Distributed Statistical Estimation and Rates of Convergence in Normal Approximation

Stanislav Minsker [*,1]

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

Abstract: This paper presents a class of new algorithms for distributed statistical estimation that exploit divide-and-conquer approach. We show that one of the key benefits of the divide-and-conquer strategy is robustness, an important characteristic for large distributed systems. We establish connections between performance of these distributed algorithms and the rates of convergence in normal approximation, and prove non-asymptotic deviations guarantees, as well as limit theorems, for the resulting estimators. Our techniques are illustrated through several examples: in particular, we obtain new results for the median-of-means estimator, and provide performance guarantees for distributed maximum likelihood estimation.

MSC 2010 subject classifications: Primary 6F35; secondary 68W15.
Keywords and phrases: distributed estimation, robust estimation, median-of-means estimator, normal approximation.

Received January 0000.

1. Introduction.

This paper introduces new statistical estimation methods that exhibit *scalability*, a necessary characteristic of modern methods designed to perform statistical analysis of large datasets, as well as *robustness* that guarantees stable performance of distributed systems when some of the nodes exhibit abnormal behavior. The computational power of a single computer is often insufficient to store and process modern data sets, and instead data is stored and analyzed in a distributed way by a cluster consisting of several machines. We consider a distributed estimation framework wherein data is assumed to be randomly assigned to computational nodes that produce intermediate results. We assume that no communication between the nodes is allowed at this first stage. On the second stage, these intermediate results are used to compute some statistic on the whole dataset; see figure 1 for a graphical illustration. Often, such a distributed

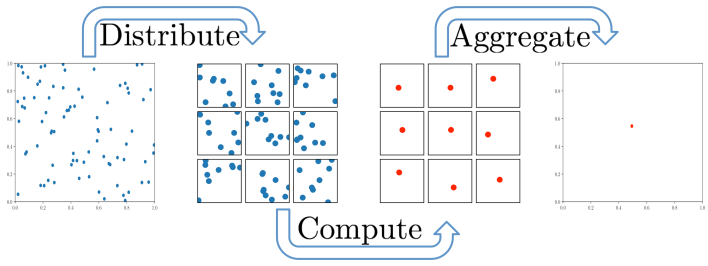


Fig 1: Distributed estimation protocol where data is randomly distributed across nodes to obtain "local" estimates that are aggregated to compute a "global" estimate.

*Supported in part by the National Science Foundation grants DMS-1712956 and CCF-1908905.

setting is unavoidable in applications, whence interactions between subsamples stored on different machines are inevitably lost. Most previous research focused on the following question: how significantly does this loss affect the quality of statistical estimation when compared to an "oracle" that has access to the whole sample? The question that we ask in this paper is different: what can be gained from randomly splitting the data across several subsamples? What are the statistical advantages of the divide-and-conquer framework? Our work indicates that one of the key benefits of an appropriate merging strategy is robustness. In particular, the quality of estimation attained by the distributed estimation algorithm is preserved even if a subset of machines stops working properly. At the same time, the resulting estimators admit tight probabilistic guarantees (expressed in the form of exponential concentration inequalities) even when the distribution of the data has heavy tails – a viable model of real-world samples contaminated by outliers.

We establish connections between a class of randomized divide-and-conquer strategies and the rates of convergence in normal approximation. Using these connections, we provide a new analysis of the "median-of-means" estimator which often yields significant improvements over the previously available results. We further illustrate the implications of our results by constructing novel algorithms for distributed maximum likelihood estimation that admit strong performance guarantees under weak assumptions on the underlying distribution.

1.1. Background and related work.

Let us introduce a simple model for distributed statistical estimation. Assume that $X_1, \dots, X_N$ is a sequence of independent random variables with values in a measurable space $(S, \mathcal{S})$ that represent the data available to a statistician. We will assume that N is large, and that that the sample $\mathcal{X} = (X_1, \dots, X_N)$ is partitioned into k disjoint subsets $G_1, \dots, G_k$ of cardinalities $n_j := \mathrm{card}(G_j)$ respectively, where the partitioning scheme is independent of the data. Let P_j be the distribution of X_j, $j = 1, \dots, N$. The goal is to estimate an unknown parameter $\theta_* = \theta_*(P_j)$, $j = 1, \dots, N$ shared by $P_1, \dots, P_N$ and taking values in a separable Hilbert space $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$; for example, if $S = \mathbb{H}$, θ_* could be the common mean of $X_1, \dots, X_N$. Distributed estimation protocol proceeds via performing "local" computations on each subset G_j, $j \le k$, and the local estimators $\bar{\theta}_j := \bar{\theta}_j(G_j)$, $j \le k$ are then pieced together to produce the final "global" estimator $\hat{\theta}^{(k)} = \hat{\theta}^{(k)}(\bar{\theta}_1, \dots, \bar{\theta}_k)$. We are interested in the statistical properties of such distributed estimation protocols, and our main focus is on the final step that combines the local estimators. Let us mention that the condition requiring the sets G_j, $1 \le j \le k$ to be disjoint can be relaxed; we discuss the extensions related to U-quantiles in section 2.6 below. The problem of distributed and communication - efficient statistical estimation has recently received significant attention from the research community. While our review provides only a subsample of the abundant literature in this field, it is important to acknowledge the works by Mcdonald et al. (2009); Zhang, Wainwright and Duchi (2012); Fan, Han and Liu (2014); Shafieezadeh-Abadeh, Esfahani and Kuhn (2015); Battey et al. (2015); Duchi et al. (2014); Lee et al. (2015); Cheng and Shang (2015); Rosenblatt and Nadler (2016); Zinkevich et al. (2010). Li, Srivastava and Dunson (2016); Scott et al. (2016); Shang and Cheng (2015); Minsker et al. (2014) have investigated closely related problems for distributed Bayesian inference. Applications to important algorithms such as Principal Component Analysis were investigated in (Fan et al., 2017; Liang et al., 2014), among others. Jordan (2013), author provides an overview of recent trends in the intersection of the statistics and computer science communities, describes popular existing strategies such as the "bag of little bootstraps", as wells as successful applications of the divide-and-conquer paradigm to problems such as matrix factorization.

The majority of the aforementioned works propose *averaging* of local estimators as a final merging step. Indeed, averaging reduces variance, hence, if the bias of each local estimator is sufficiently small, their average often attains optimal rates of convergence to the unknown parameter θ_*. For example, when $\theta_*(P) = \mathbb{E}_P X$ is the mean of X and $\bar{\theta}_j$ is the sample mean evaluated over the subsample G_j, $j = 1, \dots, k$, then the average of local estimators $\tilde{\theta} = \frac{1}{k}\sum_{j=1}^k \bar{\theta}_j$ is just a empirical mean evaluated over the whole sample. More generally, it has been shown by Battey et al. (2015); Zhang, Duchi and Wainwright (2013) that in many problems (for instance, linear regression), k can be taken as large as $O(\sqrt{N})$ without negatively affecting the estimation rates; similar guarantees hold for a variety of M-estimators (see Rosenblatt and Nadler, 2016). However, if the number of nodes k itself is large (the case we are mainly interested in), then the averaging scheme has a drawback: if one or more among the local estimators $\bar{\theta}_j$'s is anomalous (for example, due to data corruption or a computer system malfunctioning), then statistical properties of the average will be negatively affected as well. For large distributed systems, this drawback can be costly.

One way to address this issue is to replace averaging by a more robust procedure, such as the median or a robust M-estimator; this approach is investigated in the present work. In the univariate case ($\theta_* \in \mathbb{R}$), the merging strategies we study can be described as solutions of the optimization problem

$$\widehat{\theta}^{(k)} = \operatorname*{argmin}_{z\in\mathbb{R}} \sum_{j=1}^k \rho\left(|\bar{\theta}_j - z|\right) \tag{1}$$

for an appropriately defined convex function ρ; we investigate this class of estimators in detail. A natural extension to the case $\theta_* \in \mathbb{R}^m$ is to consider

$$\widehat{\theta}^{(k)} = \operatorname*{argmin}_{y\in\mathbb{R}^m} \sum_{j=1}^k \rho\left(\left\|\bar{\theta}_j - y\right\|_\circ\right)$$

for some convex function ρ and norm $\|\cdot\|_\circ$. For example, if $\rho(x) = x$, then $\widehat{\theta}^{(k)}$ becomes the spatial, also known as geometric or Haldane's, median (Haldane, 1948; Small, 1990) of $\bar{\theta}_1, \dots, \bar{\theta}_k$. Since the median remains stable as long as at least a half of the nodes in the system perform as expected, such model for distributed estimation is robust. The merging approach based on the various notions of the multivariate median has been previously considered by Minsker (2015) and Hsu and Sabato (2016); here, we analyze the setting when $\rho(x) = x$ and $\|\cdot\|_\circ$ is the L_1-norm using a novel approach.

Existing results for the median-based merging strategies have several pitfalls related to the deviation rates, and in most cases known guarantees are suboptimal. In particular, these guarantees suggest that estimators obtained via the median-based approach are very sensitive to the choice of k, the number of partitions. For instance, consider the problem of univariate mean estimation, where $X_1, \dots, X_N$ are i.i.d. copies of $X \in \mathbb{R}$, and $\theta_* = \mathbb{E}X$ is the expectation of X. Assume that $\mathrm{card}(G_j) \geq n := \lfloor N/k \rfloor$ for all j, let $\bar{\theta}_j = \frac{1}{|G_j|}\sum_{i:X_i\in G_j} X_i$ be the empirical mean evaluated over the subsample G_j, and define the "median-of-means" estimator via

$$\widehat{\theta}^{(k)} = \mathrm{med}\left(\bar{\theta}_1, \dots, \bar{\theta}_k\right), \tag{2}$$

where $\mathrm{med}\,(\cdot)$ is the usual univariate median. This estimator has been introduced by Nemirovski and Yudin (1983) in the context of stochastic optimization, and later appeared in (Jerrum, Valiant and Vazirani, 1986) and (Alon, Matias and Szegedy, 1996). If $\mathrm{Var}(X) = \sigma^2 < \infty$, it has been shown (for example, by Lerasle and Oliveira, 2011) that the median-of-means estimator

$\widehat{\theta}^{(k)}$ satisfies

$$\left|\widehat{\theta}^{(k)} - \theta_*\right| \leq 2\sigma\sqrt{6e}\sqrt{\frac{\log(1/\alpha)}{N}} \tag{3}$$

with probability $\geq 1-\alpha$ if $k = \lfloor \log(1/\alpha) \rfloor + 1$. However, this bound does not provide insight at what happens at the confidence levels other than $1-\alpha$. For example, if $k = \lfloor \sqrt{N} \rfloor$, the only conclusion we can make is that $\left|\widehat{\theta}^{(k)} - \theta_*\right| \lesssim N^{-1/4}$ with high probability, which is far from the "parametric" rate $N^{-1/2}$. And if we want the bound to hold with confidence 99% instead of $1 - e^{-\sqrt{N}}$, then, according to (3), we should take $k = \lfloor \log 100 \rfloor + 1 = 5$, in which case the beneficial effect of parallel computation is very limited. The natural questions to ask is the following: is it possible to "decouple" parameter k, the number of subgroups, from α that controls the deviation probability? Is the median-based merging step suboptimal for large values of k (e.g., $k = \lfloor \sqrt{N} \rfloor$), or is the problem related to the suboptimality of existing bounds? We claim that in many situations the latter is the case, and that previously known results can be strengthened: for instance, the statement of Corollary 1 below implies that whenever $\mathbb{E}|X-\theta_*|^3 < \infty$, the median-of-means estimator satisfies

$$|\widehat{\theta}^{(k)} - \theta_*| \leq 3\sigma\left(\frac{\mathbb{E}\,|X-\theta_*|^3}{\sigma^3}\frac{k}{N-k} + \sqrt{\frac{s}{N-k}}\right) \tag{4}$$

with probability $\geq 1 - 4e^{-2s}$, for *all* $s \lesssim k$ simulnateously. [1] Inequality (4) shows that the estimator (2) has "typical" deviations of order $N^{-1/2}$ whenever $k = O(\sqrt{N})$, hence the "statistical cost" of employing a large number of computational nodes is minor. Moreover, we will prove that $\sqrt{N}\left(\widehat{\theta}^{(k)} - \theta_*\right) \xrightarrow{d} N\left(0, \frac{\pi}{2}\sigma^2\right)$ if $k \to \infty$ and $k = o(\sqrt{N})$ as $N \to \infty$. It will also be demonstrated that improved bounds hold in other important scenarios, such as maximum likelihood estimation, even when the subgroups have different sizes and the observations are not identically distributed.

1.2. Organization of the paper.

Section 1.3 describes notation used throughout the paper. Sections 2 and 3 present main results and examples for the cases of univariate and multivariate parameter respectively. Outcomes of numerical simulation are discussed in section 4, and proofs of the main results are contained in section 5.

1.3. Notation.

Everywhere below, $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ stand for the L_1, L_2 and L_∞ norms of a vector.

Given a probability measure P, $\mathbb{E}_P(\cdot)$ will stand for the expectation with respect to P, and we will write $\mathbb{E}(\cdot)$ when P is clear from the context. Convergence in distribution will be denoted by $\xrightarrow{d}$.

For two sequences $\{a_j\}_{j\geq 1} \subset \mathbb{R}$ and $\{b_j\}_{j\geq 1} \subset \mathbb{R}$ for $j \in \mathbb{N}$, the expression $a_j \lesssim b_j$ means that there exists a constant $c > 0$ such that $a_j \leq c b_j$ for all $j \in \mathbb{N}$. Absolute constants will be

[1] Another known approach (Devroye et al., 2016) is based on a variant Lepski's method; we compare our bounds to the guarantees implied by this method in section 2.4.1.

denoted c, C, c_1, etc., and may take different values in different parts of the paper. For a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, we define

$$\operatorname*{argmin}_{z\in\mathbb{R}^d} f(z) = \{z \in \mathbb{R}^d : f(z) \le f(x) \text{ for all } x \in \mathbb{R}^d\},$$

and $\|f\|_\infty := \operatorname{ess\,sup}\{|f(x)| : x \in \mathbb{R}^d\}$. Finally, $f'_+(x) = \lim_{t\searrow 0} \frac{f(x+t)-f(x)}{t}$ and $f'_-(x) = \lim_{t\nearrow 0} \frac{f(x+t)-f(x)}{t}$ will denote the right and left derivatives of f respectively (whenever these limits exist). Additional notation and auxiliary results are introduced on demand for the proofs in section 5.

1.4. Main results.

As we have argued above, existing guarantees for the estimator (2) are sensitive to the choice of k, the number of partitions. In the following sections, we demonstrate that these bounds are often suboptimal, and show that large values of k often do not have a significant negative effect on the statistical performance of resulting algorithms.

The key observation underlying the subsequent exposition is the following: assume that the "local estimators" $\bar\theta_j$, $1 \le j \le k$, are asymptotically normal with asymptotic mean equal to θ_*. In particular, distributions of $\bar\theta_j$'s are approximately symmetric, with θ_* being the center of symmetry. The location parameters of symmetric distributions admits many robust estimators of the form (1), the sample median being a notable example.

This intuition allows us to establish a parallel between the non-asymptotic deviation guarantees for distributed estimation procedures of the form (1) and the degree of symmetry of "local" estimators quantified by the rates of convergence to normal approximation. Results for the univariate case are presented in section 2, and extensions to the multivariate case are presented in section 3.

2. The univariate case.

We assume that $X_1, \dots, X_N$ is a collection of independent (but not necessarily identically distributed) S-valued random variables with distributions $P_1, \dots, P_N$ respectively. The data are partitioned into disjoint groups $G_1, \dots, G_k$ of cardinality $n_j := \operatorname{card}(G_j)$ each, and such that $\sum_{j=1}^k n_j = N$. Let $\bar\theta_j := \bar\theta_j(G_j)$, $1 \le j \le k$ be a sequence of independent estimators of the parameter $\theta_* \in \mathbb{R}$ shared by $P_1, \dots, P_N$. Our main assumption will be that $\bar\theta_1, \dots, \bar\theta_k$ are asymptotically normal as quantified by the following condition.

