Journal of Machine Learning Research 8 (2007) 2671-2699 Submitted 11/06; Revised 2/07; Published 12/07

Ranking the Best Instances

Stéphan Clémençon CLEMENCO@ENST.FR
Département TSI Signal et Images - LTCI UMR GET/CNRS 5141
Ecole Nationale Supérieure des Télécommunications
37-39, rue Dareau
75014 Paris, France

Nicolas Vayatis VAYATIS@CMLA.ENS-CACHAN.FR
Centre de Mathématiques et de Leurs Applications - UMR CNRS 8536
Ecole Normale Supérieure de Cachan - UniverSud
61, avenue du Président Wilson
94 235 Cachan cedex, France

Editor: Yoram Singer

Abstract

We formulate a local form of the bipartite ranking problem 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. Next, we develop special performance measures for the local ranking problem which extend the Area Under an ROC Curve (AUC) 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.

Keywords: ranking, ROC curve and AUC, empirical risk minimization, fast rates

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) criterion as a measure of performance of the

ranking rule (Cortes and Mohri, 2004; Freund et al., 2003; Rudin et al., 2005; Agarwal et al., 2005). 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 (Clémençon et al., 2005). 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 (Järvelin and Kekäläinen, 2000) but providing theory for such criteria and developing empirical risk minimization strategies still is a very open issue. Recent works by Rudin (2006), Cossock and Zhang (2006), and Li et al. (2007) reveal that there are several possibilities when designing ranking algorithms with focus on the top-rated instances. In the present paper, we focus on statistical aspects rather than algorithmic. We extend the results of our previous work in Clémençon et al. (2005) 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 (2002) in the context of *R*-estimation where similar statistics, known as linear signed rank statistics, arise. By exploiting the structure of such statistics, we are able to establish noise conditions in a similar way as in Clémençon et al. (To appear) where we had to deal with performance criteria based on *U*-statistics. Under such conditions, we prove that rates of convergence up to $n^{-2/3}$ can be guaranteed for the empirical risk minimizer in the classification problem with mass constraint. Another contribution of the paper lies in our study of the optimality issue for the local ranking problem. We discuss how focusing on best instances affects the ROC curve and the AUC criterion. We propose a family of possible performance measures for the problem of ranking the best instances. In particular, we show that widespread ideas in the biostatistics literature about the partial AUC (see Dodd and Pepe, 2003) turn out to be questionable with respect to optimality considerations. We also point out that the empirical risks for local ranking are closely related to generalized Wilcoxon statistics.

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 with 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 $\mu\big(C^*_{u_0}\big) = \mathbb{P}\left\{X \in C^*_{u_0}\right\} = u_0$,
- INVARIANCE PROPERTY: as a functional of η, the set $C^*_{u_0}$ is invariant to 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*

$$L^*_{u_0} \doteq L\big(g^*_{u_0}\big) \leq L(g) .$$

Furthermore, we have

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

and

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

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 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 2 (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 (2005) (see also Scott and Nowak, 2005).

Remark 3 (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 (Cavalier, 1997) (see also Tsybakov, 1997 and Willett and Nowak, 2006) as well as in the learning literature, for instance in the context of anomaly detection (Steinwart et al., 2005; Scott and Davenport, 2006, to appear; Vert and Vert, 2006). In our framework of classification with mass constraint, the threshold defining the level set involves the quantile of the random variable $\eta(X)$.*

Remark 4 (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 Scott and Nowak (2006). 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 Scott and Nowak (2006) 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 $\widehat{Q}(s,1-u_0)$ which corresponds to the empirical quantile. We will thus consider, instead of $L_n(s)$, the *empirical error*:

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

Note that $\widehat{L}_n(s)$ is a complicated statistic since the empirical quantile involves all the instances $X_1,\ldots,X_n$. We also mention that $\widehat{L}_n(s)$ 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 $t\in\mathbb{R}$:

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

The functions F_s (respectively G_s, H_s) denote the cumulative distribution function (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 $v \in (0,1)$:

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

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 5 *We assume that*

(i) the class $\mathcal{S}$ is symmetric (that is, 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.

The following remarks provide some insights on conditions (i) and (ii) of the theorem.

Remark 6 (ON THE COMPLEXITY ASSUMPTION) *On the terminology of major sets and major classes, we refer to Dudley (1999). 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 7 (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, for instance, Papoulis, 1965):*

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

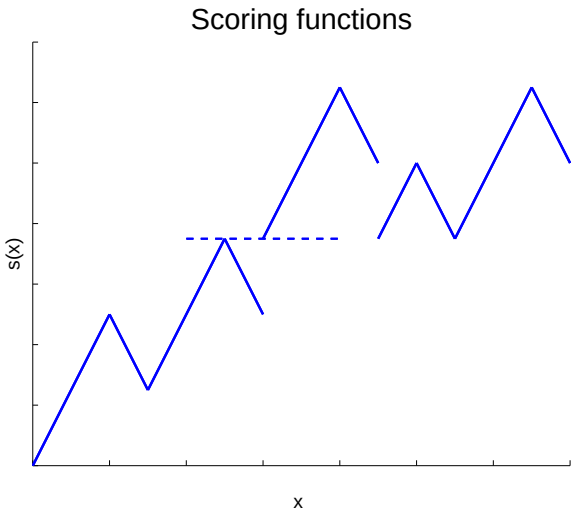

Figure 1: Typical example of a scoring function.

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 $s \in \mathcal{S}$.*

PROOF. Set $v_0 = 1-u_0$. By a standard argument (see, for instance, Devroye et al., 1996), 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, for instance, Lugosi), 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'.

