Bernoulli **13**(4), 2007, 1000–1022
DOI: 10.3150/07-BEJ6044

Optimal rates of aggregation in classification under low noise assumption

GUILLAUME LECUÉ

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

In the same spirit as Tsybakov, 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.

Keywords: aggregation of classifiers; classification; optimal rates; margin

1. Introduction

Let $(\mathcal{X},\mathcal{A})$ be a measurable space. We consider a random variable (X,Y) on $\mathcal{X}\times\{-1,1\}$ with probability distribution denoted by π. Denote by P^X the marginal of π on $\mathcal{X}$ and by $\eta(x)\stackrel{\text{def}}{=}\mathbb{P}(Y=1|X=x)$ the conditional probability function of $Y=1$, knowing that $X=x$. We have n i.i.d. observations of the couple (X,Y) denoted by $D_n=((X_i,Y_i))_{i=1,\dots,n}$. The aim is to predict the output label Y for any input X in $\mathcal{X}$ from the observations D_n.

We recall some usual notation for the classification framework. A **prediction rule** is a measurable function $f:\mathcal{X}\longmapsto\{-1,1\}$. The **misclassification error** associated with f is

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

It is well known (see, e.g., Devroye *et al.* [14]) that

$$\min_{f:\mathcal{X}\longmapsto\{-1,1\}} R(f)=R(f^*)\stackrel{\text{def}}{=}R^*,$$

where the prediction rule f^*, called the **Bayes rule**, is defined by

$$f^*(x)\stackrel{\text{def}}{=}\operatorname{sign}\big(2\eta(x)-1\big)\qquad\forall x\in\mathcal{X}.$$

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 $\hat{f}_n(\cdot,D_n):\mathcal{X}\longmapsto\{-1,1\}$. A key characteristic of $\hat{f}_n$ is the **generalization error** $\mathbb{E}[R(\hat{f}_n)]$, where

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

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 the **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, $\mathbb{E}[R(\hat{f}_n) - R^*] \le C\psi(n)$.

Given a convergence rate, Theorem 7.2 of Devroye *et al.* [14] shows that no classifier can learn at least as fast as this rate for any arbitrary underlying probability distribution π. To achieve rates of convergence, we need a complexity assumption on the set which the Bayes rule f^* belongs to. For instance, Yang [36,37] provide examples of classifiers learning with a given convergence rate under complexity assumptions. These rates cannot be faster than $n^{-1/2}$ (cf. Devroye *et al.* [14]). Nevertheless, they can be as fast as n^{-1} if we add a control on the behavior of the conditional probability function η at the level 1/2 (the distance $|\eta(\cdot) - 1/2|$ is sometimes called the **margin**). For the problem of discriminant analysis, which is close to our classification problem, Mammen and Tsybakov [25] and Tsybakov [34] have introduced the following assumption.

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

$$\mathbb{E}[|f(X) - f^*(X)|] \le c_0\big(R(f) - R^*\big)^{1/\kappa}, \tag{1}$$

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

According to Tsybakov [34] and Boucheron *et al.* [7], this assumption is equivalent to a control on the margin given by

$$\mathbb{P}[|2\eta(X) - 1| \le t] \le ct^{\alpha} \qquad \forall 0 \le t < 1.$$

Several example of **fast rates**, that is, rates faster than $n^{-1/2}$, can be found in Blanchard *et al.* [5], Steinwart and Scovel [31,32], Massart [26], Massart and Nédélec [28], Massart [27] and Audibert and Tsybakov [1].

The paper is organized as follows. In Section, 2 we introduce definitions and 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 postponed to Section 5.

2. Definitions and procedures

2.1. Loss functions

Convex surrogates ϕ for the classification loss are often used in algorithm (Cortes and Vapnic [13], Freund and Schapire [15], Lugosi and Vayatis [24], Friedman *et al.* [16], Bühlman and Yu [8], Bartlett *et al.* [2,3]). Let us introduce some notation. Take ϕ to be a measurable function from $\mathbb{R}$ to $\mathbb{R}$. The risk associated with the loss function ϕ is called the ϕ**-risk** and is defined by

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

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

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

and we denote by $A^{(\phi)*}$ the infimum over all real-valued functions $\inf_{f:\mathcal{X}\longmapsto\mathbb{R}} A^{(\phi)}(f)$.

Classifiers obtained by minimization of the empirical ϕ-risk, for different convex losses, have been proven to have very good statistical properties (cf. Lugosi and Vayatis [24], Blanchard *et al.* [6], Zhang [39], Steinwart and Scovel [31,32] and Bartlett *et al.* [3]). A wide variety of classification methods in machine learning are based on this idea, in particular, on using the convex loss $\phi(x)\stackrel{\text{def}}{=}\max(1-x,0)$ associated with support vector machines (Cortes and Vapnik [13], Schölkopf and Smola [30]), called the **hinge loss**. The corresponding risk is called the **hinge risk** and is defined by

$$A(f)\stackrel{\text{def}}{=}\mathbb{E}\big[\max\big(1-Yf(X),0\big)\big],$$

for any measurable function $f:\mathcal{X}\longmapsto\mathbb{R}$. The **optimal hinge risk** is defined by

$$A^*\stackrel{\text{def}}{=}\inf_{f:\mathcal{X}\longmapsto\mathbb{R}} A(f). \tag{2}$$

It is easy to check that the Bayes rule f^* attains the infimum in (2) and that

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

for any measurable function f with values in $\mathbb{R}$ (cf. Lin [23] and generalizations in Zhang [39] and Bartlett *et al.* [3]), where we extend the definition of R to the class of real-valued functions by $R(f)=R(\operatorname{sign}(f))$. Thus, minimization of the **excess hinge risk**, $A(f)-A^*$, provides a reasonable alternative for minimization of the 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 $M\ge 2$ different classifiers $\hat f_1,\dots,\hat f_M$ taking values in $\{-1,1\}$. The problem of model selection type aggregation, as studied in Nemirovski [29], Yang [38], Catoni [10,11] and Tsybakov [33], consists of the construction of a new classifier $\tilde f_n$ (called an **aggregate**) which approximately mimics the best classifier among $\hat f_1,\dots,\hat f_M$. In most of these papers the aggregation is based on splitting the sample into two independent subsamples, D_m^1 and D_l^2, of sizes m and l, respectively, where $m+l=n$. The first subsample, D_m^1, is used to construct the classifiers $\hat f_1,\dots,\hat f_M$ and the second subsample, D_l^2, is used to aggregate them, that is to construct a new classifier that mimics, in a certain sense, the behavior of the best among the classifiers $\hat f_j$, $j=1,\dots,M$.

