Data-Driven Confidence Intervals with Optimal Rates for the Mean of Heavy-Tailed Distributions

Ambrus Tamás[1,2] **Szabolcs Szentpéteri**[1] **Balázs Csanád Csáji**[1,2]

[1] Institute for Computer Science and Control (SZTAKI), Hungarian Research Network (HUN-REN)
[2] Institute of Mathematics, Eötvös Loránd Universtiy (ELTE), Budapest, Hungary

Abstract

Estimating the expected value is one of the key problems of statistics, and it serves as a backbone for countless methods in machine learning. In this paper we propose a new algorithm to build non-asymptotically exact confidence intervals for the mean of a symmetric distribution based on an independent, identically distributed sample. The method combines resampling with median-of-means estimates to ensure optimal subgaussian bounds for the sizes of the confidence intervals under mild, heavy-tailed moment conditions. The scheme is completely data-driven: the construction does not need any information about the moments, yet it manages to build exact confidence regions which shrink at the optimal rate. We also show how to generalize the approach to higher dimensions and prove dimension-free, subgaussian PAC bounds for the exclusion probabilities of false candidates. Finally, we illustrate the method and its properties for heavy-tailed distributions with numerical experiments.

1 Introduction

Mean estimation is a fundamental problem in statistics, but the typical solutions only provide point estimates. On the other hand, confidence regions are often also needed in practice, especially when safety, stability and other risk factors concern us. Classical theory supports the application of the well-known empirical mean estimator and central limit theorem (CLT) based confidence intervals using the sample variance, nevertheless, in practice CLT only provides heuristic solutions and the required assumptions often do not hold. In many real-world problems we need to cope with heavy-tailed distributions and handle outliers in the data, in which cases the empirical mean and standard CLT based confidence intervals can perform poorly, which calls for more robust techniques (Huber and Ronchetti, 2009; Hampel et al., 2011).

In statistical learning the accuracy-confidence trade-off is a key phenomenon (Vapnik, 1998; Györfi et al., 2002; Shalev-Shwartz and Ben-David, 2014). Classical methods offer a variety of *probably approximately correct* (PAC) type bounds for the mean. The celebrated Hoeffding's inequality provides strong bounds for the mean of bounded variables (Hoeffding, 1963). Under the finite variance assumption tighter inequalities are proved in (Bernstein, 1937; Bennett, 1962), however, one needs to know the variance to construct confidence intervals based on these results. Under the boundedness assumption recently an adaptive martingale approach was developed in (Waudby-Smith and Ramdas, 2024) to derive concentration bounds, which empirically outperforms the classical concentration inequalities. For the mean of heavy-tailed distributions neat PAC-bounds are presented in (Catoni, 2012; Lugosi and Mendelson, 2019a), however, these inequalities suffer from similar limitations as the bounds of Bennett (1962) and Bernstein (1937): they assume the a priori knowledge of some moment parameters.

In this paper, motivated by non-asymptotic system identification methods, such as the Sign-Perturbed Sums (SPS) algorithm (Csáji et al., 2015; Szentpéteri and Csáji, 2023) and the Leave-out Sign-dominant Correlation Regions (LSCR) approach (Campi and Weyer, 2005), we introduce the *resampled median-of-means* (RMM) method to construct *exact* confidence intervals for the mean of a symmetric variable based on an i.i.d. sample without making any other a priori assumption. We prove that our method builds confi-

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

dence intervals with *optimal* sizes w.r.t. the confidence parameter (up to constant factors) without using any a priori knowledge about the moments. The main advantage of RMM w.r.t. the original SPS method lies in its robustness, i.e., we prove subgaussian PAC bounds for RMM under heavy-tailed distributional assumptions.

2 Heavy-Tailed Mean Estimation

The problem can be specified as follows. We are given a finite sample of independent and identically distributed (i.i.d.) random variables $\mathcal{D}_0 \doteq \{Y_i\}_{i=1}^n$ from an unknown distribution Q_Y which is *symmetric* about an unknown parameter μ. First, we aim at constructing a *non-asymptotically* valid hypothesis test for the null hypothesis $\mu = \theta$, then we construct a confidence interval for μ based on this test. One of the main advantages of RMM is that it needs no further assumptions other than symmetry to reach any user-chosen (rational) confidence level. Beside the finite sample validity, we prove the *exponential decay* of the rejection probability for $\mu \neq \theta$ under mild moment conditions. We also show that the test induces *exact* confidence intervals for μ and prove *optimal* non-asymptotic PAC inequalities for the sizes of these intervals.

Let $\mathrm{med}(Y)$ denote a median of random variable Y and $\mathrm{med}(\mathcal{D}_0)$ denote the empirical median of a sample. If Q_Y is symmetric, then $\mu = \mathrm{med}(Y)$. We usually assume that $\mathbb{E}Y_1 = \mu$ in which case $\mathbb{E}Y_1 = \mathrm{med}(Y_1)$, however, our results regarding the confidence level also hold without this moment assumption. We denote the well-known empirical mean (point estimate) by

$$\bar{\mu}_n \doteq \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{1}$$

It is known that if $\sigma^2 \doteq D^2 Y_1 < \infty$, then $\bar{\mu}_n$ is the "best" linear unbiased estimator, i.e., $\bar{\mu}_n$ has the lowest variance among linear unbiased estimators (Hastie et al., 2009). By the central limit theorem we have

$$\lim_{n\to\infty} \mathbb{P}\Big(\sqrt{n}\,|\,\bar{\mu}_n - \mu\,| > \sigma\sqrt{2\log(2/\delta)}\Big) \leq \delta. \tag{2}$$

if $\sigma^2 < \infty$. However, this bound holds only asymptotically. In this paper we seek to find similar bounds that hold for every finite sample size.

If the distribution of Y is σ-subgaussian, i.e., if $\mathbb{E}[\exp(\lambda(Y-\mu))] \leq \exp(\sigma^2\lambda^2/2)$ for all $\lambda \in \mathbb{R}$, then

$$\mathbb{P}\left(|\,\bar{\mu}_n - \mu\,| \leq \sigma\sqrt{\frac{2\log(2/\delta)}{n}} \right) \geq 1-\delta. \tag{3}$$

The main advantage of this inequality lies in its finite sample validity. Indeed (2) holds for all $n \in \mathbb{N}$ and $\delta > 0$. However, if the distribution is not subgaussian, Chebyshev's inequality yields a much weaker bound

$$\mathbb{P}\left(|\,\bar{\mu}_n - \mu\,| \leq \sigma\sqrt{\frac{1}{n\delta}} \right) \geq 1-\delta. \tag{4}$$

We can observe that this bound is exponentially worse in δ. Nevertheless, this is the best one can say (Catoni, 2012), because for each $\delta > 0$ there is a distribution with variance σ for which we have

$$\mathbb{P}\left(|\,\bar{\mu}_n - \mu\,| \geq \sigma\sqrt{\frac{c}{\delta n}} \right) \geq \delta. \tag{5}$$

In general, we can conclude that the empirical mean is computationally attractive and statistically efficient under subgaussian assumptions, however, sensitive for outliers which occur with high probability if the variance is large or infinite. In this paper, we aim at finding exponential PAC bounds on the *power* of our hypothesis test and also on the size of the corresponding confidence region under much milder assumptions on Q_Y than subgaussianity (cf. Assumption A3).

We assume w.l.o.g. that $n = k\cdot\tilde{n}$, where k is an odd integer and $\tilde{n} \in \mathbb{N}$. An important example of estimators for the expected value is the *median-of-means* method (Nemirovsky and Yudin, 1983; Alon et al., 1996; Huber and Ronchetti, 2009), defined by

$$\widehat{\mu}(\mathcal{D}_0) \doteq \mathrm{med}\left(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} Y_i, \ldots, \frac{1}{\tilde{n}}\sum_{i=(k-1)\tilde{n}+1}^{k\tilde{n}} Y_i \right). \tag{6}$$

One of the main advantages of the median-of-means estimate $\widehat{\mu}$ is formulated in Theorem 2.1 below (Lugosi and Mendelson, 2019a, Theorem 2).

Theorem 2.1. *Let $\{Y_1,\ldots,Y_n\}$ be an i.i.d. sample and assume that $D^2Y_1 = \sigma^2 < \infty$. Let $\delta \in (0,1)$, $k = \lceil 8\log(1/\delta)\rceil$ and $n = \tilde{n}k$, then*

$$\mathbb{P}\left(|\,\widehat{\mu} - \mu\,| \leq \sigma\sqrt{\frac{32\log(1/\delta)}{n}} \right) > 1-\delta. \tag{7}$$

Under milder moment conditions, a generalized bound can be proved (Devroye et al., 2016, Theorem 3.1), (Lugosi and Mendelson, 2019a, Theorem 3):

Theorem 2.2. *Let $\{Y_1,\ldots,Y_n\}$ be an i.i.d. sample and $\mathbb{E}Y_1 = \mu$. Assume that there is an $a \in (0,1]$ with*

$$\mathbb{E}\big[|\,Y - \mathbb{E}Y\,|^{1+a}\big] = M < \infty. \tag{8}$$

Let $\delta > 0$, $k = \lceil 8\log(2/\delta)\rceil$ and $n = k\tilde{n}$. Then, the median-of-means estimate $\widehat{\mu}$ with k blocks satisfies

$$\mathbb{P}\left(|\,\widehat{\mu} - \mu\,| \leq 8\left(\frac{12M^{1/a}\log(1/\delta)}{n}\right)^{\frac{a}{1+a}} \right) > 1-\delta. \tag{9}$$

Moreover, for any mean estimator μ_n, there exists a probability distribution with mean μ and $(1+a)$th central moment M, such that

$$\mathbb{P}\left(\,|\, \mu_n - \mu \,|\, > \left(\frac{M^{1/a} \log(2/\delta)}{n} \right)^{\frac{a}{1+a}} \right) \geq \delta. \quad (10)$$

The second part of the theorem shows that these type of bounds are optimal w.r.t. δ up to a constant factor, which implies the tightness of the presented type II error rate of the proposed test.

3 Exploiting Symmetry

Our main goal in this paper is to construct confidence intervals with optimal sizes, see Equation (9), based on a non-asymptotically exact hypothesis test, hence for a given $\theta \in \mathbb{R}$ let us consider

$$H_0 : \mu = \theta \qquad \text{and} \qquad H_1 : \mu \neq \theta \quad (11)$$

under the following assumptions:

A1. $Y_1, \ldots, Y_n$ are i.i.d.

A2. Q_Y is symmetric about μ.

We emphasise that the hypothesis test that we present is *exact* under these two conditions. We note that the assumption that variables $\{Y_1, \ldots, Y_n\}$ have the same distribution can be weakened for most of our results. The key property that we need is that each variable possesses the same *symmetry* point μ. Nevertheless, we work with an additional assumption on the moment of the observed variables to quantify the type II error probabilities or the *power* of the presented test:

A3. $\mathbb{E}\big[\,|\, Y - \mathbb{E}\, Y \,|^{1+a}\,\big] \,=\, M < \infty$ for an $a \in (0, 1]$.

We also quantify the rate of shrinkage of the corresponding confidence regions under Assumptions A1-A3. One can observe that from A2 and A3 it follows that $\mathbb{E}Y = \mu$. Furthermore, in particular if $a = 1$, then assumption A3 requires $\sigma^2 < \infty$. We emphasise that the presented test does not use the knowledge of constants a and M, yet we can prove the optimality of the sizes of the corresponding confidence regions that we build. This is one of the main advantage of our scheme compared to the methods that are used in practice, for example, see (Bubeck et al., 2013).