The novel part of the analysis lies in the control of the first term and we now show how to handle it. Following the work of Koul (2002), we set the following notations:

$$M(s,v) = \mathbb{P}\left\{Y \cdot \left(s(X) - Q(s,v)\right) < 0\right\},$$

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

and note that $U_n(s,v)$ is centered. In particular, we have:

$$L_n(s) = U_n(s,v_0) + M(s,v_0)\ .$$

As $Q(s,v) = F_s^{-1}(v)$, we have $Q(s, F_s \circ \hat{F}_s^{-1}(v)) = \hat{F}_s^{-1}(v) = \hat{Q}(s,v)$ and then

$$\hat{L}_n(s) = U_n(s, F_s \circ \hat{F}_s^{-1}(v_0)) + M(s, F_s \circ \hat{F}_s^{-1}(v_0))\ .$$

Note that $M(s, F_s \circ \hat{F}_s^{-1}(v_0)) = \mathbb{P}\left\{Y \cdot \big(s(X) - \hat{Q}(s,v)\big) < 0 \mid D_n\right\}$ where D_n denotes the sample $(X_1,Y_1),\cdots,(X_n,Y_n)$.

We then have the following decomposition, for any $s \in \mathcal{S}$ and $v_0 \in (0,1)$:

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

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

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

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

$$\begin{aligned}\left|F_s \circ \hat{F}_s^{-1}(v_0) - v_0\right| &\leq \left|F_s \circ \hat{F}_s^{-1}(v_0) - \hat{F}_s \circ \hat{F}_s^{-1}(v_0)\right| + \left|\hat{F}_s \circ \hat{F}_s^{-1}(v_0) - v_0\right| \\ &\leq \sup_{t \in (0,\lambda)} \left|F_s(t) - \hat{F}_s(t)\right| + \frac{1}{n}\ .\end{aligned}$$

Here again, we can use assumption (i) and a classical VC bound from Lugosi 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:

$$\left|U_n(s, F_s \circ \hat{F}_s^{-1}(v_0)) - U_n(s,v_0)\right| \leq \sup_{v \in (0,1)} |U_n(s,v) - U_n(s,v_0)| \leq 2 \sup_{v \in (0,1)} |U_n(s,v)|.$$

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

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

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 Mammen and Tsybakov, 1999; Tsybakov, 2004; Massart and Nédélec, 2006) 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:

- the class $\mathcal{S}$ of scoring functions is a finite class with N elements,
- 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 $D > 0$ such that, for all $t \geq 0$,

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

(iv) the family $\mathcal{K} = \{ G_s, H_s \, : \, s \in \mathcal{S} \}$ of cdfs satisfies the following property: for any $s \in \mathcal{S}$, G_s and H_s are twice differentiable at $Q(s, 1-u_0) = F_s^{-1}(1-u_0)$.

Note that condition (iii) simply extends the standard low noise assumption introduced by Tsybakov (2004) (see also Boucheron et al., 2005, for an account on this) where the level 1/2 is replaced by the $(1-u_0)$-quantile of $\eta(X)$. Condition (iv) is a technical requirement needed in order to derive an approximation of the statistics involved in empirical risk minimization.

Remark 8 (CONSEQUENCE OF CONDITION (III)) *We recall here the various equivalent formulations of condition (iii) as they are described in Section 5.2 from the survey paper by Boucheron et al. (2005). A slight variation in our setup is due to the presence of the quantile $Q(\eta, 1-u_0)$ but we can easily adapt the corresponding conditions. Hence, we have, under condition (iii), the variance control, for any $s \in \mathcal{S}$:*

$$\mathrm{Var}(\mathbb{I}\{Y \neq 2\mathbb{I}_{C_s}(X) - 1\} - \mathbb{I}\{Y \neq 2\mathbb{I}_{C^*_{u_0}}(X) - 1\}) \leq c \, (L(s) - L^*_{u_0})^{\alpha} ,$$

or, equivalently,

$$\mathbb{E}\big(\mathbb{I}_{C_s \Delta C^*_{u_0}}(X)\big) \leq c \, (L(s) - L^*_{u_0})^{\alpha} .$$

Recall that $L_n(s) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{Y_i \cdot (s(X_i) - Q(s, 1-u_0)) < 0\}$. We point out that $L_n(s)$ is not an empirical criterion since the quantile $Q(s, 1-u_0)$ depends on the distribution. However, we can introduce the minimizer of this functional:

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

for which we can use the same argument as in the classification setup. We then have, by a standard argument based on Bernstein's inequality (which will be provided for completeness in the proof of Theorem 10 below), 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. We will show below how to obtain a similar rate when the true quantile $Q(s,1-u_0)$ is replaced by the empirical quantile $\widehat{Q}(s,1-u_0)$ in the criterion to be minimized.

We point out that conditions (ii) and (iii) are not completely independent. We offer the following proposition which will be useful in the sequel.

Proposition 9 *If (G_η, H_η) belongs to the class $\mathcal{K}$ fulfilling condition (ii), then F_η is Lipschitz and condition (iii) is satisfied with $\alpha = 1/2$.*

PROOF. We recall that $F_\eta = pG_\eta + (1-p)H_\eta$ and assume for simplicity that G_η and H_η are differentiable. By condition (ii), we then have $|F'_\eta| = p|G'_\eta| + (1-p)|H'_\eta| \leq pB + (1-p)B = B$. Set $q = Q(\eta, 1-u_0)$. Then, by the mean value theorem, there exists a constant c such that, for all $t \geq 0$:

$$\mathbb{P}\{|\eta(X) - q| \leq t\} = F_\eta(t+q) - F_\eta(-t+q) \leq B(t+q-(-t+q)) = 2Bt\ .$$

We have proved that condition (iii) is fulfilled with $D = 2B$ and $\alpha = 1/2$. ■

The novel part of the analysis below lies in the control of the bias induced by plugging the empirical quantile $\widehat{Q}(s,1-u_0)$ in the risk functional. The next theorem shows that faster rates of convergence up to the order of $n^{-2/3}$ can be obtained under the previous assumptions.

Theorem 10 *We assume that the class $\mathcal{S}$ 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. We recall that $\widehat{s}_n = \arg\min_{s\in\mathcal{S}} \widehat{L}_n(s)$. Then, for any $\delta > 0$, we have, with probability $1-\delta$:*

$$L(\widehat{s}_n) - L^*_{u_0} \leq c\left(\frac{log(N/\delta)}{n}\right)^{\frac{2}{3}},$$

for some constant c.

