On Medians of (Randomized) Pairwise Means

Pierre Laforgue [1] **Stephan Clémençon** [1] **Patrice Bertail** [2]

Abstract

Tournament procedures, recently introduced in Lugosi & Mendelson (2016), offer an appealing alternative, from a theoretical perspective at least, to the principle of *Empirical Risk Minimization* in machine learning. Statistical learning by Median-of-Means (MoM) basically consists in segmenting the training data into blocks of equal size and comparing the statistical performance of every pair of candidate decision rules on each data block: that with highest performance on the majority of the blocks is declared as the winner. In the context of nonparametric regression, functions having won all their duels have been shown to outperform empirical risk minimizers w.r.t. the mean squared error under minimal assumptions, while exhibiting robustness properties. It is the purpose of this paper to extend this approach, in order to address other learning problems in particular, for which the performance criterion takes the form of an expectation over pairs of observations rather than over one single observation, as may be the case in pairwise ranking, clustering or metric learning. Precisely, it is proved here that the bounds achieved by MoM are essentially conserved when the blocks are built by means of independent sampling without replacement schemes instead of a simple segmentation. These results are next extended to situations where the risk is related to a pairwise loss function and its empirical counterpart is of the form of a U-statistic. Beyond theoretical results guaranteeing the performance of the learning/estimation methods proposed, some numerical experiments provide empirical evidence of their relevance in practice.

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris [2]Modal'X, UPL, Université Paris-Nanterre. Correspondence to: Pierre Laforgue <pierre.laforgue@telecom-paristech.fr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

1. Introduction

In Lugosi & Mendelson (2016), the concept of *tournament procedure* for statistical learning has been introduced and analyzed in the context of nonparametric regression, one of the flagship problems of machine learning. The task is to predict a real valued random variable (r.v.) Y based on the observation of a random vector X with marginal distribution $\mu(dx)$, taking its values in $\mathbb{R}^d$ with $d \geq 1$, say by means of a regression function $f : \mathbb{R}^d \to \mathbb{R}$ with minimum expected quadratic risk $\mathcal{R}(f) = \mathbb{E}[(Y - f(X))^2]$. Statistical learning usually relies on a training dataset $\mathcal{S}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ formed of independent copies of the generic pair (X, Y). Following the *Empirical Risk Minimization* (ERM) paradigm, one is encouraged to build predictive rules by minimizing an empirical version of the risk $\hat{\mathcal{R}}_n(f) = (1/n)\sum_{i=1}^n (y_i - f(x_i))^2$ over a class $\mathcal{F}$ of regression function candidates of controlled complexity (*e.g.* of finite VC dimension), while being rich enough to contain a reasonable approximant of the optimal regression function $f^*(x) = \mathbb{E}[Y \mid X = x]$: for any $f \in \mathcal{F}$, the risk excess $\mathcal{R}(f) - \mathcal{R}(f^*)$ is then equal to $||f - f^*||^2_{L_2(\mu)} = \mathbb{E}[(f(X) - f^*(X))^2]$. A completely different learning strategy, recently proposed in Lugosi & Mendelson (2016), consists in implementing a *tournament procedure* based on the Median-of-Means (MoM) method (see Nemirovsky & Yudin (1983)).

Precisely, the full dataset is first divided into 3 subsamples of equal size. For every pair of candidate functions $(f_1, f_2) \in \mathcal{F}^2$, the first step consists in computing the MoM estimator of the quantity $\|f_1 - f_2\|_{L_1(\mu)} := \mathbb{E}[|f_1(X) - f_2(X)|]$ based on the first subsample: the latter being segmented into $K \geq 1$ subsets of equal size (approximately), $\|f_1 - f_2\|_{L_1(\mu)}$ is estimated by the median of the collection of estimators formed by its empirical versions computed from each of the K sub-datasets. When the MoM estimate is large enough, the match between f_1 and f_2 is allowed. The rationale behind this approach is as follows: if one of the candidate, say f_2, is equal to f^*, and the quantity $\|f_1 - f^*\|_{L_1(\mu)}$ (which is less than $\|f_1 - f^*\|_{L_2(\mu)} = \sqrt{\mathcal{R}(f_1) - \mathcal{R}(f^*)}$) is large, so is its (robust) MoM estimate (much less sensitive to atypical values than sampling averages) with high probability. Therefore, f^* is compared to distant candidates only, against which it should hopefully win its matches. The second step consists in computing the MoM estimator of $\mathcal{R}(f_1) - \mathcal{R}(f_2)$

based on the second subsample for every *distant enough* candidates f_1 and f_2. If a candidate wins all its matches, it is kept for the third round. As said before, f^* should be part of this final pool, denoted by H. Finally, matches involving all pairs of candidates in H are computed, using a third MoM estimate on the third part of the data. A champion winning again all its matches is either f^* or has a small enough excess risk anyway.

It is the purpose of the present article to extend the MoM-based statistical learning methodology. Firstly, we investigate the impact of randomization in the MoM technique: by randomization, it is meant that data subsets are built through sampling schemes, say simple random sampling without replacement (SRSWoR in abbreviated form) for simplicity, rather than partitioning. Though introducing more variability in the procedure, we provide theoretical and empirical evidence that attractive properties of the original MoM method are essentially preserved by this more flexible variant (in particular, the number of blocks involved in this alternative procedure is arbitrary). Secondly, we consider the application of the tournament approach to other statistical learning problems, namely those involving pairwise loss functions, like popular formulations of ranking, clustering or metric-learning. In this setup, natural statistical versions of the risk of low variance take the form of U-statistics (of degree two), *i.e.* averages over all pairs of observations, see *e.g.* Clémençon et al. (2008). In this situation, we propose to estimate the risk by the median of U-statistics computed from blocks obtained through data partitioning or sampling. Results showing the accuracy of this strategy, referred to as *Median of (Randomized) Pairwise Means* here, are established and application of this estimation technique to pairwise learning is next investigated from a theoretical perspective and generalization bounds are obtained. The relevance of this approach is also supported by convincing illustrative numerical experiments.

The rest of the paper is organized as follows. Section 2 briefly recalls the main ideas underlying the MoM procedure, its applications to robust machine learning as well as basic concepts pertaining to the theory of U-statistics/processes. In section 3, the variants of the MoM approach we propose are described at length and theoretical results establishing their statistical performance are stated. Illustrative numerical experiments are displayed in section 4, while proofs are deferred to the Appendix section. Some technical details and additional experimental results are postponed to the Supplementary Material (SM).

