Scaling-up Empirical Risk Minimization: Optimization of Incomplete U-statistics

Stéphan Clémençon, Igor Colin, Aurélien Bellet

Journal of Machine Learning Research 17 (2016) 1-36 Submitted 1/15; Revised 4/16; Published 6/16

Scaling-up Empirical Risk Minimization: Optimization of Incomplete U-statistics

Stephan Clémençon STEPHAN.CLEMENCON@TELECOM-PARISTECH.FR
Igor Colin IGOR.COLIN@TELECOM-PARISTECH.FR
LTCI, CNRS, Télécom ParisTech
Université Paris-Saclay, 75013, Paris, France

Aurélien Bellet AURELIEN.BELLET@INRIA.FR
Magnet Team, INRIA Lille – Nord Europe
59650 Villeneuve d'Ascq, France

Editor: Xiaotong Shen

Abstract

In a wide range of statistical learning problems such as ranking, clustering or metric learning among others, the risk is accurately estimated by U-statistics of degree $d \geq 1$, *i.e.* functionals of the training data with low variance that take the form of averages over k-tuples. From a computational perspective, the calculation of such statistics is highly expensive even for a moderate sample size n, as it requires averaging $O(n^d)$ terms. This makes learning procedures relying on the optimization of such data functionals hardly feasible in practice. It is the major goal of this paper to show that, strikingly, such empirical risks can be replaced by drastically computationally simpler Monte-Carlo estimates based on $O(n)$ terms only, usually referred to as *incomplete* U*-statistics*, without damaging the $O_{\mathbb{P}}(1/\sqrt{n})$ learning rate of *Empirical Risk Minimization* (ERM) procedures. For this purpose, we establish uniform deviation results describing the error made when approximating a U-process by its incomplete version under appropriate complexity assumptions. Extensions to model selection, fast rate situations and various sampling techniques are also considered, as well as an application to stochastic gradient descent for ERM. Finally, numerical examples are displayed in order to provide strong empirical evidence that the approach we promote largely surpasses more naive subsampling techniques.

Keywords: big data, empirical risk minimization, U-processes, rate bound analysis, sampling design, stochastic gradient descent

1. Introduction

In classification/regression, empirical risk estimates are sample mean statistics and the theory of *Empirical Risk Minimization* (ERM) has been originally developed in this context, see Devroye et al. (1996). The ERM theory essentially relies on the study of maximal deviations between these empirical averages and their expectations, under adequate complexity assumptions on the set of prediction rule candidates. The relevant tools are mainly concentration inequalities for empirical processes, see Ledoux and Talagrand (1991) for instance.

In a wide variety of problems that received a good deal of attention in the machine learning literature and ranging from clustering to image recognition through ranking or learning on graphs, natural estimates of the risk are not basic sample means but take the

form of averages of d-tuples, usually referred to as U-statistics in Probability and Statistics, see Lee (1990). In Clémençon et al. (2005) for instance, ranking is viewed as pairwise classification and the empirical ranking error of any given prediction rule is a U-statistic of order 2, just like the *within cluster point scatter* in cluster analysis (see Clémençon, 2014) or empirical performance measures in metric learning, refer to Cao et al. (2012) for instance. Because empirical functionals are computed by averaging over tuples of sampling observations, they exhibit a complex dependence structure, which appears as the price to be paid for low variance estimates. *Linearization techniques* (see Hoeffding, 1948) are the main ingredient in studying the behavior of empirical risk minimizers in this setting, allowing to establish probabilistic upper bounds for the maximal deviation of collection of centered U-statistics under appropriate conditions by reducing the analysis to that of standard empirical processes. However, while the ERM theory based on minimization of U-statistics is now consolidated (see Clémençon et al., 2008), putting this approach in practice generally leads to significant computational difficulties that are not sufficiently well documented in the machine learning literature. In many concrete cases, the mere computation of the risk involves a summation over an extremely high number of tuples and runs out of time or memory on most machines.

Whereas the availability of massive information in the Big Data era, which machine learning procedures could theoretically now rely on, has motivated the recent development of *parallelized / distributed* approaches in order to scale-up certain statistical learning algorithms, see Bekkerman et al. (2011) or Bianchi et al. (2013) and the references therein, the present paper proposes to use *sampling techniques* as a remedy to the apparent intractability of learning from data sets of explosive size, in order to break the current computational barriers. More precisely, it is the major goal of this article to study how a simplistic sampling technique (*i.e.* drawing with replacement) applied to risk estimation, as originally proposed by Blom (1976) 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 certain "fast-rate" situations). For this purpose, we investigate to which extent a U-process, that is a collection of U-statistics, can be accurately approximated by a Monte-Carlo version (which shall be referred to as an *incomplete* U*-process* throughout the paper) involving much less terms, provided it is indexed by a class of kernels of controlled complexity (in a sense that will be explained later). A maximal deviation inequality connecting the accuracy of the approximation to the number of terms involved in the approximant is thus established. This result is the key to the analysis of the statistical performance of minimizers of risk estimates when they are in the form of an incomplete U-statistic. In particular, this allows us to show the advantage of using this specific sampling technique, compared to more naive approaches with exactly the same computational cost, consisting for instance in first drawing a subsample and then computing a risk estimate of the form of a (complete) U-statistic based on it. We also show how to incorporate this sampling strategy into iterative statistical learning techniques based on stochastic gradient descent (SGD), see Bottou (1998). The variant of the SGD method we propose involves the computation of an incomplete U-statistic to estimate the gradient at each step. For the estimator thus produced, rate bounds describing its statistical performance are established under mild assumptions. Beyond theoretical results, we

present illustrative numerical experiments on metric learning and clustering with synthetic and real-world data that support the relevance of our approach.

The rest of the article is organized as follows. In Section 2, we recall basic definitions and concepts pertaining to the theory of U-statistics/processes and present important examples in machine learning where natural estimates of the performance/risk measure are U-statistics. We then review the existing results for the empirical minimization of complete U-statistics. In Section 3, we recall the notion of incomplete U-statistic and we derive maximal deviation inequalities describing the error made when approximating a U-statistic by its incomplete counterpart uniformly over a class of kernels that fulfills appropriate complexity assumptions. This result is next applied to derive (possibly fast) learning rates for minimizers of the incomplete version of the empirical risk and to model selection. Extensions to incomplete U-statistics built by means of other sampling schemes than sampling with replacement are also investigated. In Section 4, estimation by means of incomplete U-statistics is applied to stochastic gradient descent for iterative ERM. Section 5 presents some numerical experiments. Finally, Section 6 collects some concluding remarks. Technical details are deferred to the Appendix.

2. Background and Preliminaries

As a first go, we briefly recall some key notions of the theory of U-statistics (Section 2.1) and provide several examples of statistical learning problems for which natural estimates of the performance/risk measure are in the form of U-statistics (Section 2.2). Finally, we review and extend the existing rate bound analysis for the empirical minimization of (complete) generalized U-statistics (Section 2.3). Here and throughout, $\mathbb{N}^*$ denotes the set of all strictly positive integers, $\mathbb{R}_+$ the set of nonnegative real numbers.

2.1 U-Statistics/Processes: Definitions and Properties

For clarity, we recall the definition of generalized U-statistics. An excellent account of properties and asymptotic theory of U-statistics can be found in Lee (1990).

Definition 1 (Generalized U-statistic) *Let* $K \geq 1$ *and* $(d_1, \ldots, d_K) \in \mathbb{N}^{*K}$. *Let* $\mathbf{X}_{\{1,\ldots,n_k\}} = (X_1^{(k)}, \ldots, X_{n_k}^{(k)})$, $1 \leq k \leq K$, *be* K *independent samples of sizes* $n_k \geq d_k$ *and composed of i.i.d. random variables taking their values in some measurable space* $\mathcal{X}_k$ *with distribution* $F_k(dx)$ *respectively. Let* $H : \mathcal{X}_1^{d_1} \times \cdots \times \mathcal{X}_K^{d_K} \to \mathbb{R}$ *be a measurable function, square integrable with respect to the probability distribution* $\mu = F_1^{\otimes d_1} \otimes \cdots \otimes F_K^{\otimes d_K}$. *Assume in addition (without loss of generality) that* $H(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)})$ *is symmetric within each block of arguments* $\mathbf{x}^{(k)}$ *(valued in* $\mathcal{X}_k^{d_k}$*),* $1 \leq k \leq K$. *The generalized (or* K*-sample)* U*-statistic of degrees* $(d_1, \ldots, d_K)$ *with kernel* H, *is then defined as*

$$U_{\mathbf{n}}(H) = \frac{1}{\prod_{k=1}^{K} \binom{n_k}{d_k}} \sum_{I_1} \cdots \sum_{I_K} H(\mathbf{X}_{I_1}^{(1)}, \mathbf{X}_{I_2}^{(2)}, \ldots, \mathbf{X}_{I_K}^{(K)}), \tag{1}$$

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* $1 \leq i_1 < \ldots < i_{d_k} \leq n_k$ *and* $\mathbf{n} = (n_1, \ldots, n_K)$.

The above definition generalizes standard sample mean statistics, which correspond to the case $K = 1 = d_1$. More generally when $K = 1$, $U_{\mathbf{n}}(H)$ is an average over all d_1-tuples of observations, while $K \geq 2$ corresponds to the multi-sample situation with a d_k-tuple for each sample $k \in \{1, \ldots, K\}$. A U-process is defined as a collection of U-statistics indexed by a set $\mathcal{H}$ of kernels. This concept generalizes the notion of empirical process.

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

$$\begin{aligned} \mu(H) &= \mathbb{E}\left[H(X_1^{(1)}, \ldots, X_{d_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{d_K}^{(K)})\right] \\ &= \int_{\mathbf{x}^{(1)} \in \mathcal{X}_1^{d_1}} \cdots \int_{\mathbf{x}^{(K)} \in \mathcal{X}_K^{d_K}} H(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)}) dF_1^{\otimes d_1}(\mathbf{x}^{(1)}) \cdots dF_K^{\otimes d_K}(\mathbf{x}^{(K)}) = \mathbb{E}\left[U_{\mathbf{n}}(H)\right]. \end{aligned} \tag{2}$$

Classically, the limit properties of these statistics (law of large numbers, central limit theorem, *etc.*) are investigated in an asymptotic framework stipulating that, as the size of the full pooled sample

$$n \overset{\text{def}}{=} n_1 + \ldots + n_K \tag{3}$$

tends to infinity, we have:

$$n_k/n \to \lambda_k > 0 \text{ for } k = 1, \ldots, K. \tag{4}$$

Asymptotic results and deviation/moment inequalities for K-sample U-statistics can be classically established by means of specific representations of this class of functionals, see (15) and (27) introduced in later sections. Significant progress in the analysis of U-statistics and U-processes has then recently been achieved by means of decoupling theory, see de la Peña and Giné (1999). For completeness, we point out that the asymptotic behavior of (multisample) U-statistics has been investigated under weaker integrability assumptions than that stipulated in Definition 1, see Lee (1990).

2.2 Motivating Examples

In this section, we review important supervised and unsupervised statistical learning problems where the empirical performance/risk measure is of the form of a generalized U-statistics. They shall serve as running examples throughout the paper.

2.2.1 Clustering

Clustering refers to the unsupervised learning task that consists in partitioning a set of data points $X_1, \ldots, X_n$ in a feature space $\mathcal{X}$ into a finite collection of subgroups depending on their similarity (in a sense that must be specified): roughly, data points in the same subgroup should be more similar to each other than to those lying in other subgroups. One may refer to Chapter 14 in Friedman et al. (2009) for an account of state-of-the-art clustering techniques. Formally, let $M \geq 2$ be the number of desired clusters and consider a symmetric function $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ such that $D(x, x) = 0$ for any $x \in \mathcal{X}$. D measures the

dissimilarity between pairs of observations $(x, x') \in \mathcal{X}^2$: the larger $D(x, x')$, the less similar x and x'. For instance, if $\mathcal{X} \subset \mathbb{R}^d$, D could take the form $D(x, x') = \Psi(\|x - x'\|_q)$, where $q \geq 1$, $\|a\|_q = (\sum_{i=1}^d |a_i|^q)^{1/q}$ for all $a \in \mathbb{R}^d$ and $\Psi : \mathbb{R}_+ \to \mathbb{R}_+$ is any borelian nondecreasing function such that $\Psi(0) = 0$. In this context, the goal of clustering methods is to find a partition $\mathcal{P}$ of the feature space $\mathcal{X}$ in a class Π of partition candidates that minimizes the following *empirical clustering risk*:

$$\widehat{W}_n(\mathcal{P}) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} D(X_i, X_j) \cdot \Phi_{\mathcal{P}}(X_i, X_j), \tag{5}$$

where $\Phi_{\mathcal{P}}(x, x') = \sum_{\mathcal{C} \in \mathcal{P}} \mathbb{I}\{(x, x') \in \mathcal{C}^2\}$. Assuming that the data $X_1, \ldots, X_n$ are i.i.d. realizations of a generic random variable X drawn from an unknown probability distribution $F(dx)$ on $\mathcal{X}$, the quantity $\widehat{W}_n(\mathcal{P})$, also known as the *intra-cluster similarity* or *within cluster point scatter*, is a one sample U-statistic of degree two ($K = 1$ and $d_1 = 2$) with kernel given by:

$$\forall (x, x') \in \mathcal{X}^2, \;\; H_{\mathcal{P}}(x, x') = D(x, x') \cdot \Phi_{\mathcal{P}}(x, x'), \tag{6}$$

according to Definition 1 provided that $\int\int_{(x,x') \in \mathcal{X}^2} D^2(x, x') \cdot \Phi_{\mathcal{P}}(x, x') F(dx) F(dx') < +\infty$. The expectation of the empirical clustering risk $\widehat{W}_n(\mathcal{P})$ is given by

$$W(\mathcal{P}) = \mathbb{E}\left[D(X, X') \cdot \Phi_{\mathcal{P}}(X, X')\right], \tag{7}$$

where X' is an independent copy of the r.v. X, and is named the *clustering risk* of the partition $\mathcal{P}$. The statistical analysis of the clustering performance of minimizers $\widehat{\mathcal{P}}_n$ of the empirical risk (5) over a class Π of appropriate complexity can be found in Clémençon (2014). Based on the theory of U-processes, it is shown in particular how to establish rate bounds for the excess of clustering risk of any empirical minimizer, $W(\widehat{\mathcal{P}}_n) - \inf_{\mathcal{P} \in \Pi} W(\mathcal{P})$ namely, under appropriate complexity assumptions on the cells forming the partition candidates.

