Maximal Deviations of Incomplete U-statistics with Applications to Empirical Risk Sampling

Stéphan Clémençon, Sylvain Robbiano, Jessica Tressou

Maximal Deviations of Incomplete U-statistics with Applications to Empirical Risk Sampling

Stéphan Clémençon* Sylvain Robbiano† Jessica Tressou ‡

Abstract

It is the goal of this paper to extend the *Empirical Risk Minimization* (ERM) paradigm, from a practical perspective, to the situation where a natural estimate of the risk is of the form of a K-sample U-statistics, as it is the case in the K-partite ranking problem for instance. Indeed, the numerical computation of the empirical risk is hardly feasible if not infeasible, even for moderate samples sizes. Precisely, it involves averaging $O(n^{d_1+\cdots+d_K})$ terms, when considering a U-statistic of degrees $(d_1, \ldots, d_K)$ based on samples of sizes proportional to n. We propose here to consider a drastically simpler Monte-Carlo version of the empirical risk based on $O(n)$ terms solely, which can be viewed as an *incomplete generalized U-statistic*, and prove that, remarkably, the approximation stage does not damage the ERM procedure and yields a learning rate of order $O_{\mathbb{P}}(1/\sqrt{n})$. Beyond a theoretical analysis guaranteeing the validity of this approach, numerical experiments are displayed for illustrative purpose.

Keywords: Empirical risk minimization, risk sampling, incomplete U-statistics, ranking, minimum-volume set

1 Introduction

In statistical learning theory, the paradigmatic approach to predictive problems is to use data-based estimates of the prediction error to select a decision rule from a class of candidates. In classification/regression, such estimates are sample mean statistics and the theory of *Empirical Risk Minimization* (ERM in abbreviated form) has been originally developed in this situation, relying essentially on the study of maximal deviations between these empirical averages and their expectations. The tools used for this purpose are mainly concentration inequalities for empirical processes; see [18] for instance. One may refer to [8] for a recent account of the theory of classification. Recently, a variety of learning issues, where natural empirical risk estimates are no longer basic sample mean statistics, have received a good deal of attention in the machine-learning literature, requiring to extend the ERM approach. Indeed, in certain problems such as *supervised ranking* [11], *learning on graphs* [5] or *pairwise dissimilarity-based clustering* [10], statistical counterparts of the risk are of the form of (generalized) U-statistics; see [19]. Such empirical functionals are computed by averaging over tuples of sampling observations, exhibiting thus a complex dependence structure. *Linearization techniques* (see [16]) are the main ingredient in investigating the behavior of empirical risk minimizers in this setting, the latter permitting to establish probabilistic upper bounds for the maximal deviation of collection of centered U-statistics under adequate conditions by reducing the study to that of standard empirical processes. However, while the ERM theory based on minimization of U-statistics is now consolidated, putting this approach in practice generally leads to face significant computational difficulties, not sufficiently well documented in the machine-learning literature. In many concrete cases, the mere computation of the risk involves a summation which extends over an extremely high number of tuples and runs out of time or memory on most machines. It is the major purpose of this paper to study how a simplistic sampling technique (*i.e.* drawing with replacement) applied to risk estimation, as originally proposed by [7] in the context of asymptotic pointwise estimation, may efficiently remedy this issue without damaging too much the "reduced variance" property of the estimates, while preserving the learning rates (including "fast-rate" situations). Applications to *supervised ranking* and to *minimum volume set learning* are considered here in order to illustrate this remarkable phenomenon.

The paper is structured as follows. As a first go, two important situations where the empirical functional of interest in the learning problem considered is of the form of a (generalized) U-statistic, hardly or not computable in most cases encountered in practice, are described at length in section 2 in order to motivate the subsequent study. Section 3 next recalls key concepts of the theory of *incomplete* generalized U-statistics and states the main result of the paper, establishing a probabilistic upper bound for the maximal deviation related to a

*LTCI - UMR No. 5141 Telecom Paristech CNRS - stephan.clemencon@telecom-paristech.fr

†LTCI - UMR No. 5141 Telecom Paristech CNRS - sylvain.robbiano@telecom-paristech.fr

‡INRA Metarisk - UR1204 - jessica.tressou@agroparistech.fr

finite collection of incomplete U-statistics, under mild assumptions. Finally, the implications of this result for the aforementioned examples from the learning perspective are thoroughly discussed and illustrated by numerical experiments in section 4. Technical details are postponed to the Appendix section.

2 Motivation

We start off with motivating the study of maximal deviations of collections of *incomplete generalized U-statistics* in the statistical learning context, through two problems, supervised and unsupervised respectively, which shall serve as running examples in this paper.

2.1 First example: K-partite ranking In the K-partite ranking problem, one has $K \geq 1$ independent random vectors $X^{(1)}, \ldots, X^{(K)}$ taking their values in a subset of a generally high-dimensional euclidean space $\mathcal{X}$, $\mathcal{X} \subset \mathbb{R}^d$ with $d \geq 1$ say, with respective probability distributions $F_1(dx), \ldots, F_K(dx)$. Informally, the goal is to learn, based on a pooled data set (made of independent observations drawn as the X^k's), a preorder on the input space $\mathcal{X}$, characterized by a scoring function $s : \mathcal{X} \to \mathbb{R}$ transporting the natural order on the real line onto $\mathcal{X}$ ($x \leq_s x' \Leftrightarrow s(x) \leq s(x')$ for all $(x, x') \in \mathcal{X}^2$), so that the random variable $s(X^{(k)})$ stochastically increases, as much as possible, with the label $k \in \{1, \ldots, K\}$. A quantitative performance criterion (when neglecting ties for simplicity) is given by:

(2.1)
$$L(s) \stackrel{\text{def}}{=} \mathbb{P}\left\{ s(X^{(1)}) < s(X^{(2)}) < \ldots < s(X^{(K)}) \right\}.$$

