ESAIM: PROCEEDINGS, January 2014, Vol. 44, p. 99-109
SMAI Groupe MAS – Journées MAS 2012 – Exposé plénier

APPLICATIONS OF CONCENTRATION INEQUALITIES FOR STATISTICAL SCORING AND RANKING PROBLEMS

Nicolas Vayatis

Abstract. In this overview paper, we consider the scoring approach applied to the ranking problem from a nonparametric perspective. We first focus on the problem of ROC curve optimization in terms of description of optimal elements. Then, we introduce summaries of this function-valued description of performance which are related to well-known statistics that are of higher order compared to averages of i.i.d. random variables. Eventually, we consider consistency and fast convergence results that rely on applications of concentration inequalities which involve U- and R-processes. This is a joint work with Stéphan Clémençon, Marine Depecker, Gábor Lugosi, and Sylvain Robbiano.

1. Motivations

As a starting point, and to emphasize the importance of ROC curves, we introduce two simple statistical problems: the bipartite ranking problem and the homogeneity testing problem.

1.1. Example 1 - Bipartite ranking problem

Consider data with binary feedback $\{(X_i, Y_i) \ : \ i = 1, \ldots, n\}$ i.i.d. observations in $\mathbb{R}^d \times \{-1, +1\}$. The purpose of bipartite ranking is to infer a preorder relation over $\mathbb{R}^d$ where the positive $(+1)$ instances *statistically* dominate the negative (-1) ones. The scoring approach to bipartite ranking consists in inferring a scoring rule $s : \mathbb{R}^d \to \mathbb{R}$ which affects the highest scores to the positive instances. The performance metric which completely characterizes the quality of a scoring rule s is known to be the ROC Curve:

$$\mathrm{ROC}(s, \cdot) \ : \ t \in \mathbb{R} \mapsto \big(\underbrace{P\{s(X) \geq t \mid Y = -1\}}_{\text{rate of false alarms}}, \ \underbrace{P\{s(X) \geq t \mid Y = +1\}}_{\text{rate of hits}} \big) \ .$$

Among the simple properties of the ROC curve, we point out the *invariance* property with respect to strictly increasing transforms of the scoring rule s. On Figure 1, we show examples of ROC curves for different scoring rules. The one corresponding to the green curve achieves perfect discrimination (ideal case), while the first diagonal corresponds to the worst possible discrimination between positive and negative instances. The black curve on the left display indicates better performance than the blue curve as it presents larger hit rate for the same false alarm rate. The right display highlights the fact that, when considering particular cases, it may happen that no scoring rule is *uniformly better* than another. However, the blue curve shows higher power for high scores (closer to the origin) and it would be the recommended choice for most applications. This discussion also introduces the idea of localizing the measure of performance that will be discussed below.

Article published online by EDP Sciences and available at http://www.esaim-proc.org or http://dx.doi.org/10.1051/proc/201444005

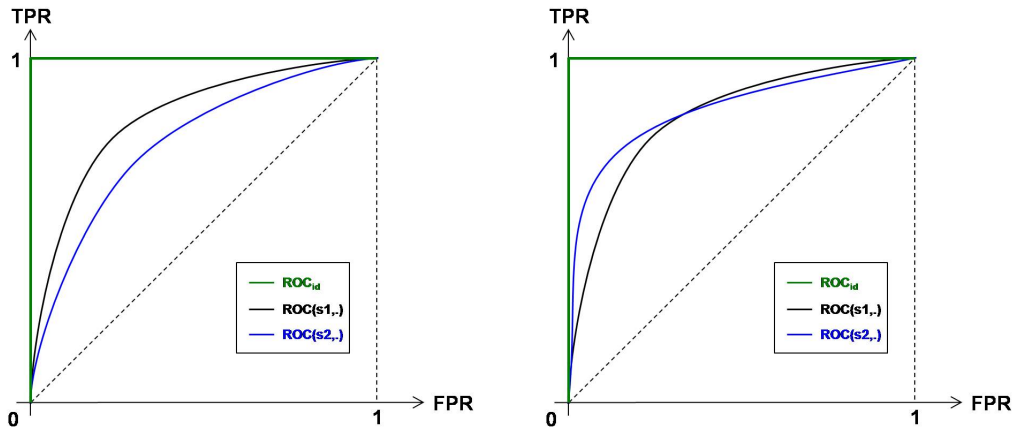

FIGURE 1. Examples of ROC curves.

