Generalization Bounds in the Presence of Outliers: a Median-of-Means Study

Pierre Laforgue [1] **Guillaume Staerman** [2] **Stephan Clémençon** [2]

Abstract

In contrast to the empirical mean, the Median-of-Means (MoM) is an estimator of the mean θ 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. Thanks to the high confidence it achieves on heavy-tailed data, MoM has found various applications in machine learning, where it is used to design training procedures that are not sensitive to atypical observations. More recently, a new line of work is now trying to characterize and leverage MoM's ability to deal with corrupted data. In this context, the present work proposes a general study of MoM's concentration properties under the contamination regime, that provides a clear understanding of the impact of the outlier proportion and the number of blocks chosen. The analysis is extended to (multisample) U-statistics, *i.e.* averages over tuples of observations, that raise additional challenges due to the dependence induced. Finally, we show that the latter bounds can be used in a straightforward fashion to derive generalization guarantees for pairwise learning in a contaminated setting, and propose an algorithm to compute provably reliable decision functions.

1. Introduction

There are undoubtedly two major reasons for the success of modern machine learning techniques: on the one hand, the increasing availability of massive datasets, on the other, the existence of computationally efficient and statistically accurate estimation procedures. If the constant improvement of data acquisition technologies, such as the Internet of Things (IoT), enables today to collect considerable datasets in an automatic fashion, it also raises numerous challenges on the estimation side, due to the heterogeneity and possible corruption of the observations acquired. From a statistical perspective, two frameworks have been introduced to model these aspects: (1) the heavy-tailed framework, where only low-order moments are assumed to be finite for the data distribution, (2) the ε-contamination model (Huber, 1964), where the available dataset is supposed to be corrupted by a proportion ε of outliers.

Univariate mean estimation plays a critical role in many statistical learning problems, ranging from classification and regression to ranking or generative modeling. Although the empirical mean appears as a natural candidate, it has been unfortunately shown to dramatically fail under either of the two models discussed above. Consider a sample $\mathcal{S}_n = \{Z_1, \ldots, Z_n\}$ composed of n independent identically distributed (i.i.d.) realizations of the real-valued random variable Z, with distribution P. It is well known that for the empirical mean $\hat{\theta} = (1/n)\sum_{i=1}^n Z_i$ to exhibit a sub-Gaussian tail behavior, it is required that distribution P must also be sub-Gaussian, *i.e.* there exists $\rho > 0$ such that $\mathbb{E}_P[e^{\lambda Z}] \leq e^{\lambda^2\rho^2/2}$ for all $\lambda \in \mathbb{R}$. In contrast, in the heavy-tailed model, one is rather interested by estimates enjoying similar guarantees but under much weaker assumptions, such as having only a finite variance, see the following assumption supposed to be verified throughout this paper.

Assumption 1. *There exist* θ *and* $\sigma^2 < +\infty$ *such that* $\mathbb{E}_P[Z] = \theta$, *and* $\mathrm{Var}_P(Z) = \sigma^2$.

The Median-of-Means (MoM) is one of the mean estimators that achieve a sub-Gaussian behavior under Assumption 1. Independently introduced during the 1980s (Nemirovsky and Yudin, 1983; Jerrum et al., 1986), the Median-of-Means is a mean estimator that is easy to compute, while exhibiting attractive robustness properties. For a predefined level of confidence $1-\delta$, with $\delta \in [e^{1-n/2}, 1)$, the MoM estimator 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 at random, independently from the data, the sample $\mathcal{S}_n$ into K disjoint blocks $\mathcal{B}_1, \ldots, \mathcal{B}_K$ 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 $\mathcal{B}_k$: $\hat{\theta}_k = (1/B)\sum_{i \in \mathcal{B}_k} Z_i$. The Median-of-Means $\hat{\theta}_{\mathrm{MoM}}$ is obtained by computing the median of the block averages (see also Figure 1):

$$\hat{\theta}_{\mathrm{MoM}} = \mathrm{median}(\hat{\theta}_1, \ldots, \hat{\theta}_K). \tag{1}$$

[1]Università degli Studi di Milano, Italy [2]LTCI, Télécom Paris, Institut Polytechnique de Paris, France. Correspondence to: Pierre Laforgue <pierre.laforgue@unimi.it>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

The recent 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. For instance, a proof of its behavior under Assumption 1 can be found in Devroye et al. (2016).

Proposition 1. *(Devroye et al., 2016) Suppose that an i.i.d. sample $\mathcal{S}_n$ is drawn from P, satisfying Assumption 1. Then, for any $\delta \in [e^{1-n/2}, 1)$, choosing $K = \lceil \log(1/\delta) \rceil$, it holds with probability at least $1-\delta$:*

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

These concentration results have further been extended to random vectors, through different generalizations of the median in a multidimensional setting (Minsker et al., 2015; Hsu and Sabato, 2016; Lugosi and Mendelson, 2019c), and to U-statistics (Joly and Lugosi (2016) for the degenerate case, Laforgue et al. (2019) with randomized blocks) among other extensions. Such interesting properties in the presence of heavy-tailed data has given birth to numerous applications in statistical learning. This includes *e.g.* an adaptation of the Upper Confidence Bound (UCB) bandit algorithm in Bubeck et al. (2013), of Empirical Risk Minimization (ERM) in Brownlees et al. (2015), or the more general framework of MoM-tournaments (Lugosi and Mendelson, 2019a) and Le Cam's approach (Lecué and Lerasle, 2019).

A recent line of work is now trying to change perspective, abandoning the heavy-tailed framework to focus on MoM's behavior within the Huber's contamination model. Formally, the assumption considered in this paper is as follows.

Assumption 2. *The sample $\mathcal{S}_n = \{Z_1, \ldots, Z_n\}$ contains $n - n_{\mathsf{O}}$ inliers drawn i.i.d. according to distribution P, and n_{O} outliers, upon which no assumption is made. We denote by $\varepsilon = n_{\mathsf{O}}/n$ the fraction of outliers among sample $\mathcal{S}_n$.*

Remark 1. *We stress that Assumption 2 does not assume independence between inliers and outliers, nor between outliers. This setting is thus more general than the standard Huber's contamination model, which assumes that $\mathcal{S}_n$ is drawn i.i.d. from the mixture $\widetilde{P} = (1-\zeta)P + \zeta A$, where $\zeta \in (0,1)$ and A is an arbitrary distribution. However, in Assumption 2 the number of outliers in fixed to n_{O}, and not a random number as in the Huber contamination's model.*

Assumption 2 has been addressed through the general angle of MoM-minimization in Lecué et al. (2018), while Lerasle et al. (2019) develops an application to Maximum Mean Discrepancy and outlier-robust mean embedding. Depersin and Lecué (2019) proposes a sub-Gaussian MoM-inspired multidimensional estimator computable in almost linear time, and Depersin (2020) studies a multivariate estimator based on one-dimensional projections. However, all these works rely on *ad-hoc* assumptions that are quite difficult to interpret. For instance, Lecué et al. (2018) uses unusual outlier-adapted Rademacher complexities, while the choice of K is based on unknown constants in Depersin (2020), or defined implicitly in Lerasle et al. (2019). In Depersin and Lecué (2019), the choice of K incidentally reduces the analysis to the case where $\varepsilon \leq 1/300$.

