arXiv:1906.04280v1 [math.ST] 10 Jun 2019

Mean estimation and regression under heavy-tailed distributions—a survey *

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

June 12, 2019

Abstract

We survey some of the recent advances in mean estimation and regression function estimation. In particular, we describe sub-Gaussian mean estimators for possibly heavy-tailed data both in the univariate and multivariate settings. We focus on estimators based on median-of-means techniques but other methods such as the trimmed mean and Catoni's estimator are also reviewed. We give detailed proofs for the cornerstone results. We dedicate a section on statistical learning problems–in particular, regression function estimation–in the presence of possibly heavy-tailed data.

AMS Mathematics Subject Classification: 62G05, 62G15, 62G35

Key words: mean estimation, heavy-tailed distributions, robustness, regression function estimation, statistical learning.

*Gábor Lugosi was supported by the Spanish Ministry of Economy and Competitiveness, Grant MTM2015-67304-P and FEDER, EU, by "High-dimensional problems in structured probabilistic models - Ayudas Fundación BBVA a Equipos de Investigación Cientifica 2017" and by "Google Focused Award Algorithms and Learning for AI". 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

[¶]Mathematical Sciences Institute, The Australian National University and LPSM, Sorbonne University, shahar.mendelson@anu.edu.au

Contents

1 Introduction

Arguably the most fundamental problem of statistics is that of estimating the expected value μ of a random variable X based on a sample of n independent, identically distributed draws from the distribution of X. The obvious choice of an estimator is, of course, the empirical mean. Its properties are well understood by classical results of probability theory. However, from the early days on, statisticians have been concerned about the quality of the empirical mean, especially when the distribution may be *heavy-tailed* or *outliers* may be present in the data. This concern gave rise to the area of *robust statistics* that addresses the problem of mean estimation (and other statistical problems) for such data. Classical references include Huber [38], Huber and Ronchetti [39], Hampel, Ronchetti, Rousseeuw, and

Stahel [30], Tukey [77].

Motivated by applications in machine learning and data science, in recent years there has been increased interest in constructing mean and regression function estimates with the requirement that the estimators should achieve *high accuracy* with a large *confidence*. The best achievable accuracy/confidence tradeoff is much better understood today and the aim of this paper is to survey some of the recent advances. We primarily focus on the mean estimation problem, both in the univariate and multivariate settings. We offer detailed discussion of what the best performance one may expect is, describe a variety of estimators, and analyze their performance. We pay special attention to a simple but powerful methodology based on *median-of-means* techniques.

We also address one of the basic problems of statistical learning theory, namely regression function estimation. We show how the technology introduced for mean estimation may be used to construct powerful learning algorithms that achieve essentially optimal performance under mild assumptions.

The paper is organized as follows. In Section 2 we address the simplest, univariate mean estimation problem. We focus on *sub-Gaussian* estimators and explore their possibilities and limitations. Section 3 is dedicated to the significantly more challenging multivariate problem. We extend the notion of sub-Gaussian estimators to the multivariate setting and analyze various estimators. In Section 4 we study the problem of estimating the mean of an entire class of random variables with the requirement that all estimators have a high accuracy simultaneously over the entire class. We show how such estimators may be constructed and use these ideas in a general framework of mean estimation. Finally, Section 5 is dedicated to applying these techniques to regression function estimation.

2 Estimating the mean of a real random variable

In this section we examine the classical problem of estimating the mean of a random variable. Let $X_1, \dots, X_n$ be independent, identically distributed real random variables with mean $\mu = \mathbb{E}X_1$. Upon observing these random variables, one would like to estimate μ. An estimator $\widehat{\mu}_n = \widehat{\mu}_n(X_1, \dots, X_n)$ is simply a measurable function of $X_1, \dots, X_n$.

The quality of an estimator may be measured in various ways. While most of the early statistical work focused on expected risk measures such as the *mean-squared error*

$$\mathbb{E}\left[(\widehat{\mu}_n - \mu)^2\right],$$

such risk measures may be misleading. Indeed, if the difference $|\widehat{\mu}_n - \mu|$ is not sufficiently concentrated, the expected value does not necessarily reflect the "typical"

behavior of the error. For such reasons, we prefer estimators $\widehat{\mu}_n$ that are close to μ *with high probability*. Thus, our aim is to understand, for any given sample size n and confidence parameter $\delta \in (0,1)$, the smallest possible value $\epsilon = \epsilon(n,\delta)$ such that

$$\mathbb{P}\left\{\left|\widehat{\mu}_n - \mu\right| > \epsilon\right\} \leq \delta \,. \tag{2.1}$$

It it important to stress that (2.1) is a non-asymptotic criterion: one would like to obtain quantitative estimates on the way the accuracy ϵ scales with the confidence parameter δ and the sample size n. This type of estimate is reminiscent to the PAC (Probably Approximately Correct) framework usually adopted in statistical learning theory, see Valiant [79], Vapnik and Chervonenkis [82], Blumer, Ehrenfeucht, Haussler, and Warmuth [9].

The most natural choice of a mean estimator is the standard empirical mean

$$\overline{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \,.$$

The behavior of the empirical mean is well understood. For example, if the X_i have a finite second moment and σ^2 denotes their variance, then the mean-squared error of $\overline{\mu}_n$ equals σ^2/n. On the other hand, the central limit theorem guarantees that this estimator has Gaussian tails, asymptotically, when $n \to \infty$. Indeed,

$$\mathbb{P}\left\{\left|\overline{\mu}_n - \mu\right| > \frac{\sigma\Phi^{-1}(1-\delta/2)}{\sqrt{n}}\right\} \to \delta \,,$$

where $\Phi(x) = \mathbb{P}\{G \leq x\}$ is the cumulative distribution function of a standard normal random variable G. One may easily see (e.g., using the fact that for $t \geq 1$, $\exp(-t^2/2) \leq t\exp(-t^2/2)$), that for all $x \geq 0$,

$$1 - \Phi(x) \leq e^{-x^2/2} \,.$$

This implies that $\Phi^{-1}(1-\delta/2) \leq \sqrt{2\log(2/\delta)}$, and the central limit theorem asserts that

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|\overline{\mu}_n - \mu\right| > \frac{\sigma\sqrt{2\log(2/\delta)}}{\sqrt{n}}\right\} \leq \delta \,.$$

However, this is an asymptotic estimate and not the quantitative one we were hoping for. Still, our goal is to obtain non-asymptotic performance bounds of the same form. In particular, we say that a mean estimator $\widehat{\mu}_n$ is *L-sub-Gaussian* if there is a constant $L > 0$, such that for all sample sizes n and with probability at least $1-\delta$,

$$\left|\widehat{\mu}_n - \mu\right| \leq \frac{L\sigma\sqrt{\log(2/\delta)}}{\sqrt{n}} \,. \tag{2.2}$$

It is worth noting here the well-known fact that if all one knows is that the unknown distribution is Gaussian, then the sample mean is optimal for all sample sizes and confidence levels δ. (See Catoni [14, Proposition 6.1] for a precise statement.) Moreover, the following observation, established by Devroye, Lerasle, Lugosi, and Oliveira [22], shows that (2.2) is essentially the best that one can hope for in general, even if one is interested in a fixed confidence level:

Theorem 1. *Let $n > 5$ be a positive integer. Let $\mu \in \mathbb{R}$, $\sigma > 0$ and $\delta \in (2e^{-n/4}, 1/2)$. Then for any mean estimator $\widehat{\mu}_n$, there exists a distribution with mean μ and variance σ^2 such that*

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

Proof. To derive the "minimax" lower bound, it suffices to consider two distributions, P_+, P_-, both concentrated on two points, defined by

$$P_+(\{0\}) = P_-(\{0\}) = 1 - p\,, \qquad P_+(\{c\}) = P_-(\{-c\}) = p\,,$$

where $p \in [0,1]$ and $c > 0$. Note that the means of the two distributions are $\mu_{P_+} = pc$ and $\mu_{P_-} = -pc$ and both have variance $\sigma^2 = c^2 p(1-p)$.

For $i = 1, \ldots, n$, let (X_i, Y_i) be independent pairs of real-valued random variables such that

$$\mathbb{P}\{X_i = Y_i = 0\} = 1 - p \quad \text{and} \quad \mathbb{P}\{X_i = c, Y_i = -c\} = p\,.$$

Note that X_i is distributed as P_+ and Y_i is distributed as P_-. Let $\delta \in (0, 1/2)$. If $\delta \geq 2e^{-n/4}$ and $p = (1/(2n))\log(2/\delta)$, then (using $1 - p \geq \exp(-p/(1-p))$),

$$\mathbb{P}\{X_1^n = Y_1^n\} = (1-p)^n \geq 2\delta\,.$$

Let $\widehat{\mu}_n$ be any mean estimator, possibly depending on δ. Then

$$\begin{aligned}
&\max\left(\mathbb{P}\left\{\left|\widehat{\mu}_n(X_1^n) - \mu_{P_+}\right| > cp\right\}, \mathbb{P}\left\{\left|\widehat{\mu}_n(Y_1^n) - \mu_{P_-}\right| > cp\right\}\right) \\
&\quad \geq \frac{1}{2}\mathbb{P}\left\{\left|\widehat{\mu}_n(X_1, \ldots, X_n) - \mu_{P_+}\right| > cp \quad \text{or} \quad \left|\widehat{\mu}_n(Y_1, \ldots, Y_n) - \mu_{P_-}\right| > cp\right\} \\
&\quad \geq \frac{1}{2}\mathbb{P}\{\widehat{\mu}_n(X_1, \ldots, X_n) = \widehat{\mu}_n(Y_1, \ldots, Y_n)\} \\
&\quad \geq \frac{1}{2}\mathbb{P}\{X_1, \ldots, X_n = Y_1, \ldots, Y_n\} \geq \delta\,.
\end{aligned}$$

From $\sigma^2 = c^2 p(1-p)$ and $p \leq 1/2$ we have that $cp \geq \sigma\sqrt{p/2}$, and therefore

$$\max\left(\mathbb{P}\left\{\left|\widehat{\mu}_n(X_1, \ldots, X_n) - \mu_{P_+}\right| > \sigma\sqrt{\frac{\log\frac{2}{\delta}}{n}}\right\}, \mathbb{P}\left\{\left|\widehat{\mu}_n(Y_1, \ldots, Y_n) - \mu_{P_-}\right| > \sigma\sqrt{\frac{\log\frac{2}{\delta}}{n}}\right\}\right) \geq \delta\,.$$

Theorem 1 follows. ∎

With Theorem 1 in mind, our aim is to consider both univariate and multivariate situations and design estimators that perform with sub-Gaussian error rate. The meaning of sub-Gaussian error rate in the multivariate case is explained in Section 3.

Naturally, the first order of business is to check whether the obvious choice of a mean estimator—the empirical mean—is L-sub-Gaussian for some L. On the one hand, it is easy to see that under certain conditions on the distribution of the X_i, it does exhibit a sub-Gaussian performance. Indeed, if the X_i are such that there exists $L > 0$ such that for all $\lambda > 0$

$$\mathbb{E}e^{\lambda(X_i-\mu)} \leq e^{\sigma^2\lambda^2/L^2} ,$$

then the empirical mean $\widehat{\mu}_n$ is L-sub-Gaussian for all $\delta \in (0,1)$, as it is easily seen by the Chernoff bound.

On the other hand, assumptions of this type are quite restrictive and impose strong conditions on the decay of the tail probabilities of the X_i. Specifically, it is equivalent to the fact that for every $p \geq 2$, $(\mathbb{E}|X_i - \mu|^p)^{1/p} \leq L'\sqrt{p}\left(\mathbb{E}|X_i - \mu|^2\right)^{1/2}$, where $c_1 L \leq L' \leq c_2 L$ for suitable absolute constants c_1 and c_2 (see, e.g., [10]). When the X_i's do not exhibit such a tail decay, the empirical mean need not be sub-Gaussian.

For example, if one only assumes that σ exists (i.e., the variance of the X_i is finite) then the bound implied by Chebyshev's inequality, that is, that with probability at least $1 - \delta$,

$$\left|\overline{\mu}_n - \mu\right| \leq \sigma\sqrt{\frac{1}{n\delta}} , \tag{2.3}$$

is essentially the best that one can hope for. Although the bound from (2.3) decays with the sample size at the optimal rate of $O(n^{-1/2})$, the dependence on the confidence parameter δ is exponentially worse than in (2.2). We refer to Catoni [14, Proposition 6.2] for a precise formulation and a simple example that (almost) saturates Chebyshev's inequality.

