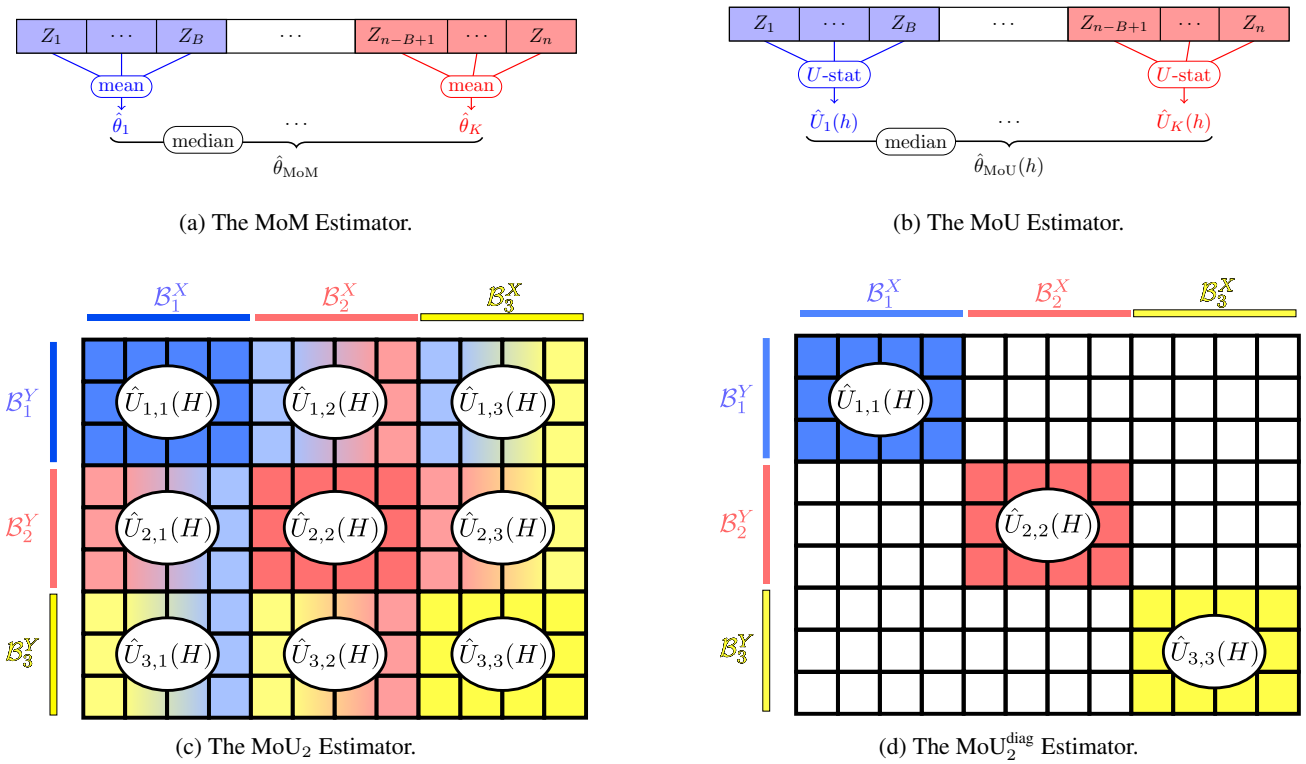
A. Summary: the different estimators considered in the present article



(a) The MoM Estimator.

(b) The MoU Estimator.

(c) The MoU_2 Estimator.

(d) The $\text{MoU}_2^{\text{diag}}$ Estimator.

Figure 8: The estimators considered in the article.

B. Additional Tables and Figures

	$\alpha(\varepsilon)$	$\beta(\varepsilon)$	$\gamma(\varepsilon)$	$\Gamma(\varepsilon)$	$\Delta(\varepsilon)$	$\eta(\varepsilon)$
	$\alpha(\varepsilon)$	$\frac{2\alpha(\varepsilon)}{\alpha(\varepsilon)-2\varepsilon}$	$\frac{\sqrt{\alpha(\varepsilon)}(\alpha(\varepsilon)-\varepsilon)}{(\alpha(\varepsilon)-2\varepsilon)^{3/2}}$	$\sqrt{\frac{\alpha(\varepsilon)}{\alpha(\varepsilon)-2\varepsilon}}$	$\sqrt{\frac{\alpha(\varepsilon)}{\varepsilon}}$	$\frac{\alpha(\varepsilon)-\varepsilon}{\alpha(\varepsilon)}$
ARITHMETIC	$\frac{1+2\varepsilon}{2}$	$\frac{2(1+2\varepsilon)}{1-2\varepsilon}$	$\frac{\sqrt{1+2\varepsilon}}{(1-2\varepsilon)^{3/2}}$	$\frac{\sqrt{1+2\varepsilon}}{\sqrt{1-2\varepsilon}}$	$\sqrt{\frac{1+2\varepsilon}{2\varepsilon}}$	$\frac{1}{1+2\varepsilon}$
GEOMETRIC	$\sqrt{2\varepsilon}$	$\frac{2(1+\sqrt{2\varepsilon})}{1-2\varepsilon}$	$\frac{(2-\sqrt{2\varepsilon})(1+\sqrt{2\varepsilon})^{3/2}}{2(1-2\varepsilon)^{3/2}}$	$\frac{\sqrt{1+\sqrt{2\varepsilon}}}{\sqrt{1-2\varepsilon}}$	$\sqrt[4]{2/\varepsilon}$	$\frac{2-\sqrt{2\varepsilon}}{2}$
HARMONIC	$\frac{4\varepsilon}{1+2\varepsilon}$	$\frac{4}{1-2\varepsilon}$	$\frac{3-2\varepsilon}{\sqrt{2}(1-2\varepsilon)^{3/2}}$	$\frac{\sqrt{2}}{\sqrt{1-2\varepsilon}}$	$\sqrt{\frac{4}{1+2\varepsilon}}$	$\frac{3-2\varepsilon}{4}$
POLYNOMIAL	$\varepsilon\left(\frac{5}{2}-\varepsilon\right)$	$\frac{2(5-2\varepsilon)}{1-2\varepsilon}$	$\frac{(3-2\varepsilon)\sqrt{5-2\varepsilon}}{(1-2\varepsilon)^{3/2}}$	$\frac{\sqrt{5-2\varepsilon}}{\sqrt{1-2\varepsilon}}$	$\sqrt{\frac{5-2\varepsilon}{2}}$	$\frac{3-2\varepsilon}{5-2\varepsilon}$

Table 1: Different upper bounds α and corresponding functions $\beta, \gamma, \Gamma, \Delta, \eta$.

