Electronic Journal of Statistics
Vol. 7 (2013) 1249–1271
ISSN: 1935-7524
DOI: 10.1214/13-EJS805

Upper bounds and aggregation in bipartite ranking

Sylvain Robbiano

LTCI UMR Institut Telecom/CNRS 5141, Telecom ParisTech, Paris cedex 13, France
e-mail: robbiano@telecom-paristech.fr

Abstract: One main focus of learning theory is to find optimal rates of convergence. In classification, it is possible to obtain optimal fast rates (faster than $n^{-1/2}$) in a minimax sense. Moreover, using an aggregation procedure, the algorithms are adaptive to the parameters of the class of distributions. Here, we investigate this issue in the bipartite ranking framework. We design a ranking rule by aggregating estimators of the regression function. We use exponential weights based on the empirical ranking risk. Under several assumptions on the class of distribution, we show that this procedure is adaptive to the margin parameter and smoothness parameter and achieves the same rates as in the classification framework. Moreover, we state a minimax lower bound that establishes the optimality of the aggregation procedure in a specific case.

AMS 2000 subject classifications: Primary 62F07, 62C20; secondary 62G08.
Keywords and phrases: Ranking, aggregation, minimax rates.

Received September 2012.

Contents

1. Introduction

The design of estimators that achieve optimal rates of convergence is a major topic in statistical learning. It has been investigated in many situations such as regression, density estimation and classification. The rates depend on the properties of the considered class of distributions. Classical conditions are on the distribution of the observations and the regularity of the regression function. In that case, the best rates are slower than $n^{-1/2}$ and the estimators depend on the regularity of the regression function. There exist adaptive estimators to get rid of the knowledge of this parameter (see [21, 17, 28, 3, 19, 4, 27]). In classification, when adding an assumption on the distribution of the regression function, rates faster than $n^{-1/2}$ and even faster than n^{-1} are achieved. The rates were obtained for plug-in classification rules in two papers. In [3], the authors estimate the regression function using the locally polynomial estimator. Moreover, the optimal rates are achieved without knowing the regularity and the margin parameters by aggregating the plug-in rules (see [18]). More recently, the local multi-resolution estimation method (see [24]) combined with the Lepski's method (see [20]), achieves the optimal and adaptive minimax rates. Both approaches firstly estimate the regression function and then threshold the estimated function at level 1/2.

In the last decade, the bipartite ranking problem, a supervised learning task, has received the attention of the statistical learning community (see [15, 26, 10] for instance). Its probabilistic framework is the same as the classification framework but the task is of a really different nature. Indeed, to solve this problem one has to order all the observations and understand the whole feature space. This task is important for many applications such as the anomaly detection in signal processing, information retrieval, design of diagnosis tools in medicine and credit-scoring in finance. The problem can be formulated as a pairwise classification problem (see [8]) where the goal is to minimize a loss based on a pair of observations called the ranking risk. In this paper, the authors show that, under a low noise assumption, the rates of convergence of the excess of ranking risk can be really close to n^{-1}. To this end, they use a procedure based on the minimization of the empirical ranking risk over a class of candidate ranking rules. The main drawback of their setup is that the target function has to belong to the class of ranking rules. In [9], minimax rates faster than $n^{-1/2}$ are achieved over class of distributions controlled by a smoothing parameter and a margin parameter. They used the same estimator of the regression function as in classification but this estimator needs the knowledge of the regularity parameter.

Here, we investigate the performance of the aggregation with exponential weights in the bipartite ranking framework in order to obtain a method that can be adaptive to the parameters. The main result is that this procedure satisfies an oracle inequality. Then we study the impact of this inequality in two settings, one with the mild density assumption over the marginal of the observation and the other with the strong assumption (see [3]). When adding assumptions on the regression function, we obtain a new adaptive upper bound in the case of the mild density assumption. Moreover, when aggregating the plug-in estimators of [9], the procedure is adaptive to the parameters of the class of distributions under the strong density assumption.

The rest of the paper is organized as follows. In section 2, we explain the notations and the bipartite ranking task. We define the ranking risk and a convexification of it using the hinge loss. Several margin assumptions are presented and equivalence links are stated. In section 3, we describe the aggregation estimator using the convexified ranking risk and we show the oracle inequalities satisfied by the procedure of aggregation. In section 4, we present two adaptive minimax upper bounds for the excess ranking risk using the aggregated estimator. Finally, we extend the minimax lower bound obtained in [9] to all dimensions. The proofs are deferred in appendix.

2. Theoretical background

Here, we introduce the main assumptions involved in the formulation of the bipartite ranking problem and recall the important results which are used in the following analysis, giving an idea of the nature of the ranking problem.

2.1. Probabilistic setup and first notations

Here and throughout, (X, Y) denotes a pair of random variables, taking its values in the product space $\mathcal{X} \times \{-1, +1\}$ where $\mathcal{X}$ is typically a subset of an Euclidean space of (very) large dimension $d \geq 1$, $\mathbb{R}^d$ say. The r.v. X is a vector of features for predicting the binary label Y. Let $p = \mathbb{P}\{Y = +1\}$ be the rate of positive instances. The joint distribution of (X, Y) is denoted by P, X's marginal distribution by μ and the posterior probability by $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$, $x \in \mathcal{X}$. For the sake of simplicity and with no loss of generality, we assume that $\mathcal{X}$ coincides with $\mu(dx)$'s support. Additionally, the r.v. $\eta(X)$ is supposed to be absolutely continuous w.r.t. the Lebesgue measure.

The indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$ and the range of any mapping Φ by $\textsc{Im}(\Phi)$. We also denote by $\mathcal{B}(x, r)$ the closed Euclidean ball in $\mathbb{R}^d$ centered in $x \in \mathbb{R}^d$ and of radius $r > 0$. For any multi-index $s = (s_1, \ldots, s_d) \in \mathbb{N}^d$ and any $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we set $|s| = \sum_{i=1}^d s_i, s! = s_1! \ldots s_d!, x^s = x_1^{s_1} \ldots x_d^{s_d}$ and $\|x\| = (x_1^2 + \cdots + x_d^2)^{1/2}$. Let D^s denote the differential operator $D^s = \frac{\partial^{s_1 + \cdots + s_d}}{\partial x_1^{s_1} \ldots \partial x_d^{s_d}}$ and $\lfloor \beta \rfloor$ the largest integer that is strictly less than $\beta \in \mathbb{R}$. For any $x \in \mathbb{R}^d$ and any $\lfloor \beta \rfloor$-times continuously differentiable real-valued function g on $\mathbb{R}^d$, we denote by g_x its Taylor polynomial expansion of degree $\lfloor \beta \rfloor$ at point x,

$$g_x(x') = \sum_{|s| \leq \lfloor \beta \rfloor} \frac{(x - x')^s}{s!} D^s g(x).$$

2.2. Bipartite ranking

The bipartite ranking task consists in learning how to order the observations according to the label Y. Specifically, from a sample $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$

with distribution P, we want to learn a scoring function $s : \mathcal{X} \to \mathbb{R}$ such as the order induced by s is the same as the order induced by η. In this case, the observations with label "+1" should have large values whereas the observations with label "−1" should have small values. The most popular tool to evaluate the accuracy of a scoring function is the ROC curve [13]. It is the plot of the false positive rate against the true positive rate

$$t \mapsto (\mathbb{P}\{s(X) > t \mid Y = -1\},\ \mathbb{P}\{s(X) > t \mid Y = +1\})$$

that corresponds to the performance of all the classifiers one can create by thresholding the scoring function s.

Pairwise classification. However, this is a functional tool and for this reason, it is complex to optimize from a theoretical and a computational perspective. For this reason, several authors have reformulated this problem as a pairwise classification problem (see [15, 1, 8]). In this setup, the goal is, given (X, Y) and (X', Y') two random couples with distribution P, to determine whether $Y > Y'$ or not. In this context, the predictor takes the form of ranking rule, namely a (measurable) function $r : \mathcal{X}^2 \to \{-1, +1\}$ such that $r(x, x') = 1$ when x' is ranked higher than x: the more pertinent a ranking rule r, the smaller the probability that it incorrectly ranks two instances drawn independently at random. Formally, optimal ranking rules are those that minimize the *ranking risk*:

$$L(r) \overset{def}{=} \mathbb{P}\left\{r(X, X') \cdot (Y' - Y) < 0\right\}. \tag{2.1}$$

A ranking rule r is said to be *transitive* iff $\forall (x, x', x'') \in \mathcal{X}^3$: "$r(x, x') = +1$ and $r(x', x'') = +1$" ⇒ "$r(x, x'') = +1$". Observe that, by standard quotient set arguments, one can see that transitive ranking rules are those induced by scoring functions: $r_s(x, x') = 2 \cdot \mathbb{I}\{s(x') \geq s(x)\} - 1$ with $s : \mathcal{X} \to$ measurable. With a slight abuse of notation, we set $L(r_s) = L(s)$ for ranking rules defined through a scoring function s.

