Sub-Gaussian estimators of the mean of a random vector *

Gábor Lugosi[†‡§] Shahar Mendelson [¶]

February 1, 2017

Abstract

We study the problem of estimating the mean of a random vector X given a sample of N independent, identically distributed points. We introduce a new estimator that achieves a purely sub-Gaussian performance under the only condition that the second moment of X exists. The estimator is based on a novel concept of a multivariate median.

2010 Mathematics Subject Classification: 62J02, 62G08, 60G25.

1 Introduction

In this paper we study the problem of estimating the mean of a random vector X taking values in $\mathbb{R}^d$. Denoting the mean by $\mu = \mathbb{E}X$, we assume throughout the paper that the covariance matrix $\Sigma = \mathbb{E}(X-\mu)(X-\mu)^T$ exists. Suppose that N independent, identically distributed samples $X_1,\ldots,X_N$ drawn from the distribution of X are available, and one wishes to estimate the mean vector μ. An estimator is simply a function of the data that we denote by $\widehat{\mu}_N = \widehat{\mu}_N(X_1,\ldots,X_N)$.

There are many possible ways of measuring the quality of an estimator. The classical statistical literature tended to focus on risk measures such as the mean

*Gábor Lugosi was supported by the Spanish Ministry of Economy and Competitiveness, Grant MTM2015-67304-P and FEDER, EU. Shahar Mendelson was supported in part by the Israel Science Foundation.

†Department of Economics and Business, Pompeu Fabra University, Barcelona, Spain, gabor.lugosi@upf.edu

‡ICREA, Pg. Llus Companys 23, 08010 Barcelona, Spain

§Barcelona Graduate School of Economics

¶Department of Mathematics, Technion, I.I.T, and Mathematical Sciences Institute, The Australian National University, shahar@tx.technion.ac.il

squared error $\mathbb{E}\|\widehat{\mu}_N - \mu\|^2$. (Here, and in the rest of the paper, $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$, $S^{d-1} = \{v \in \mathbb{R}^d : \|v\| = 1\}$ denotes the Euclidean sphere in $\mathbb{R}^d$ and $\langle \cdot, \cdot \rangle$ is the usual inner product in $\mathbb{R}^d$.) In this case the sample mean $\overline{\mu}_N = (1/N)\sum_{i=1}^N X_i$ has a mean squared error equal to $\mathrm{Tr}(\Sigma)/N$ (where $\mathrm{Tr}(\Sigma)$ denotes the trace of the covariance matrix) and, even though this estimator is not necessarily optimal even for standard normal vectors—by "Stein's paradox", see [10]—, the order of magnitude of the error cannot be improved in general.

The situation is quite different when one is interested in minimizing the value r that satisfies

$$\mathbb{P}\left\{\left\|\widehat{\mu}_N - \mu\right\| > r\right\} \le \delta$$

for some given $\delta > 0$. While one may always take $r = \sqrt{\mathrm{Tr}(\Sigma)/(N\delta)}$ for the sample mean, much better dependence on δ may be achieved if the distribution is sufficiently light tailed. For example, if X has a multivariate normal distribution with mean μ and covariance matrix Σ, then the sample mean $\overline{\mu}_N$ is also multivariate normal with mean μ and covariance matrix $(1/N)\Sigma$ and therefore, for $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left\|\overline{\mu}_N - \mu\right\| \le \sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{2\lambda_{\max}\log(1/\delta)}{N}}\,, \tag{1.1}$$

where $\lambda_{\max}$ denotes the largest eigenvalue of Σ (see Hanson and Wright [7]). Similar bounds may be proven for the performance of the sample mean if X has a sub-Gaussian distribution in the sense that for all unit vectors $v \in S^{d-1}$,

$$\mathbb{E}\exp(\lambda\langle v, X - \mathbb{E}X\rangle) \le \exp(c\lambda^2\langle v, \Sigma v\rangle)$$

for some constant c.

However, when the distribution is not necessarily sub-Gaussian and is possibly heavy-tailed, one cannot expect such a sub-Gaussian behavior of the sample mean. Thus, when is it not reasonable to assume a sub-Gaussian distribution and heavy tails may be a concern, the sample mean is a risky choice. Indeed, alternative estimators have been constructed to achieve better performance.

