International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:8, 2009

Ranking - convex risk minimization

Wojciech Rejchel

Abstract—The problem of ranking (rank regression) has become popular in the machine learning community. This theory relates to problems, in which one has to predict (guess) the order between objects on the basis of vectors describing their observed features. In many ranking algorithms a convex loss function is used instead of the $0-1$ loss. It makes these procedures computationally efficient. Hence, convex risk minimizers and their statistical properties are investigated in this paper. Fast rates of convergence are obtained under conditions, that look similarly to the ones from the classification theory. Methods used in this paper come from the theory of U-processes as well as empirical processes.

Keywords—convex loss function, empirical risk minimization, empirical process, U-process, boosting, euclidean family

I. Introduction

THE problem of ranking has become an interesting field for researchers in machine learning community. There are a few reasons for that, but, undoubtedly, the most important ones are applications of this theory in many branches of economy - whenever one compares some objects (products). For instance, a financial institution can be interested in comparing credit risks of its clients or a department of quality control in a factory can use results of this theory to indicate which machine damages earlier. Other important applications contain survival analysis or information retrieval, when one can compare documents (web pages) by the degree of relevance for a given request.

In ranking, in a nutshell, one wants to predict (guess) the order between objects on the basis of their observed features. This problem is closely related to the well-known part of machine learning - the classification theory [1], [2]. The statistical framework of ranking, that is considered here, is similar to [3]. Empirical risk minimizers, i.e. machine learning analog of M-estimators, are investigated in this paper. Since the empirical risk is a U-statistic, the theory of U-statistics (U-processes) plays a significant role in argumentation. Important facts about these objects can be found in the monographs of de la Pena [4] or Serfling [5].

At the beginning, a development of ranking (similarly to the classification theory) was slowed down by problems in applications, since one minimized the discontinuous criterion function. It created serious computational problems and was not effective [6], [7]. To overcome these problems the discontinuous loss function was replaced by a convex surrogate designed to serve a similar purpose. This trick allowed to invent effective algorithms, such as support vector machines [8], [9] or boosting [10], [11]. Except algorithmic details, statistical properties of convex risk minimizers should be better recognized - that is one of the main aims of this paper. In the paper [12] the attention was focused on linear ranking rules and conditions for the strong consistency and the asymptotic normality of convex risk minimizers were found. On the other hand, confidence intervals for the risk and the excess risk of empirically chosen ranking rules were developed in [3]. They obtained bounds for both convex risks of the order $\frac{1}{\sqrt{n}}$. It has been already mentioned, that ranking and the classification have much in common. It was noticed in the latter theory, that there are some conditions standing behind an improvement of the rate of convergence even to $\frac{1}{n}$. Namely, these assumptions concern controlling the variance of the risk by its expected value. They were obtained using "low noise" requirements [13]. And this is also the goal of this paper - developing new probabilistic inequalities for ranking, which indicate that one can get better rates than $\frac{1}{\sqrt{n}}$.

The importance of the theory of U-statistics and U-processes should be emphasized. Indeed, the empirical risk, as a U-statistic, can be split into a sum of iid random variables and a degenerate U-process. For the latter term there are some exponential inequalities [14], [15], which allow to bound this component by $\frac{1}{n}$. However, the first (empirical) element in Hoeffding decomposition of U-statistics is the leading one. It is well-known (for instance [1] or [16]) that it can be bounded by $\frac{1}{\sqrt{n}}$ (with respect to the constant), if one uses standard methods such as the bounded differences inequality, symmetrization lemma or contraction principle. With additional, but reasonable (comparing to the classification theory) conditions and basing on more powerful tools (Tallagrand inequality in place of the bounded differences one), one can obtain faster convergence, with rates up to $\frac{1}{n}$ (in fact, up to $\frac{\ln n}{n}$). Methods used in the classification theory [17], [18] were the core for studies of the empirical term. They has been transferred to the field of ranking.

The paper is organized as follows: in the next section the statistical framework of ranking is introduced. The third part is devoted to empirical terms in Hoeffding decomposition in the case of the risk of ranking rules and the excess risk. In section four tools, that are helpful to handle a degenerate U-process, are developed. Main results can be found in the fifth section, which contains also applications of proved theorems to commonly used algorithms, for instance boosting.

II. Problem of ranking

First, it is assumed that there are two objects $Z = (X, Y)$ and $Z' = (X', Y')$, which take values in $\mathcal{X} \times \mathbb{R}$, where $\mathcal{X} \subset \mathbb{R}^d$, and they are independent and identically distributed (according to the distribution P) random vectors. X and X' are considered as vectors describing observed or measured

W. Rejchel is with the Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Torun, Poland. e-mail: iggypop@mat.uni.torun.pl

features of the objects, while Y and Y' are their unseen labels. By the definition, the object Z is "better" than Z' if $Y > Y'$. The task is to predict (guess) the order between instances in the best possible way. To do it, one constructs functions (ranking rules) $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which guess the order between objects in the following way

$$\text{if } \ f(x, x') > 0 \text{ then } \ y > y'.$$

The quality of ranking rules is measured by the probability of incorrect ranking, which is defined as

$$L(f) = P(\,\mathrm{sgn}(Y - Y')f(X, X') < 0), \tag{1}$$

so the task is to find the ranking rule minimizing (1) in $f \in \mathcal{F}$, where $\mathcal{F}$ is a family of ranking rules.