Optimality. It is easy to see that an optimal ranking rule is

$$r^*(x, x') = 2 \cdot \mathbb{I}_{\{\eta(x') > \eta(x)\}} - 1 \tag{2.2}$$

defined thanks to the regression function η, see Example 1 in [8] for further details. Additionally, it should be noticed that one may derive a closed analytical form for the *excess of ranking risk* $\mathcal{E}(r) = L(r) - L^*$, with $L^* = L(r^*)$. For clarity, we recall the following result.

Lemma 1 (Ranking risk excess - [8]). *For any ranking rule r, we have:*

$$\mathcal{E}(r) = \mathbb{E}\left[|\eta(X) - \eta(X')|\, \mathbb{I}\{r(X, X')(\eta(X') - \eta(X)) < 0\}\right].$$

The accuracy of a ranking rule is here characterized by the excess of ranking risk $\mathcal{E}(r)$, the challenge from a statistical learning perspective being to build a ranking rule, based on a training sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of i.i.d. copies of the pair (X, Y), with asymptotically small excess of ranking risk for large n.

We highlight the fact that, using a basic conditioning argument, the minimum ranking risk L^* can be expressed as a function of $\eta(X)$'s Gini mean difference (where $p = \mathbb{P}\{Y = +1\}$):

$$L^* = p(1-p) - \frac{1}{2}\mathbb{E}[|\eta(X) - \eta(X')|]. \tag{2.3}$$

In binary classification, it is well-known folklore that the learning problem is all the easier when $\eta(X)$ is bounded away from $1/2$. In bipartite ranking, Eq. (2.3) roughly says that the more the r.v. $\eta(X)$ is spread, the easier is the optimal ranking of $\mathcal{X}$'s elements. Hence, the two problems are very different from this perspective.

A continuum of classification problems. In addition, we emphasize the fact that the optimal ranking rule $r^*(x, x')$ can be seen as a (nested) collection of optimal cost-sensitive classifiers: the binary rule $r^*(x, X) = 2\cdot\mathbb{I}\{\eta(X) > \eta(x)\} - 1$, related to the (regression) level set $G^*_t = \{x' \in \mathcal{X} : \ \eta(x') > t\}$ with $t = \eta(x)$, is optimal when considering the cost-sensitive risk $\mathcal{R}_\omega(C) = 2(1-p)\omega\cdot\mathbb{P}\{C(X) = +1 \mid Y = -1\} + 2p(1-\omega)\cdot\mathbb{P}\{C(X) = -1 \mid Y = -1\}$ with cost $\omega = \eta(x)$, see Proposition 15 in [11] for instance. Hence, while binary classification only aims at recovering the single level set $G^*_{1/2}$, which is made easier when $\eta(X)$ is far from $1/2$ with large probability (see [23] or [28]), the ranking task consists in finding the whole collection $\{G^*_t : \quad t \in \mathrm{I{\scriptsize M}}(\eta(X))\}$. Though of disarming simplicity, this observation describes well the main barrier for extending fast-rate analysis to the ranking setup. Indeed, the random variable $\eta(X)$ cannot be far with arbitrarily high probability from all elements of its range.

Convexification of the ranking risk. From a practical angle, optimizing the ranking risk is a real difficulty because the involved loss is not convex. In the classification framework where convex surrogates are widely used for practical purposes, it has also been used for theoretical issues ([5, 29] and [18] for instance). Here, we propose to convexify the pairwise loss and we use this loss in our aggregation procedure (see 3). Notice that minimization of convexified pairwise loss was studied in [8]. We call any measurable function $f : \mathcal{X} \times \mathcal{X}' \to [-1, 1]$ a decision rule and we set the random variable $Z = (Y - Y')/2$. With this notation, we now present the convexification of the ranking risk that we use in this paper.

Definition 2 (*Hinge ranking risk*)**.** For any decision function f, the hinge ranking risk is defined by

$$A(f) \overset{def}{=} \mathbb{E}\phi\left(-f(X, X')\cdot Z\right), \tag{2.4}$$

where $\phi(x) = \max(0, 1 + x)$.

Notice that a ranking rule is a specific kind of decision rule. The next proposition gives a justification to strategies based on the minimization of the hinge ranking risk in order to obtain accurate ranking rules.

Proposition 3. *The minimizer of the ranking risk r^* is a minimizer of the hinge ranking risk A. We call $A^* = A(r^*)$.*

As for the ranking risk, there exists a close analytical form for the hinge ranking risk. This is the purpose of the next proposition.

Lemma 4 (Hinge ranking risk excess). *For any decision rule $f : \mathcal{X} \times \mathcal{X} \to [-1, 1]$, we have:*

$$A(f) - A^* = \mathbb{E}\left[|\eta(X) - \eta(X')|\,|f(X, X') - f^*(X, X')|\right].$$

The specific use of this surrogate is not fortunate and is due to its linearity. Using this property, we see that, for any ranking rule $r : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$, we have:

$$A(r) - A^* = 2(L(r) - L^*). \tag{2.5}$$

By thresholding a decision function, we can obtain a ranking rule. More precisely, for any decision rule f, we set $r_f(x, x') = 2\mathbb{I}\{f(x, x') \geq 0\} - 1$. We now link the excess of hinge ranking risk of a decision function f with the excess of ranking risk of its associated ranking rule. Using this definition, one can easily show that, for any decision rules $f : \mathcal{X} \times \mathcal{X} \to [-1, 1]$, we have:

$$L(r_f) - L^* \leq A(f) - A^*. \tag{2.6}$$

Thus, the minimization of the excess of hinge ranking risk provides a reasonable alternative to the minimization of the excess of ranking risk.

Plug-in ranking functions. Given the form of the Bayes ranking rule $r^*(X, X')$, it is natural to consider *plug-in* ranking rules, that is to say ranking rules obtained by "plugging-in" a nonparametric estimator $\widehat{\eta}_n(x)$ of the regression function η, based on a data sample $(X_1, Y_1),\ \ldots,\ (X_n, Y_n)$, instead of $\eta(x)$ into Eq. (2.2):

$$\widehat{r}_n(x, x') \overset{def}{=} r_{\widehat{\eta}_n}(x, x'),\quad (x, x') \in \mathcal{X}^2.$$

The performance of predictive rules built via the plug-in principle has been extensively studied in the classification/regression context, under mild assumptions on the behavior of $\eta(X)$ in the vicinity of $1/2$ (see the references in [3] for instance) and on η's smoothness in particular. Similarly in the ranking situation, since one obtains as immediate corollary of Lemma 1 that $\mathcal{E}(\widehat{r}_n)$ is bounded by $\mathbb{E}[|\widehat{\eta}_n(X) - \eta(X)|]$, one should investigate under which conditions nonparametric estimators $\widehat{\eta}_n$ lead to ranking rules with fast rates of convergence of $\mathcal{E}(\widehat{r}_n)$ as the training sample size n increases to infinity.

2.3. Additional assumptions

Optimal ranking rules can be defined as those having the best possible rate of convergence of $\mathcal{E}(\widehat{r}_n)$ towards 0, as $n \to +\infty$. Therefore, the latter naturally depends on (X, Y)'s distribution. Following the footsteps of [3], we embrace the *minimax* point of view, which consists in considering a specific class $\mathcal{P}$ of joint distributions P of (X, Y) and to declare $\widehat{r}_n$ optimal if it achieves the best minimax rate of convergence over this class:

$$\sup_{P \in \mathcal{P}} \mathbb{E}\left[\mathcal{E}(\widehat{r}_n)\right] \sim \min_{r_n} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\mathcal{E}(r_n)\right] \text{ as } n \to \infty,$$

where the infimum is taken over all possible ranking rules r_n depending on $(X_1, Y_1), \ldots, (X_n, Y_n)$. In order to carry out such a study, mainly three types of hypotheses shall be used. Here, smoothness conditions related to the real-valued function $\eta : \mathcal{X} \subset \mathbb{R}^d \to (0, 1)$ together with regularity conditions on the marginal $\mu(dx)$ and assumptions that we shall interpret as "spread" conditions for $\eta(X)$'s distribution are stipulated.

Complexity assumption. In the plug-in approach, the goal is to link closeness of $\widehat{\eta}_n(x)$ to $\eta(x)$ to the rate at which $\mathcal{E}(\widehat{r}_n)$ vanishes. Complexity assumptions for the regression function (CAR) stipulating a certain degree of smoothness for η are thus quite tailored for such a study. Here, focus is on regression functions $\eta(x)$ that belong to the $(\beta, L, \mathbb{R}^d)$-Hölder class of functions, denoted $\Sigma(\beta, L, \mathbb{R}^d)$, with $\beta > 0$ and $0 < L < \infty$. The latter is defined as the set of functions $g : \mathbb{R}^d \to \mathbb{R}$ that are β times continuously differentiable and satisfy, for any x, x' in $\mathbb{R}^d$, the inequality

$$|g(x') - g_x(x')| \leq L\|x - x'\|^{\beta}.$$

Remark 1 (Alternative assumptions)**.** We point out that more general CAR assumptions could be considered (see [14] for instance), involving metric entropies or combinatorial quantities such as the VC dimension, more adapted to the study of the performance of empirical risk minimizers. Owing to space limitations, the analysis is here restricted to the Hölder assumption.