This leads to an inevitable conclusion: if one is looking for a mean estimator that is sub-Gaussian for any random variable that has a well-defined mean and variance, then one must find alternatives to the sample mean. As it happens, and perhaps surprisingly, there exist mean estimators that achieve a sub-Gaussian performance for all distributions with a finite variance. Two quite different estimators are presented and analyzed in the next two sections.

2.1 The median-of-means estimator

The median-of-means estimator presented next has been proposed in different forms in various papers, see Nemirovsky and Yudin [69], Hsu [35], Jerrum, Valiant,

and Vazirani [40], Alon, Matias, and Szegedy [1].

The definition of the median-of-means estimator calls for partitioning the data into k groups of roughly equal size, computing the empirical mean in each group, and taking the median of the obtained values.

Formally, recall that the median of k real numbers $x_1,\dots,x_k \in \mathbb{R}$ is defined as $M(x_1,\dots,x_k) = x_i$ where x_i is such that

$$\Big|\{j \in [k] : x_j \le x_i\}\Big| \ge \frac{k}{2} \quad \text{and} \quad \Big|\{j \in [k] : x_j \ge x_i\}\Big| \ge \frac{k}{2}\,.$$

(If several indices i fit the above description, we take the smallest one.)

Now let $1 \le k \le n$ and partition $[n] = \{1,\dots,n\}$ into k blocks $B_1,\dots,B_k$, each of size $|B_i| \ge \lfloor n/k \rfloor \ge 2$.

Given $X_1,\dots,X_n$, compute the sample mean in each block

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

and define the median-of-means estimator by $\widehat{\mu}_n = M(Z_1,\dots,Z_k)$.

To grasp intuitively why this estimator works, note that for each block, the empirical mean is an unbiased estimator of the mean, with controlled standard deviation $\sigma/\sqrt{n/k}$. Hence, the median of the distribution of the blockwise empirical mean lies within $\sigma/\sqrt{n/k}$ from the expectation. Now the empirical median is a highly concetrated estimator of this median.

A performance-bound of the estimator is established next. For simplicity, assume that n is divisible by k so that each block has $m = n/k$ elements.

Theorem 2. *Let $X_1,\dots,X_n$ be independent, identically distributed random variables with mean μ and variance σ^2. Let m,k be positive integers assume that $n = mk$. Then the median-of-means estimator $\widehat{\mu}_n$ with k blocks satisfies*

$$\mathbb{P}\Big\{\Big|\widehat{\mu}_n - \mu\Big| > \sigma\sqrt{4/m}\Big\} \le e^{-k/8}\,.$$

In particular, for any $\delta \in (0,1)$, if $k = \lceil 8\log(1/\delta)\rceil$, then, with probability at least $1-\delta$,

$$\Big|\widehat{\mu}_n - \mu\Big| \le \sigma\sqrt{\frac{32\log(1/\delta)}{n}}\,.$$

Proof. By Chebyshev's inequality, for each $j = 1,\dots,k$, with probability at least $3/4$,

$$\Big|Z_j - \mu\Big| \le \sigma\sqrt{\frac{4}{m}}\,.$$

Thus, $\left|\widehat{\mu}_n - \mu\right| > \sigma\sqrt{4/m}$ implies that at least $k/2$ of the means Z_j are such that $\left|Z_j - \mu\right| > \sigma\sqrt{4/m}$. Hence,

$$\begin{aligned}
\mathbb{P}\left\{\left|\widehat{\mu}_n - \mu\right| > \sigma\sqrt{4/m}\right\} &\leq \mathbb{P}\left\{\mathrm{Bin}(k,1/4) \geq \frac{k}{2}\right\} \\
&\quad \text{(where } \mathrm{Bin}(k,1/4) \text{ is a binomial } (k,1/4) \text{ random variable)} \\
&= \mathbb{P}\left\{\mathrm{Bin}(k,1/4) - \mathbb{E}\mathrm{Bin}(k,1/4) \geq \frac{k}{4}\right\} \\
&\leq e^{-k/8} \quad \text{(by Hoeffding's inequality [32]).}
\end{aligned}$$

■

Theorem 2 shows that the median-of-means estimator has a sub-Gaussian performance with $L = 8$ for all distributions with a finite variance. However, it is important to point out that the estimator $\widehat{\mu}_n$ depends on the confidence level δ as the number of blocks k is chosen as a function of δ. This is not a desirable property, since for different values of the confidence parameter δ, one obtains a different point estimator. However, as it is shown in Section 2.4 below, there do not exist sub-Gaussian estimators that are independent of the confidence level, unless one is willing to assume more than just the finiteness of the second moment of the underlying distribution.

The results of Bubeck, Cesa-Bianchi, and Lugosi [12] and Devroye, Lerasle, Lugosi, and Oliveira [22] show that the median-of-means estimator may also be used even if the distribution of the X_i has an infinite variance but has a finite moment of order $1+\alpha$ for some $\alpha \in (0,1)$.

Theorem 3. *Let $\alpha \in (0,1]$ and let $X_1, \ldots, X_n$ be independent, identically distributed random variables with mean μ and $(1+\alpha)$-th central moment $M = \mathbb{E}\left[|X_i - \mu|^{1+\alpha}\right]$. Let m, k be positive integers and assume that $n = mk$. Then the median-of-means estimator with $k = \lceil 8\log(2/\delta)\rceil$ blocks satisfies*

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

Moreover, for any mean estimator $\widehat{\mu}_n$, there exists a distribution with mean μ and $(1+\alpha)$-th central moment M such that

$$\mathbb{P}\left\{\left|\widehat{\mu}_n - \mu\right| > \left(\frac{M^{1/\alpha}\log(2/\delta)}{n}\right)^{\alpha/(1+\alpha)}\right\} \geq \delta\,.$$

The proof of the first part follows by showing that if $c(\alpha)$ is an appropriate constant that depends only on α and

$$\eta \geq c(\alpha)\left(\mathbb{E}|X_i - \mu|^{1+\alpha}\right)^{1/(1+\alpha)}\left(\frac{1}{m}\right)^{\alpha/(1+\alpha)},$$

then

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \geq \eta\right) \leq 0.2 \,.$$

The proof of the second statement goes along the lines of Theorem 1.

We finish this section by showing that if the distribution of X has a finite moment of order $2+\alpha$ for some $\alpha > 0$, then the median-of-means estimator has a sub-Gaussian performance under a much wider range of choices for the parameter k that counts the number of blocks. The following bound is due to Minsker and Strawn [68]. For simplicity of the exposition, we only consider the case $\alpha = 1$.

Theorem 4. *Let $X_1, \ldots, X_n$ be independent, identically distributed random variables with mean μ, variance σ^2, and third central moment $\rho = \mathbb{E}|X-\mu|^3$. Let m, k be positive integers and assume that $n = mk$. Assume that*

$$\sqrt{\frac{\log(2/\delta)}{2k}} + \frac{\rho}{2\sigma^3\sqrt{m}} \leq 1/4 \,. \tag{2.4}$$

Then the median-of-means estimator $\widehat{\mu}_n$ with k blocks satisfies that, with probability at least $1-\delta$,

$$\left|\widehat{\mu}_n - \mu\right| \leq \frac{1}{c}\left(\sigma\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{\rho k}{2\sigma^2 n}\right),$$

where $c = \phi(\Phi^{-1}(3/4))$ is a constant. Here ϕ and Φ denote the standard normal density and distribution functions.

Observe that the first term on the right-hand side of the bound is of the sub-Gaussian form. The second term is smaller than the first whenever the number k of blocks satisfies

$$k \leq \frac{2\sigma^3}{\rho}\sqrt{n\log(2/\delta)} \,.$$

In particular, $k \leq \frac{2\sigma^3}{\rho}\sqrt{n}$ suffices to get a sub-Gaussian performance. This is nice since with such a choice the estimator does not depend on the value of the confidence parameter δ and the estimator is sub-Gaussian simultaneously for the entire range of values of δ permitted by the condition (2.4). Also, note that the number of blocks may be chosen to be much larger than the choice suggested by Theorem 2. In particular, k can be as large as a constant multiple of $\sqrt{n}$. In that case the median-of-means estimator is sub-Gaussian simultaneously for all $\delta \geq e^{-c_0\sqrt{n}}$ for an appropriate constant c_0. The price to pay is the extra assumption of the existence of the third moment. Minsker and Strawn [68] also prove that, when $k = o(\sqrt{n})$, then, under the assumptions of Theorem 4, $\sqrt{n}\,(\widehat{\mu}_n - \mu)$ is asymptotically normal with mean zero and variance $\sigma^2\pi/2$.

Proof. Note that $\widehat{\mu}_n \in [\mu - a, \mu + a]$ if $a > 0$ is such that

$$\frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{Z_j - \mu \leq a} \geq \frac{1}{2} \quad \text{and} \quad \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{Z_j - \mu \geq -a} \geq \frac{1}{2}\,.$$

We show that, with probability at least $1 - \delta$, one may take

$$a = \frac{1}{c}\left(\sigma\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{\rho k}{2\sigma^2 n}\right).$$

To this end, note that

$$\begin{aligned}
\frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{Z_j - \mu \leq a} &= \frac{1}{k}\sum_{j=1}^{k}\Big(\mathbb{1}_{Z_j - \mu \leq a} - \mathbb{P}\Big\{Z_j - \mu \leq a\Big\}\Big) \\
&\quad + \left(\mathbb{P}\{Z_1 - \mu \leq a\} - \mathbb{P}\left\{G\frac{\sigma}{\sqrt{m}} \leq a\right\}\right) \\
&\quad + \mathbb{P}\left\{G\frac{\sigma}{\sqrt{m}} \leq a\right\} \\
&\quad \text{(where } G \text{ is a standard normal random variable).}
\end{aligned}$$

First note that, by Hoeffding's inequality, with probability at least $1 - \delta/2$,

$$\frac{1}{k}\sum_{j=1}^{k}\Big(\mathbb{1}_{Z_j - \mu \leq a} - \mathbb{P}\Big\{Z_j - \mu \leq a\Big\}\Big) \geq -\sqrt{\frac{\log(2/\delta)}{2k}}\,.$$

For the second term on the right-hand side, we may use the Berry-Esseen theorem (see Shevtsova [72]) that implies that

$$\mathbb{P}\{Z_1 - \mu \leq a\} - \mathbb{P}\left\{G\frac{\sigma}{\sqrt{m}} \leq a\right\} \geq -\frac{\rho}{2\sigma^3\sqrt{m}}\,.$$

Hence, we have that, with probability at least $1 - \delta/2$,

$$\frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{Z_j - \mu \leq a} \geq \mathbb{P}\left\{G\frac{\sigma}{\sqrt{m}} \leq a\right\} - \sqrt{\frac{\log(2/\delta)}{2k}} - \frac{\rho}{2\sigma^3\sqrt{m}}\,.$$

Thus, $(1/k)\sum_{j=1}^{k} \mathbb{1}_{Z_j - \mu \leq a} \geq \frac{1}{2}$ with probability at least $1 - \delta/2$, whenever a is such that

$$\mathbb{P}\left\{G \leq a\frac{\sqrt{m}}{\sigma}\right\} \geq \frac{1}{2} + \sqrt{\frac{\log(2/\delta)}{2k}} + \frac{\rho}{2\sigma^3\sqrt{m}}\,.$$

If $\sqrt{\frac{\log(2/\delta)}{2k}} + \frac{\rho}{2\sigma^3\sqrt{m}} \le 1/4$ then it suffices to consider values of a with $a\sqrt{m}/\sigma \le \Phi^{-1}(3/4)$. Then

$$\mathbb{P}\left\{G \le a\frac{\sqrt{m}}{\sigma}\right\} \ge \frac{1}{2} + c\frac{a\sqrt{m}}{\sigma}$$

with $c = \phi(\Phi^{-1}(3/4))$. Hence, we may take

$$a = \frac{\sigma}{c\sqrt{m}}\left(\sqrt{\frac{\log(2/\delta)}{2k}} + \frac{\rho}{2\sigma^3\sqrt{m}}\right) = \frac{1}{c}\left(\sigma\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{\rho k}{2\sigma^2 n}\right).$$

