The Annals of Statistics
2008, Vol. 36, No. 2, 844–874
DOI: 10.1214/009052607000000910

RANKING AND EMPIRICAL MINIMIZATION OF U-STATISTICS

BY STÉPHAN CLÉMENÇON, GÁBOR LUGOSI[1] AND NICOLAS VAYATIS

Ecole Nationale Supérieure des Télécommunications, ICREA and Universitat Pompeu Fabra, and Ecole Normale Supérieure de Cachan and UniverSud

The problem of ranking/ordering instances, instead of simply classifying them, has recently gained much attention in machine learning. In this paper we formulate the *ranking problem* in a rigorous statistical framework. The goal is to learn a ranking rule for deciding, among two instances, which one is "better," with minimum ranking risk. Since the natural estimates of the risk are of the form of a U-statistic, results of the theory of U-processes are required for investigating the consistency of empirical risk minimizers. We establish, in particular, a tail inequality for degenerate U-processes, and apply it for showing that fast rates of convergence may be achieved under specific noise assumptions, just like in classification. Convex risk minimization methods are also studied.

1. Introduction. Motivated by various applications including problems related to document retrieval or credit-risk screening, the ranking problem has received increasing attention both in the statistical and machine learning literature (see, e.g., Agarwal et al. [2], Cao et al. [11], Cortes and Mohri [12], Cossock and Zhang [13], Freund, Iyer, Schapire and Singer [17], Rudin [35], Usunier et al. [44] and Vittaut and Gallinari [46]). In the ranking problem, one has to compare two different observations and decide which one is "better." For example, in an application of document retrieval, one is concerned with comparing documents by degree of relevance for a particular request, rather than simply classifying them as relevant or not. Similarly, credit establishments collect and manage large databases containing the socio-demographic and credit-history characteristics of their clients to build a ranking rule which aims at indicating reliability.

In this paper we define a statistical framework for studying such ranking problems. The ranking problem defined here is closely related to the one studied by Stute [40, 41]. Indeed, Stute's results imply that certain nonparametric estimates based on local U-statistics give universally consistent ranking rules. Our approach here is different. Instead of local averages, we consider empirical minimizers of U-statistics, more in the spirit of empirical risk minimization [45] popular in statistical learning theory, see, for example, Bartlett and Mendelson [6], Boucheron,

Received March 2006; revised April 2007.

[1]Supported in part by the Spanish Ministry of Science and Technology, Grant MTM 2006-05650 and by the PASCAL Network of Excellence under EC Grant no. 506778.

AMS 2000 subject classifications. 68Q32, 60E15, 60C05, 60G25.

Key words and phrases. Statistical learning, theory of classification, VC classes, fast rates, convex risk minimization, moment inequalities, U-processes.

Bousquet and Lugosi [8], Koltchinskii [26], Massart [32] for surveys and recent development. The important feature of the ranking problem is that natural estimates of the ranking risk involve U-statistics. Therefore, our methodology is based on the theory of U-processes, and the key tools involve maximal and concentration inequalities, symmetrization tricks and a contraction principle for U-processes. For an excellent account of the theory of U-statistics and U-processes we refer to the monograph of de la Peña and Giné [15].

We also provide a theoretical analysis of certain nonparametric ranking methods that are based on an empirical minimization of convex cost functionals over convex sets of scoring functions. The methods are inspired by boosting- and support vector machine-type algorithms for classification. The main results of the paper prove universal consistency of properly regularized versions of these methods, establish a novel tail inequality for degenerate U-processes and, based on the latter result, show that fast rates of convergence may be achieved for empirical risk minimizers under suitable noise conditions.

We point out that under certain conditions, finding a good ranking rule amounts to constructing a scoring function s. An important special case is the bipartite ranking problem [2, 17] in which the available instances in the data are labeled by binary labels (good and bad). In this case, the ranking criterion is closely related to the so-called AUC [area under on ROC (receiver operating characteristic) curve] criterion (see Appendix B for more details).

The rest of the paper is organized as follows. In Section 2, the basic model and the two special cases of the ranking problem we consider are introduced. Section 3 provides some basic uniform convergence and consistency results for empirical risk minimizers. Section 4 contains the main statistical results of the paper, establishing performance bounds for empirical risk minimization for ranking problems. In Section 5, we describe the noise assumptions which guarantee fast rates of convergence in particular cases. In Section 6, a new exponential concentration inequality is established for U-processes which serves as a main tool in our analysis. In Section 7, we discuss convex risk minimization for ranking problems, laying down a theoretical framework for studying boosting and support vector machine-type ranking methods. In the Appendix A, we summarize some basic properties of U-statistics and highlight some connections of the ranking problem defined here to properties of the so-called ROC curve, appearing in related problems.

2. The ranking problem. Let (X, Y) be a pair of random variables taking values in $\mathcal{X} \times \mathbb{R}$ where $\mathcal{X}$ is a measurable space. The random object X models some observation and Y its real-valued label. Let (X', Y') denote a pair of random variables identically distributed with (X, Y), and independent of it. Denote

$$Z = \frac{Y - Y'}{2}.$$

In the ranking problem one observes X and X' but not their labels Y and Y'. We think about X being "better" than X' if $Y > Y'$, that is, if $Z > 0$. (The factor $1/2$

in the definition of Z is not significant, it is merely here as a convenient normalization.) The goal is to rank X and X' so that the probability that the better ranked of them has a smaller label is as small as possible. Formally, a *ranking rule* is a function $r : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$. If $r(x, x') = 1$ then the rule ranks x higher than x'. The performance of a ranking rule is measured by the *ranking risk*

$$L(r) = \mathbb{P}\{Z \cdot r(X, X') < 0\},$$

that is, the probability that r ranks two randomly drawn instances incorrectly. Observe that in this formalization, the ranking problem is equivalent to a binary classification problem in which the sign of the random variable Z is to be guessed based upon the pair of observations (X, X'). Now it is easy to determine the ranking rule with minimal risk. Introduce the notation

$$\rho_+(X, X') = \mathbb{P}\{Z > 0 \mid X, X'\},$$
$$\rho_-(X, X') = \mathbb{P}\{Z < 0 \mid X, X'\}.$$

Then we have the following simple fact:

PROPOSITION 1. *Define*

$$r^*(x, x') = 2\mathbb{I}_{[\rho_+(x,x') \geq \rho_-(x,x')]} - 1$$

and denote $L^* = L(r^*) = \mathbb{E}\{\min(\rho_+(X, X'), \rho_-(X, X'))\}$. *Then for any ranking rule* r,

$$L^* \leq L(r).$$

PROOF. Let r be any ranking rule. Observe that, by conditioning first on (X, X'), one may write

$$L(r) = \mathbb{E}\big(\mathbb{I}_{[r(X,X')=1]}\rho_-(X, X') + \mathbb{I}_{[r(X,X')=-1]}\rho_+(X, X')\big).$$

It is now easy to check that $L(r)$ is minimal for $r = r^*$. □

Thus, r^* minimizes the ranking risk over all possible ranking rules. In the definition of r^* ties are broken in favor of ρ_+ but obviously if $\rho_+(x, x') = \rho_-(x, x')$, an arbitrary value can be chosen for r^* without altering its risk.

The purpose of this paper is to investigate the construction of ranking rules of low risk based on training data. We assume that n independent, identically distributed copies of (X, Y), are available: $D_n = (X_1, Y_1), \ldots, (X_n, Y_n)$. Given a ranking rule r, one may use the training data to estimate its risk $L(r) = \mathbb{P}\{Z \cdot r(X, X') < 0\}$. The perhaps most natural estimate is the *U-statistic*

$$L_n(r) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}_{[Z_{i,j} \cdot r(X_i, X_j) < 0]},$$

where $Z_{i,j} = (Y_i - Y_j)/2$. In this paper we consider minimizers of the empirical estimate $L_n(r)$ over a class $\mathcal{R}$ of ranking rules and study the performance of such empirically selected ranking rules. Before discussing empirical risk minimization for ranking, a few remarks are in order.

REMARK 1. Note that the actual values of the Y_i's are never used in the ranking rules discussed in this paper. It is sufficient to know the values of the $Z_{i,j}$, or, equivalently, the ordering of the Y_i's.

REMARK 2 (*Ranking and scoring*). In many interesting cases the ranking problem may be reduced to finding an appropriate *scoring function*. These are the cases when the joint distribution of X and Y is such that there exists a function $s^* : \mathcal{X} \to \mathbb{R}$ such that

$$r^*(x, x') = 1 \quad \text{if and only if} \quad s^*(x) \geq s^*(x').$$

A function s^* satisfying the assumption is called an *optimal scoring function*. Obviously, any strictly increasing transformation of an optimal scoring function is also an optimal scoring function. Below we describe some important special cases when the ranking problem may be reduced to scoring.

REMARK 3 (*Ranking more than two items*). Throughout this paper we consider the problem of ranking just two observations X, X'. However, one may be interested in the more general problem of ranking m independent observations $X^{(1)}, \ldots, X^{(m)}$. The problem of ranking pairs is considerably simpler and has many practical applications (see, e.g., [12, 17, 46, 47], and the connection to the AUC detailed in the Appendix B) but ranking more than two items has also been considered in the literature (see Stute [40, 41], Cossock and Zhang [13]). In the general problem the value of a ranking function $r(X^{(1)}, \ldots, X^{(m)})$ is a permutation π of $\{1, \ldots, m\}$ and the goal is that π should coincide with (or at least resemble to) the permutation $\overline{\pi}$ for which $Y^{(\overline{\pi}(1))} \geq \cdots \geq Y^{(\overline{\pi}(m))}$. Given a loss function ℓ that assigns a number in $[0, 1]$ to a pair of permutations, the ranking risk is defined as

$$L(r) = \mathbb{E}\ell\big(r\big(X^{(1)}, \ldots, X^{(m)}\big), \overline{\pi}\big).$$

In this general case, natural estimates of $L(r)$ involve mth order U-statistics. Some results of this paper (such as those of Sections 3 and 7) extend in a rather straightforward manner but some others require significant additional work. The moment inequality of Theorem 11 should be possible to generalize by induction as all ingredients of the proof are available. In fact, the inequalities of Adamczak [1] and Major [31] are stated for general U-statistics of m variables. The key question is how the results of Section 4 can be generalized. In order to see this, one needs to understand what the analog of Assumption 4 means and under what conditions

such an assumption is satisfied. This depends in an essential way of how the quality of ranking is measured. This is an interesting and important problem for future research.

We now introduce the two main examples of statistical models which will serve to illustrate some of our results in Section 5.

