Journal of Machine Learning Research 13 (2012) 1373-1392 Submitted 5/11; Revised 1/12; Published 5/12

On Ranking and Generalization Bounds

Wojciech Rejchel WREJCHEL@GMAIL.COM
Faculty of Mathematics and Computer Science
Nicolaus Copernicus University
Chopina 12/18
87-100 Toruń, Poland

Editor: Nicolas Vayatis

Abstract

The problem of ranking is to predict or to guess the ordering between objects on the basis of their observed features. In this paper we consider ranking estimators that minimize the empirical convex risk. We prove generalization bounds for the excess risk of such estimators with rates that are faster than $\frac{1}{\sqrt{n}}$. We apply our results to commonly used ranking algorithms, for instance boosting or support vector machines. Moreover, we study the performance of considered estimators on real data sets.

Keywords: convex risk minimization, excess risk, support vector machine, empirical process, U-process

1. Introduction

The problem of ranking is to predict or to guess the ordering between objects on the basis of their observed features. This problem has numerous applications in practice. We can mention information retrieval, banking, quality control or survival analysis. The problem is closely related to the classification theory, however it has its own specificity. In recent years many authors have focused their attention on this subject (Freund et al., 2004; Agarwal et al., 2005; Cossock and Zhang, 2006; Rudin, 2006; Clémençon et al., 2008).

In the paper we consider a population of objects equipped with a relation of (linear) ordering. For any two distinct objects o_1 and o_2 it holds either $o_1 \preceq o_2$ or $o_1 \succeq o_2$ (or maybe both), but it is unknown which is true. We lose little generality by assuming that real numbers y_1 and y_2 are assigned to the objects o_1 and o_2 in such a way that $o_1 \preceq o_2$ is equivalent to $y_1 \leq y_2$. Moreover, let d-dimensional vectors x_1 and x_2 describe observed or measured features of the objects and let the observation space $\mathcal{X}$ be a Borel subset of $\mathbb{R}^d$. We are to construct a function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, called a ranking rule, which predicts the ordering between objects in the following way:

$$\text{if } \ f(x_1,x_2) \leq 0, \text{ then we predict that } \ y_1 \leq y_2.$$

To measure the quality of a ranking rule f we introduce a probabilistic setting. Let us assume that two objects are randomly selected from the population. They are described by a pair of independent and identically distributed (with respect to the measure P) random vectors $Z_1 = (X_1, Y_1)$ and $Z_2 = (X_2, Y_2)$ taking values in $\mathcal{X} \times \mathbb{R}$. Random vectors X_1 and X_2 are regarded as observations, while Y_1 and Y_2 are unknown variables which define the ordering as above.

Most natural approach is to look for a function f which minimizes the risk (the probability of incorrect ranking)

$$L(f) = \mathbb{P}(\text{sign}(Y_1 - Y_2) f(X_1, X_2) < 0) \tag{1}$$

in some family of ranking rules $\mathcal{F}$, where $\text{sign}(t) = 1$ for $t > 0$, $\text{sign}(t) = -1$ for $t < 0$ and $\text{sign}(t) = 0$ for $t = 0$. Since we do not know the distribution P, we cannot solve this problem directly. But if we possess a learning sample $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$, then we can consider a sample analog of (1), namely the empirical risk

$$L_n(f) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}[\text{sign}(Y_i - Y_j) f(X_i, X_j) < 0], \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function. The ranking rule that minimizes (2) can be used as an estimator of the function that minimizes (1). Notice that $L_n(f)$ is a U-statistic of the order two for a fixed $f \in \mathcal{F}$. The main difficulty in this approach lies in discontinuity of the function (2). It entails that finding its minimizer is computationally difficult and not effective. This fact is probably the main obstacle to wider use of such estimators in practice. To overcome this problem one usually replaces the discontinuous loss function by its convex analog. This trick has been successfully used in the classification theory and has allowed to invent boosting algorithms (Freund and Schapire, 1997) or support vector machines (Vapnik, 1998). Therefore, instead of $0-1$ loss function we consider a convex and nonnegative loss function $\psi : \mathbb{R} \to \mathbb{R}$. Denote the "convex" risk of a ranking rule f by

$$Q(f) = \mathbb{E}\, \psi[\,\text{sign}(Y_1 - Y_2)\, f(X_1, X_2)],$$

and the "convex" empirical risk as

$$Q_n(f) = \frac{1}{n(n-1)} \sum_{i \neq j} \psi_f(Z_i, Z_j),$$

where $\psi_f(z_1, z_2) = \psi[\text{sign}(y_1 - y_2)\, f(x_1, x_2)]$. Notice that $Q_n(f)$ is also a U-statistic of the order two for a fixed function f. Therefore, features of U-process $\{Q_n(f) : f \in \mathcal{F}\}$ are the basis for our consideration on statistical properties of the rule $f_n = \arg\min_{f \in \mathcal{F}} Q_n(f)$ as an estimator of the unknown function $f^* = \arg\min_{f \in \mathcal{F}} Q(f)$. Niemiro and Rejchel (2009) stated theorems about the strong consistency and the asymptotical normality of the estimator f_n in the linear case, that is, when we consider linear ranking rules $f(x_1, x_2) = \theta^T (x_1 - x_2)$, where $\theta \in \mathbb{R}^d$. Similar studies on the asymptotic behaviour of estimators were done in Niemiro (1992) and Bose (1998).

In this paper we are interested in the excess risk of an estimator f_n (in the general model, not necessarily linear). This is the case when one compares the convex risk of f_n with the convex risk of the best rule in the class. Generalization bounds are very popular for such studying in the learning theory. They are probabilistic inequalities of the following form: for every $\alpha \in (0, 1)$

$$\mathbb{P}(\ Q(f_n) - Q(f^*) \leq \eta\ ) \geq 1 - \alpha, \tag{3}$$

where $\eta > 0$ is some small number that depends on the level α, the number n of elements in the sample, a family of ranking rules $\mathcal{F}$ and a loss function ψ, but it is independent of an unknown distribution P. Similar objects were widely studied in the classification theory (Blanchard et al., 2003, 2008; Lugosi and Vayatis, 2004; Bartlett et al., 2006). In ranking one can find them in Clémençon

et al. (2005, 2008). In the latter two papers Authors proved that with some restrictions on the class $\mathcal{F}$ the number η in (3) is equal to $C\sqrt{\frac{\ln(1/\alpha)}{n}}$, where C is some constant. Their inequalities can be applied to ranking analogs of support vector machines or boosting algorithms. Moreover, it was shown in the classification theory that better rates than $\frac{1}{\sqrt{n}}$ are possible to obtain in similar bounds to (3). Noticing the close relation between ranking and the classification theory Clémençon et al. (2008) formulated the question if one can get generalization bounds with "fast rates" for the excess risk in ranking? They gave a positive answer (Clémençon et al., 2008, Corollary 6) but only for estimators that minimize the empirical risk with $0-1$ loss. We have already mentioned about problems with finding such minimizers. Convex loss functions and estimators that minimize the convex empirical risk are used in practice. In this paper we indicate assumptions and methods that allowed us to obtain generalization bounds with better rates than $\frac{1}{\sqrt{n}}$ for the excess convex risk of such estimators. Similar studies were done in Rejchel (2009), but here we strengthen and extend those results. The construction of inequalities of the form (3) is based on the empirical and U-process theory. Empirical processes are well-known and widely described in the literature, while U-processes are not so popular. However, there are very comprehensive monographs about this theory (see de la Peña and Giné, 1999), which originates from Hoeffding (1948).

The paper is organized as follows: Section 2 is devoted to theoretical results. We show that using Hoeffding's decomposition our problem can be divided into two parts. In the first one (Sections 2.1 and 2.2) we are interested in properties of some empirical process. The second part (Section 2.3) is devoted to a U-process that we obtain after Hoeffding's decomposition. We state the main theorem and describe its applications to commonly used ranking algorithms in Section 2.4. In Section 3 we study the practical performance of described estimators on real data sets.

