# Phase Transitions in Heavy-Tailed Mean Estimation under $\ell_p$ Norms

**Ishaq Aden-Ali**
UC Berkeley
adenali@berkeley.edu

**Yeshwanth Cherapanamjeri**
Nanyang Technological University
yesh.c@ntu.edu.sg

**Mikael Møller Høgsgaard**
University of Oxford
hogsgaard@cs.au.dk

**Kasper Green Larsen**
Aarhus University
larsen@cs.au.dk

**Nikita Zhivotovskiy**
UC Berkeley
zhivotovskiy@berkeley.edu

## Abstract

We study the problem of estimating the mean of a random vector in $\mathbb{R}^d$ under $\ell_p$ norms, assuming only the existence of a covariance matrix. In the Euclidean norm, the sample mean is already optimal in expectation, and robust estimators are usually needed only to obtain high-probability tails. Our results show that, even under identity covariance, this picture changes for $\ell_p$ norms with $p > 2$: robustification is already necessary even for the expected risk. We prove that the minimax risk exhibits a sharp phase transition, controlled by the interaction between the sample size $n$, the dimension $d$, and the norm parameter $p$. In the large-sample regime, the classical Gaussian rate remains achievable. In the complementary regime, heavy-tailed distributions create a genuinely non-Gaussian obstruction. Furthermore we show that the simple estimator: coordinate-wise median of means whose number of buckets chosen according to the geometry of the loss, optimally tracks this phase transition in its expected risk minimax. We also show that natural projection-based median-of-means estimators for general norms, despite achieving the optimal confidence trade-off relative to the sample-mean benchmark of Lugosi and Mendelson (*Probab. Theory Relat. Fields*, 2019), can be polynomially suboptimal for $\ell_p$ norms even in the constant-probability regime.

## 1 Introduction

Mean estimation, or location estimation, is one of the foundational subroutines in machine learning and statistics. A modern line of work focuses on statistical estimators that are robust to heavy-tailed distributions. Starting with the systematic study pioneered by Catoni [Cat12], this line of work seeks to obtain non-asymptotic *sub-Gaussian* rates for estimating the means of random variables, usually using estimators such as the median of means, which appeared independently in [NY83, JVV86, AMS99], Catoni's estimator [Cat12], or trimmed-mean estimators [LM21]. As we recall later, more involved versions of these estimators also work well in high-dimensional regimes, where the primary focus has traditionally been on the Euclidean norm.

However, the Euclidean norm is arguably the most benign norm to work with. Indeed, given $n$ independent vectors $X_1, \dots, X_n$ in $\mathbb{R}^d$ with common mean $\mu$ and identity covariance $I_d$, the sample mean is essentially optimal, up to universal constant factors for all $n$ and $d$:

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right\|_2 \leq \sqrt{\frac{d}{n}}, \tag{1}$$

which follows from Cauchy–Schwarz and does not require the existence of any moments above the second moment, while the lower bound follows from Anderson's lemma [And55] as discussed below.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Assuming only the existence of the covariance, and no higher moments, the main focus has therefore been on obtaining sub-Gaussian estimators, namely estimators with optimal sub-Gaussian type tails, where the sample mean is clearly the wrong object [Min15, LM19b, CG17, LM21].

When going beyond the Euclidean norm, while still keeping only the bounded covariance assumption, even the in expectation story of (1) becomes more delicate especially in the high-dimensional regime where the dimension $d$ is larger than the sample size $n$. In non-Hilbertian geometries, the vanilla sample mean is no longer necessarily a good estimator.[1] Our key contribution is to show that the Euclidean picture is special even for the simple mean estimation problem in $\mathbb{R}^d$. For $\ell_p$ norms with $p > 2$, robustification can be necessary even for *in expectation* bounds of the form (1). This phenomenon does not appear in the regime $1 \le p \le 2$, where the sample mean remains optimal up to universal constants, but it becomes crucial once the geometry of the loss is sufficiently far from $\ell_2$. As we discuss later the suboptimal of the sample mean for general norms is relevant to deviation bounds in general norms, since some of these bounds contain this quantity (see "Phase transition and the existing general norm results.").

**Theorem 1.1** (Phase transition for location estimation in $\ell_p$, informal)**.** *Under only the covariance assumption* $\mathrm{Cov}_P(X) = I_d$*, the optimal expected error for estimating the mean in* $\ell_p$ *satisfies*

$$\inf_{\widehat{\mu}} \sup_{\substack{\mu\in\mathbb{R}^d \\ P:\,\mathbb{E}_P[X]=\mu \\ \mathrm{Cov}_P(X)=I_d}} \mathbb{E}\,\|\widehat{\mu}-\mu\|_p \asymp \begin{cases} d^{\frac{1}{2\lceil n/2\rceil}}, & \bar{p} > 2\lceil n/2\rceil, \\ d^{\frac{1}{\bar{p}}}\sqrt{\frac{\bar{p}}{n}}, & \bar{p} \le 2\lceil n/2\rceil, \end{cases}$$

*where* $\hat{\mu}$ *ranges over all estimators based on the* $n$ *samples and*

$$\bar{p} = \min\{p,\, \max\{\log d,\, 1\}\}.$$

Thus, the minimax risk undergoes a sharp phase transition. In the large sample regime, the rate has the familiar Gaussian form, corresponding to the Gaussian width of the dual ball inducing the norm, as we discuss below:

$$d^{1/\bar{p}}\sqrt{\frac{\bar{p}}{n}}.$$

In contrast, in the small sample or large $p$ regime, the effective exponent is capped by the number of samples, and the optimal rate becomes

$$d^{\frac{1}{2\lceil n/2\rceil}}.$$

We show that a estimator achieving this rate is a simple coordinate-wise median of means estimator, with the choice of number of buckets depending on the geometry of the loss. Recall that median of means first splits the sample into nearly equal buckets, computes the average inside each bucket, and then takes a median of these bucket averages. In the coordinate-wise version used here, if $\bar{X}^{(1)}, \dots, \bar{X}^{(t)}$ denote the nonempty bucket averages, the estimator is defined by

$$\widehat{\mu}_j = \mathrm{Median}\left(\{\bar{X}_j^{(1)}, \dots, \bar{X}_j^{(t)}\}\right), \qquad j = 1, \dots, d.$$

The number of buckets is dictated by the geometry of the loss, namely by the effective exponent $\bar{p}$. More precisely, the estimator uses

$$t(\bar{p}) = 2\left\lceil \frac{\bar{p}}{2} \right\rceil - 1$$

buckets, with the median taken over the nonempty buckets. When $1 \le p < 2$, we have $t(\bar{p}) = 1$, so the estimator is just the ordinary sample mean. The same happens at $p = 2$, recovering exactly the Euclidean behavior in (1). When $p > 2$ and $\bar{p} > 2$, the estimator becomes a genuine coordinate-wise median of means estimator, with the number of buckets increasing with $\bar{p}$. Finally, in the extreme regime where $t(\bar{p}) > n$, every nonempty bucket contains a single observation, and the procedure reduces to the coordinate-wise median of the sample points. The technical difficulty is to show that these regimes match sharply and together give the precise minimax rate above.

[1] As an example, for matrix valued observations the natural benchmark is often the operator norm. In this setting, heavy-tailed matrices are known to require additional truncation steps even for in expectation risk bounds, before one can ask for sub-Gaussian tails; see, for example, [AZ26].

**Phase transition and the existing general norm results.** Our *in expectation* analysis complements the existing high probability theory of robust mean estimation with respect to general norms. Sub-Gaussian mean estimators for general norms were studied, in different forms, in [LM19a, Min25, DL22, OR25]. The original benchmark in [LM19a] is distribution dependent and is centered around robustly reproducing the performance of the empirical mean, together with the correct one dimensional sub-Gaussian deviation term. Informally, in our $\ell_p$ setup this benchmark contains a term of the form

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right\|_p,$$

or its symmetrized analogue. This is the natural target if the goal is to obtain optimal high probability tails around the performance of the sample mean. However, for the problem considered in this paper, this benchmark can be too large: as follows from our results, for $p > 2$, the sample mean itself is already suboptimal even in expectation.[2]

A related but different point is emphasized in [DL22]. If one takes Gaussian mean estimation as the benchmark, then the relevant complexity is the Gaussian width of the dual ball inducing the norm. In the case of the $\ell_p$ norm and identity covariance, this gives

$$\frac{\mathbb{E}\,\|G\|_p}{\sqrt{n}} \asymp d^{1/\bar{p}}\sqrt{\frac{\bar{p}}{n}},$$

where $G \sim N(0, I_d)$ and $\bar{p} = \min\{p, \max\{\log d, 1\}\}$. Thus, the second branch of our theorem,

$$d^{1/\bar{p}}\sqrt{\frac{\bar{p}}{n}},$$

coincides exactly with the Gaussian width benchmark. In this regime, our estimator achieves the same rate that the sample mean would have for Gaussian data; this may be called the sub-Gaussian regime. The genuinely new phenomenon provided by our analysis is the first branch,

$$d^{\frac{1}{2\lceil \frac{n}{2}\rceil}}.$$

This term does not appear in the Gaussian mean estimation setup. It reflects a finite sample heavy-tailed regime under only a covariance assumption. Morally, the Gaussian width rate is correct once the number of samples is large enough relative to the effective exponent $\bar{p}$. When this condition fails, the minimax risk enters a different phase.

**Phase transition and the Gaussian location estimation problem.** The Gaussian width branch of Theorem 1.1 is also closely connected to the classical minimax theory of Gaussian location estimation. In the Gaussian location model, Anderson's lemma implies an exact minimax identity for every bowl-shaped loss; see the original inequality of Anderson [And55] and the modern presentation in [PW25, Section 28.6]. In particular, for $X_1, \dots, X_n \sim N(\mu, I_d)$, Anderson's lemma implies

$$\inf_{\widehat{\mu}} \sup_{\mu\in\mathbb{R}^d} \mathbb{E}_{\mu}\,\|\widehat{\mu}-\mu\|_p = \frac{\mathbb{E}\,\|G\|_p}{\sqrt{n}} \asymp d^{1/\bar{p}}\sqrt{\frac{\bar{p}}{n}}, \qquad G\sim N(0,I_d), \quad \bar{p}=\min\{p,\max\{\log d,1\}\}.$$

Thus, as discussed above, the second branch of our theorem is exactly the classical Gaussian location obstruction. Our result shows that this Gaussian answer is not the whole story under only a covariance assumption. For the much larger class of all distributions with covariance $I_d$, the Gaussian benchmark remains sharp only when

$$\bar{p} \le 2\left\lceil \frac{n}{2}\right\rceil.$$

When this condition fails, heavy-tailed distributions create an additional obstruction, and the minimax risk becomes $d^{\frac{1}{2\lceil \frac{n}{2}\rceil}}$. In this sense, Theorem 1.1 may be viewed as a heavy-tailed, covariance-only analogue of the classical Gaussian location minimax identity.

[2] For example, let $X = \varepsilon\sqrt{d}\,e_J$, where $\varepsilon$ is Rademacher and $J$ is uniform on $[d]$. Then $\mathbb{E}\,X = 0$ and $\mathrm{Cov}(X) = I_d$. If $n \le c\sqrt{d}$, the sampled basis vectors are distinct with constant probability, so $\mathbb{E}\left\|n^{-1}\sum_{i=1}^n X_i\right\|_p \gtrsim \frac{\sqrt{d}}{n^{1-1/p}}$. When $2 < p \le \log d$ and $p \le 2\lceil n/2\rceil$, Theorem 1.1 gives the rate $d^{1/p}\sqrt{p/n}$. Their ratio is $\asymp (d/n)^{1/2-1/p}/\sqrt{p}$, which diverges for fixed $p > 2$ as $d/n \to \infty$. Thus, the sample mean is suboptimal when $n \le c\sqrt{d}$ and $2 < p \le \min\{\log d, 2\lceil n/2\rceil\}$.

**High probability implications and limitations of projection-based estimators.** As of now, the full high probability minimax picture for $\ell_p$ norms remains unclear. In particular, the aforementioned general norm results do not directly provide the matching high probability behavior for $\ell_p$ norms under only the covariance assumption. Our theorem in expectation already shows that the sample mean can be suboptimal. On the negative side, we further show that natural projection-based median of means estimators, of the type used in the general norm theory [LM19a, Min25, DL22], can also fail to achieve the correct $\ell_p$ behavior.

Informally, a projection-based median of means estimator works as follows. It searches for a point whose projection in every dual direction agrees with the median of the projected bucket means:

$$\widehat{\mu}_{\mathrm{proj}} \in \arg\min_{x} \sup_{\|v\|_q \le 1} \left| \langle v, x \rangle - \mathrm{Median}\left( \left\{ \langle v, \bar{X}^{(1)} \rangle, \dots, \langle v, \bar{X}^{(t)} \rangle \right\} \right) \right|, \qquad \frac{1}{p} + \frac{1}{q} = 1.$$

This construction is natural as a generic norm-agnostic procedure, but our result shows that it can be too crude even for the classical family of $\ell_p$ norms once $p > 2$.

**Theorem 1.2** (Suboptimality of Projection-based median-of-means)**.** *Let $p > 2$. There exist absolute constants $C, c > 0$ such that the following hold. Let $d, n \in \mathbb{N}$ be such that $n \le c\sqrt{d}$ and $d > C$. Then, there exists a distribution $D$ with mean $0$ and covariance $I_d$ such that the projection-based median-of-means estimator with any bucket size $m$ satisfying $n/m \ge 3$, $\widehat{\mu}_{\mathrm{Proj}}$, satisfies:*

$$\Pr_{\boldsymbol{X} \sim D^{\otimes n}} \left[ \|\widehat{\mu}_{\mathrm{Proj}}(\boldsymbol{X})\|_p \ge \frac{\sqrt{d}}{2n^{1-1/p}} \right] \ge 0.9.$$

This should be compared with the coordinate-wise estimator analyzed above. In the Gaussian width regime, the latter achieves a much smaller rate. For example, for $p = \infty$ and $\log d \lesssim n \ll \sqrt{d}$, the coordinate-wise estimator has error of order, with constant probability,

$$\sqrt{\frac{\log d}{n}},$$

while the projection-based median of means lower bound is of order, again with constant probability,

$$\frac{\sqrt{d}}{n}.$$

Thus, in this regime the projection-based estimator pays a polynomial price in the dimension, while the coordinate-wise estimator has only logarithmic dependence. This shows that, for structured norms such as $\ell_p$, one needs to go beyond projection-based estimators for general norms.

**A high probability result for the $\ell_\infty$ norm.** Although our main focus is on the results in expectation, as a complementary result we give a sharp high probability analysis for the $\ell_\infty$ norm. This result is the high probability analogue of the phase transition phenomenon above, and refines even the constant probability regimes implied by our analysis in expectation.