Assumption 1. *Let $\Phi(t)$ be the cumulative distribution function of the standard normal random variable $Z \sim N(0,1)$. For each $j = 1, \dots, k$,*

$$g_j(n_j) := \sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\bar\theta_j - \theta_*}{\sqrt{Var\left(\bar\theta_j\right)}} \le t \right) - \Phi(t) \right| \to 0 \text{ as } n_j \to \infty.$$

If what follows, we will set $\sigma_{n_j}^{(j)} := \sqrt{\operatorname{Var}\left(\bar\theta_j\right)}$. Furthermore, let

$$H_k := \left(\frac{1}{k} \sum_{j=1}^k \frac{1}{\sigma_{n_j}^{(j)}} \right)^{-1}$$

be the *harmonic mean* of $\sigma_{n_j}^{(j)}$'s, and set $\alpha_j = \frac{H_k}{\sigma_{n_j}^{(j)}}$. Note that $\alpha_1 = \ldots = \alpha_k = 1$ if $\sigma_{n_1}^{(1)} = \ldots = \sigma_{n_k}^{(k)}$.

2.1. Merging procedure based on the median.

In this subsection, we establish guarantees for the merging procedure based on the sample median, namely,

$$\widehat{\theta}^{(k)} = \operatorname{med}\left(\bar{\theta}_1, \ldots, \bar{\theta}_k\right).$$

This case is treated separately due to its practical importance, the fact that we can obtain better numerical constants, and a conceptually simpler proof.

Theorem 1. *Assume that $s > 0$ and $n_j = \operatorname{card}(G_j)$, $j = 1, \ldots, k$ are such that*

$$\frac{1}{k}\sum_{i=1}^{k}\left(g_i(n_i) + \sqrt{\frac{s}{k}}\right) \cdot \max_{j=1,\ldots,k} \alpha_j < \frac{1}{2}. \tag{5}$$

Moreover, let assumption 1 be satisfied, and let $\zeta_j(n_j, s)$ solve the equation

$$\Phi\left(\zeta_j(n_j,s)/\sigma_{n_j}^{(j)}\right) - \frac{1}{2} = \alpha_j \cdot \frac{1}{k}\sum_{i=1}^{k}\left(g_i(n_i) + \sqrt{\frac{s}{k}}\right).$$

Then for all s satisfying (5),

$$\left|\widehat{\theta}^{(k)} - \theta_*\right| \le \zeta(s) := \max_{j=1,\ldots,k} \zeta_j(n_j, s)$$

with probability at least $1 - 4e^{-2s}$.

Proof. See section 5.2. □

The following lemma yields a more explicit form of the bound and numerical constants.

Lemma 1. *Assume that $\frac{1}{k}\sum_{i=1}^{k}\left(g_i(n_i) + \sqrt{\frac{s}{k}}\right) \cdot \max_{j=1,\ldots,k} \alpha_j \le 0.33$. Then*

$$\zeta(s) \le 3H_k \cdot \frac{1}{k}\sum_{j=1}^{k}\left(g_j(n_j) + \sqrt{\frac{s}{k}}\right).$$

Proof. This is an immediate consequence of Lemma 4 in the Appendix. □

Remark 1. *Let $\bar{\sigma}^{(1)} \le \ldots \le \bar{\sigma}^{(k)}$ be the non-decreasing rearrangement of $\sigma_{n_1}^{(1)}, \ldots, \sigma_{n_k}^{(k)}$. It is easy to see that the harmonic mean H_k of $\sigma_{n_1}^{(1)}, \ldots, \sigma_{n_k}^{(k)}$ satisfies*

$$H_k \le \frac{k}{\lfloor k/m \rfloor} \cdot \frac{1}{\lfloor k/m \rfloor}\sum_{j=1}^{\lfloor k/m \rfloor} \bar{\sigma}^{(j)}$$

for any integer $1 \le m \le k$, hence the deviations of $\widehat{\theta}^{(k)}$ are controlled by the smallest elements among $\left\{\sigma_{n_j}^{(j)}\right\}_{j=1}^{k}$ rather than the largest.

2.2. Example: new bounds for the median-of-means estimator.

The univariate mean estimation problem is pervasive in statistics, and serves as a building block of more advanced methods such as empirical risk minimization. Early works on robust mean estimation include Tukey's "trimmed mean" (Tukey and Harris, 1946), as well as "winsorized mean" (Bickel et al., 1965); also see discussion in (Bubeck, Cesa-Bianchi and Lugosi, 2013). These techniques often produce estimators with significant bias. A different approach based on M-estimation was suggested by O. Catoni (Catoni, 2012); Catoni's estimator yields almost optimal constants, however, its construction requires additional information about the variance or the kurtosis of the underlying distribution; moreover, its computation is not easily parallelizable, therefore this technique cannot be easily employed in the distributed setting.

Here, we will focus on a fruitful idea that is commonly referred to as the "median-of-means" estimator that was formally defined in equation (2) above. Several refinements and extensions of this estimator to higher dimensions have been studied by Minsker (2015); Hsu and Sabato (2013); Devroye et al. (2016); Joly, Lugosi and Oliveira (2016); Lugosi and Mendelson (2017). Advantages of this method include the facts that that it can be implemented in parallel and does not require prior knowledge of any information about parameters of the distribution (e.g., its variance). The following result for the median-of-means estimator is the corollary of Theorem 1; for brevity, we treat only the case of identically distributed observations. Recall that $n = \lfloor N/k \rfloor$ and $\mathrm{card}(G_j) \geq n,\ j = 1, \dots, k$.

Corollary 1. *Let $X_1, \dots, X_N$ be a sequence of i.i.d. copies of a random variable $X \in \mathbb{R}$ such that $\mathbb{E}X = \theta_*$, $Var(X) = \sigma^2$, and $\mathbb{E}|X - \theta_*|^3 < \infty$. Then for all $s > 0$ and k such that $0.4748 \frac{\mathbb{E}|X-\theta_*|^3}{\sigma^3\sqrt{n}} + \sqrt{\frac{s}{k}} \leq 0.33$, the estimator $\widehat{\theta}^{(k)}$ defined in (2) satisfies*

$$|\widehat{\theta}^{(k)} - \theta_*| \leq \sigma \left(1.43 \frac{\mathbb{E}\left|X - \theta_*\right|^3 / \sigma^3}{n} + 3\sqrt{\frac{s}{kn}} \right)$$

with probability at least $1 - 4e^{-2s}$.

Remark 2. *The term $1.43\,\sigma \frac{\mathbb{E}|X-\theta_*|^3/\sigma^3}{n}$ can be thought of as the "bias" due to asymmetry of the distribution of the sample mean. Note that whenever $k \lesssim \sqrt{N}$ (so that $n \gtrsim \sqrt{N}$), the right-hand side of the inequality above is of order $(kn)^{-1/2} \simeq N^{-1/2}$. It is also not hard to see that dependence on k in the term $1.43\,\sigma \frac{\mathbb{E}|X-\theta_*|^3/\sigma^3}{n} \propto \frac{k}{N}$ can not be improved in general. Indeed, assume that X has exponential distribution $\mathrm{E}(1)$. Then the sum $\sum_{j=1}^n X_j$ has Gamma distribution $\Gamma(n, 1)$ with mean equal to n. Moreover, it is known (Choi, 1994) that for large n, the median M of $\Gamma(n, 1)$ satisfies $n - 1/3 < M < 1 - 1/3 + \frac{1}{2n}$, hence one easily checks that the median M_n of the law of $\frac{1}{n}\sum_{j=1}^n X_j - n$ satisfies $|M_n| \geq \frac{1}{4n} \propto \frac{k}{N}$ for n large enough. On the other hand, when $k \to \infty$ while n remains fixed, $\widehat{\theta}^{(k)} \to M_n$ almost surely.*

Proof. It follows from the Berry-Esseen Theorem (Fact 1 in section 5.1) that assumption 1 is satisfied with $\sigma_n^{(1)} = \ldots = \sigma_n^{(k)} = \frac{\sigma}{\sqrt{n}}$, and

$$g_j(n) \leq 0.4748 \frac{\mathbb{E}|X - \theta_*|^3}{\sigma^3\sqrt{n}}$$

for all j. Lemma 1 implies that $\max_j \zeta_j(n, s) \leq 3\frac{\sigma}{\sqrt{n}} \left(0.4748 \frac{\mathbb{E}|X-\theta_*|^3}{\sigma^3\sqrt{n}} + \sqrt{s/k} \right)$, and the claim follows from Theorem 1. □

Results similar to Corollary 1 can be obtained under weaker moment assumptions as well.

Corollary 2. *Let $X_1, \dots, X_N$ be a sequence of i.i.d. copies of a random variable $X \in \mathbb{R}$ such that $\mathbb{E}X = \theta_*$, $Var(X) = \sigma^2$, $\mathbb{E}|X - \theta_*|^{2+\delta} < \infty$ for some $\delta \in (0,1)$. Then there exist absolute constants $c_1, C_2 > 0$ such that for all $s > 0$ and k satisfying $\frac{\mathbb{E}|X-\theta_*|^{2+\delta}}{\sigma^{2+\delta} n^{\delta/2}} + \sqrt{\frac{s}{k}} \le c_1$, the following inequality holds with probability at least $1 - 4e^{-2s}$:*

$$|\widehat{\theta}^{(k)} - \theta_*| \le C_2\, \sigma \left(\frac{\mathbb{E}\,|X - \theta_*|^{2+\delta} / \sigma^{2+\delta}}{n^{\frac{1+\delta}{2}}} + \sqrt{\frac{s}{N}} \right).$$

In this case, typical deviations of $\widehat{\theta}^{(k)}$ are still of order $N^{-1/2}$ as long as $k \lesssim N^{\delta/(1+\delta)}$. The proof of this result again follows from Theorem 1 and a version of the Berry-Esseen inequality stated in section 5.1. Finally, we remark that under stronger assumptions on the distribution of X, the "bias term" can be improved.

Corollary 3. *Let $X_1, \dots, X_N$ be a sequence of i.i.d. copies of a random variable $X \in \mathbb{R}$ such that $\mathbb{E}X = \theta_*$, $Var(X) = \sigma^2$, $\mathbb{E}(X - \theta_*)^3 = 0$ and $\mathbb{E}|X - \theta_*|^{3+\delta} < \infty$ for some $\delta \in (0,1)$. Moreover, assume that the characteristic function $\phi_X(t)$ of X is such that $\limsup_{t\to\infty} |\phi_X(t)| < 1$. Then there exist positive constants c_1^X, C_2^X that depend on the distribution of X such that for all $s > 0$ and k such that $n^{-\frac{1+\delta}{2}} + \sqrt{\frac{s}{k}} \le c_1^X$, the estimator $\widehat{\theta}^{(k)}$ defined in (2) satisfies*

$$|\widehat{\theta}^{(k)} - \theta_*| \le C_2^X\, \sigma \left(\frac{1}{n^{1+\delta/2}} + \sqrt{\frac{s}{N}} \right)$$

with probability at least $1 - 4e^{-2s}$.

Proof. It follows from Theorem 2 in (Ibragimov, 1967) that under the stated assumptions, there exists $C^X > 0$ that depends on the distribution of X such that

$$\sup_{s\in\mathbb{R}} \left| \mathbb{P}\left(\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \le s \right) - \Phi(s) \right| \le \frac{C^X}{n^{\frac{1+\delta}{2}}},$$

hence $g_j(n) \le \frac{C^X}{n^{\frac{1+\delta}{2}}}$ for all j. The claim now follows from Lemma 1 and Theorem 1.

We remark the the requirement $\limsup_{t\to\infty} |\phi_X(t)| < 1$ implies that the distribution of X is not concentrated on a lattice. □

2.3. Example: distributed maximum likelihood estimation.

Let $X_1, \dots, X_N$ be i.i.d. copies of a random vector $X \in \mathbb{R}^d$ with distribution P_{θ_*}, where $\theta_* \in \Theta \subseteq \mathbb{R}$. Assume that for each $\theta \in \Theta$, P_θ is absolutely continuous with respect to a σ-finite measure μ, and let $p_\theta = \frac{dP_\theta}{d\mu}$ be the corresponding density. In this section, we state sufficient conditions for assumption 1 to be satisfied when $\bar{\theta}_1, \dots, \bar{\theta}_k$ are the maximum likelihood estimators (MLE) of θ_*. All derivatives below (denoted by $'$) are taken with respect to θ, unless noted otherwise. Pinelis (2016) proved that the following conditions suffice to guarantee that the rate of convergence of the distribution of the MLE to the normal law is $n^{-1/2}$. Assume that the the log-likelihood function $\ell_x(\theta) = \log p_\theta(x)$ is such that:

(1) $[\theta_* - \delta, \theta_* + \delta] \subseteq \Theta$ for some $\delta > 0$;
(2) "standard regularity conditions" that allow differentiation under the expectation: assume that $\mathbb{E}\ell'_X(\theta_*) = 0$, and that the Fisher information $\mathbb{E}\ell'_X(\theta_*)^2 = -\mathbb{E}\ell''_X(\theta_*) := I(\theta_*)$ is finite;
(3) $\mathbb{E}\,|\ell'_X(\theta_*)|^3 + \mathbb{E}\,|\ell''_X(\theta_*)|^3 < \infty$;

(4) for μ-almost all x, $\ell_x(\theta)$ is three times differentiable for $\theta \in [\theta_* - \delta, \theta_* + \delta]$, and

$$\mathbb{E} \sup_{|\theta-\theta_*|\le\delta} |\ell'''_X(\theta)|^3 < \infty;$$

(5) $\mathbb{P}\big(|\bar\theta_1 - \theta_*| \ge \delta\big) \le c\gamma^n$ for some positive constants c and $\gamma \in [0,1)$.

In turn, condition (5) above is implied by the following two inequalities (see Pinelis, 2016, section 6.2, for detailed discussion and examples):

1. $H^2(\theta,\theta_*) \ge 2 - \frac{2}{(1+c_0(\theta-\theta_*)^2)^\gamma}$, where $H(\theta_1,\theta_2) = \sqrt{\int_{\mathbb{R}^d} \left(\sqrt{p_{\theta_1}} - \sqrt{p_{\theta_2}}\right)^2 d\mu}$ is the Hellinger distance, and c_0, γ are positive constants;
2. $I(\theta) \le c_1 + c_2 |\theta|^\alpha$ for some positive constants c_1, c_2 and α and all $\theta \in \Theta$.

Corollary 4. *Assume that conditions (1)-(5) are satisfied, and that* $\mathrm{card}(G_j) \ge n = \lfloor N/k \rfloor$, $j = 1,\dots,k$. *Then for all* $s > 0$ *such that* $\frac{\mathfrak{C}}{\sqrt{n}} + c\gamma^n + \sqrt{\frac{s}{k}} \le 0.33$,

$$\left|\widehat\theta^{(k)} - \theta_*\right| \le \frac{3}{\sqrt{I(\theta_*)}} \left(\frac{\mathfrak{C}}{n} + \frac{c}{\sqrt{n}}\gamma^n + \sqrt{\frac{s}{kn}}\right)$$

with probability at least $1 - 4e^{-2s}$, *where* $\mathfrak{C}$ *is a positive constant that depends only on* $\{P_\theta\}_{\theta\in[\theta_*-\delta,\theta_*+\delta]}$.

