The Geometric Median and Applications to Robust Mean Estimation *

Stanislav Minsker† and Nate Strawn‡

Abstract. This paper is devoted to the statistical and numerical properties of the geometric median, and its applications to the problem of robust mean estimation via the median of means principle. Our main theoretical results include (a) an upper bound for the distance between the mean and the median for general absolutely continuous distributions in $\mathbb{R}^d$, and examples of specific classes of distributions for which these bounds do not depend on the ambient dimension d; (b) exponential deviation inequalities for the distance between the sample and the population versions of the geometric median, which again depend only on the trace-type quantities and not on the ambient dimension. As a corollary, we deduce improved bounds for the (geometric) median of means estimator that hold for large classes of heavy-tailed distributions. Finally, we address the error of numerical approximation, which is an important practical aspect of any statistical estimation procedure. We demonstrate that the objective function minimized by the geometric median satisfies a "local quadratic growth" condition that allows one to translate suboptimality bounds for the objective function to the corresponding bounds for the numerical approximation to the median itself, and propose a simple stopping rule applicable to any optimization method which yields explicit error guarantees. We conclude with the numerical experiments including the application to estimation of mean values of log-returns for S&P 500 data.

Key words. geometric median, median of means, heavy tails

AMS subject classifications. 62G35, 60E15

1. Introduction.

The geometric median, also referred to as the spatial median and the L_1 median, is one of the oldest and most popular robust estimators of location. Its roots go back to the Fermat, Toricelli and Weber [49] under the name of "Fermat-Weber" point, and to the work [19] under the name of Haldane's median; other notable early references include the paper by Gini and Galvani [17]. The geometric median is an element of a more general family of spatial quantiles that was introduced and studied in detail by Koltchinskii and Chaudhuri [26, 27, 9]: in particular, existence, uniqueness, and the asymptotic properties of spatial quantiles are well-understood. Extensions of the geometric median to the general Banach spaces were analyzed by Kempreman [25] and, more recently, by Romon [46]. Deep connections between the probability distributions and the corresponding spatial quantiles have been investigated by Konen [28].

Renewed interest in the properties of the geometric median was sparked with the re-introduction of the so-called "median of means" (MOM) estimator into high-dimensional statistics and machine learning literature. Originally appearing in the works of Nemirovski and Yudin [39, 22, 2] in a different context, the MOM estimator was shown to be a powerful

*Submitted in August 2023

Funding: S. Minsker acknowledges support by the National Science Foundation grants DMS CAREER-2045068 and CCF-1908905.

†Department of Mathematics, University of Southern California, Los Angeles, CA (minsker@usc.edu).

‡Department of Mathematics and Statistics, Georgetown University, Washington, DC (nate.strawn@georgetown.edu).

tool for the analysis of corrupted and heavy-tailed data by Lerasle and Oliveira [30]. The work by Hsu and Sabato [21] demonstrated multiple novel applications of the original estimator by Nemirovski and Yudin in general metric spaces, while Minsker [37] introduced a version of the median of means principle based on the geometric median. On a high level, the median of means estimator can be viewed as a "majority vote" among several independent estimators of the mean. Its popularity can be attributed to the fact that it is widely applicable, efficiently computable even in high dimensions, requires minimal tuning, and admits strong theoretical guarantees in many circumstances. However, as was pointed out by several authors, for instance by Lugosi and Mendelson [31], the geometric median of means estimator fails to attain optimal deviation bounds for the fundamental problem of multivariate mean estimation. Specifically, let $Y_1, \dots, Y_N$ be i.i.d. copies of a random vector $Y \in \mathbb{R}^d$ with mean $\mathbb{E}Y = \mu$ and covariance $\mathbb{E}(Y-\mu)(Y-\mu)^T = \Sigma_Y$. Then, as shown by Minsker [37], for any $1 \le t \le N/2$, there exists a version $\widehat{\mu}_N = \widehat{\mu}_N(Y_1, \dots, Y_N; t)$ of the geometric MOM estimator, formally defined in display (1.4) below, such that

$$\|\widehat{\mu}_N - \mu\| \le C\sqrt{\frac{\operatorname{tr}(\Sigma_Y)t}{N}} \tag{1.1}$$

with probability at least $1 - e^{-t}$; here, $C > 0$ is an absolute constant, $\|\cdot\|$ stands for the Euclidean norm of a vector and the spectral norm of a matrix, and $\operatorname{tr}(\cdot)$ denotes the trace of an operator. At the same time, a sub-Gaussian estimator $\widetilde{\mu}_N$ should satisfy an inequality akin to the sample mean of a Gaussian distribution, namely,

$$\|\widetilde{\mu}_N - \mu\| \le C\left(\sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{N}} + \sqrt{\|\Sigma_Y\|}\sqrt{\frac{t}{N}}\right) \tag{1.2}$$

with probability at least $1 - e^{-t}$, where $C > 0$ is an absolute constant; the advantage of the latter inequality over (1.1) is the fact that the deviation parameter t and the dimension-dependent quantity $\operatorname{tr}(\Sigma_Y)$ appear in separate additive terms. It immediately implies that the radii of the confidence balls for the true mean μ derived from the inequality (1.2) are much smaller compared to their counterpart obtained from (1.1). Lugosi and Mendelson [31] proposed an alternative to the standard median of means principle based on the notion of tournaments and showed that the resulting estimator achieves the desired sub-Gaussian deviation guarantees for distributions possessing only the finite second moment. Many improvements, extensions and refinements of sub-Gaussian estimators have been suggesteed in the mathematical statistics and theoretical computer science literature since: we refer the reader to the excellent surveys by Lugosi and Mendelson [32] and Diakonikolas and Kane [14]. While the original estimator by Lugosi and Mendelson [31] is difficult to compute, several closely related numerically feasible alternatives have been proposed by Hopkins [20], Cherapanamjeri [11], Depersin and Lecué [13], Bateni et al. [3], among others. However, to the best of our knowledge, none of these methods admit practical implementations comparable to the best algorithms for evaluating the geometric median, as those in the works by Cohen et al. [12], Beck and Sabach [4] and Cardot et al. [7]. Due to the computational advantages offered by the geometric median of means, it has become a popular tool for designing robust versions of distributed optimization methods such as Federated Learning [1, 6, 10, 44]. Therefore,

improved guarantees for the geometric MOM estimator have immediate implications for a variety of algorithms that use MOM principle as a subroutine.

1.1. Statistical error bounds. In this paper, we revisit the original geometric median of means construction and show that the inequality (1.1) can be improved for large classes of absolutely continuous, heavy-tailed distributions with sufficiently large effective rank

$$r(\Sigma_Y) := \frac{\operatorname{tr}(\Sigma_Y)}{\|\Sigma_Y\|}.$$

Indeed, if $r(\Sigma_Y)$ is bounded by a constant, then (1.1) readily provides sub-Gaussian guarantees. Specifically, we show that $\widehat{\mu}_N$ satisfies the bound of sub-exponential type: for all $t \lesssim \sqrt{N}$ (where $\lesssim$ denotes the inequality up to an absolute multiplicative constant), there exists a version of the MOM estimator $\widehat{\mu}_N$ such that

$$\|\widehat{\mu}_N - \mu\| \le C\left(\sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{N}} + \sqrt{\|\Sigma_Y\|}\frac{t}{\sqrt{N}}\right) \tag{1.3}$$

with probability at least $1 - e^{-t}$. While this bound succeeds in separating the trace and the confidence parameter t into different additive terms as in (1.2), thus making a significant improvement over the previously known bound (1.1), it fails to achieve optimal sub-Gaussian behavior. A remaining open question is whether there exists an easily described class of heavy-tailed distributions for which the geometric median admits truly sub-Gaussian deviation bounds.

While the proof of the inequality (1.1) is based on a simple "majority vote-type" argument, the present analysis leading to (1.3) blends accurate estimates for the bias and the stochastic error of the geometric median of means. The upper bound for the bias (Theorem 3.3 and section 3.3) is shown to be controlled by ratios of the negative moments of the norm that in turn depend on the "small ball" probability estimates. Control of the stochastic error relies on the deviation bounds for the geometric median (Theorem 3.9) that, to the best of our knowledge, are new. In particular, our bounds depend only on trace-type quantities and not on the dimension of the ambient space, and yield sub-Gaussian type guarantees for a wide range of confidence levels.

1.2. Numerical error bounds. Recall that the geometric median associated with the distribution P_Y of a random vector $Y \in \mathbb{R}^d$ is defined as

$$m(P_Y) := \operatorname*{argmin}_{z\in\mathbb{R}^d} \mathbb{E}\left(\|z - Y\| - \|Y\|\right).$$

Its empirical version based on an i.i.d. sample $Y_1, \dots, Y_N$ is

$$\widehat{m} = \operatorname{med}(Y_1, \dots, Y_N) := \operatorname*{argmin}_{z\in\mathbb{R}^d} \frac{1}{N}\sum_{j=1}^{N} \|z - Y_j\|. \tag{1.4}$$

In the sequel, we will frequently write $F(z) = F(z; Y_1, \dots, Y_N)$ in place of $\frac{1}{N}\sum_{j=1}^{N}\|z - Y_j\|$. It is well-known that m and $\widehat{m}$ are well-defined and are unique unless P_Y (or its empirical counterpart $\widehat{P}_N = \frac{1}{N}\sum_{j=1}^{N}\delta_{Y_j}$) is supported on a straight line.

Theoretical guarantees usually describe the performance of the "ideal" estimator $\widehat{m}$ that is never known exactly. In practical applications, the ability to quantify numerical error of the algorithms used to approximate $\widehat{m}$ gives one a litmus test for the overall performance of the estimator. However, most known results focus on suboptimality bounds for the objective function $F(z)$, while explicit error bounds for the approximation to the median itself are not available, to the best of our knowledge. More specifically, convex optimization theory usually asks questions about the computational complexity of finding a point z_ε such that $F(z_\varepsilon) \leq F(\widehat{m}) + \varepsilon$ for a given threshold $\varepsilon > 0$. Existing results are fully quantitative, but for statistical applications we require the bounds for $\|z - \widehat{m}\|$ instead. To bridge this gap, we prove (in Theorem 2.3) a quadratic growth condition of the form

$$F(z) - F(\widehat{m}) \geq C_1 \frac{\|z - \widehat{m}\|^2}{\|z - \widehat{m}\| + C_2},$$

where C_1 and C_2 are explicit functions of the data $Y_1, \dots, Y_k$. This inequality immediately promotes any sub-optimality bound for the objective function to an error bound for approximating the median. As a corollary, we deduce a practical stopping criteria for any algorithm designed to find the geometric median, and propose a simple numerical procedure with fully explicit error bounds.

The problem of minimizing $F(z)$ is a classical one, and has a long history. The most well-known numerical method is perhaps the celebrated *Weiszfeld's algorithm* [50]. Various improvements, refinements and accelerated versions of Weiszfeld's algorithm have been proposed and analyzed over the years. For example, results in this direction have been obtained in [42, 43, 24, 48], among others; an excellent review of the state of the art along with several new advances is given by Beck and Sabach [4]. The work by Cardot et al. [8] develops an online stochastic descent algorithm for minimizing $F(z)$ and provide its asymptotic convergence rate. The interior point method with the best-to-date convergence guarantees has been developed by Cohen et al. [12]; this work also provides a thorough comparison of existing alternatives.

1.3. Organization. The rest of the content is organized as follows: in Section 2, we introduce key notation and state our main results. Section 3 is devoted to the non-asymptotic analysis of the statistical properties of the geometric median, and culminates in the proof of Theorem 2.1. In section 4, we discuss numerical algorithms for computing the geometric median that admit quantifiable error bounds, and prove a so-called quadratic growth condition. Section 5 concludes the paper with the numerical experiments.

2. Main results. Let us recall the definition of the median of means estimator based on a sample $Y_1, \dots, Y_N$. Let $G_1 \cup \ldots \cup G_k \subseteq \{1, \dots, N\}$ be an arbitrary collection of $k \leq N/2$ disjoint subsets ("blocks") of cardinality $n = \lfloor N/k \rfloor$ each, $\bar{Y}_j := \frac{1}{|G_j|} \sum_{i \in G_j} Y_i$ and

$$\widehat{\mu}_N := \operatorname{med}\left(\bar{Y}_1, \dots, \bar{Y}_k\right). \tag{2.1}$$

The main goal of this work is to understand when the random variable $\|\widehat{\mu}_N - \mu\|$ admits good deviation bounds under minimal assumptions on the distribution of Y, and what is the

typical computational complexity of approximating $\widehat{\mu}_N$. We will now define the classes of distributions for which such "good bounds" can be established.

Everywhere below, it will be assumed that the distribution of a random vector Y is absolutely continuous with respect to the volume measure on a linear subspace of $\mathbb{R}^d$ (the linear span of the support of P_Y), and $M(Y)$ will stand for the sup-norm of the corresponding density p_Y. Similarly, if $X \in \mathbb{R}$ is a random variable with absolutely continuous distribution, $M(X)$ will denote the sup-norm of its density. The classes of distributions we are interested in are defined next.

