Optimal rates of aggregation in classification

Guillaume Lecué

Optimal rates of aggregation 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

In the same spirit as Tsybakov [2003], we define the optimality of an aggregation procedure in the problem of classification. Using an aggregate with exponential weights, we obtain an optimal rate of convex aggregation for the hinge risk under the margin assumption. Moreover we obtain an optimal rate of model selection aggregation under the margin assumption for the excess Bayes risk.

AMS 1991 subject classifications: 62G05, 62G20
Key words and phrases: Classification, statistical learning, aggregation of classifiers, optimal rates, margin.
Short title: Aggregation of Classifiers

1 Introduction

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. We assume that the space $\mathcal{X} \times \{-1, 1\}$ is endowed with an unknown probability measure π. We consider a random variable (X, Y) with values in $\mathcal{X} \times \{-1, 1\}$, such that π is the distribution of (X, Y). We denote by P^X the marginal of π on $\mathcal{X}$ and $\eta(x) = \mathbb{P}(Y = 1|X = x)$ the a posteriori probability of $Y = 1$ knowing that $X = x$. This setting means that in each point x of $\mathcal{X}$ we play to "heads or tails" with a biased coin such that "heads" arise with probability $\eta(x)$ and "tails" arise with probability $1-\eta(x)$. In the classification framework we have n i.i.d. observations of the couple (X, Y) denoted by $D_n = (X_i, Y_i)_{i=1,\dots,n}$, where X_i is the ith realisation of X and Y_i the result of the game at X_i (namely, $Y_i = 1$ if "heads" and $Y_i = -1$ if "tails"). The aim of classification is to predict the result Y for any X in $\mathcal{X}$. Obviously we have to make assumptions to be able to construct efficient prediction procedures. First one is on the way the coin is biased, especially in points of $\mathcal{X}$ of high P^X probability, one cannot predict a result of a "heads or tails" better than with probability $1/2$ if the coins is not biased. One assumption of this kind is called "margin assumption" and has been introduced in Tsybakov [2004].

We recall some usual notation introduced for the classification framework. A *prediction rule* is

a measurable function $f : \mathcal{X} \longmapsto \{-1, 1\}$. The *misclassification error* associated to f is

$$R(f) = \mathbb{P}(Y \neq f(X)).$$

It is well known (see, e.g., Devroye et al. [1996]) that

$$\min_f R(f) = R(f^*) = R^*,$$

where the prediction rule f^* is called *Bayes rule* associated to η and is defined by

$$f^*(x) = \text{sign}(2\eta(x) - 1), \forall x \in \mathcal{X}.$$

The minimal risk R^* is called the *Bayes risk*. A *classifier* is a function, $\hat{f}_n = \hat{f}_n(X, D_n)$, measurable with respect to D_n and X with values in $\{-1, 1\}$, that assigns to the sample D_n a prediction rule $\hat{f}_n(., D_n) : \mathcal{X} \longmapsto \{-1, 1\}$. A key characteristic of $\hat{f}_n$ is the *generalization error* $\mathbb{E}[R(\hat{f}_n)]$. Here

$$R(\hat{f}_n) = \mathbb{P}(Y \neq \hat{f}_n(X)|D_n).$$

The aim of statistical learning is to construct a classifier $\hat{f}_n$ such that $\mathbb{E}[R(\hat{f}_n)]$ is as close to R^* as possible. Accuracy of a classifier $\hat{f}_n$ is measured by the value $\mathbb{E}[R(\hat{f}_n) - R^*]$ called *excess Bayes risk* of $\hat{f}_n$. We say that the classifier $\hat{f}_n$ learns with the convergence rate $\psi(n)$, where $(\psi(n))_{n\in\mathbb{N}}$ is a decreasing sequence, if there exists an absolute constant $C > 0$ such that for any integer n, $\mathbb{E}[R(\hat{f}_n) - R^*] \leq C\psi(n)$.

The difficulty of classification is closely related to the behavior of the a posteriori probability η at the level $1/2$ (the distance $|\eta(\cdot) - 1/2|$ is sometimes called the margin). The paper Mammen and Tsybakov [1999], for the problem of discriminant analysis which is close to our classification problem, and Tsybakov [2004] have introduced the following assumption on the the margin:

(MA) Margin (or low noise) assumption. *The probability distribution π on the space $\mathcal{X} \times \{-1, 1\}$ satisfies the margin assumption MA(κ) with margin parameter $1 \leq \kappa < +\infty$ if there exists $c_0 > 0$ such that,*

$$\mathbb{E}\left\{|f(X) - f^*(X)|\right\} \leq c_0 \left(R(f) - R^*\right)^{1/\kappa}, \tag{1}$$

for all measurable functions f with values in $\{-1, 1\}$.

Under this assumption, the risk of a minimizer of the empirical risk over some fixed class $\mathcal{F}$ of decision rules can converge to R^* with *fast rates*, i.e., with the rates faster than $n^{-1/2}$. In fact, with no margin assumption, the convergence rate of the excess risk is not faster than $n^{-1/2}$ (cf. Devroye et al. [1996]). Under the margin assumption, it can be as fast as n^{-1}. Many examples of fast rates can be found in Blanchard et al. [2004], Scovel and Steinwart [2004, 2005], Massart [2000], Massart and Nédélec [2003], Massart [2004] and Audibert and Tsybakov [2005].

The aim of this paper is the following:

1. We define a notion of optimality for aggregation procedures in classification.
2. We introduce several aggregation procedures in classification and obtain exact oracle inequalities for their risks.

3. We prove lower bounds and show optimality of the suggested procedures and derive optimal rates of aggregation under the margin assumption.

The paper is organized as follows. In Section 2 we introduce definitions and the procedures which are used throughout the paper. Section 3 contains oracle inequalities for our aggregation procedures w.r.t. the excess hinge risk. Section 4 contains similar results for the excess Bayes risk. Proofs are given in Section 5.

2 Definitions and procedures

2.1 Loss functions

The quality of a classifier is often measured by a convex surrogate ϕ for the classification loss (Cortes and Vapnik [1995], Freund and Schapire [1997], Lugosi and Vayatis [2004], Friedman et al. [2000], Bühlmann and Yu [2002]).

Definition 1. *The real valued convex function ϕ on $\mathbb{R}$ is called* **convex loss for classification** *if $\phi(0) = 1$ and $\phi(x) = o(x)$ when x tends to infinity. The risk associated to the loss ϕ is called the* $\phi-$**risk** *and is defined by*

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

where $f : \mathcal{X} \longmapsto \mathbb{R}$ a measurable function. The **empirical** $\phi-$**risk** *is defined by*

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

If the minimum over all real valued functions exists, then we introduce $A_*^{(\phi)} = \min_f A^{(\phi)}(f)$.

Classifiers obtained by minimization of the empirical $\phi-$risk, for different convex losses, has been proved to have very good statistical properties (cf. Lugosi and Vayatis [2004], Blanchard et al. [2003], Zhang [2004], Scovel and Steinwart [2004, 2005] and Bartlett et al.). A wide variety of classification methods in machine learning are based on this idea, in particular, on using the convex loss associated to support vector machines (Cortes and Vapnik [1995], Schölkopf and Smola [2002]),

$$\phi(x) = (1 - x)_+,$$

called the *hinge-loss*, where $z_+ = \max(0, z)$ denotes the positive part of $z \in \mathbb{R}$. The corresponding risk is called the *hinge risk* and is defined by

$$A(f) = \mathbb{E}[(1 - Yf(X))_+],$$

for all $f : \mathcal{X} \longmapsto \mathbb{R}$ and the *optimal hinge risk* is defined by

$$A^* = \inf_f A(f), \tag{2}$$

where the infimum is taken over all measurable functions f. It is easy to check that the Bayes rule f^* attains the infimum in (2) and, moreover, Zhang [2004] has shown that,

$$R(f) - R^* \leq A(f) - A^*, \tag{3}$$

for all measurable functions f with values in $\mathbb{R}$, where we extend the definition of R to the class of real valued functions by $R(f) = R(\text{sign(f)})$. Thus minimization of the *excess hinge risk*, $A(f) - A^*$, provides a reasonable alternative for minimization of excess Bayes risk, $R(f) - R^*$.

2.2 Aggregation procedures

Now, we introduce the problem of aggregation and the aggregation procedures which will be studied in this paper.

Suppose that we have $M \geq 2$ different classifiers $\hat{f}_1, \ldots, \hat{f}_M$ taking values in $\{-1, 1\}$. The problem of model selection type aggregation, as studied in Nemirovski [2000], Yang [2000], Catoni [1999, 2001], Tsybakov [2003], consists in construction of a new classifier $\tilde{f}_n$ (called *aggregate*) which mimics approximatively the best classifier among $\hat{f}_1, \ldots, \hat{f}_M$. In most of these papers the aggregation is based on splitting of 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 first subsample D_m^1 is used to construct the classifiers $\hat{f}_1, \ldots, \hat{f}_M$ and the second subsample D_l^2 is used to aggregate them, i.e., to construct a new classifier that mimics in a certain sense the behavior of the best among the classifiers $\hat{f}_i$.