The one-dimensional case (i.e., $d = 1$) is quite well understood, see Catoni [4] and Devroye, Lerasle, Lugosi, and Oliveira [6] for recent accounts. The so-called *median-of-means* estimator is a simple and powerful univariate estimator with essentially optimal performance. This estimate was introduced independently in various papers, see Nemirovsky and Yudin [17], Jerrum, Valiant, and Vazirani [11], Alon, Matias, and Szegedy [1]. The median-of-means estimator partitions the data into $k < N$ blocks of size $m \approx N/k$ each, computes the sample mean within each block, and outputs their median. One may easily show (see, e.g., Hsu

[8]) that, for any $\delta \in (0,1)$ if $k = \lceil 8\log(1/\delta) \rceil$, then the resulting estimator $\widehat{\mu}_N^{(\delta)}$ satisfies that, with probability at least $1-\delta$,

$$\left| \widehat{\mu}_N^{(\delta)} - \mu \right| \leq 8\sigma \sqrt{\frac{\log(2/\delta)}{N}} \tag{1.2}$$

where σ^2 denotes the variance of X. In other words, in the one-dimensional case, the median-of-means estimator achieves a sub-Gaussian performance under the only condition that the variance of X exists.

The median-of-means estimator has been extended to the multivariate case by replacing the median by its natural multivariate extension, the so-called "geometric (or spatial) median" (i.e., the point that minimizes the sum of the Euclidean distances to the sample means within each block) see Lerasle and Oliveira [14], Hsu and Sabato [9], Minsker [16]. In particular, Minsker proves that for each $\delta \in (0,1)$ this generalization of the median-of-means estimator $\widehat{\mu}_N^{(\delta)}$ is such that, with probability at least $1-\delta$,

$$\left\| \widehat{\mu}_N^{(\delta)} - \mu \right\| \leq C \sqrt{\frac{\mathrm{Tr}(\Sigma)\log(1/\delta)}{N}} , \tag{1.3}$$

where C is a universal constant. This bound holds under the only assumption that the covariance matrix exists. However, it does not quite achieve a sub-Gaussian performance bound that resembles (1.1).

Joly, Lugosi, and Oliveira [12] made an attempt to construct a mean estimator with a sub-Gaussian behavior for a large class of distributions. They prove that there exists a mean estimator $\widehat{\mu}_n^{(\delta)}$ such that, if the distribution satisfies that for all $v \in S^{d-1}$

$$\mathbb{E}\left[\langle (X-\mu), v \rangle^4 \right] \leq K(\langle v, \Sigma v \rangle)^2 ,$$

for some constant K, then for all $N \geq CK \log d \, (d + \log(1/\delta))$, with probability at least $1-\delta$,

$$\left\| \widehat{\mu}_N^{(\delta)} - \mu \right\| \leq C \left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{\lambda_{\max} \log(\delta^{-1} \log d)}{N}} \right) , \tag{1.4}$$

where again C is a universal constant. This bound resembles the sub-Gaussian inequality (1.1). However, there are various caveats: the additional fourth-moment assumption, the requirement that $N = \Omega(d \log d)$, and, to a lesser extent, the extra $\log\log d$ term in the bound seem sub-optimal.

The main result of this paper is that there exists a mean estimator that achieves purely sub-Gaussian performance under the minimal condition that the covariance matrix exists. More precisely, we prove the existence of a mean estimator $\widehat{\mu}_N^{(\delta)}$ such that, for all distributions with a finite second moment, for all N, with probability at least $1-\delta$,

$$\left\|\widehat{\mu}_N^{(\delta)} - \mu\right\| \le C\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{\lambda_{\max}\log(2/\delta)}{N}}\right),$$

for an explicit universal constant C.

The proposed estimator may be interpreted as a multivariate median-of-means estimate but with a new notion of a multivariate median which may be interesting in its own right. The construction of the new estimator is inspired by the technique of "median-of-means tournament", put forward by the authors in [15].

In the next section we present the proposed estimator and the performance bound. In Section 3 we present the proofs. We finish the paper by remarks about the computation of the estimator.

2 The estimator

Here we introduce the proposed mean estimator. Recall that we are given an i.i.d. sample $X_1,\ldots,X_N$ of random vectors in $\mathbb{R}^d$. As in the case of the median-of-means estimator, we start by partitioning the set $\{1,\ldots,N\}$ into k blocks $B_1,\ldots,B_k$, each of size $|B_j| \ge m \stackrel{\text{def.}}{=} \lfloor N/k \rfloor$, where k is a parameter of the estimator whose value depends on the desired confidence level, as specified below. In order to simplify the presentation, in the rest of the paper, without loss of generality, we assume that N is divisible by k and therefore $|B_j| = m$ for all $j = 1,\ldots,k$.