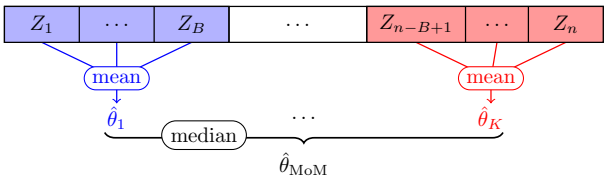


Figure 1: The MoM estimator.

In contrast, this paper proposes a unified and insightful study of the concentration properties of (univariate) MoM-based estimators under the contamination regime of Assumption 2. In particular, we show that MoM is able to handle up to 50% of outliers, at the price of a degraded constant though. Indeed, our bounds allow to encapsulate the impact of the proportion of outliers ε into constant terms only. As this performance can be achieved through a multitude of values for the number of blocks K, we also fully characterize the impact of this choice, exemplified by 4 representative strategies. Another important insight given by our analysis is that MoM may handle both outliers and heavy-tailed inliers, but on limited range of confidence levels only. Assuming instead the inliers to be sub-Gaussian, we show that MoM becomes efficient on a wide interval, allowing next to derive bounds in expectation (we are not aware of similar results for MoM) under the following assumption stipulating that the number of outliers n_{O} grows sub-linearly with n.

Assumption 3. *There exist constants $C_{\mathsf{O}} \geq 1$ and $\alpha_{\mathsf{O}} \in [0,1)$ such that: $\forall n \geq 1$, $n_{\mathsf{O}} \leq C_{\mathsf{O}}^2\, n^{\alpha_{\mathsf{O}}}$.*

The extension to multisample U-statistics raises interesting discussions about the fractions of outliers authorized by the different approaches. We then show that our bounds can be easily combined with standard class complexities (VC-dimension, entropy) to produce generalization bounds for pairwise learning in the presence of outliers. We finally detail an algorithm whose outputs satisfy these guarantees.

The rest of the article is organized as follows. In Section 2 are stated the concentration results for the MoM estimator and its extensions to (multisample) U-statistics under the regime of Assumption 2. The applications to learning theory are detailed in Section 3. Technical proofs, as well as some numerical results validating our theoretical findings, are deferred to the Supplementary Material.

Related Works. Of course, the Median-of-Means is not the sole estimator to achieve sub-Gaussian behavior under the contaminated model. One may for instance mention the trimmed mean (Oliveira and Valdora, 2019; Lugosi and Mendelson, 2019b). The existing bounds however exhibit a complex dependence with respect to ε, in contrast to our results. One of the important drawback of MoM lies in its computational intractability in high dimension, motivating an important line of research in the field of robust mean estimation (Diakonikolas et al., 2016; Lai et al., 2016; Cheng et al., 2019; Hopkins, 2018; Cherapanamjeri et al., 2019; Prasad et al., 2019; 2020). Our analysis essentially differs from these works in three ways: (i) as we ultimately target to derive learning bounds, *i.e.* bounds on risk estimates, we shall focus on univariate estimators, bypassing also the computational difficulties with multidimensional MoMs, (ii) it allows for a complete characterization of the impact of ε and K, and (iii) the extension to U-statistics is entirely new to the best of our knowledge.

2. Concentration of MoM-based Estimators in the Presence of Outliers

In this section, we study the concentration properties of MoM, and those of its recent extensions to U-statistics, under the contamination regime of Assumption 2.

2.1. Concentration Bounds for MoM

In this section, we prove an extension of bound (2) when the sample $\mathcal{S}_n$ is corrupted according to Assumption 2. As revealed by Proposition 1, when $\mathcal{S}_n$ is not corrupted, K must be set depending on the targeted confidence δ. When outliers are added, K must also be chosen according to the outlier ratio ε. Roughly, we want $K > 2n_{\mathsf{O}}$ to ensure that blocks without outliers are in majority. However, if K is too large MoM tends to the median, which is a bad estimator of the mean in general. To correctly calibrate K, we introduce a mapping $\alpha\colon [0,1/2] \to [0,1]$ upper bounding $\varepsilon \mapsto 2\varepsilon$. This way, setting $K \approx \alpha(\varepsilon)n > 2\varepsilon n = 2n_{\mathsf{O}}$ satisfies the outlier constraint, while refraining from choosing too large values if the bound α is tight enough. Based on α, we derive functions $\beta, \gamma, \Gamma, \Delta$, that appear through the computations and shape the bounds established in Proposition 2.

Assumption 4. *The mapping $\alpha\colon [0,1/2] \to [0,1]$ satisfies*

$$\forall \varepsilon \in (0,1/2), \quad 2\varepsilon < \alpha(\varepsilon) < 1.$$

From mapping α, we define the following functions:

$$\beta\colon \varepsilon \mapsto \frac{2\alpha(\varepsilon)}{\alpha(\varepsilon)-2\varepsilon}, \qquad \gamma\colon \varepsilon \mapsto \frac{\sqrt{\alpha(\varepsilon)}(\alpha(\varepsilon)-\varepsilon)}{(\alpha(\varepsilon)-2\varepsilon)^{\frac{3}{2}}},$$
$$\Gamma\colon \varepsilon \mapsto \sqrt{\frac{\alpha(\varepsilon)}{\alpha(\varepsilon)-2\varepsilon}}, \qquad \Delta\colon \varepsilon \mapsto \sqrt{\frac{\alpha(\varepsilon)}{\varepsilon}}.$$

We now give several examples of mappings α satisfying Assumption 4. Their plots can be found in Figure 2a. The reader is referred to Appendix B (Table 1, Figure 9) for details about the corresponding functions $\beta, \gamma, \Gamma, \Delta$.

Example 1. *As we want $2\varepsilon < \alpha(\varepsilon) < 1$, natural choices for α involve the means of 2ε and 1, taken either arithmetic, geometric or harmonic. The last example is a polynomial.*

	Arithmetic	Geometric	Harmonic	Polynomial
$\alpha(\varepsilon)$	$\frac{1+2\varepsilon}{2}$	$\sqrt{2\varepsilon}$	$\frac{4\varepsilon}{1+2\varepsilon}$	$\varepsilon\left(\frac{5}{2}-\varepsilon\right)$

The next proposition describes the concentration of MoM under the contamination regime of Assumption 2.

