Concentration Inequalities for Two-Sample Rank Processes with Application to Bipartite Ranking

Stéphan Clémençon, Myrto Limnios, Nicolas Vayatis

Concentration Inequalities for Two-Sample Rank Processes with Application to Bipartite Ranking

Stephan Clémençon[1], Myrto Limnios *[2], Nicolas Vayatis[2]

[1] stephan.clemencon@telecom-paris.fr, Telecom Paris, LTCI, Institut Polytechnique de Paris, 19 place Marguerite Perey, Palaiseau, 91120, France.
[2]myrto.limnios@ens-paris-saclay.fr, nicolas.vayatis@ens-paris-saclay.fr, Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, F-91190 Gif-sur-Yvette, France.

Abstract

The ROC curve is the gold standard for measuring the performance of a test/scoring statistic regarding its capacity to discriminate between two statistical populations in a wide variety of applications, ranging from anomaly detection in signal processing to information retrieval, through medical diagnosis. Most practical performance measures used in scoring/ranking applications such as the AUC, the local AUC, the p-norm push, the DCG and others, can be viewed as summaries of the ROC curve. In this paper, the fact that most of these empirical criteria can be expressed as *two-sample linear rank statistics* is highlighted and concentration inequalities for collections of such random variables, referred to as *two-sample rank processes* here, are proved, when indexed by VC classes of scoring functions. Based on these nonasymptotic bounds, the generalization capacity of empirical maximizers of a wide class of ranking performance criteria is next investigated from a theoretical perspective. It is also supported by empirical evidence through convincing numerical experiments.

1 Introduction

In the context of ranking, a variety of performance measures can be considered. In the simplest framework of bipartite ranking, where two independent *i.i.d.* samples $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, valued in the same space $\mathcal{Z}$, say $\mathbb{R}^d$ with $d \geq 1$ for instance, and drawn from probability distributions G and H respectively (referred to as the 'positive distribution' and the 'negative distribution' respectively), the goal pursued is to learn a preorder on $\mathcal{Z}$ defined through a scoring function $s : \mathcal{Z} \to \mathbb{R}$ (which transports the natural order on the real line onto the feature space $\mathcal{Z}$) such that, for any random observation $\mathbf{Z} \in \mathcal{Z}$ sampled from a distribution that is equal either to the 'positive distribution' or to the 'negative one', the larger

*Corresponding author

the score $s(z)$, the likelier it is drawn from the 'positive distribution' G. Though easy to formulate, this simple framework encompasses many practical problems from the design of search engines in Information Retrieval (in this case, for a specific request, G is the distribution of the relevant digitized documents, while H is that of the irrelevant ones) to the elaboration of decision support tools in personalized medicine for instance. In spite of its simplicity there is not one and only one natural scalar criterion for evaluating the performance of a scoring rule $s(z)$, but many possible options. The *Receiving Operator Characteric* curve (the ROC curve in abbreviated form), *i.e.* the PP-plot of the false positive rate *vs* the true positive rate:

$$t \in \mathbb{R} \mapsto \left(\mathbb{P}\{s(\mathbf{Y}) > t\},\ \mathbb{P}\{s(\mathbf{X}) > t\}\right),$$

denoting by $\mathbf{X}$ and $\mathbf{Y}$ two generic *r.v.* with distributions G and H respectively, provides an exhaustive description of the performance of any scoring rule candidate s. However, its functional nature renders direct optimization strategies rather complex, see *e.g.* [10]. *Empirical risk minimization* methods (ERM) are thus generally based on summaries of the ROC curve, which take the form of empirical risk functionals where the averages involved are no longer taken over *i.i.d.* sequences. The most popular choice is undoubtedly the AUC criterion (AUC standing for *Area Under the* ROC *Curve*), see [1] or [4] for instance, but when focus is on top-ranked instances, various choices can be considered, *e.g.* the Discounted Cumulative Gain or DCG (see [11]), the p-norm push (see [30]), the local AUC (refer to [7]) or other variants such as those recently introduced in [26]. The present paper starts from the simple observation that most of these summary criteria have a common feature: they belong to the class of *two-sample linear rank statistics*. Such statistics have been extensively studied in the mathematical statistics literature because of their optimality properties in hypothesis testing, see [19]. They are widely used in order to test whether two samples are drawn from the same distribution against the alternative stipulating that the distribution of one of the samples is stochastically larger than the other. For instance, the empirical counterpart of the AUC of a scoring function $s(z)$ corresponds to the popular Mann-Withney-Wilcoxon statistic based on the two (univariate) samples $s(\mathbf{X}_1),\ \ldots,\ s(\mathbf{X}_n)$ and $s(\mathbf{Y}_1),\ \ldots,\ s(\mathbf{Y}_m)$. Other rank statistics can be considered, corresponding to other ways of measuring how the distribution of the 'positive score' $s(\mathbf{X})$ is (possibly) stochastically larger than that of the 'negative score' $s(\mathbf{Y})$. Now, in the statistical learning view, with the importance of excess risk bounds, the *Empirical Risk Minimization* paradigm must be revisited and new problems, mainly related to the uniform control of the fluctuations of collections of two-sample linear rank statistics, termed rank processes throughout the article, and to the measure of the complexity of non-parametric classes of scoring functions, come up. The arguments required to deal with risk functionals based on two-sample linear rank statistics have been sketched in [7] in a very special case.

In the present paper, we relate two-sample linear rank statistics to performance measures relevant for the ranking problem by showing that the target of ranking algorithms corresponds to optimal ordering rules in this sense and show that the generic structure of two-sample linear rank statistics as an orthogonal decomposition after projection onto the space of sums of *i.i.d.* random variables is the key to all statistical results related to maximizers of such criteria: consistency, rates of convergence or model selection. Notice incidentally that the empirical AUC

is also a U-statistic and a decomposition method akin to that considered in this paper (though much less general) has been used in order to handle this specific dependence structure in [4]. In this article, concentration properties of two-sample rank processes (*i.e.* collections of two-sample linear rank statistics) are investigated using the linearization technique aforementioned. While interesting in themselves, the concentration inequalities established for this class of stochastic processes, when indexed by Vapnik-Chervonenkis classes (abbreviated with VC-classes) of scoring functions, are next applied to study the generalization capacity of empirical maximizers of a large collection of performance criteria based on two-sample linear rank statistics. Notice finally that a preliminary version of this work is briefly outlined in the conference paper [8]. This article presents a much deeper analysis of bipartite ranking via maximization of two-sample linear rank statistics. In particular, it offers a complete and detailed study of the concentration properties of two-sample rank processes (in a slightly different framework, stipulating that two independent *i.i.d.* samples, positive and negative, are observed, rather than classification data), provides model selection results and, from a practical perspective, tackles the issue of smoothing the risk functionals under study here with statistical learning guarantees.
The paper is organized as follows. In Section 2, the main notations are set out, the bipartite ranking problem is formulated as a statistical learning task in a rigorous probabilistic framework and the concept of two-sample linear rank statistic is briefly recalled. It is also explained that, unsurprisingly, natural performance criteria in bipartite ranking are of the form of two-sample (linear) rank statistics. Concentration results for rank processes, are established in Section 3. By means of the latter, performance of bipartite ranking rules obtained by maximizing two-sample linear rank statistics are investigated in Section 4. Finally, Section 5 displays illustrative experimental results, supporting the theoretical analysis carried out in the present article. Proofs, technical details and additional numerical results are deferred to the Appendix section.

2 Motivation and Preliminaries

We start with recalling key notions pertaining to ROC analysis and bipartite ranking, which essentially motivates the theoretical analysis carried out in the subsequent section. We next recall at length the definition of two-sample linear rank statistics, which have been intensively used to design statistical (homogeneity) testing procedures in the univariate setup, and finally highlight that many scalar summaries of empirical ROC curves, commonly used as ranking performance criteria, are precisely of this form. Here and throughout, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$, the Dirac mass at any point x by δ_x, the generalized inverse of any cumulative distribution function $W(t)$ on $\mathbb{R} \cup \{+\infty\}$ by $W^{-1}(u) = \inf\{t \in]-\infty,\ +\infty] :\ W(t) \geq u\}$, $u \in [0, 1]$. We also denote the floor and ceiling functions by $u \in \mathbb{R} \mapsto \lfloor u \rfloor$ and by $u \in \mathbb{R} \mapsto \lceil u \rceil$ respectively.

2.1 Bipartite Ranking and ROC Analysis

As recalled in the Introduction section, the goal of bipartite ranking is to learn, based on independent 'positive' and 'negative' random samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$

and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, how to score any new observations $\mathbf{Z}_1, \ldots, \mathbf{Z}_k$, being each either 'positive' or else 'negative', that is to say drawn either from G or else from H, without prior knowledge, so that positive instances are mostly at the top of the resulting list with large probability. A natural way of defining a total preorder[1] on $\mathcal{Z}$ is to map it with the natural order on $\mathbb{R} \cup \{+\infty\}$ by means of a *scoring rule*, *i.e.* a measurable mapping $s : \mathcal{Z} \to]-\infty, \infty]$. By $\mathcal{S}$ is denoted the set of all scoring rules. It is by means of ROC analysis that the capacity of a scoring rule candidate $s(z)$ to discriminate between the positive and negative statistical populations is generally evaluated.

ROC **curves.** The ROC curve is a gold standard to describe the dissimilarity between two univariate probability distributions G and H. This criterion of functional nature, $\mathrm{ROC}_{\mathrm{H,G}}$, can be defined as the parametrized curve in $[0,1]^2$:

$$t \in \mathbb{R} \mapsto (1 - H(t),\ 1 - G(t))\,,$$

where possible jumps are connected by line segments, so as to ensure that the resulting curve is continuous. With this convention, one may then see the ROC curve related to the pair of *d.f.* (H, G) as the graph of a càd-làg (*i.e.* right-continuous and left-limited) non decreasing mapping valued in $[0,1]$, defined by:

$$\alpha \in (0,1) \mapsto 1 - G \circ H^{-1}(1-\alpha)\ ,$$

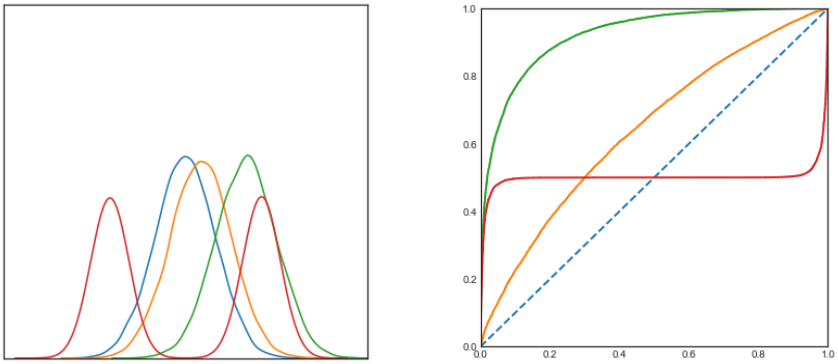
at points α such that $G \circ H^{-1}(1-\alpha) = 1-\alpha$. Denoting by $\mathcal{Z}_H$ and $\mathcal{Z}_G$ the supports of H and G respectively, observe that it connects the point $(0, 1 - G(\mathcal{Z}_H))$ to $(H(\mathcal{Z}_G), 1)$ in the unit square $[0,1]^2$ and that, in absence of plateau (which we assume here for simplicity, rather than restricting the feature space to G's support), the curve $\alpha \in (0,1) \mapsto \mathrm{ROC}_{\mathrm{G,H}}(\alpha)$ is the image of $\alpha \in (0,1) \mapsto \mathrm{ROC}_{\mathrm{H,G}}(\alpha)$ by the reflection with the main diagonal of the Euclidean plane (*i.e.* the line of equation '$\beta = \alpha$') as axis. Notice that the curve $\mathrm{ROC}_{\mathrm{H,G}}$ coincides with the main diagonal of $[0,1]^2$ if and only if the two distributions H and G are equal. Hence, the concept of ROC curve offers a visual tool to examine the differences between two distributions in a pivotal manner, see Fig. 1. For instance, the univariate distribution $G(dt)$ is stochastically larger[2] than $H(dt)$ if and only if the curve $\mathrm{ROC}_{\mathrm{H,G}}$ is everywhere above the main diagonal and $\mathrm{ROC}_{\mathrm{H,G}}$ coincides with the left upper corner of the unit square *iff* the essential supremum of the distribution H is smaller than the essential infimum of the distribution G. Another advantage of the ROC curve lies in the probabilistic interpretation of the popular ROC curve summary, referred to as the Area Under the ROC Curve (AUC in short)

$$\mathrm{AUC}_{\mathrm{H,G}} \stackrel{\mathrm{def}}{=} \int_0^1 \mathrm{ROC}_{\mathrm{H,G}}(\alpha)\mathrm{d}\alpha = \mathbb{P}\left\{\mathrm{Y} < \mathrm{X}\right\} + \frac{1}{2}\mathbb{P}\left\{\mathrm{X} = \mathrm{Y}\right\}, \tag{2.1}$$

where (X, Y) denotes a pair of independent *r.v.*'s with respective marginal distributions H and G.

[1] A preorder $\preccurlyeq$ on a set $\mathcal{Z}$ is a reflexive and transitive binary relation on $\mathcal{Z}$. It is said to be *total*, when either $z \preccurlyeq z'$ or else $z' \preccurlyeq z$ holds true, for all $(z, z') \in \mathcal{Z}^2$.

[2] Given two distribution functions $H(dt)$ and $G(dt)$ on $\mathbb{R} \cup \{+\infty\}$, it is said that $G(dt)$ is *stochastically larger* than $H(dt)$ *iff* for any $t \in \mathbb{R}$, we have $G(t) \leq H(t)$. We then write: $H \leq_{sto} G$. Classically, a necessary and sufficient condition for G to be stochastically larger than H is the existence of a coupling $(\mathbf{X},\ \mathbf{Y})$ of (G, H), *i.e.* a pair of random variables defined on the same probability space with first and second marginals equal to H and G respectively, such that $\mathbf{X} \leq \mathbf{Y}$ with probability one.



a. Probability distributions b. ROC curves

Figure 1: Examples of pairs of distributions and their related ROC curves. The 'negative' distribution H is represented in blue and three examples of 'positive' distributions are represented in red, orange and green, like the associated ROC curves.

Bipartite Ranking as ROC curve optimization. Going back to the multivariate setup, where H and G are probability distributions on $\mathcal{Z}$, say $\mathcal{Z} = \mathbb{R}^d$ with arbitrary dimension $d \geq 1$, the goal pursued in bipartite ranking can be phrased as that of building a scoring rule $s(z)$ such that the (univariate) distribution G_s of $s(\mathbf{X})$ is 'as stochastically larger as possible' than the the distribution H_s of $s(\mathbf{Y})$. Hence, the capacity of a candidate $s(z)$ to discriminate between the positive and negative statistical populations can be evaluated by plotting the ROC curve $\alpha \in (0,1) \mapsto \mathrm{ROC}(s,\alpha) = \mathrm{ROC}_{H_s,G_s}(\alpha)$: the closer to the left upper corner of the unit square the curve $\mathrm{ROC}(s,.)$, the better the scoring rule s. Therefore, the ROC curve conveys a partial preorder on the set of all scoring functions: for all pairs of scoring functions s_1 and s_2, one says that s_2 is more accurate than s_1 when $\mathrm{ROC}(s_1,\alpha) \leq \mathrm{ROC}(s_2,\alpha)$ for all $\alpha \in [0,1]$. It follows from a standard Neyman-Pearson argument that the most accurate scoring rules are increasing transforms of the likelihood ratio $\Psi(z) = dG/dH(z)$. Precisely, it is shown in [9] (see Proposition 2 therein) that the optimal scoring rules are the elements of the set:

$$\mathcal{S}^* = \left\{ s \in \mathcal{S} \;\; s.t. \text{ for all } z,\, z' \text{ in } \mathbb{R}^d : \;\; \Psi(z) < \Psi(z') \Rightarrow s^*(z) < s^*(z') \right\}. \quad (2.2)$$

We denote by $\mathrm{ROC}^*(.) = \mathrm{ROC}(\Psi, .)$ and recall incidentally that this optimal curve is non-decreasing and concave and thus always above the main diagonal of the unit square. Now, the bipartite ranking task can be reformulated in a more quantitative manner: the objective pursued is to build a scoring function $s(z)$, based on the training examples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, with a ROC curve as close as possible to ROC^*. A typical way of measuring the deviation between the two curves is to consider the distance in sup norm:

$$d_\infty(s, s^*) = \sup_{\alpha \in (0,1)} \left| \mathrm{ROC}(s,\alpha) - \mathrm{ROC}^*(\alpha) \right| . \quad (2.3)$$

Attention should be paid that this quantity is a distance between ROC curves (or between the related equivalence classes of scoring functions, the ROC curve

of any scoring function being invariant by strictly increasing transform) not between the scoring functions themselves. Since the curve ROC^* is unknown in practice, the major difficulty lies in the fact that no straightforward statistical counterpart of the (functional) loss (2.3) is available. In [9] (see also [10]), it has been however shown that bipartite ranking can be viewed as a superposition of cost-sensitive classification problems and somehow 'discretized' in an adaptive manner, so as to apply empirical risk minimization with statistical guarantees in the d_∞-sense, at the price of an additional bias term inherent to the approximation step. Alternatively, the performance of a candidate scoring rule s can be measured by means of the L_1-norm in the ROC space. Observing that, in this case, the loss can be decomposed as follows:

$$d_1(s, s^*) = \int_0^1 |\mathrm{ROC}(\mathrm{s}, \alpha) - \mathrm{ROC}^*(\alpha)| \mathrm{d}\alpha = \int_0^1 \mathrm{ROC}^*(\alpha) \mathrm{d}\alpha - \int_0^1 \mathrm{ROC}(\mathrm{s}, \alpha) \mathrm{d}\alpha \, , \tag{2.4}$$

minimizing the L_1-distance to the optimal ROC curve boils down to maximizing the area under the curve $\mathrm{ROC}(\mathrm{s}, \, .)$, that is to say

$$\mathrm{AUC}(\mathrm{s}) \stackrel{\mathrm{def}}{=} \mathrm{AUC}_{\mathrm{H_s, G_s}} = \mathbb{P}\{\mathrm{s}(\mathbf{Y}) < \mathrm{s}(\mathbf{X})\} + \frac{1}{2}\mathbb{P}\{\mathrm{s}(\mathbf{Y}) = \mathrm{s}(\mathbf{X})\} \, , \tag{2.5}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are random variables defined on the same probability space, independent, with respective distributions G and H, denoting by G_s and H_s the distributions of $s(\mathbf{X})$ and $s(\mathbf{Y})$ respectively. The scalar performance criterion $\mathrm{AUC}(\mathrm{s})$ defines a total preorder on $\mathcal{S}$ and its maximal value is denoted by $\mathrm{AUC}^* = \mathrm{AUC}(\mathrm{s}^*)$, with $s^* \in \mathcal{S}^*$. Bipartite ranking through maximization of empirical versions of the AUC criterion has been studied in several articles, including [1] or [4]. Extension to *multipartite ranking* (*i.e.* when the number of samples/distributions under study is larger than 3) is considered in [6], see also [5]. In contrast to [9] or [10], where the task of learning scoring rules with statistical guarantees in sup norm in the ROC space is considered, the present article focuses on optimization of summary scalar empirical criteria generalizing the AUC that takes the form of two-sample linear *rank statistics*, as could be naturally expected when addressing ranking problems.

2.2 Two-Sample Linear Rank Statistics

If the curve $\mathrm{ROC}_{\mathrm{H,G}}$ is the appropriate tool to examine to which extent a univariate distribution G is stochastically larger than another one H, practical decisions are generally made on the basis of the observations of two univariate independent random *i.i.d.* samples $\{X_1, \, \ldots, \, X_n\}$ and $\{Y_1, \, \ldots, \, Y_m\}$, drawn from G and H respectively. Computing the empirical cumulative distribution functions $\widehat{H}_m(t) = (1/m) \sum_{j=1}^m \mathbb{I}\{Y_j \leq t\}$ and $\widehat{G}_n(t) = (1/n) \sum_{i=1}^n \mathbb{I}\{X_i \leq t\}$ for $t \in \mathbb{R}$, one can plot the empirical ROC curve:

$$\widehat{\mathrm{ROC}} = \mathrm{ROC}_{\widehat{\mathrm{H}}_\mathrm{m}, \, \widehat{\mathrm{G}}_\mathrm{n}} \, . \tag{2.6}$$

Observe that the ROC curve (2.6) is an increasing broken line connecting $(0, 0)$ to $(1, 1)$ in the unit square $[0, 1]^2$ and is fully determined by the set of ranks occupied by the positive instances within the pooled sample $\{\mathrm{Rank}(X_i) : \, i = 1, \, \ldots, \, n\}$, where:

$$\forall i \in \{1, \ldots, n\} \quad \mathrm{Rank}(X_i) = N\widehat{F}_N(X_i) \ , \tag{2.7}$$

with $\widehat{F}_N(t) = (1/N)\sum_{i=1}^{n} \mathbb{I}\{X_i \leq t\} + (1/N)\sum_{j=1}^{m} \mathbb{I}\{Y_j \leq t\}$ and $N = n + m$. Breakpoints of the piecewise linear curve (2.6) necessarily belong to the set of gridpoints $\{(j/m,\ i/n) : j \in \{1, \ldots, m-1\} \text{ and } i \in \{1, \ldots, n-1\}\}$. Denote by $X_{(i)}$ the order statistics related to the sample $\{X_1, \ldots, X_n\}$, *i.e.* $\mathrm{Rank}(X_{(n)}) > \cdots > \mathrm{Rank}(X_{(1)})$, and by $Y_{(j)}$ those related to the sample $\{Y_1, \ldots, Y_m\}$. Consider the càd-làg step function:

$$\alpha \in [0,1] \mapsto \sum_{j=1}^{m} \widehat{\gamma}_j \cdot \mathbb{I}\{\alpha \in [(j-1)/m,\ j/m[\} \ , \tag{2.8}$$

where, for all $j \in \{1, \ldots, m\}$, we set:

$$\begin{aligned}\widehat{\gamma}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i > Y_{(m-j+1)}\} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\mathrm{Rank}(X_{(n-i+1)}) > \mathrm{Rank}(Y_{(m-j+1)})\} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{j \geq N - \mathrm{Rank}(X_{(n-i+1)}) - i + 2\} \ .\end{aligned}$$

The ROC curve (2.6) is the continuous broken line that connects the jump points of the step curve (2.8) and can thus be expressed as a function of the 'positive ranks' *i.e.* the $\mathrm{Rank}(X_i)$'s only. As a consequence, any summary of the empirical ROC curve, is a two-sample rank statistic, that is a measurable function of the 'positive ranks'. In particular, the empirical AUC, *i.e.* the AUC of the empirical ROC curve (2.6), also termed the rate of concording pairs or the *Mann-Whitney statistic*, can be easily shown to coincide, up to an affine transform, with the sum of 'positive ranks', the well-known *rank-sum Wilcoxon statistic* [37]:

$$\widehat{W}_{n,m} = \sum_{i=1}^{n} \mathrm{Rank}(X_i) \ . \tag{2.9}$$

Indeed, we have:

$$\widehat{W}_{n,m} = nm\mathrm{AUC}_{\widehat{\mathrm{H}}_\mathrm{m},\widehat{\mathrm{G}}_\mathrm{n}} + \frac{\mathrm{n(n+1)}}{2} \ .$$

However, two-sample rank statistics (*i.e.* functions of the $\mathrm{Rank}(X_i)$'s) form a very rich collection of statistics and this is by no means the sole possible choice to summarize the empirical ROC curve.