**Theorem 1.3** ($\ell_\infty$ high probability phase transition, informal)**.** *Under only the covariance assumption $\mathrm{Cov}_P(X) = I_d$, the optimal $\ell_\infty$ error at confidence level $1 - \delta$ is, up to universal constants,*

$$\begin{cases} \left(\frac{d}{\delta}\right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}}, & n \lesssim \log(d/\delta), \\ \sqrt{\frac{\log(d/\delta)}{n}}, & n \gtrsim \log(d/\delta). \end{cases}$$

*The upper bound is achieved by the coordinate-wise median of means estimator with $t = 2\lceil \log(d/\delta) \rceil$ buckets, and the lower bound holds for every estimator.*

As before, the first regime is genuinely non-Gaussian: the error is polynomial in $d/\delta$, although the exponent decreases with the sample size. Once $n$ is of order $\log(d/\delta)$, the usual sub-Gaussian rate takes over. Thus, even for the classical $\ell_\infty$ norm, the Gaussian width or sub-Gaussian benchmark captures only the large sample side of the theory, while under only a covariance assumption there is an additional finite sample heavy-tailed phase. This also highlights a limitation of projection-based estimators: in the regimes where the coordinate-wise estimator attains the sharp $\ell_\infty$ rate, projection-based median of means estimators can be polynomially suboptimal, as shown by Theorem 1.2.

**On computation and coordinate-wise medians.** A significant line of work on heavy tailed mean estimation in the Euclidean norm studies polynomial time estimators with sub-Gaussian guarantees under finite covariance assumptions; see, for example, [Hop20, CFB19, HLZ20, DKP20]. In the main regimes in expectation studied here, computation is essentially trivial: the optimal estimator is coordinate-wise and reduces to bucket averages and one dimensional medians. At the same time, coordinate-wise median of means is not universally optimal. In the high probability Euclidean problem, no single distribution independent choice of the number of buckets is optimal for all distributions with identity covariance. We prove the following result.

**Theorem 1.4** (Suboptimality of coordinate-wise median-of-means)**.** *There are universal constants $c, C \geq 0$ such that the following holds. For any $\delta \in (0, c)$ and $n, d, t \in \mathbb{N}$ with $t \leq n$, there exists a distribution $P$ over $\mathbb{R}^d$ with mean $0$ and covariance $I_d$ such that if Algorithm 1 is ran with $t$ buckets and sample $X_1, \ldots, X_n \sim P$ as input, it returns an estimate $\hat{\mu}$ satisfying*

$$\mathop{\mathbf{Pr}}_{X_1,\ldots,X_n \sim P}\left[\|\hat{\mu}\|_2 \geq C\sqrt{\frac{d\log(1/\delta)}{n\log\log(1/\delta)}}\right] > \delta.$$

This is expected, and probably folklore, but we include it for completeness in Section D. The complete high probability theory for general norms therefore remains substantially more delicate than the picture in expectation developed here.[3]

## 2 Notation and organization

We begin with some notation. We use $\log$ to denote the natural logarithm. For a real number $a$, we will use $\lfloor a \rfloor$ and $\lceil a \rceil$ to denote its floor and ceiling respectively. For a positive integer $a$, we will often use the fact that $\lfloor \frac{a+1}{2} \rfloor$ is a short form for $\frac{a+1}{2}$ when $a$ is odd and $\frac{a}{2}$ when $a$ is even. For a multiset of real numbers, $\{a_i\}_{i=1}^n$, we use $a_{(1)} \geq a_{(2)} \cdots \geq a_{(n)}$ to denote the decreasing rearrangement of its elements with the median being $a_{\lfloor (n+1)/2 \rfloor}$ if $n$ is odd or $(a_{(n/2)} + a_{(n/2+1)})/2$ when $n$ is even. For positive integer $a$ we define $[a] = \{1, 2, \ldots, a\}$. We use $c, C$ for universal constants, and their value may change from line to line. For $x \in \mathbb{R}^d$ and $p \in [1, \infty)$, we use $\|x\|_p$ to denote the $\ell_p$ norm of $x$, with $q$ typically used to denote the dual norm exponent satisfying $\frac{1}{p} + \frac{1}{q} = 1$. We use $e_1, \ldots, e_d \in \mathbb{R}^d$ to denote the standard basis vectors over $\mathbb{R}^d$. We use $I_d$ for identity matrix whose dimension will be clear from context.

For a distribution $P$ over $\mathcal{X}$, event $A : \mathcal{X} \to \{0, 1\}$ and function $f : \mathcal{X} \to \mathbb{R}$, we will use $\mathbf{Pr}_{X \sim P}[A(X)]$, and $\mathbb{E}_{X \sim P}[f(X)]$ to be the probability of event $A$ and the expectation of $f$ when $X$ is drawn from $P$ respectively. When a random variable or its distribution is clear from context, we often drop the subscript and simply write $\mathbf{Pr}_P$ ($\mathbb{E}_P$) or $\mathbf{Pr}$ ($\mathbb{E}$). Furthermore, for a natural number $n$, we use $P^{\otimes n}$ to denote the product distribution over $n$ i.i.d. samples from $P$. For a distribution $P$ over $\mathbb{R}^d$ with mean $\mu$, we define the covariance matrix as $\mathrm{Cov}_P(X) = \mathbb{E}_{X \sim P}[XX^\top] - \mu\mu^\top$.

For quantities $a, b$, we write $a \lesssim b$ ($a \gtrsim b$) if there exists a universal constant $C > 0$ ($c > 0$) such that $a \leq Cb$ ($a \geq cb$) uniformly over the relevant parameter ranges which will be clear from context. We write $a = O(b)$ ($a = \Omega(b)$) if $a \lesssim b$ ($a \gtrsim b$) asymptotically in the relevant parameters and $a = \Theta(b)$ if both hold. Finally, $a = o(b)$ ($a = \omega(b)$) if $\frac{a}{b}$ ($\frac{b}{a}$) goes to 0 asymptotically.

**Organization:** The remainder of the paper is organized as follows: Theorem 1.1 is formalized and proved in Section 3, Theorem 1.3 in Section A, Theorem 1.2 in Section C, and Theorem 1.4 in Section D.

## 3 Optimal in-expectation guarantees

In this section we prove sharp bounds for the mini-max risk of heavy-tailed isotropic mean estimation. The minimax risk of sample size $n$, dimension $d$ and norm $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is defined as follows:

$$\mathfrak{R}(n, d, \|\cdot\|) = \inf_{\mathcal{A}:(\mathbb{R}^d)^n \to \mathbb{R}^d} \sup_{\substack{\mu \in \mathbb{R}^d \\ P:\mathbb{E}_P[X]=\mu,\ \mathrm{Cov}_P(X)=I_d}} \mathbb{E}_{X_1,\ldots,X_n \sim P}\big[\|\mathcal{A}(X_1, \ldots, X_n) - \mu\|\big]. \quad (2)$$

[3] Recent generic chaining based approaches are particularly relevant here; see [AM26, Theorem 5.1] and [BM26]. We note that these works do not provide a sharp high probability minimax theory for $\ell_p$ norms under the covariance only assumption.

**Algorithm 1** Coordinate-wise Median-of-Means

**Require:** Number of buckets $t \le n$, samples $X_1, \dots, X_n \in \mathbb{R}^d$

Partition $[n]$ into $t$ size-balanced buckets $B_1, \dots, B_t$ by setting $B_i = \{m \in [n] : m \equiv i \pmod t\}$

**for** $i = 1$ to $t$ **do**

  $\hat{\mu}^{(i)} \leftarrow \frac{1}{|B_i|} \sum_{m \in B_i} X_m$

**end for**

**return** $\hat{\mu} \leftarrow \text{Median}\big(\hat{\mu}^{(1)}, \dots, \hat{\mu}^{(t)}\big)$ ▷ coordinate-wise median

In words, this is the smallest expected error when estimating the mean, measured in the norm $\|\cdot\|$, that an estimator can guarantee based on $n$ samples, uniformly over all $d$-dimensional distributions with identity covariance. For $\ell_p$ norms, we will simply write $\mathfrak{R}(n, d, p) = \mathfrak{R}(n, d, \|\cdot\|_p)$.

To state our main result, it will be useful to setup some notation that will be used throughout this section. We will make frequent use of the truncated version of $p \in [1, \infty]$ defined as

$$\bar{p} = \min\{p, \max\{\log d, 1\}\}. \tag{3}$$

The truncation at $\log d$ is because for any $p, q$ such that $\log d \le p \le q$, we have that $e^{-1}\|\cdot\|_p \le \|\cdot\|_q \le \|\cdot\|_p$, i.e. the norms are equivalent up to a constant factor. Thus, the minimax risk with respect to $\ell_p$ and $\ell_q$ are equivalent up to a constant factor so it suffices to consider the range $p \in [1, \log d]$.[4]

We can now state our main result.

**Theorem 3.1.** *For any $n, d \in \mathbb{N}$ and $p \in [1, \infty]$ we have that*

$$\mathfrak{R}(n, d, p) \asymp \begin{cases} d^{\frac{1}{2\lceil n/2 \rceil}}, & \bar{p} > 2\left\lceil \frac{n}{2} \right\rceil, \\ d^{\frac{1}{\bar{p}}}\sqrt{\frac{\bar{p}}{n}}, & \bar{p} \le 2\left\lceil \frac{n}{2} \right\rceil. \end{cases}$$

*Equivalently, if we define $s = \min\{\bar{p}, 2\left\lceil \frac{n}{2} \right\rceil\}$, then*

$$\mathfrak{R}(n, d, p) \asymp d^{\frac{1}{s}}\sqrt{\frac{s}{n}}.$$

*Moreover, this rate is achieved by the coordinate-wise median-of-means estimator.*

**Remark 3.2.** Observe that a phase transition occurs exactly at $\bar{p} = 2\left\lceil \frac{n}{2} \right\rceil$, where the small sample size regime $2\left\lceil \frac{n}{2} \right\rceil < \bar{p}$ rate $d^{\frac{1}{2\lceil \frac{n}{2} \rceil}}$ changes to the large sample size regime $\bar{p} \le 2\left\lceil \frac{n}{2} \right\rceil$ rate $d^{\frac{1}{\bar{p}}}\sqrt{\frac{\bar{p}}{n}}$.

### 3.1 Upper bound

In this section we prove upper bounds on the expected error of the coordinate-wise median-of-means estimator (Algorithm 1) when used to estimate the mean with respect to $\ell_p$ norms for $p \in [1, \infty]$. We will make frequent use of the function $t(\bar{p})$:

$$t(\bar{p}) = 2\left\lceil \frac{\bar{p}}{2} \right\rceil - 1. \tag{4}$$

We will use $t(\bar{p})$ as the number of buckets in coordinate-wise median-of-means estimator Algorithm 1 when estimating the mean with respect to the $\ell_p$ norm. Notice that in the case that $p \in [1, 2]$, the number of buckets is just 1 and Algorithm 1 reduces to the usual sample mean that is optimal in-expectation for $p \in [1, 2]$.

We now state our upper bound on the performance of Algorithm 1 for estimating the mean of a random vector in $\mathbb{R}^d$ under $\ell_p$ norm for any $p \ge 1$.

**Theorem 3.3.** *Let $X_1, \dots, X_n \in \mathbb{R}^d$ be i.i.d. random vectors with finite mean $\mu$ and covariance $\Sigma$. Fix $p \ge 1$. Given the number of buckets $t(\bar{p})$ and the sample $X_1, \dots, X_n$ as input, Algorithm 1*

[4] The truncation with 1 is just to avoid technicalities when $d \le 2$.

*returns an estimate $\hat{\mu}$ satisfying*

$$\mathbb{E}\left[\|\hat{\mu}-\mu\|_p\right] \leq \begin{cases} 2\left(\sum_{j=1}^{d} \Sigma_{jj}^{\left\lceil \frac{n}{2} \right\rceil}\right)^{\frac{1}{2\left\lceil \frac{n}{2} \right\rceil}}, & \bar{p} > 2\left\lceil \frac{n}{2} \right\rceil, \\ 2\,\frac{\left(\sum_{j=1}^{d} \Sigma_{jj}^{\bar{p}/2}\right)^{\frac{1}{\bar{p}}}}{\sqrt{\lfloor n/t(\bar{p}) \rfloor}}, & \bar{p} \leq 2\left\lceil \frac{n}{2} \right\rceil. \end{cases}$$

*Moreover, if we define $s = \min\{\bar{p}, 2\left\lceil \frac{n}{2} \right\rceil\}$ then*

$$\mathbb{E}\left[\|\hat{\mu}-\mu\|_p\right] \leq 4\left(\sum_{j=1}^{d} \Sigma_{jj}^{s/2}\right)^{\frac{1}{s}} \sqrt{\frac{s}{n}}.$$

To prove Theorem 3.3, we will need the following moment bound on the deviation of the empirical median from the population mean which is a slight generalization of a result due to [LPT21]. For completeness, we include a proof in Section B.

**Theorem 3.4.** *Let $X_1, \ldots, X_t$ be $t$ independent random variables with a common mean $\mu$ and let $M = M(X_1, \ldots, X_t)$ denote their median. For all $q \geq 0$ it holds that*

$$\mathbb{E}\left[|M-\mu|^{q\left\lceil \frac{t}{2} \right\rceil}\right] \leq \binom{t}{\left\lceil \frac{t}{2} \right\rceil} \max_{i\in[t]} \mathbb{E}\left[|X_i-\mu|^q\right]^{\left\lceil \frac{t}{2} \right\rceil}.$$

*Proof of Theorem 3.3.* We consider two cases. First, let us consider the case that $\bar{p} \leq 2\left\lceil \frac{n}{2} \right\rceil$. Notice that this condition is equivalent to the condition that $n \geq t(\bar{p})$, which will be more convenient to use in what follows. Indeed, we have that

$$t(\bar{p}) \leq n \iff \left\lceil \frac{\bar{p}}{2} \right\rceil \leq \left\lfloor \frac{n+1}{2} \right\rfloor \iff \left\lceil \frac{\bar{p}}{2} \right\rceil \leq \left\lceil \frac{n}{2} \right\rceil \iff \frac{\bar{p}}{2} \leq \left\lceil \frac{n}{2} \right\rceil \iff \bar{p} \leq 2\left\lceil \frac{n}{2} \right\rceil. \tag{5}$$

Now, using the fact that $\|\cdot\|_p \leq \|\cdot\|_{\bar{p}}$ together with Jensen's inequality yields

$$\mathbb{E}\left[\|\hat{\mu}-\mu\|_p\right] \leq \mathbb{E}\left[\|\hat{\mu}-\mu\|_{\bar{p}}\right] \leq \mathbb{E}\left[\|\hat{\mu}-\mu\|_{\bar{p}}^{\bar{p}}\right]^{\frac{1}{\bar{p}}} = \left(\sum_{j=1}^{d} \mathbb{E}\left[|\bar{\mu}_j-\mu_j|^{\bar{p}}\right]\right)^{\frac{1}{\bar{p}}}. \tag{6}$$

