A. Technical Details

A.1. Remark on Proposition 1's Bound

We point out that K and B are chosen so that (8) and (9) are both bounded by $\delta/2$. Given τ, setting $B = \lfloor 2(\tau - \lambda)^2 n/\log(2/\delta)\rfloor$ for $\lambda \in]0, \tau[$ yields a minimal constant for $\lambda = \tau/3$. Interestingly, B involves a *floor* function, even if it is not constrained by K. An interpretation is that building large blocks increases the risk of selecting extreme values, and thus of deteriorating the performance.

A.2. Variance Computations

A.2.1. MoRM

By virtue of Chebyshev's inequality, one gets:

$$p^{\varepsilon} \leq \frac{\mathbb{E}\left[(\bar{\theta}_1 - \theta)^2\right]}{\varepsilon^2} = \frac{\mathbb{E}_{\mathcal{S}_n}\left[\mathbb{E}\left[(\bar{\theta}_1 - \theta)^2 \mid \mathcal{S}_n\right]\right]}{\varepsilon^2}.$$

Observing that $\mathbb{E}[\bar{\theta}_1 | \mathcal{S}_n] = \hat{\theta}_n$ and that

$$\mathbb{E}\left[(\bar{\theta}_1 - \theta)^2 | \mathcal{S}_n\right] = \mathrm{Var}\left(\bar{\theta}_1 \mid \mathcal{S}_n\right) + (\hat{\theta}_n - \theta)^2 = (\hat{\theta}_n - \theta)^2 + \frac{1}{B}\frac{n-B}{n}\hat{\sigma}_n^2,$$

where $\hat{\sigma}_n^2 = (1/(n-1))\sum_{i=1}^n (Z_i - \hat{\theta}_n)^2$, we deduce that

$$p^{\varepsilon} \leq \left(\frac{1}{n} + \frac{n-B}{nB}\right)\frac{\sigma^2}{\varepsilon^2} = \frac{\sigma^2}{B\epsilon^2}.$$

□

A.2.2. MoRU

Observe first that

$$\mathrm{Var}\left(\bar{U}_1(h)\right) = \mathbb{E}\left[\mathrm{Var}(\bar{U}_1(h) \mid \mathcal{S}_n)\right] + \mathrm{Var}\left(\mathbb{E}\left[\bar{U}_1(h) \mid \mathcal{S}_n\right]\right). \tag{14}$$

Recall that $\mathbb{E}[\bar{U}_1(h) \mid \mathcal{S}_n] = U_n(h)$, so that

$$\mathrm{Var}\left(\mathbb{E}\left[\bar{U}_1(h) \mid \mathcal{S}_n\right]\right) = \frac{4\sigma_1^2(h)}{n} + \frac{2\sigma_2^2(h)}{n(n-1)}. \tag{15}$$

In addition, we have, for $B \geq 4$,

$$\mathrm{Var}(\bar{U}_1(h) \mid \mathcal{S}_n) = \frac{4}{B^2(B-1)^2}\sum_{i<j} h^2(X_i, X_j)\mathrm{Var}(\epsilon_{1,i}\epsilon_{1,j}) \quad + \sum_{\substack{i<j,\ k<l \\ (i,j)\neq(k,l)}} \mathrm{Cov}(\epsilon_{1,i}\epsilon_{1,j},\ \epsilon_{1,l}\epsilon_{1,k}) h(X_i, X_j) h(X_k, X_l).$$

Let $i \neq j$, one may check that

$$\mathrm{Var}(\epsilon_{1,i}\epsilon_{1,j}) = \frac{B(B-1)(n-B)(n+B-1)}{n^2(n-1)^2}.$$

And, for any $k \neq l$, we have

$$\mathrm{Cov}(\epsilon_{1,i}\epsilon_{1,j},\ \epsilon_{1,l}\epsilon_{1,k}) = -\frac{B(B-1)}{n(n-1)}\frac{(n-B)(4nB - 6n - 6B + 6)}{n(n-1)(n-2)(n-3)}$$

when $\{i, j\} \cap \{k, l\} = \emptyset$, as well as

$$\mathrm{Cov}(\epsilon_{1,i}\epsilon_{1,j},\ \epsilon_{1,i}\epsilon_{1,k}) = \frac{B(B-1)}{n(n-1)}\frac{(n-B)(nB - 2n - 2B + 2)}{n(n-1)(n-2)}$$

when $k \neq j$ and $k \neq i$. Hence, observing that $\mathbb{E}[h(X_1, X_2)h(X_1, X_3)] = \sigma_1^2(h) + \theta^2(h)$, we obtain:

$$\begin{aligned}\mathbb{E}\left[\mathrm{Var}(\bar{U}_1(h) \mid \mathcal{S}_n)\right] = \; & \frac{2(n-B)(n+B-1)}{n(n-1)B(B-1)}\left(\sigma^2(h) + \theta^2(h)\right) \;-\; \frac{(n-B)(4nB-6n-6B+6)}{n(n-1)B(B-1)}\theta^2(h) \\ & + \; \frac{4(n-B)(nB-2n-2B+2)}{n(n-1)B(B-1)}(\sigma_1^2(h) + \theta^2(h)). \end{aligned} \tag{16}$$

Combining (14), (15) and (16), we get:

$$\begin{aligned}\mathrm{Var}\left(\bar{U}_1(h)\right) = \; & \frac{4\sigma_1^2(h)}{n} \;+\; \frac{2\sigma_2^2(h)}{n(n-1)} \;+\; \frac{2(n-B)(n+B-1)}{n(n-1)B(B-1)}\left(2\sigma_1^2(h) + \sigma_2^2 + \theta^2(h)\right) \\ & - \; \frac{(n-B)(4nB-6n-6B+6)}{n(n-1)B(B-1)}\theta^2(h) \;+\; \frac{4(n-B)(nB-2n-2B+2)}{n(n-1)B(B-1)}(\sigma_1^2(h) + \theta^2(h)), \\ \mathrm{Var}\left(\bar{U}_1(h)\right) = \; & \frac{4\sigma_1^2(h)}{B} + \frac{2\sigma_2^2(h)}{B(B-1)}. \end{aligned}$$

Chebyshev inequality permits to conclude. □

A.3. Remark on the Term $\log(2/\delta)$ in the Rate Bounds

In all results related to randomized versions (namely Proposition 1 and Proposition 3), the term $\log(2/\delta)$ appears, instead of $\log(1/\delta)$. We point out that this limitation can be easily overcome by means of a more careful analysis in (5) and (13). Indeed, K and B have been chosen so that both exponential terms are equal to $\delta/2$, but one could of course consider splitting the two terms into $(1-\kappa)\delta$ and $\kappa\delta$ for any $\kappa \in]0,1[$. This, way, choosing

$$K = \left\lceil \log\left(\frac{1}{(1-\kappa)\delta}\right) / (2(1/2-\tau)^2) \right\rceil \text{ and } B = \left\lfloor 8\tau^2 n / (9 \log\left(\frac{1}{\kappa\delta}\right)) \right\rfloor$$

leads to $\log(1/\kappa\delta)$ instead.

