Ranking the Best Instances

Stéphan Clémençon
MODALX - Université Paris X Nanterre
&
Laboratoire de Probabilités et Modèles Aléatoires
UMR CNRS 7599 - Universités Paris VI et Paris VII

Nicolas Vayatis
Laboratoire de Probabilités et Modèles Aléatoires
UMR CNRS 7599 - Universités Paris VI et Paris VII

November 6, 2006

Abstract

We formulate the local ranking problem in the framework of *bipartite ranking* where the goal is to focus on the best instances. We propose a methodology based on the construction of real-valued scoring functions. We study empirical risk minimization of dedicated statistics which involve empirical quantiles of the scores. We first state the problem of *finding* the best instances which can be cast as a classification problem with mass constraint. We explore the conditions under which fast rates of convergence can be achieved for this problem. The main innovation is the presence of an additional term in the rate of convergence which accounts for quantile estimation. Next, we develop special performance measures for the local ranking problem which extend the Area Under an ROC Curve (AUC/AROC) criterion and describe the optimal elements of these new criteria. We also highlight the fact that the goal of ranking the best instances cannot be achieved in a stage-wise manner where first, the best instances would be tentatively identified and then a standard AUC criterion could be applied. Eventually, we state preliminary statistical results for the local ranking problem.

1 Introduction

The first takes all the glory, the second takes nothing. In applications where ranking is at stake, people often focus on the best instances. When scanning the results from a query on a search engine, we rarely go beyond the one or two first pages on the screen. In the different context of credit risk screening, credit establishments elaborate scoring rules as reliability indicators and their main concern is to identify risky prospects especially among the top scores. In medical diagnosis, test scores indicate the odds for a patient to be healthy given a series of measurements (age, blood pressure, ...). There again a particular attention is given to the "best" instances not to miss a possible diseased patient among the highest scores. These various situations can be formulated in the setup of bipartite ranking where one observes i.i.d. copies of a random pair (X, Y) with X being an observation vector describing the instance (web page, debtor, patient) and Y a binary label assigning to one population or the other (relevant vs. non relevant, good vs. bad, healthy vs. diseased). In this problem, the goal is to rank the instances instead of simply classifying them. There is a growing literature on the ranking problem in the field of Machine Learning but most of it considers the Area under the ROC Curve (also known as the AUC or AROC) criterion as a measure of performance of the ranking rule [6, 12, 22, 1]. In a previous work, we have mentioned that the bipartite ranking problem under the AUC criterion could be interpreted as a classification problem with pairs of observations [4]. But the limit of this approach is that it weights uniformly the pairs of items which are badly ranked. Therefore it does not permit to distinguish between ranking rules making the same number of mistakes but in very different parts of the ROC curve. The AUC is indeed a global criterion which does not allow to concentrate on the "best" instances. Special performance measures, such as the Discounted Cumulative Gain (DCG) criterion, have been introduced by practitioners in order to weight instances according to their rank [15] (see also [21, 7]) but providing theory for such criteria and developing empirical risk minimization strategies still is a very open issue. In the present paper, we extend the results of our previous work in [4] and set theoretical grounds for the problem of local ranking. The methodology we propose is based on the selection of a real-valued scoring function for which we formulate appropriate performance measures generalizing the AUC criterion. We point out that ranking the best instances is an involved task as it is a two-fold problem: (i) find the best instances and (ii) provide a good ranking on these instances. The fact that these two goals cannot be considered independently will be highlighted in the paper. Despite this observation, we will first formulate the issue of finding the best instances which is to be understood as a toy problem for our purpose. This problem corresponds to a *binary classification problem with a mass constraint* (where the proportion u_0 of $+1$ labels predicted by the classifiers is fixed) and it might present an interest *per se*. The main complication here has to do with the necessity of performing

quantile estimation which affects the performance of statistical procedures. Our proof technique was inspired by the former work of Koul [16] in the context of R-estimation where similar statistics arise.

The rest of the paper is organized as follows. We first state the problem of finding the best instances and study the performance of empirical risk minimization in this setup (Section 2). We also explore the conditions on the distribution in order to recover fast rates of convergence. In Section 3 we formulate performance measures for local ranking and provide extensions of the AUC criterion. Eventually (Section 4), we state some preliminary statistical results on empirical risk minimization of these new criteria.

2 Finding the best instances

In the present section, we have a limited goal which is only to determine the best instances without bothering of their order in the list. By considering this subproblem, we will identify the main technical issues involved in the sequel. It also permits to introduce the main notations of the paper.

Just as in standard binary classification, we consider the pair of random variables (X, Y) where X is an observation vector in a measurable space $\mathcal{X}$ and Y is a binary label in $\{-1, +1\}$. The distribution of (X, Y) can be described by the pair (μ, η) where μ is the marginal distribution of X and η is the a posteriori distribution defined by $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$, $\forall x \in \mathcal{X}$. We define the *rate of best instances* as the proportion of best instances to be considered and denote it by $u_0 \in (0, 1)$. We denote by $Q(\eta, 1 - u_0)$ the $(1 - u_0)$-quantile of the random variable $\eta(X)$. Then the *set of best instances at rate* u_0 is given by:

$$C^*_{u_0} = \{x \in \mathcal{X} \mid \eta(x) \geq Q(\eta, 1 - u_0)\} .$$

We mention two trivial properties of the set $C^*_{u_0}$ which will be important in the sequel:

- Mass constraint: we have $\mu(C^*_{u_0}) = \mathbb{P}\left\{X \in C^*_{u_0}\right\} = u_0$,
- Invariance property: as a functional of η, the set $C^*_{u_0}$ is invariant by strictly increasing transforms of η.

The problem of finding a proportion u_0 of the best instances boils down to the estimation of the unknown set $C^*_{u_0}$ on the basis of empirical data. Before turning to the statistical analysis of the problem, we first relate it to binary classification.

2.1 A classification problem with a mass constraint

A classifier is a measurable function $g : \mathcal{X} \to \{-1, +1\}$ and its performance is measured by the classification error $L(g) = \mathbb{P}\{Y \neq g(X)\}$. Let $u_0 \in (0, 1)$ be fixed. Denote by

$g^*_{u_0} = 2\mathbb{I}_{C^*_{u_0}} - 1$ the classifier predicting $+1$ on the set of best instances $C^*_{u_0}$ and -1 on its complement. The next proposition shows that $g^*_{u_0}$ is an optimal element for the problem of minimization of $L(g)$ over the family of classifiers g satisfying the *mass constraint* $\mathbb{P}\{g(X) = 1\} = u_0$.

Proposition 1 *For any classifier* $g : \mathcal{X} \to \{-1, +1\}$ *such that* $g(x) = 2\mathbb{I}_C(x) - 1$ *for some subset* C *of* $\mathcal{X}$ *and* $\mu(C) = \mathbb{P}\{g(X) = 1\} = u_0$*, we have*

$$L^*_{u_0} \stackrel{\circ}{=} L(g^*_{u_0}) \leq L(g) \ .$$

Furthermore, we have

$$L(g) - L(g^*_{u_0}) = 2\mathbb{E}\left(|\eta(X) - Q(\eta, 1 - u_0)|\, \mathbb{I}_{C^*_{u_0} \Delta C}(X)\right) ,$$

where Δ *denotes the symmetric difference operation between two subsets of* $\mathcal{X}$*.*

PROOF. For simplicity, we temporarily change the notation and set $q = Q(\eta, 1 - u_0)$. Then, for any classifier g satisfying the the constraint $\mathbb{P}\{g(X) = 1\} = u_0$, we have

$$L(g) = \mathbb{E}\left((\eta(X) - q)\mathbb{I}_{[g(X)=-1]} + (q - \eta(X))\mathbb{I}_{[g(X)=+1]}\right) + (1 - u_0)q + (1 - q)u_0 \ .$$

The statements of the proposition immediately follow. ∎

There are several progresses in the field of classification theory where the aim is to introduce constraints in the classification procedure or to adapt it to other problems. We relate our formulation to other approaches in the following remarks.