Define the sample mean within each block by

$$Z_j = \frac{1}{m}\sum_{i\in B_j} X_i .$$

For each $a \in \mathbb{R}^d$, let

$$S_a = \left\{ x \in \mathbb{R}^d : \exists J \subset [k] : |J| > k/2 \text{ such that } \min_{j\in J}\Big(\|Z_j - x\| - \|Z_j - a\|\Big) > 0 \right\} \tag{2.1}$$

and define the mean estimator by

$$\widehat{\mu}_N \in \operatorname*{argmin}_{a\in\mathbb{R}^d} \operatorname{diam}(S_a^c) .$$

Thus, $\widehat{\mu}_N$ is chosen to minimize, over all $a \in \mathbb{R}^d$, the diameter of the complement of set S_a defined as the set of points $x \in \mathbb{R}^d$ for which $\|Z_j - x\| > \|Z_j - a\|$ for the majority of the blocks, and if there are several minimizers, one may pick any one of them.

Note that the minimum is always achieved. This follows from the fact that $\mathrm{diam}(S_a^c)$ is a continuous function of a (since, for each a, S_a^c is the intersection of a finite union of closed balls, and the centers and radii of the closed balls are continuous in a).

One may interpret $\mathrm{argmin}_{a\in\mathbb{R}^d}\,\mathrm{diam}(S_a^c)$ as a new multivariate notion of the median of $Z_1,\dots,Z_k$. Indeed, when $d=1$, it is a particular choice of the median and the proposed estimator coincides with the median-of-means estimator.

The main result of this paper is the following performance bound:

Theorem 1. *Let $\delta\in(0,1)$ and consider the mean estimator $\widehat{\mu}_N$ with parameter $k=\lceil 360\log(2/\delta)\rceil$. If $X_1,\dots,X_N$ are i.i.d. random vectors in $\mathbb{R}^d$ with mean $\mu\in\mathbb{R}^d$ and covariance matrix Σ, then for all N, with probability at least $1-\delta$,*

$$\left\|\widehat{\mu}_N-\mu\right\|\le 2\max\left(400\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}},240\sqrt{\frac{\lambda_{max}\log(2/\delta)}{N}}\right).$$

Thus, the proposed estimator achieves a purely sub-Gaussian performance under minimal conditions. Just like in the case of the median-of-means estimator for the univariate case, the estimator depends on the desired level of confidence δ. As it is shown in [6], such a dependence cannot be avoided without imposing additional conditions on the distribution. However, following the route laid down in [6], one may construct sub-Gaussian estimators that work for a wide range of confidence levels simultaneously under more assumptions on the distribution. Since this issue is beyond the scope of this paper and will not be pursued further here.

Just like Minsker's bound (1.3)—but unlike the bound (1.4)—, the performance bound of Theorem 1 is "infinite-dimensional" in the sense that the bound does not depend on the dimension d explicitly. Indeed, the same estimator may be defined for Hilbert-space valued random vectors and Theorem 1 remains valid as long as $\mathrm{Tr}(\Sigma)=\mathbb{E}\|X-\mu\|^2$ is finite.

Theorem 1 is an outcome of the following observation which is of interest in its own right on the geometry of a typical collection $\{X_1,\dots,X_N\}$.

Theorem 2. *Using the same notation as above and setting*

$$r=\max\left(400\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}},240\sqrt{\frac{\lambda_{max}\log(2/\delta)}{N}}\right),$$

with probability at least $1-\delta$, for any $a\in\mathbb{R}^d$ such that $\|a-\mu\|\ge r$, one has $\|Z_j-a\|>\|Z_j-\mu\|$ for more than $k/2$ indices j.

Theorem 2 implies that for a 'typical' collection $X_1,...,X_N$, μ is closer to a majority of the Z_j's when compared to any $a \in \mathbb{R}^d$ that is sufficiently far from μ. Obviously, for an arbitrary collection $x_1,...,x_N \subset \mathbb{R}^d$ such a point need not exist, and it is rather surprising that for a typical i.i.d. configuration, this property is satisfied by μ.

