How Robust is the Median-of-Means?
Concentration Bounds in Presence of Outliers

Pierre Laforgue Guillaume Staerman Stephan Clémençon
LTCI, Télécom Paris, Institut Polytechnique de Paris, France
Correspondence to: pierre.laforgue@telecom-paris.fr

Abstract
In contrast to the empirical mean, the Median-of-Means (MoM) is an estimator of the mean \( \theta \) of a square integrable r.v. \( Z \), around which accurate nonasymptotic confidence bounds can be built, even when \( Z \) does not exhibit a sub-Gaussian tail behavior. Because of the high confidence it achieves when applied to heavy-tailed data, MoM has recently found applications in statistical learning, in order to design training procedures that are not sensitive to atypical nor corrupted observations. For the first time, we provide concentration bounds for the MoM estimator in presence of outliers, that depend explicitly on the fraction of contaminated data present in the sample. These results are also extended to “Medians-of-\( U \)-statistics” (i.e. averages over tuples of observations), and are shown to furnish generalization guarantees for pairwise learning techniques (e.g. ranking, metric learning) based on contaminated training data. Beyond the theoretical analysis carried out, numerical results are displayed, that provide strong empirical evidence of the robustness properties claimed by the learning rate bounds established.

1 Introduction
Independently introduced for convex optimization (Nemirovsky and Yudin, 1983), computational complexity theory (Jerrum et al., 1986), or memory-efficient frequency moments estimation (Alon et al., 1999), the Median-of-Means (MoM in abbreviated form) is a mean estimator that is easy to compute, while exhibiting attractive robustness properties. Indeed, in contrast to the sample mean estimator, it is not sensitive to the presence of observations far away from the mean on a small fraction of the data sample from which it is computed, as it is the case when considering heavy-tailed data. Precisely, consider a collection \( D_n = \{ Z_1, \ldots, Z_n \} \) of \( n \geq 1 \) independent copies of a generic real-valued square integrable r.v. \( Z \) with mean \( \theta \) and variance \( \sigma^2 \). Whereas the sample mean estimator \( \bar{\theta}_n = (1/n) \sum_{i=1}^{n} Z_i \) exhibits a sub-Gaussian tail behavior in the sole case where the r.v. \( Z \) has sub-Gaussian tails as well, the tails of the MoM estimator are always “well-behaved” (and in particular even if the r.v. \( Z \) is heavy-tailed) and thus permit to build accurate non asymptotic confidence intervals. For a given confidence \( 1 - \delta \), \( \delta \in [\exp(1 - n/2), 1] \), MoM is built as follows. Set \( K = \lceil \log(1/\delta) \rceil \leq n \), denoting by \( x \in \mathbb{R} \mapsto \lceil x \rceil \) the ceiling function, and partition the dataset \( D_n \) into \( K \) disjoint blocks \( B_1, \ldots, B_K \) of size \( B = \lfloor n/K \rfloor \), denoting by \( x \in \mathbb{R} \mapsto \lfloor x \rfloor \) the floor function. For \( k \leq K \), compute the empirical mean based on block \( B_k \): \( \hat{\theta}_k = (1/B) \sum_{i \in B_k} Z_i \). The MoM estimator is then obtained by computing the median of the block averages (see Figure 1):

\[
\hat{\theta}_{\text{MoM}} = \text{median}(\hat{\theta}_1, \ldots, \hat{\theta}_K).
\] (1)

In the case where the median is not uniquely defined, one may choose the smallest one by convention. One may then show (see e.g. Devroye et al. (2016)) that with probability at least \( 1 - \delta \), we have:

\[
|\hat{\theta}_{\text{MoM}} - \theta| \leq 2\sqrt{2e} \sigma \sqrt{\frac{1 + \log(1/\delta)}{n}}.
\] (2)
Remarkably, the tails of MoM are sub-Gaussian under the sole assumption that $Z$ has finite variance. Hence, the accuracy of MoM is not damaged by the presence of atypical observations on a small part of the sample $D_n$, as may be the case when considering data drawn from a heavy-tailed distribution. The robustness properties of MoM do not only consist in coping with heavy-tailed situations, and some recent works point out the capacity of MoM to resist to the presence of outliers within the sample, i.e., data drawn from a different distribution than that of $Z$, possibly adversarial (Lecué and Lerasle, 2017). However, no concentration result for the MoM estimator in a contaminated setup is documented in the literature, to the best of our knowledge. It is the first goal of this paper to bridge this gap and develop a dedicated framework, inspired from Lecué et al. (2018). The dataset $D_n$ is now supposedly contaminated with $n_0 < n/2 - 1$ outliers (by definition, the majority of the data are not outliers), on which we make no assumption. In particular, their order of magnitude may be arbitrarily large, jeopardizing the use of the empirical mean. The rest of the sample is assumed to be composed of $n - n_0$ i.i.d. realizations of the r.v. $Z$. In this paper, we shall work under the very general assumption, stipulating no restriction on the distribution of the outliers (except independence).

**Assumption 1.** The sample $D_n = \{Z_1, \ldots, Z_n\}$ is composed of $n \geq 1$ independent observations, including $n - n_0 > n/2$ realizations of the r.v. $Z$ and $n_0$ outliers with arbitrary distributions.

We prove concentration bounds for the estimator (1), generalizing that stated in (2) (which corresponds to the case $n_0 = 0$), as well as bounds in expectation under the additional hypothesis that the number of outliers grows sub-linearly with $n$ (i.e. $n_0 = O(n^{\alpha_0})$ with $\alpha_0 \in [0, 1]$). We next extend these results to the Medians-of-$U$-statistics (i.e. medians of averages over tuples computed on blocks of observations), proposed and analyzed in Joly and Lugosi (2016) and Laforgue et al. (2019). Recently, the MoM approach has found applications in statistical learning, where it is used to design methods that can cope with heavy-tailed data in a sound validity framework. This includes e.g. the adaptation of bandit algorithms in Bubeck et al. (2013), or the MoM-tournament procedure for regression introduced in Lugosi and Mendelson (2019). The case of contaminated data is addressed through the general angle of MoM-minimization in Lecué et al. (2018), while an application to Maximum Mean Discrepancy is developed in Lerasle et al. (2019). We show that the concentration inequalities primarily established further allow to derive generalization bounds for MoM-minimization of pairwise learning criteria. This provides guarantees for a wide class of problems, ranging from ranking to metric-learning, even when considered on contaminated training data. Numerical experiments have also been carried out, providing strong empirical evidence of the relevance of the approach promoted.

The article is organized as follows. The main results of this paper are stated in Section 2, that investigates the concentration properties of the MoM estimator, and its extensions to $U$-statistics, in presence of outliers. These results are next used in Section 3 in order to establish statistical guarantees for the generalization capacity of pairwise learning techniques based on contaminated data. Due to space limitation, numerical results and technical details are deferred to the Supplementary Material.

## 2 Revisiting the MoM Concentration Properties in Presence of Outliers

The resurgence of interest for MoM in the statistical literature dates back to the seminal deviation studies by Audibert and Catoni (2011) and Catoni (2012), that propose to assess an estimator through its deviation probabilities, rather than by computing its quadratic risk. Extensively studied since then, MoM now benefits from a large corpus of concentration results, ranging from the standard scalar scenario (Devroye et al., 2016) to the more involved case of random vectors (Minsker et al., 2015; Hsu and Sabato, 2016; Lugosi and Mendelson, 2017)), or that of (randomized) tuples of observations (Joly and Lugosi, 2016; Laforgue et al., 2019). In this section, we study the concentration properties of the MoM estimator, and those of its recent extensions to $U$-statistics, in presence of outliers.
2.1 Concentration Bounds for the MoM Estimator based on Contaminated Data

As a first go, we prove an extension of bound (2) when the sample $D_n$ from which MoM is computed includes a proportion of outliers $\tau = n_0/n < 1/2$. The following two assumptions are involved in the subsequent analysis.

**Assumption 2.** The mapping $\alpha : [0, 1/2] \rightarrow [0, 1]$ is such that: $\forall \tau \in [0, 1/2]$, $2\tau < \alpha(\tau) < 1$. From mapping $\alpha$, we define $\beta : \tau \mapsto 2\alpha(\tau)/(\alpha(\tau) - 2\tau), \gamma : \tau \mapsto \sqrt{\alpha(\tau)/(\alpha(\tau) - 2\tau)}, \Delta : \tau \mapsto \sqrt{\alpha(\tau)/(\alpha(\tau) - 2\tau)}, \xi : \tau \mapsto (\alpha(\tau) - \tau)/\alpha(\tau)$.

The functions listed above are involved in the formulation of the bounds given below, the mappings $\beta, \gamma, \Delta, \xi$ being entirely determined by the initial choice of function $\alpha$. This choice shapes the constants in the upper bounds, as well as the range of confidence levels for which they hold true (however, it does not affect the rate). This yields a subtle balance between accuracy and the possible area of variation for the confidence level when $\alpha(\tau) \geq 2\tau$, as discussed after Proposition 1.

**Example 1.** Examples of mappings $\alpha$ fulfilling Assumption 2 and the related functions $\beta, \gamma, \Delta, \xi$ are given in the Table below. One may refer to Figure 5 for a visual representation.

|             | $\alpha(\tau)$ | $\beta(\tau)$ | $\gamma(\tau)$ | $\Delta(\tau)$ | $\xi(\tau)$ |
|-------------|----------------|---------------|----------------|----------------|-------------|
| ARITHMETIC  | $\frac{1 + 2\tau}{2}$ | $\frac{2(1 + 2\tau)}{1 - 2\tau}$ | $\sqrt{1 + 2\tau}/(1 - 2\tau)^{3/2}$ | $\sqrt{1 + 2\tau}/\alpha(\tau) - 2\tau$ | $\sqrt{1 + 2\tau}/\alpha(\tau) - 2\tau$ |
| GEOMETRIC   | $\sqrt{2\tau}$ | $\frac{2(1 + \sqrt{2}\tau)}{1 - 2\tau}$ | $\frac{(2 - \sqrt{2}\tau)(1 + \sqrt{2}\tau)^{3/2}}{2(1 - 2\tau)^{3/2}}$ | $\frac{1 + \sqrt{2}\tau}{\sqrt{1 - 2\tau}}$ | $\frac{\sqrt{2\tau}}{\alpha(\tau) - 2\tau}$ |
| HARMONIC    | $\frac{4\tau}{1 + 2\tau}$ | $\frac{4}{1 - 2\tau}$ | $\sqrt{\frac{3 - 2\tau}{2(1 - 2\tau)^{3/2}}}$ | $\sqrt{\frac{\sqrt{3 - 2\tau}}{\sqrt{1 - 2\tau}}}$ | $\frac{\sqrt{4}}{\alpha(\tau) - 2\tau}$ |
| POLYNOMIAL  | $\frac{\tau(5/2 - \tau)}{1 - 2\tau}$ | $\frac{2(5 - 2\tau)}{1 - 2\tau}$ | $\frac{(3 - 2\tau)(5 - 2\tau)^{3/2}}{2(1 - 2\tau)^{3/2}}$ | $\frac{\sqrt{5 - 2\tau}}{\sqrt{1 - 2\tau}}$ | $\frac{\sqrt{5 - 2\tau}}{\alpha(\tau) - 2\tau}$ |

**Assumption 3.** There exist constants $C_0 \geq 1$ and $\alpha_0 \in [0, 1]$ such that: $\forall n \geq 1$, $n_0 \leq C_0 n^{\alpha_0}$.