In this paper we will not consider the sample splitting and concentrate only on the construction of aggregates (following Juditsky and Nemirovski [2000], Tsybakov [2003], Birgé [2005], Bunea et al. [2004]). Thus, the first subsample is fixed and instead of classifiers $\hat{f}_1, \ldots, \hat{f}_M$, we have fixed prediction rules $f_1, \ldots, f_M$. Rather than working with a part of the initial sample we will suppose, for notational simplicity, that the whole sample D_n of size n is used for the aggregation step instead of a subsample D_l^2.

Let $\mathcal{F} = \{f_1, \ldots, f_M\}$ be a finite set of real-valued functions, where $M \geq 2$. An **aggregate** is a real valued statistic of the form:

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

where the weights $(w^{(n)}(f))_{f \in \mathcal{F}}$ satisfy

$$w^{(n)}(f) \geq 0, \quad \sum_{f \in \mathcal{F}} w^{(n)}(f) = 1.$$

Let ϕ be a convex loss for classification. The Empirical Risk Minimization **ERM** aggregate is defined by the weights,

$$\forall f \in \mathcal{F}: \quad w^{(n)}(f) = \begin{cases} 1 & \text{for one } f \in \mathcal{F} \text{ such that } A_n^{(\phi)}(f) = \min_{g \in \mathcal{F}} A_n^{(\phi)}(g), \\ 0 & \text{for other } f \in \mathcal{F}. \end{cases}$$

The ERM aggregate is denoted by $\tilde{f}_n^{(ERM)}$.

The **averaged ERM** aggregate is defined by the weights

$$\forall f \in \mathcal{F}, w^{(n)}(f) = \begin{cases} 1/N & \text{if} A_n^{(\phi)}(f) = \min_{g\in\mathcal{F}} A_n^{(\phi)}(g), \\ 0 & otherwise, \end{cases}$$

where N is the number of functions in $\mathcal{F}$ minimizing the empirical $\phi-$risk. The averaged ERM aggregate is denoted by $\tilde{f}_n^{(AERM)}$.

The Aggregation with Exponential Weights (**AEW**) aggregate is defined by the weights:

$$w^{(n)}(f) = \frac{\exp\left(-nA_n^{(\phi)}(f)\right)}{\sum_{g\in\mathcal{F}} \exp\left(-nA_n^{(\phi)}(g)\right)}, \quad \forall f \in \mathcal{F}. \tag{4}$$

The AEW aggregate is denoted by $\tilde{f}_n^{(AEW)}$.

The **cumulative AEW** aggregate is an on-line procedure defined by the weights:

$$w^{(n)}(f) = \frac{1}{n}\sum_{k=1}^{n} \frac{\exp\left(-kA_k^{(\phi)}(f)\right)}{\sum_{g\in\mathcal{F}} \exp\left(-kA_k^{(\phi)}(g)\right)}, \quad \forall f \in \mathcal{F}.$$

The cumulative AEW aggregate is denoted by $\tilde{f}_n^{(CAEW)}$.

There is a link between ERM, AERM and AEW aggregates. The following proposition states that the AEW aggregate is almost an ERM aggregate up to the residual term $\frac{\log M}{n}$, and the AERM aggregate is not worse than the ERM aggregate.

Proposition 1. *For any finite set $\mathcal{F}$ of real valued functions with cardinality M, and for any integers $M, n \geq 1$,*

$$A_n^{(\phi)}(\tilde{f}_n^{(AEW)}) \leq A_n^{(\phi)}(\tilde{f}_n^{(ERM)}) + \frac{\log M}{n},$$

and

$$A_n^{(\phi)}(\tilde{f}_n^{(AERM)}) \leq A_n^{(\phi)}(\tilde{f}_n^{(ERM)}).$$

When $\mathcal{F}$ is a class of prediction rules, intuitively, the AEW aggregate is more robust than the ERM aggregate w.r.t. the problem of overfitting. If the classifier with smallest empirical risk is overfitted, i.e., it fits too much to the observations, then the ERM aggregate will be overfitted. But, if other classifiers in $\mathcal{F}$ are good classifiers, the aggregate with exponential weights will consider their "opinions" in the final decision procedure and these opinions can balance with the opinion of the overfitted classifier in $\mathcal{F}$ which can be false because of its overfitting property. The ERM only considers the "opinion" of the classifier with the smallest risk, whereas the AEW takes into account all the opinions of the classifiers in the set $\mathcal{F}$. Moreover, the AEW aggregate does not need any minimization algorithm contrarily to the ERM aggregate.

The aggregation weights can be found in several situations. First, one can check that the solution of the following minimization problem

$$\min\left(\sum_{j=1}^{M} \lambda_j A_n^{(\phi)}(f_j) + \epsilon \sum_{j=1}^{M} \lambda_j \log\lambda_j : \sum_{j=1}^{M} \lambda_j \leq 1, \lambda_j \geq 0, j = 1, \ldots, M\right), \tag{5}$$

for all $\epsilon > 0$, is

$$\lambda_j = \frac{\exp\left(-\frac{A_n^{(\phi)}(f_j)}{\epsilon}\right)}{\sum_{k=1}^M \exp\left(-\frac{A_n^{(\phi)}(f_k)}{\epsilon}\right)}, \forall j = 1, \ldots, M.$$

Thus, for $\epsilon = 1/n$, we find the exponential weights used for the AEW aggregate. Second, these weights can also be found in the theory of prediction of individual sequences, cf. Vovk [1990].

2.3 Optimal Rates of Aggregation

In the same spirit as in Tsybakov [2003], where the regression problem is treated, we introduce a notion of optimality for an aggregation procedure and for rates of aggregation, in the classification framework. Our aim is to prove that the aggregates introduced above are optimal in the following sense. All the results are given under the margin assumption. We denote by $\mathcal{P}_\kappa$ the set of all probability measures π on $\mathcal{X} \times \{-1, 1\}$ satisfying the margin assumption with margin parameter $\kappa \geq 1$.

Definition 2. *Let ϕ be a convex loss for classification. The remainder term $\gamma(n, M, \kappa, \mathcal{F}, \pi)$ is called* **optimal rate of model selection type aggregation (MS-aggregation) for the $\phi-$risk**, *if the two following inequalities hold:*

(i) $\forall \mathcal{F} = \{f_1, \ldots, f_M\}$, there exists a statistic $\tilde{f}_n$, depending on $\mathcal{F}$, such that $\forall \pi \in \mathcal{P}_\kappa$, $\forall n \geq 1$,

$$\mathbb{E}\left[A^{(\phi)}(\tilde{f}_n) - A^{(\phi)*}\right] \leq \min_{f \in \mathcal{F}} \left(A^{(\phi)}(f) - A^{(\phi)*}\right) + C_1 \gamma(n, M, \kappa, \mathcal{F}, \pi). \tag{6}$$

(ii) $\exists \mathcal{F} = \{f_1, \ldots, f_M\}$ such that for any statistic $\bar{f}_n$, $\exists \pi \in \mathcal{P}_\kappa$, $\forall n \geq 1$

$$\mathbb{E}\left[A^{(\phi)}(\bar{f}_n) - A^{(\phi)*}\right] \geq \min_{f \in \mathcal{F}} \left(A^{(\phi)}(f) - A^{(\phi)*}\right) + C_2 \gamma(n, M, \kappa, \mathcal{F}, \pi). \tag{7}$$

Here, C_1 and C_2 are positive constants. Moreover, when these two inequalities are satisfied, we say that the procedure $\tilde{f}_n$, appearing in (6), is an **optimal MS-aggregate for the $\phi-$risk**. *If $\mathcal{C}$ denotes the convex hull of $\mathcal{F}$ and if (6) and (7) are satisfied with $\min_{f \in \mathcal{F}} \left(A^{(\phi)}(f) - A^{(\phi)*}\right)$ replaced by $\min_{f \in \mathcal{C}} \left(A^{(\phi)}(f) - A^{(\phi)*}\right)$ then, we say that $\gamma(n, M, \kappa, \mathcal{F}, \pi)$ is an* **optimal rate of convex aggregation type for the $\phi-$risk** *and $\tilde{f}_n$ is an* **optimal convex aggregation procedure for the $\phi-$risk**.

3 Optimal rates of convex aggregation for the hinge-loss.

Take M real valued functions $f_1, \ldots, f_M$. Consider the convex hull $\mathcal{C} = \text{Conv}(f_1, \ldots, f_M)$. We want to mimic the best function in $\mathcal{C}$ using the hinge risk and working under the margin assumption. Since we consider the hinge-risk, it suffices to use functions with values in $[-1, 1]$. In fact, for any real valued function f, we have $(1 - y\psi(f(x)))_+ \leq (1 - yf(x))_+$ for all $y \in \{-1, 1\}$ and $x \in \mathcal{X}$, so:

$$A(\psi(f)) - A^* \leq A(f) - A^*,$$

where

$$\psi(x) = \begin{cases} 1 & \text{if } x \geq 1 \\ x & \text{if } -1 \leq x \leq 1 \\ -1 & \text{if } x \leq -1, \end{cases} \qquad \forall x \in \mathbb{R}. \tag{8}$$

We first introduce the margin assumption with respect to the hinge-risk.