The same argument shows that, with probability at least $1 - \delta/2$,

$$\frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{Z_j - \mu \ge -a} \ge \frac{1}{2}$$

for the choice of a above. ∎

2.2 Catoni's estimator

Next we present a completely different approach for constructing a mean estimator, introduced and analyzed by Catoni [14]. To introduce Catoni's idea, note first that the empirical mean $\overline{\mu}_n$ is just the solution $y \in \mathbb{R}$ of the equation

$$\sum_{i=1}^{n} (X_i - y) = 0\,.$$

Catoni proposed to replace the left-hand side of the equation above by another strictly decreasing function of y of the form

$$R_{n,\alpha}(y) = \sum_{i=1}^{n} \psi\left(\alpha(X_i - y)\right)\,,$$

where $\psi : \mathbb{R} \to \mathbb{R}$ is an antisymmetric increasing function and $\alpha \in \mathbb{R}$ is a parameter. The idea is that if $\psi(x)$ increases much slower than x, then the effect of "outliers" present due to heavy tails is diminished. Catoni offers a whole range of "influence" functions ψ. For the ease of exposition, we single out one specific choice, namely

$$\psi(x) = \begin{cases} \log(1 + x + x^2/2) & \text{if } x \ge 0 \\ -\log(1 - x + x^2/2) & \text{if } x < 0\,. \end{cases}$$

We define Catoni's mean estimator $\widehat{\mu}_{\alpha,n}$ as the unique value y such that $R_{n,\alpha}(y) = 0$ with this choice of ψ. Since $\psi(x) \le \log(1 + x + x^2/2)$ for all $x \in \mathbb{R}$, we have, for all $y \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[e^{R_{n,\alpha}(y)}\right] &\le \left(\mathbb{E}\left[1 + \alpha(X-y) + \frac{\alpha^2(X-y)^2}{2}\right]\right)^n \\
&= \left(1 + \alpha(\mu-y) + \frac{\alpha^2\left(\sigma^2 + (\mu-y)^2\right)}{2}\right)^n \\
&\le \exp\left(n\alpha(\mu-y) + \frac{n\alpha^2\left(\sigma^2 + (\mu-y)^2\right)}{2}\right),
\end{aligned}$$

whenever the X_i have a finite variance σ^2. Thus, by Markov's inequality, we have that, for any fixed $y \in \mathbb{R}$ and $\delta \in (0,1)$,

$$\mathbb{P}\left\{R_{n,\alpha}(y) \ge n\alpha(\mu-y) + \frac{n\alpha^2\left(\sigma^2 + (\mu-y)^2\right)}{2} + \log(1/\delta)\right\} \le \delta\,.$$

Suppose that the parameter α is such that $\alpha^2\sigma^2 + 2\log(1/\delta)/n \le 1$. Then the quadratic polynomial of y

$$n\alpha(\mu-y) + \frac{n\alpha^2\left(\sigma^2 + (\mu-y)^2\right)}{2} + \log(1/\delta)$$

has at least one root. In particular, taking the smaller root

$$y_+ = \mu + \frac{\frac{\alpha\sigma^2}{2} + \frac{\log(1/\delta)}{n\alpha}}{\frac{1}{2} + \frac{1}{2}\sqrt{1 - \alpha^2\sigma^2 - \frac{2\log(1/\delta)}{n}}}\,,$$

we have that $R_{n,\alpha}(y_+) < 0$ with probability at least $1-\delta$. Since $R_{n,\alpha}(y)$ is strictly decreasing, this implies that $\widehat{\mu}_{\alpha,n} < y_+$ with probability at least $1-\delta$. A symmetric argument shows that $\widehat{\mu}_{\alpha,n} > y_-$ with probability at least $1-\delta$, where

$$y_- = \mu - \frac{\frac{\alpha\sigma^2}{2} + \frac{\log(1/\delta)}{n\alpha}}{\frac{1}{2} + \frac{1}{2}\sqrt{1 - \alpha^2\sigma^2 - \frac{2\log(1/\delta)}{n}}}\,.$$

Now by straightforward bounding, and choosing the parameter α to optimize the bounds, we obtain the following performance estimate.

Theorem 5. *Let $X_1, \dots, X_n$ be independent, identically distributed random variables with mean μ and variance σ^2. Let $\delta \in (0,1)$ be such that $n > 2\log(1/\delta)$. Catoni's mean estimator $\widehat{\mu}_{n,\alpha}$ with parameter*

$$\alpha = \sqrt{\frac{2\log(1/\delta)}{n\sigma^2\left(1 + \frac{2\log(1/\delta)}{n-2\log(1/\delta)}\right)}}$$

satisfies that, with probability at least $1-2\delta$,

$$\left|\widehat{\mu}_{n,\alpha} - \mu\right| < \sqrt{\frac{2\sigma^2\log(1/\delta)}{n - 2\log(1/\delta)}}\,. \tag{2.5}$$

The theorem highlights that, with an appropriately chosen parameter α, Catoni's estimator has a sub-Gaussian performance. Quite remarkably, the constant $\sqrt{2}$ is the best possible. A disadvantage of Catoni's estimator with respect to median-of-means is that the estimator—at least in the form given in the theorem—depends on the variance σ^2. In general, it is unrealistic to assume knowledge of σ^2. If one substitutes σ^2 in the formula of α by an upper bound v, then the bound (2.5) still holds with v replacing σ^2. In case no good upper bound for σ^2 is available, Catoni [14] shows how to use Lepski's method to select α from the data that has near-optimal performance. Huber [37] combines the median-of-means estimator with Catoni's estimator into a two-step procedure that to obtain an estimator with the optimal leading constant in the sub-Gaussian bound when $|\sigma/\mu|$ is bounded by a known constant.

Another problem—shared with the median-of-means estimator—is that Catoni's estimator also depends on the required confidence level δ. Such a dependence is necessary as it is shown in Section 2.4 below. A quick fix is to use the estimator with a δ-independent parameter, though then the resulting estimate, naturally, cannot be sub-Gaussian. One reasonable choice is $\alpha = \sqrt{2/(n\sigma^2)}$. In this case, it is easy to see that, whenever $n > 2(1+\log(1/\delta))$, Catoni's estimator satisfies, with probability at least $1-2\delta$,

$$\left|\widehat{\mu}_{n,\alpha} - \mu\right| < \sqrt{\frac{\sigma^2}{2n}} \cdot \frac{1+\log(1/\delta)}{1 - \frac{1+\log(1/\delta)}{n}}\,.$$

This is not a sub-Gaussian bound because of an extra factor of $\sqrt{\log(1/\delta)}$ but the "sub-exponential" tail probabilities are still non-trivial and useful.

2.3 Trimmed mean

Perhaps the most natural attempt to improve the performance of the empirical mean is removing possible outliers using a truncation of X. Indeed, the so-called

trimmed-mean (or *truncated*-mean) estimator is defined by removing a fraction of the sample, consisting of the ϵn largest and smallest points for some parameter $\epsilon \in (0,1)$, and then averaging over the rest. This idea is one of the most classical tools in robust statistics and we refer to Tukey and McLaughlin [78], Huber and Ronchetti [39], Bickel [8], Stigler [74] for early work on the theoretical properties of the trimmed-mean estimator. However, it was only recently that the non-asymptotic sub-Gaussian property of the trimmed mean was established. Indeed, Oliveira and Orenstein [70] proved that if ϵ is chosen proportionally to $\log(1/\delta)/n$, then the trimmed-mean estimator has a sub-Gaussian performance for all distributions with a finite variance.

To show how this works in the simplest way, here we analyze a simple variant of the trimmed-mean estimator.

The estimator splits the data in two equal parts. One half is used to determine the correct truncation level. The points from the other half are averaged, except for the data points that fall outside of the truncation region, which are ignored. For convenience of the notation, we assume that the data consists of $2n$ independent copies of the random variable X, denoted by $X_1, \dots, X_n, Y_1, \dots, Y_n$.

For $\alpha \le \beta$, define the truncation function

$$\phi_{\alpha,\beta}(x) = \begin{cases} \beta & \text{if } x > \beta, \\ x & \text{if } x \in [\alpha, \beta], \\ \alpha & \text{if } x < \alpha, \end{cases}$$

and for $x_1, \dots, x_m \in \mathbb{R}$ let $x_1^* \le x_2^* \le \cdots \le x_m^*$ be its non-decreasing rearrangement.

With this notation in place, the definition of the estimator is as follows:

(1) Given the confidence level $\delta \ge 8e^{-3n/16}$, set

$$\varepsilon = \frac{16 \log(8/\delta)}{3n}.$$

(2) Let $\alpha = Y^*_{\varepsilon n}$ and $\beta = Y^*_{(1-\varepsilon)n}$ (assuming, for simplicity, that εn is an integer) and set

$$\widehat{\mu}_{2n} = \frac{1}{n} \sum_{i=1}^{n} \phi_{\alpha,\beta}(X_i).$$

Theorem 6. *Let $X_1, \dots, X_n, Y_1, \dots, Y_n$ be independent, identically distributed random variables with mean μ and variance σ^2. Let $\delta \in (0,1)$ be such that $n > (16/3)\log(8/\delta)$. Then, with probability at least $1 - \delta$,*

$$|\widehat{\mu}_{2n} - \mu| \le 9\sigma \sqrt{\frac{\log(8/\delta)}{n}}.$$

Proof. We start by showing that the truncation level is close to the appropriate quantiles of the distribution. To this end, for $p \in (0,1)$, introduce the quantiles

$$Q_p = \sup\{M \in \mathbb{R} : \mathbb{P}\{X \geq M\} \geq 1-p\} .$$

For ease of exposition, assume that X has a nonatomic distribution. (This assumption is not necessary, but simplifies notation.) In that case $\mathbb{P}\{X > Q_p\} = \mathbb{P}\{X \geq Q_p\} = 1-p$.

By a straightforward application of Bernstein's inequality, with probability at least $1 - 2\exp(-(3/16)\varepsilon n)$, we have both

$$|\{i \in [n] : Y_i \geq Q_{1-2\varepsilon}\}| \geq \varepsilon n$$

and

$$\left|\{i \in [n] : Y_i \leq Q_{1-\varepsilon/2}\}\right| \geq (1-\varepsilon)n .$$

This implies that, with probability at least $1 - 2\exp(-(3/16)\varepsilon n)$,

$$Q_{1-2\varepsilon} \leq Y^*_{(1-\varepsilon)n} \leq Q_{1-\varepsilon/2} . \tag{2.6}$$

By the same argument, with probability at least $1 - 2\exp(-(3/16)\varepsilon n)$,

$$Q_{\varepsilon/2} \leq Y^*_{\varepsilon n} \leq Q_{2\varepsilon} , \tag{2.7}$$

From here, we simply need to show that $|\mathbb{E}\phi_{\alpha,\beta}(X) - \mu|$ is small and that $(1/n)\sum_{i=1}^n \phi_{\alpha,\beta}(X_i)$ concentrates around its mean.