2.2.2 Metric Learning

Many problems in machine learning, data mining and pattern recognition (such as the clustering problem described above) rely on a metric to measure the distance between data points. Choosing an appropriate metric for the problem at hand is crucial to the performance of these methods. Motivated by a variety of applications ranging from computer vision to information retrieval through bioinformatics, metric learning aims at adapting the metric to the data and has attracted a lot of interest in recent years (see for instance Bellet et al., 2013, for an account of metric learning and its applications). As an illustration, we consider the metric learning problem for supervised classification. In this setting, we observe independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of a random couple (X, Y), where the r.v. X takes values in some feature space $\mathcal{X}$ and Y in a finite set of labels, $\mathcal{Y} = \{1, \ldots, C\}$ with $C \geq 2$ say. Consider a set $\mathcal{D}$ of distance measures $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. Roughly speaking, the goal of metric learning in this context is to find a metric under which pairs of points with the same label are close to each other and those with different labels are far away. The risk of a metric D can be expressed as:

$$R(D) = \mathbb{E}\left[\phi\left((1 - D(X, X') \cdot (2\mathbb{I}\{Y = Y'\} - 1)\right)\right], \tag{8}$$

where $\phi(u)$ is a convex loss function upper bounding the indicator function $\mathbb{I}\{u \geq 0\}$, such as the hinge loss $\phi(u) = \max(0, 1-u)$. The natural empirical estimator of this risk is

$$R_n(D) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \phi\left((D(X_i, X_j) - 1) \cdot (2\mathbb{I}\{Y_i = Y_j\} - 1)\right), \tag{9}$$

which is a one sample U-statistic of degree two with kernel given by:

$$H_D\left((x, y), (x', y)\right) = \phi\left((D(x, x') - 1) \cdot (2\mathbb{I}\{y = y'\} - 1)\right). \tag{10}$$

The convergence to (8) of a minimizer of (9) has been studied in the frameworks of algorithmic stability (Jin et al., 2009), algorithmic robustness (Bellet and Habrard, 2015) and based on the theory of U-processes under appropriate regularization (Cao et al., 2012).

2.2.3 Multipartite Ranking

Given objects described by a random vector of attributes/features $X \in \mathcal{X}$ and the (temporarily hidden) ordinal labels $Y \in \{1, \ldots, K\}$ assigned to it, the goal of *multipartite ranking* is to rank them in the same order as that induced by the labels, on the basis of a training set of labeled examples. This statistical learning problem finds many applications in a wide range of fields (*e.g.* medicine, finance, search engines, e-commerce). Rankings are generally defined by means of a scoring function $s : \mathcal{X} \to \mathbb{R}$, transporting the natural order on the real line onto the feature space and the gold standard for evaluating the ranking performance of $s(x)$ is the ROC manifold, or its usual summary the VUS criterion (VUS standing for *Volume Under the* ROC *Surface*), see Clémençon and Robbiano (2014) and the references therein. In Clémençon et al. (2013), optimal scoring functions have been characterized as those that are optimal for all bipartite subproblems. In other words, they are increasing transforms of the likelihood ratio dF_{k+1}/dF_k, where F_k denotes the class-conditional distribution for the k-th class. When the set of optimal scoring functions is non-empty, the authors also showed that it corresponds to the functions which maximize the volume under the ROC surface

$$VUS(s) = \mathbb{P}\{s(X_1) < \ldots < s(X_K) | Y_1 = 1, \ldots, Y_K = K\}.$$

Given K independent samples $(X_1^{(k)}, \ldots, X_{n_k}^{(k)}) \overset{\text{i.i.d.}}{\sim} F_k(dx)$ for $k = 1, \ldots, K$, the empirical counterpart of the VUS can be written in the following way:

$$\widehat{VUS}(s) = \frac{1}{\prod_{k=1}^K n_k} \sum_{i_1=1}^{n_1} \cdots \sum_{i_K=1}^{n_K} \mathbb{I}\{s(X_{i_1}^{(1)}) < \ldots < s(X_{i_K}^{(K)})\}. \tag{11}$$

The empirical VUS (11) is a K-sample U-statistic of degree $(1, \ldots, 1)$ with kernel given by:

$$H_s(x_1, \ldots, x_K) = \mathbb{I}\{s(x_1) < \ldots < s(x_K)\}. \tag{12}$$

2.3 Empirical Minimization of U-Statistics

As illustrated by the examples above, many learning problems can be formulated as finding a certain rule g in a class $\mathcal{G}$ in order to minimize a risk of the same form as (2), $\mu(H_g)$, with

kernel $H = H_g$. Based on $K \geq 1$ independent i.i.d. samples

$$\mathbf{X}^{(k)}_{\{1,\,\ldots,\,n_k\}} = (X^{(k)}_1,\,\ldots,\,X^{(k)}_{n_k}) \text{ with } 1 \leq k \leq K,$$

the ERM paradigm in statistical learning suggests to replace the risk by the U-statistic estimation $U_{\mathbf{n}}(H_g)$ in the minimization problem. The study of the performance of minimizers $\widehat{g}_{\mathbf{n}}$ of the empirical estimate $U_{\mathbf{n}}(H_g)$ over the class $\mathcal{G}$ of rule candidates naturally leads to analyze the fluctuations of the U-process

$$\{U_{\mathbf{n}}(H_g) - \mu(H_g) :\ g \in \mathcal{G}\}. \tag{13}$$

Given the bound

$$\mu(H_{\widehat{g}_{\mathbf{n}}}) - \inf_{g\in\mathcal{G}} \mu(H_g) \leq 2 \sup_{g\in\mathcal{G}} |U_{\mathbf{n}}(H_g) - \mu(H_g)|, \tag{14}$$

a probabilistic control of the maximal deviation $\sup_{g\in\mathcal{G}} |U_{\mathbf{n}}(H_g) - \mu(H_g)|$ naturally provides statistical guarantees for the generalization ability of the empirical minimizer $\widehat{g}_{\mathbf{n}}$. As shown at length in the case $K = 1$ and $d_1 = 2$ in Clémençon et al. (2008) and in Clémençon (2014) for specific problems, this can be achieved under adequate complexity assumptions of the class $\mathcal{H}_{\mathcal{G}} = \{H_g :\ g \in \mathcal{G}\}$. These results rely on the *Hoeffding's representation* of U-statistics, which we recall now for clarity in the general multisample U-statistics setting. Denote by $\mathfrak{S}_m$ the symmetric group of order m for any $m \geq 1$ and by $\sigma(i)$ the i-th coordinate of any permutation $\sigma \in \mathfrak{S}_m$ for $1 \leq i \leq m$. Let $\lfloor z \rfloor$ be the integer part of any real number z and set

$$N = \min\{\lfloor n_1/d_1 \rfloor,\,\ldots,\,\lfloor n_K/d_K \rfloor\}.$$

Observe that the K-sample U-statistic (1) can be expressed as

$$U_{\mathbf{n}}(H) = \frac{1}{\prod_{k=1}^K n_k!} \sum_{\sigma_1\in\mathfrak{S}_{n_1}} \cdots \sum_{\sigma_K\in\mathfrak{S}_{n_K}} V_H\left(X^{(1)}_{\sigma_1(1)},\,\ldots,\,X^{(K)}_{\sigma_K(n_K)}\right), \tag{15}$$

where

$$\begin{aligned} V_H\left(X^{(1)}_1,\,\ldots,\,X^{(1)}_{n_1},\,\ldots,\,X^{(K)}_1,\,\ldots,\,X^{(K)}_{n_K}\right) = \frac{1}{N}\Big[& H\left(X^{(1)}_1,\,\ldots,\,X^{(1)}_{d_1},\,\ldots,\,X^{(K)}_1,\,\ldots,\,X^{(K)}_{d_K}\right) \\ & + H\left(X^{(1)}_{d_1+1},\,\ldots,\,X^{(1)}_{2d_1},\,\ldots,\,X^{(K)}_{d_K+1},\,\ldots,\,X^{(K)}_{2d_K}\right) + \ldots \\ & + H\left(X^{(1)}_{(N-1)d_1+1},\,\ldots,\,X^{(1)}_{Nd_1},\,\ldots,\,X^{(K)}_{(N-1)d_K+1},\,\ldots,\,X^{(K)}_{Nd_K}\right)\Big]. \end{aligned}$$

This representation, sometimes referred to as the *first Hoeffding's decomposition* (see Hoeffding, 1948), allows to reduce a first order analysis to the case of sums of i.i.d. random variables. The following result extends Corollary 3 in Clémençon et al. (2008) to the multisample situation.

Proposition 2 *Let $\mathcal{H}$ be a collection of bounded symmetric kernels on $\prod_{k=1}^K \mathcal{X}_k^{d_k}$ such that*

$$M_{\mathcal{H}} \stackrel{\text{def}}{=} \sup_{(H,x)\in\mathcal{H}\times\mathcal{X}} |H(x)| < +\infty. \tag{16}$$

Suppose also that $\mathcal{H}$ is a VC *major class of functions with finite Vapnik-Chervonenkis dimension $V < +\infty$. For all $\delta \in (0,1)$, we have with probability at least $1-\delta$,*

$$\sup_{H \in \mathcal{H}} |U_{\mathbf{n}}(H) - \mu(H)| \leq \mathcal{M}_{\mathcal{H}} \left\{ 2\sqrt{\frac{2V\log(1+N)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} \right\}, \tag{17}$$

where $N = \min\{\lfloor n_1/d_1 \rfloor, \ldots, \lfloor n_K/d_K \rfloor\}$.

Observe that, in the usual asymptotic framework (4), the bound (17) shows that the learning rate is, as expected, of order $O_{\mathbb{P}}(\sqrt{\log n/n})$, where n denotes the size of the pooled sample.

Remark 3 (Uniform boundedness) *We point out that condition* (16) *is clearly satisfied for the class of kernels considered in the multipartite ranking situation, whatever the class of scoring functions considered. In the case of the clustering example, it is fulfilled as soon as the essential supremum of $D(X,X') \cdot \Phi_{\mathcal{P}}(X,X')$ is uniformly bounded over $\mathcal{P} \in \Pi$, whereas in the metric learning example, it is satisfied when the essential supremum of the r.v. $\phi((D(X,X')-1)\cdot(2\mathbb{I}\{Y=Y'\}-1))$ is uniformly bounded over $D \in \mathcal{D}$. We underline that this simplifying condition can be easily relaxed and replaced by appropriate tail assumptions for the variables $H(X_1^{(1)}, \ldots, X_{d_K}^{(K)})$, $H \in \mathcal{H}$, combining the arguments of the subsequent analysis with the classical "truncation trick" originally introduced in Fuk and Nagaev (1971).*

Remark 4 (Complexity assumptions) *Following in the footsteps of Clémençon et al. (2008) which considered 1-sample U-statistics of degree 2, define the Rademacher average*

$$\mathcal{R}_N = \sup_{H \in \mathcal{H}} \frac{1}{N} \left| \sum_{l=1}^{N} \epsilon_l H\left(X_{(l-1)d_1+1}^{(1)}, \ldots, X_{ld_1}^{(1)}, \ldots, X_{(l-1)d_K+1}^{(K)}, \ldots, X_{ld_K}^{(K)}\right) \right|, \tag{18}$$

where $\epsilon_1, \ldots, \epsilon_N$ are independent Rademacher random variables (random symmetric sign variables), independent from the $X_i^{(k)}$'s. As can be seen by simply examining the proof of Proposition 2 (Appendix A), a control of the maximal deviations similar to (17) *relying on this particular complexity measure can be obtained: the first term on the right hand side is then replaced by the expectation of the Rademacher average $\mathbb{E}[\mathcal{R}_N]$, up to a constant multiplicative factor. This expected value can be bounded by standard metric entropy techniques and in the case where $\mathcal{H}$ is a* VC *major class of functions of dimension V, we have:*

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

See Appendix A for further details.

3. Empirical Minimization of Incomplete U-Statistics

We have seen in the last section that the empirical minimization of U-statistics leads to a learning rate of $O_{\mathbb{P}}(\sqrt{\log n/n})$. However, the computational cost required to find the empirical minimizer in practice is generally prohibitive, as the number of terms to be summed up to compute the U-statistic (1) is equal to:

$$\binom{n_1}{d_1} \times \cdots \times \binom{n_K}{d_K}.$$

In the usual asymptotic framework (4), it is of order $O(n^{d_1+\ldots+d_K})$ as $n \to +\infty$. It is the major purpose of this section to show that, in the minimization problem, the U-statistic $U_{\mathbf{n}}(h_g)$ can be replaced by a Monte-Carlo estimation, referred to as an *incomplete* U-*statistic*, whose computation requires to average much less terms, without damaging the learning rate (Section 3.1). We further extend these results to model selection (Section 3.2), fast rates situations (Section 3.3) and alternative sampling strategies (Section 3.4).

3.1 Uniform Approximation of Generalized U-Statistics

As a remedy to the computational issue mentioned above, the concept of *incomplete generalized* U-*statistic* has been introduced in the seminal contribution of Blom (1976). The calculation of such a functional involves a summation 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 independently with replacement*, leading to the following definition.

Definition 5 (Incomplete Generalized U-statistic) *Let* $B \geq 1$. *The incomplete version of the* U-*statistic* (1) *based on* B *terms is defined by:*

$$\widetilde{U}_B(h) = \frac{1}{B} \sum_{I=(I_1,\ldots,I_K)\in\mathcal{D}_B} h(\mathbf{X}_{I_1}^{(1)}, \ldots, \mathbf{X}_{I_K}^{(K)}) = \frac{1}{B} \sum_{I\in\mathcal{D}_B} h(\mathbf{X}_I), \tag{19}$$

where $\mathcal{D}_B$ *is a set of cardinality* B *built by sampling with replacement in the set*

$$\Lambda = \{((i_1^{(1)}, \ldots, i_{d_1}^{(1)}), \ldots, (i_1^{(K)}, \ldots, i_{d_K}^{(K)})) : 1 \leq i_1^{(k)} < \ldots < i_{d_k}^{(k)} \leq n_k,\ 1 \leq k \leq K\}, \tag{20}$$

and $\mathbf{X}_I = (\mathbf{X}_{I_1}^{(1)}, \ldots, \mathbf{X}_{I_K}^{(K)})$ *for all* $I = (I_1, \ldots, I_K) \in \Lambda$.

We stress that the distribution of a complete U-statistic built from subsamples of reduced sizes n_k' drawn uniformly at random is quite different from that of an incomplete U-statistic based on $B = \prod_{k=1}^K \binom{n_k'}{d_k}$ terms sampled with replacement in Λ, although they involve the summation of the same number of terms, as depicted by Fig. 1.