A.4. Extension to Generalized U-statistics

As noticed in Remark 5, Propositions 2 and 3 have been established for U-statistics of degree 2, but remain valid for generalized ones. For clarity, we recall the definition of generalized U-statistics. An excellent account of properties and asymptotic theory of U-statistics can be found in Lee (1990).

Definition 1. *Let $T \geq 1$ and $(d_1, \ldots, d_T) \in \mathbb{N}^{*T}$. Let $\boldsymbol{X}_{\{1,\ldots,n_t\}} = (X_1^{(t)}, \ldots, X_{n_t}^{(t)})$, $1 \leq t \leq T$, be T independent samples of sizes $n_t \geq d_t$ and composed of i.i.d. random variables taking their values in some measurable spaces $\mathcal{X}_t$ with distribution $F_t(dx)$ respectively. Let $H : \mathcal{X}_1^{d_1} \times \ldots \times \mathcal{X}_T^{d_T} \to \mathbb{R}$ be a measurable function, square integrable with respect to the probability distribution $\mu = F_1^{\otimes d_1} \otimes \ldots \otimes F_T^{\otimes d_T}$. Assume in addition (without loss of generality) that $H(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)})$ is symmetric within each block of argument $\boldsymbol{x}^{(t)}$ valued in $\mathcal{X}_t^{d_t}$, $1 \leq t \leq T$. The generalized (or T-sample) U-statistic of degrees $(d_1, \ldots, d_T)$ with kernel H is then defined as*

$$U_{\boldsymbol{n}}(H) = \frac{1}{\Pi_{t=1}^T \binom{n_t}{d_t}} \sum_{I_1} \ldots \sum_{I_T} H(\boldsymbol{X}_{I_1}^{(1)}, \ldots, \boldsymbol{X}_{I_T}^{(T)}),$$

where the symbol $\sum_{I_t}$ refers to the summation over all $\binom{n_t}{d_t}$ subsets $\boldsymbol{X}_{I_t}^{(t)} = (X_{i_1}^{(t)}, \ldots, X_{i_{d_t}}^{(t)})$ related to a set I_t of d_t indexes $1 \leq i_1 < \ldots < i_{d_t} \leq n_t$ and $\boldsymbol{n} = (n_1, \ldots, n_T)$.

Within this framework, we aim at estimating $\theta(h) = \mathbb{E}\left[H(X_1^{(1)}, \ldots, X_{d_1}^{(1)}, \ldots, X_1^{(T)}, \ldots, X_{d_T}^{(T)})\right]$, and an analog of the standard MoM estimator for generalized U-statistics can be defined as follows.

Definition 2. *With the notation introduced in Definition 1, let $1 \leq K \leq \min_t n_t/(d_t+1)$. Partition each sample $\boldsymbol{X}^{(t)}$ into K blocks $B_1^{(t)}, \ldots, B_K^{(t)}$ of sizes $\lfloor n_t/K \rfloor$. Compute $\hat{\theta}_k$ the complete U-statistics based on $B_k^{(1)}, \ldots, B_k^{(T)}$ for $1 \leq k \leq K$. The Median-of-Generalized-U-statistics if then given by $\hat{\theta}_{MoGU} = median(\hat{\theta}_1, \ldots, \hat{\theta}_K)$.*

Please note that this estimator is very different from that considered in Minsker & Wei (2018). Robust versions of U-statistics have already been considered in the literature (for the purpose of covariance estimation, rather than the design of statistical learning methods), from a completely different angle however. The U-statistic is viewed as a M-estimator minimizing a criterion involving the quadratic loss, and the proposed estimator is the M-estimator solving the same criterion except that a different loss function is used. This loss function is designed to induce robustness, while being close enough to the square loss to derive guarantees.

The definition above is closer to that given in Joly & Lugosi (2016). Indeed, in the particular setting of a 1-sample U-statistic of degree d_1, Definition 2 coincides with the *diagonal blocks* estimate mentioned on page 5 of Joly & Lugosi (2016). However, as noticed therein, this estimator only considers a small fraction of possible d_t-tuples, namely those whose items are all in the block $B_k^{(t)}$. In order to overcome this limitations, an alternative strategy involving decoupled U-statistics is adopted in Joly & Lugosi (2016). The approach pursued here is rather to consider randomized U-statistics, which is another way to introduce variability into the tuples considered to build the estimator.

Proposition 4. *Using the notation of Definitions 1 and 2, let $n_{min} = \min_t n_t$, and $\underline{n}, \underline{d}$ such that $\underline{n}/\underline{d} = \min_t n_t/d_t$. Let $\delta \in [e^{1-2\underline{n}/9\underline{d}}, 1[$. Choosing $K = \lceil 9/2 \log(1/\delta) \rceil$, we have with probability at least $1-\delta$:*

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

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

Proposition 5. *Now let $\hat{\theta}_{MoRGU}$ (Median-of-Randomized-Generalized-U-statistics) be the estimator of $\theta(h)$ such that the blocks $B_k^{(t)}$ are no longer partitions of the samples $\boldsymbol{X}^{(t)}$, but rather drawn from SWoR. For any $\tau \in]0, 1/2[$, for any $\delta \in [2e^{-8\tau^2\underline{n}/9\underline{d}}, 1[$, choosing $K = \lceil \log(2/\delta)/(2(1/2-\tau)^2) \rceil$ and $B_t = \lfloor 8\tau^2 n_t/(9\log(2/\delta)) \rfloor$, it holds with probability larger than $1-\delta$:*

$$\left| \hat{\theta}_{MoRGU} - \theta(h) \right| \leq \sqrt{\frac{C_1(\tau) \log \frac{2}{\delta}}{n_{min}} + \frac{C_2(\tau) \log^2(\frac{2}{\delta})}{n_{min} \left(8n_{min} - 9 \log \frac{2}{\delta} \right)}},$$

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

Proofs. The proofs are analogous to that of Propositions 2 and 3, except that concentration results for generalized U-statistics are used (see *e.g.* Hoeffding (1963)).