2. Generalization Bounds

First, let us write conditions on a family of ranking rules $\mathcal{F}$ that we need in later work. For simplicity, assume that $f(x_1,x_2) = -f(x_2,x_1)$ for every $f \in \mathcal{F}$ which implies that the kernel of a U-statistic $Q_n(f)$ is symmetric. Moreover, let the class $\mathcal{F}$ be uniformly bounded which means that there exists some constant $A_1 > 0$ such that for every $x_1, x_2 \in \mathcal{X}$ and $f \in \mathcal{F}$ we have $|f(x_1,x_2)| \leq A_1$. We will not repeat these conditions later.

Furthermore, we need some restrictions on the "richness" of a family of ranking rules $\mathcal{F}$. They are bounds for the covering number of $\mathcal{F}$ and are similar to conditions that can be often found in the literature (Pollard, 1984; de la Peña and Giné, 1999; Mendelson, 2002). Thus, let μ be a probability measure on $\mathcal{X}\times\mathcal{X}$ and let ρ_μ be a $\mathbb{L}^2$-pseudometric on $\mathcal{F}$ defined as

$$\rho_\mu(f_1,f_2) = \frac{1}{A_1}\sqrt{\int_{\mathcal{X}\times\mathcal{X}} [f_1(x_1,x_2) - f_2(x_1,x_2)]^2 \, d\mu(x_1,x_2)}. \tag{4}$$

The covering number $N(t,\mathcal{F},\rho_\mu)$ of the class $\mathcal{F}$ with a pseudometric ρ_μ and a radius $t>0$ is the minimal number of balls (with respect to ρ_μ) with centers in $\mathcal{F}$ and radii t needed to cover $\mathcal{F}$. Thus, $N(t,\mathcal{F},\rho_\mu)$ is the minimal number m with the property

$$\exists_{\bar{\mathcal{F}}\subset\mathcal{F},\,|\bar{\mathcal{F}}|=m} \quad \forall_{f\in\mathcal{F}} \quad \exists_{\bar{f}\in\bar{\mathcal{F}}} \quad \rho_\mu(f,\bar{f}) \leq t.$$

Consider the marginal distribution P^X of the vector X and two empirical measures: $P_n^X = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\nu_n = \frac{1}{n(n-1)}\sum_{i\neq j}\delta_{(X_i,X_j)}$, where $\delta_{(\cdot)}$ is the counting measure. The family $\mathcal{F}$ that we consider satisfies one of the following conditions:

Assumption A There exist constants $D_i, V_i > 0,\ i = 1,2$ such that for every measures of the form: $\mu_1 = P^X \otimes P_n^X$, $\mu_2 = \nu_n$ and each $t \in (0,1]$ we have

$$N(t, \mathcal{F}, \rho_{\mu_i}) \leq D_i t^{-V_i} \qquad i = 1,2.$$

Assumption B There exist constants $D_i > 0$, $V_i \in (0,1),\ i = 1,2$ such that for every measures of the form: $\mu_1 = P^X \otimes P_n^X$, $\mu_2 = \nu_n$ and each $t \in (0,1]$ we have

$$\ln N(t, \mathcal{F}, \rho_{\mu_i}) \leq D_i t^{-V_i} \qquad i = 1,2.$$

Families satisfying similar conditions to Assumption A are often called VC-classes or Euclidean (Nolan and Pollard, 1987; Pakes and Pollard, 1989), while classes that fulfill Assumption B are known as satisfying the uniform entropy condition (van der Vaart and Wellner, 1996). As we will see in Section 2.4 more restrictive Assumption A leads to better results.

The first tool that we use is Hoeffding's decomposition (de la Peña and Giné, 1999) of a U-statistic $Q_n(f) - Q_n(f^*)$ that allows to obtain the equality

$$Q(f) - Q(f^*) - [Q_n(f) - Q_n(f^*)] = 2P_n[Q(f) - Q(f^*) - P\psi_f + P\psi_{f^*}] - U_n(h_f - h_{f^*}),$$

where

$$\begin{aligned} P\psi_f(z_1) &= \mathbb{E}[\psi_f(Z_1, Z_2) | Z_1 = z_1], \\ P_n(g) &= \frac{1}{n}\sum_{i=1}^{n} g(Z_i), \\ U_n(h_f - h_{f^*}) &= \frac{1}{n(n-1)}\sum_{i \neq j} [h_f(Z_i, Z_j) - h_{f^*}(Z_i, Z_j)], \\ h_f(z_1, z_2) &= \psi_f(z_1, z_2) - P\psi_f(z_1) - P\psi_f(z_2) + Q(f). \end{aligned}$$

Therefore, Hoeffding's decomposition breaks a difference between a U-statistic $Q_n(f) - Q_n(f^*)$ and its expectation into the sum of iid random variables and a degenerate U-statistic $U_n(h_f - h_{f^*})$. The degeneration of a U-statistic means that the conditional expectation of its kernel is the zero-function, that is, $\mathbb{E}[h_f(Z_1, Z_2) - h_{f^*}(Z_1, Z_2) \,|\, Z_1 = z_1] = 0$ for each $z_1 \in \mathcal{X} \times \mathbb{R}$. In what follows, we will separately look for probabilistic inequalities of the appropriate order for the empirical and degenerate term.

2.1 Empirical Term

The empirical process theory is the basis for our consideration concerning the first component in Hoeffding's decomposition of U-statistics. To get better rates in this case one has to be able to uniformly bound second moments of functions from an adequate class by their expectations. This fact combined with some consequence of Talagrand's inequality (Talagrand, 1994) was the key to obtain fast rates in the classification theory. In this subsection we want to apply this method to ranking. First, we need a few preliminaries: let $\mathcal{G}$ be a class of real functions that is uniformly bounded by a constant $G > 0$. Moreover, let us introduce an additional sequence of iid random variables $\varepsilon_1, \ldots, \varepsilon_n$ (the Rademacher sequence). Variables ε_i's take values 1 or -1 with probability $\frac{1}{2}$ and are independent of the sample $Z_1, \ldots, Z_n$. Having the Rademacher sequence let us denote

$$R_n(\mathcal{G}) = \sup_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i g(Z_i),$$

and call an expression $\mathbb{E}R_n(\mathcal{G})$ the Rademacher average of the class $\mathcal{G}$. The expectation in the Rademacher average is taken with respect to both samples $Z_1,\ldots,Z_n$ and $\varepsilon_1,\ldots,\varepsilon_n$.

Besides, we should also introduce so called sub-root functions, which are nonnegative and non-decreasing functions $\phi : [0,\infty) \to [0,\infty)$ such that for each $r > 0$ the function $r \longmapsto \phi(r)/\sqrt{r}$ is non-increasing. They have a lot of useful properties, for example they are continuous and have the unique positive fixed point r^* (the positive solution of the equation $\phi(r) = r$). Proofs of these facts can be easily found in the literature (Bartlett et al., 2005). Finally, let the class

$$\mathcal{G}^* = \{\alpha g : g \in \mathcal{G}, \alpha \in [0,1]\}$$

denote a star-hull of $\mathcal{G}$ and $Pg = \mathbb{E}g(Z_1)$. Now we can state the aforementioned theorem for empirical processes which can be also found in Massart (2000) and Bartlett et al. (2005).

Theorem 1 *Let the class $\mathcal{G}$ be such that for some constant $B > 0$ and every $g \in \mathcal{G}$ we have $Pg^2 \leq BPg$. Moreover, if there exists a sub-root function ϕ with the fixed point r^*, which satisfies*

$$\phi(r) \geq B\mathbb{E}R_n(g \in \mathcal{G}^* : Pg^2 \leq r)$$

for each $r \geq r^*$, then for every $K > 1$ and $\alpha \in (0,1)$

$$\mathbb{P}\left(\forall_{g\in\mathcal{G}} \quad Pg \leq \frac{K}{K-1}P_n(g) + \frac{6K}{B}r^* + [22G + 5BK]\frac{\ln(1/\alpha)}{n}\right) \geq 1-\alpha.$$

The proof of this theorem is based on Talagrand's inequality applied to properly rescaled class $\mathcal{G}$ and can be found in Bartlett et al. (2005). Theorem 1 says that to get better bounds for the empirical term one needs to study properties of the fixed point r^* of a sub-root ϕ. However, it gives no general method for choosing ϕ, but it suggests to relate it to $\mathbb{E}R_n(g \in \mathcal{G}^* : Pg^2 \leq r)$. We will follow this suggestion, similar reasoning was carried out in Bartlett et al. (2005) or Boucheron et al. (2005). Of course, for every n the function

$$r \to \mathbb{E}R_n(g \in \mathcal{G}^* : Pg^2 \leq r)$$

is nonnegative and non-decreasing. Replacing a class $\mathcal{G}$ by its star-hull is needed to prove the last property from the definition of the sub-root function.