The fact that Theorem 2 implies Theorem 1 is straightforward. Indeed, Theorem 2 implies that $\mathrm{diam}(S_\mu^c) \le 2r$ and that if $\|a-\mu\| \ge r$, then $\mu \in S_a^c$. By the definition of S_a, one always has $a \in S_a^c$, and thus if $\|a-\mu\| > 2r$ then $\mathrm{diam}(S_a^c) > 2r$. Therefore, the minimizer $\widehat{\mu}$ must satisfy that $\|\widehat{\mu}-\mu\| \le 2r$, as required.

We do not claim that the values of the constants appearing in Theorem 1 are optimal. They were obtained with the goal of making the proof transparent, nothing more, and it is likely that they may be improved by more careful calculations.

The proof of Theorem 2 is based on the idea of "median-of-means tournaments" which was introduced by Lugosi and Mendelson [15] is the context of regression function estimation.

3 Proof

The proof of Theorem 2 is based on the following idea. The mean μ is the minimizer of the function $f(x) = \mathbb{E}\|X-\mu\|^2$. A possible approach is to use the available data to guess, for any pair $a, b \in \mathbb{R}^d$, whether $f(a) < f(b)$. To this end, we may set up a "tournament" as follows.

Recall that $[N]$ is partitioned into k disjoint blocks $B_1, \ldots, B_k$ of size $m = N/k$. For $a, b \in \mathbb{R}^d$, we say that *a defeats b* if

$$\frac{1}{m}\sum_{i \in B_j}\left(\|X_i-b\|^2 - \|X_i-a\|^2\right) > 0$$

on more than $k/2$ blocks B_j. The main technical lemma is the following.

Lemma 1. *Let $\delta \in (0,1)$, $k = \lceil 360\log(2/\delta)\rceil$, and define*

$$r = \max\left(400\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}}, 240\sqrt{\frac{\lambda_{max}\log(2/\delta)}{N}}\right).$$

With probability at least $1-\delta$, μ defeats all $b \in \mathbb{R}^d$ such that $\|b-\mu\| \ge r$.

Proof. Note that

$$\|X_i-b\|^2 - \|X_i-\mu\|^2 = \|X_i-\mu+\mu-b\|^2 - \|X_i-\mu\|^2 = -2\langle X_i-\mu, b-\mu\rangle + \|b-\mu\|^2,$$

set $\overline{X} = X - \mu$ and put $v = b - \mu$. Thus, for a fixed b that satisfies $\|b-\mu\| \ge r$, μ defeats b if

$$-\frac{2}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, v\right\rangle + \|v\|^2 > 0$$

on the majority of blocks B_j.

Therefore, to prove our claim we need that, with probability at least $1-\delta$, for every $v \in \mathbb{R}^d$ with $\|v\| \ge r$,

$$-\frac{2}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, v\right\rangle + \|v\|^2 > 0 \tag{3.1}$$

for more than $k/2$ blocks B_j. Clearly, it suffices to show that (3.1) holds when $\|v\| = r$.

Consider a fixed $v \in \mathbb{R}^d$ with $\|v\| = r$. By Chebyshev's inequality, with probability at least 9/10,

$$\left|\frac{1}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, v\right\rangle\right| \le \sqrt{10}\sqrt{\frac{\mathbb{E}\left\langle \overline{X}, v\right\rangle^2}{m}} \le \sqrt{10}\|v\|\sqrt{\frac{\lambda_{\max}}{m}}\ ,$$

where recall that $\lambda_{\max}$ is the largest eigenvalue of the covariance matrix of X. Thus, if

$$r = \|v\| \ge 4\sqrt{10}\sqrt{\frac{\lambda_{\max}}{m}} \tag{3.2}$$

then with probability at least 9/10,

$$-\frac{2}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, v\right\rangle \ge \frac{-r^2}{2}. \tag{3.3}$$

Applying a standard binomial tail estimate, we see that (3.3) holds for a single v with probability at least $1-\exp(-k/180)$ on at least 8/10 of the blocks B_j.