A.5. Proof of Theorem 1 (sketch of)

The proof follows the path of Theorem 2.11's proof in Lugosi & Mendelson (2016), with adjustments every time U-statistics are involved instead of standard means. The first one deals with the constants involved in the propositions.

Definition 3. *Let $\lambda_{\mathbb{Q}}(\kappa, \eta, h)$ and $\lambda_{\mathbb{M}}(\kappa, \eta, h)$ defined as in Lugosi & Mendelson (2016) (see Definitions 2.2 and 2.3 therein).*

Definition 4. *A difference however occurs on $r_E(\kappa, h)$ and $\bar{r}_{\mathbb{M}}(\kappa, h)$. Indeed, let*

$$r_E(\kappa, h) = \inf \left\{ r : \mathbb{E} \sup_{u \in \mathcal{F}_{h,r}} \left| \sqrt{\frac{2}{B(B-1)}} \sum_{i<j} \sigma_{i,j} \, u(X_i, X_j) \leq \kappa \sqrt{\frac{B(B-1)}{2}} r \right| \right\},$$

and

$$\bar{r}_{\mathbb{M}}(\kappa, h) = \inf \left\{ r : \mathbb{E} \sup_{u \in \mathcal{F}_{h,r}} \left| \sqrt{\frac{2}{B(B-1)}} \sum_{i<j} \sigma_{i,j} \, u(X_i, X_j) \cdot h(X_i, X_j) \leq \kappa \sqrt{\frac{B(B-1)}{2}} r^2 \right| \right\}.$$

While the difference on $r_E(\kappa, h)$ is only due to the double summation, the change in $\bar{r}_{\mathbb{M}}(\kappa, h)$ also comprises the removal of Y, as the general framework for pairwise learning does not involve any label. As a consequence, any Y in the older definitions is replaced by 0. Moreover, one can assess a 0 noise W such that $\|W\|_{L_2} = 0 \leq 1 = \sigma$. This way, all σ's encountered in older definitions and propositions can be replaced by 1.

Lemma 1. *For every $q > 2$ and $L \geq 1$, there are constants B and κ_0 that depend only on q and L for which the following holds. If $\|h\|_{L_q} \leq L\|h\|_{L_2}$ and $X_1, \ldots, X_B$ are independent copies of X, then*

$$\mathbb{P}\left\{\frac{2}{B(B-1)} \sum_{1 \leq i < j \leq B} |h(X_i, X_j)| \geq \kappa_0 \|h\|_{L_2}\right\} \geq 0.9.$$

Proof. The proof is analogous to that of Lemma 3.4 in Mendelson (2017), except that a version of Berry-Esseen theorem for U-statistics (Callaert & Janssen, 1978) is used instead of the standard one. □

Lemma 2. *For every $q > 2$ and $L \geq 1$, there is a constant κ_1 that depends only on q and L for which the following holds. If $X_1, \ldots, X_B$ are independent copies of X, then*

$$\mathbb{P}\left\{\frac{2}{B(B-1)} \sum_{1 \leq i < j \leq B} |h(X_i, X_j)| \leq \kappa_1 \|h\|_{L_2}\right\} \geq 0.9.$$

Proof. As $\left\{\frac{2}{B(B-1)} \sum_{i<j} |h(X_i, X_j)| \geq \kappa_1 \|h\|_{L_2}\right\} \subset \{\exists i < j,\ |h(X_i, X_j)| \geq \kappa_1 \|h\|_{L_2}\}$, Chebyshev inequality gives

$$\mathbb{P}\left\{\frac{2}{B(B-1)} \sum_{1 \leq i < j \leq B} |h(X_i, X_j)| \geq \kappa_1 \|h\|_{L_2}\right\} \leq \frac{B(B-1)}{2} \mathbb{P}\left\{|h(X_i, X_j)| \geq \kappa_1 \|h\|_{L_2}\right\} \leq \frac{B(B-1)}{2\kappa_1^2}.$$

Since B only depends on q and L (see proof of Lemma 1), so does κ_1. □

Proposition 6. *There are constants $\kappa, \eta, B, c > 0$ and $0 < \alpha < 1 < \beta$ depending only on q and L for which the following holds. For a fixed $f^* \in \mathcal{F}$, let $r^* = \max\{\lambda_{\mathbb{Q}}(\kappa, \eta, H_{f^*}), r_E(\kappa, H_{f^*})\}$. For any $r \geq 2r^*$, with probability at least $1 - 2\exp(-cn)$, $\forall\, H_f \in \mathcal{H}_{\mathcal{F}}$,*

- *If $\Phi_{\mathcal{S}}(f, f^*) \geq \beta r$, then $\beta^{-1}\Phi_{\mathcal{S}}(f, f^*) \leq \|H_f - H_{f^*}\|_{L_2} \leq \alpha^{-1}\Phi_{\mathcal{S}}(f, f^*)$.*
- *If $\Phi_{\mathcal{S}}(f, f^*) \leq \beta r$, then $\|H_f - H_{f^*}\|_{L_2} \leq (\beta/\alpha) r$.*

Proof. Using Lemma 1 and Lemma 2 with $h = H_f - H_{f^*}$, together with the union bound, it holds that for every block $\mathcal{B}_k$ one has with probability at least 0.8

$$\kappa_0 \|H_f - H_{f^*}\|_{L_2} \leq \bar{U}_k(|H_f - H_{f^*}|) \leq \kappa_1 \|H_f - H_{f^*}\|_{L_2}. \tag{17}$$

Denoting by I_k the indicator of this event, and by $\bar{I}_k$ its complementary, we have $\mathbb{E}[\bar{I}_k] \leq 0.2$. Moreover,

$$\mathbb{P}\left\{\sum_{k=1}^{K} I_k \geq 0.7K\right\} = 1 - \mathbb{P}\left\{\frac{1}{K}\sum_{k=1}^{K} \bar{I}_k \geq 0.3\right\}.$$

