The geometric median and its applications to robust mean estimation

Stanislav Minsker [1,*] **and Nate Strawn** [2,**]

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

[2] *Department of Mathematics and Statistics, Georgetown University*
e-mail: [**] nate.strawn@georgetown.edu

Abstract: This paper is devoted to the statistical properties of the geometric median, a robust measure of centrality for multivariate data, as well as its applications to the problem of mean estimation via the median of means principle. Our main theoretical results include (a) the 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 the improved bounds for the multivariate median of means estimator that hold for large classes of heavy-tailed distributions.

MSC 2010 subject classifications: Primary 62G35; secondary 60E15.
Keywords and phrases: median-of-means estimator, heavy tails.

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 (Weber, 1929) under the name of "Fermat-Weber" point, and to Haldane (Haldane, 1948) under the name of Haldane's median; other notable early references include the work by Gini and Galvani (1929). The geometric median is an element of a more general family of spatial quantiles that was introduced and studied in detail by Koltchinskii (1994, 1997); Chaudhuri (1996): 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 Kemperman (1987) and, more recently, by Romon (2022). Deep connections between the probability distributions and the corresponding spatial quantiles have been investigated by Konen (2022).

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 (1983); Jerrum et al. (1986); Alon et al. (1996) in different contexts, MOM estimator was shown to be a powerful tool for the analysis of corrupted and heavy-tailed data by Lerasle and Oliveira (2011). Hsu and Sabato (2016) demonstrated multiple novel applications of the original estimator by Nemirovski and Yudin (1983) in general metric spaces, while Minsker (2015) 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,

[*]S. Minsker acknowledges support by the National Science Foundation grants DMS CAREER-2045068 and CCF-1908905.

efficiently computable even in high dimensions, requires minimal tuning, and admits strong theoretical guarantees in many circumstances. However, as was pointed out by Lugosi and Mendelson (2017), 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 (2015), 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 (3) below, such that

$$\|\widehat{\mu}_N - \mu\| \le C\sqrt{\frac{\operatorname{tr}(\Sigma_Y)t}{N}} \tag{1}$$

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 the 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,

$$\|\widetilde{\mu}_N - \mu\| \le C\left(\sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{N}} + \sqrt{\|\Sigma_Y\|}\sqrt{\frac{t}{N}}\right) \tag{2}$$

with probability at least $1-e^{-t}$, where $C>0$ is an absolute constant; the advantage of the latter inequality over (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 (2) are much smaller compared to their counterpart obtained from (1). Lugosi and Mendelson (2017) 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 proposed in the mathematical statistics and theoretical computer science literature since: we refer the reader to the excellent surveys by Lugosi and Mendelson (2019) and Diakonikolas and Kane (2021). While the original estimator by Lugosi and Mendelson (2017) is difficult to compute, several closely related numerically feasible alternatives have been proposed by Hopkins (2020); Cherapanamjeri et al. (2019); Depersin and Lecué (2022); Bateni et al. (2022), among others. However, to the best of our knowledge, none of these methods have a practical implementation comparable to the best algorithms for evaluating the geometric median (Cohen et al., 2016; Beck and Sabach, 2015; Cardot et al., 2017), with (Mathieu, 2022) being a notable exception; the latter requires slightly stronger assumptions on the underlying distribution however. 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 the Federated learning (see Pillutla et al., 2022; Alistarh et al., 2018; Chen et al., 2017; Bouhata and Moumen, 2022, and references therein). Therefore, improved guarantees for the geometric MOM estimator have immediate implications for a variety of algorithms that use MOM principle as a subroutine.

In this paper, we revisit the original geometric median of means construction and show that the inequality (1) can be improved for large classes of absolutely continuous, heavy-tailed distributions with large effective rank $r(\Sigma_Y) := \frac{\operatorname{tr}(\Sigma_Y)}{\|\Sigma_Y\|}$ (indeed, if $r(\Sigma)$ is bounded by a constant, then (1) readily provides desired sub-Gaussian guarantees). Specifically, we show that $\widehat{\mu}_N$ satisfies the bound of sub-exponential type: for all $t \lesssim N$ (where $\lesssim$ denotes the inequality up to an absolute multiplicative constant), there exists a version of the MOM estimator $\widehat{\mu}_N$ such that

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

with probability at least $1 - e^{-t}$. While the proof of the inequality (1) is based on a simple "majority vote-type" argument, the present analysis blends accurate estimates for the bias and the stochastic error of the geometric median of means. The upper bound for the bias (Theorem 4 and section 2.3) 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 5) 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.