1.2. **Example 2 - Two-sample homogeneity test in $\mathbb{R}^d$**

Consider now two independent samples of i.i.d. observations over $\mathbb{R}^d$: $\mathcal{X}_k = \{X_1, \ldots, X_k\}$ with distribution P_X, and $\mathcal{X}'_m = \{X'_1, \ldots, X'_m\}$ with distribution P'_X. The question here is to assess if the two underlying distributions P_X and P'_X are the same. We claim that this issue can be tackled from the same angle as in **Example 1**. Consider the null assumption $\mathcal{H}_0 : P_X = P'_X$. We propose to turn the multivariate homogeneity test into a collection of univariate tests as follows. Consider $\mathcal{S}$ a class of scoring rules $s : \mathbb{R}^d \to \mathbb{R}$. Set: $P_s = \mathcal{L}(s(X_1))$, and $P'_s = \mathcal{L}(s(X'_1))$. For each $s \in \mathcal{S}$, consider the homogeneity test with null assumption: $\mathcal{H}_{s,0} : P_s = P'_s$. Then, we should reject the null assumption $\mathcal{H}_0$ if there exists an $s \in \mathcal{S}$ such that $\mathcal{H}_{s,0}$ is rejected. The practical strategy then consists in finding the most discriminative scoring rule s based on pretesting data. The test statistic can be based on the ROC curve since the case where $P_s = P'_s$ corresponds to the first diagonal on the ROC curve (*e.g.* the Wilcoxon rank statistic permits to assess discrepancy from the first diagonal).

1.3. **Main questions**

The two previous examples highlight the central role of ROC curve as a tool to measure performance of statistical methods for the inference of preorder relations, and for high dimensional sample comparison. A natural strategy to build estimators is to consider optimizers of ROC curves but considering that it is a function-valued criterion, this requires some preparation. For the ROC curve and related criteria, as well as variations along **Example 1**, it is important to describe the *optimal elements*. Variations on the theme include taking into account more accurate *feedback* than just binary (e.g. Y_i taking values in $\{1, \ldots, K\}$) but also the nature of *sampling schemes* (pointwise, pairwise, listwise). From a statistical theory perspective, the application of Empirical Risk Minimization principles raises classical questions, such as conditions for *uniform convergence*, *consistency of* M*-estimators*, *(fast) rates of convergence*, in a setup which involves processes of higher order than plain empirical processes. More practical issues concern the *design of efficient algorithms*, as well as *meta-algorithms* based on the *aggregation* principle (refer, for instance to bagging-type strategies). The latter will not be considered in the present paper.

2. OPTIMAL ELEMENTS IN RANKING

We consider here the optimality issue from a probabilistic perspective (assuming that the underlying distribution of the data is known).

2.1. The bipartite case

We consider the random pair (X, Y) and we first introduce some notations:

- Joint distribution: P over $\mathbb{R}^d \times \{-1, +1\}$
- Mixture parameter: $p = \mathbb{P}\{Y = +1\}$
- Posterior probability: $\forall x \in \mathbb{R}^d, \eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$
- Class-conditional distributions: $P_+ = \mathcal{L}(X \mid Y = +1)$ and $P_- = \mathcal{L}(X \mid Y = -1)$.

In the case of bipartite ranking, it is easy to conclude about optimal elements by reformulating the problem of optimal ROC curve as a simple hypothesis testing problem. Indeed, the ROC curve can be interpreted as the *power curve* of the test statistic $s(X)$ when testing

$$\mathcal{H}_0 : X \sim P_- \quad \text{against} \quad \mathcal{H}_1 : X \sim P_+ \ .$$

By Neyman-Pearson lemma, it is well known that the test based on the likelihood ratio $\phi(X)$

$$\phi(X) = \frac{dP_+}{dP_-}(X) = \frac{1-p}{p} \cdot \frac{\eta(X)}{1-\eta(X)} \ .$$

yields a *uniformly most powerful* test. From this simple observation, we easily derive the following proposition.

Proposition 1. *The class $\mathcal{S}^*$ of* ROC*-optimal scoring rules contains all the compositions of strictly increasing functions with the posterior probability η:*

$$\mathcal{S}^* = \{T \circ \eta \mid T \ : \ [0,1] \to \mathbb{R} \quad \textit{strictly increasing}\} \ .$$

Furthermore, we also provide a generic expression for the optimal scoring rules.

Proposition 2. *Consider an element $s^* \in \mathcal{S}^*$, then:*

$$\forall x \in \mathcal{X} \ , \quad s^*(x) = c + \mathbb{E}\left(w(V) \cdot \mathbb{I}\{\eta(x) > V\}\right)$$

for some: $c \in \mathbb{R}$, V continuous random variable in $[0,1]$, $w : [0,1] \to \mathbb{R}_+$ integrable.

