U-statistics of growing order and sub-Gaussian mean estimators with sharp constants

Stanislav Minsker

Abstract. This paper addresses the following question: given a sample of i.i.d. random variables with finite variance, can one construct an estimator of the unknown mean that performs nearly as well as if the data were normally distributed? One of the most popular examples achieving this goal is the median of means estimator. However, it is inefficient in a sense that the constants in the resulting bounds are suboptimal. We show that a permutation-invariant modification of the median of means estimator admits deviation guarantees that are sharp up to $1 + o(1)$ factor if the underlying distribution possesses more than $\frac{3+\sqrt{5}}{2} \approx 2.62$ moments and it is absolutely continuous with respect to the Lebesgue measure. This result yields potential improvements for a variety of algorithms that rely on the median of means estimator as a building block. At the core of our argument are the new deviation inequalities for the U-statistics of order that is allowed to grow with the sample size, a result that could be of independent interest.

1. Introduction

Let $X_1, \ldots, X_N$ be i.i.d. random variables with distribution P having mean μ and finite variance σ^2. At the core of this paper is the following: Given $1 \leq t \leq t_{\max}(N)$, construct an estimator $\widetilde{\mu}_N = \widetilde{\mu}_N(X_1, \ldots, X_N)$ such that

$$\mathbb{P}\left(|\widetilde{\mu}_N - \mu| \geq \sigma \sqrt{\frac{t}{N}}\right) \leq 2e^{-t/L} \tag{1.1}$$

for some absolute positive constant L. Estimators that satisfy this deviation property are called sub-Gaussian. For example, the sample mean $\bar{X}_N = \frac{1}{N} \sum_{j=1}^{N} X_j$ is sub-Gaussian for $t_{\max} \asymp q(N, P)$, where $q(N, P) \to \infty$ as $N \to \infty$ and the constant L equals 2; this immediately follows from the fact that convergence of the distribution functions is uniform in the central limit theorem. However, $q(N, P)$ can grow arbitrarily slow in general, and it grows as $\log^{1/2}(N)$ if $\mathbb{E}|X|^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$ in

---

view of the Berry–Esseen theorem (for instance, see the book by Petrov [28]). At the same time, the so-called median of means (MOM) estimator, originally introduced in [1, 16, 26] and studied recently in relation to the problem at hand satisfies inequality (1.1) with t_{max} of order N and $L = 24e$ (see [21]), although the latter can be improved. A large body of existing work used the MOM estimator as a core subroutine to relax underlying assumptions for a variety of statistical problems, in particular, the methods based on the empirical risk minimization; we refer the reader to an excellent survey paper by Lugosi and Mendelson [22] for a detailed overview of the recent advances.

The exact value of constant L in inequality (1.1) is less important in problems where only the minimax rates are of interest, but it becomes crucial in terms of practical value and sample efficiency of the algorithms. The benchmark here is the situation when observations are normally distributed: Catoni [4] showed that no estimator can outperform the sample mean in this situation. The latter satisfies the relation

$$\mathbb{P}\left(|\bar{X}_N - \mu| \geq \sigma \, \frac{\Phi^{-1}(1 - e^{-t/2})}{\sqrt{N}}\right) = 2e^{-t/2},$$

where $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard normal law. As we have $\Phi^{-1}(1 - e^{-t/2}) = (1 + o(1))\sqrt{t}$ as $t \to \infty$, the best guarantee of the form (1.1) one can hope for is attained for $L = 2$. It is therefore natural to ask whether there exist sharp sub-Gaussian estimators of the mean, that is, estimators satisfying (1.1) with $L = 2(1 + o(1))$, where $o(1)$ is a sequence that converges to 0 as $N \to \infty$, under minimal assumptions on the underlying distribution. This question has previously been posed by Devroye et al. [8] as an open problem, and several results appeared since then that give partial answers. We proceed with a brief review of the state of the art.

1.1. Overview of the existing results

Catoni [4] presented the first known example of a sharp sub-Gaussian estimator with $t_{max} = o(N/\kappa)$ for distributions with finite fourth moment and a known upper bound on the kurtosis κ (or, alternatively, for distribution with finite but known variance). Devroye et al. [8] introduced an alternative estimator that also required finite fourth moment but did not explicitly depend on the value of the kurtosis as an input while satisfying required guarantees for $t_{max} = o((N/\kappa)^{2/3})$. Minsker and Ndaoud [25] designed an asymptotically efficient sub-Gaussian estimator $\widetilde{\mu}_N$ that satisfies $\sqrt{N}(\widetilde{\mu}_N - \mu) \xrightarrow{d} N(0, \sigma^2)$ assuming only the finite second moment plus a mild, "small-ball" type condition. However, the constants in the non-asymptotic version of their bounds were not sharp. Finally, Lee and Valiant [20] constructed an estimator

with required properties assuming just the finite second moment, however, their guarantees hold with optimal constants only for $t_{\min} \le t \le t_{\max}$, where $t_{\max} = o(N)$ and $t_{\min} \to \infty$ as $N \to \infty$. In particular, this range excludes t in the neighborhood of 0, which is often the region of most practical interest.

1.2. Summary of the main contributions

The reasons for the popularity of the MOM estimator are plenty: it is simple to define and to compute, it admits strong theoretical guarantees, moreover, it is scale-invariant and therefore essentially tuning-free. Thus, we believe that any quantifiable improvements to its performance are worth investigating.

We start by showing that the standard MOM estimator achieves bound (1.1) with $L = \pi(1 + o(1))$, where $o(1) \to 0$ as $N \to \infty$; this fact is formally stated in Theorem 2.1. We then define a permutation-invariant version of MOM, denoted $\widehat{\mu}_N$, and show in Corollary 3.2 that, surprisingly, it is asymptotically optimal in a sense that $\sqrt{N}(\widehat{\mu}_N - \mu) \xrightarrow{d} N(0, \sigma^2)$ under minimal assumptions; compare this to the standard MOM estimator that has a limiting variance $\frac{\pi}{2}\sigma^2$. The main result of the paper, Theorem 5.1, demonstrates that optimality of $\widehat{\mu}_N$ holds in the stronger sense, namely, that inequality (1.1) is valid for a wide range of the confidence parameters, assuming the distribution of X_1 possesses q moments for some possibly unknown $q > \frac{3+\sqrt{5}}{2} \approx 2.62$ and that its characteristic function satisfies a mild decay bound.

Analysis of the estimator $\widehat{\mu}_N$ requires new inequalities for U-statistics of order that grows with the sample size. Detailed discussion and comparison with existing bounds is given in Section 4. In particular, we prove novel bounds for large deviations of the degenerate, higher order terms of the Hoeffding decomposition (Theorem 4.1), and deduce sub-Gaussian deviation guarantees for the non-degenerate U-statistics (Corollary 4.2) with the "correct" sub-Gaussian parameter. These bounds could be of independent interest.

1.3. Notation

Unspecified absolute constants will be denoted C, c, C_1, c', etc., and may take different values in different parts of the paper. Given $a, b \in \mathbb{R}$, we will write $a \wedge b$ for $\min(a, b)$ and $a \vee b$ for $\max(a, b)$. For a positive integer M, $[M]$ denotes the set $\{1, \ldots, M\}$.

We will frequently use the standard big-O and small-o notation for asymptotic relations between functions and sequences. Moreover, given two sequences $\{a_n\}_{n \ge 1}$ and $\{b_n\}_{n \ge 1}$, where $b_n \ne 0$ for all n, we will write that $a_n \ll b_n$ if $\frac{a_n}{b_n} = o(1)$ as $n \to \infty$. Note that $o(1)$ may denote different functions/sequences from line to line.

For a function $f : \mathbb{R} \mapsto \mathbb{R}$, $f^{(m)}$ will denote its m-th derivative whenever it exists. Similarly, given $g : \mathbb{R}^d \mapsto \mathbb{R}$, $\partial_{x_j} g(x_1, \ldots, x_d)$ will stand for the partial derivative of g with respect to the j-th variable. Finally, the sup-norm of g is defined via $\|g\|_\infty :=$ ess $\sup\{|g(y)| : y \in \mathbb{R}^d\}$ and the convolution of f and g is denoted $f * g$.

Given i.i.d. random variables $X_1, \ldots, X_N$ distributed according to P, we will denote the corresponding empirical measure via $P_N := \frac{1}{N} \sum_{j=1}^N \delta_{X_j}$, where $\delta_X(f) := f(X)$. For a real-valued function f and a signed measure Q, we will write Qf for $\int f \, dQ$, assuming that the last integral is well-defined. Additional notation and auxiliary results will be introduced on demand.

2. Optimal constants for the median of means estimator

Recall that we are given an i.i.d. sample $X_1, \ldots, X_N$ from distribution P with mean μ and variance σ^2. The median of means estimator of μ is constructed as follows: let $G_1 \cup \cdots \cup G_k \subseteq [N]$ be an arbitrary (possibly random but independent from the data) collection of $k \leq N/2$ disjoint subsets ("blocks") of cardinality $\lfloor N/k \rfloor$ each, $\bar{X}_j := \frac{1}{|G_j|} \sum_{i \in G_j} X_i$ and

$$\widehat{\mu}_{\mathrm{MOM}} = \mathrm{med}(\bar{X}_1, \ldots, \bar{X}_k).$$

It is known (e.g., [8, 21]) that $\widehat{\mu}_{\mathrm{MOM}}$ satisfies inequality (1.1) for $t = k$ and $L = 8e^2$. This value of L appears to be overly pessimistic however: it follows from Theorem 5 in [24] that if $k \to \infty$ sufficiently slow so that the bias of $\widehat{\mu}_{\mathrm{MOM}}$ is of order $o(N^{-1/2})$, then

$$\sqrt{N}(\widehat{\mu}_{\mathrm{MOM}} - \mu) \xrightarrow{d} N\left(0, \frac{\pi}{2}\sigma^2\right) \tag{2.1}$$

as $k, N/k \to \infty$. In particular, if $\mathbb{E}|X|^{2+\delta} < \infty$ for some $0 < \delta \leq 1$, then $k = o(N^{\delta/(1+\delta)})$ suffices for the asymptotic unbiasedness and asymptotic normality to hold. Asymptotic relation (2.1) suggests that the best value of the constant L in the deviation inequality (1.1) for the estimator $\widehat{\mu}_{\mathrm{MOM}}$ is $\pi + o(1)$. We will demonstrate that this is indeed the case. Denote

$$g(m) := \frac{1}{\sqrt{m}} \mathbb{E}\left[\left(\frac{X_1 - \mu}{\sigma}\right)^2 \min\left(\left|\frac{X_1 - \mu}{\sigma}\right|, \sqrt{m}\right)\right]. \tag{2.2}$$

Clearly, $g(m) \to 0$ as $m \to \infty$ for distributions with finite variance. Feller [10] proved that $\sup_{t \in \mathbb{R}} |\Phi_m(t) - \Phi(t)| \leq 6g(m)$, where Φ_m and Φ are the distribution functions of $\frac{1}{\sigma\sqrt{m}} \sum_{j=1}^m X_j - \mu$ and the standard normal law, respectively. It is well known that

$$g(m) \leq C \mathbb{E}\left|\frac{X_1 - \mu}{\sigma}\right|^q m^{-(q-2)/2}$$

whenever $\mathbb{E}|X_1 - \mu|^q < \infty$ for some $q \in (2, 3]$. The next result can be viewed as a non-asymptotic analogue of relation (2.1).

Theorem 2.1. *The following bound holds*:

$$\mathbb{P}\big(|\sqrt{N}(\widehat{\mu}_{\mathrm{MOM}} - \mu)| \ge \sigma\sqrt{t}\big) \le 2\exp\Big(-\frac{t}{\pi}(1 + o(1))\Big). \qquad (2.3)$$

Here, $o(1)$ is a function that goes to 0 as $k, N/k \to \infty$, uniformly over $t \in [l_{k,N}, u_{k,N}]$ for any sequences $l_{k,N} \gg k\, g^2(N/k)$ and $u_{k,N} \ll k$.

Remark 2.2. (1) Note that the bound of the theorem holds in some range of the confidence parameter (such estimators are often called "multiple-δ" in the literature, e.g., see [8]), however, this range is distribution-dependent. In particular, if $\sqrt{k}\, g(N/k) \to 0$ as $k, N \to \infty$, the previous bound holds in the range $1 \le t \ll k$, but the function $g(\cdot)$ depends on P and may converge to 0 arbitrarily slow. Under additional assumptions, more concrete bounds can be deduced: for instance, if $\mathbb{E}|X/\sigma|^{2+\varepsilon} < \infty$ for some $0 < \varepsilon \le 1$, the condition $\sqrt{k}\, g(N/k) \to 0$ is satisfied if $k = o(N^{\frac{\varepsilon}{1+\varepsilon}})$ as $N \to \infty$. In general, by choosing k appropriately, we can construct a version of the median of means estimator that satisfies required guarantees for any $1 \le t \ll N$.

(2) The exact expression for the function $o(1)$ appearing in the statement of Theorems 2.1, as well as other results in the paper (e.g., Theorem 5.1), is not made explicit. We remark that it depends on the distribution of X_1 through the function $g(\cdot)$ defined in (2.2), and on the ratios $\frac{kg^2(N/k)}{l_{k,N}}$ and $\frac{u_{k,N}}{k}$.

Proof of Theorem 2.1. As $\widehat{\mu}_{\mathrm{MOM}}$ is scale-invariant, we can assume without loss of generality that $\sigma^2 = 1$. Denote $m = \lfloor N/k \rfloor$ for brevity, let $\rho(x) = |x|$, and note that the equivalent characterization of $\widehat{\mu}_{\mathrm{MOM}}$ is

$$\widehat{\mu}_{\mathrm{MOM}} \in \operatorname*{argmin}_{z \in \mathbb{R}} \sum_{j=1}^{k} \rho(\sqrt{m}(\bar{X}_j - z)).$$

The necessary conditions for the minimum of $F(z) := \sum_{j=1}^{k} \rho(\sqrt{m}(\bar{X}_j - z))$ imply that $0 \in \partial F(\widehat{\mu}_{\mathrm{MOM}})$ – the subgradient of F, hence the left derivative $F'_-(\widehat{\mu}_{\mathrm{MOM}}) \le 0$. Therefore, if $\sqrt{N}(\widehat{\mu}_{\mathrm{MOM}} - \mu) \ge \sqrt{t}$ for some $t > 0$, then $\widehat{\mu}_{\mathrm{MOM}} \ge \mu + \sqrt{t/N}$ and, due to F'_- being non-decreasing, $F'_-(\mu + \sqrt{t/N}) \le 0$. This implies that

$$\mathbb{P}\big(\sqrt{N}(\widehat{\mu}_{\mathrm{MOM}} - \mu) \ge \sqrt{t}\big)$$

$$\le \mathbb{P}\Big(\sum_{j=1}^{k} \rho'_-(\sqrt{m}(\bar{X}_j - \mu - \sqrt{t/N})) \ge 0\Big)$$

$$= \mathbb{P}\Big(\frac{1}{\sqrt{k}} \sum_{j=1}^{k}(\rho'_-(\sqrt{m}(\bar{X}_j - \mu - \sqrt{t/N})) - \mathbb{E}\rho'_-) \ge -\sqrt{k}\,\mathbb{E}\rho'_-\Big), \qquad (2.4)$$

where we used the shortcut $\mathbb{E}\rho'_-$ in place of $\mathbb{E}\rho'_-(\sqrt{m}(\bar{X}_j - \mu - \sqrt{t/N}))$. Note that

$$
\begin{aligned}
&- \sqrt{k}\, \mathbb{E}\rho'_-(\sqrt{m}(\bar{X}_j - \mu - \sqrt{t/N})) \\
&\quad = -\sqrt{k}\big(1 - 2\mathbb{P}\big(\sqrt{m}(\bar{X}_j - \mu - \sqrt{t/N}) \leq 0\big)\big) \\
&\quad = 2\sqrt{k}\Big(\Phi\Big(\frac{\sqrt{t}}{\sqrt{k}}\Big) - \Phi(0)\Big) - 2\sqrt{k}\Big(\Phi\Big(\frac{\sqrt{t}}{\sqrt{k}}\Big) - \mathbb{P}\Big(\sqrt{m}(\bar{X}_j - \mu) \leq \frac{\sqrt{t}}{\sqrt{k}}\Big)\Big) \\
&\quad \leq 2\sqrt{k}\cdot g(m) + 2\sqrt{t}\,\frac{1}{\sqrt{t}/\sqrt{N/m}}\Big(\Phi\Big(\frac{\sqrt{t}}{\sqrt{N/m}}\Big) - \Phi(0)\Big).
\end{aligned}
$$

Since

$$
\begin{aligned}
2\sqrt{t}\,\frac{1}{\sqrt{t}/\sqrt{N/m}}\Big(\Phi\Big(\frac{\sqrt{t}}{\sqrt{N/m}}\Big) - \Phi(0)\Big) &= 2\sqrt{t}(\phi(0) + O(t/\sqrt{N/m})) \\
&= \sqrt{t}\Big(\sqrt{\frac{2}{\pi}} + O(t/\sqrt{N/m})\Big),
\end{aligned}
$$

where $\phi(t) = \Phi'(t)$, we see that

$$
- \sqrt{k}\, \mathbb{E}\rho'_-(\sqrt{m}(\bar{X}_j - \mu - \sqrt{t/N})) \leq 2\sqrt{k}\cdot g(m) + \sqrt{t}\Big(\sqrt{\frac{2}{\pi}} + O(\sqrt{t/k})\Big),
$$

which is $\sqrt{t}\sqrt{\frac{2}{\pi}}(1 + o(1))$ whenever $t \ll k$ and $t \gg k\, g^2(m)$. It remains to apply Bernstein's inequality to the right-hand side in (2.4). Observe that

$$
\begin{aligned}
\mathrm{Var}\big(\rho'_-(\sqrt{m}(\bar{X}_j - \mu - \sqrt{t/N}))\big) &= 4\mathrm{Var}\big(I\{\sqrt{m}(\bar{X}_j - \mu) \leq \sqrt{t/k}\}\big) \\
&= 4\mathbb{P}\big(\sqrt{m}(\bar{X}_j - \mu) \leq \sqrt{t/k}\big)\big(1 - \mathbb{P}\big(\sqrt{m}(\bar{X}_j - \mu) \leq \sqrt{t/k}\big)\big) \leq 1,
\end{aligned}
$$

therefore

$$
\begin{aligned}
\mathbb{P}\big(\sqrt{N}(\hat{\mu}_{\mathrm{MOM}} - \mu) \geq \sqrt{t}\big) &\leq \exp\Big(-\frac{t}{\pi(1 + o(1)) + \frac{2\sqrt{t}\sqrt{2\pi}}{3}\frac{1}{\sqrt{k}}(1 + o(1))}\Big) \\
&= \exp\Big(-\frac{t}{\pi}(1 + o(1))\Big)
\end{aligned}
$$

whenever $\sqrt{k}\, g(m) \ll t \ll \sqrt{k}$. Similar reasoning gives a matching bound for the lower tail $\mathbb{P}(\sqrt{N}(\hat{\mu}_{\mathrm{MOM}} - \mu) \leq -\sqrt{t})$, and the result follows. ∎

One may ask whether the median of means estimator admits a more sample-efficient modification, one that would satisfy inequality (1.1) with a constant L smaller than π. A natural idea is to require that the estimator is invariant with respect to permutations of the data or, equivalently, it is a function of order statistics only.