If K independent samples, of independent copies of the r.v. $X^{(k)}$ respectively, are available,

$$X_1^{(k)} \ldots, X_{n_k}^{(k)} \text{ with } n_k \geq 1 \text{ for } 1 \leq k \leq K, \tag{2.2}$$

a natural empirical counterpart of the ranking performance criterion is:

$$\widehat{L}_{\mathbf{n}}(s) = \frac{\sum_{i_1=1}^{n_1} \cdots \sum_{i_K=1}^{n_K} \mathbb{I}_{\left\{ s(X_{i_1}^{(1)}) < \ldots < s(X_{i_K}^{(K)}) \right\}}}{n_1 \times \cdots \times n_K}, \tag{2.3}$$

where $\mathbf{n} = (n_1, \ldots, n_K)$ and $\mathbb{I}_{\{E\}}$ denotes the indicator function of any event E. The performance of empirical maximizers of the quantity (2.3) (or of variants of the latter performance measure) over a class $\mathcal{S}$ of scoring function candidates has been investigated in several papers, mainly in the *bipartite* context (*i.e.* for $K = 2$), under various complexity assumptions for $\mathcal{S}$; see [2, 11] among others. In a variety of applications (information retrieval, design of recommender systems for instance), the number of classes K and/or the sample sizes n_k are fairly large, so that the number of terms to be summed in (2.3), $n_1 \times \cdots \times n_K$ namely, is prohibitive. As an illustration, one may refer to the public databases LETOR (available at `http://research.microsoft.com/~letor/`), which can be used to evaluate search engines for ranking documents according to their degree of pertinence for specific requests in particular (see [20]), where $K = 5$ and the sample sizes are very huge for most queries. Datasets released for recent competitions, such as the *Yahoo! Labs "Learning to Rank"* challenge in 2010 or the *KDD Cup Orange* challenge in 2009, provide other examples of such situations. In the *KDD Cup Orange* challenge, where submissions were evaluated based on the AUC performance (*i.e.* the statistic (2.3) when $K = 2$), the computation of the empirical version of the criterion required to average over 10^{12} pairs approximately, making "pairwise classification" approaches inapplicable (unless the sampling technique promoted here and analyzed in the subsequent section is used).

2.2 Second example: risk exposure assessment Our second example relates to unsupervised learning. Suppose that $M \geq 1$ hazards may arise from $K \geq 1$ different sources, which can combine in an additive manner, as in many environmental or health problems. To fix ideas, suppose that, weekly say, dietary contamination by M different chemicals through the possible consumption of $P \geq 1$ food items indexed by $p \in \{1, \ldots, P\}$ over a certain statistical population of interest is under study. The joint dietary risk exposure can be described by a random vector $\mathcal{E} = (\mathcal{E}_1, \ldots, \mathcal{E}_M)$, where:

$$\mathcal{E}_m = \sum_{p=1}^{P} c_{m,p} \cdot Q_p. \tag{2.4}$$

for $1 \leq m \leq M$, denoting by Q_p the quantity of food item No. p consumed per week by an individual drawn at random in the population studied and by $c_{m,p}$ the (random) contamination level related to food item No. p and pollutant No. m. In the field of food safety, risk assessors are interested in building confidence regions for the risk exposure $\mathcal{E}$ in $\mathbb{R}_+^M$:

(2.5)
$$R_\alpha = \arg\min \left\{ \ \lambda(R) : \mathbb{P}\left\{ \mathcal{E} \in R \right\} \geq \alpha, R \in \mathcal{B}(\mathbb{R}_+^M) \ \right\},$$

where Lebesgue measure on $\mathbb{R}_+^M$ is denoted by λ and the set of borelian subsets of $\mathbb{R}_+^M$ by $\mathcal{B}(\mathbb{R}_+^M)$. For values of the level α close to 1, such *minimum volume sets* (MV-sets in short; see [22]) describe regions where the exposure distribution is most concentrated, exposures lying in their complementary sets being possibly interpreted

as "abnormal". The construction of confidence regions for the risk exposure is based on the observation of the dietary behavior of J individuals independently drawn from the population, yielding an i.i.d. sample $\{\mathbf{Q}_i = (Q_{i,1},\ \ldots,\ Q_{i,P}) :\ 1 \leq i \leq J\}$ and on a database where a number $L_{m,p}$ of measures of contamination in pollutant m for food item p, for $1 \leq p \leq P$ and $1 \leq m \leq M$, are gathered, $\{\mathbf{c}_{m,p} = (c_{m,p,1},\ \ldots,\ c_{m,p,L_{m,p}})\}$. Based on these data, the probability involved in the constraint of the MV-set problem (2.5) is estimated by the empirical quantity given by:

$$\widehat{\mathbb{P}}\{\mathcal{E} \in R\} = \left(J \prod_{m=1}^{M} \prod_{p=1}^{P} L_{m,p}\right)^{-1} \times$$
$$\sum_{i=1}^{J} \sum_{l_{1,1}=1}^{L_{1,1}} \cdots \sum_{l_{M,P}=1}^{L_{M,P}} \mathbb{I}_{\left\{\left(\sum_{p=1}^{P} c_{m,p,l_{m,p}} \cdot Q_{i,p}\right)_{1 \leq m \leq M} \in R\right\}},$$

and the level α is replaced by $\alpha - \phi$, where ϕ is some *tolerance level*, depending, roughly speaking, on the order of magnitude of $\sup_{R \in \mathcal{R}} |\widehat{\mathbb{P}}\{\mathcal{E} \in R\} - \mathbb{P}\{\mathcal{E} \in R\}|$, where $\mathcal{R}$ is the class of Borelian sets over which the search is performed. Refer to [23] for precise results following in the footsteps of those in ERM theory. In practice, averaging over the $J \times \prod_{m=1}^{M} \prod_{p=1}^{P} L_{m,p}$ terms appearing in the formula above is generally infeasible. In [4] for instance, where estimation of the probability that the risk exposure to Ochratoxin A exceeds a critical threshold is considered, this corresponds to 4×10^{21} terms!

3 Uniform approximation of Generalized U-statistics through sampling

As shall be seen below, the statistics considered in the previous section are (generalized) U-statistics, which can be *uniformly* approximated by Monte-Carlo versions whose computation cost is drastically reduced. This will be next proved to be an essential tool for investigating the performance of decision rules learnt through optimization of such empirical quantities.

3.1 Definitions and key properties For clarity, we recall the definition of generalized U-statistics, the simplest extensions of standard sample mean statistics. Properties and asymptotic theory of U-statistics can be found in [19].