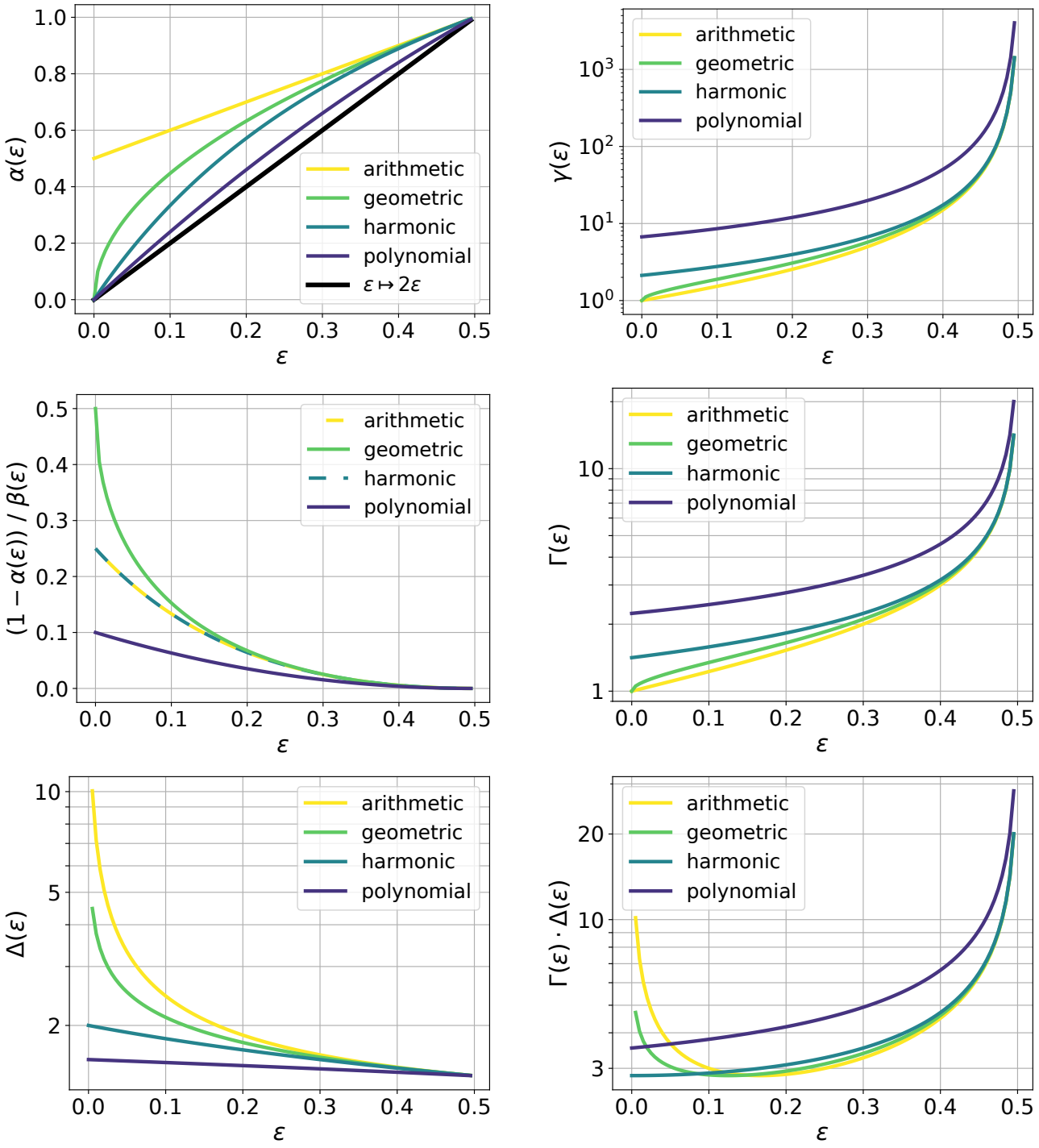


Figure 9: Influence of the chosen mapping α on the constants.

C. Technical Proofs

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

C.1. Proof of Proposition 2

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 *sane* 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 sane 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 $\eta > 1/2$ (that eventually depends on ε) such that $K_{\mathsf{S}} \geq \eta K$. 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:

$$2n_{\mathsf{O}} = 2\varepsilon n < K \leq n. \tag{8}$$

Let $\alpha\colon [0, 1/2] \to [0, 1]$ such that: $\forall \varepsilon \in (0, 1/2)$, $2\varepsilon < \alpha(\varepsilon) < 1$. Several choices of acceptable function α are detailed in Table 1, and illustrated in Figure 9. They include among others:

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

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

$$\alpha(\varepsilon)n \leq K \leq n.$$

It directly follows that

$$K_{\mathsf{S}} = K - K_{\mathsf{O}} \geq K - n_{\mathsf{O}} \geq K - \varepsilon n \geq \left(1 - \frac{\varepsilon}{\alpha(\varepsilon)}\right) K = \frac{\alpha(\varepsilon) - \varepsilon}{\alpha(\varepsilon)}\, K,$$

and one then may use

$$\eta = \eta(\varepsilon) = \frac{\alpha(\varepsilon) - \varepsilon}{\alpha(\varepsilon)}.$$

Once $\eta(\varepsilon)$ 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:

$$\begin{aligned}
\mathbb{P}\Big\{\big|\hat{\theta}_{\mathrm{MoM}} - \theta\big| > t\Big\} &\leq \mathbb{P}\left\{\sum_{\text{blocks without outlier}} \mathbb{1}\Big\{\big|\hat{\theta}_{\text{block}} - \theta\big| > t\Big\} \geq K_{\mathsf{S}} - \frac{K}{2}\right\}, \\
&\leq \mathbb{P}\left\{\sum_{\text{blocks without outlier}} \mathbb{1}\Big\{\big|\hat{\theta}_{\text{block}} - \theta\big| > t\Big\} \geq \frac{2\eta(\varepsilon) - 1}{2\eta(\varepsilon)} K_{\mathsf{S}}\right\},
\end{aligned} \tag{9}$$

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

$$\begin{aligned}\sum_{k=\lceil \frac{2\eta(\varepsilon)-1}{2\eta(\varepsilon)} K_{\mathsf{S}} \rceil}^{K_{\mathsf{S}}} \binom{K_{\mathsf{S}}}{k} p_t^k \, (1-p_t)^{K_{\mathsf{S}}-k} \;&\leq\; p_t^{\frac{2\eta(\varepsilon)-1}{2\eta(\varepsilon)} K_{\mathsf{S}}} \sum_{k=1}^{K_{\mathsf{S}}} \binom{K_{\mathsf{S}}}{k}, \\ &\leq\; p_t^{\frac{2\eta(\varepsilon)-1}{2\eta(\varepsilon)} K_{\mathsf{S}}} \, 2^{K_{\mathsf{S}}}, \\ &\leq\; p_t^{\frac{2\eta(\varepsilon)-1}{2} K} \, 2^{\eta(\varepsilon) K}.\end{aligned}$$

By virtue of Chebyshev's inequality, it holds that $p_t \leq \sigma^2/(Bt^2)$, with $B = \lfloor n/K \rfloor$ denoting the size of the blocks. The right-hand side can then be rewritten as

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

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

$$\begin{aligned} t &= \sqrt{e}\sigma \sqrt{\frac{2^{\frac{2\eta(\varepsilon)}{2\eta(\varepsilon)-1}}}{B}}, \\ &\leq \sqrt{e}\sigma \sqrt{\frac{4\eta^2(\varepsilon)}{(2\eta(\varepsilon)-1)^2} \frac{2K}{n}}, \\ &\leq 4\sqrt{e}\sigma \, \frac{\eta(\varepsilon)}{(2\eta(\varepsilon)-1)^{\frac{3}{2}}} \sqrt{\frac{1+\log(1/\delta)}{n}}, \end{aligned} \tag{10}$$

where we have used $2^{\frac{1}{x}} \leq 1/x^2$ for $x \leq 1/2$, and $\lfloor x \rfloor \geq x/2$ for $x \geq 1$.

The final writing is obtained by setting

$$\beta(\varepsilon) = \frac{2}{2\eta(\varepsilon)-1} = \frac{2\alpha(\varepsilon)}{\alpha(\varepsilon) - 2\varepsilon},$$

and

$$\gamma(\varepsilon) = \frac{\eta(\varepsilon)}{(2\eta(\varepsilon)-1)^{\frac{3}{2}}} = \frac{\sqrt{\alpha(\varepsilon)}(\alpha(\varepsilon)-\varepsilon)}{(\alpha(\varepsilon)-2\varepsilon)^{\frac{3}{2}}}.$$

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 δ's. Indeed, it is enough for δ to satisfy:

$$\alpha(\varepsilon) n \;\leq\; \beta(\varepsilon) \log(1/\delta) \;\leq\; n,$$

$$e^{-n/\beta(\varepsilon)} \;\leq\; \delta \;\leq\; e^{-n\alpha(\varepsilon)/\beta(\varepsilon)}.$$

The limitation on the range of δ 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 $2n_{\mathsf{O}} < K$ (or $\alpha(\varepsilon) n \leq K$), and is specific to our outlier framework. The purpose of the second part of Proposition 2 is precisely to remove the left limitation, under the assumption that Z is ρ sub-Gaussian.