When a MoU estimate is used, the I_k are independent, since built on disjoint blocks, and the concentration of Binomial random variables allows to finish. But interestingly, when a MoCU is used, it is straightforward to see that the last term is exactly the same quantity as the one involved in (10). The same method can thus be used since an upper bound of $\mathbb{E}[\bar{I}_k]$ is already available. Precisely, choosing $\tau = 0.25 < 0.3$ and recalling $B = \lfloor n/K \rfloor$, it holds

$$\mathbb{P}\left\{\frac{1}{K}\sum_{k=1}^{K} \bar{I}_k \geq 0.3\right\} \leq 2\exp\left(-2(0.05)^2 K\right).$$

So the number of blocks which satisfy (17) is larger than $0.7K$ with probability at least $1 - 2\exp(-c_1 K)$ for some positive constant c_1. The rest of the proof is similar to that of Proposition 3.2 in Lugosi & Mendelson (2016). □

Proposition 7. *Under the assumptions of Theorem 1, and using its notation, with probability at least*

$$1 - 2\exp(-c_0 n \min\{1, r^2\}),$$

$\forall\, f \in \mathcal{F}$ if $\Phi_{\mathcal{S}}(f, f^*) \geq \beta r$ then f^* defeats f. In particular $f^* \in H$, and $\forall\, f \in H$, $\Phi_{\mathcal{S}}(f, f^*) = \|H_f - H_{f^*}\|_{L_2} \leq \beta r$.

Proof. This proof carefully follows that of Proposition 3.5 in Lugosi & Mendelson (2016) (see Section 5.1 therein), so that only changes induced by pairwise objectives are detailed here. As discussed in Definition 4, every Y can be replaced by 0, and every σ by 1. Attention must also be paid to the fact that in the context of means, m, the cardinal of each block, is also equal to $\binom{m}{1}$, the number of possible 1-combinations. In our notation, it thus may sometimes be identified to B, the cardinal of the blocks, and sometimes to $\frac{B(B-1)}{2}$, the number of pairs per block.

Proof of pairwise Lemma 5.1 First, one may rewrite

$$\mathbb{Q}_{f,g} = \frac{2}{B(B-1)} \sum_{i<j} \left(H_f(X_i, X_j) - H_g(X_i, X_j)\right)^2,$$

$$\mathbb{M}_{f,g} = \frac{4}{B(B-1)} \sum_{i<j} \left(H_f(X_i, X_j) - H_g(X_i, X_j)\right) \cdot H_g(X_i, X_j),$$

and

$$R_k(u,t) = \left|\{(i,j) \in \mathcal{B}_k^2 : i < j, |u(X_i, X_j)| \geq t\}\right| = \sum_{i<j \in \mathcal{B}_k^2} \mathbb{1}\{|u(X_i, X_j) \geq t|\}.$$

Since all pairs are not independent, even if the X_i's are, one cannot use directly the proposed method. Instead, the Hoeffding inequality for U-statistics gives that the probability of each $R_k(H_f - H_{f^*}, \kappa_0 r)$ to be greater than $\frac{B(B-1)\rho_0}{4}$ is greater than $1 - \exp(-\frac{B\rho_0^2}{4})$. For τ small enough, we still have that this probability is greater than $1 - \tau/12$. Aggregating the Bernoulli may be done in two ways. If we deal with a MoU estimate, the independence between blocks leads to the same conclusion. If a MoRU estimate is used instead, the remark made for Proposition 6's proof is again valid, and one can conclude.

The next difficulty arises with the bounded differences inequality for Ψ. If a MoU estimate is used, changing one sample X_i' only affects one block, and generates a $1/K$ difference at most, exactly like with MoM, so that the bound holds the same way. On the contrary, if a MoRU is used, there is no guarantee that the replaced sample contaminates all K blocks. The analysis of the MoRU behavior in that case is a bit trickier, and we restrict ourselves to MoU estimates for the matches.

The end of the proof uses a standard symmetrization argument. This kind of arguments still apply to U-statistics (see *e.g.* p.150 of Peña & Giné (1999)), and the proof is completed in the pairwise setting.

Proof of pairwise Lemma 5.2

$$\begin{aligned}
\mathbb{P}\left\{\left|\frac{2}{B(B-1)}\sum_{i<j} U_{i,j} - \mathbb{E}U\right| \geq t\right\} &\leq \frac{2}{B(B-1)t}\mathbb{E}\left|\sum_{i<j} U_{i,j} - \mathbb{E}U\right| \\
&\leq \frac{2}{B(B-1)t}\sqrt{\mathbb{E}\left|\sum_{i<j} U_{i,j} - \mathbb{E}U\right|^2} \\
&\leq \frac{2}{B(B-1)t}\sqrt{\sum_{i<j,\ k<l} \mathbb{E}[U_{i,j}U_{j,k}] - (\mathbb{E}U)^2} \\
&\leq \frac{2}{B(B-1)t}\sqrt{\frac{B(B-1)}{2}\left(\mathbb{E}[U^2] - (\mathbb{E}U)^2\right) + B(B-1)(B-2)\sigma_1^2} \\
&\leq \frac{\sqrt{2(B-2)}}{\sqrt{B(B-1)}t}\|U\|_{L_2}
\end{aligned}$$

$$\mathbb{P}\left\{\left|\frac{2}{B(B-1)}\sum_{i<j} U_{i,j} - \mathbb{E}U\right| \geq t\right\} \leq \frac{\sqrt{2}}{\sqrt{B}t}\|U\|_{L_2}$$

After that, every case needing a pairwise investigation has already been treated earlier in the section: Binomial concentration, bounded differences inequality, symmetrization arguments. So is Proposition 7 proved.