In this paper, we will not consider the sample splitting and will concentrate only on the construction of aggregates (following Juditsky and Nemirovski [18], Tsybakov [33], Birgé [4], Bunea *et al.* [9]). Thus, the first subsample is fixed and, instead of classifiers $\hat{f}_1, \dots, \hat{f}_M$, we have fixed prediction rules $f_1, \dots, f_M$. 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 $\mathcal{F} = \{f_1, \dots, f_M\}$ 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

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

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

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

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

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

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

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

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

$$w^{(n)}(f) = \frac{\exp(-n A_n^{(\phi)}(f))}{\sum_{g \in \mathcal{F}} \exp(-n A_n^{(\phi)}(g))} \qquad \forall f \in \mathcal{F}. \tag{4}$$

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

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

$$w^{(n)}(f) = \frac{1}{n} \sum_{k=1}^{n} \frac{\exp(-k A_k^{(\phi)}(f))}{\sum_{g \in \mathcal{F}} \exp(-k A_k^{(\phi)}(g))} \qquad \forall f \in \mathcal{F}.$$

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

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, that is, if it fits too many to the observations, then the ERM aggregate will be overfitted. But, if other classifiers in $\mathcal{F}$ are good classifiers, then 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 of the opinions of the classifiers in the set $\mathcal{F}$.

The exponential weights, defined in (4), can be found in several situations. First, one can check that the solution of the minimization problem

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

for all $\epsilon>0$ is

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

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 [35]).

2.3. Optimal rates of aggregation

Now, we introduce a concept of optimality for an aggregation procedure and for rates of aggregation, in the same spirit as in Tsybakov [33] (where the regression problem is treated). Our aim is to prove that the aggregates introduced above are optimal in the following sense. We denote by $\mathcal{P}_\kappa$ the set of all probability measures π on $\mathcal{X}\times\{-1,1\}$ satisfying MA(κ).

Definition 1. *Let ϕ be a loss function. The remainder term $\gamma(n,M,\kappa,\mathcal{F},\pi)$ is called an* ***optimal rate of model selection type aggregation (MS-aggregation) for the ϕ-risk*** *if the two following inequalities hold*:

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

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

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

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

Here, C_1 and C_2 are positive constants which may depend on κ. 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 ϕ-risk***. *If $\mathcal{C}$ denotes the convex hull of $\mathcal{F}$ and if* (6) *and* (7) *are satisfied with $\min_{f\in\mathcal{F}}(A^{(\phi)}(f)-A^{(\phi)*})$ replaced by $\min_{f\in\mathcal{C}}(A^{(\phi)}(f)-A^{(\phi)*})$, then we say that $\gamma(n,M,\kappa,\mathcal{F},\pi)$ is an* ***optimal rate of convex aggregation type for the ϕ-risk*** *and $\tilde{f}_n$ is an* ***optimal convex aggregation procedure for the ϕ-risk***.

In Tsybakov [33], the optimal rate of aggregation depends only on M and n. In our case, the residual term may be a function of the underlying probability measure π, of the class $\mathcal{F}$ and of the margin parameter κ. Note that, without any margin assumption, we obtain $\sqrt{(\log M)/n}$ for the residual, which is free from π and $\mathcal{F}$. Under the margin assumption, we obtain a residual term dependent of π and $\mathcal{F}$ and it should be interpreted as a normalizing factor in the ratio

$$\frac{\mathbb{E}[A^{(\phi)}(\bar{f}_n) - A^{(\phi)*}] - \min_{f\in\mathcal{F}}(A^{(\phi)}(f) - A^{(\phi)*})}{\gamma(n, M, \kappa, \mathcal{F}, \pi)}.$$

In that case, our definition does not imply the uniqueness of the residual.

Remark 1. Observe that a linear function achieves its maximum over a convex polygon at one of the vertices of the polygon. The hinge loss is linear on $[-1, 1]$ and $\mathcal{C}$ is a convex set, thus MS-aggregation or convex aggregation of functions with values in $[-1, 1]$ are identical problems when we use the hinge loss. That is, we have

$$\min_{f\in\mathcal{F}} A(f) = \min_{f\in\mathcal{C}} A(f). \tag{8}$$

3. Optimal rates of convex aggregation for the hinge risk

Take M functions $f_1, \ldots, f_M$ with values in $[-1, 1]$. Consider the convex hull $\mathcal{C} = \mathrm{Conv}(f_1, \ldots, f_M)$. We want to mimic the best function in $\mathcal{C}$ using the hinge risk and working under the margin assumption. We first introduce a margin assumption w.r.t. the hinge loss.

(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* $\mathrm{MAH}(\kappa)$ *with parameter* $1 \le \kappa < +\infty$ *if there exists $c > 0$ such that*

$$\mathbb{E}[|f(X) - f^*(X)|] \le c\big(A(f) - A^*\big)^{1/\kappa} \tag{9}$$

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

Proposition 1. *The assumption* $\mathrm{MAH}(\kappa)$ *is equivalent to the margin assumption* $\mathrm{MA}(\kappa)$.

In what follows, we will assume that $\mathrm{MA}(\kappa)$ holds and thus also that $\mathrm{MAH}(\kappa)$ holds.

The AEW aggregate of M functions $f_1, \ldots, f_M$ with values in $[-1, 1]$, introduced in (4) for a general loss, has a simple form for the case of the hinge loss, given by

$$\begin{aligned} \tilde{f}_n &= \sum_{j=1}^{M} w^{(n)}(f_j) f_j, \\ &\text{where } w^{(n)}(f_j) = \frac{\exp(\sum_{i=1}^n Y_i f_j(X_i))}{\sum_{k=1}^M \exp(\sum_{i=1}^n Y_i f_k(X_i))} \quad \forall j = 1, \ldots, M. \end{aligned} \tag{10}$$

In Theorems 1 and 2, we state the optimality of our aggregates in the sense of Definition 1.