Assume now that Z is ρ 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

$$\begin{aligned}\mathbb{P}\Big\{ \big|\hat{\theta}_{\text{MoM}} - \theta\big| > t \Big\} &\leq \exp\left(\frac{2\eta(\varepsilon)-1}{2} K \cdot \log\left[2^{\frac{4\eta(\varepsilon)-1}{2\eta(\varepsilon)-1}} e^{-Bt^2/2\rho^2} \right] \right), \\ &\leq \exp\left(-\frac{2\eta(\varepsilon)-1}{16\rho^2} n t^2 \right),\end{aligned}$$

for all t such that

$$t^2 \geq \frac{4\rho^2}{B}\,\frac{4\eta(\varepsilon)-1}{2\eta(\varepsilon)-1}\log 2,$$

Reverting in δ gives that it holds with probability at least $1-\delta$

$$\left|\hat{\theta}_{\mathrm{MoM}}-\theta\right| \leq \frac{4\rho}{\sqrt{2\eta(\varepsilon)-1}}\sqrt{\frac{\log(1/\delta)}{n}},$$

for all δ that satisfies

$$\delta \leq e^{-\frac{\log 2}{4}(4\eta(\varepsilon)-1)\frac{n}{B}}, \quad \text{and in particular} \quad \delta \leq e^{-4n\alpha(\varepsilon)}. \tag{11}$$

Indeed it holds $B=\lfloor n/K\rfloor \geq n/(2K)$, so that $n/B \leq 2K = 2\lceil \alpha(\varepsilon)n\rceil \leq 2(\alpha(\varepsilon)n+1) \leq 4\alpha(\varepsilon)n$, since $1 \leq 2n_{\mathsf{O}} = 2\varepsilon n \leq \alpha(\varepsilon)n$. When $n_{\mathsf{O}}=\varepsilon=0$, one may choose $K=1$, $B=n$, and $\delta \leq 1/e$.

The final writing is obtained by setting:

$$\Gamma(\varepsilon) = \frac{1}{\sqrt{2\eta(\varepsilon)-1}} = \sqrt{\frac{\alpha(\varepsilon)}{\alpha(\varepsilon)-2\varepsilon}}.$$

To get the expectation bound, one may simply integrate the previously found deviation probabilities. Reverting the inequality gives that it holds

$$\mathbb{P}\left\{\left|\hat{\theta}_{\mathrm{MoM}}-\theta\right| > t\right\} \leq e^{-\frac{nt^2}{16\rho^2\Gamma^2(\varepsilon)}},$$

for all t such that (using Assumption 3):

$$t \geq 8\rho\,\Gamma(\varepsilon)\sqrt{\alpha(\varepsilon)}, \quad \text{and in particular} \quad t \geq 8\rho\,\Gamma(\varepsilon)\sqrt{\frac{\alpha(\varepsilon)}{\varepsilon}}\frac{C_{\mathsf{O}}}{n^{(1-\alpha_{\mathsf{O}})/2}}. \tag{12}$$

One finally gets

$$\begin{aligned}
\mathbb{E}\left[\left|\hat{\theta}_{\mathrm{MoM}}-\theta\right|\right] &= \int_0^\infty \mathbb{P}\left\{\left|\hat{\theta}_{\mathrm{MoM}}-\theta\right| > t\right\} dt, \\
&\leq \int_0^{8\rho\,\Gamma(\varepsilon)\sqrt{\frac{\alpha(\varepsilon)}{\varepsilon}}\frac{C_{\mathsf{O}}}{n^{(1-\alpha_{\mathsf{O}})/2}}} 1dt + \int_0^\infty e^{-\frac{nt^2}{16\rho^2\Gamma^2(\varepsilon)}}dt, \\
&\leq 8\rho\,\Gamma(\varepsilon)\sqrt{\frac{\alpha(\varepsilon)}{\varepsilon}}\frac{C_{\mathsf{O}}}{n^{(1-\alpha_{\mathsf{O}})/2}} + \frac{2\sqrt{\pi}\rho\,\Gamma(\varepsilon)}{\sqrt{n}}, \\
&\leq 2\rho\,\Gamma(\varepsilon)\left(4C_{\mathsf{O}}\,\frac{\Delta(\varepsilon)}{n^{(1-\alpha_{\mathsf{O}})/2}} + \sqrt{\frac{\pi}{n}}\right),
\end{aligned}$$

with the notation

$$\Delta(\varepsilon) = \sqrt{\frac{\alpha(\varepsilon)}{\varepsilon}}.$$

□

Remark 2. *Coming back to Equation (9), one may also use Hoeffding's inequality to get:*

$$\begin{aligned}
\mathbb{P}\left\{\left|\hat{\theta}_{\mathrm{MoM}}-\theta\right| > t\right\} &\leq \mathbb{P}\left\{\frac{1}{K_{\mathsf{S}}}\sum_{\textit{blocks without outlier}} \mathbb{1}\left\{\left|\hat{\theta}_{\textit{block}}-\theta\right| > t\right\} - p_t \geq \frac{2\eta(\varepsilon)-1}{2\eta(\varepsilon)} - \frac{\sigma^2}{Bt^2}\right\}, \\
&\leq \exp\left(-2\eta(\varepsilon)K\left(\frac{2\eta(\varepsilon)-1}{2\eta(\varepsilon)} - \frac{\sigma^2}{Bt^2}\right)^2\right).
\end{aligned} \tag{13}$$

The right-hand side can be set to δ by choosing $K = \left\lceil \frac{9}{2} \frac{\eta(\varepsilon)}{(2\eta(\varepsilon)-1)^2} \log(1/\delta) \right\rceil$, and t's that satisfy:

$$\begin{aligned}
\frac{2\eta(\varepsilon)-1}{6\eta(\varepsilon)} &= \frac{\sigma^2}{Bt^2}, \\
t &= \sqrt{6}\sigma \sqrt{\frac{\eta(\varepsilon)}{2\eta(\varepsilon)-1}} \frac{1}{\sqrt{B}}, \\
t &\leq \sqrt{6}\sigma \sqrt{\frac{\eta(\varepsilon)}{2\eta(\varepsilon)-1}} \sqrt{\frac{2K}{n}}, \\
t &\leq 3\sqrt{6}\sigma \frac{\eta(\varepsilon)}{(2\eta(\varepsilon)-1)^{\frac{3}{2}}} \sqrt{\frac{1+\log(1/\delta)}{n}}.
\end{aligned}$$

Up to the constant term which is bigger ($3\sqrt{6}$ instead of $4\sqrt{e}$), 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 ρ sub-Gaussian r.v. Z. Even if this sharper bound for p_t is plugged in Equation (13), the quantity $(2\eta(\varepsilon)-1)/(2\eta(\varepsilon)) - 2\exp(-Bt^2/(2\rho^2))$ may never go to 0, making it impossible to improve the confidence range similarly to what has been done in Proposition 2. Notice that the same problem arises in the proof of Proposition 4.