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 the fact 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: the focus is on the number of terms involved in the summation solely. As an estimator of $\mu(h)$, the statistic (19) is still unbiased, *i.e.* $\mathbb{E}[\widetilde{U}_B(h)] = \mu(h)$, but its variance is naturally larger than that of the complete U-statistic $U_n(h)$. Precisely, writing the variance of the r.v. $\widetilde{U}_B(h)$ as the expectation of its conditional variance given $(\mathbf{X}_I)_{I\in\Lambda}$ plus the variance of its conditional expectation given $(\mathbf{X}_I)_{I\in\Lambda}$, we obtain

$$\mathrm{Var}(\widetilde{U}_B(h)) = \left(1 - \frac{1}{B}\right) \mathrm{Var}(U_{\mathbf{n}}(h)) + \frac{1}{B}\mathrm{Var}(h(X_1^{(1)}, \ldots, X_{d_K}^{(K)})). \tag{21}$$

One may easily check that $\mathrm{Var}(\widetilde{U}_B(h)) \geq \mathrm{Var}(U_{\mathbf{n}}(h))$, and the difference vanishes as B increases. Refer to Lee (1990) for further details (see p. 193 therein). Incidentally,

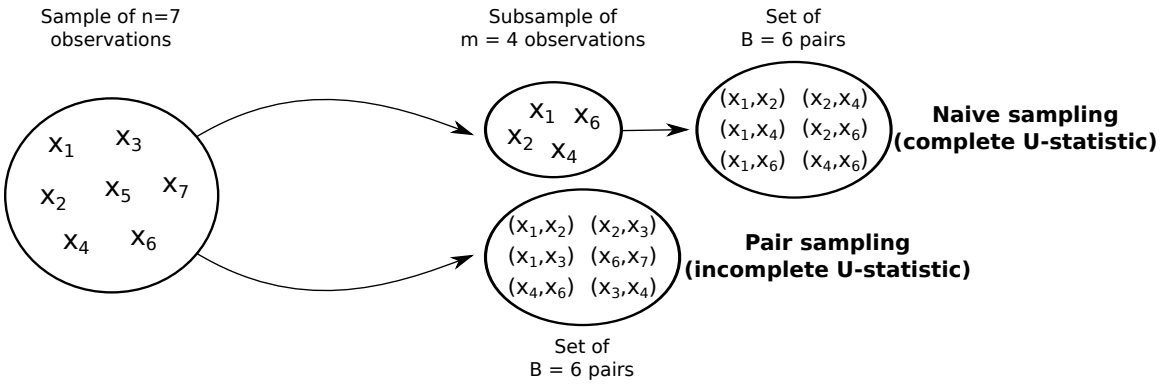


Figure 1: Illustration of the difference between an incomplete U-statistic and a complete U-statistic based on a subsample. For simplicity, we focus on the case $K = 1$ and $d_1 = 2$. In this simplistic example, a sample of $n = 7$ observations is considered. To construct a complete U-statistic of reduced complexity, we first sample a set of $m = 4$ observations and then form all possible pairs from this subsample, *i.e.* $B = m(m-1)/2 = 6$ pairs in total. In contrast, an incomplete U-statistic with the same number of terms is obtained by sampling B pairs directly from the set Λ of all possible pairs based on the original statistical population.

we underline that the empirical variance of (19) is not easy to compute either since it involves summing approximately $\#\Lambda$ terms and bootstrap techniques should be used for this purpose, as proposed in Bertail and Tressou (2006). The asymptotic properties of incomplete U-statistics have been investigated in several articles, see Janson (1984); Brown and Kildea (1978); Enqvist (1978). The angle embraced in the present paper is of very different nature: the key idea we promote here is to use incomplete versions of collections of U-statistics in learning problems such as that described in Section 2.2. The result stated below shows that this approach solves the numerical problem, while not damaging the learning rates under appropriate complexity assumptions on the collection $\mathcal{H}$ of (symmetric) kernels H considered, the complexity being described here in terms of VC dimension for simplicity. In particular, it reveals that concentration results established for U-processes (*i.e.* collections of U-statistics) such as Proposition 2 may extend to their incomplete versions, as shown by the following theorem.

Theorem 6 (Maximal deviation) *Let $\mathcal{H}$ be a collection of bounded symmetric kernels on $\prod_{k=1}^{K} \mathcal{X}_k^{d_k}$ that fulfills the assumptions of Proposition 2. Then, the following assertions hold true.*

(i) For all $\delta \in (0,1)$, with probability at least $1-\delta$, we have: $\forall \mathbf{n} = (n_1, \ldots, n_K) \in \mathbb{N}^{*K}$, $\forall B \geq 1$,

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

(ii) For all $\delta \in (0,1)$*, with probability at least* $1-\delta$*, we have:* $\forall \mathbf{n} \in \mathbb{N}^{*K}$, $\forall B \geq 1$,

$$\frac{1}{\mathcal{M}_{\mathcal{H}}} \sup_{H \in \mathcal{H}} \left| \widetilde{U}_B(H) - \mu(H) \right| \leq 2\sqrt{\frac{2V\log(1+N)}{N}} + \sqrt{\frac{\log(2/\delta)}{N}} + \sqrt{2\frac{V\log(1+\#\Lambda)+\log(4/\delta)}{B}},$$

where $N = \min\{\lfloor n_1/d_1 \rfloor, \ldots, \lfloor n_K/d_K \rfloor\}$.

Remark 7 (Complexity assumptions continued) *We point out that a bound of the same order as that stated above can be obtained under standard metric entropy conditions by means of classical chaining arguments, or under the assumption that the Rademacher average defined by*

$$\widetilde{\mathcal{R}}_B = \sup_{H \in \mathcal{H}} \frac{1}{B} \left| \sum_{b=1}^{B} \epsilon_b \left\{ \sum_{I \in \Lambda} \zeta_b(I) H(\mathbf{X}_I) \right\} \right| \tag{22}$$

has an expectation of the order $O(1/\sqrt{B})$*. The quantity* $\zeta_b(I)$ *indicates whether the subset of indexes* I *has been picked at the* b*-th draw (*$\zeta_b(I) = +1$*) or not (*$\zeta_b(I) = 0$*), see the calculation at the end of Appendix C. Equipped with this notation, notice that the* ζ_b*'s are i.i.d. multinomial random variables such that* $\sum_{I \in \Lambda} \zeta_b(I) = +1$*. This assumption can be easily shown to be fulfilled in the case where* $\mathcal{H}$ *is a* VC *major class of finite* VC *dimension (see the proof of Theorem 6 in Appendix B). Notice however that although the variables* $\sum_{I \in \Lambda} \zeta_b(I) H(\mathbf{X}_I)$*,* $1 \leq b \leq B$*, are conditionally i.i.d. given* $(\mathbf{X}_I)_{I \in \Lambda}$*, they are not independent and the quantity* (22) *cannot be related to complexity measures of the type* (18) *mentioned in Remark 4.*

Remark 8 *We underline that, whereas* $\sup_{H \in \mathcal{H}} |U_{\mathbf{n}}(H) - \mu(H)|$ *can be proved to be of order* $O_{\mathbb{P}}(1/n)$ *under adequate complexity assumptions in the specific situation where* $\{U_{\mathbf{n}}(H) : H \in \mathcal{H}\}$ *is a collection of degenerate* U*-statistics (see Section 3.3), the bound* (i) *in Theorem 6 cannot be improved in the degenerate case. Observe indeed that, conditioned upon the observations* $X_i^{(k)}$*, the deviations of the approximation* (19) *from its mean are of order* $O_{\mathbb{P}}(1/\sqrt{B})$*, since it is a basic average of* B *i.i.d. terms.*

From the theorem stated above, one may straightforwardly deduce a bound on the excess risk of kernels $\widehat{H}_B$ minimizing the incomplete version of the empirical risk based on B terms, *i.e.* such that

$$\widetilde{U}_B\left(\widehat{H}_B\right) = \min_{H \in \mathcal{H}} \widetilde{U}_B(H). \tag{23}$$

Corollary 9 *Let* $\mathcal{H}$ *be a collection of symmetric kernels on* $\prod_{k=1}^{K} \mathcal{X}_k^{d_k}$ *that satisfies the conditions stipulated in Proposition 2. Let* $\delta > 0$*. For any minimizer* $\widehat{H}_B$ *of the statistical estimate of the risk* (19)*, the following assertions hold true*

(i) We have with probability at least $1-\delta$*:* $\forall \mathbf{n} \in \mathbb{N}^{*K}$, $\forall B \geq 1$,

$$\mu(\widehat{H}_B) - \inf_{H \in \mathcal{H}} \mu(H) \leq 2\mathcal{M}_{\mathcal{H}} \times \left\{ 2\sqrt{\frac{2V\log(1+N)}{N}} + \sqrt{\frac{\log(2/\delta)}{N}} + \sqrt{2\frac{V\log(1+\#\Lambda)+\log(4/\delta)}{B}} \right\}.$$

Empirical risk criterion	Nb of terms	Rate bound
Complete U-statistic	$O(n^{d_1+\ldots+d_K})$	$O_{\mathbb{P}}(\sqrt{\log(n)/n})$
Complete U-statistic based on subsamples	$O(n)$	$O_{\mathbb{P}}\left(\sqrt{\log(n)/n^{\frac{1}{d_1+\ldots+d_K}}}\right)$
Incomplete U-statistic (our result)	$\mathbf{O(n)}$	$\mathbf{O_{\mathbb{P}}(\sqrt{\log(n)/n})}$

Table 1: Rate bound for the empirical minimizer of several empirical risk criteria *versus* the number of terms involved in the computation of the criterion. For a computational budget of $O(n)$ terms, the rate bound for the incomplete U-statistic criterion is of the same order as that of the complete U-statistic, which is a huge improvement over a complete U-statistic based on a subsample.

(ii) We have: $\forall \mathbf{n} \in \mathbb{N}^{*K}$, $\forall B \geq 1$,

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\widetilde{U}_B(h)-\mu(h)\right|\right] \leq \mathcal{M}_{\mathcal{H}}\left\{2\sqrt{\frac{2V\log(1+N)}{N}}+\sqrt{\frac{2(\log 2+V\log(1+\#\Lambda))}{B}}\right\}.$$

The first assertion of Theorem 6 provides a control of the deviations between the U-statistic (1) and its incomplete counterpart (19) uniformly over the class $\mathcal{H}$. As the number of terms B increases, this deviation decreases at a rate of $O(1/\sqrt{B})$. The second assertion of Theorem 6 gives a maximal deviation result with respect to $\mu(h)$. Observe in particular that, with the asymptotic settings previously specified, $N = O(n)$ and $\log(\#\Lambda) = O(\log n)$ as $n \to +\infty$. The bounds stated above thus show that, for a number $B = B_n$ of terms tending to infinity at a rate $O(n)$ as $n \to +\infty$, the maximal deviation $\sup_{h\in\mathcal{H}} |\widetilde{U}_B(h) - \mu(h)|$ is asymptotically of the order $O_{\mathbb{P}}((\log(n)/n)^{1/2})$, just like $\sup_{h\in\mathcal{H}} |U_{\mathbf{n}}(h) - \mu(h)|$, see bound (17) in Proposition 2. In short, when considering an incomplete U-statistic (19) with $B = O(n)$ terms only, the learning rate for the corresponding minimizer is of the same order as that of the minimizer of the complete risk (1), whose computation requires to average $\#\Lambda = O(n^{d_1+\ldots+d_K})$ terms. Minimizing such incomplete U-statistics thus yields a significant gain in terms of computational cost while fully preserving the learning rate. In contrast, as implied by Proposition 2, the minimization of a complete U-statistic involving $O(n)$ terms, obtained by drawing subsamples of sizes $n'_k = O(n^{1/(d_1+\ldots+d_K)})$ uniformly at random, leads to a rate of convergence of $O(\sqrt{\log(n)/n^{1/(d_1+\ldots+d_K)}})$, which is much slower except in the trivial case where $K = 1$ and $d_1 = 1$. These striking results are summarized in Table 1.

The important practical consequence of the above is that when n is too large for the complete risk (1) to be used, one should instead use the incomplete risk (19) (setting the number of terms B as large as the computational budget allows).

3.2 Model Selection Based on Incomplete U-Statistics

Automatic selection of the model complexity is a crucial issue in machine learning: it includes the number of clusters in cluster analysis (see Clémençon, 2014) or the choice of the number of possible values taken by a piecewise constant scoring function in multipartite ranking for instance (*cf.* Clémençon and Vayatis, 2009). In the present situation, this boils down to choosing the adequate level of complexity of the class of kernels $\mathcal{H}$, measured through its (supposedly finite) VC dimension for simplicity, in order to minimize the (theoretical) risk of the empirical minimizer. It is the purpose of this subsection to show that the incomplete U-statistic (19) can be used to define a penalization method to select a prediction rule with nearly minimal risk, avoiding procedures based on data splitting/resampling and extending the celebrated *structural risk minimization* principle, see Vapnik (1999). Let $\mathcal{H}$ be the collection of all symmetric kernels on $\prod_{k=1}^{K} \mathcal{X}_k^{d_k}$ and set $\mu^* = \inf_{H\in\mathcal{H}} \mu(H)$. Let $\mathcal{H}_1, \mathcal{H}_2, \ldots$ be a sequence of uniformly bounded major subclasses of $\mathcal{H}$, of increasing complexity (VC dimension). For any $m \geq 1$, let V_m denote the VC dimension of the class $\mathcal{H}_m$ and set $\mathcal{M}_{\mathcal{H}_m} = \sup_{(H,x)\in\mathcal{H}_m\times\mathcal{X}} |H(x)| < +\infty$. We suppose that there exists $\mathcal{M} < +\infty$ such that $\sup_{m\geq 1} \mathcal{M}_{\mathcal{H}_m} \leq \mathcal{M}$. Given $1 \leq B \leq \#\Lambda$ and $m \geq 1$, the complexity penalized empirical risk of a solution $\widetilde{U}_{B,m}$ of the ERM problem (23) with $\mathcal{H} = \mathcal{H}_m$ is

$$\widetilde{U}_B(\widehat{H}_{B,m}) + \mathrm{pen}(B, m), \tag{24}$$

where the quantity $\mathrm{pen}(B, m)$ is a *distribution free* penalty given by:

$$\begin{aligned} \mathrm{pen}(B, m) &= 2\mathcal{M}_{\mathcal{H}_m} \left\{ \sqrt{\frac{2V_m \log(1+N)}{N}} + \sqrt{\frac{2(\log 2 + V_m \log(1+\#\Lambda))}{B}} \right\} \\ &+ 2\mathcal{M}\sqrt{\frac{(B+n)\log m}{B^2}}. \end{aligned} \tag{25}$$

As shown in Assertion (ii) of Corollary 9, the quantity above is an upper bound for the expected maximal deviation $\mathbb{E}[\sup_{H\in\mathcal{H}_m} |\widetilde{U}_B(H) - \mu(H)|]$ and is thus a natural penalty candidate to compensate the overfitting within class $\mathcal{H}_m$. We thus propose to select

$$\widehat{m}_B = \arg\min_{m\geq 1} \left\{ \widetilde{U}_B(\widehat{H}_{B,m}) + \mathrm{pen}(B, m) \right\}. \tag{26}$$

As revealed by the theorem below, choosing $B = O(n)$, the prediction rule $\widehat{H}_{\widehat{m}_B}$ based on a penalized criterion involving the summation of $O(n)$ terms solely, achieves a nearly optimal trade-off between the bias and the distribution free upper bound (25) on the variance term.