Marginal density assumption. In this paper, we use the same terminology as in [3] to define the assumption over the density of the marginal of X. Let strictly positive constants c_0 and r_0 be fixed. Recall first that a Lebesgue measurable set $A \subset \mathbb{R}^d$ is said to be (c_0, r_0)-regular iff $\forall r \in]0, r_0[$, $\forall x \in A$:

$$\lambda(A \cap \mathcal{B}(x, r)) \geq c_0 \lambda(\mathcal{B}(x, r)),$$

where $\lambda(B)$ denotes the Lebesgue measure of any Lebesgue measurable set $B \subset \mathbb{R}^d$. The following assumption on the marginal distribution μ will be used in the sequel. Fix constants $c_0, r_0 > 0$ and $0 < \mu_{min} < \mu_{max} < \infty$ and suppose that a compact set $C \subset \mathbb{R}^d$ is given.

The *strong density assumption* is said to be satisfied if the marginal distribution $\mu(dx)$ is supported on a compact and (c_0, r_0)-regular set $A \subset C$ and has a density f (w.r.t. the Lebesgue measure) bounded away from zero and infinity on A: $\mu_{min} \leq f(x) \leq \mu_{max}$ if $x \in A$ and $\mu(x) = 0$ otherwise.

The *mild density assumption* is said to be satisfied if the marginal distribution $\mu(dx)$ is supported on a compact and (c_0, r_0)-regular set $A \subset C$ and has a density f (w.r.t. the Lebesgue measure) bounded away from infinity on A: $f(x) \leq \mu_{max}$ for all $x \in A$.

Global low noise assumption. Here, we introduce an additional assumption for the function η. In classification, to obtain rates faster than $n^{1/2}$, one has to assume that the regression function η satisfies a low noise assumption in addition to the classical properties of the space of the distribution. In ranking, such assumption was used in [8] and [9]. We introduce two margin assumptions in the context of bipartite ranking and we make the link with the assumption previously made. Let $\alpha \in [0,1]$. The following conditions describe the behavior of the r.v. $\eta(X)$.

Assumption $\mathbf{MA(}\alpha\mathbf{)}$. The distribution P verifies the margin assumption $\mathbf{MA(}\alpha\mathbf{)}$ with parameter $0 \leq \alpha \leq 1$ if there exists $C < \infty$ such that:

$$\mathbb{E}\left[|\mathbb{I}\{r(X,X') \neq r^*(X,X')\}|\right] \leq C(L(r) - L^*)^{\alpha/(1+\alpha)}, \tag{2.7}$$

for all measurable ranking rules $r : \mathcal{X} \times \mathcal{X} \to \{-1,+1\}$.

Assumption $\mathbf{MAK(}\alpha\mathbf{)}$. The distribution P verifies the margin assumption $\mathbf{MAK(}\alpha\mathbf{)}$ with parameter $0 \leq \alpha \leq 1$ if there exists $C < \infty$ such that:

$$\mathbb{E}\left[|f(X,X') - r^*(X,X')|\right] \leq C(A(f) - A^*)^{\alpha/(1+\alpha)}, \tag{2.8}$$

for all measurable decision functions $f : \mathcal{X} \times \mathcal{X} \to [-1,+1]$.

These conditions are introduced to control the variance of $\mathbb{I}\{r(X,X') \neq (Y - Y')\} - \mathbb{I}\{r^*(X,X') \neq (Y - Y')\}$. In particular, we use this control to state the oracle inequality 8. This type of conditions have been studied in classification in order to obtain fast rate (see [6] for further details).

In the bipartite ranking framework, a condition was introduced in [9].

Assumption $\mathbf{NA(}\alpha\mathbf{)}$. We have: $\forall (t,x) \in [0,1] \times \mathcal{X}$,

$$\mathbb{P}\left\{|\eta(X) - \eta(x)| \leq t\right\} \leq C \cdot t^{\alpha}, \tag{2.9}$$

for some constant $C < \infty$.

Equipped with these notations, we state the link between these assumptions.

Proposition 5. *If $\eta(X)$ fulfills Assumption $\mathbf{NA(}\alpha\mathbf{)}$ for $\alpha \in [0,1]$ then Assumption $\mathbf{MA(}\alpha\mathbf{)}$ and $\mathbf{MAK(}\alpha\mathbf{)}$ hold.*

The theoretical results of this paper are always stated using the condition $\mathbf{NA(}\alpha\mathbf{)}$. This is why, we do not need the inverse statement. Since in classification, such conditions are equivalent, it may be the same in ranking. Condition (2.9) above is void for $\alpha = 0$ and more and more restrictive as α grows. It clearly echoes Tsybakov's noise condition, introduced in [28], which boils down to (2.9) with $1/2$ instead of $\eta(x)$. Whereas Tsybakov's noise condition is related to the behavior of $\eta(X)$ near the level $1/2$, condition (2.9) implies global properties for $\eta(X)$'s distribution, as shown by the following result.

Lemma 6 (Low noise and continuity)**.** *[9] Let $\alpha \in]0,1]$. Suppose that assumption* $\mathbf{NA}(\alpha)$ *is fulfilled, $\eta(X)$'s distribution is then absolutely continuous w.r.t. the Lebesgue measure on $[0,1]$. In addition, in the case where $\alpha = 1$, the related density is bounded by $C/2$.*

It is important to note that, in ranking, Assumption $\mathbf{NA}(\alpha)$ can be fulfilled for $\alpha \leq 1$ solely (see the proof in [9]), whereas, in classification, α in Tsybakov's noise condition can be very large, up to $+\infty$, recovering in the limit Massart's margin condition [21]. Indeed, as may be shown by a careful examination of Lemma 6's proof, bound (2.9) for $\alpha > 1$ implies that $F'(\eta(x)) = 0$, denoting by F the cdf of $\eta(X)$. Therefore, it is obvious that the (probability) density of the r.v. $\eta(X)$ cannot be zero on its whole range $Im(\eta) = \{\eta(x),\ x \in \mathcal{X}\}$.

In the context of binary classification, by combining the CAR assumptions described above and Tsybakov's noise condition, optimal rates of convergence were obtained in [3] and adaptive optimal rates in [19]. In particular, it was shown that, with the additional assumption that $\mu(dx)$ satisfies the *mild density assumption*, the minimax rate of convergence is $n^{-\beta(1+\alpha)/(d+\beta(2+\alpha))}$ and may be thus faster than $n^{-1/2}$ or even very close to n^{-1}, depending on the values taken by the parameters α and β. With the additional assumption that $\mu(dx)$ satisfies the *strong density assumption*, the minimax rate of convergence is $n^{-\beta(1+\alpha)/(2\beta+d)}$ and may be thus faster than $n^{-1/2}$ or even than n^{-1}. We shall now attempt to determine whether similar results hold in the ranking setup.

3. Oracle inequalities for the aggregation procedure

In this section, we describe how to aggregate ranking rules into an accurate decision rule for the hinge ranking risk. We propose a procedure that uses exponentials weights. This kind of procedure is very popular in machine learning and was studied in many contexts such as regression (see [25, 12] and [2]), aggregation of experts (see [7] for instance) and classification (see [18]). We show that the obtained decision rule satisfies an oracle inequality which can be used to achieve minimax upper bounds (see 4). The proof of the theorem is an adaptation to the ranking case of the one in [18].

3.1. Aggregation via exponential weights

The ranking rules $r_1, \ldots, r_M$ are given and the goal of the aggregation method is to mimic the performance of the best of them according to the excess risk and under the low noise assumption. We define the exponential aggregate decision rule as

$$\tilde{f}_n = \sum_{m=1}^{M} w_m^{(n)} r_m \tag{3.1}$$

where the weights w_j^n are

$$w_m^{(n)} = \frac{\exp(\sum_{i \neq j} -Z_{ij} r_m(X_i, X_j))}{\sum_{k=1}^{M} \exp(\sum_{i \neq j} -Z_{ij} r_m(X_i, X_j))}, \forall j = 1, \ldots M.$$

Notice that we call it $\tilde{f}_n$ because this function takes its values in $[-1, 1]$. The functions $r_1, \dots, r_M$ take their values in $\{1; -1\}$, we have,

$$w_m^{(n)} = \frac{\exp(-n(n-1)A_n(r_m))}{\sum_{k=1}^{M} \exp(-n(n-1)A_n(r_k))}, \forall j = 1, \dots M, \tag{3.2}$$

where $A_n(r_m) = \frac{1}{n(n-1)} \sum_{i \neq j} \max(0, 1 - Z_{ij} r_m(X_i, X_j))$ is the empirical hinge ranking risk of the ranking rule r_m. Using the equality (2.5), the weights can be rewritten in terms of the empirical risks of r_m's

$$w_m^{(n)} = \frac{\exp(-2n(n-1)2L_n(r_m))}{\sum_{k=1}^{M} \exp(-2n(n-1)L_n(r_k))}, \forall j = 1, \dots M,$$

We call this procedure aggregation with exponential weights (AEW). The idea behind this procedure is to give more weight to the ranking rules that have the smaller empirical performance in order to mimic the accuracy of the empirical (hinge ranking) risk minimizer (ERM). The next result states that the AEW has similar performance as the ERM estimator up to a $(\log M)/n$ term.

