Efficient median of means estimator

Stanislav Minsker MINSKER@USC.EDU
Department of Mathematics, University of Southern California

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

The goal of this note is to present a modification of the popular median of means estimator that achieves sub-Gaussian deviation bounds with nearly optimal constants under minimal assumptions on the underlying distribution. We build on the recent work on the topic and prove that desired guarantees can be attained under weaker requirements.

Keywords: Median of means estimator; U-statistics; heavy tails; robustness.

1. Introduction.

Let X be a random variable with mean μ and variance σ^2. A sub-Gaussian estimator of μ based on a sample $\mathcal{X} = \{X_1, \dots, X_N\}$ of i.i.d. copies of X is a measurable function $\widetilde{\mu} := \widetilde{\mu}(\mathcal{X}; t)$ such that $\mathbb{P}\Big(|\widetilde{\mu} - \mu| \geq C\sigma\sqrt{\frac{t}{N}}\Big) \leq ce^{-t}$ for a absolute constants $c, C > 0$ and all $t \in [1, t_{\max}(N)]$. It is known (for instance, see the work by Catoni (2012)) that $C \geq \sqrt{2}$. A natural question, posed previously by Devroye et al. (2016), is whether sub-Gaussian estimators with $C = \sqrt{2} + o(1)$, where $o(1)$ is a function that goes to 0 as N (and possibly t) tend to infinity, exist.

Several authors showed that such estimators can indeed be constructed under various additional assumptions. In one of the earliest works on the topic, Catoni (2012) presented the first known example of sharp sub-Gaussian estimators for distributions with finite fourth moment and a known upper bound on the kurtosis, as well as for distributions with known variance. Construction by Devroye et al. (2016) similarly required the fourth moment to be finite. One of the strongest results is the one by Lee and Valiant (2020): their estimator attains required guarantees uniformly over the class of distributions with finite variance, assuming just the finite second moment, albeit with $C = \sqrt{2}$ only in the limit as $t \to \infty$. Minsker (2023) proposed a permutation-invariant version of the well known median of means (MOM) estimator (Nemirovski and Yudin, 1983; Jerrum et al., 1986; Alon et al., 1996) and proved that it achieves desired guarantees for the class of distributions with more than $\frac{3+\sqrt{5}}{2}$ finite moments and "sufficiently regular" probability density functions.

The main goal of this essay is to present a modification of the "permutation-invariant" MOM estimator that attains sub-Gaussian guarantees with asymptotically optimal constants for distributions possessing $2 + \varepsilon$ moments for some $\varepsilon > 0$. This result could yield improvements for a variety of robust algorithms (e.g., see the survey by Lugosi and Mendelson (2019)) that rely on the classical MOM estimator serves as a subroutine.

1.1. Notation.

For a positive integer N, $[N]$ will denote the set $\{1, \dots, N\}$. We employ standard big-O and small-o notation for asymptotic relations between functions and sequences; it will be implicitly assumed that $o(1)$ and $O(1)$ may denote different functions from line to line. Moreover, given two sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$ where $b_n \neq 0$ for all n, we will write that $a_n \ll b_n$ if $\frac{a_n}{b_n} = o(1)$ as $n \to \infty$. Additional notation will be introduced in the main text whenever necessary.