(MAH) Margin (or low noise) assumption for hinge-risk. *The probability distribution π on the space $\mathcal{X} \times \{-1, 1\}$ satisfies the margin assumption for hinge-risk MAH(κ) with parameter $1 \leq \kappa < +\infty$ if there exists $c > 0$ such that,*

$$\mathbb{E}\left[|f(X) - f^*(X)|\right] \leq c\left(A(f) - A^*\right)^{1/\kappa}, \tag{9}$$

for all functions f on $\mathcal{X}$ with values in $[-1, 1]$.

Proposition 2. *The assumption MAH(κ) is equivalent to the margin assumption MA(κ).*

In what follows we will assume that MA(κ) holds and thus also MAH(κ) holds.

The AEW aggregate, introduced in (4) for a general convex loss, has a simple form for the special case of the hinge-risk:

$$\tilde{f}_n = \sum_{j=1}^{M} w^{(n)}(f_j) f_j, \tag{10}$$

where

$$w^{(n)}(f_j) = \frac{\exp\left(\sum_{i=1}^{n} Y_i f_j(X_i)\right)}{\sum_{k=1}^{M} \exp\left(\sum_{i=1}^{n} Y_i f_k(X_i)\right)}, \quad \forall j = 1, \ldots, M, \tag{11}$$

where $f_1, \ldots, f_M$ are functions with values in $[-1, 1]$.

We want to prove optimality of our aggregates in the sense of Definition 2. Therefore, we need to show an exact oracle inequality of type (6) for our aggregates and a lower bound inequality of type (7). These inequalities are given in Theorems 1 and 2.

Theorem 1 (Oracle inequality). *Let $\kappa \geq 1$. We assume that π satisfies the margin assumption MA(κ). We denote by $\mathcal{C}$ the convex hull of a finite set of functions with values in $[-1, 1]$, $\mathcal{F} = \{f_1, \ldots, f_M\}$. Let $\tilde{f}_n$ be either of the four aggregates introduced in Section 2.2. Then, for any integers $M \geq 3, n \geq 1$, $\tilde{f}_n$ satisfies the following inequality*

$$\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] \leq \min_{f \in \mathcal{C}}(A(f) - A^*) + C\left(\sqrt{\frac{\min_{f \in \mathcal{C}}(A(f) - A^*)^{\frac{1}{\kappa}} \log M}{n}} + \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa - 1}}\right),$$

where $C = 32(6 \vee 537c \vee 16(2c + 1/3))$ for the ERM, AERM and AEW aggregates with $\kappa \geq 1$ and $c > 0$ is the constant in (9) and $C = 32(6 \vee 537c \vee 16(2c + 1/3))(2 \vee (2\kappa - 1)/(\kappa - 1)$ for the CAEW aggregate with $\kappa > 1$. For $\kappa = 1$ the CAEW aggregate satisfies

$$\mathbb{E}\left[A(\tilde{f}_n^{(CAEW)}) - A^*\right] \leq \min_{f \in \mathcal{C}}(A(f) - A^*) + 2C\left(\sqrt{\frac{\min_{f \in \mathcal{C}}(A(f) - A^*) \log M}{n}} + \frac{\log M \log n}{n}\right).$$

Remark 1. *The hinge loss is linear on* $[-1,1]$*, thus, MS-aggregation or convex aggregation of functions with values in* $[-1,1]$ *are identical problems if we use the hinge risk. Namely,*

$$\min_{f\in\mathcal{F}} A(f) = \min_{f\in\mathcal{C}} A(f).$$

Theorem 2 (Lower bound). *Let* $\kappa \geq 1$*,* M, n *be integer such that* $\log M \leq n$*. There exists an absolute constant* $C > 0$*, depending only on* κ *and* c*, and a set of prediction rules* $\mathcal{F} = \{f_1, \ldots, f_M\}$ *such that for any procedure* $\bar{f}_n$ *with values in* $\mathbb{R}$*, there exists a probability measure* π *satisfying the margin assumption MA(*κ*) for which*

$$\mathbb{E}\left[A(\bar{f}_n) - A^*\right] \geq \min_{f\in\mathcal{C}}(A(f) - A^*) + C\left(\sqrt{\frac{(\min_{f\in\mathcal{C}} A(f) - A^*)^{\frac{1}{\kappa}}\log M}{n}} + \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}\right),$$

where $C = c^{\kappa}(4e)^{-1}2^{-2\kappa(\kappa-1)/(2\kappa-1)}(\log 2)^{-\kappa/(2\kappa-1)}$ *and* $c > 0$ *is the constant in (9).*

Combining the exact oracle inequality of Theorem 1 and the lower bound of Theorem 2, we see that the residual

$$\sqrt{\frac{(\min_{f\in\mathcal{C}} A(f) - A^*)^{\frac{1}{\kappa}}\log M}{n}} + \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}},$$

is optimal rate of convex aggregation of M functions with values in $[-1,1]$ for the hinge-loss. Moreover, by aggregating $\psi(f_1), \ldots, \phi(f_M)$, it is easy to check that

$$\sqrt{\frac{(\min_{f\in\mathcal{F}} A(\psi(f)) - A^*)^{\frac{1}{\kappa}}\log M}{n}} + \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}},$$

is optimal rate of model-selection aggregation of M real valued functions w.r.t. the hinge loss. In both cases, the aggregate with exponential weights as well as ERM and AERM attain these optimal rates. Learning properties of the AEW procedure can be found in Lecué [2005] and Lecué [2006]. In Theorem 1 the AEW procedure satisfies an exact oracle inequality with an optimal residual whereas in Lecué [2005] and Lecué [2006] the oracle inequalities satisfied by the AEW procedure are not exact and in Lecué [2005] the residual is not optimal. The CAEW aggregate attains the optimal rate if $\kappa > 1$. It is interesting to note that these rates depend on both the class $\mathcal{F}$ and π. Namely, in the convex case, the term

$$\min_{f\in\mathcal{C}} A(f) - A^*$$

appears in the rate. This is different from the regression problem (cf. Tsybakov [2003]), where the optimal aggregation rates depends only on M and n. We denote by $\mathcal{M}(\mathcal{F},\pi)$ the minimum $\min_{f\in\mathcal{C}}(A(f) - A^*)$. Three cases can be considered:

1. If $\mathcal{M}(\mathcal{F},\pi) \leq a\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}$, for an absolute constant $a > 0$, then the hinge risk of our aggregates attains $\min_{f\in\mathcal{C}} A(f) - A^*$ with the rate $\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}$, which can be $\log M/n$ in the case $k = 1$.

2. If $a\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}} \leq \mathcal{M}(\mathcal{F},\pi) \leq b$, for some absolute constants $a, b > 0$, then our aggregates mimics the best prediction rule in $\mathcal{C}$ with a rate slower than $\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}$ but faster than $\sqrt{\frac{\log M}{n}}$.

3. If $\mathcal{M}(\mathcal{F},\pi) \geq a > 0$, where $a > 0$ is a constant, then the rate of aggregation is $\sqrt{\frac{\log M}{n}}$, as in the case of no margin assumption.

We can explain this behavior by the fact that not only κ but also $\min_{f\in\mathcal{C}} A(f) - A^*$ measures the difficulty of classification. For instance, in the extreme case where $\min_{f\in\mathcal{C}} A(f) - A^* = 0$, which means that $\mathcal{C}$ contains the Bayes rule, we have the fastest rate $\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}$. In the worst cases, which are realized when κ tends to ∞ or $\min_{f\in\mathcal{C}}(A(f) - A^*) \geq a > 0$, where $a > 0$ is a constant, the optimal rate of aggregation is a slow rate $\sqrt{\frac{\log M}{n}}$. Optimal rates of aggregation obtained in Tsybakov [2003] depends only on M and n.

4 Optimal rates of model selection aggregation for the excess risk.

Now, we provide oracle inequalities and lower bounds for the excess Bayes risk. First, we can deduce from Theorem 1 and 2, "almost optimal rates of aggregation" for the excess Bayes risk achieved by the AEW aggregate. Second, using the ERM aggregate, we obtain optimal rates of model selection aggregation for the excess Bayes risk.

Using Zhang's inequality we can derive from Theorem 1, an oracle inequality for the excess Bayes risk. The lower bound is obtained using the same proof as in Theorem 2.

Corollary 1. *Let $\mathcal{F} = \{f_1, \ldots, f_M\}$ be a finite set of prediction rules for an integer $M \geq 3$. Let $\kappa \geq 1$. We assume that π satisfies MA(κ). Denote by $\tilde{f}_n$ either the ERM or the AERM or the AEW aggregate. Then, $\tilde{f}_n$ satisfies for any number $a > 0$ and any integer n:*

$$\mathbb{E}\left[R(\tilde{f}_n) - R^*\right] \leq 2(1+a) \min_{j=1,\ldots,M}(R(f_j) - R^*) + \left(\frac{1}{a}\right)^{\frac{1}{2\kappa-1}} C\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}, \qquad (12)$$

where $C = 32(6 \vee 537c \vee 16(2c + 1/3))$. The CAEW aggregate satisfies the same inequality with $C = 32(6 \vee 537c \vee 16(2c + 1/3))(2 \vee (2\kappa - 1)/(\kappa - 1))$ when $\kappa > 1$. For $\kappa = 1$ the CAEW aggregate satisfies (12) where we need to multiply by $\log n$ the residual.