Using Theorem 1 we can state the following fact concerning the empirical term in Hoeffding's decomposition. The modulus of convexity of a function ψ that appears in this theorem is described in the next subsection.

Theorem 2 *Let the family of ranking rules $\mathcal{F}$ satisfy Assumption A and be convex. Moreover, if the modulus of convexity of a loss function ψ fulfills on the interval $[-A_1,A_1]$ the condition $\delta(t) \geq Ct^p$ for some constants $C > 0$ and $p \leq 2$, then for every $\alpha \in (0,1)$ and $K > 1$*

$$\mathbb{P}\left(\forall_{f\in\mathcal{F}} \quad Q(f) - Q(f^*) \leq \frac{K}{K-1}P_n(P\psi_f - P\psi_{f^*}) + C_1V_1\frac{\ln n + \ln(1/\alpha)}{n}\right) \geq 1-\alpha,$$

where the constant C_1 depends on K.

If the family $\mathcal{F}$ satisfies Assumption B instead of Assumption A, then for every $\alpha \in (0,1)$ and $K > 1$ with probability at least $1-\alpha$

$$\forall_{f\in\mathcal{F}} \quad Q(f) - Q(f^*) \leq \frac{K}{K-1}P_n(P\psi_f - P\psi_{f^*}) + C_2\max\left(\frac{\ln n}{n}, \frac{1}{n^\beta}\right) + C_3\frac{\ln(1/\alpha)}{n}$$

where $\frac{2}{3} < \beta = \frac{2}{2+V_1} < 1$. Constants C_2, C_3 depend on K.

Remark 3 *Although the constants C_1, C_2, C_3 can be recovered from the proofs we do not write their explicit formulas, because our task is to prove bounds with better rates, that is, which decrease fast with $n \to \infty$. For the same reason we do not attempt to optimize C_1, C_2 and C_3.*

Proof Consider the family of functions

$$P\psi_{\mathcal{F}} - P\psi_{f^*} = \{P\psi_f - P\psi_{f^*} : f \in \mathcal{F}\}.$$

A loss function ψ is convex so it is locally Lipschitz with constant L_ψ. Since $\mathcal{F}$ is uniformly bounded, then $P\psi_{\mathcal{F}} - P\psi_{f^*}$ is also uniformly bounded by $2L_\psi A_1$. Moreover, we show in the next subsection that if $\mathcal{F}$ is convex and the modulus of convexity of ψ satisfies the assumption given in Theorem 2, then one can prove that for some constant B and every function $f \in \mathcal{F}$

$$\mathbb{E}\,[P\psi_f(Z_1) - P\psi_{f^*}(Z_1)]^2 \leq B[Q(f) - Q(f^*)]. \tag{5}$$

The precise value of the constant B is given in Lemma 5 in Section 2.2. Therefore, the relation that is demanded in Theorem 1 between second moments and expectations of functions from the considered class holds. Applying this theorem to the class of functions $\mathcal{G} = \left\{ \frac{P\psi_f - P\psi_{f^*}}{2L_\psi A_1} : f \in \mathcal{F} \right\}$ and the sub-root function

$$\phi(r) = \frac{B}{2L_\psi A_1} \mathbb{E} R_n(g \in \mathcal{G}^* : Pg^2 \leq r)$$

we get the following probabilistic inequality

$$\mathbb{P}\left(\forall_{f \in \mathcal{F}} \quad Q(f) - Q(f^*) \leq \frac{K}{K-1} P_n\left(P\psi_f - P\psi_{f^*}\right) + C_1 r^* + C_2 \frac{\ln(1/\alpha)}{n} \right) \geq 1 - \alpha\,.$$

Constants C_1, C_2 and others that appear in this proof may change from line to line. To finish the proof of the first part of the theorem we have to bound the fixed point of the sub-root ϕ by $\frac{\ln n}{n}$. Described method is similar to consideration contained in Mendelson (2003) or Bartlett et al. (2005).

First we need two additional notations:

$$\mathcal{G}_r^* = \{g \in \mathcal{G}^* : Pg^2 \leq r\}$$

for some $r > 0$ and

$$\xi = \sup_{g \in \mathcal{G}_r^*} \frac{1}{n} \sum_{i=1}^n g^2(Z_i).$$

Using Chaining Lemma for empirical processes (Pollard, 1984) we obtain

$$\mathbb{E} R_n(\mathcal{G}_r^*) \leq \frac{C_1}{\sqrt{n}} \mathbb{E} \int_0^{\sqrt{\xi}/4} \sqrt{\ln N(t, \mathcal{G}_r^*, \rho_{P_n})}\, dt, \tag{6}$$

where

$$\rho_{P_n}(g_1, g_2) = \sqrt{\frac{1}{n} \sum_{i=1}^n [g_1(Z_i) - g_2(Z_i)]^2}\,.$$

Notice that $N(t, \mathcal{G}_r^*, \rho_{P_n}) \leq N(t, \mathcal{G}^*, \rho_{P_n}) \leq N(t/2, \mathcal{G}, \rho_{P_n}) \left\lceil \frac{1}{t} \right\rceil$ since from a cover of a family $\mathcal{G}$ with radius $t/2$ and a cover of the interval $[0,1]$ with radius $t/2$ one can easily construct a cover of a family $\mathcal{G}^*$. Besides, it is not difficult to show that

$$N(t, P\psi_{\mathcal{F}}, \rho_{P_n}) \leq N(t, \psi_{\mathcal{F}}, \rho_{P \otimes P_n}) \leq N\left(\frac{t}{L_\psi}, \mathcal{F}, \rho_{P^X \otimes P_n^X}\right),$$

since the first inequality follows from Nolan and Pollard (1987, Lemma 20) and to prove the second one we use the fact that ψ is locally Lipschitz. Thus, Assumption A and above properties of covering numbers imply that for some positive constants C and C_1

$$\ln N(t, \mathcal{G}_r^*, \rho_{P_n}) \leq C_1 V_1 \ln \frac{C}{t}.$$

So the right side of (6) can be bounded by $C_1\sqrt{\frac{V_1}{n}}\,\mathbb{E}\int_0^{\sqrt{\xi}/4}\sqrt{\ln\frac{C}{t}}\,dt$. Using Mendelson (2003, Lemma 3.8) and Jensen's inequality we obtain

$$C_1\sqrt{\frac{V_1}{n}}\,\mathbb{E}\int_0^{\sqrt{\xi}/4}\sqrt{\ln\frac{C}{t}}\,dt \leq C_1\sqrt{\frac{V_1}{n}}\sqrt{\mathbb{E}\xi}\sqrt{\ln\left(\frac{C}{\mathbb{E}\xi}\right)}.$$

Furthermore, applying Talagrand (1994, Corollary 3.4) to the family $\mathcal{G}_r^*$ we have

$$\mathbb{E}\xi \leq 8\mathbb{E}R_n(\mathcal{G}_r^*) + r.$$

Summarizing we have just shown that

$$\mathbb{E}R_n(\mathcal{G}_r^*) \leq C_1\sqrt{\frac{V_1}{n}}\sqrt{8\mathbb{E}R_n(\mathcal{G}_r^*) + r}\sqrt{\ln\frac{C}{r}}$$

which for the fixed point r^* implies

$$r^* \leq \frac{C_1 V_1}{n}\ln\frac{C}{r^*},$$

and now it is easy to get that $r^* \leq CV_1\frac{\ln n}{n}$.