Remark 1 (CONNECTION TO HYPOTHESIS TESTING). The implicit asymmetry in the problem due to the emphasis on the best instances is reminiscent of the statistical theory of hypothesis testing. We can formulate a test of simple hypothesis by taking the *null assumption* to be $H_0 : Y = +1$ and the *alternative assumption* being $H_1 : Y = -1$. We want to decide which hypothesis is true given the observation X. Each classifier g provides a test statistic $g(X)$. The performance of the test is then described by its type I error $\alpha(g) = \mathbb{P}\{g(X) = 1 \mid Y = -1\}$ and its power $\beta(g) = \mathbb{P}\{g(X) = 1 \mid Y = +1\}$. We point out that if the classifier g satisfies a mass constraint, then we can relate the classification error with the type I error of the test defined by g through the relation:

$$L(g) = 2(1 - p)\alpha(g) + p - u_0$$

where $p = \mathbb{P}\{Y = 1\}$, and similarly, we have: $L(g) = 2p(1 - \beta(g)) - p - u_0$. Therefore, the optimal classifier minimizes the type I error (maximizes the power) among all classifiers with the same mass constraint. In some applications, it is more relevant to fix a constraint on the probability of a false alarm (type I error) and maximize the power. This question is explored in a recent paper by Scott [23] (see also [25]).

Remark 2 (Connection with regression level set estimation) We mention that the estimation of the level sets of the regression function has been studied in the statistics literature [3] (see also [29], [35]) as well as in the learning literature, for instance in the context of anomaly detection ([28, 24, 34]). In our framework of classification with mass constraint, the threshold defining the level set involves the quantile of the random variable $\eta(X)$.

Remark 3 (Connection with the minimum volume set approach) Although the point of view adopted in this paper is very different, the problem described above may be formulated in the framework of *minimum volume sets* learning as considered in [26]. As a matter of fact, the set $C^*_{u_0}$ may be viewed as the solution of the constrained optimization problem:

$$\min_C \mathbb{P}\{X \in C \mid Y = -1\}$$

over the class of measurable sets C, subject to

$$\mathbb{P}\{X \in C\} \geq u_0 \ .$$

The main difference in our case comes from the fact that the constraint on the volume set has to be estimated using the data while in [26] it is computed from a known reference measure. We believe that learning methods for minimum volume set estimation may hopefully be extended to our setting. A natural way to do it would consist in replacing conditional distribution of X given $Y = -1$ by its empirical counterpart. This is beyond the scope of the present paper but will be the subject of future investigation.

2.2 Empirical risk minimization

We now investigate the estimation of the set $C^*_{u_0}$ of best instances at rate u_0 based on training data. Suppose that we are given n i.i.d. copies $(X_1, Y_1), \cdots, (X_n, Y_n)$ of the pair (X, Y). Since we have the ranking problem in mind, our methodology will consist in building the candidate sets from a class $\mathcal{S}$ of real-valued scoring functions $s : \mathcal{X} \to \mathbb{R}$. Indeed, we consider sets of the form

$$C_s \stackrel{\circ}{=} C_{s,u_0} = \{x \in \mathcal{X} \mid s(x) \geq Q(s, 1 - u_0)\} \ ,$$

where s is an element of $\mathcal{S}$ and $Q(s, 1-u_0)$ is the $(1-u_0)$-quantile of the random variable $s(X)$. Note that such sets satisfy the same properties of $C^*_{u_0}$ with respect to mass constraint and invariance to strictly increasing transforms of s.

From now on, we will take the simplified notation:

$$L(s) \stackrel{\circ}{=} L(s, u_0) \stackrel{\circ}{=} L(C_s) = \mathbb{P}\{Y \cdot (s(X) - Q(s, 1 - u_0)) < 0\} \ .$$

A scoring function minimizing the quantity

$$L_n(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{Y_i \cdot (s(X_i) - Q(s, 1-u_0)) < 0\}.$$

is expected to approximately minimize the true error $L(s)$, but the quantile depends on the unknown distribution of X. In practice, one has to replace $Q(s, 1-u_0)$ by its empirical counterpart $\hat{Q}(s, 1-u_0)$ which corresponds to the empirical quantile. We will thus consider, instead of $L_n(s)$, the *truly empirical* error:

$$\hat{L}_n(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{Y_i \cdot (s(X_i) - \hat{Q}(s, 1-u_0)) < 0\}.$$

Note that $\hat{L}_n(s)$ is a complicated statistic since the empirical quantile involves all the instances $X_1, \ldots, X_n$. We also mention that $\hat{L}_n(s)$ is a biased estimate of the classification error $L(s)$ of the classifier $g_s(x) = 2\mathbb{I}\{s(x) \geq Q(s, 1-u_0)\} - 1$.

We introduce some more notations. Set, for all $t \in \mathbb{R}$:

- $F_s(t) = \mathbb{P}\{s(X) \leq t\}$
- $G_s(t) = \mathbb{P}\{s(X) \leq t \mid Y = +1\}$
- $H_s(t) = \mathbb{P}\{s(X) \leq t \mid Y = -1\}$

to be the cumulative distribution functions (cdf) of $s(X)$ (respectively, given $Y = 1$, given $Y = -1$). We recall that the definition of the quantiles of (the distribution of) a random variable involves the notion of generalized inverse F^{-1} of a function F:

$$F^{-1}(z) = \inf\{t \in \mathbb{R} \mid F(t) \geq z\}\ .$$

Thus, we have, for all $\nu \in (0, 1)$:

$$Q(s, \nu) = F_s^{-1}(\nu) \quad \text{and} \quad \hat{Q}(s, \nu) = \hat{F}_s^{-1}(\nu)$$

where $\hat{F}_s$ is the empirical cdf of $s(X)$: $\hat{F}_s(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{s(X_i) \leq t\}$, $\forall t \in \mathbb{R}$.

Without loss of generality, we will assume that all scoring functions in $\mathcal{S}$ take their values in $(0, \lambda)$ for some $\lambda > 0$. We now turn to study the performance of minimizers of $\hat{L}_n(s)$ over a class $\mathcal{S}$ of scoring functions defined by

$$\hat{s}_n = \arg\min_{s \in \mathcal{S}} \hat{L}_n(s).$$

Our first main result is an excess risk bound for the empirical risk minimizer $\hat{s}_n$ over a class $\mathcal{S}$ of uniformly bounded scoring functions. In the following theorem, we consider that the level sets of scoring functions from the class $\mathcal{S}$ form a Vapnik-Chervonenkis (VC) class of sets.

Theorem 2 *We assume that*

(i) the class $\mathcal{S}$ is symmetric (i.e. if $s \in \mathcal{S}$ then $\lambda - s \in \mathcal{S}$) and is a VC major class of functions with VC dimension V.

(ii) the family $\mathcal{K} = \{ G_s, H_s \ : \ s \in \mathcal{S} \}$ of cdfs satisfies the following property: any $K \in \mathcal{K}$ has left and right derivatives, denoted by K'_+ and K'_-, and there exist strictly positive constants b, B such that $\forall (K,t) \in \mathcal{K} \times (0,\lambda)$,

$$b \leq \left|K'_+(t)\right| \leq B \quad \textit{and} \quad b \leq \left|K'_-(t)\right| \leq B \ .$$

For any $\delta > 0$, we have, with probability larger than $1-\delta$,

$$L(\hat{s}_n) - \inf_{s \in \mathcal{S}} L(s) \leq c_1 \sqrt{\frac{V}{n}} + c_2 \sqrt{\frac{\ln(1/\delta)}{n}},$$

for some positive constants c_1, c_2.

We now provide some insights on conditions (i) and (ii) of the theorem.

Remark 4 (ON THE COMPLEXITY ASSUMPTION) On the terminology of major sets and major classes, we refer to Dudley [10]. In the proof, we need to control empirical processes indexed by sets of the form $\{x \ : \ s(x) \geq t\}$ or $\{x \ : \ s(x) \leq t\}$. Condition (i) guarantees that these sets form a VC class of sets.

Remark 5 (ON THE CHOICE OF THE CLASS $\mathcal{S}$ OF SCORING FUNCTIONS) In order to grasp the meaning of condition (ii) of the theorem, we consider the one-dimensional case with real-valued scoring functions. Assume that the distribution of the random variable X_i has a bounded density f with respect to Lebesgue measure. Assume also that scoring functions s are differentiable except, possibly, at a finite number of points, and derivatives are denoted by s'. Denote by f_s the density of $s(X)$. Let $t \in (0,\lambda)$ and denote by x_1, ..., x_p the real roots of the equation $s(x) = t$. We can express the density of $s(X)$ thanks to the change-of-variable formula (see e.g. [20]):

$$f_s(t) = \frac{f(x_1)}{s'(x_1)} + \ldots + \frac{f(x_p)}{s'(x_p)} \ .$$

This shows that the scoring functions should not present neither flat nor steep parts. We can take for instance, the class $\mathcal{S}$ to be the class of linear-by-parts functions with a finite number of local extrema and with uniformly bounded left and right derivatives: $\forall s \in \mathcal{S}$, $\forall x$, $m \leq s'_+(x) \leq M$ and $m \leq s'_-(x) \leq M$ for some strictly positive constants m, and M (see Figure 1). Note that any subinterval of $[0,\lambda]$ has to be in the range of scoring functions s (if not, some elements of $\mathcal{K}$ will present a *plateau*). In fact, the proof requires such a behavior only in the vicinity of the points corresponding to the quantiles $Q(s, 1-u_0)$ for all $s \in \mathcal{S}$.