2. Background - Preliminaries

As a first go, we briefly describe the main ideas underlying the tournament procedure for robust machine learning, and next recall basic notions of the theory of U-statistics, as well as crucial results related to their efficient approximation. Here and throughout, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$, the variance of any square integrable r.v. Z by $\mathrm{Var}(Z)$, the cardinality of any finite set A by $\#A$. If $(a_1, \ldots, a_n) \in \mathbb{R}^n$, the median (sometimes abbreviated med) of $a_1, \ldots, a_n$ is defined as $a_{\sigma((n+1)/2)}$ when n is odd and $a_{\sigma(n/2)}$ otherwise, σ denoting a permutation of $\{1, \ldots, n\}$ such that $a_{\sigma(1)} \leq \ldots \leq a_{\sigma(n)}$. The floor and ceiling functions are denoted by $u \mapsto \lfloor u \rfloor$ and $u \mapsto \lceil u \rceil$.

2.1. Medians of Means based Statistical Learning

First introduced independently by Nemirovsky & Yudin (1983), Jerrum et al. (1986), and Alon et al. (1999), the Median-of-Means (MoM) is a mean estimator dedicated to real random variables. It is now receiving a great deal of attention in the statistical learning literature, following in the footsteps of the results established in Audibert & Catoni (2011), Catoni (2012), where mean estimators are studied through the angle of their deviation probabilities, rather than on their traditional mean square errors, for robustness purpose. Indeed, Devroye et al. (2016) showed that the MoM provides an optimal δ-dependent subgaussian mean estimator, under the sole assumption that a second order moment exists. The MoM estimator has later been extended to random vectors, through different generalizations of the median (Minsker et al., 2015; Hsu & Sabato, 2016; Lugosi & Mendelson, 2017). In Bubeck et al. (2013), it is used to design robust bandits strategies, while Lerasle & Oliveira (2011) and Brownlees et al. (2015) advocate minimizing a MoM, respectively Catoni, estimate of the risk, rather than performing ERM, to tackle different learning tasks. More recently, Lugosi & Mendelson (2016) introduced a tournament strategy based on the MoM approach.

The MoM estimator. Let $\mathcal{S}_n = \{Z_1, \ldots, Z_n\}$ be a sample composed of $n \geq 1$ independent realizations of a square integrable real valued r.v. Z, with expectation θ and finite variance $\mathrm{Var}(Z) = \sigma^2$. Dividing $\mathcal{S}_n$ into K disjoint blocks, each with same cardinality $B = \lfloor n/K \rfloor$, $\tilde{\theta}_k$ denotes the empirical mean based on the data lying in block k for $k \leq K$. The MoM estimator $\hat{\theta}_{\mathrm{MoM}}$ of θ is then given by

$$\hat{\theta}_{\mathrm{MoM}} = \mathrm{median}(\tilde{\theta}_1, \ldots, \tilde{\theta}_K).$$

It offers an appealing alternative to the sample mean $\hat{\theta}_n = (1/n)\sum_{i=1}^n Z_i$, much more robust, *i.e.* less sensitive to the presence of atypical values in the sample. Exponential concentration inequalities for the MoM estimator can be established in heavy tail situations, under the sole assumption that the Z_i's are square integrable. For any $\delta \in [e^{1-n/2}, 1[$, choosing $K = \lceil \log(1/\delta) \rceil$ and $B = \lfloor n/K \rfloor$, we have (see *e.g.* Devroye et al. (2016), Lugosi & Mendelson (2016)):

$$\mathbb{P}\left\{ \left| \hat{\theta}_{\mathrm{MoM}} - \theta \right| > 2\sqrt{2} e \sigma \sqrt{\frac{1 + \log(1/\delta)}{n}} \right\} \leq \delta, \quad (1)$$

The tournament procedure. Placing ourselves in the distribution-free regression framework, recalled in the Introduction section, it has been shown that, under appropriate complexity conditions and in a possibly non sub-Gaussian setup, the tournament procedure outputs a candidate $\hat{f}$ with optimal accuracy/confidence tradeoff, outperforming thus ERM in heavy-tail situations. Namely, there exist c, c_0, and $r > 0$ such that, with probability at least $1-\exp(c_0 n \min\{1, r^2\})$, it holds both at the same time (see Theorem 2.10 in Lugosi & Mendelson (2016)):

$$\|\hat{f} - f^*\|_{L_2} \leq cr, \quad \text{and} \quad \mathcal{R}(\hat{f}) - \mathcal{R}(f^*) \leq (cr)^2,$$

2.2. Pairwise Means and U-Statistics

Rather than the mean of an integrable r.v., suppose now that the quantity of interest is of the form $\theta(h) = \mathbb{E}[h(X_1, X_2)]$, where X_1 and X_2 are i.i.d. random vectors, taking their values in some measurable space $\mathcal{X}$ with distribution $F(dx)$ and $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a measurable mapping, square integrable w.r.t. $F \otimes F$. For simplicity, we assume that $h(x_1, x_2)$ is symmetric (*i.e.* $h(x_1, x_2) = h(x_2, x_1)$ for all $(x_1, x_2) \in \mathcal{X}^2$). A natural estimator of the parameter $\theta(h)$ based on an i.i.d. sample $\mathcal{S}_n = \{X_1, \ldots, X_n\}$ drawn from F is the average over all pairs

$$U_n(h) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} h(X_i, X_j). \tag{2}$$

The quantity (2) is known as the U-statistic of degree two[1], with kernel h, based on the sample $\mathcal{S}_n$. One may refer to Lee (1990) for an account of the theory of U-statistics. As may be shown by a Lehmann-Scheffé argument, it is the unbiased estimator of $\theta(h)$ with minimum variance. Setting $h_1(X_1) = \mathbb{E}[h(X_1, X_2) \mid X_1] - \theta(h)$, $h_2(X_1, X_2) = h(X_1, X_2) - \theta(h) - h_1(X_1) - h_1(X_2)$, $\sigma_1^2(h) = \text{Var}\,(h_1(X_1))$ and $\sigma_2^2(h) = \text{Var}\,(h_2(X_1, X_2))$ and using the orthogonal decomposition (usually referred to as *second Hoeffding decomposition*, see Hoeffding (1948))

$$U_n(h) - \theta(h) = \frac{2}{n} \sum_{i=1}^{n} h_1(X_i) + \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} h_2(X_i, X_j),$$

one may easily see that

$$\text{Var}\,(U_n(h)) = \frac{4\sigma_1^2(h)}{n} + \frac{2\sigma_2^2(h)}{n(n-1)}. \tag{3}$$

Of course, an estimator of the parameter $\theta(h)$ taking the form of an i.i.d. average can be obtained by splitting the dataset into two halves and computing

$$M_n(h) = \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} h(X_i, X_{i+\lfloor n/2 \rfloor}).$$

One can check that its variance, $\text{Var}(M_n(h)) = \sigma^2(h)/\lfloor n/2 \rfloor$, with $\sigma^2(h) = \text{Var}(h(X_1, X_2) = 2\sigma_1^2(h) + \sigma_2^2(h)$, is however significantly larger than (3). Regarding the difficulty of the analysis of the fluctuations of (2) (uniformly over a class of kernels possibly), the reduced variance property has a price: the variables summed up being far from independent, linearization tricks (*i.e.* Hajek/Hoeffding projection) are required to establish statistical guarantees for the minimization of U-statistics. Refer to Clémençon et al. (2008) for further details.