Theorem 1 (Oracle inequality). *Let $\kappa \geq 1$. We assume that π satisfies* MA(κ). *We denote by $\mathcal{C}$ the convex hull of a finite set $\mathcal{F}$ of functions $f_1, \dots, f_M$ with values in $[-1, 1]$. 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 inequality*

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

where $C = 32(6 \vee 537c \vee 16(2c + 1/3))$ for the ERM, AERM and AEW aggregates with $\kappa \geq 1$, $c > 0$ 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

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

Theorem 2 (Lower bound). *Let $\kappa \geq 1$ and let M, n be two integers such that $2\log_2 M \leq n$. We assume that the input space $\mathcal{X}$ is infinite. There exists an absolute constant $C > 0$, depending only on κ and c, and a set of prediction rules $\mathcal{F} = \{f_1, \dots, f_M\}$ such that for any real-valued procedure $\bar{f}_n$, there exists a probability measure π satisfying* MA(κ), *for which*

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

where $C = c^\kappa(4\mathrm{e})^{-1}2^{-2\kappa(\kappa-1)/(2\kappa-1)}(\log 2)^{-\kappa/(2\kappa-1)}$ 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

$$\sqrt{\frac{(\min_{f\in\mathcal{C}} A(f) - A^*)^{1/\kappa}\log M}{n}} + \left(\frac{\log M}{n}\right)^{\kappa/(2\kappa-1)} \tag{11}$$

is an optimal rate of convex aggregation of M functions with values in $[-1, 1]$ for the hinge loss. Moreover, for any real-valued function f, we have $\max(1 - y\psi(f(x)), 0) \leq \max(1 - yf(x), 0)$ for all $y \in \{-1, 1\}$ and $x \in \mathcal{X}$, thus

$$A(\psi(f)) - A^* \leq A(f) - A^*, \qquad \text{where } \psi(x) = \max\big(-1, \min(x, 1)\big),\ \forall x \in \mathbb{R}. \tag{12}$$

Thus, by aggregating $\psi(f_1), \dots, \psi(f_M)$, it is easy to check that

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

is an optimal rate of model-selection aggregation of M real-valued functions $f_1, \dots, f_M$ w.r.t. the hinge loss. In both cases, the aggregate with exponential weights, as well as ERM and AERM, attains these optimal rates and the CAEW aggregate attains the optimal rate if $\kappa > 1$. Applications and learning properties of the AEW procedure can be found in Lecué [20,21] (in particular, adaptive SVM classifiers are constructed by aggregating only $(\log n)^2$ SVM estimators). In Theorem 1, the AEW procedure satisfies an exact oracle inequality with an optimal residual term whereas in Lecué [21] and Lecué [20] the oracle inequalities satisfied by the AEW procedure are not exact (there is a multiplying factor greater than 1 in front of the bias term) and in Lecué [21], the residual is not optimal. In Lecué [20], it is proved that for any finite set $\mathcal{F}$ of functions $f_1, \dots, f_M$ with values in $[-1, 1]$ and any $\epsilon > 0$, there exists an absolute constant $C(\epsilon) > 0$ such that, for $\mathcal{C}$ the convex hull of $\mathcal{F}$,

$$\mathbb{E}\big[A\big(\tilde{f}_n^{(\mathrm{AEW})}\big) - A^*\big] \le (1+\epsilon)\min_{f\in\mathcal{C}}\big(A(f) - A^*\big) + C(\epsilon)\left(\frac{\log M}{n}\right)^{\kappa/(2\kappa-1)}. \tag{13}$$

This oracle inequality is good enough for several applications (see the examples in Lecué [20]). Nevertheless, (13) can be easily deduced from Theorem 1 using Lemma 3 and may be inefficient for constructing adaptive estimators with exact constants (because of the factor greater than 1 in front of $\min_{f\in\mathcal{C}}(A(f) - A^*)$). Moreover, oracle inequalities with a factor greater than 1 in front of the oracle $\min_{f\in\mathcal{C}}(A(f) - A^*)$ do not characterize the real behavior of the technique of aggregation which we are using. For instance, for any strictly convex loss ϕ, the ERM procedure satisfies (cf. Chesneau and Lecué [12])

$$\mathbb{E}\big[A^{(\phi)}\big(\tilde{f}_n^{(\mathrm{ERM})}\big) - A^{(\phi)*}\big] \le (1+\epsilon)\min_{f\in\mathcal{F}}\big(A^{(\phi)}(f) - A^{(\phi)*}\big) + C(\epsilon)\frac{\log M}{n}. \tag{14}$$

But, it has been recently proven, in Lecué [22], that the ERM procedure cannot mimic the oracle faster than $\sqrt{(\log M)/n}$, whereas, for strictly convex losses, the CAEW procedure can mimic the oracle at the rate $(\log M)/n$ (cf. Juditsky *et al.* [19]). Thus, for strictly convex losses, it is better to use the aggregation procedure with exponential weights than ERM (or even penalized ERM procedures (cf. Lecué [22])) to mimic the oracle. Non-exact oracle inequalities of the form (14) cannot tell us which procedure is better to use since both ERM and CAEW procedures satisfy this inequality.

It is interesting to note that the rate of aggregation (11) depends on both the class $\mathcal{F}$ and π through the term $\min_{f\in\mathcal{C}} A(f) - A^*$. This is different from the regression problem (cf. Tsybakov [33]), where the optimal aggregation rates depend only on M and n. Three cases can be considered, where $\mathcal{M}(\mathcal{F}, \pi)$ denotes $\min_{f\in\mathcal{C}}(A(f) - A^*)$ and M may depend on n (i.e., for function classes $\mathcal{F}$ depending on n):

1. If $\mathcal{M}(\mathcal{F},\pi) \le a(\frac{\log M}{n})^{\kappa/(2\kappa-1)}$, 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 $(\frac{\log M}{n})^{\kappa/(2\kappa-1)}$, which can be $\log M/n$ in the case $k=1$;
2. If $a(\frac{\log M}{n})^{\kappa/(2\kappa-1)} \le \mathcal{M}(\mathcal{F},\pi) \le b$ for some constants $a,b>0$, then our aggregates mimic the best prediction rule in $\mathcal{C}$ with a rate slower than $(\frac{\log M}{n})^{\kappa/(2\kappa-1)}$, but faster than $((\log M)/n)^{1/2}$;
3. If $\mathcal{M}(\mathcal{F},\pi) \ge a > 0$, 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 $(\frac{\log M}{n})^{\kappa/(2\kappa-1)}$. In the worst cases, which are realized when κ tends to ∞ or $\min_{f\in\mathcal{C}}(A(f)-A^*) \ge a > 0$, where $a>0$ is an absolute constant, the optimal rate of aggregation is the slow rate $\sqrt{\frac{\log M}{n}}$.