Proposition 7. *Let $M \geq 2$ be an integer, $f_1, \dots, f_M$ be M decision rules on $\mathcal{X} \times \mathcal{X}$. For any $n \in \mathcal{N}^*$, the aggregate $\tilde{f}_n$ estimator defined in 3.1 with weights 3.2 satisfies*

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

The main benefits of the AEW procedure are that it does not need a minimization algorithm and is less sensitive to overfitting because the output decision rule is a mixture of several ranking rules whereas ERM only involves one ranking rule.

3.2. An oracle inequality

We now provide the main tool of this paper, an oracle inequality for the excess of hinge ranking risk. The goal of an oracle inequality is to show that an estimator is nearly as good as the best one of a given collection (see [22] for example in model selection). Here, the goal of this oracle inequality is to state that the procedure AEW 3.1 has asymptotically the same performance as the best one among the convex hull formed by a finite set of decision functions.

Theorem 8 (Oracle inequality)**.** *Let $\alpha \in (0, 1]$. We assume that $NA(\alpha)$ holds. 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 the aggregate estimator introduced in 3.1. Then, for any integers $M \geq 3, n \geq 1$ and any $a > 0$, $\tilde{f}_n$ satisfies the inequality*

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

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

In [18], the author shows that the rate $\left(\frac{\log M}{n}\right)^{\frac{\alpha+1}{\alpha+2}}$ is optimal in a minimax sense. For the moment, we do not have such result of optimality, however, the rate in the oracle inequality is the same. Using this tool, we can state an oracle inequality for the excess of ranking risk.

Corollary 9 (Oracle inequality for the ranking risk)**.** *Let $\alpha \in (0,1]$, $M \geq 3$ and $\{r_1, \dots, r_M\}$ be a finite set of prediction rules. We assume that $NA(\alpha)$ holds. Let $\tilde{f}_n$ be the aggregate estimator introduced in 3.1. Then, for any integers $M \geq 3, n \geq 1$ and any $a > 0$, $\tilde{f}_n$ satisfies the inequality*

$$\mathbb{E}[L(r_{\tilde{f}_n}) - L^*] \leq 2(1+a) \min_{f \in \mathcal{C}} (L(r_f) - L^*) + C \left(\frac{\log M}{n}\right)^{\frac{\alpha+1}{\alpha+2}},$$

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

Proof. Using inequalities 2.5 and 2.6 combine with Theorem 8, we immediately get the desired result. □

This oracle is the main tool to obtain the minimax rates in Theorem 11, 12 and 15 using an estimator based on the AEW procedure.

4. Minimax rates

Here, we present the adaptive minimax upper bounds in bipartite ranking in two cases, specifically under the mild assumption and the strong assumption. The estimators of the regression function used are the same as in classification (see [18] and [3]).

4.1. The "mild" case

In this section, we assume that the regression function η belongs to a Hölder class of functions. An important result from [16], on the complexity of Hölder classes, says that:

$$\mathcal{N}\left(\Sigma(\beta, L, [0,1]^d), \epsilon, L^\infty([0,1]^d)\right) \leq C\epsilon^{-\frac{d}{\beta}}, \forall \epsilon > 0$$

where the left hand side is the ϵ-entropy of the $(\beta, L, [0,1]^d)$-Hölder class w.r.t. to the $L^\infty([0,1]^d)$ norm and C is a constant depending only on β and d. We now introduce the first class of distributions for the random couple (X, Y).

Definition 10. Let $\alpha \leq 1$, β and L be strictly positive constants. The collection of probability distributions $P(dx, dy)$ such that

1. the marginal $\mu(dx) = \int_y P(dx, dy)$ satisfies the mild density assumption with μ_{max},
2. the global noise assumption $\mathbf{NA}(\alpha)$ holds,
3. the regression function belongs to Hölder class $\Sigma(\beta, L, \mathbb{R}^d)$,

is denoted by $\mathcal{P}_{\alpha,\beta,\mu_{max}}$ (omitting to index it by the constants involved in the noise assumption for notational simplicity).

Let $\alpha \leq 1, \beta > 0$. For $\epsilon > 0$, $\Lambda_\epsilon(\beta)$ is an ϵ-net on $\Sigma(\beta, L, [0;1]^d)$ for the L^∞-norm, such that $\ln(\text{Card}(\Lambda_\epsilon(\beta))) \leq C\epsilon^{-d/\beta}$. We use the procedure 3.1 over the net $\Lambda_\epsilon(\beta)$ to define the estimator:

$$\tilde{f}_n^{\epsilon,\beta} = \sum_{g \in \Lambda_\epsilon(\beta)} w^n(r_g) r_g \tag{4.1}$$

where $r_g(x, x') = 2\mathbb{I}\{g(x) > g(x')\} - 1$ and we call the associated ranking rule $\tilde{r}_n^{\epsilon,\beta}$. This is an adaptation to the ranking case of the estimator in [18]. We now state the minimax upper bound for the excess of ranking risk over the class of distributions that satisfy the mild assumption.

Theorem 11 (Upper bound: mild case). *There exists a constant $C > 0$ such that for all $n \geq 1$, the maximum expected excess of ranking risk of the aggregation rule defines in 4.1 $\epsilon_n = n^{-\alpha\beta/(d+\beta(2+\alpha))}$, is bounded as follows:*

$$\sup_{P \in \mathcal{P}_{\alpha,\beta,\mu_{max}}} \mathcal{E}(\tilde{r}_n^{\epsilon,\beta}) \leq C \cdot n^{-\frac{\beta(1+\alpha)}{d+\beta(2+\alpha)}}. \tag{4.2}$$

where C depends on d, β and α.

To obtain an estimator adaptive to the smoothness and the margin coefficients, we aggregate classifiers $\tilde{r}_n^{(\epsilon,\beta)}$ for (ϵ, β) in a finite grid. We split the sample in two sets, the first set $D_m^{(1)}$ of size $m = n - \lfloor n/\ln n \rfloor$ is used to build the plug-in classifiers and the second one $D_l^{(2)}$ of size $l = \lfloor n/\ln n \rfloor$ to obtain the weights. We define the grid $\mathcal{G}$ of values for (ϵ, β):

$$\mathcal{G} = \left\{ (\epsilon_k, \beta_p) = \left(m^{-\phi_k}, \frac{p}{\ln n} \right) | \phi_k = \frac{k}{\ln n}, k \in \{1, \dots, \lfloor \ln(n)/2 \rfloor\}, \right.$$
$$\left. p \in \left\{ 1, \dots, \lfloor \ln(n) \rfloor^2 \right\} \right\}.$$

We propose the ranking rule $\tilde{r}_n^{adp}$ which is the sign of the decision function

$$\tilde{f}_n^{adp} = \sum_{(\epsilon,\beta) \in \mathcal{G}} w^{(l)}(r_m^{\epsilon,\beta}) r_m^{\epsilon,\beta}$$

where the weights $w^{(l)}(r)$ are those defined in 3.2 using the dataset $D_l^{(2)}$ and $r_m^{\epsilon,\beta}$ is the ranking rule associated to the decision function introduced in equation 4.1 using the dataset $D_m^{(1)}$.

Theorem 12 (Adaptivity in α and β). *Let K be a compact subset of $]0;1[\times]0;\infty[$. There exists a constant $C > 0$ such that for all $n \geq 1$, for any $(\alpha, \beta) \in K$, the maximum expected excess of ranking risk of the estimator $\tilde{r}_n^{adp}$ is bounded as follows:*

$$\sup_{P \in \mathcal{P}_{\alpha,\beta,\mu_{max}}} \mathcal{E}(\tilde{r}_n^{adp}) \leq C \cdot n^{-\frac{\beta(1+\alpha)}{d+(2+\alpha)\beta}}. \tag{4.3}$$

The cardinality of Σ_{ϵ_n} is an exponential of n so the estimators $\tilde{r}^{\epsilon_n,\beta}$, for a given (ϵ_n, β), are not easily implementable. However the procedure is very interesting from a theoretical standpoint since it is adaptive to the parameters and it achieves fast rates when $\alpha\beta > d$. Finally, notice that this estimator can achieve fast rates when $\alpha\beta > d$ i.e. when the regression function is very smooth.

4.2. The "strong" case

Now, we introduce the second case, namely the strong density assumption. The class of distributions is given in the next definition.

Definition 13. Let $\alpha \leq 1$, β and L be strictly positive constants. The collection of distributions probabilities $P(dx, dy)$ such that

1. the marginal $\mu(dx) = \int_y P(dx, dy)$ satisfies the strong density assumption with μ_{max} and with μ_{min},
2. the global noise assumption $\mathbf{NA(}\alpha\mathbf{)}$ holds,
3. the regression function belongs to Hölder class $\Sigma(\beta, L, \mathbb{R}^d)$,

is denoted by $\mathcal{P}_{\alpha,\beta,\mu_{max},\mu_{min}}$ (omitting to index it by the constant involved in the noise assumption for notational simplicity).

We recall the non-adaptive upper bound for the excess of ranking risk.