The following proposition describes how the confidence bounds related to the MoM estimator and its expected error are affected by the fraction $\tau = n_0/n$ of outliers within the sample $D_n$.

**Proposition 1.** Suppose that Assumption 1 is fulfilled, that $(\alpha, \beta, \gamma, \Delta)$ satisfy Assumption 2, and let $\tau = n_0/n < 1/2$. Then, for all $\delta \in [\exp(-n/\beta(\tau)), \exp(-n\alpha(\tau)/\beta(\tau))]$, choosing $K = [\beta(\tau)\log(1/\delta)]$, we have, with probability at least $1 - \delta$:

$$|\hat{\theta}_\text{MoM} - \theta| \leq 4\sqrt{\tau}\sigma \gamma(\tau) \sqrt{\frac{1 + \log(1/\delta)}{n}}. \quad (3)$$

If in addition the r.v. $Z$ is sub-Gaussian with parameter $\rho > 0$ (i.e. $\forall \lambda \in \mathbb{R}, \mathbb{E}[\exp(\lambda X)] \leq \exp(\rho^2\lambda^2/2)$), then for all $\delta \in [0, \exp(-4n\alpha(\tau))]$, with $K = \lceil \alpha(\tau)n \rceil$, it holds w.p.a.l. $1 - \delta$:

$$|\hat{\theta}_\text{MoM} - \theta| \leq 4\rho \Delta(\tau) \sqrt{\frac{\log(1/\delta)}{n}}. \quad (4)$$

Finally, if $n_0$ additionally satisfies Assumption 3, we have:

$$\mathbb{E} \left[ |\hat{\theta}_\text{MoM} - \theta| \right] \leq 2\rho \Delta(\tau) \left( 4C_0 \frac{\Delta(\tau)}{\eta(1-\alpha_0)/2} + \sqrt{\frac{\pi}{n}} \right).$$

The technical proof is detailed in the Supplementary Material. Its argument essentially consists in using that the MoM estimator (1) has a behavior similar to that of a majority of block means. The condition $K > 2n_0$ is strengthened into $K \geq \alpha(\tau)n$, where the function $\alpha$ is a strict upper bound of the mapping $\tau \mapsto 2\tau$ on $[0, 1/2]$; ensuring that a fraction $\eta(\tau) = (\alpha(\tau) - \tau)/\alpha(\tau) > 1/2$ of "sane"
blocks (i.e. including none of the \( n_\sigma \) outliers) actually constitutes a majority of blocks. One may then focus on the same blocks deviations only, which is controlled by means of the concentration properties of a Binomial random variable. The rest of the proof is quite similar to that used for establishing (2).

The sub-Gaussian assumption allows for a sharper analysis of what happens on the sane blocks, resulting in an improved confidence interval (notice that the choice of \( K \) then becomes independent from \( \delta \)), while the expectation bound is classically obtained by integrating the tail probability bound.

If the number \( n_\sigma \) of outliers is generally unknown in practice, observe that the above stated bounds may still be used with an overestimation of the fraction \( \tau \) of outliers, at the price of a loss of accuracy though. The bounds indeed explode when \( \tau \) tends to 1/2, while Equation (2) is recovered for \( \tau = 0 \). However, the very flexible formulation proposed, which involves various constants depending both on \( \tau \) and the choice of upper bound \( \alpha \), calls for an in-depth discussion.

A \( \delta \)-limited sub-Gaussian Tail. We first point out that the main price to pay for extending the sub-Gaussian tail behavior of MoM to the contaminated framework considered here is the limited range of acceptable confidence levels \( 1 - \delta \). This type of limitation is typical of MoM’s concentration results. The lower limit value for \( \delta \) is due to the constraint \( K \leq n \), and is not very compelling in practice as it decays to zero exponentially fast as \( n \) increases. The upper limit value for \( \delta \) comes from the constraint \( 2n_\sigma < K \) (or \( \sigma(\tau)n_\sigma < K \)) and is specific to the contaminated framework. It should be noticed that this restriction vanishes (i.e. the upper limit value is equal to 1) when \( \tau = 0 \) for all upper bound functions given in Example 1 except the arithmetic mean. Notice also that the lower limit restriction is removed when assuming that the r.v. \( Z \) is sub-Gaussian. We incidentally underline that the sub-Gaussian assumption only applies to \( Z \) and not to the \( n_\sigma \) outliers. Since the latter are possibly unbounded, any hope of using reliably the usual empirical mean estimator must be abandoned.

Accuracy vs Range of Confidence Levels. As previously mentioned, the choice of \( \alpha \) determines the range \([\exp(-n/\beta(\tau)), \exp(-n\alpha(\tau)/\beta(\tau))]\) for which Equation (3) holds true with probability at least \( 1 - \delta \), and the constant \( \gamma(\tau) \) at the same time. One may easily see that, when \( \tau \in [0, 1/2] \) is fixed, the quantity \( \gamma(\tau) \) monotonically decreases as \( \alpha(\tau) \) increases from 2\( \tau \) to 1/2: \( (\partial \gamma^2/\partial \alpha)(\alpha) = -4\tau(\alpha - \tau)^2/(\alpha - 2\tau)^4 < 0 \), with the notation \( \gamma^2(\alpha) = \alpha(\alpha - \tau)^2/(\alpha - 2\tau)^3 \). Hence, the larger \( \alpha(\tau) \), the smaller the constant in the upper bound. Similarly, when \( \tau \in [0, 1/2] \) is fixed, it can easily be seen that the range size increases with \( \alpha(\tau) \) on \([2\tau, \sqrt{2\tau}] \), and decreases on \([\sqrt{2\tau}, 1] \). Indeed, at the log scale, it is equal to \( s_\tau(\alpha) \) where \( s_\tau(\alpha) = n(\alpha - 2\tau)(1 - \alpha)/(2/\alpha) \), and \( (\partial s_\tau/\partial \alpha)(\alpha) = n(2\tau - \alpha^2)/(2/\alpha^2) \) for \( \alpha \in [0, 1/2] \). Consequently, choosing \( \alpha(\tau) \) larger than \( \sqrt{2\tau} \) (i.e. the geometric mean) still reduces the constant \( \gamma(\tau) \), but at the price of a smaller range for the confidence levels. A similar phenomenon is also true for the bound (4): there is a trade-off between the size of the range for the confidence levels and the order of magnitude of the constant \( \Gamma(\tau) \), both decreasing with \( \alpha(\tau) \). After integration, this tradeoff translates into the opposition between constants \( \Gamma(\tau) \) and \( \Delta(\tau) \), which have inverse monotonicity w.r.t. \( \alpha(\tau) \). The fact that \( \Delta(\tau) \to \infty \) when \( \tau \to 0 \) for some choices of \( \alpha \) may reflect an artifact of the proof technique. Indeed, if \( \tau = n_\sigma = 0 \), it is not allowed to multiply/divide by \( \tau \) in Equation (12). In contrast, one may use \( \delta \leq 1/\epsilon \) instead of Equation (11), which then gives a \( 1/\sqrt{n} \) term, with no dependence in \( \Delta \). Details are left to the reader. Finally notice that for all choices in Example 1, functions \( \gamma \) and \( \Gamma \) express as \( c(\tau)/(1 - 2\tau)^{3/2} \) and \( C(\tau)/\sqrt{1 - 2\tau} \) respectively, with \( c \) and \( C \) bounded on \([0, 1/2] \), nicely exhibiting the dependency w.r.t. the critical point \( \tau = 1/2 \) while making it easier to appreciate the differences, see Figure 5.

Rate bound. We underline that the rate \( 1/\sqrt{n^{1-\alpha}} \) for the mean deviation is in accordance with the expectations. Indeed, MoM trades the ability of discarding outliers for the degradation of its statistical guarantees to those of one single sane block, of order \( 1/\sqrt{B} \sim \sqrt{K/n} \sim \sqrt{n_\sigma}/n \), as \( K \) is roughly of the order of \( n_\sigma \). Hence, if \( n_\sigma \) grows linearly with \( n \), then \( B \) stays bounded and guarantees do not improve with \( n \). This also highlights the importance of not choosing a too rough upper bound \( \alpha \).

Related Work. Although they are quite similar in spirit, six critical points distinguish Proposition 1 from Theorem 1 in Lerasle et al. (2019). (1) It is important to notice first that Proposition 1 focuses on the deviations of scalar MoMs, while Theorem 1 in Lerasle et al. (2019) addresses that of kernel mean embeddings, seen as MoM minimizers. (2) This being said, our choice of \( K \) can be computed explicitly from the total proportion of outliers \( \tau \), and the targeted confidence \( \delta \). In contrast, the number of blocks in Lerasle et al. (2019) depends on the proportion of outliers with respect to the number of blocks itself, resulting in a recursive definition, hard to disambiguate. This inherent difficulty is typically overcome here through the \( \eta(\tau) \) reparametrization. (3) As a consequence, our bound features the true and fixed proportion of outliers \( \tau \) within the sample, while Lerasle et al. (2019) use the
where, for $\tau$ and $n$ only, for each possible choice of the upper bound function $\alpha$. (5) Lerasle et al. (2019) require $2n\alpha \leq K \leq n/2$, meaning they allow at most $25\%$ of outliers, while we can handle up to $50\%$. This may change when considering multi-sample $U$-statistics however. (6) Only a rough estimate of $K$ is prescribed in Lerasle et al. (2019), that might not be an integer.