Examples. In machine learning, various empirical performance criteria are of the form of a U-statistic.

• In clustering, the goal is to find a partition $\mathcal{P}$ of the feature space $\mathcal{X}$ so that pairs of observations independently drawn from a certain distribution F on $\mathcal{X}$ within a same cell of $\mathcal{P}$ are more similar w.r.t. a certain metric $D : \mathcal{X}^2 \to \mathbb{R}_+$ than pairs lying in different cells. Based on an i.i.d. training sample $X_1, \ldots, X_n$, this leads to minimize the U-statistic, referred to as *empirical clustering risk*:

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

where $\Phi_{\mathcal{P}}(x, x') = \sum_{\mathcal{C} \in \mathcal{P}} \mathbb{I}\{(x, x') \in \mathcal{C}^2\}$, over a class of partition candidates (see Clémençon (2014)).

• In pairwise ranking, the objective is to learn from independent labeled data $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn as a generic random pair (X, Y), where the real valued random label Y is assigned to an object described by a r.v. X taking its values in a measurable space $\mathcal{X}$, a ranking rule $r : \mathcal{X}^2 \to \{-1, 0, +1\}$ that permits to predict, among two objects (X, Y) and (X', Y') chosen at random, which one is preferred: (X, Y) is preferred to (X', Y') when $Y > Y'$ and, in this case, one would ideally have $r(X, X') = +1$, the rule r being supposed anti-symmetric (*i.e.* $r(x, x') = -r(x', x)$ for all $(x, x') \in \mathcal{X}^2$). This can be formulated as the problem of minimizing the U-statistic known as the *empirical ranking risk* (see Clémençon et al. (2005)) for a given loss function $\ell : \mathbb{R} \to \mathbb{R}_+$:

$$\widehat{\mathcal{L}}_n(r) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \ell\left(-r(X_i, X_j) \cdot (Y_i - Y_j)\right).$$

Other examples of U-statistics are naturally involved in the formulation of metric/similarity-learning tasks, see Bellet et al. (2013) or Vogel et al. (2018). We also point out that the notion of U-statistic is much more general than that considered above: U-statistics of degree higher than two (*i.e.* associated to kernels with more than two arguments)

[1] Let $d \geq n$ and $H : \mathcal{X}^d \to \mathbb{R}$ be measurable, square integrable with respect to $F^{\otimes k}$. The statistic $(n!/(n-d)!) \sum_{(i_1, \ldots, i_d)} H(X_{i_1}, \ldots, X_{i_d})$, where the sum is taken over all d-tuples of $(1, \ldots, n)$, is a U-statistic of degree d.

and based on more than one sample can be defined, see *e.g.* Chapter 14 in Van der Vaart (2000) for further details. The methods proposed and the results proved in this paper can be straightforwardly extended to this more general framework.

3. Theoretical Results

Mainly motivated by pairwise learning problems such as those mentioned in subsection 2.2, it is the goal of this section to introduce and study several extensions of the MoM approach for robust statistical learning.

3.1. Medians of Randomized Means

As a first go, we place ourselves in the setup of Section 2.1, and use the notation introduced therein. But instead of dividing the dataset into disjoint blocks, an arbitrary number K of blocks, of arbitrary size $B \leq n$, are now formed by sampling without replacement (SWoR), independently from $\mathcal{S}_n$. Each randomized data block $\mathcal{B}_k$, $k \leq K$, is fully characterized by a random vector $\boldsymbol{\epsilon}_k = (\epsilon_{k,1}, \ldots, \epsilon_{k,n})$, such that $\epsilon_{k,i}$ is equal to 1 if the i-th observation has been selected in the k-th block, and to 0 otherwise. The $\boldsymbol{\epsilon}_k$'s are i.i.d. random vectors, uniformly distributed on the set $\Lambda_{n,B} = \{\boldsymbol{\epsilon} \in \{0,1\}^n : \sum_{i=1}^n \epsilon_i = B\}$ of cardinality $\binom{n}{B}$. Equipped with this notation, the empirical mean computed from the k-th randomized block, for $k \leq K$, can be written as $\bar{\theta}_k = (1/B) \sum_{i=1}^n \epsilon_{k,i} Z_i$. The *Median-of-Randomized Means* (MoRM) estimator $\bar{\theta}_{\text{MoRM}}$ is then given by

$$\bar{\theta}_{\text{MoRM}} = \text{median}(\bar{\theta}_1, \ldots, \bar{\theta}_K). \tag{4}$$

We point out that the number K and size $B \leq n$ of the randomized blocks are arbitrary in the MoRM procedure, in contrast with the usual MoM approach, where $B = \lfloor n/K \rfloor$. However, choices for B and K very similar to those leading to (1) lead to an analogous exponential bound, as revealed by Proposition 1's proof. Because the randomized blocks are not independent, the argument used to establish (1) cannot be applied in a straightforward manner to investigate the accuracy of (4). Nevertheless, as can be seen by examining the proof of the result stated below, a concentration inequality can still be derived, using the conditional independence of the draws given $\mathcal{S}_n$, and a closed analytical form for the conditional probabilities $\mathbb{P}\{|\bar{\theta}_k - \theta| > \varepsilon \mid \mathcal{S}_n\}$, seen as U-statistics of degree B. Refer to the Appendix for details.

Proposition 1. *Suppose that $Z_1, \ldots, Z_n$ are independent copies of a square integrable r.v. Z with mean θ and variance σ^2. Then, for any $\tau \in]0, 1/2[$, for any $\delta \in [2e^{-8\tau^2 n/9}, 1[$, choosing $K = \lceil \log(2/\delta)/(2(1/2-\tau)^2) \rceil$ and $B = \lfloor 8\tau^2 n/(9\log(2/\delta)) \rfloor$, we have:*

$$\mathbb{P}\left\{ \left|\bar{\theta}_{\textit{MoRM}} - \theta\right| > \frac{3\sqrt{3}\,\sigma}{2\,\tau^{3/2}} \sqrt{\frac{\log(2/\delta)}{n}} \right\} \leq \delta. \tag{5}$$

The bound stated above presents three main differences with (1). Recall first that the number K of randomized blocks is completely arbitrary in the MoRM procedure and may even exceed n. Consequently, it is always possible to build $\lceil \log(2/\delta)/(2\,(1/2-\tau)^2) \rceil$ blocks, and there is no restriction on the range of admissible confidence levels δ due to K. Second, the size B of the blocks can also be chosen completely arbitrarily in $\{1, \ldots, n\}$, and independently from K. Proposition 1 exhibits their respective dependence with respect to δ and n. Still, B needs to be greater than 1, which results in a restriction on the admissible δ's, such as specified. Observe finally that B never exceeds n. Indeed for all $\tau \in]0, 1/2[$, $8\tau^2/(9\log(2/\delta))$ does not exceeds 1 as long as δ is lower than $2\exp(-2/9) \approx 1.6$, which is always true. Third, the proposed bound involves an additional parameter τ, that can be arbitrarily chosen in $]0, 1/2[$. As may be revealed by examination of the proof, the choice of this extra parameter reflects a trade-off between the order of magnitude of K and that of B: the larger τ, the larger K, the larger the confidence range, the lower B and the lower the constant in (5) as well. Since one can pick K arbitrarily large, τ can be chosen as large as possible in $]0, 1/2[$. This way, one asymptotically achieves a $3\sqrt{6}$ constant factor, which is the same than that obtained in Hsu & Sabato (2016) for a comparable confidence range. However, the price of such an improvement is the construction of a higher number of blocks in practice (for a comparable number of blocks, the constant in (5) becomes $27\sqrt{2}$).