Theorem 14 ([9])**.** *There exists a constant $C > 0$ such that for all $n \geq 1$, the maximum expected excess of ranking risk of the plug-in rule $\widehat{r}_n^{(\beta)}(x, x') = 2 \cdot \mathbb{I}\{\widehat{\eta}_{n,h_n}(x') > \widehat{\eta}_{n,h_n}(x)\} - 1$, with $h_n = n^{-1/(2\beta+d)}$ and $l = \lfloor \beta \rfloor$, is bounded as follows:*

$$\sup_{P \in \mathcal{P}_{\alpha,\beta,\mu_{max},\mu_{min}}} \mathcal{E}(\widehat{r}_n^{\beta}) \leq C \cdot n^{-\frac{\beta(1+\alpha)}{d+2\beta}}. \tag{4.4}$$

The plug-in estimator defined in the last theorem depends only on β. To obtain an estimator adaptive to the smoothness coefficient, we aggregate classifiers $\widehat{r}_n^{(\beta)}$ for β in a finite grid. We split the sample in two sets, the first set of size $m = n - \lfloor n/\ln n \rfloor$ is used to build the plug-in classifiers and the second one of size $l = \lfloor n/\ln n \rfloor$to obtain the weights. We define the set $\mathcal{F}$ of of plug-in classifiers using the training sample $D_m^1 = (X_i, Y_i)_{1\leq i\leq m}$:

$$\mathcal{F} = \left\{ \widehat{r}_n^{(\beta_k)} | \beta_k = \frac{kd}{\ln(n) - 2k}, k \in \{1, \dots, \lfloor \ln(n)/2 \rfloor\} \right\}.$$

Using the validation sample $D_l^2 = (X_i, Y_i)_{m+1\leq i\leq n}$, we build the weights, for all $r \in \mathcal{F}$

$$w_n^{(l)}(r) = \frac{\exp(\sum_{i=m+1}^{n} Y_i r(X_i))}{\sum_{\bar{r}\in\mathcal{F}} \exp(\sum_{i=m+1}^{n} Y_i \bar{r}(X_i))}$$

Finally, our ranking rule is $\widehat{r}^{adp} = \text{sign}(\widehat{\text{f}}^{\text{adp}})$, where $\widehat{f}^{adp} = \sum_{r\in\mathcal{F}} w_n^{(l)}(r) r$.

Theorem 15 (Adaptivity in β)**.** *Let K be a compact subset of $]0;1[\times]0;\infty[$. There exists a constant $C>0$ such that for all $n\geq 1$, for any $(\alpha,\beta)\in K$ such that $\alpha\beta\leq d$, the maximum expected excess of ranking risk of the estimator $\tilde{r}^{adp}$ is bounded as follows:*

$$\sup_{P\in\mathcal{P}_{\alpha,\beta,\mu_{max},\mu_{min}}} \mathcal{E}(\widehat{r}^{adp}) \leq C\cdot n^{-\frac{\beta(1+\alpha)}{d+2\beta}}. \tag{4.5}$$

5. Lower bounds

For completeness, we now state a lower bound for the minimax rate of the expected excess of ranking risk in the strong density case. It holds in a specific situation, namely when $\alpha\beta\leq 1$. When $d=1$, the result can be found in [9].

Theorem 16 (A minimax lower bound)**.** *Let $(\alpha,\beta)\in]0,1]\times\mathbb{R}_+^*$ such that $\alpha\beta\leq 1$. There exists a constant $C>0$ such that, for any ranking rule r_n based on n independent copies of the pair (X,Y), we have: $\forall n\geq 1$,*

$$\sup_{P\in\mathcal{P}_{\alpha,\beta,\mu_{max},\mu_{min}}} \mathcal{E}(r_n) \geq C\cdot n^{-\frac{\beta(1+\alpha)}{d+2\beta}}.$$

When $d\geq 2$ the rate of convergence is always slower than $n^{-1/2}$. That means that we are not able to prove optimal fast rates for the excess ranking risk. In classification, the limitation is $\alpha\beta\leq d$, so optimal fast rates can be achieved in this situation (but not hyper fast).

For the mild case and the oracle inequality, mimicking the proof of theorem 16 does not give the same rates as the upper bounds. An explanation of the difficulties as well as the rates are given in Appendix B.

6. Conclusion

In this paper, we investigate the aggregation with exponential weights of ranking rules. In order to aggregate, we convexify the ranking loss using the hinge loss. We state an oracle inequality for the aggregation procedure under a low noise assumption that achieves the same rate as in classification. This is the crucial point to obtain the adaptive upper bounds for the excess of ranking risk. In the mild density case, we establish a new upper bound that is adaptive to the margin and the regularity parameters, with the same rates as in classification. In the strong density case, we aggregate plug-in classifiers in order to obtain minimax adaptive rates of convergence, under a restrictive assumption over the parameters for all dimensions. Moreover, in dimension 1, the aggregation procedure attains minimax adaptive fast rates. These results are in the continuity of [9] and there are still a lot of issues, in particular to obtain the lower bounds, that require a better understanding of the nature of the bipartite ranking problem.

Appendix A: Proofs

Proof of Proposition 3

Proof.

$$\begin{aligned} A(f) &= \mathbb{E}[1-(f(X,X')\cdot Z)] \\ &= 1-\mathbb{E}[f(X,X')(\eta(X)(1-\eta(X'))) - f(X,X')(\eta(X')(1-\eta(X)))] \\ &= 1-\mathbb{E}[f(X,X')(\eta(X)-\eta(X'))] \end{aligned}$$

Finally to minimize A, we have to set $f^*(x,x')=1$ when $\eta(x)\geq\eta(x')$ and $f^*(x,x')=-1$ otherwise. □

Proof of Lemma 4

Proof. Because f values are in $[-1,1]$

$$\begin{aligned} &A(f)-A^* = \mathbb{E}\left[-f(X,X')\cdot Z + f^*(X,X')\cdot Z\right] \\ &= \mathbb{E}\left[-f(X,X')\eta(X)+f(X,X')\eta(X')+f^*(X,X')\eta(X)-f^*(X,X')\eta(X')\right] \\ &\qquad\qquad\qquad\qquad = \mathbb{E}\left[(f(X,X')-f^*(X,X'))(\eta(X')-\eta(X))\right] \end{aligned}$$

By definition of $f^*(X,X')$, we get the desired result. □

Proof of Proposition 5

Proof. Recall that

$$(L(r)-L^*) = \mathbb{E}\left[|\eta(X)-\eta(X')|\,\mathbb{I}\{r(X,X')(\eta(X')-\eta(X))<0\}\right]$$

Lower bounding $\eta(X)-\eta(X')$ by t we obtain the lower bound

$$t\mathbb{E}\mathbb{I}\{r(X,X')(\eta(X')-\eta(X))<0\}\mathbb{I}\{\eta(X')-\eta(X)>t\}$$

which is greater (using the noise assumption) than

$$t\mathbb{E}[\mathbb{I}\{r(X,X')\neq r^*(X,X')\}] - Ct^{1+\alpha}$$

Optimizing in the parameter t, we obtain (for $t_0 = \big(\frac{\mathbb{E}\mathbb{I}\{r(X,X')\neq r^*(X,X')\}}{C(1+\alpha)}\big)^{1/\alpha}$):

$$\mathbb{E}[\mathbb{I}\{r(X,X')\neq r^*(X,X')\}] \leq \frac{C(1+\alpha)}{C\alpha^{\alpha/(1+\alpha)}}(L(r)-L^*)^{\alpha/(1+\alpha)}$$

□

Proof of Proposition 7

Proof. Using the convexity of the hinge loss, we have $A_n(\tilde{f}_n)\leq\sum_{M}^{j=1}\omega_j A_n(f_j)$. Let $j_0 = \arg\min_{j=1,\ldots,M} A_n(f_j)$, we have $A_n(f_j) = A_n(f_{j_0}) + \frac{1}{n}(\log(\omega_{j_0}) -$

$\log(\omega_j))$ for all $j = 1, \ldots, M$ and by averaging over the ω_j, we obtain:

$$A_n(\tilde{f}_n) \leq \min_{j=1,\ldots,M} A_n(f_j) + \frac{1}{n}\sum_{j=1}^{M} \omega_j(\log(\omega_{j_0}) - \log(\omega_j)),$$

Using that $\sum_{M}^{j=1} \omega_j \frac{\log \omega_j}{1/M} = K(w|u) \geq 0$ where $K(w|u)$ denotes the Kullback-Leiber divergence between the weights $\omega = (\omega_j)_{j=1,\ldots,M}$ and the uniform weights $u = (1/M)_{j=1,\ldots,M}$ and $w_{j_0} \leq 1$, we obtain the desired result. □

Proof of Theorem 8

Proof. Let $a > 0$. Adding and subtracting $(1+a)(A_n(\tilde{f}_n) - A_n(f^*)$ to $A(\tilde{f}_n) - A^*$ and then using proposition 7, we have for any $f \in \mathcal{F}$:

$$A(\tilde{f}_n) - A^* \leq (1+a)(A_n(f) - A_n(f^* + \frac{\log M}{n}) + A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*).$$

Taking the expectation, we upper bound $\mathbb{E}[A(\tilde{f}_n) - A^*]$ by

$$(1+a)\min_{f \in \mathcal{F}}(A(f) - A(f^*)) + \frac{\log M}{n} + \mathbb{E}[A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*))].$$

