Electronic Journal of Statistics
Vol. 15 (2021) 6036–6070
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1925

# Robust and efficient mean estimation: an approach based on the properties of self-normalized sums*

**Stanislav Minsker**

*Department of Mathematics,*
*University of Southern California*
*e-mail:* minsker@usc.edu

**Mohamed Ndaoud**

*Department of Decisions Sciences,*
*ESSEC Business School*
*e-mail:* ndaoud@essec.edu

**Abstract:** Let $X$ be a random variable with unknown mean and finite variance. We present a new estimator of the mean of $X$ that is robust with respect to the possible presence of outliers in the sample, provides tight sub-Gaussian deviation guarantees without any additional assumptions on the shape or tails of the distribution, and moreover is asymptotically efficient. This is the first estimator that provably combines all these qualities in one package. Our construction is inspired by robustness properties possessed by the self-normalized sums. Theoretical findings are supplemented by numerical simulations highlighting strong performance of the proposed estimator in comparison with previously known techniques.

**Keywords and phrases:** robust estimation, sub-Gaussian deviations, self-normalized sums, efficiency.

Received October 2020.

## Contents



*Authors acknowledge support by the National Science Foundation grants DMS-1712956 and CIF-1908905. M.N. was partially supported by a James H. Zumberge Faculty Research and Innovation Fund at the University of Southern California.

## 1. Introduction.

Let $X$ be a random variable with mean $\mathbb{E}X = \mu$ and variance $\mathrm{Var}(X) = \sigma^2$, where both $\mu$ and $\sigma^2$ are unknown; in what follows, $P$ will denote the distribution of $X$ and $\mathcal{P}_{2,\sigma}$ – the class of all distributions possessing 2 finite moments and having variance $\sigma^2$. We will be interested in robust estimators $\widehat{\mu}$ of $\mu$ constructed from the data $X_1, \dots, X_N$ generated as follows: the initial non-corrupted sample $X_1, \dots, X_{N'}$ of independent, identically distributed copies of $X$ is merged with a set of $O < N'$ outliers that are independent from the initial sample, and the combined sample of cardinality $N := N' + O$ is given as an input to an algorithm responsible for construction of the estimator. This contamination framework is more general than Huber's contamination model [Hub64, CGR16] where the outliers are assumed to be identically distributed, but weaker than the framework allowing adversarial outliers [KL93, Val85] that may for instance depend on the initial sample. Robustness will be quantified by two properties: first, in the situation when $O = 0$, the estimators should admit tight non-asymptotic deviation bounds of the form

$$|\widehat{\mu} - \mu| \le C\sigma\sqrt{\frac{s}{N}} \tag{1.1}$$

with probability at least $1 - 2e^{-s}$, where $C > 0$ is an absolute constant. In particular, we will be interested in the estimators that attain such deviation guarantees uniformly over $0 < s < \psi_P(N)$ where $\psi_P(N)$ is an increasing function that might depend on the law of $X$ [1]; guarantees of type (1.1) can be informally labeled as "robustness to heavy tails." Second, the estimators of interest should perform optimally with respect to the degree of outlier contamination characterized by the quantity $\varepsilon := \frac{O}{N}$.

Another important property that we focus on is *asymptotic efficiency*. Informally speaking, efficiency measures how "wasteful" an estimator is: an efficient

[1] It follows from results in [DLLO16] that the function $\psi_P(N)$ can not be chosen to be independent of $P$, no matter how slow its growth is. At the same time, our results show that for every $\sigma > 0$ and $P \in \mathcal{P}_{2,\sigma}$, such a function exists.

estimator will capture all the information available in the sample; alternatively, in many cases it is possible to conclude that the confidence intervals centered at an efficient estimator will have (at least asymptotically) smallest possible diameter. It is difficult to quantify efficiency using only finite-sample guarantees of type (1.1) as the constants in these bounds are rarely sharp, at least, for practical considerations, and therefore a common approach is to take an asymptotic viewpoint. Specifically, we will be looking for the estimators that are asymptotically normal and have asymptotic variance that is as small as possible in the minimax sense, that is, $\sqrt{N}\,(\widehat{\mu}-\mu) \xrightarrow{d} \mathcal{N}(0,\nu^2)$ as $N\to\infty$, where $\xrightarrow{d}$ denotes convergence in distribution and $\nu^2 := \nu^2(\widehat{\mu}, P)$ is such that

$$\sup_{P\in\mathcal{P}_{2,\sigma}} \nu^2(\widehat{\mu}, P) = \inf_{\widetilde{\mu}} \sup_{P\in\mathcal{P}_{2,\sigma}} \nu^2(\widetilde{\mu}, P).$$

Here, the infimum is taken over all asymptotically normal (after rescaling by $\sqrt{N}$) estimators $\widetilde{\mu}$ of $\mu$. It is easy to see that $\inf_{\widetilde{\mu}} \sup_{P\in\mathcal{P}_{2,\sigma}} \nu^2(\widetilde{\mu}, P) = \sigma^2$ (for reader's convenience, the proof of this simple fact is given in Lemma 6.5), therefore, it suffices to find a robust estimator that satisfies $\sqrt{N}\,(\widehat{\mu}-\mu) \xrightarrow{d} \mathcal{N}(0,\sigma^2)$ for all $P\in\mathcal{P}_{2,\sigma}$. For instance, the sample mean is an example of the estimator with required asymptotic properties that is not robust, while the popular median-of-means estimator [NY83] is robust but not asymptotically efficient [Min19].

In this paper we construct the first, to the best of our knowledge, example of an estimator of the mean that is provably **(a)** robust to the heavy tails of the data-generating distribution $P$; **(b)** admits optimal error bounds with respect to the outlier contamination proportion $\varepsilon = \frac{O}{N}$; **(c)** is asymptotically efficient and **(d)** is almost tuning-free, meaning that it does not require information about any parameters of the distribution besides the upper bound for the contamination proportion $\varepsilon$. We also show how to make our procedure fully adaptive. Our construction is novel and is inspired by the properties of self-normalized sums.

The rest of the paper is organized as follows: section 2 introduces the estimator and explains the main ideas behind its construction; the key results are presented in section 3, while comparison of our estimator with existing robust estimation techniques in the context of properties **(a)** - **(d)** is presented in section 3.4. Finally, a fully adaptive procedure is outlined in section 4 while the supporting numerical simulations are included in section 5. The proofs of the main results are contained in section 6. All notation and auxiliary results will be introduced on demand.

## 2. Construction of the estimator.

We restrict our attention to the estimators that are obtained via aggregating the sample means evaluated over disjoint subsets (also referred to as "blocks") of the data. Specifically, assume that $\{1,\dots,N\} = \bigcup_{j=1}^{k} G_j$ where $G_i\cap G_j = \emptyset$ for $i\neq j$ and $|G_j| = n = N/k$ is an integer, and let $\bar{\mu}_j := \frac{1}{|G_j|}\sum_{i\in G_j} X_i$ be the

sample mean of the observations indexed by $G_j$. We consider estimators $\widehat{\mu}_N$ of the form

$$\widehat{\mu}_N = \sum_{j=1}^{k} \alpha_j \bar{\mu}_j \tag{2.1}$$

for some (possibly random and data-dependent) nonnegative weights $\alpha_1, \dots, \alpha_k$ such that $\sum_{j=1}^{k} \alpha_j = 1$. For example, the well known median-of-means estimator [NY83, AMS96, LO11] corresponds to the case $\alpha_{\widehat{j}} = 1$ for $\widehat{j}$ such that $\bar{\mu}_{\widehat{j}} = \text{median}\,(\bar{\mu}_1, \dots, \bar{\mu}_k)$ and $\alpha_{\widehat{j}} = 0$ otherwise. Construction proposed in this paper starts with an observation that it is natural to choose the weights that are inversely proportional to some increasing function of the standard deviation of each block. Indeed, the estimation error of the sample mean $\bar{\mu}_j$ in each block of the data is essentially controlled by the corresponding sample standard deviation $\widehat{\sigma}_j := \sqrt{\frac{1}{|G_j|}\sum_{i\in G_j}(X_i - \bar{\mu}_j)^2}$. To understand why, consider the following obvious identity:

$$|\bar{\mu}_j - \mu| = \left|\frac{\bar{\mu}_j - \mu}{\widehat{\sigma}_j}\right| \widehat{\sigma}_j.$$

The random variable $\frac{\bar{\mu}_j - \mu}{\widehat{\sigma}_j}$, which is equal up to normalization to the Student's t-statistic, is known to be tightly concentrated around 0: namely, it is bounded by $\left|\sqrt{\frac{t}{n}}\right|$ with probability at least $1 - e^{-ct}$ for $t \le c'n$ where $c, c'$ are positive constants, even if data are heavy-tailed (a more precise version of this fact is stated below). Therefore, $|\bar{\mu}_j - \mu|$ is bounded by a multiple of $\frac{\widehat{\sigma}_j}{\sqrt{n}}$ with high probability. And, while the error $|\bar{\mu}_j - \mu|$ is unknown, the quantity $\widehat{\sigma}_j$ is fully data-dependent. This motivates the choice of the weights of the form

$$\alpha_j = \frac{1/\widehat{\sigma}_j^p}{\sum_{i=1}^{k} 1/\widehat{\sigma}_i^p} \tag{2.2}$$

for some $p \ge 1$; in what follows, the estimator (2.1) with weights (2.2) will be denoted $\widehat{\mu}_{N,p}$. When we need to emphasize the specific value of $k$ used in the construction, we will write $\widehat{\mu}_{N,p}(k)$. Observe that when $p = 1$, the estimation error satisfies

$$\widehat{\mu}_{N,1} - \mu = \frac{\frac{1}{k}\sum_{j=1}^{k} \frac{\bar{\mu}_j - \mu}{\widehat{\sigma}_j}}{\frac{1}{k}\sum_{j=1}^{k} \frac{1}{\widehat{\sigma}_j}}, \tag{2.3}$$

which is proportional to the average of t-statistics evaluated over $k$ independent subsamples. It is therefore natural to expect that $\widehat{\mu}_{N,1} - \mu$ will satisfy strong deviation guarantees.

Let us present now an example where the weights corresponding to $p = 2$ arise naturally. Observe that one can model outliers by assuming that the variances of the data differ across $k$ groups, where large variance corresponds to a corrupted subsample: $X_i,\ i \in G_j \sim N(\mu, \sigma_j^2)$ for some $\mu \in \mathbb{R}$ and positive but unknown $\sigma_1, \dots, \sigma_k$. The maximum likelihood estimator $\widetilde{\mu}$ in this model is easily seen to

satisfy $\widetilde{\mu} = \operatorname{argmin}_{z\in\mathbb{R}} \sum_{j=1}^{k} |G_j| \log\left(\sum_{i\in G_j} (X_i - z)^2\right)$. Equivalently, $\widetilde{\mu}$ can be defined via

$$\widetilde{\mu} = \operatorname{argmin}_{z\in\mathbb{R}} \sum_{j=1}^{k} |G_j| \log\left(1 + \left(\frac{\bar{\mu}_j - z}{\widehat{\sigma}_j}\right)^2\right).$$

An approximate solution can be obtained via minimizing the first-order approximation of the loss function $z \mapsto \sum_{j=1}^{k} |G_j| \left(\frac{\bar{\mu}_j - z}{\widehat{\sigma}_j}\right)^2$ that attains its minimum at the point

$$\sum_{j=1}^{k} \bar{\mu}_j \frac{|G_j|/\widehat{\sigma}_j^2}{\sum_{i=1}^{k} |G_i|/\widehat{\sigma}_i^2} = \sum_{j=1}^{k} \bar{\mu}_j \frac{1/\widehat{\sigma}_j^2}{\sum_{i=1}^{k} 1/\widehat{\sigma}_i^2},$$

which is the estimator (2.1) with weights defined in (2.2) for $p = 2$. In the following sections we will present non-asymptotic deviation bounds for the estimator $\widehat{\mu}_{N,p}$ for all values of $p \geq 1$ and will establish its asymptotic efficiency in the absence of outliers.

## 3. Main results.

The goal of this section is to prove the deviation inequality for the estimation error $\widehat{\mu}_{N,p} - \mu$ for any $p \geq 1$, where the estimator $\widehat{\mu}_{N,p}$ corresponds to the weights defined by (2.2).

### *3.1. Preliminaries.*

In this section, we consider the simple framework of i.i.d data without outliers. We will start with a brief review of concentration inequalities for the self-normalized sums. It is known (for example, see the book [PLS08]) that the properties of the t-statistics

$$T_j := \begin{cases} \frac{\bar{\mu}_j - \mu}{\widehat{\sigma}_j}, & \widehat{\sigma}_j > 0, \\ 0, & \widehat{\sigma}_j = 0 \end{cases} \tag{3.1}$$

evaluated over subsamples indexed by $G_1, \ldots, G_k$ are closely related to the behavior of the self-normalized sums $Q_j := \frac{\bar{\mu}_j - \mu}{V_j}$ where $V_j^2 := \frac{1}{|G_j|} \sum_{i\in G_j} (X_i - \mu)^2$. Indeed, it is easy to see that $T_j = f(Q_j)$ where $f(z) = \frac{z}{\sqrt{1-z^2}}$. The following inequality is well known (cf. Theorem 2.16 [PLS08]): for any $j = 1, \ldots, k$ and any $x > 0$, [2]

$$|Q_j| \leq \frac{x}{\sqrt{|G_j|}} \left(1 + \frac{4\sigma}{V_j}\right) \tag{3.2}$$

with probability at least $1 - 4e^{-x^2/2}$, as long as $\mathbb{E}(X - \mu)^2 < \infty$. In order to deduce a non-random upper bound from (3.2), it suffices to control the ratio

[2] Since $|Q_j| \leq 1$, the inequality is nontrivial only for $x < \sqrt{|G_j|}$.

$\frac{1}{V_j}$. To this end, define

$$\zeta(X) := \inf\left\{a > 0 : \mathbb{E}\left(|X-\mu|^2 \mathbf{1}\{|X-\mu| \le \sigma \cdot a\}\right) \ge \sigma^2/2\right\}.$$

As long as $\mathrm{Var}(X)$ is finite, it is clear that $\zeta(X) < \infty$.

**Lemma 3.1.** *With probability at least* $1 - e^{-\frac{n}{40\zeta^2(X)\vee 6}}$, $V_j \ge \frac{\sigma}{2}$.

Combining this inequality with the bound (3.2), we deduce that for any $1 \le j \le k$ and any $x > 0$,

$$|Q_j| \le \frac{9x}{\sqrt{n}}, \tag{3.3}$$

with probability at least $1 - 4e^{-x^2/2} - e^{-\frac{n}{40\zeta^2(X)\vee 6}}$. If moreover $x \le \sqrt{n}/18$, then the relation $T_j = f(Q_j)$ immediately implies that t-statistics $T_j$ satisfy the bound

$$|T_j| \le \frac{11x}{\sqrt{n}} \tag{3.4}$$

with probability at least $1 - 4e^{-x^2/2} - e^{-\frac{n}{40\zeta^2(X)\vee 6}}$ for each $1 \le j \le k$. Alternatively, the previous argument also implies that the random variables $T_j \mathbf{1}\{|Q_j| \le 1/2,\ V_j \ge \sigma/2\}$ satisfy the deviation inequality

$$\Pr\left(|T_j \mathbf{1}\{|Q_j| \le 1/2,\ V_j \ge \sigma/2\}| \ge \frac{11x}{\sqrt{n}}\right) \le 4e^{-x^2/2}.$$

Therefore, we conclude that the random variable $T_j$, truncated at the right level, behaves like a sub-Gaussian random variable.[3] This fact is formalized in Lemma 6.1 and is one of the key ingredients used to show that proposed estimators have sub-Gaussian deviations.