The representation of optimal scoring rules highlights the fact that solving the bipartite ranking problem amounts to recovering the level sets of the posterior probability η:

$$\{x : \eta(x) > q\}_{q \in (0,1)} \ .$$

In terms of complexity, this problem lies between binary classification (recovering one single level set at $q = 1/2$) and regression function estimation (estimating η itself).

For further details, we refer to [CV09b, CV10] and references therein.

2.2. The K-partite case

We now explore the same question on a generalization of the previous problem called the K-partite problem with $K > 2$. This setup has been recently studied in [CRV13]. The difference with the bipartite setup is that Y takes values in a finite and ordered set such as $\{1, \dots, K\}$. We need to introduce further notations:

- Joint distribution: P over $\mathbb{R}^d \times \{1, \dots, K\}$
- Mixture parameter: $p = (p_1, \dots, p_K)$ where $p_k = \mathbb{P}\{Y = k\}$, $\forall k \in \{1, \dots, K\}$
- Posterior probability: $\forall k \in \{1, \dots, K\}$, $\forall x \in \mathbb{R}^d$, $\eta_k(x) = \mathbb{P}\{Y = k \mid X = x\}$
- Class-conditional distributions: $\forall k \in \{1, \dots, K\}$, $P_k = \mathcal{L}(X \mid Y = k)$.

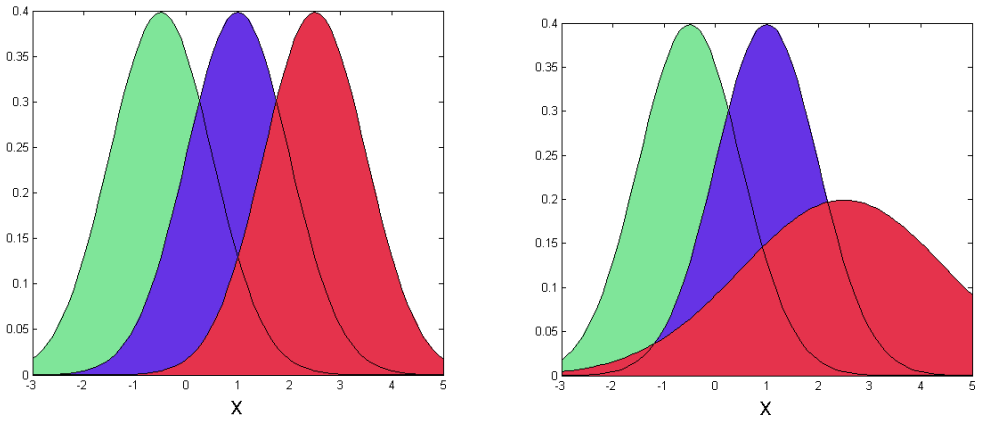

FIGURE 2. (Left) $m_1 < m_2 < m_3$, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$ - (Right) $m_1 < m_2 < m_3$, $\sigma_1^2 = \sigma_2^2 = 1$, $\sigma_3^2 = 2$.

In order to introduce the corresponding notion of optimality in this generalized setup, we build it after the description of optimal scoring rules in the bipartite case.

Definition 3. *A scoring rule* $s^* : \mathbb{R}^d \to \mathbb{R}$ *is said to be* optimal for the K-partite ranking problem *if and only if it satisfies the following condition: for any* l, k *satisfying* $1 \leq l < k \leq K$, *there exists a function* $T_{l,k} : [0,1] \to \mathbb{R}$ *strictly increasing such that*

$$s^* = T_{l,k} \circ \left(\frac{\eta_k}{\eta_l} \right) .$$

Hence, a scoring rule is optimal for the K-partite problem if and only if it is optimal w.r.t. all bipartite subproblems. We have proved that this concept of optimal element is equivalent to optimality in the sense of the ROC surface.

At this point, the important question of the existence of optimal scoring rules arises since different expressions of s^* with respect to the pairs (l,k) need to coincide at each point. We have formulated a necessary and sufficient condition for the existence of an optimal scoring rule.