1. **Linear transformations of the independent factors:** let $Y \in \mathbb{R}^d$ be given by a linear transformation $Y = AX$ where $X = (X_1, \dots, X_k) \in \mathbb{R}^k$ is a centered random vector with independent coordinates such that Σ_X is the identity matrix I_k and $M_0 := \max_{j=1,\dots,k} M(X_j) < \infty$. Moreover, assume that $\max_{j=1,\dots,k} \mathbb{E}|X_j - \mathbb{E}X_j|^q = K(q) < \infty$ for some $q > 2$. The class of corresponding distributions P_Y will be denoted $\mathcal{P}_1 := \mathcal{P}_1(M_0, K)$.
2. **Distributions with well-conditioned covariance matrices:** let $Y \in \mathbb{R}^d$ be a random vector with support contained in a k-dimensional subspace L such that its distribution is absolutely continuous with respect to the volume measure on L. Assume that [1]
   (a) $M^{1/k}\left(\Sigma_Y^{-1/2} Y\right) \le M_0$;
   (b) $\frac{\operatorname{tr}(\Sigma_Y)}{k \cdot \det^{1/k}(\Sigma_Y)} \le R$;
   (c) For some $q > 2$ and all unit vectors u,
   $$\mathbb{E}^{2/q} |\langle Y, u\rangle|^q \le K \langle \Sigma_Y u, u\rangle .$$
   The class of all such distributions will be denoted $\mathcal{P}_2 := \mathcal{P}_2(k, M_0, K, R)$.
3. **Signal plus noise:** let $Y = X + \xi \in \mathbb{R}^d$ where $P_X \in \mathcal{P}_2(k, M_0, K, R)$, ξ is independent from X and is such that $\operatorname{tr}(\Sigma_\xi) \le h \operatorname{tr}(\Sigma_X)$. This class of distributions is a natural generalization of $\mathcal{P}_2(k, M_0, K, R)$ and will be denoted $\mathcal{P}_3 := \mathcal{P}_3(k, M_0, K, R, h)$. Distributions from the class $\mathcal{P}_3$ can naturally be viewed as perturbations of the elements of the class $\mathcal{P}_2$.

The first main result of the paper is the following high-probability bound for the estimator $\widehat{\mu}_N$.

Theorem 2.1. *Assume that the distribution of Y belongs to the class $\mathcal{P}_j$, $j \in \{1, 2, 3\}$. Then for all $k_0 \le k \le N/2$, the median of means estimator $\widehat{\mu}_N$ defined in (2.1) satisfies the inequality*

$$\|\widehat{\mu}_N - \mu\| \le C\left(\sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{N}} + \sqrt{\|\Sigma_Y\|}\sqrt{\frac{k}{N}}\right) \tag{2.2}$$

with probability at least $1 - e^{-\sqrt{k}}$, where k_0 and C depend only on the parameters of the corresponding class $\mathcal{P}_j$.

[1] Here, we implicitly view Σ_Y as an operator $\Sigma_Y : L \mapsto$L.

Remark 2.2. Let us discuss the main assumptions of the theorem.

1. Note that $\frac{\mathrm{tr}\,(\Sigma_Y)}{k\cdot \det^{1/k}(\Sigma_Y)}$ is the ratio of the arithmetic and the geometric means of the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_k$ of Σ_Y: this quantity behaves well when the eigenvalues are of "similar" magnitude. For example, if $\lambda_j = \frac{C}{j^\alpha}$ for $\alpha < 1$, then it is easy to check that
$$\frac{\sum_{j=1}^k \lambda_j}{k\left(\prod_{i=1}^k \lambda_i\right)^{1/k}} \leq C(\alpha).$$
In fact, it is known [18] that for most (with respect to the uniform distribution on a sphere) sequences, the ratio of arithmetic and geometric means is well-behaved.
2. Moment equivalence conditions similar to (2c) are well known in the literature - for example, it has been employed in [36, 33, 52, 41], among others, in the contexts of robust estimation and random matrix theory. It is known to hold (Lemma 4.2 in [35]) for random vectors of the form $Y = AX$ where X is either a vector with independent coordinates, or an unconditional vector with coordinates possessing finite moments of order q (recall that a random vector has unconditional distribution when the distribution of $(\varepsilon_1 X_1, \ldots, \varepsilon_d X_d)$ is the same as the distribution of $X = (X_1, \ldots, X_d)$ for any sequence $\varepsilon_1, \ldots, \varepsilon_d \in \{\pm 1\}^d$. Many elliptically symmetric distributions, for example multivariate Student's t-distribution, also satisfy (2c) under appropriate restrictions on the number of degrees of freedom.
Define the spatial sign covariance matrix via
$$D_Y := \mathbb{E}\left[\frac{(Y-m)}{\|Y-m\|}\frac{(Y-m)^T}{\|Y-m\|}\right], \tag{2.3}$$
where $m = m(P_Y)$ is the geometric median of Y. The role of assumption (2c) is in showing that $\Delta := \|D_Y\| \leq \frac{C}{r(\Sigma_Y)}$. When (2c) does not hold, inequality (2.2) is still valid with $\sqrt{\|\Sigma_Y\|}$ replaced by $\max\left(\sqrt{\|\Sigma_Y\|}, \sqrt{\frac{\mathrm{tr}\,(\Sigma_Y)\Delta}{\sqrt{k}}}\right)$.

In the following sections, we develop the technical tools needed to prove Theorem 2.1 and discuss the numerical methods used to approximate the estimator $\widehat{\mu}_N$. The proof of Theorem 2.1 is based on the error decomposition
$$\|\widehat{\mu}_N - \mu\| \leq \|m_n - \mu\| + \|\widehat{\mu}_N - m_n\| \tag{2.4}$$
where m_n is the geometric median of the distribution $P^{(n)}$ of the average $\frac{1}{n}\sum_{j=1}^n Y_j$ (recall that $n = \lfloor N/k \rfloor$). The term $\|m_n - \mu\|$ is the main contribution to the bias of the estimator $\widehat{\mu}_N$ and is controlled by the size of the block n, while $\|\widehat{\mu}_N - m_n\|$ is the stochastic error that depends on the number of blocks k. We show that under various conditions encoded by the classes $\mathcal{P}_j$, $j \in \{1,2,3\}$, the "bias" admits a dimension-free upper bound of the form $\sqrt{\|\Sigma_Y\|}\sqrt{\frac{k}{N}}$ while
$$\|\widehat{\mu}_N - m_n\| \lesssim \sqrt{\frac{\mathrm{tr}\,(\Sigma_Y)}{N}} + \sqrt{\Delta\,\mathrm{tr}\,(\Sigma_Y)}\sqrt{\frac{s}{N}} + \sqrt{\frac{\mathrm{tr}\,(\Sigma_Y)}{N}}\frac{s}{\sqrt{k}} \tag{2.5}$$

with probability at least $1 - 4e^{-s}$ for all $s \lesssim k$. The combination of (2.4) and (2.5) yields the desired inequality.

Our second main result, stated below, is a quadratic growth condition which ensures that any sub-optimality guarantees for the objective function $F(z; y_1, \dots, y_k)$ translate into the corresponding bounds for the numerical approximation to the geometric median.
Given a collection of points $y_1, \dots, y_k \in \mathbb{R}^d$ and a positive integer p, set $\nu_p = \frac{1}{k}\sum_{j=1}^k \|y_j\|^p$.

Theorem 2.3. *Let $y_1, \dots, y_k \in \mathbb{R}^d$ be such that $\sum_{j=1}^k y_j = 0$. Moreover, assume that the matrix $\widehat{\Sigma} = \frac{1}{k}\sum_{j=1}^k y_j y_j^T$ satisfies the condition*

$$a := \sum_{j=2}^{d} \lambda_j(\widehat{\Sigma}) > 0$$

where $\lambda_j(\widehat{\Sigma})$ are the eigenvalues of $\widehat{\Sigma}$ listed in non-increasing order. Then for all $z \in \mathbb{R}^d$,

$$F(z) - F(\widehat{m}) \geq \frac{1}{2}\frac{a\|z - \widehat{m}\|^2}{b^2(\|z - \widehat{m}\| + b)}$$

where $\widehat{m}$ is the geometric median of $y_1, \dots, y_k$ and

$$b = \frac{20\nu_1^3 + 6\nu_1\nu_2 + \nu_3}{a}.$$

Note that we do not make any assumptions on the nature of the points $y_1, \dots, y_k$.

Remark 2.4. Observe that whenever $\|z - \widehat{m}\|$ is small, the leading term in the lower bound is $\frac{a}{2b^3}\|z - \widehat{m}\|^2$. If the points are drawn from the uniform distribution on a sphere of radius $\sqrt{d}$, the factor $\frac{a}{2b^3}$ scales like $d^{-1/2}$. Numerical experiments in Section 5 verify that this rate of dimensional dependence is asymptotically sharp. To see that the bound is of correct form in general, consider the collection of points in R^2 given by

$$\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}-1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}0\\-1\end{pmatrix}$$

The geometric median of these data is the origin, but the function $F(z)$ restricted to the x axis is

$$F((x,0)) = 4 + 2\frac{x^2}{\sqrt{x^2+1}+1},$$

indicating that local quadratic growth bound is optimal in general.

Theorem 2.3 immediately implies the following global error bound.

Corollary 2.5. *Under the assumptions of Theorem 2.3, for all $z \in \mathbb{R}^d$,*

$$\|\nabla F(z)\| \geq \frac{1}{2}\frac{a\|z - \widehat{m}\|}{b^2(\|z - \widehat{m}\| + b)}.$$

This bound provides a test for early termination given any iterative method. In particular, the right-hand side is monotone increasing, hence

$$\|\nabla F(z)\| < \frac{1}{2}\frac{a\varepsilon}{b^2(\varepsilon+b)},$$

is a sufficient condition for the inequality $\|z-\widehat{m}\| < \varepsilon$ to hold.

3. Statistical error bounds. In this section, we develop the technical background needed prove Theorem 2.1. The theorem itself is proved in subsection 3.5.

3.1. Preliminaries: small ball probabilities. Recall that, given a centered random vector $Z \in \mathbb{R}^d$ with a distribution that is absolutely continuous with respect to the Lebesgue measure, the sup-norm of the density p_Z of Z is denoted $M(Z)$. The following "small-ball" inequality is immediate: for any $z \in \mathbb{R}^d$ and $R > 0$,

$$\mathbb{P}(\|Z-z\| \le R) \le M(Z)V_d(R)$$

where $V_d(R) = \frac{\left(\sqrt{\pi}R\right)^d}{\Gamma(d/2+1)}$ is the volume of a ball $B(R)$ of radius R in $\mathbb{R}^d$. Assuming that the covariance matrix $\Sigma_Z = \mathbb{E}(Z-\mathbb{E}Z)(Z-\mathbb{E}Z)^T$ exists (note that it must be non-degenerate for the density p_Z to be well-defined), it is easy to see using the change-of-variables formula that $M(Z) = \frac{M\left(\Sigma_Z^{-1/2}Z\right)}{\sqrt{\det(\Sigma_Z)}}$, hence

$$\mathbb{P}(\|Z-z\| \le R) \le M\left(\Sigma_Z^{-1/2}Z\right)\frac{V_d(R)}{\sqrt{\det(\Sigma_Z)}}. \tag{3.1}$$

The advantage of the latter expression is that the quantity $M\left(\Sigma_Z^{-1/2}Z\right)$ is invariant with respect to the affine transformations of Z. Let us also recall that $V_d(R)$ satisfies the following inequalities for some absolute positive constants c_1 and c_2:

$$\frac{c_1}{\sqrt{d}}\left(\frac{\sqrt{2\pi e}\,R}{\sqrt{d}}\right)^d \le V_d(R) \le \frac{c_2}{\sqrt{d}}\left(\frac{\sqrt{2\pi e}\,R}{\sqrt{d}}\right)^d. \tag{3.2}$$

For special classes of distributions, better estimates for the small ball probabilities are available. Next, we will recall several results in this direction.

Theorem 3.1 (Theorem 4 in [29]). *Let Z have multivariate normal distribution $N(0,\Sigma)$ and let $m(\|Z\|)$ be the median corresponding to the distribution of $\|Z\|$. Then for all $x \in \mathbb{R}^d$,*

$$\mathbb{P}(\|Z-x\| \le tm(\|Z\|)) \le \frac{1}{2}(2t)^{\frac{m^2(\|Z\|)}{4\|\Sigma\|}}.$$

It is helpful to recall that $c_1\sqrt{\operatorname{tr}(\Sigma)} \le m(\|Z\|) \le c_2\sqrt{\operatorname{tr}(\Sigma)}$ for absolute constants $0 < c_1 < c_2 < \infty$, implying that the size of small balls is essentially controlled by the effective rank $r(\Sigma)$. A more general result, stated below, is due to Rudelson and Vershynin [47].

Theorem 3.2 (Theorem 1.5 in [47]). *Assume that $Z \in \mathbb{R}^d$ is given by a linear transformation $Z = AX$ where $X = (X_1, \dots, X_k) \in \mathbb{R}^k$ is a centered random vector with independent coordinates such that the covariance matrix $\Sigma_X = I_k$ and $M_0 := \max_{j=1,\dots,k} M(X_j) < \infty$. Then for any $\varepsilon > 0$, there exists a positive constant C_ε such that for all $x \in \mathbb{R}^d$ and $t > 0$,*

$$\mathbb{P}\Big(\|Z - x\| \le t\sqrt{\operatorname{tr}(\Sigma_Z)}\Big) \le \left(C_\varepsilon M_0 t\right)^{(1-\varepsilon)\widetilde{r}(\Sigma_Z)},$$

where $\Sigma_Z = AA^T$ and $\widetilde{r}(\Sigma_Z) = \left\lfloor \frac{\operatorname{tr}(\Sigma_Z)}{\|\Sigma_Z\|} \right\rfloor = \lfloor r(\Sigma_Z) \rfloor$.

In the following sections, we will be especially interested in the small ball probabilities associated with $Z_n = \frac{1}{\sqrt{n}} \sum_{j=1}^n (Y_j - \mathbb{E}Y_j)$ where $Y_1, \dots, Y_n$ are i.i.d. copies of a random vector Y with covariance matrix Σ_Y. To make the inequality (3.1) useful, we need a non-asymptotic estimate for $M\left(\Sigma_Y^{-1/2} Z_n\right)$.