### *3.2. Non-asymptotic deviation inequalities.*

In the simplest case $p = 1$, equation (2.3) suggests that in order to bound the estimation error $\widehat{\mu}_{N,p} - \mu$, it suffices to control the average $\frac{1}{k}\sum_{j=1}^k \frac{\bar{\mu}_j - \mu}{\widehat{\sigma}_j}$ and the harmonic mean $\left(\frac{1}{k}\sum_{j=1}^k \frac{1}{\widehat{\sigma}_j}\right)^{-1}$ separately. Similar intuition holds for other values of $p$ as well. In what follows, we will always assume that $O \le Ck$ for some $C < 1$, where $O$ is the number of outliers in the sample. Define the event

$$\mathcal{E}_p := \left\{\left(\frac{1}{k}\sum_{j=1}^k \frac{1}{\widehat{\sigma}_j^p}\right)^{-1} \le \left(\frac{4\sigma}{1-C}\right)^p\right\}. \tag{3.5}$$

$\mathcal{E}_p$ holds whenever the harmonic mean of the (powers of) sample variances does not exceed the corresponding power of the true variance $\sigma^2$ by too much. In

[3] $X$ has sub-Gaussian distribution if $\exists K > 0$ such that $\forall p \ge 1$ $\mathbb{E}(|X-\mu|^p)^{1/p} \le K\sqrt{p}$). See section 2.5 in [Ver18] for the details.

particular, in the absence of outliers, we can replace $C$ by 0 in the previous event. Informally speaking, the harmonic mean of a set of numbers is controlled by its smallest elements, therefore, it is natural to expect that the event $\mathcal{E}_p$ holds with overwhelming probability; this claim will be formalized in the following lemma whose proof is deferred to Section 6.3.

**Lemma 3.2.** *Recall the contamination framework defined in Section 1. Suppose that* $\mathbb{E}|X-\mu|^{1+\delta} < \infty$ *for some* $1 \le \delta \le 2$ *and that* $O \le Ck$ *for some* $C < 1$. *Then*

$$\Pr(\mathcal{E}_p) \ge 1 - e^{-ck(1-C)(1+(\delta-1)\log n)}$$

*for some constant* $c > 0$ *that depends on* $\delta$ *and* $\mathbb{E}|X-\mu|^{1+\delta}$. *Moreover, if* $X$ *has sub-Gaussian distribution, then*

$$\Pr(\mathcal{E}_p) \ge 1 - e^{-c(1-C)N}$$

*for some constant* $c > 0$ *that depends on the distribution of* $X$.

Note that the condition $O < Ck$ only requires $C$ to be smaller than 1: it means that for our technique to reliably estimate the true mean, it suffices that any constant positive fraction of subsamples indexed by $G_1, \dots, G_k$ are free from the outliers, while the popular median-of-means estimator requires at least 50% of the subsamples to be "clean". In practical applications, this difference can be substantial, and our simulation results (see section 5) confirm this observation.

Our first result presents non-asymptotic deviation bounds for the case when the sample does not contain outliers.

**Theorem 3.1.** *Suppose that* $\mathbb{E}|X-\mu|^{1+\delta} < \infty$ *for some* $1 \le \delta \le 2$. *Then with probability at least* $1 - 2e^{-s} - ke^{-cn} - \Pr(\mathcal{E}_p^c)$,

$$|\widehat{\mu}_{N,p} - \mu| \le C_p \sigma \left( \sqrt{\frac{s+1}{N}} + \phi(\delta, n) \right) \tag{3.6}$$

*where* $c > 0$ *depends only on* $\zeta(X)$, $C_p > 0$ *depends only on* $p$, *and*

$$\phi(\delta, n) = \begin{cases} o(n^{-\delta/2}), & \delta < 2, \\ O(n^{-1}), & \delta = 2 \end{cases}$$

*as* $n \to \infty$.

Combination of Theorem 3.1 with Lemma 3.2 readily implies that $\widehat{\mu}_{N,p}$ admits sub-Gaussian deviation guarantees for $s = k \lesssim \sqrt{N/\log N}$. Indeed, in that case we get with probability at least $1 - 3e^{-k}$ that

$$|\widehat{\mu}_{N,p} - \mu| \le C_p' \sigma \sqrt{\frac{k+1}{N}}.$$

As we explain in the remark below, if $k$ is chosen appropriately, this statement can often be strengthened to yield uniform deviation guarantees holding in the range $0 \le s \le k$.

**Remark 3.1.** *Dependence of the constant $c$ on $\zeta(X)$ is inherited from Lemma 3.1. The constant $\zeta(X)$ can be arbitrary large, therefore the inequality of Theorem 3.1 does not hold with overwhelming probability uniformly over the class of distributions $\mathcal{P}_{2,\sigma}$. To achieve uniformity, we need to assume slightly more about the distribution of $X$ – for example, one may impose the "small ball" condition $Q(u) := \Pr(|X - \mathbb{E}X| \geq u) \geq \tilde{c} > 0$, or the equivalence of moments of order 2 and $2 + \delta$ for some $\delta > 0$, namely that $\mathbb{E}|X - \mathbb{E}X|^{2+\delta} \leq \tilde{C}(\mathbb{E}|X - \mathbb{E}X|^2)^{1+\delta/2}$ for some fixed $\tilde{C} > 0$. Then our bounds will depend on the constant $\tilde{c}$ or $\tilde{C}$ instead, and dependence on $\zeta(X)$ can be suppressed: for instance, when the moments of order 2 and $2 + \delta$ are equivalent, we have that $\mathbf{E}(|X - \mu|^2 \mathbf{1}\{|X - \mu| \geq \sigma(2C)^{1/\delta}\}) \leq \sigma^2/2$ in view of Markov's inequality, thus $\zeta(X) \leq (2C)^{1/\delta}$. This justifies the claim that assuming $\zeta(X)$ to be "small" is a relatively mild requirement. In simple terms, we ask that the distributions in question assign nontrivial mass to a fixed neighborhood of their means. It is also interesting to take a viewpoint that assumes the distribution of $X$ to be fixed while the parameters $n, k \to \infty$: in this case, one can establish stronger claims about mean estimation – for instance, the deviations in (1.1) can be shown to be uniform over a range of values of parameter $s$.*

**Remark 3.2.** *A more precise bound for the "bias term" $\phi(\delta, n)$ has the form*

$$\phi(\delta, n) = n^{-\delta/2} \cdot \left( n^{-\frac{2-\delta}{4}} \vee g^{\frac{2-\delta}{2+\delta}}(n^{1/4}) \right),$$

*where $g(u) = \mathbb{E}\left(|X - \mu|^{1+\delta} \mathbf{1}_{\{|X-\mu| \geq u\}}\right)$. It is therefore easy to see that whenever $k = o\left(N^{\frac{\delta-1}{\delta}}\right)$, the term $\phi(\delta, n)$ is $o(N^{-1/2})$ and the sub-Gaussian deviation guarantees (3.6) hold uniformly over $s \lesssim k$ (the latter restriction appears due to the fact that the probability of event $\mathcal{E}_p$ depends on $k$ as $e^{-ck}$).*

*In the case when $\delta = 1$, $\phi(\delta, n) = o\left(\sqrt{\frac{k}{N}}\right)$ so that sub-Gaussian deviation guarantees hold with $s = k \lesssim n$. However, if $k$ is large enough, namely, if $k\left(n^{-\frac{1}{4}} \vee g^{\frac{1}{3}}(n^{1/4})\right)^2 = O(1)$, we can still achieve the situation when $\phi(\delta, n) = O\left(N^{-1/2}\right)$. In this case, deviation guarantees hold uniformly over $s \lesssim k$. The price that we have pay however is the fact that $k$ can grow arbitrarily slowly as a function of $N$, but this is unavoidable in general as shown in [DLLO16].*

Next, we discuss the more general contamination framework described in the introduction. For each block $G_j$, we denote by $W_j$ the number of outliers in $G_j$ and by $\bar{\mu}_j^I$ (respectively $\bar{\mu}_j^C$) the sample mean corresponding to the inliers (respectively outliers) within $G_j$. For every set of outliers $\mathcal{O}$ we define

$$\alpha(\mathcal{O}) := 1 + \min_{j: W_j \neq 0} \frac{W_j(\bar{\mu}_j^I - \bar{\mu}_j^C)^2}{n\sigma^2}. \tag{3.7}$$

Informally speaking, $\alpha(\mathcal{O})$ can be viewed as a proxy for the magnitude of the outliers. The following extension of Theorem 3.1 holds.

**Theorem 3.2.** *Suppose that* $\mathbb{E}|X-\mu|^{1+\delta} < \infty$ *for some* $1 \le \delta \le 2$, *and* $O \le Ck$ *for some* $C < 1$. *Then with probability at least* $1 - 2e^{-s} - ke^{-cn} - \Pr(\mathcal{E}_p^c)$,

$$|\widehat{\mu}_{N,p} - \mu| \le \frac{C_p \sigma}{(1-C)^p}\left(\sqrt{\frac{s+1}{N}} + \phi(\delta, n) + \alpha(\mathcal{O})^{-(p-1)/2}\frac{O}{k\sqrt{n}}\right)$$

*for* $c > 0$ *depending only on* $\zeta(X)$, $C_p > 0$ *depending only on* $p$, $\alpha(\mathcal{O})$ *defined via* (3.7), *and* $\phi(\delta, n)$ *defined as in Theorem 3.1.*

One may notice that $\alpha(\mathcal{O})^{-(p-1)/2} \le 1$, and this quantity gets smaller as $p$ grows, suggesting that the estimator $\widehat{\mu}_{N,p}$ is more robust to the outliers of large magnitude as $p$ increases. Next, let us discuss the term $\frac{O}{k\sqrt{n}} = \varepsilon\sqrt{n}$ that quantifies dependence of the estimation error on the fraction of outliers $\varepsilon = \frac{O}{N}$. It is easy to see that the "best" choice of $k$ for which the terms $\phi(\delta, n)$ and $\frac{O}{k\sqrt{n}}$ are of the same order is $k \propto N\varepsilon^{\frac{2}{1+\delta}}$ yielding the error rate of $\varepsilon^{\frac{\delta}{1+\delta}}$ that is known to be optimal with respect to $\delta$ (e.g. see section 1.2 in [SCV17] or Lemma 5.4 in [Min18]). However, as the upper bound depends explicitly on the magnitude of outliers through $\alpha(\mathcal{O})$, in some scenarios it can be much smaller than the worst case given by $\frac{O}{k\sqrt{n}}$.

### 3.3. Asymptotic efficiency.

The following result establishes asymptotic efficiency (in a sense defined in section 1) of the estimator $\widehat{\mu}_{N,p}$ for any $p \ge 1$ in the absence of outliers, implying that the estimator can not be uniformly improved in general.

**Theorem 3.3.** *Suppose that* $\mathbb{E}|X-\mu|^{1+\delta} < \infty$ *for some* $1 \le \delta \le 2$. *Let* $\{k_j\}_{j\ge1} \subset \mathbb{N}$, $\{n_j\}_{j\ge1} \subset \mathbb{N}$ *be two non-decreasing, unbounded sequences satisfying* $\sqrt{N_j}\phi(\delta, n_j) = o(1)$ *as* $j \to \infty$, *where* $N_j := k_j n_j$ *and* $\phi(\delta, n)$ *was defined in remark 3.2. Then for any* $p \ge 1$,

$$\sqrt{N_j}\left(\widehat{\mu}_{N_j,p} - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right) \text{ as } j \to \infty.$$

Condition $\sqrt{N_j}\phi(\delta, n_j) = o(1)$ is essentially a requirement that the bias of estimator $\widehat{\mu}_{N_j,p}$ is asymptotically of order $o\left(N_j^{-1/2}\right)$. It is not difficult to check that the sequences $\{k_j\}_{j\ge1}$, $\{n_j\}_{j\ge1}$ with required properties exist for any distribution $P \in \mathcal{P}_{2,\sigma}$, see remark 3.2 for the details. For example, if $\mathbb{E}|X-\mu|^3 < \infty$, it suffices to require that $k_j = o(n_j)$.

Together, results of section 3 imply that the estimator $\widehat{\mu}_{N,p}$ can be viewed as a true robust alternative to the sample mean – it preserves its desirable properties such as asymptotic efficiency while being robust at the same time.

### 3.4. Comparison with existing techniques.

One of the most well-known consistent, robust estimators of the mean in the class $\mathcal{P}_{2,\sigma}$ is the median-of-means estimator [NY83, AMS96, LO11]. While it

is robust to heavy tails, adversarial contamination, and is tuning-free, it is not asymptotically efficient: indeed, according to Theorem 4 in [Min19], the asymptotic variance of the median-of-means estimator is $\frac{\pi}{2}\sigma^2$. This fact is illustrated in our numerical experiments in section 5. Another family of estimators belonging to the broad class defined via equation (2.1) is discussed in section 2.4 in [Min19] and is defined via

$$\widetilde{\mu}_N = \operatorname{argmin}_{z\in\mathbb{R}} \sum_{j=1}^{k} \rho\left(\frac{\sqrt{n}}{\Delta}(\bar{\mu}_j - z)\right)$$

where $\rho$ is Huber's loss $\rho(z) = \min\left(\frac{z^2}{2}, |z| - \frac{1}{2}\right)$ and $\Delta > 0$. The asymptotic variance of this estimator can be made arbitrarily close to $\sigma^2$, however, achieving this requires $\sigma^2$ to be known.

Construction of Catoni's estimator [Cat12] again requires knowledge of $\sigma^2$ (or its tight upper bound), moreover, it is not robust to adversarial contamination. Finally, deviation bounds for the trimmed mean estimator obtained in [LM19b] are not uniform with respect to the confidence parameter $s$ (meaning that different choices of $s$ require the estimator to be re-computed), and its asymptotic efficiency, while plausible, has not been formally established. Moreover, construction employed in [LM19b] requires sample splitting. Recently, Lee and Valiant [LV20] showed that it is possible to construct a mean estimator that achieves sub-Gaussian guarantees with essentially optimal constants, however, their estimator explicitly depends on the desired confidence level, and its asymptotic behavior is not discussed.

