# Minimax Learning Rates for Bipartite Ranking and Plug-in Rules

**Stéphan Clémençon** STEPHAN.CLEMENCON@TELECOM-PARISTECH.FR
**Sylvain Robbiano** SYLVAIN.ROBBIANO@TELECOM-PARISTECH.FR
LTCI UMR Telecom ParisTech/CNRS No. 5141, 46 rue Barrault, 75634 Paris cedex 13, France

## Abstract

While it is now well-known in the standard binary classification setup, that, under suitable margin assumptions and complexity conditions on the regression function, fast or even super-fast rates (*i.e.* rates faster than $n^{-1/2}$ or even faster than $n^{-1}$) can be achieved by *plug-in* classifiers, no result of this nature has been proved yet in the context of bipartite ranking, though akin to that of classification. It is the main purpose of the present paper to investigate this issue, by considering bipartite ranking as a nested continuous collection of cost-sensitive classification problems. A global *low noise* condition is exhibited under which certain (plug-in) ranking rules are proved to achieve fast (but not super-fast) rates over a wide nonparametric class of models. A lower bound result is also stated in a specific situation, establishing that such rates are optimal from a minimax perspective.

## 1. Introduction

The study of (minimax) learning rates in the context of classification/regression has been the subject of a good deal of attention in the machine-learning and statistical literature, see (Massart, 2000; Tsybakov, 2004; Audibert & Tsybakov, 2007; Audibert, 2009; Lecué, 2008; Koltchinskii & Beznosova, 2005; Srebro et al., 2010) for instance. Under adequate smoothness/complexity assumptions on the regression function combined with a margin (or low noise) condition, minimax rates for the excess of misclassification risk have been proved in a variety of situations. Such analyses of best achievable rates of classification take into account the bias in the excess of misclassification risk and establish that plug-in classifiers (*i.e.* classifiers directly built from a nonparametric estimate of the regression function) may be optimal in the *minimax* sense.

In parallel, a supervised learning problem termed *bipartite ranking*, akin to binary classification in the sense that it involves exactly the same probabilistic setup but of very different nature (it is *global* and not *local*), has recently received much interest in the statistical learning community, see (Freund et al., 2003; Rudin, 2006; Clémençon & Vayatis, 2009c) for instance, mainly because of its ubiquity in the applications: anomaly detection in signal processing, information retrieval, design of diagnosis tools in medicine, credit-scoring in finance among others. A rigorous formulation of the goal of bipartite ranking is given in (Clémençon et al., 2008), where it is cast in terms of minimization of a *pairwise classification error*, called the *ranking risk*. Minimization of this error measure can be shown as equivalent to maximization of the so-termed "AUC criterion" (Hanley & McNeil, 1982), a widely used ranking criterion in practice. In the latter paper, a *low noise* assumption has been proposed, under which *Empirical Risk Minimization* (ERM) is shown to yield rates close to $n^{-1}$, under the restrictive assumption that an optimal ranking rule belongs to the set of candidates over which ERM is performed (*i.e.* assuming zero bias for the ranking method considered). In (Clémençon & Vayatis, 2009a), plug-in ranking rules based on partitions (grids) of the input space have been considered in a less specific framework (relaxing the "zero bias" assumption namely), and have been proved to achieve rates slower than $n^{-1/2}$. It is the major purpose of this paper to pursue this analysis by considering more general *low noise* conditions together with smoothness/complexity assumptions for the regression function and study the rates attained by plug-in ranking rules, providing thus upper bounds for the *minimax rate of the expected excess of ranking risk*. Although the contribution of the present paper is mostly theoretical, given the difficulties one may face when trying to compute nonparametric estimates of the regression function (and thus plug-in predictors) in

Appearing in *Proceedings of the 28$^{\text{th}}$ International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

high dimension, it hopefully shed light on the nature of the ranking task, underlining the major differences between classification and bipartite ranking.

The article is organized as follows. In section 2, basic notions related to the bipartite ranking issue are briefly recalled and the main notations are set out. Crucial assumptions on the regression function and global low noise conditions are next described and thoroughly discussed. Preliminary results, based on the low noise conditions and linking the accuracy of nonparametric estimators of the regression function to the ranking risk of the related plug-in ranking rules are stated in section 3. The analysis of the rates achieved by plug-in ranking rules carried out in section 4 relies on the latter. Finally, a preliminary lower bound result for the minimax rate of the expected excess of ranking risk is stated in a specific situation in section 5, showing incidentally the minimax optimality of the plug-in rule studied in the preceding section in this particular case. Technical proofs are deferred to the Appendix.

## 2. Theoretical Background

For clarity, we first set out the main assumptions involved in the formulation of the bipartite ranking problem and recall important results that shall be used in the subsequent analysis, giving incidentally an insight into 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 euclidian space of (very) large dimension $d \geq 1$, $\mathbb{R}^d$ say. The r.v. $X$ is viewed as a random observation 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 $\mu$ and the posterior probability by $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$, $x \in \mathcal{X}$. For 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 continuous.

The indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$ and the range of any mapping $\Phi$ by $\mathrm{Im}(\Phi)$. We also denote by $\mathcal{B}(x, r)$ the closed Euclidean ball in $\mathbb{R}^d$ centered at $x \in \mathbb{R}^d$ and of radius $r > 0$. For any multi-index $s = (s_1, ..., s_d) \in \mathbb{N}^d$ and any $x = (x_1, ..., x_d) \in \mathbb{R}^d$, we set $|s| = \sum_{i=1}^d s_i$, $s! = s_1! ... s_d!$, $x^s = x_1^{s_1} ... x_d^{s_d}$ and $\|x\| = (x_1^2 + ... + x_d^2)^{1/2}$. Let $D^s$ denote the differential operator $D^s = \frac{\partial^{s_1 + ... + s_d}}{\partial x_1^{s_1} ... \partial x_d^{s_d}}$ 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 of degree $\lfloor \beta \rfloor$ at point $x$,

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

Finally, for $1 \leq q \leq \infty$, we denote by $\| \cdot \|_q$ the $L_q(\mathbb{R}^d, \mu)$ norm.

### 2.2. Bipartite Ranking

In contrast to binary classification, where the goal is to guess, for a given $x \in \mathcal{X}$, the likeliest label $C^*(x) = 2 \cdot \mathbb{I}\{\eta(x) > 1/2\} - 1$, the ranking task consists in sorting all the instances $x \in \mathcal{X}$ by increasing order of the posterior probability $\eta(x)$. A natural way of defining a pre-order on $\mathcal{X}$ is to transport the usual order on the real line onto $\mathcal{X}$ through a (measurable) *scoring function* $s : \mathcal{X} \to \mathbb{R}$: $\forall (x, x') \in \mathcal{X}^2$, $x \preceq_s x' \Leftrightarrow s(x) \leq s(x')$. The gold standard for evaluating the accuracy of such a preorder is of functional nature, the so-termed ROC curve (D.M.Green & Swets, 1966), namely the plot of the false positive rate against the true positive rate

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

**Pairwise classification.** As considering a performance criterion taking its values in a function space naturally leads to great difficulties in regards to mathematical analysis and computational implementation both at the same time, many authors have addressed the ranking issue from the perspective of *pairwise classification*, (Agarwal et al., 2005; Clémençon et al., 2005; Freund et al., 2003). In this setup, the objective is to determine, given two independent pairs $(X, Y)$ and $(X', Y')$ drawn from $P$, whether $Y' > Y$ or not. In this context, the predictor takes the form of a *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*:

$$L(r) \stackrel{def}{=} \mathbb{P}\{r(X, X') \cdot (Y' - Y) < 0\}. \quad (1)$$

A ranking rule $r$ is said *transitive* iff $\forall (x, x', x'') \in \mathcal{X}^3$: "$r(x, x') = +1$ and $r(x', x'') = +1$" $\Rightarrow$ "$r(x, x'') = +1$". 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.** In regards to the performance criterion above, the rule

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

defined by the regression function $\eta(x)$ (*i.e.* $r^* = r_\eta$) is unsurprisingly optimal, see Example 1 in (Clémençon et al., 2008) 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 - (Clémençon et al., 2008)) *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 $(X_1, Y_1)$, ..., $(X_n, Y_n)$ 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:

$$L^* = p(1-p) - \frac{1}{2}\mathbb{E}[|\eta(X) - \eta(X')|]. \tag{3}$$

Hence, in contrast to binary classification where 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. (3) roughly says that the more spread the r.v. $\eta(X)$, the easier the optimal ranking of $\mathcal{X}$'s elements.

**A continuum of classification problems.** In addition, we emphasize the fact that the optimal ranking rule $r^*(x, x')$ can be viewed 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 $\mathcal{R}_\omega(C) = 2(1-p)\omega \cdot \mathbb{P}\{Y = -1\} + 2p(1-\omega) \cdot \mathbb{P}\{Y = -1\}$ with cost $\omega = \eta(x)$, see Proposition 15 in (Clémençon & Vayatis, 2009b) for instance. Hence, while binary classification only aims at recovering the single level set $G^*_{1/2}$, which problem is made easier when $\eta(X)$ is far from $1/2$ with large probability (see (Massart & Nédélec, 2006) or (Tsybakov, 2004)), the ranking task consists in finding the whole collection $\{G^*_t : t \in \mathrm{Im}(\eta(X))\}$. 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.