To this end, we will rely on two facts. The first is the generalization of Rogozin's inequality proved by Juvskevivcius and Lee [23]: let $U_1, \dots, U_n$ be i.i.d. copies of a random vector U with uniform distribution over a ball centered at the origin and with radius R_U such that $M(U) = M\left(\Sigma_Y^{-1/2} Y\right)$. Then

$$M\left(\Sigma_Y^{-1/2} Z_n\right) \le M\left(\frac{1}{\sqrt{n}} \sum_{j=1}^n U_j\right). \tag{3.3}$$

The second estimate, established by Madiman et al. [34], page 17, states that

$$M\left(\frac{1}{\sqrt{n}} \sum_{j=1}^n \widetilde{U}_j\right) \le c(d) := \frac{(1 + d/2)^{d/2}}{\Gamma(1 + d/2)}$$

where $\widetilde{U}_1, \dots, \widetilde{U}_n$ are i.i.d. with uniform distribution over a ball in $\mathbb{R}^d$ of unit volume. The definition of R_U yields that $\operatorname{vol}\left(B\left(R_U \cdot M^{1/d}\left(\Sigma^{-1/2} X\right)\right)\right) = 1$, hence

$$M\left(\frac{M^{1/d}\left(\Sigma_Y^{-1/2} Y\right)}{\sqrt{n}} \sum_{j=1}^n U_j\right) \le c(d).$$

As $M(cY) = c^{-d} M(Y)$ for any random vector $Y \in \mathbb{R}^d$, we conclude using (3.3) that

$$M\left(\Sigma_Y^{-1/2} Z_n\right) \le M\left(\Sigma_Y^{-1/2} X\right) \frac{(1 + d/2)^{d/2}}{\Gamma(1 + d/2)}.$$

Employing the inequality $\Gamma(1 + d/2) \ge \sqrt{2\pi d/2}\left(\frac{d}{2e}\right)^{d/2}$, we get a simple bound

$$M\left(\Sigma_Y^{-1/2} Z_n\right) \le M\left(\Sigma_Y^{-1/2} Y\right) (2e)^{d/2} \tag{3.4}$$

and a small ball estimate

$$\mathbb{P}(\|Z_n - z\| \le R) \le c_2 \frac{M\left(\Sigma_Y^{-1/2} Y\right)}{\sqrt{\det(\Sigma_Y)}} \left(\frac{2e\sqrt{\pi}\, R}{\sqrt{d}}\right)^d. \tag{3.5}$$

3.2. Upper bounds for the difference between the mean and the median. In this section, for the ease of notation we will assume that $Y \in \mathbb{R}^d$ is centered and that m is the geometric median of P_Y. Our goal is to estimate the distance between the mean and the median (which equals $\|m\|$ under our assumptions), hence we will exclude the trivial case $m = 0$. We are especially interested in the situation when the size of $\|m\|$ is independent of or is weakly dependent on the ambient dimension d.

Theorem 3.3. *Assume that the distribution of Y is absolutely continuous with respect to Lebesgue measure on some linear subspace of $\mathbb{R}^d$. Then $m := m(P_Y)$ satisfies the inequality*

$$\|m\| \le \min\left(\sqrt{\operatorname{tr}(\Sigma_Y)}, \sqrt{\|\Sigma_Y\|}\,\frac{\mathbb{E}^{1/2}\|Y-m\|^{-2}}{\mathbb{E}\|Y-m\|^{-1}}\right).$$

Proof. The first part of the bound is straightforward: indeed, since m minimizes the function $z \mapsto \mathbb{E}\|Y - z\|$,

$$\|m\| = \|m - \mathbb{E}Y\| \le \mathbb{E}\|Y - m\| \le \mathbb{E}\|Y\| \le \mathbb{E}^{1/2}\|Y\|^2.$$

To deduce the second inequality, note that under the stated assumptions the median m satisfies the equation $\mathbb{E}\left[\frac{Y-m}{\|Y-m\|}\right] = 0$, which implies that $m = \left(\mathbb{E}\frac{1}{\|Y-m\|}\right)^{-1}\mathbb{E}\frac{Y}{\|Y-m\|}$. Therefore, for any unit vector u,

$$\langle m, u\rangle = \left(\mathbb{E}\frac{1}{\|Y-m\|}\right)^{-1}\mathbb{E}\left[\frac{\langle Y, u\rangle}{\|Y-m\|}\right] \le \frac{\mathbb{E}^{1/2}\|Y-m\|^{-2}}{\mathbb{E}\|Y-m\|^{-1}}\mathbb{E}^{1/2}\langle Y, u\rangle^2,$$

implying that $\|m\| \le \sqrt{\|\Sigma_Y\|}\cdot\frac{\mathbb{E}^{1/2}\|Y-m\|^{-2}}{\mathbb{E}\|Y-m\|^{-1}}$. ∎

The inequality $\|m\| \le \sqrt{\operatorname{tr}(\Sigma_Y)}$ is useful when the effective rank $\mathrm{r}(\Sigma_Y)$ is small. When $\mathrm{r}(\Sigma_Y)$ is large, it is often possible to find a bound for the ratio of negative moments. This problem will be discussed in the following section.

3.3. Equivalence of the negative moments of the norm. In view of the inequality stated in Theorem 3.3, it is interesting to understand when the ratio $\frac{\mathbb{E}^{1/2}\|Y-m\|^{-2}}{\mathbb{E}\|Y-m\|^{-1}}$ of negative moments is "small," in particular, when it does not depend on the ambient dimension. We will present several sufficient conditions in this section that cover many typical situations. We state the examples in the order of increasing generality: (a) the case of Gaussian random vectors; (b) the case of linear transformations of a vector with absolutely continuous independent coordinates and (c) the case of absolutely continuous distributions with bounded density.

Lemma 3.4. *Assume that Y has normal distribution $N(0, \Sigma_Y)$ such that the effective rank of the covariance matrix $r(\Sigma_Y) > 10$. Then $\frac{\mathbb{E}^{1/2}\|Y-m\|^{-2}}{\mathbb{E}\|Y-m\|^{-1}} \le C$ for an absolute constant C.*

Proof. The claim follows from Theorem 3.1 (see Corollary 1 in [29]) once we notice that the median $m(\|Y\|)$ of $\|Y\|$ satisfies $m(\|Y\|) \ge 0.08\sqrt{\operatorname{tr}(\Sigma_Y)}$. Indeed, recall that $Y = \Sigma^{1/2}Z$ where Z has standard normal distribution. Therefore, $\mathbb{E}\|Y\| = \mathbb{E}\sqrt{Z^T\Sigma_Y Z} =$

$\mathbb{E}\sqrt{\sum_{j=1}^d \lambda_j(\Sigma_Y)Z_j^2} =: f(\lambda_1,\dots,\lambda_d)$. Observe that the function f is concave, hence its minimum in the set

$$\left\{(\lambda_1,\dots,\lambda_d):\ \lambda_j \geq 0\ \forall j,\ \sum_{j=1}^d \lambda_j = \operatorname{tr}(\Sigma_Y)\right\}$$

is achieved at an extreme point $(\operatorname{tr}(\Sigma_Y),0,\dots,0)$, implying that $\mathbb{E}\|Y\| \geq \sqrt{\operatorname{tr}(\Sigma_Y)}\sqrt{\frac{2}{\pi}}$. It remains to apply Paley-Zygmund inequality to deduce that

$$\mathbb{P}\left(\|Y\| \geq t\sqrt{\frac{2}{\pi}}\sqrt{\operatorname{tr}(\Sigma_Y)}\right) \geq \mathbb{P}(\|Y\| \geq t\mathbb{E}\|Y\|) \geq (1-t)^2\frac{(\mathbb{E}\|Y\|)^2}{\mathbb{E}\|Y\|^2} \geq (1-t)^2\frac{2}{\pi}$$

which equals 0.5 for $t = 1-\sqrt{\pi}/2 > 0.11$, and the claim follows. To apply Corollary 1 in [29], we require that $\frac{m^2(\|Y\|)}{4\|\Sigma_Y\|} > 2$, which holds in view of the previous bound whenever $r(\Sigma_Y) > 10$. ∎

Next, we show that the equivalence of negative moments holds for a larger class of distributions given by linear transformations of a vector with independent coordinates. This class, denoted $\mathcal{P}_1$, was formally defined in section 2. Since any multivariate normal vector is a linear transformation of the standard normal distribution, Lemma 3.5 below also implies a version of Lemma 3.4. Recall that $M(Y)$ stands for the sup-norm of the probability density function of a random vector Y.

Lemma 3.5. *Assume that $Y \in \mathbb{R}^d$ has distribution P_Y that belongs to the class $\mathcal{P}_1(M_0,K)$. Moreover, suppose that the effective rank $r(AA^T) \geq 4$. Then*

$$\frac{\mathbb{E}^{1/2}\|Y-m\|^{-2}}{\mathbb{E}\|Y-m\|^{-1}} \leq CM_0$$

for an absolute constant $C > 0$.

Proof. Note that $\Sigma_Y = AA^T$. Therefore,

$$\left(\mathbb{E}\|Y-m\|^{-1}\right)^{-1} \leq \mathbb{E}\|Y-m\| \leq \mathbb{E}\|Y\| \leq \sqrt{\operatorname{tr}\left(\Sigma_Y\right)}$$

in view of Jensen's and Cauchy-Schwarz inequalities. Next, we will prove a general upper bound for $\mathbb{E}\|Y-x\|^{-q}$. To this end, we will use Theorem 1.5 from the work by Rudelson and Vershynin [47] which states that for any $\varepsilon > 0$, there exists a positive constant C_ε such that for all $x \in \mathbb{R}^d$ and $t > 0$, $\mathbb{P}\Big(\|Y-x\| \leq t\sqrt{\operatorname{tr}(\Sigma_Y)}\Big) \leq (C_\varepsilon M_0 t)^{(1-\varepsilon)\widetilde{r}(\Sigma_Y)}$, where $\widetilde{r}(\Sigma_Y) = \left\lfloor \frac{\operatorname{tr}(\Sigma_Y)}{\|\Sigma_Y\|}\right\rfloor = \lfloor r(\Sigma_Y)\rfloor$. Employing this "small ball" bound and letting $r := \widetilde{r}(\Sigma_Y)$ for brevity, we deduce that for any $\delta > 0$ and $q < r$,

$$\mathbb{E}\|Y-x\|^{-q} = \int_0^\infty \mathbb{P}\big(\|Y-x\|^{-q} \leq z\big)dz = \int_0^\infty \mathbb{P}\Big(\|Y-x\| \leq t^{1/q}\Big)\frac{dt}{t^2},$$

where we made a change of variables $t = z^{-1}$. Making another change of variables $t = \left(s\sqrt{\mathrm{tr}\,(\Sigma_Y)}\right)^q$, we deduce that

$$\mathbb{E}\|Y-x\|^{-q} = \frac{q}{(\mathrm{tr}\,(\Sigma_Y))^{q/2}} \int_0^\infty \mathbb{P}\Big(\|Y-x\| \le s\sqrt{\mathrm{tr}\,(\Sigma_Y)}\Big)\frac{ds}{s^{q+1}}$$
$$\le \frac{q}{(\mathrm{tr}\,(\Sigma_Y))^{q/2}}\left(\int_{1/\delta}^\infty \frac{ds}{s^{q+1}} + \int_0^{1/\delta} (C_\varepsilon M_0 s)^{r(1-\varepsilon)}\frac{ds}{s^{q+1}}\right).$$

Choosing δ to make the sum above small (e.g. $\delta = C_\varepsilon M_0\left(\frac{q}{r(1-\varepsilon)-q}\right)^{1/r(1-\varepsilon)}$), it is easy to deduce the inequality

$$\mathbb{E}\|Y-x\|^{-q} \le \frac{2(C_\varepsilon M_0)^q}{(\mathrm{tr}\,(\Sigma_Y))^{q/2}}\left(\frac{q}{r(1-\varepsilon)-q}\right)^{q/r(1-\varepsilon)}.$$

If $\varepsilon = \frac{r-q-1/2}{r}$, then $\frac{q}{r(1-\varepsilon)-q} = 2q$ and $\frac{q}{r(1-\varepsilon)} \le 1$. For small values of q, say, $q \le r/2$, this choice of ε entails the inequality $\varepsilon > \frac{r-1}{2r} \ge \frac{3}{8}$ for $r \ge 4$, so that C_ε can be treated as an absolute constant. The claim of the lemma corresponds to the case $q = 2$. ∎

Finally, we discuss the most general situation of absolutely continuous distributions.

Lemma 3.6. *Assume $Y \in \mathbb{R}^d$ has distribution P_Y that belongs to the class $\mathcal{P}_2(k, M_0, K, R)$. Then for any x in the range L of Σ_Y and $q < k = \dim(L)$,*

$$\mathbb{E}\|Y-x\|^{-q} \le c(q)\frac{M\left(\Sigma_Y^{-1/2}Y\right)^{q/k}}{\left(k\cdot \det{}^{1/k}(\Sigma_Y)\right)^{q/2}}$$

for some constant $c(q) > 0$.