2. Main results.

Let us recall the definition of the classical median of means estimator. Given an i.i.d. sample $\mathcal{X} = \{X_1, \dots, X_N\}$ from distribution P with mean μ and variance σ^2, let $G_1 \cup \ldots \cup G_k \subseteq [N]$ be a collection of k disjoint subsets of cardinality $\lfloor N/k \rfloor$ each, $\bar{X}_j := \frac{1}{|G_j|}\sum_{i \in G_j} X_i$ and $\widehat{\mu}_{\mathrm{MOM}} = \mathrm{med}\left(\bar{X}_1, \dots, \bar{X}_k\right)$, where $\mathrm{med}\,(\cdot)$ stands for the "median." It is known that $\widehat{\mu}_{\mathrm{MOM}}$ satisfies the inequality $\mathbb{P}\Big(|\widehat{\mu}_{\mathrm{MOM}} - \mu| \geq C\sigma\sqrt{\frac{t}{N}}\Big) \leq 2e^{-t}$ with $C = \sqrt{\pi} + o(1)$, where $o(1)$ goes to 0 as $k, N/k \to \infty$. Minsker (2023) proved that allowing the overlapping subsets of data improves the constant: given $J \subseteq [N]$ of cardinality $|J| = \lfloor N/k \rfloor$, set $\bar{X}_J := \frac{1}{|J|}\sum_{j \in J} X_j$ and define $\widehat{\mu}_U = \mathrm{med}\left(\bar{X}_J,\ |J| = \lfloor N/k \rfloor\right)$, where $\left\{\bar{X}_J,\ |J| = \lfloor N/k \rfloor\right\}$ denotes the set of sample averages computed over all possible subsets of $[N]$ of cardinality $\lfloor N/k \rfloor$. Then $\widehat{\mu}_U$ attains sub-Gaussian deviations with $C = \sqrt{2} + o(1)$ under the assumptions described in section 1. Essentially, $\widehat{\mu}_U$ is a function of the order statistics which are complete and sufficient for the family of all distributions with finite variance.

Our construction, presented below, shows that it is not necessary to use all possible sample means, and that a much smaller collection of averages suffices: not only this makes computation easier, but the theoretical guarantees for the resulting estimator hold under weaker assumptions. The main idea is to split the data into subsets of size smaller than $\lfloor N/k \rfloor$, and construct all possible sample means using these subsets as "building blocks". The size of the overlap is then naturally proportional to the size of the block. For instance, the estimator $\widehat{\mu}_U$ corresponds to the blocks of size 1, resulting in the sample means over all possible subsets of a given size. Our results show that allowing the block size to be slowly growing with the the sample size could be beneficial. Formally, let k, l be positive integers such that $\frac{\lfloor N/k \rfloor}{l} \in \mathbb{N}$. Assume that $G_1 \cup \ldots \cup G_{lk} \subseteq [N]$ are disjoint subsets of cardinality $\lfloor \frac{N}{lk} \rfloor$ each, and $Z_j := \bar{X}_j = \frac{1}{|G_j|}\sum_{i \in G_j} X_i,\ j = 1, \dots, lk$. It will be convenient to set $n = lk$, $m = \lfloor \frac{N}{k} \rfloor$, and to view $Z_1, \dots, Z_n$ is a new i.i.d. sample; clearly, Z_1 has mean μ and variance $\frac{\sigma^2}{m/l}$. Given $J \subseteq [n]$ of cardinality $|J| = l$, set $\bar{Z}_J := \frac{1}{l}\sum_{j \in J} Z_j$; note that $\bar{Z}_J$ is an average of m observations from the original sample $\mathcal{X}_N$, same as in the definition of the standard MOM estimator. Define $\mathcal{A}_n^{(l)} = \{J \subset [n] :\ |J| = l\}$ and

$$\widehat{\mu}_N := \mathrm{med}\left(\bar{Z}_J,\ J \in \mathcal{A}_n^{(l)}\right),$$

where $\left\{\bar{X}_J,\ J \in \mathcal{A}_n^{(l)}\right\}$ denotes the set of sample averages computed over all possible subsets of $[n]$ of cardinality l. In other words, $\widehat{\mu}_N$ is the median of means computed over overlapping subsets of data, where the size of the overlap is proportional to $\lfloor N/lk \rfloor$, the size of the block G_1. We remark here that all explicit, non-asymptotic deviations guarantees that are valid for the classical MOM estimator $\widehat{\mu}_{\mathrm{MOM}}$ automatically extend to $\widehat{\mu}_N$ in view of the so-called "Hoeffding representation" of U-statistics (Lee, 2019) as the average of averages of independent random variables; pursuit of optimal constant however appears to require the bounds that include asymptotic terms. Everywhere below, it is assumed that k, m, l and functions of the sample size N. We proceed with the statement of our main result. Denote

$$g(m) := \frac{6}{\sqrt{m}}\mathbb{E}\left[\left(\frac{X_1 - \mu}{\sigma}\right)^2 \min\left(\left|\frac{X_1 - \mu}{\sigma}\right|, \sqrt{m}\right)\right].$$