Such an extension of the MOM estimator was proposed by Minsker [24], however, no provable improvements for the performance over the standard MOM estimator were established rigorously. The question of such improvements, especially the guarantees expressed in the form (1.1), is addressed next. Let us recall the proposed construction. Assume that $2 \leq m < N$ and, given $J \subseteq [N]$ of cardinality $|J| = m$, set $\bar{X}_J := \frac{1}{m} \sum_{j \in J} X_j$. Define $\mathcal{A}_N^{(m)} = \{J \subset [N] : |J| = m\}$ and

$$\widehat{\mu}_N := \mathrm{med}(\bar{X}_J, J \in \mathcal{A}_N^{(m)}), \tag{2.5}$$

where $\{\bar{X}_J, J \in \mathcal{A}_N^{(m)}\}$ denotes the set of sample averages computed over all possible subsets of $[N]$ of cardinality m; in particular, unlike the standard median-of-means estimator, $\widehat{\mu}_N$ is uniquely defined. Note that for $m = 2$, $\widehat{\mu}_N$ coincides with the well known Hodges–Lehmann estimator of location [13]. When m is a fixed integer greater than 2, $\widehat{\mu}_N$ is known as the generalized Hodges–Lehmann estimator. Its asymptotic properties are well-understood and can be deduced from results by Serfling [31], among other works. For example, its breakdown point is $1 - (1/2)^{1/m}$ and, in case of normally distributed data, the asymptotic distribution of $\sqrt{N}(\widehat{\mu}_N - \mu)$ is centered normal with variance $\Delta_m^2 = m\sigma^2 \arctan(\frac{1}{\sqrt{m^2-1}})$. In particular, $\Delta_m^2 = \sigma^2(1 + o(1))$ as $m \to \infty$. When the underlying distribution is not symmetric however, $\widehat{\mu}_N$ is biased for the mean, and the properties of this estimator in the regime $m \to \infty$ have not been investigated in the robust statistics literature (to the best of our knowledge). Only very recently, DiCiccio and Romano [9] proved that whenever $m \to \infty$, $m = o(\sqrt{N})$ and the sample is normally distributed, $\sqrt{N}(\widehat{\mu}_N - \mu) \to N(0, \sigma^2)$. We will extend this result in several directions: first, by allowing a much wider class of underlying distributions, second, by including the case when $\sqrt{N} \ll m \ll N$, which is interesting as bias$(\widehat{\mu}_N)$ is $o(N^{-1/2})$ in this regime, and finally by presenting sharp sub-Gaussian deviation inequalities for $\widehat{\mu}_N$ that hold for heavy-tailed data.

Let us remark that an argument behind Theorem 2.1 combined with a version of Bernstein's inequality for U-statistics due to Hoeffding [15] immediately implies that $\widehat{\mu}_N$ satisfies relation (2.3). Similar reasoning applies to other deviation guarantees for the classical median of means estimator that exist in the literature, so in this sense $\widehat{\mu}_N$ always performs at least as good as $\widehat{\mu}_{\mathrm{MOM}}$.

Analysis of the estimator $\widehat{\mu}_N$ is most naturally carried out using the language of U-statistics. The following section introduces the necessary background, while additional useful facts are summarized in Section 7.1.

3. Asymptotic normality of U-statistics and the implications for $\widehat{\mu}_N$

Let $Y_1, \ldots, Y_N$ be i.i.d. random variables with distribution P_Y and assume that $h_m : \mathbb{R}^m \mapsto \mathbb{R}$, $m \geq 1$, are permutation-symmetric and square-integrable with respect to

P_Y^m functions, i.e., $h_m(x_{\pi(1)}, \ldots, x_{\pi(m)}) = h_m(x_1, \ldots, x_m)$ for any $x_1, \ldots, x_m \in \mathbb{R}$ and any permutation $\pi : [m] \mapsto [m]$, and $\mathbb{E}h_m^2(Y_1, \ldots, Y_m) < \infty$. Without loss of generality, we will also assume that $\mathbb{E}h_m := \mathbb{E}h_m(Y_1, \ldots, Y_m) = 0$. Recall that $\mathcal{A}_N^{(m)} = \{J \subseteq [N] : |J| = m\}$. The U-statistic with kernel h_m is defined as

$$U_{N,m} = \frac{1}{\binom{N}{m}} \sum_{J \in \mathcal{A}_N^{(m)}} h_m(Y_i, \ i \in J).$$

For $i \in [N]$, let

$$h_m^{(1)}(Y_i) = \mathbb{E}[h_m(Y_1, \ldots, Y_m) \mid Y_i]. \tag{3.1}$$

We will assume that $\mathbb{P}(h_m^{(1)}(Y_1) \neq 0) > 0$ for all m, meaning that the kernels h_m are non-degenerate. The random variable

$$S_{N,m} := \sum_{j=1}^N \mathbb{E}[U_{N,m} \mid Y_j] = \frac{m}{N} \sum_{j=1}^N h_m^{(1)}(Y_j),$$

known as the Hájek projection of $U_{N,m}$, is essentially the best approximation of $U_{N,m}$ in terms of the sum of i.i.d. random variables of the form $f(Y_1) + \cdots + f(Y_m)$. We are interested in the sufficient conditions guaranteeing that $\frac{U_{N,m} - S_{N,m}}{\sqrt{\operatorname{Var}(S_{N,m})}} = o_P(1)$ as $N, m \to \infty$. Such asymptotic relation immediately implies that the limiting behavior of $U_{N,m}$ is defined by the Hájek projection $S_{N,m}$. Results of these type for U-statistics of fixed order m are standard and well-known [14, 18, 30]. However, we are interested in the situation when m is allowed to grow with N, possibly up to the order $m = o(N)$. U-statistics of growing order were studied, for example, by Frees [11], however, existing results are not readily applicable in our framework. Very recently, such U-statistics have been investigated in relation to performance of Breiman's random forests algorithm (e.g., see [27, 34]). The following theorem is essentially due to Peng, Coleman and Mentch [27]; we give a different proof of this fact in Section 7.2 as we rely on parts of the argument elsewhere in the paper.

Theorem 3.1. *Assume that* $\frac{\operatorname{Var}(h_m(Y_1, \ldots, Y_m))}{\operatorname{Var}(h_m^{(1)}(Y_1))} = o(N)$ *as* $N, m \to \infty$.[1] *Then*

$$\frac{U_{N,m} - S_{N,m}}{\sqrt{\operatorname{Var}(S_{N,m})}} = o_P(1) \quad \text{as } N, m \to \infty.$$

It is easy to see that asymptotic normality of $U_{N,m} / \sqrt{\operatorname{Var}(S_{N,m})}$ immediately follows from the previous theorem whenever its assumptions are satisfied. Next, we will apply this result to establish asymptotic normality of the estimator $\widehat{\mu}_N$ defined via (2.5).

---

[1]It is well known [14] that $\operatorname{Var}(h^{(1)}(Y_1)) \leq \operatorname{Var}(h_m)/m$, therefore the condition imposed on the ratio of variances implies that $m = o(N)$.

Corollary 3.2. *Let $X_1, \ldots, X_N$ be i.i.d. with finite variance σ^2. Moreover, assume that $\sqrt{N/m}\, g(m) \to 0$ as N/m and $m \to \infty$. Then*

$$\sqrt{N}(\hat{\mu}_N - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \quad \text{as } N/m \text{ and } m \to \infty.$$

Remark 3.3. The requirement $\sqrt{N/m}\, g(m) \to 0$ guarantees $\text{bias}(\hat{\mu}_N) = o(N^{-1/2})$. Without this requirement, asymptotic normality can be established for the debiased estimator $\hat{\mu}_N - \mathbb{E}\hat{\mu}_N$.

Proof. Let $\rho(x) = |x|$ and note that the equivalent characterization of $\hat{\mu}_N$ is

$$\hat{\mu}_N \in \operatorname*{argmin}_{z \in \mathbb{R}} \sum_{J \in \mathcal{A}_N^{(m)}} \rho(\sqrt{m}(\bar{X}_J - z)).$$

The necessary conditions for the minimum of this problem imply that for any fixed $t \geq 0$,

$$\mathbb{P}\Big(\sum_{J \in \mathcal{A}_N^{(m)}} \rho'_-(\sqrt{m}(\bar{X}_J - \mu - tN^{-1/2})) > 0 \Big) \leq \mathbb{P}\big(\sqrt{N}(\hat{\mu}_N - \mu) \geq t\big)$$

and

$$\mathbb{P}\big(\sqrt{N}(\hat{\mu}_N - \mu) \geq t\big) \leq \mathbb{P}\Big(\sum_{J \in \mathcal{A}_N^{(m)}} \rho'_-(\sqrt{m}(\bar{X}_J - \mu - tN^{-1/2})) \geq 0 \Big).$$

Thus, it suffices to show that the upper and lower bounds for $\mathbb{P}(\sqrt{N}(\hat{\mu}_N - \mu) \geq t)$ converge to the same limit. To this end, we see that

$$\mathbb{P}\Big(\sum_{J \in \mathcal{A}_N^{(m)}} \rho'_-(\sqrt{m}(\bar{X}_J - \mu - tN^{-1/2})) \geq 0 \Big)$$

$$= \mathbb{P}\Big(\frac{\sqrt{N/m}}{\binom{N}{m}} \sum_{J \in \mathcal{A}_N^{(m)}} \big(\rho'_-(\sqrt{m}(\bar{X}_J - \mu - tN^{-1/2})) - \mathbb{E}\rho'_-\big) \geq -\sqrt{\frac{N}{m}}\,\mathbb{E}\rho'_- \Big), \quad (3.2)$$

where $\mathbb{E}\rho'_-$ stands for $\mathbb{E}\rho'_-(\sqrt{m}(\bar{X}_J - \mu - tN^{-1/2}))$. As it the proof of Theorem 2.1, we deduce that $-\sqrt{N/m}\,\mathbb{E}\rho'_-(\sqrt{m}(\bar{X}_J - \mu - tN^{-1/2})) \to \frac{t}{\sigma}\sqrt{2/\pi}$ whenever we have $\sqrt{N/m}\, g(m) \to 0$ and $N/m \to \infty$. It remains to analyze the U-statistic

$$\sqrt{\frac{N}{m}}\, U_{N,m} = \frac{\sqrt{N/m}}{\binom{N}{m}} \sum_{J \in \mathcal{A}_N^{(m)}} \big(\rho'_-(\sqrt{m}(\bar{X}_J - \mu - tN^{-1/2})) - \mathbb{E}\rho'_-\big).$$

As the expression above is invariant with respect to the shift $X_j \mapsto X_j - \mu$, we can assume $\mu = 0$. To complete the proof, we will verify the conditions of Theorem 3.1,

allowing one to reduce the asymptotic behavior of $U_{N,m}$ to the analysis of sums of i.i.d. random variables. For $i \in [N]$, let

$$h^{(1)}(X_i) = \sqrt{\frac{N}{m}} \, \mathbb{E}\left[\rho'_- \left(\frac{1}{\sqrt{m}} \sum_{j=1}^{m-1} \tilde{X}_j + \frac{X_i}{\sqrt{m}} - \frac{t}{\sqrt{N/m}} \right) \Big| X_i \right] - \sqrt{\frac{N}{m}} \, \mathbb{E} \rho'_-,$$

where $(\tilde{X}_1, \ldots, \tilde{X}_m)$ is an independent copy of $(X_1, \ldots, X_m)$. Our goal is to understand the size of $\mathrm{Var}(h^{(1)}(X_1))$. Specifically, we will show that $\mathrm{Var}(\frac{m}{\sqrt{N}} h^{(1)}(X_1)) \to 2/\pi$ as both m and $N/m \to \infty$. Given an integer $l \geq 1$, let $\tilde{\Phi}_l(t)$ be the cumulative distribution function of $\sum_{j=1}^{l} X_j$. Then

$$\frac{m}{\sqrt{N}} h^{(1)}(X_1) = \sqrt{m} \left(2\tilde{\Phi}_{m-1}\left(\frac{tm}{\sqrt{N}} - X_1 \right) - 1 \right) - \sqrt{m} \mathbb{E} \, \rho'_-$$

$$= 2\sqrt{m} \left(\tilde{\Phi}_{m-1}\left(\frac{tm}{\sqrt{N}} - X_1 \right) - \mathbb{E}\tilde{\Phi}_{m-1}\left(\frac{tm}{\sqrt{N}} - X_1 \right) \right)$$

$$= 2\sqrt{m} \int_{\mathbb{R}} \left(\tilde{\Phi}_{m-1}\left(\frac{tm}{\sqrt{N}} - X_1 \right) - \tilde{\Phi}_{m-1}\left(\frac{tm}{\sqrt{N}} - x \right) \right) dP(x).$$

We will apply the dominated convergence theorem to analyze this expression. Consider first the situation when the distribution of X_1 is non-lattice.[2] Then the local limit theorem for non-lattice distributions [32, Theorem 2] implies that

$$\tilde{\Phi}_{m-1}(a + h) - \tilde{\Phi}_{m-1}(a) = \frac{h}{\sqrt{2\pi(m-1)}\sigma} \exp\left(-\frac{a^2}{2(m-1)\sigma^2} \right) + o(m^{-1/2}),$$

where $\sqrt{m} \cdot o(m^{-1/2})$ converges to 0 as $m \to \infty$ for every h and uniformly in a. Therefore, we see that conditionally on X_1 and for every x,

$$\tilde{\Phi}_{m-1}\left(\frac{tm}{\sqrt{N}} - x + (x - X_1) \right) - \tilde{\Phi}_{m-1}\left(\frac{tm}{\sqrt{N}} - x \right)$$

$$= \frac{x - X_1}{\sqrt{2\pi(m-1)}\sigma} \exp\left(-\frac{(tm/\sqrt{N} - x)^2}{2(m-1)\sigma^2} \right) + o(m^{-1/2}) \qquad (3.3)$$

uniformly in m. Since $m = o(N)$ by assumption,

$$\exp\left(-\frac{(tm/\sqrt{N} - x)^2}{2(m-1)\sigma^2} \right) = 1 + o(1) \quad \text{as } m, N \to \infty,$$

hence

$$2\sqrt{m} \left(\tilde{\Phi}_{m-1}\left(\frac{tm}{\sqrt{N}} - x + (x - X_1) \right) - \tilde{\Phi}_{m-1}\left(\frac{tm}{\sqrt{N}} - x \right) \right) = 2\frac{x - X_1}{\sqrt{2\pi}\sigma} + o(1)$$

---

[2] We say that X_1 has lattice distribution if $\mathbb{P}(X_1 \in \alpha + k\beta, k \in \mathbb{Z}) = 1$ and there is no arithmetic progression $A \subset Z$ such that $\mathbb{P}(X_1 \in \alpha + k\beta, k \in A) = 1$.