2. Main results.

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

$$\widehat{m} = \operatorname{med}(Y_1, \dots, Y_N) := \operatorname*{argmin}_{z \in \mathbb{R}^d} \sum_{j=1}^{N} \|z - Y_j\|. \tag{3}$$

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 line. Next, 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

$$\widehat{\mu}_N := \operatorname{med}\left(\bar{Y}_1, \dots, \bar{Y}_k\right). \tag{4}$$

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 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.

(A) **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 $\max_{j=1,\dots,k} M(X_j) =: M_0 < \infty$. 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 $\mathcal{P}_1 := \mathcal{P}_1(M_0, K)$.

(B) **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) $M^{1/k}\left(\Sigma_Y^{-1/2} Y\right) \leq M_0$;

(b) $\frac{\operatorname{tr}(\Sigma_Y)}{k \cdot \det^{1/k}(\Sigma_Y)} \leq R$;

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

(c) For some $q > 2$ and all unit vectors u,

$$\mathbb{E}^{2/q} |\langle Y, u\rangle|^q \le K(q) \langle \Sigma_Y u, u\rangle . \tag{5}$$

The class of all such distributions will be denoted $\mathcal{P}_2 := \mathcal{P}_2(k, M_0, K, R)$.

(C) **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 main result of the paper is the following bound.

Theorem 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 (4) satisfies the inequality*

$$\|\widehat{\mu}_N - \mu\| \le C \left(\sqrt{\frac{tr(\Sigma_Y)}{N}} + \sqrt{\|\Sigma_Y\|} \sqrt{\frac{k}{N}} \right) \tag{6}$$

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

Remark 1. *Let us discuss the main assumptions of the theorem.*

1. *Note that $\frac{tr(\Sigma_Y)}{k \cdot \det^{1/k}(\Sigma_Y)}$ is the ratio of the arithmetic and the geometric means of the eigenvalues $\lambda_1 \ge \ldots \ge \lambda_k$ 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*

$$\frac{\sum_{j=1}^k \lambda_j}{k \left(\prod_{i=1}^k \lambda_i \right)^{1/k}} \le C(\alpha).$$

   In fact, it is known that for most (with respect to the uniform distribution on a sphere) sequences, the ratio of arithmetic and geometric means is well-behaved (Gluskin and Milman, 2003).

2. *Moment equivalence conditions similar to (5) are well known in the literature - for example, it has been employed by Mendelson and Zhivotovskiy (2020); Lugosi and Mendelson (2020); Zhivotovskiy (2021); Oliveira (2016), among others, in the contexts of robust estimation and random matrix theory. It is known to hold (Mendelson, 2015, Lemma 4.2) 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 (5) under appropriate restrictions on the number of degrees of freedom. Define the spatial sign covariance matrix via $D_Y := \mathbb{E}\left[\frac{(Y-m)}{\|Y-m\|} \frac{(Y-m)^T}{\|Y-m\|} \right]$, where $m = m(P_Y)$ is the geometric median of Y. The role of assumption (5) is in showing that $\Delta := \|D_Y\| \le \frac{C}{r(\Sigma_Y)}$. When (5) does not hold, inequality (6) is still valid with $\sqrt{\|\Sigma_Y\|}$ replaced by $\max\left(\sqrt{\|\Sigma_Y\|}, \sqrt{\frac{tr(\Sigma_Y)\Delta}{\sqrt{k}}} \right)$.*