Definition 1. (Generalized U-statistic) *Let $K \geq 1$ and $(d_1,\ \ldots,\ d_K) \in \mathbb{N}^{*K}$. Let $(X_1^{(k)},\ \ldots,\ X_{n_k}^{(k)})$, $1 \leq k \leq K$, be K independent samples of i.i.d. random variables, taking their values in some space $\mathcal{X}_k$ with distribution $F_k(dx)$ respectively. The generalized (or K-sample) U-statistic of degrees $(d_1,\ \ldots,\ d_K)$ with kernel $H : \mathcal{X}_1^{d_1} \times \cdots \times \mathcal{X}_K^{d_K} \to \mathbb{R}$, square integrable with respect to the probability distribution $\mu = F_1^{\otimes d_1} \otimes \cdots \otimes F_K^{\otimes d_K}$, is defined as*

$$U_{\mathbf{n}}(H) = \frac{\sum_{I_1} \ldots \sum_{I_K} H(\mathbf{X}_{I_1}^{(1)}; \mathbf{X}_{I_2}^{(2)}; \ldots; \mathbf{X}_{I_K}^{(K)})}{\binom{n_1}{d_1} \times \cdots \binom{n_K}{d_K}}, \quad (3.6)$$

where the symbol $\sum_{I_k}$ refers to summation over all $\binom{n_k}{d_k}$ subsets $\mathbf{X}_{I_k}^{(k)} = (X_{i_1}^{(k)},\ \ldots,\ X_{i_{d_k}}^{(k)})$ related to a set I_k of d_k indexes $1 \leq i_1 < \ldots < i_{d_k} \leq n_k$. It is said symmetric when H is permutation symmetric in each set of d_k arguments $\mathbf{X}_{I_k}^{(k)}$.

Returning to the first example of the previous section, we observe that, for a fixed scoring function $s(x)$, the quantity (2.3) is a K-sample U-statistic of degree $(1,\ 1,\ \ldots,\ 1)$ with kernel given by:

$$\mathrm{H}_s(x_1,\ \ldots,\ x_K) = \mathbb{I}_{\{s(x_1) < s(x_2) < \cdots < s(x_K)\}}$$

for $(x_1,\ \ldots,\ x_K) \in \mathcal{X}^K$. In a similar manner, considering the functional involved in the second example, this corresponds to a $K = (M \times P + 1)$-sample U-statistic of degree $(1,\ 1,\ \ldots,\ 1)$, with kernel given by:

$$\mathrm{H}_R(\mathbf{q}, \mathbf{c}) = \mathbb{I}_{\left\{\left(\sum_{p=1}^{P} c_{m,p} \cdot q_p\right)_{1 \leq m \leq M} \in R\right\}},$$

for $\mathbf{q} = (q_1, \ldots, q_P) \in \mathbb{R}_+^P$ and $\mathbf{c} = ((c_{m,1}, \ldots, c_{m,P}), m = 1, \ldots, M) \in \mathbb{R}_+^{P \times M}$.

Beyond these two examples, many statistics used for pointwise estimation or hypothesis testing are actually U-statistics (*e.g.* the sample variance, the Gini mean difference, the Wilcoxon Mann-Whitney statistic, Kendall tau), their popularity mainly arise from their "reduced variance" property: the statistic $U_{\mathbf{n}}(H)$ has minimum variance among all unbiased estimators of the parameter

$$\theta(H) = \mathbb{E}[H(X_1^{(1)},\ \ldots,\ X_{d_1}^{(1)},\ \ldots,\ X_1^{(K)},\ \ldots,\ X_{d_k}^{(K)})].$$

Asymptotics. Classically, the limit properties of these statistics (LLN, CLT, *etc.*) are investigated in an asymptotic framework stipulating that, as the full sample size

$$n \stackrel{def}{=} n_1 + \ \ldots + n_K$$

tends to infinity, we have: $n_k/n \to \lambda_k > 0$ for $k = 1,\ \ldots,\ K$. They can be established by means of a linearization technique (see [16]), permitting to write $U_{\mathbf{n}}(H)$ as a sum of K basic sample mean statistics (of the order $O_{\mathbb{P}}(1/\sqrt{n})$ each, after recentering), plus possible degenerate terms (termed *degenerate* U-statistics). This method is extensively used in [11] for instance.

As previously seen on the running examples considered in this paper, in practice, the number $\prod_{k=1}^{K} \binom{n_k}{d_k}$ of terms to be summed up to compute (3.6) is generally prohibitive. As a remedy to this computational issue, in the seminal contribution [7], the concept of *incomplete generalized U-statistic* has been introduced, where the summation in formula (3.6) is replaced by a summation involving much less terms, extending over low cardinality subsets of the $\binom{n_k}{d_k}$ d_k-tuples of indices, $1 \leq k \leq K$, solely. In the simplest formulation, the subsets of indices are obtained by sampling with replacement, leading to the following definition.

DEFINITION 2. (INCOMPLETE GENERALIZED U-STATISTIC) *Let $B \geq 1$. The incomplete version of the U-statistic* (3.6) *based on B terms is defined by:*

$$\widetilde{U}_B(H) = \frac{1}{B} \sum_{(I_1, \ldots, I_K) \in \mathcal{D}_B} H(X_{I_1}^{(1)}, \ldots, X_{I_K}^{(K)}), \tag{3.7}$$

where $\mathcal{D}_B$ is a set of cardinality B built by sampling with replacement in the set $\Lambda = \{((i_1^{(1)}, \ldots, i_{d_1}^{(1)}), \ldots, (i_1^{(K)}, \ldots, i_{d_K}^{(K)})) : 1 \leq i_1^{(k)} < \ldots < i_{d_k}^{(k)} \leq n_k,\ 1 \leq k \leq K\}$.

REMARK 1. (ALTERNATIVE SAMPLING SCHEMES.) We point out that, as proposed in [17], other sampling schemes could be considered, sampling without replacement or Bernoulli sampling in particular. The results of this paper could be extended to these situations. Due to space limitation, we restrict our attention here to the sampling with replacement scheme.