P-almost everywhere. Next, we will show that

$$q_m(x, X_1) := \sqrt{m}\Big(\tilde{\Phi}_{m-1}\Big(\frac{tm}{\sqrt{N}} - X_1\Big) - \tilde{\Phi}_{m-1}\Big(\frac{tm}{\sqrt{N}} - x\Big)\Big)$$

admits an integrable majorant that does not depend on m. Note that

$$|q_m(x, X_1)| \le \sup_z \sqrt{m}\, \mathbb{P}\Big(\sum_{j=1}^{m-1} X_j \in (z, z + |x - X_1|]\Big) \le C|x - X_1|,$$

where the last inequality follows from the well known bound for the concentration function [29, Theorem 2.20]; here, $C = C(P) > 0$ is a constant that may depend on the distribution of X_1. We conclude that, by the dominated convergence theorem,

$$\frac{m}{\sqrt{N}}\, h^{(1)}(X_1) \to \sqrt{\frac{2}{\pi}}\, \frac{X_1}{\sigma}$$

as $m, N/m \to \infty$, P-almost everywhere. As

$$\Big|\frac{m}{\sqrt{N}}\, h^{(1)}(X_1)\Big| \le 2\Big|\int_{\mathbb{R}} q_m(x, X_1)\, dP(x)\Big| \le C\int_{\mathbb{R}} |x - X_1|\, dP(x)$$

and $\mathbb{E}(\int_{\mathbb{R}} |x - X_1|\, dP(x))^2 < \infty$, the second application of the dominated convergence theorem yields that $\mathrm{Var}(\frac{m}{\sqrt{N}} h^{(1)}(X_1)) \to \mathrm{Var}(\sqrt{\frac{2}{\pi}} \frac{X_1}{\sigma}) = \frac{2}{\pi}$ as $N/m \to \infty$ and $m \to \infty$.

It remains to consider the case when X_1 has a lattice distribution. In this case, a version of the local limit theorem [29] states that

$$\mathbb{P}\Big(\sum_{j=1}^{m-1} X_j = (m-1)\alpha + q\beta\Big) = \frac{\beta}{\sqrt{2\pi(m-1)}\,\sigma}\, e^{-\frac{((m-1)\alpha + q\beta)^2}{2\sigma^2(m-1)}} + o(m^{-1/2}),$$

where the $o(m^{-1/2})$ term is uniform in $q \in Z$. Now, for any given y in the interval $(tm/\sqrt{N} - x,\, tm/\sqrt{N} - x + (x - X_1)]$ of the form $y = (m-1)\alpha + q\beta$, we have that $e^{-y^2/(2\sigma^2(m-1))} = 1 + o(1)$ as $m/N \to 0$. Therefore, similarly to (3.3), in this case,

$$2\sqrt{m}\Big(\tilde{\Phi}_{m-1}\Big(\frac{tm}{\sqrt{N}} - x + (x - X_1)\Big) - \tilde{\Phi}_{m-1}\Big(\frac{tm}{\sqrt{N}} - x\Big)\Big) = 2\frac{x - X_1}{\sqrt{2\pi}\sigma} + o(1)$$

P-almost everywhere, where we also used the fact that the number of points of the form $(m-1)\alpha + q\beta$ in the interval of interest equals $(x - X_1)/\beta$. The rest of the proof proceeds exactly as in the case of non-lattice distributions, and concludes the part of the argument related to $\mathrm{Var}(\frac{m}{\sqrt{N}} h^{(1)}(X_1))$.

To finish the proof, note that $\mathrm{Var}(\sqrt{N/m}\,\rho'_-(\sqrt{m}(\bar{X}_J - \mu - tN^{-1/2}))) \leq N/m$ since $\|\rho'_-\|_\infty = 1$, hence

$$\frac{\mathrm{Var}(\sqrt{N/m}\,\rho'_-(\sqrt{m}(\bar{X}_J - \mu - tN^{-1/2})))}{\mathrm{Var}(h^{(1)}(X_1))} \leq \frac{N/m}{\frac{2}{\pi}(1 + o(1))N/m^2}$$
$$= \frac{m}{\frac{2}{\pi}(1 + o(1))} = o(N)$$

as $m \to \infty$ and $N/m \to \infty$. Therefore, Theorem 3.1 applies and yields that

$$\frac{\sqrt{\frac{N}{m}}\,U_{N,m} - \frac{m}{N}\sum_{j=1}^N h^{(1)}(X_j)}{\frac{m^2}{N}\mathrm{Var}(h^{(1)}(X_j))} = o_P(1),$$

where $\frac{m^2}{N}\mathrm{Var}(h^{(1)}(X_j)) = \frac{2}{\pi}(1 + o(1))$. In view of the central limit theorem,

$$\frac{m}{N}\sum_{j=1}^N h^{(1)}(X_j) \xrightarrow{d} N\left(0, \frac{2}{\pi}\right),$$

and we conclude that $\sqrt{\frac{N}{m}}\,U_{N,m} \xrightarrow{d} N(0, \frac{2}{\pi})$. Recalling (3.2), we see that

$$\mathbb{P}\left(\sqrt{\frac{N}{m}}\,U_{N,m} \leq \sqrt{\frac{N}{m}}\,\mathbb{E}\rho'_-\right) \to 1 - \tilde{\Phi}\left(\frac{t}{\sigma}\right),$$

or $\limsup_{m, N/m \to \infty} \mathbb{P}(\sqrt{N}(\hat{\mu}_N - \mu) \geq t) \leq 1 - \tilde{\Phi}(t/\sigma)$. On the other hand, repeating the preceding argument for the lower bound for $\mathbb{P}(\sqrt{N}(\hat{\mu}_N - \mu) \geq t)$, we get that $\liminf_{m, N/m \to \infty} \mathbb{P}(\sqrt{N}(\hat{\mu}_N - \mu) \geq t) \geq 1 - \tilde{\Phi}(t/\sigma)$, whence the claim of the theorem follows. ∎

Corollary 3.2 implies that asymptotically, the estimator $\hat{\mu}_N$ improves upon $\hat{\mu}_{\mathrm{MOM}}$. The more interesting, and difficult, question is whether non-asymptotic sub-Gaussian deviation bounds for $\hat{\mu}_N$ with improved constant can be established, and to understand the range of the deviation parameter in which such bounds are valid.

4. Deviation inequalities for U-statistics of growing order

The ultimate goal of this section is to establish a non-asymptotic analogue of Corollary 3.2. Recall that its proof relied on the classical strategy of showing that the higher-order terms in the Hoeffding decomposition of certain U-statistics are asymptotically negligible. To prove the desired non-asymptotic extension, one has to be able to show that these higher-order terms are sufficiently small with exponentially high probability. However, classical tools used to prove such bounds rely on decoupling

inequalities due to de la Pena and Montgomery-Smith [7]. Unfortunately, the constants appearing in decoupling inequalities grow very fast with respect to the order m of U-statistics, at least like m^m. As m is allowed to grow with the sample size N in our examples, such tools become insufficient to get the desired bounds in our framework. Arcones [2] derived an improved version of Bernstein's inequality for non-degenerate U-statistics where the sub-Gaussian deviations regime is controlled by $m\mathrm{Var}(h_m^{(1)}(X))$ defined in equation (3.1), rather than the larger quantity $\mathrm{Var}(h_m)$ appearing in the inequality due to Hoeffding [15]; however, this result is only useful when m is essentially fixed. Maurer [23] used different techniques that yield improvements over Arcones' result, in particular, with respect to the order m; bounds obtained in this work are non-trivial for m up to the order of $N^{1/3}$, however, this does not suffice for the applications required in the present paper. Moreover, unlike Theorem 4.1 below, results in [23] do not capture the correct behavior of degenerate U-statistics. Recently, Song, Chen and Kato [34] made significant progress in studying U-statistics of growing order and developed tools that avoid using decoupling inequalities, however, their techniques apply when $m = o(\sqrt{N})$, while we only require that $m = o(N)$.

We will be interested in U-statistics with kernels of special structure that assumes "weak" dependence on each of the individual variables. Let the kernel be centered and written in the form $h_m(\frac{x_1}{\sqrt{m}}, \ldots, \frac{x_m}{\sqrt{m}})$, whence the corresponding U-statistic is

$$U_{N,m} = \frac{1}{\binom{N}{m}} \sum_{J \in \mathcal{A}_N^{(m)}} h_m\left(\frac{X_i}{\sqrt{m}}, i \in J\right).$$

The Hoeffding decomposition of $U_{N,m}$ is defined as the sum

$$U_{N,m} = \frac{m}{N} \sum_{j=1}^{N} h_m^{(1)}(X_j) + \sum_{j=2}^{m} \frac{\binom{m}{j}}{\binom{N}{j}} \sum_{J \in \mathcal{A}_N^{(j)}} h_m^{(j)}(X_i, i \in J),$$

where $h_m^{(j)}(x_1, \ldots, x_j) = (\delta_{x_1} - P) \times \cdots \times (\delta_{x_j} - P) \times P^{m-j} h_m$. We refer the reader to Section 7.1 where the Hoeffding decomposition and related background material is reviewed in more detail.

We will assume that $U_{N,m}$ is non-degenerate, in particular, one can expect that the behavior of $U_{N,m}$ is determined by the first term $\frac{m}{N} \sum_{j=1}^{N} h_m^{(1)}(X_j)$ in the decomposition. In order to make this intuition rigorous, we need to prove that the higher-order terms are of smaller order with exponentially high probability. It is shown in the course of the proof of Theorem 3.1 that

$$\mathrm{Var}\left(\frac{\binom{m}{j}}{\binom{N}{j}} \sum_{J \in \mathcal{A}_N^{(j)}} h_m^{(j)}(X_i, i \in J)\right) \leq \mathrm{Var}(h_m)\left(\frac{m}{N}\right)^j.$$

However, to achieve our current goal, bounds for the moments of higher order are required. More specifically, the key technical difficulty lies in establishing the correct rate of decay of the higher moments with respect to the order m of the U-statistic. We will show that under suitable assumptions,

$$\mathbb{E}^{1/q} \left| \frac{\binom{m}{j}}{\binom{N}{j}} \sum_{J \in \mathcal{A}_N^{(j)}} h_m^{(j)}(X_i, i \in J) \right|^q = O\left(j^{\eta_1} q^{\eta_2} \left(\frac{m}{N} \right)^{j/2} \right)$$

for some $\eta_1 > 0$, $\eta_2 > 0$ and for all $q \geq 2$, $2 \leq j \leq j_{\max}$ for a sufficiently large $j_{\max}$. The crucial observation is that the upper bound for the higher-order L_q norms is still proportional to $(m/N)^{j/2}$, same as the L_2 norm. The following result, essentially implied by the moment inequalities of this form, is a main technical novelty and a key ingredient needed to control large deviations of the higher order terms in the Hoeffding decomposition.

Theorem 4.1. *Let*

$$V_{N,j} = \frac{\binom{m}{j}^{1/2}}{\binom{N}{j}^{1/2}} \sum_{J \in \mathcal{A}_N^{(j)}} h_m^{(j)} \left(\frac{X_i}{\sqrt{m}}, i \in J \right),$$

$$f_j(x_1, \ldots, x_j) = \mathbb{E} h_m \left(\frac{x_1}{\sqrt{m}}, \ldots, \frac{x_j}{\sqrt{m}}, \frac{X_{j+1}}{\sqrt{m}}, \ldots, \frac{X_m}{\sqrt{m}} \right)$$

and $v_k = \mathbb{E}^{1/k} |X_1 - \mathbb{E} X_1|^k$. If the kernel h_m is uniformly bounded, then there exists an absolute constant $c > 0$ such that

$$\mathbb{P}(|V_{N,j}| \geq t) \leq \exp \left(- \min \left(\frac{1}{c} \left(\frac{t^2}{\text{Var}(h_m)} \right)^{1/j}, \frac{\left(\frac{t}{\|h_m\|_\infty} \sqrt{N/j} \right)^{2/(j+1)}}{c(m/j)^{j/(j+1)}} \right) \right)$$

whenever $\min(\frac{1}{c}(\frac{t^2}{\text{Var}(h_m)})^{1/j}, \frac{\left(\frac{t}{\|h_m\|_\infty} \sqrt{N/j} \right)^{2/(j+1)}}{c(m/j)^{j/(j+1)}}) \geq 2$. Alternatively, suppose that

(i) $\|\partial_{x_1} \ldots \partial_{x_j} f_j\|_\infty \leq (C_1(P)/m)^{j/2} j^{\gamma_1 j}$ *for some $\gamma_1 \geq 1/2$,*

(ii) $v_k \leq k^{\gamma_2} M$ *for all integers $k \geq 2$ and some $\gamma_2 \geq 0$, $M > 0$.*

Then there exist constants $c_1(P), c_2(P) > 0$ that depend on γ_1 and γ_2 only such that

$$\mathbb{P}(|V_{N,j}| \geq t)$$
$$\leq \exp \left(- \min \left(\frac{1}{c_1} \left(\frac{t^2}{\text{Var}(h_m)} \right)^{1/j}, \left(\frac{t \sqrt{N/j}}{(c_2 M j^{\gamma_1 - 1/2})^j} \right)^{2/(1+j(2\gamma_2+1))} \right) \right) \quad (4.1)$$

whenever $\min(\frac{1}{c_1}(\frac{t^2}{\text{Var}(h_m)})^{1/j}, (\frac{t \sqrt{N/j}}{(c_2 M j^{\gamma_1 - 1/2})^j})^{2/(1+j(2\gamma_2+1))}) \geq \max(2, \frac{\log(N/j)}{\gamma_2 j})$. [3]

---

[3] In the course of the proof, we show that whenever $\gamma_2 = 0$, corresponding to the case of a.s. bounded X_1, inequality (4.1) is valid for all $t > 0$.

The proof of the theorem is given in Section 7.3. Let us briefly discuss the imposed conditions. The first inequality requires only boundedness of the kernel and follows from a standard argument; it is mostly useful for the degenerate kernels of higher order j, for instance, when $j \geq Cm/\log(m))$. The main result is the second inequality of the theorem that provides a much better dependence of the tails on m for small and moderate values of j. Assumption (ii) is a standard one; for instance, it holds with $\gamma_2 = 0$ for bounded random variables, $\gamma_2 = 1/2$ for sub-Gaussian and with $\gamma_2 = 1$ for sub-exponential random variables. As for assumption (i), suppose that the kernel h_m is sufficiently smooth. In this case,

$$\partial_{x_1} \cdots \partial_{x_j} f_j(x_1, \ldots, x_j)$$
$$= m^{-j/2} \mathbb{E}\left[(\partial_{x_1} \cdots \partial_{x_j} h_m)\left(\frac{x_1}{\sqrt{m}}, \ldots, \frac{x_j}{\sqrt{m}}, \frac{X_{j+1}}{\sqrt{m}}, \ldots, \frac{X_m}{\sqrt{m}}\right)\right],$$

which is indeed of order $m^{-j/2}$ with respect to m. However, the functions f_j are often smooth even if the kernel h_m is not, as we will show later for the case of an indicator function (specifically, we will prove that required inequalities hold with $\gamma_1 = 1/2$ for all $j \ll m/\log(m)$ under mild assumptions on the distribution of X_1). Next, we state a corollary – a deviation inequality that takes a particularly simple form and suffices for most of the applications discussed later. It can be viewed as an extension of Arcones version of Bernstein's inequality for the case of U-statistics of growing order, see [2].

Corollary 4.2. *Suppose that*

(i)   *the assumptions of Theorem 4.1 hold for all $2 \leq j \leq j_{\max}$ with $\gamma_1 = 1/2$,*

(ii)   *the kernel h_m is uniformly bounded,*

(iii) $\liminf_{m \to \infty} \mathrm{Var}(\sqrt{m}\, h_m^{(1)}(X_1)) > 0,$

(iv) $mM^2 = o(N^{1-\delta})$ *for some $\delta > 0$.*

Moreover, let $q(N, m)$ be an increasing function such that

$$q(N, m) = o\left(\min\left(\left(\frac{N}{mM^2}\right)^{1/(1+2\gamma_2)}, j_{\max} \log\left(\frac{N}{m}\right)\right)\right) \quad as\ N/m \to \infty.$$

Then, for all $2 \leq t \leq q(N, m)$,

$$\mathbb{P}\left(|U_{N,m}| \geq \sqrt{\frac{tm}{N}}\right) \leq 2\exp\left(-\frac{t}{2(1 + o(1))\mathrm{Var}(\sqrt{m}\, h_m^{(1)}(X_1))}\right),$$

where $o(1) \to 0$ as $N/m \to \infty$ uniformly over $2 \leq t \leq q(N, m)$. If $m = o(\frac{N^{1/2}}{\log(N)})$, we can instead choose $q(N, m)$ such that

$$q(N, m) = o\left(\min\left(\left(\frac{N}{mM^2}\right)^{1/(1+2\gamma_2)}, \frac{Nj_{\max}}{m^2}\right)\right).$$

Remark 4.3. The key point of the inequality is that the sub-Gaussian deviations are controlled by $\text{Var}(\sqrt{m}\, h_m^{(1)}(X_1))$ rather than the sub-optimal quantity $\text{Var}(h_m)$ appearing in Hoeffding's version of Bernstein's inequality for U-statistics. Moreover, the range in which $U_{N,m}$ admits sub-Gaussian deviations is much wider compared to the implications of Arcones' inequality when m is allowed to grow with N. Several comments regarding the additional assumptions are in order:

(1) The assumption of uniform boundedness of the kernel h_m is needed to ensure that we can apply Bernstein's concentration inequality to the first term of the Hoeffding decomposition. This suffices for our purposes but in general this condition can be relaxed.