**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" a nonparametric estimator $\widehat{\eta}_n(x)$ of the regression function $\eta$, based on a data sample $(X_1, Y_1)$, ..., $(X_n, Y_n)$, instead of $\eta(x)$ into Eq. (2):

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

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 (Audibert & Tsybakov, 2007) for instance) and on $\eta$'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 $\widehat{\eta}_n$ leads to ranking rules with fast rates of convergence of $\mathcal{E}(\widehat{r}_n)$ as the training sample size $n$ increases to infinity. This paper is hence devoted to the study of the convergence rates of plug-in ranking rules under specific assumptions on $(X, Y)$'s distribution, that are described/discussed in the next section.

### 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 in the footsteps of (Audibert & Tsybakov, 2007), we embrace the *minimax* point of view, that 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 $(X_1, Y_1)$, ..., $(X_n, Y_n)$. In order to carry out such a study, mainly three types of hypotheses shall be used. Following in the footsteps of (Audibert & Tsybakov, 2007), 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 "spreadness" conditions for $\eta(X)$'s distribution are stipulated.

**Complexity assumption.** In the plug-in view, 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 $\eta$ 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 $\beta$ 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 (Dudley, 1999) 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.** 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 borelian set $B \subset \mathbb{R}^d$ . The following assumption on the marginal distribution $\mu$ 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.

**Global low noise assumption.** Let $\alpha \in [0, 1]$. The condition stated below describes the behavior of $\eta(X)$.

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

$$\mathbb{P}\{|\eta(X) - \eta(x)| \leq t\} \leq C \cdot t^\alpha, \tag{4}$$

for some constant $C < \infty$.

Condition (4) above is void for $\alpha = 0$ and more and more restrictive as $\alpha$ grows. It clearly echoes Tsybakov's noise condition, introduced in (Tsybakov, 2004), which boils down to (4) 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 (4) implies global properties for $\eta(X)$'s distribution, as shown by the following result.

**Lemma 2** (Low noise and continuity) *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 $C/2$.*

Another low noise assumption has been proposed in (Clémençon et al., 2008) in the context of the study of the performance of empirical (ranking) risk minimizers. The latter may be formulated as follows.

Assumption $\mathbf{LN}(\alpha)$. There exists $C < \infty$ such that:

$$\forall x \in \mathcal{X}, \ \ \mathbb{E}[|\eta(x) - \eta(X)|^{-\alpha}] \leq C. \tag{5}$$

Under the hypothesis above, it has been proved that minimizers of an empirical version of the ranking risk (1) of the form of a $U$-statistic have an excess of risk of the order $O_{\mathbb{P}}((\log n/n)^{1/(2-\alpha)})$ when optimization is performed over classes of ranking functions of controlled complexity (VC major classes of finite VC dimension for instance), that contains an optimal ranking rule (assuming thus zero bias for the ERM method), see Proposition 5 and Corollary 6 in (Clémençon et al., 2008). The following result describes the connection between these assumptions.

**Proposition 3** (Noise assumptions) *The following assertions hold true.*

*(i) If $\eta(X)$ fulfills Assumption $\mathbf{LN}(\alpha)$ for $\alpha \in [0, 1]$ then Assumption $\mathbf{NA}(\alpha)$ holds.*

*(ii) Conversely, if $\eta(X)$ satisfies Assumption $\mathbf{NA}(\alpha)$ then Assumption $\mathbf{LN}(\alpha - \epsilon)$ holds for all $\epsilon > 0$.*

In contrast to what happens for Tsybakov's noise condition, where $\alpha$ can be very large, up to $+\infty$, recovering in the limit Massart's margin condition (Massart, 2000), Assumption $\mathbf{NA}(\alpha)$ can be fulfilled for $\alpha \leq 1$ solely. Indeed, as may be shown by a careful examination of Lemma 2's proof, bound (4) for $\alpha > 1$ implies that $F'(\eta(x)) = 0$, denoting by $F$ the cdf of $\eta(X)$. Therefore, the (probability) density of the r.v. $\eta(X)$ cannot be zero on its whole range $Im(\eta) = \{\eta(x), \ x \in \mathcal{X}\}$. Condition $\mathbf{LN}(\alpha)$ may looks rather technical and restrictive at first glance, but at first glance only. Indeed, it simply asks $\eta(X)$'s distribution to be sufficently spread. In addition, it far from restrictive: as shown by Proposition 3 combined with Corollary 8 in (Clémençon et al., 2008), assumption $\mathbf{LN}(1 - \epsilon)$ is fulfilled for any $\epsilon > 0$ as soon as $\eta(X)$ has a bounded density.

In the context of binary classification, by combining the CAR assumptions described above and Tsybakov's noise condition, optimal rates of convergence

have been obtained in (Audibert & Tsybakov, 2007). In particular, it has been shown that, 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 $n^{-1}$, depending on the values taken by the parameters $\alpha$ and $\beta$. We shall now attempt to determine whether similar results hold in ranking.