In practice, B should be chosen much smaller than the cardinality of Λ, namely $\#\Lambda = \prod_{k=1}^{K} \binom{n_k}{d_k}$, in order to overcome the computational issue previously mentioned. We emphasize that the cost related to the computation of the value taken by the kernel H at a given point $(x_{I_1}^{(1)}, \ldots, x_{I_K}^{(K)})$ depending on the form of H is not considered here, focus is on the number of terms involved in the summation solely. As an estimator of $\theta(H)$, the statistic (3.7) is still unbiased but its variance is naturally larger than that of (3.6). Precisely, we have

$$\mathrm{Var}(\widetilde{U}_B(H)) = (1 - 1/B)\mathrm{Var}(U_{\mathbf{n}}(H)) + O(1/B),$$

as $B \to +\infty$; refer to [19] (see p. 193 therein). Incidentally, we underline that the empirical variance of (3.6) is not easy to compute neither since it involves summing approximately $\#\Lambda$ terms and bootstrap techniques should be used for this purpose, as proposed in [4]. The asymptotic properties of incomplete U-statistics have been investigated in several articles; see [9, 13, 17]. The angle embraced in the present paper is of quite different nature, the key idea we promote here is to use incomplete versions of collections of U-statistics in learning problems such as those described in section 2. The result established in the next section shows that this approach solves the numerical problem, while not damaging the learning rates.

3.2 Main result - Uniform approximation of U-statistics by incomplete U-statistics Under certain assumptions on the collection $\mathcal{H}$ of (symmetric) kernels H considered, concentration results established for U-processes (*i.e.* collections of U-statistics) may extend to their incomplete versions, as revealed by the following theorem. As the goal of this paper is to present the main ideas rather than formulating results at a high level of generality owing to space limitations, we consider the (not that restrictive) situation where the class $\mathcal{H}$ of kernels is a VC major class of functions of finite Vapnik-Chervonenkis dimension; see [12].

THEOREM 3.1. (MAXIMAL DEVIATION) *Let $\mathcal{H}$ be a collection of bounded symmetric kernels on $\Omega = \prod_{k=1}^{K} \mathcal{X}_k^{d_k}$ of finite* VC *dimension $\mathcal{V} < +\infty$. We set $\mathcal{M}_{\mathcal{H}} = \sup_{(H,x) \in \mathcal{H} \times \mathcal{X}} |H(x)|$. Then, the following assertions hold.*

(i) For all $\eta > 0$, we have: $\forall \mathbf{n} = (n_1, \ldots, n_K) \in \mathbb{N}^{*K}$, $\forall B \geq 1$,

$$\mathbb{P}\left\{ \sup_{H \in \mathcal{H}} \left| \widetilde{U}_B(H) - U_{\mathbf{n}}(H) \right| > \eta \right\} \leq 2(1+\#\Lambda)^{\mathcal{V}} \times e^{-B\eta^2/\mathcal{M}_{\mathcal{H}}^2}.$$

(ii) For all $\delta \in (0,1)$, with probability at least $1-\delta$, we have: $\forall n_k \geq 1$, $1 \leq k \leq K$,

$$\frac{1}{\mathcal{M}_{\mathcal{H}}} \sup_{H \in \mathcal{H}} \left| \widetilde{U}_B(H) - \mathbb{E}\left[\widetilde{U}_B(H)\right] \right| \leq 2\sqrt{\frac{2\mathcal{V}\log(1+\kappa)}{\kappa}} + \sqrt{\frac{\log(2/\delta)}{\kappa}} + \sqrt{\frac{\mathcal{V}\log(1+\#\Lambda) + \log(4/\delta)}{B}}, \tag{3.8}$$

where $\kappa = \min\{\lfloor n_1/d_1 \rfloor, \ldots, \lfloor n_K/d_K \rfloor\}$ and $\lfloor x \rfloor$ denotes the integer part of any real number x.

Refer to the Appendix for the proof. The bounds stated above show that, for a number $B = B_n$ of terms tending to infinity as $n \to +\infty$ at a rate $O(n)$, the maximal deviation $\sup_{H \in \mathcal{H}} |\widetilde{U}_B(H) - \theta(H)|$ is asymptotically of the order $O_{\mathbb{P}}(n^{-1/2})$, just like $\sup_{H \in \mathcal{H}} |U_{\mathbf{n}}(H) - \theta(H)|$. Remarkably, except in the case $K = 1$ and $d_K = 1$ solely, using such incomplete U-statistics thus yields

a significant gain in terms of computational cost and preserves the order of the probabilistic upper bounds for the uniform deviation.

4 Applications

We now discuss the consequences of Theorem 3.1 through the examples introduced in section 2 (notice that, in both cases, we have $\mathcal{M}_{\mathcal{H}} = 1$). Beyond theoretical guarantees, the performance of algorithms based on incomplete versions of the empirical counterpart of the functional of interest is illustrated by numerical results, supporting the efficiency of the sampling approach promoted in this paper in the machine-learning context.

4.1 Sampling the risk in K-partite ranking We place ourselves in the framework described in subsection 2.1. Here, the full sample size is $n = n_1 + \ldots + n_K$. Let $(b_1,\ b_2,\ \ldots,\ b_K)$ be a sequence of nonnegative integers such that:

$$\forall k \in \{1,\ \ldots,\ K\},\ \ b_k \sim n_k^{1/K} \sim n^{1/K} \text{ as } n \to +\infty.$$

The sampling scheme consists, for $1 \leq k \leq K$, of drawing with replacement b_k observations in the sample No. k: $X_{i_1}^{(k)},\ \ldots,\ X_{i_{b_k}}^{(k)}$. Set $B = b_1 \times \cdots \times b_K$. Based on the sampled data, we compute the following estimate of the ranking performance criterion

$$\widetilde{L}_B(s) = \frac{1}{B} \sum_{l_1=1}^{b_1} \cdots \sum_{l_K=1}^{b_K} \mathbb{I}_{\left\{ s(X_{i_{l_1}^{(1)}}^{(1)}) < \ldots < s(X_{i_{l_K}^{(K)}}^{(K)}) \right\}},$$

and consider the maximizer over a class $\mathcal{S}$ of scoring function candidates:

$$\widehat{s}_B = \arg\max_{s \in \mathcal{S}} \widetilde{L}_B(s). \tag{4.9}$$

The following result provides a rate bound for the ranking performance of the scoring function above (neglecting the bias term).

COROLLARY 4.1. *Suppose that $\mathcal{S}$ is a* VC *major class of functions of finite* VC *dimension $\mathcal{V} < +\infty$. Then, for all $\delta \in (0,1)$, we have with probability at least $1-\delta$: $\forall \mathbf{n} \in \mathbb{N}^{*K}$,*

$$\max_{s \in \mathcal{S}} L(s) - L(\widehat{s}_B) \leq c\sqrt{\frac{\mathcal{V}\log(\#\Lambda/\delta)}{n}}, \tag{4.10}$$

for some constant $c < +\infty$.