Theorem 10 (Oracle inequality) *Suppose that Theorem 6's assumptions are fulfilled for all* $m \geq 1$ *and that* $\sup_{m\geq 1} \mathcal{M}_{\mathcal{H}_m} \leq \mathcal{M} < +\infty$. *Then, we have:* $\forall \mathbf{n} \in \mathbb{N}^{*K}$, $\forall B \in \{1, \ldots, \#\Lambda\}$,

$$\mu(\widehat{H}_{B,\widehat{m}}) - \mu^* \leq \inf_{k\geq 1} \left\{ \inf_{H\in\mathcal{H}_m} \mu(H) - \mu^* + \mathrm{pen}(B, m) \right\} + \mathcal{M}\frac{\sqrt{2\pi(B+n)}}{B}.$$

We point out that the argument used to obtain the above result can be straightforwardly extended to other (possibly data-dependent) complexity penalties (*cf.* Massart, 2006), see the proof in Appendix D.

3.3 Fast Rates for ERM of Incomplete U-Statistics

In Clémençon et al. (2008), it has been proved that, under certain "low-noise" conditions, the minimum variance property of the U-statistics used to estimate the ranking risk (corresponding to the situation $K = 1$ and $d_1 = 2$) leads to learning rates faster than $O_{\mathbb{P}}(1/\sqrt{n})$. These results rely on the *Hajek projection*, a linearization technique originally introduced in Hoeffding (1948) for the case of one sample U-statistics and next extended to the analysis of a much larger class of functionals in Hájek (1968). It consists in writing $U_{\mathbf{n}}(H)$ as the sum of the orthogonal projection

$$\widehat{U}_{\mathbf{n}}(H) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathbb{E}\left[U_{\mathbf{n}}(H) \mid X_i^{(k)}\right] - (n-1)\mu(H), \tag{27}$$

which is itself a sum of K independent basic sample means based on i.i.d. r.v.'s (of the order $O_{\mathbb{P}}(1/\sqrt{n})$ each, after recentering), plus a possible negligible term. This representation was used for instance by Grams and Serfling (1973) to refine the CLT in the multisample U-statistics framework. Although useful as a theoretical tool, it should be noticed that the quantity $\widehat{U}_{\mathbf{n}}(H)$ is not of practical interest, since the conditional expectations involved in the summation are generally unknown.

Although incomplete U-statistics do not share the minimum variance property (see Section 3.1), we will show that the same fast rate bounds for the excess risk as those reached by ERM of U-statistics (corresponding to the summation of $O(n^2)$ pairs of observations) can be attained by empirical ranking risk minimizers, when estimating the ranking risk by incomplete U-statistics involving the summation of $o(n^2)$ terms solely.

For clarity (and comparison purpose), we first recall the statistical learning framework considered in Clémençon et al. (2008). Let (X, Y) be a pair of random variables defined on the same probability space, where Y is a real-valued label and X models some input information taking its values in a measurable space $\mathcal{X}$ hopefully useful to predict Y. Denoting by (X', Y') an independent copy of the pair (X, Y). The goal pursued here is to learn how to rank the input observations X and X', by means of an antisymmetric *ranking rule* $r : \mathcal{X}^2 \to \{-1, +1\}$ (*i.e.* $r(x, x') = -r(x'x)$ for any $(x, x') \in \mathcal{X}^2$), so as to minimize the *ranking risk*

$$L(r) = \mathbb{P}\{(Y - Y') \cdot r(X, X') < 0\}. \tag{28}$$

The minimizer of the ranking risk is the ranking rule $r^*(X, X') = 2\mathbb{I}\{\mathbb{P}\{Y > Y' \mid (X, X')\} \geq \mathbb{P}\{Y < Y' \mid (X, X')\} - 1$ (see Proposition 1 in Clémençon et al., 2008). The natural empirical counterpart of (28) based on a sample of independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of the pair (X, Y) is the 1-sample U-statistic $U_n(H_r)$ of degree two with kernel $H_r((x, y), (x', y')) = \mathbb{I}\{(y - y') \cdot r(x, x') < 0\}$ for all (x, y) and (x', y')) in $\mathcal{X} \times \mathbb{R}$ given by:

$$L_n(r) = U_n(H_r) = \frac{2}{n(n-1)} \sum_{i<j} \mathbb{I}\{(Y_i - Y_j) \cdot r(X_i, X_j) < 0\}. \tag{29}$$

Equipped with these notations, a statistical version of the excess risk $\Lambda(r) = L(r) - L(r^*)$ is a U-statistic $\lambda_n(r)$ with kernel $q_r = H_r - H_{r^*}$. The key "noise-condition", which allows to exploit the Hoeffding/Hajek decomposition of $\Lambda_n(r)$, is stated below.

Assumption 1 *There exist constants* $c > 0$ *and* $\alpha \in [0, 1]$ *such that:*

$$\forall r \in \mathcal{R}, \quad \mathbb{V}\mathrm{ar}\left(h_r(X, Y)\right) \leq c\Lambda(r)^{\alpha},$$

where we set $h_r(x, y) = \mathbb{E}[q_r((x, y), (X', Y'))]$.

Recall incidentally that very general sufficient conditions guaranteeing that this assumption holds true have been exhibited, see Section 5 in Clémençon et al. (2008) (notice that the condition is void for $\alpha = 0$). Since our goal is to explain the main ideas rather than achieving a high level of generality, we consider a very simple setting, stipulating that the cardinality of the class of ranking rule candidates $\mathcal{R}$ under study is finite, $\#\mathcal{R} = M < +\infty$, and that the optimal rule r^* belongs to $\mathcal{R}$. The following proposition is a simplified version of the fast rate result proved in Clémençon et al. (2008) for the empirical minimizer $\widehat{r}_n = \arg\min_{r \in \mathcal{R}} L_n(r)$.

Proposition 11 *(Clémençon et al. (2008),* Corollary 6*) Suppose that Assumption 1 is fulfilled. Then, there exists a universal constant* $C > 0$ *such that for all* $\delta \in (0, 1)$*, we have:* $\forall n \geq 2$*,*

$$L(\widehat{r}_n) - L(r^*) \leq C \left(\frac{\log(M/\delta)}{n} \right)^{\frac{1}{2-\alpha}}. \tag{30}$$

Consider now the minimizer $\widetilde{r}_B$ of the incomplete U-statistic risk estimate

$$\widetilde{U}_B(H_r) = \frac{1}{B} \sum_{k=1}^{B} \sum_{(i,j):1 \leq i < j \leq n} \epsilon_k((i, j)) \mathbb{I}\{(Y_i - Y_j) \cdot r(X_i, X_j) < 0\} \tag{31}$$

over $\mathcal{R}$, where $\epsilon_k((i, j))$ indicates whether the pair (i, j) has been picked at the k-th draw ($\epsilon_k((i, j)) = 1$ in this case, which occurs with probability $1/\binom{n}{2}$) or not (then, we set $\epsilon_k((i, j)) = 0$). Observe that $\widetilde{r}_B$ also minimizes the empirical estimate of the excess risk $\widetilde{\Lambda}_B(r) = \widetilde{U}_B(q_r)$ over $\mathcal{R}$.

Theorem 12 *Let* $\alpha \in [0, 1]$ *and suppose that Assumption 1 is fulfilled. If we set* $B = O(n^{2/(2-\alpha)})$*, there exists some constant* $C < +\infty$ *such that, for all* $\delta \in (0, 1)$*, we have with probability at least* $1 - \delta$*:* $\forall n \geq 2$*,*

$$L(\widetilde{r}_B) - L(r^*) \leq C \left(\frac{\log(M/\delta)}{n} \right)^{\frac{1}{2-\alpha}}.$$

As soon as $\alpha < 1$, this result shows that the same fast rate of convergence as that reached by $\widehat{r}_n$ can be attained by the ranking rule $\widetilde{r}_B$, which minimizes an empirical version of the ranking risk involving the summation of $O(n^{2/(2-\alpha)})$ terms solely. For comparison purpose, minimization of the criterion (28) computed with a number of terms of the same order leads to a rate bound of order $O_{\mathbb{P}}(n^{1/(2-\alpha)^2})$.

Finally, we point out that fast rates for the clustering problem have been also investigated in Clémençon (2014), see Section 5.2 therein. The present analysis can be extended to the clustering framework by means of the same arguments.

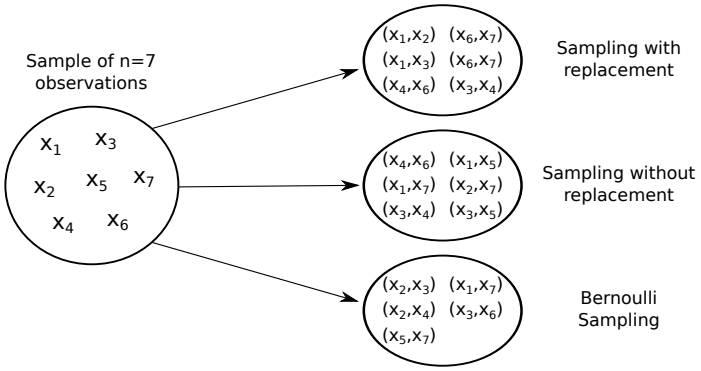


Figure 2: Illustration of different sampling schemes for approximating a U-statistic. For simplicity, consider again the case $K = 1$ and $d_1 = 2$. Here $n = 7$ and the expected number of terms is $B = 6$. Sampling with or without replacement results in exactly B terms, with possible repetitions when sampling with replacement, *e.g.* (x_6, x_7) in this example. In contrast, Bernoulli sampling with $\pi_I = B/\#\Lambda$ results in B terms only in expectation, with individual realizations that may exhibit more or fewer terms.

3.4 Alternative Sampling Schemes

Sampling with replacement is not the sole way of approximating generalized U-statistics with a controlled computational cost. As proposed in Janson (1984), other sampling schemes can be considered, Bernoulli sampling or sampling without replacement in particular (see Figure 2 for an illustration). We now explain how the results of this paper can be extended to these situations. The population of interest is the set Λ and a *survey sample* of (possibly random) size $b \leq n$ is any subset s of cardinality $b = b(s)$ less than $\#\Lambda$ in the power set $\mathcal{P}(\Lambda)$. Here, a general *survey scheme without replacement* is any conditional probability distribution R on the set of all possible samples $s \in \mathcal{P}(\Lambda)$ given $(\mathbf{X}_I)_{I\in\Lambda}$. For any $I \in \Lambda$, the first order *inclusion probability* $\pi_I(R) = \mathbb{P}_R\{I \in S\}$, is the probability that the unit I belongs to a random sample S drawn from distribution R. We set $\boldsymbol{\pi}(R) = (\pi_I(R))_{I\in\Lambda}$. The second order inclusion probabilities are denoted by $\pi_{I,J}(R) = \mathbb{P}_R\{(I, J) \in S^2\}$ for any $I \neq J$ in Λ. When no confusion is possible, we omit to mention the dependence in R when writing the first/second order probabilities of inclusion. The information related to the observed sample $S \subset \Lambda$ is fully enclosed in the random vector $\boldsymbol{\Delta} = (\Delta(I))_{I\in\Lambda}$, where $\Delta(I) = \mathbb{I}\{I \in S\}$ for all $I \in \Lambda$. The 1-d marginal distributions of the sampling scheme $\boldsymbol{\Delta}_n$ are the Bernoulli distributions with parameters π_I, $I \in \Lambda$, and the covariance matrix of the r.v. $\boldsymbol{\Delta}_n$ is given by $\Gamma = \{\pi_{I,J} - \pi_I\pi_J\}_{I,J}$ with the convention $\pi_{I,I} = \pi_I$ for all $I \in \Lambda$. Observe that, equipped with the notations above, $\sum_{I\in\Lambda} \Delta(I) = b(S)$.

One of the simplest survey plans is the Poisson scheme (without replacement), for which the $\Delta(I)$'s are independent Bernoulli random variables with parameters π_I, $I \in \Lambda$, in $(0, 1)$. The first order inclusion probabilities fully characterize such a plan. Observe in addition that the size $b(\mathcal{S})$ of a sample generated this way is random with expectation $B = \mathbb{E}[b(S) \mid (\mathbf{X}_I)_{I\in\Lambda}] = \sum_{I\in\Lambda} \pi_I$. The situation where the π_I's are all equal corresponds to the Bernoulli

sampling scheme: $\forall I \in \Lambda$, $\pi_I = B/\#\Lambda$. The Poisson survey scheme plays a crucial role in sampling theory, inso far as a wide range of survey schemes can be viewed as conditional Poisson schemes, see Hájek (1964). For instance, one may refer to Cochran (1977) or Deville (1987) for accounts of survey sampling techniques.

Following in the footsteps of the seminal contribution of Horvitz and Thompson (1951), an estimate of (1) based on a sample drawn from a survey scheme R with first order inclusion probabilities $(\pi_I)_{I\in\Lambda}$ is given by:

$$\bar{U}_{HT}(H) = \frac{1}{\#\Lambda}\sum_{I\in\Lambda}\frac{\Delta(I)}{\pi_I}H(\mathbf{X}_I), \tag{32}$$

with the convention that $0/0 = 0$. Notice that it is an unbiased estimate of (1):

$$\mathbb{E}[\bar{U}_{HT}(H) \mid (\mathbf{X}_I)_{I\in\Lambda}] = U_{\mathbf{n}}(H).$$

In the case where the sample size is deterministic, its conditional variance is given by:

$$\mathrm{Var}(\bar{U}_{HT}(H) \mid (\mathbf{X}_I)_{I\in\Lambda}) = \frac{1}{2}\sum_{I\neq J}\left(\frac{H(\mathbf{X}_I)}{\pi_I} - \frac{H(\mathbf{X}_J)}{\pi_J}\right)^2(\pi_{I,J} - \pi_I\pi_J).$$

We point out that the computation of (32) involves summing over a possibly random number of terms, equal to $B = \mathbb{E}[b(S)] = \sum_{I\in\Lambda}\pi_I$ in average and whose variance is equal to $\mathrm{Var}(b(S)) = \sum_{I\in\Lambda}\pi_I(1-\pi_I) + \sum_{I\neq J}\{\pi_{I,J} - \pi_I\pi_J\}$.

Here, we are interested in the situation where the $\Delta(I)$'s are independent from $(X_I)_{I\in\Lambda}$, and either a sample of size $B \leq \#\Lambda$ fixed in advance is chosen uniformly at random among the $\binom{\#\Lambda}{B}$ possible choices (this survey scheme is sometimes referred to as *rejective sampling* with equal first order inclusion probabilities), or else it is picked by means of a Bernoulli sampling with parameter $B/\#\Lambda$. Observe that, in both cases, we have $\pi_I = B/\#\Lambda$ for all $I \in \Lambda$. The following theorem shows that in both cases, similar results as those obtained for *sampling with replacement* can be derived for minimizers of the Horvitz-Thompson risk estimate (32).