The only other robust, tuning free estimator that is asymptotically efficient, albeit only for a *subclass* of $\mathcal{P}_{2,\sigma}$, is a permutation-invariant version of the median-of-means estimator (which is also the higher order Hodges-Lehmann estimator). It is defined as follows: let $\mathcal{A}_N^{(n)} := \{J : \ J \subseteq \{1,\dots,N\}, \operatorname{Card}(J) = n\}$ be a collection of all distinct subsets of $\{1,\dots,N\}$ of cardinality $n$, $\bar{\theta}_J := \frac{1}{n}\sum_{j\in J} X_j$, and $\widetilde{\mu}_U := \operatorname{median}\left(\bar{\theta}_J, \ J \in \mathcal{A}_N^{(n)}\right)$. We note that $\operatorname{Card}\left(\mathcal{A}_N^{(n)}\right) = \binom{N}{n}$, so that for large $N$ and $n$ exact evaluation of $\widetilde{\mu}_U$ is not computationally feasible. The following result was established recently in [DR20]: assume that $N_j = n_j k_j$ is the sample size where $n_j, \ k_j \to \infty$ as $j \to \infty$ such that $n_j = o\left(\sqrt{N_j}\right)$. Moreover, suppose that $X$ is normally distributed with mean $\mu$ and variance $\sigma^2$. Then $\sqrt{N}\left(\widetilde{\mu}_U - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$. While is likely that the result still holds for other symmetric distributions, the condition $n_j = o\left(\sqrt{N_j}\right)$ is restrictive: for example, for non-symmetric distribution possessing 3 finite moments, the bias of the estimator $\widetilde{\mu}_U$ is of order $n_j^{-1}$, and the requirement $n_j = o\left(\sqrt{N_j}\right)$ implies that this bias is asymptotically larger than $N_j^{-1/2}$.

Finally, there is a growing body of literature related to sub-Gaussian mean estimators in $\mathbb{R}^d$, for example see the papers [DM20, LM19a, Hop20], and references therein. These works are mainly concerned with rate optimality, and questions related to asymptotic efficiency have not been investigated in detail.

## 4. Adaptation to the contamination proportion $\varepsilon$.

The number of outliers $O$ is usually unknown in practice, therefore, it is desirable to have a procedure that can adapt to this unknown quantity. Fortunately, the proposed method admits a natural adaptive version. This extension is based on the following observation: assume that $p = 1$, and consider the estimation error $\widehat{\mu}_{N,1} - \mu = \frac{\frac{1}{k}\sum_{j=1}^{k} \frac{\bar{\mu}_j - \mu}{\widehat{\sigma}_j}}{\frac{1}{k}\sum_{j=1}^{k} \frac{1}{\widehat{\sigma}_j}}$. Then the numerator of this expression admits an upper bound $\left|\frac{1}{k}\sum_{j=1}^{k} \frac{\bar{\mu}_j - \mu}{\widehat{\sigma}_j}\right| \le C_p\sigma\left(\sqrt{\frac{s}{N}} + \sqrt{\frac{O}{N}}\right)$ that holds *for all* choices of $k$ with probability at least $1 - 2e^{-s} - ke^{-cn}$, as shown in the proof of Theorem 3.2. Therefore, it suffices to choose $k$ such that the harmonic mean $\frac{k}{\sum_{j=1}^{k} \frac{1}{\widehat{\sigma}_j}}$ is a good, in a relative sense, estimator of $\sigma$. Fortunately, the harmonic mean of standard deviations is a fully data-dependent quantity that can be evaluated for any $k$; similar intuition holds for other values of $p$ as well.

Based on the previous observation, we propose an adaptive estimator $\widetilde{\mu}_p$ defined as follows. We will choose $k$ as the smallest integer, on a logarithmic scale, which guarantees that $\frac{k}{\sum_{j=1}^{k} \frac{1}{\widehat{\sigma}_j^p}}$ is not too large compared to $\sigma^p$, in a sense defined by (3.5). To this end, we only need to obtain a good preliminary estimator of $\sigma$ that we can compare the harmonic means to. Assume that we are already given an estimator $\widetilde{\sigma}$ such that

$$1/20 \le \frac{\widetilde{\sigma}}{\sigma} \le 4 \tag{4.1}$$

with large probability. The above assumption is not restrictive since, as we will show in section 6.8, one can construct $\widetilde{\sigma}$ such that (4.1) holds with probability at least $1 - e^{-cN}$ for some absolute $c > 0$, under mild conditions. Next, for each positive integer $k$, set

$$\widetilde{\mathcal{E}}_p(k) := \left\{ \left(\frac{1}{k}\sum_{j=1}^{k} \frac{1}{\widehat{\sigma}_j^p}\right)^{-1} \le \left(\frac{80\widetilde{\sigma}}{1-C}\right)^p \right\}.$$

Finally, define $\tilde{k}$ via $\log_2 \tilde{k} := \inf\left\{ i \in \{1, \dots, \lfloor \log_2 N \rfloor\} : \widetilde{\mathcal{E}}_p(2^i) \text{ holds} \right\} \vee 1$, [4] and the corresponding estimator $\widetilde{\mu}_p(s) := \widehat{\mu}_{N,p}(\tilde{k} \vee \lfloor s \rfloor + 1)$. The following bound is the main result of this section; essentially, it states that $\widetilde{\mu}_p(s)$ is a robust estimator that is fully adaptive and provides sub-Gaussian deviation guarantees.

**Theorem 4.1.** *Suppose that* $\mathbb{E}|X - \mu|^2 < \infty$. *Assume that* $2 \le O \le N/4$ *and that* $\widetilde{\sigma}$ *satisfies* (4.1). *Then with probability at least* $1 - 2\log_2(3O)e^{-s} - (O \vee s)e^{-cN/(O \vee s)}$,

$$|\widetilde{\mu}_p(s) - \mu| \le C_p\sigma\left(\sqrt{\frac{s}{N}} + \sqrt{\frac{O}{N}}\right),$$

[4] We assume that the infimum over the empty set is equal to $-\infty$.

*where $c > 0$ depends only on the distribution of $X$ and $C_p > 0$ depends only on $p$.*

## 5. Numerical simulation results.

The goal of this section is to compare performance of the estimators $\widehat{\mu}_{N,p}$ for different values of $p \geq 1$, as well as evaluate their performance against the benchmarks given by other popular techniques such as the median-of-means estimator and the "oracle" trimmed mean (labeled "trim" in the figures) estimator that takes the contamination proportion $\varepsilon$ as its input.

Our simulation setup was defined as follows: $N = 2500$ observations from half-t distribution[5] with 4 degrees of freedom (d.f.). This distribution is asymmetric, therefore, results allow us to evaluate the degree to which the bias affects performance of different robust estimators; linear transformation has been applied so that the mean and variance of generated data are 0 and 1 respectively. Next, $O \in \{0, 50, 100, 150\}$ randomly selected observations have been replaced by the outliers given by the point mass at $x_0 = 10^3$; this type of outliers appears to be most challenging for the trimmed mean estimator as it creates bias due to "inliers" being removed only from one of the tails of the distribution. We compared 4 estimators: the median-of-means (MOM) estimator defined after equation (2.1), estimators $\widehat{\mu}_{N,1}$ and $\widehat{\mu}_{N,2}$ corresponding to the choice of weights (2.2) with $p = 1$ and $p = 2$, as well as the "oracle" trimmed mean estimator [LM19b] that knows the number of outliers. Specifically, trimmed mean was computed by removing the smallest $\lfloor \varepsilon N \rfloor + 5$ and well as largest $\lfloor \varepsilon N \rfloor + 5$ observations, where 5 was added to account for the outliers due to the heavy tails, and averaging over the rest. Estimators $\widehat{\mu}_{N,1}$, $\widehat{\mu}_{N,2}$ as well as MOM were evaluated for various values of parameter $k \in \{25, 50, 75, 100, 125, 150, 175, 200\}$ that controls the number of subgroups.

For each combination of values of $O$ and $k$, simulation was repeated 1000 times; we present 3 summary statistics in the plots below: the average error (Figure 1), the standard deviation (Figure 2) and the maximal (over 1000 repetitions) absolute error (Figure 3).

Overall, numerical experiments confirm our theoretical findings. Here is the summary of our simulation results:

1. In the setup with no contamination ($O = 0$), all estimators showed good performance, with $\widehat{\mu}_{N,1}$ slightly but consistently beating $\widehat{\mu}_{N,2}$ on average, but $\widehat{\mu}_{N,2}$ had the smallest maximal error among all estimators; empirical standard deviations of $\widehat{\mu}_{N,1}$ and $\widehat{\mu}_{N,2}$ were consistent with theory-predicted values;
2. as $O$ increased, $\widehat{\mu}_{N,2}$ was performing better that $\widehat{\mu}_{N,1}$, while both estimators were significantly better than MOM;

[5] $X$ has half-t distribution with $\nu$ d.f. if $X = |Y|$ where $Y$ has Student's t-distribution with $\nu$ d.f.

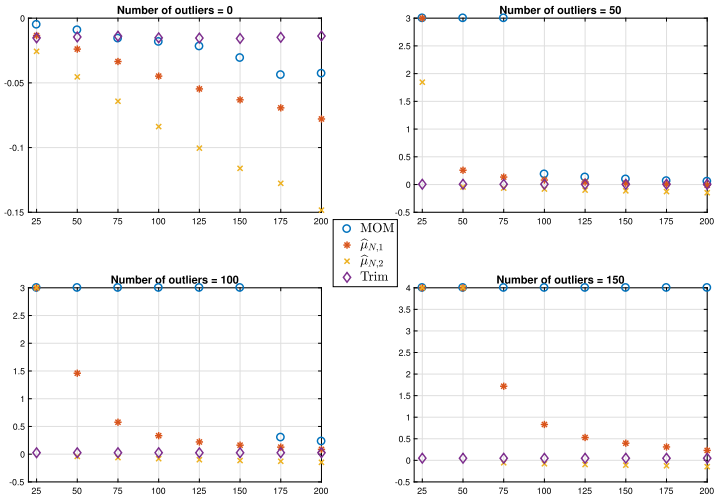


FIG 1. *Average estimation error over 1000 runs of the experiment; large values were truncated to show results on appropriate scale.*

3. both $\widehat{\mu}_{N,1}$ and $\widehat{\mu}_{N,2}$ showed consistent performance as the number of blocks $k$ increased; moreover, unlike MOM, the estimators performed well even in the challenging setup where $O \simeq k$.

## 6. Proofs.

This section contains detailed proofs of the main results of the paper.

### *6.1. Results related to the deviations of self-normalized sums.*

*6.1.1. Proof of Lemma 3.1.*

Let $Z_i := \frac{X_i - \mu}{\sigma}$, and observe that

$$\begin{aligned}
\Pr\left(\sum_{i=1}^n Z_i^2 \le \frac{n}{4}\right) &\le \Pr\left(\sum_{i=1}^n Z_i^2 \mathbf{1}\{|Z_i| \le \zeta(X)\} \le \frac{n}{4}\right) \\
&= \Pr\left(\sum_{i=1}^n \left(Z_i^2 \mathbf{1}\{|Z_i| \le \zeta(X)\} - \mathbb{E} Z_i^2 \mathbf{1}\{|Z_i| \le \zeta(X)\}\right) \le \frac{n}{4} - nH\right) \\
&\le \Pr\left(\sum_{i=1}^n \left(Z_i^2 I\{|Z_i| \le \zeta(X)\} - \mathbb{E} Z_i^2 \mathbf{1}\{|Z_i| \le \zeta(X)\}\right) \le -\frac{n}{4}\right),
\end{aligned}$$

where $H = \mathbb{E} Z^2 \mathbf{1}\{|Z| \le \zeta(X)\}$. The last inequality follows from the inequality $H = \mathbb{E} Z^2 \mathbf{1}\{|Z| \le \zeta(X)\} \ge 1/2$ implied by the definition of $\zeta(X)$. The right

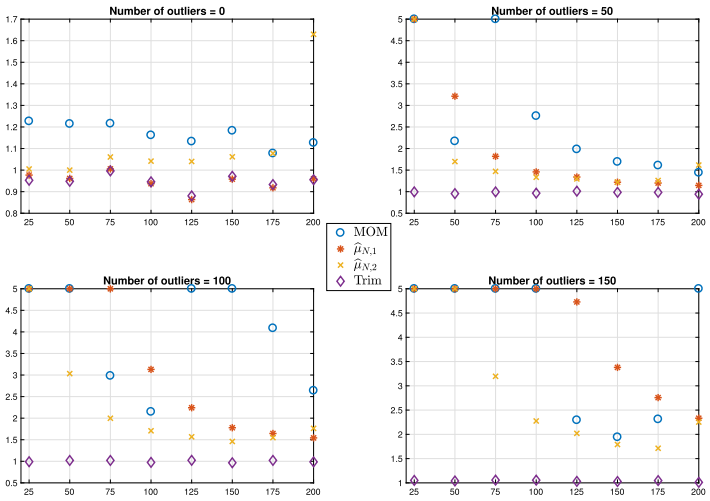


FIG 2. *Standard deviation (rescaled by $\sqrt{N}$) over 1000 runs of the experiment; large values were truncated to show results on appropriate scale.*

side of the previous display can be upper bounded via Bernstein's inequality by $e^{-\frac{n}{32(\zeta^2(X)+\zeta(X)/12)}}$ once we notice that

$$\mathbb{E}\left(Z^4\mathbf{1}(|Z|\le\zeta(X)\right)\le\zeta^2(X).$$

The claim follows from an algebraic inequality $12\zeta^2(X)+\zeta(X)\le 15\zeta^2(X)\vee 2$ entailing that $e^{-\frac{n}{32(\zeta^2(X)+\zeta(X)/12)}}\le e^{-\frac{n}{40\zeta^2(X)\vee 6}}$.

### *6.1.2. Bounds for the moment generating function of the t-statistic.*

Recall that $T_j:=\frac{\bar{\mu}_j-\mu}{\widehat{\sigma}_j}$ and $Q_j:=\frac{\bar{\mu}_j-\mu}{V_j}$ where $V_j^2=\frac{1}{|G_j|}\sum_{i\in G_j}(X_i-\mu)^2$, $j=1,\dots,k$. For all $p\ge 1$ define

$$w_j=\frac{\sigma^{p-1}T_j}{\widehat{\sigma}_j^{p-1}}\mathbf{1}\{\mathcal{E}_j\}-\mathbb{E}\left(\frac{\sigma^{p-1}T_j}{\widehat{\sigma}_j^{p-1}}\mathbf{1}\{\mathcal{E}_j\}\right)$$

where $\mathcal{E}_j=\{|Q_j|\le 1/2\}\cap\{V_j\ge\sigma/2\}$.