Proposition 2. *Suppose that sample $\mathcal{S}_n$ and mapping α satisfy Assumptions 2 and 4 respectively. Define functions $\beta, \gamma, \Gamma, \Delta$ according to Assumption 4. Then, for any $\delta \in [e^{-n/\beta(\varepsilon)}, e^{-n\alpha(\varepsilon)/\beta(\varepsilon)}]$, choosing $K = \lceil \beta(\varepsilon)\log(1/\delta) \rceil$, it holds with probability at least $1-\delta$:*

$$\left|\hat{\theta}_{\mathrm{MoM}} - \theta\right| \;\leq\; 4\sqrt{e}\sigma\,\gamma(\varepsilon)\sqrt{\frac{1+\log(1/\delta)}{n}}. \tag{3}$$

If in addition distribution P is ρ sub-Gaussian, then for all $\delta \in (0, e^{-4n\alpha(\varepsilon)}]$, with $K = \lceil \alpha(\varepsilon)n \rceil$, it holds w.p.a.l. $1-\delta$:

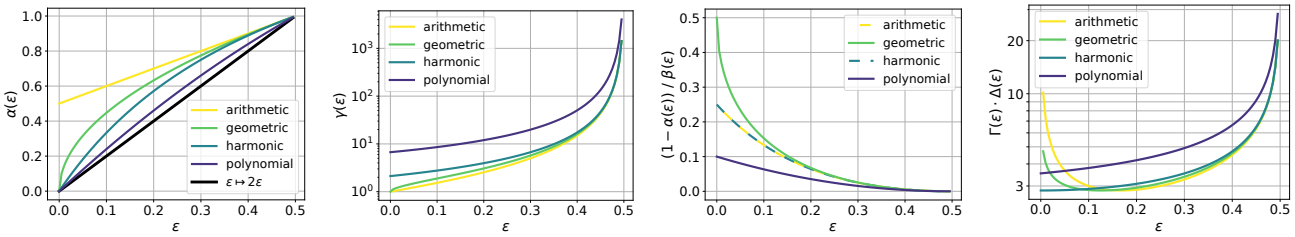
$$\left|\hat{\theta}_{\mathrm{MoM}} - \theta\right| \;\leq\; 4\rho\,\Gamma(\varepsilon)\sqrt{\frac{\log(1/\delta)}{n}}. \tag{4}$$

If furthermore n_{O} satisfies Assumption 3, the same K gives:

$$\mathbb{E}\left[\left|\hat{\theta}_{\mathrm{MoM}} - \theta\right|\right] \;\leq\; 2\rho\,\Gamma(\varepsilon)\left(4C_{\mathsf{O}}\,\frac{\Delta(\varepsilon)}{n^{(1-\alpha_{\mathsf{O}})/2}} + \sqrt{\frac{\pi}{n}}\right).$$

Note that for all mappings of Example 1, we have $\gamma(\varepsilon) \leq 3\sqrt{5}/(1-2\varepsilon)^{3/2}$ and $\Gamma(\varepsilon) \leq \sqrt{5}/\sqrt{1-2\varepsilon}$, see Table 1.

The technical proof is given in Appendix C.1. Its argument essentially consists in using that the MoM estimator (1) has a similar behavior to that of a majority of block means. The condition $K > 2n_{\mathsf{O}}$ is strengthened into $K \geq \alpha(\varepsilon)n$, where the function α is a strict upper bound of the mapping $\varepsilon \mapsto 2\varepsilon$ on $(0,1/2)$, ensuring that a fraction $\eta(\varepsilon) = (\alpha(\varepsilon)-\varepsilon)/\alpha(\varepsilon) > 1/2$ of "sane" blocks (*i.e.* including none of the n_{O} outliers) actually constitutes a majority of blocks. One may then focus on the sane blocks deviations only, which is controlled by means of the concentration properties of a Binomial random variable. 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 δ). The expectation bound is finally obtained by integrating the tail probability bound derived in Equation (4).



(a) Upper bounds $\alpha(\varepsilon)$ (b) Constants $\gamma(\varepsilon)$ (c) Range sizes $s(\varepsilon)$ in log scale (d) Constants $\Gamma(\varepsilon) \cdot \Delta(\varepsilon)$

Figure 2: Influence of the chosen mapping α on the constants.

As revealed by Proposition 2, the choice of α shapes the constant terms 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 subtle balance calls for in depth discussions to determine the optimal mapping α.

A δ-limited sub-Gaussian tail bound. We first point out that the main price to pay for extending the sub-Gaussian tail behavior of MoM to the contaminated framework of Assumption 2 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 δ 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 comes from the constraint $2n_{\mathsf{O}} < K$ (or $\alpha(\varepsilon)n \leq K$), and is specific to the contaminated framework. It should be noticed that this restriction vanishes (*i.e.* the upper limit value is 1) when $\varepsilon = 0$ for all mappings α given in Example 1, except for the arithmetic mean. Observe also that the lower limit restriction is removed when assuming that P is sub-Gaussian. We incidentally underline that this assumption only applies to the inlier distribution P, so that any hope of using reliably the empirical mean remains vain.

About the constants. An interesting property of the bounds derived in Proposition 2 is that they fully encapsulate the impact of the proportion of outliers ε into the constants $\gamma(\varepsilon)$ and $\Gamma(\varepsilon)$. Naturally, the latter increase with ε, and tend to infinity as ε goes to $1/2$, see Figure 2b. This dependence w.r.t. ε can be further explicited, as one may note that for all mappings presented in Example 1 and $\varepsilon \in [0, 1/2]$, we have $\gamma(\varepsilon) \leq 3\sqrt{5}/(1-2\varepsilon)^{3/2}$ and $\Gamma(\varepsilon) \leq \sqrt{5}/\sqrt{1-2\varepsilon}$, see Table 1. Another way to gain intuition about the bounds is to consider the limit case $\varepsilon = 0$. In the sub-Gaussian scenario, all mappings except the arithmetic mean then suggest to choose $K = n$, *i.e.* to compute the standard mean, which is optimal. In the heavy-tail scenario, the harmonic mapping suggests to set $K = \lceil 4 \log 1/\delta \rceil$, which is known to be almost optimal. We highlight that optimality of the constants must be understood here with respect to the use of the MoM technique. In the uncontaminated setting, other approaches may exhibit sharper constants, see *e.g.* Lee and Valiant (2020). Finally, we highlight that the multiplicative nature of the constants $\gamma(\varepsilon)$ and $\Gamma(\varepsilon)$ makes Proposition 2 somehow incomparable to the results discussed in Diakonikolas et al. (2020), which feature an additive term in ε.