Proposition 4. *There exists a scoring rule* $s^* : \mathbb{R}^d \to \mathbb{R}$ *which is* optimal for the K-partite ranking problem *if and only if the following condition holds:*

(C) *For any pair* (l,k) *such that* $1 \leq l < k \leq K-1$, *and for any* x, x', *we have*

$$\frac{\eta_{k+1}}{\eta_k}(x) < \frac{\eta_{k+1}}{\eta_k}(x') \Rightarrow \frac{\eta_{l+1}}{\eta_l}(x) < \frac{\eta_{l+1}}{\eta_l}(x') .$$

In particular, if the condition is satisfied, the regression function $\eta(x) = \mathbb{E}(Y \mid X = x) = \sum_{k=1}^{K} k\eta_k(x)$ *is optimal for the* K*-partite ranking problem.*

Figure 2 illustrates condition (C) on two examples in the case of a one-dimensional gaussian mixture with $K = 3$ components: GREEN = class 1, BLUE = class 2, and RED = class 3. The condition is valid on the example at the left display and nonvalid on the one at the right.

3. From performance metrics to higher order statistics

We first consider the bipartite case. We aim at M-estimation strategies for the ranking problem, but the function-valued nature of the performance metric requires some preparation before building estimators. For

typical performance assessment in medical diagnosis for instance, practitioners often refer to summaries of the ROC Curve, known as the Area Under the ROC Curve (AUC criterion) or truncated versions of the AUC (partial AUC).

Definition 5. *Consider* $X^+ \sim P_+$ *and* $X^- \sim P_-$ *independent random variables. For any scoring rule* $s : \mathbb{R}^d \to \mathbb{R}$*, define the AUC as:*

$$\mathrm{AUC}(s) = \mathbb{P}\{s(X^-) < s(X^+)\} ,$$

or considering that (X, Y) *and* (X', Y') *are i.i.d. pairs with distribution* P*:*

$$\mathrm{AUC}(s) = \mathbb{P}\{s(X') < s(X) \mid Y = +1, Y' = -1\} .$$

In credit screening applications, the AUC is also known as the *rate of concording pairs*, that is the proportion of pairs of observations that are correctly ordered by the scoring rule.

Now in order to derive the empirical counterpart of such a quantity, two sampling schemes may be considered: either i.i.d. sampling n copies from $(X, Y) \sim P$, or independent sampling from P_+, P_- with fixed sample size k and m for each subsample ($n = k + m$). Depending on the approach, we can consider, as an empirical AUC, the Mann-Whitney statistic for fixed scoring rule s under two different forms:

$$\widehat{\mathrm{AUC}}(s) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}\{(Y_j - Y_i)\big(s(X_j) - s(X_i)\big) > 0\}$$

or

$$\widehat{\mathrm{AUC}}(s) = \frac{1}{km} \sum_{i=1}^{k} \sum_{j=1}^{m} \mathbb{I}\{(s(X_j^-) < s(X_i^+)\} .$$

The two expressions are very similar and their behavior can be described using standard arguments for U-statistics. The asymptotic analysis may require additional assumptions on the sampling scheme and technical details when k and m become random quantities. We will come back to the analysis of U-statistics in the next section. For the moment, we propose to explore variations and generalizations starting from the connection to the Wilcoxon rank statistic:

$$km\widehat{\mathrm{AUC}}(s) + k(k+1)/2 = \sum_{i=1}^{k} \mathrm{Rank}(s(X_i^+)) .$$

where the rank of an observation is defined by:

$$\mathrm{Rank}(X_i^+) = \sum_{j=1}^{n} \mathbb{I}\{X_j \leq X_i^+\}$$

where $X_1, \ldots, X_n$ is the pooled sample $\mathcal{X}_k^+ \cup \mathcal{X}_m^-$ and $n = k + m$.

The connection to rank statistics actually allows to establish bridges with standard measures used in the information retrieval community such as:

- Average precision:

$$\widehat{W}(s) = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n + 1 - \mathrm{Rank}(s(X_i^+))}$$

- The top-@u%

$$\widehat{W}(s) = \sum_{i=1}^{k} \mathbb{I}\{\mathrm{Rank}(s(X_i^+))/(n+1) > u\}$$

- Discounted Cumulative Gain

$$\widehat{W}(s) = \sum_{i=1}^{k} \frac{1}{\log_2(\mathrm{Rank}(s(X_i^+)) + 1)}$$

Interestingly, one can consider a generic setup for such empirical functionals based on linear rank statistics. We thus define the concept of W-ranking functional (the W stands for Wilcoxon).

