Ranking and scoring using empirical risk minimization ⋆

Stéphan Clémençon[1,3], Gábor Lugosi[2], and Nicolas Vayatis[3]

[1] MODALX - Université Paris X
92001 Nanterre Cedex, France
sclemenc@u-paris10.fr

[2] Department of Economics, Universitat Pompeu Fabra
Ramon Trias Fargas 25-27, 08005 Barcelona, Spain
lugosi@upf.es

[3] Laboratoire de Probabilités et Modèles Aléatoires - Université Paris VI
4, place Jussieu, 75252 Paris Cedex, France
vayatis@ccr.jussieu.fr

Abstract. A general model is proposed for studying ranking problems. We investigate learning methods based on empirical minimization of the natural estimates of the ranking risk. The empirical estimates are of the form of a U-statistic. Inequalities from the theory of U-statistics and U-processes are used to obtain performance bounds for the empirical risk minimizers. Convex risk minimization methods are also studied to give a theoretical framework for ranking algorithms based on boosting and support vector machines. Just like in binary classification, fast rates of convergence are achieved under certain noise assumption. General sufficient conditions are proposed in several special cases that guarantee fast rates of convergence.

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. In the ranking problem one has to compare two (or more) different observations and decide which one is "better". For example, in document retrieval applications, one may be concerned with comparing documents by degree of relevance for a particular request, rather than simply classifying them as relevant or not.

In this paper we establish a statistical framework for studying such ranking problems. We discuss a general model and point out that the problem may be approached by empirical risk minimization methods thoroughly studied in statistical learning theory with the important novelty that natural estimates of the ranking risk involve U-statistics. Therefore, the methodology is based on the

⋆ This research was supported in part by Spanish Ministry of Science and Technology and FEDER, grant BMF2003-03324, and by the PASCAL Network of Excellence under EC grant no. 506778.

theory of 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 [9].

In this paper we establish basic performance bounds for empirical minimization of the ranking risk. We also investigate conditions under which significantly improved results may be given. 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 rest of the paper is organized as follows. In Section 2, the basic models and the two versions of the ranking problem we consider are introduced. In Sections 3 and 4, we provide the basic uniform convergence and consistency results for empirical risk and convex risk minimizers. In Section 5 we describe the noise assumptions which take advantage of the structure of the U-statistics in order to obtain fast rates of convergence.

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 necessarily their labels Y and Y'. We think about X being "better" than X' if $Y > Y'$, that is, if $Z > 0$. The goal is to rank X and X' such 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

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

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, $L^* \le L(r)$.

The purpose of this paper is to investigate the construction of ranking rules of low risk based on training data. We assume that given 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*

$$L_n(r) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}_{[Z_{i,j} \cdot r(X_i,X_j) < 0]} \quad \text{where } Z_{i,j} = \frac{Y_i - Y_j}{2} .$$

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 [14] which implies that, for all $t > 0$, if $\sigma^2 = \mathrm{Var}(\mathbb{I}_{[Z\cdot r(X,X')<0]}) = L(r)(1-L(r))$, then

$$\mathbb{P}\{|L_n(r) - L(r)| > t\} \le 2\exp\left(-\frac{\lfloor (n/2) \rfloor t^2}{2\sigma^2 + 2t/3}\right) . \tag{1}$$

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

Hoeffding's decomposition. Hoeffding's decomposition (see [21] for more details) is a basic tool for studying U-statistics. Consider the i.i.d. random variables $X, X_1, ..., X_n$ 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 q (the so-called *kernel*) is a symmetric real-valued function. 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. r.v's plus a *degenerate* U-statistic W_n. In order to write this decomposition, consider the following function of one variable

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

and the function of two variables

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

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

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

W_n is called a *degenerate* U-statistic because $\mathbb{E}\left(\tilde{h}(X_i,X) \mid X_i\right) = 0$. Clearly,

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

Note that $\mathrm{Var}(\mathbb{E}(q(X_1, X) \mid X_1))$ 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 $1/n^2$. Thus, T_n is the leading term in this orthogonal decomposition. Indeed, the limit distribution of $\sqrt{n}(U_n - \mathbb{E}U_n)$ is the normal distribution $\mathcal{N}(0, 4\mathrm{Var}(\mathbb{E}(q(X_1, X) \mid X_1))$. This suggests that inequality (1) may be quite loose.