Accuracy vs range of confidence levels. As previously mentioned, the choice of mapping α determines at the same time the range $[\exp(-n/\beta(\varepsilon)), \exp(-n\alpha(\varepsilon)/\beta(\varepsilon))]$ for which Equation (3) holds true with probability at least $1 - \delta$, and the constant $\gamma(\varepsilon)$. When $\varepsilon \in [0, 1/2)$ is fixed, the quantity $\gamma(\varepsilon)$ monotonically decreases as $\alpha(\varepsilon)$ increases. Indeed, one may easily check that it holds $(\partial \gamma_\varepsilon^2 / \partial \alpha)(\alpha) = -4\varepsilon(\alpha - \varepsilon)^2/(\alpha - 2\varepsilon)^4 < 0$, with the notation $\gamma_\varepsilon^2(\alpha) = \alpha(\varepsilon)(\alpha(\varepsilon) - \varepsilon)^2/(\alpha(\varepsilon) - 2\varepsilon)^3$. Hence, the larger $\alpha(\varepsilon)$, the smaller the constant in the upper bound, encouraging the practitioner to choose the arithmetic upper bound, see Figure 2b. However, the choice of α also impacts the confidence range, mitigating this incentive. Precisely, when $\varepsilon \in (0, 1/2)$ is fixed, its size $s(\varepsilon)$ increases with $\alpha(\varepsilon)$ on $(2\varepsilon, \sqrt{2\varepsilon}]$, and decreases on $[\sqrt{2\varepsilon}, 1]$. Indeed, at the log scale, it is equal to $s_\varepsilon(\alpha) = n(\alpha - 2\varepsilon)(1-\alpha)/(2/\alpha)$, and $(\partial s_\varepsilon / \partial \alpha)(\alpha) = n(2\varepsilon - \alpha^2)/(2/\alpha^2)$ for $\alpha \in (0,\ 1/2)$. As a consequence, starting from $\alpha(\varepsilon) = \sqrt{2\varepsilon}$ (*i.e.* the geometric mean), increasing $\alpha(\varepsilon)$ indeed reduces $\gamma(\varepsilon)$, but at the price of a smaller range of the confidence levels, see Figure 2c. A similar phenomenon occurs 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 $\Gamma(\varepsilon)$, both decreasing with $\alpha(\varepsilon)$. After integration, this tradeoff can be seen in the opposition between constants $\Gamma(\varepsilon)$ and $\Delta(\varepsilon)$, which have inverse monotonicity w.r.t. $\alpha(\varepsilon)$, see Figure 2d for plots of their product. The fact that $\Delta(\varepsilon) \to \infty$ when $\varepsilon \to 0$ for some choices of α may reflect an artifact of the proof technique. Indeed, if $\varepsilon = n_{\mathsf{O}} = 0$, it is not allowed to multiply/divide by ε in Equation (12). In contrast, one may use $\delta \leq 1/e$ instead of Equation (11), which then gives a $1/\sqrt{n}$ term, with no dependence with respect to Δ.

Rate bound. We underline that the rate $1/\sqrt{n^{1-\alpha_{\mathsf{O}}}}$ 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 $1/\sqrt{B} \sim \sqrt{K/n} \sim \sqrt{n_{\mathsf{O}}/n}$, as K is roughly of the order of n_{O}. Hence, if n_{O} 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 α. We finally highlight that this rate is optimal. Indeed, our bounds are obtained after conditioning upon the observations and, as can be seen by examining proofs, they cannot be refined, insofar as they simply rely on exact computations of the binomial distribution.

Unknown ε. In practice, the proportion of outliers ε is generally unknown, preventing from using it to calibrate K. We emphasize that the above stated bounds may still be used with an overestimation of ε, at the price of a deterioration of $\gamma(\varepsilon), \Gamma(\varepsilon)$ and $s(\varepsilon)$ though.

Related work. Although they are quite similar in spirit, six critical points distinguish Proposition 2 from Theorem 1 in Lerasle et al. (2019). (1) It is important to notice first that Proposition 2 focuses on the deviations of scalar MoMs, while Theorem 1 in Lerasle et al. (2019) addresses that of particular kernel mean embeddings, defined as MoM minimizers. (2) This being said, our choice of K can be computed explicitly from the total proportion of outliers ε, and the targeted confidence δ. 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 by reparameterizing using $\eta(\varepsilon)$. (3) As a consequence, our bound features the true and fixed proportion of outliers ε within the sample, while Lerasle et al. (2019) use the proportion w.r.t. the number of blocks, that may change with it. (4) Additionally, their range of admissible confidence levels $1-\delta$ is defined implicitly, whereas we provide an explicit interval, that depends only on ε and n. (5) Lerasle et al. (2019) require $2n_{\mathsf{O}} \leq K \leq n/2$, meaning they allow at most 25% of outliers, while we can handle up to 50%. (6) They only prescribe a rough estimate of K, that might not be an integer.

2.2. Concentration Bounds for MoU

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 (Clémençon et al., 2008), clustering, see *e.g.* Clémençon (2014), or metric-learning (Vogel et al., 2018) among others. We recall that the U-statistic of degree $d \in \{1, \ldots, n\}$ with kernel $h : \mathbb{R}^d \to \mathbb{R}$, symmetric (*i.e.* invariant under permutation of its arguments), square integrable w.r.t. $P^{\otimes d}$, denoting by P the distribution of the random variable Z, and based on independent copies $Z_1, \ldots, Z_n$ of Z is given by:

$$\bar{U}_n(h) = \frac{1}{\binom{n}{d}} \sum_{1 \leq i_1 < \ldots < i_d \leq n} h(Z_{i_1}, \ldots, Z_{i_d}). \quad (5)$$

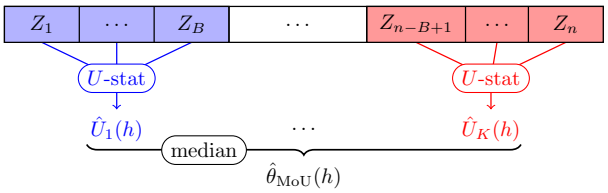


Figure 3: The MoU estimator.

As may be shown by a Lehmann-Scheffé argument, it is the unbiased estimator of the parameter $\theta(h) = \int h(z_1, \ldots, z_d) P(dz_1) \ldots P(dz_d)$ with minimal variance, given by (see *e.g.* van der Vaart (2000)):

$$\frac{1}{\binom{n}{d}} \sum_{c=1}^{d} \binom{d}{c} \binom{n-d}{d-c} \zeta_c(h) \leq \frac{d!}{n} \sum_{c=1}^{d} \binom{d}{c} \zeta_c(h),$$

where, for $1 \leq c \leq d$, we have set $\zeta_c(h) = \mathrm{Var}(h_c(Z_1, \ldots, Z_c))$, with $h_c(z_1, \ldots, z_c) = \mathbb{E}[h(z_1, \ldots, z_c, Z_{c+1}, \ldots, Z_d)]$ for all $(z_1, \ldots, z_c) \in \mathbb{R}^c$. As a single outlier affects $\binom{n-1}{d-1}$ 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 $\mathcal{B}_1, \ldots, \mathcal{B}_K$ 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 on randomized, possibly overlapping, blocks). The MoU estimator of $\theta(h)$ is defined as $\hat{\theta}_{\mathrm{MoU}}(h) = \mathrm{median}\big(\hat{U}_k(h),\ k \leq K\big)$, with

$$\hat{U}_k(h) = \frac{1}{\binom{B}{d}} \sum_{i_1 < \ldots < i_d \in \mathcal{B}_k} h(Z_{i_1}, \ldots, Z_{i_d}) \quad \text{for } k \leq K.$$

See Figure 3 for a depiction. The next proposition details the concentration guarantees of $\hat{\theta}_{\mathrm{MoU}}(h)$ when the sample $\mathcal{S}_n$ it is based upon is contaminated according to Assumption 2. The technical proof is detailed in Appendix C.2.

