Robust estimation of U-statistics

Emilien Joly * Gábor Lugosi†

April 17, 2015

This paper is dedicated to the memory of Evarist Giné.

Abstract

An important part of the legacy of Evarist Giné is his fundamental contributions to our understanding of U-statistics and U-processes. In this paper we discuss the estimation of the mean of multivariate functions in case of possibly heavy-tailed distributions. In such situations, reliable estimates of the mean cannot be obtained by usual U-statistics. We introduce a new estimator, based on the so-called median-of-means technique. We develop performance bounds for this new estimator that generalizes an estimate of Arcones and Giné (1993), showing that the new estimator performs, under minimal moment conditions, as well as classical U-statistics for bounded random variables. We discuss an application of this estimator to clustering.

1 Introduction

Motivated by numerous applications, the theory of U-statistics and U-processes has received considerable attention in the past decades. U-statistics appear naturally in *ranking* (Clémençon et al., 2008), *clustering* (Clémençon, 2014) and *learning on graphs* (Biau and Bleakley, 2006) or as components of higher-order terms in expansions of smooth statistics, see, for example, Robins et al. (2009). The general setting may be described as follows. Let X be a random variable taking values in some measurable space $\mathcal{X}$ and let $h : \mathcal{X}^m \to \mathbb{R}$ be a measurable function of $m \geq 2$ variables. Let P be the probability measure of X. Suppose we have access to $n \geq m$ independent random variables $X_1, \ldots, X_n$, all distributed as X. We define the U-statistics of order m and kernel h based on the sequence $\{X_i\}$ as

$$U_n(h) = \frac{(n-m)!}{n!} \sum_{(i_1,\ldots,i_m) \in I_n^m} h(X_{i_1}, \ldots, X_{i_m}) \; , \tag{1}$$

*École Normale Supérieure, Paris

†ICREA and Department of Economics and Business, Pompeu Fabra University. Supported by the Spanish Ministry of Science and Technology grant MTM2012-37195.

where

$$I_n^m = \{(i_1, \ldots, i_m) : 1 \le i_j \le n,\ i_j \ne i_k \ \text{ if } \ j \ne k\}$$

is the set of all m-tuples of different integers between 1 and n. U-statistics are unbiased estimators of the mean $m_h = \mathbb{E}h(X_1, \ldots, X_m)$ and have minimal variance among all unbiased estimators (Hoeffding, 1948). Understanding the concentration of a U-statistics around its expected value has been subject of extensive study. de la Peña and Giné (1999) provide an excellent summary but see also Giné et al. (2000) for a more recent development.

By a classical inequality of Hoeffding (1963), for a bounded kernel h, for all $\delta > 0$,

$$\mathbb{P}\left\{ |U_n(h) - m_h| > \|h\|_\infty \sqrt{\frac{\log(\frac{2}{\delta})}{2\lfloor n/m \rfloor}} \right\} \le \delta \ , \tag{2}$$

and we also have the "Bernstein-type" inequality

$$\mathbb{P}\left\{ |U_n(h) - m_h| > \sqrt{\frac{4\sigma^2 \log(\frac{2}{\delta})}{2\lfloor n/m \rfloor}} \vee \frac{4\,\|h\|_\infty \log(\frac{2}{\delta})}{6\lfloor n/m \rfloor} \right\} \le \delta \ ,$$

where $\sigma^2 = \mathrm{Var}\,(h(X_1, \ldots, X_m))$.

However, under certain degeneracy assumptions on the kernel, significantly sharper bounds have been proved. Following the exposition of de la Peña and Giné (1999), for convenience, we restrict out attention to symmetric kernels. A kernel h is *symmetric* if for all $x_1, \ldots, x_m \in \mathbb{R}$ and all permutations s,

$$h(x_1, \ldots, x_m) = h(x_{s_1}, \ldots, x_{s_m}) \ .$$

A symmetric kernel h is said to be P-*degenerate* of order $q - 1$, $1 < q \le m$, if for all $x_1, \ldots, x_{q-1} \in \mathcal{X}$,

$$\int h(x_1, \ldots, x_m) dP^{m-q+1}(x_q, \ldots, x_m) = \int h(x_1, \ldots, x_m) dP^m(x_1, \ldots, x_m)$$

and

$$(x_1, \ldots, x_q) \mapsto \int f(x_1, \ldots, x_m) dP^{m-q}(x_{q+1}, \ldots, x_m)$$

is not a constant function. In the special case of $m_h = 0$ and $q = m$ (i.e., when the kernel is $(m-1)$-degenerate, h is said to be P-*canonical.* P-canonical kernels appear naturally in the Hoeffding decomposition of a U-statistic, see de la Peña and Giné (1999).

Arcones and Giné (1993) proved the following important improvement of Hoeffing's inequalities for canonical kernels: If $h - m_h$ is a bounded, symmetric P-canonical kernel of

m variables, there exist finite positive constants c_1 and c_2 depending only on m such that for all $\delta \in (0,1)$,

$$\mathbb{P}\left\{ |U_n(h) - m_h| \geq c_1 \|h\|_\infty \left(\frac{\log(\frac{c_2}{\delta})}{n} \right)^{m/2} \right\} \leq \delta \ , \tag{3}$$

and also

$$\mathbb{P}\left\{ |U_n(h) - m_h| > \left(\frac{\sigma^2 \log(\frac{c_1}{\delta})}{c_2 n} \right)^{m/2} \vee \frac{\|h\|_\infty}{\sqrt{n}} \left(\frac{\log(\frac{c_1}{\delta})}{c_2} \right)^{(m+1)/2} \right\} \leq \delta \ . \tag{4}$$

In the special case of P-canonical kernels of order $m = 2$, (3) implies that

$$|U_n(h) - m_h| \leq \frac{c_1 \|h\|_\infty}{n} \log\left(\frac{c_2}{\delta}\right) \ , \tag{5}$$

with probability at least $1 - \delta$. Note that this rate of convergence is significantly faster than the rate $O_p(n^{-1/2})$ implied by (2).