Remark 11 (ON THE RATE $n^{-2/3}$) *This result highlights the fact that rates faster than the one obtained in Theorem 5 can be obtained in this setup with additional regularity assumptions. However, it is noteworthy that the standard low noise assumption (iii) is already contained, by Proposition 9, in assumption (ii) which is required in proving the typical $n^{-1/2}$ rate. The consequence of this observation is that there is no hope of getting rates up to n^{-1} unless assumption (ii) is weakened.*

Remark 12 (ON THE ASSUMPTION $s^* \in \mathcal{S}$) *This assumption is not important and can be removed. For a neat argument, check the proof of Theorem 5 from Clémençon et al. (To appear) which uses a result by Massart (2006).*

The proof of the previous theorem is based on two arguments: the structure of *linear signed rank statistics* and the variance control assumption. The situation is similar to the one we encountered in Clémençon et al. (To appear) where we were dealing with U-statistics and we had to invoke Hoeffding's decomposition in order to grasp the behavior of the underlying U-processes. Here we require a similar argument to describe the structure of the empirical risk functional $\widehat{L}_n(s)$ under study. This statistic can be interpreted as a linear signed rank statistic and the key decomposition has been used in the context of nonparametric hypotheses testing and R-estimation. We mainly

refer to Hájek and Sidák (1967), Dupac and Hájek (1969), Koul (1970), and Koul and R.G. Staudte (1972) for an account on rank statistics.

We now prepare for the proof by stating the main ideas in the next propositions, but first we need to introduce some notations. Set:

$$\forall v \in [0,1]\ , \quad K(s,v) = \mathbb{E}\left(Y \mathbb{I}\{s(X) \leq Q(s,v)\}\right) = pG_s(Q(s,v)) - (1-p)H_s(Q(s,v)),$$
$$\hat{K}_n(s,v) = \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{I}\{s(X_i) \leq \widehat{Q}(s,v)\}\ .$$

Then we can write:

$$L(s) = 1 - p + K(s, 1-u_0),$$
$$\hat{L}_n(s) = \frac{n_-}{n} + \hat{K}_n(s, 1-u_0)\ ,$$

where $n_- = \sum_{i=1}^{n} \mathbb{I}\{Y_i = -1\}$.

We note that the statistic $\hat{L}_n(s)$ is related to linear signed rank statistics.

Definition 13 (Linear signed rank statistic) *Consider $Z_1, \ldots, Z_n$ an i.i.d. sample with distribution F and a real-valued score generating function Φ. Denote by $R_i^+ = rank(|Z_i|)$ the rank of $|Z_i|$ in the sample $|Z_1|, \ldots, |Z_n|$. Then the statistic*

$$\sum_{i=1}^{n} \Phi\left(\frac{R_i^+}{n+1}\right)\ \mathrm{sgn}(Z_i)$$

is a linear signed rank statistic.

Proposition 14 *For fixed s and v, the statistic $\hat{K}_n(s,v)$ is a linear signed rank statistic.*

PROOF. Take $Z_i = Y_i s(X_i)$. The random variables Z_i have their absolute value distributed according to F_s and have the same sign as Y_i. It is easy to see that the statistic $\hat{K}_n(s,v)$ is a linear signed rank statistic with score generating function $\Phi(x) = \mathbb{I}_{[x \leq v]}$. ■

A decomposition of Hoeffding's type for such statistics can be formulated. Set first:

$$Z_n(s,v) = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - K'(s,v)\right) \mathbb{I}\{s(X_i) \leq Q(s,v)\} - K(s,v) + vK'(s,v)\ ,$$

where $K'(s,v)$ denotes the derivative of the function $v \mapsto K(s,v)$. Note that $Z_n(s,v)$ is a centered random variable with variance:

$$\sigma^2(s,v) = v - K(s,v)^2 + v(1-v)K'^2(s,v) - 2(1-v)K'(s,v)K(s,v)\ .$$

The next result is due to Koul (1970) and we provide an alternate proof in the Appendix.

Proposition 15 *Assume that condition (iv) holds. We have, for all $s \in \mathcal{S}$ and $v \in [0,1]$:*

$$\hat{K}_n(s,v) = K(s,v) + Z_n(s,v) + \Lambda_n(s) \ .$$

with

$$\Lambda_n(s) = O_{\mathbb{P}}(n^{-1}) \ \ as \ \ n \to \infty \ .$$

This asymptotic expansion highlights the structure of the statistic $\hat{L}_n(s)$ for fixed s:

$$\hat{L}_n(s) = \frac{n_-}{n} + K(s,1-u_0) + Z_n(s,1-u_0) + \Lambda_n(s) \ .$$

Once centered, the leading term $Z_n(s,1-u_0)$ is an empirical average of i.i.d. random variables (of a stochastic order of $n^{-1/2}$) and the remainder term $\Lambda_n(s)$ is of a stochastic order of n^{-1}. The nature of the decomposition of $\hat{L}_n(s)$ is certainly unexpected because the leading term contains an additional derivative term given by $K'(s,1-u_0)\,(v - \mathbb{I}\{s(X_i) \leq Q(s,1-u_0)\})$. The revelation of this fact is one of the major contributions in the work of Koul (2002).

Now, in order to establish consistency and rates-of-convergence-type results, we need to focus only on the leading term which carries most of the statistical information, while the remainder needs to be controlled uniformly over the candidate class $\mathcal{S}$. As a consequence, the variance control assumption will only concern the variance of the kernel h_s involved in the empirical average $Z_n(s,1-u_0)$ and defined as follows:

$$h_s(X_i,Y_i) = \big(Y_i - K'(s,v)\big)\,\mathbb{I}\{s(X_i) \leq Q(s,v)\} - K(s,v) + vK'(s,v) \ ,$$

We then have

$$Z_n(s,v) - Z_n(s^*,v) = \frac{1}{n}\sum_{i=1}^{n} \big(h_s(X_i,Y_i) - h_{s^*}(X_i,Y_i)\big) \ .$$

Proposition 16 *Fix $v \in [0,1]$. Assume that condition (iii) holds. Then, we have, for all $s \in \mathcal{S}$:*

$$\mathrm{Var}\big(h_s(X_i,Y_i) - h_{s^*}(X_i,Y_i)\big) \leq c\big(L(s) - L(s^*)\big)^{\alpha} \ ,$$

for some constant c.