Definition 6. *Given two independent samples $X_1^+, \ldots, X_k^+$ and $X_1^-, \ldots, X_m^-$ of i.i.d. observations, of distribution P_+ and P_- respectively, we call a W-ranking functional the following statistic for fixed scoring rule s:*

$$\widehat{W}_{k,m}(s) = \sum_{i=1}^{k} \phi\left(\frac{\mathrm{Rank}(s(X_i^+))}{k+m+1}\right) ,$$

where $\phi : [0,1] \to [0,1]$ is a nondecreasing function (called score-generating function*).*

Here are a few examples of score-generating functions which correspond to known functionals which have been studied in the machine learning or statistical literature:

- $\phi(x) = x \Rightarrow$ empirical AUC
- $\phi(x) = x\, \mathbb{I}\{x \geq 1-u\} \Rightarrow$ (empirical) local AUC
- $\phi(x) = x^p$, $p > 1 \Rightarrow$ p-norm push
- $\phi(x) = c\,((n+1)\,x) \cdot \mathbb{I}_{\{x \geq k/(n+1)\}} \Rightarrow$ Discounted Cumulative Gain
- smooth ϕ.

We refer to [CV07, CV09a] and references therein for additional details on the various criteria related to rank statistics.

4. Global criteria: The U-statistic case

The results of this section come from the two references [CLV08, CRV13].

4.1. Hoeffding's decomposition

In the case of the AUC criterion, the analysis applies thanks to the U-statistic nature of the criterion and the structure is well described by Hoeffding's decomposition. Consider $Z_1, \ldots, Z_n$ an i.i.d. sample over a measurable space $\mathcal{Z}$ and $h_s : \mathcal{Z}^m \to \mathbb{R}$ a measurable function for fixed s. The corresponding U-*statistic* of order m and *kernel* h_s indexed by scoring rules s is the following quantity:

$$U_{n,m}(h_s) = \frac{(n-m)!}{n!} \sum_{(i_1,\ldots,i_m) \in \mathcal{I}_{n,m}} h_s(Z_{i_1}, \ldots, Z_{i_m})$$

where $\mathcal{I}_{n,m} = \{(i_1, \ldots, i_m) \,:\, 1 \leq i_k \leq n,\ i_k \neq i_j \text{ for } k \neq j\}$.

Assuming a symmetric kernel h_s, Hoeffding decomposition relies on a simple projection argument recursively applied to subsets of the collection of m indices:

$$U_{n,m}(h_s) = \sum_{k=0}^{m} \binom{m}{k} U_{n,k}(\pi_k h_s)$$

where:

$$(\pi_k h_s)(Z_1, \ldots, Z_k) = \mathbb{E}(h_s(Z_1, \ldots, Z_m) \mid Z_1, \ldots, Z_k)$$

for $0 \leq k \leq m$. For $k = 0$, $\pi_0 h_s$ coincides with the expectation of h_s and for $k > 0$, we have:

$$\mathbb{E}\big((\pi_k h_s)(Z_1, \ldots, Z_k) \mid Z_1, \ldots, Z_{k-1}\big) = 0 \ , \quad \text{a.s.}$$

that is $\pi_k h_s$ are *completely degenerate kernels.*

4.2. **Consistency of AUC maximization**

We can set $Z_i = (X_i, Y_i)$ and consider U-statistics of order $m = 2$ with kernel:

$$h_s(Z_1, Z_2) = \mathbb{I}\{(Y_j - Y_i)\big(s^*(X_j) - s^*(X_i)\big) > 0\} - \mathbb{I}\{(Y_j - Y_i)\big(s(X_j) - s(X_i)\big) > 0\} \ .$$

Hence, applying Hoeffding's decomposition to the variation of empirical AUC compared to an optimal scoring rule, we can write:

$$\widehat{\mathrm{AUC}}(s^*) - \widehat{\mathrm{AUC}}(s) = \mathrm{AUC}(s^*) - \mathrm{AUC}(s) + \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}(h_s(Z_i, Z) \mid Z_i) + R_n(s) \ .$$

where the remainder term $R_n(s)$ is a U-statistic of order 2 with completely degenerate kernel $\tilde{h}_s$.

Therefore, the consistency of the empirical AUC maximizer relies on the joint analysis of a leading term which has the form of an empirical process and a remainder term which is a degenerate U-statistic. The leading term can be entirely managed with Talagrand's concentration inequality and a variance control assumption. The remainder term requires to be finely controlled in order not to spoil the fine rates of convergence of the leading term. We now state the consistency result for the empirical AUC maximizer under a VC major class complexity assumption.

Theorem 7. *Assume we have:*

- *a class $\mathcal{S}$ of candidate scoring rules which is a VC major class with dimension V*
- *for all $s \in \mathcal{S}$,*

$$\mathit{Var}(\mathbb{E}(h_s(Z_1, Z) \mid Z_1)) \leq c\,(\mathrm{AUC}(s^*) - \mathrm{AUC}(s))^{\alpha} \qquad \textbf{(V)}$$

with some constants $c > 0$ and $\alpha \in [0, 1]$.