The following sections are devoted to the proof of Theorem 1 and the required background, followed by the discussion of numerical methods used to approximate the estimator $\widehat{\mu}_N$. The

proof of Theorem 1 is based on the error decomposition

$$\|\widehat{\mu}_N - \mu\| \le \|m_n - \mu\| + \|\widehat{\mu}_N - m_n\| \tag{7}$$

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

$$\|\widehat{\mu}_N - m_n\| \lesssim \sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{N}} + \sqrt{\Delta \operatorname{tr}(\Sigma_Y)}\sqrt{\frac{s}{N}} + \sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{N}}\frac{s}{\sqrt{k}} \tag{8}$$

with probability at least $1 - 4e^{-s}$ for all $s \lesssim k$. The combination of (7) and (8) yields the desired inequality.

2.1. Preliminaries: small ball probabilities.

For a centered random vector $Z \in \mathbb{R}^d$ with a distribution that is absolutely continuous with respect to the Lebesgue measure, let $M(Z)$ denote the sup-norm of the density p_Z of 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 $V_d(R) = \frac{(\sqrt{\pi}R)^d}{\Gamma(d/2+1)}$ 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

$$\mathbb{P}(\|Z - z\| \le R) \le M\left(\Sigma_Z^{-1/2} Z\right) \frac{V_d(R)}{\sqrt{\det(\Sigma_Z)}}. \tag{9}$$

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:

$$\frac{c_1}{\sqrt{d}}\left(\frac{\sqrt{2\pi e}\, R}{\sqrt{d}}\right)^d \le V_d(R) \le \frac{c_2}{\sqrt{d}}\left(\frac{\sqrt{2\pi e}\, R}{\sqrt{d}}\right)^d. \tag{10}$$

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

Theorem 2 (Theorem 4 in Latala and Oleszkiewicz (2005))**.** *Let Z have multivariate normal distribution $N(0,\Sigma)$ and let M be the median of $\|Z\|$. Then for all $x \in \mathbb{R}^d$,*

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

It is helpful to recall that $c_1\sqrt{\mathrm{tr}\,(\Sigma)} \le M \le c_2\sqrt{\mathrm{tr}\,(\Sigma)}$ 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 (2015).

Theorem 3 (Theorem 1.5 in Rudelson and Vershynin (2015)). *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 $t > 0$,*

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

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

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 (9) 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 Juškevičius and Lee (2015): 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

$$M\left(\Sigma_Y^{-1/2} Z_n\right) \le M\left(\frac{1}{\sqrt{n}}\sum_{j=1}^n U_j\right). \tag{11}$$

The second estimate, established by Madiman et al. (2017), 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 $\mathrm{vol}\left(B\left(R_U \cdot M^{1/d}\left(\Sigma^{-1/2} X\right)\right)\right) = 1$, 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 (11) 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

$$M\left(\Sigma_Y^{-1/2} Z_n\right) \le M\left(\Sigma_Y^{-1/2} Y\right)(2e)^{d/2} \tag{12}$$

and a small ball estimate

$$\mathbb{P}(\|Z_n - z\| \le R) \le c_2 \frac{M\left(\Sigma_Y^{-1/2} Y\right)}{\sqrt{\det(\Sigma_Y)}}\left(\frac{2e\sqrt{\pi}\, R}{\sqrt{d}}\right)^d. \tag{13}$$

2.2. 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 4. *Assume that the distribution of Y is absolutely continuous with respect to Lebesgue measure on some linear subspace of $\mathbb{R}^d$. Then*

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

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 $\mathbb{E}\frac{Y-m}{\|Y-m\|} = 0$, 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,

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

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

The inequality $\|m\| \le \sqrt{\mathrm{tr}\,(\Sigma_Y)}$ 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.

2.3. Equivalence of the negative moments of the norm.

In view of the inequality stated in Theorem 4, 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 1. *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 2 (see Latala and Oleszkiewicz (2005, Corollary 1)) once we notice that the median $M(\|Y\|)$ of $\|Y\|$ satisfies $M(\|Y\|) \ge 0.08\sqrt{\mathrm{tr}\,(\Sigma_Y)}$. 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

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

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

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

which equals 0.5 for $t = 1-\sqrt{\pi}/2 > 0.11$, and the claim follows. To apply Latala and Oleszkiewicz (2005, Corollary 1), 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 2 below also implies a version of Lemma 1. Recall that $M(Y)$ stands for the sup-norm of the probability density function of a random vector Y.