Proposition 3. *Suppose that sample $\mathcal{S}_n$ and mapping α satisfy Assumptions 2 and 4 respectively. Define functions $\beta, \gamma, \Gamma, \Delta$ according to Assumption 4, and set $\Sigma^2(h)$ as follows: $\Sigma^2(h) = d! \sum_{c=1}^{d} \binom{d}{c} \zeta_c(h)$. Then, for all $\delta \in [e^{-n/\beta(\varepsilon)}, e^{-n\alpha(\varepsilon)/\beta(\varepsilon)}]$, choosing $K = \lceil \beta(\varepsilon) \log(1/\delta) \rceil$, it holds with probability larger than $1-\delta$:*

$$\left|\hat{\theta}_{\mathrm{MoU}}(h) - \theta(h)\right| \leq 4\sqrt{e}\, \Sigma(h)\, \gamma(\varepsilon) \sqrt{\frac{1 + \log(1/\delta)}{n}}.$$

If in addition the essential supremum $\|h(Z_1, \ldots, Z_d)\|_\infty = \inf\{t \geq 0 : \mathbb{P}\{|h(Z_1, \ldots, Z_d)| > t\} = 0\}$ of the r.v. $|h(Z_1, \ldots, Z_d)|$ is finite and bounded by M, then for all $\delta \in (0, e^{-4n\alpha(\varepsilon)}]$, choosing $K = \lceil \alpha(\varepsilon) n \rceil$, it holds with probability at least $1-\delta$:

$$\left|\hat{\theta}_{\mathrm{MoU}}(h) - \theta(h)\right| \leq 4\sqrt{d}\, M\, \Gamma(\varepsilon) \sqrt{\frac{\log(1/\delta)}{n}}.$$

If furthermore n_{O} satisfies Assumption 3, the same K gives:

$$\mathbb{E}\Big[\big|\hat{\theta}_{\mathrm{MoU}}(h)-\theta(h)\big|\Big] \leq 2\sqrt{d}\, M\, \Gamma(\varepsilon)\left(4C_{\mathsf{O}}\,\frac{\Delta(\varepsilon)}{n^{(1-\alpha_{\mathsf{O}})/2}}+\sqrt{\frac{\pi}{n}}\right).$$

2.3. Concentration Bounds for Multisample MoU

The notion of U-statistic can be readily extended to the multisample 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 direct and detailed in Appendix C.5. The U-statistic of degrees $(1,1)$ with kernel $H:\mathbb{R}^2\to\mathbb{R}$, square integrable w.r.t. $P\otimes Q$, denoting by P and Q the distributions of r.v. X and Y respectively, and based on two independent samples $\mathcal{S}_n^X=\{X_1,\dots,X_n\}$, and $\mathcal{S}_m^Y=\{Y_1,\dots,Y_m\}$, composed respectively of $n\geq 1$ and $m\geq 1$ independent copies of X and Y, is given by:

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

It is the unbiased estimator of $\theta(H)=\int\int H(x,y)P(dx)Q(dy)$ with minimal variance, given by:

$$\begin{aligned}\sigma^2_{n,m}(H)&=\frac{1}{nm}\sigma^2(H)+\frac{m-1}{nm}\sigma_1^2(H)+\frac{n-1}{nm}\sigma_2^2(H),\\&\leq\frac{\sigma^2(H)+\sigma_1^2(H)+\sigma_2^2(H)}{n\wedge m},\end{aligned}\tag{6}$$

where $\sigma^2(H)=\mathrm{Var}(H(X,Y))$, $\sigma_1^2(H)=\mathrm{Var}(H_1(X))$ and $\sigma_2^2(H)=\mathrm{Var}(H_2(Y))$, with $H_1(x)=\mathbb{E}\left[H(x,Y)\right]$ and $H_2(y)=\mathbb{E}\left[H(X,y)\right]$. 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 is then given by $\hat{\theta}_{\mathrm{MoU}_2}(H)=\mathrm{median}\big(\hat{U}_{k,l}(H),\ k,l\leq K_X,K_Y\big)$, with

$$\hat{U}_{k,l}(H)=\sum_{i,j\in\mathcal{B}_k^X\times\mathcal{B}_l^Y}\frac{H(X_i,Y_j)}{B_XB_Y},\quad\text{for } k,l\leq K_X,K_Y.$$

Refer to Figure 4 for a visual interpretation in the particular case $K_X=K_Y=3$. For MoU_2, the total number of blocks created is thus K_XK_Y, while the number of corrupted ones is always lower than $n_{\mathsf{O}}K_Y+m_{\mathsf{O}}K_X-n_{\mathsf{O}}m_{\mathsf{O}}$. As we still want at least twice more blocks than possibly corrupted ones, the constraint on K_X and K_Y can be expressed as:

$$2(\varepsilon_X+\varepsilon_Y-\varepsilon_X\varepsilon_Y)nm<K_XK_Y\leq nm.$$

The proportions of outliers ε_X and ε_Y for which we are able to derive statistical guarantees should therefore satisfy $\varepsilon_X+\varepsilon_Y-\varepsilon_X\varepsilon_Y<1/2$. This is a stronger requirement than for MoM, see Figure 6. The next proposition then details the concentration properties of MoU_2 under this assumption.

Proposition 4. *Suppose that both samples $\mathcal{S}_n^X$ and $\mathcal{S}_m^Y$ and mapping α satisfy Assumptions 2 and 4 respectively. Define functions $\beta,\gamma,\Gamma,\Delta$ according to Assumption 4. Let ε_X and ε_Y be such that $\tilde{\varepsilon}:=\varepsilon_X+\varepsilon_Y-\varepsilon_X\varepsilon_Y$ is strictly smaller than $1/2$. Then, for all $\delta\in[2\max(e^{-n\beta_X},e^{-m\beta_Y}),\ 2\min(e^{-n\sqrt{\alpha}(\tilde{\varepsilon})/\beta_X},e^{-m\sqrt{\alpha}(\tilde{\varepsilon})/\beta_Y})]$, choosing $K_X=\lceil\beta_X\log(2/\delta)\rceil$, and $K_Y=\lceil\beta_Y\log(2/\delta)\rceil$, it holds with probability at least $1-\delta$:*

$$\big|\hat{\theta}_{\mathrm{MoU}_2}(H)-\theta(H)\big|\ \leq\ 12\sqrt{3}\,\Sigma(H)\gamma(\tilde{\varepsilon})\sqrt{\frac{1+\log(2/\delta)}{n\wedge m}},$$

with the notation $\Sigma^2(H)=\sigma^2(H)+\sigma_1^2(H)+\sigma_2^2(H)$, $\beta_Z=\frac{18\,\eta^2(\tilde{\varepsilon})}{\eta_Z(2\eta(\tilde{\varepsilon})-1)^2}$, and $\eta_Z=1-\frac{\varepsilon_Z}{\sqrt{\alpha}(\tilde{\varepsilon})}$, for $Z=X,\ Y$.