Proof. It follows from results in (Pinelis, 2016), in particular equation (5.5), that whenever conditions (1)-(5) hold, assumption 1 is satisfied for all j with $\sigma_n^{(j)} = (nI(\theta_*))^{-1/2}$, where $I(\theta_*)$ is the Fisher information, and $g_j(n) \le \frac{\mathfrak{C}}{\sqrt{n}} + c\gamma^n$, where $\mathfrak{C}$ is a constant that depends only on $\{P_\theta\}_{\theta\in[\theta_*-\delta,\theta_*+\delta]}$. Lemma 1 implies that

$$\max_{j=1,\dots,k} \zeta_j(n,s) \le 3\left(\frac{\mathfrak{C}}{\sqrt{n}} + c\gamma^n + \sqrt{s/k}\right),$$

and the claim follows from Theorem 1. □

Remark 3. *Results of this section can be extended to include other M-estimators besides MLEs, as Bentkus, Bloznelis and Götze (1997) have shown that M-estimators satisfy a variant of Berry-Esseen bound under rather general conditions.*

2.4. Merging procedures based on robust M-estimators.

In this section, we establish performance guarantees for a distributed algorithms based on the robust M-estimators. Let ρ be a convex, even function such that $\rho(z) \to \infty$ as $|z| \to \infty$ and $\|\rho'_+\|_\infty < \infty$. Moreover, it will be assumed that $\rho'_-(z) \ge z/2$ for $0 < z \le 2$, where ρ'_- is the left derivative of ρ. For instance, $\rho(z) = |z|$ and the Huber's loss

$$\rho_M(z) = \begin{cases} z^2/2, & |z| \le M, \\ M|z| - M^2/2, & |z| > M, \end{cases} \tag{6}$$

where $M \ge 1$, satisfy these assumptions. We study the family of merging procedures based on the M-estimators

$$\widehat\theta_\rho^{(k)} := \operatorname*{argmin}_{z\in\mathbb{R}} \frac{1}{\sqrt{N}} \sum_{j=1}^k \gamma_j \,\rho\left(\tau_j(z - \bar\theta_j)\right),$$

where γ_j, $j = 1,\dots,k$ are the nonnegative weights and τ_j, $j = 1,\dots,k$ are nonnegative "scaling factors." The sample median $\mathrm{med}\left(\bar\theta_1,\dots,\bar\theta_k\right)$ corresponds to the choice of $\rho(x) = |x|$, equal

weights $\gamma_j = 1$ and $\tau_j = 1$, $j = 1, \dots, k$. Results below demonstrate that different choice of weights leads to potentially better bounds. We will also assume that for all $1 \le j \le k$,

$$0 < \liminf_{n\to\infty} \sigma_n^{(j)}\, n^{\beta_j} \le \limsup_{n\to\infty} \sigma_n^{(j)}\, n^{\beta_j} < \infty$$

for some *known* constants $\beta_1, \dots, \beta_k > 0$. In this case, we will set

$$\tau_j := \frac{n^{\beta_j}}{\Delta}, \quad \gamma_j := n_j^{1/2-\beta_j},$$

where $\Delta > 0$. Moreover, let

$$V_j := n_j^{\beta_j}\, \sigma_{n_j}^{(j)}, \quad \bar{\Delta}_j := \max(\Delta, V_j).$$

The following result quantifies non-asymptotic performance of the estimator $\widehat{\theta}_\rho^{(k)}$.

Theorem 2. *Let assumption 1 be satisfied, and suppose that $s > 0$ and $n_1, \dots, n_k$ are such that*

$$\sqrt{2s} + 2\sum_{j=1}^k \sqrt{\frac{n_j}{N}}\, g_j(n_j) \le \frac{\min\big(0.1364, 0.09\rho'_+(2)\big)}{2\,\|\rho'_+\|_\infty} \left(\frac{1}{\sqrt{N}} \sum_{j=1}^k \frac{n_j^{1/2+\beta_j}}{\bar{\Delta}_j}\right) \cdot \min_{j=1,\dots,k} \frac{\bar{\Delta}_j}{n_j^{\beta_j}}. \tag{7}$$

Then for all s satisfying (7),

$$\left|\widehat{\theta}_\rho^{(k)} - \theta_*\right| \le \frac{\|\rho'_+\|_\infty}{\min\big(0.1364, 0.09\rho'_+(2)\big)} \times \left(\frac{1}{\sqrt{N}} \sum_{j=1}^k \frac{n_j^{1/2+\beta_j}}{\bar{\Delta}_j}\right)^{-1} \left(\sqrt{2s} + 2\sum_{j=1}^k \sqrt{\frac{n_j}{N}}\, g_j(n_j)\right) \tag{8}$$

with probability at least $1 - 2e^{-s}$.

Proof. See section 5.3. □

To understand the implications of this technical bound, we consider the special case when the expressions can be simplified significantly. Let $\rho(z) = |z|$, and assume that $\beta_1 = \ldots = \beta_k = 1/2$ and $g_j(n) \le C^X n^{-1/2}$ for all j and some $C^X > 0$ that depends on the distribution of X. Moreover, let

$$\tilde{H}_k := \left(\sum_{j=1}^k \frac{n_j}{N} \frac{1}{V_j}\right)^{-1}$$

be the weighted harmonic mean of $V_1, \dots, V_k$, and set $\tilde{\alpha}_j = \frac{\tilde{H}_k}{V_j}$.

Corollary 5. *There exist positive constants c_1, C_2 such that for all $s > 0$ and $n_1, \dots, n_k$ satisfying*

$$\left(\sqrt{s} + C^X \frac{k}{\sqrt{N}}\right) \max_{j=1,\dots,k} \tilde{\alpha}_j \le c_1 \sqrt{\frac{N}{\max\limits_{j=1,\dots,k} n_j}}, \tag{9}$$

the following inequality holds with probability at least $1 - 2e^{-s}$:

$$\left|\widehat{\theta}_\rho^{(k)} - \theta_*\right| \le C_2\, \tilde{H}_k \Big(C^X \frac{k}{N} + \sqrt{\frac{s}{N}}\Big). \tag{10}$$

Proof. Observe that for $\rho(z) = |z|$, the estimator $\widehat{\theta}_\rho^{(k)}$ does not depend on the choice of Δ, hence $\bar{\Delta}_j = V_j$ for all j. Next, note that

$$\left(\frac{1}{\sqrt{N}}\sum_{j=1}^k \frac{n_j^{1/2+\beta_j}}{V_j}\right)\min_{j=1,\dots,k}\frac{V_j}{n_j^{\beta_j}} = \left(\sum_{j=1}^k \frac{n_j}{N}\frac{1}{V_j}\right)\min_{j=1,\dots,k} V_j\sqrt{\frac{N}{n_j}}$$
$$\geq \min_{j=1,\dots,k} V_j \left(\sum_{j=1}^k \frac{n_j}{N}\frac{1}{V_j}\right)\sqrt{\frac{N}{\max_j n_j}} = \frac{1}{\max_j \tilde{\alpha}_j}\sqrt{\frac{N}{\max_j n_j}},$$

hence (11) implies (7). It is also straightforward to check that (8) implies (10) for an appropriate choice of the constant C_2. □

Let us compare the previous bound with the result of Theorem 1 when the observations are i.i.d. with variance σ^2: Theorem 1 yields that

$$\left|\widehat{\theta}^{(k)} - \theta_*\right| \leq C_3\sigma\frac{1}{\frac{1}{k}\sum_{j=1}^k\sqrt{n_j}}\left(\sqrt{\frac{s}{k}} + \frac{\mathbb{E}|X-\theta_*|^3}{\sigma^3}\frac{1}{k}\sum_{j=1}^k\frac{1}{\sqrt{n_j}}\right),$$

while (10) implies that

$$\left|\widehat{\theta}^{(k)} - \theta_*\right| \leq C_4\sigma\left(\frac{\mathbb{E}|X-\theta_*|^3}{\sigma^3}\frac{k}{N} + \sqrt{\frac{s}{N}}\right)$$

with probability at least $1-2e^{-s}$. By concavity of $x\mapsto\sqrt{x}$, $\frac{1}{\frac{1}{k}\sum_{j=1}^k\sqrt{n_j}} \geq \sqrt{\frac{k}{N}}$, and by the inequality between the harmonic mean and arithmetic mean,

$$\frac{1}{k}\sum_{j=1}^k\frac{1}{\sqrt{n_j}} \geq \sqrt{\frac{k}{N}},$$

hence the second inequality is stronger than the first.

Remark 4. *Assume that ρ is Huber's loss defined in (6) with $M=1$, the data are i.i.d.,and that $|G_j| = n$ for all j. The bound of Theorem 2 implies that one should pick the scaling factor Δ that is not too large, as the quantity $\max(\Delta, V_n)$ controls the estimation error, where $V_n = n^\beta\sigma_n$. On the other hand, it will be shown in section 2.5 that to get an estimator with small asymptotic variance, one should choose Δ that is not too small, and the "optimal" choice is $\Delta = V_n$. While V_n is typically unknown, it can be estimated from the data. Indeed, since $\bar{\theta}_j$'s are approximately normal, their standard deviation can be estimated by the* median absolute deviation *as*

$$\widehat{\sigma}_{n,k} = \frac{1}{\Phi^{-1}(0.75)}med\left(|\bar{\theta}_1 - med\left(\bar{\theta}_1,\dots,\bar{\theta}_k\right)|,\dots,|\bar{\theta}_k - med\left(\bar{\theta}_1,\dots,\bar{\theta}_k\right)|\right),$$

where the factor $1/\Phi^{-1}(0.75)$ is introduced to make the estimator consistent (Hampel et al., 2011). At the same time, when $\rho(x) = |x|$, the estimator is invariant with respect to Δ, but its asymptotic variance is larger than the variance of the estimator based on Huber's loss with optimally chosen scale parameter.

2.4.1. Adversarial contamination.

One of the advantages of allowing the number of subgroups k to be large is improved robustness with respect to adversarial contamination. Assume that the initial sample $X_1, \dots, X_N$ is merged with a set of $\mathcal{O} < N$ outliers that are generated by an adversary who has an opportunity to inspect the data in advance; combined dataset of cardinality $\tilde{N} = N + \mathcal{O}$ is then presented to a statistician. We would like to understand performance of proposed estimators in this framework. To highlight the dependence of the estimation error on the number $\mathcal{O}$ of outliers, we consider only the simplest scenario of i.i.d. data and equal group sizes satisfying $\mathrm{card}(G_j) = n \geq \lfloor \tilde{N}/k \rfloor$. Moreover, suppose that $\beta_1 = \ldots = \beta_k = 1/2$ and that $g(n) \leq C^X n^{-1/2}$. In this case, the estimator we are interested in is defined as

$$\widehat{\theta}_\rho^{(k)} := \operatorname*{argmin}_{z \in \mathbb{R}} \frac{1}{\sqrt{\tilde{N}}} \sum_{j=1}^{k} \rho\left(\sqrt{n}\frac{z - \bar{\theta}_j}{\Delta}\right)$$

where $\Delta > 0$. In what follows, we will also assume that $k > 2\mathcal{O}$. The following result holds.

Theorem 3. *There exist positive constants $c_1(\rho)$, $C_2(\rho)$ such that for all $s > 0$ and $n_1, \dots, n_k$ satisfying*

$$\sqrt{\frac{s}{k}} + \frac{C^X}{\sqrt{n}} + \frac{\mathcal{O}}{k} \leq c_1(\rho), \tag{11}$$

the following inequality holds with probability at least $1 - 2e^{-s}$:

$$\left|\widehat{\theta}_\rho^{(k)} - \theta_*\right| \leq C_2(\rho)\,\widetilde{\Delta}\left(C^X\frac{k}{N} + \sqrt{\frac{s}{N}} + \frac{\mathcal{O}\sqrt{n}}{N}\right),$$

where $\widetilde{\Delta} = \max\left(\Delta, \sqrt{n}\sigma_n^{(1)}\right)$.

Proof. See section 5.4. □

The display above implies that, if $k \geq C \cdot \mathcal{O}$ for a sufficiently large constant C, the error $\left|\widehat{\theta}_\rho^{(k)} - \theta_*\right|$ behaves like the maximum of 2 terms: the first term is the error bound for the case $\mathcal{O} = 0$, and the second term is of order $\widetilde{\Delta}\sqrt{n}\frac{\mathcal{O}}{N}$. Dependence on $\mathcal{O}$ in Theorem 3 can not be improved in general: indeed, if θ_* is the mean and X has 3 finite moments, it is known (Steinhardt, Charikar and Valiant, 2017) that the estimation error can not be of order smaller than $\max\left((\mathcal{O}/N)^{2/3}, N^{-1/2}\right)$. At the same time, if $k \asymp N \cdot \left(\frac{\mathcal{O}}{N}\right)^{2/3}$, the bound of Theorem 3 is exactly of the form $(\mathcal{O}/N)^{2/3} + O\left(N^{-1/2}\right)$.

Remark 5. *An important characteristic of Theorem 3 is that its guarantees still hold uniformly over all $0 < s \lesssim k$, as long as $\mathcal{O}/k$ is not too large. Another method for obtaining bounds that hold uniformly over the wide range of confidence parameters s was suggested by Devroye et al. (2016), and is based on the ability to construct, for each $0 < s \leq c_1 N$, a "sub-Gaussian" confidence interval with coverage probability at least $1 - e^{-s}$. While our bounds rely on stronger moment assumptions to obtain uniformity over a wider range of confidence parameters, they have two important advantages in the context of the median-of-mean estimators: first, they do not require prior information about the variance that is needed to construct confidence intervals. Second, in the framework of adversarial contamination, our bounds are uniform over all $0 < s \lesssim k$, while the guarantees obtained in (Devroye et al., 2016) can only be made uniform over the range $\mathcal{O} \lesssim s \lesssim N$; when $\mathcal{O}$ is relatively large, this difference becomes noticeable.*

2.5. Asymptotic results.

In this section, we complement the previously discussed non-asymptotic deviation bounds for $\widehat{\theta}_\rho^{(k)}$ by the asymptotic results that partially explain the difference that the choice of the function ρ makes. For the benefits of clarity, we make some simplifying assumptions, and the complete list is presented below:

(1) $X_1, \dots, X_N$ are i.i.d., $n = \lfloor N/k \rfloor$ and $\mathrm{card}(G_j) = n,\ j = 1, \dots, k$; result for subgroups of different sizes is presented in Appendix 5.8.
(2) Assumption 1 is satisfied for some function $g(n)$ (note that there is no dependence on index j due to the i.i.d. assumption);
(3) k and n are such that $k \to \infty$ and $\sqrt{k} \cdot g(n) \to 0$ as $N \to \infty$;
(4) ρ is a convex, even function, such that $\rho(z) \to \infty$ as $|z| \to \infty$ and $\|\rho'_+\|_\infty < \infty$;
(5) $\widehat{\theta}_\rho^{(k)}$ is defined as

$$\widehat{\theta}_\rho^{(k)} := \operatorname*{argmin}_{z \in \mathbb{R}} \sum_{j=1}^{k} \rho\left(\frac{z - \bar{\theta}_j}{\sigma_n}\right),$$

where $\sigma_n^{(1)} = \ldots = \sigma_n^{(k)} \equiv \sigma_n$ is the standard deviation of $\bar{\theta}_j$.

For $z \in \mathbb{R}$, define

$$L(z) := \mathbb{E}\rho'(z + Z),$$

where $Z \sim N(0,1)$. Note that, since ρ is differentiable almost everywhere, $L(z) = \mathbb{E}\rho'_-(z+Z) = \mathbb{E}\rho'_+(z+Z)$.

Theorem 4. *Under assumptions (1)-(5) above,*

$$\sqrt{k}\,\frac{\widehat{\theta}_\rho^{(k)} - \theta_*}{\sigma_n} \xrightarrow{d} N(0, \Omega^2),$$

where $\Omega^2 = \frac{\mathbb{E}\left(\rho'(Z)\right)^2}{\left(L'(0)\right)^2}$.

Proof. See section 5.5. □

For example, if $\rho(x) = |x|$, Theorem 4 implies that under appropriate assumptions, the median-of-means estimator $\widehat{\theta}^{(k)}$ defined in (2) satisfies

$$\sqrt{N}\left(\widehat{\theta}^{(k)} - \theta_*\right) \xrightarrow{d} N\left(0, \frac{\pi}{2}\sigma^2\right).$$

Indeed, in this case $\sigma_n = \sigma/\sqrt{n}$, where $\sigma^2 = \mathrm{Var}(X_1)$, and

$$\rho'(x) = \begin{cases} -1, & x < 0, \\ 0, & x = 0, \\ 1, & x > 0, \end{cases}$$

hence a simple calculation yields $\Omega^2 = 1/(L'(0))^2 = \pi/2$.