Moreover, let $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$, be a learning sample, that is a collection of independent and identically distributed (also according to P) random vectors for which the ordering of components Y_i is observable. The sample analog of (1)

$$L_n(f) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}[\mathrm{sgn}(Y_i - Y_j) f(\mathbf{X}_i, \mathbf{X}_j) < 0]$$

can be considered and one wants to find the ranking rule, which minimizes this expression. This approach to the ranking problem is hardly used, because the function $L_n(f)$, that is to minimize, is discontinuous, so the task is computationally difficult. This inconvenience can be overcome using instead of the $0-1$ loss function its convex surrogate.

Therefore, let ψ be a convex, nonnegative real function, which bounds the $0-1$ loss from above. Denote the convex risk of a ranking rule f by

$$A(f) = \mathbf{E}\,\psi[\,\mathrm{sgn}(Y - Y')\,f(X, X')]. \tag{2}$$

Now one looks for a minimizer of (2) in $f \in \mathcal{F}$. The typical approach to this problem is to find a minimizer of an empirical analog of (2) of the form

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

where $\psi_f(z, z')$ denotes $\psi[\mathrm{sgn}(y - y')\ f(x, x')]$. Note that $A_n(f)$ is, for each f, a U-statistic of order two, so properties of U-processes should be used to study the performance of $f_n = \mathrm{argmin}_f A_n(f)$. Studies on the quality of f_n will go twofold. On the one hand, bounds for the risk of f_n in terms of its empirical analog are looked for. On the second hand, the excess risk, i.e. when one compares the risk of f_n to the risk of the best ranking rule in the class, is considered.

First, it can be assumed, for simplicity, that the class $\mathcal{F}$ is uniformly bounded, so for every $x, x' \in \mathcal{X}$ and $f \in \mathcal{F}$ one has $|f(x, x')| \leq A_1$ for some constant $A_1 > 0$. Moreover, it is considered here and in the sequel, for convenience, that $f(x, x') = -f(x', x)$, which implies that the kernel of $A_n(f)$ is symmetric. The first tool, that is needed, is Hoeffding decomposition of U-statistics [5]:

$$A(f) - A_n(f) = 2P_n[A(f) - P\,\psi_f] - U_n(h_f),$$

where

$$\begin{aligned} P\,\psi_f(z) &= \mathbf{E}[\psi_f(Z, Z')|Z = z], \\ P_n(g) &= \frac{1}{n}\sum_{i=1}^{n} g(Z_i), \\ h_f(Z_1, Z_2) &= \psi_f(Z_1, Z_2) - P\psi_f(Z_1) - P\psi_f(Z_2) + A(f) \\ U_n(h_f) &= \frac{1}{n(n-1)} \sum_{i \neq j} h_f(Z_i, Z_j). \end{aligned}$$

Hoeffding decomposition breaks a U-statistic into the sum of iid random variables and a degenerate U-statistic $U_n(h_f)$. The degeneration of a U-statistic means that the conditional expectation of its kernel is the zero-function. Using exponential inequalities one can bound the U-process $U_n(h_f)$ by $\frac{1}{n}$ with high probability. Moreover, some methods from the classification theory will be borrowed to develop probabilistic inequalities for the empirical term with proper rates.

Now the excess risk of a ranking rule, i.e. $A(f_n) - \inf_{f \in \mathcal{F}} A(f)$, is considered. One more assumption will be made: there exists $f^* \in \mathcal{F}$ satisfying $A(f^*) = \inf_{f \in \mathcal{F}} A(f)$. This condition can be weakened, because one can consider functions that are only close to achieving minimum. This generalization is not very hard, but makes results less transparent. Hoeffding decomposition can be applied to the U-statistic $A_n(f) - A_n(f^*)$:

$$\begin{aligned} A(f) - A(f^*) - [A_n(f) - A_n(f^*)] = \\ 2P_n[A(f) - A(f^*) - P\psi_f + P\psi_{f^*}] - [U_n(h_f) - U_n(h_{f^*})]. \end{aligned} \tag{3}$$

The procedure is similar to the one in the previous case.

III. Empirical term

In order to obtain the proper rate for the empirical term in Hoeffding decomposition one has to assume that variances of functions from an adequate family are upper bounded by the linear transformation of their expectations. The general theorem for empirical processes is presented and applied to the appropriate class of functions. At the beginning, one needs a few preliminaries - the first one refers to the sub-root function:

Definition 1: A function $\phi : [0, \infty) \to [0, \infty)$ is a sub-root function if it is nonnegative, non-decreasing and for each $r > 0$ the function $r \longmapsto \phi(r)/\sqrt{r}$ is non-increasing.

There are a lot of useful properties of sub-root functions, for example they are continuous and have the unique fixed point r^* (the positive solution of the equation $\phi(r) = r$). Proofs of these facts can be easily find in the literature [1], [17].

Thus, let $\mathcal{G} = \{g : \mathcal{X} \to \mathbb{R}\}$ be a class of real functions. Let two iid sequences be given: $Z_1, \ldots, Z_n$ as before and $\sigma_1, \ldots, \sigma_n$ such that $P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2$ for $i = 1, \ldots, n$ (Z's and σ's are also independent). Rademacher average of a class $\mathcal{G}$ is defined as

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

where the expectation is taken with respect to both samples. Finally, let Pg denote the expectation of g, so $Pg = \mathbf{E}g(Z)$. The above-mentioned theorem, which can be also found for instance in [17] or [19], is stated:

Theorem 2: Let the class $\mathcal{G}$ of functions with ranges in $[a,b]$ be such that $Pg^2 \leq B\,Pg$ for some constant B and every $g \in \mathcal{G}$. Moreover, if there exists a sub-root function ϕ with the fixed point r^*, which satisfies

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

for every $r \geq r^*$, then for every $K > 1$ and $x > 0$, with probability at least $1 - e^{-x}$

$$\forall g \in \mathcal{G} \quad Pg \leq \frac{K}{K-1}P_n(g) + \frac{6K}{B}r^* + [11(b-a)+5BK]\frac{x}{n}.$$

The class $\mathcal{G}^*$ is just a star-hull of $\mathcal{G}$, i.e.