4. Optimal rates of MS-aggregation for the excess risk

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

$$\mathbb{E}[R(\tilde{f}_n)-R^*] \le 2(1+a)\min_{j=1,\dots,M}\bigl(R(f_j)-R^*\bigr) + \bigl[C+(C^{2\kappa}/a)^{1/(2\kappa-1)}\bigr]\left(\frac{\log M}{n}\right)^{\kappa/(2\kappa-1)}, \tag{15}$$

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 (15), where we need to multiply the residual by $\log n$.

Moreover, there exists a finite set of prediction rules $\mathcal{F}=\{f_1,\dots,f_M\}$ such that, for any classifier $\bar{f}_n$, there exists a probability measure π on $\mathcal{X}\times\{-1,1\}$ satisfying MA(κ), *such that, for any $n\ge 1, a>0$,*

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

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

Due to Corollary 1,

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

is an almost optimal rate of MS-aggregation for the excess risk and the AEW aggregate achieves this rate. The word "almost" is used here because $\min_{f\in\mathcal{F}}(R(f) - R^*)$ is multiplied by a constant greater than 1. Oracle inequality (15) is not exact since the minimal excess risk over $\mathcal{F}$ is multiplied by the constant $2(1+a) > 1$. This is not the case when using the ERM aggregate, as explained in the following theorem.

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

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

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

Using Lemma 3, we can deduce the results of Herbei and Wegkamp [17] from Theorem 3. Oracle inequalities under $\mathrm{MA}(\kappa)$ have already been stated in Massart [27] (cf. Boucheron *et al.* [7]), but the remainder term obtained is worse than the one obtained in Theorem 3.

According to Definition 1, combining Theorem 3 and the following theorem, the rate

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

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* $2\log_2 M \leq n$ *and* $\kappa \geq 1$. *Assume that* $\mathcal{X}$ *is infinite. There exists an absolute constant* $C > 0$ *and a set of prediction rules* $\mathcal{F} = \{f_1, \dots, f_M\}$ *such that for any procedure* $\bar{f}_n$ *with values in* $\mathbb{R}$, *there exists a probability measure* π *satisfying* $\mathrm{MA}(\kappa)$, *for which*

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

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

5. Proofs

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

Assume that MA(κ) holds. We first explore the case $\kappa > 1$, where MA(κ) implies that there exists a constant $c_1 > 0$ such that $\mathbb{P}(|2\eta(X)-1| \le t) \le c_1 t^{1/(\kappa-1)}$ for any $t > 0$ (cf. Boucheron *et al.* [7]). Let f be a function from $\mathcal{X}$ to $[-1,1]$. We have, for any $t > 0$,

$$\begin{aligned}
A(f)-A^* &= \mathbb{E}[|2\eta(X)-1||f(X)-f^*(X)|] \\
&\ge t\mathbb{E}\big[|f(X)-f^*(X)|\mathbb{1}_{|2\eta(X)-1|\ge t}\big] \\
&\ge t\big(\mathbb{E}[|f(X)-f^*(X)|] - 2\mathbb{P}\big(|2\eta(X)-1| \le t\big)\big) \\
&\ge t\big(\mathbb{E}[|f(X)-f^*(X)|] - 2c_1 t^{1/(\kappa-1)}\big).
\end{aligned}$$

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

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

For the case $\kappa = 1$, MA(1) implies that there exists $h > 0$ such that $|2\eta(X)-1| \ge h$ a.s. Indeed, if for any $N \in \mathbb{N}^*$ (the set of all positive integers), there exists $A_N \in \mathcal{A}$ (the σ-algebra on $\mathcal{X}$) such that $P^X(A_N) > 0$ and $|2\eta(x)-1| \le N^{-1}, \forall x \in A_N$, then, for

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

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

Proof of Theorem 1. We start with a general result which says that if ϕ is a convex loss, then the aggregation procedures with the weights $w^{(n)}(f)$, $f \in \mathcal{F}$, introduced in (4) satisfy

$$A_n^{(\phi)}\big(\tilde{f}_n^{(\mathrm{AEW})}\big) \le A_n^{(\phi)}\big(\tilde{f}_n^{(\mathrm{ERM})}\big) + \frac{\log M}{n} \quad \text{and} \quad A_n^{(\phi)}\big(\tilde{f}_n^{(\mathrm{AERM})}\big) \le A_n^{(\phi)}\big(\tilde{f}_n^{(\mathrm{ERM})}\big). \tag{16}$$

Indeed, take ϕ to be a convex loss. We have $\phi(Y\tilde{f}_n(X)) \le \sum_{f\in\mathcal{F}} w^{(n)}(f)\phi(Yf(X))$, thus

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

Any $f \in \mathcal{F}$ satisfies

$$A_n^{(\phi)}(f) = A_n^{(\phi)}\big(\tilde{f}_n^{(\mathrm{ERM})}\big) + n^{-1}\big(\log\big(w^{(n)}\big(\tilde{f}_n^{(\mathrm{ERM})}\big)\big) - \log\big(w^{(n)}(f)\big)\big),$$

thus, by averaging this equality over the $w^{(n)}(f)$ and using $\sum_{f\in\mathcal{F}} w^{(n)}(f)\log(\frac{w^{(n)}(f)}{M^{-1}}) = K(w|u) \geq 0$, where $K(w|u)$ denotes the Kullback–Leibler divergence between the weights $w = (w^{(n)}(f))_{f\in\mathcal{F}}$ and the uniform weights $u = (1/M)_{f\in\mathcal{F}}$, we obtain the first inequality of (16). Using the convexity of ϕ, we obtain a similar result for the AERM aggregate.

Let $\tilde{f}_n$ be either the ERM, the AERM or the AEW aggregate for the class $\mathcal{F} = \{f_1, \ldots, f_M\}$. In all cases, we have, according to (16),

$$A_n(\tilde{f}_n) \leq \min_{i=1,\ldots,M} A_n(f_i) + \frac{\log M}{n}. \tag{17}$$

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

$$\begin{aligned}
&\mathbb{P}\Big[\inf_{f\in\mathcal{D}}\big(A_n(f) - A_n(f^*)\big) < A_{\mathcal{C}} - A^* + \epsilon\Big] \qquad (18)\\
&\quad \leq \mathbb{P}\Big[\sup_{f\in\mathcal{D}} \frac{A(f) - A^* - (A_n(f) - A_n(f^*))}{A(f) - A^* + x} > \frac{\epsilon}{A_{\mathcal{C}} - A^* + 2\epsilon + x}\Big].
\end{aligned}$$