2.2 Extension - Concentration Bounds for the Median of $U$-statistics in Presence of Outliers

Many machine learning problems can be formulated as the minimization of a certain $U$-statistic, an average over tuples of observations, generalizing the basic sample mean (one may refer to Lee (1990) for an account of the theory of $U$-statistics): ranking (see e.g. Clémençon et al. (2008)), clustering (see e.g. Clémençon (2014)) or metric-learning (see Vogel et al. (2018)) among others. We recall that $\hat{\theta}_{\text{MoU}}(h) = \frac{1}{2} \sum_{i,j=1}^{d} h(z_1, \ldots, z_d)$ is defined implicitly, whereas we provide an explicit interval, square integrable w.r.t. $\mu$ and based on independent copies $Z_1, \ldots, Z_n$ for all $(z_1, \ldots, z_d) \in \mathbb{R}^d$. A single outlier affecting $(n-1)$ terms among those averaged in (5), it is essential to design robust alternatives. Medians-of-$U$-statistics (MoU) naturally extend the MoM approach by considering the median of $U$-statistics built on disjoint blocks $B_1, \ldots, B_k$ of size $B \geq d$ (see Joly and Lugosi (2016) for the case of degenerate $U$-statistics, or Laforgue et al. (2019) for a general study based on randomized, possibly overlapping, blocks).

The MoU estimator $\hat{\theta}_{\text{MoU}}(h)$ of parameter $\theta(h)$ is defined as follows:

$$\hat{\theta}_{\text{MoU}}(h) = \text{median} \left( \hat{U}_k(h), \ k \leq K \right), \quad \text{with} \quad \hat{U}_k(h) = \frac{1}{\binom{d}{k}} \sum_{i_1 < \cdots < i_k \in B_k} h(Z_{i_1}, \ldots, Z_{i_d}).$$

See Figure 2 for a visual illustration of $\hat{\theta}_{\text{MoU}}(h)$’s construction. The proposition below then extends Proposition 1 to the latter, when it is based on a contaminated sample $\mathcal{D}_n$.

**Proposition 2.** Suppose that Assumption 1 is fulfilled, and that $(\alpha, \beta, \gamma, \Gamma, \Delta)$ satisfy Assumption 2. Let $\tau \leq n\alpha/n < 1/2$, and let $\Sigma(h)$ denote $\Sigma^2(h) = d! \sum_{i=1}^{d} \binom{d}{i} \varsigma_i(h)$. Then, for all $\delta \in \left[\exp(-n/\beta(\tau)), \exp(-n\alpha(\tau)/\beta(\tau))\right]$, choosing $K = \lceil \beta(\tau) \log(1/\delta) \rceil$, we have with probability larger than $1 - \delta$:

$$\left| \hat{\theta}_{\text{MoU}}(h) - \theta(h) \right| \leq 4\sqrt{\epsilon} \Sigma(h) \gamma(\tau) \sqrt{\frac{1 + \log(1/\delta)}{n}}.$$
Finally, if the essential supremum \( \|h(Z_1, \ldots, Z_d)\|_\infty = \inf\{t \geq 0 : P\{h(Z_1, \ldots, Z_d) > t\} = 0\} \) of the r.v. \( h(Z_1, \ldots, Z_d) \) is finite, then for all \( \delta \in]0, \exp(-4n\alpha(\tau))[, \) with \( K = \lceil \alpha(\tau)n \rceil \), we have w.p.a.l. \( 1 - \delta \):

\[
|\hat{\theta}_{\text{MoU}}(h) - \theta(h)| \leq 4\sqrt{d} \|h(Z_1, \ldots, Z_d)\|_\infty \Gamma(\tau) \frac{\log(1/\delta)}{n}.
\]

Finally, if \( n_0 \) additionally satisfies Assumption 3, we have:

\[
E \left[ |\hat{\theta}_{\text{MoU}}(h) - \theta(h)| \right] \leq 2\sqrt{d} \|h(Z_1, \ldots, Z_d)\|_\infty \Gamma(\tau) \left( 4C_0 \frac{\Delta(\tau)}{n^{1-\alpha_0}/2} + \sqrt{\frac{\pi}{n}} \right).
\]

The proof adapts that of Proposition 1 to \( U \)-statistics, see the Supplementary Material for details.

**Multi-sample generalization.** The notion of \( U \)-statistic can be readily extended to the multi-sample framework, see Lee (1990). For notational simplicity, we restrict ourselves to 2-sample \( U \)-statistics of degrees \((1, 1)\). Extensions to \( U \)-statistics of arbitrary degrees and/or based on more than two samples are straightforward and left to the reader. The \( U \)-statistic of degrees \((1, 1)\) with kernel \( H : \mathbb{R}^2 \to \mathbb{R} \), square integrable w.r.t. \( \mu \otimes \nu \), denoting by \( \mu \) and \( \nu \) the distributions of r.v. \( X \) and \( Y \) respectively, based on two independent samples, i.e. \( n \geq 1 \) independent copies \( X_1, \ldots, X_n \) of \( X \) and \( m \geq 1 \) independent copies \( Y_1, \ldots, Y_m \) of \( Y \), is given by:

\[
\hat{U}_{n,m}(H) = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m H(X_i, Y_j).
\]

It is the unbiased estimator of \( \theta(H) = \int \int H(x,y)\mu(dx)\nu(dy) \) with minimum variance, given by:

\[
\sigma^2_{n,m}(H) = \frac{1}{nm} \sigma^2(H) + \frac{m-1}{nm} \sigma_1^2(H) + \frac{n-1}{nm} \sigma_2^2(H) - \frac{\sigma^2(H) + \sigma_1^2(H) + \sigma_2^2(H)}{n \wedge m},
\]

(6)

where \( \sigma^2(H) = \text{Var}(H(X,Y)) \), \( \sigma_1^2(H) = \text{Var}(H_1(X)) \) and \( \sigma_2^2(H) = \text{Var}(H_2(Y)) \), with \( H_1(x) = E[H(x,Y)] \) and \( H_2(y) = E[H(X,y)] \). Similarly to MoM, each sample is divided into \( K_X \) (respectively \( K_Y \)) disjoint blocks of size \( B_X = \lfloor n/K_X \rfloor \) (respectively \( B_Y = \lfloor m/K_Y \rfloor \)). The Median-of-(two-sample) \( U \)-statistics estimator of \( \theta(H) \) is given by:

\[
\hat{\theta}_{\text{MoU}_2}(H) = \text{median} \left( \hat{U}_{k,l}(H), k \leq K_X, l \leq K_Y \right), \quad \text{with} \quad \hat{U}_{k,l}(H) = \frac{\sum_{i,j \in B_X \times B_Y} H(X_i, X_j)}{B_X B_Y}.
\]

Refer to Figure 3 for a visual interpretation in the particular case \( K_X = K_Y = 3 \). For \( \text{MoU}_2 \), the total number of blocks created is thus \( K_X K_Y \), while the number of corrupted ones is always lower than \( n_0 K_Y + m_0 K_X - n_0 m_0 \). As we still want at least twice more blocks than possibly corrupted ones, the constraint on \( K_X \) and \( K_Y \) can then be expressed as:

\[
2(n_0 K_Y + m_0 K_X - n_0 m_0) \leq 2(\tau_X + \tau_Y - \tau_X \tau_Y)nm < K_X K_Y \leq nm.
\]

The proportions of outliers \( \tau_X \) and \( \tau_Y \) for which we can derive guarantees should therefore satisfy \( \tau_X + \tau_Y - \tau_X \tau_Y < 1/2 \), which is a stronger requirement than for MoM, see Figure 6. The next proposition then details the concentration properties of \( \text{MoU}_2 \) under this assumption.
Proposition 3. Suppose that both samples are verifying Assumption 1, and that \((\alpha, \gamma, \eta)\) satisfy Assumption 2. Let \(\tilde{\tau} = \tau_X + \tau_Y - \tau_X \tau_Y\), and assume that \(\tilde{\tau} < 1/2\). Then, for all \(\delta \in [2 \max(\exp(-n \beta_X), \exp(-m \beta_Y)), 2 \min(\exp(-n \sqrt{\alpha(\tilde{\tau})} / \beta_X), \exp(-m \sqrt{\alpha(\tilde{\tau})} / \beta_Y))\], choosing \(K_X = \lfloor \beta_X \log(2/\delta) \rfloor\), and \(K_Y = \lfloor \beta_Y \log(2/\delta) \rfloor\), we have with probability at least \(1 - \delta\):

\[
|\hat{\theta}_{\text{MoU}_2}(H) - \theta(H)| \leq 12 \sqrt{3} \Sigma(H) \gamma(\tilde{\tau}) \sqrt{\frac{1 + \log(2/\delta)}{n \wedge m}},
\]

with \(\Sigma^2(H) = \sigma^2(H) + \sigma^2_1(H) + \sigma^2_2(H), \beta_Z = \frac{18 \eta^2(\tilde{\tau})}{n \eta(2 \eta(\tilde{\tau}) - 1)}, \text{ and } \eta_Z = 1 - \frac{\eta(\tilde{\tau})}{\sqrt{n(\tilde{\tau})}}, \text{ for } Z = X, Y.

The proof involves the same ingredients as that of Proposition 1, except that non-independent random variables are considered (see Figure 3). The conditional Hoeffding’s inequality then provides an alternative to the Binomial concentration, with the major drawback that it does not allow for a sharp analysis if one further assumes that \(\|H(X, Y)\|_\infty\) is finite (see also Remark 1’s discussion). As a result, we were not able to derive guarantees for MoU2 on an extended range, nor to bound its expected deviation, as in Proposition 1. Notice that randomized extensions considered in Laforgue et al. (2019) rely on Hoeffding’s inequality as well, and consequently suffer from the same limitation.

A block-diagonal variant. An alternative to get independent \(U\)-statistics, referred to as \(\text{MoU}_2^{\text{diag}}\), consists in considering the diagonal blocks only, cf. Figure 4. Then, one must set \(K_X = K_Y = K\), and the resulting estimator is

\[
\hat{\theta}_{\text{MoU}_2^{\text{diag}}}(H) = \text{median}\left(\tilde{U}_{k,k}(H), \ k \leq K\right),
\]

(7)

the constraint on \(K\) becoming: \(2(n_0 + m_0) < K \leq \min(n, m)\). Obviously, as soon as \(m \leq 2n_0\) this cannot be satisfied. To avoid such problems, we shall assume that \(n = m\) (see the discussion at the end of the section), which allows to analyze the concentration properties of estimator (7).