**Lemma 6.1.** *There exists $c_p>0$ such that, for all $\lambda\in\mathbf{R}$. we have*

$$\mathbb{E}(e^{\lambda w_1})\le e^{c_p\lambda^2/(2n)}.$$

*Proof.* We start by observing that on event $\mathcal{E}_1$, $\widehat{\sigma}_1=V_1\sqrt{1-Q_1^2}\ge\frac{\sqrt{3}\sigma}{4}$. Hence for all $t>0$, the discussion following (3.4) and the inequality $|T_1\mathbf{1}\{\mathcal{E}_1\}|\le\frac{\sqrt{3}}{3}$

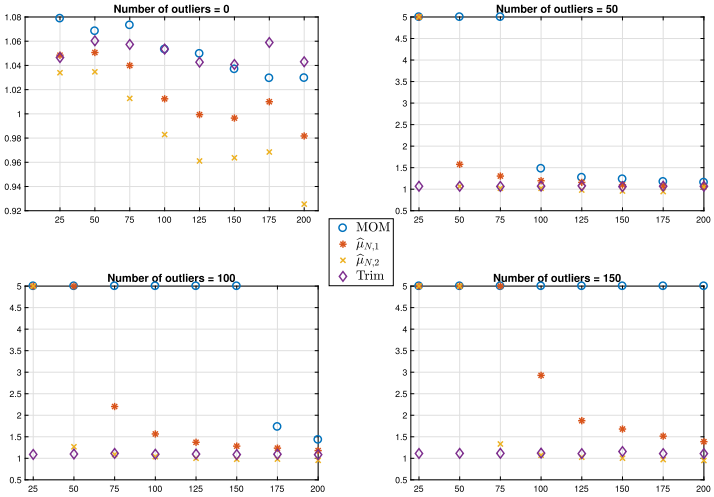


FIG 3. *Maximal absolute error over 1000 runs of the experiment; large values were truncated to show results on appropriate scale.*

imply that

$$\Pr\left(\left|\frac{\sigma^{p-1}T_1}{\widehat{\sigma}_1^{p-1}}\mathbf{1}\{\mathcal{E}_1\}\right| \geq t\right) \leq \Pr\left(|T_1\mathbf{1}\{\mathcal{E}_1\}| \geq t\left(\frac{\sqrt{3}}{4}\right)^{p-1}\right) \leq 4e^{-c_p' nt^2}$$

where $c_p' = \frac{1}{2\cdot 11^2}\left(\frac{\sqrt{3}}{4}\right)^{2p-2}$. Next, let $\widetilde{w}_1$ be an independent copy of $w_1$, and note that

$$\Pr\left(|w_1 - \widetilde{w}_1| \geq t\right) \leq 8e^{-c_p' nt^2/4}. \tag{6.1}$$

It follows from Jensen's inequality that

$$\mathbb{E}e^{\lambda w_1} \leq \mathbb{E}e^{\lambda(w_1 - \widetilde{w}_1)}.$$

Finally, is well-known that, in view of (6.1), the latter is bounded by $e^{\lambda^2 c_p/(2n)}$ for some $c_p > 0$ only depending on $p$ (for instance, this follows from Proposition 2.5.2 in [Ver18]). □

### *6.2. Auxiliary technical results.*

**Lemma 6.2.** *Let $p \geq 1$ and $\delta \geq 1$. Assume that $\mathbb{E}(|X-\mu|^{1+\delta}) < \infty$. Then for any $1 \leq j \leq k$*

$$\sigma^{p-1}\mathbb{E}\left(\frac{\bar{\mu}_j - \mu}{V_j^p}\mathbf{1}\{V_j^2 \geq \sigma^2/4\}\right) = o\left(\frac{1}{n^{\delta/2}}\right),$$

*for $\delta < 2$. At the same time, for $\delta \geq 2$, we have*

$$\sigma^{p-1}\mathbb{E}\left(\frac{\bar{\mu}_j - \mu}{V_j^p}\mathbf{1}\{V_j^2 \geq \sigma^2/4\}\right) = O\left(\frac{1}{n}\right).$$

*Proof.* Due to homogeneity, we can assume that $\sigma = 1$ without loss of generality. We will also assume that $\mu = 0$, otherwise $X_j$ should be replaced by $X_j - \mu$ for all $j$. Observe that

$$\sigma^{p-1}\mathbb{E}\left(\frac{\bar{\mu}_j - \mu}{V_j^p}\mathbf{1}\{V_j^2 \geq \sigma^2/4\}\right) = n^{p/2}\mathbb{E}\left(\frac{X_1}{(X_1^2 + Z^2)^{p/2}}\mathbf{1}\{X_1^2 + Z^2 \geq n/4\}\right), \tag{6.2}$$

where $Z = \sqrt{\sum_{i=2}^n X_i^2}$. Consider the event $\mathcal{O}_1 = \{Z^2 \geq n/4\}$, and recall that in view of Lemma 3.1

$$\Pr(\mathcal{O}_1^c) \leq e^{-cn}$$

for some $c = c(P) > 0$ that depends on the distribution $P$ of $X$. Consider the event

$$\mathcal{O}_2 = \{|X_1| \leq \alpha_n\sqrt{n}/2\},$$

where the sequence $\{\alpha_n\}_{n\geq 1}$ is defined as follows: consider a non-increasing function $g(u) = \mathbb{E}(|X|^{1+\delta}\mathbf{1}_{\{|X|\geq u\}})$, and observe that $\lim_{u\to\infty} g(u) = 0$. Therefore, taking $\alpha_n := g(n^{1/4})^{1/(2+\delta)} \vee n^{-1/4}$, we get that $\alpha_n \to 0$, and moreover

$$\lim_{n\to\infty}\frac{g(\alpha_n\sqrt{n})}{\alpha_n^{1+\delta}} \leq \lim_{n\to\infty}\frac{g(n^{1/4})}{\alpha_n^{1+\delta}} \leq \lim_{n\to\infty} g^{\frac{1}{2+\delta}}(n^{1/4}) = 0.$$

It is easy to see that

$$\Pr(\mathcal{O}_2^c) = o\left(\frac{1}{n^{(1+\delta)/2}}\right),$$

and that

$$\mathbb{E}(|X_1|\mathbf{1}_{\{\mathcal{O}_2^c\}}) = o\left(\frac{1}{n^{\delta/2}}\right).$$

Indeed, Markov's inequality implies that

$$\Pr(\mathcal{O}_2^c) \leq \frac{2^{1+\delta}\mathbb{E}(|X|^{1+\delta}\mathbf{1}_{\{\mathcal{O}_2^c\}})}{\alpha_n^{1+\delta}n^{(1+\delta)/2}} = \frac{2^{1+\delta}g(\alpha_n\sqrt{n}/2)}{\alpha_n^{1+\delta}n^{(1+\delta)/2}},$$

and

$$\mathbb{E}(|X_1|\mathbf{1}_{\{\mathcal{O}_2^c\}}) \leq \frac{2^{\delta}\mathbb{E}(|X|^{1+\delta}\mathbf{1}_{\{\mathcal{O}_2^c\}})}{\alpha_n^{\delta}n^{\delta/2}} = \frac{2^{\delta}g(\alpha_n\sqrt{n}/2)}{\alpha_n^{\delta}n^{\delta/2}}$$
$$\leq 2^{\delta}\left(g(n^{1/4})\right)^{2/(2+\delta)}n^{-\delta/2},$$

where we used Hölder's inequality. Next, we will reduce the problem to the case where $X$ and $Z$ are bounded. Define the event $\widetilde{\mathcal{O}} := \mathcal{O}_1 \cap \mathcal{O}_2$. Then $1 = \mathbf{1}_{\widetilde{\mathcal{O}}} + \mathbf{1}_{\mathcal{O}_1 \setminus \mathcal{O}_2} + \mathbf{1}_{\mathcal{O}_2 \setminus \mathcal{O}_1}$, and

$$\begin{aligned}
&\left|\mathbb{E}\left(\frac{X_1}{(X_1^2+Z^2)^{p/2}}\mathbf{1}\{X_1^2+Z^2 \geq n/4\}\right)\right| \\
&\leq \left|\mathbb{E}\left(\frac{X_1}{(X_1^2+Z^2)^{p/2}}\mathbf{1}_{\{\widetilde{\mathcal{O}}\}}\right)\right| + (n/4)^{-p/2}\left(\mathbb{E}(|X_1|\mathbf{1}\{\mathcal{O}_2^c\} + \frac{\alpha_n\sqrt{n}}{2}\Pr(\mathcal{O}_1^c)\right) \\
&\leq \left|\mathbb{E}\left(\frac{X_1}{(X_1^2+Z^2)^{p/2}}\mathbf{1}_{\{\widetilde{\mathcal{O}}\}}\right)\right| + o\left(\frac{1}{n^{(p+\delta)/2}}\right). \quad (6.3)
\end{aligned}$$

Letting $F$ be the distribution function of $X$, we deduce that conditionally on $Z$

$$\begin{aligned}
&\left|\mathbb{E}\left(\frac{X_1}{(X_1^2+Z^2)^{p/2}}\mathbf{1}_{\{\widetilde{\mathcal{O}}\}}\right)\right| = \left|\int_{-\alpha_n\sqrt{n}/2}^{\alpha_n\sqrt{n}/2} \frac{x}{(x^2+Z^2)^{p/2}}\mathbf{1}_{\{\mathcal{O}_1\}}dF(x)\right| \\
&\leq \left|\int_{-\alpha_n\sqrt{n}/2}^{\alpha_n\sqrt{n}/2}\left(\frac{x}{(x^2+Z^2)^{p/2}} - \frac{x}{Z^p}\right)\mathbf{1}_{\{\mathcal{O}_1\}}dF(x)\right| + \left(\frac{2}{\sqrt{n}}\right)^p \mathbb{E}\left(|X_1|\mathbf{1}_{\{\mathcal{O}_2^c\}}\right) \\
&\leq \left|\int_{-\alpha_n\sqrt{n}/2}^{\alpha_n\sqrt{n}/2} \frac{xZ^p(1-(1+x^2/Z^2)^{p/2})}{(x^2+Z^2)^{p/2}Z^p}\mathbf{1}_{\{\mathcal{O}_1\}}dF(x)\right| + o\left(\frac{1}{n^{(p+\delta)/2}}\right) \\
&\leq \int_{-\alpha_n\sqrt{n}/2}^{\alpha_n\sqrt{n}/2} \frac{p|x|^3}{2Z^{p+2}}\mathbf{1}_{\{\mathcal{O}_1\}}dF(x) + o\left(\frac{1}{n^{(p+\delta)/2}}\right) \\
&\leq C(p)\mathbb{E}|X|^{1+\delta}\frac{\alpha_n^{2-\delta}}{n^{(p+\delta)/2}} + o\left(\frac{1}{n^{(p+\delta)/2}}\right). \quad (6.4)
\end{aligned}$$

In the derivation above, we used the elementary inequality

$$\frac{(1+t)^{p/2}-1}{(1+t)^{p/2}} = \frac{\int_1^{1+t}\frac{p}{2}y^{p/2-1}dy}{(1+t)^{p/2}} \leq \frac{pt}{2}\frac{(1+t)^{p/2-1}}{(1+t)^{p/2}} \leq pt/2 \quad (6.5)$$

for $0 < t := \frac{x^2}{Z^2}$ and the fact that $\mathbb{E}|X|^{1+\delta} < \infty$. Combining (6.4) with (6.2), (6.3), we see that

$$\sigma^{p-1}\mathbb{E}\left(\frac{\bar{\mu}_j - \mu}{V_j^p}\mathbf{1}\{V_j^2 \geq \sigma^2/4\}\right) = o\left(\frac{1}{n^{\delta/2}}\right)$$

whenever $\delta < 2$ and that

$$\sigma^{p-1}\mathbb{E}\left(\frac{\bar{\mu}_j - \mu}{V_j^p}\mathbf{1}\{V_j^2 \geq \sigma^2/4\}\right) = O\left(\frac{1}{n}\right),$$

for $\delta = 2$ (in fact, in this case all the terms are of order $o\left(n^{-1}\right)$ besides $C(p)n^{p/2}\mathbb{E}|X|^{1+\delta}\frac{\alpha_n^{2-\delta}}{n^{(p+\delta)/2}}$ which is $O\left(n^{-1}\right)$). □

**Remark 6.1.** *It follows from the previous argument that the term* $o\left(\frac{1}{n^{\delta/2}}\right)$ *takes the form*

$$n^{-\delta/2} \cdot \left( n^{-\frac{2-\delta}{4}} \vee g^{\frac{2-\delta}{2+\delta}}(n^{1/4}) \right).$$

**Remark 6.2.** *The key quantity of interest in the previous proof is given by the expression*

$$\left| \mathbb{E}\left( \frac{X_1}{(X_1^2+Z^2)^{p/2}} \mathbf{1}\{X_1^2+Z^2 \geq n/4\} \right) \right|$$

*that was then estimated from above. Let us present a counterexample showing that one cannot improve the result of Lemma 6.2 when* $\delta \geq 2$ *for* $p=1$*. To this end, let* $X$ *be a random variable such that* $\Pr(X=a) = 1/(1+a^2)$ *and* $\Pr(X=-1/a) = a^2/(a^2+1)$ *for some* $1 < a^2 \leq 2$ *and assume that* $n \geq 8$*. Observe that* $X$ *is a.s. bounded by* $a$*, centered, and has variance* $1$*.*

*Given* $x, y > 0$*, we say that* $x \asymp y$ *when* $c \leq x/y \leq C$ *for some absolute constants* $c, C > 0$*. Let* $\mathbb{E}_Z$ *denote the conditional expectation with respect to* $Z$*. It is easy to check that on the event* $\mathcal{A} := \{Z^2 \geq n/4\}$ *we have*

$$\begin{aligned}
\mathbb{E}_Z \left( \frac{X_1}{\sqrt{X_1^2+Z^2}} \mathbf{1}\{X_1^2+Z^2 \geq n/4\} \mathbf{1}\{\mathcal{A}\} \right) &= \mathbb{E}_Z \left( \frac{X_1}{\sqrt{X_1^2+Z^2}} \right) \mathbf{1}\{\mathcal{A}\} \\
&= \frac{a}{\sqrt{a^2+Z^2}} \frac{1}{1+a^2} \mathbf{1}\{\mathcal{A}\} - \frac{1}{a\sqrt{1/a^2+Z^2}} \frac{a^2}{1+a^2} \mathbf{1}\{\mathcal{A}\} \\
&= \frac{a}{1+a^2} \frac{\sqrt{1/a^2+Z^2} - \sqrt{a^2+Z^2}}{\sqrt{a^2+Z^2}\sqrt{1/a^2+Z^2}} \mathbf{1}\{\mathcal{A}\} \\
&= \frac{a}{1+a^2} \frac{1/a^2 - a^2}{\sqrt{a^2+Z^2}\sqrt{1/a^2+Z^2}(\sqrt{1/a^2+Z^2}+\sqrt{a^2+Z^2})} \mathbf{1}\{\mathcal{A}\} \\
&\asymp \frac{a}{1+a^2} \frac{a^2-1/a^2}{Z^3} \mathbf{1}\{\mathcal{A}\} \asymp \frac{1}{n^{3/2}} \mathbf{1}\{\mathcal{A}\},
\end{aligned}$$

*where we have used that on* $\mathcal{A}$ *both* $a^2$ *and* $1/a^2$ *are smaller than* $Z^2$ *and that* $Z^2 \asymp n$*. Since* $a$ *does not depend on* $n$*,* $X$ *is a.s. bounded by an absolute constant, and* $\Pr(\mathcal{A}) \geq 1 - e^{-cn}$ *for some absolute constant* $c > 0$*. Hence*

$$\mathbb{E}\left( \frac{X_1}{\sqrt{X_1^2+Z^2}} \mathbf{1}\{X_1^2+Z^2 \geq n/4\} \right) \asymp \frac{1}{n^{3/2}} \Pr(\mathcal{A}) + \Pr(\mathcal{A}^c) \asymp \frac{1}{n^{3/2}}.$$

