j = 1, \ldots, 2^M\right\} = \{-1, 1\}^M \stackrel{\text{def}}{=} \mathcal{S}_M$.

Consider the lexicographic order on $\mathcal{S}_M$:

$$(-1, \ldots, -1) \preccurlyeq (-1, \ldots, -1, 1) \preccurlyeq (-1, \ldots, -1, 1, -1) \preccurlyeq \ldots \preccurlyeq (1, \ldots, 1).$$

Take j in $\{1, \ldots, 2^M\}$ and denote by x'_j the element in $\{x_1, \ldots, x_{2^M}\}$ such that $(f_1(x'_j), \ldots, f_M(x'_j))$ is the $j-$th element of $\mathcal{S}_M$ for the lexicographic order. We denote by φ the bijection between $\mathcal{S}_M$ and $\{x_1, \ldots, x_{2^M}\}$ such that the value of φ at the $j-$th element of $\mathcal{S}_M$ is x'_j. By using the bijection φ we can work independently either on the set $\mathcal{S}_M$ or on $\{x_1, \ldots, x_{2^M}\}$. Without any assumption on the space $\mathcal{X}$, we consider, in what follows, functions and probability measures on $\mathcal{S}_M$. Remark that for the bijection φ we have

$$f_j(\varphi(x)) = x^j, \quad \forall x = (x^1, \ldots, x^M) \in \mathcal{S}_M, \forall j \in \{1, \ldots, M\}.$$

With a slight abuse of notation, we still denote by $\mathcal{F}$ the set of functions $f_1, \ldots, f_M$ defined by $f_j(x) = x^j$, for any $j = 1, \ldots, M$.

First remark that for any f, g from $\mathcal{X}$ to $\{-1, 1\}$, using $\mathbb{E}[\phi(Yf(X))|X] = \mathbb{E}[\phi(Y)|X]\mathbb{1}_{(f(X)=1)} + \mathbb{E}[\phi(-Y)|X]\mathbb{1}_{(f(X)=-1)}$, we have

$$\mathbb{E}[\phi(Yf(X))|X] - \mathbb{E}[\phi(Yg(X))|X] = a_\phi(1/2 - \eta(X))(f(X) - g(X)).$$

Hence, we obtain $A^\phi(f) - A^\phi(g) = a_\phi(A_0(f) - A_0(g))$. So, we have for any $j = 1, \ldots, M$,

$$A^\phi(f_j) - A^\phi(f^*) = a_\phi(A_0(f_j) - A_0^*).$$

Moreover, for any $f : \mathcal{S}_M \longmapsto \{-1, 1\}$ we have $A_n^\phi(f) = \phi(1) + a_\phi A_n^{\phi_0}(f)$ and $a_\phi > 0$ by assumption, hence,

$$\tilde{f}_n^{pERM} \in \operatorname{Arg}\min_{f \in \mathcal{F}}(A_n^{\phi_0}(f) + \operatorname{pen}(f)).$$

Thus, it suffices to prove Theorem 3, when the loss function ϕ is the classical $0-1$ loss function ϕ_0.

We denote by $\mathcal{S}_{M+1}$ the set $\{-1, 1\}^{M+1}$ and by $X^0, \ldots, X^M$, $M+1$ independent random variables with values in $\{-1, 1\}$ such that X^0 is distributed according to a Bernoulli $\mathcal{B}(w, 1)$ with parameter w (that is $\mathbb{P}(X^0 = 1) = w$ and $\mathbb{P}(X^0 = -1) = 1 - w$) and the M other variables $X^1, \ldots, X^M$ are distributed

according to a Bernoulli $\mathcal{B}(1/2,1)$. The parameter $0 \leq w \leq 1$ will be chosen wisely in what follows.

For any $j \in \{1,\dots,M\}$, we consider the probability distribution $\pi_j = (P^X, \eta^{(j)})$ of a couple of random variables (X,Y) with values in $S_{M+1} \times \{-1,1\}$, where P^X is the probability distribution on S_{M+1} of $X = (X^0,\dots,X^M)$ and $\eta^{(j)}(x)$ is the regression function at the point $x \in \mathcal{S}_{M+1}$, of $Y=1$ knowing that $X=x$, given by

$$\eta^{(j)}(x) = \begin{cases} 1 & \text{if } x^0 = 1 \\ 1/2 + h/2 & \text{if } x^0 = -1, x^j = -1 \\ 1/2 + h & \text{if } x^0 = -1, x^j = 1 \end{cases}, \quad \forall x = (x^0, x^1, \dots, x^M) \in \mathcal{S}_{M+1},$$

where $h > 0$ is a parameter chosen wisely in what follows. The Bayes rule f^*, associated with the distribution $\pi_j = (P^X, \eta^{(j)})$, is identically equal to 1 on $\mathcal{S}_{M+1}$.