Indeed, exploiting further Hoeffding's decomposition, de la Peña and Giné [9] established a Bernstein's type inequality of the form (1) but with σ^2 replaced by the variance of the conditional expectation (see Theorem 4.1.13 in [9]). This remarkable improvement is not exploited in our "first-order" analysis (Sections 3 and 4) but will become crucial when establishing fast rates of convergence in Section 5.

Remark 1. (A MORE GENERAL FRAMEWORK.) One may consider a generalization of the setup described above. Instead of ranking just two observations X, X', one may be interested in ranking m independent observations $X^{(1)}, \ldots, X^{(m)}$. In this case 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

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

In this general case, natural estimates of $L(r)$ involve m-th order U-statistics. All results of this paper extend in a straightforward manner to this general setup. In order to lighten the notation, we restrict the discussion to the case described above, that is, to the case when $m = 2$ and the loss function is $\ell(\pi, \overline{\pi}) = \mathbb{I}_{[\pi \neq \overline{\pi}]}$.

Another formalization of this problem is the so-called *ordinal regression* approach (see Herbrich, Graepel, and Obermayer [13]) in which the relation between ranking and pairwise classification is also made clear. However, the fact that a sequence of pairs (X_i, X_j) of i.i.d. individual data (X_i) is no longer independent was not considered there.

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.

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

$$L^* = \mathbb{E} \min\{\eta(X)(1-\eta(X')), \eta(X')(1-\eta(X))\}$$

and also,

$$L^* = \mathrm{Var}\left(\frac{Y+1}{2}\right) - \frac{1}{2}\mathbb{E}\,|\eta(X)-\eta(X')| \le 1/4$$

where the equality $L^* = \mathrm{Var}\left(\frac{Y+1}{2}\right)$ holds when 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 $\eta(X) = \mathbb{P}\{Y = 1 \mid X\}$ through the quantity $\mathbb{E}\{|\eta(X)-\eta(X')|\}$ which is a classical measure of concentration, known as *Gini's mean difference*. It is clear from the form of the Bayes ranking rule that the optimal ranking rule is given by a scoring function s^* which 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 $L(s) \stackrel{\text{def}}{=} L(r)$, one has

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

Observe 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, e.g., [11]). More precisely, we have:

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

where $p = \mathbb{P}(Y = 1)$, so maximizing the AUC criterion is equivalent to minimizing the ranking risk.

Example 2. (A REGRESSION MODEL). Assume now that Y is real-valued and the joint distribution of X and Y is such that $Y = m(X) + \epsilon\sigma(X)$ where $m(x) = \mathbb{E}(Y|X = x)$ is the regression function and ϵ has a symmetric distribution around zero and is independent of X. Then clearly the optimal ranking rule r^* may be obtained by a scoring function s^* which 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., [10])

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

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. In our first-order approach it suffices to use the next simple inequality which reduces the problem to the study of ordinary empirical processes.

Lemma 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,*

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

assuming the suprema are measurable and the expected values exist.

The proof uses a similar trick Hoeffding's above-mentioned inequality are based on. The details are omitted.

Using the lemma with $\psi(x) = e^{\lambda x}$, we bound the moment generating function of the U-process by that of an ordinary empirical process. Then standard methods of handling empirical processes may be used directly. For example, the bounded differences inequality (see McDiarmid [20]) implies that

$$\log \mathbb{E} \exp\left(\lambda \sup_{r \in \mathcal{R}} |L_n(r) - L(r)|\right) \leq \lambda \mathbb{E} \sup_{r \in \mathcal{R}} |\tilde{L}_n(r) - L(r)| + \frac{\lambda^2}{4(n-1)} \ ,$$

where we have set $\tilde{L}_n(r) = \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{I}_{\left[Z_{i,\lfloor n/2 \rfloor + i} \cdot r(X_i, X_{\lfloor n/2 \rfloor + i}) < 0\right]}$. The expected value on the right-hand side may now be bounded by standard methods. For example, if the class $\mathcal{R}$ of indicator functions has finite VC dimension V, then