Moreover there exists a finite set of prediction rules $\mathcal{F} = \{f_1, \ldots, f_M\}$ such that for any classifier $\bar{f}_n$, there exists a probability measure π on $\mathcal{X} \times \{-1, 1\}$ satisfying the Margin Assumption with margin parameter κ, such that for any $n \geq 1, a > 0$,

$$\mathbb{E}\left[R(\bar{f}_n) - R^*\right] \geq 2(1+a) \min_{f\in\mathcal{F}}(R(f) - R^*) + C(a)\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}},$$

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

Due to Corollary 1,

$$\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}$$

is an almost optimal rate of MS-aggregation for the excess risk and the AEW aggregate achieves this rate. The word "almost" is here because $\min_{f\in\mathcal{F}}(R(f)-R^*)$ is multiplied by a constant which is greater than 1.

Remark 2. *Some applications of Corollary 1, can be found in Lecué [2005] and Lecué [2006]. In particular, adaptive SVM classifiers are constructed by aggregating SVM estimators (this procedure requires the construction of only* $(\log n)^2$ *SVM estimators).*

The last oracle inequality of Theorem 1 is not an exact one since the minimal excess risk over $\mathcal{F}$ is multiplied by the constant $2(1+a)>1$. This is not the case while using the ERM aggregate as explained in the following Theorem.

Theorem 3. *Let $\kappa\geq 1$. We assume that π satisfies MA(κ). We denote by $\mathcal{F}=\{f_1,\ldots,f_M\}$ a set of prediction rules. The ERM aggregate over $\mathcal{F}$ satisfies for any integer $n\geq 1$:*

$$\mathbb{E}\left[R(\tilde{f}_n^{(ERM)})-R^*\right]\leq \min_{f\in\mathcal{F}}(R(f)-R^*)+C\left(\sqrt{\frac{\min_{f\in\mathcal{F}}(R(f)-R^*)^{\frac{1}{\kappa}}\log M}{n}}+\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}\right),$$

where $C=32(6\vee 537c_0\vee 16(2c_0+1/3))$ and c_0 is the constant appearing in MA(κ).

Using Lemma 4, we can deduce the results of Herbei and Wegkamp [2005] from Theorem 3. Oracle inequalities under the margin assumption have already been stated in Massart [2004] (cf. Boucheron et al. [2005]). But the remainder term obtained is worse than the one obtain here or in Herbei and Wegkamp [2005].

According to Definition 2, combining Theorem 3 and the following Theorem, the rate

$$\sqrt{\frac{\min_{f\in\mathcal{F}}(R(f)-R^*)^{\frac{1}{\kappa}}\log M}{n}}+\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}$$

is an optimal rate of MS-aggregation w.r.t. the excess Bayes risk. The ERM aggregate achieves this rate.

Theorem 4 (Lower bound). *Let $M\geq 3$ and n be two integers such that $\log M\leq n$ and $\kappa\geq 1$ an integer. There exists an absolute constant $C>0$ and a set of prediction rules $\mathcal{F}=\{f_1,\ldots,f_M\}$ such that for any procedure $\bar{f}_n$ with values in $\mathbb{R}$, there exists a probability measure π satisfying the margin assumption MA(κ) for which*

$$\mathbb{E}\left[R(\bar{f}_n)-R^*\right]\geq \min_{f\in\mathcal{F}}(R(f)-R^*)+C\left(\sqrt{\frac{(\min_{f\in\mathcal{C}}R(f)-R^*)^{\frac{1}{\kappa}}\log M}{n}}+\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}\right),$$

where $C=c_0{}^{\kappa}(4e)^{-1}2^{-2\kappa(\kappa-1)/(2\kappa-1)}(\log 2)^{-\kappa/(2\kappa-1)}$ and c_0 is the constant appearing in the margin assumption MA(κ).

5 Proofs

Proof of Proposition 1: We start by showing that the AEW aggregate is almost as good as the ERM aggregate up to a $(\log M)/n$ term.

Since ϕ is convex we have $\phi(Y\tilde{f}_n(X)) \leq \sum_{f\in\mathcal{F}} w^{(n)}(f)\phi(Yf(X))$, thus

$$A_n^{(\phi)}(\tilde{f}_n) \leq \sum_{f\in\mathcal{F}} w^{(n)}(f) A_n^{(\phi)}(f).$$

We have for all $f \in \mathcal{F}$, $A_n^{(\phi)}(f) = A_n^{(\phi)}(\tilde{f}_n^{(ERM)}) + \frac{1}{n}\left(\log(w^{(n)}(\tilde{f}_n^{(ERM)})) - \log(w^{(n)}(f))\right)$ and by averaging over the $w^{(n)}(f)$ we get :

$$A_n^{(\phi)}(\tilde{f}_n) \leq \min_{f\in\mathcal{F}} A_n^{(\phi)}(f) + \frac{\log(M)}{n}. \tag{13}$$

Since $\sum_{f\in\mathcal{F}} w^{(n)}(f)\log\left(\frac{w^{(n)}(f)}{1/M}\right) = K(w|u) \geq 0$ where $K(w|u)$ denotes the Kullback-Leibler divergence between the weights $w = (w^{(n)}(f))_{f\in\mathcal{F}}$ and the uniform weights $u = (1/M)_{f\in\mathcal{F}}$.

The convexity of ϕ leads directly to the result for the AERM aggregate.

Proof of Proposition 2: Since for any function f from $\mathcal{X}$ to $\{-1,1\}$ we have $2(R(f) - R^*) = A(f) - A^*$, then, MA($\kappa$) is implied by MAH($\kappa$).

Assume that MA(κ) holds. We first explore the case $\kappa > 1$, then, MA(κ) implies that there exist a constant $c_1 > 0$ such that $\mathbb{P}(|2\eta(X) - 1| \leq t) \leq c_1 t^{1/(\kappa-1)}$ for any $t > 0$. Let f from $\mathcal{X}$ to $[-1,1]$. We have for any $0 \leq t$:

$$A(f) - A^* = \mathbb{E}\left[|2\eta(X) - 1||f(X) - f^*(X)|\right] \geq t\mathbb{E}\left[|f(X) - f^*(X)|\mathbb{1}_{|2\eta(X)-1|\geq t}\right]$$
$$\geq t\left(\mathbb{E}\left[|f(X) - f^*(X)|\right] - 2\mathbb{P}\left(|2\eta(X) - 1| \leq t\right)\right) \geq t\left(\mathbb{E}\left[|f(X) - f^*(X)|\right] - 2c_1 t^{1/(\kappa-1)}\right).$$

For $t_0 = ((\kappa-1)/(2c_1\kappa))^{\kappa-1}\mathbb{E}\left[|f(X) - f^*(X)|\right]^{\kappa-1}$, we obtain:

$$A(f) - A^* \geq ((\kappa-1)/(2c_1\kappa))^{\kappa-1}\kappa^{-1}\mathbb{E}\left[|f(X) - f^*(X)|\right]^{\kappa}.$$

For the case $\kappa = 1$, assumption MA(κ) implies that there exists $h > 0$ such that $|2\eta(X) - 1| \geq h$ a.s.. Indeed, if for any $N \in \mathbb{N}^*$, there exists $A_N \in \mathcal{A}$ such that $P^X(A_N) > 0$ and $|2\eta(x) - 1| \leq 1/N, \forall x \in A_N$, then, for

$$f_N(x) = \begin{cases} -f^*(x) & \text{if } x \in A_N \\ f^*(x) & \text{otherwise,} \end{cases}$$

we obtain $R(f_N) - R^* \leq 2P^X(A_N)/N$ and $\mathbb{E}\left[|f_N(X) - f^*(X)|\right] = 2P^X(A_N)$, and there is no constant $c_0 > 0$ such that $P^X(A_N) \leq c_0 P^X(A_N)/N$ for all $N \in \mathbb{N}^*$. So, assumption MA(1) does not hold if no $h > 0$ satisfies $|2\eta(X) - 1| \geq h$ a.s.. Thus, for any f from $\mathcal{X}$ to $[-1,1]$, we have $A(f) - A^* = \mathbb{E}\left[|2\eta(X) - 1||f(X) - f^*(X)|\right] \geq h\mathbb{E}\left[|f(X) - f^*(X)|\right]$.