C.2. Proof of Proposition 3

The proof of Proposition 2 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 $\zeta_c(h) = \mathrm{Cov}(h(Z_{i_1}, \ldots, Z_{i_d}), h(Z_{i'_1}, \ldots, Z_{i'_d}))$ when c variables are common. Noticing that $\zeta_0(h) = 0$, it holds:

$$\begin{aligned}
\mathrm{Var}\left(\bar{U}_B(h)\right) &= \mathrm{Cov}\left(\frac{1}{\binom{B}{d}} \sum_{i_1<\ldots<i_d} h\left(Z_{i_1}, \ldots, Z_{i_d}\right), \frac{1}{\binom{B}{d}} \sum_{i'_1<\ldots<i'_d} h\left(Z_{i'_1}, \ldots, Z_{i'_d}\right) \right), \\
&= \frac{1}{\binom{B}{d}^2} \sum_{\substack{i_1<\ldots<i_d \\ i'_1<\ldots<i'_d}} \mathrm{Cov}\left(h\left(Z_{i_1}, \ldots, Z_{i_d}\right), h\left(Z_{i'_1}, \ldots, Z_{i'_d}\right) \right), \\
&= \frac{1}{\binom{B}{d}} \sum_{c=1}^{d} \binom{d}{c} \binom{B-d}{d-c} \zeta_c(h), \\
&= \sum_{c=1}^{d} \frac{d!^2}{c!(d-c)!^2} \frac{(B-d)(B-d-1)\ldots(B-2d+c+1)}{B(B-1)\ldots(B-d+1)} \zeta_c(h), \\
&\leq d! \frac{\sum_{c=1}^{d} \binom{d}{c} \zeta_c(h)}{B}, \\
&= \frac{\Sigma^2(h)}{B},
\end{aligned}$$

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 ρ sub-Gaussian. If kernel h is bounded, then Hoeffding's inequality for U-statistics (Hoeffding, 1963) gives instead that $p_t \leq 2e^{-Bt^2/2d\|h\|_\infty^2}$. The rest of the proof is similar to that of Proposition 2. 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. □

C.3. Proof of Proposition 4

Let us first recall the notation needed to the analysis of $\hat{\theta}_{\mathrm{MoU}_2}(H)$. 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 $K_{X,\mathsf{S}}$ and $K_{Y,\mathsf{S}}$, and for $k \leq K_X$ and $l \leq K_Y$, we set:

$$\hat{U}_{k,l}(H) = \frac{1}{B_X B_Y} \sum_{i \in \mathcal{B}_k^X} \sum_{j \in \mathcal{B}_l^Y} H(X_i, Y_j),$$

the (two-sample) U-statistic built upon blocks $\mathcal{B}_k^X$ and $\mathcal{B}_l^Y$. Let $I_{k,l}^t = \mathbb{1}\{|\hat{U}_{k,l}(H) - \theta(H)| > t\}$ 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:

$$\alpha(\varepsilon_X + \varepsilon_Y - \varepsilon_X\varepsilon_Y)nm \leq K_X K_Y \leq nm. \tag{14}$$

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

$$\begin{cases} \sqrt{\alpha(\varepsilon_X + \varepsilon_Y - \varepsilon_X\varepsilon_Y)}n \leq K_X \leq n, \\ \sqrt{\alpha(\varepsilon_X + \varepsilon_Y - \varepsilon_X\varepsilon_Y)}m \leq K_Y \leq m. \end{cases} \tag{15}$$

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

$$K_{X,\mathsf{S}} K_{Y,\mathsf{S}} \geq \left(1 - \frac{\varepsilon_X + \varepsilon_Y - \varepsilon_X\varepsilon_Y}{\alpha(\varepsilon_X + \varepsilon_Y - \varepsilon_X\varepsilon_Y)}\right) K_X K_Y \coloneqq \eta_{XY} \cdot K_X K_Y,$$

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

$$\begin{cases} K_{X,\mathsf{S}} \geq \left(1 - \frac{\varepsilon_X}{\sqrt{\alpha(\varepsilon_X + \varepsilon_Y - \varepsilon_X\varepsilon_Y)}}\right) K_X \coloneqq \eta_X \cdot K_X, \\ K_{Y,\mathsf{S}} \geq \left(1 - \frac{\varepsilon_Y}{\sqrt{\alpha(\varepsilon_X + \varepsilon_Y - \varepsilon_X\varepsilon_Y)}}\right) K_Y \coloneqq \eta_Y \cdot K_Y, \end{cases}$$

with a slight abuse of notation since η_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:

$$\begin{aligned} \mathbb{P}\Big\{\big|\hat{\theta}_{\mathrm{MoU}_2}(H) - \theta(H)\big| > t\Big\} &\leq \mathbb{P}\left\{\sum_{k=1}^{K_X}\sum_{l=1}^{K_Y} I_{k,l}^t \geq \frac{K_X K_Y}{2}\right\}, \\ &\leq \mathbb{P}\left\{\sum_{\text{blocks without outlier}}\sum I_{k,l}^t \geq \frac{2\eta_{XY} - 1}{2\eta_{XY}} K_{X,\mathsf{S}} K_{Y,\mathsf{S}}\right\}. \end{aligned}$$

However, unlike Equation (9), the above equation does not relate to a binomial random variable, as the $I_{k,l}^t$ are not independent, see Figure 4. An elegant alternative then consists in leveraging the independence between samples X and Y and using Hoeffding's inequality. Equation (6) gives $\sigma^2_{B_X,B_Y}(H) \leq \Sigma^2(H)/(B_X \wedge B_Y)$, with $\Sigma^2(H) = \sigma^2(H) + \sigma_1^2(H) + \sigma_2^2(H)$, so that:

$$
\begin{aligned}
&\leq \mathbb{P}\Bigg\{ \frac{1}{K_{X,\mathsf{S}} K_{Y,\mathsf{S}}} \sum\sum_{\text{blocks w/o outlier}} I^t_{k,l} - \mathbb{E}\left[I^t_{k,l} \mid \boldsymbol{X}\right] + \mathbb{E}\left[I^t_{k,l} \mid \boldsymbol{X}\right] - \mathbb{E}\left[I^t_{k,l}\right] \\
&\qquad\qquad \geq \frac{2\eta_{XY}-1}{2\eta_{XY}} - \frac{\Sigma^2(H)}{(B_X \wedge B_Y)t^2} \Bigg\}, \\
&\leq \mathbb{P}\left\{ \frac{1}{K_{Y,\mathsf{S}}} \sum_{l=1}^{K_{Y,\mathsf{S}}} J^t_l - \mathbb{E}\left[J^t_l \mid \boldsymbol{X}\right] \geq \frac{2\eta_{XY}-1}{4\eta_{XY}} - \frac{\Sigma^2(H)}{2(B_X \wedge B_Y)t^2} \right\} + \\
&\quad \mathbb{P}\left\{ \frac{1}{K_{X,\mathsf{S}}} \sum_{k=1}^{K_{X,\mathsf{S}}} \mathbb{E}\left[I^t_{k,l} \mid \boldsymbol{X}\right] - \mathbb{E}\left[I^t_{k,l}\right] \geq \frac{2\eta_{XY}-1}{4\eta_{XY}} - \frac{\Sigma^2(H)}{2(B_X \wedge B_Y)t^2} \right\}, \\
&\leq \exp\left(-2\eta_Y K_Y \left(\frac{2\eta_{XY}-1}{4\eta_{XY}} - \frac{\Sigma^2(H)}{2(B_X \wedge B_Y)t^2} \right)^2 \right) + \\
&\quad \exp\left(-2\eta_X K_X \left(\frac{2\eta_{XY}-1}{4\eta_{XY}} - \frac{\Sigma^2(H)}{2(B_X \wedge B_Y)t^2} \right)^2 \right),
\end{aligned}
$$

with the notation $J^t_l = \dfrac{1}{K_{X,\mathsf{S}}} \displaystyle\sum_{k=1}^{K_{X,\mathsf{S}}} I^t_{k,l}$, and $\boldsymbol{X} = (X_1, \ldots, X_n)$.