Feller (1968) proved that $g(m)$ controls the rate of convergence in the central limit theorem, namely that $\sup_{t\in\mathbb{R}} |\Phi_m(t) - \Phi(t)| \leq g(m)$ where Φ_m and Φ are the distribution functions of $\frac{\sum_{j=1}^m X_j - \mu}{\sigma\sqrt{m}}$ and the standard normal law respectively. It is well known that $g(m) \to 0$ as $m \to \infty$ for distributions with finite variance. Moreover, $g(m)$ admits an upper bound of the form $g(m) \leq C\mathbb{E}\left|\frac{X_1-\mu}{\sigma}\right|^{2+\varepsilon} m^{-\varepsilon/2}$ whenever $\mathbb{E}|X_1 - \mu|^{2+\varepsilon} < \infty$ for some $\varepsilon \in (0, 1]$. In the context of the median of mean estimation, the role of $g(m)$ is to control the difference between the mean and the median corresponding to the distribution of $\frac{1}{m}\sum_{j=1}^m X_j$, which can be seen as the main contribution to the the bias of $\widehat{\mu}_N$.

Theorem 1 *Assume that $\mathbb{E}|X_1 - \mu|^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$. Suppose that $l = o(m^\varepsilon)$ and let $L(n, l)$ and $M(n, l)$ be any sequences such that $L(n, l) \gg \frac{n}{l} g^2(m)$ and $M(n, l) \ll \frac{n}{l^2}$. Then for all $L(n, l) \leq t \leq M(n, l)$,*

$$\mathbb{P}\left(|\widehat{\mu}_N - \mu| \geq \sigma\sqrt{\frac{t}{N}}\right) \leq 3\exp\left(-\frac{t}{2(1 + o(1))}\right),$$

where $o(1) \to 0$ as $l, k \to \infty$ uniformly over all $t \in [L(n, l), M(n, l)]$.

Remark 2

(a) A possible choice of parameters is $l = \log(m)$, $L(n, l) = \frac{n}{l}\frac{\log(m)}{m^\varepsilon}$ and $M(n, l) = \frac{n}{l^2\log(l)}$. By varying k, the deviation guarantees can be attained in the desired range of the confidence parameter.

(b) The question of uniformity of the bounds with respect to the underlying distribution is not explicitly addressed in this note. In particular, the $o(1)$ quantities appearing in the inequalities are distribution-dependent. With additional effort, it should be possible to prove uniformity with respect to the classes of distributions $\mathcal{P}_N$ of X satisfying moment conditions of the form $\mathbb{E}\left|\frac{X-\mu}{\sigma}\right|^{2+\varepsilon} \leq a_N$ for a sequence a_N that grows sufficiently slow.

(c) Exact computation of $\widehat{\mu}_N$ is still prohibitively expensive from a numerical standpoint, as the naive upper bound for evaluating the estimator exactly is $O\left((n/l)^l \log(n/l)\right)$. Instead, one may select a collection of T subsets among $J \in \mathcal{A}_n^{(l)}$ uniformly at random and compute the median of the corresponding sample means: in view of Theorem 1 in section 4.3.3 of the book by (Lee, 2019) implies that the asymptotic *distribution of the estimator constructed in this way coincides with the asymptotic distribution $N(0, \sigma^2)$ of $\widehat{\mu}_N$ as soon as $T \gg n/l$. However, this asymptotic equivalence does not automatically imply sharp non-asymptotic bounds of the estimator computed from subsampled blocks any more: results of such nature are currently unknown to us and require further investigation.*

Proof As $\widehat{\mu}_N$ is scale-invariant, we can and will assume that $\sigma^2 = 1$. Set $\rho(x) = |x|$, and note that the equivalent characterization of $\widehat{\mu}$ as an M-estimator is

$$\widehat{\mu} \in \operatorname*{argmin}_{z\in\mathbb{R}} \sum_{J\in\mathcal{A}_n^{(l)}} \rho\left(\sqrt{m}\left(\bar{Z}_J - z\right)\right).$$