If the probability distribution of (X,Y) is π_j for a $j \in \{1,\dots,M\}$ then, for any $0 < t < 1$, we have $\mathbb{P}[|2\eta(X)-1| \leq t] \leq (1-w)\mathbb{1}_{h \leq t}$. Now, we take

$$1 - w = h^{\frac{1}{\kappa-1}},$$

then, we have $\mathbb{P}[|2\eta(X)-1| \leq t] \leq t^{\frac{1}{\kappa-1}}$ and so $\pi_j \in \mathcal{P}_\kappa$.

We extend the definition of the f_j's to the set $\mathcal{S}_{M+1}$ by $f_j(x) = x^j$ for any $x = (x^0,\dots,x^M) \in \mathcal{S}_{M+1}$ and $j = 1,\dots,M$. Consider $\mathcal{F} = \{f_1,\dots,f_M\}$. Assume that (X,Y) is distributed according to π_j for a $j \in \{1,\dots,M\}$. For any $k \in \{1,\dots,M\}$ and $k \neq j$, we have

$$A_0(f_k) - A_0^* = \sum_{x \in \mathcal{S}_{M+1}} |\eta(x) - 1/2||f_k(x) - 1|\mathbb{P}[X = x] = \frac{3h(1-w)}{8} + \frac{w}{2}$$

and the excess risk of f_j is given by $A_0(f_j) - A_0^* = (1-w)h/4 + w/2$. Thus, we have

$$\min_{f \in \mathcal{F}} A_0(f) - A_0^* = A_0(f_j) - A_0^* = (1-w)h/4 + w/2.$$

First, we prove the lower bound for any selector. Let $\tilde{f}_n$ be a selector with values in $\mathcal{F}$. If the underlying probability measure is π_j for a $j \in \{1,\dots,M\}$ then,

$$\begin{aligned} \mathbb{E}_n^{(j)}[A_0(\tilde{f}_n) - A_0^*] &= \sum_{k=1}^{M} (A_0(f_k) - A_0^*)\pi_j^{\otimes n}[\tilde{f}_n = f_k] \\ &= \min_{f \in \mathcal{F}}(A_0(f) - A_0^*) + \frac{h(1-w)}{8}\pi_j^{\otimes n}[\tilde{f}_n \neq f_j], \end{aligned}$$

where $\mathbb{E}_n^{(j)}$ denotes the expectation w.r.t. the observations D_n when (X,Y) is distributed according to π_j. Hence, we have

$$\max_{1 \leq j \leq M}\{\mathbb{E}_n^{(j)}[A_0(\tilde{f}_n) - A_0^*] - \min_{f \in \mathcal{F}}(A_0(f) - A_0^*)\} \geq \frac{h(1-w)}{8} \inf_{\hat{\phi}_n} \max_{1 \leq j \leq M} \pi_j^{\otimes n}[\hat{\phi}_n \neq j],$$

where the infimum $\inf_{\hat{\phi}_n}$ is taken over all tests valued in $\{1,\dots,M\}$ constructed from one observation in the model $(\mathcal{S}_{M+1}\times\{-1,1\},\mathcal{A}\times\mathcal{T},\{\pi_1,\dots,\pi_M\})^{\otimes n}$, where $\mathcal{T}$ is the natural $\sigma-$algebra on $\{-1,1\}$. Moreover, for any $j\in\{1,\dots,M\}$, we have

$$K(\pi_j^{\otimes n}|\pi_1^{\otimes n})\leq\frac{nh^2}{4(1-h-2h^2)},$$

where $K(P|Q)$ is the Kullback-Leibler divergence between P and Q (that is $\int\log(dP/dQ)dP$ if $P<<Q$ and $+\infty$ otherwise). Thus, if we apply Lemma 2 with $h=((\log M)/n)^{(\kappa-1)/(2\kappa-1)}$, we obtain the result.