For the first step, consider the event E that both (2.6) and (2.7) hold. This event has probability at least $1 - 4\exp(-(3/16)\varepsilon n) = 1 - \delta/2$. On the event E,

$$\begin{aligned}
&\left|\mathbb{E}\left[\phi_{\alpha,\beta}(X)|Y_1,\ldots,Y_n\right] - \mu\right| \\
&\quad\leq \left|\mathbb{E}\left[(X-\alpha)\mathbb{1}_{X\leq\alpha}|Y_1,\ldots,Y_n\right]\right| + \left|\mathbb{E}\left[(X-\beta)\mathbb{1}_{X\geq\beta}|Y_1,\ldots,Y_n\right]\right| \\
&\quad\leq \left|\mathbb{E}(X-Q_{2\varepsilon})\mathbb{1}_{X\leq Q_{2\varepsilon}}\right| + \left|\mathbb{E}(X-Q_{1-2\varepsilon})\mathbb{1}_{X\geq Q_{1-2\varepsilon}}\right| .
\end{aligned}$$

To bound these two terms, forst notice that, by Chebyshev's inequality,

$$2\varepsilon = \mathbb{P}\{X \geq Q_{1-2\varepsilon}\} \leq \frac{\sigma_X^2}{(Q_{1-2\varepsilon}-\mu)^2} ,$$

and in particular,

$$Q_{1-2\varepsilon} \leq \mu + \frac{\sigma}{\sqrt{2\varepsilon}} .$$

Hence, by the Cauchy-Schwarz inequality,

$$\begin{aligned}|\mathbb{E}(X-Q_{1-2\varepsilon})\mathbb{1}_{X\geq Q_{1-2\varepsilon}}| &= |\mathbb{E}(X-\mu)-(Q_{1-2\varepsilon}-\mu))\mathbb{1}_{X\geq Q_{1-2\varepsilon}}| \\ &\leq \mathbb{E}|(X-\mu)|\mathbb{1}_{X\geq Q_{1-2\varepsilon}}+(Q_{1-2\varepsilon}-\mu)\mathbb{P}\{X\geq Q_{1-2\varepsilon}\} \\ &\leq \sigma\sqrt{\mathbb{P}\{X\geq Q_{1-2\varepsilon}\}}+2\varepsilon(Q_{1-2\varepsilon}-\mu) \\ &\leq \sigma\sqrt{8\varepsilon}\,.\end{aligned}$$

A symmetric argument shows $|\mathbb{E}(X-Q_{2\varepsilon})\mathbb{1}_{X\leq Q_{2\varepsilon}}|\leq\sigma\sqrt{8\varepsilon}$, and therefore, on the event E, we have

$$\left|\mathbb{E}\left[\phi_{\alpha,\beta}(X)|Y_1,\ldots,Y_n\right]-\mu\right|\leq\sigma\sqrt{32\varepsilon}\leq 6\sigma\sqrt{\frac{\log(8/\delta)}{n}}$$

by our choice of ϵ. Next, let

$$Z=\frac{1}{n}\sum_{i=1}^{n}\phi_{\alpha,\beta}(X_i)-\mathbb{E}\left[\phi_{\alpha,\beta}(X)|Y_1,\ldots,Y_n\right]$$

and observe that

$$Z=\frac{1}{n}\sum_{i=1}^{n}\phi_{\alpha-\mu,\beta-\mu}(X_i-\mu)-\mathbb{E}\left[\phi_{\alpha-\mu,\beta-\mu}(X-\mu)|Y_1,\ldots,Y_n\right].$$

Hence, on the event E (that only depends on $Y_1,\ldots,Y_n$), Z is an average of centered random variables that is bounded point-wise by $M=\max\{|Q_{\varepsilon/2}-\mu|,|Q_{1-\varepsilon/2}-\mu|\}\leq\sigma\sqrt{2/\varepsilon}$ and whose variance is at most σ^2. Therefore, by Bernstein's inequality, with probability at least $1-\delta/2$,

$$Z\leq\sigma\sqrt{\frac{2\log(2/\delta)}{n}}+\frac{\log(2/\delta)\sigma\sqrt{2/\varepsilon}}{n}\leq 3\sigma\sqrt{\frac{\log(2/\delta)}{n}}\,.$$

Putting the pieces together, we obtain the announced bound. ∎

Besides its conceptual simplicity, an important advantage of the trimmed mean compared to other estimators with sub-Gaussian performance is that it is robust to adversarial contamination of the data. This statement is formalized and proved in [55] where a multivariate extension is also introduced and analyzed.

2.4 Multiple-δ estimators

We have constructed various estimators–such as median-of-means and Catoni's estimator–that are sub-Gaussian under the only assumption that the underlying distribution has a finite second moment. However, both estimators depend on the

knowledge of the desired confidence parameter δ. We show next that is not a coincidence because without further information on the distribution, it is impossible to construct a single estimator that is sub-Gaussian for a nontrivial range of values of the confidence parameter δ. Next we reproduce a simplified version of an argument of Devroye, Lerasle, Lugosi, and Oliveira [22] who proved results of this kind. The theorem below shows that it is impossible to construct an estimator that is L-sub-Gaussian for some specified values of δ_1 and δ_2, at the same time. The particular values of δ_1 and δ_2 are of no special importance. We present this result to show the basic ideas in a simple form. For more general versions we refer to [22].

Theorem 7. *For every $L \geq 50$ and for every sample size n, no estimator can be simultaneously L-sub-Gaussian for both $\delta_1 = 1/(2e\sqrt{L^3+1})$ and $\delta_2 = 2e^{-L^4/4}$ for all distributions with finite second moment.*

Proof. We show that not only it is impossible to construct a single L-sub-Gaussian estimator for both $\delta_1 = 1/(2e\sqrt{L^3+1})$ and $\delta_2 = e^{-L^4/4}$ for all distributions with finite second moment but it is also the case for the restricted class of Poisson distributions.

Assume, on the contrary, that there exists an estimator $\widehat{\mu}_n$ that is L-sub-Gaussian for both δ_1 and δ_2 for all Poisson distributions. Let $X_1, \ldots, X_n$ be independent Poisson random variables with parameter $1/n$ and let $Y_1, \ldots, Y_n$ be independent Poisson random variables with parameter c/n, where we set $c = L^3 + 1$. We assume, for the sake of simplicity, that c is an integer. By the sub-Gaussian property of $\widehat{\mu}_n$,

$$\mathbb{P}\left\{\widehat{\mu}_n(Y_1, \ldots, Y_n) < \frac{c}{n} - \frac{L}{n}\sqrt{c \log \frac{1}{\delta_1}}\right\} \leq \delta_1 \ . \tag{2.8}$$

Now note that the left-hand side of the inequality may be lower bounded as folows:

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{\mu}_n(Y_1, \ldots, Y_n) < \frac{c}{n} - \frac{L}{n}\sqrt{c \log \frac{1}{\delta_1}}\right\} \\
&\quad \geq \quad \mathbb{P}\left\{\widehat{\mu}_n(Y_1, \ldots, Y_n) < \frac{c}{n} - \frac{L}{n}\sqrt{c \log \frac{1}{\delta_1}}, \sum_{i=1}^n Y_i = c\right\} \\
&\quad \geq \quad \frac{1}{e\sqrt{c}}\mathbb{P}\left\{\widehat{\mu}_n(Y_1, \ldots, Y_n) < \frac{c}{n} - \frac{L}{n}\sqrt{c \log \frac{1}{\delta_1}} \Big| \sum_{i=1}^n Y_i = c\right\} \\
&\qquad \text{(from the fact that } \sum_{i=1}^n Y_i \text{ is Poisson with parameter } c \text{ and Stirling's formula)}
\end{aligned}$$

Next we use the fact that the conditional joint distribution of n independent Poisson(λ) random variables, conditioned on the event that their sum equals c, only depends

on c but not λ. In particular,

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{\mu}_n(Y_1,\ldots,Y_n) < \frac{c}{n} - \frac{L}{n}\sqrt{c\log\frac{1}{\delta_1}}\Big|\sum_{i=1}^n Y_i = c\right\} \\
&\quad = \mathbb{P}\left\{\widehat{\mu}_n(X_1,\ldots,X_n) < \frac{c}{n} - \frac{L}{n}\sqrt{c\log\frac{1}{\delta_1}}\Big|\sum_{i=1}^n X_i = c\right\} .
\end{aligned}$$

Thus, together with (2.8), and the choice $\delta_1 = 1/(2e\sqrt{c})$. we have that

$$\begin{aligned}
\frac{1}{2} &= 1 - e\sqrt{c}\delta_1 \\
&\leq \frac{\mathbb{P}\left\{\widehat{\mu}_n(X_1,\ldots,X_n) \geq \frac{c}{n} - \frac{L}{n}\sqrt{c\log\frac{1}{\delta_1}}, \sum_{i=1}^n X_i = c\right\}}{\mathbb{P}\{\sum_{i=1}^n X_i = c\}} \\
&\leq ec!\mathbb{P}\left\{\widehat{\mu}_n(X_1,\ldots,X_n) \geq \frac{c}{n} - \frac{L}{n}\sqrt{c\log\frac{1}{\delta_1}}\right\} \\
&\leq ec!\mathbb{P}\left\{\widehat{\mu}_n(X_1,\ldots,X_n) \geq \frac{1}{n} + \frac{c-1}{n} - \frac{L}{n}\sqrt{c\log\frac{1}{\delta_1}}\right\} \\
&\leq ec!\mathbb{P}\left\{\widehat{\mu}_n(X_1,\ldots,X_n) \geq \frac{1}{n} + \frac{c-1}{2n}\right\} ,
\end{aligned}$$

where we used the fact that

$$\frac{L}{n}\sqrt{c\log\frac{1}{\delta_1}} \leq \frac{c-1}{2n} ,$$

that follows from our choice of δ_1 whenever $L \geq 10$. Now since $\widehat{\mu}_n$ is L-sub-Gaussian for $\delta_2 = 2e^{-L^4/4}$, we have that

$$\mathbb{P}\left\{\widehat{\mu}_n(X_1,\ldots,X_n) \geq \frac{1}{n} + \frac{c-1}{2n}\right\} = \mathbb{P}\left\{\widehat{\mu}_n(X_1,\ldots,X_n) \geq \frac{1}{n} + \frac{L}{n}\sqrt{\log(2/\delta_2)}\right\} \leq \delta_2 .$$

Summarizing, we have $1/2 \leq ec!\delta_2 = 2ec!e^{-L^4/4}$. However, the expression on the right-hand side is less than $1/2$ for $L \geq 50$, leading to a contradiction. ∎

We refer to [22] for a more complete version of Theorem 7 and for an extensive discussion on constructing estimators that do not require knowledge of the desired confidence parameter (i.e., estimators that are sub-Gaussian for a wide range of values of δ). In [22] it is shown how Lepski's method may be used to construct such estimators if some additional information, other than finiteness of the variance, is available on the underlying distribution. In particular, if nontrivial upper and lower bounds on the variance are available, then such "δ-independent" estimators exist for a wide range of values of δ. Existence of higher moments or certain weak symmetry assumptions may also be used.

3 Estimating the mean of a random vector

In what follows, we discuss extensions of the mean estimation problem to the multivariate setting. To set up the problem, let X be a random vector taking values in $\mathbb{R}^d$. Assume that the mean vector $\mu = \mathbb{E}X$ and covariance matrix $\Sigma = \mathbb{E}(X-\mu)(X-\mu)^T$ exist. Given n independent, identically distributed samples $X_1, \ldots, X_n$ drawn from the distribution of X, one wishes to estimate the mean vector.

Just like in the univariate case, a natural choice is the sample mean $\overline{\mu}_n = (1/n)\sum_{i=1}^n X_i$ and it has a near-optimal behavior whenever the distribution is sufficiently light tailed. However, as is the case in the univariate case, whenever heavy tails are a concern, the sample mean is to be avoided as it may have a sub-optimal performance.

3.1 Sub-Gaussian performance

For the univariate problem, we constructed mean estimators with a sub-Gaussian performance. In order to properly set up our goal for the d-dimensional case, first we need to understand what "sub-Gaussian performance" means. Just like in the univariate case, one would like to construct estimators that are "close" to the true mean μ, with "high probability". The first question is how one measures distance in $\mathbb{R}^d$. Arguably, the most natural distance measure is the Euclidean norm. In this section we focus on this choice and we denote by $\|\cdot\|$ the Euclidean norm. We explore mean estimation of a random vector with respect to an arbitrary norm in Section 4.1.