Let p be the desired (rational) significance level. Let us find integers r and m such that $p = r/m$. In this paper we propose a new resampling method to test nullhypothesis H_0. Let $Y(\theta) = \alpha(Y - \theta) + \theta$ be a parameterized random variable defined with a Rademacher variable α independent of Y. Clearly $\mathbb{E}[Y(\theta)] = \theta$ and $Y(\theta)$ is symmetric about θ. By Jensen's inequality

$$\begin{aligned} &\mathbb{E}[|Y(\theta) - \mathbb{E}[Y(\theta)]|^{1+a}] = \mathbb{E}[|\alpha|^{1+a}|Y-\theta|^{1+a}] \\ &\leq \mathbb{E}\left[\left(\frac{2|Y-\mu| + 2|\mu-\theta|}{2}\right)^{1+a}\right] \\ &\leq \mathbb{E}\left[\frac{1}{2}(2|Y-\mu|)^{1+a} + \frac{1}{2}(2|\mu-\theta|)^{1+a}\right] \\ &= 2^a\big(\,\mathbb{E}[|Y-\mu|^{1+a}] + d^{1+a}\,\big) \leq 2(M + d^{1+a}), \end{aligned} \quad (12)$$

where $d \doteq |\,\theta - \mu\,|$. In particular, for $a = 1$ we can compute the exact variance of $Y(\theta)$ by

$$\begin{aligned} &D^2[Y(\theta)] = \mathbb{E}[\alpha^2(Y-\theta)^2] \\ &= \mathbb{E}[(Y-\mu)^2] + \mathbb{E}[(\mu-\theta)^2] = \sigma^2 + d^2. \end{aligned} \quad (13)$$

For notational simplicity let $W \doteq Y - \mu$ be the centered version of Y and $\mathcal{W}_0 \doteq \{W_i\}_{i=1}^n$, where $W_i = Y_i - \mu$ for $i \in [n]$. Keep in mind, however, that $\{W_i\}_{i=1}^n$ are not observed. Note that $Y = \mu + W$ and $Y(\mu) = \mu + \alpha \cdot W$, where α is a Rademacher variable, have the same distribution, because W is symmetric about zero. Moreover, one can prove that Y and $Y(\mu)$ are conditionally i.i.d. w.r.t. $|W|$, hence they are also exchangeable (Csáji et al., 2015). On the other hand if $\theta \neq \mu$, then the distribution of Y differs from the distribution of $Y(\theta)$, e.g., a key difference is that $Y(\theta)$ is symmetric about θ whereas Y is symmetric about μ. We aim at detecting this difference. For our test we generate alternative samples using H_0 and compare the new (resampled) datasets to the original one, i.e., for $i \in [n]$ and $j \in [m-1]$ let $\{\alpha_{i,j}\}$ be i.i.d. random signs (Rademacher variables), for $j \in [m-1]$, and let

$$\mathcal{D}_j(\theta) \doteq \{\alpha_{1,j}(Y_1 - \theta) + \theta, \ldots, \alpha_{n,j}(Y_n - \theta) + \theta\} \quad (14)$$

be parameter-dependent *alternative samples*. Further, let $\mathcal{D}_0(\theta) \doteq \mathcal{D}_0$, for $\theta \in \mathbb{R}$. It is easy to see that $\mathcal{D}_j(\theta)$ is an i.i.d. sample from the distribution of $Y(\theta)$ for $j \neq 0$. We decide about H_0 by comparing $\mathcal{D}_0(\theta)$ to $\mathcal{D}_j(\theta)$ for $j \in [m-1]$ with a *ranking function*, see Definition 3.1. If $\mathcal{D}_0(\theta)$ differs significantly from $\mathcal{D}_j(\theta)$, then we reject H_0, otherwise we accept the null hypothesis.

Definition 3.1 (ranking function)**.** *Let $\mathbb{A}$ be a measurable space (with some σ-algebra), a (measurable) function $\psi : \mathbb{A}^m \to [\,m\,]$ is called a ranking function if for all $(a_1, \ldots, a_m) \in \mathbb{A}^m$ it satisfies P1 and P2.*

P1 For all permutation τ on set $\{2, \ldots, m\}$, we have

$$\psi\big(\,a_1, a_2, \ldots, a_m\,\big) \;=\; \psi\big(\,a_1, a_{\tau(2)}, \ldots, a_{\tau(m)}\,\big),$$

that is function ψ is invariant w.r.t. reordering the last $m-1$ terms of its arguments.

P2 For all $i, j \in [\,m\,]$, if $a_i \neq a_j$, then we have

$$\psi\big(\,a_i, \{a_k\}_{k\neq i}\,\big) \;\neq\; \psi\big(\,a_j, \{a_k\}_{k\neq j}\,\big),$$

where the simplified notation is justified by P1.

The value of a ranking function is called the *rank*. The main observation about the rank is given by the lemma that follows (Csáji and Tamás, 2019, Lemma 1):

Lemma 3.1. *Let $\xi_1, \ldots, \xi_m$ be almost surely pairwise different exchangeable random elements in a measurable space and let ψ be a ranking function, then $\psi(\xi_1, \ldots, \xi_m)$ is uniformly distributed on $[m]$.*

The original sample and the alternative samples are random vectors in $\mathbb{R}^n$. We can observe that the datasets can be identical, for example, if every sign that we generate equals to $+1$. This poses a technical challenge in ranking. In order to resolve this problem, we use a tie-breaking permutation, cf. (Csáji et al., 2015). Let π be a random permutation on $[m-1]_0 \doteq \{0, \ldots, m-1\}$ generated uniformly from the permutation group on $[m-1]_0$, independently from $\mathcal{D}_0$ and $\{\alpha_{i,j}\}$. We define the a.s. pairwise different extended datasets as $\mathcal{D}_j^\pi(\theta) \doteq (\mathcal{D}_j(\theta), \pi(j))$ for $j = 0, \ldots, m-1$. It is easy to prove that $\mathcal{D}_0^\pi(\mu), \ldots, \mathcal{D}_{m-1}^\pi(\mu)$ are exchangeable, hence we obtain an exact hypothesis test with significance level r/m if we reject H_0 if and only if we have $\psi(\mathcal{D}_0^\pi(\theta), \ldots, \mathcal{D}_{m-1}^\pi(\theta)) > m - r$.

Theorem 3.1. *For any ranking function ψ, if assumptions A1 and A2 hold, then we have*

$$\mathbb{P}\big(\, \psi\big(\, \mathcal{D}_0^\pi(\mu), \ldots, \mathcal{D}_{m-1}^\pi(\mu) \,\big) > m - r \,\big) \;=\; \frac{r}{m}. \quad (15)$$

We can observe that (15) provides the *exact* probability of type I error for every sample size $n \in \mathbb{N}$ and for every symmetric distribution Q_Y. Our conditions on ψ are also very mild. In fact at this point we allow degenerate rankings, as well, e.g., rankings that only depend on the tie-breaking permutation. Nevertheless, we would like to avoid these "pathological" choices and apply rankings that admit finite sample PAC bounds.

One of our main ideas is to compare the empirical estimate of the distance between parameter θ and the median-of-means estimator computed from each sample. If $\theta = \mu$, then each estimate has the same distribution, otherwise $\widehat{\mu}(\mathcal{D}_0)$ is "farther" from θ than $\widehat{\mu}(\mathcal{D}_j(\theta))$ with high probability for $j = 1, \ldots, m-1$. Let us consider *reference variable functions*

$$S_j(\theta) \doteq |\, \widehat{\mu}(\mathcal{D}_j(\theta)) - \theta \,| \quad \text{for} \quad j = 0, \ldots, m-1. \quad (16)$$

Because of the reasoning above $S_0(\theta)$ should be greater than $S_j(\theta)$ for $j \in [m-1]$ for those $\theta \in \mathbb{R}$ that are "far" from μ. However, $S_0(\theta)$ and $S_j(\theta)$ have the same distribution if $\theta = \mu$, and they are exchangeable. In conclusion we define the ranking function by

$$\begin{aligned} \mathcal{R}(\theta) \;&=\; \psi(\mathcal{D}_0^\pi(\theta), \ldots, \mathcal{D}_{m-1}^\pi(\theta)) \\ &\doteq\; 1 + \sum_{i=1}^{m-1} \mathbb{I}\big(\, S_0(\theta) \succ_\pi S_j(\theta) \,\big), \end{aligned} \quad (17)$$

Algorithm 1: RMM Hypothesis Test (for $\mu = \theta$)

Inputs: i.i.d. sample $\mathcal{D}_0$, rational significance level p, tie-breaking permutation π on $[m-1]_0$

1: Choose integers $1 \leq r < m$ such that $p = r/m$.
2: For $(i, j) \in [n] \times [m-1]_0$ generate $n(m-1)$ independent Rademacher signs $\{\alpha_{i,j}\}$.
3: Construct $m-1$ alternative datasets for $j \in [m-1]$:

$$\mathcal{D}_j(\theta) \doteq \{\alpha_{1,j}(Y_1 - \theta) + \theta, \ldots, \alpha_{n,j}(Y_n - \theta) + \theta\}$$

and let $\mathcal{D}_j^\pi(\theta) \doteq (\mathcal{D}_j(\theta), \pi(j))$ for $j \in [m-1]_0$.
4: Compute the reference variables

$$S_j(\theta) \doteq |\, \widehat{\mu}(\mathcal{D}_j(\theta)) - \theta \,| \quad \text{for} \quad j = 0, \ldots, m-1.$$

5: Compute the rank $\mathcal{R}(\theta)$ according to (17).
6: Reject H_0 if and only if $\mathcal{R}(\theta) > m - r$.

where $\prec_\pi$ is defined as the standard $<$ ordering with tie-breaking, i.e., $S_j(\theta) \prec_\pi S_k(\theta)$ if and only if $S_j(\theta) < S_k(\theta)$ or $\big(\, S_j(\theta) = S_k(\theta) \text{ and } \pi(j) < \pi(k) \,\big)$ (Csáji et al., 2015). We reject θ if and only if $\mathcal{R}(\theta) > m - r$. The test is summarized in Algorithm 1.

Theorem 3.2. *Assume A1 and A2, then the test defined in Algorithm 1 is exact, i.e., for every $1 \leq r < m$*

$$\mathbb{P}\,(\, \mathcal{R}(\mu) > m - r \,) = \frac{r}{m}. \quad (18)$$

Proof. The claim is a corollary of Theorem 3.1. One can show that $\mathcal{D}_0, \mathcal{D}_1(\mu), \ldots, \mathcal{D}_{m-1}(\mu)$ are exchangeable (Csáji and Tamás, 2019) and use Lemma 3.1. □

We prove a non-asymptotic pointwise bound for the type II error probability of the presented test.

Theorem 3.3. *Assume A1, A2 and A3. Let $\delta > 0$, $k = \lceil 8 \log(2(m-r+1)/\delta) \rceil$, $1 \leq r < m$ be user-chosen integers and $\mathcal{R}$ be defined by (17). If*

$$8\left(\frac{24(M + d^{1+a})^{1/a} \log\big(\tfrac{m-r+1}{\delta}\big)}{n} \right)^{\frac{a}{1+a}} < \frac{d}{2}, \quad (19)$$

holds for $d = |\theta - \mu|$, then we have

$$\mathbb{P}\,(\, \mathcal{R}(\theta) \leq m - r \,) \leq \delta. \quad (20)$$

We can observe that if we fix δ, m, r and θ, then Equation (19) holds for n large enough. In addition, for a given sample size we can find the smallest δ for which (19) holds. In this case the smallest upper bound for (20) will be exponentially small with respect to the sample size. Moreover, if Equation (19) is satisfied for a parameter θ with $D = |\mu - \theta|$, then it also holds for every $\theta \in \mathbb{R}$ such that $|\mu - \theta| \geq D$.