If we consider the mean estimation problem with Huber's loss $\rho_M(x)$ (6) instead of $\rho(x) = |x|$, we similarly deduce that

$$\rho'(x) = \begin{cases} -M & x \le -M, \\ x, & |x| < M, \\ M, & x \ge M, \end{cases}$$

and we get the well-known (Huber, 1964) expression $\Omega^2 = \frac{\int_{-M}^{M} x^2 d\Phi(x) + 2M^2(1-\Phi(M))}{(2\Phi(M)-1)^2}$; in particular, $\Omega^2 \to 1$ as $M \to \infty$. For instance, $\Omega^2 \simeq 1.15$ for $M = 2$ and $\Omega^2 \simeq 1.01$ for $M = 3$.

Remark 6. *The key assumptions in the list (1)-(5) governing the regime of growth of k and n are (2) and (3). For instance, if the random variables possess finite moments of order $(2+\delta)$ for some $\delta \in (0,1]$, then it follows from the Berry-Esseen bound (Fact 1 in section 5.1) that $\sqrt{k}\, g(n) \to 0$ if $k = o\left(N^{\frac{\delta}{1+\delta}}\right)$ as $N \to \infty$.*

2.6. Connection to U-quantiles.

In this section, we discuss connections of proposed algorithms to U-quantiles and the assumption requiring the groups $G_1, \dots, G_k$ to be disjoint. We assume that the data $X_1, \dots, X_N$ are i.i.d. with common distribution P, and let $\theta_* = \theta_*(P) \in \mathbb{R}$ be a real-valued parameter of interest. It is clear that the estimators produced by distributed algorithms considered above depend on the random partition of the sample. A natural way to avoid such dependence is to consider the U-quantile (in this case, the median)

$$\widetilde{\theta}^{(k)} = \operatorname{med}\left(\bar{\theta}_J,\ J \in \mathcal{A}_N^{(n)}\right),$$

where $\mathcal{A}_N^{(n)} := \{J :\ J \subseteq \{1,\dots,N\}, \operatorname{card}(J) = n := \lfloor N/k \rfloor\}$ is a collection of all distinct subsets of $\{1,\dots,N\}$ of cardinality n, and $\bar{\theta}_J := \bar{\theta}(X_j,\ j \in J)$ is an estimator of θ_* based on $\{X_j,\ j \in J\}$. For instance, when $\operatorname{card}(J) = 2$ and $\bar{\theta}_J = \frac{1}{\operatorname{card}(J)} \sum_{j \in J} X_j$, $\widetilde{\theta}^{(k)}$ is the well-known Hodges-Lehmann estimator of the location parameter, see (Hodges and Lehmann, 1963; Lehmann and D'Abrera, 2006); for a comprehensive study of U-quantiles, see (Arcones, 1996). The main result of this section is an analogue of Theorem 1 for the estimator $\widetilde{\theta}^{(k)}$; it implies that theoretical guarantees for the performance of $\widetilde{\theta}^{(k)}$ are at least as good as for the estimator $\widehat{\theta}^{(k)}$. Since the data are i.i.d., it is enough to impose the assumption 1 on $\bar{\theta}\,(X_1, \dots, X_n)$ only, hence we drop the index j and denote the normalizing sequence $\{\sigma_n\}_{n \in \mathbb{N}}$ and the corresponding error function $g(n)$.

Theorem 5. *Assume that $s > 0$ and $n = \lfloor N/k \rfloor$ are such that*

$$g(n) + \sqrt{\frac{s}{k}} < \frac{1}{2}. \tag{12}$$

Moreover, let assumption 1 be satisfied, and let $\zeta(n,s)$ solve the equation

$$\Phi\left(\zeta(n,s)/\sigma_n\right) = \frac{1}{2} + g(n) + \sqrt{\frac{s}{k}}.$$

Then for any s satisfying (12),

$$\left|\widetilde{\theta}^{(k)} - \theta_*\right| \le \zeta(n,s)$$

with probability at least $1 - 4e^{-2s}$.

Proof. See section 5.6. As before, a more explicit form of the bound immediately follows from Lemma 1. □

A drawback of the estimator $\widetilde{\theta}^{(k)}$ is the fact that its exact computation requires evaluation of $\binom{n}{N}$ estimators $\bar{\theta}_J$ over subsamples $\left\{\{X_j,\ j \in J\},\ J \in \mathcal{A}_N^{(n)}\right\}$. For large N and n, such task

becomes intractable. However, an approximate result can be obtained by choosing ℓ subsets $J_1, \dots, J_\ell$ from $\mathcal{A}_N^{(n)}$ uniformly at random, and setting $\widetilde{\theta}_\ell^{(k)} := \operatorname{med}\left(\bar{\theta}_{J_1}, \dots, \bar{\theta}_{J_\ell}\right)$.
We note that Theorem 2 admits a similar extension for the estimator defined as

$$\widetilde{\theta}_\rho^{(k)} := \operatorname*{argmin}_{z\in\mathbb{R}} \sum_{J\in\mathcal{A}_N^{(n)}} \rho\left(\sqrt{n}\frac{z-\bar{\theta}_J}{\Delta}\right).$$

Namely, if the data are i.i.d., then under the assumptions on ρ made in section 2.4,

$$\left|\widetilde{\theta}_\rho^{(k)} - \theta_*\right| \le C_1(\rho)\max(\sigma,\Delta)\left(\sqrt{\frac{s}{N}} + \frac{g(n)}{\sqrt{n}}\right) \tag{13}$$

with probability at least $1-2e^{-s}$, whenever $s>0$ and $n = \lfloor N/k \rfloor$ are such that

$$\sqrt{\frac{s}{k}} + g(n) \le c_2(\rho)$$

for some positive constants $C_1(\rho),\ c_2(\rho)$. We omit the proof of (13) since the required modifications in the argument of Theorem 2 are exactly the same as those explained in the proof of Theorem 5.

3. Estimation in higher dimensions.

Results presented above admit natural extension to higher dimensions. In this section, it will be assumed that $\theta_* = (\theta_{*,1}, \dots, \theta_{*,m}) \in \mathbb{R}^m,\ m \ge 2$, is a vector-valued parameter of interest. Let $X_1, \dots, X_N$ be i.i.d. random variables that are randomly partitioned into disjoint groups $G_1, \dots, G_k$ with cardinalities $n_1, \dots, n_k$, and let $\bar{\theta}_j := \bar{\theta}_j(G_j) \in \mathbb{R}^m,\ 1 \le j \le k$ be a sequence of estimators of θ_*, the common parameter of the distributions of X_j's. Define

$$g_j^{(m)}(n_j) := \max_{i=1,\dots,m} \sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\bar{\theta}_{j,i} - \theta_{*,i}}{\sqrt{\operatorname{Var}\left(\bar{\theta}_{j,i}\right)}} \le t\right) - \Phi(t)\right|$$

and $V_j^{(i)} := n_j^{1/2}\sqrt{\operatorname{Var}\left(\bar{\theta}_{j,i}\right)}$. Moreover, we will assume that for all $1 \le i \le m$ and $1 \le j \le k$,

$$0 < \liminf_{n_j\to\infty} V_j^{(i)} \le \limsup_{n_j\to\infty} V_j^{(i)} < \infty.$$

We will be interested in the estimator given by weighted L_1 median

$$\widehat{\theta}^{(k)} := \operatorname*{argmin}_{z\in\mathbb{R}^m} \sum_{j=1}^k \sqrt{\frac{n_j}{N}}\left\|z - \bar{\theta}_j\right\|_1.$$

Theorem 6. *There exist absolute constants $c_1, C_2 > 0$ such that for all $s>0$ and all $n_1, \dots, n_k$ satisfying*

$$\sqrt{s} + \sum_{j=1}^k \sqrt{\frac{n_j}{N}} g_j^{(m)}(n_j) \le c_1\sqrt{\frac{N}{\max_{j=1,\dots,k} n_j}}, \tag{14}$$

the following inequality holds with probability at least $1-2e^{-s}$:

$$\left\|\widehat{\theta}^{(k)} - \theta_*\right\|_\infty \le C_2 \max_{i,j} V_j^{(i)}\left(\sqrt{\frac{s+\log m}{N}} + \sum_{j=1}^k \frac{\sqrt{n_j}}{N} g_j^{(m)}(n_j)\right).$$

Proof. See section 5.7. □

Similar results can be established under the more general setting of Theorem 2, albeit at the cost of bulkier statements.

3.1. Example: multivariate median-of-means estimator.

Consider the special case of Theorem 6 when $\theta_* = \mathbb{E}X$ is the mean of $X \in \mathbb{R}^m$, and $\bar\theta_j(X) := \frac{1}{|G_j|}\sum_{X_i \in G_j} X_i$ are the sample means evaluated over the subsamples indexed by $G_1, \dots, G_k$. The problem of finding the mean estimator that admits sub-Gaussian concentration around $\mathbb{E}X$ under weak moment assumptions on the underlying distribution has recently been investigated in several works. For instance, Joly, Lugosi and Oliveira (2016) constructs an estimator that admits "almost optimal" behavior under the assumption that the entries of X possess 4 moments. Recently, Lugosi and Mendelson (2017, 2018) proposed new estimators that attains optimal bounds and requires existence of only 2 moments. More specifically, the aforementioned papers show that, for any s such that $\frac{2}{N} < e^{-s} < 1$, there exists an estimator $\widehat\theta_{(s)}$ such that with probability at least $1 - c\,e^{-s}$,

$$\left\|\widehat\theta_{(s)} - \theta_*\right\|_2 \le C\left(\sqrt{\frac{\mathrm{tr}\,(\Sigma)}{N}} + \sqrt{\frac{s\,\lambda_{\max}(\Sigma)}{N}}\right),$$

where $c, C > 0$ are numerical constants, Σ is the covariance matrix of X, $\mathrm{tr}\,(\Sigma)$ is its trace and $\lambda_{\max}(\Sigma)$ - its largest eigenvalue. However, construction of these estimators explicitly depends on the desired confidence level s, and (more importantly) they are numerically difficult to compute. On the other hand, Theorem 6 demonstrates that performance of the multivariate median-of-means estimator is robust with respect to the choice of the number of subgroups k, and the resulting deviation bounds hold simultaneously over the range of confidence parameter s under mild assumptions, for example when the coordinates of X possess $2+\delta$ moments for some $\delta > 0$. The following corollary summarizes these claims.

Corollary 6. *Let $X_1, \dots, X_N$ be i.i.d. random vectors such that $\theta_* = \mathbb{E}X_1$ is the unknown mean, $\Sigma = \mathbb{E}\left[(X_1 - \theta_*)(X_1 - \theta_*)^T\right]$ is the covariance matrix, $\sigma_i^2 = \Sigma_{i,i}$, and $\max_{i=1,\dots,m}\mathbb{E}|X_{1,i} - \theta_{*,i}|^{2+\delta} < \infty$ for some $\delta \in (0,1]$. Moreover, assume that $n_j \ge n := \lfloor N/k \rfloor$ for all $1 \le j \le k$. Then there exist absolute constants $c_1, C_2 > 0$ such that for all $s > 0$ and k satisfying*

$$\sqrt{\frac{s}{k}} + \max_{i=1,\dots,m}\frac{\mathbb{E}|X_{1,i} - \theta_{*,i}|^{2+\delta}}{\sigma_i^{2+\delta}}\frac{1}{n^{\delta/2}} \le c_1,$$

the following inequality holds with probability at least $1 - 2e^{-s}$:

$$\left\|\widehat\theta^{(k)} - \theta_*\right\|_\infty \le C_2 \max_{i=1,\dots,m}\sigma_i\left(\max_{i=1,\dots,m}\frac{\mathbb{E}|X_{1,i} - \theta_{*,i}|^{2+\delta}}{\sigma_i^{2+\delta}}\frac{1}{n^{\frac{1+\delta}{2}}} + \sqrt{\frac{s+\log m}{N}}\right).$$

Proof. Result follows immediately from Theorem 6 and Fact 1 in section 5.1. □

Remark 7. *Let us compare the bound achieved in Corollary 6 with the deviation guarantees for the sample mean of gaussian random vectors. It follows from the general deviation inequalities for suprema of Gaussian processes (Ledoux and Talagrand, 1991) that if $Z_1, \dots, Z_N$ are i.i.d. copies of $N(\theta_*, \Sigma)$ random vector Z, then the sample mean $\bar Z_N$ satisfies*

$$\left\|\bar Z_N - \theta_*\right\|_\infty \le C\left[\frac{\mathbb{E}\|Z\|_\infty}{\sqrt{N}} + \sup_{\|v\|_1 \le 1}\left(\mathbb{E}\left(\langle Z, v\rangle\right)^2\right)^{1/2}\sqrt{\frac{s}{N}}\right]$$

with probability at least $1-e^{-s}$. *It is easy to check that* $\mathbb{E}\|Z\|_\infty \le C\max_{j=1,\dots,m}\sqrt{\Sigma_{j,j}}\sqrt{\log m}$ *for an absolute constant* $C>0$, *and this bound is tight when* Σ *is an identity matrix. Moreover, as the maximum of a convex function* $v\mapsto\langle\Sigma v,v\rangle$ *over the* ℓ_1 *ball is attained at one of the extreme points,* $\sup_{\|v\|_1\le 1}\left(\mathbb{E}\left(\langle Z,v\rangle\right)^2\right)^{1/2} = \sqrt{\sup_{\|v\|_1\le 1}\langle\Sigma v,v\rangle} = \sqrt{\max\limits_{j=1,\dots,m}\Sigma_{j,j}}$. *Hence, as long as* $k\ll N^{\delta/(1+\delta)}$ *(so that the term* $\max_{i=1,\dots,m}\frac{\mathbb{E}|X_{1,i}-\theta_{*,i}|^{2+\delta}}{\sigma_i^{2+\delta}}\left(\frac{k}{N}\right)^{\frac{1+\delta}{2}}$ *is of order* $o(N^{-1/2})$*), the deviation inequality of Corollary 6 provides sub-Gaussian guarantees in the range* $0<s\lesssim k$.

4. Simulation results.

We illustrate results of the previous sections with numerical simulations that compare performance of the median-of-means estimator with the usual sample mean, see figure 2 below. Moreover, we compared the theoretical guarantees for the median-of-means estimator (described in section 2.2) against the empirical outcomes for the Lomax distribution with shape parameter $\alpha=4$ and scale parameter $\lambda=1$; the corresponding probability density function is

$$p(x)=\frac{\alpha}{\lambda}\left(1+\frac{x}{\lambda}\right)^{-(\alpha+1)}\quad\text{for } x\ge 0$$

In particular, the Lomax distribution with $\alpha=4$ and $\lambda=1$ has mean $1/3$ and median $\sqrt[4]{2}-1\approx 0.1892$. Since the mean and median do not coincide, the error of the median-of-means estimator has a significant bias component for large values of k. Figure 3 depicts the impact of the bias beyond $k=\sqrt{N}$ (equivalently, $\log_N k=1/2$), and also the fact that the median error is mostly flat for $k<\sqrt{N}$.

Finally, we assessed empirical coverage of the confidence intervals constructed using Theorem 4 and centered at the median-of-means estimator; results are presented in figure 4. The sample of size $N=10^5$ was generated from the half-t distribution with 3 degrees of freedom; recall that a random variable ξ has half-t distribution with ν degrees of freedom if $\xi\overset{\mathrm{d}}{=}|\eta|$ where η has usual t-distribution with ν degrees of freedom. It is clear that half-t distribution is both asymmetric and heavy-tailed. Each sample was further corrupted by outliers sampled from the normal distribution with mean 0 and standard deviation 10^5; the number of outliers ranged from 0 to $\sqrt{N}=100$ with increments of 20. The median-of-means estimator was constructed for $k=\sqrt{N}=100$. For comparison, we present empirical coverage levels attained by the sample mean in the same framework.