Proposition 4. Suppose that both samples are verifying Assumption 1, that \(\alpha\) satisfies Assumption 2, and that \(\tau_X + \tau_Y < 1/2\). Then, for all \(\delta \in [\exp(-n / \beta(\tau_X + \tau_Y)), \exp(-n \alpha(\tau_X + \tau_Y) / \beta(\tau_X + \tau_Y))]\), choosing \(K = \lfloor \beta(\tau_X + \tau_Y) \log(1/\delta) \rfloor\), we have with probability at least \(1 - \delta\):

\[
|\hat{\theta}_{\text{MoU}_2^{\text{diag}}}(H) - \theta(H)| \leq 4 \sqrt{e} \Sigma(H) \gamma(\tau_X + \tau_Y) \sqrt{\frac{1 + \log(1/\delta)}{n}}.
\]

If in addition \(\|H(X, Y)\|_\infty\) is finite, then for all \(\delta \in [0, \exp(-4n \alpha(\tau_X + \tau_Y))]\), choosing \(K = \lfloor \alpha(\tau_X + \tau_Y) n \rfloor\), it holds with probability at least \(1 - \delta\):

\[
|\hat{\theta}_{\text{MoU}_2^{\text{diag}}}(H) - \theta(H)| \leq 8 \|H(X, Y)\|_\infty \Gamma(\tau_X + \tau_Y) \sqrt{\frac{\log(1/\delta)}{n}}.
\]

Finally, if \(n_0\) additionally satisfies Assumption 3, we have:

\[
\mathbb{E}\left[|\hat{\theta}_{\text{MoU}_2^{\text{diag}}}(H) - \theta(H)|\right] \leq 4 \|H(X, Y)\|_\infty \Gamma(\tau_X + \tau_Y) \left(4C_0 \frac{\Delta(\tau_X + \tau_Y)}{n(1 - \alpha \gamma)/2} + \sqrt{\frac{\pi}{n}} \right).
\]

The proof is an adaptation of that of Proposition 1, technical details can be found in the Supplementary Material. Notice that the constraint \(n = m\) can be relaxed, as long as \(2(n_0 + m_0) \leq \min(n, m)\) still holds. However, the case \(n = m\) is the only one documented in MoM’s literature to our knowledge (Lerasle et al., 2019), while it nicely exhibits in Proposition 4 the critical point \(\tau_X + \tau_Y = 1/2\). When estimating Integral Probability Metrics (Sriperumbudur et al., 2012), two-sample \(U\)-statistics arise, with kernels of the form \(H_\phi(X, Y) = \phi(X) - \phi(Y)\), for \(\phi\) in the functional set considered. Hence, one might use a MoM-MoM estimate instead of MoU2 (see Staereman et al. (2020) for an application to the Wasserstein distance). The resulting proportions of outliers admitted would be \(\tau_X < 1/2\), and \(\tau_Y < 1/2\), which is looser than MoU2’s condition. Both constraints are depicted in Figure 6, and would write for \(p\)-sample \(U\)-statistics: \(\|\tau\|_\infty < 1/2\) for a MoM-based estimate, and \(\|\tau\|_1 < 1/2\) for MoUp, with \(\tau = (\tau_1, \ldots, \tau_p)\) the vector containing the \(p\) samples proportions of outliers.

3 Statistical Guarantees for Pairwise Learning in Presence of Outliers

A simple and meaningful way to illustrate the relevance of MoM-based estimators in presence of outliers is to use them for revisiting the Empirical Risk Minimization paradigm (ERM, see e.g.
when based on contaminated training data. Consider a generic supervised learning problem, defined by a pair of input/output random variables $Z = (X, Y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with unknown distribution $P$, and a loss function $\ell: \mathcal{G} \times \mathcal{Z} \to \mathbb{R}_+$. ERM then consists in substituting the unknown risk $\mathbb{E}_P \{\ell(g, Z)\}$ by its empirical version based on a sample $\mathcal{D}_n = \{Z_1, \ldots, Z_n\}$ of i.i.d. realizations of $Z$, and solving next $\min_{g \in \mathcal{G}} (1/n) \sum_{i=1}^n \ell(g, Z_i)$. There exists $\min_{g \in \mathcal{G}} \text{MoM}_{\mathcal{D}_n} \{\ell(g, Z)\}$. This framework, explored in Lecué et al. (2018) for standard MoMs by means of ad hoc Rademacher complexities tailored to outliers, is referred to as MoM-minimization. This section builds upon Section 2’s concentration bounds to extend these ideas to pairwise learning problems, with a simpler formalism based on the Vapnik-Chervonenkis dimension. Consider now a hypothesis set $\mathcal{G} \subset \{-1, +1\}^{\mathcal{X} \times \mathcal{X}}$, a symmetric loss function on the product space $\ell: \mathcal{G} \times \mathcal{Z} \rightarrow \mathbb{R}$, denote by $Z'$ an independent copy of $Z$, and set $\ell_g(Z, Z') = \ell(g, Z, Z')$. Our goal is to find a decision rule $g^*$ that minimizes $\mathcal{R}(g) = \mathbb{E}_{Z, Z'} \{\ell_g(Z, Z')\}$ over $\mathcal{G}$. A classical example covered by this setting is ranking, where one is typically interested in predicting if an object $X$ is preferred over object $X'$. We study the performance of the MoU-minimizer $\hat{g}_{MoU} = \arg\min_{g \in \mathcal{G}} \text{MoU}_{\mathcal{D}_n}(\ell_g)$, where

$$\text{MoU}_{\mathcal{D}_n}(\ell_g) = \text{median} \left( \sum_{i<j \in B_i} \ell_g(Z_i, Z_j), \ldots, \sum_{i<j \in B_k} \ell_g(Z_i, Z_j) \right).$$

The following two assumptions on the hypothesis set and the loss are required to our analysis.

**Assumption 4.** The hypothesis set $\mathcal{G}$ has finite VC dimension $\text{VC}_{\text{dim}}(\mathcal{G})$.

**Assumption 5.** There exists $M > 0$ such that $\ell(g, Z, Z') \leq M$ almost surely.

Assumptions 4 and 5 are standard in statistical learning. One typically has $M = 1$ for the 0-1 loss $\ell: (g, Z, Z') \mapsto 1\{g(X, X') (Y - Y') \leq 0\}$. If $\mathcal{Y}$ is bounded, any convex relaxation of the latter also fits. We again stress that Assumption 5 only applies to the inliers, i.e. to the realizations of $Z$ and $Z'$, not necessarily to the outliers. The next theorem characterizes $\hat{g}_{MoU}$’s generalization capacity.

**Theorem 1.** Suppose that Assumption 1 is fulfilled, that $(\alpha, \Delta)$ satisfy Assumption 2, and let $\tau = n_{Q}/n < 1/2$. Assume furthermore that $\mathcal{G}$ and $\ell$ satisfy Assumptions 4 and 5 respectively. Then, for all $\delta \in [0, \exp(-4\Delta^2(\tau)n_{Q})]$, choosing $K = \lfloor \alpha(\tau)n \rfloor$, we have with probability larger than $1 - \delta$:

$$\mathcal{R}(\hat{g}_{MoU}) \leq \mathcal{R}(g^*) + 8\sqrt{2M} \Gamma(\tau) \sqrt{\frac{\text{VC}_{\text{dim}}(\mathcal{G})(1 + \log(n)) + \log(1/\delta)}{n}}.$$
**Broader Impact**

In the Big Data era, the availability of massive digitized information to train predictive rules is an undeniable opportunity for the widespread deployment of machine-learning solutions. However, the poor control of the data acquisition process one is confronted with in many applications puts practitioners at risk of jeopardizing the validity, the “generalization ability”, of the rules produced by the algorithms implemented. The Median-of-Means methodology has precisely received attention in the machine learning literature because of its capacity to be robust to the presence of atypical values in training datasets. Of theoretical nature essentially, the present paper offers sound guarantees for a reliable use of this methodology when the data used to train a predictor are contaminated by outliers. Hopefully, it may participate to increase globally the trust of machine-learning practitioners in the deployment of such methods. However, the predictive performance attained depending on the use-case considered, the theoretical and experimental results documented in this article cannot be expected to predict it for specific applications.

**References**

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

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

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

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

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

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

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

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

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

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

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

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

Laforgue, P., Clémençon, S., and Bertail, P. (2019). On medians of (Randomized) pairwise means. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*.

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

Lecué, G., Lerasle, M., and Mathieu, T. (2018). Robust classification via mom minimization. *arXiv preprint arXiv:1808.03106*.
Lee, A. J. (1990). *U-statistics: Theory and practice*. Marcel Dekker, Inc., New York.

Lerasle, M., Szabo, Z., Mathieu, T., and Lecué, G. (2019). Monk – outlier-robust mean embedding estimation by median-of-means. In *Proceedings of the 36th International Conference on Machine Learning* (ICML 2019).

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

Lugosi, G. and Mendelson, S. (2019). Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*.

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

Nemirovsky, A. S. and Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons Ltd.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. (2012). On the empirical estimation of integral probability metrics. *Electron. J. Statist.*, 6:1550–1599.

Staerman, G., Laforgue, P., Mozharovskyi, P., and d’Alché-Buc Florence (2020). When OT meets MoM: Robust estimation of wasserstein distance. *arXiv preprint*.

van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge university press.

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

|           | \(\alpha(\tau)\) | \(\beta(\tau)\) | \(\gamma(\tau)\) | \(\Gamma(\tau)\) | \(\Delta(\tau)\) | \(\eta(\tau)\) |
|-----------|------------------|------------------|------------------|------------------|------------------|------------------|
| ARITHMETIC| \(\frac{2\alpha(\tau)}{\alpha(\tau) - 2\tau}\) | \(\frac{2(1 + 2\tau)}{1 - 2\tau}\) | \(\frac{\sqrt{\alpha(\tau)}(\alpha(\tau) - \tau)}{(\alpha(\tau) - 2\tau)^{3/2}}\) | \(\frac{\sqrt{\alpha(\tau)}}{\alpha(\tau) - 2\tau}\) | \(\frac{\alpha(\tau)}{(\alpha(\tau) - 2\tau)^{3/2}}\) | \(\frac{\alpha(\tau)}{\alpha(\tau) - 2\tau}\) |
| GEOMETRIC | \(\frac{\sqrt{\alpha(\tau)}}{\alpha(\tau) - 2\tau}\) | \(\frac{(2 - \sqrt{2}) (1 + \sqrt{2\tau})^{3/2}}{2(1 - 2\tau)^{3/2}}\) | \(\frac{\sqrt{1 + \sqrt{2\tau}}}{\sqrt{1 - 2\tau}}\) | \(\frac{\sqrt{2\tau}}{2\tau}\) | \(\frac{\tau}{2\tau}\) | \(\frac{3 - 2\tau}{4}\) |
| HARMONIC  | \(\frac{4\tau}{1 + 2\tau}\) | \(\frac{3 - 2\tau}{1 - 2\tau}\) | \(\frac{\sqrt{2}}{\sqrt{(1 - 2\tau)^{3/2}}\) | \(\sqrt{1 - 2\tau}\) | \(\sqrt{\frac{3 - 2\tau}{2}\) | \(\frac{3 - 2\tau}{5 - 2\tau}\) |
| POLYNOMIAL| \(\frac{\tau}{2(5 - 2\tau)}\) | \(\frac{(3 - 2\tau)\sqrt{5 - 2\tau}}{(1 - 2\tau)^{3/2}}\) | \(\frac{\sqrt{\tau}}{\sqrt{1 - 2\tau}}\) | \(\sqrt{\frac{5 - 2\tau}{2}\) | \(\frac{3 - 2\tau}{5 - 2\tau}\) | \(\frac{3 - 2\tau}{5 - 2\tau}\) |

Figure 5: Upper bound functions \(\alpha\) and associated functions \(\gamma, \Gamma, \Delta\).