$$\mathbb{E} \sup_{r \in \mathcal{R}} \frac{1}{\lfloor n/2 \rfloor} \left| \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{I}_{\left[Z_{i,\lfloor n/2 \rfloor + i} \cdot r(X_i, X_{\lfloor n/2 \rfloor + i}) < 0\right]} - L(r) \right| \leq c\sqrt{\frac{V}{n}}$$

for a universal constant c (see, e.g., Lugosi [17]). By the Chernoff bound $P\{X > t\} \leq E \exp(\lambda X - \lambda t)$ we immediately obtain the following corollary:

Proposition 2. *Let $\mathcal{R}$ be a class of ranking rules of* VC *dimension V. Then for any $t > 0$,*

$$\mathbb{P}\left\{\sup_{r \in \mathcal{R}} |L_n(r) - L(r)| > c\sqrt{\frac{V}{n}} + t\right\} \leq e^{-(n-1)t^2} \ .$$

A similar result is proved in the bipartite ranking case by Agarwal, Har-Peled, and Roth ([1], [2]) with the restriction that their bound holds conditionally on a label sequence. Their analysis relies on a particular complexity measure called the rank-shatter coefficient but the core of the argument is the same (since they implicitly make use of the permutation argument to recover a sum of independent quantities).

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 (2) 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. However, in the above analysis of the ranking problem there is an additional weakness due to the fact that estimators based on U-statistics have an even smaller variance as we pointed it out above. 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 estimate $L(r)$ by

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

(The same holds for the results of Section 4 as well.) Thus, in the analysis above one looses the advantage of using U-statistics. In Section 5 it is shown that under certain, not uncommon, circumstances significantly smaller risk bounds are achievable. There it will have an essential importance to use the sharp exponential bounds for U-statistics.

4 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, e.g., Bartlett, Jordan, and McAuliffe [4], Blanchard, Lugosi and Vayatis [6], Breiman [8], Jiang [15], Lugosi and Vayatis [18], Zhang [23].

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 [11]. In what follows we adapt the arguments of Lugosi and Vayatis [18] (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 $L(f) \stackrel{\text{def}}{=} \mathbb{P}\{\text{sgn}(Z) \cdot f(X, X') < 0\} = L(r)$ the risk of the ranking rule induced by f. (Here $\text{sgn}(x) = 1$ if $x > 0$, $\text{sgn}(x) = -1$ if $x < 0$, and $\text{sgn}(x) = 0$ if $x = 0$.) Let $\phi : \mathbb{R} \to [0, \infty)$ a convex *cost function* satisfying $\phi(0) = 1$ and $\phi(x) \geq \mathbb{I}_{[x \geq 0]}$. 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

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

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

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

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

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

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 recent result of Bartlett, Jordan, and McAuliffe [4]. To this end, introduce the function

$$\begin{aligned} H(\rho) &= \inf_{\alpha \in \mathbb{R}} (\rho\phi(-\alpha) + (1 - \rho)\phi(\alpha)) \\ H^-(\rho) &= \inf_{\alpha : \alpha(2\rho - 1) \leq 0} (\rho\phi(-\alpha) + (1 - \rho)\phi(\alpha)) \ . \end{aligned}$$

Defining ψ over $\mathbb{R}$ by $\psi(x) = H^-((1 + x)/2) - H((1 + x)/2)$, Theorem 3 of [4] implies that for all functions $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,

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

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:

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

To bound the estimation error, 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 1 and the bounded differences inequality which imply that for any $\lambda > 0$,

$$\log \mathbb{E} \exp \left(\lambda \sup_{f\in\mathcal{F}} |A_n(f) - A(f)| \right) \le \lambda \mathbb{E} \sup_{f\in\mathcal{F}} \left(\tilde{A}_n(f) - A(f) \right) + \frac{\lambda^2 B^2}{2(n-1)} ,$$

where $\tilde{A}_n(f) = \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \phi\left(-\operatorname{sgn}(Z_{i,\lfloor n/2\rfloor+i}) \cdot f(X_i, X_{\lfloor n/2\rfloor+i})\right)$. 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 [16] (see also the proof of Lemma 2 in Lugosi and Vayatis [18]), the expectation on the right-hand side may be bounded by

$$4B\phi'(B)\mathbb{E} \sup_{f\in\mathcal{F}} \left(\frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \sigma_i \cdot f(X_i, X_{\lfloor n/2\rfloor+i}) \right)$$

where $\sigma_1, \ldots, \sigma_{\lfloor n/2 \rfloor}$ are i.i.d. Rademacher random variables independent of D_n. We summarize our findings:

Proposition 3. *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 $1-\delta$,*

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

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

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

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

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 1. *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

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

Classes $\mathcal{R}$ satisfying the conditions of the corollary exist, we refer the reader to Lugosi and Vayatis [18] for several examples.

Proposition 3 can also be used for establishing performance bounds for kernel methods such as support vector machines. The details are omitted for the lack of space.

5 Fast rates

As we have mentioned at the end of Section 3, the bounds obtained there may be significantly improved under certain conditions. It is well known (see, e.g., §5.2 in the survey [7] 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 combined with the fact that the optimal (Bayes) classifier is in the class of candidate classification rules (see, e.g., Massart and Nédélec [19], Tsybakov [22]).

The purpose of this section is to examine possibilities of obtaining such improved performance bounds for empirical ranking risk minimization. The main message is that in the ranking problem one also may 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.

Below we establish improved bounds for the excess risk for empirical ranking risk minimization introduced in Section 3 above. Similar results also hold for the estimator based on the convex risk $A(s)$ though some assumptions may be more difficult to interpret (see [6] for classification), and here we restrict our attention to the minimizer r_n of the empirical ranking risk $L_n(r)$ over a class $\mathcal{R}$ of ranking rules.

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* $\Lambda(r) \stackrel{\text{def}}{=} L(r) - L^* = \mathbb{E}q_r((X,Y),(X',Y'))$ given by:

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

which is a U-statistic of degree 2 with symmetric kernel q_r. 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

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

is a degenerate U-statistic with symmetric kernel

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

Now consider the estimator r_n obtained as the minimizer of

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

over all $r \in \mathcal{R}$.

In this section we work under the following basic assumptions:

(a) The class $\mathcal{R}$ of ranking rules has a finite VC dimension V.
(b) The optimal ranking rule r^* is in the class $\mathcal{R}$.
(c) For all $r \in \mathcal{R}$,

$$\mathrm{Var}(h_r(X,Y)) \le c\,\Lambda(r)^{\alpha}$$

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

The basic tools we need are an exponential inequality for U-processes indexed by a VC class of degenerate kernels due to Arcones and Giné [3] and a general inequality for empirical risk minimizers of Bartlett and Mendelson [5]. The Arcones-Giné inequality, simplified to the case we need states that there exists a universal constant C such that, with probability at least $1 - \delta$,

$$\sup_{r \in \mathcal{R}} |W_n(r)| \leq \frac{CV}{n-1} \log\left(\frac{1}{\delta}\right) . \tag{3}$$

Theorem 1. *Consider the minimizer of the empirical ranking risk $L_n(r)$ over a class $\mathcal{R}$ of ranking rules and assume that conditions (a),(b), and (c) listed above hold. Then there exists a universal constant C such that, with probability at least $1 - \delta$, the ranking risk of r_n satisfies*