The technical proof is detailed in Appendix C.3, and is made significantly more involved due to the introduction of dependent random variables, see the $\hat{U}_{k,l}(H)$ in Figure 4. 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 the discussion in Remark 2. As a result, Proposition 4 must be restricted to guarantees on the restricted range of confidence levels. Notice that randomized extensions considered in Laforgue et al. (2019) rely on Hoeffding's inequality as well, and consequently suffer from the same restriction. To overcome this limitation, an alternative consists in removing the dependence between the U-statistics, at the cost of a loss of information though.

Indeed, getting independent U-statistics might be easily achieved, by considering only the diagonal blocks as in Figure 5. This procedure however results in an important loss of information, since a large portion of the grid remains unexplored. Another drawback of this approach is that it forces to set $K_X=K_Y=K$. Overall, this estimator, denoted $\hat{\theta}_{\mathrm{MoU}_2^{\mathrm{diag}}}(H)$ is given by

$$\hat{\theta}_{\mathrm{MoU}_2^{\mathrm{diag}}}(H)=\mathrm{median}\left(\hat{U}_{k,k}(H),\ \ k\leq K\right).\tag{7}$$

The constraint on K then becomes: $2(n_{\mathsf{O}}+m_{\mathsf{O}})<K\leq\min(n,m)$. Obviously, as soon as $m\leq 2n_{\mathsf{O}}$ this cannot be satisfied. To avoid such problems, we shall assume that $n=m$, see the discussion at the end of the section. We now analyze the concentration properties of estimator (7).

Proposition 5. *Suppose that samples $\mathcal{S}_n^X$ and $\mathcal{S}_m^Y$ and mapping α satisfy Assumptions 2 and 4 respectively. Define functions $\beta,\gamma,\Gamma,\Delta$ according to Assumption 4, and assume that $\varepsilon_X+\varepsilon_Y<1/2$. Then, for all $\delta\in[e^{-n/\beta(\varepsilon_X+\varepsilon_Y)},e^{-n\alpha(\varepsilon_X+\varepsilon_Y)/\beta(\varepsilon_X+\varepsilon_Y)}]$, with $K=$*

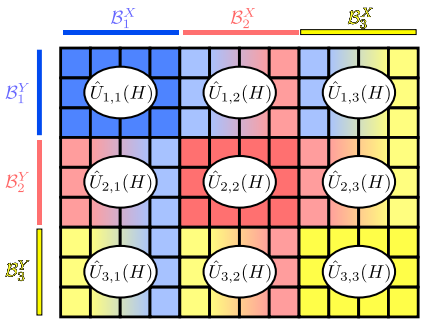


Figure 4: The MoU_2 estimator.

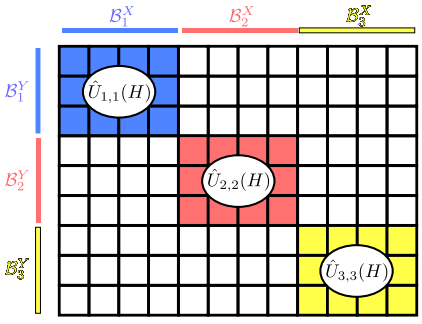


Figure 5: The $\mathrm{MoU}_2^{\mathrm{diag}}$ estimator.

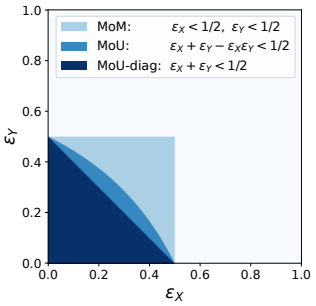


Figure 6: Outliers accepted.

$\lceil \beta(\varepsilon_X + \varepsilon_Y)\log(1/\delta)\rceil$, *it holds w.p.a.l.* $1-\delta$*:*

$$\left|\hat{\theta}_{\mathrm{MoU}_2^{diag}}(H) - \theta(H)\right| \leq 4\sqrt{e}\,\Sigma(H)\,\gamma(\varepsilon_X+\varepsilon_Y)\sqrt{\frac{1+\log(1/\delta)}{n}}.$$

If in addition $\|H(X,Y)\|_\infty$ *is finite and upper bounded by* M*, then for all* $\delta \in (0, e^{-4n\alpha(\varepsilon_X+\varepsilon_Y)}]$*, choosing* $K = \lceil \alpha(\varepsilon_X+\varepsilon_Y)n\rceil$*, it holds with probability at least* $1-\delta$*:*

$$\left|\hat{\theta}_{\mathrm{MoU}_2^{diag}}(H) - \theta(H)\right| \leq 8M\,\Gamma(\varepsilon_X+\varepsilon_Y)\sqrt{\frac{\log(1/\delta)}{n}}.$$

If furthermore n_O *and* m_O *satisfy Assumption 3, the same* K *gives:*

$$\mathbb{E}\left[\left|\hat{\theta}_{\mathrm{MoU}_2^{diag}}(H) - \theta(H)\right|\right] \leq 4M\,\Gamma(\varepsilon_X+\varepsilon_Y)\left(4\sqrt{2}\,C_\mathsf{O}\,\frac{\Delta(\varepsilon_X+\varepsilon_Y)}{n^{(1-\alpha_\mathsf{O})/2}} + \sqrt{\frac{\pi}{n}}\right).$$

The proof can be found in Appendix C.4. Notice that the constraint $n = m$ can be relaxed, as long as $2(n_\mathsf{O}+m_\mathsf{O}) \leq \min(n,m)$ 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 the critical point $\varepsilon_X + \varepsilon_Y = 1/2$. When estimating Integral Probability Metrics (Sriperumbudur et al., 2012), one typically relies on two-sample U-statistics, built upon kernels of the form $H_\phi(X,Y) = \phi(X) - \phi(Y)$, for ϕ in the functional set considered. Hence, one might use a MoM-MoM estimate, instead of a MoU_2 or a $\mathrm{MoU}_2^{\mathrm{diag}}$ estimate (see Staerman et al. (2021) for an application to the estimation of the 1-Wasserstein distance). The corresponding proportions of outliers admitted would be $\varepsilon_X < 1/2$, and $\varepsilon_Y < 1/2$, that represents a less stringent constraint, as shown in Figure 6. For p-sample U-statistics this constraints would write as $\|\boldsymbol{\varepsilon}\|_\infty < 1/2$ for a MoM-based estimate, and $\|\boldsymbol{\varepsilon}\|_1 < 1/2$ for MoU_p, with $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)$ the vector containing the p samples proportions of outliers.