Now the goal is to control the expectation in the RHS. For that we use the Bernstein's inequality. First, notice that, using the linearity of the hinge loss on $[-1, 1]$ we have:

$$A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*)) \leq \max_{f \in \mathcal{F}} A(f) - A^* - (1+a)(A_n(f) - A_n(f^*)),$$

using the union bound we deduce that, for all $\delta \in]0, 4+2a[$, the probability $\mathbb{P}\{A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*))\} \geq \delta\}$ is bounded by the sum $\sum_{f \in \mathcal{F}} \mathbb{P}\{A(f) - A^* - (1+a)(A_n(f) - A_n(f^*)) \geq \delta\}$. The MA($\alpha$) assumption implies that the variance of $\mathbb{I}\{Z \neq f(X, X')\} - \mathbb{I}\{Z \neq f^*(X, X')\}$ is bounded by $(A(f) - A^*)^{\frac{\alpha}{1+\alpha}}$. Now, using the Bernstein's inequality on $\mathbb{P}\{A(f) - A^* - (A_n(f) - A_n(f^*)) \geq \frac{\delta + a(A(f) - A^*)}{1+a}\delta\}$, $\mathbb{P}\{A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*)) \geq \delta\}$ is bounded for all by $\delta \in]0, 4+2a[$

$$\sum_{f \in \mathcal{F}} \exp\left(-\frac{n(\delta + a(A(f) - A^*))^2}{2(1+a)^2(A(f) - A^*)^{\frac{\alpha}{1+\alpha}} + 2(1+a)(\delta + a(A(f) - A^*))/3}\right).$$

The quantity inside the exponential is lower for all $\delta \in]0, 4+2a[$ and $f \in \mathcal{F}$ than $-c\delta^{2-\frac{\alpha}{1+\alpha}}$ where c depends only on a. Using the fact that $\int_u^{+\infty} \exp(-bt^\kappa)dt \leq \frac{\exp(-bu^\kappa)}{\kappa bu^{\kappa-1}}$ and the inequality obtained, we get

$$\mathbb{E}[A(\tilde{f}_n) - A^* - (1+a)(A_n(\tilde{f}_n) - A_n(f^*))] \leq 2t + M\frac{\exp(-nct^{\frac{2+\alpha}{2+2\alpha}})}{ncbt^{\frac{1}{1+\alpha}}}$$

Optimizing in t the RHS, we obtain the desired result. □

Proof of Theorem 11

Proof. Using Corollary 9 with $a = 1$, we get, for any $\epsilon > 0$:

$$\mathcal{E}(r_{\tilde{f}^{\epsilon,n}}) \leq 4 \min_{g \in \Lambda_\epsilon(\beta)} (L(r_g - L^*) + C \left(\frac{\log \Lambda_\epsilon(\beta)}{n} \right)^{\frac{\alpha+1}{\alpha+2}}.$$

Using that $L(r_g) - L^* \leq C \|g - \eta\|_{L^\infty}^{1+\alpha}$ (see [9]) we obtain that

$$\mathcal{E}(r_{\tilde{f}^{\epsilon,n}}) \leq D \left(\epsilon^{1+\alpha} + \left(\frac{\epsilon^{-d/\beta}}{n} \right)^{\frac{\alpha+1}{\alpha+2}} \right).$$

Taking $\epsilon_n = n^{-\alpha\beta/(d+\beta(2+\alpha))}$, we obtain the result. □

Proof of Theorem 12

Proof. We introduce the function $\phi :]0;1[\times]0;\infty[\to]0;1/2[, (\alpha, \beta) \mapsto \frac{\beta}{d+\beta(2+\alpha)}$. There exists n_1 depending on K such that for any n greater than n_1 we have, for all $(\alpha, \beta) \in K$

$$\ln(n)^{-1} \leq \phi(\alpha, \beta) \leq \lfloor \ln(n)/2 \rfloor \ln(n)^{-1}.$$

Let $(\alpha_0, \beta_0) \in K$. For $n \geq n_1$, we denote $a_0 \in \{1, \dots, \lfloor \ln(n)/2 \rfloor\}$ the integer such that $\phi_{a_0} = a_0 \ln(n)^{-1} \leq \phi(\alpha_0, \beta_0) \leq (a_0 + 1) \ln(n)^{-1}$ and $q_0 \in \{1, \dots, \lfloor \ln(n) \rfloor^2 - 1\}$ such that $\beta_{q_0} = q_0 \ln(n)^{-1} \leq \beta_0 \leq (q_0 + 1) \ln(n)^{-1}$. Denote by $g_{\beta_{q_0}}(.)$ the decreasing function $\phi(., \beta_{q_0})$ from $[0, 1]$ to $[0, 1/2]$ and we set $\alpha_{0,n} = g_{\beta_{q_0}}^{-1}(\phi_{a_0})$. There exists A such that $A|\alpha_{0,n} - \alpha_0| \leq |g_{\beta_{q_0}}(\alpha_{0,n}) - g_{\beta_{q_0}}(\alpha_0)| \leq \ln(n)^{-1}$. Let P be a probability distribution belonging to $\mathcal{P}_{\alpha_0,\beta_0,\mu_{max}}$. Applying the Corollary 9 with $a = 1$, we get

$$\mathbb{E}\left[\mathcal{E}(r_{\tilde{f}^{adp}}) | D_m^1\right] \leq 4 \min_{(\epsilon,\beta) \in \mathcal{G}} (L(r) - L^*) + C \left(\frac{\ln \mathrm{Card}(\mathcal{G})}{l} \right)^{\frac{\alpha+1}{\alpha+2}}$$

Using that $l = n/\ln(n)$ and that $\mathrm{Card}(\mathcal{G}) \leq \ln(\mathrm{n})^3$ combined with the definition of a_0 and $\epsilon_m^0 = m^{-\frac{a_0}{\ln n}}$, we have

$$\mathbb{E}_P\left[\mathcal{E}(\tilde{r}^{adp})\right] \leq C \left(\mathbb{E}_P\left[\mathcal{E}\left(r_m^{\epsilon_m^0, \beta_{a_0}}\right)\right] + C \left(\frac{\ln^2 n}{n} \right)^{\frac{\alpha+1}{\alpha+2}} \right)$$

where C is independent of n. Since $\beta_{a_0} \leq \beta_0$ and that there exists a constant A_1 such that $\alpha_0 > \alpha_{0,n} - A_1 \ln(n)^{-1} = \alpha'_{0,n}$, we have $\mathcal{P}_{\alpha_0,\beta_0,\mu_{max}} \subset \mathcal{P}_{\alpha'_{0,n},\beta_{a_0},\mu_{max}}$. Using theorem 11 we can upper bound $\mathbb{E}_P[\mathcal{E}(r^{\epsilon_m^0 \beta_{a_0}})]$ by $Cm^{-\psi(\alpha_0,\beta_{k_0})}$ where C depend on K and d and $\psi(\alpha, \beta) = \frac{\beta(1+\alpha)}{\beta(2+\alpha)+d}$. By construction, there exist A_2

such that $|\psi(\alpha_0, \beta_{a_0}) - \psi(\alpha_0, \beta_0)| \leq A_2 \ln(n)^{-1}$ and using that $n^{A_2/\ln(n)} = e^{A_2}$, we get

$$\mathbb{E}_P[\mathcal{E}(\tilde{r}^{adp})] \leq C \left(n^{-\psi(\alpha_0,\beta_0)} + C \left(\frac{\ln^2 n}{n} \right)^{\frac{\alpha_0+1}{\alpha_0+2}} \right)$$

We conclude the proof using that $\psi(\alpha_0, \beta_0) \leq \frac{\alpha_0+1}{\alpha_0+2}$. □

Proof of Theorem 15

Proof. We introduce the function Θ $:]0; 1[\times]0; \infty[\to]0; 1/2[, (\alpha, \beta) \mapsto \frac{\beta(1+\alpha)}{d+2\beta}$. There exists n_1 depending on K such that for any n greater than n_1 we have, for all $(\alpha, \beta) \in K$

$$\min_{(\alpha,\beta)\in K}(1+\alpha)\ln(n)^{-1} \leq \Theta(\alpha,\beta) \leq \max_{(\alpha,\beta)\in K}(1+\alpha)\lfloor \ln(n)/2 \rfloor \ln(n)^{-1}.$$

Let $(\alpha_0, \beta_0) \in K$ be such that $\alpha_0\beta_0 \leq d$. For $n \geq n_1$, we denote $k_0 \in \{1, \ldots, \lfloor \ln(n)/2 \rfloor\}$ the integer such that

$$(1+\alpha_0)k_0 \ln(n)^{-1} \leq \Theta(\alpha,\beta) \leq (1+\alpha_0)(k_0+1)\ln(n)^{-1}.$$

Let P be a probability distribution belonging to $\mathcal{P}_{\alpha_0,\beta_0,\mu_{max},\mu_{min}}$. Applying the Corollary 9 with $a = 1$, we get

$$\mathbb{E}\left[\mathcal{E}(r_{\tilde{f}^{adp}})|D_m^1\right] \leq 4\min_{r\in\mathcal{F}}(L(r) - L^*) + C\left(\frac{\ln \mathrm{Card}(\mathcal{F})}{l}\right)^{\frac{\alpha+1}{\alpha+2}}$$