Proof of Theorem 1: Let $\tilde{f}_n$ be either the ERM or the AERM or the AEW aggregate for the class $\mathcal{F} = \{f_1, \ldots, f_M\}$. We have in all the cases:

$$A_n(\tilde{f}_n) \leq \min_{i=1,\ldots,M} A_n(f_i) + \frac{\log(M)}{n}. \tag{14}$$

Let $\epsilon > 0$. We consider $\mathcal{D} = \{f \in \mathcal{C} : A(f) > A_{\mathcal{C}} + 2\epsilon\}$, where $A_{\mathcal{C}} = \min_{f\in\mathcal{C}} A(f)$. Let $x > 0$. If

$$\sup_{f\in\mathcal{D}} \frac{A(f) - A^* - (A_n(f) - A_n(f^*))}{A(f) - A^* + x} \leq \frac{\epsilon}{A_{\mathcal{C}} - A^* + 2\epsilon + x},$$

then for any $f \in \mathcal{D}$, we have

$$A_n(f) - A_n(f^*) \geq A(f) - A^* - \frac{\epsilon(A(f) - A^* + x)}{(A_{\mathcal{C}} - A^* + 2\epsilon + x)} \geq A_{\mathcal{C}} - A^* + \epsilon,$$

because $A(f) - A^* \geq A_{\mathcal{C}} - A^* + 2\epsilon$. Hence,

$$\begin{aligned}
&\mathbb{P}\left[\inf_{f\in\mathcal{D}} (A_n(f) - A_n(f^*)) < A_{\mathcal{C}} - A^* + \epsilon\right] \\
&\quad \leq \quad \mathbb{P}\left[\sup_{f\in\mathcal{D}} \frac{A(f) - A^* - (A_n(f) - A_n(f^*))}{A(f) - A^* + x} > \frac{\epsilon}{A_{\mathcal{C}} - A^* + 2\epsilon + x}\right].
\end{aligned}$$

Take $f' \in \{f_1, \ldots, f_M\}$ such that $A(f') = \min_{j=1,\ldots,M} A(f_j)$. Observe that a linear function achieves its maximum over a convex polygon at one of the vertices of the polygon. Thus, since we are working in the linear part of the hinge-loss (f_j's take their values in $[-1, 1]$), we have $A_{\mathcal{C}} = \inf_{f\in\mathcal{C}} A(f) = \inf_{f\in\{f_1,..,f_M\}} A(f) = A(f')$. If $A(\tilde{f_n}) > A_{\mathcal{C}} + 2\epsilon$ then $\tilde{f_n} \in \mathcal{D}$. On the other hand, $A(\tilde{f_n}) \leq \min_{j=1,\ldots,M} A_n(f_j) + \frac{\log(M)}{n} \leq A_n(f') + \frac{\log(M)}{n}$. Hence, there exists $f \in \mathcal{D}$ such that $A_n(f) - A_n(f^*) \leq A_n(f') - A_n(f^*) + \frac{\log M}{n}$. We have:

$$\begin{aligned}
&\mathbb{P}\left[A(\tilde{f_n}) > A_{\mathcal{C}} + 2\epsilon\right] \leq \mathbb{P}\left[\inf_{f\in\mathcal{D}} A_n(f) - A_n(f^*) \leq A_n(f') - A_n(f^*) + \frac{\log M}{n}\right] \\
&\quad \leq \quad \mathbb{P}\left[\inf_{f\in\mathcal{D}} A_n(f) - A_n(f^*) \leq A_{\mathcal{C}} - A^* + \epsilon\right] + \mathbb{P}\left[A_n(f') - A_n(f^*) \geq A_{\mathcal{C}} - A^* + \epsilon - \frac{\log M}{n}\right] \\
&\quad \leq \quad \mathbb{P}\left[\sup_{f\in\mathcal{C}} \frac{A(f) - A^* - (A_n(f) - A_n(f^*))}{A(f) - A^* + x} > \frac{\epsilon}{A_{\mathcal{C}} - A^* + 2\epsilon + x}\right] \\
&\qquad + \mathbb{P}\left[A_n(f') - A_n(f^*) \geq A_{\mathcal{C}} - A^* + \epsilon - \frac{\log M}{n}\right].
\end{aligned}$$

If we assume that

$$\sup_{f\in\mathcal{C}} \frac{A(f) - A^* - (A_n(f) - A_n(f^*))}{A(f) - A^* + x} > \frac{\epsilon}{A_{\mathcal{C}} - A^* + 2\epsilon + x},$$

then, there exists $f = \sum_{j=1}^M w_j f_j \in \mathcal{C}$ (where $w_j \geq 0$ and $\sum w_j = 1$), such that

$$\frac{A(f) - A^* - (A_n(f) - A_n(f^*))}{A(f) - A^* + x} > \frac{\epsilon}{A_{\mathcal{C}} - A^* + 2\epsilon + x}.$$

The linearity of the hinge loss on $[-1, 1]$ leads to

$$\frac{A(f) - A^* - (A_n(f) - A_n(f^*))}{A(f) - A^* + x} = \frac{\sum_{j=1}^M w_j[A(f_j) - A^* - (A_n(f_j) - A_n(f^*))}{\sum_{j=1}^M w_j[A(f_j) - A^* + x]}$$

and according to Lemma 3, we have

$$\max_{j=1,\dots,M} \frac{A(f_j) - A^* - (A_n(f_j) - A_n(f^*))}{A(f_j) - A^* + x} > \frac{\epsilon}{A_{\mathcal{C}} - A^* + 2\epsilon + x}.$$

We now use the relative concentration inequality of Lemma 1 to obtain:

$$\begin{aligned}
&\mathbb{P}\left[\max_{j=1,\dots,M} \frac{A(f_j) - A^* - (A_n(f_j) - A_n(f^*))}{A(f_j) - A^* + x} > \frac{\epsilon}{A_{\mathcal{C}} - A^* + 2\epsilon + x}\right] \\
&\leq \quad M\left(1 + \frac{8c(A_{\mathcal{C}} - A^* + 2\epsilon + x)^2 x^{1/\kappa}}{n(\epsilon x)^2}\right) \exp\left(-\frac{n(\epsilon x)^2}{8c(A_{\mathcal{C}} - A^* + 2\epsilon + x)^2 x^{1/\kappa}}\right) \\
&\quad + M\left(1 + \frac{16(A_{\mathcal{C}} - A^* + 2\epsilon + x)}{3n\epsilon x}\right) \exp\left(-\frac{3n\epsilon x}{16(A_{\mathcal{C}} - A^* + 2\epsilon + x)}\right).
\end{aligned}$$

Using the assumption MAH(κ) (which is implied by MA(κ) to upper bound the variance term and applying Bernstein's inequality we get

$$\mathbb{P}\left[A_n(f') - A_n(f^*) \geq A_{\mathcal{C}} - A^* + \epsilon - \frac{\log M}{n}\right] \leq \exp\left(-\frac{n(\epsilon - (\log M)/n)^2}{4c(A_{\mathcal{C}} - A^*)^{1/\kappa} + (8/3)(\epsilon - (\log M)/n)}\right),$$

for any $\epsilon > (\log M)/n$. We take $x = A_{\mathcal{C}} - A^* + 2\epsilon$, then, for any $(\log M)/n < \epsilon < 1$, we have:

$$\mathbb{P}\left(A(\tilde{f}_n) > A_{\mathcal{C}} + 2\epsilon\right) \leq \exp\left(-\frac{n(\epsilon - \log M/n)^2}{4c(A_{\mathcal{C}} - A^*)^{1/\kappa} + (8/3)(\epsilon - (\log M)/n)}\right)$$

$$+M\left(1 + \frac{32c(A_{\mathcal{C}} - A^* + 2\epsilon)^{1/\kappa}}{n\epsilon^2}\right) \exp\left(-\frac{n\epsilon^2}{32c(A_{\mathcal{C}} - A^* + 2\epsilon)^{1/\kappa}}\right) + M\left(1 + \frac{32}{3n\epsilon}\right) \exp\left(-\frac{3n\epsilon}{32}\right).$$

Thus, for $2(\log M)/n < u < 1$, we have:

$$\mathbb{E}\left[A(\tilde{f}_n) - A_{\mathcal{C}}\right] \leq 2u + 2\int_{u/2}^{1} \left[T_1(\epsilon) + M(T_2(\epsilon) + T_3(\epsilon))\right] d\epsilon, \tag{15}$$

where

$$T_1(\epsilon) = \exp\left(-\frac{n(\epsilon - (\log M)/n)^2}{4c((A_{\mathcal{C}} - A^*)/2)^{1/\kappa} + (8/3)(\epsilon - (\log M)/n)}\right),$$

$$T_2(\epsilon) = \left(1 + \frac{64c(A_{\mathcal{C}} - A^* + 2\epsilon)^{1/\kappa}}{2^{1/\kappa} n\epsilon^2}\right) \exp\left(-\frac{2^{1/\kappa} n\epsilon^2}{64c(A_{\mathcal{C}} - A^* + 2\epsilon)^{1/\kappa}}\right)$$

and

$$T_3(\epsilon) = \left(1 + \frac{16}{3n\epsilon}\right) \exp\left(-\frac{3n\epsilon}{16}\right).$$

Set $\beta_1 = \min(32^{-1}, (2148c)^{-1}, (64(2c + 1/3))^{-1})$ where the constant $c > 0$ appears in MAH(κ). We recall that $A_{\mathcal{C}} = \min_{f \in \mathcal{C}} A(f)$. Consider separately the following cases (1) and (2).

(1) The case $A_{\mathcal{C}} - A^* \geq (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$.