Theorem 13 *Let $\mathcal{H}$ be a collection of bounded symmetric kernels on $\prod_{k=1}^{K}\mathcal{X}_k^{d_k}$ that fulfills the assumptions involved in Proposition 2. Let $B \in \{1, \ldots, \#\Lambda\}$. Suppose that, for any $H \in \mathcal{H}$, $\bar{U}_{HT}(H)$ is the incomplete U-statistic based on either a Bernoulli sampling scheme with parameter $B/\#\Lambda$ or else a sampling without replacement scheme of size B. For all $\delta \in (0,1)$, we have with probability at least $1-\delta$: $\forall \mathbf{n} \in \mathbb{N}^{*K}$, $\forall B \in \{1, \ldots, \#\Lambda\}$,*

$$\sup_{H\in\mathcal{H}}\left|\bar{U}_{HT}(H) - U_{\mathbf{n}}(H)\right| \leq 2\mathcal{M}_{\mathcal{H}}\sqrt{\frac{\log(2(1+\#\Lambda)^V/\delta)}{B}} + \frac{2\log(2(1+\#\Lambda)^V/\delta)\mathcal{M}_{\mathcal{H}}}{3B},$$

in the case of the Bernoulli sampling design, and

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

in the case of the sampling without replacement design.

We highlight the fact that, from a computational perspective, sampling with replacement is undoubtedly much more advantageous than Bernoulli sampling or sampling without replacement. Indeed, although its expected value is equal to B, the size of a Bernoulli sample is stochastic and the related sampling algorithm requires a loop through the elements I of Λ and the practical implementation of sampling without replacement is generally based on multiple iterations of sampling with replacement, see Tillé (2006).

4. Application to Stochastic Gradient Descent for ERM

The theoretical analysis carried out in the preceding sections focused on the properties of empirical risk minimizers but ignored the issue of finding such a minimizer. In this section, we show that the sampling technique introduced in Section 3 also provides practical means of scaling up iterative statistical learning techniques. Indeed, large-scale training of many machine learning models, such as SVM, Deep Neural Networks or soft K-means among others, is based on stochastic gradient descent (SGD in abbreviated form), see Bottou (1998). When the risk is of the form (2), we now investigate the benefit of using, at each iterative step, a gradient estimate of the form of an incomplete U-statistic, instead of an estimate of the form of a complete U-statistic with exactly the same number of terms based on subsamples drawn uniformly at random.

Let $\Theta \subset \mathbb{R}^q$ with $q \geq 1$ be some parameter space and $H : \prod_{k=1}^{K} \mathcal{X}_k^{d_k} \times \Theta \to \mathbb{R}$ be a loss function which is convex and differentiable in its last argument. Let $(X_1^{(k)}, \ldots, X_{d_k}^{(k)})$, $1 \leq k \leq K$, be K independent random vectors with distribution $F_k^{\otimes d_k}(dx)$ on $\mathcal{X}_k^{d_k}$ respectively such that the random vector $H(X_1^{(1)}, \ldots, X_{d_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{d_K}^{(K)}; \theta)$ is square integrable for any $\theta \in \Theta$. For all $\theta \in \Theta$, set

$$L(\theta) = \mathbb{E}[H(X_1^{(1)}, \ldots, X_{d_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{d_K}^{(K)}; \theta)] = \mu(H(\cdot; \theta))$$

and consider the *risk minimization* problem $\min_{\theta \in \Theta} L(\theta)$. Based on K independent i.i.d. samples $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ with $1 \leq k \leq K$, the empirical version of the risk function is $\theta \in \Theta \mapsto \widehat{L}_{\mathbf{n}}(\theta) \overset{\text{def}}{=} U_{\mathbf{n}}(H(\cdot; \theta))$. Here and throughout, we denote by ∇_θ the gradient operator w.r.t. θ.

Gradient descent Many practical machine learning algorithms use variants of the standard gradient descent method, following the iterations:

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta \widehat{L}_{\mathbf{n}}(\theta_t), \tag{33}$$

with an arbitrary initial value $\theta_0 \in \Theta$ and a learning rate (step size) $\eta_t \geq 0$ such that $\sum_{t=1}^{+\infty} \eta_t = +\infty$ and $\sum_{t=1}^{+\infty} \eta_t^2 < +\infty$.

Here we place ourselves in a large-scale setting, where the sample sizes $n_1, \ldots, n_K$ of the training data sets are so large that computing the gradient of $\widehat{L}_{\mathbf{n}}$

$$\widehat{g}_{\mathbf{n}}(\theta) = \frac{1}{\prod_{k=1}^{K} \binom{n_k}{d_k}} \sum_{I_1} \cdots \sum_{I_K} \nabla_\theta H(\mathbf{X}_{I_1}^{(1)}; \mathbf{X}_{I_2}^{(2)}; \ldots; \mathbf{X}_{I_K}^{(K)}; \theta) \tag{34}$$

at each iteration (33) is computationally too expensive. Instead, Stochastic Gradient Descent uses an unbiased estimate $\widetilde{g}(\theta)$ of the gradient (34) that is cheap to compute. A

natural approach consists in replacing (34) by a complete U-statistic constructed from subsamples of reduced sizes $n'_k << n_k$ drawn uniformly at random, leading to the following gradient estimate:

$$\widetilde{g}_{\mathbf{n}'}(\theta) = \frac{1}{\prod_{k=1}^{K} \binom{n'_k}{d_k}} \sum_{I_1} \dots \sum_{I_K} \nabla_\theta H(\mathbf{X}^{(1)}_{I_1}; \mathbf{X}^{(2)}_{I_2}; \dots; \mathbf{X}^{(K)}_{I_K};\, \theta), \tag{35}$$

where the symbol $\sum_{I_k}$ refers to summation over all $\binom{n'_k}{d_k}$ subsets $\mathbf{X}^{(k)}_{I_k} = (X^{(k)}_{i_1}, \dots, X^{(k)}_{i_{d_k}})$ related to a set I_k of d_k indexes $1 \leq i_1 < \dots < i_{d_k} \leq n'_k$ and $\mathbf{n}' = (n'_1, \dots, n'_K)$.

We propose an alternative strategy based on the sampling scheme described in Section 3, *i.e.* a gradient estimate in the form of an *incomplete* U-statistic:

$$\widetilde{g}_B(\theta) = \frac{1}{B} \sum_{(I_1, \dots, I_K) \in \mathcal{D}_B} \nabla_\theta H(\mathbf{X}^{(1)}_{I_1}, \dots, \mathbf{X}^{(K)}_{I_K};\, \theta), \tag{36}$$

where $\mathcal{D}_B$ is built by sampling with replacement in the set Λ.

It is well-known that the variance of the gradient estimate negatively impacts on the convergence of SGD. Consider for instance the case where the loss function H is $(1/\gamma)$-smooth in its last argument, *i.e.* $\forall \theta_1, \theta_2 \in \Theta$:

$$\|\nabla_\theta H(\cdot;\, \theta_1) - \nabla_\theta H(\cdot;\, \theta_2)\| \leq \frac{1}{\gamma} \|\theta_1 - \theta_2\|.$$

Then one can show that if $\widetilde{g}$ is the gradient estimate:

$$\begin{aligned}
\mathbb{E}[\widehat{L}_{\mathbf{n}}(\theta_{t+1})] &= \mathbb{E}[\widehat{L}_{\mathbf{n}}(\theta_t - \eta_t \widetilde{g}(\theta_t))] \\
&\leq \mathbb{E}[\widehat{L}_{\mathbf{n}}(\theta_t)] - \eta_t \|\mathbb{E}[\widehat{g}_{\mathbf{n}}(\theta_t)]\|^2 + \frac{\eta_t^2}{2\gamma} \mathbb{E}[\|\widetilde{g}(\theta_t)\|^2] \\
&\leq \mathbb{E}[\widehat{L}_{\mathbf{n}}(\theta_t)] - \eta_t \left(1 - \frac{\eta_t}{2\gamma}\right) \mathbb{E}[\|\widehat{g}_{\mathbf{n}}(\theta_t)\|^2] + \frac{\eta_t^2}{2\gamma} \mathrm{Var}[\widetilde{g}(\theta_t)].
\end{aligned}$$

In other words, the smaller the variance of the gradient estimate, the larger the expected reduction in objective value. Some recent work has focused on variance-reduction strategies for SGD when the risk estimates are basic sample means (see for instance Le Roux et al., 2012; Johnson and Zhang, 2013).

In our setting where the risk estimates are of the form of a U-statistic, we are interested in comparing the variance of $\widetilde{g}_{\mathbf{n}'}(\theta)$ and $\widetilde{g}_B(\theta)$ when $B = \prod_{k=1}^{K} \binom{n'_k}{d_k}$ so that their computation requires to average over the same number of terms and thus have similar computational cost.[1] Our result is summarized in the following proposition.

Proposition 14 *Let* $B = \prod_{k=1}^{K} \binom{n'_k}{d_k}$ *for* $n'_k \ll n_k$, $k = 1, \dots, K$. *In the asymptotic framework* (4)*, we have:*

$$\mathrm{Var}[\widetilde{g}_{\mathbf{n}'}(\theta)] = O\left(\frac{1}{\sum_{k=1}^{K} n'_k}\right), \quad \mathrm{Var}[\widetilde{g}_B(\theta)] = O\left(\frac{1}{\prod_{k=1}^{K} \binom{n'_k}{d_k}}\right),$$

as $n' = n'_1 + \dots + n'_K \to +\infty$.

1. Note that sampling B sets from Λ to obtain (36) is potentially more efficient than sampling n'_k points from $\mathbf{X}_{\{1,\dots,n_k\}}$ for each $k = 1, \dots, K$ and then forming all combinations to obtain (35).

Proposition 14 shows that the convergence rate of $\mathrm{Var}[\widetilde{g}_B(\theta)]$ is faster than that of $\mathrm{Var}[\widetilde{g}_{\mathbf{n}'}(\theta)]$ except when $K = 1$ and $d_1 = 1$. Thus the expected improvement in objective function at each SGD step is larger when using a gradient estimate in the form of (36) instead of (35), although both strategies require to average over the same number of terms. This is also supported by the experimental results reported in the next section.

5. Numerical Experiments

We show the benefits of the sampling approach promoted in this paper on two applications: metric learning for classification, and model selection in clustering.

5.1 Metric Learning

In this section, we focus on the metric learning problem (see Section 2.2.2). As done in much of the metric learning literature, we restrict our attention to the family of pseudo-distance functions $D_{\mathbf{M}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ defined as

$$D_{\mathbf{M}}(x, x') = (x - x')\mathbf{M}(x - x')^T,$$

where $\mathbf{M} \in \mathbb{S}_+^d$, and $\mathbb{S}_+^d$ is the cone of $d \times d$ symmetric positive-semidefinite (PSD) matrices.

Given a training sample $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{1, \ldots, C\}$, let $y_{ij} = 1$ if $y_i = y_j$ and 0 otherwise for any pair of samples. Given a threshold $b \geq 0$, we define the empirical risk as follows:

$$R_n(D_{\mathbf{M}}) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} [y_{ij}(b - D_{\mathbf{M}}(x_i, x_j))]_+ , \tag{37}$$

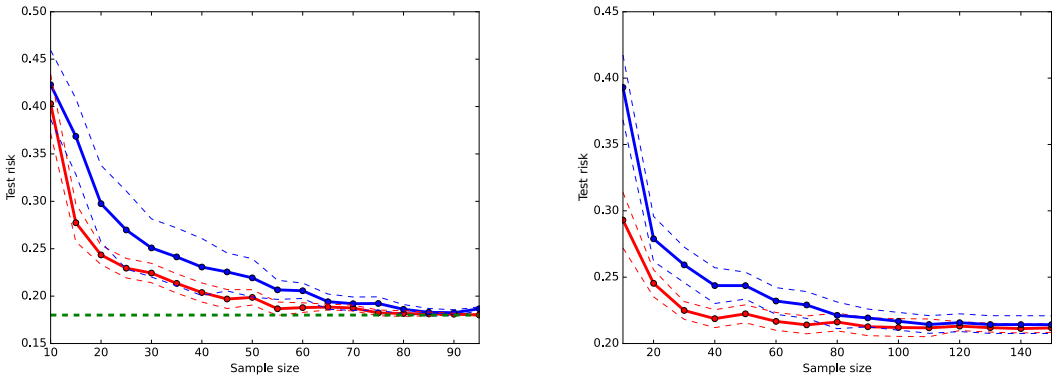
where $[u]_+ = \max(0, 1 - u)$ is the hinge loss. This risk estimate is convex and was used for instance by Jin et al. (2009) and Cao et al. (2012). Our goal is the find the empirical risk minimizer among our family of distance functions, i.e.:

$$\widehat{\mathbf{M}} = \operatorname*{arg\,min}_{\mathbf{M} \in \mathbb{S}_+^d} R_n(D_{\mathbf{M}}). \tag{38}$$

In our experiments, we use the following two data sets:

- **Synthetic data set**: some synthetic data that we generated for illustration. X is a mixture of 10 gaussians in $\mathbb{R}^{40}$ – each one corresponding to a class – such that all gaussian means are contained in an subspace of dimension 15 and their shared covariance matrix is proportional to identity with a variance factor such that some overlap is observed. That is, the solution to the metric learning problem should be proportional to the linear projection over the subspace containing the gaussians means. Training and testing sets contain respectively 50,000 and 10,000 observations.
- **MNIST data set**: a handwritten digit classification data set which has 10 classes and consists of 60,000 training images and 10,000 test images.[2] This data set has been used extensively to benchmark metric learning (Weinberger and Saul, 2009). As done by previous authors, we reduce the dimension from 784 to 164 using PCA so as to retain 95% of the variance, and normalize each sample to unit norm.

2. See `http://yann.lecun.com/exdb/mnist/`.



(a) Synthetic data set

(b) MNIST data set

Figure 3: Test risk with respect to the sample size n of the ERM when the risk is approximated using complete (blue) or incomplete (red) U-statistics. Solid lines represent means and dashed ones represent standard deviation. For the synthetic data set, the green dotted line represent the performance of the true risk minimizer.

Note that for both data sets, merely computing the empirical risk (37) for a given M involves averaging over more than 10^9 pairs.

We conduct two types of experiment. In Section 5.1.1, we subsample the data before learning and evaluate the performance of the ERM on the subsample. In Section 5.1.2, we use Stochastic Gradient Descent to find the ERM on the original sample, using subsamples at each iteration to estimate the gradient.