EXAMPLE 1 (*The bipartite ranking problem*). In the bipartite ranking problem the label Y is binary, it takes values in $\{-1, 1\}$. Writing $\eta(x) = \mathbb{P}\{Y = 1|X = x\}$, it is easy to see that the Bayes ranking risk equals

$$\begin{aligned} L^* &= \mathbb{E}\min\bigl\{\eta(X)\bigl(1-\eta(X')\bigr), \eta(X')\bigl(1-\eta(X)\bigr)\bigr\} \\ &= \mathbb{E}\min\{\eta(X), \eta(X')\} - (\mathbb{E}\eta(X))^2 \end{aligned}$$

and also,

$$L^* = \mathrm{Var}\biggl(\frac{Y+1}{2}\biggr) - \frac{1}{2}\mathbb{E}|\eta(X) - \eta(X')|.$$

In particular,

$$L^* \le \mathrm{Var}\biggl(\frac{Y+1}{2}\biggr) \le 1/4,$$

where the equality $L^* = \mathrm{Var}(\frac{Y+1}{2})$ holds if and only if X and Y are independent and the maximum is attained when $\eta \equiv 1/2$. Observe that the difficulty of the bipartite ranking problem depends on the concentration properties of the distribution of $\eta(X) = \mathbb{P}(Y = 1|X)$ through the quantity $\mathbb{E}(|\eta(X) - \eta(X')|)$ which is a classical measure of concentration, known as *Gini's mean difference*. For given $p = \mathbb{E}(\eta(X))$, Gini's mean difference ranges from a minimum value of zero, when $\eta(X) \equiv p$, to a maximum value of $\frac{1}{2}p(1-p)$ in the case when $\eta(X) = (Y+1)/2$. It is clear from the form of the Bayes ranking rule that the optimal ranking rule is given by a scoring function s^* where s^* is any strictly increasing transformation of η. Then one may restrict the search to ranking rules defined by scoring functions s, that is, ranking rules of form $r(x, x') = 2\mathbb{I}_{[s(x)\ge s(x')]} - 1$. Writing $L(s) \stackrel{\mathrm{def}}{=} L(r)$, one has

$$L(s) - L^* = \mathbb{E}\bigl(|\eta(X') - \eta(X)|\mathbb{I}_{[(s(X)-s(X'))(\eta(X)-\eta(X'))<0]}\bigr).$$

We point out that the ranking risk in this case is closely related to the AUC criterion which is a standard performance measure in the bipartite setting (see [17] and Appendix B). More precisely, if $\mathbb{P}\{s(X) = s(X')\} = 0$, then we have

$$\mathrm{AUC}(s) = \mathbb{P}\{s(X) \ge s(X')|Y = 1, Y' = -1\} = 1 - \frac{1}{2p(1-p)}L(s),$$

where $p = \mathbb{P}(Y = 1)$, so that maximizing the AUC criterion boils down to minimizing the ranking error.

EXAMPLE 2 (*A regression model*). Assume now that Y is real-valued and the joint distribution of X and Y is such that $Y = m(X) + \epsilon$, where $m(x) = \mathbb{E}(Y|X = x)$ is the regression function, ϵ is independent of X and has a symmetric distribution around zero. Then clearly the optimal ranking rule r^* may be obtained by a scoring function s^* where s^* may be taken as any strictly increasing transformation of m.

3. Empirical risk minimization. Based on the empirical estimate $L_n(r)$ of the risk $L(r)$ of a ranking rule defined above, one may consider choosing a ranking rule by minimizing the empirical risk over a class $\mathcal{R}$ of ranking rules $r : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$. Define the empirical risk minimizer, over $\mathcal{R}$, by

$$r_n = \arg\min_{r \in \mathcal{R}} L_n(r).$$

(Ties are broken in an arbitrary way.) In a "first-order" approach, we may study the performance $L(r_n) = \mathbb{P}\{Z \cdot r_n(X, X') < 0 | D_n\}$ of the empirical risk minimizer by the standard bound (see, e.g., [16])

$$L(r_n) - \inf_{r \in \mathcal{R}} L(r) \le 2 \sup_{r \in \mathcal{R}} |L_n(r) - L(r)|. \tag{3.1}$$

This inequality points out that bounding the performance of an empirical minimizer of the ranking risk boils down to investigating the properties of *U-processes*, that is, suprema of U-statistics indexed by a class of ranking rules. For a detailed and modern account of U-process theory, we refer to the book of de la Peña and Giné [15]. In a first-order approach we basically reduce the problem to the study of ordinary empirical processes.

By using the simple Lemma A.1 given in Appendix A, we obtain the following:

PROPOSITION 2. *Define the Rademacher average*

$$R_n = \sup_{r \in \mathcal{R}} \frac{1}{\lfloor n/2 \rfloor} \left| \sum_{i=1}^{\lfloor n/2 \rfloor} \epsilon_i \mathbb{I}_{[Z_{i,\lfloor n/2 \rfloor + i} r(X_i, X_{\lfloor n/2 \rfloor + i}) < 0]} \right|,$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables (*i.e.*, *random symmetric sign variables*), *independent of D_n. Then for any convex nondecreasing function ψ,*

$$\mathbb{E}\psi\left(L(r_n) - \inf_{r \in \mathcal{R}} L(r) \right) \le \mathbb{E}\psi(4R_n).$$

PROOF. The inequality follows immediately from (3.1), Lemma A.1, and a standard symmetrization inequality; see, for example, Giné and Zinn [19]. □

One may easily use this result to derive probabilistic performance bounds for the empirical risk minimizer. For example, by taking $\psi(x) = e^{\lambda x}$ for some $\lambda > 0$, and using the bounded differences inequality (see McDiarmid [34]), we have

$$\mathbb{E}\exp\left(\lambda\left(L(r_n) - \inf_{r\in\mathcal{R}} L(r)\right)\right) \le \mathbb{E}\exp(4\lambda R_n) \le \exp\left(4\lambda\mathbb{E}R_n + \frac{4\lambda^2}{n-1}\right)$$

where we used the fact that R_n may be considered as a function of $\lfloor n/2\rfloor$ independent random vectors $(\epsilon_i, Z_{i,\lfloor n/2\rfloor+i}, X_i, X_{\lfloor n/2\rfloor+i})$ and changing any of them can change the value of R_n by at most $n-1$. By using Markov's inequality and choosing λ to minimize the bound, we readily obtain:

COROLLARY 3. *Let $\delta > 0$. With probability at least $1-\delta$,*

$$L(r_n) - \inf_{r\in\mathcal{R}} L(r) \le 4\mathbb{E}R_n + 4\sqrt{\frac{\ln(1/\delta)}{n-1}}\,.$$

The expected value of the Rademacher average R_n may now be bounded by standard metric entropy methods, see, for example, Lugosi [29], Boucheron, Bousquet and Lugosi [8]. For example, if the class $\mathcal{R}$ of indicator functions has finite VC dimension V, then

$$\mathbb{E}R_n \le c\sqrt{\frac{V}{n}}$$

for a universal constant c.

This result is similar to the one proved in the bipartite ranking case by Agarwal, Graepel, Herbrich, Har-Peled and Roth [2] with the restriction that their bound holds conditionally on a label sequence. The analysis of [2] relies on a particular complexity measure called rank-shatter coefficient but the core of the argument is the same.

The proposition above is convenient, simple, and, in a certain sense, not improvable. However, it is well known from the theory of statistical learning and empirical risk minimization for classification that the bound (3.1) is often quite loose. In classification problems the looseness of such a "first-order" approach is due to the fact that the variance of the estimators of the risk is ignored and bounded uniformly by a constant. Therefore, the main interest in considering U-statistics precisely consists in the fact that they have minimal variance among all unbiased estimators. However, the reduced-variance property of U-statistics plays no role in the above analysis of the ranking problem. Observe that all upper bounds obtained in this section remain true for an empirical risk minimizer that, instead of using

estimates based on U-statistics, estimates the risk of a ranking rule by splitting the data set into two halves and estimates $L(r)$ by

$$\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{I}_{[Z_{i,\lfloor n/2 \rfloor + i} \cdot r(X_i, X_{\lfloor n/2 \rfloor + i}) < 0]}.$$

Hence, in the argument of this section one loses the advantage of using U-statistics. In Section 4 it is shown that under certain, not uncommon, circumstances significantly smaller risk bounds are achievable. There it will have an essential importance to use sharp exponential bounds for U-processes, involving their reduced variance.

4. Fast rates. The main results of this paper show that the bounds obtained in the previous section may be significantly improved under certain conditions. It is well known (see, e.g., Section 5.2 in the survey [8] and the references therein) that tighter bounds for the excess risk in the context of binary classification may be obtained if one can control the variance of the excess risk by its expected value. In classification this can be guaranteed under certain "low-noise" conditions (see Tsybakov [43], Massart and Nédélec [33], Koltchinskii [26]).

Next we examine the possibilities of obtaining such improved performance bounds for empirical ranking risk minimization. The main message is that in the ranking problem one may also obtain significantly improved bounds under some conditions that are analogous to the low-noise conditions in the classification problem, though quite different in nature.

Here we will greatly benefit from using U-statistics (as opposed to splitting the sample) as the small variance of the U-statistics used to estimate the ranking risk gives rise to sharper bounds. The starting point of our analysis is the Hoeffding decomposition of U-statistics (see Appendix A).

Set first

$$q_r((x, y), (x', y')) = \mathbb{I}_{[(y-y') \cdot r(x,x') < 0]} - \mathbb{I}_{[(y-y') \cdot r^*(x,x') < 0]}$$

and consider the following estimate of the *excess risk* $\Lambda(r) = L(r) - L^* = \mathbb{E} q_r((X, Y), (X', Y'))$:

$$\Lambda_n(r) = \frac{1}{n(n-1)} \sum_{i \neq j} q_r((X_i, Y_i), (X_j, Y_j)),$$

which is a U-statistic of degree 2 with kernel q_r. If the ranking rules r and r^* are symmetric in the sense that $r(x, x') = -r(x', x)$ for all $x, x' \in \mathcal{X}$, then the kernel q_r is symmetric. This can always be achieved if we define $r(x, x) = 0$ for all x. In the analysis it is convenient to work with symmetric kernels, so we assume that all ranking rules are symmetric in the sequel.

Clearly, the minimizer r_n of the empirical ranking risk $L_n(r)$ over $\mathcal{R}$ also minimizes the empirical excess risk $\Lambda_n(r)$. To study this minimizer, consider the Hoeffding decomposition of $\Lambda_n(r)$:

$$\Lambda_n(r) - \Lambda(r) = 2T_n(r) + W_n(r),$$

where

$$T_n(r) = \frac{1}{n}\sum_{i=1}^{n} h_r(X_i, Y_i)$$

is a sum of i.i.d. random variables with

$$h_r(x, y) = \mathbb{E}q_r((x, y), (X', Y')) - \Lambda(r)$$

and

$$W_n(r) = \frac{1}{n(n-1)}\sum_{i \neq j} \widehat{h}_r((X_i, Y_i), (X_j, Y_j))$$

is a *degenerate* U-statistic with symmetric kernel

$$\widehat{h}_r((x, y), (x', y')) = q_r((x, y), (x', y')) - \Lambda(r) - h_r(x, y) - h_r(x', y').$$

In the analysis we show that the contribution of the degenerate part $W_n(r)$ of the U-statistic is negligible compared to that of $T_n(r)$. This means that minimization of Λ_n is approximately equivalent to minimizing $T_n(r)$. But since $T_n(r)$ is an average of i.i.d. random variables, this can be studied by known techniques worked out for empirical risk minimization.