$$L(r_n) - L^* \leq C \left(\frac{V \log(n/\delta)}{n}\right)^{1/(2-\alpha)} .$$

SKETCH OF PROOF. Let A be the event on which $\sup_{r \in \mathcal{R}} |W_n(r)| \leq \rho$, where $\rho = \frac{CV}{n-1} \log\left(\frac{2}{\delta}\right)$ and C denotes the constant in (3). Then by (3), $\mathbb{P}[A] \geq 1 - \delta/2$. By the Hoeffding decomposition of the U-statistic $\Lambda_n(r)$, it is clear that, on A, r_n is an ρ-minimizer of $(1/n)\sum_{i=1}^n f_r(X_i, Y_i)$ over $r \in \mathcal{R}$ (in the sense that the average calculated for $r = r_n$ exceeds the minimum by not more than ρ) where, for every $r \in \mathcal{R}$, we write $f_r(x, y) = \mathbb{E}q_r((X, Y), (x, y))$. Define $\tilde{r}_n$ as r_n on A and an arbitrary minimizer of $(1/n)\sum_{i=1}^n f_r(X_i, Y_i)$ 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 ρ-minimizer of $(1/n)\sum_{i=1}^n f_r(X_i, Y_i)$. Thus, we can use a general result of Bartlett and Mendelson [5] to bound the excess ranking risk $\Lambda(\tilde{r}_n) = \mathbb{E}(f_{\tilde{r}_n}(X, Y)|D_n)$ of $\tilde{r}_n$. To this end, we need an estimate on the L_2 covering numbers of the class of functions $\{f_r : r \in \mathcal{R}\}$. Now observe that for any pair $r, r' \in \mathcal{R}$, by Jensen's inequality,

$$\begin{aligned} d(f_r, f_{r'}) &= \sqrt{\mathbb{E}(f_r(X, Y) - f_{r'}(X, Y))^2} \\ &\leq \sqrt{\mathbb{E}(\mathbb{I}_{[(Y-Y')\cdot r(X,X')<0]} - \mathbb{I}_{[(Y-Y')\cdot r'(X,X')<0]})^2} . \end{aligned}$$

Thus, the L_2 covering numbers of the class $\{f_r : r \in \mathcal{R}\}$ are not more than those of the class of indicator functions $\{\mathbb{I}_{[(y-y')\cdot r(x,x')<0]} : r \in \mathcal{R}\}$. However, since $\mathcal{R}$ has VC dimension V, by Haussler's inequality [12], the covering numbers of this class satisfy $\log N(\epsilon) \leq cV \log(1/\epsilon)$. Then an argument similar to Theorem 2.12 of [5] may be used to complete the proof. ∎

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 $x \in \mathcal{X}$,

$$\mathbb{E}_{X'}(|\eta(x) - \eta(X')|^{-\alpha}) \leq c . \tag{4}$$

Proposition 4. *Under (4), we have, for all $s \in \mathcal{F}$,* $\mathrm{Var}(h_s(X,Y)) \le c\,\Lambda(s)^\alpha$.

PROOF.

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

Condition (4) 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 $\alpha < 1$, poses quite mild restrictions as it is highlighted in the following example:

Corollary 2. *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* $\epsilon > 0$,

$$\mathbb{E}_{X'}(\left|\eta(x) - \eta(X')\right|^{-1+\epsilon}) \le \frac{2B}{\epsilon}$$

and therefore, by Theorem 1 and Proposition 4, for every $\delta \in (0,1)$ *there is a constant* C *such that the excess ranking risk of the empirical minimizer* r_n *satisfies*

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

PROOF. The corollary follows simply by checking that (4) is satisfied for any $\alpha = 1 - \epsilon < 1$. The details are omitted. $\blacksquare$

The condition (4) 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 $\epsilon > 0$.

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 has a Gaussian density and is independent of X. 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 $L^* = \mathbb{E}\Phi\left(-\left|\Delta(X, X')\right|\right)$ whose maximal value is attained when the regression function $m(x)$ is constant. Furthermore, we have

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

where Φ is the distribution function of ϵ.

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

$$\mathbb{E}_{X'}(|\Delta(x, X')|^{-\alpha}) \leq c . \tag{5}$$

Proposition 5. *Under (5), we have, for all $s \in \mathcal{F}$, $\mathrm{Var}(h_s(X, Y)) \leq (2\Phi(c) - 1)\, \Lambda(s)^\alpha$.*

PROOF. By symmetry, $\left|2\Phi\left(\Delta(X, X')\right) - 1\right| = 2\Phi\left(\left|\Delta(X, X')\right|\right) - 1$. Then, using the concavity of the distribution function Φ on $\mathbb{R}_+$, we have, by Jensen's inequality,

$$\mathbb{E}_{X'}\Phi(|\Delta(x, X')|^{-\alpha}) \leq \Phi(\mathbb{E}_{X'}|\Delta(x, X')|^{-\alpha}) \leq \Phi(c) ,$$

where we have used (5) together with the fact that Φ is increasing. Now the result follows following the argument given in the proof of Proposition 4. ∎