It now suffices to bound the terms $\mathbb{E}\left[|\hat{\mu}_j-\mu_j|^{\bar{p}}\right]$. Notice that $\left\lceil \frac{t(\bar{p})}{2} \right\rceil = \lfloor \frac{t(\bar{p})+1}{2} \rfloor = \lceil \frac{\bar{p}}{2} \rceil \geq 1$, so there is some $q \in [1,2]$ such that $q\left\lceil \frac{t(\bar{p})}{2} \right\rceil = \bar{p}$. Thus, we can invoke the moment bound of Theorem 3.4 with $t = t(\bar{p})$ and $q \in [1,2]$ as above together with Jensen's inequality to get

$$\begin{aligned} \mathbb{E}\left[|\hat{\mu}_j-\mu_j|^{\bar{p}}\right] &\leq \binom{t(\bar{p})}{\left\lceil \frac{t(\bar{p})}{2} \right\rceil} \max_{i\in[t(\bar{p})]} \mathbb{E}\left[|\hat{\mu}_j^{(i)}-\mu_j|^q\right]^{\left\lceil \frac{t(\bar{p})}{2} \right\rceil} \\ &\leq \binom{t(\bar{p})}{\left\lceil \frac{t(\bar{p})}{2} \right\rceil} \max_{i\in[t(\bar{p})]} \mathbb{E}\left[|\hat{\mu}_j^{(i)}-\mu_j|^2\right]^{\frac{q}{2}\left\lceil \frac{t(\bar{p})}{2} \right\rceil} \\ &\leq \binom{t(\bar{p})}{\left\lceil \frac{t(\bar{p})}{2} \right\rceil} \left(\frac{\Sigma_{jj}}{\lfloor \frac{n}{t(\bar{p})} \rfloor}\right)^{\frac{q}{2}\left\lceil \frac{t(\bar{p})}{2} \right\rceil} = \binom{t(\bar{p})}{\lfloor \frac{t(\bar{p})+1}{2} \rfloor} \left(\frac{\Sigma_{jj}}{\lfloor \frac{n}{t(\bar{p})} \rfloor}\right)^{\frac{\bar{p}}{2}}, \end{aligned}$$

where the last inequality follows from the fact that each bucket gets at least $\lfloor n/t(\bar{p}) \rfloor$ samples and that the variance of the $j$th coordinate is $\Sigma_{jj}$. Plugging the above back into (6) yields

$$\mathbb{E}\left[\|\hat{\mu}-\mu\|_p\right] \leq \binom{t(\bar{p})}{\left\lceil \frac{t(\bar{p})}{2} \right\rceil}^{\frac{1}{\bar{p}}} \frac{\left(\sum_{j=1}^{d} (\Sigma_{jj})^{\frac{\bar{p}}{2}}\right)^{\frac{1}{\bar{p}}}}{\sqrt{\lfloor n/t(\bar{p}) \rfloor}} \leq 2\frac{\left(\sum_{j=1}^{d} (\Sigma_{jj})^{\frac{\bar{p}}{2}}\right)^{\frac{1}{\bar{p}}}}{\sqrt{\lfloor n/t(\bar{p}) \rfloor}},$$

where the final inequality follows from the simple bound

$$\binom{t(\bar{p})}{\left\lceil \frac{t(\bar{p})}{2} \right\rceil} = \binom{2\lceil \bar{p}/2 \rceil - 1}{\lceil \bar{p}/2 \rceil} = \binom{2\lceil \bar{p}/2 \rceil - 1}{\lceil \bar{p}/2 \rceil - 1} \leq \sum_{k=0}^{\lceil \bar{p}/2 \rceil - 1} \binom{2\lceil \bar{p}/2 \rceil - 1}{k} = 2^{2\lceil \frac{\bar{p}}{2} \rceil - 2} \leq 2^{\bar{p}}.$$

Now consider the case that $\bar{p} > 2\left\lceil \frac{n}{2} \right\rceil$, which is equivalent to $n < t(\bar{p})$ by (5). In this case, every non-empty bucket has a single sample and the estimator is the coordinate-wise median. Since $n < t(\bar{p}) = 2\lceil \bar{p}/2 \rceil - 1$, a simple case analysis on the parity of $n$ shows that $p > 2\lceil \frac{n}{2} \rceil =: r(n)$. Thus, using the fact that $\|\cdot\|_p \leq \|\cdot\|_{2\lceil \frac{n}{2} \rceil}$ and Jensen's inequality yields

$$\mathbb{E}\left[\|\hat{\mu} - \mu\|_p\right] \leq \mathbb{E}\left[\|\hat{\mu} - \mu\|_{2\lceil \frac{n}{2} \rceil}\right] \leq \mathbb{E}\left[\|\hat{\mu} - \mu\|_{r(n)}^{r(n)}\right]^{\frac{1}{r(n)}} = \left(\sum_{j=1}^{d} \mathbb{E}\left[|\bar{\mu}_j - \mu_j|^{r(n)}\right]\right)^{\frac{1}{r(n)}}. \quad (7)$$

So, it suffices to bound the terms $\mathbb{E}\left[|\hat{\mu}_j - \mu_j|^{r(n)}\right]$. As in the previous case, we can invoke Theorem 3.4 with $q = 2$ and $t = n$ to get the bound

$$\mathbb{E}\left[|\hat{\mu}_j - \mu_j|^{r(n)}\right] = \mathbb{E}\left[|\hat{\mu}_j - \mu_j|^{2\lceil \frac{n}{2} \rceil}\right] \leq \binom{n}{\lceil \frac{n}{2} \rceil} \Sigma_{jj}^{\lceil \frac{n}{2} \rceil}.$$

Plugging this into the bound (7) and using a similar approximation for the binomial term as before yields the final bound

$$\mathbb{E}\left[\|\hat{\mu} - \mu\|_p\right] \leq 2\left(\sum_{j=1}^{d} (\Sigma_{jj})^{\lceil \frac{n}{2} \rceil}\right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}},$$

as desired. We now derive the uniform bound based on $s = \min\{\bar{p},\, 2\left\lceil \frac{n}{2} \right\rceil\}$. Notice that in the case that $n \geq t(\bar{p})$ we have $s = \bar{p}$ and we have the simple bound

$$\mathbb{E}\left[\|\hat{\mu} - \mu\|_p\right] \leq 2\frac{\left(\sum_{j=1}^{d} (\Sigma_{jj})^{\frac{\bar{p}}{2}}\right)^{\frac{1}{\bar{p}}}}{\sqrt{\lfloor n/t(\bar{p}) \rfloor}} = 2\frac{\left(\sum_{j=1}^{d} (\Sigma_{jj})^{\frac{s}{2}}\right)^{\frac{1}{s}}}{\sqrt{\lfloor n/t(\bar{p}) \rfloor}} \leq 4\left(\sum_{j=1}^{d} (\Sigma_{jj})^{\frac{s}{2}}\right)^{\frac{1}{s}} \sqrt{\frac{s}{n}},$$

where the final inequality follows from the facts that $n \geq t(\bar{p})$ implies $\lfloor \frac{n}{t(\bar{p})} \rfloor \geq \frac{n}{2t(\bar{p})}$ and $t(\bar{p}) = 2\left\lceil \frac{\bar{p}}{2} \right\rceil - 1 \leq 2\bar{p} = 2s$ and so $\frac{1}{\sqrt{\lfloor n/t(\bar{p}) \rfloor}} \leq \sqrt{\frac{2t(\bar{p})}{n}} \leq 2\sqrt{\frac{s}{n}}$. Now in the case that $n < t(\bar{p})$ we have $s = 2\left\lceil \frac{n}{2} \right\rceil = r(n)$ and hence

$$\mathbb{E}\left[\|\hat{\mu} - \mu\|_p\right] \leq 2\left(\sum_{j=1}^{d} (\Sigma_{jj})^{\lceil \frac{n}{2} \rceil}\right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}} = 2\left(\sum_{j=1}^{d} (\Sigma_{jj})^{\frac{s}{2}}\right)^{\frac{1}{s}} \leq 2\left(\sum_{j=1}^{d} (\Sigma_{jj})^{\frac{s}{2}}\right)^{\frac{1}{s}} \sqrt{\frac{s}{n}},$$

where the final inequality uses the fact that $s = 2\left\lceil \frac{n}{2} \right\rceil \geq n$. This completes the proof. □

### 3.2 Lower bounds

In this section we prove that the rate achieved in Theorem 3.3 is optimal. The lower bound has two components, proved as two separate lemmas, corresponding to the two regimes discussed in the introduction. The first obstruction is already present in the Gaussian location model and gives the Gaussian-width branch. The second obstruction is genuinely non-Gaussian and is witnessed by heavy-tailed distributions. We begin with the Gaussian lower bound. It matches Theorem 3.3 up to constants in the regime $\bar{p} \leq 2\left\lceil \frac{n}{2} \right\rceil$.

**Lemma 3.5** (Gaussian-width lower bound)**.** *There exists a universal constant $c > 0$ such that for any $d, n \in \mathbb{N}$, any $p \in [1, \infty]$, and any estimator $\mathcal{A} : (\mathbb{R}^d)^n \to \mathbb{R}^d$,*

$$\sup_{\mu \in \mathbb{R}^d} \mathbb{E}_{X_1, \ldots, X_n \sim \mathcal{N}(\mu, I_d)} \left[\|\mathcal{A}(X_1, \ldots, X_n) - \mu\|_p\right] \geq c\, d^{\frac{1}{p}} \sqrt{\frac{\bar{p}}{n}}.$$

*Consequently, the same lower bound holds for the minimax risk $\mathfrak{R}(n, d, p)$.*

*Proof.* Let $P_\mu = \mathcal{N}(\mu, I_d)$ and let $G \sim \mathcal{N}(0, I_d)$. Since the Gaussian family $\{P_\mu : \mu \in \mathbb{R}^d\}$ is contained in the class of distributions with covariance $I_d$, it suffices to prove the lower bound over this subfamily. By [PW25, Theorem 28.7], which is a consequence of Anderson's lemma [And55], applied with $\rho(x) = \|x\|_p$, the minimax risk in the Gaussian location model is

$$\inf_{\widehat{\mu}} \sup_{\mu \in \mathbb{R}^d} \mathop{\mathbb{E}}_{X_1,\dots,X_n \sim P_\mu} \left[\|\widehat{\mu}(X_1, \dots, X_n) - \mu\|_p\right] = \mathbb{E}\left\|\frac{G}{\sqrt{n}}\right\|_p = \frac{\mathbb{E}\,\|G\|_p}{\sqrt{n}}.$$

Indeed, the function $\rho(x) = \|x\|_p$ is lower-semicontinuous and bowl-shaped (its sublevel sets are symmetric and convex). Therefore every estimator $\mathcal{A}$ satisfies

$$\sup_{\mu \in \mathbb{R}^d} \mathop{\mathbb{E}}_{X_1,\dots,X_n \sim P_\mu} \left[\|\mathcal{A}(X_1, \dots, X_n) - \mu\|_p\right] \geq \frac{\mathbb{E}\,\|G\|_p}{\sqrt{n}}. \tag{8}$$

It remains to lower bound the Gaussian width term. The standard Gaussian norm estimate gives

$$\mathbb{E}\,\|G\|_p \asymp d^{\frac{1}{\bar{p}}}\sqrt{\bar{p}}, \qquad \bar{p} = \min\{p,\, \max\{\log d,\, 1\}\}.$$

Plugging this estimate into (8) yields the claim. □

We now prove the second lower bound. Unlike Lemma 3.5, this obstruction is not present in the Gaussian location model. It is a genuinely heavy-tailed obstruction and gives the small-sample branch of the minimax rate. Since $\|x\|_p \geq \|x\|_\infty$ for all $x \in \mathbb{R}^d$ and all $p \geq 1$, it suffices to prove a lower bound in the $\ell_\infty$ norm.

**Lemma 3.6** (Heavy-tailed small-sample lower bound)**.** *There exists a universal constant $c > 0$ satisfying the following. For any $d, n \in \mathbb{N}$ with $\log d \geq 2\left\lceil \frac{n}{2} \right\rceil$ and any estimator $\mathcal{A} : (\mathbb{R}^d)^n \to \mathbb{R}^d$, there exists a distribution $P$ with mean $\mu \in \mathbb{R}^d$ and covariance $I_d$ such that*

$$\mathop{\mathbb{E}}_{X_1,\dots,X_n \sim P} \left[\|\mathcal{A}(X_1, \dots, X_n) - \mu\|_\infty\right] \geq c\, d^{\frac{1}{2\left\lceil \frac{n}{2} \right\rceil}}.$$

*Consequently, in the regime $\log d \geq 2\left\lceil \frac{n}{2} \right\rceil$, the same lower bound holds for the minimax risk $\mathfrak{R}(n, d, p)$ for every $p \in [1, \infty]$.*

*Proof.* We use the high-probability $\ell_\infty$ lower bound proven in Theorem A.4. Apply Theorem A.4 with $\sigma = 1$ and $\delta_0 = 1/4$. Then for any estimator $\mathcal{A}$, there exists a distribution $P$ on $\mathbb{R}^d$ with mean $\mu$ and covariance $I_d$ such that

$$\mathop{\mathbf{Pr}}_{X_1,\dots,X_n \sim P}\left[\|\mathcal{A}(X_1, \dots, X_n) - \mu\|_\infty \geq \frac{1}{4}\left(1 - 2d^{-\frac{1}{\left\lceil \frac{n}{2} \right\rceil}}\right)(4d)^{\frac{1}{2\left\lceil \frac{n}{2} \right\rceil}}\right] > \frac{1}{4}.$$

Here we used that, for $\delta_0 = 1/4$, the auxiliary parameter appearing in Theorem A.4 is $\left(\frac{4\delta_0}{d}\right)^{\frac{1}{\left\lceil \frac{n}{2} \right\rceil}} = d^{-\frac{1}{\left\lceil \frac{n}{2} \right\rceil}}$. By the assumption $\log d \geq 2\left\lceil \frac{n}{2} \right\rceil$, we have $d^{-\frac{1}{\left\lceil \frac{n}{2} \right\rceil}} \leq e^{-2}$. Therefore,

$$\frac{1}{4}\left(1 - 2d^{-\frac{1}{\left\lceil \frac{n}{2} \right\rceil}}\right)(4d)^{\frac{1}{2\left\lceil \frac{n}{2} \right\rceil}} \geq \frac{1 - 2e^{-2}}{4} d^{\frac{1}{2\left\lceil \frac{n}{2} \right\rceil}}.$$

Using the elementary inequality $\mathbb{E}\, Z \geq a\, \mathbf{Pr}[Z \geq a]$ for nonnegative random variables $Z$, we conclude that

$$\mathop{\mathbb{E}}_{X_1,\dots,X_n \sim P} \left[\|\mathcal{A}(X_1, \dots, X_n) - \mu\|_\infty\right] \geq \frac{1 - 2e^{-2}}{16} d^{\frac{1}{2\left\lceil \frac{n}{2} \right\rceil}}.$$

This proves the lemma with $c = \frac{1-2e^{-2}}{16}$. □

### 3.3 Proof of Theorem 3.1

We begin by showing that guarantee of Algorithm 1 provided in Theorem 3.3 achieves the claimed minimax risk in the isotropic case. First, notice that in the isotropic case $\Sigma = I_d$, so the guarantee reduces to

$$\mathfrak{R}(n, d, p) \leq \begin{cases} 2d^{\frac{1}{2\left\lceil \frac{n}{2} \right\rceil}}, & \bar{p} > 2\left\lceil \frac{n}{2} \right\rceil, \\ 2\,\dfrac{d^{\frac{1}{\bar{p}}}}{\sqrt{\lfloor n/t(\bar{p}) \rfloor}}, & \bar{p} \leq 2\left\lceil \frac{n}{2} \right\rceil. \end{cases}$$

Notice that guarantee of Theorem 3.3 based on $s = \min\{\bar{p}, 2\left\lceil \frac{n}{2} \right\rceil\}$ implies the upper bound

$$\mathfrak{R}(n, d, p) \leq 4d^{\frac{1}{s}} \sqrt{\frac{s}{n}}.$$

We now combine the lower bounds from Lemmas 3.5 and 3.6. Consider the regime where $\bar{p} > 2\left\lceil \frac{n}{2} \right\rceil$. In this case, we certainly have that $\log d \geq 2\left\lceil \frac{n}{2} \right\rceil$, so Lemma 3.6 implies that

$$\mathfrak{R}(n, d, p) \gtrsim d^{\frac{1}{2\left\lceil \frac{n}{2} \right\rceil}}.$$

We also have the lower bound on the risk implied by Lemma 3.5 that tells us that

$$\mathfrak{R}(n, d, p) \gtrsim d^{\frac{1}{\bar{p}}} \sqrt{\frac{\bar{p}}{n}}.$$

Thus, all together we conclude the lower bound

$$\mathfrak{R}(n, d, p) \gtrsim \begin{cases} d^{\frac{1}{2\lceil n/2 \rceil}}, & \bar{p} > 2\left\lceil \frac{n}{2} \right\rceil, \\ d^{\frac{1}{\bar{p}}} \sqrt{\frac{\bar{p}}{n}}, & \bar{p} \leq 2\left\lceil \frac{n}{2} \right\rceil, \end{cases}$$

which is equivalent to the following lower bound based on $s$:

$$\mathfrak{R}(n, d, p) \gtrsim d^{\frac{1}{s}} \sqrt{\frac{s}{n}}.$$

This concludes the proof. □

## Acknowledgments and Disclosure of Funding

Mikael was while carrying out this work, supported by an Internationalisation Fellowship from the Carlsberg Foundation, Independent Research Fund Denmark (DFF) Sapere Aude Research Leader grant No. 9064-00068B, and the European Union (ERC, TUCLA, 101125203).