Proof. Let us fix $\theta \neq \mu$, integers $r < m$ and use notation $\widehat{\mu}^j \doteq \widehat{\mu}(\mathcal{D}_j(\theta))$ for $j \in [m-1]_0$. First we observe that if $|\widehat{\mu}^0 - \mu| < d/2$, then $|\widehat{\mu}^0 - \theta| \geq d/2$, thus

$$\begin{aligned}\mathbb{P}(\,\mathcal{R}(\theta) > m - r\,) \geq \mathbb{P}\big(\,|\,\widehat{\mu}^0 - \mu\,| < d/2 \text{ and} \\ |\,\widehat{\mu}^j - \theta\,| < d/2 \text{ for } j = 1, \ldots, m-r\,\big).\end{aligned} \tag{21}$$

Then, we define ancillary events as

$$\begin{aligned}&B_0 \doteq \{\,|\,\widehat{\mu}^0 - \mu\,| < {}^{d}\!/_{2}\,\},\\ &B_j \doteq \{\,|\,\widehat{\mu}^j - \theta\,| < {}^{d}\!/_{2}\,\} \text{ for } j = 1, \ldots, m-1.\end{aligned} \tag{22}$$

By the union bound, we have

$$\begin{aligned}&\mathbb{P}\Big(\bigcup_{j=0}^{m-r} \bar{B}_j\Big) \leq \sum_{j=0}^{m-r} \mathbb{P}(\bar{B}_j)\\ &= \mathbb{P}\big(\,|\widehat{\mu}^0 - \mu| \geq {}^{d}\!/_{2}\,\big) + \sum_{j=1}^{m-r} \mathbb{P}\big(\,|\widehat{\mu}^j - \theta| \geq {}^{d}\!/_{2}\,\big).\end{aligned} \tag{23}$$

Because of (19), the first term can be bounded as

$$\begin{aligned}&\mathbb{P}(|\,\widehat{\mu}^0 - \mu\,| \geq d/2) \\ &\leq \mathbb{P}\bigg(|\,\widehat{\mu}^0 - \mu\,| > 8\Big(\frac{12M^{1/a}\log(1/\tilde{\delta})}{n}\Big)^{\frac{a}{1+a}}\bigg) \leq \tilde{\delta},\end{aligned} \tag{24}$$

where $\delta = (m - r + 1)\tilde{\delta}$. The other terms can be bounded similarly by

$$\begin{aligned}&\mathbb{P}(|\widehat{\mu}^j - \theta| \geq d/2) \leq \\ &\mathbb{P}\bigg(|\,\widehat{\mu}^j - \theta\,| > 8\Big(\frac{24(M + d^{1+a})^{1/a}\log(1/\tilde{\delta})}{n}\Big)^{\frac{a}{1+a}}\bigg) \leq \tilde{\delta}.\end{aligned} \tag{25}$$

In conclusion, under (19) we have

$$\mathbb{P}(\,\mathcal{R}(\theta) \leq m - r\,) \;\leq\; (m - r + 1)\,\tilde{\delta} \;=\; \delta, \tag{26}$$

which proves the desired bound. □

If $a = 1$, the constants of the previous theorem can be strengthened to obtain Theorem 3.4.

Theorem 3.4. *Assume A1, A2 and A3. Let $\delta > 0$, $k = \lceil 8\log((m - r + 1)/\delta)\rceil$, $1 \leq r < m$ be user-chosen integers and $\mathcal{R}$ be defined by* (17)*. If*

$$4(\sigma^2 + d^2)\frac{32\log((m - r + 1)/\delta)}{n} \leq d^2 \tag{27}$$

holds for $d = |\mu - \theta|$, then we have

$$\mathbb{P}(\,\mathcal{R}(\theta) \leq m - r\,) \leq \delta. \tag{28}$$

The proof is included in the supplementary materials.

4 Confidence Intervals

In this section we apply the presented hypothesis test to define the RMM method to construct confidence regions for μ. We include those parameters in the confidence set that are accepted by Algorithm 1, that is

$$\Theta_n \;\doteq\; \{\,\theta : \mathcal{R}(\theta) \leq m - r\,\}. \tag{29}$$

Note that we do not need to regenerate the random signs for each θ, the same set of signs can be used. Moreover, as we will see, Θ_n is a (possibly degenerate) *interval* and its endpoints can be explicitly computed.

Because of Theorem 3.2, the following claim holds:

Corollary 4.1. *Assume A1 and A2, then Θ_n is an* exact *confidence region for μ, i.e., for $1 \leq r < m$*

$$\mathbb{P}\big(\,\mu \in \Theta_n\,\big) \;=\; 1 - \frac{r}{m}. \tag{30}$$

Additionally, Theorem 3.3 implies that the inclusion probability for any $\theta \neq \mu$ goes to zero with an exponential rate as the sample size tends to infinity. In this section we seek to give a PAC bound on the size of Θ_n under the finite moment condition, A3. Let

$$\operatorname{diam}(\Theta_n) \;\doteq\; \sup_{\theta_1,\theta_2} \big\{\,|\,\theta_1 - \theta_2\,| : \theta_1, \theta_2 \in \Theta_n\,\big\} \tag{31}$$

as usually. It can be shown that $\operatorname{diam}(\Theta_n)$ is a random variable, which can be infinite. We prove optimal PAC bounds on the shrinkage of $\operatorname{diam}(\Theta_n)$.

Theorem 4.1. *Assume A1, A2 and A3. Let $m = 2$, $\delta > 0$, $k = \lceil 8\log({}^{20}\!/_{\delta})\rceil$, $\mathcal{R}$ be defined by* (17) *and $\Theta_n \doteq \{\,\theta : \mathcal{R}(\theta) \leq 1\,\}$, then for $n \geq k(k + 8\log(k))$*

$$\mathbb{P}\bigg(\,\operatorname{diam}(\Theta_n) > 8\Big(\frac{12M^{1/a}\log(10/\delta)}{n}\Big)^{\frac{a}{1+a}}\,\bigg) \leq \delta. \tag{32}$$

Proof. The proof consists of several parts. First, we fix $\varepsilon > 0$ and bound the probability

$$\mathbb{P}(\operatorname{diam}(\Theta_n) > \varepsilon) \tag{33}$$

from above with two terms. We handle the case when there are too many or too few $+1$ in a random sign vector realization separately, because the probability of these events is exponentially small. Let

$$Z_\ell \doteq \frac{1}{2}\sum_{i=(\ell-1)\tilde{n}+1}^{\ell\tilde{n}} (1 - \alpha_i) \quad \text{for} \quad \ell = 1, \ldots, k. \tag{34}$$

We can see that $\{Z_\ell\}_{\ell=1}^k$ are independent binomial variables with parameters $\tilde{n}$ and ${}^1\!/_2$ and expected value ${}^{\tilde{n}}\!/_2$. Let us define the events that follow

$$\begin{aligned}&A_\ell \doteq \big\{\,|\,Z_\ell - \mathbb{E}Z_\ell\,| \leq {}^{\tilde{n}}\!/_{4}\,\big\}, \;\text{ for } \ell = 1, \ldots, k,\\ &A \doteq \bigcap_{\ell=1}^{k} A_\ell = \big\{\,\max_{1\leq\ell\leq k} |Z_\ell - \mathbb{E}Z_\ell| \leq {}^{\tilde{n}}\!/_{4}\,\big\}.\end{aligned} \tag{35}$$

By the law of total probability, we have

$$\begin{aligned}&\mathbb{P}(\mathrm{diam}(\Theta_n) > \varepsilon) = \mathbb{P}(\mathrm{diam}(\Theta_n) > \varepsilon \,|\, A)\,\mathbb{P}(A)\\&+ \mathbb{P}(\mathrm{diam}(\Theta_n) > \varepsilon \,|\, \bar{A})\,\mathbb{P}(\bar{A})\\&\leq \mathbb{P}(\{\mathrm{diam}(\Theta_n) > \varepsilon\} \cap A) + \mathbb{P}(\bar{A}).\end{aligned} \tag{36}$$

We bound the second term with the union bound and Hoeffding's inequality as

$$\begin{aligned}&\mathbb{P}\big(\max_{1\leq\ell\leq k} |Z_\ell - \mathbb{E}Z_\ell| > \tilde{n}/4\big) = \mathbb{P}\bigg(\bigcup_{\ell=1}^{k} \bar{A}_\ell\bigg)\\&\leq \sum_{\ell=1}^{k} \mathbb{P}(\bar{A}_\ell) = k\,\mathbb{P}(\bar{A}_\ell)\\&\leq k\,\mathbb{P}(|Z_\ell - \mathbb{E}Z_\ell| > \tilde{n}/4) \leq 2\,k\,\exp\bigg(-\frac{n}{8k}\bigg).\end{aligned} \tag{37}$$

In the next step, we show that if A occurs, then Θ_n is a bounded interval. Reference variable function

$$S_0(\theta) = |\,\widehat{\mu}(\mathcal{D}_0) - \theta\,| \tag{38}$$

is the absolute value of a linear function with slope -1. Let us define the sub-sample alternative lines as

$$S_1^{(\ell)}(\theta) \doteq \frac{1}{\tilde{n}} \sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1}(Y_i - \theta) \text{ for } \ell = 1,\ldots,k. \tag{39}$$

Then, the alternative reference function equals to

$$\begin{aligned}S_1(\theta) &= |\widehat{\mu}(\mathcal{D}_1(\theta)) - \theta|\\&= \big|\,\mathrm{med}\,\big(S_1^{(1)}(\theta) + \theta, \ldots, S_1^{(k)}(\theta) + \theta\big) - \theta\big|\\&= \big|\,\mathrm{med}\,\big(S_1^{(1)}(\theta), \ldots, S_1^{(k)}(\theta)\big)\big|,\end{aligned} \tag{40}$$

which is the median of linear functions with slopes strictly between -1 and $+1$. Without taking the absolute value, the median of linear functions have exactly one intersection with $\widehat{\mu}(\mathcal{D}_0) - \theta$, i.e., equation

$$\widehat{\mu}(\mathcal{D}_0) - \theta = \mathrm{med}\,\big(S_1^{(1)}(\theta), \ldots, S_1^{(k)}(\theta)\big) \tag{41}$$

has exactly one solution. It can be shown that the solution of this equation equals to the median of the intersections of $\widehat{\mu}(\mathcal{D}_0) - \theta$ with sub-sample lines $\{S_1^{(\ell)}(\theta)\}$. These intersections can be computed as

$$\begin{aligned}\nu_\ell^- &= \frac{\widehat{\mu}(\mathcal{D}_0) - \frac{1}{\tilde{n}}\sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1} Y_i}{1 - \frac{1}{\tilde{n}}\sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1}}\\&= \mu + \frac{\widehat{\mu}(\mathcal{W}_0) - \frac{1}{\tilde{n}}\sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1} W_i}{\frac{1}{\tilde{n}}\sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} (1 - \alpha_{i,1})}\\&= \mu + \frac{\widehat{\mu}(\mathcal{W}_0) - \frac{1}{\tilde{n}}\sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1} W_i}{\frac{2}{\tilde{n}} Z_\ell}\end{aligned} \tag{42}$$

for $\ell = 1,\ldots,k$. Hence, the median of intersections

$$\begin{aligned}\nu^- &= \mathop{\mathrm{med}}_{\ell\in[k]} \nu_\ell^-\\&= \mu + \mathop{\mathrm{med}}_{\ell\in[k]} \left(\frac{\widehat{\mu}(\mathcal{W}_0) - \frac{1}{\tilde{n}}\sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1} W_i}{\frac{2}{\tilde{n}} Z_\ell}\right)\\&= \mu + V^-,\end{aligned} \tag{43}$$

is an intersection of $S_0(\theta)$ and $S_1(\theta)$, where

$$V^- \doteq \mathop{\mathrm{med}}_{\ell\in[k]} \left(\frac{\widehat{\mu}(\mathcal{W}_0) - \frac{1}{\tilde{n}}\sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1} W_i}{\frac{2}{\tilde{n}} Z_\ell}\right). \tag{44}$$