According to (8), for $f' \in \{f_1, \ldots, f_M\}$ such that $A(f') = \min_{j=1,\ldots,M} A(f_j)$, we have $A_{\mathcal{C}} = \inf_{f\in\mathcal{C}} A(f) = \inf_{f\in\{f_1,\ldots,f_M\}} A(f) = A(f')$. According to (17), we have

$$A_n(\tilde{f}_n) \leq \min_{j=1,\ldots,M} A_n(f_j) + \frac{\log M}{n} \leq A_n(f') + \frac{\log M}{n}.$$

Thus, if we assume that $A(\tilde{f}_n) > A_{\mathcal{C}} + 2\epsilon$, then, by definition, we have $\tilde{f}_n \in \mathcal{D}$ and thus there exists $f \in \mathcal{D}$ such that $A_n(f) - A_n(f^*) \leq A_n(f') - A_n(f^*) + (\log M)/n$. According to (18), we have

$$\begin{aligned}
&\mathbb{P}[A(\tilde{f}_n) > A_{\mathcal{C}} + 2\epsilon]\\
&\quad \leq \mathbb{P}\Big[\inf_{f\in\mathcal{D}} A_n(f) - A_n(f^*) \leq A_n(f') - A_n(f^*) + \frac{\log M}{n}\Big]
\end{aligned}$$

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

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

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

and, according to Lemma 2, 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 5 to obtain

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

Using Proposition 1 and Lemma 4 to upper bound the variance term and applying Bernstein's inequality, we get

$$\begin{aligned}
&\mathbb{P}\biggl[A_n(f') - A_n(f^*) \geq A_{\mathcal{C}} - A^* + \epsilon - \frac{\log M}{n}\biggr] \\
&\quad \leq \exp\biggl(-\frac{n(\epsilon - (\log M)/n)^2}{4c(A_{\mathcal{C}} - A^*)^{1/\kappa} + (8/3)(\epsilon - (\log M)/n)}\biggr)
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}\bigl(A(\tilde{f}_n) > A_{\mathcal{C}} + 2\epsilon\bigr) \\
&\quad \leq \exp\biggl(-\frac{n(\epsilon - \log M/n)^2}{4c(A_{\mathcal{C}} - A^*)^{1/\kappa} + (8/3)(\epsilon - (\log M)/n)}\biggr) \\
&\qquad + M\biggl(1 + \frac{32c(A_{\mathcal{C}} - A^* + 2\epsilon)^{1/\kappa}}{n\epsilon^2}\biggr)\exp\biggl(-\frac{n\epsilon^2}{32c(A_{\mathcal{C}} - A^* + 2\epsilon)^{1/\kappa}}\biggr) \\
&\qquad + M\biggl(1 + \frac{32}{3n\epsilon}\biggr)\exp\biggl(-\frac{3n\epsilon}{32}\biggr).
\end{aligned}$$

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

$$\mathbb{E}[A(\tilde{f}_n) - A_{\mathcal{C}}] \leq 2u + 2\int_{u/2}^{1}\bigl[T_1(\epsilon) + M\bigl(T_2(\epsilon) + T_3(\epsilon)\bigr)\bigr]\,\mathrm{d}\epsilon, \tag{19}$$

where

$$\begin{aligned}
T_1(\epsilon) &= \exp\biggl(-\frac{n(\epsilon - (\log M)/n)^2}{4c((A_{\mathcal{C}} - A^*)/2)^{1/\kappa} + (8/3)(\epsilon - (\log M)/n)}\biggr), \\
T_2(\epsilon) &= \biggl(1 + \frac{64c(A_{\mathcal{C}} - A^* + 2\epsilon)^{1/\kappa}}{2^{1/\kappa}n\epsilon^2}\biggr)\exp\biggl(-\frac{2^{1/\kappa}n\epsilon^2}{64c(A_{\mathcal{C}} - A^* + 2\epsilon)^{1/\kappa}}\biggr)
\end{aligned}$$

and

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

Set $\beta_1 = \min(32^{-1}, (2148c)^{-1}, (64(2c + 1/3))^{-1})$, where the constant $c > 0$ appears in MAH(κ). Consider separately the following cases, (C1) and (C2).

(C1) *The case* $A_{\mathcal{C}} - A^* \geq (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$. Denote by $\mu(M)$ the solution of $\mu = 3M\exp(-\mu)$. We have $(\log M)/2 \leq \mu(M) \leq \log M$. Take u such that $(n\beta_1 u^2)/(A_{\mathcal{C}} - A^*)^{1/\kappa} = \mu(M)$. Using the definitions of case (C1) and $\mu(M)$, we get $u \leq A_{\mathcal{C}} - A^*$.

Moreover, $u \geq 4(\log M)/n$, thus

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

Using Lemma 1 and the inequality $u \leq A_{\mathcal{C}} - A^*$, we obtain

$$\int_{u/2}^{1} T_1(\epsilon)\,\mathrm{d}\epsilon \leq \frac{64(2c+1/3)(A_{\mathcal{C}}-A^*)^{1/\kappa}}{nu} \times \exp\biggl(-\frac{nu^2}{64(2c+1/3)(A_{\mathcal{C}}-A^*)^{1/\kappa}}\biggr). \tag{20}$$

We have $128c(A_{\mathcal{C}} - A^* + u) \leq nu^2$. Thus, using Lemma 1, we get

$$\begin{aligned}\int_{u/2}^{1} T_2(\epsilon)\,\mathrm{d}\epsilon &\leq 2\int_{u/2}^{(A_{\mathcal{C}}-A^*)/2} \exp\biggl(-\frac{n\epsilon^2}{64c(A_{\mathcal{C}}-A^*)^{1/\kappa}}\biggr)\,\mathrm{d}\epsilon \\ &\quad + 2\int_{(A_{\mathcal{C}}-A^*)/2}^{1} \exp\biggl(-\frac{n\epsilon^{2-1/\kappa}}{128c}\biggr)\,\mathrm{d}\epsilon \\ &\leq \frac{2148c(A_{\mathcal{C}}-A^*)^{1/\kappa}}{nu} \exp\biggl(-\frac{nu^2}{2148c(A_{\mathcal{C}}-A^*)^{1/\kappa}}\biggr).\end{aligned} \tag{21}$$

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

$$\begin{aligned}\int_{u/2}^{1} T_3(\epsilon)\,\mathrm{d}\epsilon &\leq \frac{64}{3n}\exp\biggl(-\frac{3nu}{64}\biggr) \\ &\leq \frac{64(A_{\mathcal{C}}-A^*)^{1/\kappa}}{3nu} \exp\biggl(-\frac{3nu^2}{64(A_{\mathcal{C}}-A^*)^{1/\kappa}}\biggr).\end{aligned} \tag{22}$$

From (20), (21), (22) and (19), we obtain

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

The definitions of u leads to $\mathbb{E}[A(\tilde{f}_n) - A_{\mathcal{C}}] \leq 4\sqrt{\frac{(A_{\mathcal{C}}-A^*)^{1/\kappa}\log M}{n\beta_1}}$.