The main tool for handling the degenerate part is a new general moment inequality for U-processes that may be interesting on its own right. This inequality is presented in Section 6. We mention here that for VC classes one may use an inequality of Arcones and Giné [4] and its significant improvement due to Major [31].

It is well known from the theory of empirical risk minimization (see Tsybakov [43], Bartlett and Mendelson [6], Koltchinskii [26], Massart [32]), that in order to improve the rates of convergence [such as the bound $O(\sqrt{V/n})$ obtained for VC classes in Section 3], it is necessary to impose some conditions on the joint distribution of (X, Y). In our case, the key assumption takes the following form:

ASSUMPTION 4. *There exist constants $c > 0$ and $\alpha \in [0, 1]$ such that for all $r \in \mathcal{R}$,*

$$\operatorname{Var}(h_r(X, Y)) \leq c\Lambda(r)^{\alpha}.$$

The improved rates of convergence will depend on the value of α. We will see in some examples that this assumption is satisfied for a surprisingly large family of

distributions, guaranteeing improved rates of convergence. For $\alpha = 0$ the assumption is always satisfied and the corresponding performance bound does not yield any improvement over those of Section 3. However, we will see that in many natural examples Assumption 4 is satisfied with values of α close to one, providing significant improvements in the rates of convergence.

Now we are prepared to state and prove the main result of the paper. In order to state the result, we need to introduce some quantities related to the class $\mathcal{R}$. Let $\epsilon_1, \dots, \epsilon_n$ be i.i.d. Rademacher random variables independent of the (X_i, Y_i). Let

$$Z_\epsilon = \sup_{r\in\mathcal{R}} \left| \sum_{i,j} \epsilon_i \epsilon_j \widehat{h}_r((X_i, Y_i), (X_j, Y_j)) \right|,$$

$$U_\epsilon = \sup_{r\in\mathcal{R}} \sup_{\alpha\,:\,\|\alpha\|_2\le 1} \sum_{i,j} \epsilon_i \alpha_j \widehat{h}_r((X_i, Y_i), (X_j, Y_j)),$$

$$M = \sup_{r\in\mathcal{R}, k=1,\dots,n} \left| \sum_{i=1}^{n} \epsilon_i \widehat{h}_r((X_i, Y_i), (X_k, Y_k)) \right|.$$

Introduce the "loss function"

$$\ell(r, (x, y)) = 2\mathbb{E}\, \mathbb{I}_{[(y-Y)\cdot r(x,X)<0]} - L(r)$$

and define

$$\nu_n(r) = \frac{1}{n} \sum_{i=1}^{n} \ell(r, (X_i, Y_i)) - L(r).$$

[Observe that $\nu_n(r)$ has zero mean.] Also, define the pseudo-distance

$$d(r, r') = \left(\mathbb{E}\left(\mathbb{E}\left[\mathbb{I}_{[r(X,X')\neq r'(X,X')]} | X\right]\right)^2\right)^{1/2}.$$

Let $\phi : [0, \infty) \to [0, \infty)$ be a nondecreasing function such that $\phi(x)/x$ is nonincreasing and $\phi(1) \ge 1$ such that for all $r \in \mathcal{R}$,

$$\sqrt{n}\mathbb{E} \sup_{r'\in\mathbb{R}, d(r,r')\le\sigma} |\nu_n(r) - \nu_n(r')| \le \phi(\sigma).$$

THEOREM 5. *Consider a minimizer r_n of the empirical ranking risk $L_n(r)$ over a class $\mathcal{R}$ of ranking rules and assume Assumption 4. Then there exists a universal constant C such that, with probability at least $1 - \delta$, the ranking risk of r_n satisfies*

$$\begin{aligned} L(r_n) - L^* \le 2\left(\inf_{r\in\mathcal{R}} L(r) - L^*\right) \\ + C\left(\frac{\mathbb{E}Z_\epsilon}{n^2} + \frac{\mathbb{E}U_\epsilon\sqrt{\log(1/\delta)}}{n^2} \right. \\ \left. + \frac{\mathbb{E}M\log(1/\delta)}{n^2} + \frac{\log(1/\delta)}{n} + \rho^2\log(1/\delta)\right) \end{aligned}$$

where $\rho > 0$ is the unique solution of the equation

$$\sqrt{n}\rho^2 = \phi(\rho^\alpha).$$

The theorem provides a performance bound in terms of expected values of certain Rademacher chaoses indexed by $\mathcal{R}$ and local properties of an ordinary empirical process. These quantities have been thoroughly studied and well understood, and may be easily bounded in many interesting cases. Below we will work out an example when $\mathcal{R}$ is a VC class of indicator functions.

PROOF OF THEOREM 5. We consider the Hoeffding decomposition of the U-statistic $\Lambda_n(r)$ that is minimized over $r \in \mathcal{R}$. The idea of the proof is to show that the degenerate part $W_n(r)$ is of a smaller order and becomes negligible compared to the part $T_n(r)$. Therefore, r_n is an approximate minimizer of $T_n(r)$ which can be handled by recent results on empirical risk minimization when the empirical risk is defined as a simple sample average.

Let A be the event on which

$$\sup_{r\in\mathcal{R}} |W_n(r)| \le \kappa,$$

where

$$\kappa = C\left(\frac{\mathbb{E}Z_\epsilon}{n^2} + \frac{\mathbb{E}U_\epsilon\sqrt{\log(1/\delta)}}{n^2} + \frac{\mathbb{E}M\log(1/\delta)}{n^2} + \frac{\log(1/\delta)}{n}\right)$$

for an appropriate constant C. Then by Theorem 11, $\mathbb{P}[A] \ge 1 - \delta/2$. By the Hoeffding decomposition of the U-statistics $\Lambda_n(r)$ it is clear that, on A, r_n is a ρ-minimizer of

$$\frac{2}{n}\sum_{i=1}^n \ell(r, (X_i, Y_i))$$

over $r \in \mathcal{R}$ in the sense that the value of this latter quantity at its minimum is at most κ smaller than at r_n.

Define $\tilde{r}_n$ as r_n on A and an arbitrary minimizer of $(2/n)\sum_{i=1}^n \ell(r, (X_i, Y_i))$ on A^c. Then clearly, with probability at least $1 - \delta/2$, $L(r_n) = L(\tilde{r}_n)$ and $\tilde{r}_n$ is a κ-minimizer of $(2/n)\sum_{i=1}^n \ell(r, (X_i, Y_i))$. But then we may use Theorem 8.3 of Massart [32] to bound the performance of $\tilde{r}_n$ which implies the theorem. □

Observe that the only condition for the distribution is that the variance of h_r can be bounded in terms of $\Lambda(r)$. In Section 5 we present examples in which Assumption 4 is satisfied with $\alpha > 0$. We will see below that the value of α in this assumption determines the magnitude of the last term which, in turn, dominates the right-hand side (apart from the approximation error term).

The factor of 2 in front of the approximation error term $\inf_{r\in\mathcal{R}} L(r) - L^*$ has no special meaning. It can be replaced by any constant strictly greater than one

at the price of increasing the value of the constant C. Notice that in the bound for $L(r_n) - L^*$ derived from Corollary 3, the approximation error appears with a factor of 1. Thus, the improvement of Theorem 5 is only meaningful if $\inf_{r\in\mathcal{R}} L(r) - L^*$ does not dominate the other terms in the bound. Ideally, the class $\mathcal{R}$ should be chosen such that the approximation error and the other terms in the bound are balanced. If this was the case, the theorem would guarantee faster rates of convergence. Based on the bounds presented here, one may design penalized empirical minimizers of the ranking risk that select the class $\mathcal{R}$ from a collection of classes achieving this objective. We do not give the details here, we just mention that the techniques presented in Massart [32] and Koltchinskii [26] may be used in a relatively straightforward manner to derive such "oracle inequalities" for penalized empirical risk minimization in the present framework.

In order to illustrate Theorem 5, we consider the case when $\mathcal{R}$ is a VC class, that is, it has a finite VC dimension V.

COROLLARY 6. *Consider the minimizer r_n of the empirical ranking risk $L_n(r)$ over a class $\mathcal{R}$ of ranking rules of finite VC dimension V and assume Assumption 4. Then there exists a universal constant C such that, with probability at least $1-\delta$, the ranking risk of r_n satisfies*

$$L(r_n) - L^* \le 2\left(\inf_{r\in\mathcal{R}} L(r) - L^*\right) + C\left(\frac{V\log(n/\delta)}{n}\right)^{1/(2-\alpha)}.$$

PROOF. In order to apply Theorem 5, we need suitable upper bounds for $\mathbb{E}Z_\epsilon$, $\mathbb{E}U_\epsilon$, $\mathbb{E}M$ and ρ. To bound $\mathbb{E}Z_\epsilon$, observe that Z_ϵ is a Rademacher chaos indexed by $\mathcal{R}$ for which Propositions 2.2 and 2.6 of Arcones and Giné [3] may be applied. In particular, by using Haussler's [21] metric entropy bound for VC classes, it is easy to see that there exists a constant C such that

$$\mathbb{E}Z_\epsilon \le CnV.$$

Similarly, $\mathbb{E}_\epsilon M$ is just an expected Rademacher average that may be bounded by $C\sqrt{Vn}$ (see, e.g., [8]).