Similarly, another intersection of $S_0(\theta)$ and $S_1(\theta)$ is

$$\nu^+ = \mu + V^+, \tag{45}$$

where

$$V^+ \doteq \mathop{\mathrm{med}}_{\ell\in[k]} \left(\frac{\widehat{\mu}(\mathcal{W}_0) + \frac{1}{\tilde{n}}\sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1} W_i}{\frac{2}{\tilde{n}} (\tilde{n} - Z_\ell)}\right). \tag{46}$$

Since $S_1(\theta)$ is "flatter" than $S_0(\theta)$, these are the only two intersections and

$$\Theta_n \;=\; \{\,\theta : S_0(\theta) \prec_\pi S_1(\theta)\,\} \;=\; [\lambda, \varrho]_\pi, \tag{47}$$

where $\lambda \doteq \min(\nu^-, \nu^+)$, $\varrho \doteq \max(\nu^-, \nu^+)$ and $[\cdot,\cdot]_\pi$ denotes the interval of which endpoints are included if $\pi(0) < \pi(1)$. We can also observe that V^- and V^+ have the same symmetric distribution and both ν^- and ν^+ are well-defined, because $\tilde{n}/4 \leq Z_\ell \leq 3\tilde{n}/4$. Thus, the size of the confidence interval can be computed as

$$\mathrm{diam}(\Theta_n) = |\,\nu^+ - \nu^-\,| = |\,V^+ - V^-\,|. \tag{48}$$

We note that the formulas for $\nu^\pm$ remain valid if all $\{Z_\ell, \tilde{n} - Z_\ell\}_{\ell=1}^k$ are positive and are extended for the degenerate cases in Algorithm 2. Using symmetry and the application of the union bound yield

$$\begin{aligned}&\mathbb{P}(\mathrm{diam}(\Theta_n) > \varepsilon \,|\, A) \leq \mathbb{P}(|V^+| > \varepsilon/2 \text{ or } |V^-| > \varepsilon/2 \,|\, A)\\&\leq \mathbb{P}(|V^+| > \varepsilon/2 \,|\, A) + \mathbb{P}(|V^-| > \varepsilon/2 \,|\, A)\\&= 2 \cdot \mathbb{P}(|V^-| > \varepsilon/2 \,|\, A) \leq 4 \cdot \mathbb{P}(V^- > \varepsilon/2 \,|\, A).\end{aligned} \tag{49}$$

We can see that Z_ℓ is independent of the nominator in (44), because $\alpha_{i,1}W_i$ and $\alpha_{i,1}$ are independent for each $i \in [n]$. We consider the events that follow

$$B = \{V^- > \varepsilon/2\}, \tag{50}$$

$$\tilde{B} = \left\{\mathop{\mathrm{med}}_{\ell\in[k]} \left(\frac{\widehat{\mu}(\mathcal{W}_0) - \frac{1}{\tilde{n}}\sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1} W_i}{\frac{1}{2}}\right) > \varepsilon/2\right\}.$$

Our key observation is that $B \cap A \subseteq \tilde{B} \cap A$, because when V^- is positive, then decreasing Z_ℓ in the denominator for every $\ell \in [k]$ to $\tilde{n}/4$, increases the median. Consequently similarly as in (49)

$$\mathbb{P}(B \cap A) \;\leq\; \mathbb{P}(\tilde{B} \cap A) \;\leq\; 2 \cdot \mathbb{P}\big(\widehat{\mu}(\mathcal{W}_0) > \varepsilon/8\big). \tag{51}$$

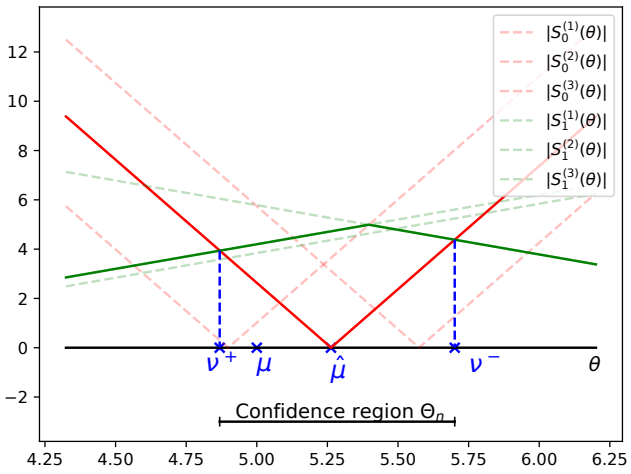


Figure 1: Construction of the confidence interval with $m = 2$ (50 % confidence), $n = 30$ and $k = 3$.

The details of this calculation can be found in the supplementary material. In conclusion, if $k = \lceil 8\log(2/\tilde{\delta})\rceil$, $\varepsilon = 64\Big(\frac{12M^{1/a}\log(1/\tilde{\delta})}{n}\Big)^{\frac{a}{1+a}}$ and $n \geq k(8\log(k)+k)$:

$$\mathbb{P}\Big(\ \operatorname{diam}(\Theta_n) > 8\Big(\frac{12M^{1/a}\log(1/\tilde{\delta})}{n}\Big)^{\frac{a}{1+a}}\Big) \leq 2\,k\exp\Big(-\frac{n}{8k}\Big) + 8\,\tilde{\delta} \leq 10\,\tilde{\delta}. \tag{52}$$

from which the theorem follows. □

We proved that $\Theta_n = [\lambda, \varrho\,]_\pi$, where random permutation π "decides" about the endpoints, and Θ_n is bounded with high probability. In addition, $\widehat{\mu}(\mathcal{D}_0)$ is always included in Θ_n if $\Theta_n \neq \emptyset$. The proof provided explicit formulas, cf. (42), (43), and (45), to construct the confidence interval for $m = 2$. The time and space complexities of the algorithm are linear in n and k.

An illustrative example of the confidence interval (Θ_n) construction for $m = 2$ and $k = 3$ is presented in Figure 1. In this example, the dashed lines are the subsample reference and alternative lines, while the solid lines are the reference variable functions for $j = 0$ and $j = 1$. Then, the confidence region is given by the intersections of the reference variables for $p = 0.5$.

We also show that $\Theta_n = [\lambda, \varrho\,]_\pi$ for every $m \geq 2$ and present an explicit formula for the interval for all sign realizations $\{\alpha_{i,j}\}$ and $m \geq 2$ in Algorithm 2. We prove, as well, that Θ_n admits the special form of (60).

In (53), we calculate, for each j and ℓ, the points in which $S_j^{(\ell)}(\theta)$ is equal to $\widehat{\mu} - \theta$ and $\theta - \widehat{\mu}$, respectively. With high probability, these functions have only one intersection defined by the formula in (53) independently from s, i.e., typically $\nu^-_{j,-1} = \nu^-_{j,+1}$. However, a technical challenge is posed by the tie-breaking. If for some $\ell \in [k]$ the corresponding signs $\{\alpha_{i,j}\}_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}}$ are all equal to $+1$, $S_j^{(\ell)}(\theta)$ is parallel with $\widehat{\mu} - \theta$. In this case, we define the extended intersection values as $+\infty$ or $-\infty$, respectively. For $0/0$, the extended intersection values are determined based on the tie-breaking permutation π. Index s is introduced to include the dependence of these extended values on π. Then, the interval in which we have $S_0(\theta) \prec_\pi S_j(\theta)$ is determined by (54), (55), (56) and (57), and the RMM confidence regions are built up from these intervals, see (60).

Algorithm 2: RMM Confidence Interval

Inputs: i.i.d. sample $\mathcal{D}_0$, rational significance level p, tie-breaking permutation π on $[m-1]_0$, odd median-of-means parameter k

1: Choose integers $1 \leq r < m$ such that $p = r/m$.

2: Randomly generate $n \cdot (m-1)$ independent Rademacher signs $\{\alpha_{i,j}\}$ for $(i,j) \in [n] \times [m-1]_0$.

3: For all $(\ell, j) \in [k] \times [m-1]$ and $s \in \{\pm 1\}$, compute the extended values

$$\nu^{\pm}_{\ell,j,s} = \frac{\widehat{\mu}(\mathcal{D}_0) \pm \frac{1}{\tilde{n}}\sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,j} Y_i}{1 \pm \frac{1}{\tilde{n}}\sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,j}}, \tag{53}$$

where $\frac{\pm c}{0} = \pm\infty$ for all $c > 0$ and $\frac{0}{0} = s \cdot \infty$.

4: For all $j \in [m-1]$ and $s \in \{\pm 1\}$, let

$$\nu^+_{j,s} \doteq \operatorname*{med}_{\ell\in[k]} \nu^+_{\ell,j,s}, \qquad \nu^-_{j,s} \doteq \operatorname*{med}_{\ell\in[k]} \nu^-_{\ell,j,s}. \tag{54}$$

5: For all $j \in [m-1]$, let

$$v^j \doteq [\nu^+_{j,-1}, \nu^+_{j,+1}, \nu^-_{j,-1}, \nu^-_{j,+1}] \tag{55}$$

6: For $j \in [m-1]$ if $\pi(j) > \pi(0)$, let

$$\lambda_j \doteq v^j_{(1)}, \qquad \varrho_j \doteq v^j_{(4)} \tag{56}$$

otherwise let

$$\lambda_j \doteq v^j_{(2)}, \qquad \varrho_j \doteq v^j_{(3)}, \tag{57}$$

where $v^j_{(1)}, v^j_{(2)}, v^j_{(3)}, v^j_{(4)}$ are ordered.

7: Let $\lambda_{(1)}, \ldots, \lambda_{(m-1)}$ and $\varrho_{(1)}, \ldots, \varrho_{(m-1)}$ be the order statistics w.r.t. $\prec_\pi$ and let

$$\lambda \doteq \lambda_{(r)}, \qquad \varrho \doteq \varrho_{(m-r)}. \tag{58}$$

8: Return

$$[\lambda, \varrho\,]_\pi \doteq (\lambda, \varrho) \cup \{\lambda : \pi(0) < \pi(\eta_1)\} \cup \{\varrho : \pi(0) < \pi(\eta_2)\}, \tag{59}$$

where $\eta_1, \eta_2 \in [m-1]$ are the original indices of $\lambda_{(r)}$ and $\varrho_{(m-r)}$, i.e., their indices before ordering.

Theorem 4.2. *For $m \geq 2$ we have*

$$\Theta_n = \bigcup_{\substack{J\subseteq[m-1],\, j\in J \\ |J|=m-r}} \bigcap \big\{\, \theta : S_0(\theta) \prec_\pi S_j(\theta) \,\big\}, \qquad (60)$$

thus Θ_n is an interval that contains $\widehat{\mu}(\mathcal{D}_0)$, if $\Theta_n \neq \emptyset$.