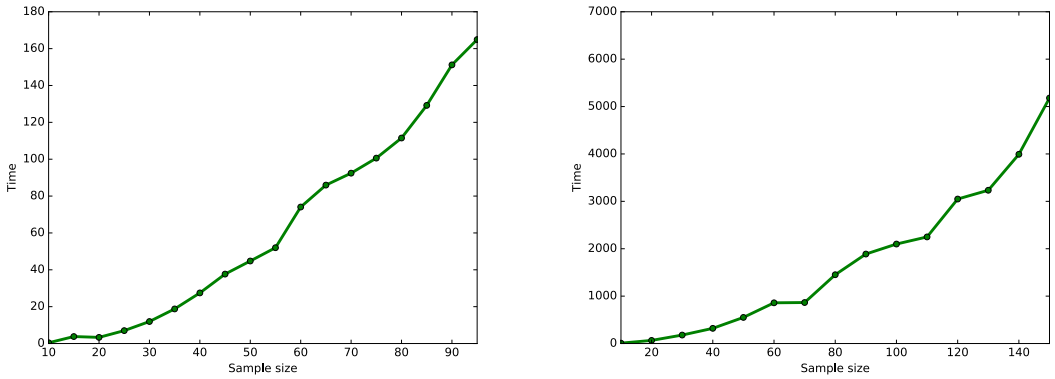
5.1.1 One-Time Sampling

We compare two sampling schemes to approximate the empirical risk:

- Complete U-statistic: n indices are uniformly picked at random in $\{1, \dots, n\}$. The empirical risk is approximated using any possible pair formed by the n indices, that is $\frac{n(n-1)}{2}$ pairs.
- Incomplete U-statistic: the empirical risk is approximated using $\frac{n(n-1)}{2}$ pairs picked uniformly at random in $\{1, \dots, n\}^2$.

For each strategy, we use a projected gradient descent method in order to solve (38), using several values of n and averaging the results over 50 random trials. As the testing sets are large, we evaluate the test risk on 100,000 randomly picked pairs.

Figure 3(a) shows the test risk of the ERM with respect to the sample size n for both sampling strategies on the synthetic data set. As predicted by our theoretical analysis, the incomplete U-statistic strategy achieves a significantly smaller risk on average. For instance, it gets within 5% error of the true risk minimizer for $n = 50$, while the complete U-statistic needs $n > 80$ to reach the same performance. This represents twice more computational time, as shown in Figure 4(a) (as expected, the runtime increases roughly quadratically with



(a) Synthetic data set

(b) MNIST data set

Figure 4: Average training time (in seconds) with respect to the sample size p.

p). The incomplete U-statistic strategy also has the advantage of having a much smaller variance between the runs, which makes it more reliable. The same conclusions hold for the MNIST data set, as can be seen in Figure 3(b) and Figure 4(b).

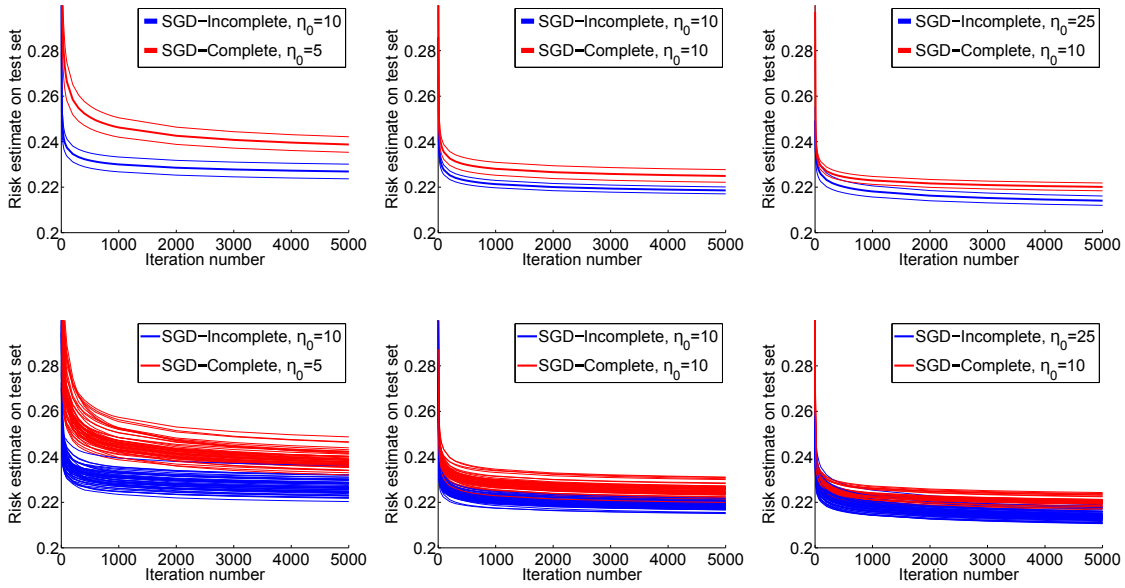
5.1.2 Stochastic Gradient Descent

In this section, we focus on solving the ERM problem (38) using Stochastic Gradient Descent and compare two approaches (analyzed in Section 4) to construct a mini-batch at each iteration. The first strategy, SGD-Complete, is to randomly draw (with replacement) a subsample and use the complete U-statistic associated with the subsample as the gradient estimate. The second strategy, SGD-Incomplete (the one we promote in this paper), consists in sampling an incomplete U-statistic with the same number of terms as in SGD-Complete.

For this experiment, we use the MNIST data set. We set the threshold in (37) to $b = 2$ and the learning rate of SGD at iteration t to $\eta_t = 1/(\eta_0 t)$ where $\eta_0 \in \{1, 2.5, 5, 10, 25, 50\}$. To reduce computational cost, we only project our solution onto the PSD cone at the end of the algorithm, following the "one projection" principle used by Chechik et al. (2010). We try several values m for the mini-batch size, namely $m \in \{10, 28, 55, 105, 253\}$.[3] For each mini-batch size, we run SGD for 10,000 iterations and select the learning rate parameter η_0 that achieves the minimum risk on 100,000 pairs randomly sampled from the training set. We then estimate the generalization risk using 100,000 pairs randomly sampled from the test set.

For all mini-batch sizes, SGD-Incomplete achieves significantly better test risk than SGD-Complete. Detailed results are shown in Figure 5 for three mini-batch sizes, where we plot the evolution of the test risk with respect to the iteration number.[4] We make several comments. First, notice that the best learning rate is often larger for SGD-Incomplete than for SGD-Complete ($m = 10$ and $m = 253$). This confirms that gradient estimates from the

3. For each m, we can construct a complete U-statistic from n' samples with $n'(n'-1)/2 = m$ terms.
4. We point out that the figures look the same if we plot the runtime instead of the iteration number. Indeed, the time spent on computing the gradients (which is the same for both variants) largely dominates the time spent on the random draws.



(a) $m = 10$ (b) $m = 55$ (c) $m = 253$

Figure 5: SGD results on the MNIST data set for various mini-batch size m. The top row shows the means and standard deviations over 50 runs, while the bottom row shows each run separately.

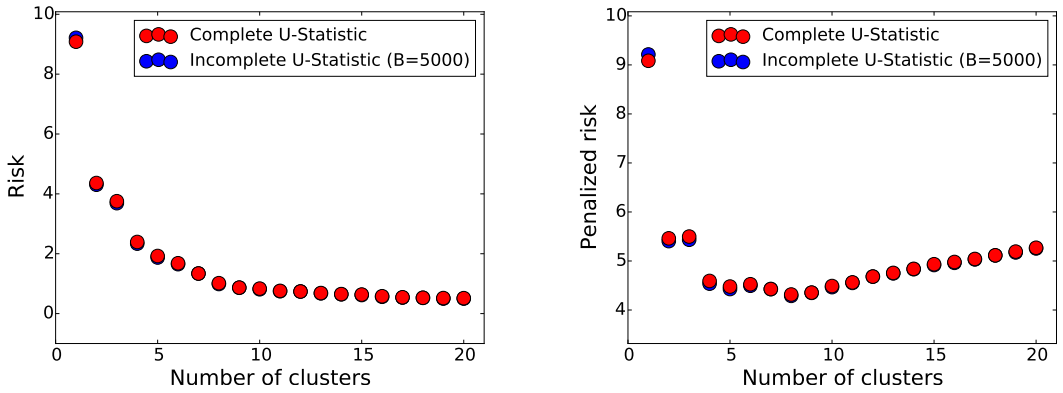
former strategy are generally more reliable. This is further supported by the fact that even though larger learning rates increase the variance of SGD, in these two cases SGD-Complete and SGD-Incomplete have similar variance. On the other hand, for $m = 55$ the learning rate is the same for both strategies. SGD-Incomplete again performs significantly better on average and also has smaller variance. Lastly, as one should expect, the gap between SGD-Complete and SGD-Incomplete reduces as the size of the mini-batch increases. Note however that in practical implementations, the relatively small mini-batch sizes (in the order of a few tens or hundreds) are generally those which achieve the best error/time trade-off.

5.2 Model Selection in Clustering

In this section, we are interested in the clustering problem described in Section 2.2.1. Specifically, let $X_1, \ldots, X_n \in \mathbb{R}^d$ be the set of points to be clustered. Let the clustering risk associated with a partition $\mathcal{P}$ into M groups $\mathcal{C}_1, \ldots, \mathcal{C}_M$ be:

$$\widehat{W}_n(\mathcal{P}) = \frac{2}{n(n-1)} \sum_{m=1}^{M} \sum_{1 \leq i < j \leq n} D(X_i, X_j) \cdot \mathbb{I}\{(X_i, X_j) \in \mathcal{C}_m^2\}. \tag{39}$$

In this experiment, given a set of candidate partitions, we want to perform model selection by picking the partition which minimizes the risk (39) plus some term penalizing the complexity of the partition. When the number of points n is large, the complete risk is very



(a) Risk (b) Penalized risk

Figure 6: Clustering model selection results on the forest cover type data set. Figure 6(a) shows the risk (complete and incomplete with $B = 5,000$ terms) for the first 20 partitions, while Figure 6(b) shows the penalized risk for $c = 1.1$.

expensive to compute. Our strategy is to replace it with an incomplete approximation with much fewer terms. Like in the approach theoretically investigated in Section 3.2, the goal here is to show that using the incomplete approximation instead of the complete version as the goodness-of-fit measure in a complexity penalized criterion does not damage the selection, while reducing the computational cost. For simplicity, the complexity penalty we use below is not of the same type as the structural VC dimension-based penalty considered in Theorem 10, but we will see that the incomplete approximation is very accurate and can thus effectively replace the complete version regardless of the penalty used.

The experimental setup is as follows. We used the forest cover type data set,[5] which is popular to benchmark clustering algorithms (see for instance Kanungo et al., 2004). To be able to evaluate the complete risk, we work with $n = 5,000$ points subsampled at random from the entire data set of 581,012 points in dimension 54. We then generated a hierarchical clustering of these points using agglomerative clustering with Ward's criterion (Ward, 1963) as implemented in the `scikit-learn` Python library (Pedregosa et al., 2011). This defines n partitions $\mathcal{P}_1, \ldots, \mathcal{P}_n$ where $\mathcal{P}_m$ consists of m clusters ($\mathcal{P}_1$ corresponds to a single cluster containing all points, while in $\mathcal{P}_n$ each point has its own cluster).

For each partition size, we first compare the value of the complete risk (39) with $n(n-1) = 24,995,000$ terms with that of an incomplete version with only $B = n = 5,000$ pairs drawn at random. As shown in Figure 6(a), the incomplete U-statistic is a very accurate approximation of the complete one, despite consisting of 5000 times less terms. It will thus lead to similar results in model selection. To illustrate, we use a simple penalty term of the form $\mathrm{pen}(\mathcal{P}_m) = c \cdot \log(m)$ where c is a scaling constant. Figure 6(b) shows that both selection criteria choose the same model $\mathcal{P}_8$. Performing this model selection over $\mathcal{P}_1, \ldots, \mathcal{P}_{20}$

5. See `https://archive.ics.uci.edu/ml/datasets/Covertype`.

took about 66 seconds for the complete U-statistic, compared to only 0.1 seconds for the incomplete version.[6]

Finally, we generated 100 incomplete U-statistics with different random seeds ; all of them correctly identified $\mathcal{P}_8$ as the best model. Using $B = 5,000$ pairs is thus sufficient to obtain reliable results with an incomplete U-statistic for this data set. In contrast, the complete U-statistics based on a subsample (leading to the same number of pairs) selected the correct model in only 57% of cases.

6. Conclusion

In a wide variety of statistical learning problems, U-statistics are natural estimates of the risk measure one seeks to optimize. As the sizes of the samples increase, the computation of such functionals involves summing a rapidly exploding number of terms and becomes numerically unfeasible. In this paper, we argue that for such problems, *Empirical Risk Minimization* can be implemented using statistical counterparts of the risk based on much less terms (picked randomly by means of sampling with replacement), referred to as *incomplete U-statistics*. Using a novel deviation inequality, we have shown that this approximation scheme does not deteriorate the learning rates, even preserving fast rates in certain situations where they are proved to occur. Furthermore, we have extended these results to U-statistics based on different sampling schemes (Bernoulli sampling, sampling without replacement) and shown how such functionals can be used for the purpose of model selection and for implementing ERM iterative procedures based on stochastic gradient descent. Beyond theoretical rate bounds, the efficiency of the approach we promote is illustrated by several numerical experiments.

Acknowledgments

This work is supported by the Chair "Machine Learning for Big Data" of Télécom ParisTech, and was conducted while A. Bellet was affiliated with Télécom ParisTech. The authors are grateful to the reviewers for their careful reading of the paper, which permitted to improve significantly the presentation of the results.

Appendix A. Proof of Proposition 2

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

$$V_H\left(X_1^{(1)}, \ldots, X_{n_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{n_K}^{(K)}\right) = \frac{1}{N}\Big[H\left(X_1^{(1)}, \ldots, X_{d_1}^{(1)}, \ldots, X_1^{(K)}, \ldots, X_{d_K}^{(K)}\right) + H\left(X_{d_1+1}^{(1)}, \ldots, X_{2d_1}^{(1)}, \ldots, X_{d_K+1}^{(K)}, \ldots, X_{2d_K}^{(K)}\right) + \ldots + H\left(X_{Nd_1-d_1+1}^{(1)}, \ldots, X_{Nd_1}^{(1)}, \ldots, X_{Nd_K-d_K+1}^{(K)}, \ldots, X_{Nd_K}^{(K)}\right)\Big],$$

6. The $n \times n$ distance matrix was precomputed before running the agglomerative clustering algorithm. The associated runtime is thus not taken into account in these timing results.