Algorithm 1 MoU Gradient Descent (MoU-GD)

**input:** \(D_n, K, T \in \mathbb{N}^+, (\gamma_t)_{t \leq T} \in \mathbb{R}^T, u_0 \in \mathbb{R}^p\)

for epoch from 1 to \(T\) do

// Randomly partition the data
Choose a random permutation \(\pi\) of \([1, n]\)
Build a partition \(B_1, \ldots, B_k\) of \(\{\pi(1), \ldots, \pi(n)\}\)

// Select block with median risk
for \(k \leq K\) do

\[ \hat{U}_{B_k} = \sum_{i \in B_k} \ell(g_{u_i}, Z_i, Z_i) \]
Set \(B_{med}\) s.t. \(\hat{U}_{med} = \text{median}(\hat{U}_{B_1}, \ldots, \hat{U}_{B_K})\)

// Gradient step
\[ u_{t+1} = u_t - \gamma_t \sum_{i \in B_k} \nabla_{u_t} \ell(g_{u_i}, Z_i, Z_i) \]

return \(u_T\)
B Technical Proofs

In this section are detailed the proofs of the theoretical claims stated in the core article.

B.1 Proof of Proposition 1

Roughly speaking, the median has the same behavior as that of a majority of observations. Similarly, the MoM has the same behavior as that of a majority of blocks. In presence of outliers, the key point consists in focusing on same blocks only, i.e. on blocks that do not contain a single outlier, since no prediction can be made about blocks hit by an outlier, in absence of any structural assumption concerning the contamination. One simple way to ensure the same blocks to be in (almost) majority is to consider twice more blocks than outliers. Indeed, in the worst case scenario each outlier contaminates one block, but the sane ones remain more numerous. Let \( K \) denote the total number of blocks chosen, \( K_O \) the number of blocks containing at least one outlier, and \( K_S \) the number of sane blocks containing no outlier. The crux of our proofs then consists in determining some \( K \) such that \( K_S \geq \eta K \). As discussed before, we thus need to consider at least twice more blocks than outliers. On the other hand, \( K \) is by design upper bounded by \( n \). The global constraint can be written:

\[
2n_O = 2\tau n < K \leq n. \tag{8}
\]

Let \( \alpha : [0, 1/2] \to [0, 1] \) such that: \( \forall \tau \in [0, 1/2], 2\tau < \alpha(\tau) < 1 \). Several choices of acceptable function \( \alpha \) are detailed in Example 1, and illustrated in Figure 5. They include among others:

- the arithmetic mean: \( \alpha(\tau) = \frac{1+2\tau}{2} \).
- the geometric mean: \( \alpha(\tau) = \sqrt{2\tau} \).
- the harmonic mean: \( \alpha(\tau) = \frac{4\tau}{1+2\tau} \).
- the polynomial: \( \alpha(\tau) = \tau(5/2 - \tau) \).

Once the function \( \alpha \) is selected, Equation (8) is satisfied as soon as \( K \) verifies:

\[ \alpha(\tau)n \leq K \leq n. \]

It directly follows that

\[ K_S = K - K_O \geq K - n_O \geq K - \tau n \geq \left( 1 - \frac{\tau}{\alpha(\tau)} \right) K = \frac{\alpha(\tau) - \tau}{\alpha(\tau)} K, \]

and one then may use

\[ \eta = \eta(\tau) = \frac{\alpha(\tau) - \tau}{\alpha(\tau)}. \]

Once \( \eta(\tau) \) is determined, a standard MoM deviation study can be carried out. If at least \( K/2 \) sane blocks have an empirical estimate that is \( t \) close to the expectation, then so is the MoM. Reversing the implication gives:

\[
\mathbb{P}\left\{ |\hat{\theta}_{\text{MoM}} - \theta| > t \right\} \leq \mathbb{P}\left\{ \sum_{\text{blocks w/o outlier}} \mathbb{1}\left\{ |\hat{\theta}_{\text{block}} - \theta| > t \right\} \geq K_S - \frac{K}{2} \right\},
\]

\[
\leq \mathbb{P}\left\{ \sum_{\text{blocks w/o outlier}} \mathbb{1}\left\{ |\hat{\theta}_{\text{block}} - \theta| > t \right\} \geq \frac{2\eta(\tau) - 1}{2\eta(\tau)} K_S \right\}, \tag{9}
\]

with \( \hat{\theta}_{\text{block}} = (1/B) \sum_{i \in \text{block}} Z_i \) the block empirical mean. Now observe that Equation (9) describes the deviation of a binomial random variable, with \( K_S \) trials and parameter \( p_t = \mathbb{P}\{ |\hat{\theta}_{\text{block}} - \theta| > t \} \).

It can thus be upper bounded by

\[
\sum_{k=\left[ \frac{2n(\tau) - 1}{2\eta(\tau)} K_S \right]}^{K_S} \left( K_S \right)_k p_t^k (1 - p_t)^{K_S - k} \leq p_t \frac{2n(\tau) - 1}{2\eta(\tau)} K_S \sum_{k=1}^{K_S} \left( K_S \right)_k, \]

\[
\leq p_t \frac{2n(\tau) - 1}{2\eta(\tau)} K_S 2^{K_S}, \]

\[
\leq p_t \frac{2n(\tau) - 1}{2} K 2^{\eta(\tau) K}. \]
By virtue of Chebyshev’s inequality, it holds that \( p_t \leq \sigma^2/(Bt^2) \), with \( B = [n/K] \) denoting the size of the blocks. The right-hand side can then be rewritten as

\[
\exp \left( \frac{2\eta(\tau) - 1}{2} K \cdot \log \left[ \frac{2^{\eta(\tau)}}{2^{\eta(\tau)-1}} \right] \right).
\]

It can be set to \( \delta \) by choosing \( K = \left[ \frac{2\eta(\tau)}{2^{\eta(\tau)-1}} \log(1/\delta) \right] \), we will see later how this is compatible with the initial constraint \( \alpha(\tau)n \leq K \leq n \), and \( t \) such that \( 2^{\eta(\tau)} \sigma^2/(Bt^2) = 1/\delta \), or again:

\[
t = \sqrt{e\sigma} \sqrt{\frac{2\eta(\tau)}{2^{\eta(\tau)-1} B}}.
\]

\[
\leq \sqrt{e\sigma} \sqrt{\frac{4\eta^2(\tau)}{(2\eta(\tau) - 1)^2 n}},
\]

\[
\leq 4\sqrt{e\sigma} \frac{\eta(\tau)}{(2\eta(\tau) - 1)^2} \sqrt{\frac{1 + \log(1/\delta)}{n}}, \tag{10}
\]

where we have used \( 2^{1/2} \leq 1/x^2 \) for \( x \leq 1/2 \), and \( |x| \geq x/2 \) for \( x \geq 1 \).

The final writing is obtained by setting

\[
\beta(\tau) = \frac{2}{2\eta(\tau) - 1} = \frac{2\alpha(\tau)}{\alpha(\tau) - 2\tau},
\]

and

\[
\gamma(\tau) = \frac{\eta(\tau)}{(2\eta(\tau) - 1)^2} = \sqrt{\frac{\alpha(\tau)(\alpha(\tau) - \tau)}{(\alpha(\tau) - 2\tau)^2}}.
\]

Finally, the first part of the proof is achieved by ensuring that \( K \) satisfies the initial constraint. To do so, one may restrict the interval of acceptable \( \delta \)'s. Indeed, it is enough for \( \delta \) to satisfy:

\[
\alpha(\tau)n \leq \beta(\tau) \log(1/\delta) \leq n,
\]

\[
e^{-\eta(\tau)/\beta(\tau)} \leq \delta \leq e^{-\eta(\tau)/\beta(\tau)}.
\]

The limitation on the range of \( \delta \) is typical of MoM’s concentration proofs. The left limitation is due to the constraint \( K \leq n \), and is not very compelling in practice. The right limitation comes from the constraint \( 2n_0 < K \), or \( \alpha(\tau)n < K \), and is specific to our outlier framework. The purpose of the second part of Proposition 1 is precisely to remove the left limitation, under the assumption that \( Z \) is \( \rho \) sub-Gaussian.

Assume now that \( Z \) is \( \rho \) sub-Gaussian. Chernoff’s bound now gives that \( p_t \leq 2e^{-Bt^2/2\rho^2} \). Plugging this bound into MoM’s deviation yields

\[
\mathbb{P} \left\{ |\hat{\theta}_{\text{MoM}} - \theta| > t \right\} \leq \exp \left( \frac{2\eta(\tau) - 1}{2} K \cdot \log \left[ \frac{2^{\eta(\tau)}}{2^{\eta(\tau)-1}} e^{-Bt^2/2\rho^2} \right] \right),
\]

\[
\leq \exp \left( -\frac{2\eta(\tau) - 1}{16\rho^2} - nt^2 \right),
\]

for all \( t \) such that

\[
t^2 \geq \frac{4\rho^2}{B} \frac{4\eta(\tau) - 1}{2\eta(\tau) - 1} \log 2,
\]

Reverting in \( \delta \) gives that it holds with probability at least \( 1 - \delta \)

\[
|\hat{\theta}_{\text{MoM}} - \theta| \leq \frac{4\rho}{\sqrt{2\eta(\tau) - 1}} \sqrt{\frac{\log(1/\delta)}{n}},
\]

13
for all \( \delta \) that satisfies
\[
\delta \leq e^{-\frac{1}{2\tau^2}(4\eta(\tau)-1)}, \quad \text{and in particular} \quad \delta \leq e^{-4n\alpha(\tau)}.
\] (11)

Indeed it holds \( B = \lfloor n/K \rfloor \geq n/(2K) \), so that \( n/B \leq 2K = 2^\lceil \alpha(\tau)n \rceil \leq 2(\alpha(\tau)n+1) \leq 4\alpha(\tau)n \), since \( 1 \leq 2n_0 = 2r n \leq \alpha(\tau)n \). When \( n_0 = \tau = 0 \), one may choose \( K = 1, B = n \), and \( \delta \leq 1/e \).