Remark 1. (ALTERNATIVE SAMPLING SCHEMES) *We point out that other procedures than the SWoR scheme above (e.g. Poisson/Bernoulli/Monte-Carlo sampling) can be considered to build blocks and estimates of the parameter θ. However, as discussed in the SM, the theoretical analysis of such variants is much more challenging, due to possible replications of the same original observation in a block.*

Remark 2. (EXTENSION TO RANDOM VECTORS) *Among approaches extending MoMs to random vectors, that of Minsker et al. (2015) could be readily adapted to MoRM. Indeed, once Lemma 2.1 therein has been applied, the sum of indicators can be bounded exactly as in Proposition 1's proof. Computationally, MoRM only differs from MoM in the sampling, adding no difficulty, while multivariate medians can be computed efficiently (Hopkins, 2018).*

Remark 3. (RANDOMIZATION MOTIVATION) *Theoretically, randomization being a natural alternative to data segmentation, it appeared interesting to study its impact on MoMs. On the practical side, when performing a MoM Gradient Descent (GD), it is often needed shuffling the blocks at each step (see e.g. Remark 5 in Lecué et al. (2018)). While this shuffling may seem artificial and "ad hoc" in a MoM GD, it is already included and controlled with MoRM. Finally, extending MoU to incomplete U-statistics like in subsection 3.4 first requires a MoM randomization's study.*

3.2. Medians of (Randomized) U-statistics

We now consider the situation described in subsection 2.2, where the parameter of interest is $\theta(h) = \mathbb{E}[h(X_1, X_2)]$ and investigate the performance of two possible approaches for extending the MoM methodology.

Medians of U-statistics. The most straightforward way of extending the MoM approach is undoubtedly to form complete U-statistics based on K subsamples corresponding to sets of indexes $I_1, \ldots, I_K$ of size $B = \lfloor n/K \rfloor$ built by segmenting the full sample, as originally proposed: for $k \in \{1, \ldots, K\}$,

$$\hat{U}_k(h) = \frac{2}{B(B-1)} \sum_{(i,j) \in I_k^2,\, i<j} h(X_i, X_j).$$

The *median of U-statistics* estimator (MoU in abbreviated form) of the parameter $\theta(h)$ is then defined as

$$\hat{\theta}_{\text{MoU}}(h) = \text{median}(\hat{U}_1(h), \ldots, \hat{U}_K(h)).$$

The following result provides a bound analogous to (1), revealing its accuracy.

Proposition 2. *Let $\delta \in [e^{1-2n/9}, 1[$. Choosing $K = \lceil 9/2 \log(1/\delta) \rceil$, we have with probability at least $1-\delta$:*

$$\left| \hat{\theta}_{MoU}(h) - \theta(h) \right| \leq \sqrt{\frac{C_1 \log \frac{1}{\delta}}{n} + \frac{C_2 \log^2(\frac{1}{\delta})}{n\left(2n - 9\log \frac{1}{\delta}\right)}},$$

with $C_1 = 108\sigma_1^2(h)$ and $C_2 = 486\sigma_2^2(h)$.

We point out that another robust estimator $\hat{\theta}_{\text{MoM}}(h)$ of θ could also have been obtained by applying the classic Mo(R)M methodology recalled in subsection 2.1 to the set of $\lfloor n/2 \rfloor$ i.i.d. observations $\{h(X_i, X_{i+\lfloor n/2 \rfloor}) : 1 \leq i \leq \lfloor n/2 \rfloor\}$, see the discussion in subsection 2.2. In this context, we deduce from Eq. (1) with $K = \lceil \log(1/\delta) \rceil$ and $B = \lfloor \lfloor n/2 \rfloor / K \rfloor$ that, for any $\delta \in [e^{1-\lfloor n/2 \rfloor/2}, 1[$

$$\left| \hat{\theta}_{\text{MoM}}(h) - \theta(h) \right| \leq 2\sqrt{2} e \sigma(h) \sqrt{\frac{1 + \log(1/\delta)}{\lfloor n/2 \rfloor}}$$

with probability at least $1-\delta$. The Mo(R)M strategies on independent pairs lead to constants respectively equal to $4e\sigma(h)$ and $6\sqrt{3}\sigma(h)$. On the other hand, MoU reaches a $6\sqrt{3}\sigma_1(h)$ constant factor on its dominant term. Recalling that $\sigma^2(h) = 2\sigma_1^2(h) + \sigma_2^2(h)$, beyond the $\sqrt{2}$ constant factor, MoU provides an improvement all the more significant that $\sigma_2^2(h)$ is large. Another difference between the bounds is the restriction on δ, which is looser in the MoU case. This is due to the fact that the MoU estimator can possibly involve any pair of observations among the $n(n-1)/2$ possible ones, in contrast to $\hat{\theta}_{\text{MoM}}(h)$ that relies on the $\lfloor n/2 \rfloor$ pairs set once and for all at the beginning only. MoU however exhibits a more complex *two rates* formula, but the second term being negligible the performance are not affected, as shall be confirmed empirically.

As suggested in subsection 3.1, the data blocks used to compute the collection of K U-statistics could be formed by means of a SRSWoR scheme. Confidence bounds for such a median of randomized U-statistics estimator, comparable to those achieved by the MoU estimator, are stated below.

Medians of Randomized U-statistics. The alternative we propose consists in building an arbitrary number K of data blocks $\mathcal{B}_1, \ldots, \mathcal{B}_K$ of size $B \leq n$ by means of a SRSWoR scheme, and, for each data block $\mathcal{B}_k$, forming all possible pairs of observations in order to compute

$$\bar{U}_k(h) = \frac{1}{B(B-1)} \sum_{i<j} \boldsymbol{\epsilon}_{k,i} \boldsymbol{\epsilon}_{k,j} \cdot h(X_i, X_j),$$

where $\boldsymbol{\epsilon}_k$ denotes the random vector characterizing the $k-th$ randomized block, just like in subsection 3.1. Observe that, for all $k \in \{1, \ldots, K\}$, we have: $\mathbb{E}[\bar{U}_k(h) \mid \mathcal{S}_n] = U_n(h)$. The *Median of Randomized U-statistics* estimator of $\theta(h)$ is then defined as

$$\bar{\theta}_{\text{MoRU}}(h) = \text{median}(\bar{U}_1(h), \ldots, \bar{U}_K(h)). \tag{6}$$

The following proposition establishes the accuracy of the estimator (6), while emphasizing the advantages of the greater flexibility it offers when choosing B and K.