If X has a multivariate normal distribution with mean vector μ and covariance matrix Σ, then the sample mean $\overline{\mu}_n$ is also multivariate normal with mean μ and covariance matrix $(1/n)\Sigma$. Thus, for all $t > 0$,

$$\mathbb{P}\left\{\|\overline{\mu}_n - \mu\| \geq \mathbb{E}\|\overline{\mu}_n - \mu\| + t\right\} = \mathbb{P}\left\{\left\|\overline{X}\right\| - \mathbb{E}\left\|\overline{X}\right\| \geq t\sqrt{n}\right\},$$

where $\overline{X}$ is a Gaussian vector in $\mathbb{R}^d$ with zero mean and covariance matrix Σ. A key property of Gaussian vectors is that $\overline{X}$ has the same distribution as $\Sigma^{1/2}Y$ where Y is a standard normal vector (i.e., with zero-mean and identity covariance matrix) and $\Sigma^{1/2}$ is the positive semidefinite square root of Σ. Also, observe that for all $y, y' \in \mathbb{R}^d$,

$$\left|\left\|\Sigma^{1/2}y\right\| - \left\|\Sigma^{1/2}y'\right\|\right| \leq \left\|\Sigma^{1/2}(y-y')\right\| \leq \left\|\Sigma^{1/2}\right\|_{2\to 2} \cdot \|y-y'\|,$$

where $\left\|\Sigma^{1/2}\right\|_{2\to 2}$ is the spectral norm of $\Sigma^{1/2}$. Thus, $\Sigma^{1/2}y$ is a Lipschitz function of $y \in \mathbb{R}^d$ with Lipschitz constant $\|\Sigma^{1/2}\|_{2\to 2} = \sqrt{\lambda_{\max}}$, with $\lambda_{\max} = \lambda_{\max}(\Sigma)$ denoting

the largest eigenvalue of the covariance matrix Σ. Now it follows from the Gaussian concentration inequality of Tsirelson, Ibragimov, and Sudakov [75] (see also Ledoux [51] and Boucheron, Lugosi, and Massart [10] for more information) that

$$\mathbb{P}\left\{\left\|\overline{X}\right\| - \mathbb{E}\left\|\overline{X}\right\| \geq t\sqrt{n}\right\} \leq e^{-nt^2/(2\lambda_{\max})} .$$

Noting that

$$\mathbb{E}\left\|\overline{X}\right\| \leq \sqrt{\mathbb{E}\left\|\overline{X}\right\|^2} = \sqrt{\mathrm{Tr}(\Sigma)} ,$$

the trace of the covariance matrix Σ, we have that, for $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|\overline{\mu}_n - \mu\| \leq \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{\frac{2\lambda_{\max}\log(1/\delta)}{n}} . \tag{3.1}$$

Thus, in the multivariate case, we will say that a mean estimator is *sub-Gaussian* if, with probability at least $1-\delta$, it satisfies an inequality of the form (3.1) (with possibly different constant factors). Note that for any distribution with mean μ and covariance matrix Σ, the mean-squared error of the empirical mean equals

$$\mathbb{E}\|\overline{\mu}_n - \mu\|^2 = \frac{\mathrm{Tr}(\Sigma)}{n} .$$

In particular, $\mathbb{E}\|\overline{\mu}_n - \mu\| \leq \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}$. An important feature of the sub-Gaussian property (3.1) is that the random fluctuations are controlled by the spectral norm $\lambda_{\max}$ of the covariance matrix, which is possibly much smaller than $\mathrm{Tr}(\Sigma)$, the sum of all eigenvalues of Σ.

3.2 Multivariate median-of-means

For non-Gaussian and possibly heavy-tailed distributions, one cannot expect a sub-Gaussian behavior of the sample mean similar to (3.1).

As an alternative, one may try to extend the median-of-means estimator to the multivariate case. An obvious idea is to divide the data into disjoint blocks, calculate the empirical mean within each block, and compute a multivariate median of the obtained empirical means. However, there is no standard notion of a median for multivariate data, and it is not entirely clear what definition of a multivariate median works best for median-of-means mean estimators. Among the numerous possibilities, we mention the *coordinate-wise median*, the *geometric (or spatial) median*, the *Tukey (or halfspace) median*, the *Oja median*, and the *Liu median*, see Small [73] for a survey and relevant references.

Regardless of what notion of a multivariate median we decide to adopt, we start by partitioning $[n] = \{1,\dots,n\}$ into k blocks $B_1,\dots,B_k$, each of size $|B_i| \geq$

$\lfloor n/k \rfloor \geq 2$. Here k is a parameter of the estimator to be chosen later. For simplicity, we assume that $km = n$ for some positive integer m. Just like before, we compute the sample mean of the random vectors within each block: for $j = 1, \ldots, k$, let

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

Perhaps the most natural first try is to define $\widehat{\mu}_n$ as the vector of coordinate-wise medians of the Z_j (i.e., the ℓ-th component of the vector $\widehat{\mu}_n$ is the median of the ℓ-th components of $Z_1, \ldots, Z_k$, for $\ell \in [d]$). Then Theorem 2 and the union bound imply that, for any $\delta \in (0,1)$, taking $k = \lceil 8 \log(1/\delta) \rceil$, with probability at least $1 - \delta$,

$$\left\| \widehat{\mu}_n - \mu \right\| \leq \sqrt{\frac{32 \mathrm{Tr}(\Sigma) \log(d/\delta)}{n}} \, ,$$

where we used the fact that $\mathrm{Tr}(\Sigma) = \mathbb{E}\|X - \mathbb{E}X\|^2$ is the sum of the variances of the d components of X. Clearly, this bound is far from the sub-Gaussian inequality (3.1) for several reasons. First, it is not "dimension-free" as d appears explicitly in the bound. Perhaps more importantly, $\log(1/\delta)$ is multiplied by $\mathrm{Tr}(\Sigma)$ instead of $\lambda_{\max}(\Sigma)$ and that may be a major difference in high-dimensional problems, especially when one is interested in small failure probabilities. An instructive example is when all eigenvalues of Σ are identical and equal to $\lambda_{\max}$. If the dimension d is large, (3.1) is of the order of $\sqrt{(\lambda_{\max}/n)(d + \log(1/\delta))}$ while the bound above only gives the order $\sqrt{(\lambda_{\max}/n)(d \log(d/\delta))}$.

One may quite easily improve on this by using a different (non-standard) notion of median in the definition of the estimate: choose $\widehat{\mu}_n$ to be the point in $\mathbb{R}^d$ with the property that the Euclidean ball centered at $\widehat{\mu}_n$ that contains more than $k/2$ of the points Z_j has minimal radius. Since $\mathbb{E}\|Z_j - \mu\|^2 = \mathrm{Tr}(\Sigma)/m$, by Chebyshev's inequality, $\|Z_j - \mu\| \leq r \stackrel{\text{def.}}{=} 2\sqrt{\mathrm{Tr}(\Sigma)/m}$ with probability at least 3/4. Thus, by choosing $k = \lceil 8 \log(1/\delta) \rceil$, we have that, with probability at least $1 - \delta$, more than half of the points Z_j satisfy

$$\|Z_j - \mu\| \leq r \, .$$

Denote this event by E. (Thus, $\mathbb{P}\{E\} \geq 1 - \delta$.) On the event E, this radius is at most r. Hence, at least one of the Z_j is within distance r to both μ and $\widehat{\mu}_n$. Thus, by the triangle inequality, $\|\widehat{\mu}_n - \mu\| \leq 2r$. We have obtained the following proposition.

Proposition 1. *Let $X_1, \ldots, X_n$ be i.i.d. random vectors in $\mathbb{R}^d$ with mean μ and covariance matrix Σ. Let $\delta \in (0,1)$ and let $\widehat{\mu}_n$ be the estimator defined above with $k = \lceil 8 \log(1/\delta) \rceil$. Then, with probability at least $1 - \delta$,*

$$\left\| \widehat{\mu}_n - \mu \right\| \leq 4 \sqrt{\frac{\mathrm{Tr}(\Sigma)(8 \log(1/\delta) + 1)}{n}} \, .$$

The bound of Proposition 1 is quite remarkable as it is "dimension-free" and no assumption other than the existence of the covariance matrix is made. However, it still does not achieve a sub-Gaussian performance bound that resembles (3.1). Moreover, the notion of median used here (i.e., the center of the smallest ball that contains at least half of the points) is problematic from a computational point of view, since computing such a median is a nontrivial problem.

An efficiently computable version of a multivariate median is the so-called *geometric median*, defined as

$$\widehat{\mu}_n = \operatorname*{argmin}_{m \in \mathbb{R}^d} \sum_{j=1}^{k} \|Z_i - m\| .$$

This estimator was proposed by Minsker [65] and independently by Hsu and Sabato [36] (see also Lerasle and Oliveira [53]). In particular, Minsker [65] proved that this version of the multivariate median-of-means estimator achieves a similar performance bound as Proposition 1. Moreover, computing the geometric median—and therefore the multivariate median-of-means estimator—involves solving a convex optimization problem. Thus, the geometric median may be approximated efficiently, see Cohen, Lee, Miller, Pachocki, and Sidford [20] for the most recent result and for the rich history of the problem. We refer to Aloupis [2] for a survey of computational aspects of various other notions of multivariate medians.

For a quite different mean estimator based on the median-of-means idea with "almost" sub-Gaussian guarantees but with a serious computational burden, see Joly, Lugosi, and Oliveira [41].

In order to achieve a truly sub-Gaussian performance, we need to define a new estimator. In what follows we define two that achieve the desired performance: the first, introduced in [56] is based on the idea of median-of-means tournaments and the second, from [57], is defined using the intersection of random slabs. The former leads to an error estimate with respect to the Euclidean norm (see Section 3.4), and the latter, described in Section 4.1 holds with respect to an arbitrary norm. However, before presenting these estimates, we recall a very different estimator introduced by Catoni and Giulini [16].

3.3 Thresholding the norm: the Catoni-Giulini estimator

In this section we briefly discuss a remarkably simple estimator, suggested and analyzed by Catoni and Giulini [16]. The Catoni-Giulini estimator is

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \min\left(1, \frac{1}{\alpha \|X_i\|}\right) , \tag{3.2}$$

where $\alpha > 0$ is a (small) parameter. Thus, $\widehat{\mu}_n$ is simply an empirical average of the X_i, with the data points with large norm shrunk towards zero. This estimate is trivial to compute, as opposed to the more complex estimators that we discuss in Sections 3.4 and 4.1. On the other hand, shrinking to zero is somewhat arbitrary and unnatural. In fact, the estimator is not invariant under translations of the data in the sense that $\widehat{\mu}_n(X_1 + a, \dots, X_n + a)$ is not necessarily equal to $\widehat{\mu}_n(X_1, \dots, X_n) + a$ when $a \neq 0$.

Catoni and Giulini prove that if one chooses the parameter as

$$\alpha = \sqrt{\frac{c \log(1/\delta)}{v n}} \,,$$

where $v \geq \lambda_{\max}$ and $c > 0$ is a numerical constant, then the estimator (3.2) satisfies, with probability at least $1 - \delta$,

$$\left\| \widehat{\mu}_n - \mu \right\| \leq C \sqrt{\frac{(\mathrm{Tr}(\Sigma) + v + \|\mu\|^2) \log(1/\delta)}{n}} \,,$$

where C is a constant depending on c only. This bound is similar to but weaker than that of Proposition 1, principally due to two facts. First, the estimator requires prior knowledge of (a good upper bound of) $\lambda_{\max}$ whereas the geometric median-of-means estimator assumes no such prior information. Second, $\|\mu\|^2$ appears in the upper bound and a priori this can be arbitrarily large compared to $\mathrm{Tr}(\Sigma)$. The presence of this term is due to the lack of translation invariance of the estimator. This second issue may be fixed by defining a two-stage estimator: first one may use a translation-invariant estimator like geometric-median-of-means defined in the previous section to get a rough estimate of the mean. Then, using a new batch of independent data, one may center the data at the estimated mean and then use the Catoni-Giulini estimator for the centered data. This new estimator is translation invariant, and the term $\|\mu\|^2$ may be replaced by the squared error of the estimator of the first step, that is, by $\mathrm{Tr}(\Sigma)\log(1/\delta)/n$. But even with this modification, the bound is not sub-Gaussian in the sense of (3.1).

Remarkably, however, the performance of the Catoni-Giulini estimator comes close to being sub-Gaussian in the desired sense under just a small extra assumption. In particular, if $\mathbb{E}\|X\|^\beta < \infty$ for some $\beta > 2$, then, with the same choice of α as above, one has

$$\left\| \widehat{\mu}_n - \mu \right\| \leq C \left(\sqrt{\frac{v \log(1/\delta)}{n}} + \sqrt{\frac{(\mathrm{Tr}(\Sigma) + v)}{n}} + \frac{\kappa_\beta}{n^{(\beta-1)/2}} \right) ,$$

where κ_β is a constant that depends on β and the β-th raw moment of $\|X\|$. Thus, if the prior parameter v is a good estimate of $\lambda_{\max}$ in the sense that it is bounded by

a constant multiple of it, then the first two terms of the bound are of the desired sub-Gaussian form. The third term is of smaller order though again, it can be arbitrarily large if the mean is far from the origin, which may be remedied by making the estimator more complex. We refer to Catoni and Giulini [15] for other estimates of a similar spirit and more discussion. The proof techniques of [15, 16] rely on so-called *PAC-Bayesian* inequalities whose details go beyond the scope of this survey.