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

$$U_{\mathbf{n}}(h) = \frac{1}{n_1! \cdots n_K!} \sum_{\sigma_1 \in \mathfrak{S}_{n_1}, \ldots, \sigma_K \in \mathfrak{S}_{n_K}} V_h\left(X^{(1)}_{\sigma_1(1)}, \ldots, X^{(1)}_{\sigma_1(n_1)}, \ldots, X^{(K)}_{\sigma_K(1)}, \ldots, X^{(K)}_{\sigma_K(n_K)}\right), \tag{40}$$

where $\mathfrak{S}_m$ denotes the symmetric group of order m for any $m \geq 1$. This representation as an average of sums of N independent terms is known as the (first) Hoeffding's decomposition, see Hoeffding (1948). Then, using Jensen's inequality in particular, one may easily show that, for any nondecreasing convex function $\psi : \mathbb{R}_+ \to \mathbb{R}$, we have:

$$\mathbb{E}\left[\psi\left(\sup_{h \in \mathcal{H}} \left|U_{\mathbf{n}}(\bar{h})\right|\right)\right] \leq \mathbb{E}\left[\psi\left(\sup_{h \in \mathcal{H}} \left|V_{\bar{h}}(X^{(1)}_1, \ldots, X^{(1)}_{n_1}, \ldots, X^{(K)}_1, \ldots, X^{(K)}_{n_K})\right|\right)\right], \tag{41}$$

where we set $\bar{h} = h - \mu(h)$ for all $h \in \mathcal{H}$. Now, using standard symmetrization and randomization arguments (see Giné and Zinn (1984) for instance) and (41), we obtain that

$$\mathbb{E}\left[\psi\left(\sup_{h \in \mathcal{H}} \left|U_{\mathbf{n}}(\bar{h})\right|\right)\right] \leq \mathbb{E}\left[\psi\left(2\mathcal{R}_N\right)\right], \tag{42}$$

where

$$\mathcal{R}_N = \sup_{h \in \mathcal{H}} \frac{1}{N} \sum_{l=1}^{N} \epsilon_l h\left(X^{(1)}_{(l-1)d_1+1}, \ldots, X^{(1)}_{ld_1}, \ldots, X^{(K)}_{(l-1)d_K+1}, \ldots, X^{(K)}_{ld_K}\right),$$

is a Rademacher average based on the Rademacher chaos $\epsilon_1, \ldots, \epsilon_N$ (independent random symmetric sign variables), independent from the $X_i^{(k)}$'s. We now apply the bounded difference inequality (see McDiarmid (1989)) to the functional $\mathcal{R}_N$, seen as a function of the i.i.d. random variables $(\epsilon_l, X^{(1)}_{(l-1)d_1+1}, \ldots, X^{(1)}_{ld_1}, \ldots, X^{(K)}_{(l-1)d_K+1}, \ldots, X^{(K)}_{ld_K})$, $1 \leq l \leq N$: changing any of these random variables change the value of $\mathcal{R}_N$ by at most $\mathcal{M}_{\mathcal{H}}/N$. One thus obtains from (42) with $\psi(x) = \exp(\lambda x)$, where $\lambda > 0$ is a parameter which shall be chosen later, that:

$$\mathbb{E}\left[\exp\left(\lambda \sup_{h \in \mathcal{H}} \left|U_{\mathbf{n}}(\bar{h})\right|\right)\right] \leq \exp\left(2\lambda \mathbb{E}[\mathcal{R}_N] + \frac{\mathcal{M}_{\mathcal{H}}^2 \lambda^2}{4N}\right). \tag{43}$$

Applying Chernoff's method, one then gets:

$$\mathbb{P}\left\{\sup_{h \in \mathcal{H}} \left|U_{\mathbf{n}}(\bar{h})\right| > \eta\right\} \leq \exp\left(-\lambda\eta + 2\lambda \mathbb{E}[\mathcal{R}_N] + \frac{\mathcal{M}_{\mathcal{H}}^2 \lambda^2}{4N}\right). \tag{44}$$

Using the bound (see Eq. (6) in Boucheron et al. (2005) for instance)

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

and taking $\lambda = 2N(\eta - 2\mathbb{E}[\mathcal{R}_N])/\mathcal{M}_{\mathcal{H}}^2$ in (44), one finally establishes the desired result.

Appendix B. Proof of Theorem 6

For convenience, we introduce the random sequence $\zeta = ((\zeta_k(I))_{I\in\Lambda})_{1\leq k\leq B}$, where $\zeta_k(I)$ 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 ζ_k's are i.i.d. random vectors and, for all $(k, I) \in \{1, \ldots, B\} \times \Lambda$, the r.v. $\zeta_k(I)$ 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 $\mathcal{Z}_k(H) = \sum_{I\in\Lambda}(\zeta_k(I) - 1/\#\Lambda)H(\mathbf{X}_I)$ for any $(k, I) \in \{1, \ldots, B\} \times \Lambda$. It follows from the independence between the $\mathbf{X}_I$'s and the $\zeta(I)$'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 $\sum_{I\in\Lambda}\zeta_k(I) = 1$ for all $k \geq 1$). By virtue of Sauer's lemma, since $\mathcal{H}$ is a VC major class with finite VC dimension 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:

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

Taking the expectation, this 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 $1 - \delta$:

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

Turning to the second part of the theorem, it straightforwardly results from the first part combined with Proposition 2.

Appendix C. Proof of Corollary 9

Assertion (i) is a direct application of Assertion (ii) in Theorem 6 combined with the bound $\mu(\widehat{H}_B) - \inf_{H\in\mathcal{H}}\mu(H) \leq 2\sup_{H\in\mathcal{H}}|\widetilde{U}_B(H) - \mu(H)|$.

Turning next to Assertion (ii), observe that by triangle inequality we have:

$$\mathbb{E}\left[\sup_{H\in\mathcal{H}_m}|\widetilde{U}_B(H) - \mu(H)|\right] \leq \mathbb{E}\left[\sup_{H\in\mathcal{H}_m}|\widetilde{U}_B(H) - U_{\mathbf{n}}(H)|\right] + \mathbb{E}\left[\sup_{H\in\mathcal{H}_m}|U_{\mathbf{n}}(H) - \mu(H)|\right]. \quad (45)$$

The same argument as that used in Theorem 6 (with $\psi(u) = u$ for any $u \geq 0$) yields a bound for the second term on the right hand side of Eq. (45):

$$\mathbb{E}\left[\sup_{H\in\mathcal{H}}|U_{\mathbf{n}}(H) - \mu(H)|\right] \leq 2\mathcal{M}_{\mathcal{H}}\sqrt{\frac{2V\log(1 + N)}{N}}. \quad (46)$$

The first term can be controlled by means of the following lemma, whose proof can be found for instance in Lugosi (2002, Lemmas 1.2 and 1.3).

Lemma 15 *The following assertions hold true.*

(i) Hoeffding's lemma. Let Z *be an integrable r.v. with mean zero such that* $a \leq Z \leq b$ *almost-surely. Then, we have:* $\forall s > 0$

$$\mathbb{E}[\exp(sZ)] \leq \exp\left(s^2(b-a)^2/8\right).$$

(ii) Let $M \geq 1$ *and* $Z_1, \ldots, Z_M$ *be real valued random variables. Suppose that there exists* $\sigma > 0$ *such that* $\forall s \in \mathbb{R}$*:* $\mathbb{E}[\exp(sZ_i)] \leq e^{s^2\sigma^2/2}$ *for all* $i \in \{1, \ldots, M\}$*. Then, we have:*

$$\mathbb{E}\left[\max_{1\leq i\leq M} |Z_i|\right] \leq \sigma\sqrt{2\log(2M)}. \tag{47}$$

Assertion (i) shows that, since $-\mathcal{M}_{\mathcal{H}} \leq \mathcal{Z}_k(H) \leq \mathcal{M}_{\mathcal{H}}$ almost surely,

$$\mathbb{E}\left[\exp(s\sum_{k=1}^{B} \mathcal{Z}_k(H)) \mid (\mathbf{X}_I)_{I\in\Lambda}\right] \leq e^{\frac{1}{2}Bs^2\mathcal{M}_{\mathcal{H}}^2}.$$

With $\sigma = \mathcal{M}_{\mathcal{H}}\sqrt{B}$ and $M = \#\{H(\mathbf{X}_I) : H \in \mathcal{H}\} \leq (1+\#\Lambda)^V$, conditioning upon $(\mathbf{X}_I)_{I\in\Lambda}$, this result yields:

$$\mathbb{E}\left[\sup_{H\in\mathcal{H}} \left|\frac{1}{B}\sum_{k=1}^{B} \mathcal{Z}_k(H)\right| \mid (\mathbf{X}_I)_{I\in\Lambda}\right] \leq \mathcal{M}_{\mathcal{H}}\sqrt{\frac{2(\log 2 + V\log(1+\#\Lambda))}{B}}. \tag{48}$$

Integrating next over $(\mathbf{X}_I)_{I\in\Lambda}$ and combining the resulting bound with (45) and (46) leads to the inequality stated in (ii).

A bound for the expected value. For completeness, we point out that the expected value of $\sup_{H\in\mathcal{H}} |(1/B)\sum_{k=1}^{B} \mathcal{Z}_k(H)|$ can also be bounded by means of classical symmetrization and randomization devices. Considering a "ghost" i.i.d. sample $\zeta_1', \ldots, \zeta_B'$ independent from $((\mathbf{X}_I)_{I\in\Lambda}, \zeta)$, distributed as ζ, Jensen's inequality yields:

$$\begin{aligned}
\mathbb{E}\left[\sup_{H\in\mathcal{H}} \left|\frac{1}{B}\sum_{k=1}^{B} \mathcal{Z}_k(H)\right|\right] &= \mathbb{E}\left[\sup_{H\in\mathcal{H}} \left\{\mathbb{E}\left[\left|\frac{1}{B}\sum_{k=1}^{B}\sum_{I\in\Lambda} H(\mathbf{X}_I)\left(\zeta_k(I) - \zeta_k'(I)\right)\right| \mid (\mathbf{X}_I)_{I\in\Lambda}\right]\right\}\right] \\
&\leq \mathbb{E}\left[\sup_{H\in\mathcal{H}} \left|\frac{1}{B}\sum_{k=1}^{B}\sum_{I\in\Lambda} H(\mathbf{X}_I)\left(\zeta_k(I) - \zeta_k'(I)\right)\right|\right].
\end{aligned}$$

Introducing next independent Rademacher variables $\epsilon_1, \ldots, \epsilon_B$, independent from $((\mathbf{X}_I)_{I\in\Lambda}, \zeta, \zeta')$, we have:

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{B}\sum_{k=1}^{B}\sum_{I\in\Lambda} h(\mathbf{X}_I)\left(\zeta_k(I)-\zeta_k'(I)\right)\right| \mid (\mathbf{X}_I)_{I\in\Lambda}\right] =$$
$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{B}\sum_{k=1}^{B}\epsilon_k\sum_{I\in\Lambda} h(\mathbf{X}_I)\left(\zeta_k(I)-\zeta_k'(I)\right)\right| \mid (\mathbf{X}_I)_{I\in\Lambda}\right]$$
$$\leq 2\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{B}\sum_{k=1}^{B}\epsilon_k\sum_{I\in\Lambda} h(\mathbf{X}_I)\zeta_k(I)\right| \mid (\mathbf{X}_I)_{I\in\Lambda}\right].$$

We thus obtained:

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{B}\sum_{k=1}^{B}\mathcal{Z}_k(h)\right|\right] \leq 2\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{B}\sum_{k=1}^{B}\epsilon_k\sum_{I\in\Lambda} h(\mathbf{X}_I)\zeta_k(I)\right|\right].$$

Appendix D. Proof of Theorem 10

We start with proving the intermediary result, stated below.

Lemma 16 *Under the assumptions stipulated in Theorem 10, we have:* $\forall m \geq 1$, $\forall \epsilon > 0$,

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}_m}|\mu(h)-\widetilde{U}_B(h)| > 2\mathcal{M}_{\mathcal{H}_m}\left\{\sqrt{\frac{2V_m\log(1+N)}{N}}+\sqrt{\frac{2(\log 2+V_m\log(1+\#\Lambda))}{B}}\right\}+\epsilon\right\}$$
$$\leq \exp\left(-B^2\epsilon^2/\left(2(B+n)\mathcal{M}_{\mathcal{H}_m}^2\right)\right).$$

Proof This is a direct application of the bounded difference inequality (see McDiarmid (1989)) applied to the quantity $\sup_{h\in\mathcal{H}_m}|\mu(h)-\widetilde{U}_B(h)|$, viewed as a function of the $(B+n)$ independent random variables $(X_1^{(1)}, X_{n_K}^{(K)}, \epsilon_1, \ldots, \epsilon_B)$ (jumps being bounded by $2\mathcal{M}_h/B$), combined with Assertion (ii) of Corollary 9. ■

Let $m \geq 1$ and decompose the expected excess of risk of the rule picked by means of the complexity regularized incomplete U-statistic criterion as follows:

$$\mathbb{E}\left[\mu(\widehat{h}_{B,\widehat{m}})-\mu_m^*\right] = \mathbb{E}\left[\mu(\widehat{h}_{B,\widehat{m}})-\widetilde{U}_B(\widehat{h}_{B,\widehat{m}})-\mathrm{pen}(B,\widehat{m})\right]$$
$$+\mathbb{E}\left[\inf_{j\geq 1}\left\{\widetilde{U}_B(\widehat{h}_{B,j})+\mathrm{pen}(B,j)\right\}-\mu_m^*\right],$$

where we set $\mu_m^* = \inf_{H \in \mathcal{H}_m} \mu(H)$. In order to bound the first term on the right hand side of the equation above, observe that we have: $\forall \epsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left\{\mu(\widehat{H}_{B,\widehat{m}}) - \widetilde{U}_B(\widehat{H}_{B,\widehat{m}}) - \mathrm{pen}(B,\widehat{m}) > \epsilon\right\} &\leq \mathbb{P}\left\{\sup_{j\geq 1}\left\{\mu(\widehat{H}_{B,j}) - \widetilde{U}_B(\widehat{H}_{B,j}) - \mathrm{pen}(B,j)\right\} > \epsilon\right\} \\
&\leq \sum_{j\geq 1} \mathbb{P}\left\{\mu(\widehat{H}_{B,j}) - \widetilde{U}_B(\widehat{H}_{B,j}) - \mathrm{pen}(B,j) > \epsilon\right\} \\
&\leq \sum_{j\geq 1} \mathbb{P}\left\{\sup_{H\in\mathcal{H}_j} |\mu(\widehat{H}) - \widetilde{U}_B(H)| - \mathrm{pen}(B,j) > \epsilon\right\} \\
&\leq \sum_{j\geq 1} \exp\left(-\frac{-B^2}{2(B+n)\mathcal{M}^2}\left(\epsilon + 2\mathcal{M}\sqrt{\frac{(B+n)\log j}{B^2}}\right)^2\right) \\
&\leq \exp\left(-\frac{B^2\epsilon^2}{2(B+n)\mathcal{M}^2}\right)\sum_{j\geq 1} 1/j^2 \leq 2\exp\left(-\frac{B^2\epsilon^2}{2(B+n)\mathcal{M}^2}\right),
\end{aligned}$$

using successively the union bound and Lemma 16. Integrating over $[0,+\infty)$, we obtain that:

$$\mathbb{E}\left[\mu(\widehat{H}_{B,\widehat{m}}) - \widetilde{U}_B(\widehat{H}_{B,\widehat{m}}) - \mathrm{pen}(B,\widehat{m})\right] \leq \mathcal{M}\frac{\sqrt{2\pi(B+n)}}{B}. \tag{49}$$

Considering now the second term, notice that

$$\mathbb{E}\left[\inf_{j\geq 1}\left\{\widetilde{U}_B(\widehat{H}_{B,j}) + \mathrm{pen}(B,j)\right\} - \mu_m^*\right] \leq \mathbb{E}\left[\widetilde{U}_B(\widehat{H}_{B,m}) + \mathrm{pen}(B,m) - \mu_m^*\right] \leq \mathrm{pen}(B,m).$$

Combining the bounds, we obtain that: $\forall m \geq 1$,

$$\mathbb{E}\left[\mu(\widehat{H}_{B,\widehat{m}})\right] \leq \mu_m^* + \mathrm{pen}(B,m) + \mathcal{M}\frac{\sqrt{2\pi(B+n)}}{B}.$$

The oracle inequality is thus proved.