*It follows that, for* $p=1$*, we have*

$$\begin{aligned}
&\mathbb{E}\left( \frac{\bar{\mu}_j - \mu}{V_j} \mathbf{1}\{V_j^2 \geq \sigma^2/4\} \right) \\
&\qquad = n^{1/2} \mathbb{E}\left( \frac{X_1}{(X_1^2+Z^2)^{1/2}} \mathbf{1}\{X_1^2+Z^2 \geq n/4\} \right) \asymp n^{-1}.
\end{aligned}$$

*Although* $X$ *admits infinitely many moments, the previous bound cannot be improved beyond three moments due to the asymmetry of the distribution of* $X$*.*

**Lemma 6.3.** *Let $p \geq 1$. If $Var(X) < \infty$, then*

$$\lim_{n\to\infty} \sigma^{2p-2} n\,\mathbb{E}\left(\frac{\bar{\mu}_1 - \mu}{V_1^p}\right)^2 \mathbf{1}\{V_1^2 \geq \sigma^2/4\} = 1.$$

*Proof.* Again, we can assume without loss of generality that $\sigma^2 = 1$ and that $\mathbb{E}X = 0$. Observe that

$$n\,\mathbb{E}\left(\frac{\bar{\mu}_1 - \mu}{V_1^p}\right)^2 \mathbf{1}\{V_1^2 \geq \sigma^2/4\} = n^p\mathbb{E}\left(\frac{X_1^2}{(X_1^2+Z^2)^p}\mathbf{1}\{X_1^2+Z^2 \geq n/4\}\right) + n^p\mathbb{E}\left(\frac{X_1 Y}{(X_1^2+Z^2)^p}\mathbf{1}\{X_1^2+Z^2 \geq n/4\}\right),$$

where $Y = \sum_{i=2}^n X_i$ and $Z = \sqrt{\sum_{i=2}^n X_i^2}$. It is clear that

$$n^p \frac{X_1^2}{(X_1^2+Z^2)^p}\mathbf{1}\{X_1^2+Z^2 \geq n/4\} \to X_1^2 \text{ in probability.}$$

Indeed, $\mathbf{1}\{X_1^2+Z^2 \geq n/4\} \to 1$ in probability in view of Lemma 3.1, while $\left(\frac{X_1^2+Z^2}{n}\right)^p \to 1$ in probability by the Law of Large Numbers.

Moreover, $n^p \frac{X_1^2}{(X_1^2+Z^2)^p}\mathbf{1}\{X_1^2+Z^2 \geq n/4\} \leq 4^p X_1^2$. Therefore

$$n^p\mathbb{E}\left(\frac{X_1^2}{(X_1^2+Z^2)^p}\mathbf{1}\{X_1^2+Z^2 \geq n/4\}\right) \xrightarrow{n\to\infty} 1.$$

It remains to prove that

$$n^p\mathbb{E}\left(\frac{X_1 Y}{(X_1^2+Z^2)^p}\mathbf{1}\{X_1^2+Z^2 \geq n/4\}\right) \xrightarrow{n\to\infty} 0.$$

Consider the event $\mathcal{O}_1 = \{Z^2 \geq n/4\}$, and recall that $\Pr(\mathcal{O}_1^c) \leq e^{-cn}$ for some $c > 0$ that depends on the distribution of $X$ as given in Lemma 3.1. We will also need to consider the event

$$\mathcal{O}_2 = \{|X_1| \leq \alpha_n\sqrt{n}/2\}$$

where $(\alpha_n)_n$ is defined as in Lemma 6.2 with $\delta = 1$. Namely, consider the non-increasing function $g(u) = \mathbb{E}(|X|^2\mathbf{1}_{\{|X|\geq u\}})$, and define $\alpha_n = g(n^{1/4})^{1/3} \vee n^{-1/4}$, so that $\alpha_n \to 0$ and

$$\lim_{n\to\infty} \frac{g(\alpha_n\sqrt{n})}{\alpha_n^2} \leq \lim_{n\to\infty} \frac{g(n^{1/4})}{\alpha_n^2} = 0.$$

As in the proof of Lemma 6.2, we deduce that $\Pr(\mathcal{O}_2^c) = o\left(\frac{1}{n}\right)$ and that

$$\mathbb{E}(|X_1|\mathbf{1}_{\{\mathcal{O}_2^c\}}) = o\left(\frac{1}{\sqrt{n}}\right). \tag{6.6}$$

Next, we will reduce the problem to the case where $X$ and $Z$ are bounded. Let $\widetilde{\mathcal{O}} := \mathcal{O}_1 \cap \mathcal{O}_2$. Then

$$n^p \left| \mathbb{E} \frac{X_1 Y}{(X_1^2+Z^2)^p} \mathbf{1}\{X_1^2+Z^2 \geq n/4\} \right| \leq n^p \left| \mathbb{E}\left( \frac{X_1 Y}{(X_1^2+Z^2)^p} \mathbf{1}_{\{\widetilde{\mathcal{O}}\}} \right) \right| + c_1(p)\sqrt{n}\mathbb{E}(|X_1|\mathbf{1}_{\mathcal{O}_2^c}) + c_2(p) n \left(\Pr(\mathcal{O}_1^c)\right)^{1/2}. \quad (6.7)$$

Indeed, $1 = \mathbf{1}_{\widetilde{\mathcal{O}}} + \mathbf{1}_{\mathcal{O}_1 \setminus \mathcal{O}_2} + \mathbf{1}_{\mathcal{O}_2 \setminus \mathcal{O}_1}$, and

$$\begin{aligned} &\left| \mathbb{E} \frac{X_1 Y}{(X_1^2+Z^2)^p} \mathbf{1}\{X_1^2+Z^2 \geq n/4\} \mathbf{1}_{\{\mathcal{O}_1 \setminus \mathcal{O}_2\}} \right| \\ &\qquad \leq \left(\frac{4}{n}\right)^{p-1/2} \mathbb{E}\left| X_1 \mathbf{1}_{\{\mathcal{O}_2^c\}} \frac{2Y}{\sqrt{n}} \right| \\ &\qquad = \left(\frac{4}{n}\right)^{p-1/2} \mathbb{E}\left| X_1 \mathbf{1}_{\{\mathcal{O}_2^c\}} \right| \mathbb{E}\left| \frac{2Y}{\sqrt{n}} \right| \\ &\qquad \leq 2\left(\frac{4}{n}\right)^{p-1/2} \mathbb{E}\left| X_1 \mathbf{1}_{\{\mathcal{O}_2^c\}} \right| := \frac{c_1(p)}{n^{p-1/2}} \mathbb{E}\left| X_1 \mathbf{1}_{\{\mathcal{O}_2^c\}} \right| \end{aligned}$$

as $\mathbb{E}\left|\frac{Y}{\sqrt{n}}\right| \leq \frac{\mathbb{E}^{1/2} Y^2}{\sqrt{n}} \leq 1$. Moreover,

$$\begin{aligned} \left| \mathbb{E} \frac{X_1 Y}{(X_1^2+Z^2)^p} \mathbf{1}\{X_1^2+Z^2 \geq n/4\} \mathbf{1}_{\{\mathcal{O}_2 \setminus \mathcal{O}_1\}} \right| &\leq \left(\frac{4}{n}\right)^{p-1/2} \frac{\alpha_n \sqrt{n}}{2} \mathbb{E}\left| \frac{2Y}{\sqrt{n}} \mathbf{1}_{\mathcal{O}_1^c} \right| \\ &\leq 2\left(\frac{4}{n}\right)^{p-1/2} \frac{\alpha_n \sqrt{n}}{2} \left(\Pr(\mathcal{O}_1^c)\right)^{1/2} := \frac{c_2(p)}{n^{p-1}} \left(\Pr(\mathcal{O}_1^c)\right)^{1/2}, \end{aligned}$$

thus (6.7) follows. Next, letting $F$ be the distribution function of $X$, we deduce that conditionally on $(Y, Z)$,

$$\begin{aligned} \left| \mathbb{E}\left[ \frac{X_1 Y}{(X_1^2+Z^2)^p} \mathbf{1}_{\{\widetilde{\mathcal{O}}\}} \middle| Y, Z \right] \right| &= \left| \int_{-\alpha_n\sqrt{n}/2}^{\alpha_n\sqrt{n}/2} \frac{xY}{(x^2+Z^2)^p} \mathbf{1}_{\{\mathcal{O}_1\}} dF(x) \right| \\ &= \left| \int_{-\alpha_n\sqrt{n}/2}^{\alpha_n\sqrt{n}/2} \left( \frac{xY}{(x^2+Z^2)^p} - \frac{xY}{Z^{2p}} \right) \mathbf{1}_{\{\mathcal{O}_1\}} dF(x) \right| \\ &\quad + \left(\frac{4}{n}\right)^{p-1/2} \frac{|Y|\mathbf{1}_{\mathcal{O}_1}}{Z} \mathbb{E}(|X_1|\mathbf{1}_{\{\mathcal{O}_2^c\}}) \\ &\leq \left| \int_{-\alpha_n\sqrt{n}/2}^{\alpha_n\sqrt{n}/2} \frac{xYZ^{2p}(1-(1+x^2/Z^2)^p)}{(x^2+Z^2)^p Z^{2p}} \mathbf{1}_{\{\mathcal{O}_1\}} dF(x) \right| \\ &\quad + \left(\frac{4}{n}\right)^{p-1/2} \frac{|Y|\mathbf{1}_{\mathcal{O}_1}}{Z} \mathbb{E}(|X_1|\mathbf{1}_{\{\mathcal{O}_2^c\}}) \end{aligned}$$

$$\leq \int_{-\alpha_n\sqrt{n}/2}^{\alpha_n\sqrt{n}/2} \frac{p|x|^3|Y|}{Z^{2p+2}} \mathbf{1}_{\{\mathcal{O}_1\}} dF(x) + \left(\frac{4}{n}\right)^{p-1/2} \frac{|Y|\mathbf{1}_{\mathcal{O}_1}}{Z} \mathbb{E}(|X_1|\mathbf{1}_{\{\mathcal{O}_2^c\}}).$$

In the derivation above, we used the bound

$$\left|\int_{-\alpha_n\sqrt{n}/2}^{\alpha_n\sqrt{n}/2} \frac{xY}{Z^{2p}} \mathbf{1}_{\{\mathcal{O}_1\}} dF(x)\right| \leq \left(\frac{4}{n}\right)^{p-1/2} \frac{|Y|\mathbf{1}_{\mathcal{O}_1}}{Z} \left|\int_{\mathbb{R}} x dF(x) - \int_{-\alpha_n\sqrt{n}/2}^{\alpha_n\sqrt{n}/2} x dF(x)\right| = \left(\frac{4}{n}\right)^{p-1/2} \frac{|Y|\mathbf{1}_{\mathcal{O}_1}}{Z} \left|\mathbb{E}X_1\mathbf{1}_{\mathcal{O}_2^c}\right|$$

and relation

$$\left|\frac{xYZ^{2p}(1-(1+x^2/Z^2)^p)}{(x^2+Z^2)^pZ^{2p}}\right| = \left|\frac{xY(1-(1+x^2/Z^2)^p)}{(x^2/Z^2+1)^pZ^{2p}}\right| \leq \frac{p|x|^3|Y|}{Z^{2p+2}},$$

where the last inequality follows from an elementary bound (6.5). Moreover,

$$\mathbb{E}\left(\int_{-\alpha_n\sqrt{n}/2}^{\alpha_n\sqrt{n}/2} \frac{p|x|^3|Y|}{Z^{2p+2}} \mathbf{1}_{\{\mathcal{O}_1\}} dF(x)\right) \leq 2p\left(\frac{4}{n}\right)^{p+1/2} \mathbb{E}\left(\frac{|Y|}{\sqrt{n}}\right) \int_{-\alpha_n\sqrt{n}/2}^{\alpha_n\sqrt{n}/2} |x|^3 dF(x) \leq 4p\alpha_n\left(\frac{4}{n}\right)^p \int_{\mathbb{R}} x^2 dF(x) = o\left(\frac{1}{n^p}\right)$$

and

$$\mathbb{E}\left(\left(\frac{4}{n}\right)^{p-1/2} \frac{|Y|\mathbf{1}_{\mathcal{O}_1}}{Z}\right) \mathbb{E}\left(|X_1|\mathbf{1}_{\{\mathcal{O}_2^c\}}\right) \leq \left(\frac{4}{n}\right)^p \mathbb{E}\left(\frac{|Y|}{\sqrt{n}}\right) \sqrt{n}\mathbb{E}(|X_1|\mathbf{1}_{\{\mathcal{O}_2^c\}}) = o\left(\frac{1}{n^p}\right)$$

in view of (6.6). Therefore, we see that

$$\lim_{n\to\infty} n^p \left|\mathbb{E}\frac{X_1Y}{(X_1^2+Z^2)^p}\mathbf{1}\{X_1^2+Z^2 \geq n/4\}\right| = 0,$$

concluding the proof. □

**Lemma 6.4.** *Let $p \geq 1$, assume that $\mathbb{E}|X-\mu|^{1+\delta} < \infty$ for some $\delta \geq 1$. Consider the event $\widetilde{\mathcal{O}} = \{|Q_1| \leq 1/2\} \cap \{V_1 \geq \sigma/2\}$. Then*

$$\sigma^{p-1}\left|\mathbb{E}\left(\frac{\bar{\mu}_1-\mu}{\widehat{\sigma}_1^p}\mathbf{1}\{\widetilde{\mathcal{O}}\}\right)\right| \leq \phi(\delta,n) + 2^{p-1}\sqrt{\frac{ke^{-cn}}{N}},$$

*where $c>0$ depends only on $\zeta(X)$, $\phi(\delta,n) = o(n^{-\delta/2})$ for $\delta < 2$ and $\phi(\delta,n) = O(n^{-1})$ otherwise. Moreover, if $Var(X) < \infty$, then*

$$Var\left(\frac{\sqrt{n}\sigma^{p-1}(\bar{\mu}_1-\mu)}{\widehat{\sigma}_1^p}\mathbf{1}\{\widetilde{\mathcal{O}}\}\right) \xrightarrow{n\to\infty} 1.$$