We consider the M-estimate $\widehat{s}_n$ based on the empirical AUC. Then, with probability larger than $1 - \delta$:

$$\mathrm{AUC}(s^*) - \mathrm{AUC}(\widehat{s}_n) \leq 2\left(\mathrm{AUC}(s^*) - \inf_{s \in \mathcal{S}} \mathrm{AUC}(s)\right) + C\left(\frac{V \log(n/\delta)}{n}\right)^{1/(2-\alpha)}$$

If we assume, for instance, that the posterior probability η is such that the random variable $\eta(X)$ is absolutely continuous on $[0, 1]$ with bounded density, then condition **(V)** holds. In the previous theorem, the fine control of the remainder term is guaranteed by the complexity assumption of a VC major class for the candidate scoring rules, as we discuss below.

Additional Complexity Measures. The control of leading term does not present major differences with the standard framework of empirical processes which allows to deal for instance with the binary classification problem. However, the analysis on the remainder term introduces some technical complications. In particular, the derivation of general upper bounds on degenerate U-processes requires some additional complexity measures.

Definition 8. *Denote by* $\epsilon_1, \ldots, \epsilon_n$ *i.i.d. Rademacher random variables (*$\{-1,+1\}$*-valued with uniform probability distribution) which are also independent of the* (X_i, Y_i)*'s. We introduce the following complexity measures:*

$$(1) \quad Z_\epsilon = \sup_{s\in\mathcal{S}} \left| \sum_{i,j} \epsilon_i \epsilon_j \tilde{h}_s(Z_i, Z_j) \right|$$

$$(2) \quad U_\epsilon = \sup_{s\in\mathcal{S}} \sup_{\alpha:\|\alpha\|_2\leq 1} \sum_{i,j} \epsilon_i \alpha_j \tilde{h}_s(Z_i, Z_j)$$

$$(3) \quad M_\epsilon = \sup_{s\in\mathcal{S}} \max_{k=1\ldots n} \left| \sum_{i=1}^{n} \epsilon_i \tilde{h}_s(Z_i, Z_k) \right|$$

The key result for the control of the remainder term is the following moment inequality.

Theorem 9. *If* R_n *is a degenerate* U*-statistic, then there exists a universal constant* $C > 0$ *such that for all* n *and* $q \geq 2$*,*

$$\left(\mathbb{E}\big(\sup_{s\in\mathcal{S}} R_n(s)\big)^q \right)^{1/q} \leq C \left(\mathbb{E}Z_\epsilon + q^{1/2}\mathbb{E}U_\epsilon + q(\mathbb{E}M_\epsilon + n) + q^{3/2}n^{1/2} + q^2 \right)$$

The main tools to derive this moment inequalities are typical of U-processes techniques, namely symmetrization, decoupling and concentration inequalities. Optimizing in q the previous inequality allows to turn the moment inequality into an exponential inequality

Corollary 10. *With probability* $1-\delta$*,*

$$\sup_{s\in\mathcal{S}} R_n(s) \leq C \left(\frac{\mathbb{E}Z_\epsilon}{n^2} + \frac{\mathbb{E}U_\epsilon \sqrt{\log(1/\delta)}}{n^2} + \frac{\mathbb{E}M_\epsilon \log(1/\delta)}{n^2} + \frac{\log(1/\delta)}{n} \right)$$

Eventually, under the VC major class assumption, this upper bound can be simplified to

$$\sup_{s\in\mathcal{S}} R_n(s) \leq \frac{1}{n} \left(V + \log(1/\delta) \right) \ , \quad \text{with probability } 1-\delta \ ,$$

since we have

$$\mathbb{E}Z_\epsilon \leq CnV \ , \qquad \mathbb{E}U_\epsilon \leq Cn\sqrt{V} \ , \qquad \mathbb{E}_\epsilon M_\epsilon \leq C\sqrt{Vn} \ .$$

We conclude that the rate obtained in the remainder term will not affect the one guaranteed for the empirical process even in the case of fast rates of the order of $O(1/n)$.