Proof. The proof is similar to the argument behind Lemma 3.5. Note that for any $\delta > 0$

$$\mathbb{E}\|Y-x\|^{-q} = \int_0^\infty \mathbb{P}(\|Y-x\| \le t^{1/q})t^{-2}\,dt$$
$$\le \int_{1/\delta}^\infty t^{-2}\,dt + \int_0^{1/\delta}\mathbb{P}(\|Y-x\| \le t^{1/q})\,\frac{dt}{t^2}$$
$$= \delta + \int_0^{1/\delta}\mathbb{P}(\|Y-x\| \le t^{1/q})t^{-2}\,dt$$
$$\le \delta + c_2\frac{M\left(\Sigma^{-1/2}Y\right)}{\sqrt{\det(\Sigma_Y)}}\int_0^{1/\delta}\left(\frac{\sqrt{2\pi e}\,t^{1/q}}{\sqrt{d}}\right)^k\frac{dt}{t^2}$$

in view of (3.5). For the choice of $\delta = c_3(q)\frac{M\left(\Sigma_Y^{-1/2}Y\right)^{q/k}}{\left(k\det{}^{1/k}(\Sigma_Y)\right)^{q/2}}$, the latter is bounded by $c_4(q)\frac{M\left(\Sigma_Y^{-1/2}Y\right)^{q/k}}{\left(k\cdot\det{}^{1/k}(\Sigma_Y)\right)^{q/2}}$ for some constant $c_4 > 0$ that depends only on q. ∎

Since $\left(\mathbb{E}\|Y-m\|^{-1}\right)^{-1} \leq \sqrt{\mathrm{tr}\,(\Sigma_Y)}$, we immediately get from the previous result that whenever $k \geq 3$, then for some absolute constant $C > 0$

$$\frac{\mathbb{E}^{1/2}\|Y-m\|^{-2}}{\mathbb{E}\|Y-m\|^{-1}} \leq CM^{1/k}\left(\Sigma_Y^{-1/2}Y\right)\sqrt{\frac{\mathrm{tr}\,(\Sigma_Y)}{k\cdot \det^{1/k}(\Sigma_Y)}}. \tag{3.6}$$

Lemma 3.6 is robust to small perturbations: for example, assume that $\Sigma_Y = \lambda\sum_{j=1}^k e_j e_j^T + \delta I_d$ where $d\cdot\delta \leq Ck\cdot\lambda$. In this case, $\frac{\mathrm{tr}\,(\Sigma_Y)}{d\cdot\det^{1/d}(\Sigma_Y)}$ can be very large, and direct application of Lemma 3.6 yields a suboptimal bound. The following simple observation often yields a better result: for any linear subspace H of $\mathbb{R}^d$,

$$\mathbb{E}\|Y-x\|^{-q} \leq \mathbb{E}\|\Pi_H(Y-x)\|^{-q}, \tag{3.7}$$

where $\Pi_H(\cdot)$ stands for the orthogonal projection onto H. We formalize this observation in the following statement.

Lemma 3.7. *Assume that $Y = X+\xi \in \mathbb{R}^d$ has distribution P_Y that belongs to the class $\mathcal{P}_3(k, M_0, K, R, h)$ and that $k \geq 3$. Then for any $x \in \mathbb{R}^d$,*

$$\frac{\mathbb{E}^{1/2}\|Y-x\|^{-2}}{\mathbb{E}\|Y-x\|^{-1}} \leq C(1+h)M^{1/k}\left(\Sigma_X^{-1/2}X\right)\sqrt{\frac{\mathrm{tr}\,(\Sigma_X)}{k\det^{1/k}(\Sigma_X)}}.$$

Proof. Let H be the range of Σ_X, where, according to the definition of the class $\mathcal{P}_3$, $Y = X+\xi$. The dimension of H equals k by assumption. Employing the previously stated observation (3.7), we deduce that

$$\begin{aligned}\mathbb{E}\|Y-x\|^{-2} \leq \mathbb{E}\|\Pi_H(Y-x)\|^{-2} &= \mathbb{E}\|(X+\Pi_H\xi)-\Pi_H x\|^{-2}\\ &\leq c\frac{M^{2/k}(\widetilde{Y})}{k},\end{aligned}$$

where $\widetilde{Y} = X+\Pi_H\xi$. It remains to note that $M(\widetilde{Y}) \leq M(X)$ by the elementary properties of the convolution operator, and that $M(X) = \frac{M\left(\Sigma_X^{-1/2}X\right)}{\sqrt{\det(\Sigma_X)}}$. ∎

In the case when $\Sigma_X = \lambda\sum_{j=1}^k e_j e_j^T$, $\Sigma_\xi = \delta I_d$ and $d\cdot\delta \leq Ck\cdot\lambda$, the previous result yields that the ratio of moments is at most $O(1)M^{1/k}\left(\Sigma_X^{-1/2}X\right)$.

Remark 3.8. It should be noted that there exist examples where estimates based on the ratios of the arithmetic and geometric means provide only crude bounds: for instance, if $\lambda_j(\Sigma_Y) = \frac{1}{m+j}$, $j = 1,\dots,d$ for some positive integer m, then $\frac{\sum_{j=1}^d \lambda_j}{\max_{j\geq 1} j\left(\prod_{i=1}^j \lambda_i\right)^{1/j}}$ can be made arbitrary large by varying m and d (more specifically, it is large when m/d is large). However, under additional assumptions on the distribution (e.g. in the framework of Lemmas 3.5 - 3.7), better bounds become possible.

3.4. The geometric median: bounds for the stochastic error. Our goal in this section is to establish high-confidence deviation bounds for the distance between the empirical geometric median and its population counterpart. Recall that

$$D_Y = \mathbb{E}\left[\frac{(Y-m)}{\|Y-m\|}\frac{(Y-m)^T}{\|Y-m\|}\right], \quad \Delta = \|D_Y\|.$$

Note that $\operatorname{tr}\left(\mathbb{E}\left[\frac{(Y-m)}{\|Y-m\|}\frac{(Y-m)^T}{\|Y-m\|}\right]\right) = 1$. Therefore, if the random vector Y is sufficiently "spread out," we expect that Δ will be small. To get a rigorous bound supporting this intuition, we will assume that Y satisfies the following conditions:

(a) Condition (2c) is satisfied. We are especially interested in the situation when K is a constant that does not depend on the ambient dimension d.

(b) $\mathbb{E}^{(q-2)/q}\|Y-m\|^{-\frac{q}{2(q-2)}} \le \frac{C(q)}{\operatorname{tr}(\Sigma_Y)}$.

When (a) and (b) hold, Hölder's inequality implies that

$$\Delta = \sup_{\|u\|=1}\mathbb{E}\left[\frac{\langle Y-m,u\rangle^2}{\|Y-m\|^2}\right] \le \sup_{\|u\|=1}\mathbb{E}^{2/q}|\langle Y-m,u\rangle|^q\,\mathbb{E}^{(q-2)/q}\|Y-m\|^{-\frac{q}{2(q-2)}} \tag{3.8}$$

$$\le K\,C(q)\frac{\|\Sigma_Y\|}{\operatorname{tr}(\Sigma_Y)} = \frac{K\,C(q)}{r(\Sigma_Y)}. \tag{3.9}$$

Moment equivalence condition (a) has been discussed in detail in section in remark 2.2. Condition (b) holds for the classes of distributions discussed in section 3.3 when the effective rank of Σ_Y is sufficiently large relative to $\frac{q}{q-2}$. For instance, it holds for linear transformations of random vectors with independent coordinates as well as for random vectors with "well-conditioned" covariance matrices, in a sense that the geometric mean of their eigenvalues is equivalent to the arithmetic mean. We conclude that for large classes of distributions, $\operatorname{tr}(\Sigma_Y)\Delta \asymp \|\Sigma_Y\|$: indeed, if $r(\Sigma_Y)$ is small, then it follows since $\Delta \le 1$, and if $r(\Sigma_Y)$ is large, it follows from the previous discussion. We are ready to state the main result of this section.

Theorem 3.9. *Let $m := m(P_Y)$ be the geometric median associated with the distribution P_Y and $\widehat{m}$ - its empirical counterpart based on an i.i.d. sample $Y_1,\dots,Y_k$ from P_Y. Assume that $\Delta < 1$. Then there exist constants $c_1(\Delta)$, $c_2(\Delta)$ that depend only on Δ such that, if*

$$\mathbb{E}^{1/2}\frac{1}{\|Y-m\|^2}\left(\sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{k}} + \sqrt{\Delta\operatorname{tr}(\Sigma_Y)}\sqrt{\frac{s}{k}} + \sqrt{\operatorname{tr}(\Sigma_Y)}\frac{s}{k}\right) < c_1(\Delta), \tag{3.10}$$

then

$$\|\widehat{m}-m\| \le K(\Delta)\left(\sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{k}} + \sqrt{\Delta\operatorname{tr}(\Sigma_Y)}\sqrt{\frac{s}{k}} + \sqrt{\operatorname{tr}(\Sigma_Y)}\frac{s}{k}\right)$$

that holds with probability at least $1-2e^{-s}-2e^{-k/4}$ for all $s \le c_2(\Delta)k$.

Remark 3.10. ·

1. In view of the discussion preceding the theorem, we are mainly interested in the situation when $r(\Sigma_Y)$ is bounded from below by a sufficiently large absolute constant and when Δ is not too close to 1, e.g. $\Delta \le 1/2$.
2. Assumption (3.10) is rather mild: indeed, we showed in section 3.3 that in many common situations, $\mathbb{E}^{1/2}\frac{1}{\|Y-m\|^2} \asymp \frac{1}{\sqrt{\operatorname{tr}(\Sigma_Y)}}$.

When the confidence parameter s is not too large ($s \lesssim \sqrt{k}$), Theorem 3.9 implies the deviation guarantees of sub-Gaussian type. This fact is formally stated below.

Corollary 3.11. *Assume that assumption (3.10) is satisfied. If* $s \le \min\left(\sqrt{k}, c_2(\Delta)k\right)$, *then*

$$\|\widehat{m} - m\| \le K(\Delta)\left(\sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{k}} + \sqrt{\Delta \operatorname{tr}(\Sigma_Y)}\sqrt{\frac{s}{k}}\right)$$

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

Proof of the theorem. Recall that, in view of Theorem 3.1 in [37], $\|\widehat{m} - m\| \le 2\sqrt{\operatorname{tr}(\Sigma_Y)}$ on event $\mathcal{E}$ of probability at least $1 - e^{-k/4}$ (it suffices to take $p = 1/8$ and $\alpha = 5/12$ in the aforementioned result). In what follows, we will assume that event $\mathcal{E}$ occurs. Define $\widehat{u} := \frac{m-\widehat{m}}{\|m-\widehat{m}\|}$ (for absolutely continuous distributions, $\widehat{m} \ne m$ with probability 1, so $\widehat{u}$ is well-defined) and

$$G_k(s) := \frac{1}{k}\sum_{j=1}^{k} \|m + s\widehat{u} - Y_j\|.$$

Then $G_k(s)$ is convex, achieves its minimum at $\widehat{s} = \|\widehat{m} - m\|$, and its derivative $G_k'(s)$ is non-decreasing and satisfies $G_k'(s) \le 0$ for $s \in [0, \widehat{s}]$. It implies that $\|\widehat{m} - m\| \ge t$ is true only if $G_k'(t) \le 0$. In view of convexity of G_k,

$$0 \ge G_k'(t) \ge G_k'(0) + \inf_{0\le z\le t} G_k''(z)\cdot t,$$

where $G_k''(z) = \frac{1}{k}\sum_{j=1}^{k} \frac{1}{\|m+z\widehat{u}-Y_j\|}\left(1 - \left\langle \frac{m+z\widehat{u}-Y_j}{\|m+z\widehat{u}-Y_j\|}, \widehat{u}\right\rangle^2\right)$. Therefore, a necessary condition for the inequality $\|\widehat{m} - m\| \ge t$ to hold is

$$\frac{1}{k}\sum_{j=1}^{k}\left\langle \frac{m-Y_j}{\|m-Y_j\|}, \widehat{u}\right\rangle \ge t \inf_{0\le z\le t}\frac{1}{k}\sum_{j=1}^{k}\frac{1}{\|m+z\widehat{u}-Y_j\|}\left(1 - \left\langle \frac{m+z\widehat{u}-Y_j}{\|m+z\widehat{u}-Y_j\|}, \widehat{u}\right\rangle^2\right),$$

which is possible only if

$$\left\|\frac{1}{k}\sum_{j=1}^{k}\frac{m-Y_j}{\|m-Y_j\|}\right\| \ge t \inf_{0\le z\le t}\frac{1}{k}\sum_{j=1}^{k}\frac{1}{\|m+z\widehat{u}-Y_j\|}\left(1 - \left\langle \frac{m+z\widehat{u}-Y_j}{\|m+z\widehat{u}-Y_j\|}, \widehat{u}\right\rangle^2\right).$$

Next, we will find high confidence bounds for both sides of the inequality above. Note that we can assume that $t \le 2\sqrt{\operatorname{tr}(\Sigma_Y)}$ on event $\mathcal{E}$.

Lemma 3.12. *With probability at least* $1-e^{-s}$,

$$\left\|\frac{1}{k}\sum_{j=1}^{k}\frac{Y_j-m}{\|Y_j-m\|}\right\| \leq \frac{2}{\sqrt{k}}+\sqrt{\Delta}\sqrt{\frac{2s}{k}}+\frac{4s}{3k}. \tag{3.11}$$

Moreover, if that random vector $\frac{Y-m}{\|Y-m\|}$ *has sub-Gaussian distribution, then*

$$\left\|\frac{1}{k}\sum_{j=1}^{k}\frac{Y_j-m}{\|Y_j-m\|}\right\| \leq C\left(\frac{1}{\sqrt{k}}+\sqrt{\Delta}\sqrt{\frac{s}{k}}\right) \tag{3.12}$$

for an absolute constant $C>0$ *and with probability at least* $1-e^{-s}$.