All the results cited above require boundedness of the kernel. If the kernel is unbounded but $h(X_1, \ldots, X_m)$ has sufficiently light (e.g., sub-Gaussian) tails, then some of these results may be extended, see, for example, Giné et al. (2000). However, if $h(X_1, \ldots, X_m)$ may have a heavy-tailed distribution, exponential inequalities do not hold anymore (even in the univariate $m = 1$ case). However, even though U-statistics may have an erratic behavior in the presence of heavy tails, in this paper we show that under minimal moment conditions, one may construct estimators of m_h that satisfy exponential inequalities analogous to (2) and (3). These are the main results of the paper. In particular, in Section 2 we introduce a robust estimator of the mean m_h. Theorems 1 and 3 establish exponential inequalities for the performance of the new estimator under minimal moment assumptions. More precisely, Theorem 1 only requires that $h(X_1, \ldots, X_m)$ has a finite variance and establishes inequalities analogous to (3) for P-degenerate kernels. In Theorem 3 we further weaken the conditions and only assume that there exists $1 < p \leq 2$ such that $\mathbb{E}|h|^p < \infty$.

The next example illustrates why classical U-statistics fail under heavy-tailed distributions.

Example. Consider the special case $m = 2$, $\mathbb{E}X_1 = 0$ and $h(X_1, X_2) = X_1 X_2$. Note that this kernel is P-canonical. We define $Y_1, \ldots, Y_n$ as independent copies of $X_1, \ldots, X_n$. By decoupling inequalities for the tail of U-statistics given in Theorem 3.4.1 in de la Peña and Giné (1999) (see also Theorem 7 in the Appendix), $U_n(h)$ has a similar tail behavior to $\left(\frac{1}{n}\sum_{i=1}^n X_i\right)\left(\frac{1}{n-1}\sum_{j=1}^{n-1} Y_j\right)$. Thus, $U_n(h)$ behaves like a product of two independent empirical mean estimators of the same distribution. When the X_i are heavy tailed, the empirical mean is known to be a poor estimator of the mean. As an example, assume that

X follows an α-stable law $S(\gamma, \alpha)$ for some $\alpha \in (1, 2)$ and $\gamma > 0$. Recall that a random variable X has an α-*stable* law $S(\gamma, \alpha)$ if for all $u \in \mathbb{R}$,

$$\mathbb{E} \exp(iuX) = \exp(-\gamma^\alpha |u|^\alpha)$$

(see Zolotarev (1986), Nolan (2015)). Then it follows from the properties of α-stable distributions (summarized in Proposition 9 in the Appendix) that there exists a constant $c > 0$ depending only on α and γ such that

$$\mathbb{P}\left\{ U_n(h) \geq n^{2/\alpha - 2} \right\} \geq c \ ,$$

and therefore there is no hope to reproduce an upper bound like (5). Below we show how this problem can be dealt with by replacing the U-statistics by a more robust estimator.

Our approach is based on robust mean estimators in the univariate setting. Estimation of the mean of a possibly heavy-tailed random variable X from i.i.d. sample $X_1, \ldots, X_n$ has recently received increasing attention. Introduced by Nemirovsky and Yudin (1983), the *median-of-means* estimator takes a confidence level $\delta \in (0, 1)$ and divides the data into $V \approx \log \delta^{-1}$ blocks. For each block $k = 1, \ldots, V$, one may compute the empirical mean $\widehat{\mu}_k$ on the variables in the block. The median $\overline{\mu}$ of the $\widehat{\mu}_k$ is the so-called median-of-means estimator. A short analysis of the resulting estimator shows that

$$|\overline{\mu} - m_h| \leq c\sqrt{\mathrm{Var}\,(X)} \sqrt{\frac{\log(1/\delta)}{n}}$$

with probability at least $1 - \delta$ for a numerical constant c. For the details of the proof see Lerasle and Oliveira (2011). When the variance is infinite but a moment of order $1 < p \leq 2$ exists, the median-of means estimator is still useful, see Bubeck et al. (2013). This estimator has recently been studied in various contexts. M-estimation based on this technique has been developed by Lerasle and Oliveira (2011) and generalizations in a multivariate context have been discussed by Hsu and Sabato (2013) and Minsker (2015). A similar idea was used in Alon et al. (2002). An interesting alternative of the median-of-means estimator has been proposed by Catoni (2012).

The rest of the paper is organized as follows. In Section 2 we introduce a robust estimator of the mean m_h and present performance bounds. In particular, Section 2.1 deals with the finite variance case. Section 2.2 is dedicated to case when h has a finite p-th moment for some $1 < p < 2$ for P-degenerate kernels. Finally, in Section 3, we present an application to clustering problems.

2 Robust U-estimation

In this section we introduce a "median-of-means"-style estimator of $m_h = \mathbb{E}h(X_1, \ldots, X_m)$. To define the estimator, one divides the data into V blocks. For any m-tuple of different blocks, one may compute a (decoupled) U-statistics. Finally, one computes the median of all the obtained values. The rigorous definition is as follows.

The estimator has a parameter $V \le n$, the number of blocks. A partition $\mathcal{B} = (B_1, \ldots, B_V)$ of $\{1, \ldots, n\}$ is called *regular* if for all $K = 1, \ldots, V$,

$$\left| |B_K| - \frac{n}{V} \right| \le 1 \ .$$

For any $B_{i_1}, \ldots, B_{i_m}$ in $\mathcal{B}$, we set

$$I_{B_{i_1},\ldots,B_{i_m}} = \left\{ (k_1, \ldots, k_m) : k_j \in B_{i_j} \right\}$$

and

$$U_{B_{i_1},\ldots,B_{i_m}}(h) = \frac{1}{|B_{i_1}| \cdots |B_{i_m}|} \sum_{(k_1,\ldots,k_m) \in I_{B_{i_1},\ldots,B_{i_m}}} h(X_{k_1}, \ldots, X_{k_m}) \ .$$

For any integer N and any vector $(a_1, \ldots, a_N) \in \mathbb{R}^N$, we define the median $\mathrm{Med}(a_1, \ldots, a_N)$ as any number b such that

$$\left| \{ i \le N \ : \ a_i \le b \} \right| \ge \frac{N}{2} \quad \text{and} \quad \left| \{ i \le N \ : \ a_i \ge b \} \right| \ge \frac{N}{2} \ .$$

Finally, we define the robust estimator:

$$\overline{U}_{\mathcal{B}}(h) = \mathrm{Med}\{ U_{B_{i_1},\ldots,B_{i_m}}(h) : i_j \in \{1, \ldots, V\}, 1 \le i_1 < \ldots < i_m \le V \} \ . \tag{6}$$

Note that, mostly in order to simplify notation, we only take those values of $U_{B_{i_1},\ldots,B_{i_m}}(h)$ into account that correspond to distinct indices $i_1 < \cdots < i_m$. Thus, each $U_{B_{i_1},\ldots,B_{i_m}}(h)$ is a so-called decoupled U-statistics (see the Appendix for the definition). One may incorporate all m-tuples (not necessarily with distinct indices) in the computation of the median. However, this has a minor effect on the performance. Similar bounds may be proven though with a more complicated notation.