Also, by the Cauchy–Schwarz inequality,

$$\begin{aligned}\mathbb{E}U_\epsilon^2 &\le \mathbb{E}\sup_{r\in\mathcal{R}}\sqrt{\sum_j\left(\sum_i \epsilon_i \widehat{h}_r((X_i,Y_i),(X_j,Y_j))\right)^2}\\ &= \mathbb{E}\sup_{r\in\mathcal{R}}\left\{\sum_j\sum_i \widehat{h}_r((X_i,Y_i),(X_j,Y_j))^2\right.\\ &\qquad\left.+\sum_j\sum_{i,k}\epsilon_i\epsilon_k\widehat{h}_r((X_i,Y_i),(X_j,Y_j))\widehat{h}_r((X_j,Y_j),(X_k,Y_k))\right\}\end{aligned}$$

$$\leq n^2 + \mathbb{E}\sup_{r\in\mathcal{R}} \sum_j \sum_{i,k} \epsilon_i \epsilon_k \widehat{h}_r((X_i, Y_i), (X_j, Y_j)) \widehat{h}_r((X_j, Y_j), (X_k, Y_k)).$$

Observe that the second term on the right-hand side is a Rademacher chaos of order 2 that can be handled similarly to $\mathbb{E}Z_\epsilon$. Indeed, defining

$$h'_r((X_i, Y_i), (X_k, Y_k)) = \frac{1}{n} \sum_j \widehat{h}_r((X_i, Y_i), (X_j, Y_j)) \widehat{h}_r((X_j, Y_j), (X_k, Y_k)),$$

the second term has the same form as Z_ϵ so repeating the same argument, one obtains

$$\mathbb{E}U_\epsilon^2 \leq n^2 + CVn^2.$$

Thus,

$$\mathbb{E}(U_\epsilon) \leq \sqrt{\mathbb{E}(U_\epsilon^2)} \leq CnV^{1/2}.$$

This shows that the value of κ defined in the proof of Theorem 5 is of the order of $n^{-1}(V + \log(1/\delta))$. The main term in the bound of Theorem 5 is ρ^2. By mimicking the argument of Massart [32], pages 297–298, we get

$$C\left(\frac{V \log n}{n}\right)^{1/(2-\alpha)}$$

as desired. □

5. Examples.

5.1. *The bipartite ranking problem.* Next we derive a simple sufficient condition for achieving fast rates of convergence for the bipartite ranking problem. Recall that here it suffices to consider ranking rules of the form $r(x, x') = 2\mathbb{I}_{[s(x)\geq s(x')]} - 1$ where s is a scoring function. With some abuse of notation we write h_s for h_r.

NOISE ASSUMPTION. There exist constants $c > 0$ and $\alpha \in [0, 1]$ such that for all $x \in \mathcal{X}$,

$$\mathbb{E}_{X'}\big(|\eta(x) - \eta(X')|^{-\alpha}\big) \leq c. \tag{5.2}$$

PROPOSITION 7. *Under* (5.2), *we have, for all* $s \in \mathcal{F}$

$$\mathrm{Var}(h_s(X, Y)) \leq c\Lambda(s)^\alpha.$$

PROOF.

$$\begin{aligned}
&\operatorname{Var}(h_s(X,Y)) \\
&\quad\le \mathbb{E}_X\big[\big(\mathbb{E}_{X'}\big(\mathbb{I}_{[(s(X)-s(X'))(\eta(X)-\eta(X'))<0]}\big)\big)^2\big] \\
&\quad\le \mathbb{E}_X\big[\mathbb{E}_{X'}\big(\mathbb{I}_{[(s(X)-s(X'))(\eta(X)-\eta(X'))<0]}|\eta(X)-\eta(X')|^{\alpha}\big) \\
&\qquad\qquad \times \big(\mathbb{E}_{X'}(|\eta(X)-\eta(X')|^{-\alpha})\big)\big] \\
&\qquad\qquad \text{(by the Cauchy–Schwarz inequality)} \\
&\quad\le c\big(\mathbb{E}_X\mathbb{E}_{X'}\big(\mathbb{I}_{[(s(X)-s(X'))(\eta(X)-\eta(X'))<0]}|\eta(X)-\eta(X')|\big)\big)^{\alpha} \\
&\qquad\qquad \text{(by Jensen's inequality and the noise assumption)} \\
&\quad= c\Lambda(s)^{\alpha}. \qquad \square
\end{aligned}$$

Condition (5.2) is satisfied under quite general circumstances. If $\alpha = 0$ then clearly the condition poses no restriction, but also no improvement is achieved in the rates of convergence. On the other hand, at the other extreme, when $\alpha = 1$, the condition is quite restrictive as it excludes η to be differentiable, for example, if X has a uniform distribution over $[0,1]$. However, interestingly, for any $\alpha < 1$, it poses quite mild restrictions as it is highlighted in the following example:

COROLLARY 8. *Consider the bipartite ranking problem and assume that* $\eta(x) = \mathbb{P}\{Y = 1|X = x\}$ *is such that the random variable* $\eta(X)$ *has an absolutely continuous distribution on* $[0,1]$ *with a density bounded by* B. *Then for any* $\epsilon > 0$,

$$\forall x \in \mathcal{X} \qquad \mathbb{E}_{X'}\big(|\eta(x)-\eta(X')|^{-1+\epsilon}\big) \le \frac{2B}{\epsilon}$$

and therefore, by Propositions 4 and 7, there is a constant C *such that for every* $\delta, \epsilon \in (0,1)$, *the excess ranking risk of the empirical minimizer* r_n *satisfies, with probability at least* $1-\delta$,

$$L(r_n) - L^* \le 2\left(\inf_{r\in\mathcal{R}} L(r) - L^*\right) + CB\epsilon^{-1}\left(\frac{V\log(n/\delta)}{n}\right)^{1/(1+\epsilon)}.$$

PROOF. The corollary follows simply by checking that (5.2) is satisfied for any $\alpha = 1-\epsilon < 1$. Denoting the density of $\eta(X)$ by f, we have

$$\begin{aligned}
\mathbb{E}_{X'}\big(|\eta(x)-\eta(X')|^{-\alpha}\big) &= \int_0^1 \frac{1}{|\eta(x)-u|^{\alpha}} f(u)\,du \\
&\le B\int_0^1 \frac{1}{|\eta(x)-u|^{\alpha}}\,du \\
&= B\frac{\eta(x)^{1-\alpha} + (1-\eta(x))^{1-\alpha}}{1-\alpha} \le \frac{2B}{1-\alpha}. \qquad \square
\end{aligned}$$

Condition (5.2) of the corollary requires that the distribution of $\eta(X)$ is sufficiently spread out, for example, it cannot have atoms or infinite peaks in its density. Under such a condition a rate of convergence of the order of $n^{-1+\epsilon}$ is achievable for any $\epsilon > 0$.

REMARK 4. Note that we crucially used the reduced variance of the U-statistic $L(r_n)$ to derive fast rates from the rather weak condition (5.2). Applying a similar reasoning for the variance of $q_s((X, Y), (X', Y'))$ (which would be the case if one considered a risk estimate based on independent pairs by splitting the training data into two halves, see Section 3), would have led to the condition

$$|\eta(x) - \eta(x')| \geq c, \tag{5.3}$$

for some constant c, and $x \neq x'$. This condition is satisfied only when $\eta(X)$ has a discrete distribution.

5.2. *Noiseless regression model.* Next we consider the *noise-free regression model* in which $Y = m(X)$ for some (unknown) function $m : \mathcal{X} \to \mathbb{R}$. Here obviously $L^* = 0$ and the Bayes ranking rule is given by the scoring function $s^* = m$ (or any strictly increasing transformation of it). Clearly, in this case

$$q_r(x, x') = \mathbb{I}_{[(m(x)-m(x'))\cdot r(x,x')<0]}$$

and therefore

$$\mathrm{Var}(h_r(X, Y)) \leq \mathbb{E}q_r^2(X, X') = L(r),$$

and therefore the condition of Proposition 4 is satisfied with $c = 1$ and $\alpha = 1$. Thus, the risk of the empirical risk minimizer r_n satisfies, with probability at least $1 - \delta$,

$$L(r_n) \leq 2 \inf_{r \in \mathcal{R}} L(r) + C\frac{V \log(n/\delta)}{n}$$

provided $\mathcal{R}$ has finite VC dimension V.

5.3. *Regression model with noise.* Now we turn to the *general regression model with heteroscedastic errors* in which $Y = m(X) + \sigma(X)\epsilon$ for some (unknown) functions $m : \mathcal{X} \to \mathbb{R}$ and $\sigma : \mathcal{X} \to \mathbb{R}$, where ϵ is a standard Gaussian random variable, independent of X.

We set

$$\Delta(X, X') = \frac{m(X) - m(X')}{\sqrt{\sigma^2(X) + \sigma^2(X')}}.$$

We have again $s^* = m$ (or any strictly increasing transformation of it) and the optimal risk is

$$L^* = \mathbb{E}\Phi(-|\Delta(X, X')|)$$

where Φ is the distribution function of a standard Gaussian random variable. The maximal value of L^* is attained when the regression function $m(x)$ is constant. Furthermore, we have

$$L(s) - L^* = \mathbb{E}\big(|2\Phi(\Delta(X, X')) - 1| \cdot \mathbb{I}_{[(m(x)-m(x'))\cdot(s(x)-s(x'))<0]}\big).$$

Noise assumption. There exist constants $c > 0$ and $\alpha \in [0, 1]$ such that for all $x \in \mathcal{X}$,

$$\mathbb{E}_{X'}(|\Delta(x, X')|^{-\alpha}) \le c. \tag{5.4}$$

PROPOSITION 9. *Under* (5.4), *we have*, *for all* $s \in \mathcal{F}$,

$$\operatorname{Var}(h_s(X, Y)) \le \big(2\Phi(c) - 1\big)\Lambda(s)^{\alpha}.$$

PROOF. By symmetry, we have

$$|2\Phi(\Delta(X, X')) - 1| = 2\Phi(|\Delta(X, X')|) - 1.$$

Then, using the concavity of the distribution function Φ on $\mathbb{R}_+$, we have, by Jensen's inequality,

$$\forall x \in \mathcal{X} \qquad \mathbb{E}_{X'}\Phi(|\Delta(x, X')|^{-\alpha}) \le \Phi(\mathbb{E}_{X'}|\Delta(x, X')|^{-\alpha}) \le \Phi(c),$$

where we have used (5.4) together with the fact that Φ is increasing. Now the result follows following the argument given in the proof of Proposition 7. □

The preceding noise condition is satisfied in many cases, as illustrated by the example below.

COROLLARY 10. *Suppose that* $m(X)$ *has a bounded density and the conditional variance* $\sigma(x)$ *is bounded over* $\mathcal{X}$. *Then the noise condition* (5.4) *is satisfied for any* $\alpha < 1$.

REMARK 5. The argument above still holds if we drop the Gaussian noise assumption. Indeed we only need the random variable ϵ to have a symmetric density decreasing over $\mathbb{R}_+$.

6. A moment inequality for U-processes. In this section we establish a general exponential inequality for U-processes. This result is based on moment inequalities obtained for empirical processes and Rademacher chaoses in Bousquet, Boucheron, Lugosi and Massart [9] and generalizes an inequality due to Arcones and Giné [4]. We also mention an essential improvement of the results of [4] due to Major [31] for VC and other "nice" classes. We also refer to the corresponding results obtained for U-statistics by Adamczak [1], Giné, Latala and Zinn [18] and Houdré and Reynaud-Bouret [24]. We

point out here that the recent work of Adamczak [1] establishes very general moment inequalities for Banach space-valued degenerate U-statistics of arbitrary order. Adamczak's inequality has a natural counterpart for suprema of U-processes. When specialized to the case of U-processes of order 2, Adamczak's Theorem 1 takes a form very similar to Theorem 11 below. However, Adamczak's result is given in terms of various operator norms corresponding to the kernel while the relevant quantities in the theorem below are defined in terms of expectations of certain associated Rademacher averages and chaoses. For our applications we find the latter quantities easier to handle.