(2) Assumption on the asymptotic behavior of the variance is made to simplify the statement and the proof; if it does not hold, the result is still valid once the definition of $q(N, m)$ is modified to reflect the different behavior of the this quantity. We include the following heuristic argument which shows that $\lim_{m\to\infty} \text{Var}(\sqrt{m}\, h_m^{(1)}(X_1))$ often admits a simple closed-form expression. Indeed, $\sqrt{m}(h_m^{(1)}(X_1) - h_m^{(1)}(0)) = \int_0^{X_1} \sqrt{m}\, \partial_u h_m^{(1)}(u)\, du$. Therefore, if the relation $\|\partial_u^2 h_m^{(1)}\|_\infty = o(m^{-1/2})$ holds, then

$$\sqrt{m}\,|\partial_u h_m^{(1)}(u) - \partial_u h_m^{(1)}(0)| \le \sqrt{m}\,\|\partial_u^2 h_m^{(1)}\|_\infty\, u \to 0$$

pointwise as $m \to \infty$. If the limit $\sqrt{m}\, \partial_u h_m^{(1)}(0)$ exists, then

$$\sqrt{m}(h_m^{(1)}(X_1) - h_m^{(1)}(0)) \to \lim_{m\to\infty} \sqrt{m}\, \partial_u h_m^{(1)}(0) X_1 \quad P\text{-almost everywhere.}$$

Moreover, as $\sqrt{m}\|\partial_u h_m^{(1)}\|_\infty$ admits an upper bound independent of m by assumption (i) of Theorem 4.1 and X_1 is sufficiently integrable, Lebesgue's dominated convergence theorem applies and yields that $\text{Var}(\sqrt{m}\, h_m^{(1)}(X_1)) \to (\lim_{m\to\infty} \partial_u h_m^{(1)}(0))^2 \text{Var}(X_1)$. For instance, this heuristic argument can often be made precise for kernels of the form $h(\sum_{j=1}^m x_j/\sqrt{m})$.

(3) Finally, the condition requiring that $mM^2 = o(N^{1-\delta})$ is used to ensure that $(\frac{N}{mM^2})^\tau \gg \log(m)$ for any fixed $\tau > 0$, which simplifies the statement and the proof.

Proof. The union bound together with Hoeffding's decomposition entails that for any $t > 0$ and $0 < \varepsilon < 1$ (to be chosen later),

$$\mathbb{P}\Big(|U_{N,m}| \ge \sqrt{\frac{tm}{N}}\Big) \le \mathbb{P}\Big(\Big|\frac{m}{N}\sum_{j=1}^N h_m^{(1)}(X_j)\Big| \ge (1-\varepsilon)\sqrt{t}\,\sqrt{\frac{m}{N}}\Big)$$

$$+ \mathbb{P}\Big(\Big|\sum_{j=2}^m \frac{\binom{m}{j}}{\binom{N}{j}}\sum_{J\in\mathcal{A}_N^{(j)}} h_m^{(j)}(X_i,\, i\in J)\Big| \ge \varepsilon\sqrt{t}\,\sqrt{\frac{m}{N}}\Big).$$

Bernstein's inequality yields that

$$\mathbb{P}\left(\left|\frac{m}{N}\sum_{j=1}^{N}h_m^{(1)}(X_j)\right| \ge (1-\varepsilon)\sqrt{t}\sqrt{\frac{m}{N}}\right)$$

$$\le 2\exp\left(-\frac{(1-\varepsilon)^2\,t/2}{\mathrm{Var}(\sqrt{m}\,h_m^{(1)}(X_1)) + (1-\varepsilon)\frac{1}{3}\sqrt{m/N}\|h_m\|_\infty\,t^{1/2}}\right)$$

$$= 2\exp\left(-\frac{(1-\varepsilon)^2\,t}{2\,\mathrm{Var}(\sqrt{m}\,h_m^{(1)}(X_1))(1+o(1))}\right),$$

where $o(1) \to 0$ as $N/m \to \infty$ uniformly over $s \le q(N/m)$. It remains to control the expression involving higher order Hoeffding decomposition terms. Specifically, we will show that under our assumptions, this expression is bounded from above by $\exp(-t/[2\,\mathrm{Var}(\sqrt{m}\,h_m^{(1)}(X_1))]) \cdot o(1)$, where $o(1) \to 0$ uniformly over the range of t. To this end, denote $t_\varepsilon := \varepsilon^2 t$ and $j_* := \min(j_{\max}, \lfloor \log(N/m)\rfloor + 1)$. Observe that

$$\mathbb{P}\left(\left|\sum_{j=2}^{m}\frac{\binom{m}{j}}{\binom{N}{j}}\sum_{J\in\mathcal{A}_N^{(j)}}h_m^{(j)}(X_i,\,i\in J)\right| \ge \sqrt{t_\varepsilon}\sqrt{\frac{m}{N}}\right)$$

$$\le \mathbb{P}\left(\left|\sum_{j=2}^{j_*}\frac{\binom{m}{j}}{\binom{N}{j}}\sum_{J\in\mathcal{A}_N^{(j)}}h_m^{(j)}(X_i,\,i\in J)\right| \ge \frac{\sqrt{t_\varepsilon}}{3}\sqrt{\frac{m}{N}}\right)$$

$$+ \mathbb{P}\left(\left|\sum_{j=j_*+1}^{j_{\max}}\frac{\binom{m}{j}}{\binom{N}{j}}\sum_{J\in\mathcal{A}_N^{(j)}}h_m^{(j)}(X_i,\,i\in J)\right| \ge \frac{\sqrt{t_\varepsilon}}{3}\sqrt{\frac{m}{N}}\right)$$

$$+ \mathbb{P}\left(\left|\sum_{j>j_{\max}}\frac{\binom{m}{j}}{\binom{N}{j}}\sum_{J\in\mathcal{A}_N^{(j)}}h_m^{(j)}(X_i,\,i\in J)\right| \ge \frac{\sqrt{t_\varepsilon}}{3}\sqrt{\frac{m}{N}}\right), \qquad (4.2)$$

where the second sum may be empty depending on the value of j_*. First, we estimate the last term using Chebyshev's inequality. Repeating the reasoning leading to equation (7.5) in the proof of Theorem 3.1, we see that

$$\mathrm{Var}\left(\sum_{j>j_{\max}}\frac{\binom{m}{j}}{\binom{N}{j}}\sum_{J\in\mathcal{A}_N^{(j)}}h_m^{(j)}(X_i,\,i\in J)\right) \le \mathrm{Var}(h_m)\left(\frac{m}{N}\right)^{j_{\max}+1}\left(1-\frac{m}{N}\right)^{-1},$$

hence

$$\mathbb{P}\left(\left|\sum_{j>j_{\max}}\frac{\binom{m}{j}}{\binom{N}{j}}\sum_{J\in\mathcal{A}_N^{(j)}}h_m^{(j)}(X_i,\,i\in J)\right| \ge \frac{\sqrt{t_\varepsilon}}{3}\sqrt{\frac{m}{N}}\right)$$

$$\le \frac{18\,\mathrm{Var}(h_m)}{t_\varepsilon}\left(\frac{m}{N}\right)^{j_{\max}} = 18\,\mathrm{Var}(h_m)\exp\left(-j_{\max}\log\left(\frac{N}{m}\right)+\log(t_\varepsilon)\right)$$

whenever $N/m \geq 2$. Alternatively, we can apply the first inequality of Theorem 4.1 instead of Chebyshev's inequality to each term corresponding to $j > j_{\max}$ individually, with $t = t_{j,\varepsilon} := \frac{\sqrt{t_\varepsilon}}{3j^2} \left(\frac{N}{m}\right)^{(j-1)/2}$. This implies that

$$
\mathbb{P}\left(\left| \sum_{j > j_{\max}} \frac{\binom{m}{j}}{\binom{N}{j}} \sum_{J \in \mathcal{A}_N^{(j)}} h_m^{(j)}(X_i, i \in J) \right| \geq \frac{\sqrt{t_\varepsilon}}{3} \sqrt{\frac{m}{N}} \right)
$$

$$
\leq \sum_{j > j_{\max}} \mathbb{P}\left(\left| \frac{\binom{m}{j}}{\binom{N}{j}} \sum_{J \in \mathcal{A}_N^{(j)}} h_m^{(j)}(X_i, i \in J) \right| \geq \frac{\sqrt{t_{j,\varepsilon}}}{3} \sqrt{\frac{m}{N}} \right)
$$

$$
\leq m \max_{j > j_{\max}} \exp\left(-c \min\left(t_\varepsilon^{1/j} \left(\frac{N}{m}\right)^{\frac{j-1}{j}}, \left(\frac{t_\varepsilon}{\|h\|_\infty^2}\right)^{\frac{1}{j+1}} \left(\frac{Nj}{m^2}\right)^{\frac{j}{j+1}} \right)\right).
$$

This bound is useful when $(Nj_{\max}/m^2)^{j_{\max}/(j_{\max}+1)} \gg j_{\max} \log(N/m)$, which is true whenever $m^2 \ll N/\log^2(N)$. If moreover $\varepsilon \gg 1/\sqrt{\log(N)}$, then the last probability is bounded from above by

$$
\max_{j > j_{\max}} \exp\left(-c' \min\left(t_\varepsilon^{1/j} \left(\frac{N}{m}\right)^{\frac{j-1}{j}}, \left(\frac{t_\varepsilon}{\|h\|_\infty^2}\right)^{\frac{1}{j+1}} \left(\frac{Nj}{m^2}\right)^{\frac{j}{j+1}} \right)\right).
$$

Now, to estimate the middle term (the probability involving the terms indexed by $j_* + 1 \leq j \leq j_{\max}$), we apply Theorem 4.1 to each term individually for $t = t_{j,\varepsilon} := \frac{\sqrt{t_\varepsilon}}{3j^2} \left(\frac{N}{m}\right)^{(j-1)/2}$, keeping in mind that $\sum_{j \geq j_*+1} t_{j,\varepsilon} \leq \frac{\pi^2}{18} \left(\frac{N}{m}\right)^{(j-1)/2} \sqrt{t_\varepsilon}$. Note that for any $2 \leq t \leq N/m$, $\varepsilon > m/N$ and $j \geq \lfloor \log(N/m) \rfloor + 1$,

$$
\min\left(\frac{t_{j,\varepsilon}^{2/j}}{c}, \left(\frac{t_{j,\varepsilon}\sqrt{N/j}}{(cMj^{\gamma_1-1/2})^j}\right)^{\frac{1}{1+j(2\gamma_2+1)}}\right) \geq \frac{c_1}{M^{2/(1+2\gamma_2)}} \left(\frac{N}{m}\right)^{\frac{1}{1+2\gamma_2}},
$$

whence

$$
\mathbb{P}\left(\left| \sum_{j=j_*+1}^{j_{\max}} \frac{\binom{m}{j}}{\binom{N}{j}} \sum_{J \in \mathcal{A}_N^{(j)}} h_m^{(j)}(X_i, i \in J) \right| \geq \frac{\sqrt{t_\varepsilon}}{3} \sqrt{\frac{m}{N}} \right)
$$

$$
\leq j_{\max} \exp\left(-\frac{c_1}{M^{2/(1+2\gamma_2)}} \left(\frac{N}{m}\right)^{\frac{1}{1+2\gamma_2}}\right) \leq \exp\left(-\frac{c_2}{M^{2/(1+2\gamma_2)}} \left(\frac{N}{m}\right)^{\frac{1}{1+2\gamma_2}}\right).
$$

Finally, to estimate the first term in the right-hand side of inequality (4.2), we again apply Theorem 4.1. With $t_{j,\varepsilon}$ defined as above,

$$
\mathbb{P}\left(\left| \sum_{j=2}^{j_*} \frac{\binom{m}{j}}{\binom{N}{j}} \sum_{J \in \mathcal{A}_N^{(j)}} h_m^{(j)}(X_i, i \in J) \right| \geq \frac{\sqrt{t_\varepsilon}}{3} \sqrt{\frac{m}{N}} \right)
$$

$$
\leq \sum_{j=2}^{j_*} \mathbb{P}\left(\left| \frac{\binom{m}{j}}{\binom{N}{j}} \sum_{J \in \mathcal{A}_N^{(j)}} h_m^{(j)}(X_i, i \in J) \right| \geq \frac{6}{\pi^2} \frac{\sqrt{t_{j,\varepsilon}}}{3} \sqrt{\frac{m}{N}} \right)
$$

$$\leq \sum_{j=2}^{j_*} \exp\left(-c \min\left(t_\varepsilon^{1/j}\left(\frac{N}{m}\right)^{\frac{j-1}{j}}, \frac{t_\varepsilon^{1/(1+j(1+2\gamma_2))}}{M^{2j/(1+j(1+2\gamma_2))}}\left(\frac{N}{m}\right)^{\frac{j}{1+j(1+2\gamma_2)}}\right)\right)$$

$$\leq j_* \max_{2\leq j\leq j_*} \exp\left(-c \min\left(t_\varepsilon^{1/j}\left(\frac{N}{m}\right)^{\frac{j-1}{j}}, \frac{t_\varepsilon^{1/(1+j(1+2\gamma_2))}}{M^{2j/(1+j(1+2\gamma_2))}}\left(\frac{N}{m}\right)^{\frac{j}{1+j(1+2\gamma_2)}}\right)\right).$$

Whenever $\varepsilon \geq \frac{1}{\sqrt{N/m}}$, the last expression is upper bounded by

$$\max_{2\leq j\leq j_*} \exp\left(-c_3 \min\left(t_\varepsilon^{1/j}\left(\frac{N}{m}\right)^{\frac{j-1}{j}}, \frac{t_\varepsilon^{1/(1+j(1+2\gamma_2))}}{M^{2j/(1+j(1+2\gamma_2))}}\left(\frac{N}{m}\right)^{\frac{j}{1+j(1+2\gamma_2)}}\right)\right)$$

for c_3 small enough. Combining all the estimates, we obtain the inequality

$$\mathbb{P}\left(\left|\sum_{j=2}^{m} \frac{\binom{m}{j}}{\binom{N}{j}} \sum_{J\in\mathcal{A}_N^{(j)}} h_m^{(j)}(X_i,\, i\in J)\right| \geq \sqrt{t_\varepsilon}\sqrt{\frac{m}{N}}\right)$$

$$\leq \max_{2\leq j\leq j_*} \exp\left(-c_3 \min\left(t_\varepsilon^{1/j}\left(\frac{N}{m}\right)^{\frac{j-1}{j}}, t_\varepsilon^{\frac{1}{1+j(1+2\gamma_2)}}\left(\frac{N}{mM^2}\right)^{\frac{j}{1+j(1+2\gamma_2)}}\right)\right)$$

$$+ \exp\left(-c_2\left(\frac{N}{mM^2}\right)^{\frac{1}{1+2\gamma_2}}\right) + c_4 \mathrm{Var}(h_m)\exp\left(-j_{\max}\log\left(\frac{N}{m}\right)+\log(t_\varepsilon)\right) \quad (4.3)$$

that holds if $\varepsilon \geq \frac{1}{\sqrt{N/m}}$ and $2 \leq t \leq \frac{N}{m}$. If $t < (\frac{N}{mM^2})^{1/(1+2\gamma_2)}\varepsilon^4$, then the first two terms on the right-hand side of the previous display are bounded by $e^{-ct_\varepsilon/\varepsilon^3} = e^{-ct/\varepsilon}$ each, and if $t < \varepsilon(j_{\max}-1)\log(N/m)$, the same is true for the last term. Therefore, if

$$t < \varepsilon^4 \min\left(\left(\frac{N}{mM^2}\right)^{\frac{1}{1+2\gamma_2}}, (j_{\max}-1)\log\left(\frac{N}{m}\right)\right),$$

then

$$\mathbb{P}\left(\left|\sum_{j=2}^{m} \frac{\binom{m}{j}}{\binom{N}{j}} \sum_{J\in\mathcal{A}_N^{(j)}} h_m^{(j)}(X_i,\, i\in J)\right| \geq \sqrt{t_\varepsilon}\sqrt{\frac{m}{N}}\right)$$

$$\leq 3\exp\left(-\frac{ct}{\varepsilon}\right) = \exp\left(-\frac{t}{2\,\mathrm{Var}(\sqrt{m}\,h_m^{(1)}(X_1))}\right)\cdot o(1),$$

where the last equality holds whenever we choose $\varepsilon := \varepsilon(N,m)$ such that $\varepsilon(N,m)\to 0$ as $N/m \to \infty$. Specifically, take

$$\varepsilon = \left(\frac{q(N,m)}{\min\left(\left(\frac{N}{mM^2}\right)^{1/(1+2\gamma_2)},\, j_{\max}\log(N/m)\right)}\right)^{1/4},$$

where the function $q(N,m)$ was defined in the statement of the corollary, and the conclusion follows immediately. If $m^2 \ll N/\log^2(N)$, we can replace the last term

in equation (4.3) by

$$\max_{j > j_{\max}} \exp\left(-c' \min\left(t_\varepsilon^{1/j} \left(\frac{N}{m}\right)^{\frac{j-1}{j}}, \left(\frac{t_\varepsilon}{\|h\|_\infty^2}\right)^{\frac{1}{j+1}} \left(\frac{Nj}{m^2}\right)^{\frac{j}{j+1}}\right)\right),$$

which is bounded by $e^{-ct/\varepsilon}$ whenever $t < \frac{Nj_{\max}}{m^2}\varepsilon^4$. The final result in this case follows similarly. ∎

5. Implications for the median of means estimator

We are going to apply the results of the previous section to deduce non-asymptotic bounds for the permutation-invariant version of the median of means estimator. Recall that it was defined as

$$\widehat{\mu}_N := \mathrm{med}(\bar{X}_J, \, J \in \mathcal{A}_N^{(m)}).$$

Theorem 5.1. *Assume that $X_1, \ldots, X_N$ are i.i.d. copies of a random variable X with mean μ and variance σ^2. Moreover, suppose that*

(i)   *the distribution of X_1 is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$ with density ϕ_1,*

(ii)   *the Fourier transform $\widehat{\phi}_1$ of the density satisfies the inequality $|\widehat{\phi}_1(x)| \leq C_1/(1 + |x|)^\delta$ for some positive constants C_1 and δ,*

(iii) $\mathbb{E}|(X_1 - \mu)/\sigma|^q < \infty$ *for some* $\frac{3+\sqrt{5}}{2} < q \leq 3$.