Proposition 3. *For any $\tau \in]0, 1/2[$, for any $\delta \in [2e^{-8\tau^2 n/9}, 1[$, choosing $K = \lceil \log(2/\delta)/(2(1/2-\tau)^2) \rceil$ and $B = \lfloor 8\tau^2 n/(9\log(2/\delta)) \rfloor$, it holds w.p.a.l. $1-\delta$:*

$$\left| \bar{\theta}_{MoRU}(h) - \theta(h) \right| \leq \sqrt{\frac{C_1(\tau) \log \frac{2}{\delta}}{n} + \frac{C_2(\tau) \log^2(\frac{2}{\delta})}{n\left(8n - 9\log \frac{2}{\delta}\right)}},$$

with $C_1(\tau) = 27\sigma_1^2(h)/(2\tau^3)$ and $C_2 = 243\sigma_2^2(h)/(4\tau^3)$.

Remark 4. *Observe that Proposition 2's constants (and bound) can be recovered asymptotically by letting $\tau \to 1/2$.*

Remark 5. *Propositions 2 and 3 remain valid for (multi-samples) U-statistics of arbitrary degree. Refer to the SM for the general statements, and discussions about related approaches (Joly & Lugosi, 2016; Minsker & Wei, 2018).*

3.3. MoU-based Pairwise Learning

We now describe a version of the tournament method tailored to *pairwise learning*. Let $\mathcal{X}$ be a measurable space, $\mathcal{F} \subset \mathbb{R}^{\mathcal{X} \times \mathcal{X}}$ a class of decision rules, and $\ell : \mathcal{F} \times \mathcal{X}^2 \to \mathbb{R}_+$ a given loss function. The goal pursued here is to learn from $2n$ i.i.d. variables $X_1, \ldots, X_{2n}$ distributed as a generic r.v. X valued in $\mathcal{X}$ a minimizer of the risk $\mathcal{R}(f) = \mathbb{E}[\ell(f, (X, X'))]$, where X' denotes an independent copy of X. In order to benefit from the standard tournament setting, we introduce the following notation: for every

$f \in \mathcal{F}$, let $H_f(X, X') = \sqrt{\ell(f,(X,X'))}$ the kernel that maps every pair (X, X') to its (square root) loss through f. Let $\mathcal{H}_{\mathcal{F}} = \{H_f : f \in \mathcal{F}\}$. It is easy to see that for all $f \in \mathcal{F}$, $\mathcal{R}(f) = \|H_f\|^2_{L_2(\mu)}$, and that if f^* and H^*_f denote respectively the $\mathcal{R}$ and L_2 minimizers over $\mathcal{F}$ and $\mathcal{H}_{\mathcal{F}}$, then $H^*_f = H_{f^*}$. First, the dataset is split into 2 subsamples $\mathcal{S}$ and $\mathcal{S}'$, each of size n. Then, a distance oracle is used to allow matches to take place. Namely, for any $f, g \in \mathcal{F}^2$, let $\Phi_{\mathcal{S}}(f, g)$ be a MoU estimate of $\|H_f - H_g\|_{L_1}$ built on $\mathcal{S}$. If $\mathcal{B}_1, \ldots, \mathcal{B}_K$ is a partition of $\mathcal{S}$, it reads:

$$\Phi_{\mathcal{S}}(f, g) = \text{med}\left(\hat{U}_1|H_f - H_g|, \ldots, \hat{U}_K|H_f - H_g|\right).$$

If $\Phi_{\mathcal{S}}(f, g)$ is greater than βr, for β and r to be specified later, a match between f and g is allowed. As shall be seen in the SM proofs, a MoRU estimate of $\|H_f - H_g\|_{L_1}$ could also have been used instead of a MoU. The idea underlying the pairwise tournament is the same than that of the standard one: with high probability $\Phi_{\mathcal{S}}(f, g)$ is a good estimate of $\|H_f - H_g\|_{L_2}$, so that only distant candidates are allowed to confront. And if H_{f^*} is one of these two candidates, it should hopefully win its match against a distant challenger. The nature of these matches is to be specified now. For any $f, g \in \mathcal{F}^2$, let $\Psi_{\mathcal{S}'}(f, g)$ denote a MoU estimate of $\mathbb{E}[H^2_f - H^2_g]$ built on $\mathcal{S}'$. With $\mathcal{B}'_1, \ldots, \mathcal{B}'_{K'}$ a partition of $\mathcal{S}'$, $\Psi_{\mathcal{S}'}(f, g)$ reads

$$\Psi_{\mathcal{S}'}(f, g) = \text{med}\left(\hat{U}_1(H^2_f - H^2_g), \ldots, \hat{U}_{K'}(H^2_f - H^2_g)\right).$$

f is declared winner of the match if $\Psi_{\mathcal{S}'}(f, g) \leq 0$, *i.e.* if $\sum_{i<j} \ell(f,(X_i, X_j))$ is lower than $\sum_{i<j} \ell(g,(X_i, X_j))$ on more than half of the blocks. A candidate that has not lost a single match it has been allowed to participate in presents good generalization properties under mild assumptions, as revealed by the following theorem.

Theorem 1. *Let $\mathcal{F}$ be a class of prediction functions, and ℓ a loss function such that $\mathcal{H}_{\mathcal{F}}$ is locally compact. Assume that there exist $q > 2$ and $L > 1$ such that $\forall H_f \in span(\mathcal{H}_{\mathcal{F}}), \|H_f\|_{L_q} \leq L\|H_f\|_{L_2}$. Let r^* (properly defined in the SM due to space limitation) that only depends on f^*, L, q, and the geometry of $\mathcal{H}_{\mathcal{F}}$ around H_{f^*}. Set $r \geq 2r^*$. Then there exist $c_0, c > 0$, and a procedure based on $X_1, \ldots, X_{2n}, L$, and r that selects a function $\hat{f} \in \mathcal{F}$ such that with probability at least $1 - \exp(c_0 n \min\{1, r^2\})$,*

$$\mathcal{R}(\hat{f}) - \mathcal{R}(f^*) \leq cr.$$

Proof. The proof is analogous to that of Theorem 2.11 in Lugosi & Mendelson (2016), and sketched in the SM.