5. Proofs

In this section, we present the proofs of the main results.

5.1. Preliminaries.

We recall several well-known facts that are used in the proofs below. The following generalization of Berry-Esseen bound (Berry, 1941), (Esseen, 1942) is due to Petrov (1995).

Fact 1 (Berry-Esseen bound). *Assume that* $Y_1,\dots,Y_n$ *is a sequence of i.i.d. copies of a random variable* Y *with mean* μ, *variance* σ^2 *and such that* $\mathbb{E}|Y-\mu|^{2+\delta}<\infty$ *for some* $\delta\in(0,1]$. *Then there exists an absolute constant* $A>0$ *such that*

$$\sup_{s\in\mathbb{R}}\left|\mathbb{P}\left(\sqrt{n}\frac{\bar{Y}-\mu}{\sigma}\le s\right)-\Phi(s)\right|\le A\frac{\mathbb{E}|Y-\mu|^{2+\delta}}{\sigma^{2+\delta}n^{\delta/2}}.$$

Moreover, for $\delta = 1$, $A \le 0.4748$.

The upper bound on A in the case when $\mathbb{E}|X|^3 < \infty$ is due to Shevtsova (2011).

Fact 2 (Bounded difference inequality)**.** *Let* $X_1, \dots, X_k$ *be i.i.d. random variables, and assume that* $Z = g(X_1, \dots, X_k)$*, where* g *is such that for all* $j = 1, \dots, k$ *and all* $x_1, x_2, \dots, x_j, x_j', \dots, x_k$*,*

$$\left|g(x_1, \dots, x_{j-1}, x_j, x_{j+1}, \dots, x_k) - g(x_1, \dots, x_{j-1}, x_j', x_{j+1}, \dots, x_k)\right| \le c_j.$$

Then

$$\mathbb{P}(Z - \mathbb{E}Z \ge t) \le \exp\left\{-\frac{2t^2}{\sum_{j=1}^k c_j^2}\right\}$$

and

$$\mathbb{P}(Z - \mathbb{E}Z \le -t) \le \exp\left\{-\frac{2t^2}{\sum_{j=1}^k c_j^2}\right\}.$$

Finally, we recall the definition of a U-statistic. Let $h : \mathbb{R}^n \mapsto \mathbb{R}$ be a measurable function of n variables, and

$$\mathcal{A}_N^{(n)} := \{J : \ J \subseteq \{1, \dots, N\}, \mathrm{card}(J) = n\}.$$

A U-statistic of order n with kernel h based on the i.i.d. sample $X_1, \dots, X_N$ is defined as (Hoeffding, 1948)

$$U_N(h) = \frac{1}{\binom{n}{N}} \sum_{J \in \mathcal{A}_N^{(n)}} h\left(X_j, \ j \in J\right).$$

Clearly, $\mathbb{E}U_N(h) = \mathbb{E}h(X_1, \dots, X_n)$, moreover, $U_N(h)$ has the smallest variance among all unbiased estimators. The following analogue of fact 2 holds for the U-statistics:

Fact 3 (Concentration inequality for U-statistics, (Hoeffding, 1963))**.**
Assume that the kernel h *satisfies* $|h(x_1, \dots, x_n)| \le M$ *for all* $x_1, \dots, x_n$*. Then for all* $s > 0$*,*

$$\mathbb{P}(|U_N(h) - \mathbb{E}U_N(h)| \ge s) \le 2\exp\left\{-\frac{2\lfloor N/n \rfloor t^2}{M^2}\right\}.$$

5.2. Proof of Theorem 1.

Observe that

$$\left|\widehat{\theta}^{(k)} - \theta_*\right| = \left|\mathrm{med}\left(\bar{\theta}_1 - \theta_*, \dots, \bar{\theta}_k - \theta_*\right)\right|.$$

Let $\Phi^{(n_j, j)}(\cdot)$ be the distribution function of $\bar{\theta}_j - \theta_*$, $j = 1, \dots, k$, and $\widehat{\Phi}_k(\cdot)$ - the empirical distribution function corresponding to the sample $W_1 = \bar{\theta}_1 - \theta_*, \dots, W_k = \bar{\theta}_k - \theta_*$, that is,

$$\widehat{\Phi}_k(z) = \frac{1}{k}\sum_{j=1}^k I\left\{W_j \le z\right\}.$$

Suppose that $z \in \mathbb{R}$ is fixed, and note that $\widehat{\Phi}_k(z)$ is a function of the random variables $W_1, \dots, W_k$, and $\mathbb{E}\widehat{\Phi}_k(z) = \frac{1}{k}\sum_{j=1}^k \Phi^{(n_j, j)}(z)$. Moreover, the hypothesis of the bounded difference inequality (fact 2) is satisfied with $c_j = 1/k$ for $j = 1, \dots, k$, and therefore it implies that

$$\left|\widehat{\Phi}_k(z) - \frac{1}{k}\sum_{j=1}^k \Phi^{(n_j, j)}(z)\right| \le \sqrt{\frac{s}{k}} \tag{15}$$

on the draw of $W_1, \dots, W_k$ with probability $\geq 1 - 2e^{-2s}$.

Let $z_1 \geq z_2$ be such that $\frac{1}{k}\sum_{j=1}^k \Phi^{(n_j,j)}(z_1) \geq \frac{1}{2} + \sqrt{\frac{s}{k}}$ and $\frac{1}{k}\sum_{j=1}^k \Phi^{(n_j,j)}(z_2) \leq \frac{1}{2} - \sqrt{\frac{s}{k}}$. Applying (15) for $z = z_1$ and $z = z_2$ together with the union bound, we see that for $j = 1, 2$,

$$\left| \widehat{\Phi}_k(z_j) - \frac{1}{k}\sum_{j=1}^k \Phi^{(n_j,j)}(z_j) \right| \leq \sqrt{\frac{s}{k}}$$

on an event $\mathcal{E}$ of probability $\geq 1 - 4e^{-2s}$. It follows that on $\mathcal{E}$, $\widehat{\Phi}_k(z_1) \geq 1/2$ and $1 - \widehat{\Phi}_k(z_2) \geq 1/2$ simultaneously, hence

$$\operatorname{med}(W_1, \dots, W_k) \in [z_2, z_1] \tag{16}$$

by the definition of the median. It remains to estimate z_1 and z_2. Assumption 1 implies that

$$\begin{aligned}\frac{1}{k}\sum_{j=1}^k \Phi^{(n_j,j)}(z_1) &\geq \frac{1}{k}\sum_{j=1}^k \Phi\left(\frac{z_1}{\sigma_{n_j}^{(j)}}\right) - \left|\frac{1}{k}\sum_{j=1}^k \left(\Phi^{(n_j,j)}(z_1) - \Phi\left(\frac{z_1}{\sigma_{n_j}^{(j)}}\right)\right)\right| \\ &\geq \frac{1}{k}\sum_{j=1}^k \Phi\left(\frac{z_1}{\sigma_{n_j}^{(j)}}\right) - \frac{1}{k}\sum_{j=1}^k g_j(n_j).\end{aligned}$$

Hence, it suffices to find z_1 such that $\frac{1}{k}\sum_{j=1}^k \Phi\left(\frac{z_1}{\sigma_{n_j}^{(j)}}\right) \geq \frac{1}{2} + \sqrt{\frac{s}{k}} + \frac{1}{k}\sum_{j=1}^k g_j(n_j)$. Recall that $\alpha_j = \frac{1/\sigma_{n_j}^{(j)}}{1/k\sum_{i=1}^k 1/\sigma_{n_j}^{(i)}}$, $j = 1, \dots, k$, and let $\zeta_j(n_j, s)$ solve the equation

$$\Phi\left(\zeta_j(n_j,s)/\sigma_n^{(j)}\right) - \frac{1}{2} = \alpha_j \cdot \frac{1}{k}\sum_{i=1}^k \left(g_i(n_i) + \sqrt{\frac{s}{k}}\right).$$

Note that $\zeta_j(n, s)$ always exists since $\alpha_j \cdot \frac{1}{k}\sum_{i=1}^k \left(g_i(n_i) + \sqrt{\frac{s}{k}}\right) < \frac{1}{2}$ by assumption. Finally, since $\sum_{j=1}^k \alpha_j = k$, it is clear that any

$$z_1 \geq \max_{j=1,\dots,k} \zeta_j(n_j, s)$$

satisfies the requirements. Similarly,

$$\begin{aligned}\frac{1}{k}\sum_{j=1}^k \Phi^{(n_j,j)}(z_2) &\leq \frac{1}{k}\sum_{j=1}^k \Phi\left(\frac{z_2}{\sigma_{n_j}^{(j)}}\right) + \left|\frac{1}{k}\sum_{j=1}^k \left(\Phi^{(n_j,j)}(z_2) - \Phi\left(\frac{z_2}{\sigma_{n_j}^{(j)}}\right)\right)\right| \\ &\leq \frac{1}{k}\sum_{j=1}^k \Phi\left(\frac{z_2}{\sigma_{n_j}^{(j)}}\right) + \frac{1}{k}\sum_{j=1}^k g_j(n_j)\end{aligned}$$

by assumption 1, hence it is sufficient to choose z_2 such that $z_2 \leq \max_{j=1,\dots,k} \tilde{\zeta}_j(n_j, s)$, where $\tilde{\zeta}_j(n_j, s)$ satisfies $\Phi\left(\tilde{\zeta}_j(n_j,s)/\sigma_n^{(j)}\right) - \frac{1}{2} = -\alpha_j \cdot \frac{1}{k}\sum_{i=1}^k \left(g_i(n_i) + \sqrt{\frac{s}{k}}\right)$. Noting that $\tilde{\zeta}_j(n_j, s) = -\zeta_j(n_j, s)$ and recalling (16), we conclude that

$$\left|\widehat{\theta}^{(k)} - \theta_*\right| \leq \max_{j=1,\dots,k} \zeta_j(n_j, s)$$

with probability at least $1 - 4e^{-2s}$.

5.3. *Proof of Theorem 2.*

We will use notation as in the proof of Theorem 1. Let

$$F(z) = \frac{1}{\sqrt{N}} \sum_{j=1}^{k} n_j^{1/2-\beta_j} \, \rho\left(\frac{n_j^{\beta_j}}{\Delta}(z - \bar{\theta}_j) \right).$$

Clearly, $0 \in \partial F\left(\widehat{\theta}_\rho^{(k)}\right)$, where $\partial F(z)$ is the subdifferential of F at point z. In turn, it implies that $F'_+\left(\widehat{\theta}_\rho^{(k)}\right) \geq 0$ and $F'_-\left(\widehat{\theta}_\rho^{(k)}\right) \leq 0$, where

$$F'_+(z) = \frac{1}{\sqrt{N}} \sum_{j=1}^{k} \frac{n_j^{1/2}}{\Delta} \, \rho'_+\left(\frac{n_j^{\beta_j}}{\Delta}(z - \bar{\theta}_j) \right),$$

$$F'_-(z) = \frac{1}{\sqrt{N}} \sum_{j=1}^{k} \frac{n_j^{1/2}}{\Delta} \, \rho'_-\left(\frac{n_j^{\beta_j}}{\Delta}(z - \bar{\theta}_j) \right).$$

Suppose z_1, z_2 are such that $F'_+(z_1) > 0$ and $F'_-(z_2) < 0$. Since $F'_+,\ F'_-$ are increasing, it is easy to see that $\widehat{\theta}_\rho^{(k)} \in (z_2, z_1)$. To find such z_1 and z_2, we proceed in 3 steps; we provide a bound on z_1, and the bound on z_2 follows in a similar manner.
Observe that

$$\begin{aligned} F'_+(z) = {} & \frac{1}{\sqrt{N}} \sum_{j=1}^{k} \frac{n_j^{1/2}}{\Delta} \left(\rho'_+\left(\frac{n_j^{\beta_j}}{\Delta}(z - \bar{\theta}_j) \right) - \mathbb{E}\rho'_+\left(\frac{n_j^{\beta_j}}{\Delta}(z - \bar{\theta}_j) \right) \right) \\ & + \frac{1}{\sqrt{N}} \sum_{j=1}^{k} \frac{n_j^{1/2}}{\Delta} \left(\mathbb{E}\rho'_+\left(\frac{n_j^{\beta_j}}{\Delta}(z - \bar{\theta}_j) \right) - \mathbb{E}\rho'_+\left(\frac{1}{\Delta}\left(n_j^{\beta_j} z - Z_j \right) \right) \right) \\ & + \frac{1}{\sqrt{N}} \sum_{j=1}^{k} \frac{n_j^{1/2}}{\Delta} \mathbb{E}\rho'_+\left(\frac{1}{\Delta}\left(n_j^{\beta_j} z - Z_j \right) \right), \end{aligned}$$

where $Z_j \sim N\left(\theta_*, V_j^2\right),\ j = 1, \ldots, k$.

(a) First, note that the bounded difference inequality (fact 2) implies that for any fixed $z \in \mathbb{R}$,

$$\sum_{j=1}^{k} \sqrt{\frac{n_j}{N}} \left(\rho'_+\left(\frac{n_j^{\beta_j}}{\Delta}(z - \bar{\theta}_j) \right) - \mathbb{E}\rho'_+\left(\frac{n_j^{\beta_j}}{\Delta}(z - \bar{\theta}_j) \right) \right) \geq - \left\| \rho' \right\|_\infty \sqrt{2s}$$

with probability at least $1 - e^{-s}$.

(b) Next, we will find an upper bound for

$$\left| \mathbb{E}\rho'_+\left(\frac{n_j^{\beta_j}}{\Delta}(z - \bar{\theta}_j) \right) - \mathbb{E}\rho'_+\left(\frac{1}{\Delta}\left(n_j^{\beta_j} z - Z_j \right) \right) \right|.$$

Note that for any bounded non-negative function $f : \mathbb{R} \mapsto \mathbb{R}_+$ and a signed measure Q,

$$\left| \int_{\mathbb{R}} f(x) dQ \right| = \left| \int_0^{\|f\|_\infty} Q\left(x :\ f(x) \geq t \right) dt \right| \leq \|f\|_\infty \max_{t \geq 0} \left| Q\left(x :\ f(x) \geq t \right) \right|.$$

Since any bounded function f can be written as $f = \max(f,0) - \max(-f,0)$, we deduce that

$$\left|\int_{\mathbb{R}} f(x)dQ\right| \le \|f\|_\infty \left(\max_{t\ge 0}|Q\left(x:\ f(x)\ge t\right)| + \max_{t\le 0}|Q\left(x:\ f(x)\le t\right)|\right).$$

Moreover, if f is monotone, the sets $\{x:\ f(x)\ge t\}$ and $\{x:\ f(x)\le t\}$ are half-intervals. Applying this to $f(x) = \rho'_+\left(\frac{n_j^{\beta_j}}{\Delta}\left(z - \sigma^{(j)}_{n_j}(x+\theta_*)\right)\right)$ and $Q = \Phi^{(n_j,j)} - \Phi$, we deduce that

$$\begin{aligned}\left|\mathbb{E}\rho'_+\left(\frac{n_j^{\beta_j}}{\Delta}(z-\bar\theta_j)\right) - \mathbb{E}\rho'_+\left(\frac{1}{\Delta}\left(n_j^{\beta_j}z - Z_j\right)\right)\right| &\le 2\|\rho'\|_\infty \sup_{t\in\mathbb{R}}\left|\Phi^{(n_j,j)}(t) - \Phi(t)\right| \\ &= 2\|\rho'\|_\infty\, g_j(n_j).\end{aligned}$$

(c) In remains to find z_1 satisfying

$$\sum_{j=1}^k \sqrt{\frac{n_j}{N}}\mathbb{E}\rho'_+\left(\frac{1}{\Delta}\left(n_j^{\beta_j}z_1 - Z_j\right)\right) > \|\rho'\|_\infty\left(\sqrt{2s} + 2\sum_{j=1}^k\sqrt{\frac{n_j}{N}}\,g_j(n_j)\right).$$

The following bound yields the desired inequality.