A simpler alternative is obtained by taking only "diagonal" blocks into account. More precisely, let $U_{B_i}(h)$ be the U-statistics calculated using the variables in block B_i (as defined in (1)). One may simply calculate the median of the V different U-statistics $U_{B_i}(h)$. This version is easy to analyze because $\left| \{ i \le V \ : U_{B_i}(h) \ge b \} \right|$ is a sum of independent random variables. However, this simple version is wasteful in the sense that only a small fraction of possible m-tuples are taken into account.

In the next two sections we analyze the performance of the estimator $\overline{U}_{\mathcal{B}}(h)$.

2.1 Exponential inequalities for P-degenerate kernels with finite variance.

Next we present a performance bound of the estimator $\overline{U}_{\mathcal{B}}(h)$ in the case when σ^2 is finite. The somewhat more complicated case of infinite second moment is treated in Section 2.2.

Theorem 1. *Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in $\mathcal{X}$. Let $h : \mathcal{X}^m \mapsto \mathbb{R}$ be a symmetric kernel that is P-degenerate of order $q-1$. Assume $Var(h(X_1, \ldots, X_m)) = \sigma^2 < \infty$. Let $\delta \in (0, \frac{1}{2})$ be such that $\lceil \log(1/\delta) \rceil \leq \frac{n}{64m}$. Let $\mathcal{B}$ be a regular partition of $\{1, \ldots, n\}$ with $|\mathcal{B}| = 32m \lceil \log(1/\delta) \rceil$. Then, with probability at least $1 - 2\delta$, we have*

$$\left|\overline{U}_{\mathcal{B}}(h) - m_h\right| \leq K_m \sigma \left(\frac{\lceil \log(1/\delta) \rceil}{n} \right)^{q/2} , \tag{7}$$

where $K_m = 2^{\frac{7}{2}m+1} m^{\frac{m}{2}}$.

When $q = m$, the kernel $h - m_h$ is P-canonical and the rate of convergence is then given by $(\log \delta^{-1}/n)^{m/2}$. Thus, the new estimator has a performance similar to standard U-statistics as in (3) and (4) but without the boundedness assumption for the kernel. It is important to note that a disadvantage of the estimator $\overline{U}_{\mathcal{B}}(h)$ is that it depends on the confidence level δ (through the number of blocks). For different confidence levels, different estimators are used.

Because of its importance in applications, we spell out the special case when $m = q = 2$. In Section 3 we use this result in an example of cluster analysis.

Corollary 2. *Let $\delta \in (0, 1/2)$. Let $h : \mathcal{X}^2 \mapsto \mathbb{R}$ be a P-canonical kernel with $\sigma^2 = Var(h(X_1, X_2))$ and let $n \geq 128(1 + \log(1/\delta))$. Then, with probability at least $1 - 2\delta$,*

$$|\overline{U}_{\mathcal{B}}(h) - m_h| \leq 512\sigma \frac{1 + \log(1/\delta)}{n} . \tag{8}$$

In the proof of Theorem 1 we need the notion of *Hoeffding decomposition* (Hoeffding, 1948) of U-statistics. For probability measures $P_1, \ldots, P_m$, define $P_1 \times \cdots \times P_m h = \int h \; d(P_1, \ldots, P_m)$. For a symmetric kernel $h : \mathcal{X}^m \mapsto \mathbb{R}$ the *Hoeffding projections* are defined, for $0 \leq k \leq m$ and $x_1, \ldots, x_k \in \mathcal{X}$, as

$$\pi_k h(x_1, \ldots, x_k) := (\delta_{x_1} - P) \times \cdots \times (\delta_{x_k} - P) \times P^{m-k} h$$

where δ_x denotes the Dirac measure at the point x. Observe that $\pi_0 h = P^m h$ and for $k > 0$, $\pi_k h$ is a P-canonical kernel. h can be decomposed as

$$h(x_1, \ldots, x_m) = \sum_{k=0}^{m} \sum_{1 \leq i_1 < \ldots < i_k \leq m} \pi_k h(x_{i_1}, \ldots, x_{i_k}) . \tag{9}$$

If h is assumed to be square-integrable (i.e., $P^m h^2 < \infty$), the terms in (9) are orthogonal. If h is degenerate of order $q-1$, then for any $1 \leq k \leq q-1$, $\pi_k h = 0$.

Proof of Theorem 1. We begin with a "weak" concentration result on each $U_{B_{i_1}, \ldots, B_{i_m}}(h)$. Let $B_{i_1}, \ldots, B_{i_m}$ be elements of $\mathcal{B}$. For any $B \in \mathcal{B}$, we have $\frac{n}{2|\mathcal{B}|} \leq |B| \leq \frac{2n}{|\mathcal{B}|}$. We denote by

$\mathbf{k} = (k_1, \ldots, k_m)$ an element of $I_{B_{i_1},\ldots,B_{i_m}}$. We have, by the above-mentioned orthogonality property,

$$
\begin{aligned}
&\mathrm{Var}\left(U_{B_{i_1},\ldots,B_{i_m}}(h)\right)\\
&= \mathbb{E}\left[(U_{B_{i_1},\ldots,B_{i_m}}(h) - P^m h)^2\right]\\
&= \frac{1}{|B_{i_1}|^2 \ldots |B_{i_m}|^2} \sum_{\substack{\mathbf{k}\in I_{B_{i_1},\ldots,B_{i_m}}\\ \mathbf{l}\in I_{B_{i_1},\ldots,B_{i_m}}}} \mathbb{E}\left[(h(X_{k_1},\ldots,X_{k_m}) - P^m h)(h(X_{l_1},\ldots,X_{l_m}) - P^m h)\right]\\
&= \frac{1}{|B_{i_1}|^2 \ldots |B_{i_m}|^2} \sum_{\substack{\mathbf{k}\in I_{B_{i_1},\ldots,B_{i_m}}\\ \mathbf{l}\in I_{B_{i_1},\ldots,B_{i_m}}}} \sum_{s=q}^{m} \binom{|\mathbf{k}\cap\mathbf{l}|}{s} \mathbb{E}\left[\pi_s h(X_1,\ldots,X_s)^2\right] \quad \text{(by orthogonality)}\\
&\leq \frac{1}{|B_{i_1}|^2 \ldots |B_{i_m}|^2} \sum_{\mathbf{k}\in I_{B_{i_1},\ldots,B_{i_m}}} \sum_{s=q}^{m}\sum_{t=0}^{m} \binom{t}{s} \mathbb{E}\left[\pi_s h(X_1,\ldots,X_s)^2\right] \times \left(\frac{2n}{|\mathcal{B}|}\right)^{m-t} .
\end{aligned}
$$