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

$$
\begin{aligned}
\frac{\Sigma^2(H)}{2(B_X \wedge B_Y)t^2} &= \frac{2\eta_{XY}-1}{12\eta_{XY}}, \\
t &= \Sigma(H) \sqrt{\frac{6\eta_{XY}}{2\eta_{XY}-1}} \sqrt{\frac{1}{B_X \wedge B_Y}}, \\
&\leq \Sigma(H) \sqrt{\frac{6\eta_{XY}}{2\eta_{XY}-1}} \sqrt{\frac{2\max(K_X, K_Y)}{n \wedge m}}, \\
&\leq 12\sqrt{3}\,\Sigma(H) \left(\frac{\eta_{XY}}{2\eta_{XY}-1} \right)^{\frac{3}{2}} \sqrt{\frac{1+\log(2/\delta)}{n \wedge m}}, \\
&\leq 12\sqrt{3}\,\Sigma(H)\,\gamma(\varepsilon_X + \varepsilon_Y - \varepsilon_X\varepsilon_Y) \sqrt{\frac{1+\log(2/\delta)}{n \wedge m}}.
\end{aligned}
$$

Constraints (15) are finally fulfilled by choosing δ such that:

$$
\begin{cases}
\sqrt{\alpha(\varepsilon_X + \varepsilon_Y - \varepsilon_X\varepsilon_Y)}n \leq \frac{18\,\eta^2_{XY}}{\eta_X(2\eta_{XY}-1)^2} \log(2/\delta) \leq n, \\
\sqrt{\alpha(\varepsilon_X + \varepsilon_Y - \varepsilon_X\varepsilon_Y)}m \leq \frac{18\,\eta^2_{XY}}{\eta_Y(2\eta_{XY}-1)^2} \log(2/\delta) \leq m,
\end{cases}
$$

$$
2\max\left(e^{-n\beta_X}, e^{-m\beta_Y}\right) \;\leq\; \delta \;\leq\; 2\min\left(e^{-n\sqrt{\alpha}/\beta_X}, e^{-m\sqrt{\alpha}/\beta_Y}\right),
$$

with the shortcut notation $\alpha = \alpha(\varepsilon_X + \varepsilon_Y - \varepsilon_X\varepsilon_Y)$, and $\beta_Z = \frac{18\,\eta^2_{XY}}{\eta_Z(2\eta_{XY}-1)^2}$ for $Z = X, Y$. □

C.4. Proof of Proposition 5

Again, the proof can be directly adapted from that of Proposition 2. The first difference lies in the constraint K needs to satisfy. It now writes: $2(n_{\mathsf{O}}+m_{\mathsf{O}}) = 2(\varepsilon_X+\varepsilon_Y)n < K \leq n$, and the reasoning can then be reused in totality with $\varepsilon_X+\varepsilon_Y$ instead of ε. The second difference is Chebyshev's inequality, but Equation (6) gives that $\sigma^2_{B_X,B_Y}(H) \leq \Sigma^2(H)/B$, with $\Sigma^2(H) = \sigma^2(H)+\sigma_1^2(H)+\sigma_2^2(H)$. Finally, when $\|H\|_\infty$ is finite, using the notation $\boldsymbol{X} = (X_1,\dots,X_n)$, one may bound p_t as follows:

$$
\begin{aligned}
p_t =&\ \mathbb{P}\left\{|\hat{U}_{1,1}(H)-\theta(H)| > t\right\},\\
=&\ \mathbb{P}\Big\{\Big|\frac{1}{B^2}\sum_{i\in\mathcal{B}_1^X}\sum_{j\in\mathcal{B}_1^Y} H(X_i,Y_j)-\theta(H)\Big| > t\Big\},\\
\leq&\ \mathbb{P}\Big\{\Big|\frac{1}{B}\sum_{j\in\mathcal{B}_1^Y}\Big(\sum_{i\in\mathcal{B}_1^X}\frac{H(X_i,Y_j)}{B}-\mathbb{E}\Big[\sum_{i\in\mathcal{B}_1^X}\frac{H(X_i,Y_j)}{B}\,\Big|\,\boldsymbol{X}\Big]\Big)\Big| > \frac{t}{2}\,\Big|\,\boldsymbol{X}\Big\}\\
&+\mathbb{P}\Big\{\Big|\frac{1}{B}\sum_{i\in\mathcal{B}_1^X}\mathbb{E}_Y\big[H(X_i,Y)\big]-\theta(H)\Big| > \frac{t}{2}\Big\},\\
\leq&\ 2e^{-Bt^2/8\|H\|_\infty^2}+2e^{-Bt^2/8\|H\|_\infty^2},
\end{aligned}
$$

where we have used Hoeffding's inequality twice: on the $\sum_{i\in\mathcal{B}_1^X}\frac{H(X_i,Y_j)}{B}$ for $j\in\mathcal{B}_1^Y$, conditionally to the X_i's, and a second time to the $\mathbb{E}_Y\big[H(X_i,Y)\big]$ for $i\in\mathcal{B}_1^X$, both random variables being bounded by $\|H\|_\infty$. The rest of the proof is similar to that of Proposition 2. □

C.5. Extension to U-statistics of Arbitrary Degrees and Number of Samples

Similarly to the extension from Proposition 2 to Proposition 3, the first important step consists in upper bounding the variance of the U-statistic. To allow an effective use of Chebyshev's inequality, the latter must be of the order $\mathcal{O}(1/n)$, where we recall that n is the number of observations in the sample (or the size of the smallest sample in the case of a multisample U-statistic). This is for instance the case in Equation (6), *i.e.* for the 2-sample U-statistic of degree $(1,1)$. As a first go, we detail here the derivation of Equation (6). We then show that with similar computations, it is direct to show that for any p-sample U-statistic of degrees $(1,\dots,1)$, the $\mathcal{O}(1/n)$ condition holds. Finally, we extend it to arbitrary degrees. Recall that we compute the variance of the 2-sample U-statistic of degrees $(1,1)$, based on the samples $\mathcal{S}_n^X = \{X_1,\dots,X_n\}$, and $\mathcal{S}_m^Y = \{Y_1,\dots,Y_m\}$. It holds:

$$
\begin{aligned}
&\mathrm{Var}\Big(\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m H(X_i,Y_j)\Big)\\
&\quad=\frac{1}{n^2m^2}\mathrm{Var}\left(\sum_{i=1}^n\sum_{j=1}^m H(X_i,Y_j)\right),\\
&\quad=\frac{1}{n^2m^2}\mathbb{E}\left[\sum_{i,i'=1}^n\sum_{j,j'=1}^m H(X_i,Y_j)H(X_{i'},Y_{j'})\right]-\theta^2(H),\\
&\quad=\frac{1}{nm}\mathbb{E}\left[H^2(X,Y)\right]+\frac{m-1}{nm}\mathbb{E}\left[H(X,Y)H(X,Y')\right]+\frac{n-1}{nm}\mathbb{E}\left[H(X,Y)H(X',Y)\right]-\frac{n+m-1}{nm}\theta^2(H),\\
&\quad=\frac{1}{nm}\sigma^2(H)+\frac{m-1}{nm}\sigma_1^2(H)+\frac{n-1}{nm}\sigma_2^2(H),\\
&\quad\leq\frac{\Sigma^2(H)}{n\wedge m},
\end{aligned}
$$

with $\Sigma^2(H) = \sigma^2(H)+\sigma_1^2(H)+\sigma_2^2(H)$, $\sigma^2(H) = \mathrm{Var}\left(H(X,Y)\right)$, $\sigma_1^2(h) = \mathrm{Cov}\left(H(X,Y), H(X,Y')\right) = \mathrm{Var}(H_1(X))$, with $H_1(x) = \mathbb{E}\left[H(x,Y)\right]$, and $\sigma_2^2(h) = \mathrm{Cov}\left(H(X,Y), H(X',Y)\right) = \mathrm{Var}(H_2(Y))$, with $H_2(y) = \mathbb{E}\left[H(X,y)\right]$.