3. Statistical Guarantees for Pairwise Learning in the Presence of Outliers

A simple and meaningful way to illustrate the relevance of MoM-based estimators in the presence of outliers is to use them for revisiting the Empirical Risk Minimization paradigm (ERM, see *e.g.* Devroye et al. (1996)). 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, a hypothesis set $\mathcal{G} \subset \mathcal{Y}^\mathcal{X}$, and a loss function $\ell\colon \mathcal{G}\times\mathcal{Z} \to \mathbb{R}_+$. ERM then consists in substituting the unknown risk $\mathbb{E}_P[\ell(g,Z)]$ by its empirical version based on sample $\mathcal{S}_n$, and solving the optimization problem $\min_{g\in\mathcal{G}}(1/n)\sum_{i=1}^n \ell(g,Z_i)$. When $\mathcal{S}_n$ is possibly contaminated, a natural idea to robustify ERM is to solve instead $\min_{g\in\mathcal{G}} \mathrm{MoM}_{\mathcal{S}_n}[\ell(g,Z)]$. This approach, 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 the concentration bounds established in Section 2 to extend these ideas to pairwise learning problems, with a simpler formalism based on the Vapnik-Chervonenkis dimension. Consider now a hypothesis set $\mathcal{G} \subset \{-1,+1\}^{\mathcal{X}\times\mathcal{X}}$, and a symmetric loss function $\ell\colon \mathcal{G}\times\mathcal{Z}^2 \to \mathbb{R}_+$. Let Z' denote 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 over $\mathcal{G}$ $\mathcal{R}(g) = \mathbb{E}_{Z,Z'}[\ell_g(Z,Z')]$. Classical examples of problems covered by this setting include *metric learning* and *ranking*. We study the performance of the MoU-minimizer $\hat{g}_{\mathrm{MoU}} = \mathrm{argmin}_{g\in\mathcal{G}}\ \mathrm{MoU}_{\mathcal{S}_n}(\ell_g)$, where

$$\mathrm{MoU}_{\mathcal{S}_n}(\ell_g) = \mathrm{median}\Big(\sum_{i<j\in\mathcal{B}_1}\ell_g(Z_i,Z_j), \ldots \sum_{i<j\in\mathcal{B}_K}\ell_g(Z_i,Z_j)\Big).$$

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

Assumption 5. *The hypothesis space* $\mathcal{G}$ *considered has finite VC dimension* $\mathrm{VC}_{dim}(\mathcal{G})$.

Assumption 6. *There exists $M > 0$ such that it holds $\ell(g, Z, Z') \leq M$ almost surely.*

Assumptions 5 and 6 are standard in statistical learning. One typically has $M = 1$ for the 0-1 loss $\ell\colon (g, Z, Z') \mapsto \mathbb{I}\{(g(X, X')(Y - Y') \leq 0\}$. Notice that if $\mathcal{Y}$ is bounded, any convex relaxation of the latter also fits. We again stress that Assumption 6 only applies to the inliers, *i.e.* to the realizations of Z and Z', not necessarily to the outliers. The next theorem characterizes $\hat{g}_{\text{MoU}}$'s generalization capacity.

Theorem 1. *Suppose that sample $\mathcal{S}_n$ and mapping α satisfy Assumptions 2 and 4 respectively. Define functions Γ, Δ according to Assumption 4. Assume furthermore that $\mathcal{G}$ and ℓ satisfy Assumptions 5 and 6 respectively. Then, for all $\delta \in [0, e^{-4\Delta^2(\varepsilon)n_{\text{O}}}]$, choosing $K = \lceil \alpha(\varepsilon)n \rceil$, it holds with probability at least $1 - \delta$:*

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

Theorem 1 is proved by combining the second claim of Proposition 3 with the complexity assumption on $\mathcal{G}$, details can be found in Appendix C.6. We emphasize on the generic nature of the bounds established in Section 2. This key property allows to efficiently combine them with various complexity assumptions on $\mathcal{G}$. A second generalization bound based upon an entropic control of $\mathcal{G}$ is for instance proposed in Appendix C.7. In contrast, the guarantees in Lecué et al. (2018) uses an *ad hoc* Rademacher complexity specifically tailored to their needs. If VC dimensions are also used in Depersin (2020), we emphasize that it is for estimation purposes, that do not relate to the learning bounds established in Theorem 1.

From an algorithmic point of view, computing decision functions with guarantees similar to that in Theorem 1 can be done through MoU Gradient Descent (MoU-GD). It is an pairwise adaptation of the algorithm proposed in Lecué et al. (2018), that can be described as follows. For simplicity, we assume that $\mathcal{G}$ is a parametric hypothesis set of dimension p, *i.e.* for every $g \in \mathcal{G}$ there exists $u \in \mathbb{R}^p$ such that $g = g_u$. MoU-GD then revisits minibatch Gradient Descent in the following way. At each step, the dataset is partitioned, and (pairwise) risk estimates are computed on each block. The block with the median risk is selected, and a minibatch Gradient Descent step is computed, with the median block acting as the minibatch. This is repeated until convergence. The approach is formally detailed in Algorithm 1. Observe that the partition needs to be randomized at each iteration in order to avoid local minima, see Remark 5 in Lecué et al. (2018). Under standard convexity assumptions (detailed in Appendix C.8, along with the proof of Theorem 2), we now show that the output of Algorithm 1 converges towards $\hat{g}_{\text{alg}}$, that enjoys the guarantees established in Theorem 1.

Algorithm 1 MoU Gradient Descent (MoU-GD)

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

for *epoch from* 1 *to* T **do**

```
// Randomly partition the data
```

Choose a random permutation π of $\{1, \ldots, n\}$

Build a partition $B_1, \ldots, B_k$ of $\{\pi(1), \ldots, \pi(n)\}$

```
// Select block with median risk
```

for $k \leq K$ **do**

$\hat{U}_{B_k} = \sum_{i<j \in B_k^2} \ell(g_{u_t}, Z_i, Z_j)$

Set B_{med} s.t. $\hat{U}_{B_{\text{med}}} = \text{median}(\hat{U}_{B_k}, \ldots \hat{U}_{B_K})$

```
// Gradient step
```

$u_{t+1} = u_t - \gamma_t \sum_{i<j \in B_k^2} \nabla_{u_t} \ell(g_{u_t}, Z_i, Z_j)$

return u_T

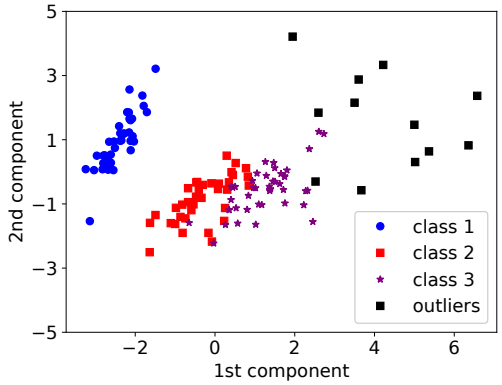
Theorem 2. *Suppose that the assumptions of Theorem 1 hold, and that pairwise adaptations of the assumptions of Theorem 3 in Lecué et al. (2018) hold. Then, the output of Algorithm 1 converges almost surely towards $\hat{g}_{\text{alg}}$, that satisfies with probability at least $1 - \delta$:*

$$\mathcal{R}(\hat{g}_{\text{alg}}) - \mathcal{R}(g^*) \leq 8\sqrt{2}M\,\Gamma(\varepsilon)\sqrt{\frac{\text{VC}_{dim}(\mathcal{G})(1 + \log(n)) + \log(1/\delta)}{n}}.$$