In the second part of the theorem we use less restrictive Assumption B. Reasoning is the same as in the previous case, we need only to notice that

$$\ln N(t, \mathcal{G}_r^*, \rho_{P_n}) \leq C\left[\ln N\left(t, \mathcal{F}, \rho_{P^X \otimes P_n^X}\right) + \ln\frac{C_1}{t}\right] \leq C\left[t^{-V_1} + \ln\frac{C_1}{t}\right].$$

Therefore, the right side of (6) can be bounded by

$$\frac{C}{\sqrt{n}}\,\mathbb{E}\int_0^{\sqrt{\xi}/4}\sqrt{t^{-V_1}}\,dt + \frac{C}{\sqrt{n}}\,\mathbb{E}\int_0^{\sqrt{\xi}/4}\sqrt{\ln\frac{C_1}{t}}\,dt. \tag{7}$$

The second component in (7) has been just considered, so we focus on the first one. Notice that it is equal to $\frac{C}{\sqrt{n}}\,\mathbb{E}\xi^{1/2-V_1/4}$. Again, using Jensen's inequality and Talagrand (1994, Corollary 3.4) it is less than

$$\frac{C}{\sqrt{n}}\left[8\mathbb{E}R_n(\mathcal{G}_r^*) + r\right]^{1/2-V_1/4}$$

which implies that

$$\mathbb{E}R_n(\mathcal{G}_r^*) \leq \frac{C}{\sqrt{n}}\left[\left(8\mathbb{E}R_n(\mathcal{G}_r^*) + r\right)^{\frac{1}{2}-\frac{V_1}{4}} + \sqrt{\left(8\mathbb{E}R_n(\mathcal{G}_r^*) + r\right)\ln\frac{C_1}{r}}\right]. \tag{8}$$

For the fixed point r^* the inequality (8) takes the form

$$r^* \leq \frac{C}{\sqrt{n}} \left[(r^*)^{\frac{1}{2}-\frac{V_1}{4}} + \sqrt{r^* \ln \frac{C_1}{r^*}} \right],$$

so

$$r^* \leq C \max \left(\frac{\ln n}{n}, \frac{1}{n^{\frac{2}{2+V_1}}} \right)$$

and $\frac{2}{3} < \frac{2}{2+V_1} < 1$, since $0 < V_1 < 1$. ■

2.2 On the Inequality (5)

Theorem 2 in the previous subsection shows that better rates can be obtained if we are able to bound second moments of functions from the family $P\psi_{\mathcal{F}} - P\psi_{f^*}$ by their expectations. In this subsection we indicate conditions that are sufficient for even stronger relationship, namely

$$\mathbb{E}\left[\psi_f(Z_1,Z_2) - \psi_{f^*}(Z_1,Z_2)\right]^2 \leq B[Q(f) - Q(f^*)]. \tag{9}$$

The key object in further analysis is the modulus of convexity of the loss ψ. This function was very helpful in proving similar relation in the classification theory (Mendelson, 2002; Bartlett et al., 2006). With minor changes we will use it in our studies.

Definition 4 *The modulus of convexity of ψ is the function $\delta : [0,\infty) \to [0,\infty]$ defined as*

$$\delta(t) = \inf \left\{ \frac{\psi(x_1)+\psi(x_2)}{2} - \psi\left(\frac{x_1+x_2}{2}\right) : |x_1 - x_2| \geq t \right\}.$$

We illustrate this object with a few examples: for the quadratic function $\psi(x) = x^2$ we obtain $\delta(t) = t^2/4$, the modulus of convexity of the exponential function defined on the interval $[-a,a]$ is equal to $\delta(t) = \frac{t^2}{8\exp(a)} + o(t^2)$, whereas for $\psi(x) = \max[0, 1-x]$ we have $\delta(t) = 0$.

If the class $\mathcal{F}$ is convex, then the risk $Q : \mathcal{F} \to \mathbb{R}$ is the convex functional. It allows to consider the modulus of convexity of Q, that is given by

$$\tilde{\delta}(t) = \inf \left\{ \frac{Q(f_1)+Q(f_2)}{2} - Q\left(\frac{f_1+f_2}{2}\right) : d(f_1,f_2) \geq t \right\},$$

where d is the $\mathbb{L}^2$-pseudometric on $\mathcal{F}$, that is,

$$d(f_1,f_2) = \sqrt{\mathbb{E}\left[f_1(X_1,X_2) - f_2(X_1,X_2)\right]^2}.$$

The important property of the modulus of convexity is the fact that it can be often lower bounded by Ct^p for some $C, p > 0$. This relation is satisfied for many interesting convex functions, for instance e^{-x}, $\log_2(1+e^{-2x})$ or $[\max(0,1-x)]^2$ (the last case needs minor changes in consideration). This property implies the similar one for the modulus of convexity of the functional Q, which is sufficient to prove the relationship (9) between second moments and expectations of functions from the family $\psi_{\mathcal{F}} - \psi_{f^*}$. The following lemma, which is based on Bartlett et al. (2006, Lemma 7 and Lemma 8), can be stated:

Lemma 5 *If the family $\mathcal{F}$ is convex and there exist constants $C, p > 0$ such that the modulus of convexity of ψ satisfies*

$$\delta(t) \geq Ct^p, \tag{10}$$

then

$$\mathbb{E}\left[\psi_f(Z_1,Z_2) - \psi_{f^*}(Z_1,Z_2)\right]^2 \leq L_\psi^2 D_p [Q(f) - Q(f^*)]_,^{\min(1,2/p)} \tag{11}$$

where

$$D_p = \begin{cases} (2C)^{-2/p} & \text{if} \quad p \geq 2, \\ 2^{1-p} A_1^{2-p} C^{-1} & \text{if} \quad p < 2. \end{cases}$$

Proof Using Lipschitz property of ψ we can obtain

$$\begin{aligned} & \mathbb{E}\left[\psi_f(Z_1,Z_2) - \psi_{f^*}(Z_1,Z_2)\right]^2 \\ \leq \;& L_\psi^2 \mathbb{E}\left[\operatorname{sign}(Y_1 - Y_2) f(X_1,X_2) - \operatorname{sign}(Y_1 - Y_2) f^*(X_1,X_2)\right]^2 \\ = \;& L_\psi^2 d^2(f, f^*). \end{aligned} \tag{12}$$

The second step of the proof is based on showing that if the modulus δ satisfies (10), then the modulus $\tilde{\delta}$ also fulfills a similar condition. Namely, let $f_1, f_2 \in \mathcal{F}$ satisfy $d(f,g) \geq t$. Then from the definition of the modulus of convexity δ and (10)

$$\begin{aligned} & \frac{Q(f_1) + Q(f_2)}{2} - Q\left(\frac{f_1 + f_2}{2}\right) = \mathbb{E}\left[\frac{\psi_{f_1}(Z_1,Z_2) + \psi_{f_2}(Z_1,Z_2)}{2} - \psi_{\frac{f_1+f_2}{2}}(Z_1,Z_2)\right] \\ \geq \;& \mathbb{E}\,\delta(|\operatorname{sign}(Y_1 - Y_2) f_1(X_1,X_2) - \operatorname{sign}(Y_1 - Y_2) f_2(X_1,X_2)|) \\ = \;& \mathbb{E}\,\delta(|f_1(X_1,X_2) - f_2(X_1,X_2)|) \geq C\,\mathbb{E}\,|f_1(X_1,X_2) - f_2(X_1,X_2)|^p. \end{aligned}$$

Easy calculation (see Bartlett et al., 2006, the proof of Lemma 8) indicates that the modulus $\tilde{\delta}$ fulfills

$$\tilde{\delta}(t) \geq C_p t^{\max(2,p)}, \tag{13}$$

where $C_p = C$ for $p \geq 2$ and $C_p = C(2A_1)^{p-2}$, otherwise. Moreover, from the definition of the modulus $\tilde{\delta}$ and the fact that f^* is the minimizer of $Q(f)$ in the convex class $\mathcal{F}$ we have

$$\frac{Q(f) + Q(f^*)}{2} \geq Q\left(\frac{f + f^*}{2}\right) + \tilde{\delta}(d(f,f^*)) \geq Q(f^*) + \tilde{\delta}(d(f,f^*)).$$