THEOREM 11. *Let $X, X_1, \dots, X_n$ be i.i.d. random variables and let $\mathcal{F}$ be a class of kernels. Consider a degenerate U-process Z of order 2 indexed by $\mathcal{F}$,*

$$Z = \sup_{f \in \mathcal{F}} \left| \sum_{i,j} f(X_i, X_j) \right|$$

where $\mathbb{E} f(X, x) = 0$, $\forall x$, f. Assume also $f(x, x) = 0$, $\forall x$ and $\sup_{f \in \mathcal{F}} \|f\|_\infty = F$. Let $\epsilon_1, \dots, \epsilon_n$ be i.i.d. Rademacher random variables and introduce the random variables

$$Z_\epsilon = \sup_{f \in \mathcal{F}} \left| \sum_{i,j} \epsilon_i \epsilon_j f(X_i, X_j) \right|,$$

$$U_\epsilon = \sup_{f \in \mathcal{F}} \sup_{\alpha : \|\alpha\|_2 \le 1} \sum_{i,j} \epsilon_i \alpha_j f(X_i, X_j),$$

$$M = \sup_{f \in \mathcal{F}, k = 1, \dots, n} \left| \sum_{i=1}^n \epsilon_i f(X_i, X_k) \right|.$$

Then there exists a universal constant $C > 0$ such that for all n and $q \ge 2$,

$$(\mathbb{E} Z^q)^{1/q} \le C\big(\mathbb{E} Z_\epsilon + q^{1/2} \mathbb{E} U_\epsilon + q(\mathbb{E} M + Fn) + q^{3/2} F n^{1/2} + q^2 F\big).$$

Also, there exists a universal constant C such that for all n and $t > 0$,

$$\mathbb{P}\{Z > C\mathbb{E} Z_\epsilon + t\} \le \exp\left(-\frac{1}{C} \min\left(\left(\frac{t}{\mathbb{E} U_\epsilon}\right)^2, \frac{t}{\mathbb{E} M + Fn}, \left(\frac{t}{F\sqrt{n}}\right)^{2/3}, \sqrt{\frac{t}{F}}\right)\right).$$

REMARK 6. Generously overestimated values of the constants may be easily deduced from the proof. We are convinced that these are far from being the best possible but do not have a good guess of what the best constants might be.

PROOF OF THEOREM 11. The proof of Theorem 11 is based on symmetrization, decoupling, and concentration inequalities for empirical processes and Rademacher chaos.

Since the f are degenerate kernels, one may relate the moments of Z to those of Z_ϵ by the randomization inequality

$$\mathbb{E}Z^q \le 4^q \mathbb{E}Z_\epsilon^q,$$

valid for $q \ge 1$; see Chapter 3 of [15]. Thus, it suffices to derive moment inequalities for the symmetrized U-process Z_ϵ. We do this by conditioning. Denote by $\mathbb{E}_\epsilon$ the expectation taken with respect to the variables ϵ_i (i.e., conditional expectation given $X_1, \dots, X_n$). Then we write $\mathbb{E}Z_\epsilon^q = \mathbb{E}\mathbb{E}_\epsilon Z_\epsilon^q$ and study the quantity $\mathbb{E}_\epsilon Z_\epsilon^q$, with the X_i fixed. But then Z_ϵ is a so-called *Rademacher chaos* whose tail behavior has been studied, see Talagrand [42], Ledoux [28], Boucheron, Bousquet, Lugosi and Massart [9]. In particular, for any $q \ge 2$,

$$\begin{aligned}(\mathbb{E}_\epsilon Z_\epsilon^q)^{1/q} &\le \mathbb{E}_\epsilon Z_\epsilon + \bigl(\mathbb{E}_\epsilon \bigl(Z_\epsilon - \mathbb{E}_\epsilon Z_\epsilon\bigr)_+^q\bigr)^{1/q} \qquad (\text{since } Z \ge 0)\\ &\le \mathbb{E}_\epsilon Z_\epsilon + 3\sqrt{q}\,\mathbb{E}_\epsilon U_\epsilon + 4qB\end{aligned}$$

with U_ϵ defined above and

$$B = \sup_{f\in\mathcal{F}} \sup_{\alpha,\alpha' :\, \|\alpha\|_2, \|\alpha'\|_2 \le 1} \left| \sum_{i,j} \alpha_i \alpha'_j f(X_i, X_j) \right|$$

where the second inequality follows by Theorem 14 of [9]. Using the inequality $(a+b+c)^q \le 3^{q-1}(a^q + b^q + c^q)$ valid for $q \ge 2$, $a, b, c > 0$, we have

$$\mathbb{E}_\epsilon Z_\epsilon^q \le 3^{q-1}\bigl((\mathbb{E}_\epsilon Z_\epsilon)^q + 3^q q^{q/2} (\mathbb{E}_\epsilon U_\epsilon)^q + 4^q q^q B^q\bigr).$$

It remains to derive suitable upper bounds for the expectation of the three terms on the right-hand side.

First term: $\mathbb{E}(\mathbb{E}_\epsilon Z_\epsilon)^q$. In order to handle the moments of $\mathbb{E}_\epsilon Z_\epsilon$, first we note that by a decoupling inequality in de la Peña and Giné [15], page 101,

$$\mathbb{E}_\epsilon Z_\epsilon \le 8\mathbb{E}_\epsilon Z'_\epsilon,$$

where

$$Z'_\epsilon = \sup_{f\in\mathcal{F}} \left| \sum_{i,j} \epsilon_i \epsilon'_j f(X_i, X_j) \right|.$$

Here $\epsilon'_1, \dots, \epsilon'_n$ are i.i.d. Rademacher variables, independent of the X_i and the ϵ_i. Note that $\mathbb{E}_\epsilon$ now denotes expectation taken with respect to both the ϵ_i and the ϵ'_i.

Thus, we have

$$\mathbb{E}(\mathbb{E}_\epsilon Z_\epsilon)^q \le 8^q \mathbb{E}(\mathbb{E}_\epsilon Z'_\epsilon)^q.$$

In order to bound the moments of the random variable $A = \mathbb{E}_\epsilon Z'_\epsilon$, we apply Corollary 3 of [9]. In order to apply this corollary, define, for $k = 1, \dots, n$, the random variables

$$A_k = \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \left| \sum_{i,j\ne k} \epsilon_i \epsilon'_j f(X_i, X_j) \right|.$$

It is easy to see that $A_k \le A$.

On the other hand, defining

$$R_k = \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^{n} \epsilon_i f(X_i, X_k) \right|,$$

we clearly have

$$A - A_k \le 2\mathbb{E}_\epsilon R_k.$$

Also, denoting by f^* the (random) function achieving the maximum in the definition of Z, we have

$$\begin{aligned}\sum_{k=1}^{n}(A - A_k) &\le \mathbb{E}_\epsilon\left(\sum_{k=1}^{n}\epsilon_k \sum_{j=1}^{n}\epsilon'_j f^*(X_k, X'_j) + \sum_{k=1}^{n}\epsilon'_k \sum_{i=1}^{n}\epsilon_i f^*(X_i, X'_k)\right)\\ &= 2A.\end{aligned}$$

Therefore,

$$\sum_{k=1}^{n}(A - A_k)^2 \le 4A\mathbb{E}_\epsilon M,$$

where $M = \max_k R_k$. Then by Corollary 3 of [9], we obtain

$$\mathbb{E}(\mathbb{E}_\epsilon Z'_\epsilon)^q = \mathbb{E}A^q \le 2^{q-1}\bigl(2^q(\mathbb{E}Z'_\epsilon)^q + 5^q q^q \mathbb{E}(\mathbb{E}_\epsilon M)^q\bigr).$$

By un-decoupling (see de la Peña and Giné [15], page 101), we have $\mathbb{E}Z'_\epsilon \le 4\mathbb{E}Z_\epsilon$.

To bound $\mathbb{E}(\mathbb{E}_\epsilon M)^q$, observe that $\mathbb{E}_\epsilon M$ is a conditional Rademacher average, for which Theorem 13 of [9] may be applied. According to this,

$$\mathbb{E}(\mathbb{E}_\epsilon M)^q \le 2^{q-1}\bigl(2^q(\mathbb{E}M)^q + 5^q q^q F^q\bigr).$$

Collecting terms, we have

$$\mathbb{E}(\mathbb{E}_\epsilon Z_\epsilon)^q \le 128^q(\mathbb{E}Z_\epsilon)^q + 320^q q^q(\mathbb{E}M)^q + 800^q F^q q^{2q}.$$

Second term: $\mathbb{E}_X(\mathbb{E}_\epsilon U_\epsilon)^q$. The moments of $\mathbb{E}_\epsilon U_\epsilon$ can be estimated by the same inequality as the one we used for $\mathbb{E}_\epsilon M$ since $\mathbb{E}_\epsilon U_\epsilon$ is also a conditional Rademacher average. Observing that

$$\sup_{f,i}\sup_{\alpha:\|\alpha\|_2\le 1}\sum_{j\ne i}\alpha_j f(X_i, X_j) \le F\sqrt{n}$$

by the Cauchy–Schwarz inequality, we have, by Theorem 13 from [9],

$$\mathbb{E}(\mathbb{E}_\epsilon U_\epsilon)^q \le 2^{q-1}(2^q\bigl(\mathbb{E}U_\epsilon)^q + 5^q q^q F^q n^{q/2}\bigr).$$

Third term: $\mathbb{E}_X B^q$. Finally, by the Cauchy–Schwarz inequality, we have $B \le nF$ so

$$\mathbb{E}_X B^q \le n^q F^q.$$

Now it remains to simply put the pieces together to obtain

$$\mathbb{E}Z^q \le 12^q\big(128^q(\mathbb{E}Z_\epsilon)^q + 12^q q^{q/2}(\mathbb{E}U_\epsilon)^q + 320^q q^q(\mathbb{E}M)^q + 4^q F^q n^q q^q + 30^q F^q n^{q/2} q^{3q/2} + 800^q F^q q^{2q}\big),$$

proving the announced moment inequality.