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

Here Rademacher average of $\mathcal{G}^*$ is used, but the similar theorem can be proved without this modification. Of course, making a class star-shaped increases it, but as it will be seen later it is not significant, for instance in the sense of covering numbers, which is essential in this paper.

Theorem 2 says that to get proper bounds for the empirical term one needs to find good bounds for the fixed point r^* of a sub-root ϕ, but there is no general method for choosing ϕ. In this paper $\mathbf{E}R_n(g \in \mathcal{G}^* : Pg^2 \leq r)$ is used as a sub-root. The nonnegativity of this function is obvious. Moreover, it is non-decreasing, because it is enough to notice that $\{g : Pg^2 \leq r\} \subset \{g : Pg^2 \leq r'\}$ for $r \leq r'$. The last property from the definition of a sub-root function is also not difficult. Using the above-mentioned sub-root one can show that the rate of the empirical process can be essentially better than $\frac{1}{\sqrt{n}}$. Namely, the fixed point of such chosen sub-root function will be bounded by $\frac{\ln n}{n}$ and this is the best known order [20], [21].

As it has been already mentioned, two cases are considered: the first one relies on rewarding bounds for the 'typical' risk $A(f)$. The second one refers to the excess risk, i.e. the difference between the risks $A(f)$ and $A(f^*)$. The former situation is quite easy, since the appropriate class is nonnegative and bounded. This trivially implies the relation between the variance and the expectation of its elements. The latter case is more complicated, since the class of functions does not have to be nonnegative. So one needs more effort to get proper bounds. The convexity of a loss function plays a crucial role in this investigation.

Let the class

$$P\psi_{\mathcal{F}} = \{P\psi_f \ :\ f \in \mathcal{F}\}$$

be considered, where $\mathcal{F}$ consists of uniformly bounded functions and ψ is a convex, nonnegative loss function (so ψ is also locally Lipschitz with respect to the euclidean distance on the real line). There is one more assumption on the class $\mathcal{F}$, namely it is euclidean. This condition is not very restrictive and often fulfilled by commonly used classes (for example a class with finite Vapnik-Chervonenkis dimension or a family used by boosting). This property of the class $\mathcal{F}$ allows to bound its covering number in a "nice" way. Its definition is introduced below. More details on euclidean families may be found in [22].

For every probability measure Q on $\mathcal{X} \times \mathcal{X}$ one can define the $\mathbb{L}^2$-distance on $\mathcal{F}$ by

$$d(f,g) = \sqrt{\int |f-g|^2 dQ}.$$

The covering number $\mathcal{N}(\varepsilon, \mathcal{F}, d)$ of the class $\mathcal{F}$ with radius $\varepsilon > 0$ and with respect to d is the smallest cardinality of a subclass $\mathcal{F}_1$ of $\mathcal{F}$ such that for every $f \in \mathcal{F}$ one can find $f_1 \in \mathcal{F}_1$ for which $d(f, f_1) \leq \varepsilon$. Thus, the definition of a euclidean class can be written:

Definition 3: The class $\mathcal{F}$ is euclidean, if there exist positive constants A and V with the property: for every $\varepsilon > 0$ and a probability measure Q

$$\mathcal{N}(\varepsilon, \mathcal{F}, d) \leq A\varepsilon^{-V}.$$

The constants must not depend on Q.

Now one can state the theorem concerning the empirical term, when the attention is focused on the risk $A(f)$:

Theorem 4: For every $K > 1$ and $x > 0$, with probability at least $1 - 2e^{-x}$

$$\forall f \in \mathcal{F}\ \ A(f) \leq \frac{K}{K-1}\,P_n(P\psi_f) + D\frac{\ln n}{n} + E\frac{x}{n}\,,$$

where $\mathcal{F}$ is a euclidean class of functions and D and E are constants.

Remark 5: The precise forms of constants D and E from Theorem 4 and Theorem 6 (below) can be given. They are not introduced here since the order of bound is more important in this paper.

Proof: The class $P\psi_{\mathcal{F}}$ is nonnegative and bounded, because ψ is nonnegative and convex, and $\mathcal{F}$ is uniformly bounded. Let the constant A_2 be a suitable bound for $P\psi_{\mathcal{F}}$, so $|P\psi_f(z)| \leq A_2$ for all $f \in \mathcal{F}$ and $z \in \mathcal{X} \times \mathbb{R}$. Using these two properties

$$\mathbf{E}\ (P\psi_f(Z))^2 = A_2^2\,\mathbf{E}\left(\frac{P\psi_f(Z)}{A_2}\right)^2$$
$$\leq \quad A_2^2\,\mathbf{E}\left(\frac{P\psi_f(Z)}{A_2}\right) \leq A_2\,\mathbf{E}(P\psi_f(Z)),$$

which confirms that there exists the wanted relation between the variance and the expectation. Theorem 2 can be used for the family $P\psi_{\mathcal{F}}$ and the sub-root function

$$\phi(r) = 10A_2\,\mathbf{E}R_n(h \in (P\psi_{\mathcal{F}})^* : Ph^2 \leq r) + 11A_2^2\frac{x}{n},$$

where x is an arbitrary positive number. Of course, ϕ is a sub-root, because it is a modification of the sub-root function $\mathbf{E}R_n(h \in (P\psi_{\mathcal{F}})^* : Ph^2 \leq r)$. The sub-root function ϕ, which is chosen in this proof, depends on n, so its fixed point too. From Theorem 2 the probabilistic inequality (with proper constants D and E) is deduced

$$A(f) \leq \frac{K}{K-1}\,P_n(P\psi_f) + Dr^* + E\frac{x}{n}$$

for every $f \in \mathcal{F}$. To finish the proof the fixed point r^* of the sub-root ϕ should be bounded by a term of the order $\frac{\ln n}{n}$. Similar methods to presented here can be found in [17] or