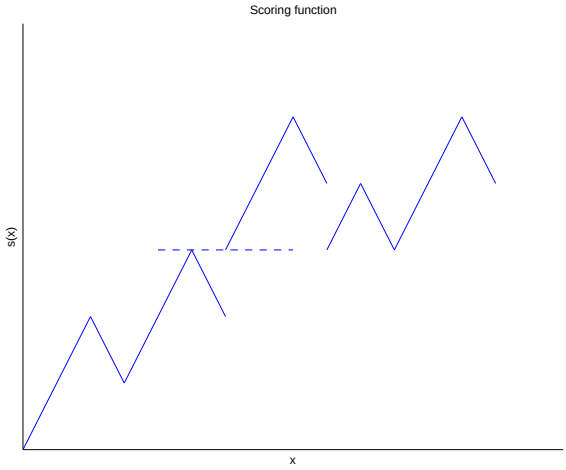

Figure 1: Typical example of a scoring function.

PROOF. Set $v_0 = 1 - u_0$. By a standard argument (see e.g. [8]), we have:

$$\begin{aligned} L(\hat{s}_n) - \inf_{s \in \mathcal{S}} L(s) &\leq 2 \sup_{s \in \mathcal{S}} \left| \hat{L}_n(s) - L(s) \right| \\ &\leq 2 \sup_{s \in \mathcal{S}} \left| \hat{L}_n(s) - L_n(s) \right| + 2 \sup_{s \in \mathcal{S}} |L_n(s) - L(s)| \ . \end{aligned}$$

Note that the second term in the bound is an empirical process whose behavior is well-known. In our case, assumption (i) implies that the class of sets $\{x : s(x) \geq Q(s, v_0)\}$ indexed by scoring functions s has a VC dimension smaller than V. Hence, we have by a concentration argument combined with a VC bound for the expectation of the supremum (see, e.g. [17]), for any $\delta > 0$, with probability larger than $1 - \delta$,

$$\sup_{s \in \mathcal{S}} |L_n(s) - L(s)| \leq c\sqrt{\frac{V}{n}} + c'\sqrt{\frac{\ln(1/\delta)}{n}}$$

for universal constants c, c'.

We now show how to handle the first term. Inspired by a work due to Koul [16], we set the following notations:

$$\begin{aligned} M(s, v) =& \mathbb{P}\left\{Y \cdot \left(s(X) - Q(s, v)\right) < 0\right\} \\ U_n(s, v) =& \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Y_i \cdot \left(s(X_i) - Q(s, v)\right) < 0\} - M(s, v) \ . \end{aligned}$$

and note that $U_n(s, v)$ is centered.

We then have the following decomposition, for any $s \in \mathcal{S}$ and $\nu_0 \in (0,1)$:

$$\left|\hat{L}_n(s) - L_n(s)\right| \leq \left|U_n(s, F_s \circ \hat{F}_s^{-1}(\nu_0)) - U_n(s,\nu_0)\right| + \left|M(s, F_s \circ \hat{F}_s^{-1}(\nu_0)) - M(s,\nu_0)\right| \ .$$

Note that $M(s, F_s \circ \hat{F}_s^{-1}(\nu_0)) = \mathbb{P}\left\{Y \cdot \left(s(X) - \hat{Q}(s,\nu)\right) < 0 \mid D_n\right\}$ where D_n denotes the sample $(X_1, Y_1), \cdots, (X_n, Y_n)$.

Recall the notation $p = \mathbb{P}\{Y = 1\}$. Since $M(s,\nu) = (1-p)(1-H_s \circ F_s^{-1}(\nu)) + pG_s \circ F_s^{-1}(\nu)$ and $F_s = pG_s + (1-p)H_s$, the mapping $\nu \mapsto M(s,\nu)$ is Lipschitz by assumption (ii). Thus, there exists a constant $\kappa < \infty$, depending only on p, b and B, such that:

$$\left|M(s, F_s \circ \hat{F}_s^{-1}(\nu_0)) - M(s,\nu_0)\right| \leq \kappa \left|F_s \circ \hat{F}_s^{-1}(\nu_0) - \nu_0)\right| .$$

Moreover, we have, for any $s \in \mathcal{S}$:

$$\begin{aligned}\left|F_s \circ \hat{F}_s^{-1}(\nu_0) - \nu_0\right| &\leq \left|F_s \circ \hat{F}_s^{-1}(\nu_0) - \hat{F}_s \circ \hat{F}_s^{-1}(\nu_0))\right| + \left|\hat{F}_s \circ \hat{F}_s^{-1}(\nu_0) - \nu_0)\right| \\ &\leq \sup_{t \in (0,\lambda)} \left|F_s(t) - \hat{F}_s(t)\right| + \frac{1}{n} \ .\end{aligned}$$

Here again, we can use assumption (i) and a classical VC bound from [17] in order to control the empirical process, with probability larger than $1 - \delta$:

$$\sup_{(s,t) \in \mathcal{S} \times (0,\lambda)} \left|F_s(t) - \hat{F}_s(t)\right| \leq c\sqrt{\frac{V}{n}} + c'\sqrt{\frac{\ln(1/\delta)}{n}}$$

for some constants c, c'.

It remains to control the term involving the process U_n:

$$\begin{aligned}\left|U_n(s, F_s \circ \hat{F}_s^{-1}(\nu_0)) - U_n(s,\nu_0)\right| &\leq \sup_{\nu \in (0,1)} |U_n(s,\nu) - U_n(s,\nu_0)| \\ &\leq 2 \sup_{\nu \in (0,1)} |U_n(s,\nu)|\end{aligned}$$

Using that the class of sets of the form $\{x : s(x) \geq Q(s,\nu)\}$ for $\nu \in (0,1)$ is included in the class of sets of the form $\{x : s(x) \geq t\}$ where $t \in (0,\lambda)$, we then have

$$\sup_{\nu \in (0,1)} |U_n(s,\nu)| \leq \sup_{t \in (0,\lambda)} \left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{Y_i \cdot (s(X_i) - t) < 0\} - \mathbb{P}\left\{Y \cdot (s(X) - t) < 0\right\}\right| \ ,$$

which leads again to an empirical process indexed by a VC class of sets and can be bounded as before. ∎

2.3 Fast rates of convergence

We now propose to examine conditions leading to fast rates of convergence (faster than $n^{-1/2}$). It has been noticed (see [18], [30], [19]) that it is possible to derive such rates of convergence in the classification setup under additional assumptions on the distribution. We propose here to adapt these assumptions for the problem of classification with mass constraint.

Our concern here is to formulate the type of conditions which render the problem easier from a statistical perspective. For this reason and to avoid technical issues, we will consider a quite restrictive setup where it is assumed that:

1. the class $\mathcal{S}$ of scoring functions is a finite class with N elements,
2. an optimal scoring rule s^* is contained in $\mathcal{S}$.

We have found that the following additional conditions on the distribution and the class $\mathcal{S}$ allow to derive fast rates of convergence for the excess risk in our problem.

(iii) There exist constants $\alpha \in (0,1)$ and $B > 0$ such that, for all $t \geq 0$,

$$\mathbb{P}\{|\eta(X) - Q(\eta, 1-u_0)| \leq t\} \leq B\, t^{\frac{\alpha}{1-\alpha}} \ .$$

(iv) The class of cumulative distribution functions F_s with $s \in \mathcal{S}$ is a subset of $\mathcal{H}(\beta, L)$, the Hölder class of functions $F : (0,1) \to \mathbb{R}$, satisfying:

$$\sup_{(x,y)\in(0,1)^2} \frac{|F(x) - F(y)|}{|x-y|^\beta} \leq L \ ,$$

with $L < \infty$ and $1 \leq \beta$.