Lemma 2. *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,

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

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 (2015) which states that for any $\varepsilon > 0$, there exists a positive constant C_ε such that for all $x \in \mathbb{R}^d$ and $t > 0$, $\mathbb{P}\Big(\|Y-x\| \leq t\sqrt{\mathrm{tr}\,(\Sigma_Y)}\Big) \leq (C_\varepsilon M_0 t)^{(1-\varepsilon)\widetilde{r}(\Sigma_Y)}$, where $\widetilde{r}(\Sigma_Y) = \left\lfloor \frac{\mathrm{tr}\,(\Sigma_Y)}{\|\Sigma_Y\|} \right\rfloor = \lfloor r(\Sigma_Y) \rfloor$. Employing this "small ball" bound and letting $r := \widetilde{r}(\Sigma_Y)$ for brevity, we deduce that for any $\delta > 0$ and $q < r$,

$$\begin{aligned}\mathbb{E}\|Y-x\|^{-q} &= \int_0^\infty \mathbb{P}\Big(\|Y-x\| \leq t^{1/q}\Big)\frac{dt}{t^2} \\ &= \frac{q}{\left(\mathrm{tr}\,(\Sigma_Y)\right)^{k/2}}\left(\int_{1/\delta}^\infty \frac{ds}{s^{q+1}} + \int_0^{1/\delta}(C_\varepsilon M_0 s)^{r(1-\varepsilon)}\frac{ds}{s^{q+1}}\right).\end{aligned}$$

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

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

If $\varepsilon = \frac{r-q-1/2}{r}$, then $\frac{q}{r(1-\varepsilon)-q} = 2q$ and $\frac{q}{r(1-\varepsilon)} \leq 1$. For small values of q, say, $q \leq r/2$, this choice of ε entails the inequality $\varepsilon > \frac{r-1}{2r} \geq \frac{3}{8}$ for $r \geq 4$, 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. *Assume that $Y \in \mathbb{R}^d$ 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 2. Note that for any $\delta > 0$

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

in view of (13). 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 $\left(\mathbb{E}\|Y - m\|^{-1}\right)^{-1} \le \sqrt{\operatorname{tr}(\Sigma_Y)}$, we immediately get from the previous result that whenever $k \ge 3$, then for some absolute constant $C > 0$

$$\frac{\mathbb{E}^{1/2}\|Y - m\|^{-2}}{\mathbb{E}\|Y - m\|^{-1}} \le C M^{1/k}\left(\Sigma_Y^{-1/2} Y\right) \sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{k \cdot \det^{1/k}(\Sigma_Y)}}. \tag{14}$$

Result of Lemma 3 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 $d \cdot \delta \le Ck \cdot \lambda$. In this case, $\frac{\operatorname{tr}(\Sigma_Y)}{d \cdot \det^{1/d}(\Sigma_Y)}$ can be very large, and direct application of Lemma 3 yields a suboptimal bound. The following simple observation often yields a better result: for any linear subspace H of $\mathbb{R}^d$,

$$\mathbb{E}\|Y - x\|^{-q} \le \mathbb{E}\|\Pi_H(Y - x)\|^{-q}, \tag{15}$$

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

Lemma 4. *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 $k \ge 3$. Then for any $x \in \mathbb{R}^d$,*

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

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 (15), we deduce that

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

where $\widetilde{Y} = X + \Pi_H\xi$. It remains to note that $M(\widetilde{Y}) \le M(X)$ 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 $d\cdot\delta \le Ck\cdot\lambda$, 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 2. *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* $\frac{\sum_{j=1}^d \lambda_j}{\max_{j\ge1} j\left(\prod_{i=1}^j \lambda_i\right)^{1/j}}$ *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 2 - 4), better bounds become possible.*