Definition. 1. (Two-sample linear rank statistics) *Let $\phi : [0,1] \to \mathbb{R}$ be a nondecreasing function. The two-sample linear rank statistics with 'score-generating function' $\phi(u)$ based on the random samples $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ is given by:*

$$\widehat{W}^{\phi}_{n,m} = \sum_{i=1}^{n} \phi\left(\frac{\mathrm{Rank}(X_i)}{N+1}\right). \tag{2.10}$$

The statistics (2.10) defined above are all distribution-free when $H = G$ and are, for this reason, particularly useful to detect differences between the distributions H and G and widely used to perform homogeneity tests in the univariate setup. Tabulating their distribution under the null assumption, they

can be used to design unbiased tests at certain levels α in $(0, 1)$. The choice of the score-generating function ϕ can be guided by the type of difference between the two distributions (*e.g.* in scale, in location) one possibly expects, and may then lead to locally most powerful testing procedures, capable of detecting 'small' deviations from the homogeneous situation. More generally, depending on the statistical test to perform, one may use particular function ϕ, Figure 2 shows classic score-generating functions broadly used for two-sample statistical tests (refer to [17]). One may refer to Chapter 9 in [31] or to Chapter 13 in [35] for an account of the (asymptotic) theory of rank statistics. In the present paper, two-sample linear rank statistics are used for a very different purpose, as empirical performance measures in bipartite ranking based on two independent multivariate samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$. The analysis of the bipartite ranking problem carried out in Section 4, based on the concentration inequalities established in Section 3, shows the relevance of evaluating the ranking performance of a scoring rule candidate $s(z)$ by computing a two-sample linear rank statistic based on the univariate samples obtained after scoring $\{s(\mathbf{X}_1), \ldots, s(\mathbf{X}_n)\}$ and $\{s(\mathbf{Y}_1), \ldots, s(\mathbf{Y}_m)\}$ and establishes statistical guarantees for the generalization capacity of scoring rules built by optimizing such an empirical criterion.

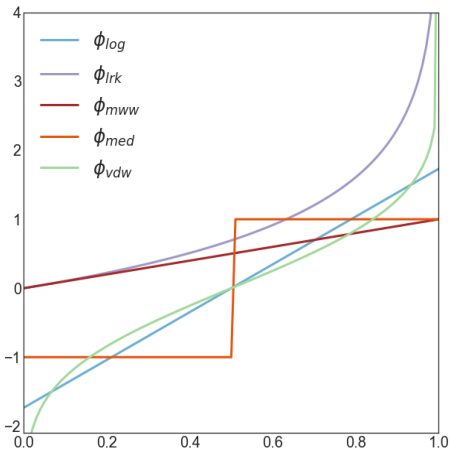


Figure 2: Curves of two-sample score-generating functions with the associated statistical test: Logistic test $\phi_{log}(u) = 2\sqrt{3}(u - 1/2)$ in blue, Logrank test $\phi_{lrk}(u) = -\log(1 - x)$ in purple, Mann-Whitney-Wilcoxon test $\phi_{mww}(u) = u$ in red, Median test $\phi_{med}(u) = \text{sgn}(u - 1/2)$ in orange, Van der Waerden test $\phi_{vdw}(u) = \Phi^{-1}(u)$ in green, Φ being the normal quantile function.

2.3 Bipartite Ranking as Maximization of Two-Sample Rank Statistics

As foreshadowed above, empirical performance measures in bipartite ranking should be unsurprisingly based on ranks. We propose here to evaluate empirically the ranking performance of any scoring function candidate $s(z)$ in $\mathcal{S}$ by

means of statistics of the type:

$$\widehat{W}^{\phi}_{n,m}(s) = \sum_{i=1}^{n} \phi\left(\frac{\mathrm{Rank}(s(\mathbf{X}_i))}{N+1}\right), \tag{2.11}$$

where $N = n + m$, $\phi : [0,1] \to \mathbb{R}$ is some Borelian nondecreasing function. This quantity is a two-sample linear rank statistic (see Definition 1) related to the score-generating function $\phi(u)$ and the samples $\{s(\mathbf{X}_1), \ldots, s(\mathbf{X}_n)\}$ and $\{s(\mathbf{Y}_1), \ldots, s(\mathbf{Y}_m)\}$. This statistic is invariant by increasing transform of the scoring function s, just like the (empirical) ROC curve and, as recalled in the previous section, it is a natural and common choice to quantify differences in distribution between the univariate samples $\{s(\mathbf{X}_1), \ldots, s(\mathbf{X}_n)\}$ and $\{s(\mathbf{Y}_1), \ldots, s(\mathbf{Y}_m)\}$, to evaluate to which extent the distribution of the first sample is stochastically larger than that of the second sample in particular. It consequently appears as legitimate to learn a scoring function s by maximizing the criterion (2.11). Whereas rigorous arguments are developed in Section 4, we highlight here that, for specific choices of the score-generating function ϕ, many relevant criteria considered in the ranking literature can be accurately approximated by statistics of this form:

- $\phi(u) = u$ - this choice leads to the celebrated Wilcoxon-Mann-Whitney statistic which is related to the empirical version of the AUC.
- $\phi(u) = u\mathbb{I}\{u \geq u_0\}$, for some $u_0 \in (0,1)$ - such a score-generating function corresponds to the local AUC criterion, introduced recently in [7]. Such a criterion is of interest when one wants to focus on the highest ranks.
- $\phi(u) = u^q$ - this is another choice which puts emphasis on high ranks but in a smoother way than the previous one. This is related to the q-norm push approach taken in [30]. However, we point out that the criterion studied in the latter work relies on a different definition of the rank of an observation. Namely, the rank of positive instances among negative instances (and not in the pooled sample) is used. This choice permits to use independence which makes the technical part much simpler, at the price of increasing the variance of the criterion.
- $\phi(u) = \phi_N(u) = c\left((N+1)u\right)\mathbb{I}\{u \geq k/(N+1)\}$ - this corresponds to the DCG criterion in the bipartite setup (see [11]), one of the 'gold standard quality measure' in information retrieval, when *grades* are binary. The $c(i)$'s denote the *discount factors*, $c(i)$ measuring the importance of rank i. The integer k denotes the number of top-ranked instances to take into account. Notice that, with our indexation, top positions correspond to the largest ranks and the sequence $\{c(i)\}_{i \leq N}$ should be chosen increasing.

Depending on the choice of the score-generating function ϕ, some specific patterns of the preorder induced by a scoring function $s(z)$ can be either enhanced by the criterion (2.11) or else completely disappear: for instance, the value of (2.11) is essentially determined by the possible presence of positive instances among top-ranked observations, when considering a score generating function ϕ that rapidly vanishes near 0 and takes much higher values near 1.

Investigating the performance of maximizers of the criterion (2.11) from a nonasymptotic perspective is however far from straightforward, due to the complexity of the latter (*i.e.* a sum of strongly dependent random variables). It

requires in particular to prove concentration inequalities for collections of two-sample linear rank statistics, indexed by classes of scoring functions of controlled complexity (*i.e.* of VC-type), referred to as two-sample rank processes throughout the article. It is the purpose of the next section to establish such results.

3 Concentration Inequalities for Two-Sample Rank Processes

This section is devoted to prove concentration bounds for collections of two-sample linear rank statistics (2.11), indexed by classes $\mathcal{S}_0 \subset \mathcal{S}$ of scoring functions. In order to study the fluctuations of (2.11) as the full sample size N increases, it is of course required to control the fraction of 'positive'/'negative' observations in the pooled dataset. Let $p \in (0,1)$ be the 'theoretical' fraction of positive instances. For $N \geq 1/p$, we suppose that $n = \lfloor pN \rfloor$ and $m = \lceil (1-p)N \rceil = N - n$. Define the mixture probability distribution $F = pG + (1-p)H$. For any $s \in \mathcal{S}$, the distribution of $s(\mathbf{X})$ (*i.e.* the image of G by s) is denoted by G_s, that of $s(\mathbf{Y})$ (*i.e.* the image of H by s) by H_s. We also denote by F_s the image of distribution F by s. For simplicity, the same notations are used to mean the related cumulative distribution functions. We also introduce their statistical versions $\widehat{G}_{s,n}(t) = (1/n)\sum_{i=1}^{n} \mathbb{I}\{s(\mathbf{X}_i) \leq t\}$ and $\widehat{H}_{s,m}(t) = (1/m)\sum_{j=1}^{m} \mathbb{I}\{s(\mathbf{Y}_j) \leq t\}$ and define:

$$\widehat{F}_{s,N}(t) = (n/N)\widehat{G}_{s,n}(t) + (m/N)\widehat{H}_{s,m}(t) \ . \tag{3.1}$$

Since $n/N \to p$ as N tends to infinity, the quantity above is a natural estimator of the *c.d.f.* F_s. Equipped with these notations, we can write:

$$\frac{1}{n}\widehat{W}^{\phi}_{n,m}(s) = \frac{1}{n}\sum_{i=1}^{n} \phi\left(\frac{N}{N+1}\widehat{F}_{s,N}(s(\mathbf{X}_i))\right). \tag{3.2}$$

Hence, the statistic (3.2) can be naturally seen as an empirical version of the quantity defined below, around which it fluctuates.

Definition. 2. *For a given score-generating function ϕ, the functional*

$$W_{\phi}(s) = \mathbb{E}[(\phi \circ F_s)(s(\mathbf{X}))] \ , \tag{3.3}$$

is referred to as the "W_{ϕ}-ranking performance measure".

Indeed, replacing $\widehat{F}_{s,N}(s(\mathbf{X}_i))$ in (3.2) by $F_s(s(\mathbf{X}_i))$ and taking next the expectation permits to recover (3.3). Observe in addition that, for $\phi(u) = u$, the quantity (3.3) is equal to AUC(s) (2.5) a soon as the distribution F_s is continuous. The next lemma reveals that the criterion (3.3) can be viewed as a scalar summary of the ROC curve.

Lemma. 3. *Let ϕ be a score-generating function. We have, for all s in $\mathcal{S}$,*

$$W_{\phi}(s) = \frac{1}{p}\int_0^1 \phi(u)du - \frac{1-p}{p}\int_0^1 \phi\left(p(1-\mathrm{ROC}(\mathrm{s},\alpha)) + (1-\mathrm{p})(1-\alpha)\right) \ d\alpha \ . \tag{3.4}$$

PROOF. Using the decomposition $F_s = pG_s + (1-p)H_s$, we are led to the following expression:

$$pW_\phi(s) = \int_0^1 \phi(u)\ du - (1-p)\mathbb{E}[(\phi \circ F_s)(s(\mathbf{Y}))]\ .$$

Then, using a change of variable, we get:

$$\mathbb{E}[(\phi \circ F_s)(s(\mathbf{Y}))] = \int_0^1 \phi(p(1-\mathrm{ROC(s},\alpha)) + (1-\mathrm{p})(1-\alpha))\ \mathrm{d}\alpha\ .$$

As revealed by Eq. (3.4), a score-generating function ϕ that takes much higher values near 1 than near 0 defines a criterion (3.3) that mainly summarizes the behavior of the ROC curve near the origin, *i.e.* the preorder on the set of instances with highest scores.

Below, we investigate the concentration properties of the process:

$$\left\{ \frac{1}{n}\widehat{W}_{n,m}^{\phi}(s) - W_\phi(s) \right\}_{s \in \mathcal{S}_0}. \tag{3.5}$$

As a first go, we prove, by means of linearization techniques, that two-sample linear rank statistics can be uniformly approximated by much simpler quantities, involving *i.i.d.* averages and two-sample U-statistics. This will be key to establish probability bounds for the maximal deviation:

$$\sup_{s \in \mathcal{S}_0} \left| \frac{1}{n}\widehat{W}_{n,m}^{\phi}(s) - W_\phi(s) \right|, \tag{3.6}$$

under adequate complexity assumptions for the class $\mathcal{S}_0$ of scoring functions considered and to study next the generalization ability of maximizers of the empirical criterion (3.2) in terms of W_ϕ-ranking performance. Throughout the article, all the suprema considered, such as (3.6), are assumed to be measurable and we refer to Chapter 2.3 in [36] for more details on the formulation in terms of outer measure/expectation that guarantees measurability.

Uniform approximation of two-sample linear rank statistics. Whereas statistical guarantees for Empirical Risk Minimization in the context of classification or regression can be directly obtained by means of classic concentration results for empirical processes (*i.e.* averages of *i.i.d.* random variables), the study of the fluctuations of the process (3.5) is far from straightforward, insofar as the terms averaged in (3.2) are not independent. For averages of non-*i.i.d.* random variables, the underlying statistical structure can be revealed by orthogonal projections onto the space of sums of *i.i.d.* random variables in many situations. This projection argument was the key for the study of empirical AUC maximization or that of within cluster point scatter, which involved U-processes, see [4] and [3]. In the case of U-statistics, this orthogonal decomposition is known as the *Hoeffding decomposition* and the remainder may be expressed as a degenerate U-statistic, see [20]. For rank statistics, a similar though more complex decomposition can be considered. We refer to [18] for a systematic use of the *projection method* for investigating the asymptotic properties of general statistics. From the perspective of ERM in statistical learning theory, through

the *projection method*, well-known concentration results for standard empirical processes and U-processes may carry over to more complex collections of random variables such as *two-sample linear rank processes*, as revealed by the approximation result stated below. It holds true under the following technical assumptions.

***Assumption* 1.** *Let $M > 0$. For all $s \in \mathcal{S}_0$, the random variables $s(\mathbf{X})$ and $s(\mathbf{Y})$ are continuous, with density functions that are twice differentiable and have Sobolev $\mathcal{W}^{2,\infty}$-norms[3] bounded by $M < +\infty$.*

***Assumption* 2.** *The score-generating function $\phi : [0,1] \mapsto \mathbb{R}$, is nondecreasing and twice continuously differentiable.*

***Assumption* 3.** *The class of scoring functions $\mathcal{S}_0$ is a* VC *class of finite* VC *dimension $\mathcal{V} < +\infty$.*

For the definition of VC classes of functions, one may refer to *e.g.* [36], see section 2.6.2 therein, and also recalled in Appendix section A.3. By means of the proposition below, the study of the fluctuations of the two-sample linear rank process (3.5) boils down to that of basic empirical processes.

Proposition. 4. *Suppose that Assumptions 1-3 are fulfilled. The two-sample linear rank process (3.5) can be linearized/decomposed as follows. For all $s \in \mathcal{S}_0$,*

$$\widehat{W}_{n,m}^{\phi}(s) = n\widehat{W}_{\phi}(s) + \left(\widehat{V}_n^X(s) - \mathbb{E}\left[\widehat{V}_n^X(s)\right]\right) + \left(\widehat{V}_m^Y(s) - \mathbb{E}\left[\widehat{V}_m^Y(s)\right]\right) + \mathcal{R}_{n,m}(s) \ , \tag{3.7}$$

where

$$\begin{aligned}
\widehat{W}_{\phi}(s) &= \frac{1}{n}\sum_{i=1}^{n} (\phi \circ F_s)\left(s(\mathbf{X}_i)\right) \ , \\
\widehat{V}_n^X(s) &= \frac{n-1}{N+1}\sum_{i=1}^{n} \int_{s(\mathbf{X}_i)}^{+\infty} (\phi' \circ F_s)(u) dG_s(u) \ , \\
\widehat{V}_m^Y(s) &= \frac{n}{N+1}\sum_{j=1}^{m} \int_{s(\mathbf{Y}_j)}^{+\infty} (\phi' \circ F_s)(u) dG_s(u) \ .
\end{aligned}$$

For any $\delta \in (0,1)$, there exist constants c_1, $c_3 > 0$, $c_2 \geq 1$, $c_4 > 6$, $c_5 > 3$, depending on ϕ and $\mathcal{V}$, such that

$$\mathbb{P}\left\{\sup_{s \in \mathcal{S}_0} |\mathcal{R}_{n,m}(s)| < t\right\} \geq 1 - \delta \ , \tag{3.8}$$

where $t = c_1 + c_2 \log(c_4/\delta)$ as soon as $N \geq (c_3/p)\log(c_5/\delta)$.

The proof of this linearization result is detailed in the Appendix section B.1 (refer to it for a description of the constants involved in the bound stated above).

[3] Recall that the Sobolev space $\mathcal{W}^{2,\infty}$ is the space of all Borelian functions $h : \mathbb{R} \to \mathbb{R}$ such that h and its first and second order weak derivatives h' and h'' are bounded almost-everywhere. Denoting by $||.||_\infty$ the norm of the Lebesgue space L_∞ of Borelian and essentially bounded functions, $\mathcal{W}^{2,\infty}$ is a Banach space when equipped with the norm $||h||_{2,\infty} = \max\{||h||_\infty, ||h'||_\infty, ||h''||_\infty\}$.

Its main argument consists in decomposing (3.2) by means of a Taylor expansion at order two of the score generating function $\phi(u)$ and applying next the Hájek orthogonal projection technique (recalled at length in the Introduction Lemma A.1 for completeness) to the component corresponding to the first order term. The quantity $\mathcal{R}_{n,m}(s)$ is then formed by bringing together the remainder of the Hájek projection and the component corresponding to the second order term of the Taylor expansion, while the probabilistic control of its order of magnitude is established by means of concentration results for (degenerate) one/two-sample U-processes (see the Appendix section A.4 for more details). It follows from decomposition (3.7) combined with triangular inequality that:

$$\sup_{s\in\mathcal{S}_0}\left|\frac{1}{n}\widehat{W}_{n,m}^{\phi}(s)-W_{\phi}(s)\right| \leq \sup_{s\in\mathcal{S}_0}\left|\widehat{W}_{\phi}(s)-W_{\phi}(s)\right| + \sup_{s\in\mathcal{S}_0}\frac{1}{n}\left|\widehat{V}_n^X(s)-\mathbb{E}\left[\widehat{V}_n^X(s)\right]\right| + \sup_{s\in\mathcal{S}_0}\frac{1}{n}\left|\widehat{V}_m^Y(s)-\mathbb{E}\left[\widehat{V}_m^Y(s)\right]\right| + \sup_{s\in\mathcal{S}_0}\frac{1}{n}\left|\mathcal{R}_{n,m}(s)\right|. \quad (3.9)$$

Hence, nonasymptotic bounds for the maximal deviation of the process (3.5) can be deduced from concentration inequalities for standard empirical processes, as shall be seen below. Before this, a few comments are in order.

Remark 1. (On the complexity assumption) *We point out that alternative complexity measures could be naturally considered, such as those based on Rademacher averages, see e.g. [22]. However, as different types of stochastic process (i.e. empirical process, degenerate one-sample U-process and degenerate two-sample U-process) are involved in the present nonasymptotic study, different types of Rademacher complexities (see e.g. [4]) should be introduced to control their fluctuations as well. For the sake of simplicity, the concept of* VC*-type class of functions is used here.*

Remark 2. (Smooth score-generating functions) *The subsequent analysis is restricted to the case of smooth score-generating functions for simplification purposes. We nevertheless point out that, although one may always build smooth approximants of irregular score generating functions, the theoretical results established below can be directly extended to non-smooth situations, at the price of a significantly greater technical complexity.*

The theorem below provides a concentration bound for the two-sample rank process (3.5). The proof is based on the uniform approximation result precedingly established, refer to the Appendix section B.3 for technical details.

Theorem. 5. *Suppose that the assumptions of Proposition 4 are fulfilled. Then, there exist constants $C_1,\ C_3>0,\ C_2\geq 24$, depending on ϕ, $\mathcal{V}$ and $C_4\geq C_1$ depending on ϕ, such that:*

$$\mathbb{P}\left\{\sup_{s\in\mathcal{S}_0}\left|\frac{1}{n}\widehat{W}_{n,m}^{\phi}(s)-W_{\phi}(s)\right|>t\right\}\leq C_2 e^{-pC_3Nt^2}, \quad (3.10)$$

as soon as $C_1/\sqrt{pN}\leq t\leq C_4\min(p,1-p)$.

The concentration inequalities stated above are extensively used in the next section to study the ranking bipartite learning problem, when formulated as W_ϕ-ranking performance maximization.

4 Performance of Maximizers of Two-Sample Rank Statistics in Bipartite Ranking

This section provides a theoretical analysis of bipartite ranking methods, based on maximization of the empirical ranking performance measure (2.11). While the concentration inequalities established in the previous section are the key technical tools to derive nonasymptotic bounds for the deficit of W_ϕ-ranking performance measure of empirical maximizers, we start by showing that the criterion (3.3) is relevant to measure ranking performance, whatever the score generating function ϕ is chosen, beyond the examples listed in Subsection 2.3.

Optimal elements. The next result states that optimal scoring functions do maximize the W_ϕ-ranking performance and form a collection that coincides with the set $\mathcal{S}_\phi^*$ of maximizers of (3.3), provided that the score-generating function ϕ is strictly increasing on $(0, 1)$.

Proposition. 6. *Let ϕ be a score-generating function. The assertions below hold true.*

(i) For all $(s,\ s^*) \in \mathcal{S} \times \mathcal{S}^*$, we have $W_\phi(s) \leq W_\phi(s^*) = W_\phi^*$, where $W_\phi^* \overset{def}{=} W_\phi(\Psi)$.

(ii) Assuming in addition that the score-generating function ϕ is strictly increasing on $(0, 1)$, we have: $\mathcal{S}_\phi^* = \mathcal{S}^*$.

The proof immediately results from (3.4) combined with the fact that the ROC curve of increasing transforms of the likelihood ratio $\Psi(z)$ dominates everywhere any other ROC curve, as recalled in Section 2.1: $\forall (s,\ s^*) \in \mathcal{S} \times \mathcal{S}^*, \forall \alpha \in (0,\ 1)$, $\mathrm{ROC}(s, \alpha) \leq \mathrm{ROC}(s^*, \alpha) = \mathrm{ROC}^*(\alpha)$. Details are left to the reader.

Remark 3. (On plug-in ranking rules) *Theoretically, a possible approach to bipartite ranking is the* plug-in *method ([12]), which consists of using an estimate $\hat{\Psi}$ of the likelihood function as a scoring function. As shown by the subsequent bound, when ϕ is differentiable with a bounded derivative, when $\hat{\Psi}$ is close to Ψ in the L_1-sense, it leads to a nearly optimal ordering in terms of W-ranking criterion:*

$$W_\phi^* - W_\phi\left(\widehat{\Psi}\right) \leq (1-p)||\phi'||_\infty \mathbb{E}[|\widehat{\Psi}(\mathbf{X}) - \Psi(\mathbf{X})|] \ .$$

However, the bound above may be loose and the plug-in approach faces computational difficulties when dealing with high-dimensional data, see [16], which provide the motivation for exploring algorithms based on W_ϕ-ranking performance maximization.

Remark 4. (Alternative probabilistic framework) *We point out that the present analysis can be extended to the alternative setup, where, rather than assuming that two samples of sizes n and m, 'positive' and 'negative', are available for the learning tasks considered in this paper, the i.i.d. observations Z are supposed to come with a random label Y either equal to $+1$ or else to -1, indicating whether Z is distributed according to G or H. If p denotes the probability that the label Y is equal to 1, the number n of positive observations among a training sample of size N is then random, distributed as a binomial of size N with parameter p.*

Consider any maximizer of the empirical W_ϕ-ranking performance measure over a class $\mathcal{S}_0 \subset \mathcal{S}$ of scoring rules:

$$\widehat{s} \in \arg\max_{s \in \mathcal{S}_0} \widehat{W}^{\phi}_{n,m}(s) \ . \tag{4.1}$$

Since we obviously have:

$$W^*_\phi - W_\phi(\widehat{s}) \leq 2 \sup_{s \in \mathcal{S}_0} \left| \frac{1}{n} \widehat{W}^{\phi}_{n,m}(s) - W_\phi(s) \right| + \left(W^*_\phi - \sup_{s \in \mathcal{S}_0} W_\phi(s) \right), \tag{4.2}$$

the control of deficit of W-ranking performance of empirical maximizers of (3.2) can be deduced from the concentration properties of the process (3.5).

4.1 Generalization Error Bounds and Model Selection

The corollary below describes the generalization capacity of scoring rules based on empirical maximization of W_ϕ-ranking performance criteria. It straightforwardly results from Theorem 5 combined with the bound (4.2).