Combining this fact with the inequality (12) and the property (13) of the modulus $\tilde{\delta}$ we get

$$\begin{aligned} Q(f) - Q(f^*) &\geq 2\tilde{\delta}\left(\frac{\sqrt{\mathbb{E}\left[\psi_f(Z_1,Z_2) - \psi_{f^*}(Z_1,Z_2)\right]^2}}{L_\psi}\right) \\ &\geq 2C_p\left(\frac{\sqrt{\mathbb{E}\left[\psi_f(Z_1,Z_2) - \psi_{f^*}(Z_1,Z_2)\right]^2}}{L_\psi}\right)^{\max(2,p)} \end{aligned}$$

which is equivalent to the inequality (11). ■

Thus, for convex functions that were mentioned before Lemma 5 we obtain in the inequality (11) the exponent equal to 1, because their modulus of convexity can be easily bounded from below with $p = 2$. However, if $p > 2$, then the exponent belongs to the interval $(0,1)$, but we can still bound the considered empirical process by an expression of the order better than $\frac{1}{\sqrt{n}}$ (Mendelson, 2002; Bartlett et al., 2006). Of course, we get better bounds if the exponent is closer to 1.

2.3 Degenerate Component

In this subsection we obtain exponential inequalities for degenerate U-processes. We bound the second term in Hoeffding's decomposition by $\frac{1}{n}$ that is sufficient to get better rates for the excess risk of ranking estimators. Let us recall that considered object has the following form

$$\left\{ U_n(h_f - h_{f^*}) = \frac{1}{n(n-1)} \sum_{i \neq j} [h_f(Z_i, Z_j) - h_{f^*}(Z_i, Z_j)] : f \in \mathcal{F} \right\}, \tag{14}$$

where

$$h_f(z_1, z_2) = \psi_f(z_1, z_2) - P\psi_f(z_1) - P\psi_f(z_2) + Q(f).$$

Moreover, kernels of the U-process (14) are symmetric, uniformly bounded and degenerate.

Similar problems were also considered in Arcones and Giné (1994); de la Peña and Giné (1999), Major (2006) and Adamczak (2007).

Theorem 6 *If a family of ranking rules $\mathcal{F}$ satisfies Assumption A, then for every $\alpha \in (0,1)$*

$$\mathbb{P}\left(\quad \forall_{f \in \mathcal{F}} \quad |U_n(h_f - h_{f^*})| \leq C_1 \max(V_1, V_2) \frac{\ln(C_2/\alpha)}{n} \right) \geq 1 - \alpha$$

for some constants $C_1, C_2 > 0$.

If a family of ranking rules $\mathcal{F}$ satisfies Assumption B, then for every $\alpha \in (0,1)$

$$\mathbb{P}\left(\quad \forall_{f \in \mathcal{F}} \quad |U_n(h_f - h_{f^*})| \leq \frac{C_3}{1 - \max(V_1, V_2)} \frac{\ln(C_4/\alpha)}{n} \right) \geq 1 - \alpha$$

for some constants $C_3, C_4 > 0$.

Remark 7 *In Assumption B we restrict to $V_1, V_2 < 1$, whereas in the empirical process theory these exponents usually belong to $(0,2)$. This restriction is needed to prove Theorem 6, namely to calculate the integral (19) in the proof of this theorem.*

Proof Our aim is to bound the expression

$$\mathbb{E} \exp\left(\lambda \sqrt{\sup_{f \in \mathcal{F}} |(n-1) U_n(h_f - h_{f^*})|} \right) \tag{15}$$

for every $\lambda > 0$. Combining it with Markov's inequality finishes the proof.

Introduce the Rademacher sequence $\varepsilon_1, \ldots, \varepsilon_n$ and the symmetrized U-process defined as

$$S_n(h_f - h_{f^*}) = \frac{1}{n(n-1)} \sum_{i \neq j} \varepsilon_i \varepsilon_j [h_f(Z_i, Z_j) - h_{f^*}(Z_i, Z_j)].$$

Using Symmetrization for U-processes (de la Peña and Giné, 1999) we can bound (15) by

$$C_2 \mathbb{E} \exp\left(C_1 \lambda \sqrt{\sup_{f \in \mathcal{F}} |(n-1) S_n(h_f - h_{f^*})|} \right). \tag{16}$$

Constants C_1, C_2 that appear in this proof may differ from line to line. If we fix $Z_1, \ldots, Z_n$, then we work with the Rademacher chaos process whose properties are well studied. By Arcones and Giné (1994, Formula 3.4 and 3.5) and Hölder's inequality we bound (16) by

$$C_2 \mathbb{E} \exp\left(C_1 \lambda^2 \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} |(n-1) S_n (h_f - h_{f^*})| \right),$$

where $\mathbb{E}_\varepsilon$ is conditional expectation with respect to the Rademacher sequence. To finish the proof we study the expression $\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} |(n-1) S_n (h_f - h_{f^*})|$. This step relies on Chaining Lemma for U-processes (Nolan and Pollard, 1987, Lemma 5), similar reasoning can be found in Arcones and Giné (1993) and Sherman (1993). For convenience let us denote $h_f - h_{f^*}$ by h. Furthermore, for fixed $Z_1, \ldots, Z_n$ consider a stochastic process

$$\left\{ J_n(h) = \frac{1}{En} \sum_{i \neq j} \varepsilon_i \varepsilon_j h(Z_i, Z_j) : h \in \mathcal{H} \right\}, \tag{17}$$

where E is the uniform bound on elements of $\mathcal{H}$. Define a pseudometric ρ on $\mathcal{H} = \{h_f - h_{f^*} : f \in \mathcal{F}\}$ as

$$\rho(h_1, h_2) = \frac{1}{E} \sqrt{\frac{1}{n(n-1)} \sum_{i \neq j} [h_1(Z_i, Z_j) - h_2(Z_i, Z_j)]^2}.$$

The process (17) satisfies assumptions of Chaining Lemma for U-processes with the function $\phi(x) = \exp\left(\frac{x}{\kappa} - 1\right)$, where κ is some positive constant. Indeed, $J_n(h_1 - h_2)$ is the Rademacher chaos of the order two, so from de la Peña and Giné (1999, Corollary 3.2.6) there exists $\kappa > 0$ such that

$$\mathbb{E}_\varepsilon \exp\left(\frac{|J_n(h_1 - h_2)|}{\kappa \sqrt{\mathbb{E}_\varepsilon [J_n(h_1 - h_2)]^2}} \right) \leq e.$$

Moreover, it is easy to calculate that $\mathbb{E}_\varepsilon [J_n(h_1 - h_2)]^2 \leq \rho^2(h_1, h_2)$, which implies that $\mathbb{E}_\varepsilon \phi\left(\frac{|J_n(h_1 - h_2)|}{\rho(h_1, h_2)} \right) \leq 1$. Therefore, we obtain the inequality

$$\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} |(n-1) S_n (h_f - h_{f^*})| \leq C_1 \int_0^{1/4} \ln N(t, \mathcal{H}, \rho)\, dt. \tag{18}$$

Besides, the covering number of the family $\mathcal{H}_1 + \mathcal{H}_2 = \{h_1 + h_2 : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$ clearly satisfies the inequality

$$N(2t, \mathcal{H}_1 + \mathcal{H}_2, \rho) \leq N(t, \mathcal{H}_1, \rho) N(t, \mathcal{H}_2, \rho).$$