In order to derive the exponential inequality, use Markov's inequality $\mathbb{P}\{Z > t\} \le t^{-q}\mathbb{E}Z^q$ and choose

$$q = C\min\left(\left(\frac{t}{\mathbb{E}U_\epsilon}\right)^2, \frac{t}{\mathbb{E}M}, \frac{t}{Fn}, \left(\frac{t}{F\sqrt{n}}\right)^{2/3}, \sqrt{\frac{t}{F}}\right)$$

for an appropriate constant C. $\square$

7. Convex risk minimization. Several successful algorithms for classification, including various versions of *boosting* and *support vector machines* are based on replacing the loss function by a convex function and minimizing the corresponding empirical convex risk functionals over a certain class of functions (typically over a ball in an appropriately chosen Hilbert or Banach space of functions). This approach has important computational advantages, as the minimization of the empirical convex functional is often computationally feasible by gradient descent algorithms. Recently significant theoretical advance has been made in understanding the statistical behavior of such methods, see, for example, Bartlett, Jordan and McAuliffe [5], Blanchard, Lugosi and Vayatis [7], Breiman [10], Jiang [25], Lugosi and Vayatis [30] and Zhang [48].

The purpose of this section is to extend the principle of convex risk minimization to the ranking problem studied in this paper. Our analysis also provides a theoretical framework for the analysis of some successful ranking algorithms such as the RANKBOOST algorithm of Freund, Iyer, Schapire and Singer [17]. In what follows we adapt the arguments of Lugosi and Vayatis [30] (where a simple binary classification problem was considered) to the ranking problem.

The basic idea is to consider ranking rules induced by real-valued functions, that is, ranking rules of the form

$$r(x, x') = \begin{cases} 1, & \text{if } f(x, x') > 0, \\ -1, & \text{otherwise,} \end{cases}$$

where $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is some measurable real-valued function. With a slight abuse of notation, we will denote by $L(f) = \mathbb{P}\{\operatorname{sgn}(Z) \cdot f(X, X') < 0\} = L(r)$ the risk of the ranking rule induced by f. [Here $\operatorname{sgn}(x) = 1$ if $x > 0$, $\operatorname{sgn}(x) = -1$ if $x < 0$ and $\operatorname{sgn}(x) = 0$ if $x = 0$.] Let $\phi : \mathbb{R} \to [0, \infty)$ be a convex *cost function* satisfying $\phi(0) = 1$ and $\phi(x) \ge \mathbb{I}_{[x \ge 0]}$. Typical choices of ϕ include the exponential cost function $\phi(x) = e^x$, the "logit" function $\phi(x) = \log_2(1 + e^x)$, or the "hinge loss" $\phi(x) = (1 + x)_+$. Define the *cost functional* associated to the cost function ϕ by

$$A(f) = \mathbb{E}\phi\big(-\operatorname{sgn}(Z) \cdot f(X, X')\big).$$

Obviously, $L(f) \le A(f)$. We denote by $A^* = \inf_f A(f)$ the "optimal" value of the cost functional where the infimum is taken over all measurable functions $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

The most natural estimate of the cost functional $A(f)$, based on the training data D_n, is the *empirical cost functional* defined by the U-statistic

$$A_n(f) = \frac{1}{n(n-1)} \sum_{i \neq j} \phi\big(-\operatorname{sgn}(Z_{i,j}) \cdot f(X_i, X_j)\big).$$

The ranking rules based on *convex risk minimization* we consider in this section minimize, over a set $\mathcal{F}$ of real-valued functions $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the empirical cost functional A_n, that is, we choose $f_n = \arg\min_{f \in \mathcal{F}} A_n(f)$ and assign the corresponding ranking rule

$$r_n(x, x') = \begin{cases} 1, & \text{if } f_n(x, x') > 0, \\ -1, & \text{otherwise.} \end{cases}$$

[Here we assume implicitly that the minimum exists. More precisely, one may define f_n as any function $f \in \mathcal{F}$ satisfying $A_n(f_n) \le \inf_{f \in \mathcal{F}} A_n(f) + 1/n$.]

By minimizing convex risk functionals, one hopes to make the excess convex risk $A(f_n) - A^*$ small. This is meaningful for ranking if one can relate the excess convex risk to the excess ranking risk $L(f_n) - L^*$. This may be done quite generally by recalling a result of Bartlett, Jordan and McAuliffe [5]. To this end, introduce the functions

$$H(\rho) = \inf_{\alpha \in \mathbb{R}} \big(\rho\phi(-\alpha) + (1-\rho)\phi(\alpha)\big)$$

and

$$H^-(\rho) = \inf_{\alpha\,:\,\alpha(2\rho-1) \le 0} \big(\rho\phi(-\alpha) + (1-\rho)\phi(\alpha)\big).$$

Defining ψ over $\mathbb{R}$ by

$$\psi(x) = H^-\left(\frac{1+x}{2}\right) - H^-\left(\frac{1-x}{2}\right),$$

Theorem 3 of [5] implies that for all functions $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,

$$L(f) - L^* \le \psi^{-1}\big(A(f) - A^*\big)$$

where ψ^{-1} denotes the inverse of ψ. Bartlett, Jordan and McAuliffe show that, whenever ϕ is convex, $\lim_{x \to 0} \psi^{-1}(x) = 0$, so convergence of the excess convex risk to zero implies that the excess ranking risk also converges to zero. Moreover, in most interesting cases $\psi^{-1}(x)$ may be bounded, for $x > 0$, by a constant multiple of $\sqrt{x}$ (such as in the case of exponential or logit cost functions) or even by x [e.g., if $\phi(x) = (1+x)_+$ is the so-called *hinge loss*].

Thus, to analyze the excess ranking risk $L(f) - L^*$ for convex risk minimization, it suffices to bound the excess convex risk. This may be done by decomposing it into "estimation" and "approximation" errors as follows:

$$A(f_n) - A^*(f) \le \left(A(f_n) - \inf_{f \in \mathcal{F}} A(f) \right) + \left(\inf_{f \in \mathcal{F}} A(f) - A^* \right).$$

Clearly, just like in Section 3, we may (loosely) bound the excess convex risk over the class $\mathcal{F}$ as

$$A(f_n) - \inf_{f \in \mathcal{F}} A(f) \le 2 \sup_{f \in \mathcal{F}} |A_n(f) - A(f)|.$$

To bound the right-hand side, assume, for simplicity, that the class $\mathcal{F}$ of functions is uniformly bounded, say $\sup_{f \in \mathcal{F}, x \in \mathcal{X}} |f(x)| \le B$. Then once again, we may appeal to Lemma A.1 and the bounded differences inequality which imply that for any $\lambda > 0$,

$$\begin{aligned}
&\mathbb{E} \exp\left(\lambda \sup_{f \in \mathcal{F}} |A_n(f) - A(f)| \right) \\
&\quad \le \mathbb{E} \exp\left(\lambda \sup_{f \in \mathcal{F}} \left(\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \phi\big(-\operatorname{sgn}\big(Z_{i,\lfloor n/2 \rfloor + i}\big) \right.\right. \\
&\qquad\qquad \left.\left. \times f\big(X_i, X_{\lfloor n/2 \rfloor + i}\big)\big) - A(f) \right) \right) \\
&\quad \le \exp\left(\lambda \mathbb{E} \sup_{f \in \mathcal{F}} \left(\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \phi\big(-\operatorname{sgn}\big(Z_{i,\lfloor n/2 \rfloor + i}\big) \right.\right. \\
&\qquad\qquad \left.\left. \times f\big(X_i, X_{\lfloor n/2 \rfloor + i}\big)\big) - A(f) \right) + \frac{\lambda^2 B^2}{2n} \right).
\end{aligned}$$

Now it suffices to derive an upper bound for the expected supremum appearing in the exponent. This may be done by standard symmetrization and contraction inequalities. In fact, by mimicking Koltchinskii and Panchenko [27] (see also the proof of Lemma 2 in Lugosi and Vayatis [30]), we obtain

$$\begin{aligned}
&\mathbb{E} \sup_{f \in \mathcal{F}} \left(\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \phi\big(-\operatorname{sgn}\big(Z_{i,\lfloor n/2 \rfloor + i}\big) \cdot f\big(X_i, X_{\lfloor n/2 \rfloor + i}\big)\big) - A(f) \right) \\
&\qquad \le 4B\phi'(B) \mathbb{E} \sup_{f \in \mathcal{F}} \left(\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \sigma_i \cdot f\big(X_i, X_{\lfloor n/2 \rfloor + i}\big) \right)
\end{aligned}$$

where $\sigma_1, \ldots, \sigma_{\lfloor n/2 \rfloor}$ i.i.d. Rademacher random variables independent of D_n, that is, symmetric sign variables with $\mathbb{P}\{\sigma_i = 1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$.

We summarize our findings:

PROPOSITION 12. *Let f_n be the ranking rule minimizing the empirical convex risk functional $A_n(f)$ over a class of functions f uniformly bounded by $-B$ and B. Then, with probability at least $1-\delta$,*

$$A(f_n) - \inf_{f\in\mathcal{F}} A(f) \le 8B\phi'(B)R_n(\mathcal{F}) + \sqrt{\frac{2B^2\log(1/\delta)}{n}}$$

where R_n denotes the Rademacher average

$$R_n(\mathcal{F}) = \mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}\sigma_i\cdot f\bigl(X_i, X_{\lfloor n/2\rfloor+i}\bigr)\right).$$

Many interesting bounds are available for the Rademacher average of various classes of functions. For example, in analogy of boosting-type classification problems, one may consider a class $\mathcal{F}_B$ of functions defined by

$$\mathcal{F}_B = \left\{f(x,x') = \sum_{j=1}^N w_j g_j(x,x') : N\in\mathbb{N}, \sum_{j=1}^N |w_j| = B, g_j\in\mathcal{R}\right\}$$

where $\mathcal{R}$ is a class of ranking rules as defined in Section 3. In this case it is easy to see that

$$R_n(\mathcal{F}_B) \le B R_n(\mathcal{R}) \le \text{const.}\frac{BV}{\sqrt{n}}$$

where V is the VC dimension of the "base" class $\mathcal{R}$.

Summarizing, we have shown that a ranking rule based on the empirical minimization $A_n(f)$ over a class of ranking functions $\mathcal{F}_B$ of the form defined above, the excess ranking risk satisfies, with probability at least $1-\delta$,

$$L(f_n) - L^* \le \psi^{-1}\left(8B\phi'(B)c\frac{BV}{\sqrt{n}} + \sqrt{\frac{2B^2\log(1/\delta)}{n}} + \left(\inf_{f\in\mathcal{F}_B} A(f) - A^*\right)\right).$$

This inequality may be used to derive the *universal consistency* of such ranking rules. For example, the following corollary is immediate.

COROLLARY 13. *Let $\mathcal{R}$ be a class of ranking rules of finite VC dimension V such that the associated class of functions $\mathcal{F}_B$ is rich in the sense that*

$$\lim_{B\to\infty}\inf_{f\in\mathcal{F}_B} A(f) = A^*$$

for all distributions of (X,Y). Then if f_n is defined as the empirical minimizer of $A_n(f)$ over $\mathcal{F}_{B_n}$ where the sequence B_n satisfies $B_n\to\infty$ and $B_n^2\phi'(B_n)/\sqrt{n}\to 0$, then

$$\lim_{n\to\infty} L(f_n) = L^* \qquad \textit{almost surely.}$$

Classes $\mathcal{R}$ satisfying the conditions of the corollary exist, we refer the reader to Lugosi and Vayatis [30] for several examples.