The necessary conditions for the minimum of $F(z) := \sum_{J\in\mathcal{A}_n^{(l)}} \rho\left(\sqrt{m}\left(\bar{Z}_J - z\right)\right)$ imply that $0 \in \partial F(\widehat{\mu}_N)$, hence the left derivative $F'_-(\widehat{\mu}_N) \le 0$. Therefore, if $\sqrt{N}\left(\widehat{\mu}_N - \mu\right) \ge \sqrt{t}$ for some $t > 0$, then $\widehat{\mu}_N \ge \mu + \sqrt{t/N}$ and, due to F'_- being nondecreasing, $F'_-\left(\mu + \sqrt{t/N}\right) \le 0$. It implies that

$$\begin{aligned}\mathbb{P}\Big(\sqrt{N}(\widehat{\mu}_N - \mu) \ge \sqrt{t}\Big) &\le \mathbb{P}\left(\sum_{J\in\mathcal{A}_n^{(l)}} \rho'_-\left(\sqrt{m}\left(\bar{Z}_J - \mu - \sqrt{t/N}\right)\right) \ge 0\right)\\ &= \mathbb{P}\left(\frac{\sqrt{k}}{\binom{n}{l}}\sum_{J\in\mathcal{A}_n^{(l)}}\left(\rho'_-\left(\sqrt{m}\left(\bar{Z}_J - \mu - \sqrt{t/N}\right)\right) - \mathbb{E}\rho'_-\right) \ge -\sqrt{k}\mathbb{E}\rho'_-\right)\end{aligned}$$

where we used the shortcut $\mathbb{E}\rho'_-$ in place of

$$\begin{aligned}\mathbb{E}\rho'_-\left(\sqrt{m}\left(\bar{Z}_J - \mu - \sqrt{t/N}\right)\right) &= I\left\{\sqrt{m}(\bar{Z}_J - \mu) \le \sqrt{\frac{t}{k}}\right\} - I\left\{\sqrt{m}(\bar{Z}_J - \mu) > \sqrt{\frac{t}{k}}\right\}\\ &= 1 - 2I\left\{\sqrt{m}(\bar{Z}_J - \mu) \le \sqrt{\frac{t}{k}}\right\}.\end{aligned}$$

Note that

$$\begin{aligned}-\sqrt{k}\mathbb{E}\rho'_-\left(\sqrt{m}\left(\bar{Z}_J - \mu - \sqrt{t/N}\right)\right) &= -\sqrt{k}\left(1 - 2\mathbb{P}\Big(\sqrt{m}\left(\bar{Z}_J - \mu - \sqrt{t/N}\right) \le 0\Big)\right)\\ &= 2\sqrt{k}\left(\Phi\left(\sqrt{\frac{t}{k}}\right) - \Phi(0)\right) - 2\sqrt{k}\left(\Phi\left(\sqrt{\frac{t}{k}}\right) - \mathbb{P}\left(\sqrt{m}\left(\bar{Z}_J - \mu\right) \le \sqrt{\frac{t}{k}}\right)\right)\\ &\ge -2\sqrt{k}\cdot g(m) + 2\sqrt{t}\frac{1}{\sqrt{t}/\sqrt{k}}\left(\Phi\left(\frac{\sqrt{t}}{\sqrt{k}}\right) - \Phi(0)\right).\end{aligned}$$

Since

$$2\sqrt{t}\frac{1}{\sqrt{t}/\sqrt{k}}\left(\Phi\left(\frac{\sqrt{t}}{\sqrt{k}}\right) - \Phi(0)\right) = 2\sqrt{t}\left(\phi(0) + O(\sqrt{t/k})\right) = \sqrt{t}\left(\sqrt{\frac{2}{\pi}} + O(\sqrt{t/k})\right)$$