2.4. 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. Denote

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

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) For some $q > 2$ and all unit vectors u,

$$\mathbb{E}^{2/q}\left|\langle Y,u\rangle\right|^q \le K\left\langle \Sigma_Y u, u\right\rangle.$$

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

$$\begin{aligned}\Delta = \sup_{\|u\|=1}\mathbb{E}\frac{\langle Y-m,u\rangle^2}{\|Y-m\|^2} \le \sup_{\|u\|=1}\mathbb{E}^{2/q}\left|\langle Y-m,u\rangle\right|^q\mathbb{E}^{(q-2)/q}\|Y-m\|^{-\frac{q}{2(q-2)}} \\ \le K\,C(q)\frac{\|\Sigma_Y\|}{\operatorname{tr}(\Sigma_Y)} = \frac{K\,C(q)}{r(\Sigma_Y)}. \quad (16)\end{aligned}$$

Moment equivalence condition (a) has been discussed in detail in section in remark 1. Condition (b) holds for the classes of distributions discussed in section 2.3 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, $\mathrm{tr}\,(\Sigma_Y)\Delta \asymp \|\Sigma_Y\|$: 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 5. *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 $\Delta < 1$ and that*

$$\mathbb{E}^{1/2}\frac{1}{\|Y-m\|^2}\left(\sqrt{\frac{tr\,(\Sigma_Y)}{k}} + \sqrt{\Delta\, tr\,(\Sigma_Y)}\sqrt{\frac{s}{k}} + \sqrt{tr\,(\Sigma_Y)}\frac{s}{k}\right) < c_1(\Delta). \tag{17}$$

Then

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

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

Remark 3.

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 (17) is rather mild: indeed, we showed in section 2.3 that in many common situations, $\mathbb{E}^{1/2}\frac{1}{\|Y-m\|^2} \asymp \frac{1}{\sqrt{tr\,(\Sigma_Y)}}$.*
3. *Note that for $s \lesssim \sqrt{k}$, the bound of the theorem yields deviations guarantees of sub-Gaussian type for the median $\widehat{m}$: namely,*

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

with probability at least $1-4e^{-s}$. Despite the fact that asymptotic properties of the geometric median have been well-understood, we believe that this bound is new.

Proof. Recall that, in view of Theorem 3.1 in Minsker (2015), $\|\widehat{m}-m\| \le 2\sqrt{\mathrm{tr}\,(\Sigma_Y)}$ 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 $\|\widehat{m}-m\| \geq t$ to hold is

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

which is possible only if

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

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

Lemma 5. *With probability at least* $1-e^{-s}$,

$$\left\|\frac{1}{k}\sum_{j=1}^k \frac{Y_j-m}{\|Y_j-m\|}\right\| \leq \frac{2}{\sqrt{k}} + \sqrt{\Delta}\sqrt{\frac{2s}{k}} + \frac{4s}{3k}. \tag{18}$$

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

$$\left\|\frac{1}{k}\sum_{j=1}^k \frac{Y_j-m}{\|Y_j-m\|}\right\| \leq C\left(\frac{1}{\sqrt{k}} + \sqrt{\Delta}\sqrt{\frac{s}{k}}\right) \tag{19}$$

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 the book by Boucheron et al., 2013) 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 Dirksen (e.g., see 2015) in place of Bousquet's inequality. □

Lemma 6. *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 $1-2e^{-k/4}-2e^{-s}$:

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

Remark 4. *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, $\left(\mathbb{E}^{1/2}\|Y_1-m\|^{-2}\right)^{-1} \asymp \sqrt{tr(\Sigma_Y)}$.*

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

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

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

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

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 $\tau > 0$,

$$\sup_{\|u\|=1,\|x-m\|\le t}\frac{1}{k}\sum_{j=1}^{k}\left\langle\frac{Y_j-x}{\|Y_j-x\|},u\right\rangle^2 \le \frac{1+\tau}{k}\sum_{j=1}^{k}\left\langle\frac{Y_j-m}{\|Y_j-m\|},u\right\rangle^2+\frac{4+1/\tau}{k}\sum_{j=1}^{k}\min\left(1,\frac{t^2}{\|Y_j-m\|^2}\right). \quad (20)$$