Theorem 1's proof. Proposition 6 and Proposition 7 gives that if any $f \in \mathcal{F}$ wins all its matches, then with probability at least $1 - 2\exp(-c_0 n \min\{1, r^2\})$ $\|H_f - H_{f^*}\|_{L_2} \leq cr$. Hence it also holds with the same probability:

$$\mathcal{R}(f) - \mathcal{R}(f^*) = \|H_f\|_{L_2} - \|H_{f^*}\|_{L_2} \leq \|H_f - H_{f^*}\|_{L_2} \leq cr.$$

□

Although this extension to the pairwise learning framework deals with any loss ℓ, it is important to notice that the extension of Theorem 2.11 in Lugosi & Mendelson (2016) is applied to H_f, and not f directly. H_f is penalized via the quadratic loss (as in Lugosi & Mendelson (2016)), so that the Theorem apply, up to technicalities induced by the U-statistics. Doing so, one achieved a control on $\|H_f - H_f^*\|_{L_2}$ (as in Lugosi & Mendelson (2016)), which is equal to $\|H_f - H_{f^*}\|_{L_2}$ thanks to the remark stated in the first paragraph of Subsection 3.3. This quantity happens to be greater than the excess risk of f, hence the conclusion. Formally, the tournament procedure outputs a $\hat{H}_f$, and one has to recover the $\hat{f}$ such that $\hat{H}_f = H_{\hat{f}}$. Knowing the dependence between f and H_f, and with the ability to evaluate $\hat{H}_f$, which is known, on any pair, this last step should not be too difficult.

About the extension to pairwise learning, one should keep in mind that the general framework does not involve any target Y. Instead, one seeks directly to minimize $\ell(f, X, X') = \sqrt{\ell(f, X, X')}^2 = (\sqrt{\ell(f, X, X')} - 0)^2 = (H_f(X, X') - 0)^2$. We recover the setting of Theorem 2.11 in Lugosi & Mendelson (2016): quadratic loss, with $Y = 0$, for the decision function H_f. The only novelty to address here is the fact that H_f depends on two random variables X and X'. And this is precisely what has been done in this subsection.

B. More Numerical Results

B.1. MoRM Estimation Results

Table 3. Quadratic Risks for the Mean Estimation, $\delta = 0.001$

	NORMAL $(0, 1)$	STUDENT (3)	LOG-NORMAL $(0, 1)$	PARETO (3)
MOM	0.00149 ± 0.00218	0.00410 ± 0.00584	0.00697 ± 0.00948	$\mathbf{1.02036} \pm \mathbf{0.06115}$
MORM$_{1/6,\text{ SWOR}}$	0.01366 ± 0.01888	0.02947 ± 0.04452	0.06210 ± 0.07876	1.12256 ± 0.14970
MORM$_{1/6,\text{ MC}}$	0.01370 ± 0.01906	0.02917 ± 0.04355	0.06167 ± 0.07143	1.13058 ± 0.14880
MORM$_{3/10,\text{ SWOR}}$	0.00255 ± 0.00361	0.00602 ± 0.00868	0.01241 ± 0.01610	1.05458 ± 0.07041
MORM$_{3/10,\text{ MC}}$	0.00264 ± 0.00372	0.00622 ± 0.00895	0.01283 ± 0.01650	1.05625 ± 0.07298
MORM$_{9/20,\text{ SWOR}}$	0.00105 ± 0.00148	$\mathbf{0.00264} \pm \mathbf{0.00372}$	$\mathbf{0.00497} \pm \mathbf{0.00668}$	1.02802 ± 0.04903
MORM$_{9/20,\text{ MC}}$	$\mathbf{0.00105} \pm \mathbf{0.00146}$	0.00265 ± 0.00374	0.00499 ± 0.00673	1.02985 ± 0.04880

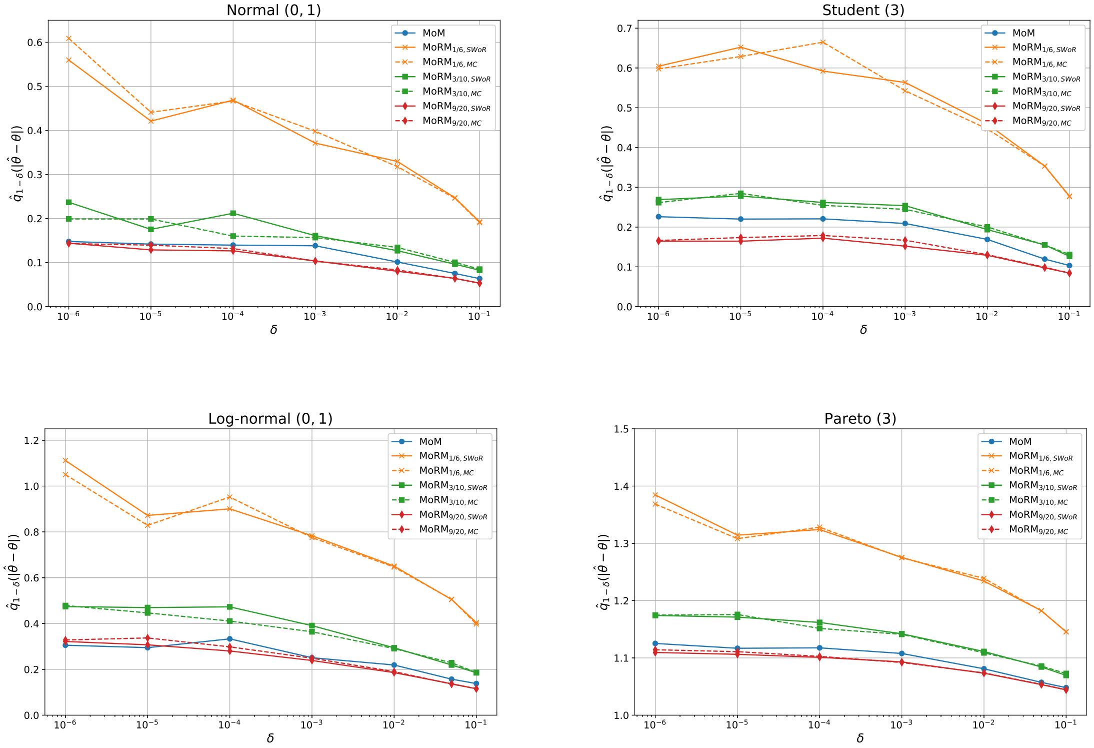

Figure 1. Empirical Quantiles for the Different Mean Estimators on 4 Laws

The empirical quantiles confirm the quadratic risks results: the τ parameter is crucial, making MoRM the worst or the best estimate depending on its value. The sampling scheme does not affect to much the performance, even if the MC scenario is much more complex to analyze theoretically.