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

$$-\sqrt{k}\,\mathbb{E}\rho'_-\left(\sqrt{m}\left(\bar{Z}_J - \mu - \sqrt{t/N}\right)\right) \ge -2\sqrt{k}\cdot g(m) + \sqrt{t}\left(\sqrt{\frac{2}{\pi}} + O(\sqrt{t/k})\right)$$

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 analyze the U-statistic

$$\sqrt{k}\,U_{n,l}(\rho'_-) = \frac{\sqrt{k}}{\binom{n}{l}}\sum_{J\in\mathcal{A}_n^{(l)}}\left(\rho'_-\left(\sqrt{m}\left(\bar{Z}_J - \mu - \sqrt{t/N}\right)\right) - \mathbb{E}\rho'_-\right). \tag{1}$$

As the expression above is invariant with respect to the shift $Z_j \mapsto Z_j - \mu$, we can assume that $\mu = 0$. For $i \in [N]$, let

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

where $(\tilde{Z}_1, \ldots, \tilde{Z}_l)$ is an independent copy of $(Z_1, \ldots, Z_l)$ based on a sample $\tilde{\mathcal{X}}_N$ that is an independent copy of $\mathcal{X}_N$. Our goal is to determine the size of $\mathrm{Var}(h^{(1)}(X_1))$.

Remark 3 *The quantity $h^{(1)}(Z)$ is related to the so-called Hájek projection that can be viewed as the best (in mean squared sense) approximation of the U-statistic $U_{n,l}(\rho'_-)$ in terms of the sums of i.i.d. random variables. For related background on U-statistics, we refer the reader to an excellent monograph by Lee (2019).*

Lemma 4 *In the framework of Theorem 1,*

$$Var\left(h^{(1)}(Z_1)\right) \to \frac{2}{\pi}$$

as $l, k \to \infty$, uniformly over all $t \in [L(n,l), M(n,l)]$.

The proof of the lemma is given in section 3. The following result, a deviation inequality for U-statistics of order that grows with the sample size, is the second key technical tool required to complete the argument.

Theorem 5 *Let $h : \mathbb{R}^l \mapsto \mathbb{R}$ be a function that is invariant with respect to permutations of its arguments, and let $U_{n,l}(h) = \frac{1}{\binom{n}{l}} \sum_{J \in \mathcal{A}_n^{(l)}} \left(h\left(X_j,\ j \in J\right) - \mathbb{E}h\left(X_1, \ldots, X_l\right)\right)$ be the corresponding U-statistic with kernel h evaluated on a sample $X_1, \ldots, X_n$. Assume that l is an increasing function of n, and that*

(a) h is uniformly bounded;

(b) $\liminf_{l\to\infty} Var\left(\sqrt{l}\,h^{(1)}(X_1)\right) > 0$, where $h^{(1)}(X_1) = \mathbb{E}\left[h(X_1, X_2, \ldots, X_l)|X_1\right]$.

Let $q(n,l)$ be increasing in n, decreasing in l, and such that $q(n,l) = o\left(\frac{n}{l^2}\right)$. Then for all $2 \le t \le q(n,l)$,

$$\mathbb{P}\left(|U_{n,l}(h)| \ge \sqrt{\frac{tl}{n}}\right) \le (2 + o(1)) \exp\left(-\frac{t}{2(1+o(1))Var\left(\sqrt{l}\,h^{(1)}(X_1)\right)}\right),$$

where $o(1) \to 0$ as $l, n/l \to \infty$ uniformly over $2 \le t \le q(n,l)$.

The proof of this result is outlined in section 4 [1]. To get the desired inequality for the estimator $\widehat{\mu}_N$, it remains to apply Theorem 5 and Lemma 4 to the U-statistic defined in (1): specifically, we deduce that

$$\mathbb{P}\left(\left|\sqrt{k}\,U_{n,l}(\rho'_-)\right| \ge \sqrt{t}\sqrt{\frac{2}{\pi}}\,(1 + o(1))\right) \le 2\exp\left(-\frac{t}{2(1+o(1))}\right),$$

1. We note that closely related results for U-statistics were obtained by Maurer (2019), and it may be possible to use Maurer's inequality in place of Theorem 5.

uniformly over $\frac{n}{l}g^2(m) \ll t \ll \frac{n}{l^2}$, and the final result follows.