[23]. Using Corollary 2.2 from the former paper one has with probability at least $1-e^{-x}$

$$\{h \in (P\psi_{\mathcal{F}})^* : Ph^2 \le r^*\} \subset \{h \in (P\psi_{\mathcal{F}})^* : P_n h^2 \le 2r^*\},$$

since

$$\begin{aligned} & \mathbf{E}R_n(h \in (P\psi_{\mathcal{F}})^* : Ph^2 \le r^*) \\ \le \quad & \mathbf{E}R_n(h \in (P\psi_{\mathcal{F}})^* : P_n h^2 \le 2r^*), \end{aligned}$$

so

$$r^* \le 10A_2\, \mathbf{E}R_n(h \in (P\psi_{\mathcal{F}})^* : P_n h^2 \le 2r^*) + 11A_2^2 \frac{x}{n}.$$

Now Chaining Lemma for empirical processes [24] can be used to get the following inequality

$$\begin{aligned} & \mathbf{E}R_n(h \in (P\psi_{\mathcal{F}})^* : P_n h^2 \le 2r^*) \\ \le \quad & \frac{C}{\sqrt{n}}\, \mathbf{E} \int_0^{\sqrt{2r^*}} \sqrt{\ln \mathcal{N}(\varepsilon, (P\psi_{\mathcal{F}})^*, L^2(P_n))}\, d\varepsilon \end{aligned}$$

for some constant C. It is straightforward that having an $\varepsilon/2$ - cover of the family $P\psi_{\mathcal{F}}$ and an $\varepsilon/2A_2$ - cover of the interval [0,1] one can create an ε-cover of the star-hull of $P\psi_{\mathcal{F}}$ and

$$\mathcal{N}(\varepsilon, (P\psi_{\mathcal{F}})^*, L^2(P_n)) \le \mathcal{N}(\varepsilon/2, P\psi_{\mathcal{F}}, L^2(P_n))\lceil 2A_2/\varepsilon \rceil.$$

$\mathcal{F}$ is euclidean and ψ is locally Lipschitz, hence $P\psi_{\mathcal{F}}$ is also euclidean, so there exist constants A and V for which

$$\mathcal{N}(\varepsilon/2, P\psi_{\mathcal{F}}, L^2(P_n)) \le A \left(\frac{\varepsilon}{2}\right)^{-V}.$$

Therefore, with constants C and D, that may change from line to line

$$\begin{aligned} & \frac{C}{\sqrt{n}}\, \mathbf{E} \int_0^{\sqrt{2r^*}} \sqrt{\ln \mathcal{N}(\varepsilon, (P\psi_{\mathcal{F}})^*, L^2(P_n))}\, d\varepsilon \\ \le \quad & \frac{C}{\sqrt{n}} \int_0^{\sqrt{2r^*}} \sqrt{\ln \frac{D}{\varepsilon}}\, d\varepsilon \le \frac{C}{\sqrt{n}} \sqrt{r^* \ln \frac{D}{r^*}}. \end{aligned}$$

In the last inequality Lemma 3.8 from [23] is used. Joining together above results one has

$$r^* \le C\sqrt{\frac{\ln n}{n}}\sqrt{r^*} + D\frac{x}{n},$$

which easily implies

$$r^* \le C\frac{\ln n}{n} + D\frac{x}{n}.$$

■

Now the excess risk of a ranking rule is considered. If one looks at Hoeffding decomposition (3), then Theorem 2 can be used to handle its empirical term. Obviously, the class

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

is bounded, because $|P\,\psi_f(z) - P\,\psi_{f^*}(z)| \le 2L_\psi A_1$ with L_ψ being the Lipschitz constant of ψ. But there is one more assumption in Theorem 2, which concerns variances of functions from the family $P\psi_{\mathcal{F}} - P\psi_{f^*}$. The following inequality is needed: for every $f \in \mathcal{F}$

$$Var[P\psi_f(Z) - P\psi_{f^*}(Z)] \le B[A(f) - A(f^*)] \qquad (4)$$

for some constant B. The class, that is considered, need not to be nonnegative, so the reasoning as in the previous case fails. It will be shown in the subsection III.A that with more assumptions (but not too restrictive) this condition is also satisfied. The following theorem can be stated:

Theorem 6: For every $K > 1$ and $x > 0$, with probability at least $1 - 2e^{-x}$ for each $f \in \mathcal{F}$

$$A(f) - A(f^*) \le \frac{K}{K-1}\, P_n(P\psi_f - P\psi_{f^*}) + D\frac{\ln n}{n} + E\frac{x}{n},$$

where $\mathcal{F}$ is a euclidean class of functions, that satisfies (4) and D and E are constants.