The final writing is obtained by setting:
\[
\Gamma(\tau) = \frac{1}{\sqrt{2n(\tau) - 1}} = \sqrt{\frac{\alpha(\tau)}{\alpha(\tau) - 2\tau}}.
\]

To get the expectation bound, one may simply integrate the previously found deviation probabilities. Reverting the inequality gives that it holds
\[
P\left\{ \left| \hat{\theta}_{\MoM} - \theta \right| > t \right\} \leq e^{-\frac{\delta^2}{n(1-\alpha)/2}},
\]
for all \( t \) such that (using Assumption 3):
\[
t \geq 8\rho \Gamma(\tau) \sqrt{\alpha(\tau)}, \quad \text{and in particular} \quad t \geq 8\rho \Gamma(\tau) \sqrt{\frac{\alpha(\tau)}{\tau} \frac{C_0}{n(1-\alpha)/2} + \frac{\Delta(\tau)}{n}}.
\] (12)

One finally gets
\[
E\left[ \left| \hat{\theta}_{\MoM} - \theta \right| \right] = \int_0^{\infty} P\left\{ \left| \hat{\theta}_{\MoM} - \theta \right| > t \right\} dt,
\]
\[
\leq \int_0^{\infty} 8\rho \Gamma(\tau) \sqrt{\frac{\alpha(\tau)}{\tau} \frac{C_0}{n(1-\alpha)/2}} dt + \int_0^{\infty} e^{-\frac{\delta^2}{n(1-\alpha)/2}} dt,
\]
\[
\leq 8\rho \Gamma(\tau) \sqrt{\frac{\alpha(\tau)}{\tau} \frac{C_0}{n(1-\alpha)/2} + \frac{2\sqrt{\pi} \rho \Gamma(\tau)}{\sqrt{n}}},
\]
\[
\leq 2\rho \Gamma(\tau) \left( 4C_0 \frac{\Delta(\tau)}{n(1-\alpha)/2} + \sqrt{\frac{\pi}{n}} \right),
\]
with the notation
\[
\Delta(\tau) = \sqrt{\frac{\alpha(\tau)}{\tau}}.
\]

\[\square\]

**Remark 1.** Coming back to Equation (9), one may also use Hoeffding’s inequality to get:
\[
P\left\{ \left| \hat{\theta}_{\MoM} - \theta \right| > t \right\} \leq \frac{1}{K} \sum_{\text{blocks w/o outlier}} I\left\{ \left| \hat{\theta}_{\text{block}} - \theta \right| > t \right\} - p_t \geq \frac{2\eta(\tau) - 1}{2\eta(\tau)} - \frac{\sigma^2}{Bt^2},
\]
\[
\leq \exp\left( -2\eta(\tau)K \left( \frac{2\eta(\tau) - 1}{2\eta(\tau)} - \frac{\sigma^2}{Bt^2} \right)^2 \right).
\] (13)

The right-hand side can be set to \( \delta \) by choosing
\[
K = \left[ \frac{\eta(\tau)}{2(\eta(\tau) - 1)^2} \log(1/\delta) \right],
\]
and \( t \)'s that satisfy:
\[
\frac{2\eta(\tau) - 1}{6\eta(\tau)} = \frac{\sigma^2}{Bt^2},
\]
\[
t = \sqrt{6\sigma} \sqrt{\frac{\eta(\tau)}{2\eta(\tau) - 1}} \frac{1}{\sqrt{B}},
\]
\[
t \leq \sqrt{6\sigma} \sqrt{\frac{\eta(\tau)}{2\eta(\tau) - 1}} \frac{2K}{n},
\]
\[
t \leq 3\sqrt{6\sigma} \frac{\eta(\tau)}{(2\eta(\tau) - 1)^2} \sqrt{1 + \log(1/\delta)}.
\]
Up to the constant term which is bigger (3\sqrt{6} instead of 4\sqrt{6}), and the number of blocks which is more important, the latter result is very similar to Equation (10). But constant factors were not the only reason motivating our choice of using the Binomial concentration. Indeed, it should be noticed that the Hoeffding bound becomes vacuous when using \( p_t \leq 2 \exp(-Bt^2/2\rho^2) \) for a \( \rho \) sub-Gaussian r.v. \( Z \). Even if this sharper bound for \( p_t \) is plugged in Equation (13), the quantity \( (2\eta(\tau) - 1)/(2\eta(\tau)) - 2 \exp(-Bt^2/(2\rho^2)) \) may never go to 0, making it impossible to improve the confidence range similarly to what has been done in Proposition 1. Notice that the same problem arises in the proof of Proposition 3.

B.2 Proof of Proposition 2

The proof of Proposition 1 can be fully reused, up to two details related to \( U \)-statistics. The first one is Chebyshev’s inequality, used to bound \( p_t \) in the general case. The latter now features the variance of the \( U \)-statistic, that can be upper bounded as follows. Using the notation of van der Vaart (2000) (see Chapter 12 therein), for \( c \leq d \) define \( \zeta_c(h) = \text{Cov}(h(Z_{i_1}, \ldots, Z_{i_d}), h(Z_{i'_1}, \ldots, Z_{i'_d})) \) when \( c \) variables are common. Noticing that \( \zeta_0(h) = 0 \), it holds:

\[
\text{Var} \left( \hat{U}_B(h) \right) = \text{Cov} \left( \frac{1}{(B_d)} \sum_{i_1 \leq \ldots \leq i_d} h(Z_{i_1}, \ldots, Z_{i_d}), \frac{1}{(B_{d'})} \sum_{i'_1 < \ldots < i'_{d'}} h(Z_{i'_1}, \ldots, Z_{i'_{d'}}) \right),
\]

\[
= \frac{1}{(B_d)^2} \sum_{i_1 < \ldots < i_d} \text{Cov} \left( h(Z_{i_1}, \ldots, Z_{i_d}), h(Z_{i'_1}, \ldots, Z_{i'_{d'}}) \right),
\]

\[
= \frac{1}{(B_d)^2} \sum_{c=1}^{d} \binom{d}{c} (B - d) \zeta_c(h),
\]

\[
= \sum_{c=1}^{d} \frac{d!}{c!(d-c)!} \left( B - d \right) \left( B - d - 1 \right) \ldots \left( B - 2d + c + 1 \right) \zeta_c(h),
\]

\[
\leq d! \sum_{c=1}^{d} \frac{\binom{d}{c} \zeta_c(h)}{B},
\]

with \( \Sigma^2(h) = d! \sum_{c=1}^{d} \binom{d}{c} \zeta_c(h) \).

The second critical point that should be adapted is the upper bound \( p_t \leq 2e^{-Bt^2/2\rho^2} \) when \( Z \) is \( \rho \) sub-Gaussian. If kernel \( h \) is bounded, then Hoeffding’s inequality for \( U \)-statistics (Hoeffding, 1963) gives instead that \( p_t \leq 2e^{-Bt^2/2d||h||_2^2} \). The rest of the proof is similar to that of Proposition 1. We stress that Hoeffding’s inequality is used on a sane block, so that we only need \( h \) to be bounded if applied to r.v. \( Z \). In particular, it needs not be bounded on the outliers. This happens e.g. for any continuous kernel \( h \) and r.v. \( Z \) with bounded support.

B.3 Proof of Proposition 3

Let us first recall the notation needed to the analysis of \( \hat{\theta}_{\text{MoU}_2}(H) \). The numbers of blocks are denoted by \( K_X \) and \( K_Y \), and the block sizes by \( B_X = \lfloor n/K_X \rfloor \) and \( B_Y = \lfloor m/K_Y \rfloor \) respectively. The number of sane blocks are denoted by \( K_{X,S} \) and \( K_{Y,S} \), and for \( k \leq K_X \) and \( l \leq K_Y \), we set:

\[
\hat{U}_{k,l}(H) = \frac{1}{B_X B_Y} \sum_{i \in B_k^X} \sum_{j \in B_l^Y} H(X_i, Y_j),
\]

the (two-sample) \( U \)-statistic built upon blocks \( B_k^X \) and \( B_l^Y \). Let \( I_{k,l} = \mathbb{1}\{ |\hat{U}_{k,l}(H) - \theta(H)| > t \} \) be the indicator random variable characterizing its \( t \)-closeness to the true parameter \( \theta(H) \).
As previously discussed, the constraint on $K_X$ and $K_Y$ now writes:

$$\alpha(\tau_X + \tau_Y - \tau_X \tau_Y)n m \leq K_X K_Y \leq nm. \quad (14)$$

In order to simplify the computation, we will however consider the following double constraint:

$$\begin{align*}
\sqrt{\alpha(\tau_X + \tau_Y - \tau_X \tau_Y)n m} & \leq K_X \leq n, \\
\sqrt{\alpha(\tau_X + \tau_Y - \tau_X \tau_Y)m n} & \leq K_Y \leq m.
\end{align*} \quad (15)$$

Equation (15) naturally implies Equation (14), and one may observe that it does not impact the limit condition $\tau_X + \tau_Y - \tau_X \tau_Y < 1/2$. Similarly to previous proofs, Equation (14) yields

$$K_{X,S} K_{Y,S} \geq \left(1 - \frac{\tau_X + \tau_Y - \tau_X \tau_Y}{\alpha(\tau_X + \tau_Y - \tau_X \tau_Y)}\right) K_X K_Y := \eta_{XY} \cdot K_X K_Y,$$

for notation simplicity. On the other hand, Equation (15) ensures both

$$\begin{align*}
K_{X,S} & \geq \left(1 - \frac{\tau_X}{\sqrt{\alpha(\tau_X + \tau_Y - \tau_X \tau_Y)}}\right) K_X := \eta_X \cdot K_X, \\
K_{Y,S} & \geq \left(1 - \frac{\tau_Y}{\sqrt{\alpha(\tau_X + \tau_Y - \tau_X \tau_Y)}}\right) K_Y := \eta_Y \cdot K_Y,
\end{align*}$$

with a slight abuse of notation since $\eta_X$ also depends on $Y$ (and conversely). Notice that it holds true $1/2 \leq \eta_X, \eta_Y \leq 1$. Using the same reasoning as before, one gets:

$$\begin{align*}
P\left\{ \left| \theta_{\text{MoU}}(H) - \theta(H) \right| > t \right\} & \leq 2 \cdot 2 \cdot \frac{\sum_{k=1}^{K_X} \sum_{l=1}^{K_Y} I_{k,l}}{2} \geq \frac{2\eta_{XY} - 1}{2\eta_{XY}} K_{X,S} K_{Y,S} \right\},
\end{align*}$$