Now we need to extend the above from a fixed vector v to all vectors with norm r. In order to show that (3.3) holds simultaneously for all $v \in r\cdot S^{d-1}$ on at least 7/10 of the blocks B_j, we first consider a maximal ε-separated set $V_1 \subset r\cdot S^{d-1}$ with respect to the $L_2(X)$ norm. In other words, V_1 is a subset of $r\cdot S^{d-1}$ of maximal cardinality such that for all $v_1, v_2 \in V_1$, $\|v_1 - v_2\|_{L_2(X)} = \langle v_1 - v_2, \Sigma(v_1-v_2)\rangle^{1/2} \ge \varepsilon$. We may estimate this cardinality by the "dual Sudakov" inequality (see [13] and also [18] for a version with the specified constant), which implies that the cardinality of V_1 is bounded by

$$\log|V_1| \le \left(\frac{\mathbb{E}\left[\langle G, \Sigma G\rangle^{1/2}\right]}{4\varepsilon/r}\right)^2\ ,$$

where G is a standard normal vector in $\mathbb{R}^d$. Notice that for any $a \in \mathbb{R}^d$, $\mathbb{E}_X \langle a, X\rangle^2 = \langle a, \Sigma a\rangle$, and therefore,

$$\begin{aligned}\mathbb{E}\left[\langle G, \Sigma G\rangle^{1/2}\right] &= \mathbb{E}_G\left[\left(\mathbb{E}_X\left[\left\langle G, \overline{X}\right\rangle^2\right]\right)^{1/2}\right] \le \left(\mathbb{E}_X \mathbb{E}_G\left[\left\langle G, \overline{X}\right\rangle^2\right]\right)^{1/2} \\ &= \left(\mathbb{E}\left[\left\|\overline{X}\right\|^2\right]\right)^{1/2} = \sqrt{\mathrm{Tr}(\Sigma)}\,.\end{aligned}$$

Hence, by setting

$$\varepsilon = 5r\left(\frac{1}{k}\mathrm{Tr}(\Sigma)\right)^{1/2}\,, \tag{3.4}$$

we have $|V_1| \le e^{k/360}$ and thus, by the union bound, with probability at least $1 - e^{-k/360} \ge 1 - \delta/2$, (3.3) holds for all $v \in V_1$ on at least 8/10 of the blocks B_j.

Next we check that property (3.1) holds simultaneously for all x with $\|x\| = r$ on at least 7/10 of the blocks B_j.

For every $x \in r \cdot S^{d-1}$, let v_x be the nearest element to x in V_1 with respect to the $L_2(X)$ norm. It suffices to show that, with probability at least $1 - \exp(-k/200) \ge 1 - \delta/2$,

$$\sup_{x \in r\cdot S^{d-1}} \frac{1}{k}\sum_{j=1}^k \mathbb{1}_{\{|m^{-1}\sum_{i\in B_j}\langle \overline{X}_i, x - v_x\rangle| \ge r^2/4\}} \le \frac{1}{10}\,. \tag{3.5}$$

Indeed, on that event it follows that for every $x \in r \cdot S^{d-1}$, on at least 7/10 of the coordinate blocks B_j, both

$$-\frac{2}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, v_x\right\rangle \ge \frac{-r^2}{2} \quad \text{and} \quad 2\left|\frac{1}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, x\right\rangle - \frac{1}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, v_x\right\rangle\right| < \frac{r^2}{2}$$

hold and hence, on those blocks, $-\frac{2}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, x\right\rangle + r^2 > 0$ as required.

It remains to prove (3.5). Observe that

$$\frac{1}{k}\sum_{j=1}^k \mathbb{1}_{\{|m^{-1}\sum_{i\in B_j}\langle \overline{X}_i, x - v_x\rangle| \ge r^2/4\}} \le \frac{4}{r^2}\frac{1}{k}\sum_{j=1}^k \left|\frac{1}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, x - v_x\right\rangle\right|\,.$$

Since $\|x - v_x\|_{L_2(X)} = (\mathbb{E}\langle X, x - v_x\rangle^2)^{1/2} \le \varepsilon$ it follows that for every j

$$\mathbb{E}\left|\frac{1}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, x - v_x\right\rangle\right| \le \sqrt{\frac{\mathbb{E}\left[\left\langle \overline{X}, x - v_x\right\rangle^2\right]}{m}} \le \frac{\varepsilon}{\sqrt{m}}\,,$$

and therefore,

$$
\begin{aligned}
&\mathbb{E} \sup_{x \in r \cdot S^{d-1}} \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}_{\{|m^{-1} \sum_{i \in B_j} \langle \overline{X}_i, x - v_x \rangle| \geq r^2/4\}} \\
&\leq \quad \frac{4}{r^2} \mathbb{E} \sup_{x \in r \cdot S^{d-1}} \frac{1}{k} \sum_{j=1}^{k} \left(\left| \frac{1}{m} \sum_{i \in B_j} \left\langle \overline{X}_i, x - v_x \right\rangle \right| - \mathbb{E} \left| \frac{1}{m} \sum_{i \in B_j} \left\langle \overline{X}_i, x - v_x \right\rangle \right| \right) + \frac{4\varepsilon}{r^2 \sqrt{m}} \\
&\stackrel{\text{def.}}{=} \quad (A) + (B) .
\end{aligned}
$$