4.3. **Empirical ranking functional for K-partite ranking**

In the case of K-partite ranking, the natural extension of the ROC curve conducts to the concept of ROC surface and the equivalent of the AUC for $K > 2$ is called the Volume Under the ROC Surface (VUS). Scoring rules with high discriminative power tend to have high VUS. The empirical version of the VUS can be represented as a U-statistic of order $m = 3$.

$$\widehat{\mathrm{VUS}}_n(s) = \frac{1}{n_1 n_2 n_3} \sum_{1\leq i,\, j,\, k\leq n} \mathbb{I}\{s(X_i) < s(X_j) < s(X_k)\} \cdot \mathbb{I}\{Y_i = 1,\ Y_j = 2,\ Y_k = 3\}, \tag{1}$$

with $n_i = |\{j \ : \ Y_j = i\}|$.

The analysis of empirical VUS maximizers can certainly be carried out based on U-processes techniques in a similar fashion as it was performed on empirical AUC maximizers (see previous section) although it remains

to be done. However, deriving concave surrogates for algorithmic purposes is not as straightforward as for the bipartite case. In a recent work [CRV13], we have explored the relationship between pairwise strategies (such as one-versus-one aggregation) and VUS maximization, but obtaining sharp convergence rates for this problem is still under investigation.

5. Weighted criteria: the R-statistics case

Results of this section are discussed in details in [CV07, CV09a].

5.1. Local AUC and signed rank statistics

We focus on a subproblem of the bipartite ranking problem which is to identify the best instances. We call best instances those observations with the highest scores with respect to an optimal rule for bipartite ranking and we set as a target a proportion u of these observations.

Definition 11. *The class of candidates for the set with best instances at level $u \in (0,1)$ is the class of sets of the form:*

$$C_{s,u} = \{x \in \mathbb{R}^d \mid s(x) > F_s^{-1}(1-u)\}$$

where s is any candidate real-valued scoring rule and we denote by $F_s^{-1}(1-u)$ the $(1-u)$-quantile of the random variable $s(X)$.

We consider a natural empirical risk functional for this problem which unfortunately presents a complicated structure.

Definition 12. *The empirical risk functional for the problem of finding the best instances is the following:*

$$\widehat{W}_n(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{Y_i \cdot (s(X_i) - \widehat{F}_s^{-1}(1-u)) < 0\} \ ,$$

where s is any candidate real-valued scoring rule.

We introduce the following closely related functional:

$$\widehat{K}_n(s,u) = \frac{1}{n}\sum_{i=1}^{n} Y_i \ \mathbb{I}\{s(X_i) \leq \widehat{F}_s^{-1}(1-u)\}$$

so that $\widehat{W}_n(s) = \dfrac{m}{n} + \widehat{K}_n(s,u)$ where $m = \sum_{i=1}^{n} \mathbb{I}\{Y_i = -1\}$.

At this point, we recall the definition of a linear signed rank statistic.

Definition 13. *Consider $Z_1, \ldots, Z_n$ i.i.d. and $\Phi : [0,1] \to [0,1]$ a score generating function. We denote by $R_i^+ = rank(|Z_i|)$. The statistic*

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

is called a linear signed rank statistic.

It turns out from the definition that, for fixed s and u, the statistic $\widehat{K}_n(s,u)$ is a linear signed rank statistic. The structure of such statistics can also be understood thanks to a different projection argument due to Koul (1970). We assume that the cdf F_s as well as its inverse function are both differentiable.

Proposition 14. *We introduce the following quantities:*

- $K(s,u) = \mathbb{E}\left(Y \ \mathbb{I}\{s(X) \leq F_s^{-1}(1-u)\}\right)$

- $K'(s,u) = \dfrac{\partial K}{\partial u}(s,u)$
- $Z_n(s,u) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \left(Y_i - K'(s,u)\right) \mathbb{I}\{s(X_i) \leq F_s^{-1}(1-u)\} - K(s,u) + uK'(s,u)$

We then have, for all s *and* $u \in [0,1]$*:*

$$\widehat{K}_n(s,u) = K(s,u) + Z_n(s,u) + \Lambda_n(s) \ .$$

with

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

The fastest known rate for recovering the set $C_{\eta,u}$ with best instances at level $u \in (0,1)$, under VC-type complexity assumption, is **of the order of** $n^{-2/3}$. Indeed, there are two competing effects here: the target scoring rule (and thus the candidates scoring rules) should be steep enough to separate the best instances from the others, but at the same time it should be flat enough in order to allow efficient quantile estimation $F_\eta^{-1}(1-u)$. It is not known if this rate is optimal as there are no lower bounds available.