3.4 Median-of-means tournaments

Here we introduce a mean estimator with a sub-Gaussian performance for all distributions whose covariance matrix exists, proposed by Lugosi and Mendelson [56]. The estimator presented below is the first and simplest instance of what we call *median-of-means tournaments*.

Recall that we are given an i.i.d. sample $X_1, \dots, X_n$ of random vectors in $\mathbb{R}^d$. As in the case of the median-of-means estimator, we start by partitioning the set $\{1, \dots, n\}$ into k blocks $B_1, \dots, B_k$, each of size $|B_j| \geq m \stackrel{\text{def.}}{=} \lfloor n/k \rfloor$, 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, we assume that n is divisible by k and therefore $|B_j| = m$ for all $j = 1, \dots, k$.

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

$$T_a = \left\{ x \in \mathbb{R}^d : \exists J \subset [k] : |J| \geq k/2 \text{ such that for all } j \in J, \ \|Z_j - x\| \leq \|Z_j - a\| \right\} \tag{3.3}$$

and define the mean estimator by

$$\widehat{\mu}_n \in \operatorname*{argmin}_{a \in \mathbb{R}^d} \operatorname{radius}(T_a) \, ,$$

where $\operatorname{radius}(T_a) = \sup_{x \in T_a} \|x - a\|$. Thus, $\widehat{\mu}_n$ is chosen to minimize, over all $a \in \mathbb{R}^d$, the radius of the set T_a defined as the set of points $x \in \mathbb{R}^d$ for which $\|Z_j - x\| \leq \|Z_j - a\|$ for the majority of the blocks. If there are several minimizers, one may pick any one of them.

The set T_a may be seen as the set of points in $\mathbb{R}^d$ that are at least as close to the point cloud $\{Z_1, \dots, Z_k\}$ as the point a. The estimator $\widehat{\mu}_n$ is obtained by minimizing the radius of T_a.

Note that the minimum is always achieved. This follows from the fact that radius(T_a) is a continuous function of a (since, for each a, T_a 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 $\operatorname{argmin}_{a\in\mathbb{R}^d} \text{radius}(T_a)$ as yet another 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 estimator coincides with the median-of-means estimator.

The following performance bound shows that the estimator has the desired sub-Gaussian performance.

Theorem 8. *(Lugosi and Mendelson [56].) Let $\delta \in (0,1)$ and consider the mean estimator $\widehat{\mu}_n$ with parameter $k = \lceil 200\log(2/\delta)\rceil$. 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 $1-\delta$,*

$$\left\|\widehat{\mu}_n - \mu\right\| \le \max\left(960\sqrt{\frac{\operatorname{Tr}(\Sigma)}{n}}, 240\sqrt{\frac{\lambda_{max}\log(2/\delta)}{n}}\right).$$

Just like the performance bound of Proposition 1, Theorem 8 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 8 remains valid as long as $\operatorname{Tr}(\Sigma) = \mathbb{E}\|X-\mu\|^2$ is finite.

Theorem 8 is an outcome of the following observation.

Theorem 9. *Using the same notation as above and setting*

$$r = \max\left(960\sqrt{\frac{\operatorname{Tr}(\Sigma)}{n}}, 240\sqrt{\frac{\lambda_{max}\log(2/\delta)}{n}}\right),$$

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. In other words, $\|a-\mu\| \ge r$ implies that $a \notin T_\mu$.

Theorem 9 implies that for a 'typical' collection $X_1,\dots,X_n$, 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 $x_1,\dots,x_n \subset \mathbb{R}^d$ such a point need not even exist, and it is surprising that for a typical i.i.d. configuration, this property is satisfied by μ.

The fact that Theorem 9 implies Theorem 8 is straightforward. Indeed, the definition of $\widehat{\mu}_n$ and Theorem 9 imply that, with probability at least $1-\delta$, radius$(T_{\widehat{\mu}_n}) \le$ radius$(T_\mu) \le r$. Since either $\mu \in T_{\widehat{\mu}_n}$ or $\widehat{\mu} \in T_\mu$, we must have $\|\widehat{\mu}_n - \mu\| \le r$, as required.

The constants appearing in Theorem 8 are certainly not optimal. They were obtained with the goal of making the proof transparent.

The proof of Theorem 9 is based on the following idea. The mean μ is the minimizer of the function $f(x) = \mathbb{E}\|X - x\|^2$. A possible approach is to use the available data to guess, for any pair $a, b \in \mathbb{R}^d$, whether $f(a) < f(b)$. A natural choice is to use a median of means estimator to decide which of the two is better. The "tournament" is simply a way of comparing every such pair, as described next. [1].

To define the tournament, 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

$$\frac{1}{m} \sum_{i \in B_j} \left(\|X_i - b\|^2 - \|X_i - a\|^2 \right) > 0 \tag{3.4}$$

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

Lemma 1. *Let $\delta \in (0,1)$, $k = \lceil 200 \log(2/\delta) \rceil$, and define*

$$r = \max\left(960 \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}, 240 \sqrt{\frac{\lambda_{max} \log(2/\delta)}{n}} \right).$$

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

The outcome of Lemma 1 stands to reason: if $\|b - \mu\|$ is large enough, that will be reflected in 'typical values' of $(\|X_i - \mu\|)_{i=1}^n$ and $(\|X_i - b\|)_{i=1}^n$. Comparing the values via (3.4) ensures 'stability', and the fact that b is far from μ is exhibited with high probability. We stress that the probability estimate has to be *uniform* in b. Such uniform estimates are a recurring theme in what follows.

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 $\|b - \mu\| \geq r$, 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 $\|v\| \geq r$,

$$-\frac{2}{m} \sum_{i \in B_j} \left\langle \overline{X}_i, v \right\rangle + \|v\|^2 > 0 \tag{3.5}$$

[1] As we explain in what follows, it suffices to ensure that the comparison is correct between μ and any point that is not too close to μ.

for more than $k/2$ blocks B_j. Clearly, it suffices to show that (3.5) 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,

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

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

$$r = \|v\| \geq 4\sqrt{10} \sqrt{\frac{\lambda_{\max}}{m}} \tag{3.6}$$

then with probability at least 9/10,

$$-\frac{2}{m} \sum_{i \in B_j} \left\langle \overline{X}_i, v \right\rangle \geq \frac{-r^2}{2}. \tag{3.7}$$

Applying Hoeffding's inequality ([32]), we see that (3.7) holds for a single v with probability at least $1 - \exp(-k/50)$ 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 (3.7) 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 $v_1, v_2 \in V_1$, $\|v_1 - v_2\|_{L_2(X)} = \langle v_1 - v_2, \Sigma(v_1 - v_2) \rangle^{1/2} \geq \epsilon$. We may estimate this cardinality by the "dual Sudakov" inequality (see Ledoux and Talagrand [52] and also Vershynin[83] for a version with the specific constant used here): the cardinality of V_1 is bounded by

$$\log(|V_1|/2) \leq \frac{1}{32} \left(\frac{\mathbb{E}\left[\langle G, \Sigma G \rangle^{1/2} \right]}{\epsilon / r} \right)^2,$$

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,

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

Hence, by setting

$$\epsilon = 2r \left(\frac{1}{k} \mathrm{Tr}(\Sigma) \right)^{1/2}, \tag{3.8}$$

we have $|V_1| \le 2e^{k/100}$ and by the union bound, with probability at least $1-2e^{-k/100} \ge 1-\delta/2$, (3.7) holds for all $v \in V_1$ on at least 8/10 of the blocks B_j.

Next we check that property (3.5) 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 $1-\exp(-k/200) \ge 1-\delta/2$,

$$\sup_{x \in r\cdot S^{d-1}} \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{\{|m^{-1}\sum_{i\in B_j}\langle \overline{X}_i, x-v_x\rangle| \ge r^2/4\}} \le \frac{1}{10}\,. \tag{3.9}$$

Indeed, on that event it follows that for every $x \in r \cdot S^{d-1}$, on at least 7/10 of the 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 (3.9). 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 $\|x-v_x\|_{L_2(X)} = \sqrt{\mathbb{E}\left\langle \overline{X}, x-v_x\right\rangle^2} \le \epsilon$ it follows that for every j

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

and therefore,

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

To bound (B), note that, by (3.8),

$$\frac{4\epsilon}{r^2\sqrt{m}} = 8\left(\frac{\mathrm{Tr}(\Sigma)}{n}\right)^{1/2} \cdot \frac{1}{r} \le \frac{1}{60}$$

provided that

$$r \ge 480\left(\frac{\mathrm{Tr}(\Sigma)}{n}\right)^{1/2}.$$

We may bound (A) by standard techniques of empirical processes such as symmetrization, contraction for Rademacher averages and de-symmetrization. Indeed, let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher random variables (i.e., $\mathbb{P}\{\sigma_i = 1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$), independent of all of the X_i. Then

$$\begin{aligned}
(A) &\le \frac{8}{r^2}\mathbb{E}\sup_{x\in r\cdot S^{d-1}} \frac{1}{k}\sum_{j=1}^{k}\sigma_j \left|\frac{1}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, x - v_x\right\rangle\right| \\
&\quad \text{(by a standard symmetrization inequality, see, e.g., [81, Lemma 2.3.6])} \\
&\le \frac{8}{r^2}\mathbb{E}\sup_{x\in r\cdot S^{d-1}} \left|\frac{1}{k}\sum_{j=1}^{k}\sigma_j \frac{1}{m}\sum_{i\in B_j}\left\langle \overline{X}_i, x - v_x\right\rangle\right| \\
&\quad \text{(by a contraction lemma for Rademacher averages, see [52])} \\
&\le \frac{16}{r^2}\mathbb{E}\sup_{x\in r\cdot S^{d-1}} \left|\frac{1}{n}\sum_{i=1}^{n}\left\langle \overline{X}_i, x - v_x\right\rangle\right| \\
&\quad \text{(see again [81, Lemma 2.3.6])} \\
&\le \frac{32}{r}\mathbb{E}\sup_{\{t:\|t\|\le 1\}} \left|\frac{1}{n}\sum_{i=1}^{n}\left\langle \overline{X}_i, t\right\rangle\right| \\
&\quad \text{(noting that } \|x - v_x\| \le 2r\text{)} \\
&\le \frac{32}{r}\cdot\frac{\sqrt{\mathbb{E}\left\|\overline{X}\right\|^2}}{\sqrt{n}} = \frac{32}{r}\left(\frac{\mathrm{Tr}(\Sigma)}{n}\right)^{1/2} \le \frac{1}{30}
\end{aligned}$$

provided that $r \ge 960\left(\frac{\mathrm{Tr}(\Sigma)}{n}\right)^{1/2}$.

Thus, for

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

we have proved that $\mathbb{E}Y \le 1/60 + 1/30 = 1/20$. Finally, in order to establish (3.9), it suffices to show that, $\mathbb{P}\{Y > \mathbb{E}Y + 1/20\} \le e^{-k/200}$, which follows from the bounded differences inequality (see, e.g., [10, Theorem 6.2]). ∎

Proof of Theorem 9

Theorem 9 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

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

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

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

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 $\|a-\mu\| \geq r$ then on more than $k/2$ blocks B_j, $\frac{1}{m}\sum_{i\in B_j}\Big(\|X_i - a\|^2 - \|X_i - \mu\|^2\Big) > 0$, which, by the above argument, is the same as saying that for at least $k/2$ indices j, $\|Z_j - a\| > \|Z_j - \mu\|$. ∎

Upon reflection it is clear that the ideas used in the proof of Theorem 8 are rather general. In fact, they are at the heart of the *small-ball method* introduced in Mendelson [61] (see also [63] for results of similar flavour). The small-ball method holds in far more general situations than Theorem 8 and will be repeated throughout this note. To explain how the argument can be extended, let us outline again the three steps that allowed us to compare every b and μ:

(1) For any *fixed* $b \in \mathbb{R}^d$ we obtain a bound that holds with high probability;

(2) Then, thanks to the high probability estimate from (1), we invoke the union bound and control a large (yet finite) collection of points.