Kasper was supported by the European Union (ERC, TUCLA, 101125203). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Yeshwanth Cherapanamjeri is supported by an NTU-NAP Start-up Grant (025661-00016).

## References

[AM26] Pedro Abdalla and Shahar Mendelson. Covariance estimation with direction dependence accuracy. *Probability Theory and Related Fields*, 194(1–2):371–409, 2026. 5

[AMS99] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999. 1

[And55] T. W. Anderson. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 6(2):170–176, 1955. 1, 3, 9, 15

[AZ26] Pedro Abdalla and Nikita Zhivotovskiy. Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails. *Journal of the European Mathematical Society*, 28(4):1809–1847, 2026. 2

[BM26] Daniel Bartl and Shahar Mendelson. Uniform mean estimation via generic chaining. *Advances in Mathematics*, 493, 2026. 5

[Cat12] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012. 1

[CFB19] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-Gaussian rates. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 786–806. PMLR, 2019. 5

[CG17] Olivier Catoni and Ilaria Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression, 2017. 2

[DKP20] Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. In *Advances in Neural Information Processing Systems*, volume 33, pages 1830–1840, 2020. 5

[DL22] Jules Depersin and Guillaume Lecué. Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms. *Probability Theory and Related Fields*, 183(3):1005–1056, 2022. 3, 4

[HLZ20] Samuel B. Hopkins, Jerry Li, and Fred Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 11902–11912, 2020. 5

[Hop20] Samuel B. Hopkins. Mean estimation with sub-Gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193–1213, 2020. 5

[JVV86] Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43(2–3):169–188, 1986. 1

[LM19a] Gábor Lugosi and Shahar Mendelson. Near-optimal mean estimators with respect to general norms. *Probability Theory and Related Fields*, 175(3–4):957–973, 2019. 3, 4

[LM19b] Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019. 2

[LM21] Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021. 1, 2

[LPT21] Kasper Green Larsen, Rasmus Pagh, and Jakub Tetek. Countsketches, feature hashing and the median of three. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pages 6011–6020. PMLR, 2021. 7

[Min15] Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015. 2

[Min25] Stanislav Minsker. Uniform bounds for robust mean estimators. *Stochastic Processes and their Applications*, 190:104724, 2025. 3, 4

[NY83] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Chichester, 1983. 1

[OR25] Roberto I. Oliveira and Lucas Resende. Trimmed sample means for robust uniform mean estimation and regression. *The Annals of Statistics*, 53(5):2153 – 2178, 2025. 3

[PW25] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025. 3, 9

## A Optimal high-probability guarantees for $\ell_\infty$

In this section, we formalize and prove Theorem 1.3 which characterizes the optimal high-probability estimation guarantees under the $\ell_\infty$ norm. First, define the following minimax risk:

$$r(n,d,\delta) = \inf_{\mathcal{A}:(\mathbb{R}^d)^n\to\mathbb{R}^d} \left\{ r>0 : \sup_{\substack{\mu\in\mathbb{R}^d \\ P:\,\mathbb{E}_P[X]=\mu \\ \mathrm{Cov}_P(X)=I_d}} \mathbf{Pr}_{X_1,\dots,X_n\sim P}\left[\|\mathcal{A}(X_1,\dots,X_n)-\mu\|_\infty \geq r\right] \leq \delta \right\}. \tag{9}$$

The formalization of Theorem 1.3 is given below.

**Theorem A.1.** *There is a universal constant $c > 0$ such that for any $n, d \in \mathbb{N}$, and $\delta \in (0, c)$, we have that:*

$$r(n,d,\delta) \asymp \begin{cases} \left(\frac{d}{\delta}\right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}}, & n \lesssim \log(d/\delta), \\ \sqrt{\frac{\log(d/\delta)}{n}}, & n \gtrsim \log(d/\delta). \end{cases}$$

*Moreover, this rate is achieved by the coordinate-wise median-of-means estimator.*

Note that the rates in the two cases coincide when $n = \Theta(\log(d/\delta))$. The proof is presented in two parts with the upper bound established in Subsection A.1 and the lower bound in Subsection A.2.

### A.1 High-probability upper bound in $\ell_\infty$

In this section we prove high probability bounds for the performance of the coordinate wise median of means estimator for estimating the mean of a random vector in $\mathbb{R}^d$ under the $\ell_\infty$ norm. Towards this end, we begin with the one dimensional case.

**Theorem A.2.** *Let $X_1, \ldots, X_n \in \mathbb{R}$ be i.i.d. random variables with finite mean $\mu$ and variance $\sigma^2$. Fix a confidence parameter $\delta \in (0,1)$. Given the number of buckets $t = 2\lceil \log(1/\delta) \rceil$ and the sample $X_1, \ldots, X_n$ as input, Algorithm 1 returns an estimate $\hat{\mu}$ satisfying*

$$|\hat{\mu} - \mu| \le \begin{cases} 2\sigma \left(\frac{1}{\delta}\right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}}, & n < 2\lceil \log\left(\frac{1}{\delta}\right) \rceil, \\ 4\sigma \sqrt{\frac{e}{n} \log\left(\frac{1}{\delta}\right)}, & n \ge 2\lceil \log\left(\frac{1}{\delta}\right) \rceil, \end{cases} \tag{10}$$

*with probability at least $1 - \delta$ over the randomness of $X_1, \ldots, X_n$.*

*Proof.* We split the analysis into two very different cases. First, let us consider the case that the sample size $n$ is smaller than the number of buckets $t$. In this case, Algorithm 1 returns the median of the $n$ sample points. Using Markov's inequality together with the moment bound of Theorem 3.4 with $q = 2$ and $t = n$ implies that our output $\hat{\mu}$ satisfies

$$\mathbf{Pr}\left[|\hat{\mu} - \mu| > 2\sigma \left(\frac{1}{\delta}\right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}}\right] \le \frac{\delta\, \mathbb{E}\left[|\hat{\mu} - \mu|^{2\lceil \frac{n}{2} \rceil}\right]}{(2\sigma)^{2\lceil \frac{n}{2} \rceil}} \le \frac{\delta \binom{n}{\lceil \frac{n}{2} \rceil}}{2^{2\lceil \frac{n}{2} \rceil}} < \frac{\delta 2^n}{2^{2\lceil \frac{n}{2} \rceil}} \le \delta,$$

where the second-to-last inequality uses a simple upper bound on centered binomial coefficients.

Now consider the (well studied) case that $n$ is at least $t$. In this case, Algorithm 1 returns the median of the $t$ bucket means. Each bucket contains at least $\lfloor \frac{n}{t} \rfloor$ samples, so the variance of each bucket mean is at most $\frac{\sigma^2}{\lfloor n/t \rfloor}$. Thus, Markov's inequality applied together with moment bound of Theorem 3.4 tells us that our output $\hat{\mu} = \mathrm{Median}(\hat{\mu}^{(1)}, \ldots, \hat{\mu}^{(t)})$ satisfies

$$\mathbf{Pr}\left[|\hat{\mu} - \mu| > 2\sigma \sqrt{e\left(\frac{1}{\lfloor n/t \rfloor}\right)}\right] \le \left(\frac{\lfloor n/t \rfloor}{4e\sigma^2}\right)^{\lceil \frac{t}{2} \rceil} \binom{t}{\lceil \frac{t}{2} \rceil} \max_{i \in [t]} \mathbb{E}\left[(\hat{\mu}^{(i)} - \mu)^2\right]^{\lceil \frac{t}{2} \rceil} \le e^{-\lceil \frac{t}{2} \rceil} \le \delta.$$

The above implies the second bound in (10) since $\lfloor \frac{n}{t} \rfloor \ge \frac{n}{4\log(1/\delta)}$. This completes the proof. □

Applying the above median-of-means estimator to each coordinate with failure parameter $\delta/d$ together with a union bound gives us the following corollary for estimation under the $\ell_\infty$ norm.

**Corollary A.3.** *Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be i.i.d. random vectors with finite mean $\mu$ and covariance $\sigma^2 I_d$. Fix a confidence parameter $\delta \in (0,1)$. Given the number of buckets $t = 2\lceil \log(d/\delta) \rceil$ and the sample $X_1, \ldots, X_n$ as input, Algorithm 1 returns an estimate $\hat{\mu}$ satisfying*

$$\|\hat{\mu} - \mu\|_\infty \le \begin{cases} 2\sigma \left(\frac{d}{\delta}\right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}}, & n < 2\lceil \log\left(\frac{d}{\delta}\right) \rceil, \\ 4\sigma \sqrt{\frac{e}{n} \log\left(\frac{d}{\delta}\right)}, & n \ge 2\lceil \log\left(\frac{d}{\delta}\right) \rceil, \end{cases} \tag{11}$$

*with probability at least $1 - \delta$ over the randomness of $X_1, \ldots, X_n$.*

### A.2 High-probability lower bound in $\ell_\infty$

In this sub-section we show a high-probability lower bound for any estimator in $\ell_\infty$. We prove our lower bound in two regimes: the small sample regime ($n \le \log(d/\delta)$) in Theorem A.4 and the large sample regime ($n \ge \log(d/\delta)$) in Lemma A.5.

**Theorem A.4.** *For any $\delta \in (0, 1/2)$, any $\sigma > 0$, any $n, d \in \mathbb{N}$ and any deterministic estimator $\mathcal{A} : (\mathbb{R}^d)^n \to \mathbb{R}^d$, there exists a distribution $P$ on $\mathbb{R}^d$ with mean $\mu$ and covariance matrix $\sigma^2 I_d$ such that*

$$\Pr_{X_1,\dots,X_n \sim P}\left[\|\mathcal{A}(X_1,\dots,X_n) - \mu\|_\infty \ge \frac{\sigma(1-2p)}{4}\left(\frac{d}{\delta}\right)^{\frac{1}{2\lceil \frac{n}{2}\rceil}}\right] > \delta,$$

*where $p = (\frac{4\delta}{d})^{\frac{1}{\lceil \frac{n}{2}\rceil}}$.*

*Proof.* If $p \ge 1/2$, then $1 - 2p \le 0$, so the lower bound in the theorem is non-positive. In this case the claim is trivial since the norm is always non-negative and the claim follows for any distribution $P$ with covariance $\sigma^2 I_d$. Thus, we assume that $p < 1/2$ throughout. Define

$$\ell = \frac{\sigma}{\sqrt{p(1-p)}}.$$

We consider two distributions $P_0$ and $P_1$ on $\{0, \ell\}$ given by

$$\Pr_{Y\sim P_0}[Y = \ell] = p, \qquad \Pr_{Y\sim P_1}[Y = \ell] = 1 - p.$$

Equivalently, under $P_0$ the point $\ell$ occurs with probability $p$, while under $P_1$ it occurs with probability $1 - p$. Write $\mu_0 = \mathbb{E}_{Y\sim P_0}[Y] = p\ell$ and $\mu_1 = \mathbb{E}_{Y\sim P_1}[Y] = (1-p)\ell$ for the mean of $P_0$ and $P_1$ respectively. It is immediate that both distributions have variance $p(1-p)\ell^2 = \sigma^2$. Moreover, we define the separation of the means to be

$$\Delta = \mu_1 - \mu_0 = (1 - 2p)\ell,$$

and since $p < 1/2$, we have $\Delta > 0$. We now construct the hard distribution. For each $b = (b_1, \dots, b_d) \in \{0,1\}^d$, let $P_b$ be the product distribution on $\mathbb{R}^d$ whose $j$th coordinate has distribution $P_{b_j}$, independently over $j = 1, \dots, d$. Notice that $P_b$ has mean $\mu_b$ and covariance matrix $\sigma^2 I_d$ where $\mu_b = (\mu_{b_1}, \dots, \mu_{b_d})$. We will show that for any deterministic estimator $\mathcal{A}$, there exists *some* choice of $b \in \{0,1\}^d$ witnessing the lower bound. Given samples $X_1, \dots, X_n \in \mathbb{R}^d$, define a decoder $\widehat{b} = \widehat{b}(X_1, \dots, X_n) \in \{0,1\}^d$ coordinate-wise by

$$\widehat{b}_j \in \arg\min_{u\in\{0,1\}} \left|\mathcal{A}(X_1,\dots,X_n)_j - \mu_u\right|,$$

with ties broken arbitrarily. Now, if $\|\mathcal{A}(X_1,\dots,X_n) - \mu_b\|_\infty < \frac{\Delta}{2}$, then every coordinate of $\mathcal{A}(X_1,\dots,X_n)$ is strictly closer to the correct mean than to the incorrect mean, hence

$$\|\mathcal{A}(X_1,\dots,X_n) - \mu_b\|_\infty < \frac{\Delta}{2} \quad\Longrightarrow\quad \widehat{b} = b. \tag{12}$$