PROOF. We first write that:

$$h_s(X_i,Y_i) - h_{s^*}(X_i,Y_i) = I + II + III + IV + V$$

where

$$\begin{aligned}
I &= Y_i\,\big(\mathbb{I}\{s(X_i) \leq Q(s,v)\} - \mathbb{I}\{s^*(X_i) \leq Q(s^*,v)\}\big), \\
II &= (K'(s^*,v) - K'(s,v))\,\mathbb{I}\{s^*(X_i) \leq Q(s^*,v)\}, \\
III &= K'(s,v)\,\big(\mathbb{I}\{s^*(X_i) \leq Q(s^*,v)\} - \mathbb{I}\{s(X_i) \leq Q(s,v)\}\big), \\
IV &= K(s^*,v) - K(s,v), \\
V &= v\,(K'(s,v) - K'(s^*,v)) \ .
\end{aligned}$$

By Cauchy-Schwarz inequality, we only need to show that the expected value of the square of these quantities is smaller than $(L(s)-L^*)^\alpha$ up to some multiplicative constant.

Note that, by definition of K, we have:

$$\begin{aligned} II &= (L'(s^*,v)-L'(s,v))\,\mathbb{I}\{s^*(X_i)\leq Q(s^*,v)\}, \\ IV &= L(s^*)-L(s), \\ V &= v\,(L'(s,v)-L'(s^*,v)) \end{aligned}$$

where $L'(s,v)$ denotes the derivative of the function $v\mapsto L(s,v)$. It is clear that, for any s, we have $L(s,v)=L(s^*,v)$ implies that $L'(s,v)=L'(s^*,v)$ otherwise s^* would not be an optimal scoring function at some level v' in the vicinity of v. Therefore, since $\mathcal{S}$ is finite, there exists a constant c such that

$$(L'(s,v)-L'(s^*,v))^2\leq c(L(s)-L^*)^\alpha$$

and then $\mathbb{E}\big(II^2\big)$ and $\mathbb{E}\big(V^2\big)$ are bounded accordingly.

Moreover, we have:

$$\begin{aligned} \mathbb{E}\big(I^2\big) &\leq \mathbb{E}\big(\mathbb{I}_{C_s\Delta C_{s^*}}(X)\big) \\ &\leq c(L(s)-L(s^*))^\alpha \end{aligned}$$

for some positive constant c, by assumption (iii).

Eventually, by assumption (ii), we have that $K'(s,v)$ is uniformly bounded and thus, the term $\mathbb{E}\big(III^2\big)$ can be handled similarly. ■

Proof of Theorem 10. Set $v_0=1-u_0$. First notice that $\hat{s}_n=\arg\min_{s\in\mathcal{S}}\hat{K}_n(s,1-u_0)$. We then have

$$\begin{aligned} L(\hat{s}_n)-L(s^*) &= K(\hat{s}_n,v_0)-K(s^*,v_0) \\ &\leq \hat{K}_n(s^*,v_0)-\hat{K}_n(\hat{s}_n,v_0)-(K(s^*,v_0)-K(\hat{s}_n,v_0)) \\ &\leq Z_n(s^*,v_0)-Z_n(\hat{s}_n,v_0))+2\sup_{s\in\mathcal{S}}|\Lambda_n(s)| \end{aligned}$$

where we used the decomposition of the linear signed rank statistic from Proposition 15 to obtain the last inequality.

By Proposition 15, we know that the second term on the right hand side is of stochastic order n^{-1} since the class $\mathcal{S}$ is of finite cardinality. It remains to control the leading term $Z_n(s^*,v_0)-Z_n(\hat{s}_n,v_0)$. At this point, we will use the same argument as in Section 5.2 from Boucheron et al. (2005).

Denote by $C=\sup_{s,x,y}|h_s(x,y)|$ and by $\sigma^2(s)=\mathrm{Var}\big(h_s(X_i,Y_i)-h_{s^*}(X_i,Y_i)\big)$. By Bernstein's inequality for averages of upper bounded and centered random variables (see Devroye et al., 1996) and the union bound, we have, with probability $1-\delta$, for all $s\in\mathcal{S}$:

$$Z_n(s^*,v_0)-Z_n(s,v_0)\leq\sqrt{\frac{2\sigma^2(s)\log(N/\delta)}{n}}+\frac{2C\log(N/\delta)}{3n}$$

$$\leq \sqrt{\frac{2c(L(s)-L^*)^{\alpha}\ \log(N/\delta)}{n}} + \frac{2C\log(N/\delta)}{3n}$$

thanks to the variance control obtained in Proposition 16. Since this inequality holds for any s, it holds in particular for $s = \hat{s}_n$. Therefore, we have obtained the following result, with probability $1-\delta$:

$$L(\hat{s}_n) - L(s^*) \leq \sqrt{\frac{2c(L(\hat{s}_n)-L^*)^{\alpha}\ \log(N/\delta)}{n}} + \frac{2c'\log(N/\delta)}{3n}$$

for some constants c, c'. At the cost of increasing the multiplicative constant factor, we can get rid of the second term and solve the inequality in the quantity $L(\hat{s}_n) - L(s^*)$ to get

$$L(\hat{s}_n) - L(s^*) \leq c\left(\frac{\log(N/\delta)}{n}\right)^{\frac{1}{2-\alpha}}$$

for some constant c. To end the proof, we plug the value of $\alpha = 1/2$ following from Proposition 9. ∎

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. In the sequel, we build on the results from Section 2 and also on our previous work on the (global) ranking problem (Clémençon et al., To appear) 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 Receiver Operating Characteristic known as the ROC curve (van Trees, 1968; Egan, 1975). If we take the notations $\overline{G}_s(z) = \mathbb{P}\{s(X) > z \mid Y = 1\}$ (true positive rate) and $\overline{H}_s(z) = \mathbb{P}\{s(X) > z \mid Y = -1\}$ (false positive rate), we can define the ROC curve, for any scoring function s, as the plot of the function:

$$z \mapsto \left(\overline{H}_s(z), \overline{G}_s(z)\right)$$

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

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

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 2) and using Neyman-Pearson's Lemma, it is easy to check that the 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 to 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:

$$\alpha(s,u) = \mathbb{P}\{s(X) \geq Q(s,1-u) \mid Y = -1\} = \bar{H}_s \circ F_s^{-1}(1-u),$$
$$\beta(s,u) = \mathbb{P}\{s(X) \geq Q(s,1-u) \mid Y = +1\} = \bar{G}_s \circ F_s^{-1}(1-u) \ .$$

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 Hanley and McNeil, 1982). 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. In order to focus on the best instances, a natural idea is to truncate the AUC (as in the approach by Dodd and Pepe (2003)).