Using that $l = n/\ln(n)$ and that $\mathrm{Card}(\mathcal{F}) \leq \ln(\mathrm{n})$ combined with the definition of k_0, we have

$$\mathbb{E}_P\left[\mathcal{E}(r_{\tilde{f}^{adp}})\right] \leq C\left(\mathbb{E}_P\left[\mathcal{E}\left(r_m^{\beta_{k_0}}\right)\right] + C\left(\frac{\ln^2 n}{n}\right)^{\frac{\alpha+1}{\alpha+2}}\right)$$

where C is independent of n. Since $\beta_{k_0} \leq \beta_0$, we have $\mathcal{P}_{\alpha_0,\beta_0,\mu_{max},\mu_{min}} \subset \mathcal{P}_{\alpha_0,\beta_{k_0},\mu_{max},\mu_{min}}$. Using theorem 14 we can upper bound $\mathbb{E}_P[\mathcal{E}(r_m^{\beta_{k_0}})]$ by $Cm^{-\Theta(\alpha_0,\beta_{k_0})}$ where C depend on K and d. By construction, we have $|\Theta(\alpha_0, \beta_{k_0}) - \Theta(\alpha_0, \beta_0)| \leq \ln(n)^{-1}$ and using that $n^{1/\ln(n)} = e$, we get

$$\mathbb{E}_P[\mathcal{E}(r_{\tilde{f}^{adp}})] \leq C\left(n^{-\Theta(\alpha_0,\beta_0)} + C\left(\frac{\ln^2 n}{n}\right)^{\frac{\alpha_0+1}{\alpha_0+2}}\right)$$

We conclude the proof using that $\Theta(\alpha_0, \beta_0) \leq \frac{\alpha_0+1}{\alpha_0+2}$ when $\alpha_0\beta_0 \leq d$. □

Proof of Theorem 16

Proof. The proof is classically based on Assouad's lemma. For $q \geq 1$, consider the regular grid on $[0;1]^d$ defined as

$$G^{(q)} = \left\{ \left(\frac{2k_1+1}{2q}, \dots, \frac{2k_d+1}{2q} \right) \text{ such as } k_1, \dots, k_d \in \{0, \dots, q-1\} \right\}.$$

Let $\eta_q(x) \in G^{(q)}$ be the closest point to $x \in [0;1]^d$ in $G^{(q)}$ (uniqueness of $\eta_q(x)$ is assumed: if it does not hold, define $\eta_q(x)$ as the one which is moreover closest to 0). Consider the partition $\mathcal{X}'_1, \dots, \mathcal{X}'_{q^d}$ of $[0,1]^d$ canonically defined using the grid $G^{(q)}$(x and y belong to the same subset iff $\eta_q(x) = \eta_q(y)$). Obviously, $\mathcal{X} = [0,1]^d = \cup_{i=1}^{q^d} \mathcal{X}'_i$. Let $u_1 : \mathbb{R}_+ \to \mathbb{R}_+$ be a non increasing infinitely differentiable function as in [3]. Let $u_2 : \mathbb{R}_+ \to \mathbb{R}_+$ be an infinitely differentiable bump function such as $u'_2 = 1$ on $[1/12, 1/6]$. Let $\phi_1, \phi_2 : \mathbb{R}^d \to \mathbb{R}_+$ be function defined as

$$\phi_i(x) = C_\phi u_i(\|x\|), \tag{A.1}$$

where the positive constant C_ϕ is taken small enough to ensure that $|\phi_i(x) - \phi_{i,x}(x')| \leq L\|x' - x\|^\beta$ for any $x, x' \in \mathbb{R}$. Thus $\phi_1, \phi_2 \in \Sigma(\beta, L, \mathbb{R})$. Now we define the hypercube $\mathcal{H}$. For this purpose, we merge together intervals: $G_k = [(k-1)K/q; kK/q] \times [0,1]^{d-1}, k \in \{1, \dots, H\}$ where K is the number of intervals we bring together relatively to the first coordinate (and it will play a role in the proof), $m = Kq^{d-1}$ is the number of cubes in a group and $H = \lfloor q/K \rfloor$. Define the hypercube $\mathcal{H} = \{\mathbb{P}_{\vec{\sigma}}, \vec{\sigma} \in S_m^H\}$, where S_m is the symmetric group of order m, of probability distributions $\mathbb{P}_{\vec{\sigma}}$ of (X, Y) as follows.

We design the marginal distribution of X that does not depend on σ and has a density μ w.r.t Lebesgue measure on $\mathbb{R}^d$. For fixed $0 < W$, we take μ as $\mu(x) = W/\lambda_d(B(z, 1/4q))$ if x belongs to a set $B(z, 1/6q) \setminus B(z, 1/12q)$ for some $z \in G^{(q)}$, and $\mu(x) = 0$ for all other x. We call $\mathcal{X}_i = \mathcal{X}_i \cap B(z, 1/6q) \setminus B(z, 1/12q)$ for $i = 1, \dots, m$. Next, the distribution of Y given X for $\mathbb{P}_{\sigma,k} \in \mathcal{H}$ is determined by the regression function, if $x \in \mathcal{X}'_i$ with $i \in \{1, \dots, m\}$,

$$\eta_{\vec{\sigma}}(x) = k(x)K/q + \sigma^{k(x)}(x)\tilde{h}\phi_1\left(q|x - \eta_q(x)|\right) + \tilde{h}\phi_2\left(q|x - \eta_q(x)|\right).$$

where $\tilde{h}$ is a function of q and $k(x) = \lfloor xK/q \rfloor$.

We now check the assumptions. Because of the design Hölder condition holds for $x, x' \in \mathcal{X}_i$ ([3]). In contrast of classification situation, we have to check whether Hölder condition holds for $x \in \mathcal{X}_i, x' \in \mathcal{X}_j$ when $i \neq j$ belong to a same group G_k. One can see that Hölder condition holds as soon as $m\tilde{h} \leq Lq^{-\beta}$ (i.e $K\tilde{h} \leq Lq^{1-d-\beta}$). Consider now the margin assumption. For $t = O(\tilde{h})$ the margin condition implies $W \leq C\tilde{h}^\alpha$. A constraint on K is also induced by the margin assumption: restricted to a group, the range of η has a measure of order $q^{-\beta}$ (because of the Hölder assumption). Hence, the margin assumption is satisfied if $mW = O(q^{-\alpha\beta})$ because of the strong density assumption $W \geq C/q^d$. Coupling the two last inequalities leads to $\alpha\beta \leq 1$, guaranteeing $K \geq 2$. So we take $\tilde{h} = C'q^{-d-\beta+\alpha\beta}$ and we verify that the margin condition holds. Indeed, if $\alpha\beta \leq d$, there exists $C' > 0$ such as $\tilde{h}^{\alpha} = C'q^{-\alpha d-\alpha\beta+\alpha^2\beta} \geq C/q^d$.

We denote $G(j)$, the set of cubes in the same group of j and for $i \in G(j), i \neq j$, $\sigma_{i,j} = +1$ if for all $x \in \mathcal{X}_i, x' \in \mathcal{X}_j$, $\eta_{\vec{\sigma}}(x) > \eta_{\vec{\sigma}}(x')$ and $\sigma_{i,j} = -1$ otherwise.

Using lemma 1, we have

$$\mathbb{E}_{\vec{\sigma}}L(r_n) - L^* = \frac{1}{2}\mathbb{E}_{\vec{\sigma}}\left[\mathbb{E}_{\vec{\sigma}}\left[\sum_{j=1}^{q^d}|\eta_{\vec{\sigma}}(X) - \eta_{\vec{\sigma}}(X')||r_{\vec{\sigma}}(X,X') - \hat{r}(X,X')|\mathbb{I}_{X'\in\mathcal{X}_j}\right]\right].$$

Using that $\mathcal{X} = \bigsqcup \mathcal{X}_i$ (i.e the disjoint union of the $\mathcal{X}_i$) combined with the definition of the margin law of X, we lower bound the excess risk by

$$\frac{W}{2}\mathbb{E}_{\vec{\sigma}}\left[\sum_{j=1}^{q^d}\mathbb{I}_{X'\in\mathcal{X}_j}\mathbb{E}_{\vec{\sigma}}\left[\sum_{i\in G(j),i\neq j}\int_{\mathcal{X}_i}|\eta_{\vec{\sigma}}(x) - \eta_{\vec{\sigma}}(X')||r_{\vec{\sigma}}(x,X') - \hat{r}(x,X')|\frac{dx}{\lambda(\mathcal{X}_i)}\right]\right].$$

We denote $d_\eta(\mathcal{X}_i, \mathcal{X}_j) = \min_{(x,x')\in(\mathcal{X}_i,\mathcal{X}_{j_0})}|\eta_\sigma(x) - \eta_\sigma(x')|$. Now, using the definition of $\sigma_{i,j}$ and $G(i)$, the last expression is lower than

$$\frac{W}{2}\mathbb{E}_{\vec{\sigma}}\left[\sum_{j=1}^{q^d}\mathbb{I}_{X'\in\mathcal{X}_j}\left[\mathbb{E}_{\vec{\sigma}}[\sum_{i\in G(j),i\neq j}d_\eta(\mathcal{X}_i,\mathcal{X}_j)\left|\sigma_{i,j} - \int_{\mathcal{X}_i}\hat{r}(x,X')\frac{dx}{\lambda(\mathcal{X}_i)}\right|\right]\right]$$