Corollary. 7. *Let $\hat{s}$ be an empirical W_ϕ-ranking performance maximizer over the class $\mathcal{S}_0$, i.e. $\hat{s} \in \arg\max_{s \in \mathcal{S}_0} \widehat{W}^{\phi}_{n,m}(s)$. Under the assumptions of Proposition 4, for any $\delta \in (0,1)$, we have with probability at least $1-\delta$:*

$$W^*_\phi - W_\phi(\hat{s}) \leq 2\sqrt{\frac{\log(C_2/\delta)}{pC_3N}} + \left(W^*_\phi - \sup_{s \in \mathcal{S}_0} W_\phi(s) \right), \tag{4.3}$$

as soon as $N \geq 1/(p \min(p, 1-p)^2 C_3 C_4^2) \log(C_2/\delta)$ and $\delta \leq C_2 e^{-C_1^2 C_3}$ where the constants C_i, $i \leq 4$, being the same as those involved in Theorem 5.

The result above establishes that maximizers of the empirical criterion (2.11) achieve a classic learning rate bound of order $O_{\mathbb{P}}(1/\sqrt{N})$ when based on a training data set of size N, just like in standard classification, see *e.g.* [12]. Refer to the Appendix section B.4 for the proof of an additional result, that provides a bound in expectation for the deficit of W_ϕ-ranking performance measure, similar to that established in the subsequent analysis, devoted to the model selection issue.

Model selection by complexity penalization. We have investigated the issue of approximately recovering the best scoring rule in a given class $\mathcal{S}_0$ in the sense of the W_ϕ-ranking performance measure (3.3), which is satisfactory only when the minimum achieved over $\mathcal{S}_0$ is close to W^*_ϕ of course. We now address the problem of model selection, that is the problem of selecting a good scoring function from one of a collection of VC classes $\mathcal{S}_k$, $k \geq 1$. A model selection method is a data-based procedure that aims at achieving a trade-off regarding two contradictory objectives, *i.e.* at finding a class $\mathcal{S}_k$ rich enough to include a reasonable approximant of an element of $\mathcal{S}^*$, while being not too complex so that the performance of the empirical minimizer over it $\hat{s}_k = \arg\max_{s \in \mathcal{S}_k} \widehat{W}^{\phi}_{n,m}(s)$ can be statistically guaranteed. We suppose that all class candidates $\mathcal{S}_k$, $k \geq 1$, fulfill the assumptions of Proposition 4 and denote by $\mathcal{V}_k$ the VC dimension of the class $\mathcal{S}_k$. Various model selection techniques, based on (re-)sampling or data-splitting procedures, could be naturally considered for this purpose. Here,

in order to avoid overfitting, we focus on a complexity regularization approach, of which study can be directly derived from the rate bound analysis previously carried out, that consists in substracting to the empirical ranking performance measure the penalty term (increasing with $\mathcal{V}_k$) given by:

$$\mathrm{pen}(N,k) = B_1\sqrt{\frac{\mathcal{V}_k}{pN}} + \sqrt{\frac{2C\log k}{p^2N}}\ , \tag{4.4}$$

for $pN \geq B_2\mathcal{V}_k$ where the constants B_1 and B_2 are those involved in Proposition 21 and $C = 6(\|\phi\|_\infty^2 + 9\|\phi'\|_\infty^2 + 9\|\phi''\|_\infty^2)$. The scoring function selected maximizes the penalized empirical ranking performance measure, it is $\hat{s}_{\hat{k}}(z)$ where:

$$\hat{k} = \arg\max_{k\geq 1}\left\{\frac{1}{n}\widehat{W}_{n,m}^{\phi}(s) - \mathrm{pen}(N,k)\right\}\ . \tag{4.5}$$

The result below shows that the scoring rule $\hat{s}_{\hat{k}}$ nearly achieves the expected deficit of W_ϕ-ranking performance that would have been attained with the help of an oracle, revealing the model minimizing $W_\phi^* - \mathbb{E}[W_\phi(\hat{s}_k)]$.

Proposition. 8. *Suppose that the assumptions of Proposition 4 are fulfilled for any class $\mathcal{S}_k$ with $k \geq 1$ and that $\sup_{k\geq 1}\mathcal{V}_k < +\infty$. Then, we have:*

$$W_\phi^* - \mathbb{E}\left[W_\phi(\hat{s}_{\hat{k}})\right] \leq \min_{k\geq 1}\left\{2pen(N,k) + \left(W_\phi^* - \sup_{s\in\mathcal{S}_k} W_\phi(s)\right)\right\} + 2\sqrt{\frac{C}{p^2N}}\ , \tag{4.6}$$

as soon as $pN \geq B_2\sup_{k\geq 1}\mathcal{V}_k$, where the constant $B_2 > 0$ is the same as that involved in Proposition 21 and $C = 6(\|\phi\|_\infty^2 + 9\|\phi'\|_\infty^2 + 9\|\phi''\|_\infty^2)$.

Refer to the Appendix section B.5 for the technical proof.

4.2 Kernel Regularization for Ranking Performance Maximization

Many successful algorithmic approaches to statistical learning (*e.g.* boosting, support vector machines, neural networks) consist in smoothing the empirical risk/performance functional to be optimized, so as to use computationally feasible techniques based on gradient descent/ascent methods. Concerning the empirical criterion (2.11), although one may choose a regular score generating function ϕ (*cf* Remark 2), smoothness issues arise when replacing F_s in (3.3) by the raw empirical *c.d.f.* (3.1). A classic remedy involves using a kernel-smoothed version of the empirical *c.d.f.* instead. Let $K : \mathbb{R} \to \mathbb{R}$ be a second-order Parzen-Rosenblatt kernel *i.e.* a Borelian symmetric function, integrable *w.r.t.* the Lebesgue measure such that $\int K(t)dt = 1$ and $\int t^2K(t)dt < +\infty$. Precisely, for any $h > 0$ and all $t \in \mathbb{R}$, define the smoothed approximation of the *c.d.f.* $F_s(t)$:

$$\widetilde{F}_{s,h}(t) = \int_{\mathbb{R}} \kappa\left(\frac{t-u}{h}\right) F_s(du)\ , \tag{4.7}$$

where $\kappa(t) = \int_{-\infty}^{t} K(u)du$ and $h > 0$ is the bandwidth that determines the degree of smoothing, see *e.g.* [27]. The uniform integrated error $\sup_{s\in\mathcal{S}_0}\int|\widetilde{F}_{s,h}(t)-$

$F_s(t)|dt$ is shown to be of order $O(h^2)$ under the assumptions recalled below, see [21].

***Assumption* 4.** *Let $R > 0$. For all s in $\mathcal{S}_0$, the cumulative distribution function F_s is differentiable with derivative f_s such that $\int (f_s'(t))^2 dt \leq R$.*

***Assumption* 5.** *The kernel function K is of the form $K_1 \circ K_2$, where K_1 is a function of bounded variation and K_2 is a polynomial.*

Notice that Assumption 4 is fulfilled as soon as Assumption 1 is satisfied with $R \geq M$. The statistical counterpart of (4.7) is then:

$$\widehat{F}_{s,N,h}(t) = \frac{1}{N}\sum_{i=1}^{n} \kappa\left(\frac{t - s(\mathbf{X}_i)}{h}\right) + \frac{1}{N}\sum_{j=1}^{m} \kappa\left(\frac{t - s(\mathbf{Y}_j)}{h}\right). \tag{4.8}$$

A smooth version of the theoretical criterion (3.3) is given by:

$$\widetilde{W}_{\phi,h}(s) = \mathbb{E}[(\phi \circ \widetilde{F}_{s,h})(s(\mathbf{X}))] \ , \tag{4.9}$$

for all $s \in \mathcal{S}$ and an empirical version of the latter is $\widehat{W}^{\phi}_{n,m,h}(s)/n$, where:

$$\widehat{W}^{\phi}_{n,m,h}(s) = \sum_{i=1}^{n} (\phi \circ \widehat{F}_{s,N,h})(s(\mathbf{X}_i)) \ . \tag{4.10}$$

For any maximizer $\tilde{s}$ of (4.10) over the class $\mathcal{S}_0$ of scoring function candidates, we almost-surely have:

$$W^*_{\phi} - W_{\phi}(\tilde{s}) \leq 2 \sup_{s \in \mathcal{S}_0} \left| \frac{1}{n}\widehat{W}^{\phi}_{n,m,h}(s) - \widetilde{W}_{\phi,h}(s) \right| + \sup_{s \in \mathcal{S}_0} \left| \widetilde{W}_{\phi,h}(s) - W_{\phi}(s) \right| + \left\{ W^*_{\phi} - \sup_{s \in \mathcal{S}_0} W_{\phi}(s) \right\} \ . \tag{4.11}$$

This decomposition is similar to that obtained in (4.2) for maximizers of the criterion (2.11), apart from the additional bias term. Since the latter can be shown to be of order $O(h^2)$ under appropriate regularity conditions and the first term on the right hand side of the equation above can be controlled like in Theorem 5, one may bound the deficit of W_{ϕ}-ranking performance measure of $\tilde{s}$ as follows.

Proposition. 9. *Suppose that the assumptions of Proposition 4 are fulfilled, as well as Assumptions 4 and 5. Let $\tilde{s}$ be any maximizer of the smoothed criterion (4.10) over the class $\mathcal{S}_0$. Then, for any $\delta \in (0,1)$, there exist constants C_1, $C_3 > 0$, $C_2 \geq 24$ depending on ϕ, K, R, $\mathcal{V}$, $C_4 \geq C_1$, and $C_5 > 0$ is a constant depending on ϕ, K and R, such that we have with probability at least $1-\delta$:*

$$W^*_{\phi} - W_{\phi}(\tilde{s}) \leq 2\sqrt{\frac{\log(C_2/\delta)}{pC_3N}} + C_5 h^2 + \left\{ W^*_{\phi} - \sup_{s \in \mathcal{S}_0} W_{\phi}(s) \right\}, \tag{4.12}$$

as soon as $N \geq 1/(p \min(p, 1-p)^2 C_3 C_4^2) \log(C_2/\delta)$ and $\delta \leq C_2 e^{-C_1^2 C_3}$.

The proof is detailed in the Appendix section B.6.

5 Numerical Experiments

It is the purpose of this section to illustrate empirically various points highlighted by the theoretical analysis previously carried out: in particular, the capacity of ranking rules obtained by maximization of empirical W_ϕ-performance measures to generalize well and the impact of the choice of the score generating function ϕ on ranking performance from the perspective of ROC analysis. Some practical issues, concerning the maximization of smoothed versions of the empirical W_ϕ-performance criterion, are also discussed through numerical experiments. Additional experimental results can be found in the Appendix section C. All experiments displayed in this article can be reproduced using the code available at https://github.com/MyrtoLimnios/grad_2sample.

5.1 A Gradient-Based Algorithmic Approach

We start by describing the gradient ascent method (GA) used in the experiments in order to maximize the smoothed criterion (4.10) obtained by kernel smoothing over the class of scoring functions $\mathcal{S}_0$ considered, as proposed in section 4.2, see Algorithm 1. Precisely, suppose that $\mathcal{S}_0$ is a parametric class, indexed by a parameter space $\Theta \subset \mathbb{R}^d$ with $d \geq 1$ say: $\mathcal{S}_0 = \{s_\theta : \mathcal{X} \to \mathbb{R}, \quad \theta \in \Theta\}$. Assume also that, for all $z \in \mathcal{Z}$, the mapping $\theta \in \Theta \mapsto s_\theta(z)$ is of class $\mathcal{C}^1$ (*i.e.* continuously differentiable) with gradient $\partial_\theta s_\theta(z)$ and that the score-generating function ϕ fulfills Assumption 2. The gradient of the smoothed ranking performance measure of s_θ *w.r.t.* to the parameter θ, is given by: for all $\theta \in \Theta$, $h > 0$,

$$\nabla_\theta \left(\widehat{W}^{\phi}_{n,m,h}(s_\theta) \right) = \sum_{i=1}^{n} \phi' \left(\widehat{F}_{s_\theta,N,h}(s_\theta(\mathbf{X}_i)) \right) \nabla_\theta \left(\widehat{F}_{s_\theta,N,h}(s_\theta(\mathbf{X}_i)) \right) \ , \quad (5.1)$$

where the gradient of $\widehat{F}_{s_\theta,N,h}(s_\theta(z))$ *w.r.t.* to θ is:

$$\begin{aligned} \nabla_\theta \left(\widehat{F}_{s_\theta,N,h}(s_\theta(z)) \right) = \frac{1}{Nh} \sum_{i=1}^{n} K \left(\frac{s_\theta(z) - s_\theta(\mathbf{X}_i)}{h} \right) (\partial_\theta s_\theta(z) - \partial_\theta s_\theta(\mathbf{X}_i)) \\ + \frac{1}{Nh} \sum_{j=1}^{m} K \left(\frac{s_\theta(z) - s_\theta(\mathbf{Y}_j)}{h} \right) (\partial_\theta s_\theta(z) - \partial_\theta s_\theta(\mathbf{Y}_j)) \ , \quad (5.2) \end{aligned}$$

for any $z \in \mathcal{Z}$, using the fact that $\kappa' = K$.

Algorithm 1: Gradient Ascent for W-ranking performance maximization

Data: Independent *i.i.d.* samples $\{\mathbf{X}_i\}_{i\leq n}$ and $\{\mathbf{Y}_j\}_{j\leq m}$.
Input: Score-generating function ϕ, kernel K, bandwidth $h>0$, number of iterations $T\geq 1$, step size $\eta>0$.
Result: Scoring rule $s_{\widehat{\theta}_{n,m}}(z)$.

Choose the initial point $\theta^{(0)}$ in Θ;
for $t=0,\ldots,\ T-1$ **do**
  compute the gradient estimate $\nabla_\theta\left(\widehat{W}^{\phi}_{n,m,h}(s_{\theta^{(t)}})\right)$;
  update the parameter $\theta^{(t+1)}=\theta^{(t)}+\eta\nabla_\theta\left(\widehat{W}^{\phi}_{n,m,h}(s_{\theta^{(t)}})\right)$;
end
Set $\widehat{\theta}_{n,m}=\theta^{(T)}$.

In practice, the iterations are continued until the order of magnitude of the variations $||\theta^{(t+1)}-\theta^{(t)}||$ becomes negligible. Then, the approximate maximum $s_{\widehat{\theta}_{n,m}}(z)$ output by Algorithm 1 is next used to rank test data. Averages over several Monte-Carlo replications are computed in order to produce the results displayed in Subsection 5.3.

5.2 Synthetic Data Generation

We now describe the data generating models used in the simulation experiments, as well as the parametric class of scoring functions, which the learning algorithm previously described is applied to.

Score-generating functions. To illustrate the importance of the function ϕ in the W_ϕ-performance ranking criterion, we successively consider $\phi_{MWW}(u)=u$ (MWW), $\phi_{Pol}(u)=u^q$, $q\in\mathbb{N}^*$ (Pol, [30]) and $\phi_{RTB}(u)=\text{SoftPlus}(u-u_0)+u_0\text{Sigmoid}(u-u_0)$, $u_0\in(0,1)$ (RTB, smoothed version of [7]), where the activation functions are defined by: $\text{SoftPlus}(u)=(1/\beta)\log(1+\exp(\beta u))$ and $\text{Sigmoid}(u)=1/(1+\exp(-\lambda u))$, $\beta,\lambda>0$ being hyperparameters to fit and control the derivative's slope.

Probabilistic models. Two classic two-sample statistical models are used here, namely the location and the scale models, where both samples are drawn from multivariate Gaussian distributions. We denote by $S_d^+(\mathbb{R})$ the set of positive definite matrices of dimension $d\times d$, by $\mathbb{I}_d$ the identity matrix.

Location model. Inspired by the optimality properties of linear rank statistics regarding shift detection in the univariate setup (*cf* Subsection 2.2), the model considered stipulates that $\mathbf{X}\sim\mathcal{N}_d(\mu_X,\Sigma)$ and $\mathbf{Y}\sim\mathcal{N}_d(\mu_Y,\Sigma)$ where $\Sigma\in S_d^+(\mathbb{R})$ and the mean/location parameters μ_X and μ_Y differ. The Algorithm 1 is implemented here with $\mathcal{Z}=\mathbb{R}^d=\Theta$ and $\mathcal{S}_0=\{s_\theta(\cdot)=\langle\cdot,\theta\rangle,\ \ \theta\in\Theta\}$ as class of scoring functions, where $\langle\cdot,\cdot\rangle$ denotes the Euclidean scalar product on the feature space $\mathbb{R}^d$, and consequently exhibits no bias caused by the model. Indeed, by computing the loglikelihood ratio, one may easily check that the function $\langle\theta^*,\ \cdot\rangle$, where $\theta^*=\Sigma^{-1}(\mu_X-\mu_Y)$, is an optimal scoring function for the related

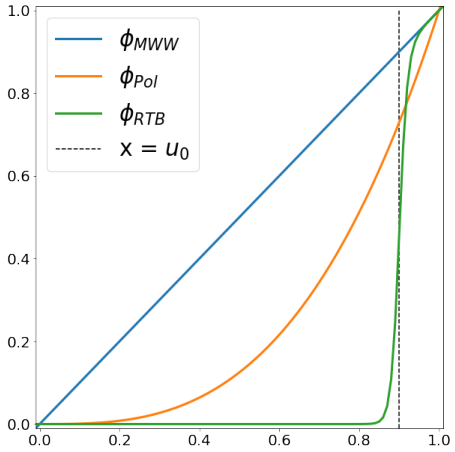


Figure 3: Curves of the three score-generating functions under study: $\phi_{MWW}(u) = u$ in blue, $\phi_{Pol}(u) = u^3$ in orange, $\phi_{RTB}(u) = \text{SoftPlus}(u - u_0) + u_0\text{Sigmoid}(u - u_0)$ the smoothed version of $u \mapsto u\mathbb{I}\{u \geq u_0\}$ in green, vertical line at $x = u_0$ in black.

bipartite ranking problem. Denoting by $\Delta(t) = (1/\sqrt{2\pi})\int_{-\infty}^{t} \exp(-u^2/2)du$, $t \in \mathbb{R}$, the *c.d.f.* of the centered standard univariate Gaussian distribution, one may immediately check that the optimal ROC curve is given by:

$$\forall \alpha \in (0,1), \quad \text{ROC}^*(\alpha) = 1 - \Delta\left(\Delta^{-1}(1-\alpha) + \sqrt{(\mu_X - \mu_Y)^T \Sigma^{-1}(\mu_X - \mu_Y)}\right).$$

Three levels of difficulty are tested through the implementations Loc1, Loc2 and Loc3. The nearly diagonal covariance matrix of the three models has its eigenvalues in $[0.5, 1.5]$ and $\mu_X = (1+\varepsilon)\mu_Y$ with $\varepsilon = 0.10$ (*resp.* $\varepsilon = 0.20$ and $\varepsilon = 0.30$) for Loc1 (*resp.* Loc2 and Loc3). The empirical ROC curves over the test pooled samples and additional curves are depicted in Fig. 10, 4, 11 for *resp.* Loc1, 2 and 3. The averaged ROC curves and the *best* one are gathered for the three models in Fig. 5. In Fig. 6, the evolution of the averaged empirical value of the W_ϕ-criteria on the train set during the algorithm is computed. Fig. 14 shows the results for Loc2 and 3 for three different parameters of the RTB model with $u_0 \in \{0.70,\ 0.90,\ 0.95\}$.

Scale model. Consider now the situation where $\mathbf{X} \sim \mathcal{N}_d(\mu, \Sigma_X)$ and $\mathbf{Y} \sim \mathcal{N}_d(\mu, \Sigma_Y)$, the distributions having the same location vector $\mu \in \mathbb{R}^d$ but different scale parameters Σ_X and Σ_Y in $S_d^+(\mathbb{R})$. The Algorithm 1 is implemented with $\mathcal{Z} = \mathbb{R}^d$, $\Theta = S_d^+(\mathbb{R})$ and $\mathcal{S}_0 = \{s_\theta(z) = \langle z, \theta^{-1}z\rangle, \quad \text{for all} \quad z \in \mathcal{Z}, \quad \theta \in \Theta\}$, with the notations previously introduced. By computing the likelihood ratio, one immediately checks that $s_{\theta^*}(\cdot)$, with $\theta^* = \Sigma_X^{-1} - \Sigma_Y^{-1}$, is an optimal scoring function for the related scale model. For models Scale1, Scale2 and Scale3, observations are centered, $\Sigma_Y = \mathbb{I}_d$ and $\Sigma_X = \mathbb{I}_d + (\varepsilon/d)H$, where ε is taken equal to 0.70, 0.80 and 0.90 respectively and H a $d \times d$ symmetric matrix with real entries such that all the eigenvalues of $\Sigma_X \in S_d^+(\mathbb{R})$ are close to 1. Similar to the location models, the empirical ROC curves over the test pooled samples and additional curves are depicted in Fig. 7, 12, 13 for *resp.* Scale1, 2

and 3. The averaged ROC curves and the *best* one are gathered for the three models in Fig. 8. In Fig. 9, the evolution of the averaged empirical value of the W_ϕ-criteria on the train set during the Algorithm is computed. Fig. 15 shows the results for Scale2 for three different parameters of the RTB model with $u_0 \in \{0.60,\ 0.70,\ 0.80\}$.

Experimental parameters. In all the experiments below, the pooled train sample is balanced, *i.e.* $n = m = 150$ and the dimension of the feature space is $d = 15$. Similarly for the test sample with $n = m = 10^6$ and $d = 15$. Concerning the score-generating functions, we consider $q = 3$ (Pol) and $u_0 = 0.9$ (RTB). We use the Gaussian smoothing kernel $K(u) = (1/\sqrt{2\pi})\exp\{-u^2/2\}$ with a bandwidth $h \sim N^{-1/5}$, yielding an (asymptotically) optimal trade-off between bias and variance. Algorithm 1 is implemented with $T = 50$ and a learning step size η of order $1/\sqrt{T}$. For each model, $B = 50$ Monte-Carlo replications of the train pooled sample. Based on the latter, a standard deviation for the test average ROC curve is computed for each model.

Evaluation of the criteria. In order to evaluate the performance of the scoring function produced by an early-stopped version of Algorithm 1 depending on the score-generating function chosen, it is used to score the test sample and the corresponding ROC curves and its average are compared to those of the optimal scoring function $s_{\theta^*}(z)$. Also we consider the *best*/*worst* curves in the sense of *resp.* the minimization/maximization of the generalization error of the set of ROC curves obtained computed over the test pooled sample. Particular attention is paid to the behavior of these curves near the origin, which reflects the ranking performance for the instances with highest score values.

5.3 Results and Discussion

We now analyze the experimental results, by commenting on the test ROC curves obtained after learning the scoring functions, using the early-stopped version of the Algorithm 1 described above, that maximize the chosen (smoothed variant of the) W_ϕ-performance measure: MWW, Pol and RTB. We compare them with ROC*. All the experiments were run using Python.

For both the location and scale models, we ran the algorithm for three increasing levels of difficulty defined by the decreasing value of the parameter ε. Figures 5 (location) and 8 (scale) show that the three methods (MWW, Pol, RTB) learn an empirical parameter $\widehat{\theta}_{n,m}$ such that the corresponding ROC curve gets close to ROC* (red curves) and the more ε increases and the more the scoring rule learned generalizes well. Fig. 6 (location) and 9 (scale) reveal the monotonicity of the evolution of the empirical criteria, as the number of iterative steps of Algorithm 1 increases. Unsurprisingly, all the results show an increasing ability to learn a scoring function that maximizes the three W_ϕ-performance measures, as ε increases (*i.e.* when the distribution G and H are significantly more different from each other).