We have complete freedom to choose the collection as we want, and we select it as an ϵ-net in the set in question.

(3) The crucial part of the argument is passing from the control we have on every point in the net to the wanted uniform control on entire class; specifically, we show that if a 'center', that is, an element of the net, is well-behaved[2],

[2] in the proof of Theorem 8, 'well-behaved' means that (3.5) holds for a majority of the blocks.

then the same is true for any point close enough to the center. To that end, we show that 'random oscillations' do not destroy the good behaviour of a center on too many blocks.

3.5 Computational considerations

An important issue that we have mostly swept under the rug so far is computational feasibility of mean estimators. While the empirical mean is trivial to compute, many of the more sophisticated estimators discussed here are far from being so. In particular, a basic requirement for any multivariate mean estimator for having a chance to being useful in practice is that it can be computed in polynomial time (i.e., in time that is a polynomial of the sample size n and the dimension d). As we already pointed it out, some of the estimators described above fall in this category. For example, the geometric median-of-means estimator or the Catoni-Giulini estimator are both efficiently computable in this sense. However, these estimators fall short from being sub-Gaussian. The median-of-means tournament estimator is sub-Gaussian but its computation poses a highly nontrivial challenge. In fact, the way the estimator is defined, it is likely to be computationally intractable (i.e., NP hard). However, in a recent beautiful paper, Hopkins [33] defines a clever semi-definite relaxation of the median-of-means tournament estimator that can be computed in time $O(nd + d\log(1/\delta)^c)$ for a dimension-independent constant and, at the same time, achieves the desired sub-Gaussian guarantee under the only assumption that the covariance matrix exists. This is the first efficiently computable sub-Gaussian multivariate mean estimator. Even more recently, Cherapanamjeri, Flammarion, and Bartlett [18] improved the running time to $O(nd + d\log(1/\delta)^2 + \log(1/\delta)^4)$ by combining Hopkins' ideas with clever gradient-descent optimization. This is likely not the last word on the subject as many exciting computational challenges arise in the context of mean estimation and regression.

In the theoretical computer science community there has been a recent important surge of results that address the problem of computationally efficient *robust* mean estimation. In this context, an estimator is defined to be robust if it performs well in the presence of a small constant fraction of (possibly adversarial) outliers. Various different models have been introduced, see Charikar, Steinhardt, and Valiant [17], Diakonikolas, Kamath, Kane, Li, Moitra, and Stewart [23, 24, 25], Diakonikolas, Kane, and Stewart [26], Diakonikolas, Kong, and Stewart [27], Hopkins and Li [34], Klivans, Kothari, and Meka [42], Kothari, Steinhardt, and Steurer [44], Lai, Rao, and Vempala [45], Loh and Tan [54], for a sample of this important growing body of literature. Surveying this area goes beyond the scope of this paper.

4 Uniform median-of-means estimators

The median-of-means tournament used in the previous section is an example of a uniform median-of-means estimator: given a class of functions $\mathcal{F}$, there is a high-probability event on which, for every f in the class, the median of means estimator based on the data $f(X_1),\dots,f(X_n)$ is close to the mean $\mathbb{E}f(X)$. Indeed, the tournament is simply a median-of-means estimator that was used to check whether a was closer to μ than b, or vice-versa, uniformly for every pair $a,b\in\mathbb{R}^n$.

In what follows we present a general version of a uniform median-of-means estimator and turn our attention to two applications: estimating the mean of a random vector with respect to an arbitrary norm, and L_2-distance oracles (the latter proves useful in regression problems, see Section 4.2 and [59] for more details).

Formally, the question we consider is as follows:

> Let $\mathcal{F}$ be a class of functions on a probability space (Ω,ν). Given an independent sample $(X_1,\dots,X_n)$ distributed according to ν, find an estimator $\widehat{\phi}_n(f)$ for each $f\in\mathcal{F}$, such that, with high probability, for every $f\in\mathcal{F}$, $|\widehat{\phi}_n(f)-\mathbb{E}f(X)|$ is small.

A natural idea is to define $\widehat{\phi}_n(f)$ to be the median-of-means estimator based on $f(X_1),\dots,f(X_n)$. It stands to reason that the bound established in Section 2.1 for the performance of the median-of-means estimator cannot simply hold uniformly for every $f\in\mathcal{F}$. Rather, the uniform error consists of two terms: the 'worst' individual estimate for a function $f\in\mathcal{F}$, and a 'global' error, taking into account the 'complexity' of the class.

To analyze uniform median-of-means estimators, it is natural to follow the path of the small-ball method outlined in the previous section. To this end, fix integers k and m and let $n=mk$. As always, we split the given sample into k blocks, each one of cardinality m, keeping in mind that the natural choice is $k\sim\log(2/\delta)$ if one wishes a confidence of $1-\delta$. For $0<\eta<1$ set

$$p_m(\eta)=\sup_{f\in\mathcal{F}}\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m f(X_i)-\mathbb{E}f(X)\right|\geq\eta\right),$$

denote by $D=\left\{f:\mathbb{E}f(X)^2\leq 1\right\}$ the unit ball in $L_2(\nu)$ and let $\mathcal{M}(\mathcal{F},rD)$ be the maximal cardinality of a subset of $\mathcal{F}$ that is r-separated with respect to the $L_2(\nu)$ norm.

The following bound was recently established in [57].

Theorem 10. *There exist absolute constants $c_0,\dots,c_4$ for which the following holds. Set η_0,η_1 and $\eta_2\geq c_0\eta_1/\sqrt{m}$ that satisfy the following:*

(1) $p_m(\eta_0) \le 0.05$;

(2) $\log \mathcal{M}(\mathcal{F}, \eta_1 D) \le c_2 k \log(e/p_m(\eta_0))$;

(3) $\mathbb{E} \sup_{w \in \overline{W}} \left| \sum_{i=1}^n \varepsilon_i w(X_i) \right| \le c_3 \eta_2 n$,

where $\varepsilon_1, \dots, \varepsilon_n$ *are independent Rademacher random variables (i.e.,* $\mathbb{P}\{\varepsilon_i = 1\} = \mathbb{P}\{\varepsilon_i = -1\} = 1/2$*) and* $W = (\mathcal{F} - \mathcal{F}) \cap \eta_1 D = \{f_1 - f_2 : f_1, f_2 \in \mathcal{F},\ \|f_1 - f_2\|_{L_2} \le \eta_1\}$ *and* $\overline{W} = \{w - \mathbb{E}w : w \in W\}$*. Let* $r = \eta_0 + \eta_2$*. Then, with probability at least* $1 - 2\exp(-c_4 k)$*, for all* $f \in \mathcal{F}$ *one has*

$$\left| \frac{1}{m} \sum_{i \in B_j} f(X_i) - \mathbb{E}f \right| \le r \ \text{ for at least } 0.6k \text{ blocks } B_j \ .$$

The error r in Theorem 10 has the two terms we expected. Indeed, η_0 is error one would have if the goal were to obtain an individual mean estimator for a fixed function in $\mathcal{F}$: writing $\sigma_f = \sqrt{\mathrm{Var}(f(X))}$, by Chebyshev's inequality, for every $f \in \mathcal{F}$,

$$\mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^m f(X_i) - \mathbb{E}f \right| \ge \eta_0 \right) \le \frac{\sigma_f^2}{m\eta_0^2} \le 0.05$$

provided that

$$\eta_0 \gtrsim \frac{\sigma_f}{\sqrt{m}} \sim \sigma_f \sqrt{\frac{\log(2/\delta)}{n}} \ .$$

As outlined in Section 2.1, this leads to the standard sub-Gaussian error estimate for the function $f \in \mathcal{F}$. On the other hand, η_2 involves the Rademacher averages associated with $\mathcal{F} - \mathcal{F}$, and captures the price one has to pay for the uniform control over the class $\mathcal{F}$.

The proof of Theorem 10 follows the same path we outlined previously: the definition of p_m allows us to show that the empirical mean of f on a block B_j of cardinality m is close to the true mean with reasonable probability, say, larger than 0.95. Thus, with probability $1 - e^{-ck}$, this property is satisfied on $0.9k$ blocks. Next, the high-probability estimate combined with the union bound allow us to control all the elements in a finite set uniformly, as long as its cardinality is at most exponential in k. The set of choice is an appropriate net in $\mathcal{F}$ and its mesh width η_1 is selected to ensure that the cardinality of the net is small enough. Finally, as always, the crucial component is to ensure that oscillations do not 'corrupt' the good behaviour on too many blocks. Since our interest is in the median of means, one can live with up to $0.4k$ of the blocks being corrupted, and the additional error of η_2 suffices to guarantee that indeed no more than $0.4k$ blocks are affected.

The technical analysis can be found in [57], where Theorem 10 is used for the study the problem of multivariate mean estimation with respect to a general norm, outlined in the next section.

We mention here that uniform estimators based on Catoni's mean estimator were studied by Brownlees, Joly, and Lugosi [11] in the context of regression function estimation. Minsker [67] discusses uniform estimators in a similar spirit to those presented here, also for adversarially contaminated data.

4.1 Multivariate mean estimation—the general case

To illustrate the power of the uniform median-of-means bounds established in the previous section, we now return to the problem of estimating the mean of a random vector. As before, let $X_1, \dots, X_n$ be independent, identically distributed random vectors in $\mathbb{R}^d$ with mean μ and covariance matrix Σ. The question we seek to answer is to what extent one can estimate μ when the error is measured by a given norm $\|\cdot\|$ that is not necessarily the Euclidean norm. An important example is the matrix operator norm, see Minsker [66], Catoni and Giulini [15], Mendelson and Zhivotovskiy [64].

One may now cast this general mean estimation problem in the framework of uniform median-of-means estimators outlined above. The natural class of functions associated with the problem is the unit ball with respect to the dual of the norm $\|\cdot\|$ (i.e., the set of norm-one linear functionals). The natural choice of a measure ν is the one induced by $X - \mu$.

Consider the event given by Theorem 10 for this class of functions and denote the resulting error by r. It follows that for each norm-one functional x^*, we have $\mathbb{E}x^*(X-\mu) = 0$ and

$$\left| \frac{1}{m} \sum_{i \in B_j} x^*(X_j - \mu) \right| \leq r$$

for a majority of the blocks B_j. Moreover,

$$\frac{1}{m} \sum_{i \in B_j} x^*(X_i - \mu) = x^*\Big(\frac{1}{m} \sum_{i \in B_j} X_i\Big) - x^*(\mu) \, .$$

Thus, setting $Z_j = \frac{1}{m} \sum_{i \in B_j} X_j$, Theorem 10 implies that for every norm-one functional x^*,

$$|x^*(Z_j) - x^*(\mu)| \leq r \, .$$

In other words, if we define the sets

$$\mathbb{S}_{x^*} = \Big\{ y \in \mathbb{R}^d : |x^*(Z_j) - x^*(y)| \leq r \ \text{ for the majority of indices } j \Big\}$$

then on the event from Theorem 10 one has that

$$\mu \in \mathbb{S} = \bigcap_{\|x^*\|=1} \mathbb{S}_{x^*} .$$

From a geometric point of view, each set $\mathbb{S}_{x^*}$ is the union of intersection of slabs: setting $\alpha_j = x^*(Z_j)$,

$$\mathbb{S}_{x^*} = \bigcup_{|I| \geq [k/2]+1} \bigcap_{i \in I} \{y : |x^*(y) - \alpha_j| \leq r\} ,$$

which is just a union of (potentially empty) slabs, defined by the functional x^*. The set $\mathbb{S}$ is the resulting intersection of the sets $\mathbb{S}_{x^*}$. Off hand, there is no reason why the intersection of the sets S_{x^*} should not be empty. The fact that it contains μ is only due to the special nature of the Z_j's.