However, unlike Equation (9), the above equation does not relate to a binomial random variable, as the $I_{k,l}$ are not independent, see Figure 3. An elegant alternative then consists in leveraging the independence between samples $X$ and $Y$ and using Hoeffding’s inequality. Equation (6) gives $\sigma_{B_X, B_Y}^2(H) \leq \sum^2(H)/(B_X \wedge B_Y)$, with $\sum^2(H) = \sum^2(H) + \sum^2(H) + \sum^2(H)$, so that:

$$\begin{align*}
& \leq P \left\{ \frac{1}{K_{X,S} K_{Y,S}} \sum_{\text{blocks w/o outlier}} \sum_{k=1}^{K_{X,S}} I_{k,l} \geq \frac{2\eta_{XY} - 1}{4\eta_{XY}} \right\} + \\
& \leq P \left\{ \frac{1}{K_{Y,S}} \sum_{l=1}^{K_{Y,S}} J_l - \mathbb{E} [J_l | X] \geq \frac{2\eta_{XY} - 1}{4\eta_{XY}} \right\} + \\
& \leq \exp \left( -2\eta_X K_X \left( \frac{2\eta_{XY} - 1}{4\eta_{XY}} - \frac{\sum^2(H)}{2(B_X \wedge B_Y)t^2} \right)^2 \right) + \\
& \exp \left( -2\eta_Y K_Y \left( \frac{2\eta_{XY} - 1}{4\eta_{XY}} - \frac{\sum^2(H)}{2(B_X \wedge B_Y)t^2} \right)^2 \right),
\end{align*}$$

with the notation $J_l = \frac{1}{K_{X,S}} \sum_{k=1}^{K_{X,S}} I_{k,l}$, and $X = (X_1, \ldots, X_n)$. 

16
Now the right-hand side is set to δ by choosing $K_Z = \left\lceil \frac{18 n_X^2 \log(2/\delta)}{\eta^2(2n_X Y - 1)^2} \right\rceil$ for $Z = X, Y$ respectively, and for $t$ that satisfies:

$$\frac{\Sigma^2(H)}{2(B_X \wedge B_Y)^2} = \frac{2\eta_{XY} - 1}{12\eta_{XY}},$$

$$t = \Sigma(H) \sqrt{\frac{6\eta_{XY}}{2\eta_{XY} - 1}} \sqrt{\frac{\Sigma(H)}{B_X \wedge B_Y}},$$

$$\leq \Sigma(H) \sqrt{\frac{6\eta_{XY}}{2\eta_{XY} - 1}} \sqrt{\frac{2\max(K_X, K_Y)}{n \wedge m}},$$

$$\leq 12\sqrt{3} \Sigma(H) \left( \frac{\eta_{XY}}{2\eta_{XY} - 1} \right) \left( \frac{1 + \log(2/\delta)}{n \wedge m} \right),$$

Constraints (15) are finally fulfilled by choosing $\delta$ such that:

$$\begin{cases} \sqrt{\alpha(H)} + \tau_H \leq \frac{18 n_X^2 \log(2/\delta)}{\eta^2(2n_X Y - 1)^2} \log(2/\delta) \leq n, \\ \sqrt{\alpha(H)} + \tau_H \leq \frac{18 n_X^2 \log(2/\delta)}{\eta^2(2n_X Y - 1)^2} \log(2/\delta) \leq m, \\ 2\max(e^{-n\beta_X}, e^{-m\beta_Y}) \leq \delta \leq 2\min(e^{-n\beta_X}, e^{-m\beta_Y}), \end{cases}$$

with the shortcut notation $\alpha = \alpha(H)$, and $\beta = \frac{18 n_X^2 \log(2/\delta)}{\eta^2(2n_X Y - 1)^2}$ for $Z = X, Y$.

**B.4 Proof of Proposition 4**

Again, the proof can be directly adapted from that of Proposition 1. The first difference lies in the constraint $K$ needs to satisfy. It now writes: $2(n_0 + n_0) = 2(\tau_X + \tau_Y)n < K \leq n$, and the reasoning can then be reused in totality with $\tau_X + \tau_Y$ instead of $\tau$. The second difference is Chebyshev’s inequality, but Equation (6) gives that $\sigma^2(H)_B(\tau_H) \leq \Sigma^2(H)/B$, with $\Sigma^2(H) = \sigma^2(H) + \sigma^2_1(H) + \sigma^2_2(H)$. Finally, when $\|H\|_{\infty}$ is finite, using the notation $X = (X_1, \ldots, X_n)$, one may bound $p_i$ as follows:

$$p_i = P \left\{ |\hat{U}_{1,i}(H) - \theta(H)| > t \right\},$$

$$= P \left\{ \left| \frac{1}{B^2} \sum_{i \in B_i^X} \sum_{j \in B_i^Y} H(X_i, Y_j) - \theta(H) \right| > t \right\},$$

$$\leq P \left\{ \left| \frac{1}{B} \sum_{j \in B_i^Y} \left( \sum_{i \in B_i^X} \frac{H(X_i, Y_j)}{B} - \mathbb{E} \left[ \sum_{i \in B_i^X} \frac{H(X_i, Y_j)}{B} \mid X \right] \right) \right| > \frac{t}{2} \right\},$$

$$+ P \left\{ \left| \frac{1}{B} \sum_{i \in B_i^X} \mathbb{E}_Y \left[ H(X_i, Y) \right] - \theta(H) \right| > \frac{t}{2} \right\},$$

$$\leq 2e^{-Bt^2/8\|H\|_{\infty}^2} + 2e^{-Bt^2/8\|H\|_{\infty}^2},$$

where we have used Hoeffding’s inequality twice: on the $\sum_{i \in B_i^X} \frac{H(X_i, Y)}{B}$ for $j \in B_i^Y$, conditionally to the $X_i$’s, and a second time to the $\mathbb{E}_Y \left[ H(X_i, Y) \right]$ for $i \in B_i^X$, both random variables being bounded by $\|H\|_{\infty}$. The rest of the proof is similar to that of Proposition 1.  

□
B.5 Proof of Theorem 1

Using the fact that \( \hat{g}_{\text{MoU}} \) minimizes \( \text{MoU}_{D_n}(\ell_g) \) over \( G \), one gets:

\[
\mathcal{R}(\hat{g}_{\text{MoU}}) - \mathcal{R}(g^*) \leq \mathcal{R}(\hat{g}_{\text{MoU}}) - \text{MoU}_{D_n}(\ell_{\hat{g}_{\text{MoU}}}) + \text{MoU}_{D_n}(\ell_{g^*}) - \mathcal{R}(g^*),
\]

\[
\leq 2 \sup_{g \in G} |\text{MoU}_{D_n}(\ell_g) - \mathcal{R}(g)|,
\]

\[
\leq 2 \sup_{g \in G} |\text{MoU}_{D_n}(\ell_g) - \mathbb{E}[\ell_g]|.
\]

For a fixed \( g \in G \), Proposition 2 and Assumption 5 gives that for all \( \delta \in [0, \exp(-4n\alpha(\tau))] \), we have with probability larger than \( 1 - \delta \):

\[
|\text{MoU}_{D_n}(\ell_g) - \mathbb{E}[\ell_g]| \leq 4\sqrt{2}M \Gamma(\tau) \sqrt{\frac{\log(1/\delta)}{n}}.
\]

By virtue of Sauer’s lemma, Assumption 4 altogether with the union bound then gives that for all \( \delta \in [0, \exp(-4\Delta^2(\tau)n\alpha)] \), it holds with probability at least \( 1 - \delta \):

\[
\sup_{g \in G} |\text{MoU}_{D_n}(\ell_g) - \mathbb{E}[\ell_g]| \leq 4\sqrt{2}M \Gamma(\tau) \sqrt{\frac{\text{VC}_{\text{dim}}(G) (1 + \log(n)) + \log(1/\delta)}{n}}.
\]

\[\square\]

Remark 2. An adaptation of Theorem 3 in Lecué et al. (2018) shows that the output of Algorithm 1 converges toward \( \hat{g}_{\text{alg}} \), minimizer of \( \mathbb{E}_{\text{part}} [\text{MoU}_{D_n}(\ell_g)] \), where the expectation is taken with respect to all possible ways of partitioning \( D_n \) into \( K \) disjoint blocks. A simple application of Jensen’s inequality then allows to transfer the guarantees of Theorem 1 to \( \hat{g}_{\text{alg}} \). Indeed it holds:

\[
\mathcal{R}(\hat{g}_{\text{alg}}) - \mathcal{R}(g^*) \leq 2 \sup_{g \in G} |\mathbb{E}_{\text{part}} [\text{MoU}_{D_n}(\ell_g)] - \mathcal{R}(g)| \leq 2 \mathbb{E}_{\text{part}} \left[ \sup_{g \in G} |\text{MoU}_{D_n}(\ell_g) - \mathbb{E}[\ell_g]| \right].
\]

C Numerical Experiments

In this section, we present numerical experiments highlighting the remarkable robustness-to-outliers of MoM-based estimators. In particular, we present mean and (multi-sample) \( U \)-statistics estimation experiments under Assumption 3, that emphasize the superiority of MoM/MoU/MoU_{2} compared to standard alternatives (see Appendix C.1). We also provide implementations of Algorithm 1 on both ranking and metric learning problems (Appendix C.2). They illustrate the good behavior of the MoU Gradient Descent (MoU-GD) when the training dataset is contaminated.

C.1 Estimation Experiments

For all our experiments, we set \( n_\alpha = \sqrt{n} \), so that Assumption 3 is fulfilled with \( C_\alpha = 1 \), \( \alpha_\alpha = 1/2 \). We next specify particular instances of Assumption 1, \textit{i.e.} a distribution for \( Z \) (or for \( X \) and \( Y \)), and a distribution for the outliers, such that standard estimators are dramatically damaged, while the MoM-based versions studied in the present article are barely impacted, corroborating the theoretical guarantees established in Propositions 1, 2 and 4. We have selected \( K \) according to the Harmonic upper bound, so that Assumption 2 is fulfilled as well.