Lemma 2. *Let $\varepsilon > 0$ be such that*

$$\varepsilon \le \frac{\min\left(0.1364, 0.09\rho'_+(2)\right)}{2\sqrt{N}}\left(\sum_{j=1}^k \frac{n_j^{1/2+\beta_j}}{\max\left(\Delta, V_j\right)}\right)\min_{j=1,\dots,k}\frac{\max\left(\Delta,V_j\right)}{n_j^{\beta_j}},$$

and set

$$z_1 = \frac{\varepsilon}{\min\left(0.1364, 0.09\rho'_+(2)\right)}\sqrt{N}\left(\sum_{j=1}^k \frac{n_j^{1/2+\beta_j}}{\max\left(\Delta,V_j\right)}\right)^{-1}.$$

Then

$$\sum_{j=1}^k\sqrt{\frac{n_j}{N}}\,\mathbb{E}\rho'_+\left(\frac{n_j^{\beta_j}z_1 - Z_j}{\Delta}\right) > \varepsilon.$$

Proof. The proof is relatively long and is presented in section 5.8. □

Finally, set $\varepsilon := \left\|\rho'_+\right\|_\infty\left(\sqrt{2s} + 2\sum_{j=1}^k\sqrt{\frac{n_j}{N}}\,g_j(n_j)\right)$. If conditions of Lemma 2 are satisfied, the result follows. The estimate for z_2 follows the same pattern, and yields that one can choose $z_2 = -z_1$, implying the claim.

5.4. *Proof of Theorem 3.*

Let $J \subset \{1,\dots,k\}$ of cardinality $|J| \ge k - \mathcal{O}$ be the subset containing all j such that the subsample $\{X_i,\ i\in G_j\}$ does not include any of the $\mathcal{O}$ outliers. Clearly, $\{X_i:\ i\in G_j,\ j\in J\}$ are still i.i.d. as the partitioning scheme is independent of the data. Set $N_J := \sum_{j\in J}|G_j|$, and note that

$$N_J \ge n|J| \ge \frac{kn}{2}.$$

All the probabilities below are evaluated conditionally on N_J. The proof closely follows the steps of the proof of Theorem 2. Let

$$F(z) = \frac{\Delta}{\sqrt{\tilde{N}}} \sum_{j=1}^{k} \rho\left(\sqrt{n}\frac{z-\bar{\theta}_j}{\Delta}\right).$$

As $0 \in \partial F\left(\widehat{\theta}_\rho^{(k)}\right)$, we have that $F'_+\left(\widehat{\theta}_\rho^{(k)}\right) \geq 0$ and $F'_-\left(\widehat{\theta}_\rho^{(k)}\right) \leq 0$. We would like to find z_1, z_2 are such that $F'_+(z_1) > 0$ and $F'_-(z_2) < 0$. Since F'_+, F'_- are increasing, it is easy to see that $\widehat{\theta}_\rho^{(k)} \in (z_2, z_1)$ in this case. Observe that

$$F'_+(z) = \frac{\sqrt{n}}{\sqrt{\tilde{N}}} \sum_{j\in J} \rho'_+\left(\sqrt{n}\frac{z-\bar{\theta}_j}{\Delta}\right) + \frac{\sqrt{n}}{\sqrt{\tilde{N}}} \sum_{j\notin J} \rho'_+\left(\sqrt{n}\frac{z-\bar{\theta}_j}{\Delta}\right)$$

The second sum can be estimated as

$$\left|\frac{\sqrt{n}}{\sqrt{\tilde{N}}} \sum_{j\notin J} \rho'_+\left(\sqrt{n}\frac{z-\bar{\theta}_j}{\Delta}\right)\right| \leq \|\rho'\|_\infty \sum_{j\notin J} \frac{\sqrt{n}}{\sqrt{\tilde{N}}} \leq \|\rho'\|_\infty \frac{\mathcal{O}}{\sqrt{k}},$$

hence, to guarantee that $F'_+(z_1) > 0$, it suffices to find z_1 satisfying

$$\frac{\sqrt{n}}{\sqrt{\tilde{N}}} \sum_{j\in J} \rho'_+\left(\sqrt{n}\frac{z_1-\bar{\theta}_j}{\Delta}\right) > \tilde{\varepsilon} := \|\rho'\|_\infty \frac{\mathcal{O}}{\sqrt{k}}.$$

By the definition of the set J, the sum in the expression

$$\frac{\sqrt{n}}{\sqrt{\tilde{N}}} \sum_{j\in J} \rho'_+\left(\sqrt{n}\frac{z_1-\bar{\theta}_j}{\Delta}\right)$$

is over the subgroups not including the adversarial contamination, hence it can be processed in exactly the same way as in the proof of Theorem 2, and the desired inequality would follow.

5.5. *Proof of Theorem 4.*

Recall that $L(z) = \mathbb{E}\rho'(z+Z)$ for $Z \sim N(0,1)$, and note that under our assumptions, equation $L(z) = 0$ has a unique solution $z = 0$ (even if ρ is not strictly convex). Next, observe that

$$\begin{aligned}\Pr\left(\sum_{j=1}^{k} \rho'_-\left(\frac{\theta_* - \bar{\theta}_j + \frac{t\Omega\sigma_n}{\sqrt{k}}}{\sigma_n}\right) < 0\right) &\leq \Pr\left(\frac{\sqrt{k}}{\sigma_n}\left(\widehat{\theta}_\rho^{(k)} - \theta_*\right) \geq t\Omega\right) \\ &\leq \Pr\left(\sum_{j=1}^{k} \rho'_-\left(\frac{\theta_* - \bar{\theta}_j + \frac{t\Omega\sigma_n}{\sqrt{k}}}{\sigma_n}\right) \leq 0\right),\end{aligned}$$

hence it suffices to show that both the left-hand side and the right-hand side of the inequality above converge to $1-\Phi(t)$ for all t. We will outline the argument for the left-hand side, and the remaining part is proven in a similar fashion. Note that

$$\Pr\left(\sum_{j=1}^{k} \rho'_-\left(\frac{\theta_* - \bar{\theta}_j + \frac{t\Omega\sigma_n}{\sqrt{k}}}{\sigma_n}\right) < 0\right) = \Pr\left(\frac{\sum_{j=1}^{k} Y_{n,j} - \mathbb{E}Y_{n,j}}{\sqrt{k\mathrm{Var}\left(Y_{n,1}\right)}} < -\frac{\sqrt{k}\,\mathbb{E}Y_{n,1}}{\sqrt{\mathrm{Var}\left(Y_{n,1}\right)}}\right), \tag{17}$$

where $Y_{n,j} = \rho'_-\left(\frac{\theta_* - \bar{\theta}_j + \frac{t\Omega\sigma_n}{\sqrt{k}}}{\sigma_n}\right)$.

Lemma 3. *Under the assumptions of Theorem 4,* $\sqrt{k}\mathbb{E}Y_{n,1} \to t\,\Omega\,L'(0)$ *and*

$$\sqrt{Var\,(Y_{n,1})} \;\to\; \sqrt{\mathbb{E}\,(\rho'(Z))^2} = \Omega \cdot L'(0) \text{ as } N \to \infty,$$

where $Z \sim N(0,1)$.

Proof of Lemma 3. Let $Z \sim N(0,1)$. Since ρ is convex, its derivative $\rho' := (\rho'_+ + \rho'_-)/2$ is monotone and continuous almost everywhere (with respect to Lebesgue measure). Together with the assumption that $\|\rho'\|_\infty < \infty$, Lebesgue dominated convergence Theorem implies that

$$\frac{d}{dz}L(z)\big|_{z=0} = \lim_{h\to 0}\frac{1}{h\sqrt{2\pi}}\int_{\mathbb{R}}\rho'(x+h)e^{-x^2/2}dx = \lim_{h\to 0}\frac{1}{h\sqrt{2\pi}}\int_{\mathbb{R}}\rho'(x)e^{-(x-h)^2/2}dx$$
$$= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} x\rho'(x)e^{-x^2/2}dx. \quad (18)$$

Next, we will prove the assertion that $\sqrt{k}\mathbb{E}Y_{n,1} \to t\,\Omega\,L'(0)$. It is easy to see that

$$\sqrt{k}\mathbb{E}Y_{n,1} = \sqrt{k}\left(\mathbb{E}\rho'\left(\frac{\theta_* - \bar\theta_1}{\sigma_n} + \frac{t\Omega}{\sqrt{k}}\right) - \mathbb{E}\rho'\left(Z + \frac{t\Omega}{\sqrt{k}}\right)\right)$$
$$+ t\Omega \cdot \frac{1}{t\Omega/\sqrt{k}}\left(\mathbb{E}\rho'\left(Z + \frac{t\Omega}{\sqrt{k}}\right) - \underbrace{\mathbb{E}\rho'\,(Z)}_{=0}\right).$$

Reasoning as in the proof of Theorem 2 (see step (b) in section 5.3), we deduce that

$$\left|\mathbb{E}\rho'\left(\frac{\theta_* - \bar\theta_1}{\sigma_n} + \frac{t\Omega}{\sqrt{k}}\right) - \mathbb{E}\rho'\left(Z + \frac{t\Omega}{\sqrt{k}}\right)\right| \le 2\,\|\rho'\|_\infty\; g(n),$$

where $g(n)$ is the function from assumption 1. Hence, recalling that $g(n)\sqrt{k} \to 0$ as $N \to \infty$, we obtain that

$$\sqrt{k}\left(\mathbb{E}\rho'\left(\frac{\theta_* - \bar\theta_1}{\sigma_n} + \frac{t\Omega}{\sqrt{k}}\right) - \mathbb{E}\rho'\left(Z + \frac{t\Omega}{\sqrt{k}}\right)\right) \to 0 \text{ as } N \to \infty.$$

On the other hand, it follows from (18) that for $t \neq 0$

$$t\Omega \cdot \frac{1}{t\Omega/\sqrt{k}}\mathbb{E}\rho'\left(Z + \frac{t\Omega}{\sqrt{k}}\right) \xrightarrow{N\to\infty} t\Omega \cdot L'(0).$$

For $t = 0$, it is also clear that $\mathbb{E}\rho'\,(Z) = 0$. To establish the fact that

$$\sqrt{\mathrm{Var}\,(Y_{n,1})} \;\to\; \sqrt{\mathbb{E}\,(\rho'(Z))^2},$$

note that weak convergence of $\frac{\bar\theta_1 - \theta_*}{\sigma_n}$ to the normal law (assumption 1) together with Lebesgue dominated convergence Theorem implies that

$$\mathbb{E}\rho'\left(\frac{\theta_* - \bar\theta_1}{\sigma_n} + \frac{t\Omega}{\sqrt{k}}\right) \to \mathbb{E}\rho'\,(Z) \;= 0,$$
$$\mathbb{E}\left(\rho'\left(\frac{\theta_* - \bar\theta_1}{\sigma_n} + \frac{t\Omega}{\sqrt{k}}\right)\right)^2 \to \mathbb{E}\,(\rho'(Z))^2.$$

Since $L'(0) > 0$, we deduce that

$$\mathbb{E}^{1/2}\,(\rho'(Z))^2 = \Omega \cdot L'(0),$$

and the claim follows. □

Lemma 3 implies that $-\frac{\sqrt{k}\,\mathbb{E}Y_{n,1}}{\sqrt{\mathrm{Var}(Y_{n,1})}} \xrightarrow{N\to\infty} t$. It remains to apply Lindeberg's Central Limit Theorem (Serfling, 1981, Theorem 1.9.3) to $Y_{n,j}$'s to deduce the result from equation (17). To this end, we only need to verify the Lindeberg condition requiring that for any $\varepsilon > 0$,

$$\mathbb{E}(Y_{n,1} - \mathbb{E}Y_{n,1})^2\, I\left\{|Y_{n,1} - \mathbb{E}Y_{n,1}| \geq \varepsilon\sqrt{k}\right\} \to 0 \text{ as } k \to \infty. \tag{19}$$

However, since $\rho'(\cdot)$ (and hence $Y_{n,1}$) is bounded, (19) easily follows.

5.6. Proof of Theorem 5.

The argument is similar to the proof of Theorem 1. Let $\Phi^{(n)}(\cdot)$ be the distribution function of $\frac{\bar\theta_1 - \theta_*}{\sigma_n}$ and $\widehat\Phi_{\binom{N}{n}}(\cdot)$ - the empirical distribution function corresponding to the sample $\left\{W_J = \frac{\bar\theta_J - \theta_*}{\sigma_n},\ J \in \mathcal{A}_N^{(n)}\right\}$ of size $\binom{N}{n}$.

Suppose that $z \in \mathbb{R}$ is fixed, and note that $\widehat\Phi_{\binom{N}{n}}(z)$ is a U-statistic with mean $\Phi^{(n)}(z)$. We will apply the concentration inequality for U-statistics (fact 3) with $M = 1$ to get that

$$\left|\widehat\Phi_{\binom{N}{n}}(z) - \Phi^{(n)}(z)\right| \leq \sqrt{\frac{s}{\lfloor N/n\rfloor}} \leq \sqrt{\frac{s}{k}} \tag{20}$$

with probability $\geq 1 - 2e^{-2s}$; here, we also used the fact that $n = \lfloor N/k\rfloor$.

Let $z_1 \geq z_2$ be such that $\Phi^{(n)}(z_1) \geq \frac12 + \sqrt{\frac{s}{k}}$ and $\Phi^{(n)}(z_2) \leq \frac12 - \sqrt{\frac{s}{k}}$. Applying (20) for $z = z_1$ and $z = z_2$ together with the union bound, we see that for $j = 1, 2$,

$$\left|\widehat\Phi_{\binom{N}{n}}(z_j) - \Phi^{(n)}(z_j)\right| \leq \sqrt{\frac{s}{k}}$$

on an event $\mathcal{E}$ of probability $\geq 1 - 4e^{-2s}$. It follows that on $\mathcal{E}$, $\mathrm{med}\left(W_J, J \in \mathcal{A}_N^{(n)}\right) \in [z_2, z_1]$. The rest of the proof repeats the argument of section 5.2.

5.7. Proof of Theorem 6.

Set $F(z) := \sum_{j=1}^k \sqrt{\frac{n_j}{N}}\left\|z - \bar\theta_j\right\|_1$. Then $\widehat{\theta}^{(k)} = \mathrm{argmin}_{z\in\mathbb{R}^m} F(z)$ by the definition. Since $F(z)$ is convex, the sufficient and necessary condition for $\widehat{\theta}^{(k)}$ to be its minimizer is that $0 \in \partial F(\widehat{\theta}^{(k)})$, the subdifferential of F at point $z = (z_1, \ldots, z_m)$. It is easy to see that

$$\partial F(z) = \left\{u \in \mathbb{R}^m :\ \sum_{j=1}^k \sqrt{\frac{n_j}{N}}\rho'_{-,i}(z_i - \bar\theta_{j,i}) \leq u_i \leq \sum_{j=1}^k \sqrt{\frac{n_j}{N}}\rho'_{+,i}(z_i - \bar\theta_{j,i}),\ i = 1, \ldots, m\right\},$$

where $\rho(x) = |x|$, $\rho'_{+,i}(x) = I\left\{z_i \geq \bar\theta_{j,i}\right\} - I\left\{z_i < \bar\theta_{j,i}\right\}$ and $\rho'_{-,i}(x) = I\left\{z_i > \bar\theta_{j,i}\right\} - I\left\{z_i \leq \bar\theta_{j,i}\right\}$ are the right and left derivative of ρ. Since the subdifferential is convex, it suffices to find points $z_{i,1}, z_{i,2},\ i = 1, \ldots, m$ such that for all i,

$$\sum_{j=1}^k \sqrt{\frac{n_j}{N}}\rho'_{-,i}(z_{i,1} - \bar\theta_{j,i}) > 0, \tag{21}$$

$$\sum_{j=1}^k \sqrt{\frac{n_j}{N}}\rho'_{+,i}(z_{i,2} - \bar\theta_{j,i}) < 0.$$

This task has already been accomplished in the proof of Theorem 2: in particular, the argument presented in section 5.3 yields that, on an event of probability at least $1-2e^{-s}$, inequalities (21) hold for fixed i with

$$z_{i,1} = \theta_{*,i} + C_2 \max_{j=1,\dots,k} V_j^{(i)} \left(\sqrt{\frac{s}{N}} + \sum_{j=1}^{k} \frac{\sqrt{n_j}}{N} g_j^{(m)}(n_j) \right),$$

$$z_{i,2} = \theta_{*,i} - C_2 \max_{j=1,\dots,k} V_j^{(i)} \left(\sqrt{\frac{s}{N}} + \sum_{j=1}^{k} \frac{\sqrt{n_j}}{N} g_j^{(m)}(n_j) \right),$$

assuming that condition (14) is satisfied. The union bound implies that for $i = 1, \dots, m$ simultaneously,

$$\left| \widehat{\theta}_i^{(k)} - \theta_{*,i} \right| \le C_2 \max_{j=1,\dots,k} V_j^{(i)} \left(\sqrt{\frac{s}{N}} + \sum_{j=1}^{k} \frac{\sqrt{n_j}}{N} g_j^{(m)}(n_j) \right) \tag{22}$$

with probability $\ge 1 - 2me^{-s}$. The result follows by taking the maximum over i on both sides of (22).