Proof: The idea is the same as in the proof of Theorem 4. One just takes a slightly different sub-root function

$$\begin{aligned} \phi(r) = {} & 20L_\psi A_1 B\, \mathbf{E}R_n(h \in (P\psi_{\mathcal{F}} - P\psi_{f^*})^* : Ph^2 \le r) \\ + \quad & 44L_\psi^2 A_1^2 \frac{x}{n} \end{aligned}$$

and bounds its unique fixed point. Since $\mathcal{F}$ is euclidean and ψ is locally Lipschitz, then the family $P\,\psi_{\mathcal{F}} - P\,\psi_{f^*}$ also has this property. ■

A. To bound variance by expectation

Theorem 6 in Section III shows that fast rates can be obtained if one can relate the variance of elements from $P\psi_{\mathcal{F}} - P\psi_{f^*}$ to their expectation. In this subsection conditions, that are sufficient for even stronger relationship

$$Var[\psi_f(Z,Z') - \psi_{f^*}(Z,Z')] \le B[A(f) - A(f^*)],$$

are found. Moreover, these conditions are not very restrictive and, as it will be seen later, they will be satisfied in the most interesting situations.

The key object in further analysis - the modulus of convexity of the function ψ - will be defined. It is known that this function is very helpful if one wants to show the similar relation in the classification theory [18], [21]. With minor changes one can use the modulus of convexity in the ranking setting.

Definition 7: The modulus of convexity of ψ is the function $\delta : [0,\infty) \to [0,\infty]$ defined as

$$\delta(\varepsilon) = \inf\left\{\frac{\psi(x)+\psi(y)}{2} - \psi\left(\frac{x+y}{2}\right) : |x-y| \ge \varepsilon\right\}.$$

If one looks at the convex risk $A(f)$ of the ranking rule f as a functional $A : \mathcal{F} \to \mathbb{R}$ and the class $\mathcal{F}$ is convex, then the functional A is also convex. It allows to consider the modulus of convexity of it, given by

$$\tilde{\delta}(\varepsilon) = \inf\left\{\frac{A(f)+A(g)}{2} - A\left(\frac{f+g}{2}\right) : d(f,g) \ge \varepsilon\right\},$$

where d is the $\mathbb{L}^2$-distance for $f, g \in \mathcal{F}$, so

$$d(f,g) = \sqrt{\mathbf{E}[f(X,X') - g(X,X')]^2}.$$

The key property of the modulus of convexity is the fact that it can be often lower bounded by $C\varepsilon^r$, for some $C, r > 0$. This is satisfied for a large family of convex loss functions, for instance e^{-x}, $\ln(1+\mathrm{e}^{-2x})$ or $(1-x)_+^2$ (the last case needs just minor changes in consideration). This property implies

the similar one for the modulus of the functional A, which is sufficient for proving the relationship between the variance and the expectation for functions from the family $P\psi_{\mathcal{F}} - P\psi_{f^*}$. The following lemma, which has its origin in [18], can be stated:

Lemma 8: If there exist constants $C, r > 0$ for which the modulus of convexity of ψ satisfies

$$\delta(\varepsilon) \geq C\varepsilon^r, \tag{5}$$

then

$$Var[\psi_f(Z,Z') - \psi_{f^*}(Z,Z')] \leq L_\psi^2 D_r [A(f) - A(f^*)]^{\min(1,2/r)}_,$$

where

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

Proof: First, the variance of functions of the form $\psi_f - \psi_{f^*}$ can be easily bounded if one uses the Lipschitz property of ψ:

$$\begin{aligned} & Var[\psi_f(Z,Z') - \psi_{f^*}(Z,Z')] \\ \leq\ & \mathbf{E}[\psi_f(Z,Z') - \psi_{f^*}(Z,Z')]^2 \\ \leq\ & L_\psi^2 \mathbf{E}[\mathrm{sgn}(Y-Y')f(X,X') - \mathrm{sgn}(Y-Y')f^*(X,X')]^2 \\ =\ & L_\psi^2 d^2(f,f^*). \end{aligned} \tag{6}$$

The second step of the proof relies on showing that if the modulus δ satisfies (5), then its analogue $\tilde{\delta}$ also fulfills a similar condition. Namely, let $f, g \in \mathcal{F}$ satisfy $d(f,g) \geq \varepsilon$. Then

$$\begin{aligned} & \frac{A(f)+A(g)}{2} - A\left(\frac{f+g}{2}\right) \\ =\ & \mathbf{E}\left[\frac{\psi_f(Z,Z') + \psi_g(Z,Z')}{2} - \psi_{\frac{f+g}{2}}(Z,Z')\right] \\ \geq\ & \mathbf{E}\,\delta(|\mathrm{sgn}(Y-Y')f(X,X') - \mathrm{sgn}(Y-Y')g(X,X')|) \\ =\ & \mathbf{E}\,\delta(|f(X,X') - g(X,X')|) \\ \geq\ & C\mathbf{E}\,[f(X,X') - g(X,X')]^r. \end{aligned}$$

This and easy calculation (as in the proof of Lemma 15 in [18]) show that the modulus of convexity of the functional A fulfills

$$\tilde{\delta}(\varepsilon) \geq C_r \varepsilon^{max(2,r)}\,, \tag{7}$$

where $C_r = C$ for $r \geq 2$ and $C_r = C(2A_1)^{r-2}$ otherwise. Moreover, from the definition of the modulus of convexity and the fact that f^* is the minimizer of $A(f)$ in the convex class $\mathcal{F}$:

$$\begin{aligned} \frac{A(f)+A(f^*)}{2} &\geq A\left(\frac{f+f^*}{2}\right) + \tilde{\delta}(d(f,f^*)) \\ &\geq A(f^*) + \tilde{\delta}(d(f,f^*)). \end{aligned}$$

The bound for the variance (6) and the property (7) of the modulus $\tilde{\delta}$ imply

$$\begin{aligned} & A(f) - A(f^*) \geq 2\tilde{\delta}\left(\frac{\sqrt{Var\,[\psi_f(Z,Z') - \psi_{f^*}(Z,Z')]}}{L_\psi}\right) \\ \geq\ & 2C_r \left(\frac{\sqrt{Var\,[\psi_f(Z,Z') - \psi_{f^*}(Z,Z')]}}{L_\psi}\right)^{\max(2,r)}_, \end{aligned}$$

which is equivalent to

$$Var[\psi_f(Z,Z') - \psi_{f^*}(Z,Z')] \leq L_\psi^2 D_r [A(f) - A(f^*)]^{\min(1,2/r)}_.$$

■

Thus, for loss functions that were mentioned at the beginning of this subsection one obtains the exponent in the above inequality equal to 1, because their modulus of convexity can be easily lower bounded with $r = 2$.