Let $U$ be the uniform distribution over $\{0,1\}^d$ and let $B = (B_1, \dots, B_d) \sim U$. Conditioned on $B = b$, let $X_1, \dots, X_n$ be i.i.d. samples from $P_b$. For each coordinate $j$, define the coordinate samples

$$Z_j = (X_{1,j}, X_{2,j}, \dots, X_{n,j}) \in \mathcal{Z},$$

where $\mathcal{Z} := \{0, \ell\}^n$. Conditioned on $B = b$, the random vectors $Z_1, \dots, Z_d$ are independent and $Z_j$ follows $P_{b_j}^n$. Finally define $\eta = \frac{1}{2}\sum_{z\in\{0,\ell\}^n} \min\{P_0^n(z), P_1^n(z)\}$, which, by restricting the sum to those $z$ satisfying $|\{i : z_i = \ell\}| = \lfloor n/2 \rfloor$, can be lower bounded as follows:

$$\eta \ge \frac{1}{2}\binom{n}{\lfloor \frac{n}{2}\rfloor} \min\{p^{\lceil \frac{n}{2}\rceil}(1-p)^{\lfloor \frac{n}{2}\rfloor}, (1-p)^{\lceil \frac{n}{2}\rceil}p^{\lfloor \frac{n}{2}\rfloor}\} \ge \frac{1}{2}\binom{n}{\lfloor \frac{n}{2}\rfloor} p^{\lceil \frac{n}{2}\rceil} 2^{-\lfloor \frac{n}{2}\rfloor} \ge \frac{1}{2}p^{\lceil \frac{n}{2}\rceil} = \frac{2\delta}{d}, \tag{13}$$

where the second inequality uses the fact that $p < 1/2$ twice, first to say the first term in the minimum is no larger than the second and then lower bound $1-p$ by $1/2$, the final inequality uses the well known

inequality $\binom{n}{r} \geq (\frac{n}{r})^r$ for $n \geq r \geq 1$ to conclude that $2^{-\lfloor \frac{n}{2} \rfloor}\binom{n}{\lfloor \frac{n}{2} \rfloor} \geq 2^{-\lfloor \frac{n}{2} \rfloor}(\frac{n}{\lfloor \frac{n}{2} \rfloor})^{\lfloor \frac{n}{2} \rfloor} \geq 1$,[5] and the equality follows from the definition of $p$. We now claim that *every* decoder $\phi : \left(\{0,\ell\}^n\right)^d \to \{0,1\}^d$ satisfies

$$\mathop{\mathbf{Pr}}_{B\sim U,\ X_1,\ldots,X_n\sim P_B}[\phi(Z) = B] \leq (1-\eta)^d.$$

Indeed, writing $Z = (Z_1, \ldots, Z_d)$ and summing over the sample space we have that

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{B\sim U,\ X_1,\ldots,X_n\sim P_B}[\phi(Z) = B] &= 2^{-d} \sum_{b\in\{0,1\}^d} \sum_{z\in\mathcal{Z}^d} \mathbf{1}\{\phi(z) = b\} \prod_{j=1}^{d} P_{b_j}^n(z_j) \\
&= 2^{-d} \sum_{z\in\mathcal{Z}^d} \prod_{j=1}^{d} P_{\phi(z)_j}^n(z_j) \\
&\leq 2^{-d} \sum_{z\in\mathcal{Z}^d} \prod_{j=1}^{d} \max_{u\in\{0,1\}} P_u^n(z_j) \\
&= 2^{-d} \sum_{z_1\in\mathcal{Z}} \cdots \sum_{z_d\in\mathcal{Z}} \prod_{j=1}^{d} \max_{u\in\{0,1\}} P_u^n(z_j) \\
&= \left(\frac{1}{2} \sum_{w\in\mathcal{Z}} \max_{u\in\{0,1\}} P_u^n(w)\right)^d.
\end{aligned}$$

Now notice that

$$\frac{1}{2} \sum_{w\in\mathcal{Z}} \max\{P_0^n(w), P_1^n(w)\} = \frac{1}{2} \sum_{w\in\mathcal{Z}} (P_0^n(w) + P_1^n(w) - \min\{P_0^n(w), P_1^n(w)\}) = 1 - \eta,$$

so,

$$\mathop{\mathbf{Pr}}_{B\sim U,\ X_1,\ldots,X_n\sim P_B}[\phi(Z) = B] \leq (1-\eta)^d.$$

The above upper bound certainly holds for our decoder $\widehat{b}$, so an averaging argument implies the existence of *some* $b^\star \in \{0,1\}^d$ such that

$$\mathop{\mathbf{Pr}}_{X_1,\ldots,X_n\sim P_{b^\star}}\left[\widehat{b}(Z) = b^\star\right] \leq (1-\eta)^d \leq \left(1 - \frac{2\delta}{d}\right)^d \leq e^{-2\delta} < 1 - \delta,$$

where we used the lower bound on $\eta$ from (13) and the inequality $e^{-2\delta} \leq 1 - \delta$ that holds for $\delta \in (0, \frac{1}{2})$. Now, using the above with (12), we conclude that

$$\mathop{\mathbf{Pr}}_{X_1,\ldots,X_n\sim P_{b^\star}}\left[\left\|\mathcal{A}(X_1,\ldots,X_n) - \mu_{b^\star}\right\|_\infty \geq \frac{\Delta}{2}\right] \geq \mathop{\mathbf{Pr}}_{X_1,\ldots,X_n\sim P_{b^\star}}\left[\widehat{b}(Z) \neq b^\star\right] > \delta.$$

All that remains is to lower bound $\Delta/2$. We have

$$\begin{aligned}
\frac{\Delta}{2} &= \frac{\sigma(1-2p)}{2\sqrt{p(1-p)}} \\
&\geq \frac{\sigma(1-2p)}{2} \left(\frac{d}{4\delta}\right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}} \\
&\geq \frac{\sigma(1-2p)}{4} \left(\frac{d}{\delta}\right)^{\frac{1}{2\lceil \frac{n}{2} \rceil}}.
\end{aligned}$$

where we used the facts $p = \left(\frac{4\delta}{d}\right)^{\frac{1}{\lceil \frac{n}{2} \rceil}}$, $\sqrt{\frac{1}{1-p}} \geq 1$ and $4^{-\frac{1}{2\lceil \frac{n}{2} \rceil}} \geq \frac{1}{2}$. This completes the proof. $\square$

[5] If $n = 1$, then the claimed lower bound trivially holds since $2^{-\lfloor \frac{n}{2} \rfloor} = 1$ and $\binom{n}{\lfloor \frac{n}{2} \rfloor} = \binom{1}{0} = 1$ by convention.

**Lemma A.5** (Gaussian lower bound for the large-sample regime)**.** *There exists a universal constant $c > 0$ such that for every $n, d \in \mathbb{N}$, every $\sigma > 0$, and every $\delta \in (0, 1/4)$,*

$$r(n, d, \sigma, \delta) \geq c\sigma\sqrt{\frac{\log(d/\delta)}{n}}.$$

*Proof.* Note that it suffices to prove the lower bound on the Gaussian distribution:

$$X_1, \ldots, X_n \mid \mu \sim \mathcal{N}(\mu, \sigma^2 I_d)^{\otimes n}, \qquad \mu \in \mathbb{R}^d.$$

Fix an arbitrary estimator

$$\mathcal{A} : (\mathbb{R}^d)^n \to \mathbb{R}^d.$$

Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

For a parameter $s > 0$ and assume the Gaussian prior on $\mu$:

$$\mu \sim \mathcal{N}(0, s^2 I_d).$$

Under this prior and the Gaussian observation model above, $\bar{X}$ is a sufficient statistic for $\mu$. Concretely, the posterior distribution of $\mu$ given the full sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ is

$$\mu \mid \boldsymbol{X} \sim \mathcal{N}\left(\alpha_s \bar{X},\, \tau_s^2 I_d\right),$$

where

$$\alpha_s = \frac{s^2}{s^2 + \sigma^2/n}, \qquad \tau_s^2 = \left(\frac{1}{s^2} + \frac{n}{\sigma^2}\right)^{-1} = \frac{s^2\sigma^2}{\sigma^2 + ns^2}.$$

Note, in particular, that:

$$\tau_s \longrightarrow \frac{\sigma}{\sqrt{n}} \qquad \text{as } s \to \infty.$$

Fix $r > 0$, and define the cube

$$K_r \coloneqq \{z \in \mathbb{R}^d : \|z\|_\infty < r\}.$$

Observe that $K_r$ is convex and centrally symmetric.

Now condition on the realized sample $\boldsymbol{X}$. The conditional success probability of the estimator is

$$\mathbf{Pr}\left[\|\mathcal{A}(\boldsymbol{X}) - \mu\|_\infty < r \mid \boldsymbol{X}\right] = \mathbf{Pr}\left[\mu \in \mathcal{A}(\boldsymbol{X}) + K_r \mid \boldsymbol{X}\right].$$

Since

$$\mu \mid \boldsymbol{X} \sim \mathcal{N}(\alpha_s \bar{X}, \tau_s^2 I_d),$$

this conditional success probability of the estimator is

$$\mathbf{Pr}\left[\alpha_s \bar{X} + \tau_s G \in \mathcal{A}(\boldsymbol{X}) + K_r \,\middle|\, \boldsymbol{X}\right] = \mathbf{Pr}\left[\tau_s G \in \mathcal{A}(\boldsymbol{X}) - \alpha_s \bar{X} + K_r \,\middle|\, \boldsymbol{X}\right],$$

where $G \sim \mathcal{N}(0, I_d)$. By Anderson's inequality, for a centered Gaussian vector and a convex centrally symmetric set $K_r$, the Gaussian measure of a translate of $K_r$ is maximized at the centered translate [And55]. Hence

$$\mathbf{Pr}\left[\tau_s G \in \mathcal{A}(\boldsymbol{X}) - \alpha_s \bar{X} + K_r \,\middle|\, \boldsymbol{X}\right] \leq \mathbf{Pr}\left[\tau_s G \in K_r \mid \boldsymbol{X}\right].$$

Averaging over $X$ under the joint prior-predictive distribution gives

$$\mathbf{Pr}\left[\|\mathcal{A}(\boldsymbol{X}) - \mu\|_\infty < r\right] \leq \mathbf{Pr}\left[\tau_s \|G\|_\infty < r\right].$$

And consequently,

$$\mathbf{Pr}\left[\|\mathcal{A}(\boldsymbol{X}) - \mu\|_\infty \geq r\right] \geq \mathbf{Pr}\left[\tau_s \|G\|_\infty \geq r\right].$$

Letting $s \to \infty$, and using continuity of the distribution of $\|G\|_\infty$, yields

$$\sup_{\mu \in \mathbb{R}^d} \mathop{\mathbf{Pr}}_{X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2 I_d)} \left[\|\mathcal{A}(X_1, \ldots, X_n) - \mu\|_\infty \geq r\right] \geq \mathbf{Pr}\left[\frac{\sigma}{\sqrt{n}} \|G\|_\infty \geq r\right]. \tag{14}$$

It remains to choose $r$. We use the following standard Gaussian maximum lower bound:

**Claim A.6.** There exists a universal constant $c_0 > 0$ such that for every $d \in \mathbb{N}$ and every $\delta \in (0, 1/4)$,

$$\mathbf{Pr}\left[\|G\|_\infty \geq c_0\sqrt{\log(d/\delta)}\right] \geq \delta.$$

*Proof.* Let

$$L = \log(d/\delta), \qquad t = c_0\sqrt{L}, \qquad p_t = \mathbf{Pr}(|g| \geq t),$$

where $g \sim \mathcal{N}(0,1)$. If $t \leq 1$, then

$$\mathbf{Pr}[\|G\|_\infty \geq t] \geq \mathbf{Pr}[|g| \geq 1] > \delta,$$

as $\mathbf{Pr}[|g| \geq 1] \geq 1/4$ and $\delta < 1/4$.

If $t > 1$, the Gaussian tail lower bound gives

$$p_t \geq c_1 t^{-1} e^{-t^2/2}$$

for a universal constant $c_1 > 0$. Thus

$$dp_t \geq \frac{c_1 d}{c_0\sqrt{L}} \exp\left(-\frac{c_0^2}{2}L\right) = \frac{c_1\delta}{c_0\sqrt{L}}\left(\frac{d}{\delta}\right)^{1-c_0^2/2}.$$

Choosing $c_0 > 0$ sufficiently small ensures that $dp_t \geq 2\delta$ for all $d \geq 1$ and $\delta \in (0, 1/4)$. Using the independence of the coordinates yields,

$$\mathbf{Pr}[\|G\|_\infty \geq t] = 1 - (1-p_t)^d \geq 1 - e^{-dp_t} \geq 1 - e^{-2\delta} \geq \delta.$$

This proves the Gaussian maximum lower bound. □

Now setting

$$r = c_0\sigma\sqrt{\frac{\log(d/\delta)}{n}}.$$

with Claim A.6 and Equation 14 concludes the proof. □

**Proof of Theorem A.1**

Write $L = \log(d/\delta)$. We first prove the upper bound. In the isotropic case, Corollary A.3 gives, for the coordinate-wise median-of-means estimator with $t = 2\lceil L\rceil$ buckets

$$r(n,d,\delta) \lesssim \begin{cases} \left(\frac{d}{\delta}\right)^{\frac{1}{2\lceil \frac{n}{2}\rceil}}, & n < 2\lceil L\rceil, \\ \sqrt{\frac{L}{n}}, & n \geq 2\lceil L\rceil. \end{cases}$$

This implies the claimed upper bound in both regimes. If $n \lesssim L$, then the first line is exactly the desired bound when it applies, while if the second line applies then $\sqrt{L/n} \leq 1 \leq (d/\delta)^{1/(2\lceil \frac{n}{2}\rceil)}$. If $n \gtrsim L$, then the second line is exactly the desired bound when it applies, while if the first line applies then $n = \Theta(L)$, and the two rates are comparable up to universal constants.

We now prove the lower bound. First, consider the large sample regime $n \gtrsim L$. We can invoke the Gaussian lower bound (Lemma A.5) in the isotropic to get

$$r(n,d,\delta) \gtrsim \sqrt{\frac{L}{n}},$$

as claimed.