Remark 6. *In pairwise learning, one seeks to minimize $\ell(f, X, X') = (\sqrt{\ell(f, X, X')} - 0)^2 = (H_f(X, X') - 0)^2$. We almost recover the setting of Lugosi & Mendelson (2016): quadratic loss, with $Y = 0$, for the decision function H_f. This is why any loss function ℓ can be considered, once technicalities induced by U-statistics are tackled. The control obtained on $\|H_f - H^*_f\|_{L_2}$ then translates into a control on the excess risk of f (see SM for further details).*

Remark 7. *As discussed at length in Lugosi & Mendelson (2016), computing the tournament winner is a nontrivial problem. However, one could alternatively consider performing a tournament on an ϵ-coverage of $\mathcal{F}$, while controlling the approximation error of this coverage.*

3.4. Discussion - Further Extensions

The computation of the U-statistic (2) is expensive in the sense that it involves the summation of $\mathcal{O}(n^2)$ terms. The concept of *incomplete U-statistic*, see Blom (1976), precisely permits to address this computational issue and achieve a trade-off between scalability and variance reduction. In one of its simplest forms, it consists in selecting a subsample of size $M \geq 1$ by sampling with replacement in the set of all pairs of observations that can be formed from the original sample. Setting $\Lambda = \{(i, j) : 1 \leq i < j \leq n\}$, and denoting by $\{(i_1, j_1), \ldots, (i_M, j_M)\} \subset \Lambda$ the subsample drawn by Monte-Carlo, the incomplete version of the U-statistic (2) is: $\widetilde{U}_M(h) = (1/M)\sum_{m \leq M} h(X_{i_m}, X_{j_m})$. $\widetilde{U}_M(h)$ is directly an unbiased estimator of θ with variance

$$\text{Var}\left(\widetilde{U}_M(h)\right) = \left(1 - \frac{1}{M}\right)\text{Var}(U_n(h)) + \frac{\sigma^2(h)}{M}.$$

The difference between its variance and that of (2) vanishes as M increases. In contrast, when $M \leq \#\Lambda = n(n-1)/2$, the variance of a complete U-statistic based on a subsample of size $\lfloor\sqrt{M}\rfloor$, and thus on $\mathcal{O}(M)$ pairs just like $\widetilde{U}_M(h)$, is of order $\mathcal{O}(1/\sqrt{M})$. Minimization of incomplete U-statistics has been investigated in Clémençon et al. (2016) from the perspective of scalable statistical learning. Hence, rather than sampling first observations and forming next pairs from data blocks in order to compute a collection of *complete U-statistics*, which the median is subsequently taken of, one could sample directly pairs of observations, compute alternatively estimates of drastically reduced variance and output a *Median of Incomplete U-statistics*. However, one faces significant difficulties when trying to analyze theoretically such a variant, as explained in the SM.

4. Numerical Experiments

Here we display numerical results supporting the relevance of the MoM variants analyzed in the paper. Additional experiments are presented in the SM for completeness.

MoRM experiments. Considering inference of the expectation of four specified distributions (Gaussian, Student, Lognormal and Pareto), based on a sample of size $n = 1000$, seven estimators are compared below: standard MoM, and

Table 1. Quadratic Risks for the Mean Estimation, $\delta = 0.001$

	NORMAL $(0,1)$	STUDENT (3)	LOG-NORMAL $(0,1)$	PARETO (3)
MOM	0.00149 ± 0.00218	0.00410 ± 0.00584	0.00697 ± 0.00948	$\mathbf{1.02036 \pm 0.06115}$
$\text{MORM}_{1/6,\,\text{SWOR}}$	0.01366 ± 0.01888	0.02947 ± 0.04452	0.06210 ± 0.07876	1.12256 ± 0.14970
$\text{MORM}_{3/10,\,\text{SWOR}}$	0.00255 ± 0.00361	0.00602 ± 0.00868	0.01241 ± 0.01610	1.05458 ± 0.07041
$\text{MORM}_{9/20,\,\text{SWOR}}$	$\mathbf{0.00105 \pm 0.00148}$	$\mathbf{0.00264 \pm 0.00372}$	$\mathbf{0.00497 \pm 0.00668}$	1.02802 ± 0.04903

Table 2. Quadratic Risks for the Variance Estimation, $\delta = 0.001$

	NORMAL $(0,1)$	STUDENT (3)	LOG-NORMAL $(0,1)$	PARETO (3)
$\text{MOU}_{1/2;\,1/2}$	0.00409 ± 0.00579	1.72618 ± 28.3563	2.61283 ± 23.5001	1.35748 ± 36.7998
$\text{MOU}_{\text{PARTITION}}$	$\mathbf{0.00324 \pm 0.00448}$	$\mathbf{0.38242 \pm 0.31934}$	$\mathbf{1.62258 \pm 1.41839}$	$\mathbf{0.09300 \pm 0.05650}$
$\text{MORU}_{\text{SWOR}}$	0.00504 ± 0.00705	0.51202 ± 3.88291	2.01399 ± 4.85311	0.09703 ± 0.07116

six MoRM estimators, related to different sampling schemes (SRSWoR, Monte-Carlo) or different values of the tuning parameter τ. Results are obtained through 5000 replications of the estimation procedures. Beyond the quadratic risk, accuracy of the estimators are assessed by means of deviation probabilities (see SM), *i.e.* empirical quantiles for a geometrical grid of confidence levels δ. As highlighted above, $\tau = 1/6$ leads to (approximately) the same number of blocks as in the MoM procedure. However, MoRM usually select blocks of cardinality lower than n/K, so that the MoRM estimator with $\tau = 1/6$ uses less examples than MoM. Proposition 1 exhibits a higher constant for MoRM in that case, and it is confirmed empirically here. The choice $\tau = 3/10$ guarantees that the number of MoRM blocks multiplied by their cardinality is equal to n. This way, MoRM uses as much samples as MoM. Nevertheless, the increased variability leads to a slightly lower performance in this case. Finally, $\tau = 9/20$ is chosen to be closer to $1/2$, as suggested by (5). In this setting, the two constant factors are (almost) equal, and MoRM even empirically shows a systematic improvement compared to MoM. Note that the quantile curves should be decreasing. However, the estimators being δ-dependent, different experiments are run for each value of δ, and the rare little increases are due to this random effect.

Mo(R)U experiments. In these experiments assessing empirically the performance of the Mo(R)U methods, the parameter of interest is the variance (*i.e.* $h(x,y) = (x-y)^2/2$) of the four laws used above. Again, estimators are assessed through their quadratic risk and empirical quantiles. A metric learning application is also proposed in the SM.

5. Conclusion

In this paper, various extensions of the Medians-of-Means methodology, which tournament-based statistical learning techniques recently proposed in the literature to handle heavy-tailed situations rely on, have been investigated at length. First, confidence bounds showing that accuracy can be fully preserved when data blocks are built through SRSWoR schemes rather than simple segmentation, giving more flexibility to the approach regarding the number of blocks and their size, are established. Second, its application to estimation of pairwise expectations (*i.e.* Medians-of-U-statistics) is studied in a valid theoretical framework, paving the way for the design of robust pairwise statistical learning techniques in clustering, ranking or similarity-learning tasks.