Then the estimator $\widehat{\mu}_N$ satisfies

$$\mathbb{P}\left(|\sqrt{N}(\widehat{\mu} - \mu)| \geq \sigma \sqrt{t}\right) \leq 2 \exp\left(-\frac{t}{2(1 + o(1))}\right),$$

where $o(1) \to 0$ as $m, N/m \to \infty$ uniformly for all $t \in [l_{N,m}, u_{N,m}]$ for any sequences $\{l_{N,m}\}, \{u_{N,m}\}$ such that $l_{N,m} \gg N/m^{q-1}$ and $u_{N,m} \ll \frac{N}{m^{q/(q-1)} \vee m \log^2(N)}$.

Remark 5.2. (1) Let us recall the Riemann–Lebesgue lemma stating that $|\widehat{\phi}_1(x)| \to 0$ as $|x| \to \infty$ for any absolutely continuous distribution, so assumption (ii) is rather mild.

(2) The inequality $q > \frac{3+\sqrt{5}}{2}$ assures that $l_{N,m}$ and $u_{N,m}$ can be chosen such that $l_{N,m} \ll u_{N,m}$.

Proof. Throughout the course of the proof, we will assume without loss of generality that $\sigma^2 = 1$; the general case follows by rescaling. Let us also recall that all asymptotic relations are defined in the limit as both m and N/m tend to infinity. Note that direct application of Corollary 4.2 requires existence of all moments of X_1, which is too prohibitive. Therefore, we will first show how to reduce the problem to the case of

bounded random variables. Specifically, we want to truncate $X_j - \mu$, $j = 1, \ldots, N$, in a way that preserves the decay rate of the characteristic function. To this end, let R be a large constant (that will later be specified as an increasing function of m), and define the standard mollifier $\kappa(x)$ via

$$\kappa(x) = \begin{cases} C_1 \exp\left(-\frac{1}{1-x^2}\right), & |x| < 1, \\ 0, & |x| \geq 1, \end{cases}$$

where C_1 is chosen so that $\int_{\mathbb{R}} \kappa(x) = 1$. Moreover, let $\chi_R(x) = (I_{2R} * \kappa_R)(x)$ be the smooth approximation of the indicator function of the interval $[-2R, 2R]$, where $I_{2R}(x) = I\{|x| \leq 2R\}$ and $\kappa_R(x) = \frac{1}{R}\kappa\left(\frac{x}{R}\right)$; in particular, $\chi_R(x) = 1$ for $|x| \leq R$ and $\chi_R(x) = 0$ for $|x| \geq 3R$. Set

$$\psi(x) = C_2 \phi_1(x + \mu) \chi_R(x)$$

where $C_2 > 0$ is such that $\int_{\mathbb{R}} \psi(x)dx = 1$. Suppose that $Y^{(R)}$ has distribution with density ψ and note that by construction the laws of $X_1 - \mu$ and $Y^{(R)}$, conditionally on the events $\{|X_1 - \mu| \leq R\}$ and $\{|Y^{(R)}| \leq R\}$, respectively, coincide. Therefore, there exists a random variable Z independent from X_1 such that

$$Y_1^{(R)} := \begin{cases} X_1 - \mu, & |X_1 - \mu| \leq R, \\ Z, & |X_1 - \mu| > R, \end{cases} \tag{5.1}$$

also has density ψ. Observe the following properties of $Y_1^{(R)}$:

(a) $|Y_1^{(R)}| \leq 3R$ almost surely.

(b) $\mathbb{E}h(Y_1^{(R)}) \leq C_2\mathbb{E}h(X_1 - \mu)$ for any nonnegative function h. Indeed, this follows from the inequality $\psi(x) \leq C_2\phi_1(x + \mu)$.

(c) $|\mathbb{E}Y_1^{(R)}| \leq (1 + C_2)\frac{\mathbb{E}|X_1 - \mu|^q I\{|X_1 - \mu| > R\}}{R^{q-1}}$. Indeed,

$$\begin{aligned} |\mathbb{E}Y_1^{(R)}| &= \left|\mathbb{E}Y_1^{(R)}I\{|X_1 - \mu| \leq R\} + \mathbb{E}Y_1^{(R)}I\{|X_1 - \mu| > R\}\right| \\ &= \left|\mathbb{E}(\mu - X_1)I\{|X_1 - \mu| > R\} + \mathbb{E}Y_1^{(R)}I\{|Y_1^{(R)}| > R\}\right| \\ &\leq \mathbb{E}|X_1 - \mu|I\{|X_1 - \mu| > R\} + C_2\mathbb{E}|X_1 - \mu|I\{|X_1 - \mu| > R\}, \end{aligned}$$

where the last bound follows from property (b) for $h(x) = |x|I\{|x| > R\}$. It remains to apply Hölder's and Markov's inequalities.

The final property of $Y_1^{(R)}$ is stated in the lemma below and is proven in Section .

Lemma 5.3. *The characteristic function $\widehat{\psi}(x)$ of $Y_1^{(R)}$ satisfies*

$$|\widehat{\psi}_1(x)| \leq \frac{C}{(1 + |x|)^\delta}$$

for all $x \in \mathbb{R}$ and a sufficiently large constant C.

Define $\rho(x) = |x|$. Proceeding as in the proof of Theorem 2.1, we observe that

$$\mathbb{P}(\sqrt{N}(\hat{\mu} - \mu) \geq \sqrt{t}) \leq \mathbb{P}\left(\frac{\sqrt{N/m}}{\binom{N}{m}} \sum_{J \in \mathcal{A}_N^{(m)}} \rho'_-(\sqrt{m}(\bar{X}_J - \mu - \sqrt{t/N})) \geq 0\right). \quad (5.2)$$

Our next goal is to show that for sufficiently large R, the U-statistic with kernel ρ'_- appearing in (5.2) and evaluated at $X_1, \ldots, X_N$ can be replaced by the U-statistic evaluated over an i.i.d. sample $Y_1^{(R)}, \ldots, Y_N^{(R)}$ where $Y_j^{(R)}$ is related to X_j according to (5.1). To this end, recall that $\mathbb{E}|X_1 - \mu|^q < \infty$, and choose R as $R = cm^{1/(2(q-1))}$ for some $c > 0$. Next, observe that

$$\sum_{J \in \mathcal{A}_N^{(m)}} \rho'_-(\sqrt{m}(\bar{X}_J - \mu - \sqrt{t/N}))$$

$$= \sum_{J \in \mathcal{A}_N^{(m)}} \left(\rho'_-(\sqrt{m}(\bar{Y}_J^{(R)} - \sqrt{t/N})) - \mathbb{E}\rho'_{-,R} + \mathbb{E}\rho'_-\right)$$

$$+ \sum_{J \in \mathcal{A}_N^{(m)}} \left(\rho'_-(\sqrt{m}(\bar{X}_J - \mu - \sqrt{t/N}))\right.$$

$$\left. - \rho'_-(\sqrt{m}(\bar{Y}_J^{(R)} - \sqrt{t/N})) - \mathbb{E}\rho'_- + \mathbb{E}\rho'_{-,R}\right), \quad (5.3)$$

where $\mathbb{E}\rho'_- = \rho'_-(\sqrt{m}(\bar{X}_J - \mu - \sqrt{t/N}))$ and $\mathbb{E}\rho'_{-,R} = \mathbb{E}\rho'_-(\sqrt{m}(\bar{Y}_J^{(R)} - \sqrt{t/N}))$. It was shown in the proof of Theorem 2.1 that

$$\sqrt{\frac{N}{m}} \mathbb{E}\rho'_- \leq C\sqrt{k} \cdot g(m) - \sqrt{t}\left(\sqrt{\frac{2}{\pi}} + O\left(\sqrt{\frac{t}{k}}\right)\right) = -\sqrt{t}\sqrt{\frac{2}{\pi}}(1 + o(1))$$

whenever $t \ll N/m$ and $t \gg \frac{N}{m}g^2(m)$. Let us remark that in view of imposed moment assumptions, $g(m) = O(m^{-(q-2)/2})$. Moreover, it follows from Hoeffding's version of Bernstein's inequality for U-statistics [15] that

$$\frac{\sqrt{N/m}}{\binom{N}{m}} \sum_{J \in \mathcal{A}_N^{(m)}} \left(\rho'_-(\sqrt{m}(\bar{X}_J - \mu - \sqrt{t/N}))\right.$$

$$\left. - \rho'_-(\sqrt{m}(\bar{Y}_J^{(R)} - \sqrt{t/N})) - \mathbb{E}\rho'_- + \mathbb{E}\rho'_{-,R}\right)$$

$$\leq 2\mathbb{E}^{1/2}\left(\rho'_-(\sqrt{m}(\bar{X}_{[m]} - \mu - \sqrt{t/N})) - \rho'_-(\sqrt{m}(\bar{Y}_{[m]}^{(R)} - \sqrt{t/N}))\right)^2 \sqrt{s} \vee \frac{16s}{3}\sqrt{\frac{m}{N}}$$

with probability at least $1 - e^{-s}$. We want to choose $s > 0$ such that $t = o(s)$ and

$$\alpha(s, R) := 2\mathbb{E}^{1/2}\left(\rho'_-(\sqrt{m}(\bar{X}_{[m]} - \mu - \sqrt{t/N}))\right.$$

$$\left. - \rho'_-(\sqrt{m}(\bar{Y}_{[m]}^{(R)} - \sqrt{t/N}))\right)^2 \sqrt{s} \vee \frac{16s}{3}\sqrt{\frac{m}{N}} = o(\sqrt{t}) \quad (5.4)$$

as $m, N/m \to \infty$. To estimate

$$\Sigma_m^2 := \mathbb{E}\big(\rho'_-(\sqrt{m}(\bar{X}_{[m]} - \mu - \sqrt{t/N})) - \rho'_-(\sqrt{m}(\bar{Y}_{[m]}^{(R)} - \sqrt{t/N}))\big)^2,$$

note that for any $a > 0$, $\rho'_-(\sqrt{m}(\bar{X}_{[m]} - \mu - \sqrt{t/N})) = \rho'_-(\sqrt{m}(\bar{Y}_{[m]}^{(R)} - \sqrt{t/N}))$ whenever we have $|\sqrt{m}(\bar{Y}_{[m]}^{(R)} - \sqrt{t/N})| > a/2$, $|\sqrt{m}(\bar{X}_{[m]} - \mu - \sqrt{t/N})| > a/2$ and $|\sqrt{m}(\bar{X}_{[m]} - \mu - \bar{Y}_{[m]}^{(R)})| \le a$, hence

$$\Sigma_m^2 \le 4\big(\mathbb{P}\big(|\sqrt{m}(\bar{Y}_{[m]}^{(R)} - \sqrt{t/N})| \le a\big) + \mathbb{P}\big(|\sqrt{m}(\bar{X}_{[m]} - \mu - \sqrt{t/N})| \le a\big)\big)$$
$$+ 4\mathbb{P}\big(|\sqrt{m}(\bar{X}_{[m]} - \mu - \bar{Y}_{[m]}^{(R)})| > a\big).$$

Up to the additive error term $Cg(m) = O(m^{-(q-2)/2})$, the distributions of $\sqrt{m}\,\bar{X}_{[m]}$ and $\sqrt{m}\,\bar{Y}_{[m]}^{(R)}$ can be approximated by the normal distribution, hence

$$\mathbb{P}\big(|\sqrt{m}(\bar{Y}_{[m]}^{(R)} - \sqrt{t/N})| \le a\big) + \mathbb{P}\big(|\sqrt{m}(\bar{X}_{[m]} - \mu - \sqrt{t/N})| \le a\big) \le C(a + g(m)).$$

Moreover, letting $\mathcal{E}_1$ denote the event

$$\mathcal{E}_1 = \Big\{\Big|\frac{1}{\sqrt{m}}\sum_{j=1}^m Y_j^{(R)} I\{|Y_j^{(R)}| > R\} - \mathbb{E}\big(Y_j^{(R)} I\{|Y_j^{(R)}| > R\}\big)\Big|$$
$$\ge a - \sqrt{m}\big|\mathbb{E}Y_j^{(R)} I\{|Y_j^{(R)}| > R\}\big|\Big\},$$

we deduce that

$$\mathbb{P}\big(|\sqrt{m}(\bar{X}_{[m]} - \mu - \bar{Y}_{[m]}^{(R)})| > a\big) = \mathbb{P}\Big(\Big|\frac{1}{\sqrt{m}}\sum_{j=1}^m Y_j^{(R)} I\{|Y_j^{(R)}| > R\}\Big| \ge a\Big)$$

$$\le \mathbb{P}(\mathcal{E}_1) \le C_2 \frac{\mathbb{E}|X_1 - \mu|^2 I\{|X_1 - \mu| > R\}}{(a - C_2\sqrt{m}|\mathbb{E}(X_1 - \mu)I\{|X_1 - \mu| > R\}|)^2}$$
$$\le \frac{\mathbb{E}|X_1 - \mu|^q I\{|X_1 - \mu| > R\}}{R^{q-2}(a - C_2\sqrt{m}|\mathbb{E}(X_1 - \mu)I\{|X_1 - \mu| > R\}|)^2}, \tag{5.5}$$

where we used property (b) of $Y_1^{(R)}$ along with Hölder's and Markov's inequalities. It is also clear that

$$\sqrt{m}|\mathbb{E}(X_1 - \mu)I\{|X_1 - \mu| > R\}| \le \frac{\sqrt{m}\,\mathbb{E}|X_1 - \mu|^q I\{|X_1 - \mu| > R\}}{R^{q-1}},$$

thus, for $R = cm^{\frac{1}{2(q-1)}}$ specified before, $\sqrt{m}|\mathbb{E}(X_1 - \mu)I\{|X_1 - \mu| > R\}| = o(1)$. Setting $a = 2C_2\frac{\sqrt{m}\,\mathbb{E}^{1/2}|X_1-\mu|^q I\{|X_1-\mu|>R\}}{R^{q-1}}$, one easily checks that the right-hand side in (5.5) is at most $CR^{-(q-2)} = C'm^{-\frac{q-2}{2(q-1)}}$. whence $\Sigma_m^2 = o(1)$. Therefore, there exists a function $o(1)$ such that setting $s = t/o(1)$ yields the stated goal, namely, that $t = o(s)$ and $\alpha(s, R) = o(\sqrt{t})$ where $\alpha(s, R)$ was defined in (5.4). Combined

with (5.3), this implies that

$$
\mathbb{P}\big(\sqrt{N}(\hat{\mu} - \mu) \geq \sqrt{t}\big)
$$
$$
\leq o(1) \cdot e^{-t} + \mathbb{P}\bigg(\frac{\sqrt{N/m}}{\binom{N}{m}} \sum_{J \in \mathscr{A}_N^{(m)}} \big(\rho'_-(\sqrt{m}(\bar{Y}_J^{(R)} - \sqrt{t/N})) - \mathbb{E}\rho'_{-,R}\big)
$$
$$
\geq \sqrt{t}\sqrt{\frac{2}{\pi}}(1 + o(1))\bigg). \tag{5.6}
$$

Note that the U-statistic in the display above is now a function of bounded random variables, hence we can apply Corollary 4.2 with $\gamma_2 = 0$. As $\|\rho'_-\|_\infty = 1$, condition (ii) of the corollary holds. Let $\sqrt{m/N} \sum_{j=1}^N h^{(1)}(Y_j^{(R)})$ be the first term in Hoeffding decomposition of the U-statistic

$$
\frac{\sqrt{N/m}}{\binom{N}{m}} \sum_{J \in \mathscr{A}_N^{(m)}} \big(\rho'_-(\sqrt{m}(\bar{Y}_J^{(R)} - \mathbb{E}Y_1^{(R)} - \sqrt{t/N} + \mathbb{E}Y_1^{(R)})) - \mathbb{E}\rho'_{-,R}\big).
$$

Following the lines of the proof of Theorem 3.2 and recalling that $\sqrt{m}|\mathbb{E}Y_1^{(R)}| = o(1)$ in view of property (c) of $Y_1^{(R)}$ and the choice of R, we deduce that

$$
\mathrm{Var}\big(\sqrt{m}\, h^{(1)}(Y_1^{(R)})\big) = \frac{2}{\pi}(1 + o(1)),
$$

where $o(1) \to 0$ as $m, N/m \to \infty$, validating assumption (iii) of the corollary. It remains to verify assumption (i) and specify the value of $j_{\max}$. Recall that $\rho'_-(x) = I\{x \geq 0\} - I\{x < 0\}$ and let $\widetilde{Y}_j^{(R)}$ stand for $Y_j^{(R)} - \mathbb{E}Y_j^{(R)}$. Therefore, the function $f_j(u_1, \ldots, u_j)$ appearing in the statement of Theorem 4.1 can be expressed as

$$
f_j(u_1, \ldots, u_j)
$$
$$
= \mathbb{E}\rho'_-\bigg(\frac{1}{\sqrt{m}} \sum_{i=1}^j u_i + \sqrt{\frac{m-j}{m}} \frac{\sum_{i=j+1}^m \widetilde{Y}_i^{(R)}}{\sqrt{m-j}} - \sqrt{\frac{tm}{N}} + \sqrt{m}\, \mathbb{E}Y_1^{(R)} \bigg)
$$
$$
= \mathbb{E}\rho'_-\bigg(\frac{1}{\sqrt{m}} \sum_{i=1}^j u_i + \sqrt{\frac{m-j}{m}} \frac{\sum_{i=j+1}^m \widetilde{Y}_i^{(R)}}{\sqrt{m-j}} - \sqrt{\frac{tm}{N}} + \sqrt{m}\, \mathbb{E}Y_1^{(R)} \bigg)
$$
$$
= 2\Phi_{m-j}\bigg(\frac{1}{\sqrt{m-j}} \sum_{i=1}^j u_i - \sqrt{\frac{m}{m-j}}\bigg(\sqrt{\frac{tm}{N}} + \sqrt{m}\, \mathbb{E}Y_1^{(R)}\bigg) \bigg) - 1,
$$

where for any integer $k \geq 1$, Φ_k stands for the cumulative distribution function of $\frac{1}{\sqrt{k}} \sum_{j=1}^k \widetilde{Y}_j^{(R)}$ and ϕ_k is the corresponding density function that exists by assumption. Consequently,

$$
\partial_{u_j} \ldots \partial_{u_1} f_1(u_1, \ldots, u_j)
$$
$$
= \frac{2}{(m-j)^{j/2}} \phi_{m-j}^{(j-1)}\bigg(\frac{1}{\sqrt{m-j}} \sum_{i=1}^j u_i - \sqrt{\frac{m}{m-j}}\bigg(\sqrt{\frac{tm}{N}} + \sqrt{m}\, \mathbb{E}Y_1^{(R)}\bigg) \bigg).
$$

The following lemma demonstrates that Theorem 4.1 applies with $\gamma_1 = 1/2$ and that $j_{\max} = \frac{m}{\log(m)} o(1)$ in the statement of Corollary 4.2.