It remains to prove the lower bound in the small sample regime. Assume that $n \lesssim L$. By taking the universal constant small in the theorem statement to be small enough, we may assume that $\delta < 1/4$ and that $L \geq 2\log 4$. Set

$$a = \left(\frac{4\delta}{d}\right)^{1/\lceil \frac{n}{2}\rceil}.$$

If $a \leq 1/4$, then $1-2a \geq 1/2$, and Theorem A.4 implies that for every estimator there is a distribution with mean $\mu$ and covariance $I_d$ such that

$$\mathbf{Pr}\left[\|\mathcal{A}(X_1,\ldots,X_n)-\mu\|_\infty \geq \frac{1}{8}\left(\frac{d}{\delta}\right)^{\frac{1}{2\left\lceil\frac{n}{2}\right\rceil}}\right] > \delta.$$

Hence, by the definition of the minimax risk, we have

$$r(n,d,\delta) \gtrsim \left(\frac{d}{\delta}\right)^{\frac{1}{2\left\lceil\frac{n}{2}\right\rceil}}.$$

On the other hand if $a > 1/4$, then

$$\left(\frac{d}{\delta}\right)^{\frac{1}{2\left\lceil\frac{n}{2}\right\rceil}} \asymp 1,$$

which we can combine with Lemma A.5 and our assumption that $n \lesssim L$ to conclude that

$$r(n,d,\delta) \gtrsim \sqrt{\frac{L}{n}} \gtrsim 1 \asymp \left(\frac{d}{\delta}\right)^{\frac{1}{2\left\lceil\frac{n}{2}\right\rceil}}.$$

Combining the upper and lower bounds completes the proof. □

# B Technical proof

**Lemma B.1.** *Let $f_1,\ldots,f_t : \mathbb{R}_+ \to \mathbb{R}_+$ be $t$, measurable, non-increasing and integrable functions. Then*

$$\int_0^\infty \prod_{i=1}^t f_i(x^{1/t})dx \leq \prod_{i=1}^t \int_0^\infty f_i(x)dx$$

*Proof.* We have that

$$\begin{aligned}
\prod_{i=1}^t \int_0^\infty f_i(x)dx &= \prod_{i=1}^t \int_0^\infty f_i(x_i)dx_i \\
&= \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^t f_i(x_i)dx_1\ldots dx_t \\
&= \sum_{\substack{x_{i_1}<\ldots<x_{i_t} \\ \{i_1,\ldots,i_t\}=[t]}} \int_0^\infty \int_0^{x_{i_t}} \cdots \int_0^{x_{i_2}} \prod_{i=1}^t f_i(x_i)dx_{i_1}\ldots dx_{i_{t-1}}dx_{i_t} \\
&\geq \sum_{\substack{x_{i_1}<\ldots<x_{i_t} \\ \{i_1,\ldots,i_t\}=[t]}} \int_0^\infty \prod_{i=1}^t f_i(x_{i_t}) \int_0^{x_{i_t}} \cdots \int_0^{x_{i_2}} 1dx_{i_1}\ldots dx_{i_{t-1}}dx_{i_t} \\
&= t!\int_0^\infty \prod_{i=1}^t f_i(x)\frac{x^{t-1}}{(t-1)!}dx \\
&= \int_0^\infty \prod_{i=1}^t f_i(x^{1/t})dx,
\end{aligned}$$

where the third equality follows from decomposing the integration domain $\mathbb{R}_+^t$ into the $t!$ disjoint regions corresponding to the possible orderings $x_{i_1} < \cdots < x_{i_t}$ of the coordinates. The inequality follows from the fact that each $f_i$ is non-increasing, so on any region where $x_{i_1} < \cdots < x_{i_t}$ we have $f_i(x_i) \geq f_i(x_{i_t})$ for every $i \in [t]$. The second last equality follows from the fact that there are $t!$ such orderings, together with an explicit evaluation of the inner integrals. The final equality follows from a change of variables. □

**Theorem 3.4.** *Let $X_1, \dots, X_t$ be $t$ independent random variables with a common mean $\mu$ and let $M = M(X_1, \dots, X_t)$ denote their median. For all $q \geq 0$ it holds that*

$$\mathbb{E}\left[|M-\mu|^{q\left\lceil \frac{t}{2}\right\rceil}\right] \leq \binom{t}{\left\lceil \frac{t}{2}\right\rceil} \max_{i\in[t]} \mathbb{E}\left[|X_i-\mu|^q\right]^{\left\lceil \frac{t}{2}\right\rceil}.$$

*Proof.* The case $q = 0$ follows from the left hand side of the inequality to be less than or equal to the right hand side of the inequality. Furthermore, if the right hand side is infinite then the inequality also holds. Thus, we can assume that $q > 0$ and that the right hand side is finite, meaning all the expectations $\mathbb{E}\left[|X_i - \mu|^q\right]$ are finite. By calculations we have that

$$\mathbb{E}\left[|M-\mu|^{q\left\lceil \frac{t}{2}\right\rceil}\right] = \int_0^\infty \mathbf{Pr}\left[|M-\mu|^{q\left\lceil \frac{t}{2}\right\rceil} \geq x\right] dx = \int_0^\infty \mathbf{Pr}\left[|M-\mu|^q \geq x^{\frac{1}{\left\lceil \frac{t}{2}\right\rceil}}\right] dx,$$

where we have used that the expectation of a positive random variable is the integral of its right tail. Now for the median $M$ of $X_1, \dots, X_t$ to be $x$ away from the mean $\mu$ it must be the case that there are $\left\lceil \frac{t}{2}\right\rceil$ $X_i$'s such that $|X_i - \mu| \geq x$. By this, a union bound and independence of $X_i$'s we then get that:

$$\begin{aligned}
\mathbb{E}\left[|M-\mu|^{q\left\lceil \frac{t}{2}\right\rceil}\right] &\leq \int_0^\infty \mathbf{Pr}\left[\exists I \in \binom{t}{\left\lceil \frac{t}{2}\right\rceil} : \forall i \in I \text{ such that } |X_i-\mu|^q \geq x^{\frac{1}{\left\lceil \frac{t}{2}\right\rceil}}\right] dx \\
&\leq \sum_{I\in\binom{t}{\left\lceil \frac{t}{2}\right\rceil}} \int_0^\infty \mathbf{Pr}\left[\forall i \in I : |X_i-\mu|^q \geq x^{\frac{1}{\left\lceil \frac{t}{2}\right\rceil}}\right] dx \\
&= \sum_{I\in\binom{t}{\left\lceil \frac{t}{2}\right\rceil}} \int_0^\infty \prod_{i\in I} \mathbf{Pr}\left[|X_i-\mu|^q \geq x^{\frac{1}{\left\lceil \frac{t}{2}\right\rceil}}\right] dx.
\end{aligned}$$

Now using that for each $i$, the function $\mathbf{Pr}\left[|X_i - \mu|^q \geq x\right]$ is non-increasing in $x$, integrable by considering the case where the expectation is finite for all $X_i$, and measureable by $X_i$ being a random variable, we can use Lemma B.1 to get that

$$\begin{aligned}
\mathbb{E}\left[|M-\mu|^{q\left\lceil \frac{t}{2}\right\rceil}\right] &\leq \sum_{I\in\binom{t}{\left\lceil \frac{t}{2}\right\rceil}} \prod_{i\in I} \int_0^\infty \mathbf{Pr}\left[|X_i-\mu|^q \geq x\right] dx \\
&= \sum_{I\in\binom{t}{\left\lceil \frac{t}{2}\right\rceil}} \prod_{i\in I} \mathbb{E}\left[|X_i-\mu|^q\right] \\
&\leq \binom{t}{\left\lceil \frac{t}{2}\right\rceil} \max_{i\in[t]} \mathbb{E}\left[|X_i-\mu|^q\right]^{\left\lceil \frac{t}{2}\right\rceil}.
\end{aligned}$$

where the first inequality follows from Lemma B.1, the first equality follows from the fact that the expectation of a positive random variable is the integral of its right tail, and the last inequality follows from taking max under the product. □

## C Proof of Theorem 1.2

In this section, we establish that projection-based Median-of-Means estimators are provably sub-optimal for $\ell_p$ norms when $p > 2$. Let $p > 2$ and $q < 2$ denote the order of the corresponding dual ball. The projection-based median is now defined as follows for a pointset $Y = \{y_i\}_{i=1}^n$ with $y_i \in \mathbb{R}^d$:

$$\mathrm{Med}_{\mathrm{proj}}(Y) = \arg\min_x \max_{v:\|v\|_q=1} |\langle v, x\rangle - \mathrm{Med}\left(\{\langle v, y_i\rangle\}_{i=1}^n\right)|. \tag{PROJ-MED}$$

**Theorem 1.2** (Suboptimality of Projection-based median-of-means)**.** *Let $p > 2$. There exist absolute constants $C, c > 0$ such that the following hold. Let $d, n \in \mathbb{N}$ be such that $n \leq c\sqrt{d}$ and $d > C$. Then, there exists a distribution $D$ with mean $0$ and covariance $I_d$ such that the projection-based median-of-means estimator with any bucket size $m$ satisfying $n/m \geq 3$, $\widehat{\mu}_{\mathrm{Proj}}$, satisfies:*

$$\mathop{\mathbf{Pr}}_{\boldsymbol{X}\sim D^{\otimes n}}\left[\left\|\widehat{\mu}_{\mathrm{Proj}}\left(\boldsymbol{X}\right)\right\|_p \geq \frac{\sqrt{d}}{2n^{1-1/p}}\right] \geq 0.9.$$

**Remark C.1.** Firstly, observe that the lower bound in Theorem 1.2 directly translates to an in-expectation lower bound for the projection-based median-of-means estimator. Secondly, observe that Theorem 3.3 when $\Sigma = I_d$ (and $n = \omega(\log(d))$) yields the in-expecation bound:

$$\mathbb{E}\left[\|\widehat{\mu} - \mu\|_p\right] \leq 4d^{1/p}\sqrt{\frac{\log d}{n}}.$$

Comparing this bound with Theorem 1.2, we see that the projection-based median-of-means estimator is strictly worse when $n = \widetilde{o}(\sqrt{d})$. In particular, for $p = \infty$, Theorem 1.2 incurs a *polynomial* dependence on $d$ while Theorem 3.3 is *logarithmic*.

We will show that the median-of-means estimator with the projection-based median defined above is sub-optimal for any choice of number of buckets, $t$, equivalently bucket size $m = n/t$, for the distribution $D$, corresponding to the uniform distribution over the signed and scaled basis vectors. Formally:

$$\forall i \in [d], s \in \pm 1 : \mathop{\mathbf{Pr}}_{X \sim D}\left(X = s\sqrt{d}e_i\right) = \frac{1}{2d}.$$

where $e_i$ denote the standard basis vectors in $\mathbb{R}^d$. Note that $D$ has mean 0 and covariance $I_d$.

Now, let $\boldsymbol{X} = \{X_i\}_{i=1}^n$ be $n$ iid samples from from $D$. Furthermore, from our definition of $D$, $X_i = s_i\sqrt{d}e_{c(i)}$ with $s_i = \pm 1$ and $c(i) \in [d]$. We first prove a simple lemma establishing that the $c(i)$ are unique with high-probability:

**Lemma C.2.** *Under the preconditions of Theorem 1.2, we have:*

$$\forall i, j \in [n], i \neq j : c(i) \neq c(j)$$

*with probability at least* 0.9.

*Proof.* We have:

$$\begin{aligned}\mathbf{Pr}\left(\forall i, j \in [n], i \neq j : c(i) \neq c(j)\right) &= \prod_{i=1}^{n}\left(1 - \frac{(i-1)}{d}\right) \geq \prod_{i=1}^{n}\left(1 - \frac{n}{d}\right) \\ &= \left(1 - \frac{n}{d}\right)^n \geq \exp\left(-\frac{2n}{d} \cdot n\right) \geq 0.9\end{aligned}$$

as $n \leq c\sqrt{d}$. □

Now, we condition on the uniqueness of $c(i)$. Define the bucketed means $Z_i$ as follows for $i \in [t]$ with $m \coloneqq n/t$:

$$Z_i = \frac{1}{m}\sum_{j=(i-1)m+1}^{im} X_j$$

with the collection denoted by $\boldsymbol{Z}$. From PROJ-MED, the projection-based median-of-means estimate is the minimizer of the following function:

$$F(x) \coloneqq \max_{v:\|v\|_q=1}\left|\langle v, x\rangle - \mathrm{Med}\left(\{\langle v, Z_i\rangle\}_{i=1}^t\right)\right|$$

We now make some observations:

**Observation C.3.** $F$ is a convex function of $x$.

*Proof.* This follows from $F$ being defined as the maximum of a set of convex functions. □

Next, $F$ satisfies the following symmetries:

**Observation C.4.** Defining:

$$\widetilde{X}_i = \sqrt{d}e_{c(i)},\ S \in \mathbb{R}^{d\times d} : S_{i,j} \coloneqq \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j, \forall \ell, c(\ell) \neq i \\ 1 & \text{if } i = j, \exists \ell, c(\ell) = i \text{ and } s_\ell = 1 \\ -1 & \text{otherwise} \end{cases},$$

$$\text{and } \widetilde{Z}_i \coloneqq \frac{1}{m} \sum_{j=(i-1)m+1}^{im} \widetilde{X}_j.$$

where uniqueness of $c(i)$ from Lemma C.2 ensures that $S$ is well-defined. Finally, defining $\widetilde{F}$ analogously to $F$, we observe the following:

$$\forall x : F(x) = \widetilde{F}(Sx).$$

Consequently,

$$\arg\min_x F(x) = S \arg\min_x \widetilde{F}(x).$$

*Proof.* From the definition of $S$ we have

$$S\tilde{X}_i = S\sqrt{d}e_{c(i)} = S_{c(i),c(i)}\sqrt{d}e_{c(i)} = s_i\sqrt{d}e_{c(i)} = X_i$$

which implies

$$\forall v, \|v\|_q = 1, i \in [t] : \langle v, Z_i \rangle = \langle Sv, \widetilde{Z}_i \rangle.$$

and hence

$$\begin{aligned} F(x) &= \max_{v:\|v\|_q=1} \left| \langle v, x \rangle - \text{Med}\left( \{\langle v, Z_i \rangle\}_{i=1}^t \right) \right| \\ &= \max_{v:\|v\|_q=1} |\langle Sv, Sx \rangle - \text{Med}\left( \left\{ \langle Sv, \tilde{Z}_i \rangle \right\}_{i=1}^t \right)| = \tilde{F}(Sx) \end{aligned}$$

where the last equality follows from the fact that $S$ is a diagonal matrix with entries in $\{\pm 1\}$, so $Sv$ ranges over all vectors with $\ell_q$ norm at most 1 as $v$ does. □

On the event of Lemma C.2, define the bucket supports

$$C_r \coloneqq \{c((r-1)m+1), \ldots, c(rm)\}, \qquad r \in [t].$$

Then the sets $C_1, \ldots, C_t$ are disjoint, $|C_r| = m$ for all $r \in [t]$, and

$$\widetilde{Z}_r = \frac{\sqrt{d}}{m} \sum_{j \in C_r} e_j.$$

Hence,

$$\widetilde{F}(x) = \max_{v:\|v\|_q=1} \left| \langle v, x \rangle - \text{Med}\left( \left\{ \frac{\sqrt{d}}{m} \sum_{j \in C_r} v_j \right\}_{r=1}^t \right) \right|.$$

**Lemma C.5.** *Fix $r \in [t]$, and let $P$ be any permutation matrix that permutes the coordinates in $C_r$ and fixes all coordinates outside $C_r$. Then*

$$\forall x \in \mathbb{R}^d : \widetilde{F}(Px) = \widetilde{F}(x).$$

*Proof.* Since

$$\widetilde{Z}_\ell = \frac{\sqrt{d}}{m} \sum_{j \in C_\ell} e_j$$

for every $\ell \in [t]$, permuting coordinates inside $C_r$ leaves every $\widetilde{Z}_\ell$ unchanged:

$$\forall \ell \in [t] : P\widetilde{Z}_\ell = \widetilde{Z}_\ell.$$

Therefore,

$$\begin{aligned} \widetilde{F}(Px) &= \max_{v:\|v\|_q=1} \left| \langle v, Px \rangle - \text{Med}\left( \left\{ \langle v, \widetilde{Z}_\ell \rangle \right\}_{\ell=1}^t \right) \right| \\ &= \max_{v:\|v\|_q=1} \left| \langle P^\top v, x \rangle - \text{Med}\left( \left\{ \langle P^\top v, \widetilde{Z}_\ell \rangle \right\}_{\ell=1}^t \right) \right|. \end{aligned}$$

Since $P$ is a permutation matrix, $\|P^\top v\|_q = \|v\|_q$. Hence, after the change of variables $u = P^\top v$,

$$\widetilde{F}(Px) = \max_{u:\|u\|_q=1} \left| \langle u, x \rangle - \text{Med}\left( \left\{ \langle u, \widetilde{Z}_\ell \rangle \right\}_{\ell=1}^t \right) \right| = \widetilde{F}(x).$$

□

**Corollary C.6.** *There exists a minimizer $x^\star$ of $\widetilde{F}$ such that for every $r \in [t]$, all coordinates of $x^\star$ in $C_r$ are equal.*

*Proof.* Let $\mathcal{G}_{\text{in}}$ denote the finite group of permutation matrices that permute coordinates inside each $C_r$ and fix all coordinates outside $\cup_{r=1}^t C_r$. For any minimizer $x$ of $\widetilde{F}$, define its symmetrization

$$\bar{x} \coloneqq \frac{1}{|\mathcal{G}_{\text{in}}|} \sum_{P \in \mathcal{G}_{\text{in}}} Px.$$

By convexity of $\widetilde{F}$ from Observation C.3 and Lemma C.5,

$$\widetilde{F}(\bar{x}) \le \frac{1}{|\mathcal{G}_{\text{in}}|} \sum_{P \in \mathcal{G}_{\text{in}}} \widetilde{F}(Px) = \frac{1}{|\mathcal{G}_{\text{in}}|} \sum_{P \in \mathcal{G}_{\text{in}}} \widetilde{F}(x) = \widetilde{F}(x).$$

Since $x$ is already a minimizer, $\bar{x}$ is also a minimizer.