Based on these observations, we can prove the optimality of our method, more precisely, we have

Theorem 4.3. *Assume A1, A2 and A3. Let $r < m$ be user-chosen integers, $\delta > 0$, $k = \lceil 8\log(20(m-r)/\delta)\rceil$, then for $n \geq k(k + 8\log(k))$, we have*

$$\mathbb{P}\bigg(\operatorname{diam}(\Theta_n) > 8\bigg(\frac{12M^{1/a}\log\big(\frac{10(m-r)}{\delta}\big)}{n}\bigg)^{\frac{a}{1+a}}\bigg) \leq \delta. \qquad (61)$$

Proof. The proof relies on the union bound and it is essentially the same as the one used in the end of the proof of (Szentpéteri and Csáji, 2023, Theorem 3). □

If $a = 1$, an easy computation yields a similar result with better constants based on Theorem 2.1.

Theorem 4.4. *Assume A1, A2 and A3. Let $r < m$ be user-chosen integers, $\delta > 0$, $k = \lceil 8\log(10(m-r)/\delta)\rceil$, then for $n \geq k(k + 8\log(k))$, we have*

$$\mathbb{P}\bigg(\operatorname{diam}(\Theta_n) > \sigma\sqrt{\frac{32\log(10(m-r)/\delta)}{n}}\bigg) \leq \delta. \qquad (62)$$

5 Multivariate Hypothesis Test

In this section we assume that Y is a q-dimensional random vector and $\{Y^i\}_{i=1}^n$ is an i.i.d. sample, cf. A1. Henceforth, we denote the coordinates of Y with lower indices. Our assumptions are as follows.

A4. $Y - \mu \overset{d}{=} \mu - Y$, *for some vector* $\mu \in \mathbb{R}^q$.

A5. $\Sigma \doteq \mathbb{E}[(Y - \mathbb{E}Y)(Y - \mathbb{E}Y)^{\mathrm{T}}]$ *exists.*

We note that the components of Y can be correlated, and A4 does not require symmetry w.r.t. each coordinate axis. It is easy to see that from A4 and A5 it follows that $\mu = \mathbb{E}Y$. Let us denote the greatest eigenvalue of covariance matrix Σ by $\lambda^* = \lambda_{\max}(\Sigma)$.

It is known that for a zero mean multidimensional normal variable Z with $\operatorname{var}(Z) = \Sigma$, we have

$$\mathbb{P}(\|Z\| - \mathbb{E}\|Z\| \geq t\sqrt{n}) \leq \exp\bigg(-\frac{nt^2}{2\lambda^*}\bigg) \qquad (63)$$

for the Euclidean norm $\|\cdot\|$, see (Boucheron et al., 2013; Cirel'son et al., 2006; Lugosi and Mendelson, 2019a). Because of

$$\mathbb{E}\|Z\| \leq \sqrt{\mathbb{E}[\|Z\|^2]} = \sqrt{\operatorname{Tr}(\Sigma)}, \qquad (64)$$

if $\bar{\mu}_n$ denotes the empirical mean of a normally distributed i.i.d. sample $\{Z^i\}_{i=1}^n$, then for all $\delta \in (0,1)$

$$\mathbb{P}\bigg(\|\bar{\mu}_n - \mu\| > \sqrt{\frac{\operatorname{Tr}(\Sigma)}{n}} + \sqrt{\frac{2\lambda^*\log(1/\delta)}{n}}\bigg) \leq \delta. \qquad (65)$$

In general, it is hard to reach this high probability bound for an unknown distribution with covariance matrix Σ, e.g., the coordinate-wise median-of-means

$$\tilde{\mu}_j \doteq \widehat{\mu}(\{Y_j^i\}_{i=1}^n) \quad \text{for} \quad j = 1,\ldots,q, \qquad (66)$$

admits a much weaker PAC bound. For all $\delta \in (0,1)$ if $k = \lceil 8\log(1/\delta)\rceil$ only a dimension dependent inequality can be proved (Lugosi and Mendelson, 2019a), where $\log(1/\delta)$ is multiplied by the trace instead of the largest eigenvalue of Σ, i.e.,

$$\mathbb{P}\bigg(\|\tilde{\mu} - \mu\| \leq \sqrt{\frac{32\operatorname{Tr}(\Sigma)\log(q/\delta)}{n}}\bigg) \geq 1 - \delta. \qquad (67)$$

One may construct estimators with stronger guarantees. Several methods (such as the geometrical median-of-means (Hsu and Sabato, 2016; Minsker, 2015) and the median-of-means tournaments (Lugosi and Mendelson, 2019b)) are analyzed in the neat survey of Lugosi and Mendelson (2019a). The median-of-means tournament estimator and also its polynomial time relaxation are proved to be in the subgaussian regime (Hopkins, 2020; Lugosi and Mendelson, 2019b).

In this paper we use *any* subgaussian estimator $\hat{\mu}$, such as the median-of-means tournament estimator (Lugosi and Mendelson, 2019b), to define a *meta test* for the mean estimation problem in the multivariate setup. In particular, we assume that

A6. *For the expected value estimator $\hat{\mu}$ and any $\delta > 0$, there exist positive constants c_1, c_2 and c_3 such that*

$$\mathbb{P}\bigg(\|\hat{\mu} - \mu\| > \sqrt{\frac{c_1\operatorname{Tr}(\Sigma)}{n}} + \sqrt{\frac{c_2\lambda^*\log(c_3/\delta)}{n}}\bigg) \leq \delta. \qquad (68)$$

For $\theta \in \mathbb{R}^q$ we consider the nullhypothesis $H_0 : \mu = \theta$ and alternative $H_1 : \mu \neq \theta$. We generalize the RMM method to the multivariate case by constructing the alternative datasets of vectors for $j \in [m-1]$ as

$$\mathcal{D}_j(\theta) \doteq \{\alpha_{1,j}\mathbb{1}\odot(Y_1 - \theta) + \theta, \ldots, \alpha_{n,j}\mathbb{1}\odot(Y_n - \theta) + \theta\},$$

where $\odot$ denotes the Hadamard (element-wise) product. By introducing $\mathcal{D}_j^\pi \doteq (\mathcal{D}_j(\theta), \pi(j))$ for $j \in [m-1]_0$, and defining the reference variables as

$$S_j(\theta) \doteq \|\, \hat{\mu}(\mathcal{D}_j(\theta)) - \theta \,\| \qquad (69)$$

for $j = 0,\ldots,m-1$, the nullhypothesis is rejected as in Algorithm 1, i.e., if and only if $\mathcal{R}(\theta) > m - r$, where

$\mathcal{R}(\theta)$ is defined as in (17). A detailed description of this algorithm can be found in the appendix.

The exact confidence level of this test is stated under assumptions 0.A1 and A4 in the following theorem:

Theorem 5.1. *Assume A1 and A4, then the defined hypothesis test is exact, i.e., for every* $1 \leq r < m$

$$\mathbb{P}\left(\mathcal{R}(\mu) > m - r\right) = \frac{r}{m}. \tag{70}$$

Finite sample guarantees that follow can be proved for the type II error probability of the presented test.

Theorem 5.2. *Assume A1, A4, A5 and A6. Let* $\delta > 0$ *and* $r < m$ *be user-chosen integers. For* $\theta \neq \mu$ *if*

$$\sqrt{\frac{c_1(\mathrm{Tr}(\Sigma) + \Delta^2)}{n}} + \sqrt{\frac{c_2(\lambda^* + \Delta^2)\log\left(\frac{c_3(m-r)}{\delta}\right)}{n}} < \frac{\Delta}{2}, \tag{71}$$

holds for $\Delta = \|\theta - \mu\|$*, then we have*

$$\mathbb{P}(\mathcal{R}(\theta) > m - r) \geq 1 - \delta. \tag{72}$$

The complete proofs are presented in the supplements.

6 Numerical Experiment

In this section we present a numerical experiment comparing our theoretical bounds on the sizes of RMM confidence intervals, given by Theorem 4.3, with the theoretical confidence bound for the median-of-means estimator, given in (Lugosi and Mendelson, 2019a, Theorem 3), and with the empirical sizes of the RMM and SPS confidence regions. We consider a scalar mean estimation problem, where Y is sampled from a symmetrized Pareto distribution with scale parameter 1 and shape parameter 1.6. Throughout our experiments we considered 0.5-level confidence regions, in particular $m = 2$ and $r = 1$ were used, a sample size of $n = 12900$ and $s = 10000$ independently simulated trajectories. We performed the experiments with $\delta = 0.1$, $a = 0.5$ and set k according to Theorem 4.3.

In Figure 2 the $(1-\delta)$-quantiles of the empirical confidence interval sizes of the proposed RMM method and the SPS algorithm are compared with the theoretical bounds of (61) and (Lugosi and Mendelson, 2019a, Theorem 3) for each sample size.

Although the theoretical bounds are a bit conservative compared to the empirical sizes, it can be observed that the difference between the theoretical bounds are negligible and that our *data-driven* RMM algorithm has a better empirical sample complexity than using the results of (Lugosi and Mendelson, 2019a, Theorem 3). Our experiments are also indicative of the phenomenon that in case of heavy-tailed distributions, the RMM method outperforms the original SPS construction regarding sample complexity.

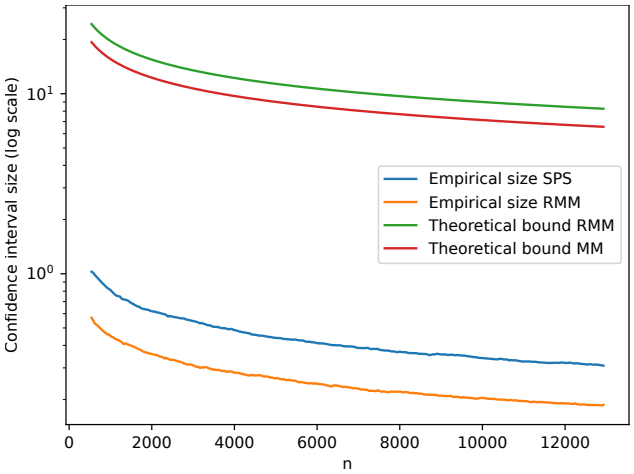


Figure 2: Comparison of confidence interval sizes.

7 Conclusions

Constructing a confidence region for the mean of a distribution is an essential and well-studied problem in statistics. However, standard approaches either build on asymptotic results or on the knowledge of some moments that might be harder to get than the mean.

In this paper the problem of estimating the mean of a symmetric, heavy-tailed distribution from an i.i.d. sample was addressed, and a new method, called the *resampled meadian-of-means* (RMM), was introduced. RMM is completely data-driven, it does not require additional a priori knowledge, e.g., information on moments. First the RMM based hypothesis test was presented, then the corresponding RMM confidence interval was constructed. It was shown that the RMM confidence interval has exact coverage probability for any finite sample size. Moreover, optimal PAC bounds were proved for the shrinkage of the RMM intervals. Finally, the construction was extended to the multivariate case and RMM was also evaluated empirically.

Acknowledgements

This research was supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory, and by the TKP2021-NKTA-01 grant of the National Research, Development and Innovation Office (NRDI), Hungary. The authors also acknowledge the support of the Doctoral Student Scholarship Program of the Cooperative Doctoral Program, financed from the National, Research, Development and Innovation Fund.