B.2. MoRU Estimation Results

Table 4. Quadratic Risks for the Variance Estimation, $\delta = 0.001$

	Normal $(0, 1)$	Student (3)	Log-normal $(0, 1)$	Pareto (3)
MoU$_{1/2;\,1/2}$	0.00409 ± 0.00579	1.72618 ± 28.3563	2.61283 ± 23.5001	1.35748 ± 36.7998
MoU$_{\text{Partition}}$	0.00324 ± 0.00448	$\mathbf{0.38242} \pm \mathbf{0.31934}$	$\mathbf{1.62258} \pm \mathbf{1.41839}$	$\mathbf{0.09300} \pm \mathbf{0.05650}$
MoRU$_{\text{SWoR}}$	0.00504 ± 0.00705	0.51202 ± 3.88291	2.01399 ± 4.85311	0.09703 ± 0.07116
MoIU$_{1/6,\ \text{SWoR}}$	0.00206 ± 0.00285	1.78161 ± 34.7216	2.50529 ± 21.8989	1.37800 ± 40.1308
MoIU$_{1/6,\ \text{MC}}$	$\mathbf{0.00205} \pm \mathbf{0.00281}$	1.65481 ± 26.2157	2.61701 ± 24.7918	1.50578 ± 42.9135
MoIU$_{3/10,\ \text{SWoR}}$	0.00216 ± 0.00301	1.13887 ± 16.9511	2.07136 ± 14.8312	0.85041 ± 21.9916
MoIU$_{3/10,\ \text{MC}}$	0.00211 ± 0.00288	1.22402 ± 17.4715	2.16590 ± 15.2378	0.89035 ± 22.2866

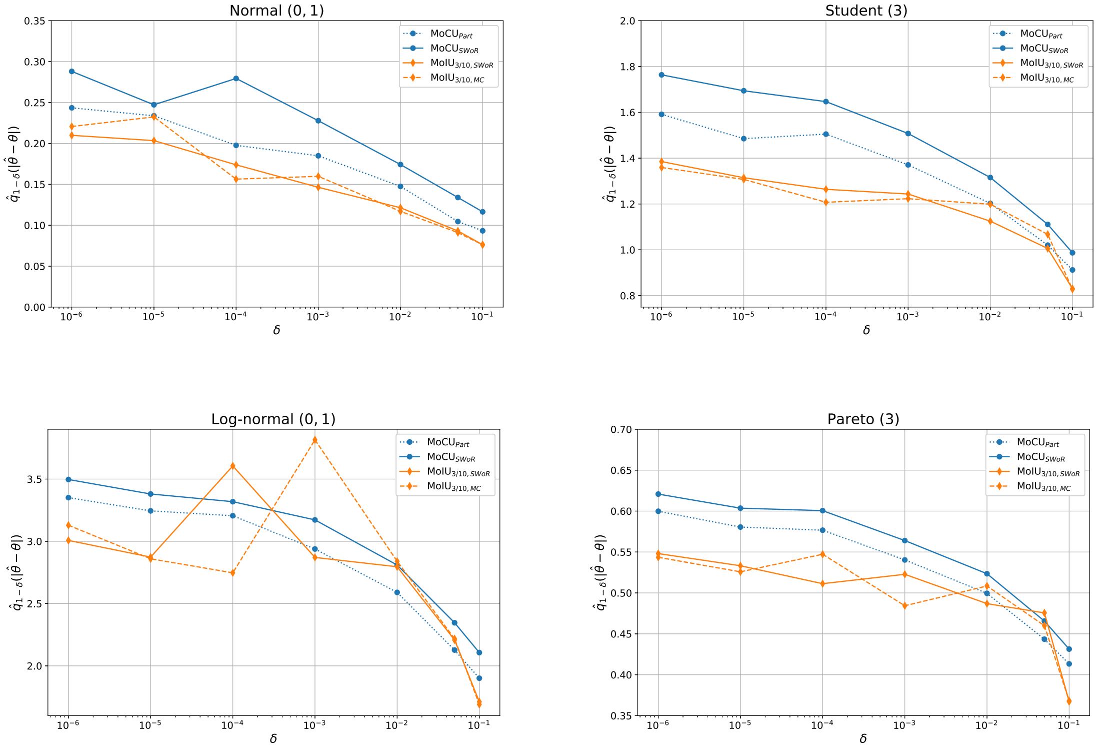

Figure 2. Empirical Quantiles for the Different Variance Estimators on 4 Laws

The partitioning MoU seems to outperform every other estimate. One explanation can be that an extreme value may *corrupt* only one block within this method, whereas randomized versions can suffer from it in several blocks.