*Proof.* We will prove the two claims separately. Recall the algebraic identity $\widehat{\sigma}_1 = V_1\sqrt{1-Q_1^2}$. To deduce the first inequality, observe that

$$\begin{aligned}\left|\frac{\bar{\mu}_1-\mu}{\widehat{\sigma}_1^p} - \frac{\bar{\mu}_1-\mu}{V_1^p}\right|\mathbf{1}\{\widetilde{\mathcal{O}}\} \\ = \frac{|\bar{\mu}_1-\mu|}{V_1^p}\left|(1-Q_1^2)^{-p/2}-1\right|\mathbf{1}\{\widetilde{\mathcal{O}}\} \leq p4^p\sigma^{1-p}|Q_1|^3\mathbf{1}\{V_1^2 \geq \sigma^2/4\},\end{aligned}$$

where we have used the elementary inequality

$$(1-t^2)^{-p/2}-1 = \left(1+\frac{t^2}{1-t^2}\right)^{p/2}-1 \leq \int_0^{\frac{4}{3}t^2}\frac{p}{2}(1+u)^{p/2-1}du \leq \frac{2^p p}{3^{p/2}}t^2 \quad (6.8)$$

that holds for all $0 \leq t \leq 1/2$. Taking (3.2) into account, we get that

$$\mathbb{E}(|Q_1|^3\mathbf{1}\{V_1^2 \geq \sigma^2/4\}) \leq \frac{C}{n^{3/2}}$$

for an absolute constant $C>0$. Indeed, it directly follows from the inequality

$$\Pr\left(|Q_1|\mathbf{1}\{V_1^2 \geq \sigma^2/4\} \geq \frac{9x}{\sqrt{n}}\right) \leq 4e^{-x^2} \quad (6.9)$$

that is valid for all $x \geq 0$. As a consequence,

$$\left|\mathbb{E}\left(\frac{\bar{\mu}_1-\mu}{\widehat{\sigma}_1^p}\mathbf{1}\{\widetilde{\mathcal{O}}\}\right)\right| \leq \left|\mathbb{E}\left(\frac{\bar{\mu}_1-\mu}{V_1^p}\mathbf{1}\{\widetilde{\mathcal{O}}\}\right)\right| + C\frac{p4^p\sigma^{1-p}}{n^{3/2}}. \quad (6.10)$$

Moreover, we have that

$$\begin{aligned}\left|\mathbb{E}\left(\frac{\bar{\mu}_1-\mu}{V_1^p}\mathbf{1}\{\widetilde{\mathcal{O}}\}\right)\right| &\leq \left|\mathbb{E}\left(\frac{\bar{\mu}_1-\mu}{V_1^p}\mathbf{1}\{V_1^2 \geq \sigma^2/4\}\right)\right| \\ &\quad + 2^{p-1}\sigma^{1-p}\mathbb{E}(|Q_1|\mathbf{1}\{|Q_1| \geq 1/2\} \cap \{V_1^2 \geq \sigma^2/4\}) \\ &\leq \left|\mathbb{E}\left(\frac{\bar{\mu}_1-\mu}{V_1^p}\mathbf{1}\{V_1^2 \geq \sigma^2/4\}\right)\right|\end{aligned}$$

$$+ 2^{p-1}\sigma^{1-p}\sqrt{\mathbb{E}(Q_1^2\mathbf{1}\{V_1^2 \geq \sigma^2/4\})\Pr(\{|Q_1| \geq 1/2\} \cap \{V_1^2 \geq \sigma^2/4\})}$$
$$\leq \left|\mathbb{E}\left(\frac{\bar{\mu}_1 - \mu}{V_1^p}\mathbf{1}\{V_1^2 \geq \sigma^2/4\}\right)\right| + 2^{p-1}\sigma^{1-p}\sqrt{\frac{ke^{-cn}}{N}}, \quad (6.11)$$

where we have used (6.9) in the last inequality. We conclude using Lemma 6.2 as long as (6.10) and (6.11) that

$$\left|\mathbb{E}\left(\frac{\bar{\mu}_1 - \mu}{\widehat{\sigma}_1^p}\mathbf{1}\{\widetilde{\mathcal{O}}\}\right)\right| \leq \sigma^{1-p}\left(\phi(\delta, n) + 2^{p-1}\sqrt{\frac{ke^{-cn}}{N}}\right). \quad (6.12)$$

The first claim is a consequence of both (6.10) and (6.12) since $n^{-3/2}$ is always less than $\phi(\delta, n)$.

Next, we establish the second claim of the lemma. Since, due to the first inequality of the lemma, $\sqrt{n}\left|\mathbb{E}\frac{\sigma^{p-1}(\bar{\mu}_1-\mu)}{\widehat{\sigma}_1^p}\mathbf{1}\{\widetilde{\mathcal{O}}\}\right|$ vanishes as $n$ goes to infinity, it is enough to prove that the second moment converges to 1. We follow the same steps as in the first part to deduce that

$$n\sigma^{2p-2}\mathbb{E}\left|\left(\frac{\bar{\mu}_1 - \mu}{\widehat{\sigma}_1^p}\right)^2 - \left(\frac{\bar{\mu}_1 - \mu}{V_1^p}\right)^2\right|\mathbf{1}\{\widetilde{\mathcal{O}}\}$$
$$= n\sigma^{2p-2}\frac{|\bar{\mu}_1 - \mu|^2}{V_1^{2p}}\left|(1 - Q_1^2)^{-p} - 1\right|\mathbf{1}\{\widetilde{\mathcal{O}}\}$$
$$\leq 2p8^p n\mathbb{E}\left(|Q_1|^4\mathbf{1}\{V_1^2 \geq \sigma^2/4\}\right) \leq \frac{pC8^p}{n}, \quad (6.13)$$

where we have used (6.8) in the first inequality and (6.9) in the second one. Moreover, we also have that

$$n\sigma^{2p-2}\mathbb{E}\left|\left(\frac{\bar{\mu}_j - \mu}{V_j^p}\right)^2\mathbf{1}\{\widetilde{\mathcal{O}}\} - \left(\frac{\bar{\mu}_j - \mu}{V_j^p}\right)^2\mathbf{1}\{V_j^2 \geq \sigma^2/4\}\right|$$
$$\leq n4^{p-1}\mathbb{E}(Q_1^2\mathbf{1}\{|Q_1| \geq 1/2\} \cap \{V_1^2 \geq \sigma^2/4\})$$
$$\leq n4^{p-1}\sqrt{\mathbb{E}(Q_1^4\mathbf{1}\{V_1^2 \geq \sigma^2/4\})\Pr(\{|Q_1| \geq 1/2\} \cap \{V_1^2 \geq \sigma^2/4\})}$$
$$\leq 4^{p-1}e^{-c'n}, \quad (6.14)$$

where we again used (6.9). Combining (6.13) and (6.14), we get that

$$\lim_{n\to\infty}\mathrm{Var}\left(\frac{\sqrt{n}\sigma^{p-1}(\bar{\mu}_1 - \mu)}{\widehat{\sigma}_1^p}\mathbf{1}\{\widetilde{\mathcal{O}}\}\right)$$
$$= \lim_{n\to\infty} n\sigma^{2p-2}\mathbb{E}\left(\left(\frac{\bar{\mu}_j - \mu}{V_j^p}\right)^2\mathbf{1}\{V_j^2 \geq \sigma^2/4\}\right).$$

The conclusion follows immediately from Lemma 6.3. □

**Lemma 6.5.** *In the framework of section 1,*

$$\inf_{\widetilde{\mu}} \sup_{P \in \mathcal{P}_{2,\sigma}} \nu^2(\widetilde{\mu}, P) = \sigma^2.$$

*Proof.* Let $\widetilde{\mathcal{P}}$ be the family of normal distributions $\{N(\mu, \sigma^2),\ \mu \in \mathbb{R}\}$. Then we deduce from the almost-everywhere convolution theorem (Theorem 8.9 in [VdV00]) that for any $\widetilde{\mu}$, $\sup_{P \in \mathcal{P}_{2,\sigma}} \nu^2(\widetilde{\mu}, P) \geq \sigma^2$. On the other hand, letting $\widetilde{\mu}$ be the sample mean $\widetilde{\mu} = \frac{1}{N}\sum_{j=1}^{N} X_j$, we obtain the reverse inequality $\inf_{\widetilde{\mu}} \sup_{P \in \mathcal{P}_{2,\sigma}} \nu^2(\widetilde{\mu}, P) \leq \sigma^2$. □

### *6.3. Proof of Lemma 3.2.*

We will first consider the outlier-free case, meaning that $O = 0$. It is easy to see that

$$\left(\frac{1}{k}\sum_{j=1}^{k} \frac{1}{\widehat{\sigma}_j^p}\right)^{-1} \leq 2\,\mathrm{median}\left(\widehat{\sigma}_1^p, \ldots, \widehat{\sigma}_k^p\right).$$

Hence, Bennett's inequality yields that

$$\begin{aligned}\Pr\left(\frac{1}{k}\sum_{j=1}^{k} \frac{1}{\widehat{\sigma}_j^p} \leq \frac{1}{(4\sigma)^p}\right) &\leq \Pr\left(\mathrm{median}\left(\widehat{\sigma}_1, \ldots, \widehat{\sigma}_k\right) \geq 2\sigma\right) \\ &\leq \Pr\left(\sum_{j=1}^{k} (\mathbf{1}\{\widehat{\sigma}_j^2 \geq 4\sigma^2\} - \pi) \geq k/4\right) \leq e^{-ck\left(\log\frac{1}{\pi}+1\right)}, \quad (6.15)\end{aligned}$$

for some absolute constant $c > 0$, where $\pi := \Pr\left(\widehat{\sigma}_1^2 \geq 4\sigma^2\right) \leq \Pr\left(V_1^2 \geq 4\sigma^2\right) \leq \frac{1}{4}$. Alternatively, if $X$ possesses more than 2 moments, we can apply von Bahr-Esseen inequality [vBE+65] to deduce that

$$\pi \leq \frac{\mathbb{E}|X-\mu|^{1+\delta}/\sigma^{1+\delta}}{n^{\frac{\delta-1}{2}}},$$

for any $\delta \geq 1$. It yields that

$$\Pr\left(\frac{1}{k}\sum_{j=1}^{k} \frac{1}{\widehat{\sigma}_j^p} \leq \frac{1}{(4\sigma)^p}\right) \leq e^{-c'k(1+(\delta-1)\log n)}, \quad (6.16)$$

for $c' > 0$ depending only on the ratio $\mathbb{E}|X-\mu|^{1+\delta}/\sigma^{1+\delta}$. When $X$ has sub-Gaussian distribution, we instead use the Hanson-Wright inequality [HW71] and deduce that

$$\pi \leq e^{-cn\frac{\sigma^4}{\|X\|_{\psi_2}^4}}$$

where $c > 0$ is an absolute constant and $\|X\|_{\psi_2}$ is the $\psi_2$ norm of $X$ [6]. In this case, (6.15) yields that

$$\Pr\left(\frac{1}{k}\sum_{j=1}^{k}\frac{1}{\widehat{\sigma}_j^p} \le \frac{1}{(4\sigma)^p}\right) \le e^{-c(P)kn} \le e^{-\widetilde{c}(P)N} \tag{6.17}$$

$c(P) := c_1 \frac{\sigma^4}{\|X\|_{\psi_2}^4}$ for an absolute constant $c_1 > 0$.

Next, we consider the case $O > 0$. Let $\widehat{\sigma}_{(1)}, \dots, \widehat{\sigma}_{(k)}$ be the increasing order statistics corresponding to $\widehat{\sigma}_1, \dots, \widehat{\sigma}_k$. If $O \le Ck$ for $C < 1$, then at least a fraction of data buckets is outlier-free. Let us call the index set of these buckets $J$ so that $\mathrm{Card}(J) \ge \lfloor (1-C)k \rfloor$, whence

$$\frac{1}{k}\sum_{i=1}^{k}\frac{1}{\widehat{\sigma}_i^p} \ge \frac{\lfloor \frac{(1-C)k}{2} \rfloor}{k}\frac{1}{\widehat{\sigma}^p_{(\lfloor (1-C)k/2) \rfloor)}}.$$

Hence, we get that

$$\Pr\left(\frac{1}{k}\sum_{i=1}^{k}\frac{1}{\widehat{\sigma}_i^p} \le \left(\frac{1-C}{4\sigma}\right)^p\right) \le \Pr\left(\widehat{\sigma}^p_{(\lfloor (1-C)k/2) \rfloor)} \ge 2\sigma\right).$$

The final result follows from (6.16) and (6.17) replacing $k$ by $\lfloor (1-C)k \rfloor$.

**Lemma 6.6.** *Let $\widehat{\sigma}_n$ be such that $\widehat{\sigma}_n = V_n\sqrt{1-Q_n^2}$, and let $\widetilde{\mathcal{O}} = \{|Q_n| \le 1/2\} \cap \{V_n^2 \ge \sigma^2/4\}$ using previous notations. Then*

$$\lim_{n\to\infty} \mathbb{E}\left|\frac{\sigma^p}{\widehat{\sigma}_n^p}\mathbf{1}\{\widetilde{\mathcal{O}}\} - 1\right| = 0.$$

*Proof.* We have that $Q_n \le 1/2$ and $\widehat{\sigma}_n^2 \ge 3/4 V_n^2 \ge 3/16\sigma^2$ on $\widetilde{\mathcal{O}}$. Therefore,

$$\mathbb{E}\left|\frac{\sigma^p}{\widehat{\sigma}_n^p}\mathbf{1}\{\widetilde{\mathcal{O}}\} - 1\right| \le \mathbb{E}\left|\frac{\widehat{\sigma}_n^p - \sigma^p}{\widehat{\sigma}_n^p}\right|\mathbf{1}\{\widetilde{\mathcal{O}}\} + \Pr(\widetilde{\mathcal{O}}^c) \le c_p\mathbb{E}\left|\frac{\widehat{\sigma}_n - \sigma}{\widehat{\sigma}_n}\right|\mathbf{1}\{\widetilde{\mathcal{O}}\} + \Pr(\widetilde{\mathcal{O}}^c)$$

where we have used that for $x \ge 3y/16 > 0$,

$$\frac{|x^p - y^p|}{x^p} = \frac{|x-y|}{x}\sum_{i=0}^{p-1}\left(\frac{y}{x}\right)^{p-i} \le \left(\frac{16}{3}\right)^p\frac{|x-y|}{x}.$$

Moreover,

$$\begin{aligned}\mathbb{E}\left|\frac{\sigma^p}{\widehat{\sigma}_n^p}\mathbf{1}\{\widetilde{\mathcal{O}}\} - 1\right| &\le C_p\mathbb{E}\left|\frac{V_n^2 - \sigma^2}{V_n^2 + \sigma^2}\right|\mathbf{1}\{\widetilde{\mathcal{O}}\} + c_p'\mathbb{E}\left(\frac{Q_n^2V_n^2}{V_n^2+\sigma^2}\mathbf{1}\{\widetilde{\mathcal{O}}\}\right) + \Pr(\widetilde{\mathcal{O}}^c)\\ &\le C_p\mathbb{E}\left|\frac{V_n^2 - \sigma^2}{V_n^2 + \sigma^2}\right|\mathbf{1}\{\widetilde{\mathcal{O}}\} + \frac{c_p'}{n} + e^{-cn},\end{aligned}$$

[6] The $\psi_2$ norm of $X$ is defined via $\|X\|_{\psi_2} := \inf\{C > 0 : \mathbb{E}\exp(|X/C|^2) \le 2\}$.

where we employed inequality (6.9) and Lemma 3.1. Observing that the random variable $\left|\frac{V_n^2-\sigma^2}{V_n^2+\sigma^2}\right| \mathbf{1}\{\widetilde{\mathcal{O}}\}$ converges to 0 in probability (in view of the Law of Large Numbers) and is bounded, hence the convergence holds also in $L^1$. This completes the proof. □