The last inequality is obtained by counting, for any fixed $\mathbf{k}$ and t, the number of elements $\mathbf{l}$ such that $|\mathbf{k}\cap\mathbf{l}| = t$. Thus,

$$
\begin{aligned}
\mathrm{Var}\left(U_{B_{i_1},\ldots,B_{i_m}}(h)\right) &\leq \frac{1}{|B_{i_1}| \ldots |B_{i_m}|} \sum_{s=q}^{m}\sum_{t=q}^{m} \binom{t}{s} \mathbb{E}\left[\pi_s h(X_1,\ldots,X_s)^2\right] \times \left(\frac{2n}{|\mathcal{B}|}\right)^{m-t}\\
&\leq \frac{1}{|B_{i_1}| \ldots |B_{i_m}|} \sum_{s=q}^{m} \binom{m}{s} \mathbb{E}\left[\pi_s h(X_1,\ldots,X_s)^2\right] \times \sum_{t=q}^{m} \left(\frac{2n}{|\mathcal{B}|}\right)^{m-t}\\
&\leq \frac{1}{\left(\frac{n}{2|\mathcal{B}|}\right)^m} \sum_{s=q}^{m} \binom{m}{s} \mathbb{E}\left[\pi_s h(X_1,\ldots,X_s)^2\right] \times 2\left(\frac{2n}{|\mathcal{B}|}\right)^{m-q}\\
&\leq \frac{2^{2m-q+1}|\mathcal{B}|^q}{n^q} \sum_{s=q}^{m} \binom{m}{s} \mathbb{E}\left[\pi_s h(X_1,\ldots,X_s)^2\right] .
\end{aligned}
$$

On the other hand, we have, by (9),

$$
\begin{aligned}
\mathrm{Var}(h) &= \mathbb{E}\left[\left(\sum_{s=q}^{m} \sum_{1\leq i_1<\ldots<i_s\leq m} \pi_s h(X_{i_1},\ldots,X_{i_s})\right)^2\right]\\
&= \sum_{s=q}^{m} \sum_{1\leq i_1<\ldots<i_s\leq m} \mathbb{E}\left[(\pi_s h(X_{i_1},\ldots,X_{i_s}))^2\right]\\
&= \sum_{s=q}^{m} \binom{m}{s} \mathbb{E}\left[(\pi_s h(X_1,\ldots,X_s))^2\right] .
\end{aligned}
$$

Combining the two displayed equations above,

$$\mathrm{Var}\left(U_{B_{i_1},\ldots,B_{i_m}}(h)\right) \leq \frac{2^{2m-q+1}|\mathcal{B}|^q}{n^q}\sigma^2 \leq \frac{2^{2m}|\mathcal{B}|^q}{n^q}\sigma^2 \ .$$

By Chebyshev's inequality, for all $r \in (0,1)$,

$$\mathbb{P}\left\{U_{B_{i_1},\ldots,B_{i_m}}(h) - P^m h > 2^m \sigma \frac{|\mathcal{B}|^{q/2}}{n^{q/2} r^{1/2}}\right\} \leq r \ . \tag{10}$$

We set $x = 2^m \sigma \frac{|\mathcal{B}|^{q/2}}{n^{q/2} r^{1/2}}$, and

$$N_x = \left|\left\{(i_1, \ldots, i_m) \in \{1, \ldots, V\}^m : 1 \leq i_1 < \ldots < i_m \leq |B|,\ U_{B_{i_1},\ldots,B_{i_m}}(h) - P^m h > x\right\}\right| \ .$$

The random variable $\frac{1}{\binom{|\mathcal{B}|}{m}} N_x$ is a U-statistics of order m with the symmetric kernel $g : (i_1, \ldots, i_m) \mapsto \mathbb{1}_{\{U_{B_{i_1},\ldots,B_{i_m}}(h) - P^m h > x\}}$. Thus, Hoeffding's inequality for centered U-statistics (2) gives

$$\mathbb{P}\left\{N_x - \mathbb{E}N_x \geq t\binom{|\mathcal{B}|}{m}\right\} \leq \exp\left(-\frac{|\mathcal{B}|t^2}{2m}\right) \ . \tag{11}$$

By (10) we have $\mathbb{E}N_x \leq \binom{|\mathcal{B}|}{m} r$. Taking $t = r = \frac{1}{4}$ in (11), by the definition of the median, we have

$$\begin{aligned}
\mathbb{P}\left\{\overline{U}_{\mathcal{B}}(h) - P^m(h) > x\right\} &\leq \mathbb{P}\left\{N_x \geq \frac{\binom{|\mathcal{B}|}{m}}{2}\right\} \\
&\leq \exp\left(-\frac{|\mathcal{B}|}{32m}\right) \ .
\end{aligned}$$

Since $|\mathcal{B}| \geq 32m \log(\delta^{-1})$, with probability at least $1 - \delta$, we have

$$\overline{U}_{\mathcal{B}}(h) - P^m h \leq K_m \sigma \left(\frac{\lceil \log \delta^{-1} \rceil}{n}\right)^{q/2}$$

with $K_m = 2^{\frac{7}{2}m+1} m^{\frac{m}{2}}$. The upper bound for the lower tail holds by the same argument. $\square$

2.2 Bounded moment of order p with $1 < p \leq 2$

In this section, we weaken the assumption of finite variance and only assume the existence of a centered moment of order p for some $1 < p \leq 2$. The outline of the argument is similar as in the case of finite variance. First we obtain a "weak" concentration inequality for the U-statistics is each block and then use the property of the median to boost the weak inequality. While for the case of finite variance weak concentration could be proved by a direct calculation of the variance, here we need the randomization inequalities for convex functions of U-statistics established by de la Peña (1992) and Arcones and Giné (1993). Note that, here, a P-canonical technical assumption is needed.