If the family $\mathcal{F}$ fulfills Assumption A, then similarly to the proof of Theorem 2 we have $N(t, P\psi_{\mathcal{F}}, \rho_{P_n}) \leq C_1 t^{-V_1}$ and $N(t, \psi_{\mathcal{F}}, \rho) \leq C_2 t^{-V_2}$. Therefore, for some constants $C, C_1 > 0$

$$N(t, \mathcal{H}, \rho) \leq C t^{-C_1 \max(V_1, V_2)}$$

and the right-hand side of (18) is bounded (for some constants $C, C_1, C_2 > 0$) by

$$C_1 \max(V_1, V_2) \int_0^{1/4} \ln \frac{C}{t}\, dt \leq C_2 \max(V_1, V_2).$$

If the family satisfies Assumption B, then the right-hand side of (18) is bounded (for some constants $C, C_1 > 0$) by

$$C\int_0^{1/4} t^{-\max(V_1,V_2)}\,dt \le \frac{C_1}{1-\max(V_1,V_2)}\,. \tag{19}$$

Summarizing we obtain for every $\lambda > 0$

$$\mathbb{E}\exp\left(\lambda\sqrt{\sup_{f\in\mathcal{F}}|(n-1)U_n(h_f - h_{f^*})|}\right) \le C_2\exp(C_1\lambda^2)$$

and the form of the constant C_1 depends on the assumption (A or B) that is satisfied by the family $\mathcal{F}$. Finally, we take $\lambda = \sqrt{\frac{\ln(C_2/\alpha)}{C_1}}$ and use Markov's inequality. ■

2.4 Main Result and Examples

Our task relied on showing that in ranking, similarly to the classification theory, the convex excess risk can be bounded with better rates than $\frac{1}{\sqrt{n}}$ which were proved in Clémençon et al. (2008). By Hoeffding's decomposition the effort was divided into the empirical term (Sections 2.1 and 2.2) and the degenerate U-process (Section 2.3). Taking results of these three parts together we can state the main theorem.

Theorem 8 *Let the family of ranking rules $\mathcal{F}$ satisfy Assumption A and be convex. Moreover, if the modulus of convexity of a function ψ fulfills on the interval $[-A_1, A_1]$ the condition $\delta(t) \ge Ct^p$ for some constants $C > 0$ and $p \le 2$, then for every $\alpha \in (0,1)$*

$$\mathbb{P}\left(Q(f_n) - Q(f^*) \le C_1\max(V_1,V_2)\frac{\ln n + \ln(C_2/\alpha)}{n}\right) \ge 1-\alpha \tag{20}$$

for some constants C_1, C_2.

If the family $\mathcal{F}$ satisfies Assumption B instead of Assumption A, then for every $\alpha \in (0,1)$

$$\mathbb{P}\left(Q(f_n) - Q(f^*) \le C_3\max\left(\frac{\ln n}{n}, \frac{1}{n^\beta}\right) + \frac{C_4}{1-\max(V_1,V_2)}\frac{\ln(C_5/\alpha)}{n}\right) \ge 1-\alpha$$

for some constants C_3, C_4, C_5 and $\beta = \frac{2}{2+V_1} \in \left(\frac{2}{3}, 1\right)$.

Remark 9 *The dependence on exponents V_1, V_2 in the inequality (20) is the same as in Clémençon et al. (2008, Corollary 6), where one considered minimizers of the empirical risk with $0-1$ loss and the family $\mathcal{F}$ with finite Vapnik-Chervonenkis dimension.*

Proof Let us slightly modify Hoeffding's decomposition of the U-statistic $Q_n(f) - Q_n(f^*)$, namely for each $K > 2$

$$\begin{aligned}&(K-2)[Q(f) - Q(f^*)] - K[Q_n(f) - Q_n(f^*)]\\ &= 2P_n\left\{(K-1)[Q(f) - Q(f^*)] - K(P\psi_f - P\psi_{f^*})\right\} - K[U_n(h_f) - U_n(h_{f^*})].\end{aligned}$$

Therefore, the first part of Theorem 2, Lemma 5 and Theorem 6 are sufficient to prove that for every $\alpha \in (0,1)$ and $K > 2$ with probability at least $1-\alpha$

$$\forall_{f\in\mathcal{F}} \quad Q(f) - Q(f^*) \leq \frac{K}{K-2}[Q_n(f) - Q_n(f^*)] + C_1 \max(V_1, V_2)\left[\frac{\ln n + \ln(C_2/\alpha)}{n}\right]$$

for some constants C_1, C_2. Moreover, for the rule f_n that minimizes the empirical convex risk we have $Q_n(f_n) - Q_n(f^*) \leq 0$. If the family $\mathcal{F}$ satisfies Assumption B we use the second thesis of Theorem 2 and further reasoning is the same. ■

Now we give three examples of ranking procedures that we can apply Theorem 8 to.

Example 1 *Consider the family $\mathcal{F}$ containing linear ranking rules*

$$\mathcal{F} = \{ f(x_1,x_2) = \theta^T(x_1 - x_2) : \theta, x_1, x_2 \in \mathbb{R}^d \}$$

In this case our prediction of the ordering between objects depends on the hyperplane that the vector $x_1 - x_2$ belongs to. The family $\mathcal{F}$ is convex. Moreover, the class $\{subgraph(f) : f \in \mathcal{F}\}$, where

$$subgraph(f) = \{(x_1,x_2,t) \in \mathcal{X}^2 \times \mathbb{R} : 0 < t < f(x_1,x_2) \text{ or } f(x_1,x_2) < t < 0\},$$

is by Pakes and Pollard (1989, Lemma 2.4 and 2.5) a VC-class of sets. Thus, Pakes and Pollard (1989, Lemma 2.12) implies that the family $\mathcal{F}$ satisfies Assumption A. If we take a "good" function ψ (for example one of functions mentioned in Section 2.2), then we obtain generalization bounds for the excess risk of the estimator f_n of the order $\frac{\ln n}{n}$.

Theorem 8 can be also applied to a popular ranking procedure called "boosting". Here we are interested in a ranking version of AdaBoost that uses the exponential loss function.

Example 2 *Let $\mathcal{R} = \{r : \mathcal{X} \times \mathcal{X} \to \{-1,1\}\}$ be a family of "base" ranking rules with finite Vapnik-Chervonenkis dimension. The output of the algorithm is an element of a convex T-hull of $\mathcal{R}$, where T is the number of iterations of the procedure. Namely, it belongs to the family*

$$conv_T(\mathcal{R}) = \{ f(x_1,x_2) = \sum_{j=1}^{T} w_j r_j(x_1,x_2) : \sum_{j=1}^{T} w_j = A_1, \\ w_j \geq 0, r_j \in \mathcal{R} \text{ for } j = 1,\ldots,T \}.$$

This class is obviously convex. The family $\mathcal{R}$ has finite VC dimension, so a class

$$\{A_r = \{(x_1,x_2) : r(x_1,x_2) = 1\} : r \in \mathcal{R}\}$$

is a VC-class of sets. The subgraph of each $r \in \mathcal{R}$ has the following form

$$\{(x_1,x_2) \in A_r \text{ and } t \in (0,1)\} \cup \{(x_1,x_2) \in A_r^c \text{ and } t \in (-1,0)\}.$$

Again using Pakes and Pollard (1989, Lemma 2.5 and 2.12) we obtain that $N(t,\mathcal{R},\rho_\mu) \leq Ct^{-V}$ for some constants $C, V > 0$ and every probability measure μ on $\mathcal{X} \times \mathcal{X}$. Quick calculation shows that

$$N(t, conv_T(\mathcal{R}), \rho_\mu) \leq C_1 t^{-T(V+1)},$$

so $\mathcal{F}$ satisfies Assumption A. Furthermore, the modulus of convexity of $\psi(x) = \exp(-x)$ fulfills on the interval $[-A_1, A_1]$ the condition $\delta(t) > \frac{t^2}{8\exp(A_1)}$. Thus, in this example we also obtain generalization bounds for the excess convex risk of f_n of the order $\frac{\ln n}{n}$.

The last example is a ranking version of support vector machines.