To highlight the mechanism at stake, we reproduce the above computations for a 3-sample U-statistic of degrees $(1,1,1)$. It is then direct to see that for any p-sample U-statistic of degrees $(1,\dots,1)$, the $\mathcal{O}(1/n)$ condition holds. We have now at disposal a new sample $\mathcal{S}_q^Z = \{Z_1,\dots,Z_q\}$, and the variance of the U-statistic writes:

$$
\begin{aligned}
&\mathrm{Var}\Big(\frac{1}{nmq}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q} H(X_i,Y_j,Z_k)\Big) \\
&\quad= \frac{1}{n^2m^2q^2}\mathrm{Var}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q} H(X_i,Y_j,Z_k)\right), \\
&\quad= \frac{1}{n^2m^2q^2}\mathbb{E}\left[\sum_{i,i'=1}^{n}\sum_{j,j'=1}^{m}\sum_{k,k'} H(X_i,Y_j,Z_k)H(X_{i'},Y_{j'},Z_{k'})\right] - \theta^2(H), \qquad (16)\\
&\quad= \frac{1}{nmq}\mathbb{E}\left[H^2(X,Y,Z)\right] + \frac{(m-1)(q-1)}{nmq}\mathbb{E}\left[H(X,Y,Z)H(X,Y',Z')\right] \\
&\quad\quad+ \frac{(n-1)(q-1)}{nmq}\mathbb{E}\left[H(X,Y,Z)H(X',Y,Z')\right] + \frac{(n-1)(m-1)}{nmq}\mathbb{E}\left[H(X,Y,Z)H(X',Y',Z)\right] \\
&\quad\quad+ \frac{n-1}{nmq}\mathbb{E}\left[H(X,Y,Z)H(X',Y,Z)\right] + \frac{m-1}{nmq}\mathbb{E}\left[H(X,Y,Z)H(X,Y',Z)\right] + \frac{q-1}{nmq}\mathbb{E}\left[H(X,Y,Z)H(X,Y,Z')\right] \\
&\quad\quad- \frac{nmq-(n-1)(m-1)(q-1)}{nmq}\theta^2(H), \\
&\quad= \frac{1}{nmq}\sigma^2(H) + \frac{(m-1)(q-1)}{nmq}\sigma_1^2(H) + \frac{(n-1)(q-1)}{nmq}\sigma_2^2(H) + \frac{(n-1)(m-1)}{nmq}\sigma_3^2(H) \\
&\quad\quad+ \frac{n-1}{nmq}\sigma_{23}^2(H) + \frac{m-1}{nmq}\sigma_{13}^2(H) + \frac{q-1}{nmq}\sigma_{12}^2(H) \\
&\quad\leq \frac{\Sigma^2(H)}{n\wedge m\wedge q},
\end{aligned}
$$

with $\Sigma^2(H) = \sigma^2(H) + \sigma_1^2(H) + \sigma_2^2(H) + \sigma_3^2(H) + \sigma_{23}^2(H) + \sigma_{13}^2(H) + \sigma_{12}^2(H)$, and with a notation abuse $\sigma_{i/ij}^2 = \mathrm{Var}\big(H_{i/ij}(X,Y,Z)\big)$, with $H_{i/ij}(X_1,X_2,X_3) = \mathbb{E}[H(X_1,X_2,X_3) \mid X_i]$ or $\mathbb{E}[H(X_1,X_2,X_3) \mid X_i, X_j]$ respectively.

From this second example we can extrapolate the mechanism that generates the variance of the U-statistic. Coming back to Equation (16), we have to compute a certain number of covariance terms. The important thing that distinguishes the different covariances is the number of variables shared between $H(X_i,Y_j,Z_k)$ and $H(X_{i'},Y_{j'},Z_{k'})$. Depending on this number, and on which variable(s) is (are) shared, one of the $\sigma_{i/ij}^2$ variances appears. This variance is multiplied by the number of times a suitable combination arise. For a shared variable, this is n (respectively, m or q, *i.e.* the size of the associated sample). For non-shared variables, this is $n(n-1)$. As at least one variable is shared (otherwise the two terms are independent, and the expectation is then equal to $\theta^2(H)$, that cancels with the last term of Equation (16)), we end up with variance terms, multiplied by $1/n_{\min}$ at most (because of the $1/(n^2m^2q^2)$ factor). This reasoning validates the $\mathcal{O}(1/n)$ condition discussed earlier, and is applicable to an arbitrary number of samples. Notice finally that it can be shown that all partial variance terms are smaller than $\sigma^2(H) = \mathrm{Var}\big(H(X_1,\dots,X_p)\big)$, so that a simple condition for all the variance terms to be finite is $\sigma^2(H) < +\infty$. The same analysis also applies to arbitrary numbers of samples **and** degrees. Combining it to the variance computation of appendix C.2, it is direct to show that the $\mathcal{O}(1/n)$ remains valid in this setting.

The second important step is the generalization of Hoeffding's inequality when the essential supremum is bounded. There is no particular difficulty here, since Hoeffding's inequality for U-statistics of arbitrary degrees can be used, possibly combined with the condition trick introduced in the previous section when several samples are considered.

C.6. Proof of Theorem 1

Using the fact that $\hat{g}_{\mathrm{MoU}}$ minimizes $\mathrm{MoU}_{\mathcal{S}_n}(\ell_g)$ over $\mathcal{G}$, one gets:

$$\begin{aligned}\mathcal{R}(\hat{g}_{\mathrm{MoU}}) - \mathcal{R}(g^*) &\leq \mathcal{R}(\hat{g}_{\mathrm{MoU}}) - \mathrm{MoU}_{\mathcal{S}_n}(\ell_{\hat{g}_{\mathrm{MoU}}}) + \mathrm{MoU}_{\mathcal{S}_n}(\ell_{g^*}) - \mathcal{R}(g^*),\\ &\leq 2\sup_{g\in\mathcal{G}}\left|\mathrm{MoU}_{\mathcal{S}_n}(\ell_g) - \mathcal{R}(g)\right|,\\ &\leq 2\sup_{g\in\mathcal{G}}\left|\mathrm{MoU}_{\mathcal{S}_n}(\ell_g) - \mathbb{E}[\ell_g]\right|.\end{aligned}$$

For a fixed $g \in \mathcal{G}$, Proposition 3 and Assumption 6 gives that for all $\delta \in (0, \exp(-4n\alpha(\varepsilon))]$, we have with probability larger than $1-\delta$:

$$\left|\mathrm{MoU}_{\mathcal{S}_n}(\ell_g) - \mathbb{E}[\ell_g]\right| \;\leq\; 4\sqrt{2}M\,\Gamma(\varepsilon)\,\sqrt{\frac{\log(1/\delta)}{n}}.$$

By virtue of Sauer's lemma, Assumption 5 altogether with the union bound then gives that for all $\delta \in (0, \exp(-4\Delta^2(\varepsilon)n_{\mathsf{O}})]$, it holds with probability at least $1-\delta$:

$$\sup_{g\in\mathcal{G}}\left|\mathrm{MoU}_{\mathcal{S}_n}(\ell_g) - \mathbb{E}[\ell_g]\right| \;\leq\; 4\sqrt{2}M\,\Gamma(\varepsilon)\,\sqrt{\frac{\mathrm{VC}_{\mathrm{dim}}(\mathcal{G})(1+\log(n)) + \log(1/\delta)}{n}}.$$

□

C.7. Generalization Bound via Entropic Complexity

In this section, we highlight the versatility of the concentration bounds established in Section 2 by deriving generalization guarantees through another complexity assumption than that used in Theorem 1. Namely, we use the following entropic characterization.