■

3. Proof of Lemma 4.

Note that we can rewrite $h^{(1)}(Z_1)$ as

$$h^{(1)}(Z_1) = \sqrt{l}\,\mathbb{E}\left[\rho'_-\left(\sqrt{m}\left(\frac{1}{m}\sum_{j=1}^{m-m/l}\tilde{Z}_j + \frac{1}{\sqrt{ml}}\left(Z_1\sqrt{m/l}\right) - \sqrt{t/N}\right)\right)\,\Big|\,Z_i\right] - \sqrt{l}\,\mathbb{E}\rho'_-.$$

Given an integer $r \geq 1$, let $\widetilde{\Phi}_r(t)$ be the cumulative distribution function of $\sum_{j=1}^r \tilde{X}_j$. Then

$$\begin{aligned}
h^{(1)}(Z_1) &= \sqrt{l}\left(2\widetilde{\Phi}_{m-m/l}\left(m\sqrt{\frac{t}{N}} - \sqrt{m/l}\left(Z_1\sqrt{m/l}\right)\right) - 1\right) - \sqrt{l}\,\mathbb{E}\,\rho'_- \\
&= 2\sqrt{l}\left(\widetilde{\Phi}_{m-m/l}\left(m\sqrt{\frac{t}{N}} - \sqrt{m/l}\left(Z_1\sqrt{m/l}\right)\right) - \mathbb{E}\widetilde{\Phi}_{m-m/l}\left(m\sqrt{\frac{t}{N}} - \sqrt{m/l}\left(Z_1\sqrt{m/l}\right)\right)\right) \\
&= 2\sqrt{l}\int_{\mathbb{R}}\left(\widetilde{\Phi}_{m-m/l}\left(m\sqrt{\frac{t}{N}} - \sqrt{m/l}\left(Z_1\sqrt{m/l}\right)\right)\right. \\
&\qquad\left. -\widetilde{\Phi}_{m-m/l}\left(m\sqrt{\frac{t}{N}} - x\sqrt{m/l}\right)\right) dP_{Z_1\sqrt{m/l}}(x),
\end{aligned}$$

with $P_{Z_1\sqrt{m/l}}$ being the law of $Z_1\sqrt{m/l}$. Feller's version of Berry-Esseen theorem implies that

$$\sup_{x\in\mathbb{R}}\left|\widetilde{\Phi}_{m-m/l}(x) - \Phi(x/\sqrt{m-m/l})\right| \leq 6g(m-m/l)$$

where Φ is the distribution function of standard normal law. Therefore,

$$\begin{aligned}
\Bigg|h^{(1)}(Z_1) - 2\sqrt{l}\int_{\mathbb{R}}\Bigg(&\Phi\left(\frac{m}{\sqrt{m-m/l}}\sqrt{\frac{t}{N}} - \sqrt{\frac{m/l}{m-m/l}}\left(Z_1\sqrt{m/l}\right)\right) \\
&-\Phi\left(\frac{m}{\sqrt{m-m/l}}\sqrt{\frac{t}{N}} - x\sqrt{\frac{m/l}{m-m/l}}\right)\Bigg) dP_{Z_1\sqrt{m/l}}(x)\Bigg| \leq 12\sqrt{l}\,g(m-m/l) \to 0
\end{aligned}$$

by assumption. At the same time,

$$\begin{aligned}
&\sqrt{l}\left(\Phi\left(\frac{m}{\sqrt{m-m/l}}\sqrt{\frac{t}{N}} - \sqrt{\frac{m/l}{m-m/l}}\left(Z_1\sqrt{m/l}\right)\right) - \Phi\left(\frac{m}{\sqrt{m-m/l}}\sqrt{\frac{t}{N}} - x\sqrt{\frac{m/l}{m-m/l}}\right)\right) \\
&= \frac{1}{\sqrt{2\pi}}\exp\left(-q(x)/2\right)(x - Z_1\sqrt{m/l})\sqrt{\frac{m}{m-m/l}} + \frac{C(x,Z_1)}{\sqrt{l}}(x - Z_1\sqrt{m/l})^2\frac{m/l}{m-m/l}
\end{aligned}$$