Second, we prove the lower bound for the pERM procedure $\hat{f}_n=\tilde{f}_n^{pERM}$. Now, we assume that the probability distribution of (X,Y) is π_M and we take

$$h=\Big(C^2\frac{\log M}{n}\Big)^{\frac{\kappa-1}{2\kappa}}. \tag{12}$$

We have $\mathbb{E}[A_0(\hat{f}_n)-A_0^*]=\min_{f\in\mathcal{F}}(A_0(f)-A_0^*)+\dfrac{h(1-w)}{8}\mathbb{P}[\hat{f}_n\neq f_M]$. Now, we upper bound $\mathbb{P}[\hat{f}_n=f_M]$, conditionally to $\mathcal{Y}=(Y_1,\dots,Y_n)$. We have

$$\begin{aligned}&\mathbb{P}[\hat{f}_n=f_M|\mathcal{Y}]\\&=\mathbb{P}[\forall j=1,\dots,M-1,A_n^{\phi_0}(f_M)+\mathrm{pen}(f_M)\leq A_n^{\phi_0}(f_j)+\mathrm{pen}(f_j)|\mathcal{Y}]\\&=\mathbb{P}[\forall j=1,\dots,M-1,\nu_M\leq\nu_j+n(\mathrm{pen}(f_j)-\mathrm{pen}(f_M))|\mathcal{Y}],\end{aligned}$$

where $\nu_j=\sum_{i=1}^n\mathbb{1}_{(Y_iX_i^j\leq 0)},\forall j=1,\dots,M$ and $X_i=(X_i^j)_{j=0,\dots,M}\in\mathcal{S}_{M+1},\forall i=1,\dots,n$. Moreover, the coordinates $X_i^j,i=1,\dots,n;j=0,\dots,M$ are independent, $Y_1,\dots,Y_n$ are independent of $X_i^j,i=1,\dots,n;j=1,\dots,M-1$ and $|\mathrm{pen}(f_j)|\leq h^{\kappa/(\kappa-1)},\forall j=1,\dots,M$. So, we have

$$\begin{aligned}\mathbb{P}[\hat{f}_n=f_M|\mathcal{Y}]&=\sum_{k=0}^n\mathbb{P}[\nu_M=k|\mathcal{Y}]\prod_{j=1}^{M-1}\mathbb{P}[k\leq\nu_j+n(\mathrm{pen}(f_j)-\mathrm{pen}(f_M))|\mathcal{Y}]\\&\leq\sum_{k=0}^n\mathbb{P}[\nu_M=k|\mathcal{Y}]\Big(\mathbb{P}[k\leq\nu_1+2nh^{\kappa/(\kappa-1)}|\mathcal{Y}]\Big)^{M-1}\\&\leq\mathbb{P}[\nu_M\leq\bar{k}|\mathcal{Y}]+\big(\mathbb{P}[\bar{k}\leq\nu_1+2nh^{\kappa/(\kappa-1)}|\mathcal{Y}]\big)^{M-1},\end{aligned}$$

where

$$\begin{aligned}\bar{k}&=\mathbb{E}[\nu_M|\mathcal{Y}]-2nh^{\kappa/(\kappa-1)}\\&=\frac{1}{2}\sum_{i=1}^n\Big(\frac{2-4h}{2-3h}\mathbb{1}_{(Y_i=-1)}+\frac{1+h^{1/(\kappa-1)}(h/2-1/2)}{1+h^{1/(\kappa-1)}(3h/4-1/2)}\mathbb{1}_{(Y_i=1)}\Big)-2nh^{\kappa/(\kappa-1)}.\end{aligned}$$

Using Einmahl and Masson's concentration inequality (cf. [12]), we obtain

$$\mathbb{P}[\nu_M\leq\bar{k}|\mathcal{Y}]\leq\exp(-2nh^{2\kappa/(\kappa-1)}).$$

Using Berry-Esséen's theorem (cf. p.471 in [4]), the fact that $\mathcal{Y}$ is independent of $(X_i^j; 1 \le i \le n, 1 \le j \le M-1)$ and $\bar{k} \ge n/2 - 9nh^{\kappa/(\kappa-1)}/4$, we get

$$\mathbb{P}[\bar{k} \le \nu_1 + 2nh^{\frac{\kappa}{\kappa-1}} | \mathcal{Y}] \le \mathbb{P}\left[\frac{n/2-\nu_1}{\sqrt{n}/2} \le 6h^{\frac{\kappa}{\kappa-1}}\sqrt{n}\right] \le \Phi(6h^{\frac{\kappa}{\kappa-1}}\sqrt{n}) + \frac{66}{\sqrt{n}},$$

where Φ stands for the standard normal distribution function. Thus, we have

$$\mathbb{E}[A_0(\hat{f}_n) - A_0^*] \ge \min_{f\in\mathcal{F}}(A_0(f) - A_0^*) \tag{13}$$
$$+\frac{(1-w)h}{8}\Big(1 - \exp(-2nh^{2\kappa/(\kappa-1)}) - \Big(\Phi(6h^{\kappa/(\kappa-1)}\sqrt{n}) + 66/\sqrt{n}\Big)^{M-1}\Big).$$