## 3. Comparison Inequalities

It is the purpose of this section to show how the low noise assumption **NA(α)** enables to link the accuracy of a nonparametric estimate of $\eta(x)$ in terms of $L_q$-approximation error to the excess of ranking risk of the related plug-in ranking rule. Here, $\bar{\eta}$ is a Borel function on $\mathbb{R}^d$ and $\bar{r}(x,x') = 2\cdot\mathbb{I}\{\bar{\eta}(x) \geq \bar{\eta}(x')\}-1$ denotes the corresponding (plug-in) ranking function. The following results improve upon the bound stated in (Clémençon & Vayatis, 2009a), see Corollary 9 therein.

**Proposition 4** (Risk excess and $L_q$-error) *Let $\alpha \in ]0,1[$ and assume that Assumption* **NA(α)** *is fulfilled. Then, the excess of ranking risk can be bounded as follows: there exists a constant $C < \infty$, such that for any distribution $P$ and all approximant $\bar{\eta}$, we have*

$$L(\bar{\eta}) - L^* \leq C\|\eta - \bar{\eta}\|_\infty^{1+\alpha}. \tag{6}$$

*In addition, we have:*

$$\mathbb{P}\{\bar{r}(X,X') \neq r^*(X,X')\} \leq C\|\eta - \bar{\eta}\|_\infty^{\alpha}, \tag{7}$$

*where $(X,X')$ denotes a pair of independent r.v.'s drawn from $\mu(dx)$.*

*Let $1 \leq q < \infty$. There exist finite constants $C_0(\alpha,q)$, $C_1(\alpha,q)$ such that, whatever the distribution $P$ and the approximant $\bar{\eta}$:*

$$L(\bar{\eta}) - L^* \leq C_0(\alpha,q)\|\eta - \bar{\eta}\|_q^{\frac{q(1+\alpha)}{q+\alpha}} \tag{8}$$

*and* $\mathbb{P}\{\bar{r}(X,X') \neq r^*(X,X')\} \leq C_1(\alpha,q)\|\eta - \bar{\eta}\|_q^{\frac{q}{q+\alpha}}$.

These inequalities permit to derive bounds for the expected excess of ranking risk of plug-in ranking rules directly (by taking the expectation). Considering $L_\infty(\mathbb{R}^d,\mu)$-error for instance, the existence of nonparametric *locally polynomial* estimators (LP) $\widehat{\eta}_n$, optimal in the minimax sense, such that

$$\sup_{\eta\in\Sigma(\beta,L,\mathbb{R}^d)} \mathbb{E}\left[\|\widehat{\eta}_n - \eta\|_\infty^m\right] \leq C\left(\log n/n\right)^{m\beta/(2\beta+d)}, \tag{9}$$

for any $m > 0$, has been shown in (Stone, 1982) under the strong density assumption. With $m = 1+\alpha$, this bound combined with Eq. (6) leads to an upper bound of the order $(\log n/n)^{(1+\alpha)\beta/(2\beta+d)}$ for the maximum expected excess of ranking risk of the rule $\widehat{r}_n = r_{\widehat{\eta}_n}$. Upper bound results, related to the MSE based on the $L_2(\mathbb{R}^d,\mu)$-error measure, established in (Yang, 1999) (see also (Stone, 1982) in a more restrictive framework, stipulating that the strong density assumption is fulfilled) claim that there exist nonparametric estimators of the regression function that attain the minimax rate $n^{-2\beta/(2\beta+d)}$ uniformly over the class $\Sigma(\beta,L,\mathbb{R}^d)$, yielding an upper bound of the order $n^{-2\beta(1+\alpha)/((2\beta+d)(2+\alpha))}$ for the maximum expected excess of ranking risk of the corresponding plug-in ranking functions.

However, although the comparison inequalities stated above are useful from a technical perspective (refer to the Appendix), as will be shown in the next section, such bounds are not optimal: in the $L_\infty$ case, an extra logarithm factor appears in the rate thus obtained and in the $L_2$ situation, the exponent involved in the rate is even suboptimal.

## 4. Fast Rates in Bipartite Ranking

Equipped with the intermediary results proved in the previous section, we are now ready for establishing upper bounds for the minimax rate of the expected excess of ranking risk $\inf_{r_n}\sup_{P\in\Sigma}\mathbb{E}[\mathcal{E}(r_n)]$, under the set of assumptions described in the following definition.