B.3. MoRU Learning Results: a Metric Learning Application

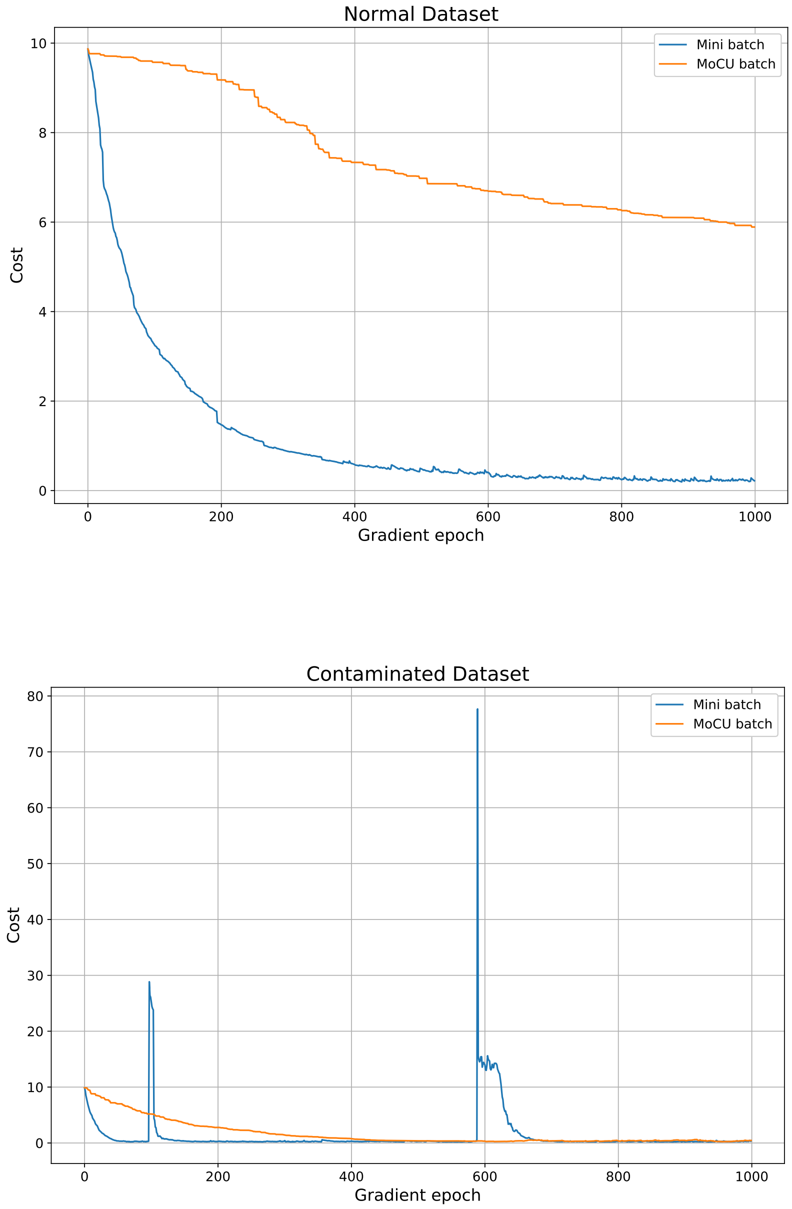

Figure 3. Gradient Descent Convergences for Normal and Contaminated Datasets

In this experiment, we try to learn from points that are known to be close or not, a distance that fits them, over the set of all possible Mahalanobis distances, *i.e.* $d(x, y) = \sqrt{(x-y)^\top M(x-y)}$ for some positive definite matrix M. It is done through mini-batch gradient descent over the parameter M. On the normal iris dataset (top), we see that standard mini-batches perform well, while MoRU mini-batches induce a much slower convergence. But in the spirit of Lecué & Lerasle (2017), the experiment is also run on a (artificially) corrupted dataset (bottom). This highlights how well can the MoM-like estimators behave in the presence of outliers. Indeed, although the convergence with the MoRU mini-batches remains slower than with the standard ones on the normal regime, they avoid peaks, presumably caused by the presence of one (or more) outlier in the mini-batch. This makes it a very interesting alternative in the context of highly corrupted data. Experiments run on the same dataset for clustering purposes show the same behavior.

C. Alternative Sampling Schemes - Possible Extensions

As pointed out in Remark 1 and in the discussion in the end of Section 3, the approaches investigated in the present paper could be implemented with sampling procedures different from the SRSWoR scheme. It is the purpose of this section to review possible alternatives and discuss the technical difficulties inherent to the study of their performance.

C.1. MoRM using a Sampling with Replacement Scheme

Rather than forming $K \geq 1$ data blocks of size $B \leq n$ from the original sample $\mathcal{S}_n$ by means of a SRSWoR scheme, one could use a Monte-Carlo procedure and draw independently K times an arbitrary number B of observations with replacement. In this case, each data block $\mathcal{B}_k$ is characterized by a random vector $\boldsymbol{e}_k = (e_{k,1},\ \dots,\ e_{k,B})$ independent from $\mathcal{S}_n$, where, for each draw $b \in \{1,\ \dots,\ B\}$, $e_{k,b} = (e_{k,b}(1),\ \dots,\ e_{k,b}(n))$ is a multinomial random vector in $\{0,1\}^n$ indicating the index of the observation randomly selected: for any $i \in \{1,\ \dots,\ n\}$, $e_{k,b}(i) = 1$ if i has been chosen, $e_{k,b}(i) = 0$ otherwise and $\mathbb{P}\{e_{k,b}(i) = 1\} = 1/n$ (notice also that $\sum_{i=1}^n e_{k,b}(i) = 1$ with probability one). In this case, the empirical mean based on block $\mathcal{B}_k$ can be written as

$$\tilde{\theta}_k = \frac{1}{B}\sum_{b=1}^{B}\langle e_{k,b},\ \mathbf{Z}_n\rangle,$$

where $\mathbf{Z}_n = (Z_1,\ \dots,\ Z_n)$ and $\langle .,\ .\rangle$ is the usual Euclidean scalar product on $\mathbb{R}^n$. Conditioned upon $\mathcal{S}_n$, the $\tilde{\theta}_k$'s are i.i.d. and we have $\mathbb{E}[\tilde{\theta}_1 \mid \mathcal{S}_n] = \hat{\theta}_n$, as well as

$$\text{Var}(\tilde{\theta}_1|\mathcal{S}_n) = \frac{1}{B}\left(\frac{1}{n}\sum_{j=1}^{n} Z_j^2 - \hat{\theta}_n^2\right).$$

The corresponding variant of the MoM estimator is

$$\tilde{\theta}_{\text{MoRM}} = \text{median}\left(\tilde{\theta}_1,\ \dots,\ \tilde{\theta}_K\right).$$

Observe that this estimation procedure offers a greater flexibility, insofar as both K and B can be arbitrarily chosen. In addition, the variance of the block estimators $\tilde{\theta}_k$:

$$\text{Var}(\tilde{\theta}_1) = \text{Var}\left(\mathbb{E}[\tilde{\theta}_1 \mid \mathcal{S}_n]\right) + \mathbb{E}\left[\text{Var}\left(\tilde{\theta}_1 \mid \mathcal{S}_n\right)\right] = \left(\frac{1}{n} + \frac{1}{B}\left(1 - \frac{1}{n}\right)\right)\sigma^2.$$

It is comparable to that of the $\bar{\theta}_k$'s for the same block size $B \leq n$, although always larger: $\text{Var}(\tilde{\theta}_1) - \text{Var}(\bar{\theta}_1) = (1 - 1/B)\sigma^2/n \geq 0$. However, investigating the accuracy of $\tilde{\theta}_{\text{MoRM}}$ is challenging, due to the fact that it is far from straightforward to study the concentration properties of the random quantity

$$\tilde{U}_n^\varepsilon = \mathbb{P}\{|\tilde{\theta}_1 - \theta| > \varepsilon \mid \mathcal{S}_n\}.$$