We point out that these two conditions are not completely independent. Indeed, if F_η belongs to the class $\mathcal{H}(\beta, L)$, then condition (iii) is satisfied with $\alpha = \beta/(1+\beta)$ and $B = 2^\beta L$. This implies that $\alpha \in (1/2, 1)$ when $\beta \geq 1$. Note that condition (iii) simply extends the standard low noise assumption introduced by Tsybakov [30] (see also [2] for an account on this) where the level $1/2$ is replaced by the $(1-u_0)$-quantile of $\eta(X)$. Indeed, we have, under condition (iii), the variance control, for any $s \in \mathcal{S}$:

$$\mathrm{Var}(\mathbb{I}\{Y \neq \mathbb{I}_{C_s}(X)\} - \mathbb{I}\{Y \neq \mathbb{I}_{C^*_{u_0}}(X)\}) \leq (L(s) - L^*_{u_0})^\alpha \ .$$

Now, if we denote

$$s_n = \arg\min_{s\in\mathcal{S}} L_n(s) \ ,$$

we have, by a standard argument based on Bernstein's inequality (see Section 5.2 in [2]), with probability $1-\delta$,

$$L(s_n) - L^*_{u_0} \leq c \left(\frac{\log(N/\delta)}{n}\right)^{\frac{1}{2-\alpha}} .$$

for some positive constant c.

The novel part of the analysis below lies in the control of the bias induced by plugging the empirical quantile $\hat{Q}(s, 1-u_0)$ in the risk functional.

Theorem 3 *We assume that the class S of scoring functions is a finite class with N elements, and that it contains an optimal scoring rule s^*. Moreover, we assume that conditions (i)-(iv) are satisfied. Set*

$$\gamma = \begin{cases} \frac{\beta}{2} & \text{if } \beta < 2 \\ 1-\epsilon & \text{if } \beta \geq 2 \ , \quad \text{with } \epsilon > 0 \ . \end{cases}$$

Then, for any $\delta > 0$, we have, with probability $1-\delta$:

$$L(\hat{s}_n) - L^*_{u_0} \leq c_1 \left(\frac{\log(N/\delta)}{n}\right)^{\frac{1}{2-\alpha}} + c_2 \left(\frac{\log(N/\delta)}{n}\right)^{\gamma} ,$$

for some constants c_1, c_2.

Remark 6 (On the interpretation of the rate of convergence) In this result, the first term corresponds to the rate when quantile estimation is easy (when we have sufficient smoothness properties for the risk functional), while the second term is governed by the difficulty of approximating (uniformly) the component of the risk involving the quantile threshold. To understand the relative contribution of each of the two terms appearing in the previous result, we can set $\beta = \alpha/(1-\alpha)$ according to our previous observation on the dependency between conditions (iii) and (iv). We then observe that the second term is faster than the first one only for $\alpha \in (0, 2-\sqrt{2})$ which restricts the range of achievable fast rates (orders between $n^{-1/\sqrt{2}}$ and $n^{-1/2}$) with this method.

Remark 7 (On the exponent γ) The tuning of the parameter γ shows that the second term related to quantile estimation cannot go to zero arbitrarily fast. Beyond $\beta = 2$, increasing the regularity of F_s has no impact on the rate of convergence for this term and it is of the order $O(n^{-1+\epsilon})$ with arbitrarily small $\epsilon > 0$.

We now provide a more compact statement showing that there is a kind of "phase transition" in the rate of convergence as function of the regularity parameter β.

Corollary 4 *Assume that conditions (i), (ii) and (iv) are satisfied. Then, for any $\delta > 0$, we have, with probability $1-\delta$:*

$$L(\hat{s}_n) - L^*_{u_0} \leq \begin{cases} C\left(\frac{\log(N/\delta)}{n}\right)^{\frac{\beta+1}{\beta+2}} & \text{if } \beta > \sqrt{2} \\ C\left(\frac{\log(N/\delta)}{n}\right)^{\frac{\beta}{2}} & \text{if } \beta \in [1, \sqrt{2}] \end{cases}$$

for some constant C.

This corollary shows that when β is large enough and that there is enough smoothness of the risk functional, then quantile estimation does not affect the rate of convergence of the excess risk and the leading term is the first term of Theorem 3. However, for small values of β, the rate is limited by the second term corresponding to quantile estimation.

Before turning to the proof of Theorem 3, we recall an exponential inequality on the deviation between empirical and true quantiles from [27].

Theorem 5 (Serfling (1980)) *Suppose that $Q(s, v_0)$ is the unique solution q of*

$$\lim_{y \to q^-} F_s(y) \leq v_0 \leq F_s(q) \ .$$

Then, for any $\epsilon > 0$, we have, for all n,

$$\mathbb{P}\left\{\left|\hat{Q}(s, v_0) - Q(s, v_0)\right| > \epsilon\right\} \leq 2\exp\{-2n\Lambda^2(\epsilon)\}$$

where $\Lambda(\epsilon) = \min\{F_s(Q(s, v_0) + \epsilon) - v_0, v_0 - F_s(Q(s, v_0) - \epsilon)\}$.

We are now ready to give the proof of Theorem 3.

PROOF. We need to refine the control of the deviation term $\sup_{s \in \mathcal{S}} |\hat{L}_n(s) - L_n(s)|$. We set $v_0 = 1 - u_0$ again and use the same decomposition as in the proof of Theorem 2:

$$\left|\hat{L}_n(s) - L_n(s)\right| \leq \left|U_n(s, F_s \circ \hat{F}_s^{-1}(v_0)) - U_n(s, v_0)\right| + \left|M(s, F_s \circ \hat{F}_s^{-1}(v_0)) - M(s, v_0)\right| \ .$$

We start with the M-term. Using the regularity condition (iv), we have:

$$\left|F_s \circ \hat{F}_s^{-1}(v_0) - v_0\right| \leq L \ \left|\hat{Q}(s, v_0) - Q(s, v_0)\right|^{\beta} \ .$$

By condition (ii), the function F_s has no flat parts, hence, we can use Theorem 5 to obtain, with probability $1-\delta$:

$$\left|\hat{Q}(s, v_0) - Q(s, v_0)\right| \leq \min\{F_s^{-1}(v_0 + z) - F_s^{-1}(v_0),\ F_s^{-1}(v_0) - F_s^{-1}(v_0 - z)\}$$

where $z = \sqrt{\dfrac{\log(2N/\delta)}{2n}}$. By condition (ii), we also have, with probability $1-\delta$:

$$\left|F_s \circ \hat{F}_s^{-1}(v_0) - v_0\right| \leq c \ \left(\frac{\log(2N/\delta)}{2n}\right)^{\frac{\beta}{2}}$$

for some constant c. By the same argument as in the proof of Theorem 2, we get:

$$\left|M(s, F_s \circ \hat{F}_s^{-1}(v_0)) - M(s, v_0)\right| \leq \kappa c \ \left(\frac{\log(2N/\delta)}{2n}\right)^{\frac{\beta}{2}}$$

where κ is a positive constant.

We now turn to the U_n-term. We denote by $A(s, \epsilon)$ the event $\left\{\left|F_s \circ \hat{F}_s^{-1}(v_0) - v_0\right| < \epsilon\right\}$. On the event $A(s, \epsilon)$, we have:

$$\left|U_n(s, F_s \circ \hat{F}_s^{-1}(v_0)) - U_n(s, v_0)\right| \leq \sup_{v \,:\, |v - v_0| < \epsilon} |U_n(s, v) - U_n(s, v_0)| \ .$$

We will make use of an argument from [31] but, first, we need to put things into the right format. Set:

$$U_n(s, v) - U_n(s, v_0) = \frac{1}{n} \sum_{i=1}^{n} \left(u_i(s, v) - u_i(s, v_0)\right) \ ,$$

where $u_i(s, v) = \mathbb{I}\{Y_i \cdot (s(X_i) - Q(s, v)) < 0\} - \mathbb{P}\left\{\mathbb{I}\{Y \cdot (s(X) - Q(s, v)) < 0\}\right\}$ for $s \in \mathcal{S}$ and $v \in (0, 1)$. We observe that

$$|u_i(s, v) - u_i(s, v_0)| \leq d_i(v, v_0),$$

where

$$d_i(v, v_0) = \mathbb{I}\{s(X_i) \in [Q(s, v \wedge v_0), Q(s, v \vee v_0)]\} + |v - v_0| \ .$$

Denote by

$$\hat{d}(v, v_0) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{s(X_i) \in [Q(s, v \wedge v_0), Q(s, v \vee v_0)]\} + |v - v_0| \ .$$

a distance over $\mathbb{R}$. Set also:

$$\hat{R}(\epsilon) = \sup_{v \,:\, |v - v_0| < \epsilon} \hat{d}(v, v_0) \, .$$

and observe that

$$\hat{R}(\epsilon) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{s(X_i) \in [Q(s, v_0 - \epsilon), Q(s, v_0 + \epsilon)]\} + \epsilon \, .$$

We then have, by applying Lemma 8.5 from [31], for $nt^2/\hat{R}^2(\epsilon)$ sufficiently large,

$$\mathbb{P}\left\{\sup_{v\,:\,|v-v_0|\leq\epsilon}|U_n(s,v)-U_n(s,v_0)|\geq t\;\middle|\;X_1,\ldots,X_n\right\}\leq C\exp\left\{-\frac{cnt^2}{\hat{R}^2(\epsilon)}\right\}\;,$$

for some positive constants c and C.