Next, for any $a > 0$, by the elementary properties of the tails of normal distribution, we have

$$1 - \Phi(a) = \frac{1}{\sqrt{2\pi}}\int_a^{+\infty} \exp(-t^2/2)dt \ge \frac{a}{\sqrt{2\pi}(a^2+1)}e^{-a^2/2}. \tag{14}$$

Besides, we have for $0 < C < \sqrt{2}/6$ (a modification for $C = 0$ is obvious) and $(3376C)^2(2\pi M^{36C^2}\log M) \le n$, thus, if we replace h by its value given in (12) and if we apply (14) with $a = 16C\sqrt{\log M}$, then we obtain

$$\Big(\Phi(6h^{\kappa/(\kappa-1)}\sqrt{n}) + 66/\sqrt{n}\Big)^{M-1} \le \exp\Big[-\frac{M^{1-18C^2}}{18C\sqrt{2\pi\log M}} + \frac{66(M-1)}{\sqrt{n}}\Big]. \tag{15}$$

Combining (13) and (15), we obtain the result with $C_4 = (C/4)\Big(1 - \exp(-8C^2) - \exp(-1/(36C\sqrt{2\pi\log 2}))\Big) > 0$.

■

The following lemma is used to establish the lower bounds of Theorem 2. It is a version of Assouad's Lemma (cf. [28]). Proof can be found in [24].

Lemma 1. *Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. Consider a set of probability $\{P_\omega/\omega \in \Omega\}$ indexed by the cube $\Omega = \{0,1\}^m$. Denote by $\mathbb{E}_\omega$ the expectation under P_ω. Let $\theta \ge 1$ be a number. Assume that:*

$$\forall \omega, \omega' \in \Omega/\rho(\omega,\omega') = 1,\ H^2(P_\omega, P_{\omega'}) \le \alpha < 2,$$

then we have

$$\inf_{\hat{w}\in[0,1]^m}\max_{\omega\in\Omega}\mathbb{E}_\omega\left[\sum_{j=1}^m |\hat{w}_j - w_j|^\theta\right] \ge m2^{-3-\theta}(2-\alpha)^2$$

where the infimum $\inf_{\hat{w}\in[0,1]^m}$ is taken over all estimator based on an observation from the statistical experience $\{P_\omega|\omega\in\Omega\}$ and with values in $[0,1]^m$.

We use the following lemma to prove the weakness of selector aggregates. A proof can be found p. 84 in [28].

Lemma 2. *Let* $\mathbb{P}_1, \ldots, \mathbb{P}_M$ *be* M *probability measures on a measurable space* $(\mathcal{Z}, \mathcal{T})$ *satisfying* $\frac{1}{M}\sum_{j=1}^{M} K(\mathbb{P}_j|\mathbb{P}_1) \leq \alpha \log M$, *where* $0 < \alpha < 1/8$. *We have*

$$\inf_{\hat{\phi}} \max_{1\leq j\leq M} \mathbb{P}_j(\hat{\phi} \neq j) \geq \frac{\sqrt{M}}{1+\sqrt{M}}\Big(1 - 2\alpha - 2\sqrt{\frac{\alpha}{\log 2}}\Big),$$

where the infimum $\inf_{\hat{\phi}}$ *is taken over all tests* $\hat{\phi}$ *with values in* $\{1, \ldots, M\}$ *constructed from one observation in the statistical model* $(\mathcal{Z}, \mathcal{T}, \{\mathbb{P}_1, \ldots, \mathbb{P}_M\})$.