We denote by $\hat{\sigma}_{i,j} = \int_{X_i} r_n(x,X')\mathbb{I}_{X'\in\mathcal{X}_j}\frac{dx}{\lambda(\mathcal{X}_i)}$. So it remains to lower bound,

$$\sup_{\vec{\sigma}\in S_m^H}\mathbb{E}_{\vec{\sigma}}\left[\sum_{i\in G(j),i\neq j}d_\eta(\mathcal{X}_i,\mathcal{X}_j)|\sigma_{i,j} - \hat{\sigma}_{i,j}|\right].$$

Using that the sup is always greater than the mean and the linearity of the expectation, we lower bound by

$$\frac{1}{m!^H}\sum_{i\in G(j),i\neq j}\sum_{\vec{\sigma}\in S_m^H}\mathbb{E}_{\vec{\sigma}}\left[d_\eta(\mathcal{X}_i,\mathcal{X}_j)|\sigma_{i,j} - \hat{\sigma}_{i,j}|\right]$$

Restricting the sum to $\vec{\sigma}$'s such that the σ corresponding at the group $G(j)$ satisfies $\sigma(i) - \sigma(j) > m/2$ or $\sigma(j) - \sigma(i) > m/2$, we have $d_\eta(\mathcal{X}_i,\mathcal{X}_j) \geq C/q^\beta$. Combining this with the triangular inequality we obtain the following lower bound

$$\frac{1}{m!^H}\sum_{i\in G(j),i\neq j}\sum_{\vec{\sigma}\in S_m^H|\sigma_{i,j}=1,\sigma(i)-\sigma(j)>m/2}\frac{1}{2q^\beta}\mathbb{E}_{\pi_\sigma,\pi_{\tau_{i,j}\sigma}}[|\sigma_{i,j} - \tau_{i,j}\sigma_{i,j}|]$$

where $\tau_{i,j}$ is the transposition (i,j). Using inequality between the divergence and the Hellinger's distance, the last expression is greater than

$$\frac{1}{m!^H}\sum_{i\in G(j),i\neq j}\sum_{\sigma\in S_m^H|\sigma_{i,j}=1,\sigma(i)-\sigma(j)>m/2}\frac{1}{2q^\beta}(1-\sqrt{1-(1-H^2(P_{\vec{\sigma}}^{\otimes n},P_{\tau_{i,j}\vec{\sigma}}^{\otimes n})/2)^2}$$

A straightforward calculus shows that $H^2(P_{\vec{\sigma}}, P_{\tau_{i,j}\vec{\sigma}}) \leq 4W(1-\sqrt{1-q^{-2\beta}} \leq 4Wq^{-2\beta}$. Using argument of [4] we have,

$$1 - \sqrt{1-(1-H^2(\mathbb{P}_{\sigma}^{\otimes n}, \mathbb{P}_{\tau_{i,i\_}\sigma}^{\otimes n})/2)^2}) \geq 1 - \sqrt{2n(W/q^{2\beta})}$$

The number of $\sigma \in S_m^H$ such that $\sigma(i) - \sigma(j) > m/2$ is greater than $m!^H/8$, so finally we have

$$\inf_{r_n} \sup_{P \in \mathcal{P}_{\alpha,\beta,\mu_{min},\mu_{max}}} L(r_n) - L^* \geq C\frac{Wm}{q^{\beta}}(1 - q^{-\beta}\sqrt{2nW})$$

Now, we take $q = C_1 n^{1/(2\beta+d)}$ combined with $W = C_2 q^{-d}$ and $m = C_3 q^{d-\alpha\beta}$ with some positive constants C_1, C_2, C_3, to conclude the proof. □

Appendix B: Lower bounds

Basically, the idea of the proof of 16 is the following: first fix $\mathcal{X}_1 \subset \mathcal{X}$ a part of the space such that $X \in \mathcal{X}_1$ then create a classification problem as in [3] around $\mathcal{X}_1$. Doing that gives the rates of convergence of classification multiplied by the measure of $\mathcal{X}_1$. So the next step is to create classification problems for a union of part of the space with a measure independent of n. For the mild case in classification (see the proof in [3]), the classification problem uses the all space $\mathcal{X}$ i.e. all the important parts of the space have an η close to $1/2$ and a density close to zero. So with our strategy we obtain the rates of classification times the measure of the space $\mathcal{X}_1$ (i.e. W in the previous proof) which is not independent of n. For information only, we give the lower bounds that are achievable with this strategy. Since we believe they are not optimal, we do not give the proofs.

Oracle inequality. Adapting the proof of Theorem 3 in [19], one can get the next proposition. Let $\mathcal{P}_\alpha$ be the set of all probability distributions such that NA(α) holds.

Proposition 17 (Lower bound)**.** *For any integers M and n such that $M \leq \exp(n)$, there exist M prediction rules $f_1, \ldots, f_M$ such that for any decision function $\hat{f}_n$ and any $a > 0$, we have*

$$\sup_{P \in \mathcal{P}_\alpha} \left[\mathbb{E}[L(\hat{f}_n) - L^*] - (1+a) \min_{j=1,\ldots,M} (L(f_j) - L^*)\right] \geq C_1 \left(\frac{\log M}{n \log\log M}\right)^{\frac{2\alpha+2}{\alpha+2}},$$

where $C > 0$ is a constant depending only on α and c_0.

Notice that the power of n is half the power of n in the upper bound. Moreover a term in $\log\log(M)$ appears and comes from the fact that, we use permutations instead of the hypercube $\{-1, +1\}^{\log(M)}$.

Mild assumption. In that case, using directly the same proof as in the strong case with the choice of the parameters as in [3], one can prove the following proposition.

Proposition 18. *Let* $(\alpha, \beta) \in]0,1] \times \mathbb{R}_+^*$. *There exists a constant* $C > 0$ *such that, for any ranking rule* r_n *based on* n *independent copies of the pair* (X, Y), *we have:* $\forall n \geq 1$,

$$\sup_{P \in \mathcal{P}_{\alpha,\beta,\mu_{max},\mu_{min}}} \mathcal{E}(r_n) \geq C \cdot n^{-\frac{\beta(1+2\alpha)}{d+(2+\alpha)\beta}}.$$

Notice that the only change here is the factor 2 in front of the α.

References

[1] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the Area Under the ROC Curve. *The Journal of Machine Learning Research*, 6:393–425, 2005. MR2249826

[2] Pierre Alquier and Karim Lounici. Pac-bayesian bounds for sparse regression estimation with exponential weights. *EJS*, 5:127–145, 2011. MR2786484

[3] J.-Y. Audibert and A. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633, 2007. MR2336861

[4] J. Y. Audibert. Fast learning rates in statistical inference through aggregation. *Annals of Statistics*, 37:1591–1646, 2009. MR2533466

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

[6] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of Classification: A Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005. MR2182250

[7] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. MR2409394

[8] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. *Ann. Statist.*, 36(2):844–874, 2008. MR2396817

[9] S. Clémençon and S. Robbiano. Minimax learning rates for bipartite ranking and plug-in rules. In *Proceedings of the 28th international Conference on Machine Learning*, ICML'11, pages 441–448, 2011.

[10] S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, 2009. MR2582885

[11] S. Clémençon and N. Vayatis. Overlaying classifiers: a practical approach to optimal scoring. *Constructive Approximation*, 32(3):619–648, 2010. MR2726448

[12] A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Mach. Learn.*, 72(1-2):39–61, 2008.

[13] D. M. Green and J. A. Swets. *Signal detection theory and psychophysics*. Wiley, 1966.

[14] R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999. MR1720712

[15] Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research*, 4:933–969, 2003. MR2125342

[16] A. N. Kolmogorov and V. M. Tikhomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *Amer. Math. Soc. Translations Ser. 2,*, 17:277–364, 1961. MR0124720

[17] V. Koltchinskii and O. Beznosova. Exponential convergence rates in classification. In *Proceedings of COLT'05*, 2005. MR2203269

[18] G. Lecué. Optimal oracle inequality for aggregation of classifiers under low noise condition. In *COLT*, 2006. MR2280618

[19] G. Lecué. Classification with minimax fast rates for classes of bayes rules with sparse representation. *Electronic Journal of Statistics*, 2:741–773, 2008. MR2430253

[20] O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Annals Statistics*, 25:929–947, 1997. MR1447734

[21] P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math.*, 9:245–303, 2000. MR1813803

[22] P. Massart. *Concentration inequalities and model selection*. Lecture Notes in Mathematics. Springer, 2006. MR2319879

[23] P. Massart and E. Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5), 2006. MR2291502

[24] J.-B Monnier. Classification via local multi-resolution projections. *EJS*, 6:382–420, 2012. MR2988413

[25] P. Rigollet and A. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, 27:558–575, 2011.

[26] C. Rudin. Ranking with a P-Norm Push. In *Proceedings of COLT*, 2006.

[27] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Proceedings of NIPS*. 2010.

[28] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004. MR2051002

[29] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization (with discussion). *Annals of Statistics*, 32:56–85, 2004. MR2051001