**Definition 5** *Let $\alpha \leq 1$, $\beta$ 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,*
2. *the global noise assumption* ***NA(α)*** *holds,*
3. *the regression function belongs to Hölder class $\Sigma(\beta,L,\mathbb{R}^d)$,*

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

An upper bound for the minimax rate is proved by exhibiting a sequence of ranking rules attaining the latter. Here we consider the same estimator as that studied in (Audibert & Tsybakov, 2007) (see section 3 therein). Let $K : \mathbb{R}^d \to \mathbb{R}$ be a Parzen-Rosenblatt kernel such that $K$ is bounded away from 0 on a neighborhood of 0 in $\mathbb{R}^d$, $\int(1+\|x\|^{4\beta})K^2(x)dx < \infty$ and $\sup_x(1+\|x\|^{2\beta})K^2(x) < \infty$. Fix $l \in \mathbb{N}$

and a bandwidth $h > 0$, set $U(u) = (u^s)_{|s|\leq l}$, $Q = (Q_{s_1,s_2})_{|s_1|,\,|s_2|\leq\lfloor l\rfloor}$ with $Q_{s_1,s_2} = \sum_{i=1}^n (X_i - x)^{s_1+s_2} K((X_i-x)/h)$ and $B_n = (B_{s_1,s_2})_{|s_1|,\,|s_2|\leq\lfloor\beta\rfloor}$ with $B_{s_1,s_2} = (nh^d)^{-1}\sum_{i=1}^n ((X_i-x)/h)^{s_1+s_2} K((X_i-x)/h)$. Consider the estimator $\widehat{\eta}_{n,h}(x)$ equal to the locally polynomial estimate

$$\widehat{\eta}_n^{LP}(x) = \sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h}\right) U^t(0) Q^{-1} U(X_i - x)$$

when $\widehat{\eta}_n^{LP}(x) \in [0,1]$ and $B_n$'s smallest eigenvalue is larger than $1/\log n$, and to 0 otherwise.

**Theorem 6** (A MINIMAX UPPER BOUND) *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 $\widehat{r}_n(x,x') = 2\cdot\mathbb{I}\{\widehat{\eta}_{n,h_n}(x') > \widehat{\eta}_{n,h_n}(x)\} - 1$, with $h_n = n^{-1/(2\beta+d)}$ and $l = \lfloor\beta\rfloor$, is bounded as follows:*

$$\sup_{P\in\mathcal{P}_{\alpha,\beta,L}} \mathcal{E}(\widehat{r}_n) \leq C\cdot n^{-\frac{\beta(1+\alpha)}{d+2\beta}}. \tag{10}$$

**Remark 2** (FAST, BUT NOT SUPER-FAST, RATES.) *Notice that, since $\alpha \leq 1$ rates faster than $n^{-1}$ cannot be achieved by the plug-in rule $\widehat{r}_n$ defined in the theorem above, in spite of the optimality of the related estimator $\widehat{\eta}_{n,h_n}$. However, for any $\alpha \in ]0,1]$, fast rates can be attained (i.e. rates faster than $n^{-1/2}$), provided that the regression function is sufficiently smooth, when $\beta > d/2\alpha$ namely. Note also that the rate highly depends on the dimension $d$ of the feature space, since the bias term is taken into account in the present analysis (in contrast to most convergence rate studies). Observe finally that building $\widehat{\eta}_{n,h_n}$ requires knowledge of the parameter $\beta$, unknown in most cases encountered in practice. The construction of plug-in for other purpose than to carry out minimax rate analysis leads to consider adaptive regression estimators, as in (Lecué, 2008) for instance.*

## 5. A Minimax Lower Bound

For completeness, we now state a lower bound for the minimax rate of the expected excess of ranking risk. It holds in a specific situation, when $d = 1$ and $\alpha\beta \leq 1$ namely, proving the results in full generality requiring significative developments of the argument sketched in the Appendix.

**Theorem 7** (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: $\forall n \geq 1$,*

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

**Remark 3** (MINIMAXITY & PLUG-IN OPTIMALITY.) *This result shows that the plug-in rule described in Section 4 is optimal in the case $\alpha\beta \leq 1$, the rates involved in Theorem 6 being minimax (and fast when, additionally, $\alpha\beta > 1/2$).*

## 6. Conclusion

The need for understanding the originality/specificity of bipartite ranking in regards to the (minimax) learning rates that can be attained, in comparison to classification rates in particular, motivates the present paper. A global low noise assumption, extending the Mammen-Tsybakov condition originally proposed in the context of binary classification, is introduced, under which novel comparison inequalities, linking approximation error of a regression estimate and ranking risk of the corresponding plug-in rule, are proved. By considering a specific (locally polynomial) regression estimator, we highlighted the fact that fast rates can be achieved (by plug-in ranking rules in particular) in certain situations. A preliminary lower bound result showed that these rates are actually optimal in a restrictive (univariate) situation. To the best of our knowledge, the present analysis, destined to be completed in regards to minimax lower bounds and adaptivity of the nonparametric estimators considered, is the first to state results of this nature.