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

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

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

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

in view of the symmetrization and Talagrand's contraction inequalities (e.g., see Giné and Nickl, 2015). To summarize, we showed that with probability at least $1 - e^{-s}$, for all unit vectors u,

$$\frac{1+\tau}{k}\sum_{j=1}^{k}\left|\left\langle \frac{Y_j - m}{\|Y_j - m\|}, u\right\rangle\right| \le (1+\tau)\left(\sqrt{\Delta}\left(1 + \sqrt{\frac{2s}{k}}\right) + \frac{4}{\sqrt{k}} + \frac{4s}{3k}\right). \tag{21}$$

In view of Bernstein's inequality, the second term in (20) is at most

$$\begin{aligned}(4 + 1/\tau)&\left(\mathbb{E}\min\left(1, \frac{z^2}{\|Y_j - m\|^2}\right) + 2\sqrt{\mathrm{Var}\left(\min\left(1, \frac{z^2}{\|Y - m\|^2}\right)\right)}\sqrt{\frac{s}{k}} + \frac{2s}{3k}\right) \\ &\le (4 + 1/\tau)\left(2t^2\mathbb{E}\frac{1}{\|Y - m\|^2} + \frac{5s}{3k}\right)\end{aligned} \tag{22}$$

with probability at least $1 - e^{-s}$. Combining (20), (21), (22), we deduce the inequality

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

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

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

uniformly over all $\|u\| = 1$ and $\|x - m\| \le t$ with probability at least $1 - 2e^{-s}$. Now we set $\kappa := \frac{1-\delta^{1/2}(k,t,\tau,\Delta;s)}{2}$ in (2.4) 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 5 and 6 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 (19) in place of (18). □

3. Implications for the median of means estimator.

In this section we prove Theorem 1. To this end, we will apply Theorems 4 and 5 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) 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 (12) 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 2, Lemma 3 or Lemma 4 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 (2009)), 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 (16) implies that

$$\Delta^{(n)} \le \frac{KC_1(q)}{r(\Sigma_Y)} \tag{23}$$

whenever $P_Y \in \mathcal{P}_j,\ j \in \{1,2,3\}$. Therefore, conditions of Theorem 5 hold for k large enough, and we deduce that

$$\begin{aligned}\|\widehat{\mu}_N - \mu\| &\le \|m_n - \mu\| + \|\widehat{\mu}_N - m_n\| \\ &\le 2\sqrt{\frac{\|\Sigma_Y\|k}{N}}\,\frac{\mathbb{E}^{1/2}\|\sqrt{n}(\bar{Y}_1 - m_n)\|^{-2}}{\mathbb{E}\|\sqrt{n}(\bar{Y}_1 - m_n)\|^{-1}} + K(\Delta^{(n)})\left(\sqrt{\frac{\operatorname{tr}(\Sigma_Y)}{N}} + \sqrt{\Delta^{(n)}\operatorname{tr}(\Sigma_Y)}\sqrt{\frac{\sqrt{k}}{N}}\right)\end{aligned} \tag{24}$$

with probability at least $1 - 4e^{-\sqrt{k}}$. The final form of the bound follows once we apply the inequality (23) and estimate the ratio of moments via one of the lemmas in section 2.3. For instance, if $P_Y \in \mathcal{P}_1(M_0, K)$, then Lemma 2 combined with (24) 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.

References

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

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

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

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

Boucheron, S., Lugosi, G. and Massart, P. (2013) *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Bouhata, D. and Moumen, H. (2022) Byzantine fault tolerance in distributed machine learning: a survey. *arXiv preprint arXiv:2205.02572*.

Cardot, H., Cénac, P. and Godichon-Baggioni, A. (2017) Online estimation of the geometric median in Hilbert spaces: Nonasymptotic confidence balls.

Chaudhuri, P. (1996) On a geometric notion of quantiles for multivariate data. *Journal of the American statistical association*, **91**, 862–872.

Chen, Y., Su, L. and Xu, J. (2017) Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, **1**, 1–25.