Theorem 3. *Let h be a symmetric kernel of order m such that $h - m_h$ is P-canonical. Assume that $M_p := \mathbb{E}\left[\left|h(X_1,\ldots,X_m) - m_h\right|^p\right]^{1/p} < \infty$ for some $1 < p \le 2$. Let $\delta \in (0, \frac{1}{2})$ be such that $\lceil \log(\delta^{-1}) \rceil \le \frac{n}{64m}$. Let $\mathcal{B}$ be a regular partition of $\{1,\ldots,n\}$ with $|\mathcal{B}| = 32m \lceil \log(\delta^{-1}) \rceil$. Then, with probability at least $1 - 2\delta$, we have*

$$|\overline{U}_{\mathcal{B}}(h) - m_h| \le K_m M_p \left(\frac{\lceil \log(\delta^{-1}) \rceil}{n}\right)^{m(p-1)/p} \tag{12}$$

where $K_m = 2^{4m+1} m^{\frac{m}{2}}$.

Proof. Define the centered version of h by $g(x_1,\ldots,x_m) := h(x_1,\ldots,x_m) - m_h$. Let $\varepsilon_1,\ldots,\varepsilon_n$ be i.i.d. Rademacher random variables (i.e., $\mathbb{P}\{\varepsilon_1 = -1\} = \mathbb{P}\{\varepsilon_1 = 1\} = 1/2$) independent of $X_1,\ldots,X_n$. By the randomization inequalities (see Theorem 3.5.3 in de la Peña and Giné (1999) and also Theorem 8 in the Appendix), we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{(k_1,\ldots,k_m)\in I_{B_{i_1},\ldots,B_{i_m}}} g(X_{k_1},\ldots,X_{k_m})\right|^p\right] \\
&\le 2^{mp}\mathbb{E}_X\mathbb{E}_\varepsilon\left[\left|\sum_{(k_1,\ldots,k_m)\in I_{B_{i_1},\ldots,B_{i_m}}} \varepsilon_{k_1}\ldots\varepsilon_{k_m} g(X_{k_1},\ldots,X_{k_m})\right|^p\right] \qquad (13)\\
&\le 2^{mp}\mathbb{E}_X\left[\left|\mathbb{E}_\varepsilon\left[\left(\sum_{(k_1,\ldots,k_m)\in I_{B_{i_1},\ldots,B_{i_m}}} \varepsilon_{k_1}\ldots\varepsilon_{k_m} g(X_{k_1},\ldots,X_{k_m})\right)^2\right]\right|^{p/2}\right] \\
&= 2^{mp}\mathbb{E}_X\left[\left|\sum_{(k_1,\ldots,k_m)\in I_{B_{i_1},\ldots,B_{i_m}}} g(X_{k_1},\ldots,X_{k_m})^2\right|^{p/2}\right] \\
&\le 2^{mp}\sum_{(k_1,\ldots,k_m)\in I_{B_{i_1},\ldots,B_{i_m}}} \mathbb{E}|g(X_{k_1},\ldots,X_{k_m})|^p \\
&= 2^{mp}|B_{i_1}|\cdots|B_{i_m}|\mathbb{E}|g|^p\,. \qquad (14)
\end{aligned}$$

Thus, we have $\mathbb{E}\left[|U_{B_{i_1},\dots,B_{i_m}}(h) - m_h|^p\right] \le 2^{mp}(|B_{i_1}|\dots|B_{i_m}|)^{1-p}\mathbb{E}|g|^p$ and by Markov's inequality,

$$\mathbb{P}\left\{U_{B_{i_1},\dots,B_{i_m}}(h) - m_h > \frac{2^m M_p}{r^{\frac{1}{p}}}\left(\frac{n}{(2|\mathcal{B}|)}\right)^{m\frac{1-p}{p}}\right\} \le r \ . \tag{15}$$

Another use of (11) with $t = r = \frac{1}{4}$ gives

$$\overline{U}_{\mathcal{B}}(h) - P^m h \le 2^{4m+1} m^{\frac{m}{2}} M_p \left(\frac{\lceil \log \delta^{-1} \rceil}{n}\right)^{m\frac{p-1}{p}} \ .$$

□

To see why the bound of Theorem 3 gives essentially the right order of magnitude, consider again the example described in the introduction, when $m = 2$, $h(X_1, X_2) = X_1 X_2$, and the X_i have an α-*stable* law $S(\gamma, \alpha)$ for some $\gamma > 0$ and $1 < \alpha \le 2$. Note that an α-stable random variable has finite moments up to (but not including) α and therefore we may take any $p = \alpha - \epsilon$ for any $\epsilon \in (0, 1 - \alpha)$. As we noted it in the introduction, there exists a constant c depending on α and γ only such that for all $1 \le i_1 < i_2 \le V$,

$$\mathbb{P}\left\{\left|U_{B_{i_1},B_{i_2}}(h) - m_h\right| \ge c\left(\frac{n}{|\mathcal{B}|}\right)^{2/\alpha - 2}\right\} \ge 2/3 \ ,$$

and therefore (15) is essentially the best rate one can hope for.

3 Cluster analysis with U-statistics

In this section we illustrate the use of the proposed mean estimator in a clustering problem when the presence of possibly heavy-tailed data requires robust techniques.

We consider the general statistical framework defined by Clémençon (2014), described as follows: Let X, X' be i.i.d. random variables taking values in $\mathcal{X}$ where typically but not necessarily, $\mathcal{X}$ is a subset of $\mathbb{R}^d$). For a partition $\mathcal{P}$ of $\mathcal{X}$ into K disjoint sets–the so-called "cells"–, define $\Phi_{\mathcal{P}}(x, x') = \sum_{\mathcal{C} \in \mathcal{P}} \mathbb{1}_{\{(x,x') \in \mathcal{C}^2\}}$ the $\{0, 1\}$-valued function that indicates whether two elements x and x' belong to the same cell $\mathcal{C}$. Given a *dissimilarity measure* $D : \mathcal{X}^2 \to \mathbb{R}_+^*$, the clustering task consists in finding a partition of $\mathcal{X}$ minimizing the *clustering risk*

$$W(\mathcal{P}) = \mathbb{E}\left[D(X, X')\Phi_{\mathcal{P}}(X, X')\right] \ .$$

Let Π_K be a finite class of partitions $\mathcal{P}$ of $\mathcal{X}$ into K cells and define $W^* = \min_{\mathcal{P} \in \Pi_K} W(\mathcal{P})$.