Proof. Let $X_k(u)=\frac{1}{k}\sum_{j=1}^{k}\left\langle\frac{Y_j-m}{\|Y_j-m\|},u\right\rangle$ and note that $\mathbb{E}X_k(u)=0$ for all u. Next, write the norm as

$$\sup_{\|u\|=1}X_k(u)=\sup_{\|u\|=1}\frac{1}{k}\sum_{j=1}^{k}\left\langle\frac{Y_j-m}{\|Y_j-m\|},u\right\rangle.$$

Bousquet's version of Talagrand's concentration inequality (see [5]) yields that

$$\sup_{\|u\|=1}X_k(u)\leq 2\mathbb{E}\sup_{\|u\|=1}X_k(u)+\sup_{\|u\|=1}\mathrm{Var}^{1/2}\left(X_k(u)\right)\sqrt{2s}+\frac{4s}{3k}$$

with probability at least $1-e^{-s}$. It remains to notice that

$$\mathbb{E}\sup_{\|u\|=1}X_k(u)\leq\mathbb{E}^{1/2}\left\|\frac{1}{k}\sum_{j=1}^{k}\frac{Y_j-m}{\|Y_j-m\|}\right\|^2=\frac{1}{\sqrt{k}}\mathbb{E}^{1/2}\left\|\frac{Y_1-m}{\|Y_1-m\|}\right\|^2=\frac{1}{\sqrt{k}}$$

and that $\sup_{\|u\|=1}\mathrm{Var}^{1/2}\left(X_k(u)\right)=\frac{1}{\sqrt{k}}\left\|\mathbb{E}\frac{Y_1-m}{\|Y_1-m\|}\frac{(Y_1-m)^T}{\|Y_1-m\|}\right\|^{1/2}=\sqrt{\frac{\Delta}{k}}$. Part (b) of the lemma follows from the standard concentration bound for sub-Gaussian processes (see [15]) in place of Bousquet's inequality. ∎

Lemma 3.13. *Let* $\tau>$ *be a positive constant, and define*

$$\delta:=\delta(k,t,\tau,\Delta;s):=(1+\tau)\left(\sqrt{\Delta}\left(1+\sqrt{\frac{2s}{k}}\right)+\frac{4}{\sqrt{k}}\right)$$
$$+2(4+1/\tau)t^2\mathbb{E}\frac{1}{\|Y-m\|^2}+\left(8+\frac{4\tau}{3}+\frac{5}{3\tau}\right)\frac{s}{k}$$

If $\delta<1$, *then the following inequality holds with probability at least* $1-2e^{-k/4}-2e^{-s}$:

$$\inf_{\|u\|=1,\|m-x\|\leq t}\frac{1}{k}\sum_{j=1}^{k}\frac{1}{\|Y_j-x\|}\left(1-\left\langle\frac{Y_j-x}{\|Y_j-x\|},u\right\rangle^2\right)\geq\frac{C\left(\delta\right)}{\sqrt{\mathrm{tr}\left(\Sigma_Y\right)}}.$$

Remark 3.14. Since $\Delta < 1$ (recall that we are mostly interested in the situation $\Delta \le 1/2$), there exist $\tau = \tau(\Delta) > 0$ and $\varepsilon = \varepsilon(\Delta) > 0$ such that $\delta < 1$ whenever

$$t < \varepsilon \left(\mathbb{E}^{1/2} \|Y_1 - m\|^{-2} \right)^{-1}$$

and k is sufficiently large; let us again recall that in many typical situations,

$$\left(\mathbb{E}^{1/2} \|Y_1 - m\|^{-2} \right)^{-1} \asymp \sqrt{\operatorname{tr}(\Sigma_Y)}.$$

Proof. Note that on event $\mathcal{E}$ that was defined at the start of the proof of the theorem, $\|Y_j - x\| \le \|Y_j - m\| + 2\sqrt{\operatorname{tr}(\Sigma_Y)}$ for all j, hence one easily gets that for any $\kappa > 0$,

$$\mathbb{P}\Big(\exists J \subset [k] : \ |J| \ge \kappa k \text{ and } \|Y_j - x\| \ge (c(\kappa) + 2)\sqrt{\operatorname{tr}(\Sigma_Y)}, \ j \in J \Big) \\ \le \binom{k}{\lfloor \kappa k \rfloor} \left(2/c(\kappa)^2 \right)^{\lfloor \kappa k \rfloor} \le e^{-k}$$

where $c(\kappa) < (e^{1/\kappa}/\kappa)^{1/2}$. Consequently, on event $\mathcal{E}_1$ of probability at least $1 - e^{-k}$,

$$\|Y_j - x\| \le (c(\kappa) + 2)\sqrt{\operatorname{tr}(\Sigma_Y)} \text{ for all } j \in J \text{ such that } |J| \ge (1 - \kappa)k.$$

Next, we will find an upper bound for $\frac{1}{k} \sum_{j=1}^k \left\langle \frac{Y_j - x}{\|Y_j - x\|}, u \right\rangle^2$ that holds uniformly over u. Recall the following elementary inequality that is valid for all vectors $y_1, y_2 \in \mathbb{R}^d$: $\left\| \frac{y_1}{\|y_1\|} - \frac{y_2}{\|y_2\|} \right\| \le 2 \frac{\|y_1 - y_2\|}{\max(\|y_1\|, \|y_2\|)}$. It implies that for all j, $1 \le j \le k$,

$$\left| \left\langle \frac{Y_j - x}{\|Y_j - x\|}, u \right\rangle - \left\langle \frac{Y_j - m}{\|Y_j - m\|}, u \right\rangle \right|^2 \le 4 \min\left(1, \frac{\|x - m\|^2}{\|Y_j - m\|^2} \right)$$

so that for any $\tau > 0$,

$$\sup_{\|u\|=1, \|x-m\| \le t} \frac{1}{k} \sum_{j=1}^k \left\langle \frac{Y_j - x}{\|Y_j - x\|}, u \right\rangle^2 \\ \le \frac{1 + \tau}{k} \sum_{j=1}^k \left\langle \frac{Y_j - m}{\|Y_j - m\|}, u \right\rangle^2 + \frac{4 + 1/\tau}{k} \sum_{j=1}^k \min\left(1, \frac{t^2}{\|Y_j - m\|^2} \right). \tag{3.13}$$

The first term in the sum above can be estimated as follows: note that

$$\frac{1}{k} \sum_{j=1}^k \left\langle \frac{Y_j - m}{\|Y_j - m\|}, u \right\rangle^2 \le \frac{1}{k} \sum_{j=1}^k \left| \left\langle \frac{Y_j - m}{\|Y_j - m\|}, u \right\rangle \right|$$

and define

$$Z_k(u) = \frac{1}{k} \sum_{j=1}^k \left| \left\langle \frac{Y_j - m}{\|Y_j - m\|}, u \right\rangle \right| - \mathbb{E} \left| \left\langle \frac{Y_1 - m}{\|Y_1 - m\|}, u \right\rangle \right|.$$

Bousquet's version of Talagrand's concentration inequality yields that

$$\sup_{\|u\|=1} Z_k(u) \leq 2\mathbb{E} \sup_{\|u\|=1} Z_k(u) + \sup_{\|u\|=1} \mathrm{Var}^{1/2}\left(Z_k(u)\right)\sqrt{2s} + \frac{4s}{3k}$$

with probability at least $1-e^{-s}$. It remains to note that $\mathbb{E}\left|\left\langle \frac{Y_1-m}{\|Y_1-m\|}, u\right\rangle\right| \leq \sqrt{\Delta}$ in view of Cauchy-Schwarz inequality, and that

$$\mathbb{E}\sup_{\|u\|=1} Z_k(u) \leq 2\mathbb{E}\sup_{\|u\|=1} \frac{1}{k}\sum_{j=1}^k \varepsilon_j \left|\left\langle \frac{Y_j-m}{\|Y_j-m\|}, u\right\rangle\right|$$
$$\leq 4\mathbb{E}\sup_{\|u\|=1} \frac{1}{k}\sum_{j=1}^k \left\langle \frac{Y_j-m}{\|Y_j-m\|}, u\right\rangle = 4\mathbb{E}\left\|\frac{1}{k}\sum_{j=1}^k \frac{Y_j-m}{\|Y_j-m\|}\right\| \leq \frac{4}{\sqrt{k}}$$

in view of the symmetrization and Talagrand's contraction inequalities (see [16]). To summarize, we showed that with probability at least $1-e^{-s}$, for all unit vectors u,

$$\frac{1+\tau}{k}\sum_{j=1}^k \left|\left\langle \frac{Y_j-m}{\|Y_j-m\|}, u\right\rangle\right| \leq (1+\tau)\left(\sqrt{\Delta}\left(1+\sqrt{\frac{2s}{k}}\right) + \frac{4}{\sqrt{k}} + \frac{4s}{3k}\right). \tag{3.14}$$

In view of Bernstein's inequality, the second term in (3.13) is at most

$$(4+1/\tau)\left(\mathbb{E}\min\left(1, \frac{z^2}{\|Y_j-m\|^2}\right) + 2\sqrt{\mathrm{Var}\left(\min\left(1, \frac{z^2}{\|Y-m\|^2}\right)\right)}\sqrt{\frac{s}{k}} + \frac{2s}{3k}\right)$$
$$\leq (4+1/\tau)\left(2t^2\mathbb{E}\frac{1}{\|Y-m\|^2} + \frac{5s}{3k}\right) \tag{3.15}$$

with probability at least $1-e^{-s}$. Combining (3.13), (3.14), (3.15), we deduce the inequality

$$\sup_{\|u\|=1, \|x-m\|\leq t} \frac{1}{k}\sum_{j=1}^k \left\langle \frac{Y_j-x}{\|Y_j-x\|}, u\right\rangle^2$$
$$\leq \delta(k,t,\tau,\Delta;s) := (1+\tau)\left(\sqrt{\Delta}\left(1+\sqrt{\frac{2s}{k}}\right) + \frac{4}{\sqrt{k}}\right)$$
$$+ 2(4+1/\tau)t^2\mathbb{E}\frac{1}{\|Y-m\|^2} + \left(8 + \frac{4\tau}{3} + \frac{5}{3\tau}\right)\frac{s}{k}$$

that holds with probability at least $1-2e^{-s}$. If $\delta(k,t,\tau,\Delta;s) < 1$, then

$$\left|\left\{j : \left\langle \frac{Y_j-x}{\|Y_j-x\|}, u\right\rangle^2 \geq \delta^{1/2}(k,t,\tau,\Delta;s)\right\}\right| \leq \delta^{1/2}(k,t,\tau,\Delta;s)k$$

uniformly over all $\|u\|=1$ and $\|x-m\| \leq t$ with probability at least $1-2e^{-s}$. Now we set $\kappa := \frac{1-\delta^{1/2}(k,t,\tau,\Delta;s)}{2}$ in (8) and deduce that for all u, there exists a subset J of cardinality

at least κk such that $\frac{1}{\|Y_j-x\|} \geq \frac{1}{C(\kappa)\sqrt{\operatorname{tr}(\Sigma_Y)}}$ and $\left\langle \frac{Y_j-x}{\|Y_j-x\|}, u \right\rangle^2 < \delta^{1/2}(k,t,\tau,\Delta;s) < 1$ for all $j \in J$. Consequently,

$$\inf_{\|u\|=1, \|m-x\|\leq z} \frac{1}{k}\sum_{j=1}^{k} \frac{1}{\|Y_j-x\|}\left(1-\left\langle \frac{Y_j-x}{\|Y_j-x\|}, u\right\rangle^2\right) \geq \frac{C(\kappa)}{\sqrt{\operatorname{tr}(\Sigma_Y)}}$$

with probability at least $1-2e^{-k/4}-2e^{-s}$, where $C(\kappa) \to \infty$ as $\kappa \to 0$. ∎

To complete the proof of the theorem, choose

$$t = \widehat{t} := K\left(\sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{k}} + \sqrt{\Delta \operatorname{tr}(\Sigma_Y)}\sqrt{\frac{s}{k}} + \sqrt{\operatorname{tr}(\Sigma_Y)}\frac{s}{k}\right)$$

where the constant K is sufficiently large (the specific requirement for the size of K is given below). If $k \geq k_0(\Delta)$ is large enough, $s \leq c_1(\Delta)k$ and $\widehat{t}\mathbb{E}^{1/2}\|Y-m\|^{-2} \leq c_2(\Delta)$, then $\delta(k,\widehat{t},\tau,\Delta) < 1$, implying that the results of Lemmas 3.12 and 3.13 hold with $t = \widehat{t}$ on event $\mathcal{E}_2$ of probability at least $1-2e^{-k/4}-2e^{-s}$. If $\|\widehat{m}\| \geq \widehat{t}$, then the following inequality must hold on $\mathcal{E}_2$:

$$\widehat{t} \leq \frac{1}{C'(\Delta)}\left(\sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{k}} + \sqrt{\Delta \operatorname{tr}(\Sigma_Y)}\sqrt{\frac{s}{k}}\right).$$

If K is set so that $K > \frac{1}{C'(\Delta)}$, this yields a contradiction. Finally, the bound for the case when $\frac{Y-m}{\|Y-m\|}$ has sub-Gaussian distribution follows with (3.12) in place of (3.11). ∎

3.5. Implications for the median of means estimator. In this section we prove Theorem 2.1. To this end, we will apply Theorems 3.3 and 3.9 to the distribution $P^{(n)}$ of the average $\frac{1}{n}\sum_{j=1}^{n} Y_j$ and the sample $\bar{Y}_1, \dots, \bar{Y}_k$, noting that the corresponding covariance matrix satisfies $\Sigma_{\bar{Y}_1} = \frac{\Sigma}{n} \preceq 2\Sigma\frac{k}{N}$ whenever $k \leq N/2$. In what follows, let m_n denote the geometric median of $P^{(n)}$.