Lemma 5.4. *Let assumptions of Theorem 5.1 hold. Then, for m large enough and $j = o(m/\log m)$,*

$$\|\phi_{m-j}^{(j-1)}\|_\infty \leq C\left(\frac{2j}{e}\right)^{j/2}$$

for a sufficiently large constant $C = C(P)$.

We postpone the proof of this lemma until Section 7.6. As all the necessary conditions have been verified, the bound of Corollary 4.2 applies. Recalling that $t \gg \frac{N}{m} g^2(m)$ and that $g(m) \leq C\mathbb{E}|X_1 - \mu|^q/m^{(q-2)/2}$, Corollary 4.2 yields that the probability in the right-hand side of inequality (5.6) can be bounded from above by $\exp(-\frac{t}{2\sigma^2(1+o(1))})$ for all

$$\frac{N}{m^{q-1}} \ll t \leq q(N, m) \tag{5.7}$$

whenever

$$q(N, m) = \min\left(\frac{N}{mR^2}, \frac{N}{m \log^2(N)}\right) \cdot o(1) \quad \text{as } N/m \to \infty.$$

To get the expression for the second term in the minimum above from the bound of the corollary, it suffices to consider the cases when $m \geq \frac{\sqrt{N}}{\log(N)} o(1)$ and $m \leq \frac{\sqrt{N}}{\log(N)} o(1)$ separately; we omit the simple algebra. Since $R = cm^{1/(2(q-1))}$, (5.7) in only possible when $q - 2 > \frac{1}{q-1}$, implying the requirement $q > \frac{3+\sqrt{5}}{2}$. The final form of the bound stating that

$$\mathbb{P}\left(\sqrt{N}(\hat{\mu} - \mu) \geq \sigma\sqrt{t}\right) \leq \exp\left(-\frac{t}{2(1 + o(1))}\right)$$

uniformly for all $\frac{N}{m^{q-1}} \ll t \ll \frac{N}{m^{q/(q-1)}\vee m \log^2(N)}$. The argument needed to estimate the lower tail $\mathbb{P}(\sqrt{N}(\hat{\mu} - \mu) \leq -\sigma\sqrt{t})$ is identical. ∎

6. Open questions

Several potentially interesting questions and directions have not been addressed in this paper. We summarize few of them below.

(i) First is the question related to assumptions in Theorem 5.1: Does it still hold for distributions with only $2 + \varepsilon$ moments? And can the assumptions requiring absolute continuity and a bound on the rate of decay of the characteristic function be dropped? For example, Corollary 3.2 holds for lattice distributions as well.

(ii)   It is known (see [12]) that the sample mean based on i.i.d. observations from the multivariate normal distribution $N(\mu, \Sigma)$ satisfies the inequality

$$\|\bar{X}_N - \mu\|_2 \leq \sqrt{\frac{\text{trace}(\Sigma)}{N}} + \sqrt{\frac{2t\|\Sigma\|}{N}}$$

with probability at least $1 - e^{-t}$. Does there exist an estimator of the mean that achieves this bound (up to $o(1)$ factors) for the heavy-tailed distributions? Partial results in this direction have been recently obtained in [19].

(iii)   Exact computation of the estimator $\hat{\mu}_N$ is infeasible, as it requires evaluation and sorting of $\asymp (N/m)^m$ sample means. Therefore, it is interesting to understand whether it can be replaced by $\text{med}(\bar{X}_J, J \in \mathcal{B})$, where $\mathcal{B}$ is a (deterministic or random) subset of $\mathcal{A}_N^{(m)}$ of much smaller cardinality, while preserving the deviation guarantees. For instance, it is easy to deduce from results on incomplete U-statistics in [18, Section 4.3], combined with the proof of Corollary 3.2, that if $\mathcal{B}$ consists of M subsets selected at random with replacement from $\mathcal{A}_N^m$, then the asymptotic distribution of $\sqrt{N}(\text{med}(\bar{X}_J, J \in \mathcal{B}) - \mu)$ is still $N(0, \sigma^2)$ as long as $M \gg N$. However, establishing results in spirit of Theorem 5.1 in this framework appears to be more difficult.

7. Remaining proofs

The proofs omitted in the main text are presented in this section.

7.1. Technical tools

Let us recall the definition of Hoeffding's decomposition [14] and closely related concepts that are at the core of many arguments related to U-statistics. Assume that $Y_1, \ldots, Y_N$ are i.i.d. random variables with distribution P_Y. Recall that $\mathcal{A}_N^{(m)} = \{J \subseteq [N] : |J| = m\}$ and that the U-statistic with permutation-symmetric kernel h_m is defined as

$$U_{N,m} = \frac{1}{\binom{N}{m}} \sum_{J \in \mathcal{A}_N^{(m)}} h_m(Y_i, i \in J),$$

where we assume that $\mathbb{E}h_m = 0$. Moreover, for $j = 1, \ldots, m$, define the projections

$$(\pi_j h_m)(y_1, \ldots, y_j) := (\delta_{y_1} - P_Y) \times \cdots \times (\delta_{y_j} - P_Y) \times P_Y^{m-j} h_m. \tag{7.1}$$

For brevity and to ease notation, we will often write $h_m^{(j)}$ in place of $\pi_j h_m$. The variances of these projections will be denoted by

$$\delta_j^2 := \text{Var}(h_m^{(j)}(Y_1, \ldots, Y_j)).$$

In particular, $\delta_m^2 = \mathrm{Var}(h_m)$. It is well known that $h_m^{(j)}$ can be viewed geometrically as orthogonal projections of h_m onto a particular subspace of $L_2(P_Y^m)$, see [18]. The kernels $h_m^{(j)}$ have the property of complete degeneracy, meaning that

$$\mathbb{E} h_m^{(j)}(y_1, \ldots, y_{j-1}, Y_j) = 0$$

for P_Y-almost all $y_1, \ldots, y_{j-1}$ while $h_m^{(j)}(Y_1, \ldots, Y_j)$ is non-zero with positive probability. One can easily check that $h(y_1, \ldots, y_m) = \sum_{j=1}^m \sum_{J \subseteq [m]:|J|=j} h_m^{(j)}(y_i, i \in J)$, in particular, the partial sum $\sum_{j=1}^k \sum_{J \subseteq [m]:|J|=j} h_m^{(j)}(y_i, i \in J)$ is the best approximation of h_m, in the mean-squared sense, in terms of sums of functions of at most k variables. The Hoeffding decomposition states that (see [14] as well as [18])

$$U_{N,m} = \sum_{j=1}^m \binom{m}{j} U_{N,m}^{(j)}, \tag{7.2}$$

where $U_{N,m}^{(j)}$ are U-statistics with kernels $h_m^{(j)}$, namely,

$$U_{N,m}^{(j)} := \frac{1}{\binom{N}{j}} \sum_{J \in \mathcal{A}_N^{(j)}} h_m^{(j)}(Y_i, i \in J).$$

Moreover, all terms in representation (7.2) are uncorrelated.

Next, we recall some useful moment bounds, found for instance in the book by de la Peña and Giné [6], for the Rademacher chaos variables. Let $\varepsilon_1, \ldots, \varepsilon_N$ be i.i.d. Rademacher random variables (random signs), $\{a_J, \ J \in \mathcal{A}_N^{(l)}\} \subset \mathbb{R}$, and $Z = \sum_{J \in \mathcal{A}_N^{(l)}} a_J \prod_{i \in J} \varepsilon_i$. Here, $\prod_{i \in J} \varepsilon_i = \varepsilon_{i_1} \cdots \varepsilon_{i_l}$ for $J = \{i_1, \ldots, i_l\}$.

Fact 1 (Bonami inequality). Let $\sigma^2(Z) = \mathrm{Var}(Z) = \sum_{J \in \mathcal{A}_N^{(l)}} a_J^2$. Then for any $q > 2$,

$$\mathbb{E}|Z|^q \leq (q-1)^{ql/2}(\sigma^2(Z))^{q/2}.$$

Now we state a version of the symmetrization inequality for completely degenerate U-statistics due to Sherman [33], also see the paper by Song, Chen and Kato [34] for the modern exposition of the proof. The main feature of this inequality, put forward in [34], is the fact that its proof does not rely on decoupling, and yields constants that do not grow too fast with the order of U-statistics.

Fact 2. Let h be a completely degenerate kernel of order l, and let Φ be a convex, nonnegative, non-decreasing function. Moreover, assume that $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. Rademacher random variables. Then

$$\mathbb{E}\Phi\left(\sum_{1 \leq j_1 < \cdots < j_l \leq N} h(Y_{j_1}, \ldots, Y_{j_l})\right) \leq \mathbb{E}\Phi\left(2^l \sum_{1 \leq j_1 < \cdots < j_l \leq N} \varepsilon_{j_1} \cdots \varepsilon_{j_l} h(Y_{j_1}, \ldots, Y_{j_l})\right).$$

Next is the well-known identity, due to Hoeffding [15], that allows to reduce many problems for non-degenerate U-statistics to the corresponding problems for the sums of i.i.d. random variables.

Fact 3. The following representation holds:

$$U_{N,m} = \frac{1}{N!} \sum_\pi W_\pi,$$

where the sum is over all permutations $\pi \colon [N] \mapsto [N]$, and

$$W_\pi = \frac{1}{k}\big(h_m(Y_{\pi(1)}, Y_{\pi(2)}, \ldots, Y_{\pi(m)})$$
$$+ \cdots + h_m(Y_{\pi((k-1)m+1)}, Y_{\pi((k-1)m+2)}, \ldots, Y_{\pi(km)})\big)$$

for $k = \lfloor N/m \rfloor$.

Finally, we state a version of Rosenthal's inequality for the moments of sums of independent, nonnegative random variables with explicit constants, see [3, 5].

Fact 4. Let $Y_1, \ldots, Y_N$ be independent random variables such that $Y_j \geq 0$ with probability 1 for all $j \in [N]$. Then for any $q \geq 1$,

$$\left(\mathbb{E}\left|\sum_{j=1}^N Y_j\right|^q\right)^{1/q} \leq \left(\left(\sum_{j=1}^N \mathbb{E}Y_j\right)^{1/2} + 2\sqrt{eq}\big(\mathbb{E}\max_{j=1,\ldots,N} Y_j^q\big)^{1/(2q)}\right)^2.$$

7.2. Proof of Theorem 3.1

Recall that

$$h_m^{(j)}(y_1, \ldots, y_j) := (\delta_{y_1} - P_Y) \times \cdots \times (\delta_{y_j} - P_Y) \times P_Y^{m-j} h_m,$$
$$\delta_j^2 := \mathrm{Var}(h_m^{(j)}(Y_1, \ldots, Y_j)).$$

It is easy to verify that

$$h_m(Y_1, \ldots, Y_m) = (\delta_{Y_1} - P_Y + P_Y) \times \cdots \times (\delta_{Y_m} - P_Y + P_Y)h_m$$
$$= \sum_{j=1}^m \sum_{J \subseteq [m]:|J|=j} h_m^{(j)}(Y_i, i \in J),$$

and that the terms in the sum above are mutually orthogonal, yielding that

$$\mathrm{Var}(h_m(Y_1, \ldots, Y_m)) = \sum_{j=1}^m \binom{m}{j}\delta_j^2. \tag{7.3}$$

Moreover, as a corollary of Hoeffding's decomposition, one can get the well-known identities

$$\mathrm{Var}(U_{N,m}) = \sum_{j=1}^{m} \frac{\binom{m}{j}^2}{\binom{N}{j}} \delta_j^2, \quad \mathrm{Var}(S_{N,m}) = \frac{m^2}{N} \delta_1^2,$$

$$\mathrm{Var}(U_{N,m} - S_{N,m}) = \mathrm{Var}(U_{N,m}) - \mathrm{Var}(S_{N,m}) = \sum_{j=2}^{m} \frac{\binom{m}{j}^2}{\binom{N}{j}} \delta_j^2.$$

See [18, Chapters 1.6 and 1.7] for detailed derivations of these facts. The simple but key observation following from equation (7.3) is that for any $j \in [m]$, we have $\mathrm{Var}(h_m) \geq \binom{m}{j} \delta_j^2$, or

$$\delta_j^2 \leq \frac{\mathrm{Var}(h_m)}{\binom{m}{j}}. \tag{7.4}$$

Therefore,

$$\mathrm{Var}(U_{N,m} - S_{N,m}) = \sum_{j=2}^{m} \frac{\binom{m}{j}^2}{\binom{N}{j}} \delta_j^2 \leq \mathrm{Var}(h) \sum_{j=2}^{m} \frac{\binom{m}{j}}{\binom{N}{j}}$$

$$\leq \mathrm{Var}(h) \sum_{j \geq 2} \left(\frac{m}{N}\right)^j = \mathrm{Var}(h) \left(\frac{m}{N}\right)^2 \left(1 - \frac{m}{N}\right)^{-1}, \tag{7.5}$$

where we used the fact that $\binom{m}{j}/\binom{N}{j} \leq (\frac{m}{N})^j$ for $m \leq N$; indeed, the latter easily follows from the identity $\binom{m}{j}/\binom{N}{j} = \frac{m(m-1)\cdots(m-j+1)}{N(N-1)\cdots(N-j+1)}$. It is well known (see [14]) that $\mathrm{Var}(h^{(1)}(Y_1)) \leq \frac{\mathrm{Var}(h_m)}{m}$, therefore the condition $\frac{\mathrm{Var}(h_m(Y_1,\ldots,Y_m))}{\mathrm{Var}(h_m^{(1)}(Y_1))} = o(N)$ imposed on the ratio of variances implies that $m = o(N)$. Therefore, for m, N large enough (so that $m/N \leq 1/2$),

$$\frac{\mathrm{Var}(U_{N,m} - S_{N,m})}{\mathrm{Var}(S_{N,m})} \leq 2 \frac{\mathrm{Var}(h_m)(m/N)^2}{\delta_1^2 m^2/N} = 2 \frac{\mathrm{Var}(h_m)}{N \delta_1^2} = o(1),$$

by assumption, yielding that $\frac{U_{N,m} - S_N}{\mathrm{Var}^{1/2}(S_N)} = o_P(1)$ as $N, m \to \infty$.