### *6.4. Proof of Theorem 3.1.*

Let $p \geq 1$. Denote $\widehat{\mu} := \widehat{\mu}_{N,p}$ and consider the events

$$\mathcal{O}_j := \{|Q_j| \leq 1/2\} \cap \{V_j \geq \sigma/2\}.$$

Set

$$\mathcal{E} := \bigcap_{j=1}^{k} \mathcal{O}_j. \tag{6.18}$$

Using Lemma 3.1 and inequality (3.3), we get that

$$\Pr(\mathcal{E}^c) \leq ke^{-cn}$$

for some constant $c > 0$ depending on the distribution of $X$. Therefore, for all $t > 0$

$$\Pr(|\widehat{\mu}_{N,p} - \mu| \geq t) \leq \Pr(\{|\widehat{\mu}_{N,p} - \mu| \geq t\} \cap \mathcal{E} \cap \mathcal{E}_p) + ke^{-cn} + \Pr(\mathcal{E}_p^c).$$

Recall the definition (3.1) of the t-statistics $T_1, \ldots, T_k$. The following chain of inequalities holds:

$$\begin{aligned}
\Pr(\{|\widehat{\mu}_{N,p} - \mu| \geq t\} \cap \mathcal{E} \cap \mathcal{E}_p) &\leq \Pr\left(\left\{\left|\sum_{j=1}^{k} \frac{T_j}{\widehat{\sigma}_j^{p-1}}\right| \geq t\frac{k}{4^p\sigma^p}\right\} \cap \mathcal{E}\right) \\
&\leq \Pr\left(\left|\sum_{j=1}^{k} \frac{T_j}{\widehat{\sigma}_j^{p-1}} \mathbf{1}\{\mathcal{O}_j\}\right| \geq t\frac{k}{4^p\sigma^p}\right) \\
&\leq \Pr\left(\left|\sum_{j=1}^{k} w_j\right| \geq t\frac{k}{4^p\sigma} - k\left|\mathbb{E}\left(\frac{\sigma^{p-1}T_j}{\widehat{\sigma}_j^{p-1}}\mathbf{1}\{\mathcal{O}_j\}\right)\right|\right)
\end{aligned}$$

where $w_j := \frac{\sigma^{p-1}T_j}{\widehat{\sigma}_j^{p-1}}\mathbf{1}\{\mathcal{O}_j\} - \mathbb{E}\left(\frac{\sigma^{p-1}T_j}{\widehat{\sigma}_j^{p-1}}\mathbf{1}\{\mathcal{O}_j\}\right)$. It is easy to check that $\sqrt{n}w_j$ is a centered sub-Gaussian random variable, since in view of Lemma 6.1 we have that for all $\lambda \in \mathbf{R}$,

$$\mathbb{E}\left(e^{\sqrt{n}\lambda w_j}\right) \leq e^{c_p\lambda^2/2}$$

for some $c_p > 0$ depending only on $p$. Choosing $t$ as

$$t = 4^p\sigma\left|\mathbb{E}\left(\frac{\sigma^{p-1}T_j}{\widehat{\sigma}_j^{p-1}}\mathbf{1}\{\mathcal{O}_j\}\right)\right| + 4^p\sigma\sqrt{\frac{2c_ps}{N}},$$

we get that

$$\Pr(\{|\widehat{\mu}-\mu|\geq t\}\cap\mathcal{E}\cap\mathcal{E}_p)\leq \Pr\left(\left|\sum_{j=1}^k w_j\right|\geq k\sqrt{\frac{2c_p s}{N}}\right)$$
$$\leq \Pr\left(\sum_{j=1}^k\sqrt{\frac{2sN}{c_p}}w_j\geq 2sk\right)+\Pr\left(-\sum_{j=1}^k\sqrt{\frac{2sN}{c_p}}w_j\geq 2sk\right)$$
$$\leq 2\left(\mathbb{E}\left[e^{\sqrt{\frac{2s}{c_p k}}\sqrt{n}w_j}\right]\right)^k e^{-2s}\leq 2e^{-s},$$

where we used Chernoff bound on the last step. Combining the display above with Lemma 6.4, we conclude that for all $s>0$

$$\Pr\left(|\widehat{\mu}-\mu|\geq C_p\sigma\left(\phi(\delta,n)+\sqrt{\frac{s+ke^{-cn}}{N}}\right)\right)\leq 2e^{-s}+ke^{-cn}+\Pr(\mathcal{E}_p^c),$$

for some $C_p>0$ depending only on $p$. When $ke^{-cn}\geq 1$, the previous bound is trivial. It follows that

$$\Pr\left(|\widehat{\mu}-\mu|\geq C_p\sigma\left(\phi(\delta,n)+\sqrt{\frac{s+1}{N}}\right)\right)\leq 2e^{-s}+ke^{-cn}+\Pr(\mathcal{E}_p^c)$$

for all $s>0$.

### 6.5. *Proof of Theorem 3.2.*

The proof follows similar steps as the argument used to establish Theorem 3.1. We will first show that with high probability the proportion of outliers in each bucket of observations is less than $1/2$. Indeed, letting $W_j$ denote the number of ouliers in the subsample indexed by $G_j$, it is straightfoward to see that $\sum_{j=1}^k W_j = O$, and that the random variables $\{W_j,\ j=1,\dots,k\}$ are negatively correlated. Consider the event

$$\mathcal{E}_2=\bigcap_{j=1}^k\{W_j\leq n/2\}.$$

Recall that $W_j=\sum_{i\in G_j}\mathbf{1}_{i\in\mathcal{O}_j}$. Since $\sum_{j=1}^k W_j=O$, the random variables $(\mathbf{1}_{i\in\mathcal{O}_j})_{i\in G_j}$ are 1-negatively correlated for each $j=1,\dots,k$, as a sub-sequence of a 1-negatively correlated sequence of random variables. Applying the Chernoff bound for negatively correlated random variables (see [Doe20, section 1.10.2.2 and Theorem 1.10.23] for the definitions and the required version of the Chernoff bound), we get that as long as $O\leq N/4$,

$$\Pr(\mathcal{E}_2^c)\leq ke^{-cn}.$$

Hence in what follows, we can restrict our attention on the event $\mathcal{E}_2$. We use the superscript $I$ to denote "clean" sample and $C$ ("corrupted") – otherwise. Notice that

$$\bar{\mu}_j - \mu = \frac{W_j}{n}(\bar{\mu}_j^C - \mu) + \left(1 - \frac{W_j}{n}\right)(\bar{\mu}_j^I - \mu) = \frac{W_j}{n}(\bar{\mu}_j^C - \bar{\mu}_j^I) + \bar{\mu}_j^I - \mu$$

where $\bar{\mu}_j^C, \bar{\mu}_j^I$ are, respectively, empirical means of the corrupted and clean part of the sub-sample indexed by $G_j$. We also have that

$$\widehat{\sigma}_j^2 = \frac{W_j}{n}(\widehat{\sigma}_j^C)^2 + \left(1 - \frac{W_j}{n}\right)(\widehat{\sigma}_j^I)^2 + \frac{W_j(n - W_j)}{n^2}(\bar{\mu}_j^C - \bar{\mu}_j^I)^2, \tag{6.19}$$

where $(\widehat{\sigma}_j^C)^2, (\widehat{\sigma}_j^I)^2$ are, respectively, empirical variances of the corrupted and clean sub-samples of $G_j$. Observe that $\widehat{\sigma}_j^2 \geq (\widehat{\sigma}_j^I)^2/2$, and, therefore, as in the previous proof we deduce that the weights $\alpha_j$ given by (2.2) can not be too large even when outliers are present in the sample. Consider the events $\mathcal{O}_j := \{|Q_j^I| \leq 1/2\} \cap \{V_j^I \geq \sigma/2\} \cap \{W_j \leq n/2\}$, and

$$\mathcal{E} := \bigcap_{j=1}^{k} \mathcal{O}_j.$$

Using Lemma 3.1 and inequality (3.3), we get that

$$\Pr(\mathcal{E}^c) \leq ke^{-cn}$$

for some constant $c > 0$ that depends only on the distribution of $X$. In the rest of the proof we assume that the event $\mathcal{E} \cap \mathcal{E}_p$ holds, with $\mathcal{E}_p$ defined in (3.5). On this event, we have that

$$\begin{aligned}
|\widehat{\mu} - \mu|\,\mathbf{1}\{\mathcal{E}\} \leq \left(\frac{4\sigma}{1-C}\right)^p & \underbrace{\left|\frac{1}{k}\sum_{j=1}^{k} \frac{W_j(\bar{\mu}_j^C - \bar{\mu}_j^I)}{n\widehat{\sigma}_j^p}\mathbf{1}\{\mathcal{O}_j\}\right|}_{(A)} \\
& + \left(\frac{4\sigma}{1-C}\right)^p \underbrace{\left|\frac{1}{k}\sum_{j=1}^{k} \frac{\bar{\mu}_j^I - \mu}{\widehat{\sigma}_j^p}\mathbf{1}\{\mathcal{O}_j\} - \mathbb{E}\left(\frac{\bar{\mu}_j^I - \mu}{\widehat{\sigma}_j^p}\mathbf{1}\{\mathcal{O}_j\}\right)\right|}_{(B)} \\
& + \left(\frac{4\sigma}{1-C}\right)^p \underbrace{\frac{1}{k}\sum_{j=1}^{k}\left|\mathbb{E}\left(\frac{\bar{\mu}_j^I - \mu}{\widehat{\sigma}_j^p}\mathbf{1}\{\mathcal{O}_j\}\right)\right|}_{(C)}.
\end{aligned}$$

We will proceed by estimating each of the terms separately.

**Control of (A):** Using (6.19), observe that on $\mathcal{O}_j$ we have

$$\widehat{\sigma}_j^2 \geq \frac{(\widehat{\sigma}_j^I)^2}{2} + \frac{W_j}{2n}(\bar{\mu}_j^C - \bar{\mu}_j^I)^2 \geq C'\left(\sigma^2 + \frac{W_j}{n}(\bar{\mu}_j^C - \bar{\mu}_j^I)^2\right),$$

for some absolute constant $C' > 0$. It comes out that

$$\begin{aligned}\left|\frac{1}{k}\sum_{j=1}^k \frac{W_j(\bar{\mu}_j^C - \bar{\mu}_j^I)}{n\widehat{\sigma}_j^p}\mathbf{1}\{\mathcal{O}_j\}\right| &\leq \frac{1}{k}\sum_{j=1}^k \frac{C_p W_j|\bar{\mu}_j^C - \bar{\mu}_j^I|}{n\sqrt{\sigma^2 + \frac{W_j}{n}(\bar{\mu}_j^C - \bar{\mu}_j^I)^2}^p} \\ &\leq \frac{C_p}{k\sigma^{p-1}}\sum_{j=1}^k \frac{W_j|\bar{\mu}_j^C - \bar{\mu}_j^I|}{n\sqrt{\sigma^2 + \frac{W_j}{n}(\bar{\mu}_j^C - \bar{\mu}_j^I)^2}} \\ &\leq \frac{C_p\alpha(\mathcal{O})^{(1-p)/2}}{k\sigma^{p-1}}\sum_{j=1}^k \sqrt{W_j/n},\end{aligned}$$

where $\alpha(\mathcal{O}) := 1 + \min_{j/W_j\neq 0}\frac{W_j(\bar{\mu}_j^C - \bar{\mu}_j^I)^2}{n\sigma^2}$. Hence it follows from Cauchy-Schwarz inequality that

$$\left|\frac{1}{k}\sum_{j=1}^k \frac{W_j(\bar{\mu}_j^C - \bar{\mu}_j^I)}{n\widehat{\sigma}_j^p}\mathbf{1}\{\mathcal{O}_j\}\right| \leq \frac{2^p\alpha(\mathcal{O})^{(1-p)/2}}{\sigma^{p-1}}\frac{\sqrt{\sum_{j=1}^k W_j}(\sqrt{O}\wedge\sqrt{k})}{k\sqrt{n}}.$$

As a consequence,

$$\left|\frac{1}{k}\sum_{j=1}^k \frac{W_j(\bar{\mu}_j^C - \bar{\mu}_j^I)}{n\widehat{\sigma}_j^p}\mathbf{1}\{\mathcal{O}_j\}\right| \leq \frac{2^p\alpha(\mathcal{O})^{(1-p)/2}}{\sigma^{p-1}}\left(\frac{O}{k\sqrt{n}}\wedge\sqrt{\frac{O}{N}}\right).$$

Observe that the previous statement holds pointwise, and is not probabilistic in nature. It also suggests that the worst scenario occurs whenever all buckets are corrupted.

**Control of (B):** Since $\widehat{\sigma}_j^2 \geq \left(\widehat{\sigma}_j^I\right)^2/2$ under $\mathcal{O}_j$, we have that

$$\Pr\left(\sigma^{p-1}\left|\frac{\bar{\mu}_i^I - \mu}{\widehat{\sigma}_i^p}\right|\mathbf{1}\{\mathcal{O}_j\} \geq x\right) \leq \Pr\left(\sigma^{p-1}\left|\frac{\bar{\mu}_i^I - \mu}{(\widehat{\sigma}_j^I)^p}\right|\mathbf{1}\{\mathcal{O}_j\} \geq 2^{p/2}x\right).$$

Hence we can show, as in Lemma 6.1, that the random variable

$$\sigma^{p-1}\frac{\bar{\mu}_i^I - \mu}{\widehat{\sigma}_i^p}\mathbf{1}\{\mathcal{O}_j\}$$

is sub-Gaussian. Following the same arguments as in Theorem 3.1, this leads to the bound

$$\sigma^{p-1}\left|\frac{1}{k}\sum_{j=1}^k \frac{\bar{\mu}_i^I - \mu}{\widehat{\sigma}_i^p}\mathbf{1}\{\mathcal{O}_j\} - \mathbb{E}\left(\frac{\bar{\mu}_i^I - \mu}{\widehat{\sigma}_i^p}\mathbf{1}\{\mathcal{O}_j\}\right)\right| \leq C_p\sqrt{\frac{s}{N}}$$

that holds with probability at least $1 - 2e^{-s}$ for some $C_p > 0$.