(C2) *The case* $A_{\mathcal{C}} - A^* \leq (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$. We now choose 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 (C2) and $\mu(M)$, we get $u \geq A_{\mathcal{C}} - A^*$. Using Lemma 1 and $u > 4(\log M)/n$, $u \geq$

$2(32c/n)^{\kappa/(2\kappa-1)}$ and $u > 32/(3n)$, respectively, we obtain

$$\begin{aligned} \int_{u/2}^{1} T_1(\epsilon)\,\mathrm{d}\epsilon &\le \frac{32(6c+1)}{3nu^{1-1/\kappa}} \exp\left(-\frac{3nu^{2-1/\kappa}}{32(6c+1)}\right), \\ \int_{u/2}^{1} T_2(\epsilon)\,\mathrm{d}\epsilon &\le \frac{128c}{nu^{1-1/\kappa}} \exp\left(-\frac{nu^{2-1/\kappa}}{128c}\right) \end{aligned} \tag{23}$$

and

$$\int_{u/2}^{1} T_3(\epsilon)\,\mathrm{d}\epsilon \le \frac{64}{3nu^{1-1/\kappa}} \exp\left(-\frac{3nu^{2-1/\kappa}}{64}\right). \tag{24}$$

From (23), (24) and (19), we obtain

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

The definition of u yields $\mathbb{E}[A(\tilde{f}_n) - A_{\mathcal{C}}] \le 4(\frac{\log M}{n\beta_2})^{\kappa/(2\kappa-1)}$.

Finally, we obtain

$$\mathbb{E}[A(\tilde{f}_n) - A_{\mathcal{C}}] \le 4 \begin{cases} \left(\dfrac{\log M}{n\beta_2}\right)^{\kappa/(2\kappa-1)}, & \text{if } A_{\mathcal{C}} - A^* \le \left(\dfrac{\log M}{n\beta_1}\right)^{\kappa/(2\kappa-1)}, \\ \sqrt{\dfrac{(A_{\mathcal{C}} - A^*)^{1/\kappa} \log M}{n\beta_1}}, & \text{otherwise.} \end{cases}$$

For the CAEW aggregate, it suffices to upper bound the sums by integrals in the following inequality to get the result:

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

Proof of Theorem 2. Let a be a positive number, $\mathcal{F}$ be a finite set of M real-valued functions and $f_1, \dots, f_M$ be M prediction rules (which will be carefully chosen in what follows). Using (8), taking $\mathcal{F} = \{f_1, \dots, f_M\}$ and assuming that $f^* \in \{f_1, \dots, f_M\}$, we obtain

$$\inf_{\hat{f}_n}\sup_{\pi\in\mathcal{P}_\kappa}\left(\mathbb{E}[A(\hat{f}_n)-A^*]-(1+a)\min_{f\in\mathrm{Conv}(\mathcal{F})}\big(A(f)-A^*\big)\right)$$
$$\geq \inf_{\hat{f}_n}\sup_{\substack{\pi\in\mathcal{P}\\ \kappa f^*\in\{f_1,\dots,f_M\}}}\mathbb{E}[A(\hat{f}_n)-A^*], \tag{25}$$

where $\mathrm{Conv}(\mathcal{F})$ is the set made of all convex combinations of elements in $\mathcal{F}$. Let N be an integer such that $2^{N-1}\leq M$, $x_1,\dots,x_N$ 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 $j=1,\dots,N-1$, and $P^X(\{x_N\})=1-(N-1)w$. We consider the cube $\Omega=\{-1,1\}^{N-1}$. Let $0<h<1$. For all $\sigma=(\sigma_1,\dots,\sigma_{N-1})\in\Omega$ we consider

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

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

Assume that $\kappa>1$. We have $\mathbb{P}(|2\eta_\sigma(X)-1|\leq t)=(N-1)w\mathbb{1}_{h\leq t}$ for any $0\leq t<1$. Thus, if we assume that $(N-1)w\leq h^{1/(\kappa-1)}$, then $\mathbb{P}(|2\eta_\sigma(X)-1|\leq t)\leq t^{1/(\kappa-1)}$ for all $0\leq t<1$. Thus, according to Tsybakov [34], π_σ belongs to $\mathcal{P}_\kappa$.

We denote by ρ the Hamming distance on Ω. Let $\sigma,\sigma'\in\Omega$ be such that $\rho(\sigma,\sigma')=1$. Denote by H the Hellinger distance. Since $H^2(\pi_\sigma^{\otimes n},\pi_{\sigma'}^{\otimes n})=2(1-(1-H^2(\pi_\sigma,\pi_{\sigma'})/2)^n)$ and

$$H^2(\pi_\sigma,\pi_{\sigma'})=w\sum_{j=1}^{N-1}\Big(\sqrt{\eta_\sigma(x_j)}-\sqrt{\eta_{\sigma'}(x_j)}\Big)^2+\Big(\sqrt{1-\eta_\sigma(x_j)}-\sqrt{1-\eta_{\sigma'}(x_j)}\Big)^2$$
$$=2w\big(1-\sqrt{1-h^2}\big),$$

the Hellinger distance between the measures $\pi_\sigma^{\otimes n}$ and $\pi_{\sigma'}^{\otimes n}$ satisfies

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

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

Let $\sigma\in\Omega$ and $\hat{f}_n$ be an estimator with values in $[-1,1]$ (according to (12), we consider only estimators in $[-1,1]$). Using $\mathrm{MA}(\kappa)$, we have, conditionally on the observations D_n and for $\pi=\pi_\sigma$,

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

Taking here the expectations, we find $\mathbb{E}_{\pi_\sigma}[A(\hat{f}_n)-A^*] \geq (cw)^\kappa \mathbb{E}_{\pi_\sigma}[(\sum_{j=1}^{N-1}|\hat{f}_n(x_j)-\sigma_j|)^\kappa]$. Using Jensen's inequality and Lemma 6, we obtain

$$\inf_{\hat{f}_n}\sup_{\sigma\in\Omega}\big(\mathbb{E}_{\pi_\sigma}[A(\hat{f}_n)-A^*]\big) \geq (cw)^\kappa\left(\frac{N-1}{4\mathrm{e}^2}\right)^\kappa. \tag{26}$$