Consider two scenarios: if $r(\Sigma_Y) \leq c\frac{q}{q-2}$, then the inequality (1.1) readily yields the result. On the other hand, if $r(\Sigma_Y) > c\frac{q}{q-2}$, then $\Delta\operatorname{tr}(\Sigma_Y) \leq C(q)\|\Sigma_Y\|$ and $\mathbb{E}^{1/2}\frac{1}{\|Y-m\|^2} \leq \frac{C'}{\sqrt{\operatorname{tr}(\Sigma_Y)}}$ for a constant C' that depends on the parameters of the class $\mathcal{P}_j$, $j \in \{1,2,3\}$. It remains to show that the relevant parameters of the distribution $P^{(n)}$ can be controlled by the corresponding parameters of the distribution P_Y. First, recall the inequality (3.4) which implies that

$$M^{1/k}\left(\Sigma_{\bar{Y}_1}^{-1/2}\sqrt{n}\bar{Y}_1\right) \leq \sqrt{2e}M^{1/k}\left(\Sigma_Y^{-1/2}Y\right).$$

Therefore, the ratio $\frac{\mathbb{E}^{1/2}\|\sqrt{n}(\bar{Y}_1-m_n)\|^{-2}}{\mathbb{E}\|\sqrt{n}(\bar{Y}_1-m_n)\|^{-1}}$ can be estimated via Lemma 3.5, Lemma 3.6 or Lemma 3.7 in terms of parameters of the distribution P_Y whenever it belongs to one of the classes $\mathcal{P}_j$, $j \in \{1,2,3\}$. Next, consider the norm of the spatial sign covariance matrix

$$\Delta^{(n)} := \left\|\mathbb{E}\left[\frac{(\bar{Y}_1-m_n)}{\|\bar{Y}_1-m_n\|}\frac{(\bar{Y}_1-m_n)^T}{\|\bar{Y}_1-m_n\|}\right]\right\|.$$

In view of the well-known moment bounds (e.g., the Marcinkiewicz-Zygmund type inequality by Rio [45]), for any unit vector u and $q > 2$,

$$\mathbb{E}\left|\left\langle \frac{1}{\sqrt{n}}\sum_{j=1}^{n} Y_j, u\right\rangle\right|^q \le (q-1)^{q/2}\mathbb{E}\left|\langle Y_1, u\rangle\right|^q,$$

thus the reasoning similar to (3.9) implies that

$$\Delta^{(n)} \le \frac{KC_1(q)}{r(\Sigma_Y)} \tag{3.16}$$

whenever $P_Y \in \mathcal{P}_j$, $j \in \{1,2,3\}$. Therefore, conditions of Theorem 3.9 hold for k large enough, and we deduce that

$$\begin{aligned}\|\widehat{\mu}_N - \mu\| &\le \|m_n - \mu\| + \|\widehat{\mu}_N - m_n\| \\ &\le 2\sqrt{\frac{\|\Sigma_Y\|k}{N}}\,\frac{\mathbb{E}^{1/2}\|\sqrt{n}(\bar{Y}_1 - m_n)\|^{-2}}{\mathbb{E}\|\sqrt{n}(\bar{Y}_1 - m_n)\|^{-1}} + K(\Delta^{(n)})\left(\sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{N}} + \sqrt{\Delta^{(n)}\operatorname{tr}(\Sigma_Y)}\sqrt{\frac{\sqrt{k}}{N}}\right)\end{aligned} \tag{3.17}$$

with probability at least $1 - 4e^{-\sqrt{k}}$. The final form of the bound follows once we apply the inequality (3.16) and estimate the ratio of moments via one of the lemmas in section 3.3. For instance, if $P_Y \in \mathcal{P}_1(M_0, K)$, then Lemma 3.5 combined with (3.17) implies that

$$\begin{aligned}\|\widehat{\mu}_N - \mu\| &\le C\left(M_0\sqrt{\frac{\|\Sigma_Y\|k}{N}} + \sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{N}} + \sqrt{KC_1(q)}\sqrt{\|\Sigma_Y\|\frac{\sqrt{k}}{N}}\right) \\ &\le C\left(M_0\sqrt{\frac{\|\Sigma_Y\|k}{N}} + \sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{N}}\right)\end{aligned}$$

with probability at least $1 - 4e^{-\sqrt{k}}$ whenever $k \ge k_0(M_0, K, q)$. Bounds for the classes $\mathcal{P}_2$ and $\mathcal{P}_3$ follow similarly.

Remark 3.15 (regarding efficiency). The values of most numerical constants appearing in our bounds were left unspecified, however, they are important in applications. In one-dimensional case, the question related to the optimality of the guarantees for the MOM estimator was explored in the paper [38]. It turns out that one can improve the performance of MOM estimator by making the estimator permutation-invariant: specifically, let $\mathcal{A}_N^{(n)} = \{J \subset \{1,\dots,N\} : \ |J| = n\}$ and $\bar{Y}_J = \frac{1}{n}\sum_{i\in J} Y_i$, and define

$$\widehat{\mu}_N^U := \operatorname{med}\left(\bar{Y}_J,\ J \in \mathcal{A}_N^{(n)}\right).$$

That is, $\widehat{\mu}_N^U$ is the geometric median of the means evaluated over all possible subsets of the data of given size n. Since $\operatorname{card}\left(\mathcal{A}_N^{(n)}\right) = \binom{N}{n}$ is too large, we can approximate $\widehat{\mu}_N^U$ via iterating the

following procedure: (1) apply a random permutation to the sample $Y_1, \dots, Y_N$; (2) partition the permuted sample into k disjoint blocks of size n; (3) compute the sample means over the resulting blocks. After repeating the process l times, we obtain a set $k \cdot l$ sample means. We then compute the geometric median of this set. We tested this procedure with $l = 10$ in our numerical simulations described in section 5 below, and confirmed its excellent performance. Theoretical justification of the advantages of this approach in the case of multivariate data are left for the future work.

4. Numerical error bounds. In this section we first propose a practical numerical procedure for approximating the geometric median which admits provable error bounds based on Theorem 2.3. The method itself is an accelerated gradient descent of a smooth relaxation of the mean norm deviation function $F(z; \{y_j\}_{j=1}^k) = \frac{1}{k}\sum_{j=1}^k \|z - y_j\|$. While this method enjoys theoretical support, we found that it can still be improved in practice. This improvement is achieved by Newton's method applied to successively weaker smooth relaxations of $F(z)$ is fast in simulations, but does not enjoy the same theoretical guarantees. Despite this theoretical gap, Theorem 2.3 ensures that we can always check if the output of any numerical routine satisfies error thresholds.

This rest of the exposition is organized as follows: subsection 4.1 details the aforementioned algorithms, subsection 4.2 provides theoretical analysis of the accelerated gradient method, and finally subsection 4.3 contains the proof of Theorem 2.3.

4.1. Algorithms. Given $\{y_i\}_{i=1}^k \subset \mathbb{R}^d$ and $\delta > 0$, a smooth relaxation of the mean norm deviation function $F(z)$ is defined via

$$F_\delta(z) = F_\delta\left(z; \{y_i\}_{i=1}^k\right) = \frac{1}{k}\sum_{i=1}^k \sqrt{\|z - y_i\|^2 + \delta^2}. \tag{4.1}$$

This relaxation involves terms reminiscent of the Charbonnier loss, so we call this function and the associated minimization program the Charbonnier relaxation of the mean norm deviation. The main advantage of the Charbonnier relaxation is that, unlike $F(z)$ the function $F_\delta(z)$ is smooth, therefore, one may perform accelerated gradient descent to approximate its critical point. At the same time, we show that for small δ, this critical point must be close to the minimizer of $F(z)$. Algorithm 4.1 below is just Nesterov's accelerated gradient descent [40] applied to the Charbonnier relaxation with $\delta = \varepsilon/2$ for $\varepsilon > 0$ being the desired error threshold.

Using the standard results for accelerated gradient descent and simple sub-optimality bounds, we establish the following estimate.

Theorem 4.1. *The output of Algorithm 4.1 satisfies $F(x^{(t)}) - F(\widehat{m}) < \varepsilon$.*

The proof is given in section 4.2. Recall that, in view of quadratic growth condition proved in Theorem 2.3, whenever $\|z - \widehat{m}\|$ is small, it behaves like

$$\sqrt{F(z) - F(\widehat{m})}.$$

In turn, it implies that the Restarted Gradient Descent algorithm [51] achieves an iteration complexity of order $\mathcal{O}(\varepsilon^{-1}\log(\varepsilon^{-1}))$ for computation of the geometric median. However, the

Algorithm 4.1 Accelerated Gradient Descent of the Charbonnier Relaxation

Require: $\varepsilon > 0$, $t = 0$, $\alpha_0 = 3/4$, $\{y_i\}_{i=1}^k \in \mathbb{R}^d$, $x^{(0)} = v^{(0)} = \bar{y}_k := k^{-1}\sum_{i=1}^k y_i$

  while $\varepsilon/2 \leq \frac{16}{9(t+1)^2}\left[F_{\varepsilon/2}(\bar{y}_k) + \frac{9}{4\varepsilon}F^2_{\varepsilon/2}(\bar{y}_k)\right]$ **do**

    $x^{(t+1)} \leftarrow v^{(t)} - \frac{\varepsilon}{2}\nabla F_{\varepsilon/2}(v^{(t)})$

    $\alpha_{t+1} \leftarrow \frac{\alpha_t}{2}\left(\sqrt{\alpha_t^2 + 4} - \alpha_t\right)$

    $v^{(t+1)} \leftarrow x^{(t+1)} + \frac{\alpha_t(1-\alpha_t)}{\alpha_t^2 + \alpha_{t+1}}(x_{t+1} - x_t)$

    $t \leftarrow t + 1$

  end while

  return $x^{(t)}$

algorithm 4.1 admits the better iteration complexity $\mathcal{O}(\varepsilon^{-1})$ due to the strong convexity of the Charbonnier relaxation.

In practice, we found that a simple second order method outperforms Algorithm 4.1. It proceeds by decreasing δ by a fixed multiplicative factor and performing a single step of Newton's method on F_δ for each iteration until satisfaction of the stopping rule described in Corollary 2.5. However, we note that, unlike algorithm 4.1, no rigorous analysis for algorithm 4.2 is currently known to us.

Algorithm 4.2 Newton's Method for Successive Charbonnier Relaxation

Require: $\varepsilon > 0$, $t = 0$, $\tau = 1$, $M > 1$, $\{y_i\}_{i=1}^k \in \mathbb{R}^d$, $x^{(0)} = \bar{y}_k$; constants a, b defined in the statement of Theorem 2.3.

  while $\frac{a\varepsilon}{2b^2(\varepsilon+b)} \leq \|\nabla F(x^{(t)})\|$ **do**

    $\tau \leftarrow \tau/M$

    $x^{(t+1)} \leftarrow x^{(t)} - \left(\nabla^2 F_\tau(x^{(t)})\right)^{-1}\nabla F_\tau(x^{(t)})$

    $t \leftarrow t + 1$

  end while

  return $x^{(t)}$

4.2. Proof of Theorem 4.1. To prove Theorem 4.1, we first exhibit simple sub-optimality bounds for the solution to the Charbonnier relaxation. Since F_δ is smooth, the Lipschitz constant for ∇F_δ equals $\sup_{x\in\mathbb{R}^d}\|\nabla^2 F_\delta(x)\|$. The following estimate is straightforward:

$$\|\nabla^2 F_\delta(x)\| = \left\|\frac{1}{k}\sum_{i=1}^k \frac{1}{\sqrt{\|x-y_i\|^2+\delta^2}}\left(I - \frac{1}{\|x-y_i\|^2+\delta^2}(y_i-x)(y_i-x)^T\right)\right\| \leq \frac{1}{\delta}.$$

Indeed, it follows from the relation

$$\left\|I - \frac{1}{\|x-y_i\|^2+\delta^2}(y_i-x)(y_i-x)^T\right\| \leq 1$$

that holds for all $i = 1, \ldots, k$. Thus, we may minimize F_δ using the "constant step scheme II" method on page 93 of the book by Nesterov [40]. In the notation of [40], we have that $q_f = 0$, so we get the following sub-optimality bound.

Lemma 4.2. *Suppose* $\{y_i\}_{i=1}^k \subset \mathbb{R}^d$, $\delta > 0$, *and let* $x_{*,\delta}$ *minimize* F_δ. *Then the updates* x_t *of algorithm 4.1 satify*

$$F_\delta(x^{(t)}) - F_\delta(x_{*,\delta}) \leq \frac{16}{9(t+1)^2}\left[F_\delta(\bar{y}_k) + \frac{9}{8\delta}F_\delta^2(\bar{y}_k)\right]$$

for all $t \geq 0$.

Proof. Theorem 2.2.3 in the book by Nesterov [40] holds with $q_f = 0$ since

$$\alpha_0 = \frac{3}{4} \leq \frac{6}{3+\sqrt{21}}$$

in accordance with condition (2.2.21) of the book. This yields the bound

$$F_\delta(x^{(t)}) - F_\delta(x_{*,\delta}) \leq \frac{4L}{\gamma_0(k+1)^2}\left[F_\delta(x^{(0)}) - F_\delta(x_{*,\delta}) + \frac{\gamma_0}{2}\|x^{(0)} - x_{*,\delta}\|^2\right]$$

with $L = \frac{1}{\delta}$ and $\gamma_0 = \frac{\alpha_0^2}{1-\alpha_0} = \frac{9}{4}L$. The desired result follows from this bound when we also invoke the inequalities $F_\delta(\bar{y}_k) - F_\delta(x_{*,\delta}) \leq F_\delta(\bar{y}_k)$ and $\|\bar{y}_k - x_{*,\delta}\| \leq F_\delta(\bar{y}_k)$. ∎

We are now ready to prove Theorem 4.1.

Proof of Theorem 4.1. Using sub-additivity of the square root, observe that

$$F(z) \leq F_\delta(z) \leq F(z) + \delta$$

for all $z \in \mathbb{R}^d$. Consequently, if $\widehat{m}$ minimizes F and $x_{*,\delta}$ minimizes F_δ, then

$$F(\widehat{m}) \leq F(x_{*,\delta}) \leq F_\delta(x_{*,\delta}) \leq F(\widehat{m}) + \delta.$$

Therefore, a $\frac{\varepsilon}{2}$ numerical approximation $\tilde{x}$ to $x_{*,\varepsilon/2}$ satisfies

$$\begin{aligned} F(\tilde{x}) - F(\widehat{m}) &\leq F_\delta(\tilde{x}) - F(\widehat{m}) \\ &= F_\delta(\tilde{x}) - F_\delta(x_{*,\varepsilon/2}) + F_\delta(x_{*,\varepsilon/2}) - F(\widehat{m}) \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

With $\tilde{x} = x^{(t)}$ being the output of Algorithm 4.1, the termination condition and the bound of Lemma 4.2 together imply the result. ∎

Remark 4.3. Note that $F_\delta(x^{(t)}) - F_\delta(x_{*,\delta}) \leq \delta$ whenever $\frac{4}{3\sqrt{\delta}}\left[f + \frac{9}{8\delta}f^2\right]^{1/2} - 1 \leq t$, where $f = F_\delta(\bar{y}_k)$. When $\delta = \varepsilon/2$, we see that the required number of iterations t is of order$1/\varepsilon$. While computer science literature (e.g. [12]) exhibits better bounds than these, the numerical constants in those algorithms can be impractical.