Given $X_1, \dots, X_n$ be i.i.d. random variables distributed as X, the goal is to find a partition $\mathcal{P} \in \Pi_K$ with risk as close to W^* as possible. A natural idea–and this is the approach of Clémençon (2014)–is to estimate $W(\mathcal{P})$ by the U-statistics

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

and choose a partition minimizing the empirical clustering risk $\widehat{W}_n(\mathcal{P})$. Clémençon (2014) uses the theory of U-processes to analyze the performance of such minimizers of U-statistics. However, in order to control uniform deviations of the form $\sup_{\mathcal{P}\in\Pi_K} |\widehat{W}_n(\mathcal{P}) - W(\mathcal{P})|$, exponential concentration inequalities are needed for U-statistics. This restricts one to consider bounded dissimilarity measures $D(X, X')$. When $D(X, X')$ may have a heavy tail, we propose to replace U-statistics by the median-of-means estimators of $W(\mathcal{P})$ introduced in this paper.

Let $\mathcal{B}$ be a regular partition of $\{1, \ldots, n\}$ and define the median-of-means estimator $\overline{W}_{\mathcal{B}}(\mathcal{P})$ of $W(\mathcal{P})$ as in (6). Then Theorem 1 applies and we have the following simple corollary.

Corollary 4. *Let Π_K be a class of partitions of cardinality $|\Pi_K| = N$. Assume that $\sigma^2 := \mathbb{E}\left[D(X_1, X_2)^2\right] < \infty$. Let $\delta \in (0, 1/2)$ be such that $n \ge 128 \lceil \log(N/\delta) \rceil$. Let $\mathcal{B}$ be a regular partition of $\{1, \ldots, n\}$ with $|\mathcal{B}| = 64 \lceil \log(N/\delta) \rceil$. Then there exists a constant C such that, with probability at least $1 - 2\delta$,*

$$\sup_{\mathcal{P}\in\Pi_K} |\overline{W}_{\mathcal{B}}(\mathcal{P}) - W(\mathcal{P})| \le C\sigma \left(\frac{\lceil \log(N/\delta) \rceil}{n} \right)^{1/2} . \tag{16}$$

Proof. Since $\Phi_{\mathcal{P}}(x, x')$ is bounded by 1, $\mathrm{Var}\left(D(X_1, X_2)\Phi_{\mathcal{P}}(X_1, X_2)\right) \le \mathbb{E}\left[D(X_1, X_2)^2\right]$. For a fixed $\mathcal{P} \in \Pi_K$, Theorem 1 applies with $m = 2$ and $q = 1$. The inequality follows from the union bound. □

Once uniform deviations of $\overline{W}_{\mathcal{B}}(\mathcal{P})$ from its expected value are controlled, it is a routine exercise to derive performance bounds for clustering based on minimizing $\overline{W}_{\mathcal{B}}(\mathcal{P})$ over $\mathcal{P} \in \Pi_K$.

Let $\widehat{\mathcal{P}} = \mathrm{argmin}_{\mathcal{P}\in\Pi_K} \overline{W}_{\mathcal{B}}(\mathcal{P})$ denote the empirical minimizer. (In case of multiple minimizers, one may select one arbitrarily.) Now for any $\mathcal{P}_0 \in \Pi_K$,

$$\begin{aligned} W(\widehat{\mathcal{P}}) - W^* &= W(\widehat{\mathcal{P}}) - \overline{W}_{\mathcal{B}}(\widehat{\mathcal{P}}) + \overline{W}_{\mathcal{B}}(\widehat{\mathcal{P}}) - W^* \\ &\le W(\widehat{\mathcal{P}}) - \overline{W}_{\mathcal{B}}(\widehat{\mathcal{P}}) + \overline{W}_{\mathcal{B}}(\mathcal{P}_0) - W(\mathcal{P}_0) + W(\mathcal{P}_0) - W^* \\ &\le 2 \sup_{\mathcal{P}\in\Pi_K} |\overline{W}_{\mathcal{B}}(\mathcal{P}) - W(\mathcal{P})| + W(\mathcal{P}_0) - W^* . \end{aligned}$$

Taking the infimum over Π_K,

$$W(\widehat{\mathcal{P}}) - W^* \le 2 \sup_{\mathcal{P}\in\Pi_K} |\overline{W}_{\mathcal{B}}(\mathcal{P}) - W(\mathcal{P})| . \tag{17}$$

Finally, (16) implies that

$$W(\widehat{\mathcal{P}}) - W^* \le 2C\sigma \left(\frac{1+\log(N/\delta)}{n}\right)^{1/2} .$$

This result is to be compared with Theorem 2 of Clémençon (2014). Our result holds under the only assumption that $D(X, X')$ has a finite second moment. (This may be weakened to assuming the existence of a finite p-th moment for some $1 < p \le 2$ by using Theorem 3). On the other hand, our result holds only for a finite class of partitions while Clémençon (2014) uses the theory of U-processes to obtain more sophisticated bounds for uniform deviations over possibly infinite classes of partitions. It remains a challenge to develop a theory to control processes of median-of-means estimators–in the style of Arcones and Giné (1993)–and not having to resort to the use of simple union bounds.

In the rest of this section we show that, under certain "low-noise" assumptions, analogous to the ones introduced by Mammen and Tsybakov (1999) in the context of classification, to obtain faster rates of convergence. In this part we need bounds for P-canonical kernels and use the full power of Corollary 2. Similar arguments for the study of minimizing U-statistics appear in Clémençon et al. (2008), Clémençon (2014).

We assume the following conditions, also considered by Clémençon (2014):

1. There exists $\mathcal{P}^*$ such that $W(\mathcal{P}^*) = W^*$
2. There exist $\alpha \in [0, 1]$ and $\kappa < \infty$ such that for all $\mathcal{P} \in \Pi_K$ and for all $x \in \mathcal{X}$,

$$\mathbb{P}\left\{\Phi_{\mathcal{P}}(x, X) \neq \Phi_{\mathcal{P}^*}(x, X)\right\} \le \kappa (W(\mathcal{P}) - W^*)^{\alpha} .$$

Note that $\alpha \le 2$ since by the Cauchy-Schwarz inequality,

$$W(\mathcal{P}) - W^* \le \mathbb{E}\left[D(X_1, X_2)^2\right]^{1/2} \mathbb{P}\left\{\Phi_{\mathcal{P}}(X_1, X_2) \neq \Phi_{\mathcal{P}^*}(X_1, X_2)\right\}^{1/2} .$$

Corollary 5. *Assume the conditions above and that $\sigma^2 := \mathbb{E}\left[D(X_1, X_2)^2\right] < \infty$. Let $\delta \in (0, 1/2)$ be such that $n \ge 128 \lceil \log(N/\delta) \rceil$. Let $\mathcal{B}$ be a regular partition of $\{1, \ldots, n\}$ with $|\mathcal{B}| = 64 \lceil \log(N/\delta) \rceil$. Then there exists a constant C such that, with probability at least $1 - 2\delta$,*

$$W(\widehat{\mathcal{P}}) - W^* \le C\sigma^{2/(2-\alpha)} \left(\frac{\lceil \log(N/\delta) \rceil}{n}\right)^{1/(2-\alpha)} . \quad (18)$$

The proof Corollary 5 is postponed to the Appendix.