By construction, $\bar{x}$ is invariant under every permutation inside each $C_r$. Hence,

$$\forall r \in [t],\ \forall i, j \in C_r : \bar{x}_i = \bar{x}_j.$$

□

**Lemma C.7.** *Let $\sigma$ be any permutation of $[t]$. For each $r \in [t]$, choose any bijection*

$$\phi_r : C_r \to C_{\sigma(r)}.$$

*Let $P_{\sigma,\phi}$ be the permutation matrix that sends each coordinate $j \in C_r$ to $\phi_r(j)$, and fixes all coordinates outside $\cup_{r=1}^t C_r$. Then*

$$\forall x \in \mathbb{R}^d : \widetilde{F}(P_{\sigma,\phi} x) = \widetilde{F}(x).$$

*Proof.* By construction,

$$P_{\sigma,\phi} \widetilde{Z}_r = \frac{\sqrt{d}}{m} \sum_{j \in C_r} P_{\sigma,\phi} e_j = \frac{\sqrt{d}}{m} \sum_{k \in C_{\sigma(r)}} e_k = \widetilde{Z}_{\sigma(r)}.$$

Therefore,

$$\begin{aligned}
\widetilde{F}(P_{\sigma,\phi} x) &= \max_{v : \|v\|_q = 1} \left| \langle v, P_{\sigma,\phi} x \rangle - \text{Med}\left( \left\{ \langle v, \widetilde{Z}_r \rangle \right\}_{r=1}^t \right) \right| \\
&= \max_{v : \|v\|_q = 1} \left| \langle P_{\sigma,\phi}^\top v, x \rangle - \text{Med}\left( \left\{ \langle P_{\sigma,\phi}^\top v, \widetilde{Z}_{\sigma(r)} \rangle \right\}_{r=1}^t \right) \right|.
\end{aligned}$$

Since $\sigma$ only reorders the multiset of bucket projections, the median is unchanged:

$$\text{Med}\left( \left\{ \langle P_{\sigma,\phi}^\top v, \widetilde{Z}_{\sigma(r)} \rangle \right\}_{r=1}^t \right) = \text{Med}\left( \left\{ \langle P_{\sigma,\phi}^\top v, \widetilde{Z}_r \rangle \right\}_{r=1}^t \right).$$

Since $P_{\sigma,\phi}$ is a permutation matrix $\|P_{\sigma,\phi}^\top v\|_q = \|v\|_q$. Hence, after the change of variables $u = P_{\sigma,\phi}^\top v$,

$$\widetilde{F}(P_{\sigma,\phi} x) = \max_{u : \|u\|_q = 1} \left| \langle u, x \rangle - \text{Med}\left( \left\{ \langle u, \widetilde{Z}_r \rangle \right\}_{r=1}^t \right) \right| = \widetilde{F}(x).$$

□

**Corollary C.8.** *There exists a minimizer $x^\star$ of $\widetilde{F}$ and a scalar $a \in \mathbb{R}$ such that*

$$\forall j \in \bigcup_{r=1}^t C_r : x_j^\star = a.$$

*Proof.* Start from the minimizer given by Corollary C.6, so that

$$\forall r \in [t],\ \forall j \in C_r : x_j = a_r$$

for some scalars $a_1, \dots, a_t$.

Let $\mathcal{G}_{\text{buck}}$ denote the finite set of all permutations $P_{\sigma,\phi}$ from Lemma C.7. Define

$$\bar{x} := \frac{1}{|\mathcal{G}_{\text{buck}}|} \sum_{P \in \mathcal{G}_{\text{buck}}} Px.$$

By convexity of $\widetilde{F}$ from Observation C.3 and Lemma C.7,

$$\widetilde{F}(\bar{x}) \le \frac{1}{|\mathcal{G}_{\text{buck}}|} \sum_{P \in \mathcal{G}_{\text{buck}}} \widetilde{F}(Px) = \widetilde{F}(x).$$

Hence $\bar{x}$ is also a minimizer.

The group $\mathcal{G}_{\text{buck}}$ acts transitively on the active coordinates $\cup_{r=1}^{t} C_r$: any active coordinate can be mapped to any other active coordinate by first permuting buckets and then using a bijection inside the target bucket. Therefore, $\bar{x}$ must take the same value on all active coordinates, i.e.,

$$\exists a \in \mathbb{R} \text{ such that } \forall j \in \bigcup_{r=1}^{t} C_r : \bar{x}_j = a.$$

□

**Lemma C.9.** *There exists a minimizer $x^\star$ of $\widetilde{F}$ such that*

$$\forall j \notin \bigcup_{r=1}^{t} C_r : x^\star_j = 0.$$

*Consequently, there exists a minimizer of the form*

$$x^\star = a \sum_{r=1}^{t} \sum_{j \in C_r} e_j$$

*for some scalar $a \in \mathbb{R}$.*

*Proof.* Fix $j \notin \cup_{r=1}^{t} C_r$, and let $D_j$ be the diagonal matrix defined by

$$(D_j)_{k,k} = \begin{cases} -1 & \text{if } k = j, \\ 1 & \text{otherwise.} \end{cases}$$

Since every $\widetilde{Z}_r$ is supported on $C_r$, we have

$$\forall r \in [t] : D_j \widetilde{Z}_r = \widetilde{Z}_r.$$

Hence, arguing exactly as in Lemma C.5,

$$\forall x \in \mathbb{R}^d : \widetilde{F}(D_j x) = \widetilde{F}(x).$$

Let $x$ be any minimizer of $\widetilde{F}$, and define

$$\bar{x} := \frac{x + D_j x}{2}.$$

By convexity of $\widetilde{F}$ from Observation C.3,

$$\widetilde{F}(\bar{x}) \le \frac{\widetilde{F}(x) + \widetilde{F}(D_j x)}{2} = \widetilde{F}(x).$$

Thus $\bar{x}$ is also a minimizer. Moreover, $\bar{x}$ agrees with $x$ on every coordinate except $j$, and

$$\bar{x}_j = 0.$$

Repeating this argument for every $j \notin \cup_{r=1}^{t} C_r$ yields a minimizer supported on $\cup_{r=1}^{t} C_r$.

Combining this with Corollary C.8 proves that there exists a minimizer of the form

$$x^\star = a \sum_{r=1}^{t} \sum_{j \in C_r} e_j = a \sum_{i=1}^{n} e_{c(i)}.$$

□

At this point, by Lemma C.9, it remains only to optimize over the scalar multiplying the active coordinates. Define

$$\rho := \left\lfloor \frac{t}{2} \right\rfloor + 1, \qquad \eta := t - \rho.$$

Let $y_{(1)} \geq \cdots \geq y_{(t)}$ denote the non-increasing rearrangement of a multiset of scalars $\boldsymbol{Y} = \{y_i\}_{i=1}^{t}$. We use the convention

$$\mathrm{Med}(\boldsymbol{Y}) = \begin{cases} y_{(k)} & \text{if } t = 2k-1, \\ \dfrac{y_{(k)} + y_{(k+1)}}{2} & \text{if } t = 2k. \end{cases}$$

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

$$x^{(a)} := a \sum_{i=1}^{n} e_{c(i)} = a \sum_{r=1}^{t} \sum_{j \in C_r} e_j.$$

**Lemma C.10.** *For every $a \in \mathbb{R}$,*

$$\widetilde{F}(x^{(a)}) = m^{1/p} \max_{u \in \mathbb{R}^t : \|u\|_q \leq 1} \left| a \sum_{r=1}^{t} u_r - \frac{\sqrt{d}}{m} \mathrm{Med}\left( \{u_r\}_{r=1}^{t} \right) \right|.$$

*Proof.* For $v \in \mathbb{R}^d$, define

$$u_r := m^{-1/p} \sum_{j \in C_r} v_j, \qquad r \in [t].$$

By Hölder,

$$|u_r| = m^{-1/p} \left| \sum_{j \in C_r} v_j \right| \leq \|v_{C_r}\|_q.$$

Hence,

$$\sum_{r=1}^{t} |u_r|^q \leq \sum_{r=1}^{t} \|v_{C_r}\|_q^q \leq \|v\|_q^q.$$

In particular, every $v$ with $\|v\|_q \leq 1$ produces a vector $u$ with $\|u\|_q \leq 1$. Moreover,

$$\sum_{j \in C_r} v_j = m^{1/p} u_r \qquad \text{for all } r \in [t].$$

Therefore,

$$\begin{aligned} \widetilde{F}(x^{(a)}) &= \max_{v : \|v\|_q = 1} \left| a \sum_{r=1}^{t} \sum_{j \in C_r} v_j - \frac{\sqrt{d}}{m} \mathrm{Med}\left( \left\{ \sum_{j \in C_r} v_j \right\}_{r=1}^{t} \right) \right| \\ &= m^{1/p} \max_{v : \|v\|_q = 1} \left| a \sum_{r=1}^{t} u_r - \frac{\sqrt{d}}{m} \mathrm{Med}\left( \{u_r\}_{r=1}^{t} \right) \right| \\ &\leq m^{1/p} \max_{u : \|u\|_q \leq 1} \left| a \sum_{r=1}^{t} u_r - \frac{\sqrt{d}}{m} \mathrm{Med}\left( \{u_r\}_{r=1}^{t} \right) \right|. \end{aligned}$$

For the reverse inequality, fix any $u \in \mathbb{R}^t$ with $\|u\|_q \le 1$, and define $v \in \mathbb{R}^d$ by setting

$$\forall r \in [t],\ \forall j \in C_r : v_j \coloneqq m^{-1+1/p} u_r,$$

and $v_j \coloneqq 0$ outside $\cup_{r=1}^t C_r$. Then

$$\sum_{j \in C_r} v_j = m^{1/p} u_r \qquad \text{for all } r \in [t],$$

and

$$\|v\|_q^q = \sum_{r=1}^t \sum_{j \in C_r} m^{-q+q/p} |u_r|^q = \sum_{r=1}^t m^{(1-q)+q/p} |u_r|^q = \sum_{r=1}^t |u_r|^q = \|u\|_q^q \le 1,$$

since $(1-q) = -q/p$ as $1/p + 1/q = 1$. Plugging $v$ into the definition of $\widetilde{F}$ now yields the reverse inequality, concluding the proof of the lemma. □

**Lemma C.11.** *For every $a \in \mathbb{R}$,*

$$\widetilde{F}(x^{(a)}) = m^{1/p} \left( \rho \left| \frac{\sqrt{d}}{\rho m} - a \right|^p + \eta |a|^p \right)^{1/p}.$$

*Proof.* Let

$$\lambda \coloneqq \frac{\sqrt{d}}{m}.$$

By Lemma C.10, it suffices to compute

$$\max_{\|u\|_q \le 1} \left| a \sum_{r=1}^t u_r - \lambda \mathrm{Med}\left( \{u_r\}_{r=1}^t \right) \right|.$$

Since $\mathrm{Med}(-u) = -\mathrm{Med}(u)$, this maximum equals

$$\max_{\|u\|_q \le 1} \left\{ \lambda \mathrm{Med}\left( \{u_r\}_{r=1}^t \right) - a \sum_{r=1}^t u_r \right\}.$$

Let

$$u_{(1)} \ge u_{(2)} \ge \cdots \ge u_{(t)}$$

be the non-increasing rearrangement of $u$. Now, observe that:

$$\mathrm{Med}(u) \le \frac{1}{\rho} \sum_{r=1}^{\rho} u_{(r)}.$$

Therefore,

$$\lambda \mathrm{Med}(u) - a \sum_{r=1}^t u_r \le \frac{\lambda}{\rho} \sum_{r=1}^{\rho} u_{(r)} - a \sum_{r=1}^t u_{(r)} = \left( \frac{\lambda}{\rho} - a \right) \sum_{r=1}^{\rho} u_{(r)} - a \sum_{r=\rho+1}^t u_{(r)}.$$

By Hölder,

$$\lambda \mathrm{Med}(u) - a \sum_{r=1}^t u_r \le \left( \rho \left| \frac{\lambda}{\rho} - a \right|^p + \eta |a|^p \right)^{1/p} \|u\|_q.$$

Since $\|u\|_q \le 1$, this establishes an upper bound.

For the lower bound, set

$$A \coloneqq \frac{\lambda}{\rho} - a, \qquad B \coloneqq -a, \qquad \alpha \coloneqq \left( \rho |A|^p + \eta |B|^p \right)^{-1/q}, \text{ and } J(z) \coloneqq \mathrm{sgn}(z) |z|^{p-1}.$$

Now, define $u \in \mathbb{R}^t$ by

$$u_1 = \cdots = u_\rho = \alpha J(A), \qquad u_{\rho+1} = \cdots = u_t = \alpha J(B).$$

Since $A - B > 0$, and $J$ is increasing, this vector is non-increasing. Also, observe

$$\|u\|_q^q = \alpha^q \left(\rho|A|^p + \eta|B|^p\right) = 1 \text{ and } \mathrm{Med}(u) = \alpha J(A).$$

Hence

$$\begin{aligned}\lambda \mathrm{Med}(u) - a\sum_{r=1}^{t} u_r &= \lambda\alpha J(A) - a\left(\rho\alpha J(A) + \eta\alpha J(B)\right) = \alpha\left(\rho A J(A) + \eta B J(B)\right) \\ &= \alpha\left(\rho|A|^p + \eta|B|^p\right) = \left(\rho|A|^p + \eta|B|^p\right)^{1/p}.\end{aligned}$$

matching the upper bound and proving the lemma. □

**Corollary C.12.** *The optimal scalar is*

$$a^\star = \frac{\sqrt{d}}{\rho m} \cdot \frac{1}{1 + (\eta/\rho)^{1/(p-1)}}.$$

*Consequently, a minimizer of $\widetilde{F}$ is*

$$x^\star = a^\star \sum_{i=1}^{n} e_{c(i)},$$

*and hence, by Observation C.4, a minimizer of $F$ is*

$$\widehat{\mu}_{\mathrm{proj}}(\boldsymbol{X}) = Sx^\star = a^\star \sum_{i=1}^{n} s_i e_{c(i)}.$$

*Proof.* Observing that $\widetilde{F}(x^{(a)})$ is convex in $a$ and computing its derivative, the corollary follows. □

**Corollary C.13.** *The resulting estimator has*

$$\|\widehat{\mu}_{\mathrm{proj}}(\boldsymbol{X})\|_p = \|x^\star\|_p = a^\star n^{1/p} \geq \frac{\sqrt{d}}{2n^{1-1/p}}.$$

*Proof.* Since $x^\star$ has exactly $n$ nonzero coordinates, each of magnitude $a^\star$,

$$\|x^\star\|_p = a^\star n^{1/p}.$$

The corollary follows from Corollary C.12. □

*Proof of Theorem 1.2.* The theorem follows from Corollary C.13 and Lemma C.2. □

## D Proof of Theorem 1.4

In this section, we show that Algorithm 1 cannot be optimal for the $\ell_2$ norm. The optimal rate for heavy-tailed mean estimation for isotropic data in this setting is known to be of order

$$\sqrt{\frac{d}{n}} + \sqrt{\frac{1}{n}\log\left(\frac{1}{\delta}\right)}. \tag{15}$$

We prove the following lower bound which demonstrates that the coordinate-wise median-of-means strategy is sub-optimal when the failure probability is small.