The proof immediately derives from Theorem 3.1, details are left to the reader. One should pay attention to the fact that the deficit of ranking performance of the rule obtained through maximization of statistics computed by averaging $O(n)$ terms is thus of the same order as that of $\arg\max_{s \in \mathcal{S}} \widehat{L}_{\mathbf{n}}(s)$, whose computation requires to evaluate averages extending over $O(n^K)$ terms.

REMARK 2. (ON FAST RATES) In the bipartite setup (*i.e.* $K = 2$), situations where fast rates of convergence can be achieved by $\arg\max_{s \in \mathcal{S}} \widehat{L}_{\mathbf{n}}(s)$ have been exhibited; see [11]. In this regard, we point out that, in these situations, the same rate bounds can be attained by $\widehat{s}_B$, at the price of a higher computational cost (*i.e.* of a larger asymptotic order for B) however.

A numerical example with $K = 5$. As an illustration, we display below some results related to the performance of the algorithm SVMRANK (implemented with default parameters, linear kernel and $C = 20$; see [15]) using the SVM-light implementation available at `http://svmlight.joachims.org/`. We simulated a mixture of 5 Gaussian distributions on $\mathbb{R}^2$ with means $m_1,\ \ldots,\ m_5$ respectively, where $m_i = (i/6,\ i/6)$ for $1 \leq i \leq 5$, and same covariance matrix $(1/15,\ 0;\ 0,\ 1/15)$, so that an optimal scoring function (w.r.t. the criterion (2.1)) is given by: $s(x,y) = x + y$ for all $(x,y) \in \mathbb{R}^2$. We independently drew 50 training samples of size $n = 10\,000$ (2 000 per class) and a test sample of size 10 000. Inside each class, we drew with replacement b observations and formed the dataset $\mathcal{D}_b$, for $b = 20,\ 100$. The results, averaged over the 50 replications, are reported in Table 1.

Table 1: Comparison of the empirical ranking performance : "ranking" experiment - $L^* = 0.1525$

% of data	1%	5%	100%
$\overline{L}$	0.1497	0.1520	0.1524
$\widehat{\sigma}$	0.0041	0.0008	0.0002
time (in seconds)	10	200	148523

Figures speak volume. We see that, even for $b = 20$ (*i.e* 1% of the data), the performance is close to the optimum L^* for a computation time reduced by a factor 10000. For $b = 100$ (*i.e* 5% of the data), it is quasi-optimal, with a gain in time of a factor greater than 500.

LETOR4.0 datasets. We also implemented the approach promoted in this paper on the benchmark LETOR datasets, (see `research.microsoft.com/en-us/um/people/letor/`), by means of the same ranking algorithm as that used in the previous experiment. To be more precise, we used the two query sets MQ2007 and MQ2008, where pairs

"page-query" assigned to a discrete label ranging from 0 to 2 (*i.e.* "non-relevant" - "relevant" - "extremely relevant") are gathered. In both datasets, 46 features are collected, over 69 623 instances in MQ2007 and over 15 211 instances in MQ2008. In each case, an estimate of the ranking risk L has been computed through 5 replications of a five-fold cross validation procedure, the results (mean and standard error) are reported in Tables 2 and 3. We also compute the Kendall τ statistic $\widehat{\tau}$ between the resulting rankings (recall that it ranges from -1 "full disagreement" to $+1$ "full agreement"), when using 1% ,5%, 10%, 20% and 100% of the data in each of the $K = 3$ samples. The results are reported in the Tables 2 and 3.

Table 2: Empirical ranking performance : "LETOR 2008".

%	1%	5%	10%	20%	100%
$\overline{L}$	0.3735	0.3939	0.3992	0.4015	0.4088
$\widehat{\sigma}$	0.0038	0.0040	0.0025	0.0027	0.0006
$\widehat{\tau}$	0.7648	0.8653	0.8937	0.9154	1

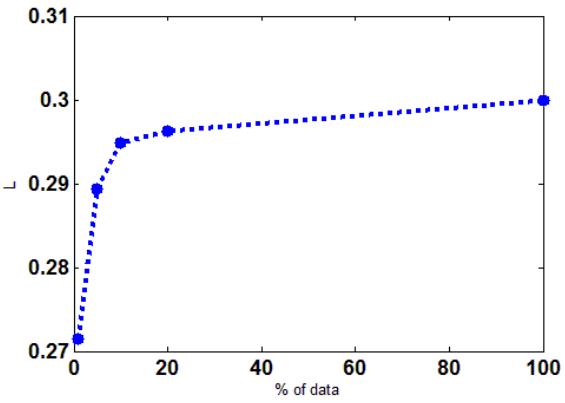

Figure 1: Empirical ranking performance for SVM-RANK based on 1%, 5%, 10%, 20% and 100% of the "LETOR 2007" dataset.

Table 3: Empirical ranking performance : "LETOR 2007".

%	1%	5%	10%	20%	100%
$\overline{L}$	0.2715	0.2894	0.2949	0.2963	0.3000
$\widehat{\sigma}$	0.0077	0.0027	0.0017	0.0019	0.0004
$\widehat{\tau}$	0.6621	0.7651	0.8328	0.8501	1

In both experiments, we observe that, as b_k/n_k increase, the ranking performance of the rules produced by the algorithm gets rapidly closer and closer to that of the ranking rule based on the whole dataset.