Appendix E. Proof of Theorem 12

We start with proving the following intermediary result, based on the U-statistic version of the Bernstein exponential inequality.

Lemma 17 *Suppose that the assumptions of Theorem 12 are fulfilled. Then, for all $\delta \in (0,1)$, we have with probability larger than $1-\delta$: $\forall r \in \mathcal{R}$, $\forall n \geq 2$,*

$$0 \leq \Lambda_n(r) - \Lambda(r) + \sqrt{\frac{2c\Lambda(r)^\alpha \log(\#\mathcal{R}/\delta)}{n}} + \frac{4\log(\#\mathcal{R}/\delta)}{3n}.$$

Proof The proof is a straightforward application of Theorem A on p. 201 in Serfling (1980), combined with the union bound and Assumption 1. ■

The same argument as that used to prove Assertion (i) in Theorem 6 (namely, freezing the

$\mathbf{X}_I$'s, applying Hoeffding inequality and the union bound) shows that, for all $\delta \in (0,1)$, we have with probability at least $1-\delta$: $\forall r \in \mathcal{R}$,

$$0 \leq \widetilde{U}_B(q_r) - U_{\mathbf{n}}(q_r) + \sqrt{\frac{M + \log(M/\delta)}{B}}$$

for all $n \geq 2$ and $B \geq 1$ (observe that $\mathcal{M}_{\mathcal{H}} \leq 1$ in this case). Now, combining this bound with the previous one and using the union bound, one gets that, for all $\delta \in (0,1)$, we have with probability larger than $1-\delta$: $\forall r \in \mathcal{R}$, $\forall n \geq 2$, $\forall B \geq 1$,

$$0 \leq \widetilde{U}_B(q_r) - \Lambda(r) + \sqrt{\frac{2c\Lambda(r)^\alpha \log(2M/\delta)}{n}} + \frac{4\log(2M/\delta)}{3n} + \sqrt{\frac{M + \log(2M/\delta)}{B}}.$$

Observing that, $\widetilde{U}_B(q_{\widetilde{r}_B}) \leq 0$ by definition, we thus have with probability at least $1-\delta$:

$$\Lambda(\widetilde{r}_B) \leq \sqrt{\frac{2c\Lambda(\widetilde{r}_B)^\alpha \log(2M/\delta)}{n}} + \frac{4\log(2M/\delta)}{3n} + \sqrt{\frac{M + \log(2M/\delta)}{B}}.$$

Choosing finally $B = O(n^{2/(2-\alpha)})$, the desired result is obtained by solving the inequality above for $\Lambda(\widetilde{r}_B)$.

Appendix F. Proof of Theorem 13

As shown by the following lemma, which is a slight modification of Lemma 1 in Janson (1984), the deviation between the incomplete U-statistic and its complete version is of order $O_{\mathbb{P}})(1/\sqrt{B})$ for both sampling schemes.

Lemma 18 *Suppose that the assumptions of 13 are fulfilled. Then, we have:* $\forall H \in \mathcal{H}$,

$$\mathbb{E}\left[\left(\bar{U}_{HT}(H) - U_{\mathbf{n}}(H)\right)^2 \mid (\mathbf{X}_I)_{I\in\Lambda}\right] \leq 2\mathcal{M}_{\mathcal{H}}^2/B.$$

Proof Observe first that, in both cases (sampling without replacement and Bernoulli sampling), we have: $\forall I \neq J$ in Λ,

$$\mathbb{E}\left[\left(\Delta(I) - \frac{B}{\#\Lambda}\right)^2\right] \leq \frac{B}{\#\Lambda} \text{ and } \mathbb{E}\left[\left(\Delta(I) - \frac{B}{\#\Lambda}\right)\left(\Delta(J) - \frac{B}{\#\Lambda}\right)\right] \leq \frac{1}{\#\Lambda}\cdot\frac{B}{\#\Lambda}.$$

Hence, as $(\Delta(I))_{I\in\Lambda}$ and $(\mathbf{X}_I)_{I\in\Lambda}$ are independent by assumption, we have:

$$B^2\mathbb{E}\left[\left(\bar{U}_{HT}(H) - U_{\mathbf{n}(H)}\right)^2 \mid (\mathbf{X}_I)_{I\in\Lambda}\right] = \mathbb{E}\left[\left(\sum_{I\in\Lambda}\left(\Delta(I) - \frac{B}{\#\Lambda}\right)H(\mathbf{X}_I)\right)^2 \mid (\mathbf{X}_I)_{I\in\Lambda}\right]$$

$$\leq \mathcal{M}_{\mathcal{H}}^2 \sum_{I\in\Lambda}\mathbb{E}\left[\left(\Delta(I) - \frac{B}{\#\Lambda}\right)^2\right] + \mathcal{M}_{\mathcal{H}}^2 \sum_{I\neq J}\mathbb{E}\left[\left(\Delta(I) - \frac{B}{\#\Lambda}\right)\left(\Delta(J) - \frac{B}{\#\Lambda}\right)\right] \leq 2B\mathcal{M}_{\mathcal{H}}^2.$$

■

Consider first the case of Bernoulli sampling. By virtue of Bernstein inequality applied to the independent variables $(\Delta(I) - B/\#\Lambda)H(\mathbf{X}_I)$ conditioned upon $(\mathbf{X}_I)_{I\in\Lambda}$, we have: $\forall H \in \mathcal{H}$, $\forall t > 0$,

$$\mathbb{P}\left\{\left|\sum_{I\in\Lambda}(\Delta(I) - B/\#\Lambda)H(\mathbf{X}_I)\right| > t \mid (\mathbf{X}_I)_{I\in\Lambda}\right\} \leq 2\exp\left(-\frac{t^2}{4B\mathcal{M}_{\mathcal{H}}^2 + 2\mathcal{M}_{\mathcal{H}}t/3}\right).$$

Hence, combining this bound and the union bound, we obtain that: $\forall t > 0$,

$$\mathbb{P}\left\{\sup_{H\in\mathcal{H}}\left|\bar{U}_{HT}(H) - U_{\mathbf{n}(H)}\right| > t \mid (\mathbf{X}_I)_{I\in\Lambda}\right\} \leq 2(1+\#\Lambda)^V \exp\left(-\frac{Bt^2}{4\mathcal{M}_{\mathcal{H}}^2 + 2\mathcal{M}_{\mathcal{H}}t/3}\right).$$

Solving

$$\delta = 2(1+\#\Lambda)^V \exp\left(-\frac{Bt^2}{4\mathcal{M}_{\mathcal{H}}^2 + 2\mathcal{M}_{\mathcal{H}}t/3}\right)$$

yields the desired bound.

Consider next the case of the sampling without replacement scheme. Using the exponential inequality tailored to this situation proved in Serfling (1974) (see Corollary 1.1 therein), we obtain: $\forall H \in \mathcal{H}$, $\forall t > 0$,

$$\mathbb{P}\left\{\frac{1}{B}\left|\sum_{I\in\Lambda}(\Delta(I) - B/\#\Lambda)H(\mathbf{X}_I)\right| > t \mid (\mathbf{X}_I)_{I\in\Lambda}\right\} \leq 2\exp\left(-\frac{Bt^2}{2\mathcal{M}_{\mathcal{H}}^2}\right).$$

The proof can be then ended using the union bound, just like above.

Appendix G. Proof of Proposition 14

For simplicity, we focus on one sample U-statistics of degree two ($K = 1$, $d_1 = 2$) since the argument easily extends to the general case. Let $U_n(H)$ be a non-degenerate U-statistic of degree two:

$$U_n(H) = \frac{2}{n(n-1)}\sum_{i<j} H(x_i, x_j).$$

In order to express the variance of $U_n(H)$ based on its second Hoeffding decomposition (see Section 2.1), we first introduce more notations: $\forall (x, x') \in \mathcal{X}_1^2$,

$$H_1(x) \stackrel{\text{def}}{=} \mathbb{E}\left[H(x, X)\right] - \mu(H) \text{ and } H_2(x, x') \stackrel{\text{def}}{=} H(x, x') - \mu(H) - H_1(x) - H_1(x').$$

Equipped with these notations, the (orthogonal) Hoeffding/Hajek decomposition of $U_n(H)$ can be written as

$$U_n(H) = \mu(H) + 2T_n(H) + W_n(H),$$

involving centered and decorrelated random variables given by

$$\begin{aligned} T_n(H) &= \frac{1}{n}\sum_{i=1}^{n} H_1(x_i), \\ W_n(H) &= \frac{2}{n(n-1)}\sum_{i<j} H_2(x_i, x_j). \end{aligned}$$

Recall that the U-statistic $W_n(H)$ is said to be degenerate, since $\mathbb{E}[H_2(x, X)] = 0$ for all $x \in \mathcal{X}_1$. Based on this representation and setting $\sigma_1^2 = \mathrm{Var}[H_1(X)]$ and $\sigma_2^2 = \mathrm{Var}[H_2(X, X')]$, the variance of $U_n(H)$ is given by

$$\mathrm{Var}[U_n(H)] = \frac{4\sigma_1^2}{n} + \frac{2\sigma_2^2}{n(n-1)}. \tag{50}$$

As already pointed out in Section 3.1, the variance of the incomplete U-statistic built by sampling with replacement is

$$\begin{aligned} \mathrm{Var}[\widetilde{U}_B(H)] &= \mathrm{Var}[U_n(H)] + \frac{1}{B}\left(1 - \frac{2}{n(n-1)}\right)\mathrm{Var}[H(X, X')] \\ &= \mathrm{Var}[U_n(H)] + \frac{1}{B}\left(1 - \frac{2}{n(n-1)}\right)(2\sigma_1^2 + \sigma_2^2). \end{aligned} \tag{51}$$

Take $B = n'(n'-1)$ for $n' \ll n$. It follows from (50) and (51) that in the asymptotic framework (4), the quantities $\mathrm{Var}[U_{n'}(H)]$ and $\mathrm{Var}[\widetilde{U}_B(H)]$ are of the order $O(1/n')$ and $O(1/n'^2)$ respectively as $n' \to +\infty$. Hence these convergence rates hold for $\widetilde{g}_{\mathbf{n}'}(\theta)$ and $\widetilde{g}_B(\theta)$ respectively.

References

R. Bekkerman, M. Bilenko, and J. Langford. *Scaling Up Machine Learning*. Cambridge, 2011.

A. Bellet and A. Habrard. Robustness and generalization for metric learning. *Neurocomputing*, 151(1):259–267, 2015.

A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *ArXiv e-prints*, June 2013.

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

P. Bianchi, S. Clémençon, J. Jakubowicz, and G. Moral-Adell. On-line learning gossip algorithm in multi-agent systems with local decision rules. In *Proceedings of the IEEE International Conference on Big Data*, 2013.

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

L. Bottou. *Online Algorithms and Stochastic Approximations: Online Learning and Neural Networks*. Cambridge University Press, 1998.

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

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

Q. Cao, Z.-C. Guo, and Y. Ying. Generalization bounds for metric and similarity learning. Technical report, University of Exeter, July 2012. arXiv:1207.5437.

G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11:1109–1135, 2010.

S. Clémençon. A statistical view of clustering performance through the theory of U-processes. *Journal of Multivariate Analysis*, 124:42–56, 2014.

S. Clémençon and S. Robbiano. Building confidence regions for the ROC surface. *To appear in Pattern Recognition Letters*, 2014.

S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, 2009.

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

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

S. Clémençon, S. Robbiano, and N. Vayatis. Ranking data with ordinal labels: optimality and pairwise aggregation. *Machine Learning*, 91(1):67–104, 2013.

W. G. Cochran. *Sampling techniques*. Wiley, NY, 1977.

V. de la Peña and E. Giné. *Decoupling: from Dependence to Independence*. Springer, 1999.

J. C. Deville. *Réplications d'échantillons, demi-échantillons, Jackknife, bootstrap dans les sondages*. Economica, Ed. Droesbeke, Tassi, Fichet, 1987.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

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

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*. Springer, 2009.

D. K. Fuk and S. V. Nagaev. Probability inequalities for sums of independent random variables. *Prob. Th. Appl.*, 16(4):643660, 1971.

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

W. Grams and R. Serfling. Convergence rates for U-statistics and related statistics. *Ann. Stat.*, 1(1):153–160, 1973.

J. Hájek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523, 1964.

J. Hájek. Asymptotic normality of simple linear rank statistics under alternatives. *Ann. Math. Stat.*, 39:325–346, 1968.

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

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *JASA*, 47:663–685, 1951.

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

R. Jin, S. Wang, and Y. Zhou. Regularized distance metric learning: theory and algorithm. In *Advances in Neural Information Processing Systems 22*, pages 862–870, 2009.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, pages 315–323, 2013.

Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. A local search approximation algorithm for k-means clustering. *Computational Geometry*, 28(2–3):89–112, 2004.

N. Le Roux, M. W. Schmidt, and F. Bach. A Stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems 25*, pages 2672–2680, 2012.

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

A. J. Lee. *U-statistics: Theory and Practice*. Marcel Dekker, Inc., New York, 1990.

G. Lugosi. Pattern classification and learning theory. In L. Györfi, editor, *Principles of Nonparametric Learning*, pages 1–56. Springer, NY, 2002.

P. Massart. *Concentration Inequalities and Model Selection*. Lecture Notes in Mathematics. Springer, 2006.

C. McDiarmid. *On the method of bounded differences*, pages 148–188. Cambridge Univ. Press, 1989.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

R. J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2(1):39–48, 1974.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, 1980.

Y. Tillé. *Sampling Algorithms*. Springer Series in Statistics, 2006.

V. N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.