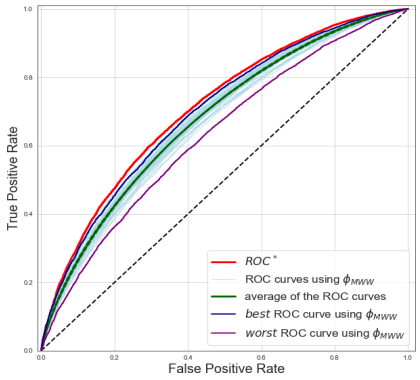
Analyzing the average of the empirical ROC curves obtained, MWW performs better for the location model as its corresponding curve converges faster to

ROC^* for all ε. This phenomenon was expected due to the well-known high power of the related Mann-Whitney-Wilcoxon test statistic in this modeling. The aggregated ROC curve for the Pol method also performs well, while RTB's presents a low performance compared to MWW, see Fig. 5. Indeed, considering only the best ranked observations at each iteration in the learning procedure, does not always achieve a good scoring parameter and is enhanced by the early-stopped rule. It results in a higher variance and a larger spectrum of the empirical curves both at the same time, see the light blue curves in Fig. 4.3. and 11.3. (Loc2 and Loc3). The slow convergence for the RTB method is illustrated with Loc1, where almost both samples are blended/coincide, for which only the ROC curves above the diagonal were kept. For the scale model, the aggregated ROC curves are comparable for the three methods with a slightly higher performance obtained by RTB and we note the faster convergence of the algorithm for this model, see Fig. 9.

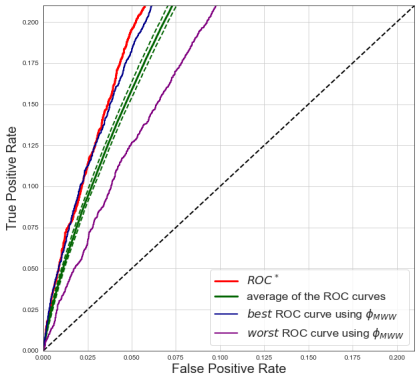
Looking at the *best* ROC curves (dark blue lines), defined as those obtained by the scoring function minimizing the generalization error for each criterion, RTB yields to a scoring function that generalizes best for most of the models. In particular, when focussing on the 'best' instances in the learning procedure, the obtained empirical scoring functions have higher performance at the beginning of the ROC curves, see the zoomed plots. Also, choosing the optimal proportion $1 - u_0$ of observations to consider for the score-generating function results in different performance measures. Figure 14 gathers the resulting plots for models Loc2 and 3 with u_0 in $\{0.7, 0.9, 0.95\}$ while Fig. 15 depicts the scale model 2 with u_0 in $\{0.6, 0.7, 0.8\}$ and a higher number of loops $T = 70$. Considering the *best* ROC curves for all models shows that when u_0 tends to one, the beginning of the curve is accurately learned. Incidentally, note that the proportion of observations considered has to be large enough, so that the optimization algorithm performs well.

6 Conclusion

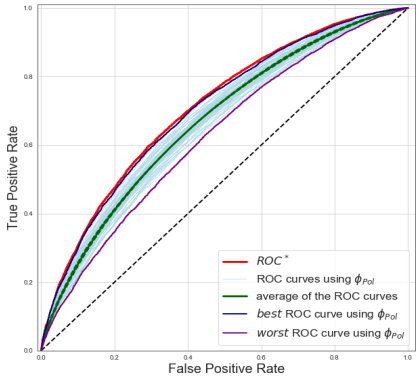
This article argues that two-sample linear rank statistics provide a very flexible and natural class of empirical performance measures for bipartite ranking. We have showed that it encompasses in particular well-known criteria used in medical diagnosis and information retrieval and proved that, in expectation, these criteria are maximized by optimal scoring functions and put the emphasis on specific parts of their ROC curves, depending on the score generating function involved in the criterion considered. We have established concentration results for collections of such statistics, referred to as *two-sample rank processes* here, under general assumptions and have deduced from them statistical learning guarantees for the maximizers of such ranking criteria in the form of a generalization bound of order $O_{\mathbb{P}}(1/\sqrt{N})$, where N means the size of the pooled training sample. Algorithmic issues concerning practical maximization have also been investigated and we have displayed numerical results supporting the theoretical analysis carried out.



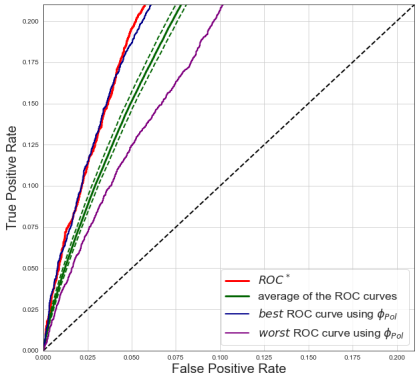
1.a. $\phi_{MWW}(u) = u$



1.a. $\phi_{MWW}(u) = u$



2.b. $\phi_{Pol}(u) = u^3$



2.b. $\phi_{Pol}(u) = u^3$

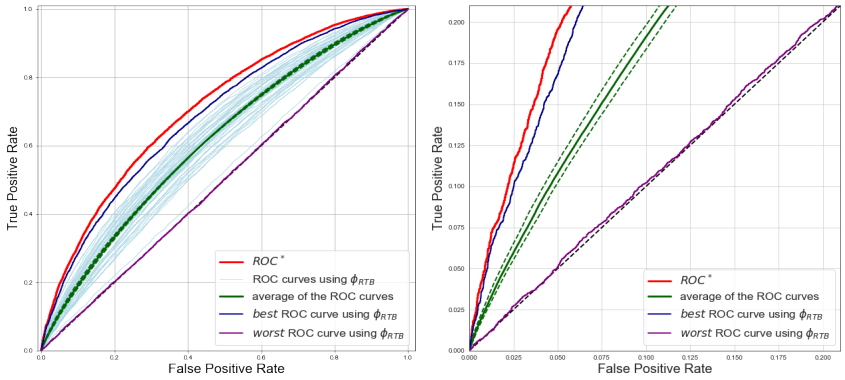
A Definitions and Preliminary Results

For the sake of clarity, crucial concepts and results extensively used in the technical analysis subsequently carried out are first recalled.

A.1 Hájek Projection Method

The Hájek projection method introduced in the seminal contribution [18] aims at decomposing (linearizing) any (possibly complex) square integrable statistic based on independent observations, so as to express it as an average of independent *r.v.*'s plus an uncorrelated term. The proof of Proposition 4 crucially relies on this technique. For completeness, it is described in the following lemma, one may refer to Chapter 11 in [35] for further details.

Lemma. 10. (HÁJEK PROJECTION, [18]) *Let $Z_1, \ldots, Z_n$ be independent r.v.'s and $T_n = T_n(Z_1, \ldots, Z_n)$ be a real-valued square integrable statistic. The Hájek projection of T_n is defined as $\widehat{T}_n = \sum_{i=1}^n \mathbb{E}[T_n \mid Z_i] - (n-1)\mathbb{E}[T]$. It is the orthogonal projection of the square integrable r.v. T_n onto the subspace of all variables of the form $\sum_{i=1}^n g_i(Z_i)$, for arbitrary measurable functions g_i s.t. $\mathbb{E}[g_i^2(Z_i)] < +\infty$.*



3.c. $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$ 3.c. $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$

Figure 4: Empirical ROC curves and average ROC curve for Loc2 ($\varepsilon = 0.20$). Samples are drawn from multivariate Gaussian distributions according to section 5.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions. Hyperparameters: $u_0 = 0.9$, $q = 3$, $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$. Figures $1, 2, 3$ correspond resp. to the models MMW, Pol, RTB. Light blue curves are the $B(= 50)$ ROC curves that are averaged in green (solid line) with $+/-$ its standard deviation (dashed green lines). The dark blue and purple curves correspond to the best and worst scoring functions in the sense of minimization and maximization of the generalization error among the B curves. The red curve corresponds to ROC*.

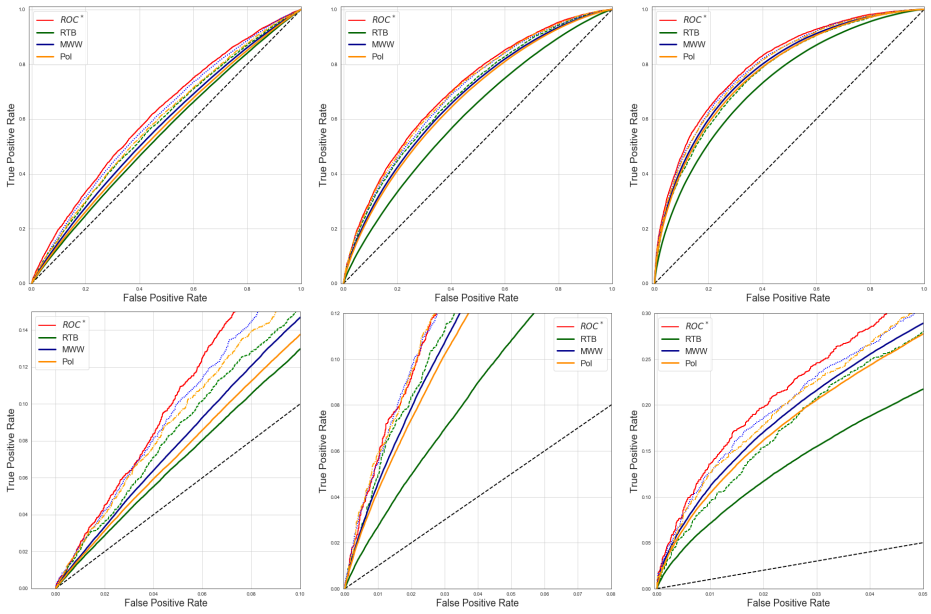
A.2 U-statistics and U-processes

As mentioned in Section 3, (degenerate) one/two-sample U-statistics are involved in the definition of the residual term introduced in Proposition 4. We recall the definition of such statistics generalizing basic *i.i.d.* sample averages, as well as some of their properties. See *e.g.* [23] for an account of the theory of U-statistics.

Definition. 11. (One-sample U-Statistic of degree two) *Let $n \geq 2$. Consider a i.i.d. sequence $X_1, \ldots, X_n$ drawn from a probability distribution μ on a measurable space $\mathcal{X}$ and $k : \mathcal{X}^2 \to \mathbb{R}$ a square integrable function w.r.t. $\mu \otimes \mu$. The one-sample U-statistic of degree 2 and kernel function k based on the X_i's is defined as:*

$$U_n(k) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} k(X_i, X_j) \ . \tag{A.1}$$

As can be shown by a basic Lehmann-Scheffé argument, the statistic $U_n(h)$ is the unbiased estimator of the parameter $\theta(k) = \int k(x_1,\ x_2)\mu(dx_1)\mu(dx_2)$ with minimum variance. Its Hájek projection can be expressed as follows: the projection of $U_n(k) - \theta(k)$ onto the space of all random variables $\sum_{i=1}^n g_i(X_i)$ with $\int g_i^2(x)\mu(dx) < +\infty$ is $\widehat{U}_n(k) = (1/n)\sum_{i=1}^n k_1(X_i)$, with $k_1 = k_{1,1} + k_{1,2}$, $k_{1,1}(x) = \mathbb{E}[k(X_1,\ x)] - \theta$ and $k_{1,2}(x) = \mathbb{E}[k(x,\ X_2)] - \theta$ for all $x \in \mathcal{X}$. The



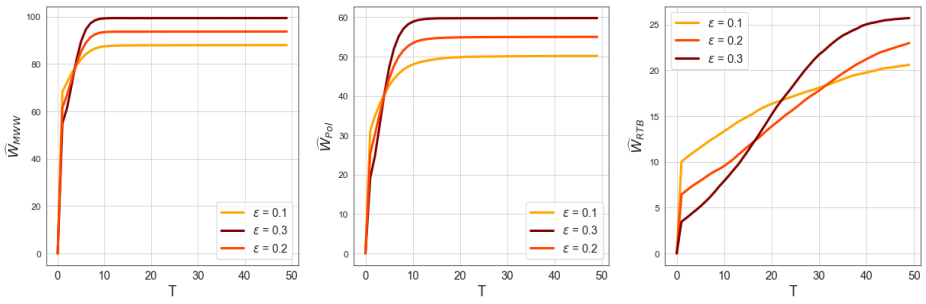
1. Loc1, $\varepsilon = 0.10$ 2. Loc2, $\varepsilon = 0.20$ 3. Loc3, $\varepsilon = 0.30$

Figure 5: Average of the ROC curves (solid line), *best* ROC curves (dashed line) for the three location models Loc1, Loc2 and Loc3. In blue for MWW, orange for Pol, green for RTB, red for ROC*. Samples are drawn from multivariate Gaussian distributions according to section 5.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions and averaged after $B = 50$ loops. Hyperparameters: $u_0 = 0.9$; $q = 3$, $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$.

U-statistic (A.1) is said to be *degenerate* when the $k_{1,l}(X_1)$'s are equal to zero with probability one, it is then of order $O_{\mathbb{P}}(1/n)$. Hence, once recentered, the U-statistic (A.1) can be written as the *i.i.d.* average $\widehat{U}_n(h)$ plus a degenerate U-statistic. This decomposition is known as the (second) Hoeffding representation of U-statistics and provides the key argument to establish limit results for such functionals, see *e.g.* [31].

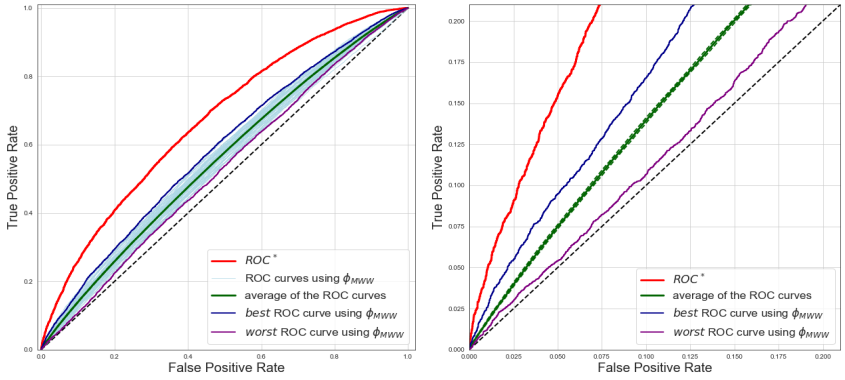
The notion of U-statistic can be generalized in several ways, by considering kernels with a number of arguments (*i.e.* degree) higher than 2 or by extending it to the multi-sample framework.

Definition. 12. (Two-sample U-Statistic of degree $(1,1)$) *Let n, m in $\mathbb{N}^*$. Consider two independent i.i.d. sequences $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ respectively drawn from probability distributions μ and ν on the measurable spaces $\mathcal{X}$ and $\mathcal{Y}$. Let $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a square integrable function w.r.t. $\mu \otimes \nu$. The two-sample U-statistic of degree $(1,1)$, with kernel function $\ell(x,y)$ and based on*

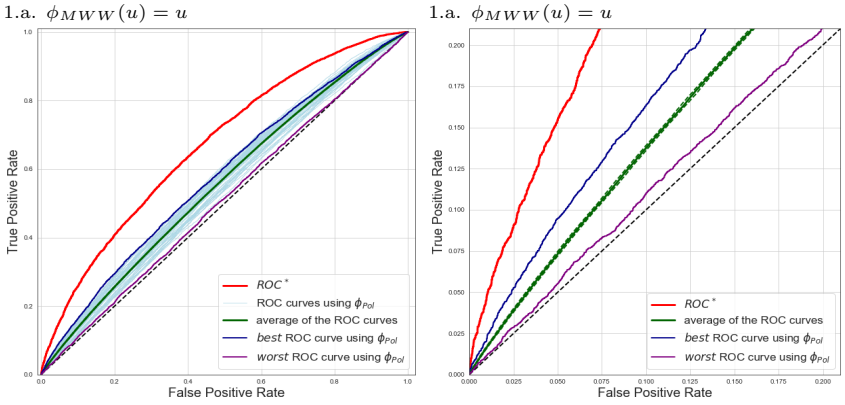


1. $\phi_{MWW}(u) = u$ 2. $\phi_{Pol}(u) = u^3$ 3. $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$

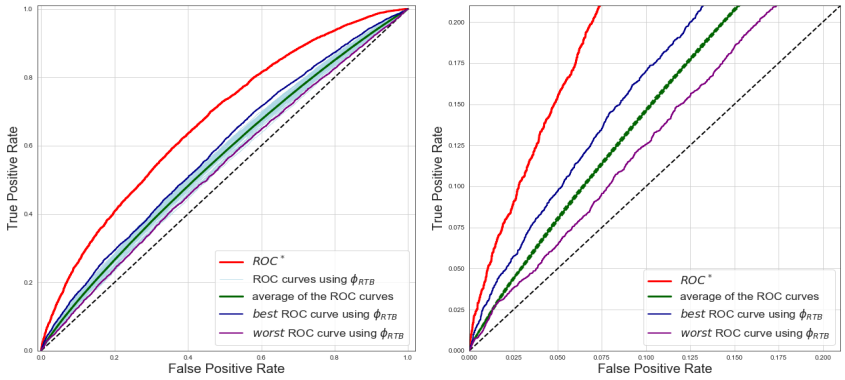
Figure 6: Average of the empirical W_ϕ-ranking performance measure over the $B = 50$ loops for the three location models Loc1, Loc2 and Loc3. Samples are drawn from multivariate Gaussian distributions according to section 5.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions and averaged after $B = 50$ loops. Hyperparameters: $u_0 = 0.9$; $q = 3$, $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$.



1.a. $\phi_{MWW}(u) = u$ 1.a. $\phi_{MWW}(u) = u$



2.b. $\phi_{Pol}(u) = u^3$ 2.b. $\phi_{Pol}(u) = u^3$



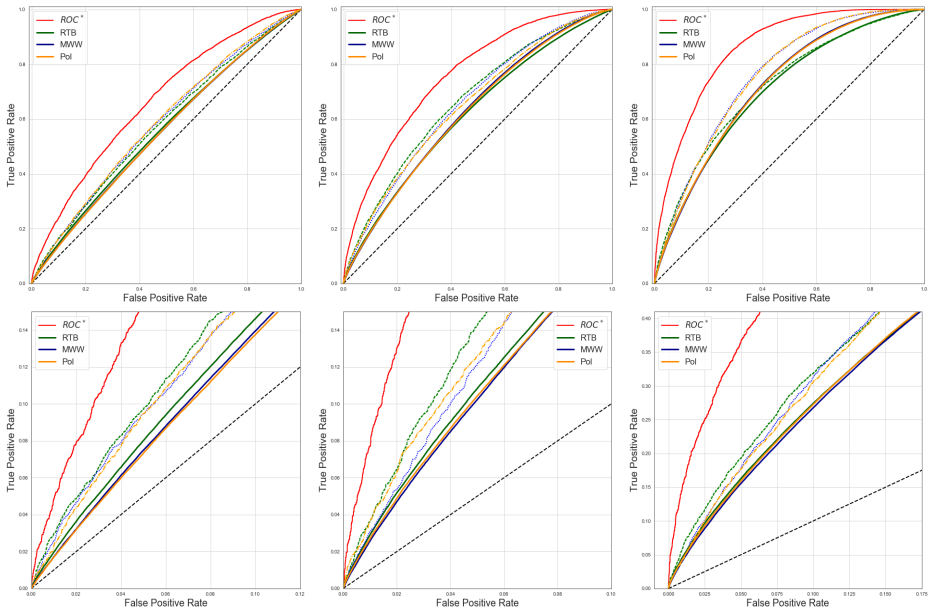
3.c. $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$ 3.c. $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$

Figure 7: Empirical ROC curves and average ROC curve for Scale1 ($\varepsilon = 0.70$). Samples are drawn from multivariate Gaussian distributions according to section 5.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions. Hyperparameters: $u_0 = 0.9$, $q = 3$, $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$. Figures $1, 2, 3$ correspond *resp.* to the models MMW, Pol, RTB. Light blue curves are the $B(= 50)$ ROC curves that are averaged in green (solid line) with $+/-$ its standard deviation (dashed green lines). The dark blue and purple curves correspond to the best and worst scoring functions in the sense of minimization and maximization of the generalization error among the B curves. The red curve corresponds to ROC*.

the X_i's and the Y_j's is defined as:

$$U_{n,m}(\ell) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \ell(X_i, Y_j) \ . \tag{A.2}$$

A classic example of two-sample U-statistic of degree $(1, 1)$ is the Mann-Whitney statistic, with symmetric kernel $\ell(x, y) = \mathbb{I}\{y < x\} + (1/2)\mathbb{I}\{y = x\}$ on $\mathbb{R}^2$ and degree $(1, 1)$. It is a natural (unbiased) estimator of the AUC: when computed from univariate samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ with distributions H and G on $\mathbb{R}$, it is equal to $\mathrm{AUC}_{\widehat{\mathrm{H}}_\mathrm{m}, \widehat{\mathrm{G}}_\mathrm{n}}$ with the notations of Subsection 2.2 and can be thus viewed as an affine transform of the rank-sum Wilcoxon statistic (2.9). The Hájek projection of (A.2) is obtained by computing the orthogonal projection of the recentered *r.v.* $U_{n,m}(\ell) - \mathbb{E}[U_{n,m}(\ell)]$ onto the subpace of L_2 composed of all random variables $\sum_{i=1}^n g_i(X_i) + \sum_{j=1}^m f_j(Y_j)$ with $\int g_i^2(x)\mu(dx) < +\infty$ and $\int f_j^2(y)\nu(dy) < +\infty$, namely $\widehat{U}_{n,m}(\ell) = (1/n)\sum_{i=1}^n \ell_{1,1}(X_i) + (1/m)\sum_{j=1}^m \ell_{1,2}(Y_j)$, with $\ell_{1,1}(x) = \mathbb{E}[\ell(x,\ Y_1)] - \mathbb{E}[U_{n,m}(\ell)]$ and $\ell_{1,2}(y) = \mathbb{E}[\ell(X_1,\ y)] - \mathbb{E}[U_{n,m}(\ell)]$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The U-statistic $U_{n,m}(\ell)$ is said to be *degenerate* when the random variables $\ell_{1,1}(X_1)$ and $\ell_{1,2}(Y_1)$ are equal to zero with probability one. Similar to (A.1), the recentered version of the two-sample U-statistic of degree $(1, 1)$ (A.2) can be written as a sum of two *i.i.d.* averages $\widehat{U}_{n,m}(\ell)$ plus a degenerate U-statistic of order $O_{\mathbb{P}}(1/n) + O_{\mathbb{P}}(1/m)$. Again, the Hoeffding decomposition is the key to directly extend limit results known for *i.i.d.* aver-



1. Scale1, $\varepsilon = 0.70$ 2. Scale2, $\varepsilon = 0.90$ 3. Scale3, $\varepsilon = 1.10$

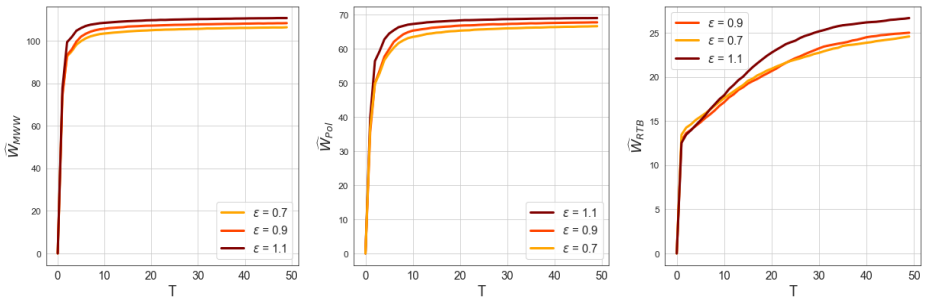
Figure 8: Average of the ROC curves (solid line), *best* ROC curves (dashed line) for the three scale models Scale1, Scale2 and Scale3. In blue for MWW, orange for Pol, green for RTB, red for ROC*. Samples are drawn from multivariate Gaussian distributions according to section 5.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions and averaged after $B = 50$ loops. Hyperparameters: $u_0 = 0.9$; $q = 3$, $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$.

ages (*e.g.* SLLN, CLT, LIL) to statistics of the type (A.2). In the subsequent technical analysis, nonasymptotic uniform results are required for U-processes, namely collections of U-statistics indexed by classes of kernels. By means of the Hoeffding decomposition, concentration bounds for U-processes can be obtained by combining classic concentration bounds for empirical processes and concentration bounds for degenerate U-processes, such as those recalled in A.4.