References

Noga Alon, Yossi Matias, and Mario Szegedy. The Space Complexity of Approximating the Frequency Moments. In *Proceedings of the 28th ACM Symposium on Theory of Computing*, pages 20–29, 1996.

George Bennett. Probability Inequalities for the Sum of Independent Random Variables. *Journal of the American Statistical Association*, 57:33–45, 1962.

Sergei N. Bernstein. On Certain Modifications of Chebyshev's Inequality. *Doklady Akademii Nauk SSSR*, 16(6):275–277, 1937.

Stephane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, 2013.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with Heavy Tail. *IEEE Transactions on Information Theory*, 59:7711–7717, 2013.

Marco C. Campi and Erik Weyer. Guaranteed Non-Asymptotic Confidence Regions in System Identification. *Automatica*, 41(10):1751–1764, 2005.

Olivier Catoni. Challenging the Empirical Mean and Empirical Variance: A Deviation Study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185, 2012.

Boris S Cirel'son, Ildar A Ibragimov, and Vladimir N. Sudakov. Norms of Gaussian Sample Functions. In *Proceedings of the Third Japan—USSR Symposium on Probability Theory*, pages 20–41. Springer, 2006.

Balázs Cs. Csáji, Marco Claudio Campi, and Erik Weyer. Sign-Perturbed Sums: A New System Identification Approach for Constructing Exact Non–Asymptotic Confidence Regions in Linear Regression models. *IEEE Transactions on Signal Processing*, 63(1):169–181, 2015.

Balázs Cs. Csáji and Ambrus Tamás. Semi-Parametric Uncertainty Bounds for Binary Classification. In *Proceedings of the 58th IEEE Conference on Decision and Control (CDC), Nice, France*, pages 4427–4432, 2019.

Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. Sub-Gaussian Mean Estimators. *Annals of Statistics*, 44(6):2695–2725, 2016.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, 2011.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2009.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Samuel B. Hopkins. Mean Estimation with Sub-Gaussian Rates in Polynomial Time. *Annals of Statistics*, 48:1193–1213, 2020.

Daniel Hsu and Sivan Sabato. Loss Minimization and Parameter Estimation with Heavy Tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. John Wiley & Sons, 2009.

Gábor Lugosi and Shahar Mendelson. Mean Estimation and Regression Under Heavy-Tailed Distributions: A Survey. *Foundations of Computational Mathematics*, 19:1145–1190, 2019a.

Gábor Lugosi and Shahar Mendelson. Sub-Gaussian Estimators of the Mean of a Random Vector. *Annals of Statistics*, 47:783–794, 2019b.

Stanislav Minsker. Geometric Median and Robust Estimation in Banach Spaces. *Bernoulli*, 21(4):2308–2335, 2015.

Arkadij Semenovič Nemirovsky and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, 1983.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Szabolcs Szentpéteri and Balázs Cs. Csáji. Sample Complexity of the Sign-Perturbed Sums Identification Method: Scalar Case. *IFAC World Congress, IFAC PapersOnLine*, 56:10363–10370, 2023.

Vladimir Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Ian Waudby-Smith and Aaditya Ramdas. Estimating Means of Bounded Random Variables by Betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

Supplementary Materials for "Data-Driven Confidence Intervals with Optimal Rates for the Mean of Heavy-Tailed Distributions"

These supplementary materials contain the proofs which were left out from the main paper due to lack of space. Some additional details for the main proofs are also included. The time and space complexity of the confidence interval construction is studied, as well, and a pseudocode is given for the multivariate version of the RMM test. Finally, additional experiments are presented comparing the sizes of the RMM and SPS confidence intervals to nonparametric tests such as the sign test and Wilcoxon signed-rank test and illustrating the distribution of the confidence interval endpoints, and the empirical coverage probabilities of the tests based on the asymptotic theory, SPS and RMM.

1 Supplementary Proofs

This section contains the proofs of Theorems 3.4, 3.6, 3.7, and 5.2; as well as additional details for Theorem 3.5.

Proof of Theorem 3.4.

In this section, we present the detailed proof of Theorem 3.4.

Let us fix $\theta \neq \mu$, integers $1 \leq r < m$ and use notation $\widehat{\mu}^j \doteq \widehat{\mu}(\mathcal{D}_j(\theta))$ for $j = 0, \ldots, m-1$. The following lower bound holds

$$\mathbb{P}(\mathcal{R}(\theta) > m - r) \geq \mathbb{P}\big(|\widehat{\mu}^0 - \mu| < d/2, |\widehat{\mu}^j - \theta| < d/2 \text{ for } j = 1, \ldots, m-r \big), \tag{1}$$

because from $|\widehat{\mu}^0 - \mu| < d/2$ it follows that $|\widehat{\mu}^0 - \theta| \geq d/2$. Let us define events

$$\begin{aligned} B_0 &\doteq \{\, |\, \widehat{\mu} - \mu \,| < d/2 \,\}, \\ B_j &\doteq \{\, |\, \widehat{\mu}^j - \theta \,| < d/2 \,\} \quad \text{for } \; j = 1, \ldots, m-1. \end{aligned} \tag{2}$$

By the union bound we have

$$\mathbb{P}\Big(\bigcup_{j=0}^{m-r} \bar{B}_j \Big) \leq \sum_{j=0}^{m-r} \mathbb{P}(\bar{B}_j) \leq \mathbb{P}\big(|\widehat{\mu}^0 - \mu| \geq {}^d\!/_2 \big) + \sum_{j=1}^{m-r} \mathbb{P}\big(|\widehat{\mu}^j - \theta| \geq {}^d\!/_2 \big). \tag{3}$$

and henceforth if

$$4(\sigma^2 + d^2) \frac{32 \log((m-r+1)/\delta)}{n} < d^2, \tag{4}$$

then we have

$$\mathbb{P}\big(|\widehat{\mu}^j - \theta| \geq {}^d\!/_2 \big) \leq \mathbb{P}\left(|\widehat{\mu}^j - \theta| > \sqrt{\frac{(\sigma^2 + d^2) 32 \log((m-r+1)/\delta)}{n}} \right) \leq \frac{\delta}{m-r+1}. \tag{5}$$

In addition because of

$$4\sigma^2 \frac{32 \log((m-r+1)/\delta)}{n} \leq 4(\sigma^2 + d^2) \frac{32 \log((m-r+1)/\delta)}{n} < d^2, \tag{6}$$

we also have

$$\mathbb{P}\big(|\widehat{\mu}^0 - \theta| \geq {}^d\!/_2 \big) \leq \mathbb{P}\left(|\widehat{\mu}^j - \theta| > \sigma \sqrt{\frac{32 \log((m-r+1)/\delta)}{n}} \right) \leq \frac{\delta}{m-r+1}. \tag{7}$$

In conclusion

$$\mathbb{P}\big(\mathcal{R}(\theta) > m - r \big) \geq \mathbb{P}\Big(\bigcap_{j=0}^{m-r} B_j \Big) \geq 1 - \sum_{j=0}^{m-1} \mathbb{P}(\bar{B}_j) \geq 1 - \delta. \tag{8}$$

Note that for any fixed θ for n large enough (4) is satisfied. The dependence on δ is logarithmic.

Equality 45 for Theorem 3.5

$$\begin{aligned}\nu^+ &= \operatorname*{med}_{\ell\in[k]} \nu_\ell^+ = \operatorname*{med}_{\ell\in[k]} \left(\frac{\widehat{\mu}(\mathcal{D}_0) + \frac{1}{\tilde{n}} \sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1} Y_i}{1 + \frac{1}{\tilde{n}} \sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1}} \right) = \operatorname*{med}_{\ell\in[k]} \left(\mu + \frac{\widehat{\mu}(\mathcal{W}_0) + \frac{1}{\tilde{n}} \sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1} W_i}{\frac{1}{\tilde{n}} \sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} (1+\alpha_{i,1})} \right) \\ &= \mu + \operatorname*{med}_{\ell\in[k]} \left(\frac{\widehat{\mu}(\mathcal{W}_0) + \frac{1}{\tilde{n}} \sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1} W_i}{\frac{2}{\tilde{n}}(\tilde{n} - Z_\ell)} \right) = \mu + V^+.\end{aligned} \tag{9}$$

Inequality 51 for Theorem 3.5.

$$\begin{aligned}&\mathbb{P}(B \cap A) \le \mathbb{P}(\tilde{B} \cap A) \le \mathbb{P}\Big(\operatorname*{med}_{\ell\in[k]} \Big(\widehat{\mu}(\mathcal{W}_0) - \frac{1}{\tilde{n}} \sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1} W_i \Big) > \varepsilon/4 \Big) \\ &= \mathbb{P}\Big(\widehat{\mu}(\mathcal{W}_0) - \operatorname*{med}_{\ell\in[k]} \Big(\frac{1}{\tilde{n}} \sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1} W_i \Big) > \varepsilon/4 \Big) \le \mathbb{P}\Big(\{\widehat{\mu}(\mathcal{W}_0) > \varepsilon/8\} \cup \Big\{ - \operatorname*{med}_{\ell\in[k]} \frac{1}{\tilde{n}} \sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} \alpha_{i,1} W_i > \varepsilon/8 \Big\} \Big) \\ &\le \mathbb{P}\big(\widehat{\mu}(\mathcal{W}_0) > \varepsilon/8\big) + \mathbb{P}\Big(\operatorname*{med}_{\ell\in[k]} \frac{1}{\tilde{n}} \sum_{i=(\ell-1)\tilde{n}}^{\ell\tilde{n}} -\alpha_{i,1} W_i > \varepsilon/8 \Big) \le 2 \cdot \mathbb{P}\big(\widehat{\mu}(\mathcal{W}_0) > \varepsilon/8\big).\end{aligned} \tag{10}$$

Proof of Theorem 3.6.

Now, we prove one of our key observations, which shows that Θ_n is in fact an interval.

Equation (53) follows from the observation that

$$\begin{aligned}&\big\{\, \theta : S_0(\theta) \prec_\pi S_j(\theta) \,\big\} = \big\{\, \theta : \mathbb{I}(S_0(\theta) \prec_j S_j(\theta)) = 1 \,\big\} \quad \text{and} \\ &\bigcup_{\substack{J\subseteq[m-1],\, j\in J \\ |J|=m-r}} \bigcap \big\{\, \theta : S_0(\theta) \prec_\pi S_j(\theta) \,\big\} = \big\{\, \theta : \mathcal{R}(\theta) \le m-r \,\big\}.\end{aligned} \tag{11}$$

We show that Θ_n is an interval in three steps. First, we prove that $I_j \doteq \{\theta : S_0(\theta) \prec_\pi S_j(\theta)\}$ is an interval for any $j \in [m-1]$. Second, we show that if I_j is nonempty, then $\widehat{\mu} = \widehat{\mu}(\mathcal{D}_0)$ is included. Finally, we use (11) to conclude that Θ_n is an interval, because it is the union of intervals that have a common element or empty.