## Technical Proofs

**Proof of Lemma 2.** Let $F$ denote $\eta(X)$'s cumulative distribution function. The first part of the lemma immediately results from the fact that $\mathbf{NA}(\alpha)$ can be rewritten as follows: $\forall (t,x) \in \mathbb{R}_+\times\mathcal{X}$, $F(\eta(x)+t) - F(\eta(x)-t) \leq C\cdot t^\alpha$. The cdf $F$ is thus absolutely continuous. Denote by $\phi$ the related density. Observe that, when $\alpha = 1$, the bound above can be written as $(F(\eta(x)+t) - F(\eta(x)-t))/t \leq C$. Letting then $t$ tend to zero, one obtains that, for all $x \in \mathrm{supp}(\mu)$, $2\phi(\eta(x)) \leq C$.

**Proof of Proposition 3.** Hölder inequality combined with condition $\mathbf{NA}(\alpha)$ shows that $\mathbb{E}[\mathbb{I}\{|\eta(X)-\eta(x)| < t\}]$ is bounded by

$$c^{1/(1+\alpha)}\mathbb{E}[\mathbb{I}\{|\eta(X)-\eta(x)| < t\}\cdot|\eta(X)-\eta(x)|]^{\alpha/(1+\alpha)},$$

which quantity is clearly less than $c^{1/(1+\alpha)}t^{\alpha/(1+\alpha)}$. This permit to prove assertion (i).

Let $x \in \mathcal{X}$ and $\epsilon > 0$ be fixed. We have

$$\mathbb{E}[|\eta(x')-\eta(X)|^{-\alpha+\epsilon}] = \int_0^{+\infty} \frac{\alpha}{t^{1+\alpha-\epsilon}}\mathbb{P}\{|\eta(X)-\eta(x)| < t\}\,dt.$$

Using $\mathbf{NA}(\alpha)$ when integrating over $[0,1]$ and bounding simply the probability by 1 otherwise, this permits to establish assertion $(ii)$.

**Proof of Lemma 4.** Lemma 1 yields

$$\mathcal{E}(r_{\bar{\eta}}) = \mathbb{E}[|\eta(X)-\eta(X')|\mathbb{I}\{(X,X')\in\Gamma_{\bar{\eta}}\}],$$

where $\Gamma_{\bar{\eta}} = \{(x,x') : (\bar{\eta}(x')-\bar{\eta}(x))(\eta(x')-\eta(x)) < 0\}$. Observe that on the event $\Gamma_{\bar{\eta}}$, we have

$$\begin{aligned}|\eta(X)-\eta(X')| &\leq |\eta(X)-\bar{\eta}(X)| + |\eta(X')-\bar{\eta}(X')| \\ &\leq 2 \|\eta-\bar{\eta}\|_\infty .\end{aligned}$$

Using now condition $\mathbf{NA}(\alpha)$, this proves the first part of the result. The same argument shows that $\mathbb{P}\{\bar{r}(X,X') \neq r^*(X,X')\} \leq \mathbb{P}\{|\eta(X)-\eta(X')| < 2\|\eta-\bar{\eta}\|_\infty\}$. Combining this bound to $\mathbf{NA}(\alpha)$ permits to finish the proof when $q=\infty$.

When $q < \infty$, decompose $\mathcal{E}(\bar{r}) = \mathbb{E}[|\eta(X)-\eta(X')|\mathbb{I}\{(X,X')\in\Gamma_{\bar{\eta}}\}]$ into a sum of two terms, depending on whether $|\eta(X)-\eta(X')| \leq t$ or not. As above, the first term is bounded by $\mathbb{E}[2|\eta(X)-\bar{\eta}(X)|\mathbb{I}\{|\eta(X)-\eta(X')| \leq t\}]$. Combining Hölder inequality with $\mathbf{NA}(\alpha)$, one gets that $2\mathbb{E}[|\eta(X)-\hat{\eta}_n(X)|\mathbb{I}\{|\eta(X)-\eta(X')|<t\}]$ is bounded by $C^{(q-1)/q}t^{\alpha(q-1)/q}\|\eta-\bar{\eta}\|_q$. The second term is bounded by the expectation of

$$\begin{aligned}&(|\eta(X)-\bar{\eta}(X)| + |\eta(X')-\bar{\eta}(X')|)\times \\ &\qquad \mathbb{I}\{|\eta(X)-\bar{\eta}(X)| + |\eta(X')-\bar{\eta}(X')| > t\},\end{aligned}$$

which term can be shown to be smaller than $4\mathbb{E}[|\eta(X)-\bar{\eta}(X)|\mathbb{I}\{|\eta(X)-\bar{\eta}(X)|>t/2\}]$. Combining Hölder and Markov inequalities, this is bounded by $2^{q+1}\|\eta-\bar{\eta}\|_q^q/t^{q-1}$. Finally, minimizing in $t$, we obtain the desired result. The same argument can be applied to $\mathbb{P}\{\bar{r}(X,X') \neq r^*(X,X')\}$, in order to decompose it into two terms, whether $|\eta(X)-\eta(X')| \leq t$ or not. The first one is bounded by $Ct^\alpha$ and the other one by $2^{q+1}\|\eta-\bar{\eta}\|_q^q/t^q$. Hence, optimizing in $t$ leads to the last bound stated in the Proposition.