Definition 17 (Partial AUC) *We define the partial AUC for a scoring function s and a rate u_0 of best instances as:*

$$\text{PARTAUC}(s,u_0) = \int_0^{\alpha(s,u_0)} \beta(s,\alpha)\, d\alpha \ .$$

We conjecture that such a criterion is not appropriate for local ranking. If it was, then we should have: $\forall s, \text{PARTAUC}(s,u_0) \leq \text{PARTAUC}(\eta,u_0)$, since the function η would provide the optimal ranking. However, there is strong evidence that this is not true as shown by a simple geometric argument which we describe below.

In order to represent the partial AUC of a scoring function s, we need to locate the cut-off point given the constraint on the rate u_0 of best instances. 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 $D(u,p)$ and call it the *control line* when $u = u_0$. 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 with the control line $D(u_0,p)$. The partial AUC is then the area under this part of the ROC curve (it corresponds to the shaded area in the left display of Figure 2).

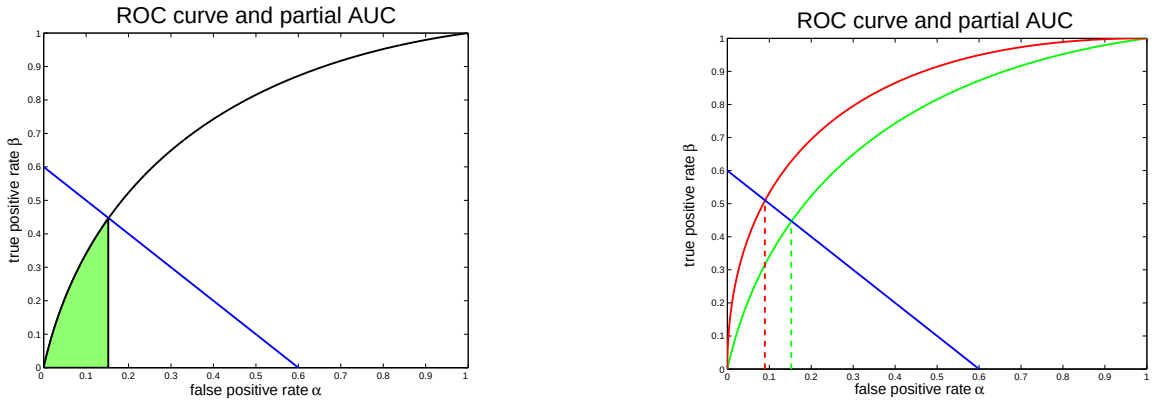

Figure 2: ROC curves, control line $D(u_0, p)$ and partial AUC at rate u_0 of best instances.

The optimality of η with respect to the partial AUC can then be questioned. Indeed, 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) so that the overall impact on the integral is not clear. Let us now put things formally in the following lemma.

Lemma 18 *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:

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

Observe that, for any scoring function s,

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

We thus have

$$\begin{aligned}p\left(G_s(Q(s,1-u)) - G_\eta(Q(\eta,1-u)\right) &= \mathbb{E}\left(\eta(X)(\mathbb{I}\{X \in C^*_u\} - \mathbb{I}\{X \in C_{s,u}\})\right)\\ &= \mathbb{E}\left(\eta(X)\mathbb{I}\{X \notin C^*_u\}(\mathbb{I}\{X \in C^*_u\} - \mathbb{I}\{X \in C_{s,u}\})\right)\\ &\quad + \mathbb{E}\left(\eta(X)\mathbb{I}\{X \in C^*_u\}(\mathbb{I}\{X \in C^*_u\} - \mathbb{I}\{X \in C_{s,u}\})\right)\\ &\geq -\mathbb{E}\left(Q(\eta,1-u)\mathbb{I}\{X \notin C^*_u\}\,\mathbb{I}\{X \in C_{s,u}\}\right)\\ &\quad + \mathbb{E}\left(Q(\eta,1-u)\mathbb{I}\{X \in C^*_u\}(1-\mathbb{I}\{X \in C_{s,u}\})\right)\end{aligned}$$

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

The second inequality simply follows from the identity below:

$$1-u = pG_s(Q(s,1-u)) + (1-p)H_s(Q(s,1-u)) \; .$$

■

The previous lemma will be important when describing the optimal rules for local ranking criteria. But, at this point, we still do not know any nice criterion for the problem of ranking the best instances. Before considering different heuristics for extending the AUC criterion in the next subsections, we will proceed backwards and define our target, that is to say, the optimal scoring functions for our problem.

Definition 19 (Class $\mathcal{S}^*$ of optimal scoring functions) *The optimal scoring functions for ranking the best instances at the rate u_0 are defined as the members of the equivalence class (functions defined up to the composition with a nondecreasing transformation) of scoring functions s^* such that:*

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

Such scoring functions fulfill the two properties of locating the best instances (indeed $C_{s^*,u_0} = C^*_{u_0}$) and ranking them as well as the regression function.

Under the light of Lemma 18, we will see that 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 Clémençon et al. (To appear), 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 $C \subset \mathcal{X}$:

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

On sets of the form $C = C_{s,u} = \{x \in \mathcal{X} \mid s(x) \geq Q(s, 1-u)\}$ with mass equal to u, the local ranking error will be denoted by $R(s,u) \stackrel{\circ}{=} R(s, C_{s,u})$.