4.3. Proof of Theorem 2.3. Fix $z \in \mathbb{R}^d$ with $z \neq \widehat{m}$, let $r = \|z - \widehat{m}\|$, and set $u = \frac{1}{r}(z - \widehat{m})$. In view of the second fundamental theorem of calculus,

$$\begin{aligned} F(z) - F(\widehat{m}) &= \int_0^r \nabla F(\widehat{m} + tu)^T u dt \\ &= \frac{1}{k} \int_0^r \sum_{i=1}^k \frac{1}{\|\widehat{m} - y_i + tu\|} (\widehat{m} - y_i + tu)^T u dt \\ &= \frac{1}{k} \int_0^r \sum_{i=1}^k \frac{(\widehat{m} - y_i)^T u + t}{\sqrt{\|\widehat{m} - y_i\|^2 + 2t(\widehat{m} - y_i)^T u + t^2}} dt \\ &= \frac{1}{k} \int_0^r \sum_{i=1}^k \frac{\gamma_i c_i + t}{\sqrt{(\gamma_i c_i + t)^2 + \gamma_i^2(1 - c_i^2)}} dt. \end{aligned}$$

In this last line, we set $\gamma_i = \|\widehat{m} - y_i\|$ and $c_i = \frac{1}{\gamma_i}(\widehat{m} - y_i)^T u$. By the Cauchy-Schwarz inequality, we have that $c_i^2 \leq 1$. If $c_i^2 = 1$, then

$$\frac{\gamma_i c_i + t}{\sqrt{(\gamma_i c_i + t)^2 + \gamma_i^2(1 - c_i^2)}} = \operatorname{sgn}(\gamma_i c_i + t) \geq c_i$$

for all $t > 0$. If $c_i^2 < 1$, then

$$\frac{\gamma_i c_i + t}{\sqrt{(\gamma_i c_i + t)^2 + \gamma_i^2(1 - c_i^2)}} = c_i + \int_0^t \frac{\gamma_i^2(1 - c_i^2)}{\left[(\gamma_i c_i + s)^2 + \gamma_i^2(1 - c_i^2)\right]^{3/2}} ds.$$

Note that $\sum_{i=1}^k c_i = \nabla F(\widehat{m})^T u = 0$ since $\widehat{m}$ is the minimizer. Consequently, we have that

$$\begin{aligned} F(z) - F(\widehat{m}) &\geq \frac{1}{k} \int_0^r \left(\sum_{i=1}^k c_i + \sum_{i: c_i^2 < 1} \int_0^t \frac{\gamma_i^2(1 - c_i^2)}{\left[(\gamma_i c_i + s)^2 + \gamma_i^2(1 - c_i^2)\right]^{3/2}} ds \right) dt \\ &= \frac{1}{k} \sum_{i: c_i^2 < 1} \int_0^r \int_0^t \frac{\gamma_i^2(1 - c_i^2)}{\left[(\gamma_i c_i + s)^2 + \gamma_i^2(1 - c_i^2)\right]^{3/2}} ds\, dt \\ &= \frac{1}{k} \sum_{i: c_i^2 < 1} \int_0^r \int_0^t \frac{1 - c_i^2}{\gamma_i} \frac{1}{\left[(c_i + \frac{s}{\gamma_i})^2 + (1 - c_i^2)\right]^{3/2}} ds\, dt. \end{aligned}$$

Given that

$$\left(c_i + \frac{s}{\gamma_i}\right)^2 + (1 - c_i^2) = \frac{s^2}{\gamma_i^2} + 2c_i \frac{s}{\gamma_i} + 1 \leq \frac{s^2}{\gamma_i^2} + 2\frac{s}{\gamma_i} + 1 = \left(1 + \frac{s}{\gamma_i}\right)^2,$$

we obtain the lower bound

$$\begin{aligned}
F(z) - F(\widehat{m}) &\geq \frac{1}{k} \sum_{i:c_i^2<1} \int_0^r \int_0^t \frac{1-c_i^2}{\gamma_i} \frac{1}{\left[(\frac{s}{\gamma_i}+1)^2\right]^{3/2}} ds\, dt \\
&= \frac{1}{k} \sum_{i:c_i^2<1} \int_0^r \int_0^t \frac{1-c_i^2}{\gamma_i} \frac{1}{(\frac{s}{\gamma_i}+1)^3} ds\, dt \\
&= \frac{1}{k} \sum_{i:c_i^2<1} \int_0^r \int_0^t \frac{\gamma_i^2(1-c_i^2)}{(s+\gamma_i)^3} ds\, dt \\
&= \frac{1}{k} \left(\sum_{j=1}^k \gamma_j^2(1-c_j^2) \right) \int_0^r \int_0^t \sum_{i=1}^k \frac{\gamma_i^2(1-c_i^2)}{\sum_{j=1}^k \gamma_j^2(1-c_j^2)} \frac{1}{(s+\gamma_i)^3} ds\, dt.
\end{aligned}$$

Noting that the inverse cubic function is convex, Jensen's inequality and straightforward integration yields

$$\begin{aligned}
F(z) - F(\widehat{m}) &\geq \frac{1}{k} \left(\sum_{j=1}^k \gamma_j^2(1-c_j^2) \right) \int_0^r \int_0^t \frac{1}{\left(s + \frac{\sum_{i=1}^k \gamma_i^3(1-c_i^2)}{\sum_{j=1}^k \gamma_j^2(1-c_j^2)} \right)^3} ds\, dt \\
&= \frac{1}{2k} \left(\sum_{j=1}^k \gamma_j^2(1-c_j^2) \right) \frac{r^2}{\left(\frac{\sum_{i=1}^k \gamma_i^3(1-c_i^2)}{\sum_{j=1}^k \gamma_j^2(1-c_j^2)} \right)^2 \left(r + \frac{\sum_{i=1}^k \gamma_i^3(1-c_i^2)}{\sum_{j=1}^k \gamma_j^2(1-c_j^2)} \right)}.
\end{aligned}$$

We now observe that

$$\sum_{i=1}^k \gamma_i^3(1-c_i^2) \leq \sum_{i=1}^k \|\widehat{m} - y_i\|^3 \leq \sum_{i=1}^k (\|\widehat{m}\| + \|y_i\|)^3 \leq \sum_{i=1}^k \left(\frac{2}{k} F(0) + \|y_i\| \right)^3$$

and also that

$$\sum_{i=1}^k \gamma_i^2(1-c_i^2) = \sum_{i=1}^k \|\widehat{m} - y_i\|^2 - \left((\widehat{m} - y_i)^T u \right)^2 = \sum_{i=1}^k \sum_{j=2}^d u_j^T (\widehat{m} - y_i)(\widehat{m} - y_i)^T u_j,$$

where $\{u, u_2, \ldots, u_d\}$ is an orthonormal basis of $\mathbb{R}^d$. We further notice that

$$\begin{aligned}
\sum_{i=1}^k (\widehat{m} - y_i)(\widehat{m} - y_i)^T &= \sum_{i=1}^k (\widehat{m} - \overline{x} + \overline{x} - y_i)(\widehat{m} - \overline{x} + \overline{x} - y_i)^T \\
&= k(\widehat{m} - \overline{x})(\widehat{m} - \overline{x})^T + \sum_{i=1}^k (y_i - \overline{x})(y_i - \overline{x})^T.
\end{aligned}$$

The Courant-Fischer characterization of the eigenvalues gives the inequality

$$\sum_{i=1}^k \gamma_i^2(1-c_i^2) \geq \sum_{j=2}^d u_j^T \left(\sum_{i=1}^k (y_i - \overline{x})(y_i - \overline{x})^T \right) u_j \geq k \sum_{j=2}^d \lambda_j(\widehat{\Sigma}),$$

where $\{\lambda_j(\widehat{\Sigma})\}_{j=1}^d$ are the eigenvalues of $\widehat{\Sigma}$ listed with multiplicity and in the non-increasing order. We therefore deduce that

$$F(z) - F(\widehat{m}) \geq \frac{1}{2} \frac{\sum_{j=2}^d \lambda_j(\widehat{\Sigma}) r^2}{\left(\frac{\frac{1}{k}\sum_{i=1}^k (2\nu_1 + \|y_i - \overline{x}\|)^3}{\sum_{j=2}^d \lambda_j(\widehat{\Sigma})}\right)^2 \left(r + \frac{\frac{1}{k}\sum_{i=1}^k (2\nu_1 + \|y_i - \overline{x}\|)^3}{\sum_{j=2}^d \lambda_j(\widehat{\Sigma})}\right)},$$

thus completing the proof of Theorem 2.3.

5. Numerical experiments. We consider 21 years worth of daily adjusted closing prices for a subset of 361 symbols from the S&P 500, and attempt to predict year-to-year average log-return vectors for these adjusted closing prices across the symbols. We retrieve the data in the date range January 1st, 2003 to December 31st, 2023 from Yahoo Finance using the python package "yfinance, " yielding 361-dimensional examples with approximately 250 examples per year for 21 years (or, for 20 pairs of consecutive years). For each estimator and each pair of consecutive years in the dataset, we use the estimator fit the first year's data as a prediction for the average log-return vector of the second year. That is, if $\mu_{t+1} \in \mathbb{R}^{361}$ is the mean of log-returns for the year $t+1$, we estimate it using the log-return data $X_j^{(t)},\ j = 1, \ldots, 250$ for the year t. This framework is based on a simple model $X_j^{(t)} = \mu_t + Z_j^{(t)},\ j = 1, \ldots, 250$ where $\mu_t \in \mathbb{R}^{361}$ and $Z_j^{(t)},\ t = 1, \ldots, 21,\ j = 1, \ldots, 250$ are i.i.d. random vectors. We assume that μ_t is varying slowly so that $\mu_t \approx \mu_{t+1}$.

While heavy tails notoriously appear in log-returns data, we must note that inclusion of a symbol in the S&P 500 generally entails lower volatility; this fact also justifies our assumption regarding the slowly varying trend. As such, our experiment only probes "robustness" of our estimators over a limited number of "shocks" (between 2007 and 2008, and also between 2008 and 2009) where the mean changes substantially, and "non-inferiority" of the estimator in less volatile regimes.

We compare several estimators: the standard mean (mean), the entry-wise median (median), the geometric median (g median), the geometric MOM with $k = 5$ blocks (gMOM5), the geometric MOM with $k = 10$ blocks (gMOM10), and the approximation of the permutation-invariant version of the MOM estimator described in remark 3.15, where we set the value of parameter l to 10 (gMOM Rep). Since the standard MOM estimators involve random partitions of the data, we perform the experiment 25 times for each estimation problem and take the mean of the resulting errors to indicate average behavior.

Figure 1 indicates the "training errors" for the different estimators in the sense that we consider the relative norm error

$$\frac{\|\widehat{\mu} - \bar{\mu}_{t+1}\|}{\|\bar{\mu}_{t+1}\|}$$

where $\widehat{\mu}$ is the output of the estimator on the first year's data and $\bar{\mu}_{t+1}$ is the empirical mean for the second year. In particular, this shows that the geometric MOM estimators are generally much closer to the standard mean than both the entry-wise median estimator and the geometric median estimator.

Figure 2 displays the prediction errors for using the different estimators to predict the next year's average log-return. In this plot, year indices are sorted so that the prediction errors

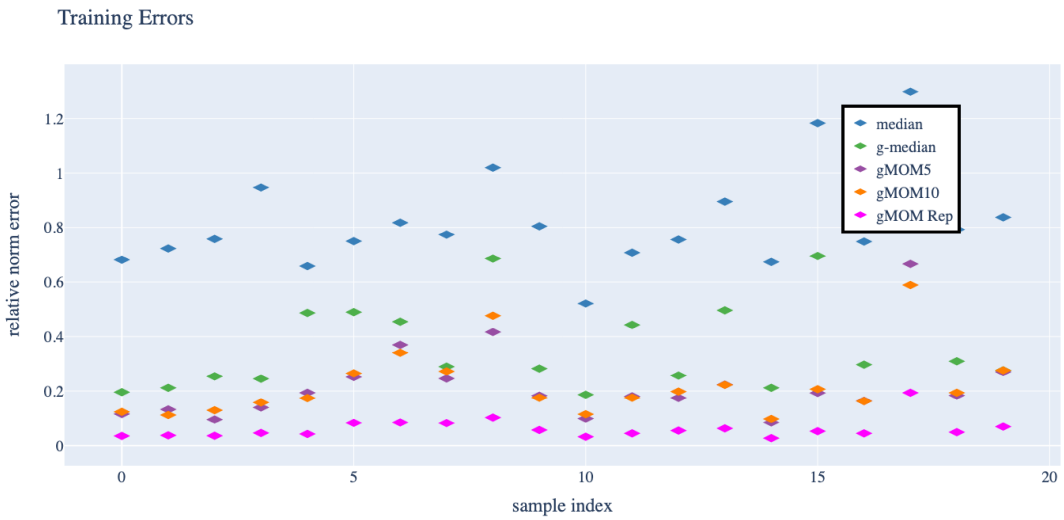

Figure 1: Relative norm differences between the mean and various estimators over 20 years of data illustrate how the MOM estimators track the standard mean more closely than the entry-wise median and geometric median.

$\|\bar{\mu}_t - \bar{\mu}_{t+1}\|$ are non-increasing. From these examples, we see that MOM estimators track $\widehat{\mu}_t$ closely across all samples, while the entry-wise median estimator and geometric median estimators exhibit some inferiority in the less volatile regime and somewhat better behavior for the two "shocks" where $\|\bar{\mu}_t - \bar{\mu}_{t+1}\|$ is largest (between 2007 and 2009). We also note that the error of the geometric median estimator often lies between the errors of the entry-wise median and the mean.