Technical Proofs

Proof of Proposition 1

Let $\varepsilon > 0$. Just like in the classic argument used to prove (1), observe that

$$\left\{\left|\bar{\theta}_{\text{MoRM}} - \theta\right| > \varepsilon\right\} \subset \left\{\sum_{k=1}^{K} I_\varepsilon(\mathcal{B}_{\boldsymbol{\epsilon}_k}) \geq K/2\right\},$$

where $I_\varepsilon(\mathcal{B}_{\boldsymbol{\epsilon}_k}) = \mathbb{I}\{\left|\bar{\theta}_k - \theta\right| > \varepsilon\}$ for $k = 1, \ldots, K$. In order to benefit from the conditional independence of the blocks given the original sample $\mathcal{S}_n$, we first condition upon $\mathcal{S}_n$ and consider the variability induced by the $\boldsymbol{\epsilon}_k$'s only:

$$\mathbb{P}\left\{\left|\bar{\theta}_{\text{MoRM}} - \theta\right| > \varepsilon \mid \mathcal{S}_n\right\} \leq \mathbb{P}\left\{\sum_{k=1}^{K} \frac{I_\varepsilon(\mathcal{B}_{\boldsymbol{\epsilon}_k})}{K} \geq \frac{1}{2}\middle| \mathcal{S}_n\right\}.$$

Now, the average $(1/K)\sum_{k=1}^{K} I_\varepsilon(\mathcal{B}_{\boldsymbol{\epsilon}_k})$ can be viewed as an approximation of the U-statistic of degree B (refer to Lee (1990)), its conditional expectation given $\mathcal{S}_n$ being

$$U_n^\varepsilon = \frac{1}{\binom{n}{B}} \sum_{\boldsymbol{\epsilon} \in \Lambda(n,B)} I_\varepsilon(\mathcal{B}_{\boldsymbol{\epsilon}}).$$

Denoting by $p^\varepsilon = \mathbb{E}[U_n^\varepsilon] = \mathbb{P}\{|\bar{\theta}_1 - \theta| > \varepsilon\}$ the expectation of the $I_\varepsilon(\mathcal{B}_{\boldsymbol{\epsilon}_k})$'s, we have $\forall \tau \in]0, 1/2[$:

$$\mathbb{P}\left\{ \left|\bar{\theta}_{\text{MoRM}} - \theta\right| > \varepsilon \right\} \leq \mathbb{P}_{\mathcal{S}_n}\left\{U_n^\varepsilon - p^\varepsilon \geq \tau - p^\varepsilon\right\} \quad (7)$$
$$+ \mathbb{E}_{\mathcal{S}_n}\left[\mathbb{P}_{\boldsymbol{\epsilon}}\left\{\frac{1}{K}\sum_{k=1}^{K} I_\varepsilon(\mathcal{B}_{\boldsymbol{\epsilon}_k}) - U_n^\varepsilon \geq \frac{1}{2} - \tau \middle| \mathcal{S}_n\right\}\right].$$

By virtue of Hoeffding inequality for i.i.d. averages (see Hoeffding (1963)) conditioned upon $\mathcal{S}_n$, we have: $\forall t > 0$,

$$\mathbb{P}_{\boldsymbol{\epsilon}}\left\{\frac{1}{K}\sum_{k=1}^{K} I_\varepsilon(\mathcal{B}_{\boldsymbol{\epsilon}_k}) - U_n^\varepsilon \geq t \middle| \mathcal{S}_n\right\} \leq \exp\left(-2Kt^2\right). \quad (8)$$

In addition, the version of Hoeffding inequality for U-statistics (*cf* Hoeffding (1963), see also Theorem A in Chapter 5 of Serfling (1980)) yields: $\forall t > 0$,

$$\mathbb{P}_{\mathcal{S}_n}\left\{U_n^\varepsilon - p^\varepsilon \geq t\right\} \leq \exp\left(-2nt^2/B\right). \quad (9)$$

One may also show (see SM A.2.1) that $p^\varepsilon \leq \frac{\sigma^2}{B\epsilon^2}$. Combining this remark with equations (7), (8) and (9), the deviation probability of $\bar{\theta}_{\text{MoRM}}$ can be bounded by

$$\exp\left(-2\frac{n}{B}\left(\tau - \frac{\sigma^2}{B\varepsilon^2}\right)^2\right) + \exp\left(-2K\left(\frac{1}{2} - \tau\right)^2\right).$$

Choosing $K = \lceil \log(2/\delta)/(2(1/2-\tau)^2 \rceil$ and $B = \lfloor 8\tau^2 n/(9\log(2/\delta)) \rfloor$ leads to the desired result. □

Proof of Proposition 2

The data blocks are built here by partitioning the original dataset into $K \leq n$ subsamples of size $B = \lfloor n/K \rfloor$. Set $I_{\varepsilon,k} = \mathbb{I}\{|\hat{U}_k(h) - \theta(h)| > \varepsilon\}$ for $k \in \{1, \ldots, K\}$. Again, observe that $\mathbb{P}\left\{\left|\hat{\theta}_{\text{MoU}}(h)(h) - \theta(h)\right| > \varepsilon\right\}$ is lower than

$$\mathbb{P}\left\{\frac{1}{K}\sum_{k=1}^{K} I_{\varepsilon,k} - q^\varepsilon \geq \frac{1}{2} - q^\varepsilon\right\},$$

where $q^\varepsilon = \mathbb{E}[I_{\varepsilon,1}] = \mathbb{P}\{|\hat{U}_1(h) - \theta(h)| > \varepsilon\}$. By virtue of Chebyshev's inequality and equation (3):

$$q^\varepsilon \leq \frac{\text{Var}\big(\hat{U}_1(h)\big)}{\varepsilon^2} = \frac{1}{\varepsilon^2}\left(\frac{4\sigma_1^2(h)}{B} + \frac{2\sigma_2^2(h)}{B(B-1)}\right).$$

Using Hoeffding inequality, the deviation probability can thus be bounded by

$$\exp\left(-2K\left(\frac{1}{2} - \left(\frac{4\sigma_1^2(h)}{B\varepsilon^2} + \frac{2\sigma_2^2(h)}{B(B-1)\varepsilon^2}\right)\right)^2\right).$$

Choosing $K = \log(1/\delta)/(2(1/2-\lambda)^2)$, $\lambda \in]0, 1/2[$ gives:

$$\varepsilon = \sqrt{C_1(\lambda)\frac{\log\frac{1}{\delta}}{n} + C_2(\lambda)\frac{\log^2(\frac{1}{\delta})}{n\left[2(\frac{1}{2}-\lambda)^2 n - \log\frac{1}{\delta}\right]}},$$

with $C_1(\lambda) = 2\sigma_1^2(h)/(\lambda(\frac{1}{2} - \lambda)^2)$ and $C_2(\lambda) = \sigma_2^2(h)/(\lambda(1/2-\lambda)^2)$. The optimal constant for the first and leading term is attained for $\lambda = 1/6$, which corresponds to $K = (9/2)\log(1/\delta)$ and gives:

$$\varepsilon = \sqrt{C_1\frac{\log\frac{1}{\delta}}{n} + C_2\frac{\log^2(\frac{1}{\delta})}{n\left(2n - 9\log\frac{1}{\delta}\right)}},$$

with $C_1 = 108\sigma_1^2(h)$ and $C_2 = 486\sigma_2^2(h)$. Finally, taking $\lceil K \rceil$ instead of K does not change the result. □