IV. Degenerate U-processes

In this section the method, which allows to get exponential inequalities for degenerate U-processes, is presented. These inequalities ensure that the second term in Hoeffding decomposition can be upper bounded by $\frac{1}{n}$, which is the right order to get better bounds for the risk and the excess risk of ranking rules.

First, a general U-process

$$\left\{ U_n(h) = \frac{1}{n(n-1)} \sum_{i \neq j} h(Z_i, Z_j) \ : \ h \in \mathcal{H} \right\}, \tag{8}$$

is considered. $\mathcal{H}$ is a euclidean family of uniformly bounded, degenerate and symmetric functions. At the end of this section it will be specialized to the case of U-processes, that are demanded in Hoeffding decompositions described in Section II.

Theorem 9: If we consider the U-process (8), then for each $x > 0$, with probability at least $1 - C_1 \exp(-x)$

$$\sup_{h \in \mathcal{H}} |U_n(h)| \leq \frac{C_2 x}{n}\,,$$

where C_1 and C_2 are positive constants.

Proof: Let λ be a positive number. Symmetrization Lemma [4] and formulas 3.5 and 3.4 from [25] imply

$$\begin{aligned} & \mathbf{E}\exp\left(\lambda \sqrt{\sup_{h \in \mathcal{H}} |(n-1)U_n(h)|}\right) \\ \leq\ & C_1 \mathbf{E} \exp\left(C_2 \lambda^2\, \mathbf{E}_\sigma \sup_{h \in \mathcal{H}} |(n-1)S_n(h)|\right). \end{aligned} \tag{9}$$

The notation $S_n(h) = \frac{1}{n(n-1)} \sum_{i \neq j} \sigma_i \sigma_j h(\mathbf{Z}_i, \mathbf{Z}_j)$ has been used in (9) and $\mathbf{E}_\sigma$ is the conditional expectation with respect to Rademacher variables. It should be indicated that the constant C_1, C_2 and other, that will appear in this proof, may differ from line to line. Therefore, the problem is reduced to finding bounds for $\mathbf{E}_\sigma \sup_{h \in \mathcal{H}} |S_n(h)|$. This step relies on Chaining Lemma ([26], Lemma 5) applied to U-processes. Namely, this lemma is used for the stochastic process

$$\{J_n(h) = \frac{1}{n} \sum_{i \neq j} \sigma_i \sigma_j h(Z_i, Z_j) : h \in \mathcal{H}\}.$$

As it has been already said the sample $Z_1, \ldots, Z_n$ is fixed now so one can denote $h(Z_i, Z_j) = h_{ij}$ in this case. Furthermore, let the pseudo-metric d on the family $\mathcal{H}$ be defined as

$$d(h,g) = \sqrt{\frac{1}{n(n-1)} \sum_{i \neq j} (h_{ij} - g_{ij})^2}\,.$$

For this stochastic process assumptions of Chaining Lemma are satisfied with the function ϕ given by $\exp(\frac{x}{s}-1)$, where s is an appropriate constant. The first and the fourth condition in Chaining Lemma are obviously fulfilled. Besides, it can be assumed that $0 \in \mathcal{H}$, which easily implies the third assumption, since the family $\mathcal{H}$ is uniformly bounded. The whole difficulty lies in the second condition. But using Corollary 3.2.6 from [4] one finds s such that

$$\mathbf{E}_\sigma \exp\left(\frac{|J_n(h-g)|}{s\sqrt{\mathbf{E}_\sigma[J_n(h-g)]^2}}\right) \leq e.$$

Furthermore, quick calculation shows that

$$\mathbf{E}_\sigma[J_n(h-g)]^2 = \frac{1}{n^2}\sum_{i\neq j}(h_{ij}-g_{ij})^2 \leq d^2(h,g),$$

which implies

$$\mathbf{E}_\sigma \phi\left(\frac{|J_n(h-g)|}{d(h,g)}\right) \leq 1.$$

Finally, using the thesis of Chaining Lemma

$$\mathbf{E}_\sigma \sup_{h\in\mathcal{H}}|S_n(h)| \leq \frac{C_1}{n-1}\int_0^H [s\ln\mathcal{N}(\varepsilon,\mathcal{H},d)+s]\,d\varepsilon,$$

where H is a bound of the class $\mathcal{H}$. Since the family $\mathcal{H}$ is euclidean, then for some constant C_3

$$\mathbf{E}_\sigma \sup_{h\in\mathcal{H}}|S_n h| \leq \frac{C_3}{n}\,. \tag{10}$$

With this inequality the right side of (9) may be bounded by $C_1\exp(C_2\lambda^2)$, because the right side of (10) does not depend on variables $Z_1,\ldots,Z_n$. This and Markov Inequality finish the proof. ■

The properties of the U-processes

$$\left\{\frac{1}{n(n-1)}\sum_{i\neq j}h_f(Z_i,Z_j)\ :\ f\in\mathcal{F}\right\} \tag{11}$$

should be studied, where

$$h_f(Z_i,Z_j) = \psi_f(Z_i,Z_j) - P\psi_f(Z_i) - P\psi_f(Z_j) + A(f).$$

One can use Theorem 9 to handle the U-process (11):

Corollary 10: For every $x>0$, with probability at least $1-C_1\exp(-x)$

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n(n-1)}\sum_{i\neq j}h_f(Z_i,Z_j)\right| \leq \frac{C_2 x}{n}\,,$$

where C_1 and C_2 are positive constants. The same probabilistic inequality, with respect to the constants, holds for the degenerate term in the case of the excess risk, i.e. when one takes $h_f - h_{f^*}$ instead of h_f.