Denote by $\mu(M)$ the unique solution of $X_0 = 3M\exp(-X_0)$. Then, clearly $\log M/2 \leq \mu(M) \leq \log M$. Take u such that $(n\beta_1 u^2)/(A_{\mathcal{C}} - A^*)^{1/\kappa} = \mu(M)$. Using the definition of case (1) and of $\mu(M)$ we get $u \leq A_{\mathcal{C}} - A^*$. Moreover, $u \geq 4\log M/n$. Then

$$\int_{\frac{u}{2}}^{1} T_1(\epsilon)d\epsilon \leq \int_{\frac{u}{2}}^{\frac{A_{\mathcal{C}} - A^*}{2}} \exp\left(-\frac{n(\epsilon/2)^2}{(4c + 4/3)(A_{\mathcal{C}} - A^*)^{1/\kappa}}\right) d\epsilon + \int_{\frac{A_{\mathcal{C}} - A^*}{2}}^{1} \exp\left(-\frac{n(\epsilon/2)^2}{(8c + 4/3)\epsilon^{1/\kappa}}\right) d\epsilon.$$

Using Lemma 2 and the inequality $u \leq A_{\mathcal{C}} - A^*$, we obtain

$$\int_{u/2}^{1} T_1(\epsilon)d\epsilon \leq \frac{64(2c+1/3)(A_{\mathcal{C}} - A^*)^{1/\kappa}}{nu} \exp\left(-\frac{nu^2}{64(2c+1/3)(A_{\mathcal{C}} - A^*)^{1/\kappa}}\right). \tag{16}$$

We have $128c(A_{\mathcal{C}} - A^* + u) \leq nu^2$ thus, using Lemma 2, we get

$$\int_{u/2}^{1} T_2(\epsilon)d\epsilon \leq 2\int_{u/2}^{(A_{\mathcal{C}}-A^*)/2} \exp\left(-\frac{n\epsilon^2}{64c(A_{\mathcal{C}} - A^*)^{1/\kappa}}\right)d\epsilon + 2\int_{(A_{\mathcal{C}}-A^*)/2}^{1} \exp\left(-\frac{n\epsilon^{2-1/\kappa}}{128c}\right)d\epsilon$$

$$\leq \frac{2148c(A_{\mathcal{C}} - A^*)^{1/\kappa}}{nu} \exp\left(-\frac{nu^2}{2148c(A_{\mathcal{C}} - A^*)^{1/\kappa}}\right). \tag{17}$$

We have $u \geq 32(3n)^{-1}$ so

$$\int_{u/2}^{1} T_3(\epsilon) \leq \frac{64}{3n} \exp\left(-\frac{3nu}{64}\right) \leq \frac{64(A_{\mathcal{C}} - A^*)^{1/\kappa}}{3nu} \exp\left(-\frac{3nu^2}{64(A_{\mathcal{C}} - A^*)^{1/\kappa}}\right). \tag{18}$$

From (16), (17), (18) and (15) we obtain

$$\mathbb{E}\left[A(\tilde{f}_n) - A_{\mathcal{C}}\right] \leq 2u + 6M\frac{(A_{\mathcal{C}} - A^*)^{1/\kappa}}{n\beta_1 u} \exp\left(-\frac{n\beta_1 u}{(A_{\mathcal{C}} - A^*)^{1/\kappa}}\right).$$

The definition of u leads to $\mathbb{E}\left[A(\tilde{f}_n) - A_{\mathcal{C}}\right] \leq 3\sqrt{\frac{(A_{\mathcal{C}}-A^*)^{1/\kappa}\log M}{n\beta_1}}$.

(2)The case $A_{\mathcal{C}} - A^* \leq (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$.

We choose now u such that $n\beta_2 u^{(2\kappa-1)/\kappa} = \mu(M)$. Where $\beta_2 = \min(3(32(6c+1))^{-1}, (256c)^{-1}, 3/64)$. Using the definition of case (2) and of $\mu(M)$ we get $u \geq A_{\mathcal{C}} - A^*$.
Using the fact that $u > 4\log M/n$ and Lemma 2, we have

$$\int_{u/2}^{1} T_1(\epsilon)d\epsilon \leq \frac{32(6c+1)}{3nu} \exp\left(-\frac{3nu^{2-1/\kappa}}{32(6c+1)}\right). \tag{19}$$

We have $u \geq 2(32c/n)^{\kappa/(2\kappa-1)}$ and using Lemma 2, we obtain:

$$\int_{u/2}^{1} T_2(\epsilon)d\epsilon \leq \frac{128c}{nu^{1-1/\kappa}} \exp\left(-\frac{nu^{2-1/\kappa}}{128c}\right). \tag{20}$$

Since $u > 32/(3n)$ we have

$$\int_{u/2}^{1} T_3(\epsilon)d\epsilon \leq \frac{64}{3nu^{1-1/\kappa}} \exp\left(-\frac{3nu^{2-1/\kappa}}{64}\right). \tag{21}$$

From (19), (20), (21) and (15) we obtain

$$\mathbb{E}\left[A(\tilde{f}_n) - A_{\mathcal{C}}\right] \leq 2u + 6M\frac{\exp\left(-n\beta_2 u^{(2\kappa-1)/\kappa}\right)}{n\beta_2 u^{1-1/\kappa}}.$$

The definition of u yields $\mathbb{E}\left[A(\tilde{f}_n) - A_{\mathcal{C}}\right] \leq 3\left(\frac{\log M}{n\beta_2}\right)^{\frac{\kappa}{2\kappa-1}}$.

To conclude, for $\beta_0 = \beta_1 \wedge \beta_2 = \beta_1$ we obtain

$$\mathbb{E}\left[A(\tilde{f}_n) - A_{\mathcal{C}}\right] \leq 3 \begin{cases} \left(\frac{\log M}{n\beta_0}\right)^{\frac{\kappa}{2\kappa-1}} & \text{if } A_{\mathcal{C}} - A^* \leq \left(\frac{\log(M)}{n\beta_1}\right)^{\frac{\kappa}{2\kappa-1}} \\ \sqrt{\frac{(A_{\mathcal{C}} - A^*)^{1/\kappa} \log M}{n\beta_0}} & \text{otherwise,} \end{cases}$$

Thus, for $C = 32(6 \vee 537c \vee 16(2c + 1/3)$, the estimator $\tilde{f}_n$ satisfies:

$$\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] \leq \min_{f \in \mathcal{C}}(A(f) - A^*) + C\left(\sqrt{\frac{\min_{f\in\mathcal{C}}(A(f) - A^*)^{\frac{1}{\kappa}} \log M}{n}} + \left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}\right).$$

For the CAEW aggregate we have:

$$\begin{aligned} \mathbb{E}\left[A(\tilde{f}_n^{(CAEW)}) - A^*\right] &\leq \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[A(\tilde{f}_k^{(AEW)}) - A^*\right] \\ &\leq \min_{f\in\mathcal{C}} A(f) - A^* + C\left\{\sqrt{(A_{\mathcal{C}} - A^*)^{\frac{1}{\kappa}} \log M}\left(\frac{1}{n}\sum_{k=1}^{n}\frac{1}{\sqrt{k}}\right) + (\log M)^{\kappa/(2\kappa-1)}\frac{1}{n}\sum_{k=1}^{n}\frac{1}{k^{\frac{\kappa}{2\kappa-1}}}\right\}, \end{aligned}$$

and, by upper bounding the sums by integrals we get the result.

Proof of Theorem 2. The linearity of the hinge loss on $[-1, 1]$ yields

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

Let $a > 0$. For all prediction rules $f_1, \ldots, f_M$, we have

$$\sup_{f_1,\ldots,f_M} \inf_{\hat{f}_n} \sup_{\pi\in\mathcal{P}_\kappa} \left(\mathbb{E}\left[A(\hat{f}_n) - A^*\right] - (1+a)\min_{j=1,\ldots,M}(A(f_j) - A^*)\right) \geq \inf_{\hat{f}_n} \sup_{\substack{\pi\in\mathcal{P}_\kappa \\ f^*\in\{f_1,\ldots,f_M\}}} \mathbb{E}\left[A(\hat{f}_n) - A^*\right].$$

Let N be an integer such that $2^{N-1} \leq M$. Let $x_1, \ldots, x_N$ be N distinct points of $\mathcal{X}$. Let $0 < w < 1/N$. 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 $0 < h < 1$. For all $\sigma = (\sigma_1, \ldots, \sigma_{N-1}) \in \Omega$ we consider

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

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 $\mathbb{P}(Y = 1|X = x) = \eta_\sigma(x), \forall x \in \mathcal{X}$.

Assume that $\kappa > 1$. We have $\mathbb{P}\left(|2\eta_\sigma(X) - 1| \leq t\right) = (N-1)w\mathbb{1}_{h\leq t}$ for any $0 \leq t < 1$. Thus, if we assume that $(N-1)w \leq h^{1/(\kappa-1)}$ then $\mathbb{P}\left(|2\eta_\sigma(X) - 1| \leq t\right) \leq t^{1/(\kappa-1)}$ for all $0 \leq t < 1$. Thus, according to Tsybakov [2004], π_σ belongs to MA(κ).

We denote by ρ the Hamming distance on Ω. Let $\sigma, \sigma' \in \Omega$ such that $\rho(\sigma, \sigma') = 1$. Then, the Hellinger's distance between the measures $\pi_\sigma^{\otimes n}$ and $\pi_{\sigma'}^{\otimes n}$ satisfies

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

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

Let $\hat{f}_n$ be a real-valued statistic. Consider the estimator$\hat{f}_n^*$ with values in $[-1,1]$ defined by $\hat{f}_n^*(x) = \psi(\hat{f}_n(x))$, where ψ is given in (8). For any underlying probability measure π, we have $A(\hat{f}_n) - A^* \geq A(\hat{f}_n^*) - A^*$. Thus to obtain minimax lower bound it is enough to consider only estimators taking values in $[-1,1]$.