Cherapanamjeri, Y., Flammarion, N. and Bartlett, P. L. (2019) Fast mean estimation with sub-Gaussian rates. In *Conference on Learning Theory*, 786–806. PMLR.

Cohen, M. B., Lee, Y. T., Miller, G., Pachocki, J. and Sidford, A. (2016) Geometric median in nearly linear time. *arXiv preprint arXiv:1606.05225*.

Depersin, J. and Lecué, G. (2022) Robust sub-Gaussian estimation of a mean vector in nearly linear time. *The Annals of Statistics*, **50**, 511–536.

Diakonikolas, I. and Kane, D. (2021) Recent advances in algorithmic high-dimensional robust statistics. In *"Beyond the worst-case analysis of algorithms"*. Cambridge University Press.

Dirksen, S. (2015) Tail bounds via generic chaining. *Electron. J. Probab*, **20**, 1–29.

Giné, E. and Nickl, R. (2015) *Mathematical foundations of infinite-dimensional statistical models*, vol. 40. Cambridge University Press.

Gini, C. and Galvani, L. (1929) Di talune estensioni dei concetti di media ai caratteri qualitativi. *Metron*, **8**, 3–209.

Gluskin, E. and Milman, V. (2003) Note on the geometric-arithmetic mean inequality. In *Geometric Aspects of Functional Analysis: Israel Seminar 2001-2002*, 131–135. Springer.

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

Hopkins, S. B. (2020) Mean estimation with sub-Gaussian rates in polynomial time.

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

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

Juškevičius, T. and Lee, J. (2015) Small ball probabilities, maximum density and rearrangements. *arXiv preprint arXiv:1503.09190*.

Kemperman, J. (1987) The median of a finite measure on a Banach space. *Statistical data analysis based on the L_1-norm and related methods*, 217–230.

Koltchinskii, V. (1994) Spatial quantiles and their Bahadur-Kiefer representations. In *Asymptotic Statistics: Proceedings of the Fifth Prague Symposium, held from September 4–9, 1993*, 361–367. Springer.

Koltchinskii, V. I. (1997) M-estimation, convexity and quantiles. *Ann. Statist.*, **25**, 435–477.

Konen, D. (2022) Recovering a probability measure from its multivariate spatial rank. *arXiv preprint arXiv:2208.11551*.

Latala, R. and Oleszkiewicz, K. (2005) Small ball probability estimates in terms of width. *Studia mathematica*, **169**, 305–314.

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

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

— (2019) Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, **19**, 1145–1190.

— (2020) Multivariate mean estimation with direction-dependent accuracy. *arXiv preprint arXiv:2010.11921*.

Madiman, M., Melbourne, J. and Xu, P. (2017) Rogozin's convolution inequality for locally compact groups. *arXiv preprint arXiv:1705.00642*.

Mathieu, T. (2022) Concentration study of M-estimators using the influence function. *Electronic Journal of Statistics*, **16**, 3695–3750.

Mendelson, S. (2015) Learning without concentration. *Journal of the ACM (JACM)*, **62**, 1–25.

Mendelson, S. and Zhivotovskiy, N. (2020) Robust covariance estimation under L_4-L_2 norm equivalence.

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

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

Oliveira, R. I. (2016) The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, **166**, 1175–1194.

Pillutla, K., Kakade, S. M. and Harchaoui, Z. (2022) Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, **70**, 1142–1154.

Rio, E. (2009) Moment inequalities for sums of dependent random variables under projective conditions. *Journal of Theoretical Probability*, **22**, 146–163.

Romon, G. (2022) Statistical properties of approximate geometric quantiles in infinite-dimensional Banach spaces. *arXiv preprint arXiv:2211.00035*.

Rudelson, M. and Vershynin, R. (2015) Small ball probabilities for linear images of high-dimensional distributions. *International Mathematics Research Notices*, **2015**, 9594–9617.

Weber, A. (1929) Uber den standort der industrien (Alfred Weber's theory of the location of industries). *University of Chicago*.

Zhivotovskiy, N. (2021) Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *arXiv preprint arXiv:2108.08198*.