References

1. Audibert, J.-Y.: A randomized online learning algorithm for better variance control. Proceedings of the 19th Annual Conference on Learning Theory (COLT), (2006). 392-407.
2. Barron, A., and Li, J.: Mixture density estimation. Biometrics **53**. (1997), 603–618.
3. Bartlett P.L., Jordan M.I. and McAuliffe J.D.: Convexity, classification, and risk bounds. Journal of the American Statistical Association, **101**(473):138-156, (2006).
4. Bickel, P., and Doksum, K.: Mathematical Statistics: Basic Ideas and Selected Topics Volume 1. Prentice Hall, (2001).
5. Boucheron, S., Bousquet, O. and Lugosi, G.: Theory of classification: some recent advances. ESAIM Probability and Statistics **9**, 323-375, (2005).
6. Bühlmann, P., and Yu, B.: Analyzing bagging. Ann. Statist. **30**, 4 (2002), 927–961.
7. Catoni, O.: Statistical Learning Theory and Stochastic Optimization. Ecole d'été de Probabilités de Saint-Flour 2001, Lecture Notes in Mathematics. Springer, N.Y., (2001).
8. Cesa-Bianchi, N. and Lugosi, G.: Prediction, Learning, and Games. Cambridge University Press, New York, (2006).
9. Chesneau, C. and Lecué, G.: Adapting to unknown smoothness by aggregation of thresholded wavelet estimators. Submitted, (2006).
10. Cortes, C. and Vapnik, V.: Support-vector networks. Machine Learning **20**, 3 (1995), 273–297.
11. Devroye, L., Györfi, L. and Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer, New York, Berlin, Heidelberg, (1996).
12. Einmahl, U. and Mason, D.: Some Universal Results on the Behavior of Increments of Partial Sums. Ann. Probab. **24**, (1996), 2626–2635.
13. Freund, Y. and Schapire, R.:A decision-theoric generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences **55** (1997), 119–139.
14. Friedman, J., Hastie, T. and Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Ann. Statist. **28** (2000), 337–407.
15. Haussler, D., Kivinen, J. and Warmuth, M.K.: Sequential prediction of individual sequences under general loss functions. IEEE Trans. on Information Theory 44(5), 1906-1925.

16. Hartigan, J.: Bayesian regression using akaike priors. Yale University, New Haven, Preprint, 2002.
17. Juditsky, A., Rigollet, P., and Tsybakov, A.: Learning by mirror averaging. Preprint n.1034, LPMA.
18. Juditsky, A., Nazin, A., Tsybakov, A. B., Vayatis, N.: Recursive Aggregation of Estimators by Mirror Descent Algorithm with averaging. Problems of Information Transmission 41(4), 368-384
19. Kivinen, J. and Warmuth, M.K.: Averaging expert predictions. Proc. Fourth. European Conf. on Computational Learning Theory, H.U. Simon and P.Fischer, eds. Lecture Notes in Artificial Intelligence, vol. 1572. Springer, Berlin, 153-167.
20. Koltchinskii, V.: Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. Ann. Statist. **34**(6) (2006) ,1–50. 2004 IMS Medallion Lecture.
21. Lecué, G.: Optimal rates of aggregation in classification. Submitted, (2005).
22. Lecué, G.: Simultaneous adaptation to the margin and to complexity in classification. To appear in Ann. Statist. (2005).
23. Lecué, G. Optimal oracle inequality for aggregation of classifiers under low noise condition. In Proceeding of the 19th Annual Conference on Learning Theory, COLT 2006, 32, 4 (2006), 364–378.
24. Lecué, G.: Suboptimality of Penalized Empirical Risk Minimization. (2006). Manuscript.
25. Leung, G. and Barron, A.: Information theory and mixing least-square regressions. IEEE Transactions on Information Theory 52 (8) (2006), 3396–3410.
26. Lugosi, G., and Vayatis, N. On the Bayes-risk consistency of regularized boosting methods. Ann. Statist. **32**, 1 (2004), 30–55.
27. Nemirovski, A.: Topics in Non-parametric Statistics, vol. 1738 of Ecole d'été de Probabilités de Saint-Flour 1998, Lecture Notes in Mathematics. Springer, N.Y., 2000.
28. Tsybakov, A.: Introduction à l'estimation non-paramétrique. Springer, 2004.
29. Tsybakov, A. B.: Optimal rates of aggregation. Computational Learning Theory and Kernel Machines. B.Schölkopf and M.Warmuth, eds. Lecture Notes in Artificial Intelligence 2777 (2003), 303–313. Springer, Heidelberg.
30. Tsybakov, A. B.: Optimal aggregation of classifiers in statistical learning. Ann. Statist. **32**, 1 (2004), 135–166.
31. Vapnik, V.N. and Chervonenkis, A.Ya.: Necessary and sufficient conditions for the uniform convergence of empirical means to their true values. Teor. Veroyatn. Primen. **26**, 543-563, (1981),
32. Vovk, V.: Aggregating Strategies. In: Proceedings of the 3rd Annual Workshop on Computational Learning Theory, COLT1990, CA: Morgan Kaufmann (1990), 371–386.
33. Yang, Y.: Mixing strategies for density estimation. Ann. Statist. **28**, 1 (2000), 75–87.
34. Zhang, T.: Statistical behavior and consistency of classification methods based on convex risk minimization. Ann. Statist. **32**, 1 (2004), 56–85.
35. Zhang, T.: Adaptive estimation in Pattern Recognition by combining different procedures. Statistica Sinica **10**, (2000), 1069-1089.
36. Zhang, T.: From epsilon-entropy to KL-complexity: analysis of minimum information complexity density estimation. , To appear in Ann. Statist. (2006).