A.3 VC-type Classes of Functions - Permanence Properties

The concentration inequalities for U-processes recalled in Appendix A.4 and involved in the proof of the main results stated in this article apply to collections of kernels that are of VC-type, a classic concept used to quantify the complexity of classes of functions. It is recalled below, see *e.g.* [36] for generalizations and further details.

Definition. 13. *A class $\mathcal{F}$ of real-valued functions defined on a measurable space $\mathcal{Z}$ is a bounded VC-type class with parameter $(A, \mathcal{V}) \in (0, +\infty)^2$ and*



1. $\phi_{MWW}(u) = u$ 2. $\phi_{Pol}(u) = u^3$ 3. $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$

Figure 9: Average of the empirical W_ϕ- ranking performance measure over the $B = 50$ loops for the three location models Loc1, Loc2 and Loc3. Samples are drawn from multivariate Gaussian distributions according to section 5.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions and averaged after $B = 50$ loops. Hyperparameters: $u_0 = 0.9$; $q = 3$, $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$.

constant envelope $L_{\mathcal{F}} > 0$ *if for all* $\varepsilon \in (0, 1)$*:*

$$\sup_Q N(\mathcal{F}, L_2(Q), \varepsilon L_{\mathcal{F}}) \leq \left(\frac{A}{\varepsilon}\right)^{\mathcal{V}}, \tag{A.3}$$

where the supremum is taken over all probability measures Q *on* $\mathcal{Z}$ *and the smallest number of* $L_2(Q)$*-balls of radius less than* ε *required to cover class* $\mathcal{F}$ *(i.e. covering number) is meant by* $N(\mathcal{F}, L_2(Q), \varepsilon)$*.*

Recall that a bounded VC class of functions with VC dimension $V < +\infty$ is of VC-type and fulfills the condition above with $\mathcal{V} = 2(V - 1)$ and $A = (cV(16e)^V)^{1/(2(V-1))}$, where c is a universal constant, see *e.g.* Theorem 2.6.7 in [36]. The lemma stated below permits to control the complexity of the classes of kernels/functions involved in the Hoeffding decompositions of a two-sample U-process of degree $(1, 1)$ or of a one-sample U-process of degree 2, *cf* subsection A.2.

Lemma. 14. *Let* X *and* Y *be two independent random variables, valued in* $\mathcal{X}$ *and* $\mathcal{Y}$ *respectively, with probability distributions* μ *and* ν*. Consider* $\mathcal{L}$ *a* VC*-type bounded class of kernels* $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ *with parameters* $(A, \mathcal{V})$ *and constant envelope* $L_{\mathcal{L}} > 0$*. Then, the sets of functions* $\{x \in \mathcal{X} \mapsto \mathbb{E}[\ell(x,\ Y)] :\ \ell \in \mathcal{L}\}$*,* $\{y \in \mathcal{Y} \mapsto \mathbb{E}[\ell(X,\ y)] :\ \ell \in \mathcal{L}\}$*,* $\{\ell(x, y) - \mathbb{E}[\ell(X,\ y)] - \mathbb{E}[\ell(x,\ Y)] :\ \ell \in \mathcal{L}\}$ *are also* VC*-type bounded classes.*

PROOF. Consider first the uniformly bounded class $\mathcal{L}_1$ composed of functions $x \in \mathcal{X} \mapsto \mathbb{E}[\ell(x,\ Y)]$ with $\ell \in \mathcal{L}$. Let $\varepsilon > 0$ and P be any probability measure on $\mathcal{X}$. Define the probability measure $P_\nu(dx, dy) = P(dx)\nu(dy)$ on $\mathcal{X} \times \mathcal{Y}$ and consider a ε-covering of the class $\mathcal{L}$ with centers $\ell_1, \ldots,\ \ell_K$ *w.r.t.* the metric

$L_2(P_\nu)$, $K \geq 1$. For all $\ell \in \mathcal{L}$, there exists $k \leq K$ such that:

$$\int_{x\in\mathcal{X}} \int_{y\in\mathcal{Y}} (\ell(x,\ y) - \ell_k(x,\ y))^2 P_\nu(dx, dy) \leq \varepsilon^2 \ .$$

By virtue of Jensen's inequality, we have

$$\begin{aligned}\int_{\mathcal{X}} (\mathbb{E}[\ell(x,\ Y)] - \mathbb{E}[\ell_k(x,\ Y)])^2 P(dx) \\ \leq \int_{\mathcal{X}} \mathbb{E}[(\ell(x,\ Y) - \ell_k(x,\ Y))^2] P(dx) \\ = \int_{\mathcal{X}} \int_{\mathcal{Y}} (\ell(x,\ y) - \ell_k(x,\ y))^2 \nu(dy) P(dx) \leq \varepsilon^2 \ .\end{aligned}$$

Hence, one gets a ε-covering of the class $\mathcal{L}_1$ with balls of centers $\{\mathbb{E}[\ell_k(\cdot,\ Y)] : k = 1,\ \ldots,\ K\}$ in $L_2(P)$. This proves that

$$N(\mathcal{L}_1, L_2(P), \varepsilon L_{\mathcal{L}}) \leq N(\mathcal{L}, L_2(P_\nu), \varepsilon L_{\mathcal{L}}).$$

As a similar reasoning can be applied to the two other classes of functions, one then gets the desired result.

A.4 Concentration Inequalities for Degenerate U-processes.

In [24] (see Theorem 2 therein), a concentration bound for one-sample degenerate U-processes of arbitrary degree indexed by L_2-dense classes of non-symmetric kernels is established. The lemma below is a formulation of the latter in the specific case of degenerate U-processes of degree 2 indexed by VC-type bounded classes of non-symmetric kernels.

Lemma. 15. *Let $n \geq 2$ and $X_1,\ \ldots,\ X_n$ be i.i.d. random variables drawn from a probability distribution μ on a measurable space $\mathcal{X}$. Let $\mathcal{K}$ be a class of measurable kernels $k : \mathcal{X}^2 \to \mathbb{R}$ such that $\sup_{x,x'\in\mathcal{X}^2} |k(x,x')| \leq D < +\infty$ and $\int_{\mathcal{X}^2} k^2(x,x')\mu(dx)\mu(dx') \leq \sigma^2 \leq D^2$, that defines a degenerate one-sample U-process of degree 2, based on the X_i's: $\{U_n(k)\ k \in \mathcal{K}\}$. Suppose in addition that the class $\mathcal{K}$ is of* VC*-type with parameters $(A, \mathcal{V})$. Then, there exist constants $C_1 > 0$, $C_2 \geq 1$ and $C_3 \geq 0$ depending on $(A, \mathcal{V})$ such that:*

$$\mathbb{P}\left\{\sup_{k\in\mathcal{K}} |U_n(k)| \geq t\right\} \leq C_2 \exp\left\{-\frac{C_3(n-1)t}{\sigma}\right\} \ , \tag{A.4}$$

as soon as $C_1 \log(2D/\sigma) \leq (n-1)t/\sigma \leq n\sigma^2/D^2$.

The next lemma provides a similar nonasymptotic result for degenerate two-sample U-processes of degree $(1,1)$.

Lemma. 16. *Let $(n,\ m) \in \mathbb{N}^*$. Consider two independent i.i.d. random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ respectively drawn from the probability distributions μ and ν on the measurable spaces $\mathcal{X}$ and $\mathcal{Y}$. Let $\mathcal{L}$ be a class of degenerate non-symmetrical kernels $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that $\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |\ell(x,y)| \leq L < +\infty$ and $\int_{\mathcal{X}\times\mathcal{Y}} \ell^2(x,y)\mu(dx)\nu(dy) \leq \sigma^2 \leq L^2$, that defines a degenerate two-sample U-process of degree $(1,1)$, based on the X_i, Y_j's: $\{U_{n,m}(\ell)\ \ell \in \mathcal{L}\}$. Suppose in addition that the class $\mathcal{L}$ is of* VC*-type with parameters $(A, \mathcal{V})$. Then, for all $t > 0$, there exists a universal constant $K > 2$ such that:*

$$\mathbb{P}\left\{\sup_{\ell\in\mathcal{L}}|U_{n,m}(\ell)|\geq t\right\}\leq K2^{\mathcal{V}}(A/L)^{2\mathcal{V}}e^{4/L^2}\exp\left\{-\frac{nmt^2}{ML^2}\right\}\ , \tag{A.5}$$

for all $nmt^2 > \max(8^4\log(2)L^2\mathcal{V}, (\log(2)L^2\mathcal{V}/2)^{1+\delta})$, $\delta\in(1,2)$ *constant and* $M = 16^3/2$.

Its proof is given in B.7 and is inspired from that of Lemma 2.14.9 in [36] and of Lemma 3.2 in [34] for empirical processes, and from Lemma 2.4 in [28] which gives a version in expectation applicable to degenerate two-sample U-processes of arbitrary degree indexed by L_p-dense classes of kernels.

B Technical Proofs

The proofs of the results stated in the paper are detailed below.

B.1 Proof of Proposition 4

Let $\theta_0\in(0,1)$. Since $\phi(u)\in\mathcal{C}^2([0,1],\mathbb{R})$ by virtue of Assumption 2, a Taylor expansion of order two yields: for all $\theta\in(0,1)$

$$\phi(\theta)=\phi(\theta_0)+(\theta-\theta_0)\phi'(\theta_0)+\int_{\theta_0}^{\theta}(\theta-u)\phi''(u)du\ . \tag{B.1}$$

Let $s\in\mathcal{S}_0$. For all $t\in\mathbb{R}$, we have

$$\begin{aligned}\phi\left(\frac{N\widehat{F}_{s,N}(t)}{N+1}\right)&=\phi\circ F_s(t)+\left(\frac{N\widehat{F}_{s,N}(t)}{N+1}-F_s(t)\right)\phi'\circ F_s(t)\\&\quad+\int_{F_s(t)}^{N\widehat{F}_{s,N}(t)/(N+1)}\left(\frac{N\widehat{F}_{s,N}(t)}{N+1}-u\right)\phi''(u)du\ ,\end{aligned} \tag{B.2}$$

with probability one. Let $i\leq n$, for $t=s(\mathbf{X}_i)$, (B.2) writes:

$$\begin{aligned}\phi\left(\frac{N\widehat{F}_{s,N}(s(\mathbf{X}_i))}{N+1}\right)&=\phi\circ F_s(s(\mathbf{X}_i))\\&\quad+\left(\frac{N\widehat{F}_{s,N}(s(\mathbf{X}_i))}{N+1}-F_s(s(\mathbf{X}_i))\right)\phi'\circ F_s(s(\mathbf{X}_i))+t_i(s)\quad a.s.\ ,\end{aligned} \tag{B.3}$$

where

$$|t_i(s)|\leq(\|\phi''\|_\infty/2)\left(N/(N+1)\widehat{F}_{s,N}(s(\mathbf{X}_i))-F_s(s(\mathbf{X}_i))\right)^2.$$

Hence, by summing over $i\in\{1,\ \ldots,\ n\}$, one gets that the approximation of $\widehat{W}_{n,m}(s)$ stated below holds true almost-surely:

$$\widehat{W}_{n,m}(s)=n\widehat{W}_\phi(s)+B_{n,m}(s)+\widehat{T}_{n,m}(s)\ , \tag{B.4}$$

where

$$B_{n,m}(s) = \sum_{i=1}^{n} \left(\frac{N\widehat{F}_{s,N}(s(\mathbf{X}_i))}{N+1} - F_s(s(\mathbf{X}_i)) \right) \phi' \circ F_s(s(\mathbf{X}_i)), \tag{B.5}$$

$$|\widehat{T}_{n,m}(s)| = \sum_{i=1}^{n} |t_i(s)| \leq \frac{\|\phi''\|_\infty}{2} \sum_{i=1}^{n} \left(\frac{N\widehat{F}_{s,N}(s(\mathbf{X}_i))}{N+1} - F_s(s(\mathbf{X}_i)) \right)^2 \tag{B.6}$$

Linearization of $B_{n,m}(\cdot)$. First, observe that

$$\begin{aligned} B_{n,m}(s) = & \frac{1}{N+1} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathbb{I}\{s(\mathbf{X}_j) \leq s(\mathbf{X}_i)\} \phi' \circ F_s(s(\mathbf{X}_i)) \\ & + \frac{1}{N+1} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}\{s(\mathbf{Y}_j) \leq s(\mathbf{X}_i)\} \phi' \circ F_s(s(\mathbf{X}_i)) \\ & + \sum_{i=1}^{n} \left(\frac{1}{N+1} - F_s(s(\mathbf{X}_i)) \right) \phi' \circ F_s(s(\mathbf{X}_i)) \ . \end{aligned} \tag{B.7}$$

Notice that the first two terms are U-processes indexed by $\mathcal{S}_0$, *cf* Section A.2, while the last term is an empirical process. Indeed, one may write

$$B_{n,m}(s) = \frac{n(n-1)}{N+1} U_n(k_s) + \frac{nm}{N+1} U_{n,m}(\ell_s) + \widehat{K}_{n,m}(s) \ , \tag{B.8}$$

where

$$U_n(k_s) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathbb{I}\{s(\mathbf{X}_j) \leq s(\mathbf{X}_i)\} \phi' \circ F_s(s(\mathbf{X}_i)) \tag{B.9}$$

is a (nondegenerate) 1-sample U-process of degree 2 based on the random sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ with nonsymmetric kernel $k_s(x, x') = \mathbb{I}\{s(x') \leq s(x)\}\phi' \circ F_s(s(x))$ on $\mathcal{X} \times \mathcal{X}$,

$$U_{n,m}(\ell_s) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}\{s(\mathbf{Y}_j) \leq s(\mathbf{X}_i)\} \phi' \circ F_s(s(\mathbf{X}_i)) \tag{B.10}$$

is a (nondegenerate) two-sample U-process of degree $(1,1)$ based on the samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ with kernel $\ell_s(x, y) = \mathbb{I}\{s(y) \leq s(x)\}\phi' \circ F_s(s(x))$ on $\mathcal{X} \times \mathcal{Y}$, and

$$\widehat{K}_{n,m}(s) = \sum_{i=1}^{n} \left(\frac{1}{N+1} - F_s(s(\mathbf{X}_i)) \right) \phi' \circ F_s(s(\mathbf{X}_i))$$

is an empirical process based on the $\mathbf{X}_i$'s. In order to write $B_{n,m}$ as an empirical process plus a (negligible) remainder term, the Hoeffding decomposition is applied to the U-processes above, *cf* Appendix A.2:

$$U_n(k_s) = \mathbb{E}[U_n(k_s)] + \widehat{U}_n(k_s) + \mathcal{R}_n(k_s) \ , \tag{B.11}$$

$$U_{n,m}(\ell_s) = \mathbb{E}[U_{n,m}(\ell_s)] + \widehat{U}_{n,m}(\ell_s) + \mathcal{R}_{n,m}(\ell_s) \ , \tag{B.12}$$

where

$$\widehat{U}_n(k_s) = \frac{1}{n}\sum_{i=1}^{n} k_{s,1,1}(\mathbf{X}_i) + \frac{1}{n}\sum_{i=1}^{n} k_{s,1,2}(\mathbf{X}_i) \ , \tag{B.13}$$

with $k_{s,1,1}(x) = \mathbb{E}[k_s(x,\mathbf{X})] - \mathbb{E}[U_n(k_s)]$ and $k_{s,1,2}(x) = \mathbb{E}[k_s(\mathbf{X},x)] - \mathbb{E}[U_n(k_s)]$, and

$$\widehat{U}_{n,m}(\ell_s) = \frac{1}{m}\sum_{j=1}^{m} \ell_{s,1,1}(\mathbf{Y}_j) + \frac{1}{n}\sum_{i=1}^{n} \ell_{s,1,2}(\mathbf{X}_i) \ , \tag{B.14}$$

with $\ell_{s,1,1}(y) = \mathbb{E}[\ell_s(\mathbf{X},y)] - \mathbb{E}[U_{n,m}(\ell_s)]$ and $\ell_{s,1,2}(x) = \mathbb{E}[\ell_s(x,\mathbf{Y})] - \mathbb{E}[U_{n,m}(\ell_s)]$. Consequently, the Hájek projection of the process $B_{n,m}(s)$ is given by

$$\widehat{B}_{n,m}(s) - \mathbb{E}[\widehat{B}_{n,m}(s)] = \frac{n(n-1)}{N+1}\widehat{U}_n(k_s) + \frac{nm}{N+1}\widehat{U}_{n,m}(\ell_s) + \widehat{K}_{n,m}(s) - \mathbb{E}[\widehat{K}_{n,m}(s)] \ . \tag{B.15}$$

The following result provides an approximation of (B.15) and is proved in Appendix B.2.2.

Lemma. 17. *Under Assumptions 1-3, the Hájek projection of the stochastic process $B_{n,m}(\cdot)$, denoted by $\widehat{B}_{n,m}(\cdot)$ and indexed by $\mathcal{S}_0$, onto the subspace generated by the random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ can be approximated as follows: for all $s \in \mathcal{S}_0$,*

$$\widehat{B}_{n,m}(s) - \mathbb{E}\left[\widehat{B}_{n,m}(s)\right] = \widehat{V}_n^X(s) + \widehat{V}_m^Y(s) + \widehat{R}_{n,m}(s) \ , \tag{B.16}$$

where

$$\widehat{V}_n^X(s) = \frac{n-1}{N+1}\sum_{i=1}^{n} k_{s,1,1}(\mathbf{X}_i), \ \widehat{V}_m^Y(s) = \frac{n}{N+1}\sum_{j=1}^{m} \ell_{s,1,1}(\mathbf{Y}_j) \ .$$

Let $\delta > 0$, there exist constants $A_1, A_3 > 0$, $A_2 \geq 1$ depending on ϕ and $\mathcal{V}$ such that for all $A_4 \geq A_1$

$$\mathbb{P}\left\{\sup_{s\in\mathcal{S}_0}\left|\widehat{R}_{n,m}(s)\right| > t\right\} \leq A_2 \exp\left\{-\frac{A_3 N t^2}{p\sigma^2}\right\} \ , \tag{B.17}$$

as soon as $A_1\sigma\sqrt{p\log(2\|\phi'\|_\infty/\sigma)/N} \leq t \leq pA_4\|\phi'\|_\infty$, with $\sigma^2 = \int_{[0,1]}\phi'^2$.

The last step relies on all previous decompositions, so as to approximate $B_{n,m}(\cdot)$ by the sum of two empirical processes $\widehat{V}_n^X(\cdot)$ and $\widehat{V}_n^Y(\cdot)$, with a uniform control of the error. All residual terms, $\widehat{R}_{n,m}(s)$ (Lemma 17) plus the remainders of the U-processes, are the components of the process $\mathcal{R}_{n,m}^B(s)$, see the following Lemma 18.

Lemma. 18. *Suppose that Assumptions 1-3 are fulfilled. The stochastic process $B_{n,m}(.)$ can be approximated as follows: for all $s \in \mathcal{S}_0$,*

$$B_{n,m}(s) - \mathbb{E}\left[B_{n,m}(s)\right] = \widehat{V}_n^X(s) + \widehat{V}_m^Y(s) + \mathcal{R}_{n,m}^B(s) \ . \tag{B.18}$$

Let $\delta > 0$. There exist $D_1 > 0$ universal constant, and constants D_3, $D_4 > 0$, $D_2 \geq 1$, d_1, $d_2 > 3$ depending on ϕ and $\mathcal{V}$, such that with probability at least $1-\delta$:

$$\sup_{s\in\mathcal{S}_0}|\mathcal{R}^B_{n,m}(s)| \leq \|\phi'\|_\infty\sqrt{p(1-p)D_1\log(d_1/\delta)} + (p\|\phi'\|_\infty D_4)\log(d_2/\delta)\ , \tag{B.19}$$

as soon as $N \geq (pD_3)^{-1}\log(D_2/\delta)$.

Refer to Appendix B.2.3 for the detailed proof.

A uniform bound for $\widehat{T}_{n,m}(\cdot)$. By virtue of (B.5), we have:

$$\sup_{s\in\mathcal{S}_0}|\widehat{T}_{n,m}(s)| \leq n\|\phi''\|_\infty\left(\sup_{(s,t)\in\mathcal{S}_0\times\mathbb{R}}\left(\widehat{F}_{s,N}(t)-F_s(t)\right)^2 + \frac{1}{(N+1)^2}\right). \tag{B.20}$$

Observe also that

$$\begin{aligned}\sup_{(s,t)\in\mathcal{S}_0\times\mathbb{R}}|\widehat{F}_{s,N}(t)-F_s(t)| \leq\ & p\sup_{(s,t)\in\mathcal{S}_0\times\mathbb{R}}|\widehat{G}_{s,n}(t)-G_s(t)| \\ & + (1-p)\sup_{(s,t)\in\mathcal{S}_0\times\mathbb{R}}|\widehat{H}_{s,m}(t)-H_s(t)| + \frac{2}{N}\ .\end{aligned} \tag{B.21}$$

A classic concentration bound for empirical processes based on the VC inequality (see *e.g.* Theorems 3.2 and 3.4 in [2]) shows that, for any $\delta\in(0,1)$, we have with probability at least $1-\delta$:

$$\sup_{(s,t)\in\mathcal{S}_0\times\mathbb{R}}|\widehat{G}_{s,n}(t)-G_s(t)| \leq c\sqrt{\frac{\mathcal{V}}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}}\ ,$$

where $c>0$ is a universal constant. In a similar fashion, we have, with probability larger than $1-\delta$,

$$\sup_{(s,t)\in\mathcal{S}_0\times\mathbb{R}}|\widehat{H}_{s,m}(t)-H_s(t)| \leq c\sqrt{\frac{\mathcal{V}}{m}} + \sqrt{\frac{2\log(1/\delta)}{m}}\ .$$

Combining the bounds above with the union bound, (B.21) and (B.20) we obtain that, for any $\delta\in(0,1)$, we have with probability larger than $1-\delta$:

$$\begin{aligned}\sup_{s\in\mathcal{S}_0}|\widehat{T}_{n,m}(s)| &\leq n||\phi''||_\infty\left(12\left(\frac{c^2\mathcal{V}+\log(2/\delta)}{N}+\frac{1}{N^2}\right)+\frac{1}{(N+1)^2}\right) \\ &\leq B_1 + B_2\log(2/\delta)\ ,\end{aligned} \tag{B.22}$$

where B_1 (*resp.* B_2) is a constant that only depends on ϕ and $\mathcal{V}$ (*resp.* on ϕ). To end the proof, it suffices to observe that the remainder process is the sum of $\mathcal{R}^B_{n,m}(s)$ and $\widehat{T}_{n,m}(s)$. Combining bounds (B.19) and (B.22), we get that, with probability at least $1-\delta$,

$$\sup_{s\in\mathcal{S}_0}|\mathcal{R}_{n,m}(s)| = \sup_{s\in\mathcal{S}_0}|\mathcal{R}^B_{n,m}(s)+\widehat{T}_{n,m}(s)| \leq B_1+\|\phi'\|_\infty\kappa_p D\log(2d/\delta)+B_2\log(4/\delta) \tag{B.23}$$

as soon as $N\geq(pD_3)^{-1}\log(D_2/\delta)$, with $D=\max(\sqrt{D_1},D_4)$, $d=\max(d_1,d_2)$, $\kappa_p=\max(\sqrt{p(1-p)},p)$. As $B_2>1$, $d\geq 3$, and for small δ, we obtain the upperbound $B_1+(\|\phi'\|_\infty\kappa_p D+B_2)\log(2d/\delta)$.

B.2 Intermediary Results

The intermediary results involved in Section B.1 are now established.

B.2.1 Permanence Properties

The lemmas below claim that the collections of kernels/functions involved in the decomposition obtained in Appendix B.1 are of VC-type and uniformly bounded.