Example 3 *Let $K : \mathcal{X}^2 \times \mathcal{X}^2 \to \mathbb{R}$ be a kernel that is symmetric, continuous and nonnegative definite function. The last property means that for every natural number m, vectors $\bar{x_1}, \ldots, \bar{x_m} \in \mathcal{X}^2$ and $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$*

$$\sum_{i,j=1}^{m} \alpha_i \alpha_j K(\bar{x_i}, \bar{x_j}) \geq 0.$$

One can show (Cucker and Smale, 2002) that for every kernel K there exists the unique Hilbert space H_K (called reproducing kernel Hilbert space) of real functions on $\mathcal{X}^2$ that the inner product of its elements is defined by K. Namely, H_K is the completion of

$$\text{span}\{K(\bar{x}, \cdot) : \bar{x} \in \mathcal{X}^2\},$$

and the inner product is defined by

$$\langle f_1, f_2 \rangle = \sum_{i=1}^{k} \sum_{j=1}^{m} \alpha_i \beta_j K(\bar{x_i}, \bar{x_j})$$

for $f_1(\cdot) = \sum_{i=1}^{k} \alpha_i K(\bar{x_i}, \cdot)$ and $f_2(\cdot) = \sum_{j=1}^{m} \beta_j K(\bar{x_j}, \cdot)$.

Similarly to SVM in the classification theory our task is to linearly separate (with possibly wide "margin") two sets: $\{(X_i, X_j) : Y_i > Y_j, 1 \leq i \neq j \leq n\}$ and $\{(X_i, X_j) : Y_i < Y_j, 1 \leq i \neq j \leq n\}$, which can be solved using Lagrange multipliers. This primary problem is "transposed" from the $\mathcal{X}^2 \subset \mathbb{R}^{2d}$ to H_K by the function $\bar{x} \longmapsto K(\bar{x}, \cdot)$ and by the "kernel trick" we obtain the nonlinear procedure - comprehensive descriptions are in Cortes and Vapnik (1995), Burges (1998), Vapnik (1998) and Blanchard et al. (2008). Finally, we are to minimize the empirical convex risk of the form

$$Q_n(f) = \frac{1}{n(n-1)} \sum_{i \neq j} \max\left[0, 1 - \text{sign}(Y_i - Y_j) f(X_i, X_j)\right] + \lambda ||f||^2$$

in some ball with radius R in the Hilbert space H_K, that is

$$\mathcal{F} = \{f \in H_K : ||f|| \leq R\}$$

and $\lambda > 0$ is a parameter. Consider a Gaussian kernel of the form

$$K(\bar{x}, \bar{x'}) = \exp\left(-\sigma^2 ||\bar{x} - \bar{x'}||_2^2\right),$$

where $\bar{x}, \bar{x'} \in \mathcal{X}^2$, $||\bar{x}||_2 = \sqrt{\sum_{i=1}^{2d} \bar{x_i}^2}$ and $\sigma > 0$ is a scale parameter. Using Scovel and Steinwart (2007, Theorem 3.1) we obtain that for every compact set $\mathcal{X}$, $\sigma \geq 1$ and $0 < V < 1$

$$\ln N(t, \mathcal{F}, C(\mathcal{X}^2)) \leq C t^{-V} \tag{21}$$

for some constant C dependent on V, d, σ and R. The covering number $N(t, \mathcal{F}, C(\mathcal{X}^2))$ denotes the minimal number of balls with centers in the space of continuous functions on $\mathcal{X}^2$ with the metric $d(f_1, f_2) = \max_{\bar{x} \in \mathcal{X}^2} |f_1(\bar{x}) - f_2(\bar{x})|$ needed to cover $\mathcal{F}$. This definition differs from ours given in the beginning of Section 2. But Steinwart (2001) proved that H_K corresponding to the Gaussian kernel is dense in $C(\mathcal{X}^2)$, so we can use the property (21) in our studies. Moreover, for every probability

measure μ on $\mathcal{X}^2$ we have $\rho_\mu(f_1,f_2) \le d(f_1,f_2)$, where ρ_μ is defined by (4). Thus, the family $\mathcal{F}$ satisfies Assumption B and is convex. The inequality (5) with $B = \frac{2(2R\lambda+1)^2}{\lambda}$ can be obtained using almost the same arguments as Scovel and Steinwart (2005, Section 6.1). Therefore, we get

$$\mathbb{P}\left(Q(f_n) - Q(f^*) \le C_1 \max\left(\frac{\ln n}{n}, \frac{1}{n^\beta}\right) + C_2 \frac{\ln(C_3/\alpha)}{n}\right) \ge 1-\alpha$$

with $\frac{2}{3} < \beta < 1$.

In the paper we consider ranking estimators that minimize the convex empirical risk. The natural question is: are these estimators also "good" in the case of the primary $0-1$ loss function? Is there any relation between the excess risk and the convex excess risk? Let us introduce, similarly to Clémençon et al. (2008), two notations

$$\rho_+(X_1,X_2) = \mathbb{P}(Y_1 > Y_2 \,|\, X_1, X_2)$$

and

$$\rho_-(X_1,X_2) = \mathbb{P}(Y_1 < Y_2 \,|\, X_1, X_2).$$

It is easy to see that the ranking rule

$$\bar{f}(x_1,x_2) = 2\,\mathbb{I}_{[\rho_+(x_1,x_2) \ge \rho_-(x_1,x_2)]} - 1$$

minimizes the risk (1) in the class of all measurable functions. Denote $L^* = L(\bar{f})$. Let Q^* be the minimal value of $Q(f)$ for every measurable functions $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Bartlett et al. (2006) proved the relation between the excess risks and the convex excess risk for the classification theory. However, Clémençon et al. (2008) noticed that those results can be applied to ranking. They obtained that for every ranking rule f

$$\gamma(L(f) - L^*) \le Q(f) - Q^*$$

for some invertible function γ that depends on ψ. Moreover, γ can be computed in most interesting cases, for instance: $\gamma(x) = 1 - \sqrt{1-x^2}$ for $\psi(x) = \exp(-x)$.

Divide the difference $Q(f) - Q^*$ into the sum of two terms

$$[Q(f) - Q(f^*)] + [Q(f^*) - Q^*]. \tag{22}$$

The first component in (22), so called "estimation error", tells us how close the risk of f is to the risk of the best element in the class $\mathcal{F}$. The second term ("approximation error") describes how much we lose using the family $\mathcal{F}$. In the paper we study the estimation error, however approximation properties of the family $\mathcal{F}$ are also important problems. For instance they were considered in Cucker and Smale (2002), Lugosi and Vayatis (2004) and Scovel and Steinwart (2007).

3. Experiments

This section is devoted to results of our experiments on real data sets (Frank and Asuncion, 2010). We compare the performance of different SVM's for ranking problems. In Section 2.4 we describe a general method to obtain such procedures, but one can propose some simplification of this idea

that is useful in practice. Consider linearly separable case which means that there exists a vector $\theta \in \mathbb{R}^d$ such that

$$\theta^T X_i > \theta^T X_j \quad \text{for} \quad Y_i > Y_j, \quad 1 \leq i \neq j \leq n.$$

Thus, our task is to assign differences $X_i - X_j$ to classes defined by $\text{sign}(Y_i - Y_j)$. We assume that the distribution of the variable Y is continuous, so $\mathbb{P}(Y_1 = Y_2) = 0$. Therefore, we can use SVM for the classification theory to solve ranking problems if we consider differences of observations in place of observations. Thus, instead of a kernel $\mathcal{K} : \mathcal{X}^2 \times \mathcal{X}^2 \to \mathbb{R}$ we can use a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if we take

$$\mathcal{K}((x_1, x_2), (x_3, x_4)) = K(x_1 - x_2, x_3 - x_4).$$

The kernel $\mathcal{K}$ is symmetric, continuous and nonnegative definite by the same properties of the kernel K. Therefore, all calculations done by a procedure are made in $\mathbb{R}^d$ instead of $\mathbb{R}^{2d}$. Similar considerations can be found in Herbrich et al. (2000) and Joachims (2006).