It remains to integrate out and, for this purpose, we introduce the event:

$$\forall x>0\;,\qquad \Delta(x)=\left\{3\epsilon-x\leq\hat{R}(\epsilon)\leq 3\epsilon+x\right\}.$$

We then have:

$$\mathbb{E}\left(\exp\left\{-\frac{cnt^2}{\hat{R}^2(\epsilon)}\right\}\right)\leq\exp\left\{-\frac{cnt^2}{(3\epsilon+x)^2}\right\}+\mathbb{P}\left\{\overline{\Delta(x)}\right\}\;.$$

Now, we have, by Bernstein's inequality:

$$\mathbb{P}\left\{\overline{\Delta(x)}\right\}=2\mathbb{P}\left\{\frac{1}{n}B(n,2\epsilon)-2\epsilon>x\right\}\leq 2\exp\left\{-\frac{3nx^2}{16\epsilon}\right\}$$

where we have used the notation $B(n,2\epsilon)$ for a binomial $(n,2\epsilon)$ random variable.

We can take $x=O(t/\sqrt{\epsilon})$ and assume also $x=o(\epsilon)$ to get, for nt^2/ϵ^2 large enough,

$$\mathbb{P}\left\{\sup_{v\,:\,|v-v_0|\leq\epsilon}|U_n(s,v)-U_n(s,v_0)|\geq t\right\}\leq C\exp\left\{-\frac{cnt^2}{\epsilon^2}\right\}\;,$$

for some positive constants c and C.

This can be reformulated, by writing that the following bound holds, with probability larger than $1-\delta/2$,

$$\sup_{v\,:\,|v-v_0|\leq\epsilon}|U_n(s,v)-U_n(s,v_0)|\leq\epsilon\sqrt{\frac{\log(2C/\delta)}{nc}}\;.$$

We recall that, if we take $\epsilon=c\,\left(\frac{\log(4N/\delta)}{2n}\right)^{\frac{\beta}{2}}$, we have $\mathbb{P}\{A(s,\epsilon)\}\geq 1-\delta/2$. It follows that, with probability larger than $1-\delta$, we have

$$\left|U_n(s,F_s\circ\hat{F}_s^{-1}(v_0))-U_n(s,v_0)\right|\leq\epsilon\sqrt{\frac{\log(2C/\delta)}{nc}}\;,$$

for any $s\in\mathcal{S}$. We use the union bound to get the result.

Eventually, we point out that, by taking t of the order $\epsilon/\sqrt{n}$ and ϵ of the order $n^{-\beta/2}$, the constraints on x become $x=O(\sqrt{\epsilon/n})$ and $x=o(n^{-\beta/2})$. We can set $x=n^{-(\beta+\eta)/2}$ for some $\eta>0$ to see that the rate of this term cannot be faster than $n^{-1+\eta}$. ∎

3 Performance measures for local ranking

Our main interest here is to develop a setup describing the problem of not only finding but also ranking the best instances. As far as we know, this problem has not been considered from a statistical perspective until now. In the sequel, we build on the results from Section 2 and also on our previous work on the (global) ranking problem [5] in order to capture some of the features of the local ranking problem. The present section is devoted to the construction of performance measures reflecting the quality of ranking rules on a restricted set of instances.

3.1 ROC curves and optimality in the local ranking problem

We consider the same statistical model as before with (X, Y) being a pair of random variables over $\mathcal{X} \times \{-1, +1\}$ and we examine ranking rules resulting from real-valued scoring functions $s : \mathcal{X} \to (0, \lambda)$. The reference tool for assessing the performance of a scoring function s in separating the two populations (positive vs. negative labels) is the Receiving Operator Characteristic known as the ROC curve ([33], [11]). If we take the notations $\bar{G}_s(z) = \mathbb{P}\{s(X) > z \mid Y = 1\}$ (true positive rate) and $\bar{H}_s(z) = \mathbb{P}\{s(X) > z \mid Y = -1\}$ (false positive rate), we can define the ROC curve, for any scoring function s, as the plot of the function:

$$z \mapsto (\bar{H}_s(z), \bar{G}_s(z))$$

for thresholds $z \in (0, \lambda)$, or equivalently as the plot of the function:

$$t \mapsto \bar{G}_s \circ H_s^{-1}(1 - t)$$

for $t \in (0, 1)$. The optimal scoring function is the one whose ROC curve dominates all the others for all $z \in (0, \lambda)$ (or $t \in (0, 1)$) and such a function actually exists. Indeed, by recalling the hypothesis testing framework in the classification model (see Remark 1) and using Neyman-Pearson's Lemma, it is easy to check that ROC curve of the function $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$ dominates the ROC curve of any other scoring function. We point out that the ROC curve of a scoring function s is invariant by strictly increasing transformations of s.

In our approach, for a given scoring function s, we focus on thresholds z corresponding to the cut-off separating a proportion $u \in (0, 1)$ of top scored instances according to s from the rest. Recall from Section 2 that the best instances according to s are the elements of the set $C_{s,u} = \{x \in \mathcal{X} \mid s(x) \geq Q(s, 1 - u)\}$ where $Q(s, 1 - u)$ is the $(1 - u)$-quantile of $s(X)$. We set the following notations:

$$\alpha(s, u) = \mathbb{P}\{s(X) \geq Q(s, 1 - u) \mid Y = -1\}$$
$$\beta(s, u) = \mathbb{P}\{s(X) \geq Q(s, 1 - u) \mid Y = +1\} \ .$$

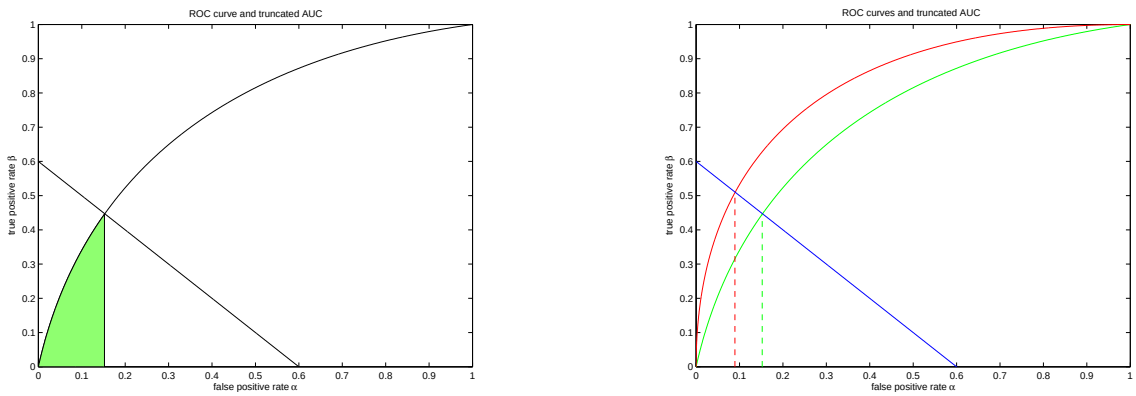

Figure 2: ROC curves, line $D(u_0, p)$ and truncated AUC at rate u_0 of best instances.

We propose to re-parameterize the ROC curve with the proportion $u \in (0, 1)$ and then describe it as the plot of the function:

$$u \mapsto (\alpha(s, u), \beta(s, u)) \ ,$$

for each scoring function s. When focusing on the best instances at rate u_0, we only consider the part of the ROC curve for values $u \in (0, u_0)$.