Proof of Proposition 3

Here we consider the situation where the estimator is the median of K randomized U-statistics, computed from data blocks built by means of independent SRSWoR schemes. We set $\mathcal{I}_\varepsilon(\boldsymbol{\epsilon}_k) = \mathbb{I}\left\{|\bar{U}_k(h) - \theta(h)| > \varepsilon\right\}$. For all $\tau \in]0, 1/2[$, we have:

$$\mathbb{P}\left\{\left|\bar{\theta}_{\text{MoRU}}(h) - \theta(h)\right| > \varepsilon\right\} \leq \mathbb{P}\left\{\frac{1}{K}\sum_{k=1}^{K}\mathcal{I}_\varepsilon(\boldsymbol{\epsilon}_k) \geq \frac{1}{2}\right\}$$
$$\leq \mathbb{E}\left[\mathbb{P}\left\{\frac{1}{K}\sum_{k=1}^{K}\mathcal{I}_\varepsilon(\boldsymbol{\epsilon}_k) - W_n^\varepsilon \geq \frac{1}{2} - \tau \mid \mathcal{S}_n\right\}\right]$$
$$+ \mathbb{P}\left\{W_n^\varepsilon - \bar{q}^\varepsilon \geq \tau - \bar{q}^\varepsilon\right\}, \quad (10)$$

where we set

$$W_n^\varepsilon = \mathbb{E}\left[\mathcal{I}_\varepsilon(\boldsymbol{\epsilon}_1) \mid \mathcal{S}_n\right] = \frac{1}{\binom{n}{B}}\sum_{\epsilon \in \Lambda_{n,B}}\mathcal{I}_\varepsilon(\epsilon),$$
$$\bar{q}^\varepsilon = \mathbb{E}[\mathcal{I}_\varepsilon(\boldsymbol{\epsilon}_1)] = \mathbb{P}\{|\bar{U}_1(h) - \theta(h)| > \varepsilon\}.$$

The conditional expectation W_n^ε, with mean $\bar{q}^\varepsilon$, is a U-statistic of degree B, so that Theorem A in Chapter 5 of Serfling (1980) yields:

$$\mathbb{P}\left\{W_n^\varepsilon - \bar{q}^\varepsilon \geq \tau - \bar{q}^\varepsilon\right\} \leq \exp\left(-2\frac{n}{B}\left(\frac{1}{2} - \bar{q}^\varepsilon\right)^2\right). \quad (11)$$

One may also show (see SM A.2.2) that

$$\bar{q}^\varepsilon \leq \frac{1}{\varepsilon^2}\left(\frac{4\sigma_1^2(h)}{B} + \frac{2\sigma_2^2(h)}{B(B-1)}\right). \quad (12)$$

Combining (10), standard Hoeffding's inequality conditioned on the data, as well as (11) together with (12) gives that the deviation of $\bar{\theta}_{\text{MoRU}}(h)$ is upper bounded by

$$\exp\left(-2K\left(\frac{1}{2} - \tau\right)^2\right) \quad (13)$$
$$+ \exp\left(-2\frac{n}{B}\left(\tau - \frac{1}{\varepsilon^2}\left(\frac{4\sigma_1^2(h)}{B} + \frac{2\sigma_2^2(h)}{B(B-1)}\right)\right)^2\right).$$

Choosing $K = \lceil \log(2/\delta)/(2(1/2-\tau)^2) \rceil$ and $B = \lfloor 8\tau^2 n/(9\log(2/\delta)) \rfloor$ leads to the desired bound. □

References

Alon, N., Matias, Y., and Szegedy, M. The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, 58(1):137–147, 1999.

Audibert, J.-Y. and Catoni, O. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

Bellet, A., Habrard, A., and Sebban, M. A Survey on Metric Learning for Feature Vectors and Structured Data. *ArXiv e-prints*, June 2013.

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

Brownlees, C., Joly, E., Lugosi, G., et al. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59 (11):7711–7717, 2013.

Callaert, H. and Janssen, P. The Berry-Esseen theorem for U-statistics. *The Annals of Statistics*, 6(2):417–421, 1978.

Catoni, O. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pp. 1148–1185. Institut Henri Poincaré, 2012.

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

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

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

Clémençon, S., Colin, I., and Bellet, A. Scaling-up Empirical Risk Minimization: Optimization of Incomplete U-statistics. *Journal of Machine Learning Research*, 17: 1–36, 2016.

Devroye, L., Lerasle, M., Lugosi, G., Oliveira, R. I., et al. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

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

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Hopkins, S. B. Sub-gaussian mean estimation in polynomial time. *arXiv preprint arXiv:1809.07425*, 2018.

Hsu, D. and Sabato, S. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

Jerrum, M., Valiant, L., and Vazirani, V. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

Joly, E. and Lugosi, G. Robust estimation of u-statistics. *Stochastic Processes and their Applications*, 126(12): 3760–3773, 2016.

Lecué, G. and Lerasle, M. Robust machine learning by median-of-means: theory and practice. *arXiv preprint arXiv:1711.10306*, 2017.

Lecué, G., Lerasle, M., and Mathieu, T. Robust classification via mom minimization. *arXiv preprint arXiv:1808.03106*, 2018.

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

Lerasle, M. and Oliveira, R. I. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*, 2011.

Lugosi, G. and Mendelson, S. Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*, 2016.

Lugosi, G. and Mendelson, S. Sub-gaussian estimators of the mean of a random vector. *arXiv preprint arXiv:1702.00482*, 2017.

McDiarmid, C. *On the method of bounded differences*, pp. 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.

Mendelson, S. On aggregation for heavy-tailed classes. *Probability Theory and Related Fields*, 168(3-4):641–674, 2017.

Minsker, S. and Wei, X. Robust modifications of u-statistics and applications to covariance estimation problems. *arXiv preprint arXiv:1801.05565*, 2018.

Minsker, S. et al. Geometric Median and Robust Estimation in Banach Spaces. *Bernoulli*, 21(4):2308–2335, 2015.

Nemirovsky, A. S. and Yudin, D. B. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, New-York, 1983.

Peña, V. H. and Giné, E. Decoupling: from dependence to independence. 1999.

Serfling, R. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1980.

Van der Vaart, A. *Asymptotic Statistics*. Cambridge university press, 2000.

Vogel, R., Clémençon, S., and Bellet, A. A Probabilistic Theory of Supervised Similarity Learning: Pairwise Bipartite Ranking and Pointwise ROC Curve Optimization. In *International Conference in Machine Learning*, 2018.