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

Corollary 3. *Suppose that $m(X)$ has a bounded density and the conditional variance $\sigma(x)$ is bounded over $\mathcal{X}$. Then the noise condition 5 is satisfied for any $\alpha < 1$.*

Acknowledgements. We thank Gilles Blanchard for his valuable comments on a previous version of this manuscript, and also Gérard Biau for his careful remarks.

References

1. S. Agarwal, T. Graepel, R. Herbrich, and D. Roth (2004). A large deviation bound for the area under the ROC curve. In Proceedings of the 18th Annual Conference on Neural Information Processing Systems, Vancouver, Canada.
2. S. Agarwal, S. Har-Peled, and D. Roth (2005). A uniform convergence bound for the area under the ROC curve. In Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, Barbados.

3. M.A. Arcones and E. Giné (1994). *U*-processes indexed by Vapnik-Chervonenkis classes of functions with applications to asymptotics and bootstrap of *U*-statistics with estimated parameters. Stochastic Processes and their Applications, **52**, pp. 17-38.
4. P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe (2003). Convexity, classification, and risk bounds. Technical Report 638, Department of Statistics, U.C. Berkeley.
5. P.L. Bartlett and S. Mendelson (2003). Empirical minimization. Technical Report, Department of Statistics, U.C. Berkeley.
6. G. Blanchard, G. Lugosi, and N. Vayatis (2003). On the rate of convergence of regularized boosting classifiers. Journal of Machine Learning Research, 4:861-894.
7. O. Bousquet, S. Boucheron, and G. Lugosi (2004). Theory of classification: a survey of recent advances. ESAIM: Probability and Statistics, to appear.
8. L. Breiman (2004). Population theory for boosting ensembles. Annals of Statistics, **32**, pp. 1–11.
9. V. de la Peña and E. Giné (1999). *Decoupling: from dependence to independence*. Springer.
10. L. Devroye, L. Györfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.
11. Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer (2003). An Efficient Boosting Algorithm for Combining Preferences. Journal of Machine Learning Research, **4**, pp. 933-969.
12. D. Haussler (1995). Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. Journal of Combinatorial Theory, Series A, **69**, pp. 217–232.
13. R. Herbrich, T. Graepel, and K. Obermayer (2000). Large margin rank boundaries for ordinal regression. In A. Smola, P.L. Bartlett, B.Schölkopf, and D.Schuurmans (eds.), *Advances in Large Margin Classifiers*, The MIT Press, pp. 115–132.
14. W. Hoeffding (1963). Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association, **58**, pp. 13-30.
15. W. Jiang (2004). Process consistency for Adaboost (with discussion). Annals of Statistics, **32**, pp. 13–29.
16. V. Koltchinskii and D. Panchenko (2002). Empirical margin distribution and bounding the generalization error of combined classifiers. Annals of Statistics, **30**, pp. 1–50.
17. G. Lugosi (2002). Pattern classification and learning theory. In L. Györfi (editor), *Principles of Nonparametric Learning*, Springer, Wien, New York, pp. 1–56.
18. G. Lugosi and N. Vayatis (2004). On the Bayes-risk consistency of boosting methods (with discussion). Annals of Statistics, **32**, pp. 30–55.
19. P. Massart and E. Nédélec (2003). Risk bounds for statistical learning. Preprint, Université Paris XI.
20. C. McDiarmid (1989). On the method of bounded differences. In *Surveys in Combinatorics 1989*, pp. 148-188, Cambridge University Press.
21. R.J. Serfling (1980). Approximation theorems of mathematical statistics. John Wiley & Sons.
22. A. Tsybakov (2004). Optimal aggregation of classifiers in statistical learning. Annals of Statistics, **32**, pp. 135–166.
23. T. Zhang (2004). Statistical behavior and consistency of classification methods based on convex risk minimization (with discussion). Annals of Statistics, **32**, pp. 56–85.