However attractive is the ROC curve as a graphical tool, it is not a practical one for developing learning procedures achieving straightforward optimization. The most natural approach is to consider risk functionals built after the ROC curve such as the Area Under an ROC Curve (known as the AUC or AROC, see [14]). Our goals in this section are:

1. to extend the AUC criterion in order to focus on restricted parts of the ROC curve,
2. to describe the optimal elements with respect to this extended criterion.

We point out the fact that extending the AUC is not trivial. Indeed, we notice that $\alpha(s, u)$ and $\beta(s, u)$ are related by a linear relation, for fixed u and p, when s varies:

$$u = p\beta(s, u) + (1 - p)\alpha(s, u)$$

where $p = \mathbb{P}\{Y = 1\}$. We denote the line plot of this relation by $D(u, p)$. Hence, the part of the ROC curve of a scoring function s corresponding to the best instances at rate u_0 is the part going from the origin $(0, 0)$ to the intersection between the line $D(u_0, p)$ and the ROC curve (shaded area in the left display of Figure 2). It follows that, the closer to η the scoring function s is, the higher the ROC curve is, but at the same time the integration domain shrinks (right display of Figure 2).

Our guideline in defining risk criteria for the problem of ranking the best instances is the form of the optimal elements. We expect the optimal scoring functions at the rate u_0 to belong to the equivalence class (functions defined up to the composition with a nondecreasing transformation) defined by scoring functions s^* such that:

$$s^*(x) = \begin{cases} \eta(x) & \text{if } x \in C^*_{u_0} \\ < \inf\limits_{C^*_{u_0}} \eta & \text{if } x \notin C^*_{u_0} \end{cases} .$$

Such scoring functions fulfill the two properties of finding the best instances (indeed $C_{s^*,u_0} = C^*_{u_0}$) and ranking them as well as the regression function. We will denote by $\mathcal{S}^*$ the set of optimal scoring functions for the problem of ranking the best instances at the rate u_0.

As a preliminary result, and before proposing an adequate criterion, we formulate a simple lemma.

Lemma 6 *For any scoring function s, we have for all $u \in (0,1)$,*

$$\begin{aligned} \beta(s,u) &\leq \beta(\eta,u) \\ \alpha(s,u) &\geq \alpha(\eta,u) \ . \end{aligned}$$

Moreover, we have equality only for those s such that $C_{s,u_0} = C^*_{u_0}$.

PROOF. We show the first inequality. By definition, we have:

$$\beta(s,u) = 1 - H_s(Q(s,1-u)) \ .$$

Observe that, for any scoring function s,

$$\begin{aligned} p(1 - H_s(Q(s,1-u)) &= \mathbb{P}\{Y = 1, s(X) > Q(s,1-u)\} \\ &= \mathbb{E}\left(\eta(X)\mathbb{I}\{X \in C_{s,u}\}\right) \ . \end{aligned}$$

We thus have

$$p\left(H_s(Q(s,1-u) - H_\eta(Q(\eta,1-u))\right.$$

$$= \mathbb{E}\left(\eta(X)(\mathbb{I}\{X \in C^*_u\} - \mathbb{I}\{X \in C_{s,u}\})\right)$$

$$= \mathbb{E}\left(\eta(X)\mathbb{I}\{X \notin C^*_u\}\left(\mathbb{I}\{X \in C^*_u\} - \mathbb{I}\{X \in C_{s,u}\}\right)\right)$$

$$+\mathbb{E}\left(\eta(X)\mathbb{I}\{X \in C^*_u\}(\mathbb{I}\{X \in C^*_u\} - \mathbb{I}\{X \in C_{s,u}\})\right)$$

$$\geq -\mathbb{E}\left(Q(\eta, 1-u)\mathbb{I}\{X \notin C^*_u\}\ \mathbb{I}\{X \in C_{s,u}\}\right) + \mathbb{E}\left(Q(\eta, 1-u)\mathbb{I}\{X \in C^*_u\}(1 - \mathbb{I}\{X \in C_{s,u}\})\right)$$

$$= Q(\eta, 1-u)(1-u-1+u) = 0\ .$$

The second inequality simply follows from the identity below:

$$1 - u = pH_s(Q(s, 1-u)) + (1-p)G_s(Q(s, 1-u))\ . \quad \blacksquare$$

In view of this result, a wide collection of criteria with the set $\mathcal{S}^*$ as the set of optimal elements could naturally be considered, depending on how one wants to weight the two types of error $1-\beta(s,u) =$ (type II error in the hypothesis testing framework) and $\alpha(s,u)$ (type I error) according to the rate $u \in [0, u_0]$. However, not all the criteria obtained in this manner can be interpreted as generalizations of the AUC criterion for $u_0 = 1$.

3.2 Generalization of the AUC criterion

In [5], we have considered the ranking error of a scoring function s as defined by:

$$R(s) = \mathbb{P}\{(Y - Y')(s(X) - s(X')) < 0\}\ ,$$

where (X', Y') is an i.i.d. copy of the random pair (X, Y).

Interestingly, it can be proved that minimizing the ranking error $R(s)$ is equivalent to maximizing the well-known AUC criterion. This is trivial once we write down the probabilistic interpretation of the AUC:

$$\mathrm{AUC}(s) = \mathbb{P}\left\{s(X) > s(X') \mid Y = 1,\ Y' = -1\right\} = 1 - \frac{1}{2p(1-p)}R(s)\ .$$

We now propose a local version of the ranking error on a measurable set $C \subset \mathcal{X}$:

$$R(s, C) = \mathbb{P}\left\{(s(X) - s(X'))(Y - Y') > 0,\ (X, X') \in C^2\right\}\ ,$$

and the local analogue of the AUC criterion:

$$\mathrm{LocAUC}(s, u) = \mathbb{P}\left\{s(X) > s(X'),\ s(X) \geq Q(s, 1-u)\ \mid\ Y = 1, Y' = -1\right\}\ .$$

This criterion obviously boils down to the standard criterion for $u = 1$. However, in the case where $u < 1$, we will see that there is no equivalence between maximizing the LocAUC criterion and minimizing the local ranking error $s \mapsto R(s,u) \stackrel{\circ}{=} R(s, C_{s,u})$. Indeed, the local ranking error is not a relevant performance measure for finding the best instances. Minimizing it would solve the problem of finding the instances that are the easiest to rank.

The following theorem states that scoring functions s^* in the set $\mathcal{S}^*$ maximize the criterion LocAUC and that the latter may be decomposed as a sum of a 'power' term and (the opposite of) a local ranking error term.

Theorem 7 *Let $u_0 \in (0,1)$. We have, for any scoring function s:*

$$\forall s^* \in \mathcal{S}^*, \quad \text{LocAUC}(s, u_0) \leq \text{LocAUC}(s^*, u_0) \ .$$

Moreover, the following relation holds:

$$\forall s, \quad \text{LocAUC}(s, u_0) = \beta(s, u_0) - \frac{1}{2p(1-p)} R(s, u_0) \ ,$$

where $R(s, u_0) = R(s, C_{s,u_0})$.

PROOF. Set $v_0 = 1 - u_0$. Observe first that:

$$\text{LocAUC}(s, u_0) = \mathbb{E}\left(H_s(s(X)) \ \mathbb{I}\{s(X) \geq Q(s, v_0)\} \ \mid Y = 1 \ \right)$$

$$= \int_{Q(s,v_0)}^{+\infty} H_s(z) \ G_s(dz) \ .$$

We use that $pG_s = F_s - (1-p)H_s$ and we obtain:

$$p\text{LocAUC}(s, u_0) = \int_{Q(s,v_0)}^{+\infty} H_s(z) \ F_s(dz) - (1-p) \int_{Q(s,v_0)}^{+\infty} H_s(z) \ H_s(dz)$$

$$= \int_{v_0}^{1} (1 - \alpha(s, v)) \ dv - \frac{1-p}{2} \left(1 - (1 - \alpha(s, v_0))^2 \right) .$$

This formula, combined with Lemma 6, establishes the first part of Theorem 7.