4 Appendix

4.1 Decoupling and randomization

Here we summarize some of the key tools for analyzing U-statistics that we use in the paper. For an excellent exposition we refer to de la Peña and Giné (1999).

Let $\{X_i\}$ be i.i.d. random variables taking values in $\mathcal{X}$ and let $\{X_i^k\}$, $k = 1, \ldots, m$, be sequences of independent copies. Let Φ be a non-negative function. As a corollary of Theorem 3.1.1 in de la Peña and Giné (1999) we have the following:

Theorem 6. *Let $h : \mathcal{X}^m \to \mathbb{R}$ be a measurable function with $\mathbb{E}|h(X_1, \ldots, X_m)| < \infty$. Let $\Phi : [0, \infty) \to [0, \infty)$ be a convex nondecreasing function such that $\mathbb{E}\Phi\left(|h(X_1, \ldots, X_m)|\right) < \infty$. Then*

$$\mathbb{E}\Phi\left(\left|\sum_{I_n^m} h(X_{i_1}, \ldots, X_{i_m})\right|\right) \leq \mathbb{E}\Phi\left(C_m\left|\sum_{I_n^m} h(X_{i_1}^1, \ldots, X_{i_m}^m)\right|\right)$$

where $C_m = 2^m(m^m - 1)((m-1)^{m-1} - 1) \times \cdots \times 3$. Moreover, if the kernel h is symmetric, then,

$$\mathbb{E}\Phi\left(c_m\left|\sum_{I_n^m} h(X_{i_1}^1, \ldots, X_{i_m}^m)\right|\right) \leq \mathbb{E}\Phi\left(\left|\sum_{I_n^m} h(X_{i_1}, \ldots, X_{i_m})\right|\right)$$

where $c_m = 1/(2^{2m-2}(m-1)!)$.

An equivalent result for tail probabilities of U-statistics is the following (see Theorem 3.4.1 in de la Peña and Giné (1999)):

Theorem 7. *Under the same hypotheses as Theorem 6, there exists a constant C_m depending on m only such that, for all $t > 0$,*

$$\mathbb{P}\left\{\left|\sum_{I_n^m} h(X_{i_1}, \ldots, X_{i_m})\right| > t\right\} \leq C_m\mathbb{P}\left\{C_m\left|\sum_{I_n^m} h(X_{i_1}^1, \ldots, X_{i_m}^m)\right| > t\right\} .$$

If moreover, the kernel h is symmetric then there exists a constant c_m depending on m only such that, for all $t > 0$,

$$c_m\mathbb{P}\left\{c_m\left|\sum_{I_n^m} h(X_{i_1}^1, \ldots, X_{i_m}^m)\right| > t\right\} \leq \mathbb{P}\left\{\left|\sum_{I_n^m} h(X_{i_1}, \ldots, X_{i_m})\right| > t\right\} .$$

The next Theorem is a direct corollary of Theorem 3.5.3 in de la Peña and Giné (1999).

Theorem 8. *Let $1 < p \leq 2$. Let $(\varepsilon_i)_{i\leq n}$ be i.i.d Rademacher random variables independent of the $(X_i)_{i\leq n}$. Let $h : \mathcal{X} \to \mathbb{R}$ be a P-degenerate measurable function such that $\mathbb{E}\left(|h(X_1, \ldots, X_m)|^p\right) < \infty$. Then*

$$\begin{aligned} c_m\mathbb{E}\left|\sum_{I_n^m} \varepsilon_{i_1} \ldots \varepsilon_{i_m} h(X_{i_1}, \ldots, X_{i_m})\right|^p &\leq \mathbb{E}\left|\sum_{I_n^m} h(X_{i_1}, \ldots, X_{i_m})\right|^p \\ &\leq C_m\mathbb{E}\left|\sum_{I_n^m} \varepsilon_{i_1} \ldots \varepsilon_{i_m} h(X_{i_1}, \ldots, X_{i_m})\right|^p , \end{aligned}$$

where $C_m = 2^{mp}$ and $c_m = 2^{-mp}$.

The same conclusion holds for decoupled U-statistics.

4.2 α-stable distributions

Proposition 9. *Let $\alpha \in (0,2)$. Let $X_1, \dots, X_n$ be i.i.d. random variables of law $S(\gamma, \alpha)$. Let $f_{\gamma,\alpha} : x \mapsto \mathbb{R}$ be the density function of X_1. Let $S_n = \sum_{1 \le i \le n} X_i$. Then*

(i) $f_{\gamma,\alpha}(x)$ is an even function.

(ii) $f_{\gamma,\alpha}(x) \underset{x \to +\infty}{\sim} \alpha \gamma^\alpha c_\alpha x^{-\alpha-1}$ with $c_\alpha = \sin(\frac{\pi\alpha}{2})\Gamma(\alpha)/\pi$.

(iii) $\mathbb{E}[X_1^p]$ is finite for any $p < \alpha$ and is infinite whenever $p \ge \alpha$.

(iv) S_n has a α-stable law $S(\gamma n^{1/\alpha}, \alpha)$.