Lemma. 19. *Suppose that Assumptions 2 and 3 are fulfilled. Then, the collections of kernels* $\{k_s(x,x') : s \in \mathcal{S}_0\}$ *and* $\{\ell_s(x,y) : s \in \mathcal{S}_0\}$ *and are bounded* VC*-type classes of functions with parameters fully determined by* $\mathcal{V}$ *and* ϕ*.*

PROOF. Recall that: $\forall (x,x') \in \mathcal{X}^2$,

$$k_s(x,x') = \mathbb{I}\{s(x') \leq s(x)\}\,(\phi' \circ F_s)\,(s(x)).$$

Hence, we have $\sup_{(x,x')\in\mathcal{X}^2} |k_s(x,x')| \leq ||\phi'||_\infty$ for all $s \in \mathcal{S}_0$. In additions, since the collections $\{(x,x') \in \mathcal{X}^2 \mapsto s(x) : s \in \mathcal{S}_0\}$ and $\{(x,x') \in \mathcal{X}^2 \mapsto s(x') : s \in \mathcal{S}_0\}$ are VC classes of functions, classic permanence properties of VC classes of functions (see *e.g.* Lemma 2.6.18) shows that $\{(x,x') \in \mathcal{X}^2 \mapsto s(x) - s(x') : s \in \mathcal{S}_0\}$ is also a VC class, as well as the class of indicator functions $\{(x,x') \in \mathcal{X}^2 \mapsto \mathbb{I}\{s(x') \leq s(x)\} : s \in \mathcal{S}_0\}$. Consequently, the argument of Lemma 14's proof permits to see easily that $\{(x,x') \in \mathcal{X}^2 \mapsto F_s(s(x)) = \mathbb{E}[\mathbb{I}\{s(X) \leq s(x)\}] : s \in \mathcal{S}_0\}$ is of VC type, just like $\{(x,x') \in \mathcal{X}^2 \mapsto (\phi' \circ F_s)(s(x)) : s \in \mathcal{S}_0\}$ using the Lipschitz property of ϕ', *cf* Assumption 2. Finally, being composed of products of a function in the bounded VC-type class $\{(x,x') \in \mathcal{X}^2 \mapsto \mathbb{I}\{s(x') \leq s(x)\} : s \in \mathcal{S}_0\}$ by a function in the bounded VC-type class $\{(x,x') \in \mathcal{X}^2 \mapsto (\phi' \circ F_s)(s(x)) : s \in \mathcal{S}_0\}$, the collection $\{k_s : s \in \mathcal{S}_0\}$ is still a bounded VC-type class of functions. A similar reasoning can be applied to show that $\{\ell_s : s \in \mathcal{S}_0\}$ is a bounded VC-type class of kernels on $\mathcal{X} \times \mathcal{Y}$. The following result is straightforwardly deduced from the lemma above combined with Lemma 14.

Lemma. 20. *Suppose that Assumptions 2 and 3 are fulfilled. Then, the collections of functions/kernels* $\{k_{s,1,1}(x) : s \in \mathcal{S}_0\}$, $\{k_{s,1,2}(x) : s \in \mathcal{S}_0\}$, $\{k_s(x,x') - k_{s,1,1}(x) - k_{s,1,2}(x') : s \in \mathcal{S}_0\}$, $\{\ell_{s,1,1}(y) : s \in \mathcal{S}_0\}$, $\{\ell_{s,1,2}(x) : s \in \mathcal{S}_0\}$ *and* $\{\ell_s(x,y) - \ell_{s,1,1}(y) - \ell_{s,1,2}(x) : s \in \mathcal{S}_0\}$ *are bounded* VC*-type classes with parameters fully determined by* $\mathcal{V}$ *and* ϕ*.*

B.2.2 Proof of Lemma 17

For $s \in \mathcal{S}_0$, by adding the diagonal term, the empirical process can be written

$$\widehat{R}_{n,m}(s) = \left(\frac{n}{N+1} - p\right) \sum_{i=1}^{n} k_{s,1,2}(\mathbf{X}_i) + \left(\frac{m}{N+1} - (1-p)\right) \sum_{i=1}^{n} \ell_{s,1,2}(\mathbf{X}_i)\ . \tag{B.24}$$

We uniformly bound all three empirical processes in probability using classic concentration bounds, see *e.g.* Theorem 2.1 in [13], as follows. Assuming Assumptions 2-3, Lemma 20 states that each class of functions $\{k_{s,1,2} : s \in \mathcal{S}_0\}$, $\{\ell_{s,1,2} : s \in \mathcal{S}_0\}$ is uniformly bounded and VC-type of parameters depending

only on ϕ and on the VC dimension $\mathcal{V}$. For the class $\{x \mapsto \phi' \circ F_s(s(x)) : s \in \mathcal{S}_0\}$, the arguments are exposed in the proof of Lemma 19. The variance of the kernels can be bounded for all $s \in \mathcal{S}_0$, by $\sigma^2 = \int_{[0,1]} \phi'^2$ and $\sigma^2 \leq ||\phi'||_\infty^2$ and notice that $|n/(N+1) - p| \leq 1/N$ and $|m/(N+1) - (1-p)| \leq 1/N$. Let $t > 0$, there exist a sequence of constants $A_{1,i} > 0, A_{2,i} \geq 1, A_{3,i} > 0$ depending on ϕ and $\mathcal{V}$, $i \in \{1, 2\}$, such that for all $A_{4,i} \geq A_{1,i}$, the following inequalities hold true.

$$\mathbb{P}\left\{\frac{1}{N}\sup_{s\in\mathcal{S}_0}\left|\sum_{i=1}^{n} k_{s,1,2}(\mathbf{X}_i)\right| > t\right\} \leq A_{2,1}\exp\left\{-\frac{A_{3,1}Nt^2}{p\sigma^2}\right\}, \tag{B.25}$$

as soon as $A_{1,1}\sigma\sqrt{p\log(2\|\phi'\|_\infty/\sigma)/N} \leq t \leq pA_{4,1}\|\phi'\|_\infty$,

$$\mathbb{P}\left\{\frac{1}{N}\sup_{s\in\mathcal{S}_0}\left|\sum_{i=1}^{n} \ell_{s,1,2}(\mathbf{X}_i)\right| > t\right\} \leq A_{2,2}\exp\left\{-\frac{A_{3,2}Nt^2}{p\sigma^2}\right\}, \tag{B.26}$$

as soon as $A_{1,2}\sigma\sqrt{p\log(2\|\phi'\|_\infty/\sigma)/N} \leq t \leq pA_{4,2}\|\phi'\|_\infty$. The union bound with threshold $t/2$ yields

$$\mathbb{P}\left\{\sup_{s\in\mathcal{S}_0}\left|\widehat{R}_{n,m}(s)\right| > t\right\} \leq A_2\exp\left\{-\frac{A_3Nt^2}{p\sigma^2}\right\}, \tag{B.27}$$

as soon as $A_1\sigma\sqrt{p\log(2\|\phi'\|_\infty/\sigma)/N} \leq t \leq 2pA_4\|\phi'\|_\infty$, with $A_1 = 2\max(A_{1,1}, A_{1,2})$, $A_2 = 2\max(A_{2,1}, A_{2,2})$, $A_3 = \min(A_{3,1}, A_{3,2})/4$, $A_4 = \min(A_{4,1}, A_{4,2})$ such that $A_4 \geq A_1$.

B.2.3 Proof of Lemma 18

The remainder of the decomposition (18) is obtained by combining Eq. (B.8), (B.15) and yields, for all $s \in \mathcal{S}_0$

$$\left|\mathcal{R}^B_{n,m}(s)\right| \leq |\widehat{R}_{n,m}(s)| + p^2N|\mathcal{R}_n(k_s)| + p(1-p)N|\mathcal{R}_{n,m}(\ell_s)| .$$

Suppose Assumptions 2-3 are fulfilled. The first process can be uniformly bounded on $\mathcal{S}_0$ as proved in Lemma 17. For the two others, we apply the results of Lemmas 15 and 16 as follows. The process $\mathcal{R}_n(k_s)$ (*resp.* $\mathcal{R}_{n,m}(\ell_s)$) is the residual term obtained by decomposing the U-process $U_n(k_s)$ (Eq. (B.11), *resp.* (B.12))), for all $s \in \mathcal{S}_0$. By Lemma 20, its class of degenerate kernels $\{(x,x') \mapsto k_s(x,x') - k_{s,1,1}(x) - k_{s,1,2}(x') : s \in \mathcal{S}_0\}$ (*resp.* $\{(x,y) \mapsto \ell_s(x,y) - \ell_{s,1,1}(y) - \ell_{s,1,2}(x) : s \in \mathcal{S}_0\}$) is uniformly bounded and VC-type of parameters depending only on ϕ and on the VC dimension $\mathcal{V}$. Notice that the three classes of functions have variances and envelopes which can be similarly bounded by $\sigma^2 = \int_{[0,1]} \phi'^2 \leq ||\phi'||_\infty^2$, up to a multiplicative constant for both residuals. Let $\delta > 0$, there exist constants $A_1, B_1 > 0, A_2, B_2 \geq 1, A_3, B_3 > 0$ depending on ϕ and $\mathcal{V}$ *s.t.* with probability at least $1 - \delta$

$$\sup_{s\in\mathcal{S}_0}\left|\widehat{R}_{n,m}(s)\right| \leq \|\phi'\|_\infty\sqrt{\frac{p\log(A_2/\delta)}{A_3N}}, \tag{B.28}$$

as soon as $N \geq (pA_3)^{-1}\log(A_2/\delta)$. Also by Lemma 15

$$p^2 N \sup_{s\in\mathcal{S}_0} |\mathcal{R}_n(k_s)| \leq (p\|\phi'\|_\infty/B_3)\log(B_2/\delta) \ , \tag{B.29}$$

when $N \geq (pB_3)^{-1}\log(B_2/\delta)$. And, by Lemma 16, there exist constants $C_1 > 0$, $C_2 > 1$ depending on $\mathcal{V}$, ϕ and a universal constant $C_3 > 0$ such that

$$p(1-p)N \sup_{s\in\mathcal{S}_0} |\mathcal{R}_{n,m}(\ell_s)| \leq \|\phi'\|_\infty \sqrt{p(1-p)C_3\log(C_2/\delta)} \ , \tag{B.30}$$

for $\log(C_2/\delta) \geq C_1(\|\phi'\|_\infty^2 C_3)^{-1}$. The union bound concludes by considering constants such that with probability at least $1-\delta$

$$\sup_{s\in\mathcal{S}_0} |\mathcal{R}^B_{n,m}(s)| \leq \|\phi'\|_\infty \sqrt{p(1-p)C_3\log(3C_2/\delta)} + (p\|\phi'\|_\infty/B_3)\log(3B_2/\delta) \ , \tag{B.31}$$

as soon as $N \geq (pD_3)^{-1}\log(D_2/\delta)$, where $D_2 = 3\max(A_2, B_2)$ and $D_3 = \min(A_3, B_3)$.

B.3 Proof of Theorem 5

Observe, by virtue of Proposition 4 and for all $s \in \mathcal{S}_0$

$$\begin{aligned}\left|\frac{1}{n}\widehat{W}^{\phi}_{n,m}(s) - W_\phi(s)\right| \leq \frac{1}{n}\left|\sum_{i=1}^n \phi\circ F_s(s(\mathbf{X}_i)) - \mathbb{E}[\phi\circ F_s(s(\mathbf{X}))]\right| \\ + \frac{1}{N}\left|\sum_{i=1}^n k_{s,1,1}(\mathbf{X}_i)\right| + \frac{1}{N}\left|\sum_{j=1}^m \ell_{s,1,1}(\mathbf{Y}_j)\right| + \frac{1}{n}\left|\mathcal{R}_{n,m}(s)\right| \ .\end{aligned}$$

Under Assumptions 2-3, we sequentially provide uniform bounds in probability for all processes. The classes of kernels $\{x \mapsto k_{s,1,1}(x) : \ s \in \mathcal{S}_0\}$ and $\{y \mapsto \ell_{s,1,1}(y) : \ s \in \mathcal{S}_0\}$, by Lemma 20, are bounded and VC-type of parameters depending on ϕ and on the VC dimension $\mathcal{V}$ of $\mathcal{S}_0$. Their variance can be bounded, for all $s \in \mathcal{S}_0$, by $\sigma^2 = \int_{[0,1]}\phi'^2$ and $\sigma^2 \leq ||\phi'||^2_\infty$. As well for the collection $\{x \mapsto \phi\circ F_s(s(x)) : s \in \mathcal{S}_0\}$ where the arguments are detailed in Lemma 19 and of variance bounded by, for all $s \in \mathcal{S}_0$, by $\Sigma^2 = \int_{[0,1]}\phi^2$ and $\Sigma^2 \leq ||\phi||^2_\infty$. Similarly to Lemma 17, we apply Theorem 2.1 in [13] to the empirical processes $\widehat{W}_\phi(s)$, $\widehat{V}^X_n(s)$ and $\widehat{V}^Y_m(s)$ as follows.
Let $t > 0$. There exist a sequence of constants $C_{1,i} > 0, C_{2,i} \geq 1, C_{3,i} > 0$ depending on ϕ and $\mathcal{V}$, such that for all $C_{4,i} \geq C_{1,i}$, $i \in \{1,2,3\}$, the following inequalities hold true.

$$\mathbb{P}\left\{\sup_{s\in\mathcal{S}_0}\left|\widehat{W}_\phi(s) - W_\phi(s)\right| > t\right\} \leq C_{2,1}\exp\left\{-\frac{C_{3,1}pNt^2}{\Sigma^2}\right\} \ , \tag{B.32}$$

as soon as $C_{1,1}\|\phi\|_\infty\sqrt{(1/pN)\log(2\|\phi\|_\infty/\Sigma)} \leq t \leq C_{4,1}\|\phi\|_\infty$.

$$\mathbb{P}\left\{\frac{1}{N}\sup_{s\in\mathcal{S}_0}\left|\sum_{i=1}^n k_{s,1,1}(\mathbf{X}_i)\right| > t\right\} \leq C_{2,2}\exp\left\{-\frac{C_{3,2}Nt^2}{p\sigma^2}\right\} \ , \tag{B.33}$$

as soon as $C_{1,2}||\phi'||_\infty\sqrt{(p/N)\log(2\|\phi'\|_\infty/\sigma)} \leq t \leq pC_{4,2}||\phi'||_\infty$.

$$\mathbb{P}\left\{\frac{1}{N}\sup_{s\in\mathcal{S}_0}\left|\sum_{j=1}^{m}\ell_{s,1,1}(\mathbf{Y}_j)\right|>t\right\}\leq C_{2,3}\exp\left\{-\frac{C_{3,3}Nt^2}{(1-p)\sigma^2}\right\}\ , \tag{B.34}$$

as soon as $C_{1,3}||\phi'||_\infty\sqrt{((1-p)/N)\log(2\|\phi'\|_\infty/\sigma)}\leq t\leq(1-p)C_{4,3}||\phi'||_\infty$. Proposition 4 provides the existence of constants $C>6$, $D>0$ and $c_3>0$, $c_5>3$ depending on ϕ and $\mathcal{V}$, such that

$$\mathbb{P}\left\{\frac{1}{n}\sup_{s\in\mathcal{S}_0}|\mathcal{R}_{n,m}(s)|>t\right\}\leq C\exp\left\{-\frac{pNt}{(\|\phi'\|_\infty\kappa_pD+B_2)}\right\}\ , \tag{B.35}$$

as soon as $N\geq(c_3/p)\log(c_5/\delta)$. The remainder process is negligible with respect to the empirical processes and we gather the four bounds to get

$$\mathbb{P}\left\{\sup_{s\in\mathcal{S}_0}\left|\frac{1}{n}\widehat{W}^{\phi}_{n,m}(s)-W_\phi(s)\right|>t\right\}\leq C_2e^{-C_3Nt^2}\ , \tag{B.36}$$

where $C_2=4\max(\{C_{2,i},i\leq3\},C)$, $C_3=(1/9)\min(C_{3,1}p/\Sigma^2,C_{3,2}/(p\sigma^2),C_{3,3}/((1-p)\sigma^2))$, as soon as (B.35) is satisfied and $C_1/\sqrt{pN}\leq t\leq C_4\min(p,1-p)$, $C_1>0$ depending on ϕ, $\mathcal{V}$ and $C_4\geq\max(C_{1,i},\ i\leq3)$ depending on ϕ, $C_4=\min(C_{4,1}\|\phi\|_\infty,C_{4,2}p\|\phi'\|_\infty,C_{4,3}(1-p)\|\phi''\|_\infty)$.

B.4 A Generalization Bound in Expectation

For the sake of completeness, we state and prove a version in expectation of the generalization result formulated in Corollary 7.

Proposition. 21. *Under the assumptions of Proposition 4, the expected risk bound is derived as follows:*

$$\mathbb{E}\left[W^*_\phi-W_\phi(\hat{s})\right]\leq B_1\sqrt{\frac{\mathcal{V}}{pN}}+W^*_\phi-\mathbb{E}\left[\sup_{s\in\mathcal{S}_0}W_\phi(s)\right]\ , \tag{B.37}$$

for $pN\geq B_2\mathcal{V}$ with constants $B_1,B_2>0$ depending on ϕ, $\mathcal{V}$.

PROOF. Following the decomposition (3.9), we bound in expectation each process recalling that they are indexed by uniformly bounded VC-type classes, refer to Proof B.3 for the details on theoretical guarantees concerning the permanence properties. For the empirical processes $\widehat{W}_\phi$, $\widehat{V}^X_n$ and $\widehat{V}^Y_m$, we use Theorem 2.1 in [13], whereas for the remainder process, we require the following result that is proved subsequently.

Lemma. 22. *Under the assumptions of Proposition 4, the remainder process can be uniformly bounded in expectation as follows:*

$$\mathbb{E}\left[\sup_{s\in\mathcal{S}_0}|\mathcal{R}_{n,m}(s)|\right]\leq D_1(1+1/p+1/\sqrt{p(1-p)})\ , \tag{B.38}$$

for $pN\geq D_2\mathcal{V}$ with constants $D_1>0$ depending on ϕ, $\mathcal{V}$ and $D_2>0$ on ϕ.

By means of [13], there exist universal constants $B_i > 0$, and $b_i > 0$, $i \in \{1, 2, 3\}$, depending on ϕ, $\mathcal{V}$ such that the inequalities below hold true.

$$\mathbb{E}\left[\sup_{s\in\mathcal{S}_0}\left|\widehat{W}_\phi(s) - W_\phi(s)\right|\right] \leq B_1\left(b_1\frac{\mathcal{V}\|\phi\|_\infty}{pN} + \|\phi\|_\infty\sqrt{b_1\frac{\mathcal{V}}{pN}}\right) , \tag{B.39}$$

and

$$\mathbb{E}\left[\frac{1}{n}\sup_{s\in\mathcal{S}_0}\left|\widehat{V}_n^X(s) - \mathbb{E}\left[\widehat{V}_n^X(s)\right]\right|\right] \leq B_2\left(b_2\frac{\mathcal{V}\|\phi'\|_\infty}{pN} + \|\phi'\|_\infty\sqrt{b_2\frac{\mathcal{V}}{pN}}\right) , \tag{B.40}$$

as well as

$$\mathbb{E}\left[\frac{1}{n}\sup_{s\in\mathcal{S}_0}\left|\widehat{V}_m^Y(s) - \mathbb{E}\left[\widehat{V}_m^Y(s)\right]\right|\right] \leq B_3\left(b_3\frac{\mathcal{V}\|\phi'\|_\infty}{pN} + \|\phi'\|_\infty\sqrt{b_3\frac{\mathcal{V}}{pN}}\right) , \tag{B.41}$$

observing that $\int_{[0,1]}\phi^2 \leq \|\phi\|_\infty^2$ and $\int_{[0,1]}\phi'^2 \leq ||\phi'||_\infty^2$.
The remainder process being of higher order, we conclude

$$\mathbb{E}\left[\sup_{s\in\mathcal{S}_0}\left|\frac{1}{n}\widehat{W}_{n,m}^\phi(s) - W_\phi(s)\right|\right] \leq B\sqrt{b\frac{\mathcal{V}}{pN}} , \tag{B.42}$$

for $pN \geq \max(b, D_2)\mathcal{V}$ with constants $B > 0$ depending on ϕ and $b > 0$ depending on ϕ, $\mathcal{V}$.

PROOF. For all $s \in \mathcal{S}_0$

$$|\mathcal{R}_{n,m}(s)| \leq |\widehat{R}_{n,m}(s)| + N|\mathcal{R}_n(k_s)| + N|\mathcal{R}_{n,m}(\ell_s)| + |\widehat{T}_{n,m}(s)| \tag{B.43}$$

The process appearing first in the remainder induced by the Hájek projection method (Lemma 17), is composed of sums of empirical processes, hence applying Theorem 2.1 in [13] to each process of (B.24) yields

$$\mathbb{E}\left[\sup_{s\in\mathcal{S}_0}\left|\widehat{R}_{n,m}(s)\right|\right] \leq D_1\left(d\frac{\mathcal{V}\|\phi'\|_\infty}{N} + \|\phi'\|_\infty\sqrt{d\frac{p\mathcal{V}}{N}}\right) , \tag{B.44}$$

with constants $D_1 > 0$ depending on ϕ and $d > 0$ on ϕ, $\mathcal{V}$.The stochastic processes $\mathcal{R}_n(k_s)$ and $\mathcal{R}_{n,m}(\ell_s)$ being both degenerate U-processes, respectively one-sample of degree 2 and two-sample of degree $(1, 1)$, we apply results in [29] (see Theorem 6 therein) and [28] (see Lemma 2.4 therein) so as to get

$$\mathbb{E}\left[\sup_{s\in\mathcal{S}_0}|\mathcal{R}_n(k_s)|\right] \leq \frac{D_2\mathcal{V}}{pN} , \tag{B.45}$$

and

$$\mathbb{E}\left[\sup_{s\in\mathcal{S}_0}|\mathcal{R}_{n,m}(\ell_s)|\right] \leq \frac{D_3\mathcal{V}}{\sqrt{p(1-p)}N} , \tag{B.46}$$

D_2, $D_3 > 0$ constants of $\phi, \mathcal{V}$. For $\widehat{T}_{n,m}(s)$, the concentration inequality proved in Eq. (B.22) holds true for all $\delta \in (0, 1)$. Hence, we have

$$\mathbb{E}\left[\sup_{s\in\mathcal{S}_0}\left|\widehat{T}_{n,m}(s)\right|\right] \leq u + \int_u^\infty \mathbb{P}\left\{\sup_{s\in\mathcal{S}_0}\left|\widehat{T}_{n,m}(s)\right| \geq x\right\} dx$$
$$= u + 2B_2 e^{-(u-B_1)/B_2} \ . \quad \text{(B.47)}$$

Minimizing the bound above *w.r.t.* $u > 0$, we obtain the point $B_1 + B_2\log(2)$ and the upperbound then writes $B_1 + B_2(1+\log(2))$, where B_1 (*resp.* B_2) is a constant that only depends on ϕ and $\mathcal{V}$ (*resp.* on ϕ). Combining all bounds together permits to conclude: for $N \geq \mathcal{V}\log(d)$, we have

$$\mathbb{E}\left[\sup_{s\in\mathcal{S}_0}|\mathcal{R}_{n,m}(s)|\right] \leq D_1\|\phi'\|_\infty + \frac{D_2\mathcal{V}}{p} + \frac{D_3\mathcal{V}}{\sqrt{p(1-p)}} + B_1 + B_2(1+\log(2))$$
$$\leq D(1 + 1/p + 1/\sqrt{p(1-p)}) \ , \quad \text{(B.48)}$$

where $D > 0$ constant depending on ϕ, $\mathcal{V}$. □

B.5 Proof of Proposition 8

We first prove the following lemma.