Clearly $\widehat{\mu}$ is a parameter that should be included in I_j, because $S_0(\widehat{\mu}) = 0$. More precisely if $\nu \in I_j$ and $\nu > \widehat{\mu}$, then we prove that $[\widehat{\mu}, \nu] \subseteq I_j$ (similarly for $\nu < \widehat{\mu}$ if $\nu \in I_j$, then we have $[\nu, \widehat{\mu}] \subseteq I_j$). Assume by contradiction that there exists $\bar{\theta} \in [\widehat{\mu}, \nu]$ such that $S_j(\bar{\theta}) \prec_\pi S_0(\bar{\theta})$. We know that S_j is a piecewise linear function, of which slopes are included in $[-1, 1]$. We have $S_j(\theta) \le \theta + S_j(\bar{\theta}) - \bar{\theta}$ for all $\theta > \bar{\theta}$, because the left derivative of S_j is at most 1. In addition for all $\theta < \nu$ we have $S_j(\theta) \ge \theta + S_j(\nu) - \nu$. If $\pi(0) < \pi(j)$, then $S_0(\nu) \prec_\pi S_0(\nu)$ implies $S_0(\nu) \le S_j(\nu)$ and $S_j(\bar{\theta}) \prec_\pi S_0(\bar{\theta})$ implies $S_j(\bar{\theta}) < S_0(\bar{\theta})$. Thus, the contradiction follows by

$$-\widehat{\mu} = \nu - \widehat{\mu} - \nu = S_0(\nu) - \nu \le S_j(\nu) - \nu \le S_j(\theta) - \theta \le S_j(\bar{\theta}) - \bar{\theta} < S_0(\bar{\theta}) - \bar{\theta} = \bar{\theta} - \widehat{\mu} - \bar{\theta} = -\widehat{\mu}. \tag{12}$$

One can proceed similarly for $\pi(0) > \pi(j)$. In conclusion we proved that I_j is an interval and if there is at least one element in I_j, then $\widehat{\mu}$ is also included. Henceforth, $\cap_{j\in J} I_j$ is also an interval for all $J \subseteq [m-1]$, which includes $\widehat{\mu}$ or is empty. Finally, the same can be said about the union of these.

Proof of Theorem 3.7.

For the sake of completeness we present the proof of Theorem 3.7. It highly relies on Theorem 1.1 which can be proved based on Theorem 3.5.

Theorem 1.1. *Assume A1, A2 and A3. Let $m = 2$, $\delta > 0$, $k = \lceil 8 \log(10/\delta) \rceil$, let $\mathcal{R}(\theta)$ be the rank of $S_0(\theta)$ among $\{S_0(\theta), S_1(\theta)\}$ w.r.t. $\succ_\pi$, and $\Theta_n \doteq \{\, \theta : \mathcal{R}(\theta) \le 1 \,\}$, then for $n \ge k(k + 8\log(k))$, we have*

$$\mathbb{P}\Big(\operatorname{diam}(\Theta_n) > \sigma \sqrt{\frac{32 \log(10/\delta)}{n}} \Big) \le \delta. \tag{13}$$

The only difference w.r.t. the proof of Theorem 3.5. is the bound on the probability of $\{\widehat{\mu}(\mathcal{W}) > \varepsilon/8\}$. We use $k = \lceil 8\log(1/\tilde{\delta})\rceil$ and $\varepsilon = 8\sigma\sqrt{\frac{32\log(1/\tilde{\delta})}{n}}$. Then for $n \geq k(8\log(k)+k)$ we have

$$\mathbb{P}\Big(\operatorname{diam}(\Theta_n) > \sigma\sqrt{\frac{32\log(1/\tilde{\delta})}{n}} \Big) \leq 2k\exp\Big(-\frac{n}{8k}\Big) + 8\tilde{\delta} \leq 10\tilde{\delta}. \tag{14}$$

Setting $\delta \doteq 10\tilde{\delta}$ completes the proof.

An auxiliary lemma for Theorem 5.2.

In this section, we present an auxiliary lemma for Theorem 4.2 and its detailed proof.

Lemma 1.1. *Let Y be a random vector in $\mathbb{R}^q$ such that $\mathbb{E}Y = \mu$, $\operatorname{var}(Y) = \Sigma_Y$ and $\lambda^* = \lambda_{\max}(\Sigma_Y)$. Let $\theta \in \mathbb{R}^q$ and α be a Rademacher variable independent of Y, then for $Z = \alpha\mathbb{1}\odot(Y-\theta)+\theta$, we have $\mathbb{E}Z = \theta$ and*

$$\operatorname{Tr}(\operatorname{var}(Z)) = \operatorname{Tr}(\Sigma_Y) + \|\mu-\theta\|^2, \tag{15}$$

$$\lambda_{\max}(\operatorname{var}(Z)) \leq \lambda_{\max}(\Sigma_Y) + \|\mu-\theta\|^2. \tag{16}$$

Proof. Because of independence we have $\mathbb{E}[\alpha\odot(Y-\theta)+\theta] = \mathbb{E}[\alpha]\odot\mathbb{E}[(Y-\theta)]+\theta = \theta$. We can show (15) by

$$\begin{aligned}\operatorname{Tr}(\operatorname{var}(Z)) &= \operatorname{Tr}(\mathbb{E}[(\alpha\mathbb{1}\odot(Y-\theta)+\theta-\mathbb{E}Z)(\alpha\mathbb{1}\odot(Y-\theta)+\theta-\mathbb{E}Z)^{\mathrm{T}}]) \\ &= \operatorname{Tr}(\mathbb{E}[(\alpha\mathbb{1}\odot(Y-\theta))(\alpha\mathbb{1}\odot(Y-\theta))^{\mathrm{T}}]) = \operatorname{Tr}(\mathbb{E}[(Y-\mu+\mu-\theta)(Y-\mu+\mu-\theta)^{\mathrm{T}}]) \\ &= \operatorname{Tr}(\Sigma_Y + (\mu-\theta)(\mu-\theta)^{\mathrm{T}}) = \operatorname{Tr}(\Sigma_Y) + \|\mu-\theta\|^2,\end{aligned} \tag{17}$$

where we also proved that $\operatorname{var}(Z) = \Sigma_Y + (\mu-\theta)(\mu-\theta)^{\mathrm{T}}$. For (16) we use the well-known fact that for a symmetric real matrix A we have

$$\lambda_{\max}(A) = \max_{x:\|x\|=1} x^{\mathrm{T}}Ax. \tag{18}$$

Therefore

$$\begin{aligned}\lambda_{\max}(\operatorname{var}(Z)) &= \max_{x:\|x\|=1} x^{\mathrm{T}}\operatorname{var}(Z)x = \max_{x:\|x\|=1} x^{\mathrm{T}}(\Sigma_Y + (\mu-\theta)(\mu-\theta)^{\mathrm{T}})x \\ &= \max_{x:\|x\|=1}\big(x^{\mathrm{T}}\Sigma_Y x + x^{\mathrm{T}}(\mu-\theta)(\mu-\theta)^{\mathrm{T}}x\big) \leq \lambda^* + \|\mu-\theta\|^2,\end{aligned} \tag{19}$$

where we used that the largest eigenvalue of vv^{T} equals to $\|v\|^2$. □

Proof of Theorem 5.2

In this section, we present the detailed proof of Theorem 5.2.

Let us fix $\theta \neq \mu$, integers $1 \leq r \leq m$ and use notation $\hat{\mu}^j \doteq \hat{\mu}(\mathcal{D}_j(\theta))$ for $j = 0,\ldots,m-1$. The following lower bound holds

$$\mathbb{P}(\mathcal{R}(\theta) > m-r) \geq \mathbb{P}\big(\,\|\hat{\mu}^0-\mu\| < \Delta/2, \|\hat{\mu}^j-\theta\| < \Delta/2 \text{ for } j = 1,\ldots,m-r\big), \tag{20}$$

because from $\|\hat{\mu}^0-\mu\| < \Delta/2$ it follows that $\|\hat{\mu}^0-\theta\| \geq \Delta/2$. Let us define events

$$\begin{aligned}&B_0 \doteq \{\|\hat{\mu}^0-\mu\| < \Delta/2\}, \\ &B_j \doteq \{\|\hat{\mu}^j-\theta\| < \Delta/2\} \text{ for } j = 1,\ldots,m-1.\end{aligned} \tag{21}$$

By the union bound we have

$$\mathbb{P}\Big(\bigcup_{j=0}^{m-r}\bar{B}_j\Big) \leq \sum_{j=0}^{m-r}\mathbb{P}(\bar{B}_j) \leq \quad \mathbb{P}\big(\,\|\hat{\mu}^0-\mu\| \geq \Delta/2\big) + \sum_{j=1}^{m-r}\mathbb{P}\big(\,\|\hat{\mu}^j-\theta\| \geq \Delta/2\big) \tag{22}$$

Because of A6 the first term can be bounded by

$$\mathbb{P}\big(\,\|\hat{\mu}^0-\mu\| \geq \Delta/2\big) \leq \mathbb{P}\Bigg(\|\hat{\mu}^0-\mu\| > \sqrt{\frac{c_1\operatorname{Tr}(\Sigma)}{n}} + \sqrt{\frac{c_2\lambda^*\log\left(\frac{c_3(m-r)}{\delta}\right)}{n}}\Bigg) \leq \tilde{\delta}, \tag{23}$$

where $\delta = (m-r+1)\tilde{\delta}$. The other terms can be bounded similarly by

$$\mathbb{P}\big(\,\|\hat{\hat{\mu}}^j - \theta\| \geq \triangle/2\,\big) \leq \mathbb{P}\left(\|\hat{\hat{\mu}}^j - \theta\| > \sqrt{\frac{c_1(\mathrm{Tr}(\Sigma)+\Delta^2)}{n}} + \sqrt{\frac{c_2(\lambda^*+\Delta^2)\log\left(\frac{c_3(m-r)}{\delta}\right)}{n}}\right) \leq \tilde{\delta}. \tag{24}$$

In conclusion under the required conditions we have

$$\mathbb{P}(\mathcal{R}(\theta) \leq m-r) \leq (m-r+1)\tilde{\delta} = \delta, \tag{25}$$

which proves the desired bound.

2 Space and Time Complexities of the Confidence Interval Construction

Algorithm 2 is computationally light. The time complexity of computing the confidence interval is in $\mathcal{O}(m \cdot n)$ flops, i.e., one can find the endpoints in linear time w.r.t. $m \cdot n$. Generating the random signs requires $(m-1)\cdot n$ steps. Then each $\alpha_{i,j}$ is used a constant times to compute $\nu^{\pm}_{\ell,j}$ for all $\ell \in [k]$ and $j \in [m-1]$ and each Y_i is used m times. Computing the median of k elements is in $\mathcal{O}(k)$, see (Blum et al., 1973), which is much smaller than $\mathcal{O}(n)$. Finally computing the quantiles of $m-1$ elements is also in $\mathcal{O}(m)$. The space complexity of the algorithm is trivially in $\mathcal{O}(m\cdot n)$, however, one does not need to memorize each $\{\alpha_{i,j}\}$ thus it can be reduced to $\mathcal{O}(\max(m,n))$. One can also reduce the space complexity by using recursive variants of median-of-means estimates. Finally we also observe that our method can be easily parallelized, because several parts can be computed simultaneously.

3 Multivariate RMM Test Algorithm

In this section the detailed description of the multivariate RMM test is presented.

Algorithm 1: Multivariate RMM Hypothesis Test (for $\mu = \theta$)

Inputs: i.i.d. sample $\mathcal{D}_0$, rational significance level p,
tie-breaking permutation π on $[m-1]_0$,
subgaussian estimator $\hat{\hat{\mu}}$

1: Choose integers $1 \leq r < m$ such that $p = r/m$.

2: Generate $n(m-1)$ independent Rademacher sign variables $\{\alpha_{i,j}\}$ for $(i,j) \in [n] \times [m-1]_0$.