Now take $w=(nh^2)^{-1}$, $N=\lceil\log M/\log 2\rceil$ and $h=(n^{-1}\lceil\log M/\log 2\rceil)^{(\kappa-1)/(2\kappa-1)}$. Replace w and N in (26) by these values. Thus, from (25), there exist $f_1,\dots,f_M$ (the first 2^{N-1} are $\operatorname{sign}(2\eta_\sigma-1)$ for $\sigma\in\Omega$ and any choice is allowed for the remaining $M-2^{N-1}$) such that, for any procedure $\bar{f}_n$, there exists a probability measure π satisfying MA(κ), such that $\mathbb{E}[A(\hat{f}_n)-A^*]-(1+a)\min_{j=1,\dots,M}(A(f_j)-A^*)\geq C_0(\frac{\log M}{n})^{\kappa/(2\kappa-1)}$, where $C_0=c^\kappa(4\mathrm{e})^{-1}2^{-2\kappa(\kappa-1)/(2\kappa-1)}(\log 2)^{-\kappa/(2\kappa-1)}$.

Moreover, according to Lemma 3, we have

$$\begin{aligned}
&a\min_{f\in\mathcal{C}}\big(A(f)-A^*\big)+\frac{C_0}{2}\left(\frac{\log M}{n}\right)^{\kappa/(2\kappa-1)}\\
&\quad\geq\sqrt{2^{-1}a^{1/\kappa}C_0}\sqrt{\frac{(\min_{f\in\mathcal{C}}A(f)-A^*)^{1/\kappa}\log M}{n}}.
\end{aligned}$$

Thus,

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

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 then suffices 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. Using inequality (3), Lemma 3 and the fact that for any prediction rule f, we have $A(f)-A^*=2(R(f)-R^*)$, for any $a>0$, with $t=a(A_{\mathcal{C}}-A^*)$ and $v=(C^2(\log M)/n)^{\kappa/(2\kappa-1)}a^{-1/(2\kappa-1)}$, we obtain the result. $\square$

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

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

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

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

We complete the proof by using Lemma 5, the fact that for any f from $\mathcal{X}$ to $\{-1, 1\}$, we have $2(R(f) - R^*) = A(f) - A^*$, and the same arguments as those developed at the end of the proof of Theorem 1. $\square$

Proof of Theorem 4. Using the same argument as the one used in the beginning of the proof of Theorem 2, we have, for all prediction rules $f_1, \ldots, f_M$ and $a > 0$,

$$\sup_{g_1,\ldots,g_M} \inf_{\hat{f}_n} \sup_{\pi\in\mathcal{P}_\kappa} \left(\mathbb{E}[R(\hat{f}_n) - R^*] - (1+a) \min_{j=1,\ldots,M}\big(R(g_j) - R^*\big)\right) \geq \inf_{\hat{f}_n} \sup_{\substack{\pi\in\mathcal{P}_\kappa \\ f^*\in\{f_1,\ldots,f_M\}}} \mathbb{E}[R(\hat{f}_n) - R^*].$$

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),

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

using Jensen's inequality and Lemma 6, we obtain

$$\inf_{\hat{f}_n} \sup_{\sigma\in\Omega} \big(\mathbb{E}_{\pi_\sigma}[R(\hat{f}_n) - R^*]\big) \geq (c_0 w)^\kappa \left(\frac{N-1}{4e^2}\right)^\kappa.$$

By taking $w = (nh^2)^{-1}$, $N = \lceil \log M / \log 2 \rceil$ and $h = (n^{-1}\lceil \log M / \log 2\rceil)^{(\kappa-1)/(2\kappa-1)}$, there exist $f_1, \ldots, f_M$ (the first 2^{N-1} are $\mathrm{sign}(2\eta_\sigma - 1)$ for $\sigma \in \Omega$ and any choice is allowed for the remaining $M-2^{N-1}$) such that for any procedure $\bar{f}_n$, there exists a probability measure π satisfying MA(κ), such that $\mathbb{E}[R(\hat{f}_n)-R^*]-(1+a)\min_{j=1,\dots,M}(R(f_j)-R^*)\geq C_0(\frac{\log M}{n})^{\kappa/(2\kappa-1)}$, where $C_0={c_0}^{\kappa}(4\mathrm{e})^{-1}2^{-2\kappa(\kappa-1)/(2\kappa-1)}(\log 2)^{-\kappa/(2\kappa-1)}$. Moreover, according to Lemma 3, we have

$$a\min_{f\in\mathcal{F}}\biggl(R(f)-R^*\biggr)+\frac{C_0}{2}\biggl(\frac{\log M}{n}\biggr)^{\kappa/(2\kappa-1)}$$
$$\geq\sqrt{a^{1/\kappa}C_0/2}\sqrt{\frac{(\min_{f\in\mathcal{F}}R(f)-R^*)^{1/\kappa}\log M}{n}}.$$

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

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

$$\int_a^{+\infty}\exp(-bt^{\alpha})\,\mathrm{d}t\leq\frac{\exp(-ba^{\alpha})}{\alpha ba^{\alpha-1}}.$$

Lemma 2. *Let $b_1,\dots,b_M$ be M positive numbers and $a_1,\dots,a_M$ some numbers. We have*

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

□

Proof.

$$\sum_{j=1}^M b_j\max_{k=1,\dots,M}\biggl(\frac{a_k}{b_k}\biggr)\geq\sum_{j=1}^M b_j\frac{a_j}{b_j}=\sum_{j=1}^M a_j.$$

□

Lemma 3. *Let $v,t>0$ and $\kappa\geq 1$. The concavity of the logarithm yields*

$$t+v\geq t^{1/(2\kappa)}v^{(2\kappa-1)/(2\kappa)}.$$

Lemma 4. *Let f be a function from $\mathcal{X}$ to $[-1,1]$ and π a probability measure on $\mathcal{X}\times\{-1,1\}$ satisfying* MA(κ) *for some $\kappa\geq 1$. Denote by $\mathbb{V}$ the symbol of variance. We have*

$$\mathbb{V}\bigl(Y\bigl(f(X)-f^*(X)\bigr)\bigr)\leq c\bigl(A(f)-A^*\bigr)^{1/\kappa}$$

and

$$\mathbb{V}\bigl(\mathbb{1}_{Yf(X)\leq 0}-\mathbb{1}_{Yf^*(X)\leq 0}\bigr)\leq c\bigl(R(f)-R^*\bigr)^{1/\kappa}.$$

Lemma 5. *Let $\mathcal{F} = \{f_1, \dots, f_M\}$ be a finite set of functions from $\mathcal{X}$ to $[-1, 1]$. Assume that π satisfies* MA(κ) *for some $\kappa \geq 1$. We have, for any positive numbers t, x and any integer n,*

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

where the constant $c > 0$ appears in MAH(κ) *and $Z_x(f) = \frac{A(f)-A_n(f)-(A(f^*)-A_n(f^*))}{A(f)-A^*+x}$.*

Proof. For any integer j, consider the set $\mathcal{F}_j = \{f \in \mathcal{F}: jx \leq A(f) - A^* < (j+1)x\}$. Using Bernstein's inequality, Proposition 1 and Lemma 4 to upper bound the variance term, we obtain

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

Lemma 1 leads to the result.