We will also consider the local analogue of the AUC criterion:

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

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 $s \mapsto R(s,u)$. 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 optimal scoring functions s^* in the set $\mathcal{S}^*$ maximize the LOCAUC criterion and that the latter may be decomposed as a sum of a 'power' term and (the opposite of) a local ranking error term.

Theorem 20 *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. We first introduce the notation for the Lebesgue-Stieltjes integral. Whenever φ is a cdf on $\mathbb{R}$ and ψ is integrable, the integral $\int \psi(z)\, d\varphi(z)$ denotes the Lebesgue-Stieltjes integral (integration with respect to the measure ν defined by $\nu[a,b) = \varphi(b) - \varphi(a)$ for any real numbers $a < b$). If φ has a density with respect to the Lebesgue measure, then the integral can be written as a Lebesgue integral: $\int \psi(z)\, d\varphi(z) = \int \psi(z)\varphi'(z)\, dz$. We shall use this convention repeatedly in the sequel. In particular, if Z is a random variable with cdf given by F_Z then we can write: $\mathbb{E}(Z) = \int z\ dF_Z(z)$. Now set $v_0 = 1 - u_0$. Observe first that, by conditioning on X, we have:

$$\begin{aligned}
\text{LOCAUC}(s,u_0) &= \mathbb{E}\left(\mathbb{I}\{s(X) > s(X')\}\ \mathbb{I}\{s(X) \geq Q(s,v_0)\}\ \mid Y = 1, Y' = -1\right) \\
&= \mathbb{E}\left(\ \mathbb{I}\{s(X) \geq Q(s,v_0)\}\, \mathbb{E}\left(\ \mathbb{I}\{s(X) > s(X')\} \mid Y' = -1, X\ \right)\ \mid Y = 1\ \right) \\
&= \mathbb{E}\left(H_s(s(X))\ \mathbb{I}\{s(X) \geq Q(s,v_0)\}\ \mid Y = 1\ \right) \\
&= \int_{Q(s,v_0)}^{+\infty} H_s(z)\ dG_s(z) .
\end{aligned}$$

The last equality is obtained by using the fact that, conditionally on $Y = 1$, the random variable $s(X)$ has cdf G_s. We now use that $pG_s = F_s - (1-p)H_s$ and we obtain:

$$p\,\text{LOCAUC}(s,u_0) = \int_{Q(s,v_0)}^{+\infty} H_s(z)\ dF_s(z) - (1-p) \int_{Q(s,v_0)}^{+\infty} H_s(z)\ dH_s(z).$$

Recall now that $\alpha(s,v) = \bar{H}_s \circ F_s^{-1}(1-v)$ and make the change of variable $1-v = F_s(z)$

$$\int_{Q(s,v_0)}^{+\infty} H_s(z)\, dF_s(z) = \int_0^{v_0} (1-\alpha(s,v))\, dv\ .$$

The second term is computed by making the change of variable $a = H_s(z)$ which leads to:

$$\int_{Q(s,v_0)}^{+\infty} H_s(z)\, dH_s(z) = \int_{1-\alpha(s,u_0)}^{1} a\, da\ .$$

We have obtained:

$$p\text{LOCAUC}(s,u_0) = \int_0^{v_0} (1-\alpha(s,v))\, dv - \frac{1-p}{2}\left(1-(1-\alpha(s,v_0))^2\right)\ .$$

From Lemma 18, we have that, for any $u \in (0,1)$, the functional $s \mapsto \alpha(s,u))$ is minimized for $s = \eta$. Hence, the first part of Theorem 20 is established.

Besides, integrating by parts, we get:

$$\int_{Q(s,v_0)}^{+\infty} H_s(z)\, dG_s(z) = [H_s(z)G_s(z)]_{Q(s,v_0)}^{+\infty} - \int_{Q(s,v_0)}^{+\infty} G_s(z)\, dH_s(z).$$

The same change of variables as before leads to:

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

We then have another expression of the $\text{LOCAUC}(s,u_0)$:

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

We develop further by expressing the product of α and β in terms of probability. Using the independence of (X,Y) and (X',Y'), we obtain:

$$\begin{aligned}
\alpha(s,u_0)\beta(s,u_0) &= \frac{1}{p(1-p)}\mathbb{P}\left\{s(X) \wedge s(X') > Q(s,v_0),\ Y=1,\ Y'=-1\right\} \\
&= \mathbb{P}\left\{s(X) > s(X'),\ s(X) \wedge s(X) > Q(s,v_0)\ \mid\ Y=1,\ Y'=-1\right\} \\
&\quad + \frac{1}{p(1-p)}\mathbb{P}\left\{s(X) < s(X'),\ (X,X') \in C^2_{s,u_0},\ Y=1,\ Y'=-1\right\} \\
&= \int_0^{\alpha(s,u_0)} \beta(s,\alpha)\, d\alpha + \frac{1}{2p(1-p)} R(s,u_0)\ .
\end{aligned}$$

Combining this with the previous formula leads to the second statement of the theorem. ■

Remark 21 (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 partial* AUC. *As a matter of fact, we have:*

$$\forall s\,, \quad \mathrm{LOCAUC}(s,u_0) = \mathrm{PARTAUC}(s,u_0) + \beta(s,u_0) - \alpha(s,u_0)\beta(s,u_0)\,.$$

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 control line* $D(u_0,p)$. *The theorem reveals that evaluating the local performance of a scoring statistic* $s(X)$ *by the truncated AUC as proposed in Dodd and Pepe (2003) is highly arguable since the maximizer of the functional* $s \mapsto \mathrm{PARTAUC}(s,u_0)$ *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 $n_+ = n - n_- = \sum_{i=1}^n \mathbb{I}\{Y_i = +1\}$.

It will be useful to have in mind the definition of a linear rank statistic.

Definition 22 (linear rank statistic) *Consider* $Z_1,\ldots,Z_n$ *an i.i.d. sample with distribution* F *and a real-valued score generating function* Φ. *Denote by* $R_i = rank(Z_i)$ *the rank of* Z_i *in the sample* $Z_1,\ldots,Z_n$. *Then the statistic*

$$\sum_{i=1}^n c_i \Phi\left(\frac{R_i}{n+1}\right)$$

is a linear rank statistic.