5.2. Smooth *W*-ranking functional and R-processes

We consider in this section summaries of the ROC curve whose empirical counterpart $\widehat{W}_n(s)$ can be expressed as a linear rank statistic with smooth score-generating function ϕ. We point out that sigmoid shapes for ϕ allow to emphasize the importance of high scores in the estimation procedure. Before stating the projection result, we introduce further notations, defined for each scoring rule s:

- F_s cdf of the random variable $s(X)$,
- G_s conditional cdf of the random variable $s(X)$ given $Y = +1$,

Proposition 15. *Consider a score-generating function* ϕ *which is twice continuously differentiable on* $[0,1]$*. We set, for all* $x \in \mathbb{R}^d$*:*

$$\Phi_s(x) = \phi(F_s(s(x))) + p\int_{s(x)}^{+\infty} \phi'(F_s(u))\mathrm{d}G_s(u) \ .$$

Let $\mathcal{S}$ *be a VC major class of functions. Then, we have:* $\forall s \in \mathcal{S}$*,*

$$\widehat{W}_n(s) = \widehat{V}_n(s) + \widehat{R}_n(s),$$

where $\widehat{V}_n(s) = \sum_{i=1}^{n} \mathbb{I}_{\{Y_i=+1\}}\Phi_s(X_i)$ *and* $\widehat{R}_n(s) = O_{\mathbb{P}}(1)$ *as* $n \to \infty$ *uniformly over* $s \in \mathcal{S}$ *.*

The asymptotic equivalent of $\widehat{W}_n(s)$ is easily seen to converge, thanks to an argument by Chernoff and Savage, to the following quantity.

Definition 16. *For a given score-generating function* ϕ*, we will call the functional*

$$W(s) = \mathbb{E}\big(\phi(F_s(s(X)) \mid Y = +1\big) \ ,$$

a W-ranking performance measure.

The class of optimal scoring rules in the sense of the ROC curves also covers the optimal elements of W-ranking performance measure for any increasing score-generating function ϕ. We have shown a consistency result for the empirical maximizer of $\widehat{W}_n(s)$ but no fast rates are presently known in this setup.

Theorem 17. *Set the empirical W-ranking performance maximizer $\hat{s}_n = \arg\max_{s\in\mathcal{S}} \widehat{W}_n(s)$. With the same notations as in the previous proposition and assuming in addition that the class of functions Φ_s induced by $\mathcal{S}$ is also a VC major class of functions with VC dimension V, we have, for any $\delta > 0$, and with probability $1-\delta$:*

$$W(s^*) - W(\hat{s}_n) \leq c_1\sqrt{\frac{V}{n}} + c_2\sqrt{\frac{\log(1/\delta)}{n}} \ ,$$

for some positive constants c_1, c_2.

6. Open issues

In this survey paper, we have presented state-of-the-art results for consistency and convergence rates for the bipartite and K-partite ranking problems. We established the strong connection between summaries of the ROC curve (or surface) and the theory of U- and $R-$ processes. A similar scheme can be proposed for establishing preliminary results based on a projection argument, concentration inequalities applied to empirical processes with variance control (Talagrand's inequality) for the leading term and *ad hoc* computations to control remainder terms. At this stage, no general argument allows to cover smooth and nonsmooth cases for the remainder term in the decomposition after projection. In particular, up to our knowledge, in the case of R-processes, there is no concentration inequality result of the same flavor as the moment inequality for degenerate U-processes. We finally summarize the open questions in the three main examples that were discussed:

(1) U-statistic case: fast rates for convex surrogate loss functions?
(2) Finding the best instances: existence of fast rates beyond the $n^{-2/3}$-rate?
(3) Smooth case: variance control assumption and fast rates?
(4) In all cases: lower bounds and optimal rates?

References

[CDV10] S. Clémençon, M. Depecker, and N. Vayatis. AUC-optimization and the two-sample problem. In *Proceedings of NIPS'09*, 2010.

[CLV08] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. *Annals of Statistics*, 36(2):844–874, 2008.

[CRV13] S. Clémençon, S. Robbiano, and N. Vayatis. Ranking data with ordinal labels: optimality and pairwise aggregation. *Machine Learning*, 91(1):67–104, 2013.

[CV07] S. Clémençon and N. Vayatis. Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699, 2007.

[CV09a] S. Clémençon and N. Vayatis. Empirical performance maximization based on linear rank statistics. In *Proceedings of NIPS'08*, *Lecture Notes in Computer Science*, 3559:1–15, Springer, 2009.

[CV09b] S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, 2009.

[CV10] S. Clémençon and N. Vayatis. Overlaying classifiers: a practical approach to optimal scoring. *Constructive Approximation*, 32(3):619-648, 2010.