**Proof of Theorem 6.** We start with establishing the following result.

**Lemma 8** *Assume that condition* $\mathbf{NA}(\alpha)$ *holds for* $\alpha > 0$. *Let* $\widehat{\eta}_n$ *be an estimator of* $\eta$. *Assume that* $\mathcal{P}$ *is a set of joint distributions such that:* $\forall n \geq 1$,

$$\sup_{P\in\mathcal{P}} \mathbb{P}\{|\widehat{\eta}_n(X)-\eta(X)| > \delta\} \leq C_1\exp(-C_2 a_n \delta^2), \tag{11}$$

*for some constants* $C_1$ *and* $C_2$. *Then, there exists a constant* $C<\infty$ *such that we have for all* $n \geq 1$*:*

$$\sup_{P\in\mathcal{P}} \mathbb{E}[\mathcal{E}(r_{\widehat{\eta}_n})] \leq C\cdot a_n^{(1+\alpha)/2}.$$

PROOF. Let $u \in (0,1)$, consider the sequence of (disjoint) subsets of $\mathbb{R}^d$ defined by

$$\begin{aligned}A_0(u) &= \{x\in\mathbb{R}^d : |\eta(x)-u| < \delta\}, \\ A_j(u) &= \{x\in\mathbb{R}^d : 2^{j-1}\delta < |\eta(x)-u| < 2^j\delta\}, \text{ for } j\geq 1.\end{aligned}$$

For any $\delta > 0$, we may write $\mathcal{E}(r_{\widehat{\eta}_n})$ as

$$\sum_{j\geq 0} \mathbb{E}_{X'}\mathbb{E}_X[|\eta(X)-\eta(X')|\mathbb{I}\{X\in A_j(X')\}\mathbb{I}\{(X,X')\in\Gamma_{\widehat{\eta}_n}\}]$$

The term corresponding to $j=0$ in the sum above is bounded by $C\delta^{1+\alpha}$ by virtue of assumption $\mathbf{NA}(\alpha)$. The one indexed by $j\geq 1$ is smaller than $2^{j+1}\delta\mathbb{E}\left[\mathbb{I}\{|\hat{\eta}_n(X)-\eta(X)| > 2^{j-2}\delta,\ X\in A_j(X')\}\right]$. Then, using the hypothesis on the class $\mathcal{P}$ plus assumption $\mathbf{NA}(\alpha)$, it is less than $2C_1 2^{j(1+\alpha)}\delta^{1+\alpha}\exp(-C_2 a_n(2^{j-2}\delta)^2)$. The proof is finished by summing all the bounds. □

It follows from Theorem 3.2 in (Audibert & Tsybakov, 2007) that (11) holds for the estimator considered with $a_n = n^{\frac{2\beta}{2\beta+d}}$ when $\mathcal{P} = \mathcal{P}_{\alpha,\beta,L}$. Using the lemma, this lead to the following upper bound for the excess risk. Now, using inequality (6) and taking $\delta = a_n^{-1/2}$, one gets the desired result.

**Proof of Theorem 7 (Sketch of).** The proof is classically based on Assouad's lemma. For $q\geq 1$, consider the regular grid on $[0,1]$ defined by

$$G^{(q)} = \left\{\frac{2k_1+1}{2q} : k\in\{0,...,q-1\}\right\}.$$

Let $\xi_q(x)\in G^{(q)}$ be the closest point to $x\in[0,1]$ in $G^{(q)}$ (uniqueness of $\eta_q(x)$ is assumed: if it does not hold, define $\xi_q(x)$ as the one which is in addition closest to 0). Consider the partition $\mathcal{X}_1',...,\mathcal{X}_q'$ of $[0,1]$ defined using the grid $G^{(q)}$: $x$ and $y$ belong to the same subset iff $\xi_q(x)=\xi_q(y)$. Obviously, $\mathcal{X} = [0,1] = \cup_{i=1}^q \mathcal{X}_i'$. Let $u_1 : \mathbb{R}_+ \to \mathbb{R}_+$ be a nonincreasing infinitely differentiable function as in (Audibert & Tsybakov, 2007). Let $u_2 : \mathbb{R}_+ \to \mathbb{R}_+$ be a infinitely differentiable bump function such as $u_2' = 1$ over $[1/12, 1/6]$. Define $\phi_1,\phi_2 : \mathbb{R}\to\mathbb{R}_+$ by $\phi_i(x) = C_\phi u_i(\|x\|)$,where the constant $C_\phi$ 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 $\phi_1,\phi_2\in\Sigma(\beta,L,\mathbb{R})$. We form groups of $K$ intervals, $G_k = [kK/q;(k+1)K/q], k\in\{1,...,\lfloor q/K\rfloor\}$, and define the hypercube $\mathcal{H} = \{\mathbb{P}_{\vec{\sigma}}, \vec{\sigma}\in\mathfrak{S}_{\lfloor q/K\rfloor}\}$, where $\mathfrak{S}_{\lfloor q/K\rfloor}$ is the symmetric group of order $\lfloor q/K\rfloor$, of probability distributions $\mathbb{P}_{\vec{\sigma}}$ on $[0,1]\times\{-1,+1\}$ as follows. We define $X$'s marginal distribution so that it does not depend on $\vec{\sigma}$ and has a density $\mu$ w.r.t Lesbesgue measure. Fix