Let $\hat{f}_n$ be an estimator with values in $[-1,1]$ and $\sigma \in \Omega$. Using the margin assumption MA(κ), we have, conditionally to the observations D_n and under π_σ:

$$A(\hat{f}_n) - A^* \geq \left(c\mathbb{E}_{\pi_\sigma}\left[|\hat{f}_n(X) - f^*(X)|\right]\right)^\kappa \geq (cw)^\kappa \left(\sum_{j=1}^{N-1} |\hat{f}_n(x_j) - \sigma_j|\right)^\kappa .$$

Taking here the expectations, we find $\mathbb{E}_{\pi_\sigma}\left[A(\hat{f}_n) - A^*\right] \geq (cw)^\kappa \mathbb{E}_{\pi_\sigma}\left[\left(\sum_{i=1}^{N-1} |\hat{f}_n(x_i) - \sigma_i|\right)^\kappa\right]$. Using Jensen's inequality and Lemma 5, we obtain:

$$\inf_{\hat{f}_n} \sup_{\pi \in \mathcal{P}_\kappa, f^* \in \{f_\sigma : \sigma \in \Omega\}} \left(\mathbb{E}_{\pi_\sigma}\left[A(\hat{f}_n) - A^*\right]\right) \geq (cw)^\kappa \left(\frac{N-1}{4}(1-\beta/2)^2\right)^\kappa .$$

Take now $w = (nh^2)^{-1}$, $N = \lceil \log M / \log 2 \rceil$, $h = \left(n^{-1}\lceil \log M/\log 2\rceil\right)^{(\kappa-1)/(2\kappa-1)}$. We can establish that there exists $f_1, \ldots, f_M$ (the 2^{N-1} first ones are sign$(2\eta_\sigma - 1)$ for $\sigma \in \Omega$ and any choice for the $M - 2^{N-1}$ remaining ones) such that for any procedure $\bar{f}_n$, there exists a probability measure π satisfying MA(κ), such that $\mathbb{E}\left[A(\hat{f}_n) - A^*\right] - (1+a)\min_{j=1,\ldots,M}(A(f_j) - A^*) \geq C_0\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}$, where $C_0 = c^\kappa (4e)^{-1} 2^{-2\kappa(\kappa-1)/(2\kappa-1)} (\log 2)^{-\kappa/(2\kappa-1)}$.
Moreover, according to Lemma 4, we have:

$$a \min_{f \in \mathcal{C}}(A(f) - A^*) + \frac{C_0}{2}\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}} \geq C_1 \sqrt{\frac{(\min_{f\in\mathcal{C}} A(f) - A^*)^{\frac{1}{\kappa}} \log M}{n}}.$$

Thus,

$$\mathbb{E}\left[A(\hat{f}_n) - A^*\right] \geq \min_{f\in\mathcal{C}}(A(f) - A^*) + \frac{C_0}{2}\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}} + C_1\sqrt{\frac{(A_\mathcal{C} - A^*)^{\frac{1}{\kappa}} \log M}{n}}.$$

For $\kappa = 1$, we take $h = 1/2$. Then $|2\eta_\sigma(X) - 1| \geq 1/2$ a.s. so $\pi_\sigma \in$MA(1). It suffices then to take $w = 4/n$ and $N = \lceil \log M / \log 2 \rceil$ to obtain the result.

Proof of Corollary 1: The result follows from Theorems 1 and 2. In fact, for any prediction rule f we have $A(f) - A^* = 2(R(f) - R^*)$ and from Zhang's inequality $A(f) - A^* \geq (R(f) - R^*)$ for all $f : \mathcal{X} \longmapsto \mathbb{R}$. Moreover, using Lemma 4, for all $a > 0$, we have $aX + (1/a)^{1/(2\kappa-1)} Y^{\kappa/(2\kappa-1)} \geq \sqrt{X^{1/\kappa} A}$, where $X = A_\mathcal{C} - A^*$ and $Y = \log M/n$.

Proof of Theorem 3: Denote by $\tilde{f}_n$ the ERM aggregate over $\mathcal{F}$. Let $\epsilon > 0$. Denote by $\mathcal{F}_\epsilon$ the set $\{f \in \mathcal{F} : R(f) > R_\mathcal{F} + 2\epsilon\}$ where $R_\mathcal{F} = \min_{f\in\mathcal{F}} R(f)$.

Let $x > 0$. If

$$\sup_{f \in \mathcal{F}_\epsilon} \frac{R(f) - R^* - (R_n(f) - R_n(f^*))}{R(f) - R^* + x} \leq \frac{\epsilon}{R_\mathcal{F} - R^* + 2\epsilon}$$

then, the same argument as in Theorem 1 yields $R_n(f) - R_n(f^*) \geq R_{\mathcal{F}} - R^* + \epsilon$, for any $f \in \mathcal{F}_\epsilon$. So, we have:

$$\begin{aligned}
&\mathbb{P}\left[\inf_{f\in\mathcal{F}_\epsilon} R_n(f) - R_n(f^*) < R_{\mathcal{F}} - R^* + \epsilon\right] \\
&\quad\leq \quad \mathbb{P}\left[\sup_{f\in\mathcal{F}_\epsilon} \frac{R(f) - R^* - (R_n(f) - R_n(f^*))}{R(f) - R^* + x} > \frac{\epsilon}{R_{\mathcal{F}} - R^* + 2\epsilon + x}\right].
\end{aligned}$$

We consider $f' \in \mathcal{F}$ such that $\min_{f\in\mathcal{F}} R(f) = R(f')$. If $R(\tilde{f}_n) > R_{\mathcal{F}} + 2\epsilon$ then $\tilde{f}_n \in \mathcal{F}_\epsilon$, so there exists $g \in \mathcal{F}_\epsilon$ such that $R_n(g) \leq R_n(f')$. Hence, using the same argument as in Theorem 1, we obtain

$$\begin{aligned}
\mathbb{P}\left[R(\tilde{f}_n) > R_{\mathcal{F}} + 2\epsilon\right] \leq \mathbb{P}\left[\sup_{f\in\mathcal{F}} \frac{R(f) - R^* - (R_n(f) - R_n(f^*))}{R(f) - R^* + x} \geq \frac{\epsilon}{R_{\mathcal{F}} - R^* + 2\epsilon + x}\right] \\
+\mathbb{P}\left[R_n(f') - R_n(f^*) > R_{\mathcal{F}} - R^* + \epsilon\right].
\end{aligned}$$

Using the fact that for any f from $\mathcal{X}$ to $\{-1, 1\}$ we have $2(R(f) - R^*) = A(f) - A^*$, Lemma 1 and the same discussion as at the end of the proof of Theorem 1, we get the result.

Proof of Theorem 4. For all prediction rules $f_1, \ldots, f_M$, we have

$$\sup_{f_1,\ldots,f_M} \inf_{\hat{f}_n} \sup_{\pi\in\mathcal{P}_\kappa} \left(\mathbb{E}\left[R(\hat{f}_n) - R^*\right] - (1+a) \min_{j=1,\ldots,M}(R(f_j) - R^*)\right) \geq \inf_{\hat{f}_n} \sup_{\substack{\pi\in\mathcal{P}_\kappa \\ f^*\in\{f_1,\ldots,f_M\}}} \mathbb{E}\left[R(\hat{f}_n) - R^*\right].$$

Consider the set of probability measures $\{\pi_\sigma, \sigma \in \Omega\}$ introduced in the proof of Theorem 2. Assume that $\kappa > 1$. Since for any $\sigma \in \Omega$ and any classifier $\hat{f}_n$, we have, by using MA(κ),

$$\mathbb{E}_{\pi_\sigma}\left[R(\hat{f}_n) - R^*\right] \geq (c_0 w)^\kappa \mathbb{E}_{\pi_\sigma}\left[\left(\sum_{i=1}^{N-1} |\hat{f}_n(x_i) - \sigma_i|\right)^\kappa\right],$$

using Jensen's inequality and Lemma (5) we obtain

$$\inf_{\hat{f}_n} \sup_{\pi\in\mathcal{P}_\kappa, f^*\in\{f_\sigma:\sigma\in\Omega\}} \left(\mathbb{E}_{\pi_\sigma}\left[R(\hat{f}_n) - R^*\right]\right) \geq (c_0 w)^\kappa \left(\frac{N-1}{4}(1-\beta/2)^2\right)^\kappa.$$

By taking $w = (nh^2)^{-1}$, $N = \lceil \log M / \log 2 \rceil$, $h = \left((n)^{-1}\lceil \log M/\log 2\rceil\right)^{\frac{\kappa-1}{2\kappa-1}}$ and $\alpha = (\kappa-1)^{-1}$, there exists $f_1, \ldots, f_M$ (the 2^{N-1} first ones are $\mathrm{sign}(2\eta_\sigma - 1)$ for $\sigma \in \Omega$ and any choice for the $M - 2^{N-1}$ remaining ones) such that for any procedure $\bar{f}_n$, there exists a probability measure π satisfying MA(κ), such that $\mathbb{E}\left[R(\hat{f}_n) - R^*\right] - (1+a)\min_{j=1,\ldots,M}(R(f_j) - R^*) \geq C_0\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}}$, where $C_0 = {c_0}^\kappa (4e)^{-1} 2^{-2\kappa(\kappa-1)/(2\kappa-1)} (\log 2)^{-\kappa/(2\kappa-1)}$.