where $q(x) := \left(\frac{m}{\sqrt{m-m/l}} \sqrt{\frac{t}{N}} - x \sqrt{\frac{m/l}{m-m/l}} \right)^2$ is such that $q(x) \to 0$ as $l \to \infty$ and $C(x, Z_1)$ is a bounded function. Therefore, $h^{(1)}(Z_1) - \sqrt{\frac{2}{\pi}} Z_1 \sqrt{m/l} \to 0$ almost surely, assuming that $\sqrt{l} g(m) = o(1)$. Finally, note that

$$\sqrt{l} \left(\widetilde{\Phi}_{m-m/l} \left(m \sqrt{\frac{t}{N}} - \sqrt{m/l} \left(Z_1 \sqrt{m/l} \right) \right) - \widetilde{\Phi}_{m-m/l} \left(m \sqrt{\frac{t}{N}} - x \sqrt{m/l} \right) \right)$$
$$\leq \sup_z \sqrt{l}\, \mathbb{P} \left(\sum_{j=1}^{m-m/l} \tilde{X}_j \in \left(z, z + \left| \sqrt{m/l} \left(x - Z_1 \sqrt{m/l} \right) \right| \right] \right) \leq C \left| x - Z_1 \sqrt{m/l} \right|,$$

where the last inequality follows from the well known bound for the concentration function (Theorem 2.20 in the book by Petrov (1995)); here, $C = C(P) > 0$ is a constant that may depend on the distribution of X_1. We therefore conclude that the sequence $\left(h^{(1)}(Z_1) - \sqrt{\frac{2}{\pi}} Z_1 \sqrt{m/l} \right)^2$ is uniformly integrable (as $Z_1 \sqrt{m/l}$ is), hence the claim follows.

4. Proof of Theorem 5.

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

$$\mathbb{P} \left(|U_{n,l}(h)| \geq \sqrt{\frac{tl}{n}} \right)$$
$$\leq \mathbb{P} \left(\left| \frac{l}{n} \sum_{j=1}^{n} h^{(1)}(Z_j) \right| \geq (1-\varepsilon) \sqrt{t} \sqrt{\frac{l}{n}} \right) + \mathbb{P} \left(\left| \sum_{j=2}^{l} \frac{\binom{l}{j}}{\binom{n}{j}} \sum_{J \in \mathcal{A}_n^{(j)}} h^{(j)}(Z_i,\, i \in J) \right| \geq \varepsilon \sqrt{t} \sqrt{\frac{l}{n}} \right),$$

where $h^{(j)}$, $2 = 1, \ldots, l$ are the degenerate kernels corresponding to the higher-order terms of Hoeffding's decomposition (not to be confused with the derivatives!). Specifically,

$$h^{(j)}(y_1, \ldots, y_j) = (\delta_{y_1} - P_Y) \times \ldots \times (\delta_{y_j} - P_Y) \times P_Y^{m-j} h,$$

where δ_y is the point measure concentrated at y; in particular, $\delta_y(h) = h(y)$. It is known that $h^{(j)}$ can be viewed geometrically as orthogonal projections of h onto a particular subspace of $L_2(P_Y^m)$. We refer the reader to the book by Lee (2019) for futher details related to the background material

on U-statistics and the Hoeffding's decomposition. Bernstein's inequality yields that

$$\mathbb{P}\left(\left|\frac{l}{n}\sum_{j=1}^{n} h^{(1)}(Z_j)\right| \geq (1-\varepsilon)\sqrt{t}\sqrt{\frac{l}{n}}\right)$$
$$\leq 2\exp\left(-\frac{(1-\varepsilon)^2\, t/2}{\mathrm{Var}\left(\sqrt{l}\, h^{(1)}(Z_1)\right) + (1-\varepsilon)\frac{1}{3}\sqrt{\frac{l}{n}}\|h\|_\infty t^{1/2}}\right)$$
$$= 2\exp\left(-\frac{(1-\varepsilon)^2\, t}{2\,\mathrm{Var}\left(\sqrt{l}\, h^{(1)}(Z_1)\right)(1+o(1))}\right)$$