Even if Chebyshev's inequality yields

$$\mathbb{E}[\tilde{U}_n^\varepsilon] = p^\varepsilon \leq \left(\frac{1}{n} + \frac{1}{B}\left(1 - \frac{1}{n}\right)\right)\frac{\sigma^2}{\epsilon^2},$$

the r.v. $\tilde{U}_n^\varepsilon$ is a complex functional of the original data $\mathcal{S}_n$, due to the possible multiple occurrence of a given observation in a single sample obtained through sampling with replacement. In particular, in contrast to U_n^ε, it is not a U-statistic of degree B. Viewing it as a function of the n i.i.d. random variables $Z_1,\ \dots,\ Z_n$ and observing that changing the value of any of them can change its value by at most $1 - (1 - 1/n)^B \leq B/n$, the bounded difference inequality (see McDiarmid (1989)) gives only: $\forall t > 0$,

$$\mathbb{P}\left\{\tilde{U}_n^\varepsilon - p^\varepsilon \geq t\right\} \leq \exp\left(-2\frac{n}{B^2}t^2\right), \tag{18}$$

while we have $\mathbb{P}\{U_n^\varepsilon - p^\varepsilon \geq t\} \leq \exp(-2(n/B)t^2)$. Hence, the bound (18) is not sharp enough to establish guarantees for the estimator $\tilde{\theta}_{\text{MoRM}}$ similar to those stated in Proposition 1.

C.2. Medians-of-Incomplete U-statistics

The computation of the U-statistic (2) is expensive in the sense that it involves the summation of $\mathcal{O}(n^2)$ terms. The concept of *incomplete U-statistic*, see Blom (1976) permits to address the computational issue raised by the expensive calculation of the U-statistic (2), which involves the summation of $\mathcal{O}(n^2)$ terms, so as to achieve a trade-off between scalability and variance reduction. In one of its simplest forms, it consists in selecting a subsample of size $M \geq$ by *sampling with replacement* (*i.e.* Monte-Carlo scheme) in the set $\Lambda = \{(i,j) : \ 1 \leq i < j \leq n\}$, of all pairs of observations that can be formed from the original sample. Denoting by $\{(i_1, j_1), \ \ldots, \ (i_M, j_M)\} \subset \Lambda$ the subsample thus drawn, the incomplete version of the U-statistic (2) is:

$$\widetilde{U}_M(h) = \frac{1}{M} \sum_{m=1}^{M} h(X_{i_m}, X_{j_m}).$$

As can be easily shown, (C.2) is an unbiased estimator of θ. More precisely, its conditional expectation given $\mathcal{S}_n$ is equal to $U_n(h)$ and its variance is

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

Repeating independently the sampling procedure K times in order to compute incomplete U-statistics $\widetilde{U}_{M,1}(h), \ \ldots, \ \widetilde{U}_{M,K}(h)$ (conditionally independent given the original dataset $\mathcal{S}_n$), one may consider the Median of Incomplete U-statistic:

$$\tilde{\theta}_{\text{MoIU}}(h) = \text{median}\left(\widetilde{U}_{M,1}(h), \ \ldots, \ \widetilde{U}_{M,K}(h)\right).$$

Although the variance of the $\widetilde{U}_{M,k}(h)$'s are smaller than that of the $\hat{U}_k(h)$'s (and that of the $\bar{U}_k(h)$'s) for the same number of pairs involved in the computation of each estimator, the argument underlying Proposition 3's proof cannot be adapted to the present situation because the concentration properties of the random quantity

$$\widetilde{W}_n^{\varepsilon} = \mathbb{P}\left\{\left|\widetilde{U}_M(h) - \theta(h)\right| > \varepsilon \mid \mathcal{S}_n\right\},$$

which has mean

$$\widetilde{q}^{\varepsilon} = \mathbb{P}\left\{\left|\widetilde{U}_M(h) - \theta(h)\right| > \varepsilon\right\} \leq \frac{\text{Var}\left(\widetilde{U}_M(h)\right)}{\varepsilon^2} = \mathcal{O}\left(\frac{1}{M\varepsilon^2}\right),$$

are very difficult to study, due to the complexity of the data functional (C.2). Like in SM C.1, a straightforward application of the bounded difference inequality gives: $\forall t > 0$,

$$\mathbb{P}\left\{\widetilde{W}_n^{\varepsilon} - \widetilde{q}^{\varepsilon} \geq t\right\} \leq \exp\left(-2\frac{n}{M^2}t^2\right).$$

Obviously, this bound is not sharp enough to yield a bound of the same order as those in Proposition 2 and Proposition 3.

Alternative sampling schemes. Other procedures than the Monte-Carlo scheme above can of course be considered to build a subset of all pairs of observations and compute an incomplete U-statistic. Refer to Lee (1990) or to subsection 3.4 in Clémençon et al. (2016) for further details. When selecting a subsample of size $M \leq n(n-1)/2$ by *simple random sampling without replacement* (SRSWoR) in the set of all pairs of observations that can be formed from the original sample, the version of (C.2)obtained has conditional expectation and variance:

$$\begin{aligned}
\mathbb{E}\left[\widetilde{U}_M(h) \mid X_1, \ \ldots, \ X_n\right] &= U_n(h), \\
\text{Var}\left(\widetilde{U}_M(h) \mid X_1, \ \ldots, \ X_n\right) &= \frac{\binom{n}{2} - M}{\binom{n}{2}} \times \frac{\hat{V}_n^2(h)}{M},
\end{aligned}$$

where

$$\hat{V}_n^2(h) = \frac{1}{\binom{n}{2} - 1} \sum_{i<j} \left(h(X_i, X_j) - U_n(h)\right)^2.$$

Hence, its variance can be expressed as

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

In contrast with the situation investigated in subsection 3.1, the conditional expectation $\widetilde{W}_n^{\varepsilon}$ cannot be viewed as a U-statistic of degree M, insofar as the $n(n-1)/2$ variables $\{h(X_i, X_j) : \; i < j\}$ are not i.i.d. r.v.'s. Here as well, the bounded difference inequality is not sufficient to get bounds of the same order as that obtained in those in Proposition 2 and Proposition 3, jumps

$$1 - \binom{(n-1)(n-2)/2}{M} / \binom{n(n-1)/2}{M}$$

being at least of order M/n, as Stirling's formula shows.