Lemma 6. *Let $\{P_\omega/\omega \in \Omega\}$ be a set of probability measures 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 and by ρ the Hamming distance on Ω. Assume that*

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

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 $\tilde{w} \in \{0, 1\}$ (e.g., the projection of $\hat{w}$ on to $\{0, 1\}$) such that $|\hat{w} - w| \geq (1/2)|\tilde{w} - w|$. We then use Theorem 2.10 of Tsybakov [33], page 103. □

References

[1] Audibert, J.-Y. and Tsybakov, A.B. (2007). Fast learning rates for plug-in classifiers under margin condition. *Ann. Statist.* **35**. To appear.

[2] Bartlett, P.L., Freund, Y., Lee, W.S. and Schapire, R.E. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist.* **26** 1651–1686. MR1673273

[3] Bartlett, P.L., Jordan, M.I. and McAuliffe, J.D. (2006). Convexity, classification and risk bounds. *J. Amer. Statist. Assoc.* **101** 138–156. MR2268032

[4] Birgé, L. (2006). Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.* **42** 273–325. MR2219712

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

[6] Blanchard, G., Lugosi, G. and Vayatis, N. (2003). On the rate of convergence of regularized boosting classifiers. *J. Mach. Learn. Res.* **4** 861–894. MR2076000

[7] Boucheron, S., Bousquet, O. and Lugosi, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM Probab. Statist.* **9** 323–375. MR2182250

[8] Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Ann. Statist.* **30** 927–961. MR1926165

[9] Bunea, F., Tsybakov, A.B. and Wegkamp, M. (2005). Aggregation for Gaussian regression. *Ann. Statist.* To appear. Available at http://www.stat.fsu.edu/~wegkamp.

[10] Catoni, O. (1999). “Universal” aggregation rules with exact bias bounds. Preprint n. 510, LPMA. Available at http://www.proba.jussieu.fr/mathdoc/preprints/index.html.

[11] Catoni, O. (2001). *Statistical Learning Theory and Stochastic Optimization. Ecole d’Été de Probabilités de Saint-Flour 2001. Lecture Notes in Math.* **1851**. New York: Springer. MR2163920

[12] Chesneau, C. and Lecué, G. (2006). Adapting to unknown smoothness by aggregation of thresholded wavelet estimators. Submitted.

[13] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning* **20** 273–297.

[14] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. New York: Springer. MR1383093

[15] Freund, Y. and Schapire, R. (1997). A decision-theoric generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55** 119–139. MR1473055

[16] Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *Ann. Statist.* **28** 337–407. MR1790002

[17] Herbei, R. and Wegkamp, H. (2006). Classification with reject option. *Canad. J. Statist.* **34** 709–721.

[18] Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric estimation. *Ann. Statist.* **28** 681–712. MR1792783

[19] Juditsky, A., Rigollet, P. and Tsybakov, A.B. (2006). Learning by mirror averaging. Preprint n. 1034, Laboratoire de Probabilités et Modèle aléatoires, Univ. Paris 6 and Paris 7. Available at http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2005.

[20] Lecué, G. (2006). Optimal oracle inequality for aggregation of classifiers under low noise condition. In *Proceedings of the 19th Annual Conference on Learning Theory, COLT 2006* **32** 364–378. MR2280618

[21] Lecué, G. (2007). Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.* To appear. Available at http://hal.ccsd.cnrs.fr/ccsd-00009241/en/.

[22] Lecué, G. (2007). Suboptimality of penalized empirical risk minimization. In *COLT07*. To appear.

[23] Lin, Y. (1999). A note on margin-based loss functions in classification. Technical Report 1029r, Dept. Statistics, Univ. Wisconsin, Madison.

[24] Lugosi, G. and Vayatis, N. (2004). On the Bayes-risk consistency of regularized boosting methods. *Ann. Statist.* **32** 30–55. MR2051000

[25] Mammen, E. and Tsybakov, A.B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618

[26] Massart, P. (2000). Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math.* (*6*) **2** 245–303. MR1813803

[27] Massart, P. (2004). Concentration inequalities and model selection. *Lectures Notes of Saint Flour*.

[28] Massart, P. and Nédélec, E. (2006). Risk bound for statistical learning. *Ann. Statist.* **34** 2326–2366.

[29] Nemirovski, A. (2000). Topics in non-parametric statistics. *Ecole d'Été de Probabilités de Saint-Flour 1998. Lecture Notes in Math*. **1738** 85–277. New York: Springer. MR1775640

[30] Schölkopf, B. and Smola, A. (2002). *Learning with Kernels*. MIT Press.

[31] Steinwart, I. and Scovel, C. (2005). Fast rates for support vector machines. In *Proceedings of the 18th Annual Conference on Learning Theory, COLT 2005*. Berlin: Springer. MR2203268

[32] Steinwart, I. and Scovel, C. (2007). Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.* **35** 575–607.

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

[34] Tsybakov, A.B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002

[35] Vovk, V.G. (1990). Aggregating strategies. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory, COLT90* 371–386. San Mateo, CA: Morgan Kaufmann.

[36] Yang, Y. (1999). Minimax nonparametric classification. I. Rates of convergence. *IEEE Trans. on Inform. Theory* **45** 2271–2284. MR1725115

[37] Yang, Y. (1999). Minimax nonparametric classification. II. Model selection for adaptation. *IEEE Trans. Inform. Theory* **45** 2285–2292. MR1725116

[38] Yang, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.* **28** 75–87. MR1762904

[39] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.* **32** 56–85. MR2051001

Received March 2006 and revised April 2007