\textbf{Ruining the mean.} In this first example, the same data is drawn according to a standard Gaussian distribution (hence \( \theta = 0 \), and the sub-Gaussian assumption is satisfied with \( \rho = 1 \)), and outliers follow a Dirac \( \delta_{n^{1/2}} \). The expected value of the empirical mean estimator \( \hat{\theta}_{\text{avg}} \) is then given by:

\[
\mathbb{E}_{\text{Dirac}}[\hat{\theta}_{\text{avg}}] = (1 - \tau) \cdot 0 + \tau \cdot \sqrt{n} = 1,
\]

always missing the true value. In contrast, MoM’s performance improves with \( n \), showing almost no perturbation due to the outliers, see Figure 7a.

\textbf{Ruining the median.} The Median-of-Means can be seen as an interpolation between the empirical mean (achieved for \( K = 1 \)) and the empirical median (\( K = n \)). If the first one is known to be very sensitive to abnormal observations, the second is however very robust. Yet, there are some cases where the median fails and MoM succeeds. Of course, MoM is a mean estimator while the empirical
median estimates the 1/2 quantile $q_{1/2}$. Hence, we need to consider a case where both coincide to ensure a fair comparison. In our second example, sane data follow a Bernoulli of parameter $\theta = 1/2$, and outliers a Dirac $\delta_1$. When applying blindly the median, one is actually estimating $q_{1/2+\tau} = 1$. The results are reported in Figure 7b. This phenomenon highlights the importance of correctly choosing $\alpha$, a too rough approximation such as the median’s leading to poor results.

**Trimmed mean.** One may argue that a fairer comparison should include the trimmed mean. However, the latter needs a threshold to be defined, which is hard to set on the basis of the proportion of outliers only. In contrast, MoM enjoys a closed form formula, depending exclusively on $\tau$, to select the number of blocks $K$ (see Proposition 1), that allows to nicely adapt to any contaminated scenario.

**Ruinng the variance.** The empirical variance $\hat{\sigma}_n^2 = 1/(n(n-1)) \sum_{i<j}(z_i - z_j)^2$ is a typical example of a (1-sample) $U$-statistic of degree 2, with kernel $h: (Z, Z') \mapsto (Z - Z')^2/2$. Our third setting is as follows: $Z$ follows a uniform law on $[0, 1]$ (so that $\theta = 1/12$, and the supremum of $h(Z, Z')$ is finite equal to $1/2$), while outliers are drawn according to the Dirac $\delta_n/4$. Similarly to the mean, one then has $\mathbb{E}_{D_n}[\hat{\sigma}_n^2]$ of the order of 1, no matter the number of observations considered. In contrast, MoU behaves almost as if the dataset were not contaminated, see Figure 7c.

**Estimating the Mann-Whitney statistic.** A classical 2-sample $U$-statistic of degrees $(1, 1)$ is the Mann-Whitney statistic. Given two random variables $X$ and $Y$, it aims at estimating $\mathbb{P}\{X \leq Y\}$. From two samples of realizations $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ of $X$ and $Y$, it is computed by: $\hat{U}_{n,m}^{MW} = 1/(nm) \sum_{i=1}^n \sum_{j=1}^m 1\{X_i \leq Y_j\}$. This example is very interesting as it highlights the importance of the bounded assumption. Indeed, to get the convergence of MoU, we only need boundedness of $H$ on the inliers. In particular, examples a) and c) above use the unboundedness of the kernel on the outliers to make the empirical mean (respectively variance) arbitrary far away from the true value. Here, since the kernel $H: (X, Y) \mapsto 1\{X \leq Y\}$ is always bounded, the empirical version actually shows more resistance, and the advantage of MoU is less important than in other configurations, see Figure 7d.

---

**Figure 7:** Performances of MoM-based estimators in presence of outliers.
C.2 Learning Experiments

Learning experiments have been run in order to highlight the good generalization capacity of MoU minimizers, theoretically established in Theorem 1 and Remark 2. We considered two pairwise learning problems, metric learning and ranking, on three benchmark datasets (iris, boston housing and wine quality). We first corrupted the datasets, in a way described below, before running Algorithm 1.

Metric Learning. In metric learning, one is interested in learning a distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{+}$, that coincides with some a priori information. We considered the set of Malahanobis distances on $\mathbb{R}^{q}$ $d_{\mathcal{M}}^{2} : (x, x') \mapsto (x - x')^{\top} M (x - x')$, with $M \in \mathbb{R}^{q \times q}$ positive semi-definite, and the iris dataset\(^1\), that gathers 4 attributes (sepal length, sepal width, petal length, and petal width) of 150 flowers issued from 3 different types of irises. The a priori information we want our distance to match is the class, and the (pairwise) criterion we want to optimize writes as follows:

$$\min_{M \in S^{+}_{q}(\mathbb{R})} \frac{2}{n(n-1)} \sum_{i<j} \max \left( 0, 1 + y_{ij} (d_{\mathcal{M}}^{2}(x_{i}, x_{j}) - 2) \right).$$

The whole dataset is first normalized and divided into a train set of size 80% and a test set of size 20%. Then, the training data is contaminated with 10% of outliers drawn uniformly over $[0, 5]^4$, and with label 2, see Figure 8a. Standard and MoU Gradient Descents are run (with a projection step on $S^{+}_{q}(\mathbb{R})$, and $K$ chosen according to the harmonic upper bound), on both the contaminated dataset and the original one of size 80%. The trajectories of the descents averaged over 100 runs are plotted in Figure 8c for the train objective, and in Figure 8d for the test one. MoU-GD remarkably resists to the presence of outliers, and shows test performance comparable to the same GD. In contrast, the contaminated GD converges towards a completely shifted parameter, degrading dramatically its test performance. The erratic convergence of MoU-GDs is due to the fact that the objective monitored is the sum of distances on the median block only, that is shuffled at each iteration. This also explains their lower values. The fact that MoU-GD performs better on the contaminated dataset might not be so surprising. MoM-based approaches discard data. When the latter is not relevant or contaminated, this is an undeniable advantage. When all data are informative, keeping the median block discards the more discriminative points, explaining the slower convergence. Notice furthermore that MoU-GD on the sane dataset has been run with a value of $K$ designed for the contaminated one. Strictly following the Harmonic upper bound one should have chosen instead $K = 1$ (since $\tau = 0$), and would have recovered the standard GD. However, since in practice the proportion of outliers is generally unknown, it appeared reasonable to apply the same $K$. This indeed provides as very interesting tradeoff: it does not affect too much the convergence if the dataset is sane, and prevents from diverging if outliers are present. The code used is in Python, and has the same computational complexity as the standard Gradient Descent. It is attached with the submission for reproducibility purpose.

Ranking. In ranking, the observations available to the practitioner are typically composed of feature vectors $X \in \mathbb{R}^{p}$ describing different objects, and labels $Y \in \mathbb{R}$ representing how much the objects are appreciated by some subject. One is then interested in learning a decision rule $g : \mathbb{R}^{p} \times \mathbb{R}^{p} \to \{-1, 1\}$ to predict if object $X$ is preferred over object $X'$ (i.e. $Y \geq Y'$). We considered the set of decision functions deriving from a scoring function $s : \mathbb{R}^{p} \to [0, 1]$ such that $g(X, X') = 2 \cdot 1 \{ s(X) \geq s(X') \} - 1$. The scoring functions themselves are indexed by vectors $w \in \mathbb{R}^{p}$ such that $s(x) = \sigma(w^{\top} x)$, with $\sigma$ the sigmoid function. ERM then consists in minimizing the disagreements among the training pairs, that writes:

$$\min_{w \in \mathbb{R}^{p}} \frac{2}{n(n-1)} \sum_{i<j} 1 \{ g_{w}(X, X')(Y - Y') \leq 0 \},$$

and can be relaxed into:

$$\min_{w \in \mathbb{R}^{p}} \frac{2}{n(n-1)} \sum_{i<j} \max \left( 0, 1 - g_{w}(X, X')(Y - Y') \right).$$

We have run Algorithm 1 with criterion (16) on two datasets: boston housing\(^2\), that gathers 506 houses described by 13 real features (e.g. number of rooms, distance to employment centers), along with a label corresponding to their prices (real, between 5 and 50), and red wine quality\(^3\), that gathers

\(^1\)https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html
\(^2\)https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html
\(^3\)https://archive.ics.uci.edu/ml/datasets/wine+quality
1,600 wines described by 12 chemical features, along with a label corresponding to a note between 0 and 10. The datasets have first been normalized, and divided into a train set of size 80%, and a test set of size 20%. The outliers have then been generated as follows. A standard GD is first run on the sane training dataset, returning an optimal vector \( \hat{w}_{\text{sane}} \). Then, 2% and 5% of outliers (for \textit{boston} and \textit{wine} respectively) have been generated by sampling \((X_{\text{outlier}}, Y_{\text{outlier}})\) uniformly around \((-\lambda \hat{w}_{\text{sane}}, \lambda)\), for some real value \(\lambda\). This way, one has:

\[
g_{\hat{w}_{\text{sane}}}(X, X_{\text{outlier}})(Y - Y_{\text{outlier}}) \approx \sigma(\hat{w}_{\text{sane}}^\top X) - \sigma(\hat{w}_{\text{sane}}^\top X_{\text{outlier}}) \right) (Y - \lambda),
\]

\[
= \left( \sigma(\hat{w}_{\text{sane}}^\top X) - \sigma(-\lambda \|\hat{w}_{\text{sane}}\|^2) \right) (Y - \lambda).
\]

Making \(\lambda\) tend to \(+\infty\) (respectively \(-\infty\)), the first term becomes always positive and the second very negative (respectively always negative and very positive), incurring important losses preventing from converging toward \(\hat{w}_{\text{sane}}\). For \textit{boston}, \(\lambda\) was set to \(-500\), and to 50 for \textit{wine}. The GD trajectories obtained are very similar to that of the metric learning example, and are thus not reproduced here. The generalization errors obtained on the test dataset of size 20% are gathered in Table 8b. Again, MoU-GD shows a remarkable resistance to the presence of outliers, and attains almost the same performance as standard GD on the sane dataset. This little gap may be partly due to the instability of MoU-GD (see \textit{e.g.} Figure 8c), which uses mini-batches.

|        | GD     | MoU-GD  |
|--------|--------|---------|
|        | sane   | 0.35 ± 0.04 | 0.36 ± 0.05 |
|        | cont.  | 0.99 ± 0.68 | 0.36 ± 0.05 |
|        | wine   | 0.73 ± 0.02 | 0.74 ± 0.02 |
|        | sane   | 0.73 ± 0.02 | 0.74 ± 0.02 |
|        | cont.  | 0.92 ± 0.11 | 0.74 ± 0.02 |

Figure 8: Performances of MoU-Gradient Descent.
D Summary: the different estimators considered in the present article

![Diagram of estimators]

Figure 9: Constructions of the different MoM-based estimators considered in the article.