**Control of (C):** As for the "bias term," it is enough to observe that for uncorrupted buckets, $\widehat{\sigma}_j = \widehat{\sigma}_j^I$, and the bias can be upper bounded exactly as in Theorem 3.1. Hence

$$\frac{\sigma^{p-1}}{k}\sum_{j\in I}\left|\mathbb{E}\left(\frac{\bar{\mu}_j^I-\mu}{\widehat{\sigma}_j^p}\mathbf{1}\{\mathcal{O}_j\}\right)\right| \le \phi(\delta,n) + C_p\sqrt{\frac{ke^{-cn}}{N}}.$$

At the same time, for the corrupted part of the bias term, we have on $\mathcal{O}_j$ that

$$\widehat{\sigma}_j^p \ge C_p'\widehat{\sigma}_j\left(\sigma^2 + \frac{W_j}{n}(\bar{\mu}_j^C-\bar{\mu}_j^I)^2\right)^{(p-1)/2},$$

for some $C_p' > 0$ depending only on $p$. Hence

$$\begin{aligned}
&\frac{\sigma^{p-1}}{k}\left|\sum_{j\in C}\mathbb{E}\left(\frac{\bar{\mu}_j^I-\mu}{\widehat{\sigma}_j^p}\mathbf{1}\{\mathcal{O}_j\}\right)\right| \\
&\quad\le \frac{C_p\sigma^{p-1}}{k}\sum_{j\in C}\mathbb{E}\left|\frac{\bar{\mu}_j^I-\mu}{\widehat{\sigma}_j^I(\sigma^2+W_j/n(\bar{\mu}_j^C-\bar{\mu}_j^I)^2)^{(p-1)/2}}\mathbf{1}\{\mathcal{O}_j\}\right| \\
&\quad\le \frac{C_p\alpha(\mathcal{O})^{(1-p)/2}(O\wedge k)}{k}\mathbb{E}|T_1^I|\mathbf{1}\{\mathcal{O}_1\} \le \frac{C_p\alpha(\mathcal{O})^{(1-p)/2}(O\wedge k)}{k\sqrt{n}} \\
&\quad\le C_p\alpha(\mathcal{O})^{(1-p)/2}\left(\frac{O}{k\sqrt{n}}\wedge\sqrt{\frac{O}{N}}\right),
\end{aligned}$$

for some $C_p > 0$, where we have used inequality (6.9) and the fact that $O\wedge k \le O\wedge\sqrt{Ok}$.

This concludes the proof of the fact that with probability at least $1-2e^{-s}-ke^{-cn}-\Pr(\mathcal{E}_p^c)$,

$$|\widehat{\mu}_{N,p}-\mu| \le \frac{C_p\sigma}{(1-C)^p}\left(\sqrt{\frac{s+1}{N}} + \phi(\delta,n) + \alpha(\mathcal{O})^{(1-p)/2}\left(\frac{\mathcal{O}}{k\sqrt{n}}\wedge\sqrt{\frac{O}{N}}\right)\right). \tag{6.20}$$

### *6.6. Proof of Theorem 3.3.*

Using the definition of $\phi$, it is easy to see that $\sqrt{N_j}\phi(\delta,n_j) = o(1)$, implying that $k_j = o(n_j)$ which in turn implies that $k_j = o\left(e^{cn_j}\right)$ for any constant $c > 0$. We recall that

$$\widehat{\mu}_{N_j,p}-\mu = \frac{\frac{1}{k_j}\sum_{i=1}^{k_j}\frac{T_i}{\widehat{\sigma}_i^{p-1}}}{\frac{1}{k_j}\sum_{i=1}^{k_j}\frac{1}{\widehat{\sigma}_i^p}}.$$

Next, we will use the following decomposition that holds on the event $\mathcal{E}$ (6.18) defined in the proof of Theorem 3.1.

$$\sqrt{N_j}(\widehat{\mu}-\mu)\mathbf{1}\{\mathcal{E}\}$$
$$= H\frac{\sqrt{n_j \mathrm{Var}\left(\frac{T_1}{\widehat{\sigma}_1^{p-1}}\mathbf{1}\{\mathcal{O}_1\}\right)}\frac{\sum_{i=1}^{k_j}\left(\frac{T_i}{\widehat{\sigma}_i^{p-1}}\mathbf{1}\{\mathcal{O}_i\}-\mathbb{E}\frac{T_i}{\widehat{\sigma}_i^{p-1}}\mathbf{1}\{\mathcal{O}_i\}\right)}{\sqrt{\sum_{i=1}^{k_j}\mathrm{Var}\left(\frac{T_i}{\widehat{\sigma}_i^{p-1}}\mathbf{1}\{\mathcal{O}_i\}\right)}} + \sqrt{N_j}\mathbb{E}\frac{T_1}{\widehat{\sigma}_1^{p-1}}\mathbf{1}\{\mathcal{O}_1\}}{\frac{1}{k_j}\sum_{i=1}^{k_j}\left(\frac{1}{\widehat{\sigma}_i^{p}}\mathbf{1}\{\mathcal{O}_i\}-\mathbb{E}\frac{1}{\widehat{\sigma}_i^{p}}\mathbf{1}\{\mathcal{O}_i\}\right)+\mathbb{E}\left(\frac{1}{\widehat{\sigma}_1^{p}}\mathbf{1}\{\mathcal{O}_1\}\right)}$$

where $H=\mathbf{1}\{\mathcal{E}\}$. Using Lemma 6.6, we have that

$$\mathbb{E}\frac{1}{\widehat{\sigma}_1^{p}}\mathbf{1}\{\mathcal{O}_1\} \xrightarrow{j\to\infty} \sigma^{-p}.$$

Moreover using Lemma 6.4 we have

$$\sqrt{N_j}\left|\mathbb{E}\frac{T_1}{\widehat{\sigma}_1^{p-1}}\mathbf{1}\{\mathcal{O}_1\}\right| \le \sqrt{N_j}\phi(\delta,n_j)+\sqrt{k_j e^{-cn_j}} \xrightarrow{j\to\infty} 0,$$

and

$$\sqrt{n_j \mathrm{Var}\left(\frac{T_1}{\widehat{\sigma}_1^{p-1}}\mathbf{1}\{\mathcal{O}_1\}\right)} \xrightarrow{j\to\infty} \sigma^{1-p}.$$

Since the independent variables $\frac{T_i}{\widehat{\sigma}_i^{p-1}}\mathbf{1}\{\mathcal{O}_i\}$ and $\frac{1}{\widehat{\sigma}_i^{p}}\mathbf{1}\{\mathcal{O}_i\}$ are uniformly bounded, they satisfy Lindeberg's condition. Therefore,

$$\frac{\sum_{i=1}^{k_j}\left(\frac{T_i}{\widehat{\sigma}_i^{p-1}}\mathbf{1}\{\mathcal{O}_i\}-\mathbb{E}\frac{T_i}{\widehat{\sigma}_i^{p-1}}\mathbf{1}\{\mathcal{O}_i\}\right)}{\sqrt{\sum_{i=1}^{k_j}\mathrm{Var}\left(\frac{T_i}{\widehat{\sigma}_i^{p-1}}\mathbf{1}\{\mathcal{O}_i\}\right)}} \xrightarrow{j\to\infty} \mathcal{N}(0,1)$$

in distribution, and

$$\frac{1}{k_j}\sum_{i=1}^{k_j}\left(\frac{1}{\widehat{\sigma}_i^{p}}\mathbf{1}\{\mathcal{O}_i\}-\mathbb{E}\left(\frac{1}{\widehat{\sigma}_i^{p}}\mathbf{1}\{\mathcal{O}_i\}\right)\right) \xrightarrow{j\to\infty} 0$$

in probability. In addition, we have that

$$\Pr(\mathcal{E}^c) \le k_j e^{-n_j} \xrightarrow{j\to\infty} 0,$$

established as in the proof of Theorem 3.1. Putting everything together, we finally conclude that

$$\sqrt{N_j}\left(\widehat{\mu}_{N_j,p}-\mu\right) \xrightarrow{d} \mathcal{N}(0,\sigma^2)$$

in distribution as $j\to\infty$.

### *6.7. Proof of Theorem 4.1.*

For any integer $m$, we denote by $\mathcal{E}_p(m)$ the event $\mathcal{E}_p$ defined via (3.5) with $m$ blocks. For every event $\mathcal{A}$, $\mathcal{A}^c$ will denote its complementary. Observe that, as long as $1/20 \leq \frac{\widetilde{\sigma}}{\sigma} \leq 4$, we have the following inclusions

$$\{(\widetilde{k} \vee s) \geq 3(O \vee s)\} \subset \{\widetilde{k} \geq 3(O \vee s)\} \subset \{\widetilde{\mathcal{E}}_p(\lfloor 3(O \vee s)/2 \rfloor)^c\} \subset \{\mathcal{E}_p(\lfloor 3(O \vee s)/2 \rfloor)^c\}$$

where $\lfloor 3(O \vee s)/2 \rfloor$ denotes the integer part of $3(O \vee O)/2$. Therefore, we deduce, using Lemma 3.2, that

$$\Pr\left((\widetilde{k} \vee s) \geq 3(O \vee s)\right) \leq e^{-c_2(O \vee s)}.$$

Finally, we recall that when $\mathcal{E}_p$ holds and $k \leq 3(O \vee s)$, then with probability at least $1 - 2e^{-s} - ke^{-cN/(O \vee s)}$

$$|\widehat{\mu}_{N,p} - \mu| \leq C_p \sigma \left(\sqrt{\frac{s}{N}} + \sqrt{\frac{O}{N}}\right),$$

as shown in (6.20). Combining the previous results, we conclude that

$$\begin{aligned}
&\Pr\left(|\widetilde{\mu}_p(s) - \mu| \geq C_p \sigma \left(\sqrt{\frac{s}{N}} + \sqrt{\frac{O}{N}}\right)\right) \\
&\leq \Pr\left(\left\{|\widetilde{\mu}_p(s) - \mu| \geq C_p \sigma \left(\sqrt{\frac{s}{N}} + \sqrt{\frac{O}{N}}\right)\right\} \cap \{\widetilde{k} \leq 3(O \vee s)\}\right) + e^{-c_2(O \vee s)} \\
&\qquad \leq \sum_{i=1}^{\log_2(3(O \vee s))} (2e^{-s} + 2^i e^{-cN/(O \vee s)}) + e^{-c_2(O \vee s)} \\
&\qquad\qquad \leq 2\log_2(3O)e^{-s} + e^{-c(O \vee s)} + (O \vee s)e^{-cN/(O \vee s)},
\end{aligned}$$

where we used in the last inequality the fact that $\{(\widetilde{k} \vee s) \leq 3(O \vee s)\} \subset \left\{\left(\frac{1}{k}\sum_{j=1}^k \frac{1}{\widehat{\sigma}_j^p}\right)^{-1} \leq \left(\frac{160\sigma}{1-C}\right)^p\right\}$, so that event $\mathcal{E}_p$ holds.

### *6.8. Construction of a robust estimator of $\sigma$.*

Let $N \geq 400$. Without loss of generality, we can assume that $N = 100k$ where $k$ is an integer and that $\{1, \ldots, N\} = \widetilde{G}_1 \cup \cdots \cup \widetilde{G}_k$ where $\widetilde{G}_j = \{100(j-1) + 1, \ldots, 100j\}$ for all $j = 1, \ldots, k$. Let $\widetilde{\sigma}$ be defined as follows:

$$\widetilde{\sigma} := \text{median}\left(\frac{1}{50}\sum_{2i \in \widetilde{G}_1} |X_{2i} - X_{2i-1}|, \ldots, \frac{1}{50}\sum_{2i \in \widetilde{G}_k} |X_{2i} - X_{2i-1}|\right)$$

Under two moments assumption, the following result holds.

**Lemma 6.7.** *Assume that* $\mathbb{E}|X-\mu|^2 < \infty$, $\mathbb{E}|X-\mu| \geq \sigma/2$ *and that* $O \leq N/400$. *Then, with probability at least* $1 - e^{-cN}$, *we have that*

$$1/20 \leq \frac{\widetilde{\sigma}}{\sigma} \leq 4,$$

*where* $c > 0$ *is an absolute constant.*

Note that Lemma 6.7 requires the new condition $\mathbb{E}|X - \mu| \geq \sigma/2$. The latter condition is mild and can be viewed as the equivalence between absolute first and second moments which is less restrictive than the equivalence between centered moments of order 2 and $2 + \delta$. This condition may also be seen as the price to pay for adaptation under only two moments.

*Proof.* Using Jensen's inequality, we get that $\mathbb{E}|X_1 - X_2| \geq \mathbb{E}|X - \mu| \geq \sigma/2$ and $\mathbb{E}|X_1 - X_2| \leq 2\sigma$. Therefore,

$$\Pr\left(1/20 \leq \frac{\widetilde{\sigma}}{\sigma} \leq 4\right) \geq \Pr\left(\left|\frac{\widetilde{\sigma}}{\mathbb{E}|X_1 - X_2|} - 1\right| \leq 9/10\right).$$

Since

$$\Pr\left(\left|\frac{1}{50\mathbb{E}|X_1 - X_2|}\sum_{2i\in\widetilde{G}_1}|X_{2i} - X_{2i-1}| - 1\right| \geq 9/10\right) \leq \frac{200\sigma^2}{4050(\mathbb{E}|X_1 - X_2|)^2} \leq 1/5$$

and that $O \leq k/4$, we conclude that

$$\Pr\left(\left|\frac{\widetilde{\sigma}}{\mathbb{E}|X_1 - X_2|} - 1\right| \geq 9/10\right) \leq \Pr\left(\sum_{j=1}^{k-O} Z_j \geq k/4\right) \leq \Pr\left(\sum_{j=1}^{k-O}(Z_j - \mathbb{E}Z_j) \geq k/20\right) \leq e^{-cN},$$

where $Z_j := \mathbf{1}\left\{\left|\frac{1}{50\mathbb{E}|X_1 - X_2|}\sum_{2i\in\widetilde{G}_1}|X_{2i} - X_{2i-1}| - 1\right| \geq 9/10\right\}$ and $c > 0$. □

## References


[AMS96] 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. MR1427494

[Cat12] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, pages 1148–1185, 2012. MR3052407

[CGR16] Mengjie Chen, Chao Gao, and Zhao Ren. A general decision theory for Huber's $\varepsilon$-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774, 2016. MR3579675

[DLLO16] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016. MR3576558

[DM20] Arnak S Dalalyan and Arshak Minasyan. All-in-one robust estimator of the Gaussian mean. *arXiv preprint arXiv:2002.01432*, 2020.

[Doe20] Benjamin Doerr. Probabilistic tools for the analysis of randomized optimization heuristics. In *Theory of Evolutionary Computation*, pages 1–87. Springer, 2020. MR4299575

[DR20] Cyrus DiCiccio and Joseph P Romano. CLT for U-statistics with growing dimension. Technical report, Standford University, 2020.

[Hop20] Samuel B Hopkins. Mean estimation with sub-Gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193–1213, 2020. MR4102693

[Hub64] P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. MR0161415

[HW71] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971. MR0279864

[KL93] Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993. MR1227763

[LM19a] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019. MR4017683

[LM19b] Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *ArXiv preprint 1907.11391*, 2019. MR4206683

[LO11] Matthieu Lerasle and Roberto I Oliveira. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*, 2011.

[LV20] Jasper CH Lee and Paul Valiant. Optimal sub-Gaussian mean estimation in R. *arXiv preprint arXiv:2011.08384*, page 2, 2020.

[Min18] Stanislav Minsker. Uniform bounds for robust mean estimators. *ArXiv preprint 1812.03523*, 2018.

[Min19] Stanislav Minsker. Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics*, 13(2):5213–5252, 2019. MR4043072

[NY83] A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley and Sons, 1983. MR0702836

[PLS08] Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008. MR2488094

[SCV17] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. *ArXiv*

*preprint 1703.04940*, 2017. MR3761781

[Val85] Leslie G Valiant. Learning disjunction of conjunctions. In *IJCAI*, pages 560–566. Citeseer, 1985.

[vBE+65] Bengt von Bahr, Carl-Gustav Esseen, et al. Inequalities for the $r$-th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303, 1965. MR0170407

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University press, 2000. MR1652247

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University press, 2018. MR3837109