To bound (B), note that, by (3.4),

$$
\frac{4\varepsilon}{r^2 \sqrt{m}} = 20 \left(\frac{\mathrm{Tr}(\Sigma)}{N} \right)^{1/2} \cdot \frac{1}{r} \leq \frac{1}{20}
$$

provided that

$$
r \geq 400 \left(\frac{\mathrm{Tr}(\Sigma)}{N} \right)^{1/2} .
$$

Turning to (A), by symmetrization, contraction for Bernoulli processes and desymmetrization (see, e.g., [13]), and noting that $\|x - v_x\| \leq 2r$, we have

$$
\begin{aligned}
(A) &\leq \frac{8}{r^2} \mathbb{E} \sup_{x \in r \cdot S^{d-1}} \left| \frac{1}{N} \sum_{i=1}^{N} \left\langle \overline{X}_i, x - v_x \right\rangle \right| \leq \frac{16}{r} \mathbb{E} \sup_{\{t : \|t\| \leq 1\}} \left| \frac{1}{N} \sum_{i=1}^{N} \left\langle \overline{X}_i, t \right\rangle \right| \\
&\leq \frac{16}{r} \cdot \frac{\mathbb{E} \left\| \overline{X} \right\|}{\sqrt{N}} = \frac{16}{r} \left(\frac{\mathrm{Tr}(\Sigma)}{N} \right)^{1/2} \leq \frac{1}{20}
\end{aligned}
$$

provided that $r \geq 320 \left(\frac{\mathrm{Tr}(\Sigma)}{N} \right)^{1/2}$.

Thus, for

$$
Y = \sup_{x \in r \cdot S^{d-1}} \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}_{\{|m^{-1} \sum_{i \in B_j} \langle \overline{X}_i, x - v_x \rangle| \geq r^2/4\}} ,
$$

we have proved that $\mathbb{E}Y \leq 1/20$. Finally, in order to prove (3.5), it suffices to prove that, $\mathbb{P}\{Y > \mathbb{E}Y + 1/20\} \leq e^{-k/200}$, which follows from the bounded differences inequality (see, e.g., [3, Theorem 6.2]). ■

Proof of Theorem 2

Theorem 2 is easily derived from Lemma 1. Fix a block B_j, and recall that $Z_j = \frac{1}{m}\sum_{i\in B_j} X_i$. Let $a, b \in \mathbb{R}^d$. Then

$$\begin{aligned}
\frac{1}{m}\sum_{i\in B_j}\left(\|X_i - a\|^2 - \|X_i - b\|^2\right) &= \frac{1}{m}\sum_{i\in B_j}\left(\|X_i - b - (a-b)\|^2 - \|X_i - b\|^2\right) \\
&= -\frac{2}{m}\sum_{i\in B_j}\langle X_i - b, a-b\rangle + \|a-b\|^2 = (*)
\end{aligned}$$

Observe that $-\frac{2}{m}\sum_{i\in B_j}\langle X_i - b, a-b\rangle = -2\left\langle \frac{1}{m}\sum_{i\in B_j} X_i - b, a-b\right\rangle = -2\left\langle Z_j - b, a-b\right\rangle$, and thus

$$\begin{aligned}
(*) &= -2\left\langle Z_j - b, a-b\right\rangle + \|a-b\|^2 \\
&= -2\left\langle Z_j - b, a-b\right\rangle + \|a-b\|^2 + \|Z_j - b\|^2 - \|Z_j - b\|^2 \\
&= \|Z_j - b - (a-b)\|^2 - \|Z_j - b\|^2 = \|Z_j - a\|^2 - \|Z_j - b\|^2 .
\end{aligned}$$

Therefore, $(*) > 0$ (i.e., b defeats a on block B_j) if and only if $\|Z_j - a\| > \|Z_j - b\|$.