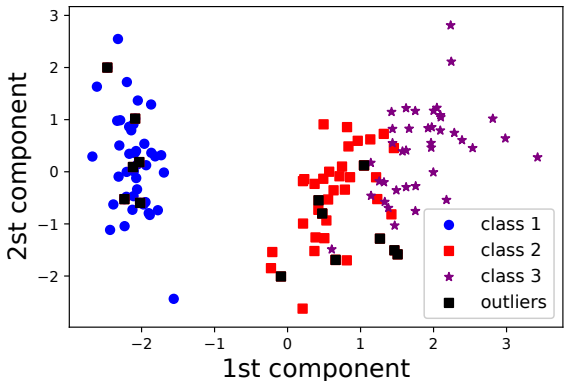
We conclude by an empirical evaluation of Theorem 2. In metric learning, one is interested in learning a distance $d\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, that should coincide with some *a priori* information. We consider the set of Malahanobis distances on $\mathbb{R}^q$ $d_M^2\colon (x, x') \mapsto (x - x')^\top M(x - x')$, with $M \in \mathbb{R}^{q \times q}$ positive semi-definite, and the *iris* dataset, 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, as we want flowers coming from the same class to be close according to our metric, and conversely. Denoting $y_{ij} = 2 \cdot \mathbb{I}\{y_i = y_j\} - 1$, the (pairwise) criterion we want to optimize writes as follows:

$$\min_{M \in S_q^+(\mathbb{R})} \frac{2}{n(n-1)} \sum_{i<j} \max\Big(0, 1 + y_{ij}(d_M^2(x_i, x_j) - 2)\Big).$$

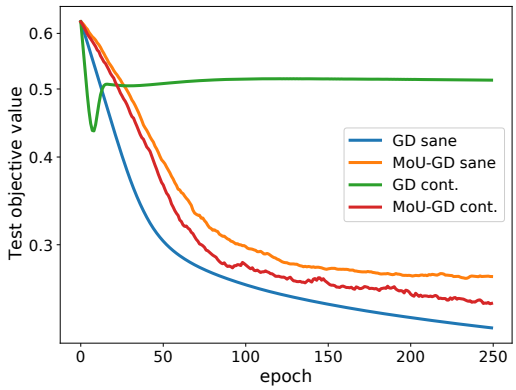
The whole dataset is first normalized and divided into a train set of size 80% and a test set of size 20%. Then, the training data is contaminated with 10% of outliers drawn uniformly over $[0, 5]^4$, and with label 2, see Figure 7a. Standard and MoU Gradient Descents are run (with a projection step 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 descent trajectories on the test data, averaged over 100 runs, are plotted in Figure 7c. MoU-GD remarkably resists to the presence of outliers, and shows test



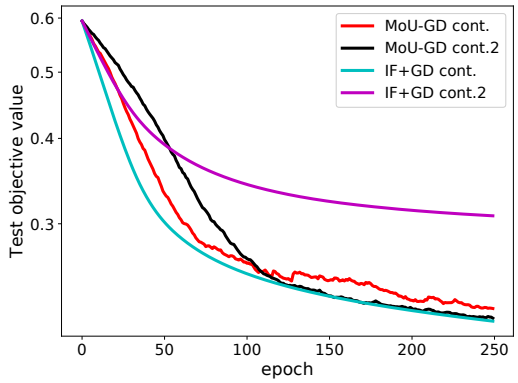
(a) Contamination 1



(b) Contamination 2



(c) MoU-GD on contamination 1



(d) MoU-GD and IF on contaminations 1 and 2

Figure 7: MoU-GD on a metric learning example.

performance comparable to the sane 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. The fact that MoU-GD performs better on the contaminated dataset is not so surprising. Indeed, 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.

However, contamination scenario 1 can be easily tackled using first an outlier detection algorithm, here Isolation Forest (IF), with the knowledge of ε, and then a standard GD on the sanitized dataset. Yet, such outlier detection algorithms are by nature unsupervised, and it is easy to elaborate contamination scenarios to fool them (Figure 7b). The corresponding IF+GD approach then also fails, as can be seen in Figure 7d. In contrast, MoU-GD performs well in all scenarios. Its ability to automatically discard all types of outliers, and without preprocessing, makes it a promising end-to-end approach for learning in the presence of outliers.

4. Conclusion

Widely analyzed and proved valid in the context of heavy-tailed data, the Median-of-Means (MoM) estimator is now the subject of numerous analyses under a variant of the Huber's contamination model. The present article offers an exhaustive view of its robustness properties under this regime, and proposes several concentration bounds with clear dependence on the proportions of outliers ε and the number of blocks K, that can be extended to (multisample) U-statistics. These bounds are incidentally shown to supply a sound theoretical basis for the reliability of MoM-based learning techniques when the training dataset is possibly contaminated in part by outliers with arbitrary distribution.

Acknowledgment. This work has been partially funded by the industrial chair *Data Science & Artificial Intelligence for Digitalized Industry and Services* from Télécom Paris. Guillaume Staerman was also supported by BPI France in the context of the PSPC Project Expresso (2017-2021). Finally, authors would like to extend a special thank to Florence d'Alché-Buc for her insightful discussions and helpful suggestions.

References

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

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

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.

Cheng, Y., Diakonikolas, I., and Ge, R. (2019). High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the thirtieth annual ACM-SIAM symposium on discrete algorithms*, pages 2755–2771. SIAM.

Cherapanamjeri, Y., Flammarion, N., and Bartlett, P. L. (2019). Fast mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, pages 786–806. PMLR.

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.

Depersin, J. (2020). Robust subgaussian estimation with vc-dimension. *arXiv preprint arXiv:2004.11734*.

Depersin, J. and Lecué, G. (2019). Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*.

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.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2016). Robust estimators in high dimensions without the computational intractability. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 655–664.

Diakonikolas, I., Kane, D. M., and Pensia, A. (2020). Outlier robust mean estimation with subgaussian rates via stability. *arXiv preprint arXiv:2007.15618*.

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

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

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

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101.

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)*.

Lai, K. A., Rao, A. B., and Vempala, S. (2016). Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674.

Lecué, G. and Lerasle, M. (2019). Learning from mom's principles: Le cam's approach. *Stochastic Processes and their applications*, 129(11):4385–4410.

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.

Lee, J. C. and Valiant, P. (2020). Optimal sub-gaussian mean estimation in $\mathbb{R}$. *arXiv preprint arXiv:2011.08384*.

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. (2019a). Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*.

Lugosi, G. and Mendelson, S. (2019b). Robust multivariate mean estimation: the optimality of trimmed mean. *arXiv preprint arXiv:1907.11391*.

Lugosi, G. and Mendelson, S. (2019c). Sub-gaussian estimators of the mean of a random vector. *Ann. Statist.*, 47(2):783–794.

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.

Oliveira, R. I. and Valdora, M. (2019). The sub-gaussian property of trimmed means estimators. *Technical report, IMPA*.

Prasad, A., Balakrishnan, S., and Ravikumar, P. (2019). A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*.

Prasad, A., Balakrishnan, S., and Ravikumar, P. (2020). A robust univariate mean estimator is all you need. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pages 4034–4044.

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, F. (2021). When ot meets mom: Robust estimation of wasserstein distance. In *International Conference on Artificial Intelligence and Statistics*, pages 136–144. PMLR.

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*.