Proof: Kernels of the U-process considered here are uniformly bounded (since $\mathcal{F}$ is uniformly bounded and ψ is convex), symmetric and degenerate. Furthermore, the class $\{h_f : f\in\mathcal{F}\}$ is euclidean, because it is a sum of euclidean classes and for families $\mathcal{H}_1$ and $\mathcal{H}_2$

$$\mathcal{N}(2\varepsilon,\mathcal{H}_1+\mathcal{H}_2,d) \leq \mathcal{N}(\varepsilon,\mathcal{H}_1,d)\,\mathcal{N}(\varepsilon,\mathcal{H}_2,d)$$

for $\mathcal{H}_1+\mathcal{H}_2 = \{h_1+h_2 : h_1\in\mathcal{H}_1, h_2\in\mathcal{H}_2\}$. ■

V. Conclusions: final results and examples

The task relied on showing that in ranking, similarly to the classification theory, the risk and the excess risk of ranking rules can be bounded with better rates than $\frac{1}{\sqrt{n}}$, which were proved in [3]. By Hoeffding decomposition the effort was divided into the empirical term (Section III) and the degenerate U-process (Section IV). If one takes results of these two parts together, then the following two theorems can be stated. Main assumptions about the class $\mathcal{F}$ and the function ψ are not repeated.

Theorem 11: If the class $\mathcal{F}$ of ranking rules is euclidean, then for each $x>0$, with probability at least $1-C\exp(-x)$

$$\forall f\in\mathcal{F}\quad A(f) \leq 3A_n(f) + D\left(\frac{\ln n + x}{n}\right),$$

where C and D are proper constants. In particular, for the ranking rule $f_n = \operatorname{argmin}_{f\in\mathcal{F}}A_n(f)$

$$A(f_n) \leq 3A_n(f_n) + D\left(\frac{\ln n + x}{n}\right).$$

Proof: Hoeffding decomposition of the U-statistic $A_n(f)$ can be slightly changed. Namely, for $K>2$

$$(K-2)A(f) - KA_n(f) = 2P_n[(K-1)A(f) - KP\psi_f] - KU_n(f).$$

Applying Theorem 4 and Corollary 10 finishes the proof. For simplicity K equals 3. ■

Theorem 12: If the class $\mathcal{F}$ of ranking rules is convex and euclidean, and the modulus of convexity $\delta(\varepsilon)$ of ψ is proportional to ε^2 on the interval $[-A_1,A_1]$, then for each $x>0$, with probability at least $1-C\exp(-x)$ for every $f\in\mathcal{F}$

$$A(f) - A(f^*) \leq 3[A_n(f) - A_n(f^*)] + D\left(\frac{\ln n + x}{n}\right), \tag{12}$$

where C and D are proper constants. In particular, for the ranking rule $f_n = \operatorname{argmin}_{f\in\mathcal{F}}A_n(f)$

$$A(f_n) - A(f^*) \leq D\left(\frac{\ln n + x}{n}\right). \tag{13}$$

Proof: First, similar changes as in the proof of Theorem 11 are needed in Hoeffding decomposition of the U-statistic $A_n(f) - A_n(f^*)$. Thus, Theorem 6, Lemma 8 and Corollary 10 are sufficient for the proof of (12). The fact that $A_n(f_n) - A_n(f) \leq 0$ for every $f\in\mathcal{F}$ implies (13). ■

Remark 13: The constant 3 before the first terms on the right side of inequalities in Theorems 11 and 12 can be decreased to the number close to 1, but in the same time the constant D would increase.

The natural problem arises: even if the rule f_n is very good in the convex case, is it accurate in the $0-1$ loss case? The relation between risks $L(f_n)$ and $A(f_n)$, and excess risks $L(f_n) - L(f^*)$ and $A(f_n) - A(f^*)$ should be found. The first situation is quite easy, since typically chosen convex loss functions are upper-bounds for the $0-1$ loss. So if $A(f_n)$ is small, then automatically $L(f_n)$ is also insignificant. The latter case is more complicated, because the excess convex risk does not need to bound the "primary" one. Results of

[18] are sufficient to formulate Theorems 11 and 12 in cases of true risks $L(f)$ and $L(f) - L(f^*)$.

Now two examples are described, where the theorems written above can be applied. The first one is very simple.

Example 14: Let

$$\mathcal{F} = \{f(x, x') = a^T(x - x') + b : a, x, x' \in \mathbb{R}^d, b \in \mathbb{R}\}$$

be a family of ranking rules. In this case the prediction of the order between objects depends on the hyperplane that the vector $x - x'$ belongs to. It is clear that $\mathcal{F}$ is euclidean and convex. If one takes a "good" convex function ψ, for instance one of the mentioned in Subsection III.A, then both Theorems 11 and 12 work. Thus, the rates of the order $\frac{1}{n}$ can be obtained for these ranking rules.

The second application concerns a very famous family of algorithms called "boosting". There are a lot of different versions of boosting, and here AdaBoost is considered, which uses the exponential loss function $\psi(x) = \exp(-x)$.