Recall that Lemma 1 states that, with probability at least $1-\delta$, if $\|a-\mu\| \ge r$ then on more than $k/2$ blocks B_j, $\frac{1}{m}\sum_{i\in B_j}\left(\|X_i - a\|^2 - \|X_i - \mu\|^2\right) > 0$, which, by the above argument, is the same as saying that for at least $k/2$ indices j, $\|Z_j - a\| > \|Z_j - \mu\|$. ■

4 Computational considerations

The problem of computing various notions of multivariate medians has been thoroughly studied in computational geometry and we refer to Aloupis [2] for a survey on this topic. For example, computing the geometric median—and therefore the multivariate median-of-means estimator proposed by Hsu and Sabato [9] and Minsker [16]—involves solving a convex optimization problem. Thus, the geometric median may be approximated efficiently, see [5] for the most recent result and for the rich history of the problem.

In contrast, efficiently computing, or even approximating, the multivariate median proposed in this paper appears to be a nontrivial challenge.

A possible approach for computing a mean estimator that approximates $\widehat{\mu}_N$ is based on a variant of a coordinate descent algorithm that works roughly as follows: starting with an arbitrary line in $\mathbb{R}^d$, one may discretize, with mesh $O(r)$, the segment on the line that supports the convex hull of $Z_1, \ldots, Z_k$. Then one uses pairwise comparisons of the discretized values, using the median-of-means estimate, to find a point that defeats every other candidate on the line that is at least distance $2r$ apart from it. (With a minor adjustment of our arguments above one may prove that such a point always exists.) Then take a line that is orthogonal to the first line and contains the "winner" and repeat the search on that line. Continue for d steps. One may prove that the point $\widetilde{\mu}_N$ obtained at the final step is such that, with probability at least $1-\delta$, $\|\widetilde{\mu}_N - \mu\|_\infty \le Cr$ for a numerical constant C. This algorithm runs in time quadratic in $1/r$ and linear in d but unfortunately it only guarantees closeness to the true mean in the ℓ_∞ sense. If one replaces orthogonal lines by random ones and keeps repeating the procedure, one eventually achieves the desired guarantee in the Euclidean distance. However, one needs to consider exponentially many (in d) directions to approach μ with the desired precision. Note that such algorithms use r as an input parameter. Naturally, the value of r is not known but the algorithm is guaranteed to work well as long as the true value of r is larger that the prior guess.

Another possibility is to start with computing the geometric median $\widetilde{\mu}^{(\delta)}$ of the Z_j. By (1.3), one may now restrict search to a ball of radius at most $r\sqrt{\log(1/\delta)}$. By exhaustively searching through this ball (after appropriately discretizing), one finds an estimate with the desired properties in additional time of order $\log^d(1/\delta)$. However, this is surely unrealistic in most interesting cases.

We leave the question of efficiently computing the proposed mean estimate (or another one with sub-Gaussian performance guarantees) as an interesting research problem.

References

[1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58:137–147, 2002.

[2] G. Aloupis. Geometric measures of data depth. *DIMACS series in discrete mathematics and theoretical computer science*, 72:147–158, 2006.

[3] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities:A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[4] O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.

[5] 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*, pages 9–21. ACM, 2016.

[6] L. Devroye, M. Lerasle, G. Lugosi, and R.I. Oliveira. Sub-Gausssian mean estimators. *Annals of Statistics*, 2016.

[7] D.L. Hanson and F.T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Annals of Mathematical Statistics*, 42:1079–1083, 1971.

[8] D. Hsu. Robust statistics. `http://www.inherentuncertainty.org/2010/12/robust-statistics.html`, 2010.

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

[10] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

[11] M. Jerrum, L. Valiant, and V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:186–188, 1986.

[12] E. Joly, G. Lugosi, and R. I. Oliveira. On the estimation of the mean of a random vector. *preprint*, 2016.

[13] M. Ledoux and M. Talagrand. *Probability in Banach Space*. Springer-Verlag, New York, 1991.

[14] M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. arXiv:1112.3914, 2012.

[15] G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *preprint*, 2016.

[16] S. Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21:23082335, 2015.

[17] A.S. Nemirovsky and D.B. Yudin. Problem complexity and method efficiency in optimization. 1983.

[18] R. Vershynin. Lectures in geometric functional analysis. *University of Michigan*, 2009.