Proposition 12 can also be used for establishing performance bounds for kernel methods such as support vector machines. A prototypical kernel-based ranking method may be defined as follows. To lighten notation, we write $\mathcal{W} = \mathcal{X} \times \mathcal{X}$.

Let $k : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ be a symmetric positive definite function, that is,

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j k(w_i, w_j) \geq 0,$$

for all choices of $n, \alpha_1, \ldots, \alpha_n \in \mathbb{R}$ and $w_1, \ldots, w_n \in \mathcal{W}$.

A kernel-type ranking algorithm may be defined as one that performs minimization of the empirical convex risk $A_n(f)$ [typically based on the hinge loss $\phi(x) = (1+x)_+$] over the class $\mathcal{F}_B$ of functions defined by a ball of the associated reproducing kernel Hilbert space of the form [where $w = (x, x')$]

$$\mathcal{F}_B = \Biggl\{ f(w) = \sum_{j=1}^{N} c_j k(w_j, w) \colon N \in \mathbb{N}, \sum_{i,j=1}^{N} c_i c_j k(w_i, w_j) \leq B^2, w_1, \ldots, w_N \in \mathcal{W} \Biggr\}.$$

In this case we have

$$R_n(\mathcal{F}_B) \leq \frac{2B}{n} \mathbb{E} \sqrt{\sum_{i=1}^{\lfloor n/2 \rfloor} k\bigl((X_i, X_{\lfloor n/2 \rfloor + i}), (X_i, X_{\lfloor n/2 \rfloor + i})\bigr)},$$

see, for example, Boucheron, Bousquet and Lugosi [8]. Once again, universal consistency of such kernel-based ranking rules may be derived in a straightforward way if the approximation error $\inf_{f \in \mathcal{F}_B} A(f) - A^*$ can be guaranteed to go to zero as $B \to \infty$. For the approximation properties of such kernel classes we refer the reader to Cucker and Smale [14], Scovel and Steinwart [36], Smale and Zhou [38], Steinwart [39] etc.

REMARK 7 (*Fast rates*). A natural question is whether the arguments of Section 4 can be extended to prove fast rates of convergence for minimizers of the convex ranking risk. For ordinary binary classification such an analysis was carried out by Blanchard, Lugosi and Vayatis [7]. It is an interesting problem to explore whether the techniques of [7] extend to the setting of this section. However, the arguments are quite technical and point beyond the scope of the present paper.

APPENDIX A: BASIC FACTS ABOUT U-STATISTICS

Here we recall some basic facts about U-statistics. Consider the i.i.d. random variables $X, X_1, \ldots, X_n$ taking values in a set $\mathcal{X}$ and denote by

$$U_n = \frac{1}{n(n-1)} \sum_{i \neq j} q(X_i, X_j)$$

a U-statistic of order 2 where the kernel $q : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric real-valued function.

U-statistics have been studied in depth and their behavior is well understood. One of the classical inequalities concerning U-statistics is due to Hoeffding [23] which implies that, for all $t > 0$,

$$\mathbb{P}\{|U_n - \mathbb{E}U_n| > t\} \leq 2e^{-2\lfloor (n/2) \rfloor t^2} \leq 2e^{-(n-1)t^2}.$$

Hoeffding also shows that, if $\sigma^2 = \mathrm{Var}(q(X_1, X_2))$, then

$$\mathbb{P}\{|U_n - \mathbb{E}U_n| > t\} \leq 2 \exp\left(- \frac{\lfloor (n/2) \rfloor t^2}{2\sigma^2 + 2t/3} \right). \tag{A.1}$$

It is important to notice here that the latter inequality may be improved by replacing σ^2 by a smaller term. This is based on the so-called Hoeffding decomposition as described below.

The U-statistic U_n is said to be *degenerate* if its kernel q satisfies

$$\mathbb{E}(q(x, X)) = 0 \qquad \text{for all } x \in \mathcal{X}.$$

There are two basic representations of U-statistics which we recall next (see Serfling [37] for more details).

Average of "sums-of-i.i.d." blocks. This representation is the key for obtaining the "first-order" results of Section 3 for nondegenerate U-statistics. The U-statistic U_n can be expressed as

$$U_n = \frac{1}{n!} \sum_{\pi} \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} q\big(X_{\pi(i)}, X_{\pi(\lfloor n/2 \rfloor + i)}\big)$$

where the sum is taken over all permutations π of $\{1, \ldots, n\}$. The idea underlying this representation is to reduce the analysis to the case of sums of i.i.d. random variables. The next simple lemma is based on this representation.

LEMMA A.1. *Let $q_\tau : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be real-valued functions indexed by $\tau \in T$ where T is some set. If $X_1, \ldots, X_n$ are i.i.d. then for any convex nondecreasing*

function ψ,

$$\mathbb{E}\psi\left(\sup_{\tau\in T}\frac{1}{n(n-1)}\sum_{i\neq j}q_\tau(X_i,X_j)\right)$$
$$\leq \mathbb{E}\psi\left(\sup_{\tau\in T}\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}q_\tau\big(X_i,X_{\lfloor n/2\rfloor+i}\big)\right),$$

assuming the suprema are measurable and the expected values exist.

PROOF. The proof uses the same trick Hoeffding's inequalities mentioned above are based on. Observe that

$$\mathbb{E}\psi\left(\sup_{\tau\in T}\frac{1}{n(n-1)}\sum_{i\neq j}q_\tau(X_i,X_j)\right)$$
$$=\mathbb{E}\psi\left(\sup_{\tau\in T}\frac{1}{n!}\sum_{\pi}\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}q_\tau\big(X_{\pi(i)},X_{\pi(\lfloor n/2\rfloor+i)}\big)\right)$$
$$\leq\mathbb{E}\psi\left(\frac{1}{n!}\sum_{\pi}\sup_{\tau\in T}\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}q_\tau\big(X_{\pi(i)},X_{\pi(\lfloor n/2\rfloor+i)}\big)\right)$$

(since ψ is nondecreasing)

$$\leq\frac{1}{n!}\sum_{\pi}\mathbb{E}\psi\left(\sup_{\tau\in T}\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}q_\tau\big(X_{\pi(i)},X_{\pi(\lfloor n/2\rfloor+i)}\big)\right)$$

(by Jensen's inequality)

$$=\mathbb{E}\psi\left(\sup_{\tau\in T}\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}q_\tau\big(X_i,X_{\lfloor n/2\rfloor+i}\big)\right)$$

as desired. □

Hoeffding's decomposition. Another way to interpret U-statistics is based on an orthogonal expansion known as Hoeffding's decomposition.

Assuming that $q(X_1,X_2)$ is square integrable, $U_n-\mathbb{E}U_n$ may be decomposed as a sum T_n of i.i.d. random variables plus a *degenerate* U-statistic W_n. In order to write this decomposition, consider the following function of one variable

$$h(X_i)=\mathbb{E}(q(X_i,X)|X_i)-\mathbb{E}U_n,$$

and the function of two variables

$$\widehat{h}(X_i,X_j)=q(X_i,X_j)-\mathbb{E}U_n-h(X_i)-h(X_j).$$

Then we have the orthogonal expansion

$$U_n = \mathbb{E}U_n + 2T_n + W_n,$$

where

$$T_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i),$$

$$W_n = \frac{1}{n(n-1)}\sum_{i\neq j} \widehat{h}(X_i, X_j).$$

W_n is a degenerate U-statistic because its kernel $\widehat{h}$ satisfies

$$\mathbb{E}(\widehat{h}(X_i, X)|X_i) = 0.$$

Clearly, the variance of T_n is

$$\mathrm{Var}(T_n) = \frac{\mathrm{Var}(\mathbb{E}(q(X_1, X)|X_1))}{n}.$$

Note that $\mathrm{Var}(\mathbb{E}(q(X_1, X)|X_1))$ is less than $\mathrm{Var}(q(X_1, X))$ (unless q is already degenerate). Furthermore, the variance of the degenerate U-statistic W_n is of the order $1/n^2$. T_n is thus the leading term in this orthogonal decomposition. Indeed, the limit distribution of $\sqrt{n}(U_n - \mathbb{E}U_n)$ is the normal distribution $\mathcal{N}(0, 4\,\mathrm{Var}(\mathbb{E}(q(X_1, X)|X_1))$ (see [22]). This suggests that inequality (A.1) may be quite loose.

Indeed, exploiting further Hoeffding's decomposition (combined with arguments related to decoupling, randomization and hypercontractivity of Rademacher chaos) de la Peña and Giné [15] established a Bernstein's type inequality of the form (A.1) but with σ^2 replaced by the variance of the conditional expectation (see Theorem 4.1.13 in [15]).

Specialized to our setting with $q(X_i, X_j) = \mathbb{I}_{[Z_{i,j}\cdot r(X_i, X_j)<0]}$ the inequality of de la Peña and Giné states that

$$\mathbb{P}\{|L_n(r) - L(r)| > t\} \leq 4\exp\left(-\frac{nt^2}{8s^2 + ct}\right),$$

where $s^2 = \mathrm{Var}(\mathbb{P}\{Z \cdot r(X, X') < 0|X\})$ is the variance of the conditional expectation and c is some constant.

APPENDIX B: CONNECTION WITH THE ROC CURVE AND THE AUC CRITERION

In the bipartite ranking problem, the ROC curve (ROC standing for *Receiver Operator Characteristic*; see [20]) and the AUC criterion are popular measures for evaluating the performance of scoring functions in applications.

Let $s : \mathcal{X} \to \mathbb{R}$ be a scoring function. The ROC curve is defined by plotting the *true positive rate*

$$\mathrm{TPR}_s(x) = \mathbb{P}\big(s(X) \geq x | Y = 1\big)$$

against the *false positive rate*

$$\mathrm{FPR}_s(x) = \mathbb{P}\big(s(X) \geq x | Y = -1\big).$$

By a straightforward change of parameter, the ROC curve may be expressed as the graph of the power of the test defined by $s(X)$ as a function of its level α:

$$\beta_s(\alpha) = \mathrm{TPR}_s(q_{s,\alpha}),$$

where $q_{s,\alpha} = \inf\{x \in (0, 1) : \mathrm{FPR}_s(x) \leq \alpha\}$.

Observe that if $s(X)$ and Y are independent (i.e., when $\mathrm{TPR}_s = \mathrm{FPR}_s$), the ROC curve is simply the diagonal segment $\beta_s(\alpha) = \alpha$. This measure of accuracy induces a partial order on the set of all scoring functions: for any s_1, s_2, we say that s_1 is more accurate than s_2 if and only if its ROC curve is above the one of s_2 for every level α, that is, if and only if $\beta_{s_2}(\alpha) \leq \beta_{s_1}(\alpha)$ for all $\alpha \in (0, 1)$.