Proof. (i) and (iv) follow directly from the definition. (ii) is proved in the introduction of Zolotarev (1986). (iii) is a consequence of (ii). □

4.3 Proof of Corollary 5

Define $\Lambda_n(\mathcal{P}) = \widehat{W}_n(\mathcal{P}) - W^*$, the U-statistics based on the sample $X_1, \dots, X_n$, with symmetric kernel

$$h_{\mathcal{P}}(x, x') = D(x, x') \left(\Phi_{\mathcal{P}}(x, x') - \Phi_{\mathcal{P}^*}(x, x') \right) .$$

We denote by $\Lambda(\mathcal{P}) = W(\mathcal{P}) - W^*$ the expected value of $\Lambda_n(\mathcal{P})$. The main argument in the following analysis is based on the Hoeffding decomposition. For all partitions $\mathcal{P}$,

$$\Lambda_n(\mathcal{P}) - \Lambda(\mathcal{P}) = 2L_n(\mathcal{P}) + M_n(\mathcal{P})$$

for $L_n(\mathcal{P}) = \frac{1}{n}\sum_{i \le n} h^{(1)}(X_i)$ with $h^{(1)}(x) = \mathbb{E}[h_{\mathcal{P}}(X, x)] - \Lambda(\mathcal{P})$ and $M_n(\mathcal{P})$ the U-statistics based on the canonical kernel given by $h^{(2)}(x, x') = h_{\mathcal{P}}(x, x') - h^{(1)}(x) - h^{(1)}(x') - \Lambda(\mathcal{P})$. Let $\mathcal{B}$ be a regular partition of $\{1, \dots, n\}$. For any $B \in \mathcal{B}$, $\Lambda_B(\mathcal{P})$ is the U-statistics on the kernel $h_{\mathcal{P}}$ restricted to the set B and $\overline{\Lambda}_B(\mathcal{P})$ is the median of the sequence $(\Lambda_B(\mathcal{P}))_{B \in \mathcal{B}}$. We define similarly $L_B(\mathcal{P})$ and $M_B(\mathcal{P})$ on the variables $(X_i)_{i \in B}$. For any $B \in \mathcal{B}$,

$$\begin{aligned} \mathrm{Var}\left(\Lambda_B(\mathcal{P})\right) &= 4\mathrm{Var}\left(L_B(\mathcal{P})\right) + \mathrm{Var}\left(M_B(\mathcal{P})\right) \\ &= \frac{4}{|B|}\mathrm{Var}\left(h^{(1)}(X)\right) + \frac{2}{|B|(|B|-1)}\mathrm{Var}\left(h^{(2)}(X_1, X_2)\right) . \end{aligned}$$

Simple computations show that $\mathrm{Var}\left(h^{(2)}(X_1, X_2)\right) = 2\mathrm{Var}\left(h^{(1)}(X)\right)$ and therefore,

$$\mathrm{Var}\left(\Lambda_B(\mathcal{P})\right) \le \frac{8}{|B|}\mathrm{Var}\left(h^{(1)}(X)\right) .$$

Moreover,

$$
\begin{aligned}
\mathrm{Var}\left(h^{(1)}(X)\right) &\le \mathbb{E}_{X'}\left[\mathbb{E}_X\left[h_{\mathcal{P}}(X,X')\right]^2\right] \\
&\le \mathbb{E}_{X'}\left[\mathbb{E}_X\left[D(X,X')^2\right]\mathbb{E}_X\left[\left(\Phi_{\mathcal{P}}(X,X')-\Phi_{\mathcal{P}^*}(X,X')\right)^2\right]\right] \\
&= \mathbb{E}_{X'}\left[\mathbb{E}_X\left[D(X,X')^2\right]\mathbb{P}_X\left\{\Phi_{\mathcal{P}}(X,X')\neq\Phi_{\mathcal{P}^*}(X,X')\right\}\right] \\
&\le \sigma^2\kappa(W(\mathcal{P})-W^*)^\alpha
\end{aligned}
$$

where $\mathbb{E}_X$ (resp. $\mathbb{E}_{X'}$) refers to the expectation taken with respect to X (resp. X'). Chebyshev's inequality gives, for $r\in(0,1)$,

$$
\mathbb{P}\left\{\Lambda_B(\mathcal{P})-\Lambda(\mathcal{P})>\sigma(W(\mathcal{P})-W^*)^{\alpha/2}\sqrt{\frac{8\kappa}{r|B|}}\right\}\le r\ .
$$

Using again (11) with $r=\frac{1}{4}$, by $|B|\ge\frac{n}{128\lceil\log(N/\delta)\rceil}$, there exists a constant C such that for any $\mathcal{P}\in\Pi_K$, with probability at least $1-2\delta/N$,

$$
|\overline{\Lambda}_{\mathcal{B}}(\mathcal{P})-\Lambda(\mathcal{P})|\le C\sigma(W(\mathcal{P})-W^*)^{\alpha/2}\sqrt{\frac{\lceil\log(N/\delta)\rceil}{n}}\ .
$$

This implies by the union bound, that

$$
|\overline{W}_{\mathcal{B}}(\widehat{\mathcal{P}})-W(\widehat{\mathcal{P}})|\le K\sigma(W(\widehat{\mathcal{P}})-W^*)^{\alpha/2}\sqrt{\frac{\lceil\log(N/\delta)\rceil}{n}}
$$

with probability at least $1-2\delta$. Using (17), we obtain

$$
(W(\widehat{\mathcal{P}})-W^*)^{1-\alpha/2}\le 2K\sigma\sqrt{\frac{\lceil\log(N/\delta)\rceil}{n}}\ ,
$$

concluding the proof.

References

Alon, N., Y. Matias, and M. Szegedy (2002). The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences 58*, 137–147.

Arcones, M. A. and E. Giné (1993). Limit theorems for U-processes. *The Annals of Probability 21*, 1494–1542.

Biau, G. and K. Bleakley (2006). Statistical inference on graphs. *Statistics & Decisions 24*(2), 209–232.

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

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

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

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

de la Peña, V. and E. Giné (1999). *Decoupling: from dependence to independence*. New York: Springer.

de la Peña, V. H. (1992). Decoupling and Khintchine's inequalities for U-statistics. *The Annals of Probability*, 1877–1892.

Giné, E., R. Latała, and J. Zinn (2000). Exponential and moment inequalities for U-statistics. In *High Dimensional Probability II—Progress in Probability*, pp. 13–38. Birkhauser.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 293–325.

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

Hsu, D. and S. Sabato (2013). Approximate loss minimization with heavy tails. *Computing Research Repository abs/1307.1827*.

Lerasle, M. and R. Oliveira (2011). Robust empirical mean estimators.

Mammen, E. and A. Tsybakov (1999). Smooth discrimination analysis. *The Annals of Statistics 27*(6), 1808–1829.

Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli*.

Nemirovsky, A. and D. Yudin (1983). Problem complexity and method efficiency in optimization.

Nolan, J. P. (2015). *Stable Distributions - Models for Heavy Tailed Data*. Boston: Birkhauser. In progress, Chapter 1 online at academic2.american.edu/∼jpnolan.

Robins, J., L. Li, E. Tchetgen, and A. van der Vaart (2009). Quadratic semiparametric von Mises calculus. *Metrika 69*(2-3), 227–247.

Zolotarev, V. (1986). *One-dimensional stable distributions*, Volume 65. American Mathematical Soc.