7.3. Proof of Theorem 4.1

We are going to estimate $\mathbb{E}|V_{N,j}|^q$ for an arbitrary $q > 2$. It follows from the symmetrization inequality (Fact 2) followed by the moment bound stated in Fact 1 that

$$\mathbb{E}|V_{N,j}|^q \leq 2^{jq} \, \mathbb{E}_X \mathbb{E}_\varepsilon \left| \frac{\binom{m}{j}^{1/2}}{\binom{N}{j}^{1/2}} \sum_{(i_1,\ldots,i_j) \in \mathcal{A}_N^{(j)}} \varepsilon_{i_1} \cdots \varepsilon_{i_j} h_m^{(j)}(X_{i_1}, \ldots, X_{i_j}) \right|^q$$

$$\leq 2^{jq}(q-1)^{jq/2}\mathbb{E}\left|\frac{\binom{m}{j}}{\binom{N}{j}}\sum_{(i_1,\dots,i_j)\in\mathcal{A}_N^{(j)}}(h_m^{(j)}(X_{i_1},\dots,X_{i_j}))^2\right|^{q/2}.$$

Next, Hoeffding's representation of the U-statistic (Fact 3), together with Jensen's inequality, yields that

$$\mathbb{E}\left|\frac{\binom{m}{j}}{\binom{N}{j}}\sum_{(i_1,\dots,i_j)\in\mathcal{A}_N^{(j)}}(h_m^{(j)}(X_{i_1},\dots,X_{i_j}))^2\right|^{q/2}\leq\mathbb{E}\left|\frac{\binom{m}{j}}{\lfloor N/j\rfloor}\sum_{i=1}^{\lfloor N/j\rfloor}W_i\right|^{q/2},$$

where $W_i := (h_m^{(j)}(X_{(i-1)j+1},\dots,X_{ij}))^2$. We are going to estimate in two different ways $\mathbb{E}\max_{j=1,\dots,\lfloor N/j\rfloor}W_j^p$. First, recall that

$$h_m^{(j)}(x_1,\dots,x_j):=(\pi_j h_m)(x_1,\dots,x_j)=(\delta_{x_1}-P_X)\times\cdots\times(\delta_{x_j}-P_X)\times P_X^{m-j}h_m.$$

Therefore, $(\pi_j h)(x_1,\dots,x_j)$ is a linear combination of 2^j terms that have the form $\prod_{i\in I}\delta_{x_i}P_X^{m-|I|}h_m$, for all choices of $I\subseteq[j]$. Consequently, $|(\pi_j h_m)(x_1,\dots,x_j)|^2\leq 2^{2j}\|h_m\|_\infty^2$, and the same bound also holds (almost surely) for the maximum of W_j's. This implies that

$$\mathbb{E}\max_{j=1,\dots,\lfloor N/j\rfloor}W_j^p\leq 2^{2jp}\|h_m\|_\infty^{2p}\text{ and }\mathbb{E}\left(\binom{m}{j}W_1\right)^p\leq(2e)^{2jp}\left(\frac{m}{j}\right)^{jp}\|h_m\|_\infty^{2p}.$$

Moreover, equation (7.4) in the proof of Theorem 3.1 implies $\mathbb{E}W_1\leq\text{Var}(h_m)/\binom{m}{j}$. Therefore, Rosenthal's inequality for nonnegative random variables (Fact 4) entails that for $q\geq 2$,

$$\mathbb{E}\left|\frac{\binom{m}{j}}{\lfloor N/j\rfloor}\sum_{i=1}^{\lfloor N/j\rfloor}W_i\right|^{q/2}$$

$$\leq C^{q/2}\left(\text{Var}^{q/2}(h_m+\left(\frac{q}{2}\right)^{q/2}\left(\frac{j}{N}\right)^{q/2}\mathbb{E}\left(\binom{m}{j}\max_{j=1,\dots,\lfloor N/j\rfloor}W_1\right)^{q/2}\right)$$

$$\leq C^{q/2}\left(\text{Var}^{q/2}(h_m)+\left(\frac{q}{2}\right)^{q/2}\left(\frac{j}{N}\right)^{q/2}(2e)^{jq}\left(\frac{m}{j}\right)^{jq/2}\|h_m\|_\infty^q\right)$$

and

$$\mathbb{E}|V_{N,j}|^q\leq(Cq^{1/2})^{qj}\left(\text{Var}^{q/2}(h_m)\vee\left(\left(\frac{qj}{N}\right)^{1/2}\left(\frac{m}{j}\right)^{j/2}\|h_m\|_\infty\right)^q\right).$$

Markov's inequality therefore yields that

$$\mathbb{P}\left(|V_{N,j}|\geq(C_1q)^{j/2}\left(\text{Var}^{1/2}(h_m)\vee\left(\frac{qj}{N}\right)^{1/2}\left(\frac{m}{j}\right)^{j/2}\|h_m\|_\infty\right)\right)\leq e^{-q}.$$

Let $A(q) = (C_1 q)^{j/2} \text{Var}^{1/2}(h_m)$ and $B(q) = \|h_m\|_\infty (qj/N)^{1/2} (C_1 q^{1/2} (m/j)^{1/2})^j$. If $t = A(q) \vee B(q)$, then $q = A^{-1}(t) \wedge B^{-1}(t)$. We can solve the inequalities explicitly to get, after some algebra, that

$$\mathbb{P}(|V_{N,j}| \geq t) \leq \exp\left(\min\left(\frac{1}{c}\left(\frac{t^2}{\text{Var}(h_m)}\right)^{1/j}, \frac{\left(\frac{t}{\|h_m\|_\infty}\sqrt{N/j}\right)^{2/(j+1)}}{(cm/j)^{j/(j+1)}}\right)\right). \quad (7.6)$$

Remark 7.1. Whenever we have that $|X_1 - \mathbb{E}X_1| \leq M$ almost surely, the inequality $|(\pi_j h_m)(x_1, \ldots, x_j)| \leq 2^j \|h_m\|_\infty$ can then be replaced by the following bound: $|(\pi_j h_m)(x_1, \ldots, x_j)| \leq C\|\partial_{u_j} \cdots \partial_{u_1} f_j\|_\infty (2M)^j$, which follows from Lemma 7.2 below. Combined with the assumption stating that

$$\|\partial_{u_j} \cdots \partial_{u_1} f_j\|_\infty \leq \left(\frac{C_1(P)}{m}\right)^{j/2} j^{\gamma_1 j},$$

one easily finds that the resulting concentration inequality reads as follows:

$$\mathbb{P}(|V_{N,j}| \geq t) \leq \exp\left(\min\left(\frac{1}{c}\left(\frac{t^2}{\text{Var}(h_m)}\right)^{1/j}, \left(\frac{t\sqrt{N/j}}{(cMj^{\gamma_1 - 1/2})^j}\right)^{2/(j+1)}\right)\right).$$

This bound holds for all $t > 0$ and is usually sharper than (7.6).

The bound (7.6) is mostly useful only when m/j is not too large. Now we will present a second way to estimate $\mathbb{E}\max_{j=1,\ldots,\lfloor N/j \rfloor} W_j^p$ that will yield much better inequalities for small values of j and that is valid when X_1 is not necessarily supported on a bounded interval. The key technical element that we rely on is the following lemma that allows one to control the growth of moments of W_1 with respect to m. Define

$$f_j(x_1, \ldots, x_j) := \mathbb{E}h_m(x_1, \ldots, x_j, X_{j+1}, \ldots, X_m).$$

Lemma 7.2. *Let the conditions of the theorem hold and let* $\sigma^2 = \text{Var}(X_1)$. *Then there exists* $C = C(P) > 0$ *such that*

$$|(\pi_j h_m)(X_1, \ldots, X_j)| \leq C\|\partial_{u_j} \cdots \partial_{u_1} f_j\|_\infty \prod_{i=1}^{j}(|X_i - \mathbb{E}X_i| + \sigma)$$

with probability 1. *Moreover, for any* $p > 2$,

$$\mathbb{E}|(\pi_j h_m)(X_1, \ldots, X_j)|^p \leq C^{pj}\|\partial_{u_j} \cdots \partial_{u_1} f_j\|_\infty^p (\mathbb{E}|X_1 - \mathbb{E}X_1|^p)^j.$$

The proof of the lemma is outlined in Section 7.4. As we have $\|\partial_{u_j} \cdots \partial_{u_1} f_j\|_\infty \leq (\frac{C_1(P)}{m})^{j/2} j^{\gamma_1 j}$ by assumption, the second bound of the lemma can be written as

$$\mathbb{E}|(\pi_j h_m)(X_1, \ldots, X_j)|^p \leq C_2^{pj} m^{-jp/2} j^{\gamma_1 pj} (\mathbb{E}|X_1 - \mathbb{E}X_1|^p)^j.$$

Recall that $\nu_k = \mathbb{E}^{1/k} |X_1 - \mathbb{E} X_1|^k$ and that under the stated assumptions, $\nu_k \leq k^{\gamma_2} M$ for all integers $k \geq 2$ and some $\gamma_2, M > 0$. Therefore,

$$\mathbb{E} W_1^p \leq C^{2pj} j^{2\gamma_1 p j} m^{-pj} \nu_{2p}^{2pj} \leq (C' M j^{\gamma_1} m^{-1/2} p^{\gamma_2})^{2pj},$$

and consequently $\mathbb{E}(\binom{m}{j} W_1)^p \leq (C' M j^{\gamma_1 - 1/2} p^{\gamma_2})^{2pj}$. The rest of the argument proceeds in a similar way as before. Recall again that $\mathbb{E} W_1 \leq \mathrm{Var}(h_m)/\binom{m}{j}$. Rosenthal's inequality for nonnegative random variables (Fact 4) implies that for $q \geq 2$,

$$\mathbb{E} \left| \frac{\binom{m}{j}}{\lfloor N/j \rfloor} \sum_{i=1}^{\lfloor N/j \rfloor} W_i \right|^{q/2}$$
$$\leq C^{q/2} \left(\mathrm{Var}^{q/2}(h_m) + \left(\frac{q}{2}\right)^{q/2} \left(\frac{j}{N}\right)^{q/2} \mathbb{E} \left(\binom{m}{j} \max_{j=1,\dots,\lfloor N/j \rfloor} W_1 \right)^{q/2} \right).$$

With the inequality for $\mathbb{E} W_1^p$ in hand, the expectation $\mathbb{E}(\binom{m}{j} \max_{j=1,\dots,\lfloor N/j \rfloor} W_1)^{q/2}$ can be upper bounded in two ways: First, trivially,

$$\mathbb{E} \left(\binom{m}{j} \max_{j=1,\dots,\lfloor N/j \rfloor} W_1 \right)^{q/2} \leq \lfloor N/j \rfloor \mathbb{E} \left(\binom{m}{j} W_1 \right)^{q/2} \leq \lfloor N/j \rfloor (C_1 M j^{\gamma_1 - 1/2} q^{\gamma_2})^{qj}.$$

On the other hand, for any identically distributed $\xi_1, \dots, \xi_k$ and any $p > 1$, we have $\mathbb{E} \max_{j=1,\dots,k} |\xi_j| \leq k^{1/p} \max_{j=1,\dots,k} \mathbb{E}^{1/p} |\xi_j|^p$. Choosing $\xi_j = \binom{m}{j} W_j$ and $p = \lfloor \log(N/j) \rfloor + 1$, we obtain the inequality

$$\mathbb{E} \left(\binom{m}{j} \max_{j=1,\dots,\lfloor N/j \rfloor} W_1 \right)^{q/2} \leq (\log(N/j))^{\gamma_2 q j} (C_1 M j^{\gamma_1 - 1/2} q^{\gamma_2})^{qj}.$$

The second bound is better for $q \leq \frac{\log(N/j)}{\gamma_2 j \log\log(N/j)}$, therefore we get the estimates

$$\mathbb{E} \left| \frac{\binom{m}{j}}{\lfloor N/j \rfloor} \sum_{i=1}^{\lfloor N/j \rfloor} W_i \right|^{q/2}$$
$$\leq C^{q/2} \left(\mathrm{Var}^{q/2}(h_m) + \left(C_3^j \left(\frac{qj}{N}\right)^{1/2} (\log^{\gamma_2}(N/j) M j^{\gamma_1 - 1/2} q^{\gamma_2})^j \right)^q \right)$$

and

$$\mathbb{E} |V_{N,j}|^q \leq (C q^{1/2})^{qj} \left(\mathrm{Var}^{q/2}(h_m) \vee \left(\left(\frac{qj}{N}\right)^{1/2} \left(\log^{\gamma_2}(N/j) M j^{\gamma_1 - 1/2} q^{\gamma_2} \right)^j \right)^q \right)$$

that we will use for $2 \leq q \leq \frac{\log(N/j)}{\gamma_2 j}$, while for larger values of q, $(N/j)^{1/q} \leq e^{\gamma_2 j}$ and

$$\mathbb{E} |V_{N,j}|^q \leq (C q^{1/2})^{qj} \left(\mathrm{Var}^{q/2}(h_m) \vee \left(\left(\frac{qj}{N}\right)^{1/2} (M j^{\gamma_1 - 1/2} q^{\gamma_2})^j \right)^q \right).$$

Markov's inequality therefore yields that for small values of q (that is, whenever $2 \leq q \leq \frac{\log(N/j)}{\gamma_2 j}$),

$$\mathbb{P}\left(|V_{N,j}| \geq (Cq)^{j/2}\left(\mathrm{Var}^{1/2}(h_m) \vee \left(\frac{qj}{N}\right)^{1/2}(\log^{\gamma_2}(N/j)Mj^{\gamma_1-1/2}q^{\gamma_2})^j\right)\right) \leq e^{-q}.$$

Let

$$A(q) = (Cq)^{j/2}\mathrm{Var}^{1/2}(h_m), \quad B(q) = \left(\frac{qj}{N}\right)^{1/2}(Cq^{1/2}\log^{\gamma_2}(N/j)Mj^{\gamma_1-1/2}q^{\gamma_2})^j.$$

If $t = A(q) \vee B(q)$, then $q = A^{-1}(t) \wedge B^{-1}(t)$. Solving these inequalities explicitly we get, after some algebra, that

$$\mathbb{P}(|V_{N,j}| \geq t)$$
$$\leq \exp\left(\min\left(\frac{1}{c}\left(\frac{t^2}{\mathrm{Var}(h_m)}\right)^{1/j}, \left(\frac{t\sqrt{N/j}}{(c\log^{\gamma_2}(N/j)Mj^{\gamma_1-1/2})^j}\right)^{\frac{2}{1+j(2\gamma_2+1)}}\right)\right)$$

for values of t satisfying

$$2 \leq \min\left(\frac{1}{c}\left(\frac{t^2}{\mathrm{Var}(h_m)}\right)^{1/j}, \left(\frac{t\sqrt{N/j}}{(c\log^{\gamma_2}(N/j)Mj^{\gamma_1-1/2})^j}\right)^{\frac{2}{1+j(2\gamma_2+1)}}\right) \leq \frac{\log(N/j)}{\gamma_2 j}.$$

Similarly, for $q \geq \max(2, \frac{\log(N/j)}{\gamma_2 j})$, the previously established bounds yield that

$$\mathbb{P}\left(|V_{N,j}| \geq (Cq)^{j/2}\left(\mathrm{Var}^{1/q}(h_m) \vee \left(\frac{qj}{N}\right)^{1/2}(Mj^{\gamma_1-1/2}q^{\gamma_2})^j\right)\right) \leq e^{-q}$$

or, equivalently,

$$\mathbb{P}(|V_{N,j}| \geq t) \leq \exp\left(\min\left(\frac{1}{c}\left(\frac{t^2}{\mathrm{Var}(h_m)}\right)^{1/j}, \left(\frac{t\sqrt{N/j}}{(cMj^{\gamma_1-1/2})^j}\right)^{\frac{2}{1+j(2\gamma_2+1)}}\right)\right) \quad (7.7)$$

whenever

$$\min\left(\frac{1}{c}\left(\frac{t^2}{\mathrm{Var}(h_m)}\right)^{1/j}, \left(\frac{t\sqrt{N/j}}{(cMj^{\gamma_1-1/2})^j}\right)\right)^{\frac{2}{1+j(2\gamma_2+1)}} \geq \max\left(2, \frac{\log(N/j)}{\gamma_2 j}\right).$$

Combination of inequalities (7.6) and (7.7) yields the final result.

7.4. Proof of Lemma 7.2

Recall that $f_j(x_1, \ldots, x_j) = \mathbb{E}h_m(\frac{x_1}{\sqrt{m}}, \ldots, \frac{x_j}{\sqrt{m}}, \frac{X_{j+1}}{\sqrt{m}}, \ldots, \frac{X_m}{\sqrt{m}})$, where $j < m$. It is easy to see from the definition of π_j that $(\pi_j h)(x_1, \ldots, x_j) = (\pi_j f_j)(x_1, \ldots, x_j)$. Next, observe that for any function $g : \mathbb{R}^{j-1} \mapsto \mathbb{R}$ of $j-1$ variables such that $\mathbb{E}g^2(X_1, \ldots, X_{j-1}) < \infty$, $\pi_j g = 0$ P^{j-1}-almost everywhere. Indeed, this follows

immediately from the definition (7.1) of the operator π_j, since g is a constant when viewed as a function of y_j. Based on this fact, it is easy to see that, for any constant $a \in \mathbb{R}$, $f_j(x_1, \ldots, x_j)$ and $f_j(x_1, \ldots, x_j) - f_j|_{x_1=a}(x_2, \ldots, x_j)$ are mapped to the same function by π_j, where $f_j|_{x_1=a}(x_2, \ldots, x_j) := f_j(a, x_2, \ldots, x_j)$,. In particular, $(\pi_j h)(x_1, \ldots, x_j) = (\pi_j(f_j - f_j|_{x_1=a}))(x_1, \ldots, x_j)$. Moreover,

$$f_j(x_1, \ldots, x_j) - f_j|_{x_1=a}(x_2, \ldots, x_j) = \int_a^{x_1} \partial_{u_1} f_j(u_1, x_2, \ldots, x_j)\, du_1.$$

Next, we repeat the same argument with f_j replaced by

$$f_{j,2}(x_2, \ldots, x_j; u_1) := \partial_{u_1} f_j(u_1, x_2, \ldots, x_j)$$

and noting that

$$f_{j,2}(x_2, \ldots, x_j; u_1) - f_{2,j}|_{x_2=a}(x_3, \ldots, x_j; u_1)$$
$$= \int_a^{x_2} \partial_{u_2} f_{j,2}(u_2, x_3, \ldots, x_j; u_1)\, du_2.$$

The expression $\int_a^{x_1} f_{j,2}|_{x_2=a}(x_3, \ldots, x_j; u_1)\, du_1$ is a function of $j - 1$ variables, hence π_j maps it to 0 so that

$$(\pi_j h_m)(x_1, \ldots, x_j) = \pi_j \left(\int_a^{x_1} \int_a^{x_2} \partial_{u_2} f_{j,2}(u_2, x_3, \ldots, x_j; u_1)\, du_2\, du_1 \right).$$