Assumption 7. *The collection of functions $\mathcal{L}_{\mathcal{G}} = \{\ell_g \colon g \in \mathcal{G}\}$ is a uniform Donsker class (relative to $\|\cdot\|_\infty$) with polynomial uniform covering numbers, i.e. there exist constants $C_{\mathcal{G}} > 0$ and $r \geq 1$ such that: $\forall \zeta > 0$,*

$$\mathcal{N}(\zeta, \mathcal{L}_{\mathcal{G}}, L_\infty(Q)) \leq C_{\mathcal{G}}(1/\zeta)^r,$$

where $\mathcal{N}(\zeta, \mathcal{L}_{\mathcal{G}}, \|\cdot\|_\infty)$ denotes the number of $\|\cdot\|_\infty$-balls of radius $\zeta > 0$ needed to cover class $\mathcal{L}_{\mathcal{G}}$.

Now, let $\zeta > 0$, and $\ell_1, \ldots, \ell_{\mathcal{N}(\zeta,\mathcal{L}_{\mathcal{G}},\|\cdot\|_\infty)}$ be a ζ-coverage of $\mathcal{L}_{\mathcal{G}}$ with respect to $\|\cdot\|_\infty$. From now on, we use $\mathcal{N} = \mathcal{N}(\zeta, \mathcal{L}_{\mathcal{G}}, \|\cdot\|_\infty)$ for notation simplicity. Let ℓ_g be an arbitrary element of $\mathcal{L}_{\mathcal{G}}$. By definition, there exists $i \leq \mathcal{N}$ such that $\|\ell_g - \ell_i\|_\infty \leq \zeta$. It holds then:

$$\begin{aligned}\left|\mathrm{MoU}_{\mathcal{S}_n}(\ell_g) - \mathbb{E}\left[\ell_g\right]\right| &\leq \left|\mathrm{MoU}_{\mathcal{S}_n}(\ell_g) - \mathrm{MoU}_{\mathcal{S}_n}(\ell_i)\right| + \left|\mathrm{MoU}_{\mathcal{S}_n}(\ell_i) - \mathbb{E}\left[\ell_i\right]\right| + \left|\mathbb{E}\left[\ell_i\right] - \mathbb{E}\left[\ell_g\right]\right|,\\ &\leq 2\zeta + \left|\mathrm{MoU}_{\mathcal{S}_n}(\ell_i) - \mathbb{E}\left[\ell_i\right]\right|. \qquad (17)\end{aligned}$$

Applying the second claim of Proposition 3 to every ℓ_i, the union bound gives that for all $\delta \in (0, e^{-4n\alpha(\varepsilon)}]$, choosing $K = \lceil \alpha(\varepsilon)n \rceil$, it holds with probability at least $1-\delta$:

$$\sup_{i\leq\mathcal{N}}\left|\mathrm{MoU}_{\mathcal{S}_n}[\ell_i] - \mathbb{E}\left[\ell_i\right]\right| \leq 4\sqrt{2}M\Gamma(\varepsilon)\sqrt{\frac{\log(\mathcal{N}/\delta)}{n}}.$$

Taking the supremum in both sides of Equation (17), it holds with probability at least $1-\delta$:

$$\sup_{g\in\mathcal{G}}\left|\mathrm{MoU}_{\mathcal{S}_n}[\ell_g] - \mathbb{E}\left[\ell_g\right]\right| \leq 2\zeta + 4\sqrt{2}M\Gamma(\varepsilon)\sqrt{\frac{\log(\mathcal{N}/\delta)}{n}}.$$

Choosing $\zeta \sim 1/\sqrt{n}$, it holds with probability at least $1-\delta$:

$$\sup_{g\in\mathcal{G}}\left|\mathrm{MoU}_{\mathcal{S}_n}[\ell_g] - \mathbb{E}\left[\ell_g\right]\right| \leq \frac{2}{\sqrt{n}} + 4\sqrt{2}M\Gamma(\varepsilon)\sqrt{\frac{(r/2)\log(n) + \log(C_{\mathcal{G}}/\delta)}{n}}.$$

We recover the bound of Theorem 1 up to a $\log(n)$ factor.

C.8. Proof of Theorem 2

First, we detail the assumptions needed to derive Theorem 2, that were not explicited in the core text due to space constraints. They are adaptations of the assumptions used to derive Theorem 3 in Lecué et al. (2018). They state as follows.

- for any $u \in \mathbb{R}^p$ and $z, z' \in \mathcal{Z}^2$, it holds: $\left\|\nabla_u \ell(g_u, z, z')\right\| \leq L$,
- for any sample $\mathcal{S}_n$, there exists a unique minimum $u_{\text{min}} = \operatorname{argmin}_{u \in \mathbb{R}^p} \mathbb{E}_{\text{part}}\left[\text{MoU}_{\mathcal{S}_n}(\ell_g) \mid \mathcal{S}_n\right]$, where the expectation is taken with respect to all possible ways of partitioning of sample $\mathcal{S}_n$,
- $\sum_{t=1}^{\infty} \gamma_t = +\infty$, and $\sum_{t=1}^{\infty} \gamma_t^2 < +\infty$,
- for any sample $\mathcal{S}_n$, model $u \in \mathbb{R}^p$, and $\epsilon > 0$, it holds: $\inf_{\|u - u_{\text{min}}\| > \epsilon} (u - u_{\text{min}})^\top \mathbb{E}_{\text{part}}\left[\nabla_u \text{MoU}_{\mathcal{S}_n}(\ell_g) \mid \mathcal{S}_n\right] < 0$,
- for any sample $\mathcal{S}_n$ and model $u \in \mathbb{R}^p$, there exists an open convex set $\mathscr{B}$ containing u such that for any equipartition of $\{1, \ldots, N\}$ into K blocks $\mathcal{B}_1, \ldots, \mathcal{B}_k$ there exists $k_{\text{med}} \leq K$ such that for all $v \in \mathscr{B}$, $\mathcal{B}_{k_{\text{med}}}$ is the median block. Note that this condition must hold almost surely (in $\mathcal{S}_n$) and almost everywhere (in u).

Under these five assumptions, a direct adaptation of Theorem 3 in Lecué et al. (2018) then gives the almost sure convergence of the output of Algorithm 1 towards u_{min}. We have now to study the excess risk of $\hat{g}_{\text{alg}} = g_{u_{\text{min}}}$. Jensen's inequality gives:

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

Applying Theorem 1 then allows to upper bound the right-hand side with high probability, and to conclude. □

D. 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 (multisample) U-statistics estimation experiments under Assumption 3, that emphasize the superiority of MoM/MoU/MoU$_2$ compared to standard alternatives (see Appendix D.1). We also provide implementations of Algorithm 1 on both ranking and metric learning problems (Appendix D.2). They illustrate the good behavior of the MoU Gradient Descent (MoU-GD) when the training dataset is contaminated.

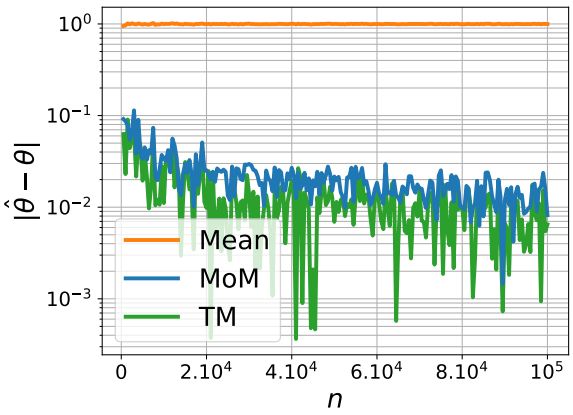
D.1. Estimation Experiments

For all our experiments, we set $n_{\mathsf{O}} = \sqrt{n}$, so that Assumption 3 is fulfilled with $C_{\mathsf{O}} = 1$, $\alpha_{\mathsf{O}} = 1/2$. We next specify particular instances of Assumption 2, *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 2, 3 and 5. We have selected K according to the Harmonic upper bound, so that Assumption 4 is fulfilled as well.