PROPOSITION B.1. *The regression function η induces an optimal ordering on $\mathcal{X}$ in the sense that its ROC curve is not below any other scoring function s:*

$$\forall \alpha \in [0, 1] \qquad \beta_\eta(\alpha) \geq \beta_s(\alpha).$$

PROOF. The result follows from the Neyman–Pearson lemma applied to the test of the null assumption "$Y = -1$" against the alternative "$Y = 1$" based on the observation X: the test based on the likelihood ratio $\eta(X)/(1 - \eta(X))$ is uniformly more powerful than any other test based on X. □

REMARK B.8. Note that the ROC curve does not characterize the scoring function. For any s and any strictly increasing function $h : \mathbb{R} \to \mathbb{R}$, s and $h \circ s$ clearly yield the same ordering on $\mathcal{X}$: $\beta_s = \beta_{h \circ s}$.

Instead of optimizing the ROC curve over a class of scoring functions which is a difficult task, a simple idea is to search for s that maximizes the Area Under the ROC Curve (known as the AUC criterion):

$$\mathrm{AUC}(s) = \int_0^1 \beta_s(\alpha)\, d\alpha.$$

This theoretical quantity may be easily interpreted in a probabilistic fashion as shown by the following proposition.

PROPOSITION B.2. *For any scoring function s,*

$$AUC(s) = \mathbb{P}\big(s(X) \geq s(X') | Y = 1, Y' = -1\big),$$

where (X, Y) and (X', Y') are independent pairs drawn from the binary classification model.

PROOF. Let U be a uniformly distributed random variable over $(0,1)$, independent of (X,Y). Denote by F_s the distribution function of $s(X)$ given $Y=-1$. Then

$$\begin{aligned}\mathrm{AUC}(s) &= \int_0^1 \mathbb{P}\big(s(X) \geq q_{s,\alpha} | Y=1\big)\, d\alpha \\ &= \mathbb{E}\big(\mathbb{P}\big(s(X) \geq F_s^{-1}(U) | Y=1\big)\big) \\ &= \mathbb{P}\big(s(X) \geq s(X') | Y=1, Y'=-1\big).\end{aligned}$$

□

Acknowledgments. We thank Gilles Blanchard and Gérard Biau for their valuable comments on a previous version of this manuscript.

REFERENCES

[1] ADAMCZAK, R. (2007). Moment inequalities for U-statistics. *Ann. Probab.* **34** 2288–2314. MR2294982

[2] AGARWAL, S., GRAEPEL, T., HERBRICH, R., HAR-PELED, S. and ROTH, D. (2005). Generalization bounds for the area under the ROC curve. *J. Machine Learning Research* **6** 393–425. MR2249826

[3] ARCONES, M. A. and GINÉ, E. (1993). Limit theorems for U-processes. *Ann. Probab* **21** 1494–1542. MR1235426

[4] ARCONES, M. A. and GINÉ, E. (1994). U-processes indexed by Vapnik–Cervonenkis classes of functions with applications to asymptotics and bootstrap of U-statistics with estimated parameters. *Stochastic Process. Appl.* **52** 17–38. MR1289166

[5] BARTLETT, P. L., JORDAN, M. I. and MCAULIFFE, J. D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101** 138–156. MR2268032

[6] BARTLETT, P. L. and MENDELSON, S. (2006). Empirical minimization. *Probab. Theory Related Fields* **135** 311–334. MR2240689

[7] BLANCHARD, G., LUGOSI, G. and VAYATIS, N. (2003). On the rates of convergence of regularized boosting classifiers. *J. Machine Learning Research* **4** 861–894. MR2076000

[8] BOUCHERON, S., BOUSQUET, O. and LUGOSI, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM Probab. Statist.* **9** 323–375. MR2182250

[9] BOUCHERON, S., BOUSQUET, O., LUGOSI, G. and MASSART, P. (2005). Moment inequalities for functions of independent random variables. *Ann. Probab.* **33** 514–560. MR2123200

[10] BREIMAN, L. (2004). Population theory for boosting ensembles. *Ann. Statist.* **32** 1–11. MR2050998

[11] CAO, Y., XU, J., LIU, T. Y., LI, H., HUANG, Y. and HON, H. W. (2006). Adapting ranking SVM to document retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 186–193. ACM Press, Seattle, WA.

[12] CORTES, C. and MOHRI, M. (2004). AUC optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems* **16** (S. Thrun, L. Saul and B. Schölkopf, eds.) 313–320. MIT Press.

[13] COSSOCK, D. and ZHANG, T. (2006). Subset ranking uning regression. *Proceedings of the 19th Annual Conference on Learning Theory COLT 2006* (G. Lugosi and H.U. Simon, eds.) 605–619. *Lecture Notes in Comput. Sci.* **4005**. Springer, Berlin. MR2280634

[14] CUCKER, F. and SMALE, S. (2002). On the mathematical foundations of learning. *Bull. Amer. Math. Soc.* **39** 1–49. MR1864085

[15] DE LA PEÑA, V. H. and GINÉ, E. (1999). *Decoupling*: *From Dependence to Independence*. Springer, New York. MR1666908

[16] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York. MR1383093

[17] FREUND, Y., IYER, R., SCHAPIRE, R. E. and SINGER, Y. (2004). An efficient boosting algorithm for combining preferences. *J. Machine Learning Research* **4** 933–969. MR2125342

[18] GINÉ, E., LATAŁA, R. and ZINN, J. (2000). Exponential and moment inequalities for U-statistics. In *High Dimensional Probability II. Progress Probab.* **47** 13–38. Birkhäuser, Boston. MR1857312

[19] GINÉ, E. and ZINN, J. (1984). Some limit theorems for empirical processes. *Ann. Probab.* **12** 929–989. MR0757767

[20] GREEN, D. M. and SWETS, J. A. (1966). *Signal Detection Theory and Psychophysics.* Wiley, New York.

[21] HAUSSLER, D. (1995). Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik–Chervonenkis dimension. *J. Combin. Theory Ser. A* **69** 217–232. MR1313896

[22] HOEFFDING, W. (1948). A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.* **19** 293–325. MR0026294

[23] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

[24] HOUDRÉ, C. and REYNAUD-BOURET, P. (2003). Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic Inequalities and Applications. Progr. Probab.* **56** 55–69. Birkhäuser, Basel. MR2073426

[25] JIANG, W. (2004). Process consistency for Adaboost (with discussion). *Ann. Statist.* **32** 13–29. MR2050999

[26] KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization (with discussion). *Ann. Statist.* **34** 2593–2706. MR2329442

[27] KOLTCHINSKII, V. and PANCHENKO, D. (2002). Empirical margin distribution and bounding the generalization error of combined classifiers. *Ann. Statist.* **30** 1–50. MR1892654

[28] LEDOUX, M. (1997). On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.* **1** 63–87. MR1399224

[29] LUGOSI, G. (2002). Pattern classification and learning theory. In *Principles of Nonparametric Learning* (L. Györfi, ed.) 5–62. Springer, Wienna. MR1987656

[30] LUGOSI, G. and VAYATIS, N. (2004). On the Bayes-risk consistency of regularized boosting methods (with discussion). *Ann. Statist.* **32** 30–55. MR2051000

[31] MAJOR, P. (2006). An estimate of the supremum of a nice class of stochastic integrals and U-statistics. *Probab. Theory Related Fields* **134** 489–537. MR2226889

[32] MASSART, P. (2007). *Concentration Inequalities and Model Selection*. Springer, Berlin. MR2319879

[33] MASSART, P. and NÉDÉLEC, E. (2006). Risk bounds for statistical learning. *Ann. Statist.* **34** 2326–2366. MR2291502

[34] MCDIARMID, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics 1989* 148–188. Cambridge Univ. Press. MR1036755

[35] RUDIN, C. (2006). Ranking with a p-norm push. In *Proceedings of COLT 2006* (P. Auer and R. Meir, eds.). *Lecture Notes in Comput. Sci.* **4005** 589–604. Springer, Berlin. MR2280633

[36] SCOVEL, S. and STEINWART, I. (2005). Fast rates for support vector machines. *Learning Theory* 279–294. *Lecture Notes in Comput. Sci.* **3559**. Springer, Berlin. MR2203268

[37] SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York. MR0595165

[38] SMALE, S. and ZHOU, D. X. (2003). Estimating the approximation error in learning theory. *Anal. Appl.* **1** 17–41. MR1959283

[39] STEINWART, I. (2001). On the influence of the kernel on the consistency of support vector machines. *J. Machine Learning Research* **2** 67–93. MR1883281

[40] STUTE, W. (1991). Conditional U-statistics. *Ann. Probab.* **19** 812–825. MR1106287

[41] STUTE, W. (1994). Universally consistent conditional U-statistics. *Ann. Statist.* **22** 460–473. MR1272093

[42] TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563. MR1419006

[43] TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002

[44] USUNIER, N., TRUONG, V., AMINI, M. and GALLINARI, P. (2005). Ranking with unlabeled data: A first study. In *Proceedings of NIPS'05 Workshop on Learning to Rank*. Whistler, Canada.

[45] VAPNIK, V. N. and CHERVONENKIS, A. YA. (1974). *Theory of Pattern Recognition*. Nauka, Moscow. (In Russian.) [German translation *Theorie der Zeichenerkennung* (1979) Akademie Verlag, Berlin.] MR0594437

[46] VITTAUT, J. N. and GALLINARI, P. (2006). Machine learning ranking for structured information retrieval. *Advances in Information Retrieval*. *Lecture Notes in Comput. Sci.* **3936** 338–349. Springer, Berlin.

[47] VU, H. T. and GALLINARI, P. (2005). Using RankBoost to compare retrieval systems. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, *CIKM'05* 309–310. ACM Press, New York.

[48] ZHANG, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization (with discussion). *Ann. Statist.* **32** 56–85. MR2051001

S. CLÉMENÇON
DÉPARTEMENT TSI—SIGNAL ET IMAGES
ECOLE NATIONALE SUPÉRIEURE DES TÉLÉCOMMUNICATIONS
37–39 RUE DAREAU
75014 PARIS
FRANCE
E-MAIL: clemenco@enst.fr

G. LUGOSI
DEPARTMENT OF ECONOMICS AND BUSINESS
UNIVERSITAT POMPEU FABRA
08005 BARCELONA
SPAIN
E-MAIL: lugosi@upf.es

N. VAYATIS
CENTRE DE MATHÉMATIQUES ET DE LEURS APPLICATIONS
ECOLE NORMALE SUPÉRIEURE DE CACHAN
61 AVENUE DU PRÉSIDENT WILSON
94235 CACHAN CEDEX
FRANCE
E-MAIL: vayatis@cmla.ens-cachan.fr