where $o(1) \to 0$ as $n/l \to \infty$ uniformly over t. It remains to control the expression involving higher order Hoeffding decomposition terms. To this end, we will show that it is bounded from above by $\exp\left(-\frac{t}{2\,\mathrm{Var}(\sqrt{l}\, h^{(1)}(X_1))}\right)\cdot o(1)$ where $o(1) \to 0$ uniformly over the range of t. To this end, we will need concentration inequality for the U-statistics of growing order established in Minsker (2023, Theorem 4.1). Set $t_{j,\varepsilon} = \left(\varepsilon\frac{\sqrt{t}}{j^2}\left(\frac{n}{l}\right)^{\frac{j-1}{2}}\right)^2$, and note that, in view of the union bound,

$$\mathbb{P}\left(\left|\sum_{j=2}^{l}\frac{\binom{l}{j}}{\binom{n}{j}}\sum_{J\in\mathcal{A}_n^{(j)}} h^{(j)}(Z_i,\, i\in J)\right| \geq \varepsilon\sqrt{t}\sqrt{\frac{l}{n}}\right)$$
$$\leq \sum_{j=2}^{l}\mathbb{P}\left(\left|\frac{\binom{l}{j}}{\binom{n}{j}}\sum_{J\in\mathcal{A}_n^{(j)}} h^{(j)}(Z_i,\, i\in J)\right| \geq \sqrt{t_{j,\varepsilon}}\sqrt{\frac{l}{n}}\right)$$
$$\leq l\max_{2\leq j\leq l}\exp\left(-c\min\left((t\varepsilon^2)^{1/j}\left(\frac{n}{l}\right)^{\frac{j-1}{j}}, \left(\frac{t\varepsilon^2}{\|h\|_\infty^2}\right)^{\frac{1}{j+1}}\left(\frac{nj}{l^2}\right)^{\frac{j}{j+1}}\right)\right),$$

where the last inequality follows from the first bound of Theorem 4.1 in Minsker (2023). Whenever $l\log^2(l) \ll k$ and $\varepsilon \gg \log^{-1/2}(l)$, the last expression is at most

$$\max_{2\leq j\leq l}\exp\left(-c_1\min\left((t\varepsilon^2)^{1/j}\left(\frac{n}{l}\right)^{\frac{j-1}{j}}, \left(\frac{t\varepsilon^2}{\|h\|_\infty^2}\right)^{\frac{1}{j+1}}\left(\frac{nj}{l^2}\right)^{\frac{j}{j+1}}\right)\right).$$

In turn, it is bounded by $e^{-\frac{c_2 t}{\varepsilon}}$ whenever $t < \frac{n}{l^2}\varepsilon^4$. Desired conclusion follows.

Acknowledgments

Author acknowledges support by the National Science Foundation grants DMS CAREER-2045068 and CCF-1908905 and appreciates the constructive feedback and insightful comments of the anonymous Referees that helped improve the quality of presentation.

References

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 Theory of computing*, pages 20–29. ACM, 1996.

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

L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

W. Feller. On the Berry-Esseen theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 10(3):261–268, 1968.

M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical computer science*, 43:169–188, 1986.

A. J. Lee. *U-statistics: Theory and Practice*. Routledge, 2019.

J. C. H. Lee and P. Valiant. Optimal sub-Gaussian mean estimation in $\mathbf{R}$. *arXiv preprint arXiv:2011.08384*, 2020.

G. Lugosi and S. Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Andreas Maurer. A Bernstein-type inequality for functions of bounded interaction. *Bernoulli*, 25 (2):1451–1471, 2019.

Stanislav Minsker. U-statistics of growing order and sub-Gaussian mean estimators with sharp constants. *Mathematical Statistics and Learning*, to appear, 2023.

A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley & Sons Inc., 1983.

V. V. Petrov. *Limit theorems of probability theory: sequences of independent random variables*. Oxford, New York, 1995.