5.8. *Proof of Lemma 2.*

For any bounded function h such that $h(-x) = -h(x)$ and $h(x) \ge 0$ for $x \ge 0$, and any $z \ge 0$,

$$\int_{\mathbb{R}} h(z+x)\phi_\sigma(x)dx = \int_0^\infty h(x)\left(\phi_\sigma(x+z) - \phi_\sigma(-x+z)\right)dx \ge 0,$$

where $\phi_\sigma(x) = (2\pi\sigma)^{-1/2} e^{-x^2/2\sigma^2}$. Recall that $\rho'_+(x) \ge \frac{x}{2}$ for $0 < x \le 2$, and take

$$h(x) := \rho'_+(x) - \frac{x}{2} I\{|x| < 2\}.$$

Observe that $h(x) \ge 0$ for $x \ge 0$ by assumptions on ρ, hence for any j,

$$\begin{aligned} \mathbb{E}\rho'_+\left(\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta}\right) &= \frac{1}{2}\mathbb{E}\left(\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta} I\left\{\left|\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta}\right| < 2\right\}\right) + \mathbb{E}h\left(\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta}\right) \\ &\ge \max\left(\frac{1}{2}\mathbb{E}\left(\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta} I\left\{\left|\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta}\right| < 2\right\}\right), \mathbb{E}h\left(\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta}\right)\right), \end{aligned} \tag{23}$$

where we used the fact that both terms are nonnegative. Next, we will find lower bounds for each of the terms in the maximum above, starting with the first.

(1) Consider two possibilities: (a) $\Delta < V_j$ and (b) $\Delta \ge V_j$. In the first case, we will use the trivial lower bound $\mathbb{E}\left(\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta} I\left\{\left|\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta}\right| < 2\right\}\right) \ge 0$. The main focus will be on the second case. To this end, note that $Z := \frac{Z_j}{V_j} \sim N(0,1)$, hence

$$\begin{aligned} &\frac{1}{2}\mathbb{E}\left(\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta} I\left\{\left|\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta}\right| < 2\right\}\right) \\ &\qquad = -\frac{V_j}{2\Delta}\mathbb{E}\left(Z\, I\left\{\left|\frac{n_j^{\beta_j} z_1}{V_j} - Z\right| < 2\frac{\Delta}{V_j}\right\}\right) + \frac{n_j^{\beta_j} z_1}{2\Delta}\mathbb{P}\left(\left|\frac{n_j^{\beta_j} z_1}{V_j} - Z\right| < 2\frac{\Delta}{V_j}\right). \end{aligned} \tag{24}$$

Direct computation shows that for any $a \in \mathbb{R},\ t > 0$,

$$\left|\mathbb{E}\left(Z\, I\left\{|a - Z| \le t\right\}\right)\right| = \frac{1}{\sqrt{2\pi}} e^{-\frac{a^2+t^2}{2}} \left|e^{at} - e^{-at}\right|. \tag{25}$$

Take $a = \frac{n_j^{\beta_j} z_1}{V_j}$, $t = 2\frac{\Delta}{V_j}$, and observe that assumptions of the Theorem imply the inequality $|a| \le \frac{t}{4}$. The minimum of the function $a \mapsto a^2 + t^2 - 2|a|t$ over the set $0 \le a \le t/4$ is attained at $a = t/4$, implying that $a^2 + t^2 - 2|a|t \ge \frac{9}{16}t^2 > \frac{t^2}{2}$. Combining this with (25), we deduce that

$$\begin{aligned}\left|\mathbb{E}\left(Z\, I\left\{|a - Z| \le t\right\}\right)\right| &\le \frac{1}{\sqrt{2\pi}} e^{-t^2/4} e^{-|at|} \left|e^{at} - e^{-at}\right| \\ &= \frac{e^{-t^2/4}}{\sqrt{2\pi}} \left(1 - e^{-2|at|}\right) \le \frac{e^{-t^2/4}}{\sqrt{2\pi}} \cdot 2|at|,\end{aligned}$$

hence

$$\left|\frac{V_j}{2\Delta}\mathbb{E}\left(Z\, I\left\{\left|\frac{n_j^{\beta_j} z_1}{V_j} - Z\right| < 2\frac{\Delta}{V_j}\right\}\right)\right| \le \frac{2}{\sqrt{2\pi}} \left|\frac{z_1 n_j^{\beta_j}}{V_j}\right| e^{-\frac{\Delta^2}{V_j^2}} = \frac{2}{\sqrt{2\pi}} \left|\frac{z_1 n_j^{\beta_j}}{\Delta}\right| \frac{\Delta}{V_j} e^{-\frac{\Delta^2}{V_j^2}}.$$

Moreover, since $|z_1| \le \frac{1}{2}\frac{\Delta}{n_j^{\beta_j}}$ by assumptions of the lemma, it follows that

$$\mathbb{P}\left(\left|\frac{n_j^{\beta_j} z_1}{V_j} - Z\right| < 2\frac{\Delta}{V_j}\right) \ge \mathbb{P}\left(|Z| < \frac{3\Delta}{2V_j}\right) \ge 1 - 2\Phi(-3/2) > 0.86.$$

Together with (23), (24), the last display yields that

$$\mathbb{E}\rho'_+\left(\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta}\right) > \left|\frac{0.86}{2}\frac{z_1 n_j^{\beta_j}}{\Delta}\right| - \frac{2}{\sqrt{2\pi}} \left|\frac{z_1 n_j^{\beta_j}}{\Delta}\right| \frac{\Delta}{V_j} e^{-\frac{\Delta^2}{V_j^2}}.$$

As $x \mapsto xe^{-x^2}$ is decreasing for $x \ge 1/\sqrt{2}$, one easily checks that $\frac{\Delta}{V_j} e^{-\frac{\Delta^2}{V_j^2}} \le e^{-1}$ as $\Delta \ge V_j$, hence

$$\mathbb{E}\rho'_+\left(\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta}\right) > \left(0.43 - \frac{2}{e\sqrt{2\pi}}\right)|z_1|\frac{n_j^{\beta_j}}{\Delta} > 0.1364|z_1|\frac{n_j^{\beta_j}}{\Delta}.$$

(2) To estimate the second term, we start with a simple inequality

$$\begin{aligned}\mathbb{E}h\left(\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta}\right) &\ge \mathbb{E}\rho'_+\left(\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta}\right) I\left\{\left|\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta}\right| \ge 2\right\} \\ &\ge \rho'_+(2)\,\mathbb{E}\left(I\left\{\frac{Z_j - n_j^{\beta_j} z_1}{\Delta} \le -2\right\} - I\left\{\frac{Z_j - n_j^{\beta_j} z_1}{\Delta} \ge 2\right\}\right)\end{aligned}$$

which follows from the definition of h and assumptions on ρ. Again, we consider two possibilities: (a) $\Delta < V_j$ and (b) $\Delta \ge V_j$. In case (b), we use the trivial bound (recalling that $z_1 \ge 0$)

$$\mathbb{E}\left(I\left\{\frac{Z_j - n_j^{\beta_j} z_1}{\Delta} \le -2\right\} - I\left\{\frac{Z_j - n_j^{\beta_j} z_1}{\Delta} \ge 2\right\}\right) \ge 0.$$

In the first case, we see that

$$\Pr\left(\frac{Z_j - n_j^{\beta_j} z_1}{\Delta} \le -2\right) - \Pr\left(\frac{Z_j - n_j^{\beta_j} z_1}{\Delta} \ge 2\right) = \Pr\left(Z \in \left[-\frac{n_j^{\beta_j} z_1}{V_j} - 2\frac{\Delta}{V_j}, \frac{n_j^{\beta_j} z_1}{V_j} - 2\frac{\Delta}{V_j}\right]\right).$$

Lemma 5 implies that

$$\Pr\left(Z \in \left[-\frac{n_j^{\beta_j} z_1}{V_j} - 2\frac{\Delta}{V_j}, \frac{n_j^{\beta_j} z_1}{V_j} - 2\frac{\Delta}{V_j}\right]\right) \ge 2e^{-\frac{2\Delta^2}{V_j^2}} \Pr\left(Z \in \left[0, \frac{n_j^{\beta_j} z_1}{V_j}\right]\right) \ge 2e^{-2} \Pr\left(Z \in \left[0, \frac{n_j^{\beta_j} z_1}{V_j}\right]\right),$$

where we used the fact that $\Delta < V_j$ by assumption. Finally, Lemma 4 implies that

$$\Pr\left(Z \in \left[0, \frac{n_j^{\beta_j} z_1}{V_j}\right]\right) > \frac{1}{3}\frac{n_j^{\beta_j} z_1}{V_j}$$

whenever $z_1 \le 0.99 \frac{V_j}{n_j^{\beta_j}}$. In conclusion, we demonstrated that in case (a)

$$\mathbb{E}h\left(\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta}\right) > \frac{2e^{-2}}{3}\rho'_+(2) z_1 \frac{n_j^{\beta_j}}{V_j} > 0.09\, \rho'_+(2) z_1 \frac{n_j^{\beta_j}}{V_j}.$$

Combining results **(1)** and **(2)** for both terms in the maximum (23), we see that for any $\Delta > 0$,

$$\mathbb{E}\rho'_+\left(\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta}\right) > \min\left(0.1364, 0.09\rho'_+(2)\right) z_1 \frac{n_j^{\beta_j}}{\max(\Delta, V_j)} \tag{26}$$

given that $|z_1| \le \frac{1}{2}\frac{\max(\Delta, V_j)}{n_j^{\beta_j}}$. Let $\varepsilon > 0$. It is easy to check that setting

$$z_1 = \frac{\varepsilon}{\min\left(0.1364, 0.09\rho'_+(2)\right)} \sqrt{N} \left(\sum_{j=1}^k \frac{n_j^{1/2+\beta_j}}{\max\left(\Delta, V_j\right)}\right)^{-1}$$

yields, in view of (26), that

$$\sum_{j=1}^k \sqrt{\frac{n_j}{N}}\, \mathbb{E}\rho'_+\left(\frac{n_j^{\beta_j} z_1 - Z_j}{\Delta}\right) > \varepsilon,$$

as long as condition $|z_1| \le \frac{1}{2}\frac{\max(\Delta, V_j)}{n_j^{\beta_j}}$ holds for all j. The latter is equivalent to requirement that

$$\varepsilon \le \frac{\min\left(0.1364, 0.09\rho'_+(2)\right)}{2\sqrt{N}} \left(\sum_{j=1}^k \frac{n_j^{1/2+\beta_j}}{\max\left(\Delta, V_j\right)}\right) \min_{j=1,\dots,k} \frac{\max\left(\Delta, V_j\right)}{n_j^{\beta_j}}.$$

Acknowledgements

I would like to thank Nate Strawn and Anatoli Juditsky for many helpful discussions and insightful suggestions, and two anonymous Referees and the Editor for their comments and remarks that helped improve the quality of presentation. I would also like to acknowledge Nate Strawn's help with the code for numerical simulations.

Appendix A: Central limit theorem in the case of unequal subgroup sizes.

We present an extension of Theorem 4 in the case of non-equal subgroup sizes for the estimator $\widehat{\theta}^{(k)} = \operatorname{med}\left(\bar{\theta}_1, \dots, \bar{\theta}_k\right)$. The following assumptions will be imposed:

1. $X_1, \dots, X_N$ are independent, $\operatorname{card}(G_j) = n_j$, and $\sum_{j=1}^k n_j = k$;
2. Assumption 1 is satisfied with some $\{\sigma_n^{(j)}\}_{n\geq 1}$ and $g_j(n)$, $j = 1, \dots, k$;
3. $k \to \infty$ and $\max_{j=1,\dots,k} \sqrt{k} \cdot g_j(n_j) \to 0$ as $N \to \infty$;
4. $\max_{j\leq k} \frac{H_k}{\sigma_{n_j}^{(j)}\sqrt{k}} \xrightarrow{N\to\infty} 0$, where $H_k := \left(\frac{1}{k}\sum_{j=1}^k \frac{1}{\sigma_{n_j}^{(j)}}\right)^{-1}$ is the harmonic mean of $\sigma_{n_j}^{(j)}$'s.

Theorem 7. *Under assumptions (a)-(e) above,*

$$\sqrt{k}\,\frac{\widehat{\theta}^{(k)} - \theta_*}{H_k} \xrightarrow{d} N\left(0, \frac{\pi}{2}\right).$$

Proof. Define $d_-(x) := I\{x > 0\} - I\{x \leq 0\}$, and $Y_{n_j,j} = d_-\left(\theta_* - \bar{\theta}_j + t\sqrt{\frac{\pi}{2}}\frac{H_k}{\sqrt{k}}\right)$. We will show that

1. $\frac{1}{k}\sum_{j=1}^k \sqrt{k}\mathbb{E}Y_{n_j,j} \to t$ as $N \to \infty$;
2. $\frac{1}{k}\sum_{j=1}^k \operatorname{Var}(Y_{n_j,j}) \to 1$ as $N \to \infty$.

To prove the first claim, first assume that $t \neq 0$ (for $t = 0$ the argument follows the same line with simplifications), and observe that

$$\sqrt{k}\mathbb{E}Y_{n_j,j} = \sqrt{k}\left(\mathbb{E}d_-\left(\frac{\theta_* - \bar{\theta}_j}{\sigma_{n_j}^{(j)}} + t\sqrt{\frac{\pi}{2}}\frac{H_k}{\sigma_{n_j}^{(j)}\sqrt{k}}\right) - \mathbb{E}d_-\left(Z + t\sqrt{\frac{\pi}{2}}\frac{H_k}{\sigma_{n_j}^{(j)}\sqrt{k}}\right)\right)$$
$$+ t\sqrt{\frac{\pi}{2}}\frac{H_k}{\sigma_{n_j}^{(j)}} \cdot \frac{1}{t\sqrt{\frac{\pi}{2}}\frac{H_k}{\sigma_{n_j}^{(j)}\sqrt{k}}}\left(\mathbb{E}d_-\left(Z + t\sqrt{\frac{\pi}{2}}\frac{H_k}{\sigma_{n_j}^{(j)}\sqrt{k}}\right) - \underbrace{\mathbb{E}d_-(Z)}_{=0}\right).$$

Moreover,

$$\left|\sqrt{k}\left(\mathbb{E}d_-\left(\frac{\theta_* - \bar{\theta}_j}{\sigma_{n_j}^{(j)}} + t\sqrt{\frac{\pi}{2}}\frac{H_k}{\sigma_{n_j}^{(j)}\sqrt{k}}\right) - \mathbb{E}d_-\left(Z + t\sqrt{\frac{\pi}{2}}\frac{H_k}{\sigma_{n_j}^{(j)}\sqrt{k}}\right)\right)\right| \leq 2g_j(n_j),$$

while under assumption (d),

$$\frac{1}{t\sqrt{\frac{\pi}{2}}\frac{H_k}{\sigma_{n_j}^{(j)}\sqrt{k}}}\left(\mathbb{E}d_-\left(Z + t\sqrt{\frac{\pi}{2}}\frac{H_k}{\sigma_{n_j}^{(j)}\sqrt{k}}\right) - \underbrace{\mathbb{E}d_-(Z)}_{=0}\right) \to \frac{2}{\sqrt{2\pi}} \text{ as } N \to \infty.$$

□

It then follows from assumption (c) that

$$\left| \frac{1}{k}\sum_{j=1}^{k} \sqrt{k}\mathbb{E}Y_{n_j,j} - t \underbrace{H_k \frac{1}{k}\sum_{j=1}^{k} \frac{1}{\sigma_{n_j}^{(j)}}}_{=1} \right| \to 0 \text{ as } N \to \infty.$$

Claim (b) follows since $\mathbb{E}\left(Y_{n_j,j}\right)^2 = 1$ and $\max_{j\le k} \mathbb{E}Y_{n_j,j} \to 0$ under assumption (d). The rest of the argument repeats the proof of Theorem 4 for $\rho(x) = |x|$.