3: Construct $m-1$ alternative datasets for $j \in [m-1]$:

$$\mathcal{D}_j(\theta) \doteq \{\alpha_{1,j}\mathbb{1} \odot (Y_1 - \theta) + \theta, \ldots, \alpha_{n,j}\mathbb{1} \odot (Y_n - \theta) + \theta\}$$

and let $\mathcal{D}^{\pi}_j \doteq (\mathcal{D}_j(\theta), \pi(j))$ for $j \in [m-1]_0$.

4: Compute the reference variables

$$S_j(\theta) \doteq \|\hat{\hat{\mu}}(\mathcal{D}_j(\theta)) - \theta\| \quad \text{for} \quad j = 0, \ldots, m-1.$$

5: Compute the rank

$$\mathcal{R}(\theta) = 1 + \sum_{j=1}^{m-1} \mathbb{I}\left(\, S_0(\theta) \succ_{\pi} S_j(\theta)\,\right).$$

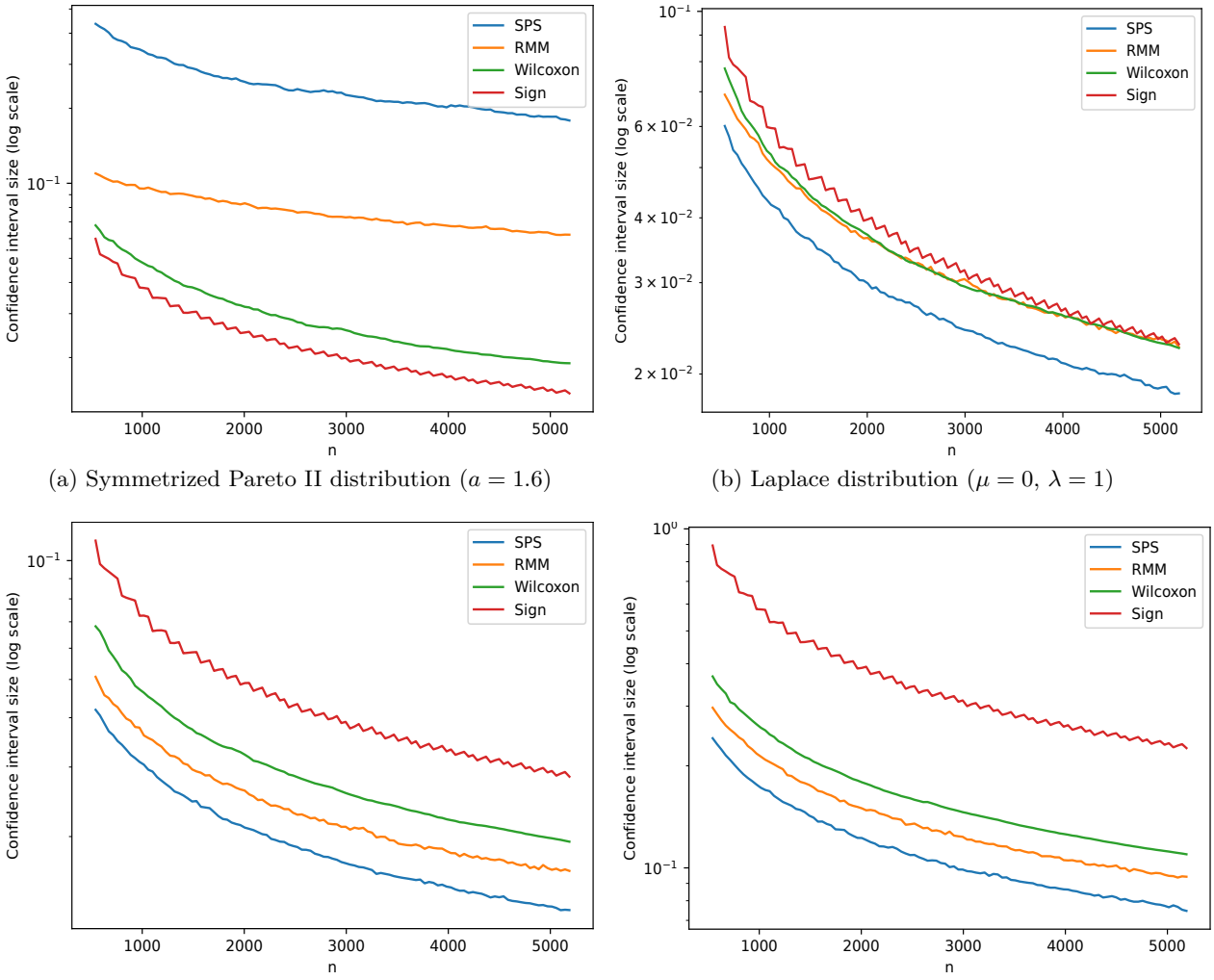
6: Reject H_0 if and only if

$$\mathcal{R}(\theta) > m - r.$$

4 Supplementary Numerical Experiments

4.1 Confidence interval size experiment

In this experiment we compare the empirical sizes of several confidence regions for different sample distributions. We consider the mean estimation problems, where Y is sampled from a *(a)* symmetrized Pareto II distribution



(a) Symmetrized Pareto II distribution ($a = 1.6$)

(b) Laplace distribution ($\mu = 0$, $\lambda = 1$)

(c) Standard normal distribution

(d) Uniform distribution on $(-10, 10)$

Figure 1: Comparison of confidence interval sizes

with shape parameter $a = 1.6$ and scale parameter 1, *(b)* zero-mean Laplace distribution with scale parameter $\lambda = 1$, *(c)* standard normal distribution and *(d)* uniform distribution on interval $(-10, 10)$. The sizes of the RMM and SPS confidence regions are compared to the sizes of the sign test-based and Wilcoxon signed-rank test-based confidence intervals (Pratt and Gibbons, 1981). For the Wilcoxon signed-rank test we used the statistical package of R with approximate confidence levels. Throughout our experiments we considered 0.5-level confidence regions, with $m = 20$ and $r = 10$ in case of the RMM and SPS, a sample size of $n = 5000$ and $s = 10000$ independently simulated trajectories. For the Wilcoxon signed-rank test we used fewer independent simulations ($s' = 100$), because constructing the confidence intervals based on this test is computationally intensive and there was a very small variance regarding the sizes. The difference between the 0.9-quantiles of the confidence interval sizes for the different sample distributions are shown in Figure 1. It can be observed that in case of the heavy-tailed Pareto II distribution the sizes of the sign test and Wilcoxon signed-rank test-based confidence regions are smaller than that of RMM and SPS. Recall that despite their smaller sizes, these order-based tests have no guarantees for their shrinkage rates, also the Wilcoxon signed-rank test has high computational burden and only has approximate guarantees for the coverage probability. These experiments also indicate that as the sample distribution becomes less heavy-tailed, the RMM and SPS methods build confidence intervals with smaller sizes than the sign and Wilcoxon signed-rank tests. We can conclude that the RMM method can be a robust solution, because it achieves a significant improvement for heavy-tailed distributions w.r.t. the SPS method and is more efficient than the tested nonparametric methods for light-tailed distributions.

4.2 Experiment: Distribution of Intersections

An empirical comparison between the SPS (Sign-Perturbed Sums) and the RMM (Resampled Median-of-Means) methods for sample distributions with different tail probabilities are presented here. In these simulations the empirical distributions of the random variables that determine the confidence intervals (the end point of the intervals) are investigated. In case of the SPS algorithm these points can be computed as in (Szentpéteri and Csáji, 2023), while in case of the RMM method these points are ν^+ and ν^-. In the experiments the observations were (symmetrically) Pareto-distributed with scale parameter 1 and shape parameter β. We set $k = 5$, considered a sample size of $n = 125$ and $s = 100000$ independently generated end points. Our results are shown for different β parameters in Figure 2. The empirical distributions indicate that for heavy-tailed noises the end points of the intersections (that determines the confidence interval) are more concentrated about 0 in case of the RMM method, while for not heavy-tailed noises they are more concentrated in case of the the SPS method.

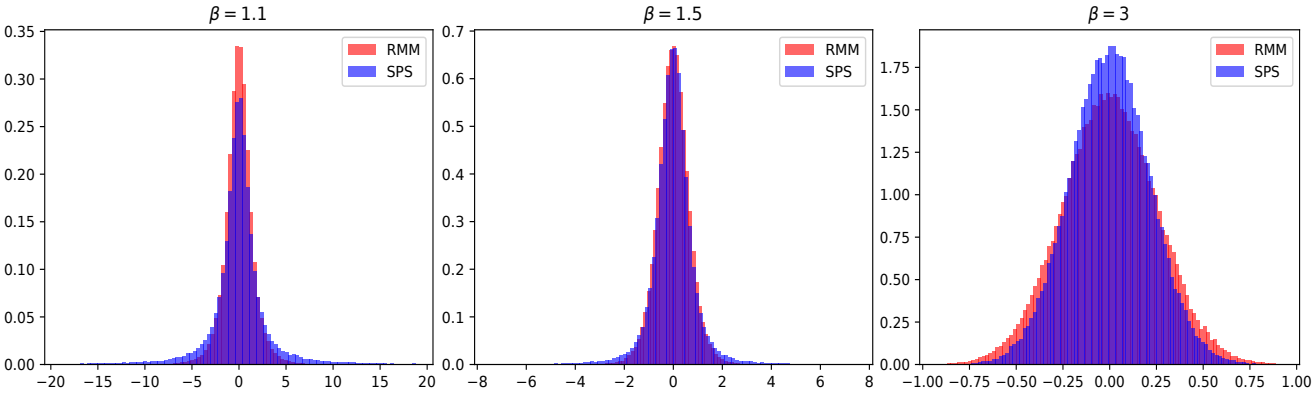


Figure 2: Comparison of confidence interval end point distributions $n = 125, k = 5, s = 100000$.

4.3 Experiment: Coverage Probability

We studied the empirical coverage probabilities of confidence intervals based on the asymptotic theory, the SPS algorithm and the RMM method. In this experiment we considered $n = 30$, $k = 3$ and $s = 100000$ independently generated confidence intervals and computed the empirical probabilities $\hat{p}$ that the true parameter is in the region. The observations were (symmetrically) Pareto-distributed with scale parameter 1 and shape parameter β as before. Our results are summarized in Table 1. We can observe that the confidence levels are always exact for the RMM and SPS methods, however, the asymptotic confidence intervals perform poorly. On the one hand for $\beta \leq 2$ the true variance is infinite, thus the asymptotic intervals constructed based on the sample variance are too large. On the other hand for $\beta = 3$ and $n = 30$ the limit distribution is a poor approximation of the distribution of the mean and the asymptotic method underestimates the uncertainty of the sample mean.

Table 1: Empirical coverage probabilities for $n = 30$, $k = 3$ and $s = 100000$.

	$\hat{p}_A$	$\hat{p}_{SPS}$	$\hat{p}_{RMM}$
$\beta = 1.1$	0.999	0.899	0.898
$\beta = 1.5$	0.987	0.901	0.901
$\beta = 2$	0.938	0.901	0.900
$\beta = 3$	0.748	0.901	0.900

References

Manuel Blum, Robert W. Floyd, Vaughan R. Pratt, Ronald L. Rivest, and Robert Endre Tarjan. Time Bounds for Selection. *Journal of Computer and System Sciences*, 7(4):448–461, 1973.

John Winsor Pratt and Jean Dickinson Gibbons. *Concepts of Nonparametric Theory*. Springer-Verlag, 1981.

Szabolcs Szentpéteri and Balázs Cs. Csáji. Sample Complexity of the Sign-Perturbed Sums Identification Method: Scalar Case. *IFAC World Congress, IFAC PapersOnLine*, 56:10363–10370, 2023.