Figure 3 illustrates these observations in aggregate. The entry-wise median enjoys better maximum error, but is generally inferior to the mean. The geometric median exhibits somewhat better outlier behavior than the mean, and comparable behavior otherwise. This trend continues with the geometric MOM estimators, which exhibit slightly better outlier behavior. The permutation-invariant geometric MOM estimator appears comparable to the mean across the entire dataset for this example.

REFERENCES

[1] D. Alistarh, Z. Allen-Zhu, and J. Li, *Byzantine stochastic gradient descent*, Advances in Neural Information Processing Systems, 31 (2018).

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

[3] A.-H. Bateni, A. Minasyan, and A. S. Dalalyan, *Nearly minimax robust estimator of the mean vector by iterative spectral dimension reduction*, arXiv preprint arXiv:2204.02323, (2022).

[4] A. Beck and S. Sabach, *Weiszfeld's method: Old and new results*, Journal of Optimization Theory and Applications, 164 (2015), pp. 1–40.

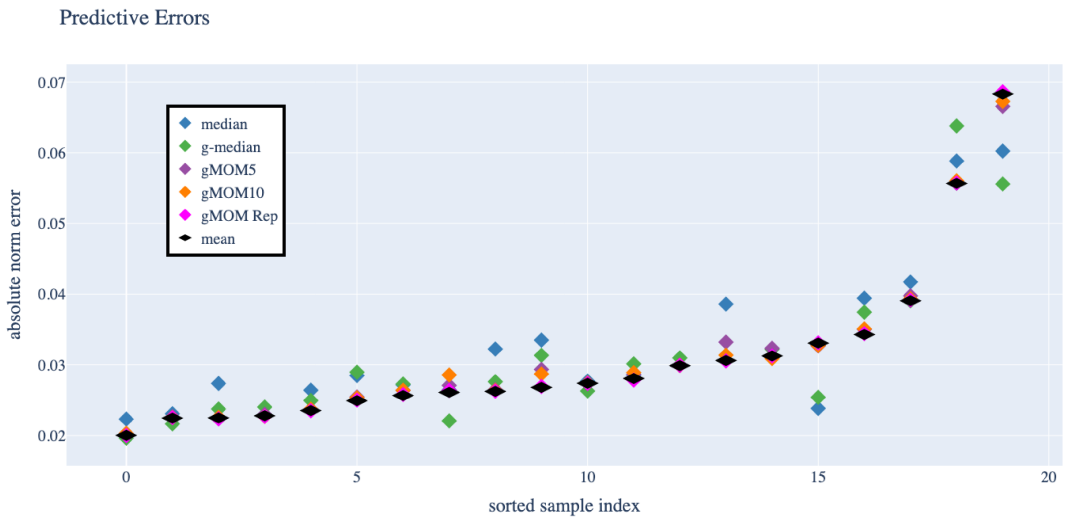

Figure 2: Plotting various estimators' prediction errors for next year's average log-returns vector over 361 symbols from the S&P 500 clearly displays two examples of outlier behavior. Here, the indices for the various years are sorted by increasing error of the mean to aid legibility of comparisons. The final two indices correspond to the period between 2007 and 2009.

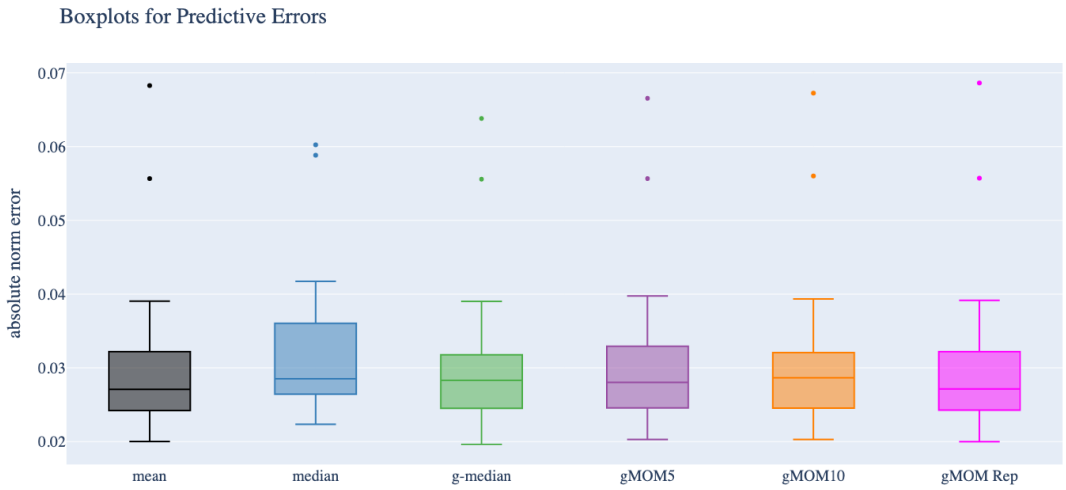

Figure 3: Box plots for the errors of various estimators generally indicate that the geometric median methods provide better control of outliers and comparable behavior otherwise.

[5] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Oxford university press, 2013.
[6] D. Bouhata and H. Moumen, *Byzantine fault tolerance in distributed machine learning: a survey*, arXiv

preprint arXiv:2205.02572, (2022).
[7] H. Cardot, P. Cénac, and A. Godichon-Baggioni, *Online estimation of the geometric median in Hilbert spaces: Nonasymptotic confidence balls*, (2017).
[8] H. Cardot, P. Cenac, P.-A. Zitt, et al., *Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm*, Bernoulli, 19 (2013), pp. 18–43.
[9] P. Chaudhuri, *On a geometric notion of quantiles for multivariate data*, Journal of the American statistical association, 91 (1996), pp. 862–872.
[10] Y. Chen, L. Su, and J. Xu, *Distributed statistical machine learning in adversarial settings: Byzantine gradient descent*, Proceedings of the ACM on Measurement and Analysis of Computing Systems, 1 (2017), pp. 1–25.
[11] Y. Cherapanamjeri, N. Flammarion, and P. L. Bartlett, *Fast mean estimation with sub-Gaussian rates*, in Conference on Learning Theory, PMLR, 2019, pp. 786–806.
[12] M. B. Cohen, Y. T. Lee, G. Miller, J. Pachocki, and A. Sidford, *Geometric median in nearly linear time*, in Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, ACM, 2016, pp. 9–21.
[13] J. Depersin and G. Lecué, *Robust sub-Gaussian estimation of a mean vector in nearly linear time*, The Annals of Statistics, 50 (2022), pp. 511–536.
[14] I. Diakonikolas and D. Kane, *Recent advances in algorithmic high-dimensional robust statistics*, in "Beyond the worst-case analysis of algorithms", Cambridge University Press, 2021.
[15] S. Dirksen, *Tail bounds via generic chaining*, Electron. J. Probab, 20 (2015), pp. 1–29.
[16] E. Giné and R. Nickl, *Mathematical foundations of infinite-dimensional statistical models*, vol. 40, Cambridge University Press, 2015.
[17] C. Gini and L. Galvani, *Di talune estensioni dei concetti di media ai caratteri qualitativi*, Metron, 8 (1929), pp. 3–209.
[18] E. Gluskin and V. Milman, *Note on the geometric-arithmetic mean inequality*, in Geometric Aspects of Functional Analysis: Israel Seminar 2001-2002, Springer, 2003, pp. 131–135.
[19] J. B. S. Haldane, *Note on the median of a multivariate distribution*, Biometrika, 35 (1948), pp. 414–417.
[20] S. B. Hopkins, *Mean estimation with sub-Gaussian rates in polynomial time*, (2020).
[21] D. Hsu and S. Sabato, *Loss minimization and parameter estimation with heavy tails*, Journal of Machine Learning Research, 17 (2016), pp. 1–40.
[22] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani, *Random generation of combinatorial structures from a uniform distribution*, Theoretical computer science, 43 (1986), pp. 169–188.
[23] T. Juškevičius and J. Lee, *Small ball probabilities, maximum density and rearrangements*, arXiv preprint arXiv:1503.09190, (2015).
[24] T. Kärkkäinen and S. Ayrämö, *On computation of spatial median for robust data mining*, Evolutionary and Deterministic Methods for Design, Optimization and Control with Applications to Industrial and Societal Problems, EUROGEN, Munich, (2005).
[25] J. Kemperman, *The median of a finite measure on a Banach space*, Statistical data analysis based on the L_1-norm and related methods, (1987), pp. 217–230.
[26] V. Koltchinskii, *Spatial quantiles and their Bahadur-Kiefer representations*, in Asymptotic Statistics: Proceedings of the Fifth Prague Symposium, held from September 4–9, 1993, Springer, 1994, pp. 361–367.
[27] V. I. Koltchinskii, *M-estimation, convexity and quantiles*, Ann. Statist., 25 (1997), pp. 435–477.
[28] D. Konen, *Recovering a probability measure from its multivariate spatial rank*, arXiv preprint arXiv:2208.11551, (2022).
[29] R. Latala and K. Oleszkiewicz, *Small ball probability estimates in terms of width*, Studia mathematica, 169 (2005), pp. 305–314.
[30] M. Lerasle and R. I. Oliveira, *Robust empirical mean estimators*, arXiv preprint arXiv:1112.3914, (2011).
[31] G. Lugosi and S. Mendelson, *Sub-Gaussian estimators of the mean of a random vector*, arXiv preprint arXiv:1702.00482, (2017).
[32] G. Lugosi and S. Mendelson, *Mean estimation and regression under heavy-tailed distributions: A survey*, Foundations of Computational Mathematics, 19 (2019), pp. 1145–1190.
[33] G. Lugosi and S. Mendelson, *Multivariate mean estimation with direction-dependent accuracy*, arXiv

preprint arXiv:2010.11921, (2020).
[34] M. Madiman, J. Melbourne, and P. Xu, *Rogozin's convolution inequality for locally compact groups*, arXiv preprint arXiv:1705.00642, (2017).
[35] S. Mendelson, *Learning without concentration*, Journal of the ACM (JACM), 62 (2015), pp. 1–25.
[36] S. Mendelson and N. Zhivotovskiy, *Robust covariance estimation under L_4-L_2 norm equivalence*, (2020).
[37] S. Minsker, *Geometric median and robust estimation in Banach spaces*, Bernoulli, 21 (2015), pp. 2308–2335.
[38] S. Minsker, *U-statistics of growing order and sub-Gaussian mean estimators with sharp constants*, arXiv preprint arXiv:2202.11842, (2022).
[39] A. Nemirovski and D. Yudin, *Problem complexity and method efficiency in optimization*, John Wiley and Sons, 1983.
[40] Y. Nesterov, *Lectures on convex optimization*, vol. 137, Springer, 2018.
[41] R. I. Oliveira, *The lower tail of random quadratic forms with applications to ordinary least squares*, Probability Theory and Related Fields, 166 (2016), pp. 1175–1194.
[42] L. M. Ostresh, *On the convergence of a class of iterative methods for solving the Weber location problem*, Operations Research, 26 (1978), pp. 597–609.
[43] M. L. Overton, *A quadratically convergent method for minimizing a sum of Euclidean norms*, Mathematical Programming, 27 (1983), pp. 34–63.
[44] K. Pillutla, S. M. Kakade, and Z. Harchaoui, *Robust aggregation for federated learning*, IEEE Transactions on Signal Processing, 70 (2022), pp. 1142–1154.
[45] E. Rio, *Moment inequalities for sums of dependent random variables under projective conditions*, Journal of Theoretical Probability, 22 (2009), pp. 146–163.
[46] G. Romon, *Statistical properties of approximate geometric quantiles in infinite-dimensional Banach spaces*, arXiv preprint arXiv:2211.00035, (2022).
[47] M. Rudelson and R. Vershynin, *Small ball probabilities for linear images of high-dimensional distributions*, International Mathematics Research Notices, 2015 (2015), pp. 9594–9617.
[48] Y. Vardi and C.-H. Zhang, *The multivariate L_1-median and associated data depth*, Proceedings of the National Academy of Sciences, 97 (2000), pp. 1423–1426.
[49] A. Weber, *Uber den standort der industrien (Alfred Weber's theory of the location of industries)*, University of Chicago, (1929).
[50] E. Weiszfeld, *Sur un problème de minimum dans l'espace*, Tohoku Mathematical Journal, (1936).
[51] T. Yang and Q. Lin, *Rsg: Beating subgradient method without smoothness and strong convexity*, arXiv preprint arXiv:1512.03107, (2015).
[52] N. Zhivotovskiy, *Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle*, arXiv preprint arXiv:2108.08198, (2021).

Appendix A. Notation. Here, we collect some of the key notation scattered throughout the paper.

1. m or $m(P)$ stand for the (geometric) median of the distribution P; $\widehat{m}$ stands for the median corresponding to the empirical distribution based on a sample from P.
2. $F(z)$ denotes the function $F(z) = \frac{1}{k}\sum_{j=1}^{k} \|z - Y_j\|$.
3. $\widehat{\mu}_N$ stands for the median of means estimator based on the sample $Y_1, \dots, Y_N$.
4. $M(Y)$ stands for the sup-norm of the density p_Y of random vector Y that is absolutely continuous with respect to the volume measure on a linear subspace of $\mathbb{R}^d$.
5. The spatial sign covariance matrix D_Y is defined via
$$D_Y := \mathbb{E}\left[\frac{(Y-m)}{\|Y-m\|}\frac{(Y-m)^T}{\|Y-m\|}\right],$$
and its spectral norm is $\Delta = \|D_Y\|$.
6. m_n stands the geometric median of the distribution $P^{(n)}$ of the average $\frac{1}{n}\sum_{j=1}^{n} Y_j$ of

i.i.d. random vectors.