4.2 Sampling the distribution of risk exposures

We now turn to the second example; see subsection 2.2. In order to avoid the computation of $\widehat{\mathbb{P}}\{\mathcal{E} \in R\}$, which involves summing over $\#\Lambda = J\prod_{m=1}^{M}\prod_{p=1}^{P} L_{m,p}$ terms and is based on $n = J+\sum_{m=1}^{M}\sum_{p=1}^{P} L_{m,p}$ observations, we draw with replacement B times in the index set $\{1, \ldots, J\} \times \prod_{m=1}^{M}\prod_{p=1}^{P}\{1, \ldots, L_{m,p}\}$, so as to get a set of indices $\mathcal{D}_B$ of cardinality B. For any borelian $R \subset \mathbb{R}^M$, the probability that exposure lies in the region R is estimated by the incomplete U-statistics:

$$\widetilde{\mathbb{P}}_B\{\mathcal{E} \in R\} = \frac{1}{B} \times \sum_{(i,(l_{m,p}))\in\mathcal{D}_B} \mathbb{I}_{\left\{\left(\sum_{p=1}^{P} \widetilde{c}_{m,p,l_{m,p}} \cdot \widetilde{Q}_{i,p}\right)_{1\leq m\leq M} \in R\right\}}. \tag{4.11}$$

Suppose that the class $\mathcal{R}$ is of finite VC dimension $\mathcal{V} < +\infty$. Let $\alpha \in (0,1)$ be the target mass and $\delta \in (0,1)$ be the desired confidence level. Define the complexity penalty by:

$$\Phi(B,n,\delta) = 2\sqrt{\frac{2\mathcal{V}\log(1+\kappa)}{\kappa}} + \sqrt{\frac{\log(2/\delta)}{\kappa}} + \sqrt{\frac{\mathcal{V}\log(1+\#\Lambda)+\log(4/\delta)}{B}}. \tag{4.12}$$

Consider the solution $\widehat{R}_\alpha$ of the constrained optimization problem:

$$\textit{maximize } \lambda(R) \textit{ over } \mathcal{R} \quad \textit{subject to } \widetilde{\mathbb{P}}\{\mathcal{E} \in R\} \geq \alpha - \Phi(B,n,\delta). \tag{4.13}$$

The result below shows that, if the number B of exposure values computed through the sampling scheme is of the order $O(n)$, the performance of $\widehat{R}_\alpha$ is then comparable to that of the region whose selection is based on the quantities $\widehat{\mathbb{P}}\{\mathcal{E} \in R\}$, $R \in \mathcal{R}$.

COROLLARY 4.2. *For all $\delta \in (0,1)$, we have with probability at least $1-\delta$:*

$$\lambda\left(\widehat{R}_\alpha\right) \leq \inf_{R\in\mathcal{R}:\ \mathbb{P}\{\mathcal{E}\in\mathcal{R}\}\geq\alpha} \lambda(R)$$

and

$$\mathbb{P}\left\{\mathcal{E} \in \widehat{R}_\alpha\right\} \geq \alpha - 2\Phi(B,n,\delta).$$

This is a straightforward consequence of Theorem 3.1, details are omitted (refer to the argument of Corollary 6 in [23] for further details).

Confidence regions for joint dietary exposure to cadmium, mercury, PCB's and sodium. Below, we present numerical results where approximate solutions of (4.13) are built by means of a *dyadic recursive partitioning* scheme of the exposure space; see [6] as well as section 7 in [23]. Dietary exposures to the four substances are built based on French data surveys. Contamination data (cadmium, mercury, PCB's) come from the second French Total Diet Study, [3], and the sodium level is extracted from the Ciqual table describing French food composition table; see `http://www.anses.fr/TableCIQUAL/`. Based on those datasets, $P = 5$ food groups containing either one or more substances were composed (dairy products, meat and eggs, fish, fruit and vegetables, other foods) according to the similarity of their contamination/composition levels totalizing $\sum_{m,p} L_{m,p} = 2341$ observations. Consumption data for French adults ($J = 2488$, aged 18-79 yo) was extracted from the INCA2 database, related to a French consumption survey conducted by the French Agency for Food, Environmental and Occupational Health Safety in 2009, [1]. Interviewed individuals reported their food consumption over 7 days, together with some sociodemographic variables including their body weights.

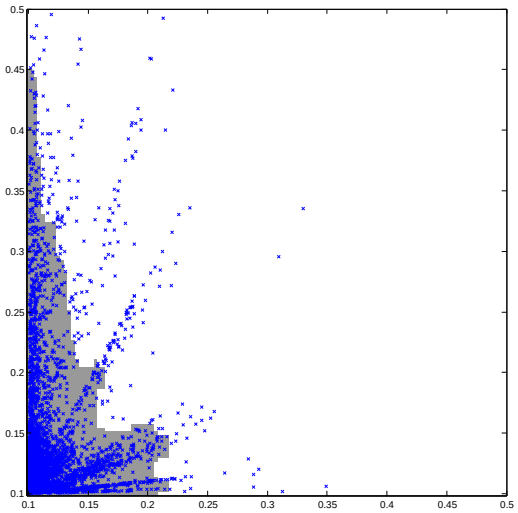

Figure 2: Cloud of rebuilt bivariate exposures to cadmium (x-axis) and mercury (y axis)and MV-set based on the incomplete criterion ($B = 10\,000$).

Table 4: Performance of the MV-set algorithm - M substances. Each cell contains the optimal volume and the test mass for a MV-set of level 95%.

M	2	3	4
n	3179	3529	4829
# Λ	2.96E+14	9.48E+21	6.89E+33
$B = 5\,000$	0.01481	0.00146	0.00002
	(95.2%)	(93.7%)	(92.3%)
$B = 10\,000$	0.01423	0.00151	0.00002
	(94.5%)	(94.1%)	(93.2%)
$B = 20\,000$	0.01412	0.01407	0.00162
	(95.0%)	(95.0%)	(94.5%)

Vectors of exposure (in $\mathbb{R}^4$) were generated by selecting data from the consumption and contamination/composition datasets at random with replacement. Whereas $\#\Lambda$ is equal to 6.9 10^{33} in this example, a MV-set of level $\alpha = 95\%$ has been learnt with only $B = 5000$, $10\,000$, or $20\,000$ exposure vectors. The resulting region was tested with a much larger sample of exposures ($1\,000\,000$). Table 4 summarizes the resulting test mass and volume obtained for the different sampling sizes considering in turn $M = 2$ (cadmium and mercury), $M = 3$ (cadmium, mercury, PCB's) or all $M = 4$ substances; the total number of observations n and the total number of terms involved in the "complete" U-statistic $\#\Lambda$ are given in Table 4. The performance of the algorithm clearly decreases when the dimension of the problem is increased. In the case $M = 2$, the optimal region related to the exposure to cadmium and mercury is illustrated in Fig. 4.2 (the exposure values are normalized to $(0, 1)$). In higher dimensions, resulting regions can simply be projected on pairs of substances (the matlab/mex code we used is available at `http://web.eecs.umich.edu/~cscott/code.html`).