Example 15: A class $\mathcal{R}$ of binary functions with the finite Vapnik-Chervonenkis dimension is considered. The output of Adaboost is an element of a convex T-hull of $\mathcal{R}$, where T is the number of components (in fact iterations). Namely, it belongs to the family

$$conv_T(\mathcal{R}) = \{f(x, x') = \sum_{j=1}^{T} w_j r_j(x, x') : \sum_{j=1}^{T} w_j = 1 \quad , \\ 0 \leq w_j \leq 1, r_j \in \mathcal{R} \text{ for } j = 1, \ldots, T\} .$$

The class $conv_T(\mathcal{R})$ is obviously convex. Furthermore, for some constants A and V

$$\mathcal{N}(\varepsilon, conv_T(\mathcal{R}), d) \leq A\varepsilon^{-TV},$$

because the class $\mathcal{R}$ is euclidean. This inequality implies that the family $conv_T(\mathcal{R})$ is also euclidean. On the other hand, the modulus of convexity of the function $\psi(x) = \exp(-x)$ on the interval [-1,1] equals to $\varepsilon^2/8e$. By Theorems 11 and 12 one again observes faster rates than previously proved (compare to [3], Section 7). Of course, these two theorems do not only work for a fixed number of components. One can deduce without much effort uniform, in T, results using a union bound for probability.

REFERENCES

[1] S. Boucheron, O. Bousquet, G. Lugosi, *Theory of classification: a survey and recent advances*, Probability and Statistics, vol. 9, pp. 323-375, 2005.
[2] L. Devroye, L. Gyorfi, G. Lugosi, *A probabilistic theory of pattern recognition.* Springer-Verlag, New York, 1996.
[3] S. Clemenon, G. Lugosi, N. Vayatis, *Ranking and empirical minimization of U-statistics*, Ann. Statist., vol. 36, pp. 844-874, 2008.
[4] V. H. de la Pena, E. Gine, *Decoupling: from dependence to independence.* Springer-Verlag, New York, 1999.
[5] R. J. Serfling, *Approximation theorems of mathematical statistics.* Wiley, New York, 1980.
[6] J. Abrevaya, *Computation of the maximum rank correlation estimator*, Economics Letters, vol. 62, pp. 279-285, 1999.
[7] P. L. Bartlett, S. Ben-David, *Hardness results for neural network approximation problems*, Theoretical Computer Science, vol. 284, pp. 53-66, 2002.
[8] V. N. Vapnik, *Statistical learning theory.* John Wiley, New York, 1998.
[9] C. Cortes, V. N. Vapnik, *Support vector networks*, Machine Learning, vol. 20, pp. 273-297, 1995.
[10] R. E. Schapire, Y. Freund, P. L. Bartlett, W. S. Lee, *Boosting the margin: a new explanation for the effectiveness of voting methods*, Ann. Statist., vol. 26, pp. 1651-1686, 1998.
[11] Y. Freund, R. Iyer, R. E. Schapire, Y. Singer, *An efficient boosting algorithm for combining preferences*, J. Machine Learning Research, vol. 4, pp. 933-969, 2004.
[12] W. Niemiro, W. Rejchel, *Rank correlation estimators and their limiting distributions*, Statistical Papers, to be published.
[13] A. B. Tsybakov, *Optimal aggregation of classifiers in statistical learning*, Ann. Statist., vol. 32, pp. 135-166, 2004.
[14] M. A. Arcones, E. Gine (1993). *Limit theorems for U-processes.* Ann. Probab., vol. 21, pp. 1494-1542.
[15] P. Major, *An estimate of the supremum of a nice class of stochastic integrals and U-statistics*, Probab. Theory Related Fields, vol. 134, pp. 489-537, 2006.
[16] V. Koltchinskii, D. Panchenko, *Empirical margin distributions and bounding the generalization error of combined classifiers*, Ann. Statist., vol. 30, pp. 1-50, 2002.
[17] P. L. Bartlett, O. Bousquet, S. Mendelson, *Local Rademacher complexities*, Ann. Statist., vol. 33, pp. 1497-1537, 2005.
[18] P. L. Bartlett, M. I. Jordan, J. D. Mcauliffe, *Convexity, Classification, and Risk Bounds,* Journal of the American Statistical Association, vol. 101, pp. 138-156, 2006.
[19] P. Massart, *Some applications of concentration inequalities to statistics*, Probability theory. Ann. Fac. Sci. Toulouse Math., vol. 9 , pp. 245-303, 2000.
[20] P. Massart, *Concentration Inequalities and Model Selection.* Springer, Berlin, 2007.
[21] S. Mendelson, *Improving the sample complexity using global data*, IEEE Trans. Inform. Theory, vol. 48, pp. 1977-1991, 2002.
[22] A. Pakes, D. Pollard, *Simulation and the asymptotics of optimization estimators*, Econometrica, vol. 57, pp. 1027-1057, 1989.
[23] S. Mendelson, *A few notes on statistical learning theory.* Advanced Lectures on Machine Learning. Lecture Notes in Comput. Sci., pp. 1-40. Springer, New York, 2003.
[24] D. Pollard, *Asymptotics via Empirical Processes.*, Statist. Sci., vol. 4, pp. 365-366, 1989.
[25] M. A. Arcones, E. Gine, *U-processes indexed by VapnikChervonenkis classes of functions with applications to asymptotics and bootstrap of U-statistics with estimated parameters*, Stochastic Process. Appl., vol. 52, pp. 17-38, 1994.
[26] D. Nolan, D. Pollard, *U-processes: rates of convergence.*, Ann. Statist., vol. 15, pp. 780-799, 1987.