Lemma. 23. *Let $\mathcal{S}_0 \subset \mathcal{S}$ and suppose that Assumptions 1-3 are fulfilled. For all $t > 0$, we have:*

$$\mathbb{P}\left\{\sup_{s\in\mathcal{S}_0}\left|W_\phi(s) - \widehat{W}^\phi_{n,m}(s)/n\right| \geq \mathbb{E}\left[\sup_{s\in\mathcal{S}_0}\left|W_\phi(s) - \widehat{W}^\phi_{n,m}(s)/n\right|\right] + t\right\}$$
$$\leq \exp\left\{-\frac{p^2Nt^2}{6(\|\phi\|^2_\infty + 9\|\phi'\|^2_\infty + 9||\phi''||^2_\infty)}\right\} \ . \quad \text{(B.49)}$$

PROOF. Recall the decomposition of $\widehat{W}^\phi_{n,m}(s)$, for all $s \in \mathcal{S}_0$, proved in Proposition 4

$$\widehat{W}_{n,m}(s) = n\widehat{W}_\phi(s) + B_{n,m}(s) + \widehat{T}_{n,m}(s) \ . \quad \text{(B.50)}$$

Considering that $\sup_{s\in\mathcal{S}_0}\left|W_\phi(s) - \widehat{W}^\phi_{n,m}(s)/n\right|$ is a function of the N independent random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_m$, observe that changing the value of any of the $\mathbf{X}_i$'s while keeping all the others fixed changes the value of the supremum by at most

$$2||\phi||_\infty + 2||\phi'||_\infty\left(1 + \frac{m + 2(n-1)}{N+1}\right) + 2||\phi''||_\infty\frac{1+2m}{N^2} \ ,$$

taking into account the jumps of each of the three terms involved in (B.50), see Eq. (B.7) and (B.20). In a similar way, changing the value of any of the $\mathbf{Y}_j$'s changes the value of the supremum by at most

$$2||\phi'||_\infty\frac{n}{N+1} + 2||\phi''||_\infty\frac{1+2n}{N^2} \ .$$

When taking the squares, both can be upperbounded by $12(\|\phi\|^2_\infty + 9\|\phi'\|^2_\infty + 9||\phi''||^2_\infty)$. The desired bound stated then straightforwardly results from the application of the bounded difference inequality, see [25]. □

Let $\varepsilon > 0$, using Proposition 21 and Lemma 23, we have, for any $k \geq 1$,

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{W}^{\phi}_{n,m}(\hat{s}_k) - B_1\sqrt{\frac{\mathcal{V}_k}{pN}} - W_\phi(\hat{s}_k) > \varepsilon\right\} \\
&\quad \leq \mathbb{P}\left\{\sup_{s\in\mathcal{S}_k}\left|W_\phi(s) - \widehat{W}^{\phi}_{n,m}(s)/n\right| > \mathbb{E}\left[\sup_{s\in\mathcal{S}_k}\left|W_\phi(s) - \widehat{W}^{\phi}_{n,m}(s)/n\right|\right] + \varepsilon\right\} \\
&\qquad\qquad \leq \exp\left\{-\frac{p^2N\varepsilon^2}{C}\right\}\ , \quad \text{(B.51)}
\end{aligned}$$

as soon as $pN \geq B_2\mathcal{V}_k$ and where $C = 6(\|\phi\|^2_\infty + 9\|\phi'\|^2_\infty + 9\|\phi''\|^2_\infty)$. For each $k \geq 1$, denote the penalized empirical ranking performance measure by

$$\widehat{W}^{\phi,k}_{n,m}(\hat{s}_k)/n = \widehat{W}^{\phi}_{n,m}(\hat{s}_k)/n - B_1\sqrt{\frac{\mathcal{V}_k}{pN}} - \sqrt{\frac{2C\log k}{p^2N}}. \quad \text{(B.52)}$$

For any $\varepsilon > 0$, we have, as soon as $pN \geq B_2 \sup_{k\geq 1}\mathcal{V}_k$,

$$\begin{aligned}
\mathbb{P}\left\{\widehat{W}^{\phi,\hat{k}}_{n,m}(\hat{s}_{\hat{k}})/n - W_\phi(\hat{s}_{\hat{k}}) \geq \varepsilon\right\} &\leq \sum_{k\geq 1}\mathbb{P}\left\{\widehat{W}^{\phi,k}_{n,m}(\hat{s}_k)/n - W_\phi(\hat{s}_k) \geq \varepsilon\right\} \\
&\leq \sum_{k\geq 1}\mathbb{P}\left\{\widehat{W}^{\phi}_{n,m}(\hat{s}_k)/n - B_1\sqrt{\frac{\mathcal{V}_k}{pN}} - W_\phi(\hat{s}_k) > \varepsilon + \sqrt{\frac{2C\log k}{p^2N}}\right\} \\
&\leq \sum_{k\geq 1}\exp\left(-\frac{p^2N}{C}\left(\varepsilon + \sqrt{\frac{2C\log k}{p^2N}}\right)^2\right) \\
&\leq \exp\left(-\frac{p^2N\varepsilon^2}{C}\right)\sum_{k\geq 1}k^{-2} < 2\exp\left\{-\frac{p^2N\varepsilon^2}{C}\right\}\ . \quad \text{(B.53)}
\end{aligned}$$

For all $k \geq 1$, $W^*_k = \sup_{s\in\mathcal{S}_k} W_\phi(s) = W_\phi(s^*_k)$ and consider the decomposition

$$W^*_k - W_\phi(\hat{s}_{\hat{k}}) = \left(W^*_k - \widehat{W}^{\phi,\hat{k}}_{n,m}(\hat{s}_{\hat{k}})/n\right) + \left(\widehat{W}^{\phi,\hat{k}}_{n,m}(\hat{s}_{\hat{k}})/n - W_\phi(\hat{s}_{\hat{k}})\right)\ .$$

The expectation of the second term of the right hand side of the equation above can be bounded by means of the tail bound (B.53)

$$\mathbb{E}\left[\widehat{W}^{\phi,\hat{k}}_{n,m}(\hat{s}_{\hat{k}})/n - W_\phi(\hat{s}_{\hat{k}})\right] \leq 2\sqrt{\frac{C}{p^2N}}\ . \quad \text{(B.54)}$$

for any $k \geq 1$, as soon as $pN \geq B_2 \sup_{k\geq 1}\mathcal{V}_k$. Concerning the expectation of the first term, observe that

$$\begin{aligned}
\mathbb{E}\left[W^*_k - \widehat{W}^{\phi,\hat{k}}_{n,m}(\hat{s}_{\hat{k}})/n\right] &\leq \mathbb{E}\left[W^*_k - \widehat{W}^{\phi,k}_{n,m}(s^*_k)\right] \\
&\leq \mathbb{E}\left[W_\phi(s^*_k) - \widehat{W}^{\phi}_{n,m}(s^*_k)\right] + \mathrm{pen}(N,k) \leq B_1\sqrt{\frac{\mathcal{V}_k}{pN}} + \mathrm{pen}(N,k)\ ,
\end{aligned}$$

for any $k \geq 1$, as soon as $pN \geq B_2 \sup_{k\geq 1}\mathcal{V}_k$. Summing the bound obtained and that in (B.54) gives the desired result.

B.6 Proof of Proposition 9

The proof consists in combining the two results stated below with the decomposition (4.11) of the W_ϕ-ranking performance deficit of the maximizer. The first result is the analogue of Theorem 5 for the smoothed criterion.

Theorem. 24. *Suppose that the assumptions of Proposition 4 are fulfilled. Then, for any $\delta \in (0,1)$, there exist constants $C_1,\ C_3 > 0,\ C_2 \geq 24$, depending on $\phi,\ K,\ R,\ \mathcal{V}$ such that with probability larger than $1-\delta$:*

$$\sup_{s\in\mathcal{S}_0} \left| \widehat{W}^{\phi}_{n,m,h}(s)/n - \widetilde{W}_{\phi,h}(s) \right| \leq \sqrt{\frac{\log(C_2/\delta)}{pC_3N}}\ , \tag{B.55}$$

as soon as $N \geq 1/(p\min(p,1-p)^2 C_3 C_4^2)\log(C_2/\delta)$ and $\delta \leq C_2 e^{-C_1^2 C_3}$.

The proof being quite similar to that of Theorem 5, it is omitted. Assumption 5 ensuring that the class $\{K((\cdot - t)/h);,\ t \in \mathbb{R}^q,\ h > 0\}$ ($q = 1$ here) is bounded VC-type (see *e.g.* Lemma 22(ii) in [29] and [14]), classic permanence properties can be used to check that all the classes of functions over which uniform bounds are taken are of finite VC dimension. The second result provides a uniform bound for the additional bias error made when approximating $W_\phi(s)$ by $\widetilde{W}_{\phi,h}(s)$ for $s \in \mathcal{S}_0$.

Lemma. 25. *Suppose that Assumptions 4 is satisfied. Then, for all $h > 0$, we have:*

$$\sup_{s\in\mathcal{S}_0} \left| \widetilde{W}_{\phi,h}(s) - W_\phi(s) \right| \leq C_5 h^2, \tag{B.56}$$

where $C_5 > 0$ is a constant depending on ϕ, K and R only.

Details are left to the reader, the proof is straightforward under Assumption 4, using the regularity of the score generating function and the uniform integrated error bound obtained in [21].

B.7 Proof of Lemma 16

We shall prove an exponential bound of Hoeffding's type for the uniformly bounded two-sample degenerate U-process $\{U_{n,m}(\ell) :\ \ell \in \mathcal{L}\}$, where

$$U_{n,m}(\ell) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} \ell(X_i, Y_j)\ . \tag{B.57}$$

In order to apply standard symmetrization arguments, see *e.g.* section 2.3 in [36], consider independent Rademacher variables $\varepsilon_1, \ldots, \varepsilon_n$ and $\eta_1, \ldots, \eta_m$ and define

$$T_{n,m}(\ell) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} \varepsilon_i \eta_j \ell(X_i, Y_j)\ , \tag{B.58}$$

for all ℓ in $\mathcal{L}$. We start by proving the following lemmas, involved in the argument.

Lemma. 26. *Let P and Q be probability distributions on measurable spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively. Consider the degenerate two-sample U-statistic of degree $(1,1)$ (B.57) with a bounded kernel $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ based on the independent i.i.d. random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, drawn from P and Q respectively. Let two sequences of i.i.d. Rademacher variables $\varepsilon_1, \ldots, \varepsilon_n$ and $\eta_1, \ldots, \eta_m$, independent of the X_i's and Y_j's, such that the randomized process (B.58) is defined. Then, for any increasing and convex function $\Phi : \mathbb{R} \to \mathbb{R}$, we have:*

$$\mathbb{E}\left[\Phi\left(\sup_{\ell\in\mathcal{L}}|U_{n,m}(\ell)|\right)\right] \leq \mathbb{E}\left[\Phi\left(4\sup_{\ell\in\mathcal{L}}|T_{n,m}(\ell)|\right)\right], \tag{B.59}$$

and

$$\mathbb{E}\left[\Phi\left(\sup_{\ell\in\mathcal{L}}U_{n,m}(\ell)\right)\right] \leq \mathbb{E}\left[\Phi\left(4\sup_{\ell\in\mathcal{L}}\ T_{n,m}(\ell)\right)\right], \tag{B.60}$$

assuming that the suprema are measurable and that the expectations exist.

PROOF. We prove the first inequality, the proof of the second one being similar. Using the independence of the two samples, Fubini's theorem and the degeneracy property, one gets that

$$\begin{aligned}
& \mathbb{E}\left[\Phi\left(\sup_{\ell\in\mathcal{L}}|U_{n,m}(\ell)|\right)\right] \\
= \ & \mathbb{E}\left[\mathbb{E}\left[\Phi\left(\sup_{\ell\in\mathcal{L}}\left|\frac{1}{nm}\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\ell(X_i,Y_j)\right)\right|\right) \mid Y_1, \ldots, Y_m\right]\right] \\
\leq \ & \mathbb{E}\left[\Phi\left(2\sup_{\ell\in\mathcal{L}}\left|\frac{1}{nm}\sum_{i=1}^{n}\varepsilon_i\left(\sum_{j=1}^{m}\ell(X_i,Y_j)\right)\right|\right)\right] \\
= \ & \mathbb{E}\left[\mathbb{E}\left[\Phi\left(2\sup_{\ell\in\mathcal{L}}\left|\frac{1}{nm}\sum_{j=1}^{m}\left(\sum_{i=1}^{n}\varepsilon_i\ell(X_i,Y_j)\right)\right|\right) \mid (X_1,\varepsilon_1), \ldots, (X_n,\varepsilon_n)\right]\right] \\
\leq \ & \mathbb{E}\left[\Phi\left(4\sup_{\ell\in\mathcal{L}}\left|\frac{1}{nm}\sum_{j=1}^{m}\eta_j\left(\sum_{i=1}^{n}\varepsilon_i\ell(X_i,Y_j)\right)\right|\right)\right] \\
= \ & \mathbb{E}\left[\Phi\left(4\sup_{\ell\in\mathcal{L}}|T_{n,m}(\ell)|\right)\right]
\end{aligned}$$

by applying Lemma 3.5.2 of [33] twice. Incidentally, notice that we can also show that

$$\mathbb{E}\left[\Phi\left(\frac{1}{4}\sup_{\ell\in\mathcal{L}}|T_{n,m}(\ell)|\right)\right] \leq \mathbb{E}\left[\Phi\left(\sup_{\ell\in\mathcal{L}}|U_{n,m}(\ell)|\right)\right].$$

by applying twice the reverse inequality in Lemma 3.5.2 of [33]. □

Next, we prove an exponential bound of Hoeffding's type for degenerate two-sample U-statistics with bounded kernels.

Lemma. 27. *Let P and Q be probability distributions on measurable spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively. Consider the degenerate two-sample U-statistic of degree*

$(1,1)$ (B.57) *with a bounded kernel* $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ *based on the independent i.i.d. random samples* $X_1, \ldots, X_n$ *and* $Y_1, \ldots, Y_m$*, drawn from* P *and* Q *respectively. For all* $t > 0$*, we then have:*

$$\mathbb{P}\left\{U_{n,m}(\ell) \geq t\right\} \leq e^{-nmt^2/(32c_\ell^2)}, \tag{B.61}$$

where $c_\ell = \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |\ell(x,y)| < +\infty$.

PROOF. Let $t > 0$. The proof is based on Chernoff's method. For all $\lambda > 0$, we have

$$\begin{aligned}\mathbb{P}\left\{U_{n,m}(\ell) \geq t\right\} &\leq \exp\left(-\lambda t + \log\left(\mathbb{E}[\exp(\lambda U_{n,m}(\ell))]\right)\right) \\ &\leq \exp\left(-\lambda t + \log\left(\mathbb{E}[\exp(4\lambda T_{n,m}(\ell))]\right)\right), \end{aligned} \tag{B.62}$$

using (B.60) with $\Phi(t) = \exp(\lambda t)$. Observe next that we almost-surely

$$\begin{aligned}\mathbb{E}[\exp(4\lambda T_{n,m}(\ell)) \mid X_1, \ldots, X_n, Y_1, \ldots, Y_m] &= \\ \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{e^{4\lambda\ell(X_i,Y_j)/(nm)} + e^{-4\lambda\ell(X_i,Y_j)/(nm)}}{2} & \\ \leq \prod_{i=1}^{n}\prod_{j=1}^{m} e^{8\lambda^2\ell^2(X_i,Y_j)/(nm)^2} \leq e^{8\lambda^2 c_\ell^2/(nm)}, &\end{aligned}$$

using the fact that $(e^u + e^{-u})/2 \leq e^{u^2/2}$ for all $u \in \mathbb{R}$. Integrating the bound over the X_i's and Y_j's and plugging it next into (B.62) yields the desired bound when choosing $\lambda = nmt/(16c_\ell^2)$. □

Finally, we prove the tail probability version of Lemma 26 stated below.

Lemma. 28. *Let* P *and* Q *be probability distributions on measurable spaces* $\mathcal{X}$ *and* $\mathcal{Y}$ *respectively. Consider the degenerate two-sample* U*-statistic of degree* $(1,1)$ (B.57) *with a bounded kernel* $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ *based on the independent i.i.d. random samples* $X_1, \ldots, X_n$ *and* $Y_1, \ldots, Y_m$*, drawn from* P *and* Q *respectively. Let two sequences of i.i.d. Rademacher variables* $\varepsilon_1, \ldots, \varepsilon_n$ *and* $\eta_1, \ldots, \eta_m$*, independent of the* Xis *and* Yjs*, such that the randomized process* (B.58) *is defined. Then we have for all* $t > 0$*,*

$$\mathbb{P}\left\{\sup_{\ell\in\mathcal{L}} |U_{n,m}(\ell)| \geq 16t\right\} \leq 16\mathbb{P}\left\{\sup_{\ell\in\mathcal{L}} |T_{n,m}(\ell)| \geq t\right\}, \tag{B.63}$$

assuming that the suprema are measurable and that the expectations exist.

PROOF. This lemma, bounding the tail probability of $\sup_{\ell\in\mathcal{L}} |U_{n,m}(\ell)|$ to that of $\sup_{\ell\in\mathcal{L}} |T_{n,m}(\ell)|$, generalizes Lemma 2.7 in [15] and Lemma 3.1 in [32] to degenerate two-sample U-processes. It is proved by applying twice a version of the latter result for independent but non necessarily identically distributed random variables. Indeed, we have: $\forall t > 0$,

$$
\begin{aligned}
&\mathbb{P}\left\{\sup_{\ell\in\mathcal{L}}|U_{n,m}(\ell)|\geq 16t\right\}\\
&\quad=\mathbb{E}\left[\mathbb{P}\left\{\sup_{\ell\in\mathcal{L}}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{m}\sum_{j=1}^{m}\ell(X_i,Y_j)\right\}\right|\geq 16t\mid Y_1,\ \ldots,\ Y_m\right\}\right]\\
&\quad\leq 4\mathbb{E}\left[\mathbb{P}\left\{\sup_{\ell\in\mathcal{L}}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{m}\sum_{j=1}^{m}\epsilon_i\ell(X_i,Y_j)\right\}\right|\geq 4t\mid Y_1,\ \ldots,\ Y_m\right\}\right]\\
&=4\mathbb{E}\left[\mathbb{P}\left\{\sup_{\ell\in\mathcal{L}}\left|\frac{1}{m}\sum_{j=1}^{m}\left\{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ell(X_i,Y_j)\right\}\right|\geq 4t\mid (X_1,\epsilon_1)\ \ldots,\ (X_n,\ \epsilon_n)\right\}\right]\\
&\quad\leq 16\mathbb{P}\left\{\sup_{\ell\in\mathcal{L}}|T_{n,m}(\ell)|\geq t\right\}.
\end{aligned}
$$

□

The proof relies on the chaining method applied to the process $U_{n,m}(\ell)$ indexed by the class of kernels $\mathcal{L}$, see *e.g.* the argument used to establish Lemma 2.14.9 in [36]. Define the random semi-metric on $\mathcal{L}$ by

$$
d_{nm}^2(\ell_1,\ell_2)=\frac{1}{nm}\sum_{i\leq n}\sum_{j\leq m}(\ell_1(X_i,Y_j)-\ell_2(X_i,Y_j))^2 \tag{B.64}
$$

for all kernels ℓ_1 and ℓ_2 in $\mathcal{L}$. For all $q\in\mathbb{N}^*$, consider a number $k_q\leq(A/\varepsilon_q)^{\mathcal{V}}$ of L_2-balls with radius $\varepsilon_q\leq L\leq 1$ and centers $\ell_{q,k}$, $1\leq k\leq k_q$, *w.r.t.* the (random) probability measure $(1/nm)\sum_{i\leq n}\sum_{j\leq m}\delta_{(X_i,Y_j)}$ covering the class $\mathcal{L}$. Assume that the sequence ε_q is decreasing as q increases, so that k_q is increasing. Let $\ell\in\mathcal{L}$, $q\geq 1$ and $\tilde{\ell}_q$ be the center of a ball s.t. $d_{nm}(\ell,\tilde{\ell}_q)\leq\varepsilon_q$. Fixing $q_0\leq q$ in $\mathbb{N}^*$, the following decomposition holds

$$
U_{n,m}(\ell)=(U_{n,m}(\ell)-U_{n,m}(\tilde{\ell}_q))+U_{n,m}(\tilde{\ell}_{q_0})+\sum_{\omega=q_0+1}^{q}\left(U_{n,m}(\tilde{\ell}_\omega)-U_{n,m}(\tilde{\ell}_{\omega-1})\right).
$$

Observe that, for all ℓ in $\mathcal{L}$, we almost-surely have

$$
|U_{n,m}(\ell)-U_{n,m}(\tilde{\ell}_q)|\leq d_{nm}(\ell,\tilde{\ell}_q)\leq\varepsilon_q\ .
$$

The triangular inequality yields

$$
\|U_{n,m}(\ell)\|_{\mathcal{L}}\leq\varepsilon_q+\max_{1\leq k\leq k_{q_0}}|U_{n,m}(\ell_{q_0,k})|+\sum_{\omega=q_0+1}^{q}||U_{n,m}(\tilde{\ell}_\omega)-U_{n,m}(\tilde{\ell}_{\omega-1})||_{\mathcal{L}}\ ,
$$

where we used the notation $||V||_{\mathcal{L}}=\sup_{\ell\in\mathcal{L}}|V(\ell)|$ for any real-valued stochastic process V indexed by $\mathcal{L}$. Considering $\eta_\omega>0$ and $\beta>0$ constants such that

$\sum_{\omega=q_0+1}^{q} \eta_\omega + \beta \leq 1$, we have for any $t > \varepsilon_q$:

$$\mathbb{P}\left\{\|U_{n,m}(\ell)\|_{\mathcal{L}} \geq 16t\right\} \leq \sum_{k=1}^{k_{q_0}} \mathbb{P}\left\{|U_{n,m}(\ell_{q_0,k}))| \geq 16t\beta\right\}$$
$$+16 \sum_{\omega=q_0+1}^{q} k_\omega^2 \mathbb{E}\left[\sup_{\ell\in\mathcal{L}} \mathbb{P}\left\{|T_{n,m}(\tilde{\ell}_\omega - \tilde{\ell}_{\omega-1})| \geq t\eta_\omega \mid X_1, \ldots, X_n, Y_1, \ldots, Y_m\right\}\right], \tag{B.65}$$

using the union bound, Lemma 28 and observing that the suprema corresponding to the terms of the series are actually maxima taken over at most $k_\omega k_{\omega-1} \leq k_\omega^2$ elements. Lemma 27 permits to bound the first term on the right hand side of (B.65):

$$\sum_{k=1}^{k_{q_0}} \mathbb{P}\left\{|U_{n,m}(\ell_{q_0,k}))| \geq 16t\beta\right\} \leq 2k_{q_0} \exp\left\{-\frac{8nm(t\beta)^2}{L^2}\right\} . \tag{B.66}$$

Concerning the second term, notice that

$$d_{nm}(\tilde{\ell}_\omega, \tilde{\ell}_{\omega-1}) \leq d_{nm}(\ell, \tilde{\ell}_{\omega-1}) + d_{nm}(\tilde{\ell}_\omega, \ell) \leq 2\varepsilon_{\omega-1} . \tag{B.67}$$

Re-using the start of the argument proving Lemma 27, we have: $\forall\lambda > 0$,

$$\mathbb{P}\left\{T_{n,m}(\tilde{\ell}_\omega - \tilde{\ell}_{\omega-1}) \geq t\eta_\omega \mid X_1, \ldots, X_n, Y_1, \ldots, Y_m\right\}$$
$$\leq \exp\left(-\lambda t\eta_\omega + \mathbb{E}\left[\exp(\lambda T_{n,m}(\tilde{\ell}_\omega - \tilde{\ell}_{\omega-1})) \mid X_1, \ldots, X_n, Y_1, \ldots, Y_m\right]\right)$$

with probability one. Like in Lemma 27's proof, we almost-surely have

$$\mathbb{E}[\exp(\lambda T_{n,m}(\tilde{\ell}_\omega - \tilde{\ell}_{\omega-1})) \mid X_1, \ldots, X_n, Y_1, \ldots, Y_m] \leq$$
$$\prod_{i=1}^{n}\prod_{j=1}^{m} e^{\lambda^2(\tilde{\ell}_\omega - \tilde{\ell}_{\omega-1})^2(X_i,Y_j)/2(nm)^2} \leq e^{2\lambda^2\epsilon_{\omega-1}^2/(nm)} .$$