We refer to Hájek and Sidák (1967) and van de Vaart (1998) for basic results related to linear rank statistics. In particular, we recall that, for fixed s, the Wilcoxon statistic $T_n(s)$ is a linear rank statistic for the sample $s(X_1),\ldots,s(X_n)$, with random weights $c_i = \mathbb{I}\{Y_i = 1\}$, score generating function $\Phi(v) = v$:

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

where $\mathrm{rank}(s(X_i))$ denotes the rank of $s(X_i)$ in the sample $\{s(X_j), 1 \le j \le n\}$. 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

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

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, we introduce the following quantity:

Definition 23 (W-ranking performance measure) *Consider the criterion related to the score generating function $\Phi_{u_0}(v) = v\,\mathbb{I}\{v > 1-u_0\}$:*

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

*It will be called the W-*ranking performance measure at rate u_0.

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

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

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

Theorem 24 *We have, for all s:*

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

Furthermore, we have:

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

PROOF. We start by the definition of W:

$$W(s,u_0) = \mathbb{E}\left(F_s(s(X))\mathbb{I}\{F_s(s(X)) > 1-u_0\}\,|\,Y=1\right)$$

$$= \int_{Q(s,1-u_0)}^{+\infty} F_s(z)\,dG_s(z).$$

We recall that: $F_s = pG_s + (1-p)H_s$ which leads to:

$$W(s,u_0) = p\int_{Q(s,1-u_0)}^{+\infty} G_s(z)\,dG_s(z) + (1-p)\int_{Q(s,1-u_0)}^{+\infty} H_s(z)\,dG_s(z)\ .$$

The second term corresponds exactly to the LOCAUC. The first term is easily computed by a change of variable $b = G_s(z)$:

$$\int_{Q(s,1-u_0)}^{+\infty} G_s(z)\,dG_s(z) = \int_{1-\beta(s,u_0)}^{1} b\,db\ .$$

Elementary computations lead to the formula in the theorem. Moreover the application $t \mapsto t(2-t)$ being nondecreasing for $t \in (0,1)$, we have, from Lemma 18:

$$\forall s^* \in \mathcal{S}^*, \quad \beta(s,u_0)(2-\beta(s,u_0)) \leq \beta(s^*,u_0)(2-\beta(s^*,u_0))\ .$$

We also use the optimality of s^* for LOCAUC established in Theorem 20 to conclude the proof. ■

Remark 25 (EVIDENCE AGAINST 'TWO-STEP' 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 20 and 24). But, in general, this remark should prevent from considering 'naive' two-step strategies for solving the local ranking problem. By 'naive' two-step strategies, we refer here to stagewise strategies which would, first, compute an estimate $\widehat{C}$ of the set containing the best instances, and then, solve the ranking problem over $\widehat{C}$ as described in Clémençon et al. (To appear). 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.*

Remark 26 (OTHER RANKING PERFORMANCE MEASURES) *The ideas expressed above suggest that several ranking criteria can be proposed. For instance, one can consider maximization of other linear rank statistics with particular score generating functions Φ and there are many possible choices which would emphasize the importance of the highest ranks. One of these choices is $\Phi(v) = v^p$ which corresponds to the p-norm push proposed by Rudin (2006) although the definition of the ranks in her work is slightly different. The Discounted Cumulative Gain criterion, studied in particular by Cossock and Zhang (2006) and Li et al. (2007), is of different nature and cannot be represented in a similar way. Other extensions can be proposed in the spirit of the tail strength measure from Taylor and Tibshirani (2006). The theoretical study of such criteria is still at an early stage, especially for the last proposal. We also point out that with such extensions, probabilistic interpretations and explicit connection to the AUC criterion seem to be lost.*

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 20 and 24, we can easily get:

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

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:

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

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

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

Note that $\hat{R}_n(s)$ 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:

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

Proposition 27 *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:*

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

with probability larger than $1 - \delta$.

PROOF. (SKETCH) The proof combines the argument used in the proof of Theorem 5 with the techniques used in establishing Proposition 2 in Clémençon et al. (2005).

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

The middle term may be bounded by applying the result stated in Theorem 5, 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 Clémençon et al. (2005) 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 5. 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 Clémençon et al. (2005). ■

Remark 28 (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 Clémençon et al. (2005) 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 Clémençon et al. (2005), 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* $O(n^{-1/2})$.

Remark 29 (ABOUT THE EXTENSION TO CONVEX RISK MINIMIZATION) *An important topic in classification theory is convex risk minimization. Understanding the connection between classification error and its convex surrogates has permitted to understand the behavior of practical algorithms such as boosting and* SVM *from a statistical perspective (see Boucheron et al., 2005 for an account on this aspect and Bartlett et al., 2006 for state-of-the-art results). A natural question which arises here is whether the consistency results on local ranking can be extended in this spirit. Note that, if we do not focus on best instances and consider the whole* AUC *as a performance criterion, it is straightforward to obtain consistency and universal rates of convergence for convex risk minimization (as explained in Clémençon et al. (To appear)). In the case of local ranking as we introduced it, this extension is less straightforward since the decision rule represented here by the scoring function* s *appears under the empirical quantile* $\hat{Q}(s, v)$ *in the criterion. We refer to Rudin (2006), Cossock and Zhang (2006) and Li et al. (2007) where convex risk minimization strategies in the context of ranking are discussed.*

5. Conclusion

In the present work, we have presented theoretical work on local ranking. In the first part of the paper (Section 2), we considered a subproblem that we called the *classification with mass constraint* problem. The scope was to establish and study an empirical risk minimization strategy for only *finding*, and not ranking, the best instances. In this case, one attempts to minimize the classification error over classifiers that contain a fixed proportion u of observations. This constraint leads to empirical risk functionals which involve an empirical quantile indexed by the class of candidate scoring functions and can be seen as linear signed rank statistics. We then provide a consistency result and discuss the noise assumptions required to derive fast rates of convergence in this setup. These assumptions require a limited regularity of the underlying distributions which prevents the fast rate from dropping below the order of $n^{-2/3}$. The second part of the paper (Section 3) is dedicated to the introduction of new performance measures for local ranking related to the ROC curve and the AUC criterion. We show that the AUC can be extended in several ways (partial AUC and local AUC) but not all these extensions are tailored for the local ranking problem. In particular

the naive extension known as the partial AUC is not appropriate and requires a correction term. We also introduce the *optimal scoring functions* which should be considered as the target of any local ranking method. We also discuss other extensions based on Wilcoxon statistics, the W-ranking performance measure, for which optimal rules can also be recovered. In the last section of the paper (Section 4), the problem of ranking the best instances is studied from a statistical perspective. A consistency result is provided for empirical risk minimization of the W-ranking performance measure.