**Theorem 1.4** (Suboptimality of coordinate-wise median-of-means)**.** *There are universal constants $c, C \geq 0$ such that the following holds. For any $\delta \in (0, c)$ and $n, d, t \in \mathbb{N}$ with $t \leq n$, there exists a distribution $P$ over $\mathbb{R}^d$ with mean $0$ and covariance $I_d$ such that if Algorithm 1 is ran with $t$ buckets and sample $X_1, \ldots, X_n \sim P$ as input, it returns an estimate $\hat{\mu}$ satisfying*

$$\mathop{\mathbf{Pr}}_{X_1,\ldots,X_n \sim P}\left[\|\hat{\mu}\|_2 \geq C\sqrt{\frac{d\log(1/\delta)}{n\log\log(1/\delta)}}\right] > \delta.$$

This result follows from combining two different lower bounds: the first demonstrates a large error when Algorithm 1 is run with many buckets, while the second shows a large error when it is run with few buckets.

**Lemma D.1.** *There is a universal constants $C > 0$ such that the following holds. For any $\delta \in (0, 1/3)$ and $n, d, t \in \mathbb{N}$ with $t \le n$, there exists a distribution $P_1$ over $\mathbb{R}^d$ with mean 0 and covariance $I_d$ such that if Algorithm 1 is given the number of buckets $t$ and the sample $X_1, \dots, X_n \sim P_1$ as input, it returns an estimate $\hat{\mu}$ satisfying*

$$\mathop{\mathbf{Pr}}_{X_1,\dots,X_n \sim P_1}\left[\|\hat{\mu}\|_2 \ge C\sqrt{\frac{dt}{n}}\right] > \delta.$$

*Proof.* Set $p = 1 - \frac{t}{8n}$ and $\ell = \frac{1}{\sqrt{p(1-p)}}$ and define random variable $Z \sim Q$ by

$$\mathop{\mathbf{Pr}}_{Z\sim Q}[Z = (1-p)\ell] = p, \qquad \mathop{\mathbf{Pr}}_{Z\sim Q}[Z = -p\ell] = 1 - p.$$

Notice that $\mathbb{E}_{Z\sim Q}[Z] = 0$ and $Z$ has variance 1. Let $P_1$ be the product distribution on $\mathbb{R}^d$ whose coordinates are independent copies of $Z$. Clearly $P_1$ has mean 0 and covariance $I_d$.

Let $X_1, \dots, X_n$ be an i.i.d. sample from $P_1$ and run Algorithm 1 using $t$ buckets. Denote the bucket sizes by $m_1, \dots, m_t$, so that

$$m_j \le \left\lceil \frac{n}{t} \right\rceil \le \frac{n}{t} + 1.$$

Fix a coordinate $k \in [d]$ and a bucket $j \in [t]$. The empirical mean of the $k$th coordinate in bucket $j$ equals $(1-p)\ell$ whenever all samples in that bucket have $k$th coordinate equal to $(1-p)\ell$. This event has probability

$$p^{m_j} = \left(1 - \frac{t}{8n}\right)^{m_j} \ge 1 - \frac{m_j t}{8n} \ge 1 - \left(\frac{n}{t} + 1\right)\frac{t}{8n} \ge \frac{3}{4},$$

where we used the inequality $(1+x)^y \ge 1 + yx$ which holds for $x \ge -1$ and $y \ge 1$, and the fact that $t \le n$. For the fixed coordinate $k \in [d]$, let $Y_k$ be the number of buckets whose coordinate-$k$ empirical mean is equal to $(1-p)\ell$. Then $Y_k$ is a sum of independent Bernoulli random variables with

$$\mathbb{E}\, Y_k \ge \frac{3t}{4}.$$

We can apply Markov's inequality to the random variable $t - Y_k$ to get

$$\mathop{\mathbf{Pr}}_{X_1,\dots,X_n \sim P_1}\left[Y_k > \frac{t}{2}\right] = 1 - \mathop{\mathbf{Pr}}_{X_1,\dots,X_n \sim P_1}\left[t - Y_k \ge \frac{t}{2}\right] \ge 1 - \frac{\frac{t}{4}}{\frac{t}{2}} = \frac{1}{2}.$$

On the event $\{Y_k > t/2\}$, strictly more than half of the bucket means in coordinate $k$ are equal to $(1-p)\ell$, and therefore the coordinate-wise median $M_k$ is $(1-p)\ell$. Now define

$$W = \sum_{k=1}^{d} \mathbf{1}\{M_k = (1-p)\ell\}.$$

Since the coordinates of $P_1$ are independent, the events $\{M_k = (1-p)\ell\}$ are independent over $k \in [d]$. Moreover since,

$$\mathbb{E}\, W \ge \frac{d}{2},$$

we can invoke Markov's inequality on the random variable $d - W$ to conclude that

$$\mathop{\mathbf{Pr}}_{X_1,\dots,X_n \sim P_1}\left[W \ge \frac{d}{4}\right] = 1 - \mathop{\mathbf{Pr}}_{X_1,\dots,X_n \sim P_1}\left[d - W \ge \frac{3d}{4}\right] \ge 1 - \frac{\frac{d}{2}}{\frac{3d}{4}} = \frac{1}{3}.$$

On this event, we have that

$$\|\hat{\mu}\|_2 \ge \sqrt{\frac{d}{4}}(1-p)\ell = \sqrt{\frac{dt}{4(8n-t)}} \ge C\sqrt{\frac{dt}{n}},$$

for some universal constant $C > 0$, where we have used the assumption $t \le n$. By setting $c = \frac{1}{3}$ it holds for every $\delta \in (0, c)$ that

$$\mathop{\mathbf{Pr}}_{X_1,\dots,X_n \sim P_1}\left[\|\hat{\mu}\|_2 \ge C\sqrt{\frac{dt}{n}}\right] > \delta,$$

as claimed. This completes the proof. □

**Lemma D.2.** *There exist universal constants $c_1, c_2$ such that the following holds. Let $n, d, t \in \mathbb{N}$ with $t \le n$, and $\delta \in (0, c_1)$ satisfying*

$$t \le \frac{\log(1/\delta)}{\log\log(1/\delta)}.$$

*There exists a distribution $P$ over $\mathbb{R}^d$ satisfying $\mathbb{E}_P X = 0$ and $\mathrm{Cov}_P(X) = I_d$, such that the following holds. If Algorithm 1 is run with $t$ buckets on $X_1, \dots, X_n \sim P$, its output $\hat{\mu}$ satisfies*

$$\mathop{\mathbf{Pr}}_{X_1,\dots,X_n \sim P}\left[\|\hat{\mu}\|_2 \ge c_2\sqrt{\frac{td\log(1/\delta)}{n}}\right] > \delta.$$

*Proof.* Our hard distribution will be of the form

$$X = R\varepsilon, \qquad \varepsilon = (\varepsilon_1, \dots, \varepsilon_d),$$

where $\varepsilon_1, \dots, \varepsilon_d$ are independent Rademacher random variables and $R$ is an independent scalar random variable to be specified subsequently. Now, let

$$L = \log(1/\delta), \qquad m = \left\lceil \frac{n}{t} \right\rceil, \qquad a = \frac{L}{8e^2}.$$

Note that $a \ge 2$ for sufficiently small $c_1$. Now, define $R$ as follows

$$\mathbf{Pr}\left[R = \sqrt{ma}\right] = \frac{1}{ma}, \qquad \mathbf{Pr}[R = 0] = 1 - \frac{1}{ma},$$

and let $X = R\varepsilon$. Then, observe that

$$\mathbb{E}\, X = 0, \qquad \mathrm{Cov}(X) = \mathbb{E}[R^2]\,\mathbb{E}[\varepsilon\varepsilon^\top] = ma \cdot \frac{1}{ma} I_d = I_d.$$

Let $m_1, \dots, m_t$ be the bucket sizes produced by Algorithm 1. Since $t \le n$,

$$m_i \in \left\{\left\lfloor \frac{n}{t} \right\rfloor, \left\lceil \frac{n}{t} \right\rceil\right\}, \qquad m_i \ge \frac{m}{2}.$$

For bucket $i$, let $E_i$ be the event that exactly one sample in the bucket has nonzero multiplier $R$. Then

$$\mathbf{Pr}[E_i] = m_i \cdot \frac{1}{ma} \cdot \left(1 - \frac{1}{ma}\right)^{m_i - 1} \ge \frac{1}{8a}.$$

Observing that $E_1, \dots, E_t$ are independent, we get for $E = \cap_{i=1}^t E_i$,

$$\mathbf{Pr}[E] \ge \left(\frac{1}{8a}\right)^t. \tag{16}$$

We condition on $E$ for the remainder of the proof. Note that each bucket contains exactly one nonzero sample, and the corresponding Rademacher vectors remain independent. Hence the $i$th bucket mean is distributed as

$$\hat{\mu}^{(i)} = \frac{\sqrt{ma}}{m_i}\varepsilon^{(i)} = s_i\varepsilon^{(i)}, \qquad i = 1, \dots, t,$$

where $\varepsilon^{(1)}, \dots, \varepsilon^{(t)}$ are independent copies of $\varepsilon$ and

$$s_i \ge s_0 \coloneqq \sqrt{\frac{a}{m}}.$$

Thus, conditioned on $E$,

$$\hat{\mu}_j = \mathrm{Median}\left(s_1\varepsilon_j^{(1)}, \dots, s_t\varepsilon_j^{(t)}\right), \qquad j \in [d].$$

We use the following elementary fact.

**Claim D.3.** For every $s_i \geq s_0$,

$$\mathbf{Pr}\left[|\mathrm{Median}(s_1\varepsilon_1, \ldots, s_t\varepsilon_t)| \geq s_0\right] \geq 1/2,$$

where $\varepsilon_1, \ldots, \varepsilon_t$ are independent Rademacher random variables.

*Proof.* If $t$ is odd, one of the two signs appears strictly more than $t/2$ times. The median then has that sign and absolute value at least $s_0$. If $t$ is even, the same conclusion holds unless exactly half of the signs are positive and half are negative. Therefore if $B \sim \mathrm{Binom}(t, 1/2)$, then

$$\mathbf{Pr}\left[|\mathrm{Median}(s_1\varepsilon_1, \ldots, s_t\varepsilon_t)| \geq s_0\right] \geq 1 - \mathbb{I}\{t \text{ even}\}\,\mathbf{Pr}\left[B = \frac{t}{2}\right] \geq \frac{1}{2}.$$

□

Let $Z = \sum_{j=1}^{d} \mathbb{I}\{|\hat{\mu}_j| \geq s_0\}$. Conditioned on $E$, the summands defining $Z$ are independent and have expectations at least $1/2$ by Claim D.3. Hence a Chernoff bound gives

$$\mathbf{Pr}\left[Z \geq \frac{d}{4} \,\middle|\, E\right] \geq c.$$

On this event, $\|\hat{\mu}\|_2 \geq s_0\sqrt{d}/2$. Therefore

$$\mathbf{Pr}\left[\|\hat{\mu}\|_2 \geq c\, s_0\sqrt{d} \,\middle|\, E\right] \geq c.$$

Since $m \leq 2n/t$,

$$s_0\sqrt{d} = \sqrt{\frac{a}{m}}\sqrt{d} \geq c\sqrt{\frac{atd}{n}}.$$

Therefore, using (16),

$$\mathbf{Pr}\left[\|\hat{\mu}\|_2 \geq c\sqrt{\frac{atd}{n}}\right] \geq c\left(\frac{1}{8a}\right)^t = c\left(\frac{e^2}{L}\right)^t.$$

By taking $c_1$ sufficently small such that $\log L - 2 \geq 0$ and since $t \leq L/\log L$,

$$\log\left[c\left(\frac{e^2}{L}\right)^t\right] \geq \log c - \frac{L}{\log L}(\log L - 2) = -L + \frac{2L}{\log L} + \log c.$$

Taking $c_1$ sufficiently small makes the last quantity larger than $-L$. Hence

$$c\left(\frac{e^2}{L}\right)^t > e^{-L} = \delta.$$

Finally, because $a = L/(8e^2)$,

$$c\sqrt{\frac{atd}{n}} = c'\sqrt{\frac{tdL}{n}} = c'\sqrt{\frac{td\log(1/\delta)}{n}},$$

for another universal constant $c' > 0$. Renaming constants establishes the lemma. □

*Proof of Theorem 1.4.* The result follows from Lemmas D.1 and D.2 for the regimes:

$$t \geq \frac{\log(1/\delta)}{\log\log(1/\delta)} \text{ and } t \leq \frac{\log(1/\delta)}{\log\log(1/\delta)}$$

respectively. □

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main theoretical contributions: optimal in-expectation guarantees for $\ell_p$ norms, and the suboptimality result for projection-based median-of-means estimators. The assumptions, including finite covariance and isotropic covariance in the minimax formulation, are stated in the theorem statements.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations of the scope of the results, including that the full high-probability minimax theory for $\ell_p$ norms under only covariance assumptions remains unresolved.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper is primarily theoretical. The main theorem and lemma statements include their assumptions. Complete proofs are provided either in the main text or in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [N/A]

Justification: The paper does not include experimental results. The main claims are theoretical and are supported by theorem statements and proofs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper does not include experimental results requiring code or data. The estimators are described mathematically.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experiments, training procedures, test splits, hyperparameter tuning, or optimizers.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments. The probability and expectation guarantees are theoretical.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include computational experiments requiring compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The paper is theoretical and does not involve human subjects, private data, deployed systems, scraped datasets, or models with dual-use release concerns. The work conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is foundational theoretical work on minimax rates and robust mean estimation. It does not introduce a deployed system with direct societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data, pretrained models, scraped datasets, or other artifacts that pose high misuse or dual-use risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing code, datasets, pretrained models, or other research assets. Prior scientific work is cited in the bibliography.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release new datasets, codebases, models, benchmarks, or other assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, human-subject experiments, or private human data.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification:

    **AI assistance disclosure.** The theorem statements, proofs, and hard constructions in this work were developed by the authors. Most of the results were obtained in summer 2025 without material AI assistance. During the preparation of the manuscript in April 2026, OpenAI GPT-5.5 assisted with formalizing and drafting parts of the proof of the lower bound for projection based median of means estimators, as well as with checking and polishing the manuscript.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.