Appendix B: Supplementary results.

Lemma 4. *Assume that $0 \le \alpha \le 0.33$ and let $z(\alpha)$ be such that $\Phi(z(\alpha)) - 1/2 = \alpha$. Then $z(\alpha) \le 3\alpha$.*

Proof. It is a simple numerical fact that whenever $\alpha \le 0.33$, $z(\alpha) \le 1$; indeed, this follows as $\Phi(1) \simeq 0.8413 > 1/2 + 0.33$. Since $e^{-y^2/2} \ge 1 - \frac{y^2}{2}$, we have

$$\sqrt{2\pi}\alpha = \int_0^{z(\alpha)} e^{-y^2/2} dy \ge z(\alpha) - \frac{1}{6}\left(z(\alpha)\right)^3 \ge \frac{5}{6} z(\alpha), \tag{27}$$

Equation (27) implies that $z(\alpha) \le \frac{6}{5}\sqrt{2\pi}\,\alpha$. Proceeding again as in (27), we see that

$$\sqrt{2\pi}\alpha \ge z(\alpha) - \frac{1}{6}\left(z(\alpha)\right)^3 \ge z(\alpha) - \frac{12\pi}{25}\alpha^2 z(\alpha) \ge z(\alpha)\left(1 - 1.51\,\alpha^2\right),$$

hence $z(\alpha) \le \frac{\sqrt{2\pi}}{1-1.51\,\alpha^2}\,\alpha$. The claim follows since $\alpha \le 0.33$ by assumption, and $\frac{\sqrt{2\pi}}{1-1.51\cdot 0.33^2} < 3$. □

Lemma 5. *Let $A \subset \mathbb{R}$ be symmetric, meaning that $A = -A$, and let $Z \sim N(0,1)$. Then for all $x \in \mathbb{R}$,*

$$\Pr(Z \in A - x) \ge e^{-x^2/2} \Pr(Z \in A).$$

Proof. The result is often known as the Cameron-Martin inequality; we give a short proof for reader's convenience. Observe that

$$\begin{aligned}
\Pr(Z \in A) &= \int_{\mathbb{R}} I\{z \in A\} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = e^{x^2/2} \int_{\mathbb{R}} I\{z \in A\} e^{-xz/2} e^{xz/2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} e^{-x^2/2} dz \\
&\le e^{x^2/2} \sqrt{\int_{\mathbb{R}} I\{z \in A\} \frac{1}{\sqrt{2\pi}} e^{-(z-x)^2/2} dz} \sqrt{\int_{\mathbb{R}} I\{z \in A\} \frac{1}{\sqrt{2\pi}} e^{-(z+x)^2/2} dz} \\
&= e^{x^2/2} \int_{\mathbb{R}} I\{z \in A\} \frac{1}{\sqrt{2\pi}} e^{-(z-x)^2/2} dz = e^{x^2/2} \Pr(Z \in A - x),
\end{aligned}$$

and the claim follows. □

References

Alon, N., Matias, Y. and Szegedy, M. (1996). The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing* 20–29. ACM.

Arcones, M. A. (1996). The Bahadur-Kiefer representation for U-quantiles. *The Annals of Statistics* **24** 1400–1422.

Battey, H., Fan, J., Liu, H., Lu, J. and Zhu, Z. (2015). Distributed Estimation and Inference with Statistical Guarantees. *arXiv preprint arXiv:1509.05457*.

Bentkus, V., Bloznelis, M. and Götze, F. (1997). A Berry–Esséen bound for M-estimators. *Scandinavian journal of statistics* **24** 485–502.

Berry, A. C. (1941). The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society* **49** 122–136.

Bickel, P. J. et al. (1965). On some robust estimates of location. *The Annals of Mathematical Statistics* **36** 847–858.

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

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

Cheng, G. and Shang, Z. (2015). Computational Limits of Divide-and-Conquer Method. *arXiv preprint arXiv:1512.09226*.

Choi, K. P. (1994). On the medians of gamma distributions and an equation of Ramanujan. *Proceedings of the American Mathematical Society* **121** 245–251.

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

Duchi, J. C., Jordan, M. I., Wainwright, M. J. and Zhang, Y. (2014). Optimality guarantees for distributed statistical estimation. *arXiv preprint arXiv:1405.0782*.

Esseen, C.-G. (1942). *On the Liapounoff limit of error in the theory of probability*. Almqvist & Wiksell.

Fan, J., Han, F. and Liu, H. (2014). Challenges of Big Data analysis. *National science review* **1** 293–314.

Fan, J., Wang, D., Wang, K. and Zhu, Z. (2017). Distributed Estimation of Principal Eigenspaces. *arXiv preprint arXiv:1702.06488*.

Haldane, J. B. S. (1948). Note on the median of a multivariate distribution. *Biometrika* **35** 414–417.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions* **196**. John Wiley & Sons.

Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics* 598–611.

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

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

Hsu, D. and Sabato, S. (2013). Loss minimization and parameter estimation with heavy tails. *arXiv preprint arXiv:1307.1827*.

Hsu, D. and Sabato, S. (2016). Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research* **17** 1–40.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35** 73–101.

Ibragimov, I. A. (1967). On the Chebyshev-Cramér asymptotic expansions. *Theory of Probability & Its Applications* **12** 455–469.

Jerrum, M. R., Valiant, L. G. and Vazirani, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science* **43** 169–188.

Joly, E., Lugosi, G. and Oliveira, R. I. (2016). On the estimation of the mean of a random vector. *arXiv preprint arXiv:1607.05421*.

Jordan, M. (2013). On statistics, computation and scalability. *Bernoulli* **19** 1378–1390.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]* **23**. Springer-Verlag, Berlin. Isoperimetry and processes. MR1102015 (93c:60001)

Lee, J. D., Sun, Y., Liu, Q. and Taylor, J. E. (2015). Communication-efficient sparse regression: a one-shot approach. *arXiv preprint arXiv:1503.04337*.

Lehmann, E. L. and D'Abrera, H. J. (2006). *Nonparametrics: statistical methods based on ranks*. Springer New York.

Lerasle, M. and Oliveira, R. I. (2011). Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*.

Li, C., Srivastava, S. and Dunson, D. B. (2016). Simple, Scalable and Accurate Posterior Interval Estimation. *arXiv preprint arXiv:1605.04029*.

Liang, Y., Balcan, M.-F. F., Kanchanapally, V. and Woodruff, D. (2014). Improved distributed Principal Component Analysis. In *Advances in Neural Information Processing Systems* 3113–3121.

Lugosi, G. and Mendelson, S. (2017). Sub-Gaussian estimators of the mean of a random vector. *arXiv preprint arXiv:1702.00482*.

Lugosi, G. and Mendelson, S. (2018). Near-optimal mean estimators with respect to general norms. *arXiv preprint arXiv:1806.06233*.

Mcdonald, R., Mohri, M., Silberman, N., Walker, D. and Mann, G. S. (2009). Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems* 1231–1239.

Minsker, S. a. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335.

Minsker, S., Srivastava, S., Lin, L. and Dunson, D. B. (2014). Robust and scalable Bayes via a median of subset posterior measures. *arXiv preprint arXiv:1403.2660*.

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

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

Pinelis, I. (2016). Optimal-order bounds on the rate of convergence to normality for maximum likelihood estimators. *arXiv preprint arXiv:1601.02177*.

Rosenblatt, J. D. and Nadler, B. (2016). On the optimality of averaging in distributed statistical learning. *Information and Inference* **5** 379–404.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I. and McCulloch, R. E. (2016). Bayes and big data: the consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management* **11** 78–88.

Serfling, R. J. (1981). Approximation theorems of mathematical statistics.

Shafieezadeh-Abadeh, S., Esfahani, P. M. and Kuhn, D. (2015). Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems* 1576–1584.

Shang, Z. and Cheng, G. (2015). A Bayesian Splitotic Theory For Nonparametric Models. *arXiv preprint arXiv:1508.04175*.

Shevtsova, I. (2011). On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands. *arXiv preprint arXiv:1111.6554*.

Small, C. (1990). A survey of multidimensional medians. *International Statistical Review* **58** 263–277.

Steinhardt, J., Charikar, M. and Valiant, G. (2017). Resilience: A criterion for learning

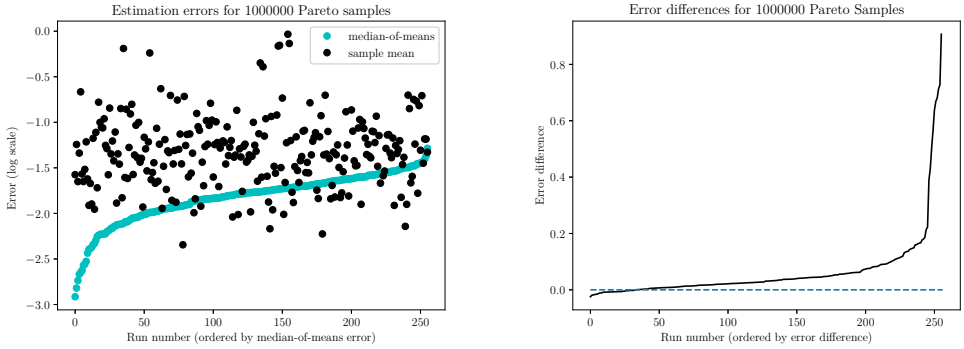
in the presence of arbitrary outliers. *arXiv preprint arXiv:1703.04940*.

Tukey, J. and Harris, T. (1946). Sampling from contaminated distributions. *Ann. Math. Statist* **17501**.

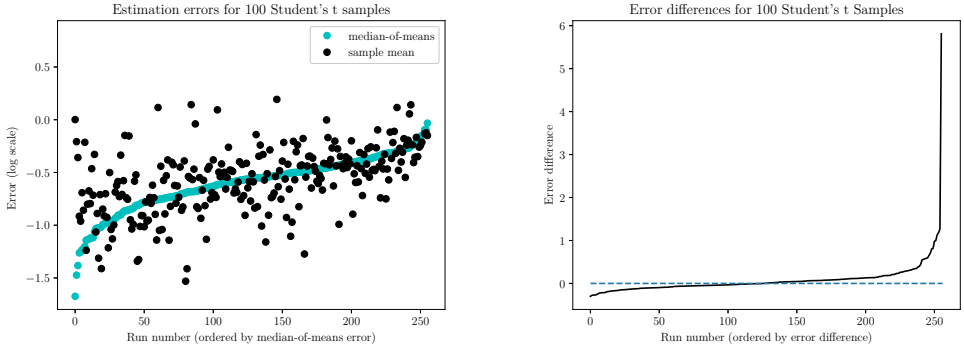
Zhang, Y., Duchi, J. and Wainwright, M. (2013). Divide and conquer kernel ridge regression. In *Conference on Learning Theory* 592–617.

Zhang, Y., Wainwright, M. J. and Duchi, J. C. (2012). Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems* 1502–1510.

Zinkevich, M., Weimer, M., Li, L. and Smola, A. J. (2010). Parallelized stochastic gradient descent. In *Advances in neural information processing systems* 2595–2603.



(a) (b)



(c) (d)

Fig 2: Comparison of errors corresponding to the median-of-means and sample mean estimator over 256 runs of the experiment. In (a) the sample of size $N = 10^6$ consists of i.i.d. random vectors in $\mathbb{R}^2$ with independent Pareto-distributed entries possessing only 2.1 moments. Each run computes the median-of-means estimator using partition into $k = 1000$ groups, as well as the usual sample mean. In (b), the ordered differences between the error of the sample mean and the median-of-means over all 256 runs illustrates robustness. Positive error differences in (b) indicate lower error for the median-of-means, and negative error differences occur when the sample mean provided a better estimate.
Images (c) and (d) illustrate a similar experiment that was performed for two-dimensional random vectors with independent entries with Student's t-distribution with 2 degrees of freedom. In this case, the sample size is $N = 100$ and the number of groups is $k = 10$.

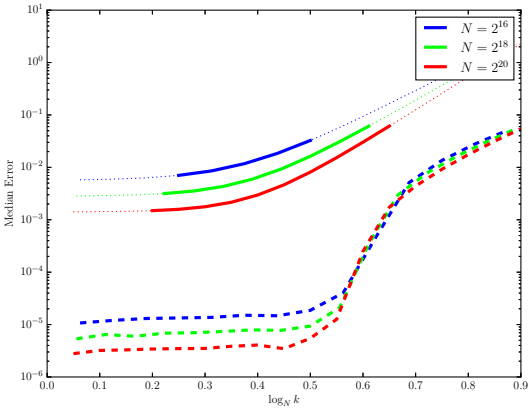


Fig 3: The solid and dotted lines indicate theoretical bounds for the different values of the sample size N, with the solid part indicating the number of subgroups k for which our estimates hold. The dashed lines indicate empirical error between the median-of-means estimator and the true mean. We consider three cases: $N = 2^{16}$ (blue), $N = 2^{18}$ (green), and $N = 2^{20}$ (red). The x-axis is $\log_N k$ taken from a uniform partition of $(0, 1)$ and the y-axis indicates the median error of the median-of-means estimator over 2^{16} runs of the experiment. For each value of N and k, we run 2^{16} simulations by drawing N i.i.d. random variables with Lomax distribution with shape parameter $\alpha = 4$ and scale parameter $\lambda = 1$, splitting into k groups, and then computing the median of the means of those groups. From the 2^{16} simulations, we display (on a logarithmic scale) the median of the absolute differences between the true mean $1/3$ and the median-of-means estimators, producing the dashed lines in the figure. The solid and dotted lines are our theoretical bounds with $4e^{-2s} = 1/2$ (that is, the probability that the solid and dotted bounds holds is guaranteed to be at least $1/2$).

Nominal confidence level	Fraction of outliers					
	0	$\frac{0.2}{\sqrt{N}}$	$\frac{0.4}{\sqrt{N}}$	$\frac{0.6}{\sqrt{N}}$	$\frac{0.8}{\sqrt{N}}$	$\frac{1}{\sqrt{N}}$
0.8	0.94	0.0008	0	0	0	0
0.95	0.99	0.001	0	0	0	0

(a)

Nominal confidence level	Fraction of outliers					
	0	$\frac{0.2}{\sqrt{N}}$	$\frac{0.4}{\sqrt{N}}$	$\frac{0.6}{\sqrt{N}}$	$\frac{0.8}{\sqrt{N}}$	$\frac{1}{\sqrt{N}}$
0.8	0.88	0.82	0.77	0.66	0.6	0.53
0.95	0.99	0.97	0.93	0.85	0.79	0.71

(b)

Fig 4: Empirical coverage levels of confidence intervals constructed using (a) the Central Limit Theorem for the sample mean and (b) Theorem 4 for the median of means; (a) reflects the results obtained for the sample mean and (b) reflects the results obtained for the median-of-means estimator.