Acknowledgments

We thank Stéphane Boucheron and Gábor Lugosi for their helpful remarks and encouragements. We also gratefully thank both referees for their comments which helped us improve the clarity of the paper.

Appendix A.

In this section, we provide the proof of Proposition 15.

PROOF. First, for all $(s,v) \in \mathcal{S} \times (0,1)$ set

$$V_n(s,v) = \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{I}\{s(X_i) \leq Q(s,v)\} - K(s,v) \ .$$

We have the following decomposition:

$$\forall v \in [0,1] \ , \quad \hat{K}_n(s,v) - K(s,v) = V_n(s, F_s \circ \hat{F}_s^{-1}(v)) + K(s, F_s \circ \hat{F}_s^{-1}(v)) - K(s,v) \ .$$

We shall first prove that

$$V_n(s, F_s \circ \hat{F}_s^{-1}(v_0)) = V_n(s,v_0) + O_{\mathbb{P}}(n^{-1}).$$

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

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

We bound the right hand side for fixed ε, by making use of an argument from van de Geer (2000). First, we need to put things into the right format. Set:

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

where $u_i(s,v) = Y_i \mathbb{I}\{s(X_i) \leq Q(s,v) < 0\} - \mathbb{E}\left(Y \mathbb{I}\{s(X) \leq Q(s,v)\}\right)$ 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:

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

and observe that

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

We then have, by applying Lemma 8.5 from van de Geer (2000), for $nt^2/\hat{R}^2(\varepsilon)$ sufficiently large,

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

for some positive constants c and C. It remains to integrate out and, for this purpose, we introduce the event:

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

We then have:

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

Now, we have, by Bernstein's inequality:

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

where we have used the notation $B(n,2\varepsilon)$ for a binomial $(n,2\varepsilon)$ random variable. We can take $x = O(t/\sqrt{\varepsilon})$ and assume also $x = o(\varepsilon)$ to get, for nt^2/ε^2 large enough,

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

for some positive constants c and C. This can be reformulated, by writing that the following bound holds, with probability larger than $1-\delta/2$,

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

We recall that, by the triangle inequality and Dvoretsky-Kiefer-Wolfowitz theorem, if we take $\varepsilon = c\,\sqrt{\frac{\log(2/\delta)}{n}}$, we have $\mathbb{P}\{A(s,\varepsilon)\} \geq 1-\delta/2$. It follows that, with probability larger than $1-\delta$, we have, for some constant κ:

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

for any $s \in \mathcal{S}$. Now it remains to deal with the second term $K(s, F_s \circ \hat{F}_s^{-1}(v_0)) - K(s, v_0)$. Therefore, by the differentiability assumption (iv), we have: $\forall s \in \mathcal{S}$,

$$\sup_{|v-v_0| \leq \delta} \{K(s,v) - K(s,v_0) - (v-v_0)K'(s,v_0)\} = O(\delta^2) \ , \quad \text{as } \delta \to 0 \ .$$

Since $|F_s \circ \hat{F}_s^{-1}(v_0)) - v_0| = O_{\mathbb{P}}(n^{-1/2})$, we get that

$$K(s, F_s \circ \hat{F}_s^{-1}(v_0)) - K(s,v_0) = K'(s,v_0)(F_s \circ \hat{F}_s^{-1}(v_0) - v_0) + O_{\mathbb{P}}(n^{-1}) \ , \quad \text{as } n \to \infty \ .$$

Moreover, as

$$F_s \circ \hat{F}_s^{-1}(v_0) - v_0 = -(\hat{F}_s \circ F_s^{-1}(v_0) - v_0) + O_{\mathbb{P}}(n^{-1}) \ ,$$

we finally obtain that

$$K(s, F_s \circ \hat{F}_s^{-1}(v_0)) - K(s,v_0) = -K'(s,v_0)(\hat{F}_s \circ F_s^{-1}(v_0) - v_0) + O_{\mathbb{P}}(n^{-1}) \ .$$

■

References

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.

P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

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

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.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. *The Annals of Statistics*, To appear.

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, 2004.

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

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

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

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

V. Dupac and J. Hájek. Asymptotic normality of simple linear rank statistics under alternatives ii. *The Annals of Mathematical Statistics*, (6):1992–2017, 1969.

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

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.

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

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

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.

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

H.L. Koul. Some convergence theorems for ranks and weighted empirical cumulatives. *The Annals of Mathematical Statistics*, (41):1768–1773, 1970.

H.L. Koul and Jr. R.G. Staudte. Weak convergence of weighted empirical cumulatives based on ranks. *The Annals of Mathematical Statistics*, (43):823–841, 1972.

P. Li, C. Burges, and Q. Wu. Learning to rank using classification and gradient boosting. Technical report MSR-TR-2007-74, Microsoft Research, 2007.

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

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

P. Massart. *Concentration Inequalities and Model Selection*. Lecture Notes in Mathematics. Springer, 2006.

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

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

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.

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.

C. Scott. Performance measures for Neyman-Pearson classification. Technical report, Department of Statistics, Rice University, 2005.

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

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

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

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

J. Taylor and R. Tibshirani. A tail strength measure for assessing the overall univariate significance in a dataset. *Biostatistics*, 7(2):167–181, 2006.

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

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

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

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

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

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.

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