Combining the two bounds above with the union bound, it holds with probability one

$$\mathbb{P}\left\{\left|T_{n,m}(\tilde{\ell}_\omega - \tilde{\ell}_{\omega-1})\right| \geq t\eta_\omega \mid X_1, \ldots, X_n, Y_1, \ldots, Y_m\right\} \leq$$
$$2\exp\left\{-\frac{nm(t\eta_\omega)^2}{8\varepsilon_{\omega-1}^2}\right\} . \tag{B.68}$$

From (B.65), (B.66) and (B.68), we deduce that

$$\mathbb{P}\left\{\|U_{n,m}(\ell)\|_{\mathcal{L}} \geq 16t\right\}$$
$$\leq 2k_{q_0} \exp\left\{-\frac{8nm(t\beta)^2}{L^2}\right\} + 32 \sum_{\omega=q_0+1}^{q} k_\omega^2 \exp\left\{-\frac{nm(t\eta_\omega)^2}{8\varepsilon_{\omega-1}^2}\right\}$$
$$\leq 2A^{\mathcal{V}}\varepsilon_{q_0}^{-\mathcal{V}} \exp\left\{-\frac{8nm(t\beta)^2}{L^2}\right\} + 32A^{2\mathcal{V}} \sum_{\omega=q_0+1}^{q} \varepsilon_\omega^{-2\mathcal{V}} \exp\left\{-\frac{nm(t\eta_\omega)^2}{8\varepsilon_{\omega-1}^2}\right\} . \tag{B.69}$$

Following Lemma 3.2 in [34] and choosing $\varepsilon_\omega = 2^{-\omega}L$, $\eta_\omega = 2^{-\omega}\sqrt{\omega}/8$, so that $\eta_{\omega+1}/\varepsilon_\omega = (1/16L)\sqrt{\omega+1}$, we have

$$\varepsilon_\omega^{-2\mathcal{V}} \exp\left\{-\frac{nm(t\eta_\omega)^2}{8\varepsilon_{\omega-1}^2}\right\} = L^{-2\mathcal{V}} \exp\left\{-(-2\mathcal{V}\log(2) + \frac{nmt^2}{4\times 8^3 L^2})\omega\right\} \quad \text{(B.70)}$$

If $nmt^2 > 8^4\log(2)L^2\mathcal{V}$, the terms of the series are decreasing *w.r.t.* ω and we upperbound by $K_1 L^{-2\mathcal{V}} \exp\left\{-nmt^2\omega/(4\times 8^3L^2)\right\}$. Problem 2.14.3 in [36] applies for $\omega \in \{q_0+1,\ldots,q\}$ with $\psi(\omega) = nmt^2\omega/(4\times 8^3L^2)$

$$\begin{aligned}\sum_{\omega=q_0+1}^{q} \varepsilon_\omega^{-2\mathcal{V}} \exp\left\{-\frac{nm(t\eta_\omega)^2}{8\varepsilon_{\omega-1}^2}\right\} &\leq K_1 L^{-2\mathcal{V}} \psi'(q_0)^{-1} \exp\left\{-\psi(q_0)\right\} \\ &\leq K_2 L^{-2(\mathcal{V}-1)} \exp\left\{-\frac{nmt^2}{4\times 8^3L^2} q_0\right\} \quad \text{(B.71)}\end{aligned}$$

K_1, $K_2 > 0$ constants and $nmt^2 \geq 1$. For $\alpha > 0$ large, setting $q_0 = 2 + \lfloor (nmt^2)^{1/(\alpha-1)} \rfloor$ yields to the upperbound $K_2 L^{-2(\mathcal{V}-1)} \exp\left\{-3nmt^2/(4\times 8^3L^2)\right\}$. For the first tail probability, by setting $\beta = 1/2 - 1/(2nmt^2)$ we obtain an upperbound of similar form

$$\begin{aligned}&A^{\mathcal{V}} \varepsilon_{q_0}^{-\mathcal{V}} \exp\left\{-\frac{8nm(t\beta)^2}{L^2}\right\} \\ &\quad\leq (A/L)^{\mathcal{V}} \exp\left\{\mathcal{V}\log(2)(2+(nmt^2)^{1/(\alpha-1)}) - \frac{2nmt^2}{L^2}(1-1/(nmt^2))^2\right\} \\ &\quad\leq (2A/L)^{\mathcal{V}} e^{4/L^2} \exp\left\{\mathcal{V}\log(2)(nmt^2)^{1/(\alpha-1)} - \frac{2nmt^2}{L^2}\right\} \\ &\quad\leq (2A/L)^{\mathcal{V}} e^{4/L^2} \exp\left\{-\frac{2nmt^2}{L^2}\right\},\end{aligned}$$

as soon as $nmt^2 > (\log(2)L^2\mathcal{V}/2)^{1+\delta}$, $\delta = 1/(\alpha-2) \in (0,1)$ for large α. Gathering both upperbounds, Eq. (B.69) yields

$$\mathbb{P}\left\{\|U_{n,m}(\ell)\|_{\mathcal{L}} \geq t\right\} \leq K 2^{\mathcal{V}+1} (A/L)^{2\mathcal{V}} e^{4/L^2} \exp\left\{-\frac{3nmt^2}{4\times 8^3L^2}\right\}, \quad \text{(B.72)}$$

for all $nmt^2 > \max(1, 8^4\log(2)L^2\mathcal{V}, (\log(2)L^2\mathcal{V}/2)^{1+\delta})$, and $K \geq 1 + 16K_2e^{-4}$ constant. Checking lastly that, for all $q \geq 1$

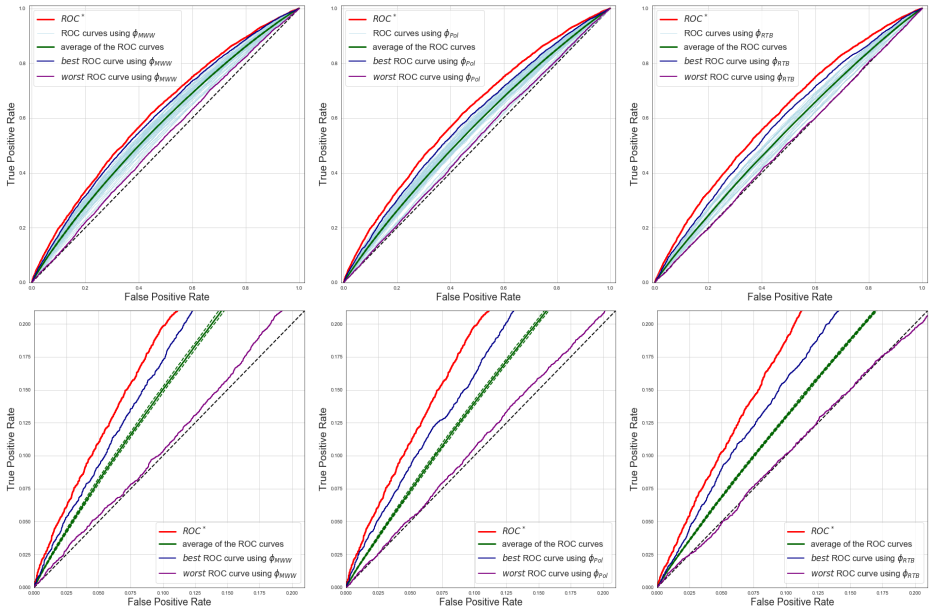
$$8\sum_{\omega=q_0+1}^{q}\eta_\omega \leq 8\sum_{\omega=1}^{q}\eta_\omega \leq 1 + \int_1^\infty 2^{-x}\sqrt{x}dx \leq 1 + (\pi/\log(2))^{1/2} \leq 4, \quad \text{(B.73)}$$

so that $\sum_{\omega=q_0+1}^{q}\eta_\omega + \beta \leq 1$ as needed.

C Additional Numerical Experiments

Following Section 5, this section gathers the numerical results of three models Loc1, Loc3 and Scale2, Scale3, as well as additional experiments regarding the difference in performance of the W-criteria for the RTB score-generating function, when we vary the rate u_0, for both the location (Fig. 14) and the scale (Fig. 15) models.

Location model. (Fig. 10, 11)

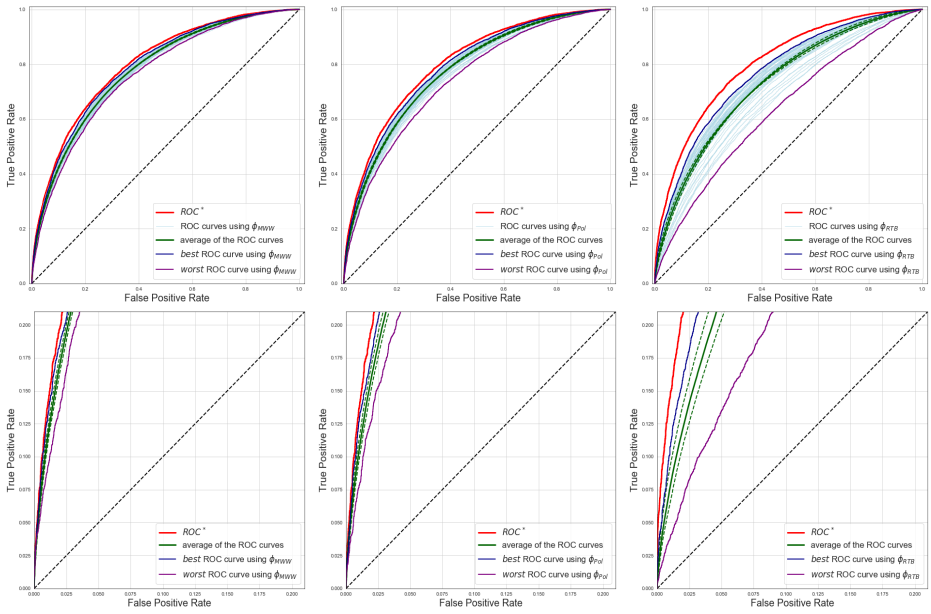


1. $\phi_{MWW}(u) = u$ 2. $\phi_{Pol}(u) = u^3$ 3. $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$

Figure 10: Empirical ROC curves and average ROC curve for Loc1 ($\varepsilon = 0.10$). Samples are drawn from multivariate Gaussian distributions according to section 5.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions. Hyperparameters: $u_0 = 0.9$, $q = 3$, $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$. Figures $1, 2, 3$ correspond *resp.* to the models MMW, Pol, RTB. Light blue curves are the $B(= 50)$ ROC curves that are averaged in green (solid line) with $+/-$ its standard deviation (dashed green lines). The dark blue and purple curves correspond to the best and worst scoring functions in the sense of minimization and maximization of the generalization error among the B curves. The red curve corresponds to ROC*.

Scale model. (Fig. 12, 13)

Comparison of three RTB score-generating functions for two location models. (Fig. 14)



1. $\phi_{MWW}(u) = u$ 2. $\phi_{Pol}(u) = u^3$ 3. $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$

Figure 11: Empirical ROC curves and average ROC curve for Loc3 ($\varepsilon = 0.30$). Samples are drawn from multivariate Gaussian distributions according to section 5.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions. Hyperparameters: $u_0 = 0.9$, $q = 3$, $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$. Figures $1, 2, 3$ correspond *resp.* to the models MMW, Pol, RTB. Light blue curves are the $B(= 50)$ ROC curves that are averaged in green (solid line) with $+/-$ its standard deviation (dashed green lines). The dark blue and purple curves correspond to the best and worst scoring functions in the sense of minimization and maximization of the generalization error among the B curves. The red curve corresponds to ROC*.

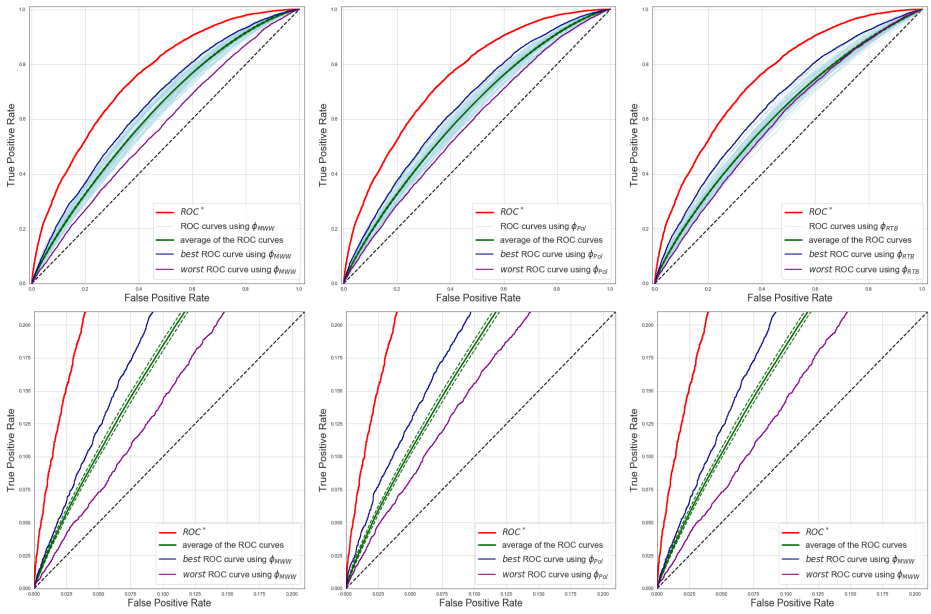
Comparison of three RTB score-generating functions for the scale model. (Fig. 15)

References

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

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

[3] S. Clémençon. A statistical view of clustering performance through the theory of U-processes. *Journal of Multivariate Analysis*, 124:42–56, 2014.



1. $\phi_{MWW}(u) = u$ 2. $\phi_{Pol}(u) = u^3$ 3. $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$

Figure 12: Empirical ROC curves and average ROC curve for Scale2 ($\varepsilon = 0.90$). Samples are drawn from multivariate Gaussian distributions according to section 5.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions. Hyperparameters: $u_0 = 0.9$, $q = 3$, $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$. Figures $1, 2, 3$ correspond *resp.* to the models MMW, Pol, RTB. Light blue curves are the $B(= 50)$ ROC curves that are averaged in green (solid line) with $+/-$ its standard deviation (dashed green lines). The dark blue and purple curves correspond to the best and worst scoring functions in the sense of minimization and maximization of the generalization error among the B curves. The red curve corresponds to ROC*.

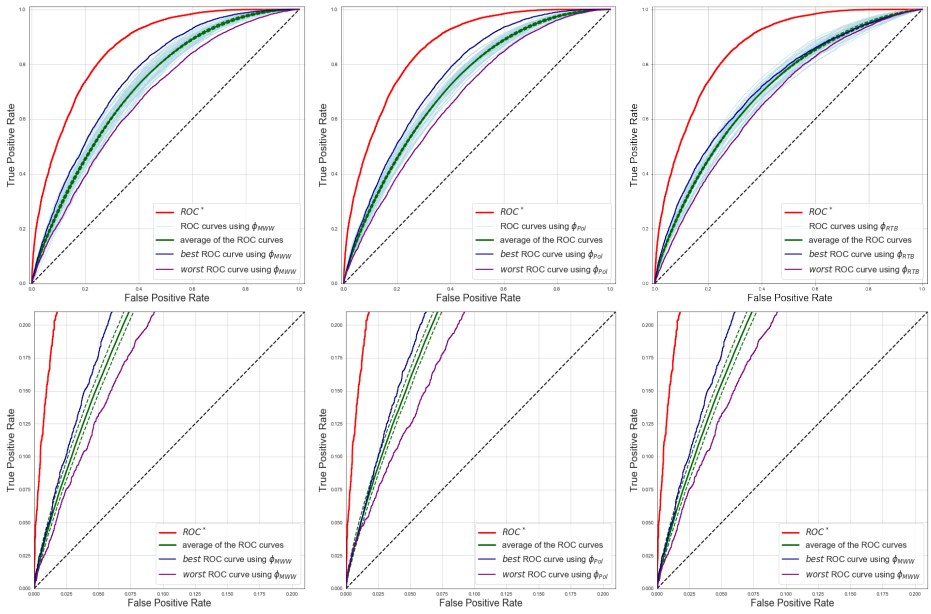
[4] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

[5] S. Clémençon and S. Robbiano. The TreeRank Tournament Algorithm for Multipartite Ranking. *Journal of Nonparametric Statistics*, 27(1):107–126, 2015.

[6] S. Clémençon, S. Robbiano, and N. Vayatis. Ranking Data with Ordinal Labels: Optimality and Pairwise Aggregation. *Machine Learning*, 93(1):67–104, 2013.

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

[8] S. Clémençon and N. Vayatis. Empirical performance maximization based on linear rank statistics. *Advances in Neural Information Processing Systems*, 3559:1–15, 2009.



1. $\phi_{MWW}(u) = u$ 2. $\phi_{Pol}(u) = u^3$ 3. $\phi_{RTB}(u) = u\mathbb{I}\{u \geq 0.9\}$

Figure 13: Empirical ROC curves and average ROC curve for Scale3 ($\varepsilon = 1.10$). Samples are drawn from multivariate Gaussian distributions according to section 5.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions. Hyperparameters: $u_0 = 0.9$, $q = 3$, $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$. Figures $1, 2, 3$ correspond *resp.* to the models MMW, Pol, RTB. Light blue curves are the $B(= 50)$ ROC curves that are averaged in green (solid line) with $+/-$ its standard deviation (dashed green lines). The dark blue and purple curves correspond to the best and worst scoring functions in the sense of minimization and maximization of the generalization error among the B curves. The red curve corresponds to ROC*.

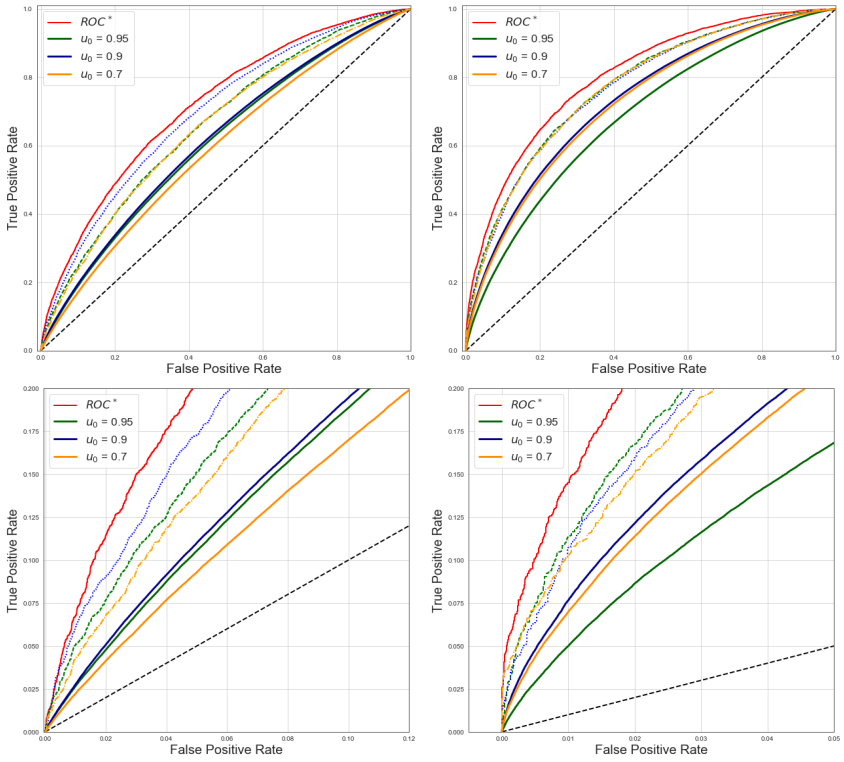
[9] S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, 2009.

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

[11] D. Cossock and T. Zhang. Subset ranking using regression. *Proceedings of COLT 2006*, 4005:605–619, 2006.

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

[13] E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6):907 – 921, 2002.



1. Loc2, $\varepsilon = 0.2$ 2. Loc3, $\varepsilon = 0.3$

Figure 14: Comparison of three RTB models. Average of the ROC curves (solid line), *best* ROC curves (dashed line) for the two location models Loc2 and Loc3. In green for $u_0 = 0.95$, blue for $u_0 = 0.90$, orange for $u_0 = 0.70$, red for ROC*. Samples are drawn from multivariate Gaussian distributions according to section 5.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions and averaged after B loops. Hyperparameters: $B = 50$, $T = 50$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$.

[14] E. Giné, V. Koltchinskii, and J. Zinn. Weighted uniform consistency of kernel density estimators. *The Annals of Probability*, 32(3B):2570–2605, 2004.

[15] E. Giné and J. Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, 12(4):929–989, 2004.

[16] L. Györfi, M. Köhler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

[17] J. Hájek. Asymptotically most powerful rank-order tests. *The Annals of Mathematical Statistics*, 33(3):112–1147, 1962.

[18] J. Hájek. Asymptotic normality of simple linear rank statistics under alternatives. *The Annals of Mathematical Statistics*, 39:325–346, 1968.

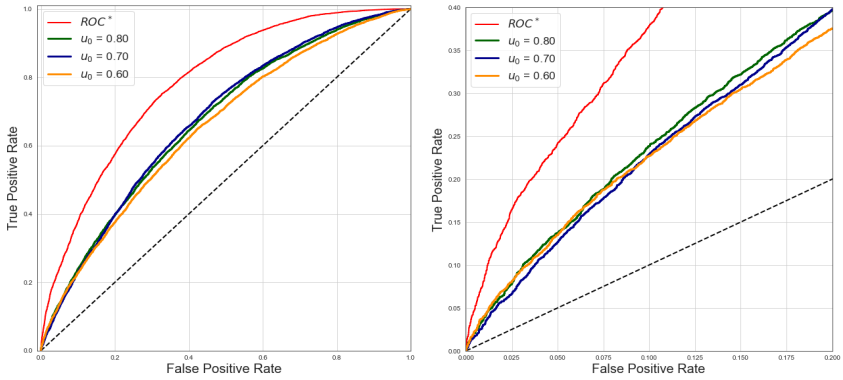


Figure 15: Comparison of three RTB models. *Best* ROC curves for the Scale2 model. In green for $u_0 = 0.80$, orange for $u_0 = 0.70$, green for $u_0 = 0.60$, red for ROC*. Samples are drawn from multivariate Gaussian distributions according to section 5.2, scored with early-stopped GA algorithm's optimal parameter for the class of scoring functions and averaged after B loops. Hyperparameters: $B = 50$, $T = 70$. Parameters for the training set: $n = m = 150$; $d = 15$; for the testing set: $n = m = 10^6$; $d = 15$.

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

[20] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19:293–325, 1948.

[21] M.C. Jones. The performance of kernel density functions in kernel distribution function estimation. *Statistics and Probability Letters*, 9(2):129–132, 1990.

[22] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593 – 2656, 2006.

[23] A. J. Lee. *U-statistics: Theory and practice*. Marcel Dekker, Inc., New York, 1990.

[24] P. Major. An estimate on the supremum of a nice class of stochastic integrals and U-statistics. *Probability Theory and Related Fields*, 134(3):489–537, 2006.

[25] C. McDiarmid. *On the method of bounded differences*, page 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.

[26] A.K. Menon and R.C. Williamsson. Bipartite ranking: A risk theoretic perspective. *Journal of Machine Learning Research*, 7:1–102, 2016.

[27] E.A. Nadaraya. Somenew estimates for distribution functions. *Theory of Probability and its Applications*, 9(3):497–500, 1964.

[28] N. Neumeyer. A central limit theorem for two-sample u-processes. *Statistics and Probability Letters*, 67(1):73 – 85, 2004.

[29] D. Nolan and D. Pollard. u-processes: Rates of convergence. *Annals of Statistics*, 15(2):780–799, 1987.

[30] C. Rudin. Ranking with a P-Norm Push. *Proceedings of COLT 2006*, 4005:589–604, 2006.

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

[32] M. Talagrand. Sharper Bounds for Gaussian and Empirical Processes. *The Annals of Probability*, 22(1):28 – 76, 1994.

[33] E. Giné V. De la Pena. *Decoupling: from dependence to independence*. Springer Science and Business Media, 1999.

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

[35] A. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[36] A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag New York, 1996.

[37] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.