To our experiments we use "e1071" package in "R" (R Development Core Team, 2009; Dimitriadou et al., 2010). We choose three types of kernels:

a) linear – $K(x_1, x_2) = < x_1, x_2 >_{\mathbb{R}^d}$,

b) polynomial – $K(x_1, x_2) = < x_1, x_2 >^3_{\mathbb{R}^d}$,

c) Gaussian – $K(x_1, x_2) = \exp\left(-\frac{1}{2}||x_1 - x_2||^2_{\mathbb{R}^d}\right)$

and two values of the parameter λ: 1 and $\frac{1}{10}$. Less value of λ corresponds to the case when the algorithm should be more adjusted to the sample. Greater value of λ has an effect in wider margin.

We divide every considered data sets into two subsets. The first one is used as a learning sample and we determine an estimator on it. On the second subset we test the estimator, that is, we take two objects and check if the ordering indicated by the estimator is the same as the true one. We repeat the experiment for every data set thirty times and average proportions of wrong decisions are presented in tables below. We denote SVM with the linear kernel and the parameter λ equal to 1 and $\frac{1}{10}$ by L(1) and L(10), respectively. Similarly, W(1) and W(10) stand for polynomial kernels, and G(1) and G(10) for Gaussian kernels.

The first data set concerns experiments that the concrete compressive strength was measured (Yeh, 1998). There are more than 1000 observations, 9 features are considered such that the age of material, contents of water, cement and other ingredients, and finally the concrete compressive strength. In Table 1 we compare errors in predicting the ordering between objects by six algorithms. Notice that in both cases (a learning sample with 100 and 300 elements) SVM with Gaussian kernels

Error	L(1)	L(10)	W(1)	W(10)	G(1)	G(10)
n=100	0,198	0,196	0,199	0,196	0,179	0,185
n=300	0,191	0,189	-	-	0,165	0,179

Table 1: Concrete compressive strength

have least errors, and among them G(1) is better. Proportions of wrong decisions of remaining four algorithms are similar. Besides, for linear and polynomial kernels greater adjustment to the sample has an effect in slightly better effectiveness, contrary to G(1) and G(10). The mark "-" in the

table means that the algorithm did not calculate an estimate for 100 minutes. Comparing to three following data sets it usually happens for polynomial SVM and n=300. Such numerical problems occur, since the number of pairs of instances, that algorithms work with, increases with the square of the sample size n. It makes these procedures inefficient for large n. Some improvements can be found in Joachims (2006).

In the second data set values of houses in the area of Boston are compared (Frank and Asuncion, 2010). Thirteen features were measured, for instance the crime rate, the distance to five Boston employment centres or pupil-teacher ratio by town. Our results are contained in Table 2. We notice

Error	L(1)	L(10)	W(1)	W(10)	G(1)	G(10)
n=100	0,153	0,157	0,148	0,153	0,133	0,132
n=300	0,132	0,133	-	-	0,107	0,123

Table 2: Boston housing data

an improvement of every procedure in recognizing the ordering. Again G(1) and G(10) have least errors. In this case estimators obtained for greater value of the parameter λ (except for G(1)) are better.

Last two experiments are carried out on data sets concerning the quality of red and white wine (Cortez et al., 2009). In both cases one measured 11 features such that the content of alcohol, citric acid, the density and pH. The quality of a wine was determined by wine experts. Results in Table

Red	L(1)	L(10)	W(1)	W(10)	G(1)	G(10)
n=100	0,226	0,227	0,281	0,271	0,257	0,285
n=300	0,214	0,216	-	-	0,232	0,270
White						
n=100	0,265	0,266	0,292	-	0,282	0,305
n=300	0,253	0,249	-	-	0,268	0,303

Table 3: Wine quality

3 indicate lower efficiency of procedures than in previous examples. For red wine as well as white one we can notice the advantage of SVM with linear kernels, whose errors are very similar. The worst algorithm is G(10) which in previous experiments has one of the least error.

Acknowledgments

The research for this paper was partially supported by the grant of Ministry of Science and Higher Education no. N N201 391237.

References

R. Adamczak. Moment inequalities for U-statistics. *Ann. Probab.*, 34:2288–2314, 2007.

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

M. A. Arcones and E. Giné. Limit theorems for U-processes. *Ann. Probab.*, 21:1494–1542, 1993.

M. A. Arcones and E. Giné. 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, 1994.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Ann. Statist.*, 33: 1497–1537, 2005.

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

G. Blanchard, G. Lugosi, and N. Vayatis. On the rates of convergence of regularized boosting classifiers. *J. Machine Learning Research*, 4:861–894, 2003.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Ann. Statist.*, 36:489–531, 2008.

A. Bose. Bahadur representation of M_m estimates. *Ann. Statist.*, 26:771–777, 1998.

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

C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and scoring using empirical risk minimization. In *Proceedings of COLT*, pages 1–15, 2005.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. *Ann. Statist.*, 36:844–874, 2008.

C. Cortes and V. N. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

P. Cortez, A. Cerdeira, F. Almeida, F. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47:547–553, 2009.

D. Cossock and T. Zhang. Subset ranking using regression. In *Proceedings of the 19th Annual Conference on Learning Theory*, 2006.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

V. H. de la Peña and E. Giné. *Decoupling: From Dependence to Independence*. Springer-Verlag, New York, 1999.

E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. *e1071: Misc Functions of the Department of Statistics (e1071)*. TU Wien, 2010. URL `http://CRAN.R-project.org/package=e1071`.

A. Frank and A. Asuncion. *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences, 2010. URL `http://archive.ics.uci.edu/ml`.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. of Computer and System Sciences*, 55:119–139, 1997.

Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *J. Machine Learning Research*, 4:933–969, 2004.

R. Herbrich, T. Graepel, and K. Obermayer. *Large Margin Rank Boundaries for Ordinal Regression*, chapter 7. Advanced Large Margin Classifiers. MIT Press, Cambridge, 2000.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, 19: 293–325, 1948.

T. Joachims. Training linear svms in linear time. In *Proceedings of the ACM KDD*, pages 217–226, 2006.

G. Lugosi and N. Vayatis. On the bayes-risk consistency of regularized boosting methods. *Ann. Statist.*, 32:30–55, 2004.

P. Major. An estimate of the supremum of a nice class of stochastic integrals and U-statistics. *Probab. Theory Related Fields*, 134:489–537, 2006.

P. Massart. Some applications of concentration inequalities to statistics. *Probability theory. Ann. Fac. Sci. Toulouse Math.*, 9:245–303, 2000.

S. Mendelson. Improving the sample complexity using global data. *IEEE Trans. Inform. Theory*, 48:1977–1991, 2002.

S. Mendelson. *A Few Notes on Statistical Learning Theory*, chapter 1. Advanced Lectures in Machine Learning. Springer, 2003.

W. Niemiro. Asymptotics for M-estimators defined by convex minimization. *Ann. Statist.*, 20: 1514–1533, 1992.

W. Niemiro and W. Rejchel. Rank correlation estimators and their limiting distributions. *Statistical Papers*, 50:887–893, 2009.

D. Nolan and D. Pollard. U-processes: rates of convergence. *Ann. Statist.*, 15:780–799, 1987.

A. Pakes and D. Pollard. Simulation and asymptotics of optimization estimators. *Econometrica*, 57:1027–1057, 1989.

D. Pollard. *Convergence of Stochastic Processes*. Springer, New York, 1984.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`.

W. Rejchel. Ranking - convex risk minimization. In *Proceedings of WASET*, pages 172–178, 2009.

C. Rudin. Ranking with a p-norm push. In *In Proceedings of the 19th Annual Conference on Learning Theory*, 2006.

C. Scovel and I. Steinwart. Fast rates for support vector machines using Gaussian kernels, 2005. Preprint.

C. Scovel and I. Steinwart. Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, 35:575–607, 2007.

R. P. Sherman. The limiting distributions of the maximum rank correlation estimator. *Econometrica*, 61:123–137, 1993.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. of Machine Learning Research*, 2:67–93, 2001.

M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, 22:28–76, 1994.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Verlag, New York, 1996.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

I. C. Yeh. Modeling of strength of high performance concrete using artificial neural networks. *Cement and Concrete Research*, 28:1797–1808, 1998.