5 Conclusion

Though of great simplicity, the results stated in this paper are of crucial importance in practice in the "big data" era. They hopefully shed light on tractable strategies for implementing learning techniques, when the (risk) functional has a statistical counterpart which is of the form of a U-statistic. Whereas the theoretical properties of decision rules based on optimizing such statistics are becoming well-documented in the machine-learning literature, computational issues related to the practical implementation of learning algorithms dedicated to these optimization problems had not been tackled, to the best of our knowledge. The essential contribution of this paper is to provide theoretical/empirical

evidence that using *incomplete U-statistics* as estimates of the criterion of interest may provide a simple and elegant way of dramatically reducing computational cost in practice, while yielding nearly optimal solutions. The analysis, carried out here in a finite VC dimension framework, suggests to investigate next the use of such statistics for model selection issues and to study concentration properties of *weighted multinomial random variables* involved in the maximal deviation between U-statistics and their incomplete versions.

Appendix - Proof of Theorem 3.1

For convenience, we introduce the random sequence $\epsilon = ((\epsilon_k(I))_{I\in\Lambda})_{1\leq k\leq B}$, where $\epsilon_k(I)$ is equal to 1 if the tuple $I = (I_1, \ldots, I_K)$ has been selected at the k-th draw and to 0 otherwise: the ϵ_k's are i.i.d. random vectors and, for all $(k, I) \in \{1, \ldots, B\} \times \Lambda$, the r.v. $\epsilon_k(I)$ has a Bernoulli distribution with parameter $1/\#\Lambda$. We also set $\mathbf{X}_I = (\mathbf{X}_{I_1}^{(1)}, \ldots, \mathbf{X}_{I_K}^{(K)})$ for any I in Λ. Equipped with these notations, observe first that one may write: $\forall B \geq 1$, $\forall \mathbf{n} \in \mathbb{N}^{*K}$,

$$\widetilde{U}_B(H) - U_{\mathbf{n}}(H) = \frac{1}{B}\sum_{k=1}^{B} \mathcal{Z}_k(H),$$

where $\mathcal{Z}_k(H) = \sum_{I\in\Lambda}(\epsilon_k(I) - 1/\#\Lambda)H(\mathbf{X}_I)$ for any $(k, I) \in \{1, \ldots, B\} \times \Lambda$. It follows from the independence between the $\mathbf{X}_I$'s and the $\epsilon(I)$'s that, for all $H \in \mathcal{H}$, conditioned upon the $\mathbf{X}_I$'s, the variables $\mathcal{Z}_1(H), \ldots, \mathcal{Z}_B(H)$ are independent, centered and almost-surely bounded by $2\mathcal{M}_{\mathcal{H}}$ (notice that $\sum_{I\in\Lambda} \epsilon_k(I) = 1$ for all $k \geq 1$). By virtue of Sauer's lemma, since $\mathcal{H}$ is a VC major class with finite VC dimension $\mathcal{V}$, we have, for fixed $\mathbf{X}_I$'s:

$$\#\{(H(\mathbf{X}_I))_{I\in\Lambda} : H \in \mathcal{H}\} \leq (1 + \#\Lambda)^V.$$

Hence, conditioned upon the $\mathbf{X}_I$'s, using the union bound and next Hoeffding's inequality applied to the independent sequence $\mathcal{Z}_1(H), \ldots, \mathcal{Z}_B(H)$, for all $\eta > 0$, we obtain that:

$$\begin{aligned}\mathbb{P}\left\{\sup_{H\in\mathcal{H}}\left|\widetilde{U}_B(H) - U_{\mathbf{n}}(H)\right| > \eta \mid (\mathbf{X}_I)_{I\in\Lambda}\right\} \\ \leq \mathbb{P}\left\{\sup_{H\in\mathcal{H}}\left|\frac{1}{B}\sum_{k=1}^{B}\mathcal{Z}_k(H)\right| > \eta \mid (\mathbf{X}_I)_{I\in\Lambda}\right\} \\ \leq 2(1 + \#\Lambda)^V e^{-B\eta^2/\mathcal{M}_{\mathcal{H}}^2},\end{aligned}$$

which proves the first assertion of the theorem. Notice that this can be formulated: for any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$:

$$\sup_{H\in\mathcal{H}}\left|\widetilde{U}_B(H) - U_{\mathbf{n}}(H)\right| \leq \mathcal{M}_{\mathcal{H}}\times \sqrt{\frac{V\log(1 + \#\Lambda) + \log(2/\delta)}{B}}.$$

The second part of the theorem straightforwardly results from the first part combined with the following result, which extends Corollary 3 in [11] to the K-sample situation.

LEMMA 5.1. *Suppose that Theorem 3.1's hypotheses are fulfilled. For all $\delta \in (0, 1)$, we have with probability at least $1 - \delta$,*

$$\frac{1}{\mathcal{M}_{\mathcal{H}}}\sup_{H\in\mathcal{H}}|U_{\mathbf{n}}(H) - \theta(H)| \leq 2\sqrt{\frac{2V\log(1 + \kappa)}{\kappa}} + \sqrt{\frac{\log(1/\delta)}{\kappa}}.$$

PROOF. Set $\kappa = \min\{\lfloor n_1/d_1\rfloor, \ldots, \lfloor n_K/d_K\rfloor\}$ and let

$$\begin{aligned}\kappa^{-1}V_H\left(X_1^{(1)}, \ldots, X_{n_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{n_K}^{(K)}\right) = \\ H\left(X_1^{(1)}, \ldots, X_{d_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{d_K}^{(K)}\right) \\ + H\left(X_{d_1+1}^{(1)}, \ldots, X_{2d_1}^{(1)}, \ldots, X_{d_K+1}^{(K)}, \ldots, X_{2d_K}^{(K)}\right) + \ldots \\ + H\left(X_{\kappa d_1 - d_1+1}^{(1)}, \ldots, X_{\kappa d_K - d_K+1}^{(K)}, \ldots, X_{\kappa d_K}^{(K)}\right),\end{aligned}$$

for any $H \in \mathcal{H}$. Recall that the K-sample U-statistic $U_{\mathbf{n}}(H)$ can be expressed as

$$U_{\mathbf{n}}(H) = \frac{1}{n_1!\cdots n_K!}\times \sum_{\sigma_1\in\mathfrak{S}_{n_1}, \ldots, \sigma_K\in\mathfrak{S}_{n_K}} V\left(X_{\sigma_1(1)}^{(1)}, \ldots, X_{\sigma_K(n_K)}^{(K)}\right),$$