Iterating this process, we arrive at the expression

$$(\pi_j h_m)(x_1, \ldots, x_j) = \pi_j \left(\int_a^{x_1} \cdots \int_a^{x_j} \partial_{u_j} \cdots \partial_{u_1} f_j(u_1, \ldots, u_j)\, du_j \cdots du_1 \right).$$

Next, observe that for any function g of j variables,

$$(\pi_j g)(x_1, \ldots, x_j) = (\delta_{x_1} - P_X) \times \cdots \times (\delta_{x_j} - P_X)g$$
$$= \mathbb{E}_{\tilde{X}}[(\delta_{x_1} - \delta_{\tilde{X}_1}) \times \cdots \times (\delta_{x_j} - \delta_{\tilde{X}_j})g],$$

where $\tilde{X}_1, \ldots, \tilde{X}_j$ are i.i.d. with the same law as X, and independent from $X_1, \ldots, X_N$. Therefore, $(\pi_j h_m)(x_1, \ldots, x_j)$ is a linear combination of 2^j terms of the form

$$\mathbb{E}_{\tilde{X}} \left(\prod_{i \in I} \delta_{x_i} \prod_{j \in I^c} \delta_{\tilde{X}_j}\, g \right) \quad \text{for all choices of } I \subseteq [j]$$

and $g(x_1, \ldots, x_j) = \int_a^{x_1} \cdots \int_a^{x_j} \partial_{u_j} \cdots \partial_{u_1} f_j(u_1, \ldots, u_j)\, du_j \cdots du_1$. Take $a := \mathbb{E}X_1$, and note that

$$|(\pi_j h_m)(x_1, \ldots, x_j)| \leq \|\partial_{u_j} \cdots \partial_{u_1} f_j\|_\infty \sum_{I \subseteq [j]} \prod_{i \in I} |x_i - a| \prod_{j \in I^c} \mathbb{E}|\tilde{X}_i - a|$$

$$\leq \|\partial_{u_j} \cdots \partial_{u_1} f_j\|_\infty \sum_{I \subseteq [j]} \prod_{i \in I} |x_i - a| \cdot \sigma^{|I^c|}$$

$$= \|\partial_{u_j} \cdots \partial_{u_1} f_j\|_\infty \prod_{i=1}^{j} (|x_i - \mathbb{E}X_1| + \sigma).$$

The first claim of the lemma follows. To deduce the moment bound, observe that since $X_1, \ldots, X_j, \tilde{X}_1, \ldots, \tilde{X}_j$ are i.i.d. and in view of convexity of the function $x \mapsto |x|^p$, for $p \geq 1$,

$$\mathbb{E}|(\pi_j h_m)(X_1, \ldots, X_j)|^p$$
$$\leq 2^{(p-1)j} \mathbb{E} \left| \int_a^{X_1} \cdots \int_a^{X_j} \partial_{u_j} \cdots \partial_{u_1} f_j(u_1, \ldots, u_j) \, du_j \cdots du_1 \right|^p$$
$$\leq 2^{(p-1)j} \|\partial_{u_j} \cdots \partial_{u_1} f_j\|_\infty^p \, \mathbb{E}|(X_1 - \mathbb{E}X_1) \cdots (X_j - \mathbb{E}X_j)|^p$$

for $a = \mathbb{E}X_1$.

7.5. Proof of Lemma 5.3

As $\psi(x)$ is integrable, its Fourier transform equals $C_2 \hat{\phi}_1 * \hat{\chi}_R$, while $\hat{\chi}_R = \hat{\kappa}_R \cdot \hat{I}_{2R}$. It is well known (see [17]) that $\hat{\kappa}(x) \leq C_3 e^{-\sqrt{|x|}}$, hence $\hat{\kappa}_R(x) = \hat{\kappa}(Rx) \leq C_3 e^{-\sqrt{R|x|}}$. Moreover, $\hat{I}_{2R}(x) = \sin(2Rx)/x$. Therefore, for $|x|$ large enough,

$$|\hat{\psi}(x)| = C_2 \left| \int_{\mathbb{R}} \hat{\phi}_1(x - y)\hat{\chi}_R(y) \, dy \right|$$
$$= C_2 \Big(\int_{y:|y-x| \geq |x|/2} \hat{\phi}_1(x - y)\hat{\chi}_R(y) \, dy$$
$$+ \int_{y:|y-x| < |x|/2} \hat{\phi}_1(x - y)\hat{\chi}_R(y) \, dy \Big).$$

To estimate the first integral, note that $\hat{\phi}_1(x - y) \leq \frac{C_1}{(1+|x|/2)^\delta} \leq \frac{C_1 2^\delta}{(1+|x|)^\delta}$ whenever $|y - x| \geq |x|/2$, and that $\hat{I}_{2R}(x) \leq 2R$, implying that

$$\left| \int_{y:|y-x| \geq |x|/2} \hat{\phi}_1(x - y)\hat{\chi}_R(y) \, dy \right| \leq \frac{C_4}{(1 + |x|)^\delta} \int_{\mathbb{R}} e^{-\sqrt{R|x|}} \, d(Rx) = \frac{C_5}{(1 + |x|)^\delta}.$$

On the other hand,

$$\left| \int_{y:|y-x| < |x|/2} \hat{\phi}_1(x - y)\hat{\chi}_R(y) \, dy \right| \leq C_6 \left| \int_{x-|x|/2}^{x+|x|/2} e^{-\sqrt{R|x|}} \frac{\sin(2Rx)}{x} \, dx \right|$$
$$\leq C_7 \int_{R|x|/2}^{3R|x|/2} e^{-\sqrt{z}} \, dz \leq C_8 e^{-\sqrt{R|x|/2}} \sqrt{R|x|}.$$

Clearly, the last expression is smaller than $\frac{C_9}{(1+|x|)^\delta}$, implying the desired result.

7.6. Proof of Lemma 5.4

The proof proceeds using the standard Fourier-analytic tools. Let $\hat{\phi}_1 := \mathcal{F}[\phi_1]$ be the Fourier transform of ϕ_1, whence $\mathcal{F}[\phi_{m-j}](t) = (\hat{\phi}_1(t/\sqrt{m-j}))^{m-j}$. Therefore,

$$\phi_{m-j}^{(j-1)}(t) = \frac{1}{2\pi} \int_{\mathbb{R}} \exp(-itx)(ix)^{j-1}\left(\hat{\phi}_1\left(\frac{x}{\sqrt{m-j}}\right)\right)^{m-j} dx$$

and

$$\|\phi_{m-j}^{(j-1)}\|_\infty \leq \frac{1}{2\pi} \int_{\mathbb{R}} |x|^{j-1}\left|\hat{\phi}_1\left(\frac{x}{\sqrt{m-j}}\right)\right|^{m-j} dx$$

$$= \frac{(m-j)^{j/2}}{2\pi} \int_{\mathbb{R}} |x|^{j-1}|\hat{\phi}_1(x)|^{m-j} dx.$$

As $|\hat{\phi}_1(x)| \leq \frac{C_1}{(1+|x|)^\delta}$ by assumption, the integral is finite when $\delta(m-j) > j$ (in particular, this inequality holds when m is large enough and $j = o(m)$ as $m \to \infty$). To get an explicit bound, we will estimate the integral over $[-\eta, \eta]$ and $\mathbb{R} \setminus [-\eta, \eta]$ separately, for a specific choice of $\eta > 0$. To this end, observe that $\hat{\phi}_1(x) = \psi_\sigma(x) + o(x^2)$, where $\psi_\sigma(x) = \exp(-\sigma^2 x^2/2)$ is the characteristic function of the normal law $N(0, \sigma^2)$. Therefore, there exists $\eta > 0$ such that for all $|x| \leq \eta$, $|\hat{\phi}_1(x)| \leq \exp(-\sigma^2 x^2/4)$, and

$$(m-j)^{j/2} \int_{-\eta}^{\eta} |x|^{j-1}|\hat{\phi}_1(x)|^{m-j} dx \leq (m-j)^{j/2} \int_{\mathbb{R}} |x|^{j-1} \exp\left(-\frac{\sigma^2 x^2 (m-j)}{4}\right) dx$$

$$= \int_{\mathbb{R}} |y|^{j-1} \exp\left(-\frac{\sigma^2 y^2}{4}\right) dy = \frac{2^j}{\sigma^j} \Gamma\left(\frac{j}{2}\right),$$

where we used the exact expression for the absolute moments of the normal distribution. As $\Gamma(x+1) \leq C_2 \sqrt{2\pi x}(x/e)^x$ for all $x \geq 1$ and an absolute constant C_2 large enough, $\frac{2^j}{\sigma^j}\Gamma(\frac{j}{2}) \leq \frac{C_2}{\sigma^j}(\frac{2j}{e})^{j/2}$. At the same time,

$$(m-j)^{j/2} \int_{\mathbb{R}\setminus[-\eta,\eta]} |x|^{j-1}|\hat{\phi}_1(x)|^{m-j} dx$$

$$= (m-j)^{j/2} \int_{\mathbb{R}\setminus[-(2C_1)^{2/\delta},(2C_1)^{2/\delta}]} |x|^{j-1}|\hat{\phi}_1(x)|^{m-j} dx$$

$$+ (m-j)^{j/2} \int_{[-(2C_1)^{2/\delta},(2C_1)^{2/\delta}]\setminus[-\eta,\eta]} |x|^{j-1}|\hat{\phi}_1(x)|^{m-j} dx,$$

where $C_1 \geq 1$ is a constant such that $|\hat{\phi}_1(x)| \leq \frac{C_1}{(1+|x|)^\delta}$. The first term can be estimated via

$$(m-j)^{j/2} \int_{\mathbb{R}\setminus[-(2C_1)^{2/\delta},(2C_1)^{2/\delta}]} |x|^{j-1}|\hat{\phi}_1(x)|^{m-j} dx$$

$$\leq C_1^{m-j}(m-j)^{j/2} \int_{\mathbb{R}\setminus[-(2C_1)^{2/\delta},(2C_1)^{2/\delta}]} \frac{|x|^{j-1}}{(1+|x|)^{\delta(m-j)}} dx$$

$$\leq \frac{2C_1^{m-j}(m-j)^{j/2}}{\delta(m-j)-j}\frac{1}{(2C_1)^{2(m-j)-2j/\delta}}.$$

When $m > 2j + 2j/\delta$, we can bound the last expression from above by $C_3 m^{j/2} 2^{-m}$. Finally, as $\sup_{|x|>\eta} |\hat{\phi}_1(x)| \leq 1 - \gamma$ for some $0 < \gamma < 1$,

$$(m-j)^{j/2} \int_{[-(2C_1)^{2/\delta},(2C_1)^{2/\delta}]\setminus[-\eta,\eta]} |x|^{j-1} |\hat{\phi}_1(x)|^{m-j}\, dx$$

$$\leq 2(m-j)^{j/2}(1-\gamma)^{m-j}\frac{(2C_1)^{2j/\delta}}{j}.$$

Putting the estimates together, we deduce that

$$\|\phi_{m-j}^{(j-1)}\|_\infty \leq \frac{(m-j)^{j/2}}{2\pi}\int_{\mathbb{R}} |x|^{j-1}|\hat{\phi}_1(x)|^{m-j}\, dx$$

$$\leq \frac{C_2}{\sigma^j}\left(\frac{2j}{e}\right)^{j/2} + C_3 m^{j/2} 2^{-m} + C_4((2C_1)^{4/\delta}m)^{j/2}(1-\gamma)^{m-j}.$$

Whenever $j = o(m/\log m)$, the last two terms in the sum above are negligible so that for m large enough,

$$\|\phi_{m-j}^{(j-1)}\|_\infty \leq \frac{C_5}{\sigma^j}\left(\frac{2j}{e}\right)^{j/2},$$

as claimed.

References

[1] N. Alon, Y. Matias, and M. Szegedy, The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996)*, pp. 20–29, ACM, New York, 1996 Zbl 0922.68057 MR 1427494

[2] M. A. Arcones, A Bernstein-type inequality for U-statistics and U-processes. *Statist. Probab. Lett.* **22** (1995), no. 3, 239–247 Zbl 0819.60021 MR 1323145

[3] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities*. Oxford University Press, Oxford, 2013 Zbl 1279.60005 MR 3185193

[4] O. Catoni, Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** (2012), no. 4, 1148–1185 Zbl 1282.62070 MR 3052407

[5] R. Y. Chen, A. Gittens, and J. A. Tropp, The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Inf. Inference* **1** (2012), no. 1, 2–20 Zbl 06242993 MR 3311439

[6] V. H. de la Peña and E. Giné, *Decoupling: From dependence to independence*. Probability and its Applications, Springer, New York, 1999 Zbl 0918.60021 MR 1666908

[7] V. H. de la Peña and S. J. Montgomery-Smith, Decoupling inequalities for the tail probabilities of multivariate U-statistics. *Ann. Probab.* **23** (1995), no. 2, 806–816 Zbl 0827.60014 MR 1334173

[8] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira, Sub-Gaussian mean estimators. *Ann. Statist.* **44** (2016), no. 6, 2695–2725 Zbl 1360.62115 MR 3576558

[9] C. DiCiccio and J. Romano, CLT for U-statistics with growing dimension. *Statist. Sinica* **32** (2022), no. 1, 323–344 Zbl 07484121 MR 4359635

[10] W. Feller, On the Berry-Esseen theorem. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **10** (1968), 261–268 Zbl 0167.17304 MR 239639

[11] E. W. Frees, Infinite order U-statistics. *Scand. J. Statist.* **16** (1989), no. 1, 29–45 Zbl 0673.62032 MR 1003967

[12] D. L. Hanson and F. T. Wright, A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.* **42** (1971), 1079–1083 Zbl 0216.22203 MR 279864

[13] J. L. Hodges, Jr. and E. L. Lehmann, Estimates of location based on rank tests. *Ann. Math. Statist.* **34** (1963), 598–611 Zbl 0203.21105 MR 152070

[14] W. Hoeffding, A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** (1948), 293–325 Zbl 0032.04101 MR 26294

[15] W. Hoeffding, Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** (1963), 13–30 Zbl 0127.10602 MR 144363

[16] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani, Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.* **43** (1986), no. 2-3, 169–188 Zbl 0597.68056 MR 855970

[17] S. G. Johnson, Saddle-point integration of C_∞ "bump" functions. 2015, arXiv:1508.04376

[18] A. J. Lee, *U-statistics: Theory and practice*. Routledge, New York, 2019

[19] J. C. H. Lee and P. Valiant, Optimal sub-Gaussian mean estimation in very high dimensions. In *13th Innovations in Theoretical Computer Science Conference*, article no. 98, LIPIcs. Leibniz Int. Proc. Inform. 215, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2022 MR 4459469

[20] J. C. H. Lee and P. Valiant, Optimal sub-Gaussian mean estimation in $\mathbb{R}$. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science—FOCS 2021*, pp. 672–683, IEEE Computer Soc., Los Alamitos, CA, 2022 MR 4399724

[21] M. Lerasle and R. I. Oliveira, Robust empirical mean estimators. 2011, arXiv:1112.3914

[22] G. Lugosi and S. Mendelson, Mean estimation and regression under heavy-tailed distributions: a survey. *Found. Comput. Math.* **19** (2019), no. 5, 1145–1190 Zbl 1431.62123 MR 4017683

[23] A. Maurer, A Bernstein-type inequality for functions of bounded interaction. *Bernoulli* **25** (2019), no. 2, 1451–1471 Zbl 1459.60052 MR 3920378

[24] S. Minsker, Distributed statistical estimation and rates of convergence in normal approximation. *Electron. J. Stat.* **13** (2019), no. 2, 5213–5252 Zbl 1434.62046 MR 4043072

[25] S. Minsker and M. Ndaoud, Robust and efficient mean estimation: an approach based on the properties of self-normalized sums. *Electron. J. Stat.* **15** (2021), no. 2, 6036–6070 Zbl 1498.62029 MR 4355703

[26] A. S. Nemirovsky and D. B. a. Yudin, *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics, John Wiley & Sons, Inc., New York, 1983 MR 702836

[27] W. Peng, T. Coleman, and L. Mentch, Rates of convergence for random forests via generalized U-statistics. *Electron. J. Stat.* **16** (2022), no. 1, 232–292 Zbl 1493.62071 MR 4359361

[28] V. V. Petrov, *Sums of independent random variables*. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 82, Springer, New York-Heidelberg, 1975 MR 0388499

[29] V. V. Petrov, *Limit theorems of probability theory*. Oxford Studies in Probability 4, The Clarendon Press, New York, 1995 Zbl 0826.60001 MR 1353441

[30] R. J. Serfling, *Approximation theorems of mathematical statistics*. Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York, 1980 Zbl 1001.62005 MR 595165

[31] R. J. Serfling, Generalized L-, M-, and R-statistics. *Ann. Statist.* **12** (1984), no. 1, 76–86 MR 733500

[32] L. A. Shepp, A local limit theorem. *Ann. Math. Statist.* **35** (1964), 419–423 Zbl 0146.39103 MR 166817

[33] R. P. Sherman, Maximal inequalities for degenerate U-processes with applications to optimization estimators. *Ann. Statist.* **22** (1994), no. 1, 439–459 Zbl 0798.60021 MR 1272092

[34] Y. Song, X. Chen, and K. Kato, Approximating high-dimensional infinite-order U-statistics: statistical and computational guarantees. *Electron. J. Stat.* **13** (2019), no. 2, 4794–4848 Zbl 1434.62071 MR 4038726

Stanislav Minsker
Department of Mathematics, USC Dornsife, 3620 S. Vermont Ave., Los Angeles, CA 90089-2532, USA; minsker@usc.edu