Besides, integrating by parts and making a change of variables, we get:

$$\int_{Q(s,v_0)}^{+\infty} H_s(z) \ G_s(dz) = 1 - (1 - \alpha(s, u_0))(1 - \beta(s, u_0)) - \int_0^{\alpha(s,u_0)} (1 - \beta(s, \alpha)) \ d\alpha$$

$$= \int_0^{\alpha(s,u_0)} \beta(s,\alpha)\, d\alpha + \beta(s,u_0)(1-\alpha(s,u_0)) \ .$$

On the other hand, one has

$$\alpha(s,u_0)\beta(s,u_0) = \frac{1}{p(1-p)}\mathbb{P}\left\{s(X) \wedge s(X') > Q(s,v_0),\ Y'=1,\ Y=-1\right\}$$

$$= \mathbb{P}\left\{s(X') > s(X),\ s(X) \wedge s(X') > Q(s,v_0)\ \mid\ Y'=1,\ Y=-1\right\}$$

$$+ \frac{1}{p(1-p)}\mathbb{P}\left\{s(X') < s(X),\ (X,X') \in C^2_{s,u_0},\ Y'=1,\ Y=-1\right\}$$

$$= \int_0^{\alpha(s,u_0)} \beta(s,\alpha)\ d\alpha + \frac{1}{2p(1-p)}R(s,u_0) \ .$$

Plugging this in the previous formula leads to the second statement of the theorem. ∎

Remark 8 (Truncating the AUC) In the theorem, we obviously recover the relation between the standard AUC criterion and the (global) ranking error when $u_0 = 1$. Besides, by checking the proof, one may relate the generalized AUC criterion to the truncated AUC. As a matter of fact, we have:

$$\forall s\ , \quad \mathrm{LocAUC}(s,u_0) = \int_0^{\alpha(s,u_0)} \beta(s,\alpha)\ d\alpha + \beta(s,u_0) - \alpha(s,u_0)\beta(s,u_0).$$

The values $\alpha(s,u_0)$ and $\beta(s,u_0)$ are the coordinates of the intersecting point between the ROC curve of the scoring function s and the line $D(u_0,p)$. Thus, the integral term represents the area of the surface delimited by the ROC curve, the horizontal x-axis and the line $x = \alpha(s,u_0)$ (see Figure 2). The theorem reveals that evaluating the local performance of a scoring statistic $s(X)$ by the truncated AUC as proposed in [9] is highly arguable since the maximizer of the functional $s \mapsto \int_0^{\alpha(s,u_0)} \beta(s,\alpha)\ d\alpha$ is usually not in $\mathcal{S}^*$.

3.3 Generalized Wilcoxon statistic

We now propose a different extension of the plain AUC criterion. Consider (X_1,Y_1), ..., (X_n,Y_n), n i.i.d. copies of the random pair (X,Y). The intuition relies on a well-known

relationship between Mann-Whitney and Wilcoxon statistics. Indeed, a natural empirical estimate of the AUC is the *rate of concording pairs*:

$$\widehat{\mathrm{AUC}}(s) = \frac{1}{n_+ n_-} \sum_{1 \le i,j \le n} \mathbb{I}\{Y_i = -1, Y_j = 1, s(X_i) < s(X_j)\} \ ,$$

with $n_+ = n - n_- = \sum_{i=1}^n \mathbb{I}\{Y_i = +1\}$. On the other hand, we recall that the Wilcoxon statistic $T_n(s)$ is the two-sample linear rank statistic associated to the *score generating function* $\Phi(v) = v$, $\forall v \in (0,1)$, obtained by summing the ranks corresponding to positive labels:

$$T_n(s) = \sum_{i=1}^n \mathbb{I}\{Y_i = 1\} \ \frac{\mathrm{rank}(s(X_i))}{n+1},$$

where $\mathrm{rank}(s(X_i))$ denotes the rank of $s(X_i)$ in the sample $\{s(X_j), 1 \le j \le n\}$. We refer to [13, 32] for basic results related to linear rank statistics. The following relation is well-known:

$$\frac{n_+ n_-}{n+1} \widehat{\mathrm{AUC}}(s) + \frac{n_+(n_+ + 1)}{2} = T_n(s) \ .$$

Moreover, the statistic $T_n(s)/n_+$ is an asymptotically normal estimate of

$$W(s) = \mathbb{E}\left(F_s(s(X)) \mid Y = 1\right) \ .$$

Note the theoretical counterpart of the previous relation may be written as

$$W(s) = (1-p)\mathrm{AUC}(s) + p/2 \ .$$

Now, in order to take into account a proportion u_0 of the highest ranks only, one may consider the criterion related to the score generating function $\Phi_{u_0}(v) = v \ \mathbb{I}\{v > 1 - u_0\}$:

$$W(s, u_0) = \mathbb{E}\left(\Phi_{u_0}(F_s(s(X))) \mid Y = 1\right)$$

which we shall call the *W-ranking error at rate* u_0.

Note that its empirical counterpart is given by $T_n(s, u_0)/n_+$, with

$$T_n(s, u_0) = \sum_{i=1}^n \mathbb{I}\{Y_i = 1\} \ \Phi_{u_0}\left(\frac{\mathrm{rank}(s(X_i))}{n+1}\right) \ .$$

Using the results from the previous subsection, we can easily check that the following theorem holds.

Theorem 8 *We have, for all s:*

$$\forall s^* \in \mathcal{S}^*, \quad W(s, u_0) \le W(s^*, u_0) \ .$$

Furthermore, we have:

$$W(s, u_0) = \frac{p}{2}\beta(s, u_0)(2 - \beta(s, u_0)) + (1-p)\mathrm{LocAUC}(s, u_0) \ .$$

PROOF. The result easily follows from the following representation of μ:

$$W(s,u_0) = \int_{Q(s,1-u_0)}^{+\infty} F_s(z)\ G_s(dz)$$

and from the fact that: $F_s = pG_s + (1-p)H_s$. ∎

Remark 9 (ON THE CHOICE OF A SCORE GENERATING FUNCTION Φ) The idea of weighting the empirical AUC criterion with non-uniform weights is equivalent to considering smooth score generating functions Φ instead of our Φ_{u_0} in the W-ranking error. Deriving optimality results for smooth criteria with our method is straightforward but we point out that, in this case, probabilistic interpretations are lost. In this approach, the stochastic processes arising are rank processes for which there is no theory available at this moment.

Remark 10 (EVIDENCE AGAINST 'DIVIDE-AND-CONQUER' STRATEGIES) It is noteworthy that not all combinations of $\beta(s,u_0)$ (or $\alpha(s,u_0)$) and $R(s,u_0)$ lead to a criterion with $\mathcal{S}^*$ being the set of optimal scoring functions. We have provided two non-trivial examples for which this is the case (Theorems 7 and 8). But, in general, this remark should prevent from considering naive 'divide-and-conquer' strategies for solving the local ranking problem. By naive 'divide-and-conquer' strategies, we refer here to stagewise strategies which would, first, compute an estimate $\hat{C}$ of the set containing the best instances, and then, solve the ranking problem over $\hat{C}$ as described in [5]. However, this idea combined with a certain amount of iterativeness might be the key to the design of efficient algorithms. In any case, we stress here the importance of making use of a global criterion, synthesizing our double goal: finding and ranking the best instances.

4 Empirical risk minimization of the local AUC criterion

In the previous section, we have seen that there are various performance measures which can be considered for the problem of ranking the best instances. In order to perform the statistical analysis, we will favor the representations of LOCAUC and W which involve the classification error $L(s,u_0)$ and the local ranking error $R(s,u_0)$. By combining Theorems 7 and 8, we can easily get:

$$2p(1-p)\text{LocAUC}(s,u_0) = (1-p)(p+u_0) - (1-p)L(s,u_0) - R(s,u_0)$$

and

$$2pW(s,u_0) = C(p,u_0) + \left(\frac{p+u_0}{2} - 1\right) L(s,u_0) - \frac{1}{4}L^2(s,u_0) - R(s,u_0)$$

where $C(p,u_0)$ is a constant depending only on p and u_0.

We exploit the first expression and choose to study the minimization of the following criterion for ranking the best instances:

$$M(s) \stackrel{\circ}{=} M(s, u_0) = R(s, C_{s,u_0}) + (1-p)L(s, u_0) \ .$$

It is obvious that the elements of $\mathcal{S}^*$ are the optimal elements of the functional $M(\ \cdot\ , u_0)$ and we will now consider scoring functions obtained through empirical risk minimization of this criterion.