where $\mathfrak{S}_m$ denotes the symmetric group of order m for any $m \geq 1$. This representation as an average of sums of κ independent terms is known as the (first) Hoeffding's decomposition; see [16]. Then, using Jensen's inequality in particular, one may easily show that, for any nondecreasing convex function $\psi : \mathbb{R}_+ \to \mathbb{R}$, the quantity $\mathbb{E}[\psi(\sup_{H\in\mathcal{H}}|U_{\mathbf{n}}(\bar{H})|)]$ is bounded by

$$\mathbb{E}\left[\psi\left(\sup_{H\in\mathcal{H}}\left|V_{\bar{H}}(X_1^{(1)}, \ldots, X_{n_K}^{(K)})\right|\right)\right],$$

where we set $\bar{H} = H - \theta(H)$ for all $H \in \mathcal{H}$. Now, using standard symmetrization and randomization arguments

(see [14] for instance) and the bound above, we obtain that

$$\mathbb{E}\left[\psi\left(\sup_{H\in\mathcal{H}}\left|U_{\mathbf{n}}(\bar{H})\right|\right)\right] \leq \mathbb{E}\left[\psi\left(2\mathcal{R}_\kappa\right)\right], \tag{5.14}$$

where

$$\mathcal{R}_\kappa = \sup_{H\in\mathcal{H}} \frac{1}{\kappa}\sum_{l=1}^{\kappa} \epsilon_l H\left(X^{(1)}_{(l-1)d_1+1},\ \ldots,\ X^{(K)}_{ld_K}\right),$$

is a Rademacher average based on the Rademacher chaos $\epsilon_1,\ \ldots,\ \epsilon_\kappa$ (independent random symmetric sign variables), independent from the $X_i^{(k)}$'s. We now apply the bounded difference inequality (see [21]) to the functional $\mathcal{R}_\kappa$, seen as a function of the i.i.d. random variables $(\epsilon_l, X^{(1)}_{(l-1)d_1+1},\ \ldots,\ X^{(1)}_{ld_1},\ \ldots,\ X^{(K)}_{(l-1)d_K+1},\ \ldots,\ X^{(K)}_{ld_K})$, $1 \leq l \leq \kappa$: changing any of these random variables change the value of $\mathcal{R}_\kappa$ by at most $\mathcal{M}_\mathcal{H}/\kappa$. One thus obtains from (5.14) with $\psi(x) = \exp(\lambda x)$, where $\lambda > 0$ is a parameter which shall be chosen later, that:

$$\mathbb{E}\left[\exp\left(\lambda \sup_{H\in\mathcal{H}}\left|U_{\mathbf{n}}(\bar{H})\right|\right)\right] \leq \exp\left(2\lambda\mathbb{E}[\mathcal{R}_\kappa] + \frac{\mathcal{M}_\mathcal{H}^2\lambda^2}{4\kappa}\right). \tag{5.15}$$

Applying Chernoff's method, one then gets:

$$\mathbb{P}\left\{\sup_{H\in\mathcal{H}}\left|U_{\mathbf{n}}(\bar{H})\right| > \eta\right\} \leq \exp\left(-\lambda\eta + 2\lambda\mathbb{E}[\mathcal{R}_\kappa] + \frac{\mathcal{M}_\mathcal{H}^2\lambda^2}{4\kappa}\right). \tag{5.16}$$

Using the bound (see Eq. (6) in [8] for instance)

$$\mathbb{E}[\mathcal{R}_\kappa] \leq \mathcal{M}_\mathcal{H}\sqrt{\frac{2V\log(1+\kappa)}{\kappa}}$$

and taking $\lambda = 2\kappa(\eta - 2\mathbb{E}[\mathcal{R}_\kappa])/\mathcal{M}_\mathcal{H}^2$ in (5.16) finally establishes the desired result.

References

[1] AFSSA. Report of the 2006/2007 Individual and National Study on Food Consumption 2 (INCA 2). Technical report, 2009. 44 pages.

[2] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the ROC curve. *J. Mach. Learn. Res.*, 6:393–425, 2005.

[3] Anses. Second french total diet study (tds 2): Report 1 and 2. Technical report, 2011. J.C. Leblanc and V. Sirot (coordinators), 304 and 354 pages.

[4] P. Bertail and J. Tressou. Incomplete generalized U-statistics for food risk assessment. *Biometrics*, 62(1):66–74, 2006.

[5] G. Biau and L. Bleakley. Statistical inference on graphs. *Statistics & Decisions*, 24:209–232, 2006.

[6] G. Blanchard, C. Schäfer, and Y. Rozenholc. Oracle bounds and exact algorithm for dyadic classification trees. pages 378–392, Heidelberg, 2004. Springer-Verlag.

[7] G. Blom. Some properties of incomplete U-statistics. *Biometrika*, 63(3):573–580, 1976.

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

[9] B.M. Brown and D.G. Kildea. Reduced U-statistics and the Hodges-Lehmann estimator. *The Annals of Statistics*, 6:828–835, 1978.

[10] S. Clémençon. On U-processes and clustering performance. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 37–45, 2011.

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

[12] R.M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

[13] E. Enqvist. *On sampling from sets of random variables with application to incomplete U-statistics*. PhD thesis, 1978.

[14] E. Giné and J. Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, 12(4):929–989, 1984.

[15] R. Herbrich, T. Graepel, and K. Obermayer. *Advances in Large Margin Classifiers*, chapter Large margin rank boundaries for ordinal regression, pages 115–132. MIT Press, 2000.

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

[17] S. Janson. The asymptotic distributions of incomplete U-statistics. *Z. Wahrsch. verw. Gebiete*, 66:495–505, 1984.

[18] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, New York, 1991.

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

[20] T.Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li. Letor: Benchmark dataset for research on learning to rank for information retrieval. In *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval*, pages 3–10, 2007.

[21] C. McDiarmid. *On the method of bounded differences*, pages 144–188. Cambridge University Press, 1989.

[22] W. Polonik. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 69(1):1–24, 1997.

[23] C. Scott and R. Nowak. Learning Minimum Volume Sets. *Journal of Machine Learning Research*, 7:665–704, 2006.