$W > 0$ and set $\mu(x) = W/\lambda_d(B(z, 1/4q)$ 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 $x$. Next, the distristribution of $Y$ given $X$ for $\mathbb{P}_{\vec{\sigma}} \in \mathcal{H}$ is defined by the regression function

$$\eta_{\vec{\sigma}}(x) = k(x)K/q + \sigma^{(k(x))}(x)\tilde{h}\phi_1(q|x - \xi_q(x)|) + \tilde{h}\phi_2(q|x - \xi_q(x)|),$$

where $\tilde{h}$ is a function of $q$ and $k(x) = \lfloor xq/K \rfloor$. We now check the assumptions. Because of the design, Hölder condition holds for $x, x' \in \mathcal{X}_i$ (Audibert & Tsybakov, 2007). In contrast to the classification situation, we have to check that the Hölder condition holds for $x \in \mathcal{X}_i$, $x' \in \mathcal{X}_j$ when $i \neq j$, $\mathcal{X}_i$ and $\mathcal{X}_j$ belong to a same group $G_k$. One can see that Hölder condition holds as soon as $K\tilde{h} \leq Lq^{-\beta}$. Consider now the margin assumption. For $t = O(\tilde{h})$, it implies that $W \leq C\tilde{h}^{\alpha}$. A constraint on $K$ is also induced by the margin assumption: restricted to a group, the range of $\eta$ has a measure of order $q^{-\beta}$ (because of the Hölder assumption). Hence, the margin assumption is satisfied if $KW = O(q^{-\alpha\beta})$. Because of the strong density assumption, we also have $W > C/q$. Combining the two last inequalities leads to $\alpha\beta \leq 1$, guaranteeing that $K \geq 2$. Now, following step by step the argument in (Mammen & Tsybakov, 1995), (Lecué, 2008) and (Audibert, 2009), we can prove that:

$$\inf_{\hat{s}_n} \sup_{\pi \in \mathcal{P}_{\Sigma,\alpha,\beta}} L(\hat{s}_n) - L^* \geq \frac{KW}{128q^{\beta}}(1 - q^{-\beta}\sqrt{2nW}).$$

Finally, taking $q = C_1 n^{\frac{1}{2\beta+1}}$, $W = C_2/q$ and $K = C_3 q^{1-\alpha\beta}$, with some positive constants $C_1, C_2, C_3$ properly chosen, ends the proof.

## References


Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., and Roth, D. Generalization bounds for the AUC. *JMLR*, 6:393–425, 2005.

Audibert, J.Y. Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37:1591–1646, 2009.

Audibert, J.Y. and Tsybakov, A.B. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 32:608–633, 2007.

Clémençon, S. and Vayatis, N. On partitioning rules for bipartite ranking. In *Proceedings of AISTATS*, number 5, pp. 97–104. JMLR: W&CP, 2009a.

Clémençon, S. and Vayatis, N. Overlaying classifiers: a practical approach to optimal scoring. *To appear in Constructive Approximation*, 2009b.

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

Clémençon, S., Lugosi, G., and Vayatis, N. Ranking and scoring using empirical risk minimization. In *Proceedings of COLT*, 2005.

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

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

Dudley, R.M. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

Freund, Y., Iyer, R. D., Schapire, R. E., and Singer, Y. An efficient boosting algorithm for combining preferences. *JMLR*, 4:933–969, 2003.

Hanley, J.A. and McNeil, J. The meaning and use of the AUC. *Radiology*, (143):29–36, 1982.

Koltchinskii, V. and Beznosova, O. Exponential convergence rates in classification. In *Proceedings of COLT*, 2005.

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

Mammen, E. and Tsybakov, A.B. Asympotical minimax recovery of the sets with smooth boundaries. *Ann. Statist.*, 23(2):502–524, 1995.

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

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

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

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

Stone, C.J. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10:1040–1053, 1982.

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

Yang, Y. Minimax nonparametric classification. I. rates of convergence. II. model selection for adaptation. *IEEE Trans. Inf. Theory*, 45:2271–2292, 1999.