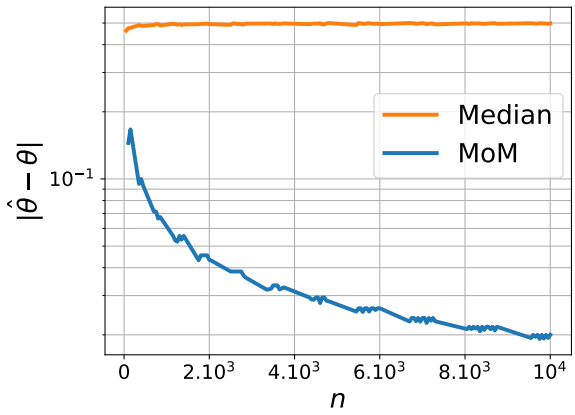
Ruining the mean. In this first example, the sane data is drawn according to a standard Gaussian distribution (hence $\theta = 0$, and the sub-Gaussian assumption is satisfied with $\rho = 1$), and outliers follow a Dirac $\delta_{n^{1/2}}$. The expected value of the empirical mean estimator $\hat{\theta}_{\text{avg}}$ is then given by: $\mathbb{E}_{\mathcal{S}_n}[\hat{\theta}_{\text{avg}}] = (1 - \varepsilon) \cdot 0 + \varepsilon \cdot \sqrt{n} = 1$, always missing the true value. In contrast, MoM's performance improves with n, showing almost no perturbation due to the outliers, see Figure 10a.

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 δ_1. When applying blindly the median, one is actually estimating $q_{1/2+\varepsilon} = 1$. The results are reported in Figure 10b. This phenomenon highlights the importance of correctly choosing α, a too rough approximation such as the median's leading to poor results.

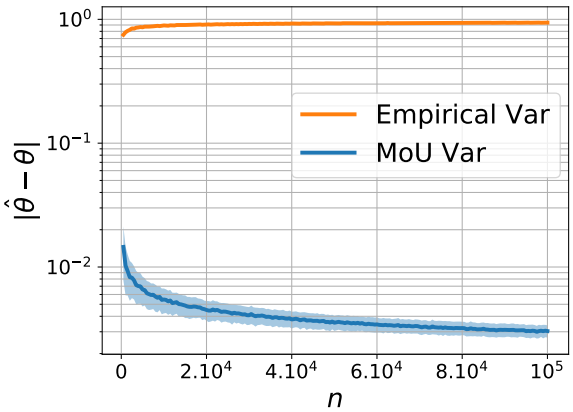
Trimmed mean. We have also benchmarked the results obtained by the Trimmed Mean (TM, Lugosi and Mendelson (2019b)), which provides similar performances for the mean estimation, see Figure 10a.



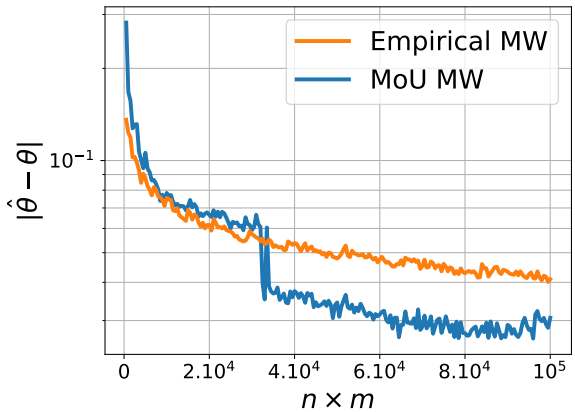
(a) Ruining the mean (avg. 500 runs).



(b) Ruining the median (avg. 500 runs).



(c) MoU estimation of the variance (avg. 500 runs).



(d) MoU estimation of MW-stat (avg. 500 runs).

Figure 10: Performances of MoM-based estimators in presence of outliers.

Ruining 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 $h\colon (Z, Z') \mapsto (Z - Z')^2/2$. 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 $\delta_{n^{1/4}}$. Similarly to the mean, one then has $\mathbb{E}_{\mathcal{S}_n}\left[\hat{\sigma}_n^2\right]$ 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 10c.

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 $(X_1, \dots X_n)$ and $(Y_1, \dots, Y_m)$ of X and Y, it is computed by: $\hat{U}_{n,m}^{\mathrm{MW}} = 1/(nm) \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}\{X_i \leq Y_j\}$. This example is very interesting as it highlights the importance of the bounded assumption. Indeed, to get the convergence of MoU_2, 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 $H\colon (X, Y) \mapsto \mathbb{1}\{X \leq Y\}$ is always bounded, the empirical version actually shows more resistance, and the advantage of MoU_2 is less important than in other configurations, see Figure 10d.

D.2. Additional Learning Experiments

Learning experiments have been run in order to highlight the good generalization capacity of MoU minimizers, theoretically established in Theorems 1 and 2. In this section, we consider a *ranking* problem, on tow benchmark datasets, *boston housing* and *wine quality*. We first corrupted the datasets, in a way described below, before running Algorithm 1.

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 $g\colon \mathbb{R}^p \times \mathbb{R}^p \to \{-1, 1\}$ 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 $s\colon \mathbb{R}^p \to [0, 1]$ such that $g(X, X') = 2 \cdot \mathbb{1}\{s(X) \geq s(X')\} - 1$. The scoring functions themselves are indexed by vectors $w \in \mathbb{R}^p$ such that $s(x) = \sigma(w^\top x)$, with σ the sigmoid function. ERM then consists in minimizing the disagreements among the training pairs, that writes:

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

and can be relaxed into:

$$\min_{w \in \mathbb{R}^p} \frac{2}{n(n-1)} \sum_{i<j} \max\Big(0, 1 - g_w(X, X')(Y - Y')\Big). \tag{18}$$

We have run Algorithm 1 with criterion (18) on two datasets: *boston housing*[1], 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*[2], that gathers $1,600$ 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 *boston* and *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 λ. This way, one has:

$$\begin{aligned} g_{\hat{w}_{\text{sane}}}(X, X_{\text{outlier}})(Y - Y_{\text{outlier}}) &\approx \big(\sigma(\hat{w}_{\text{sane}}^\top X) - \sigma(\hat{w}_{\text{sane}}^\top X_{\text{outlier}})\big)\,(Y - \lambda), \\ &= \big(\sigma(\hat{w}_{\text{sane}}^\top X) - \sigma(-\lambda \|\hat{w}_{\text{sane}}\|^2)\big)\,(Y - \lambda). \end{aligned}$$

Making λ 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 *boston*, λ was set to -500, and to 50 for *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 2. 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, empirically validating our theoretical findings.

		GD	MoU-GD
boston	sane	$\mathbf{0.35 \pm 0.04}$	0.36 ± 0.05
	cont.	0.99 ± 0.68	$\mathbf{0.36 \pm 0.05}$
wine	sane	$\mathbf{0.73 \pm 0.02}$	0.74 ± 0.02
	cont.	0.92 ± 0.11	$\mathbf{0.74 \pm 0.02}$

Table 2: Ranking test losses (avg. 50 runs).

[1] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html
[2] https://archive.ics.uci.edu/ml/datasets/wine+quality