Moreover, according to Lemma 4, we have:

$$a \min_{f\in\mathcal{C}}(R(f) - R^*) + \frac{C_0}{2}\left(\frac{\log M}{n}\right)^{\frac{\kappa}{2\kappa-1}} \geq C_1 \sqrt{\frac{(\min_{f\in\mathcal{C}} R(f) - R^*)^{\frac{1}{\kappa}} \log M}{n}}.$$

The case $\kappa = 1$ is treated in the same way as in the proof of Theorem 2.

Appendix

Lemma 1. *Let $\mathcal{F} = \{f_1, \dots, f_M\}$ a finite class of functions from $\mathcal{X}$ to $[-1,1]$. We assume that π satisfies MA(κ), for a $\kappa \geq 1$. We have for any positive numbers t, x and any integer n:*

$$\mathbb{P}\left[\max_{f\in\mathcal{F}} \frac{A(f) - A_n(f) - (A(f^*) - A_n(f^*))}{A(f) - A^* + x} > t\right] \leq M\left(\left(1 + \frac{8cx^{1/\kappa}}{n(tx)^2}\right)\exp\left(-\frac{n(tx)^2}{8cx^{1/\kappa}}\right) + \left(1 + \frac{16}{3ntx}\right)\exp\left(-\frac{3ntx}{16}\right)\right),$$

where the constant $c > 0$ is the constant of MAH(κ).

Proof. We use a "peeling device". Let $x > 0$. For all integer j, we consider

$$\mathcal{F}_j = \{f \in \mathcal{F} : jx \leq A(f) - A^* < (j+1)x\}.$$

Define the empirical process

$$Z_x(f) = \frac{A(f) - A_n(f) - (A(f^*) - A_n(f^*))}{A(f) - A^* + x}.$$

Using Bernstein's inequality and Proposition 2 to upper bound the variance term, we have:

$$\begin{aligned}
\mathbb{P}\left[\max_{f\in\mathcal{F}} Z_x(f) > t\right] &\leq \sum_{j=0}^{+\infty} \mathbb{P}\left[\max_{f\in\mathcal{F}_j} Z_x(f) > t\right] \\
&\leq \sum_{j=0}^{+\infty} \mathbb{P}\left[\max_{f\in\mathcal{F}_j} A(f) - A_n(f) - (A(f^*) - A_n(f^*)) > t(j+1)x\right] \\
&\leq M\sum_{j=0}^{+\infty} \exp\left(-\frac{n[t(j+1)x]^2}{4c((j+1)x)^{1/\kappa} + (8/3)t(j+1)x}\right) \\
&\leq M\left(\sum_{j=0}^{+\infty} \exp\left(-\frac{n(tx)^2(j+1)^{2-1/\kappa}}{8cx^{1/\kappa}}\right) + \exp\left(-(j+1)\frac{3ntx}{16}\right)\right) \\
&\leq M\left(\exp\left(-\frac{nt^2x^{2-1/\kappa}}{8c}\right) + \exp\left(-\frac{3ntx}{16}\right)\right) \\
&\quad + M\int_1^{+\infty}\left(\exp\left(-\frac{nt^2x^{2-1/\kappa}}{8c}X^{2-1/\kappa}\right) + \exp\left(-\frac{3ntx}{16}X\right)\right)dX.
\end{aligned}$$

Lemma 2 leads to the result.

Lemma 2. *Let $\alpha \geq 1$ and $a, b > 0$. An integration by part yields*

$$\int_a^{+\infty} \exp\left(-bt^\alpha\right) dt \leq \frac{\exp(-ba^\alpha)}{\alpha b a^{\alpha-1}}$$

Lemma 3. *Let $b_1, \dots, b_M$ be M positive numbers and $a_1, \dots, a_M$ some numbers. We have:*

$$\frac{\sum_{j=1}^M a_j}{\sum_{j=1}^M b_j} \leq \max_{j=1,\dots,M}\left(\frac{a_j}{b_j}\right).$$

Lemma 4. *For all positive v, t and all $\kappa \geq 1$*

$$t + v \geq v^{\frac{2\kappa-1}{2\kappa}} t^{\frac{1}{2\kappa}}.$$

Proof. Since log is concave, we have $\log(ab) = (1/x)\log(a^x) + (1/y)\log(b^y) \leq \log\left(a^x/x + b^y/y\right)$ for all positive numbers a, b and x, y such that $1/x + 1/y = 1$, thus $ab \leq a^x/x + b^y/y$. Lemma 4 follows by applying this relation with $a = t^{1/(2\kappa)}, x = 2\kappa$ and $b = v^{(2\kappa-1)/(2\kappa)}$.

We use the following version of Assouad's lemma to establish the minimax lower bound.

Lemma 5. *Let $\{P_\omega / \omega \in \Omega\}$ a statistical experience on a measurable space $(\mathcal{X}, \mathcal{A})$ indexed by the cube $\Omega = \{0,1\}^m$. Denote by $\mathbb{E}_\omega$ the expectation under P_ω. Assume that:*

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

then

$$\inf_{\hat{w} \in [0,1]^m} \max_{\omega \in \Omega} \mathbb{E}_\omega \left[\sum_{j=1}^{m} |\hat{w}_j - w_j| \right] \geq \frac{m}{4} \left(1 - \frac{\alpha}{2}\right)^2.$$

Proof: Obviously, we can replace $\inf_{\hat{w} \in [0,1]^m}$ by $(1/2)\inf_{\hat{w} \in \{0,1\}^m}$ since for all $w \in \{0,1\}$ and $\hat{w} \in [0,1]$ there exists $\tilde{w} \in \{0,1\}$ (for instance the projection of $\hat{w}$ on $\{0,1\}$) such that $|\hat{w} - w| \geq (1/2)|\tilde{w} - w|$. Then, we use Theorem 2.10 p.103 of Tsybakov [2003].

References

J.-Y. Audibert and A.B. Tsybakov. Fast learning rates for plug-in classifiers under margin condition. Available at http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2005 (Preprint PMA-998), 2005.

P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification and risk bounds. Technical Report 638, Department of Statistics, U.C. Berkeley, 2003. Available at http://stat-www.berkeley.edu/tech-reports/638.pdf.

L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. Available at http://www.proba.jussieu.fr/mathdoc/textes/PMA-862.pdf, 2005.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. Available at http//mahery.math.u-psud.fr/∼blanchard/publi/, 2004.

G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *JMLR*, 4: 861–894, 2003.

S. Boucheron, O. Bousquet, and G.Lugosi. Theory of classification: a survey of some recent advances. 2005.

P. Bühlmann and B. Yu. Analyzing bagging. *Ann. Statist.*, 30(4):927–961, 2002.

F. Bunea, A.B. Tsybakov, and M. Wegkamp. Aggregation for regression learning. 2004.

O. Catoni. "universal" aggregation rules with exact bias bounds. Preprint n.510, LPMA, available at http://www.proba.jussieu.fr/mathdoc/preprints/index.html, 1999.

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

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. URL `citeseer.ist.psu.edu/cortes95supportvector.html`.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, Berlin, Heidelberg, 1996.

Y. Freund and R. Schapire. A decision-theoric generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Ann. Statist.*, 28:337–407, 2000.

R. Herbei and H. Wegkamp. Classification with reject option. 2005.

A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric estimation. *Ann. Statist.*, 28(3):681–712, 2000.

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

G. Lecué. Optimal oracle inequality in classification for an aggregation procedure. 2006.

G. Lugosi and N. Vayatis. On the bayes-risk consistency of regularized boosting methods. *Ann. Statist.*, 32(1):30–55, 2004.

E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27:1808–1829, 1999.

P. Massart. Some applications of concentration inequalities to statistics. *Probability Theory. Annales de la Faculté des Sciences de Toulouse*, (2):245–303, 2000. volume spécial dédié à Michel Talagrand.

P. Massart. Concentration inequalities and model selection. *Lectures notes of Saint Flour*, 2004.

P. Massart and E. Nédélec. Risk bound for statistical learning. Preprint. available at http://www.math.u-psud.fr/∼massart/page5.html, 2003.

A. Nemirovski. Topics in non-parametric statistics. 1738, 2000.

B. Schölkopf and A. Smola. *Learning with kernels*. MIT press, Cambridge University, 2002.

C. Scovel and I. Steinwart. Fast rates for support vector machines using gaussian kernels. Los Alamos National Laboratory Technical Report LA-UR 04-8796, submitted to Annals of Statistics, 2004.

C. Scovel and I. Steinwart. Fast rates for support vector machines. Los Alamos National Laboratory Technical Report LA-UR 05-0451, submitted to COLT, 2005.

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. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.

V.G. Vovk. Aggregating strategies. *COLT*, pages 371–386, 1990.

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

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32(1):56–85, 2004.