More precisely, given n i.i.d. copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y), we introduce the empirical counterpart:

$$\widehat{M}_n(s) \stackrel{\circ}{=} \widehat{M}_n(s, u_0) = \widehat{R}_n(s) + \frac{n_-}{n}\widehat{L}_n(s),$$

with $n_- = \sum_{i=1}^n \mathbb{I}\{Y_i = -1\}$ and

$$\widehat{R}_n(s) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}\{(s(X_i) - s(X_j))(Y_i - Y_j) > 0,\ s(X_i) \wedge s(X_j) \geq \widehat{Q}(s, 1-u_0)\} \ .$$

Note that $\widehat{R}_n(s)$ is expected to be close to the U-statistic of degree two

$$R_n(s) = \frac{1}{n(n-1)} \sum_{i \neq j} k_s((X_i, Y_i), (X_j, Y_j)),$$

with symmetric kernel

$$k_s((x, y), (x', y')) = \mathbb{I}\{(s(x) - s(x'))(y - y') > 0,\ s(x) \wedge s(x') \geq Q(s, 1-u_0)\} \ .$$

The statistic $R_n(s)$ corresponds to an unbiased estimate of the local ranking error $R(s, u_0)$. The next result provides a standard error bound for the excess risk of the empirical risk minimizer over a class $\mathcal{S}$ of scoring functions:

$$\widehat{s}_n = \arg\min_{s \in \mathcal{S}} \widehat{M}_n(s) \ .$$

Proposition 9 *Assume that conditions (i)-(ii) of Theorem 2 are fulfilled. Then, there exist constants c_1 and c_2 such that, for any $\delta > 0$, we have:*

$$M(\widehat{s}_n) - \inf_{s \in \mathcal{S}} M(s) \leq c_1 \sqrt{\frac{V}{n}} + c_2 \sqrt{\frac{\ln(1/\delta)}{n}}$$

with probability larger than $1 - \delta$.

PROOF. (SKETCH) The proof combines the argument used in the proof of Theorem 2 with the techniques used in establishing Proposition 2 in [4].

$$M(\hat{s}_n) - \inf_{s\in\mathcal{S}} M(s) \leq 2\left(\sup_{s\in\mathcal{S}}\left|\hat{R}_n(s) - R_n(s)\right| + \sup_{s\in\mathcal{S}}|R(s) - R_n(s)|\right)$$
$$+ 2(1-p)\left(\sup_{s\in\mathcal{S}}\left|\hat{L}_n(s) - L_n(s)\right| + \sup_{s\in\mathcal{S}}|L(s) - L_n(s)|\right) + 2\left|\frac{n_+}{n} - p\right| .$$

The middle term may be bounded by applying the result stated in Theorem 2, while the last one can be handled by using Bernstein's exponential inequality for an average of Bernoulli random variables. By combining Lemma 1 in [4] with the Chernoff method, we can deal with the U-process term $\sup_{s\in\mathcal{S}}|R(s) - R_n(s)|$. Finally, the term $\sup_{s\in\mathcal{S}}\left|\hat{R}_n(s) - R_n(s)\right|$ can also be controlled by repeating the argument in the proof of Theorem 2. The only difference here is that we have to consider the U-process term

$$\sup_{(s,t)}\left|\frac{2}{n(n-1)}\sum_{i\neq j}\{K_{s,t}((X_i,Y_i),(X_j,Y_j)) - \mathbb{E}[K_{s,t}((X,Y),(X',Y'))]\}\right|$$

with

$$K_{s,t}((x,y),(x',y')) = \mathbb{I}\{(s(x)-s(x'))(y-y') > 0,\ s(x)\wedge s(x') \geq t\} .$$

For deriving first-order results with such a process, we refer to the same type of argument as used in [4]. ∎

Remark 11 (ABOUT THE POSSIBILITY OF DERIVING FAST RATES) By checking the proof sketch, it turns out that sharper bounds may be achieved for the U-process term. Indeed, it is a simple variation of our previous work in [4] where we have used Hoeffding's decomposition in order to grasp the deep structure of the underlying statistic. Here we will need, in addition, condition (iii) to hold for all $u \in (0, u_0]$. Indeed, if we localize our low-noise assumption from [4], it takes the following form: there exist constants $\alpha \in (0,1)$ and $B > 0$ such that, for all $t \geq 0$, we have

$$\forall x \in C^*_{u_0}, \qquad \mathbb{P}\{|\eta(X) - \eta(x)| \leq t\} \leq B\, t^{\frac{\alpha}{1-\alpha}} .$$

It is easy to see that this is equivalent to condition (iii) for all $u \in (0, u_0]$: there exist constants $\alpha \in (0,1)$ and $B > 0$ such that, for all $t \geq 0$, we have

$$\forall u \in (0, u_0], \qquad \mathbb{P}\{|\eta(X) - Q(\eta, 1-u)| \leq t\} \leq B\, t^{\frac{\alpha}{1-\alpha}} .$$

However, in the present formulation where p is assumed to be unknown, it looks like this improvement will be spoiled by the 'proportion term' which will still be of the order of a $O(n^{-1/2})$.

References

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

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

[3] L. Cavalier. Nonparametric estimation of regression level sets. *Statistics*, 29:131–160, 1997.

[4] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and scoring using empirical risk minimization. In P. Auer and R. Meir, editors, *Proceedings of COLT 2005*, volume 3559 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2005.

[5] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. Technical report, Laboratoire de Probabilités et Modèles Aléatoires, Université Pierre et Marie Curie, 2006.

[6] C. Cortes and M. Mohri. Auc optimization vs. error rate minimization. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2000.

[7] D. Cossock and T. Zhang. Statistical analysis of Bayes optimal subset ranking. Technical report, Yahoo! Research, 2006.

[8] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[9] L. E. Dodd and M. S. Pepe. Partial auc estimation and regression. *Biometrics*, 59(3):614–623, 2003.

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

[11] J.P. Egan. *Signal Detection Theory and ROC Analysis*. Academic Press, 1975.

[12] Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4, 2003.

[13] J. Hájek and Z. Sidák. *Theory of Rank Tests*. Academic Press, 1967.

[14] J.A. Hanley and J. McNeil. The meaning and use of the area under a ROC curve. *Radiology*, (143):29–36, 1982.

[15] K. Järvelin and J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In N.J. Belkin, P . Ingwersen, , and M.-K. Leong, editors, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, 2000.

[16] H.L. Koul. *Weighted Empirical Processes in Dynamic Nonlinear Models*, volume 166 of *Lecture Notes in Statistics*. Springer, 2nd edition, 2002.

[17] G. Lugosi. Pattern classification and learning theory. In L. Györfi, editor, *Principles of Nonparametric Learning*, pages 1–56.

[18] E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27(6):1808–1829, 1999.

[19] P. Massart and E. Nédélec. Risk bounds for statistical learning. *Annals of Statistics*, 34(5), 2006.

[20] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1965.

[21] C. Rudin. Ranking with a P-Norm Push. In H.U. Simon and G. Lugosi, editors, *Proceedings of COLT 2006*, volume 4005 of *Lecture Notes in Computer Science*, pages 589–604, 2006.

[22] C. Rudin, C. Cortes, M. Mohri, and R. E. Schapire. Margin-based ranking and boosting meet in the middle. In P. Auer and R. Meir, editors, *Proceedings of COLT 2005*, volume 3559 of *Lecture Notes in Computer Science*, pages 63–78. Springer, 2005.

[23] C. Scott. Performance measures for neyman-pearson classification. Technical report, Department of Statistics, Rice University, 2005.

[24] C. Scott and M. Davenport. Regression level set estimation via cost-sensitive classification. *IEEE Transactions on Signal Processing*, 2006, to appear.

[25] C. Scott and R. Nowak. A Neyman-Pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819, November 2005.

[26] C. Scott and R. Nowak. Learning minimum volume sets. *Journal of Machine Learning Research*, 7:665–704, April 2006.

[27] R.J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 1980.

[28] I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6:211–232, 2005.

[29] A. Tsybakov. On nonparametric estimation of density level sets. *Annals of Statistics*, 25(3):948–969, 1997.

[30] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1):135–166, 2004.

[31] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[32] A. van de Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[33] H.L. van Trees. *Detection, Estimation, and Modulation Theory, Part I*. John Wiley, 1968.

[34] R. Vert and J.-P. Vert. Consistency and convergence rates of one-class svms and related algorithms. *Journal of Machine Learning Research*, 7:817–854, May 2006.

[35] R. Willett and R. Nowak. Minimax optimal level set estimation. Technical report, Rice University, 2006.