Since each set $\mathbb{S}_{x^*}$ is data-dependent, so is $\mathbb{S}$. With that in mind, the estimator we propose is obvious: set $\widehat{\mu}_n^{(r)}$ to be any point in $\mathbb{S}$. To show that $\|\widehat{\mu}_n^{(r)} - \mu\| \leq 2r$, fix any norm-one functional x^*. Recall that if $y \in \mathbb{S}$ then $|x^*(Z_j) - x^*(y)| \leq r$ on the majority of blocks. Therefore, if $\mu, \widehat{\mu}_n^{(r)} \in \mathbb{S}$ there is some index j such that, simultaneously,

$$|x^*(Z_j) - x^*(\widehat{\mu}_n^{(r)})| \leq r \quad \text{and} \quad |x^*(Z_j) - x^*(\mu)| \leq r ,$$

and therefore $|x^*(\widehat{\mu}^{(r)} - \mu)| \leq 2r$. Thanks to Theorem 10, there is a high-probability event on which this is true for any norm-one functional, and, in particular,

$$\|\widehat{\mu}_n^{(r)} - \mu\| = \sup_{\|x^*\|=1} |x^*(\widehat{\mu}_n^{(r)} - \mu)| \leq 2r .$$

Remark. It is straightforward to verify that there is no need to select $\mathcal{F}$ to be the set of all the norm-one linear functionals. It is enough to define $\mathbb{S}$ using the functionals that are extreme points of the unit ball in the dual space.

Thanks to Theorem 10 and the argument we just outlined, the following was established in [57]:

Theorem 11. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. Suppose that the X_i have mean μ and covariance matrix Σ. There exists a mean estimator $\widehat{\mu}_n$ such that, with probability at least $1 - \delta$,*

$$\|\widehat{\mu}_n - \mu\| \leq \frac{c}{\sqrt{n}} \left(\max \left\{ \mathbb{E}\|\zeta_n\|, \ \mathbb{E}\|G\| + R\sqrt{\log(2/\delta)} \right\} \right) ,$$

where c is a numerical constant,

$$R = \sup_{\|x^*\|=1} \left(\mathbb{E}(x^*(X - \mu))^2 \right)^{1/2} ,$$

$$\zeta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i (X_i - \mu) \,,$$

$\{\epsilon_i\}$ is a sequence of i.i.d. Rademacher random variables independent of $\{X_i\}$, and G is the centered Gaussian vector with covariance matrix Σ.

As is explained in [57], Theorem 11, there is a good reason to believe that the bound of the theorem is optimal in a rather strong sense. We refer the reader to [57] for more details.

Remark. Note that the error in Theorem 11 has two types of terms: $\frac{R}{\sqrt{n}}\sqrt{\log(2/\delta)}$ is the standard one-dimensional sub-Gaussian error, and its source is the marginal $x^*(X)$ with the largest variance. At the same time, $\mathbb{E}\|G\|$ and $\mathbb{E}\|\zeta_n\|$ are 'global' parameters that calibrate the 'complexity' of the norm. When $\|\cdot\|$ is the Euclidean norm, we recover the two terms on the right-hand side of (3.1).

4.2 L_2 distance oracles

In this section we sketch how the ideas used in Theorem 10 may be used to generate a median-of-means based (isomorphic) L_2 *distance oracle*. A more accurate description of distance oracles may be found in [59].

Suppose $\mathcal{F}$ is a class of real-valued functions on a probability space (Ω, ν) and let X be a random variable distributed as ν. There are many natural situations in which, given an i.i.d. sample $X_1, \ldots, X_n$, one would like to have an accurate estimate on the L_2 distance $\|f-h\|_{L_2} \stackrel{\text{def.}}{=} \sqrt{\mathbb{E}(f(X)-h(X))^2}$ between any two class members $f, h \in \mathcal{F}$.

In some cases, the estimates are required to be almost isometric, that is, with high probability, for all $f, h \in \mathcal{F}$, the estimate should lie in the range $[(1-\eta)\|f-h\|_{L_2}, (1+\eta)\|f-h\|_{L_2}]$ for some small value of η. However, in many situations (for example, in the regression problem we describe in Section 5), a weaker property suffices: one would like to define a data-dependent functional $\widehat{\Psi}_n$ such that , with 'high' probability, for all $f, h \in \mathcal{F}$ and a 'small' value r, and some constants $0 < \alpha < 1 < \beta$,

- if $\widehat{\Psi}_n(f,h) \geq r$ then $\alpha\|f-h\|_{L_2} \leq \widehat{\Psi}_n(f,h) \leq \beta\|f-h\|_{L_2}$;
- if $\widehat{\Psi}_n(f,h) \leq r$ then $\|f-h\|_{L_2} \leq r/\alpha$.

In other words, for every $f, h \in \mathcal{F}$, based on the value of the data-dependent functional $\widehat{\Psi}_n(f,h)$ one may estimate $\|f-h\|_{L_2}$ in an isomorphic way—i.e., up to multiplicative constants. We call such a functional a *distance oracle*.

For the sake of simplicity, instead of considering simultaneous estimation of pairwise distances of functions, we address the problem of estimating L_2 norms of functions. In other words, given a class $\mathcal{F}$ of functions as above, we are interested in constructing a data-dependent functional $\widehat{\Psi}_n$ such that if $\widehat{\Psi}_n(f) \geq r$ then $\alpha\|f\|_{L_2} \leq \widehat{\Psi}_n(f) \leq \beta\|f\|_{L_2}$, and if $\widehat{\Psi}_n(f) < r$ then $\|f\|_{L_2} \lesssim r/\alpha$. Such a functional may be called a *norm oracle*. Given a norm oracle, one may construct a distance oracle in an obvious way.

In what follows we assume that there is some $q > 2$ such that the L_q and L_2 norms are equivalent on $\{f_1 - f_2 : f_1, f_2 \in \mathcal{F} \cup \{0\}\}$. In other words, there is a constant L such that $\|f_1 - f_2\|_{L_q} \leq L\|f_1 - f_2\|_{L_2}$ for all $f_1, f_2 \in \mathcal{F} \cup \{0\}$. Consider the set

$$H = \mathrm{star}(\mathcal{F}, 0) = \{\lambda f : f \in \mathcal{F},\ 0 \leq \lambda \leq 1\}$$

and let $H_\rho = H \cap \rho S(L_2)$, where $\rho S(L_2) = \{h : \|h\|_{L_2} = \rho\}$. For every $h \in H$, set

$$Z_h(j) = \frac{1}{m} \sum_{i \in B_j} |h(X_i)|$$

and our estimator $\widehat{\Psi}_n(h)$ is the median of $Z_h(1), \ldots, Z_h(k)$.

Recall that $D = \left\{f : \mathbb{E}f(X)^2 \leq 1\right\}$ denotes the unit ball in $L_2(\nu)$ and let $\mathcal{M}(\mathcal{F}, rD)$ be the maximal cardinality of a subset of $\mathcal{F}$ that is r-separated with respect to the $L_2(\nu)$ norm.

Theorem 12. *There exist constants c_1, A, B that depend on q and L, and absolute constants $c_2, \ldots, c_6$ such that the following holds. Let $m = c_1(L, q)$ and set $k = n/m$. Under the $L_q - L_2$ norm equivalence condition, if*

$$\log \mathcal{M}(H_\rho, c_2 A\rho D) \leq c_3 k\ ,$$

and

$$\mathbb{E} \sup_{w \in (H_\rho - H_\rho) \cap c_2 A\rho D} \left| \sum_{i=1}^{n} \varepsilon_i w(X_i) \right| \leq c_4 A\rho n\ ,$$

then with probability at least $1 - 2\exp(-c_5 k)$, for all $h \in H_\rho$,

- *if $\|h\|_{L_2} \geq \rho$ then $A\|h\|_{L_2} \leq \widehat{\Psi}_n(h) \leq B\|h\|_2$; and*
- *if $\|h\|_{L_2} \leq \rho$ then $\widehat{\Psi}_n(h) \leq c_6 B\rho$.*

Note that Theorem 12 shows that $\widehat{\Psi}_n$ is a desired norm oracle: if $\widehat{\Psi}_n(h) > c_6 B\rho$ then it follows that $\|h\|_{L_2} \geq \rho$, and thus

$$B^{-1}\widehat{\Psi}_n(h) \leq \|h\|_{L_2} \leq A^{-1}\widehat{\Psi}_n(h)\ .$$

On the other hand, if $\widehat{\Psi}_n(h) \le c_6 B\rho$ then one has two options: either $\|h\|_{L_2} \le \rho$, or, $\|h\|_{L_2} \ge \rho$, in which case $\|h\|_{L_2} \le A^{-1}\widehat{\Psi}_n(h) \le c_6(B/A)\rho$. Thus, $\|h\|_{L_2} \le \rho \max\{1, c_6 B/A\}$. The norm oracle is obtained by setting $r = c_6 B\rho$ and choosing α and β appropriately.

The proof of Theorem 12 follows the small-ball method: we begin by showing that for a fixed $h \in H_\rho$, and with high probability,

$$\left|\left\{ j : A\rho \le \frac{1}{m}\sum_{i\in B_j} |h|(X_i) \le B\rho \right\}\right| \ge 0.8k \tag{4.1}$$

for some constants A and B.

Then, the high-probability estimate allows us to control a satisfactory net in H_ρ, and finally, one has to control 'oscillations': a high-probability event such that if $h \in H_\rho$ and πh denotes the closest point to h in the net, then

$$\sup_{h\in H_\rho} \left|\left\{ j : \frac{1}{m}\sum_{i\in B_j} |h-\pi h|(X_i) \ge \frac{A\rho}{2} \right\}\right| \le 0.2k \ .$$

With all three components in place, it is evident that for every $h \in H_\rho$ there are at least $0.6k$ blocks B_j on which

$$A\rho \le \frac{1}{m}\sum_{i\in B_j} |\pi h|(X_i) \le B\rho \quad \text{and} \quad \frac{1}{m}\sum_{i\in B_j} |h-\pi h|(X_i) \le \frac{A\rho}{2} \ .$$

On these blocks,

$$\frac{1}{m}\sum_{i\in B_j} |h|(X_i) \ge \frac{1}{m}\sum_{i\in B_j} |\pi h|(X_i) - \frac{1}{m}\sum_{i\in B_j} |h-\pi h|(X_i) \ge \frac{A\rho}{2} \ ,$$

and a similar estimate holds for the upper bound.

Once an isomorphic estimate is established in $H_\rho = \mathrm{star}(\mathcal{F}, 0) \cap \rho S(L_2)$, the same estimate holds for any $h \in H_r$ and any $r \ge \rho$. This is evident from the fact that $H = \mathrm{star}(\mathcal{F}, 0)$ is star-shaped around 0, implying that every $h \in H_r$ has a 'scaled down' version in H_ρ. In particular, on the same event we have that if $f \in \mathcal{F}$ and $\|f\|_{L_2} \ge \rho$, then

$$\frac{A}{2}\|f\|_{L_2} \le \widehat{\Psi}_n(f) \le 2B\|f\|_{L_2} \ .$$

The second part of the claim follows the same lines (see [59] for more details).

The main question is how to ensure that (4.1) holds with high enough probability. As it happens, (4.1) can be verified under minimal assumptions, as we now explain.

Assume, for example, that the given class $\mathcal{F}$ satisfies a small-ball condition, namely, for every $\epsilon > 0$ there is a constant $\kappa(\epsilon)$ such that for every $f \in \mathcal{F}$,

$$\mathbb{P}(|f(X)| \leq \kappa(\epsilon)\|f\|_{L_2}) \leq \epsilon \, .$$

Set $\epsilon = 0.05$ and let $\kappa = \kappa(0.05)$. Then with probability at least $1 - 2\exp(-cn)$ there are at least $0.9n$ indices $i \in \{1, \ldots, n\}$ such that $|f(X_i)| \geq \kappa\|f\|_{L_2}$. At the same time,

$$\mathbb{P}(|f(X)| \geq 10\|f\|_{L_2}) \leq \frac{1}{100} \, ,$$

implying that with probability at least $1 - 2\exp(-cn)$, for at least $0.9n$ indices $1 \leq i \leq n$, $|f(X_i)| \leq 10\|f\|_{L_2}$. Thus, intersecting the two events (4.1) is established with probability at least $1 - 2\exp(-c'n)$ for $m = 1$, $A = \kappa$ and $B = 10$.

Of course it is true that not every random variable $f(X)$ satisfies the small-ball condition we use above. However, there is an additional degree of freedom that has not been exploited yet: that the random variables one truly cares about are of the form $Z_f(j) = \frac{1}{m}\sum_{i \in B_j} |f(X_i)|$, leaving us some room to generate the necessary regularity. Indeed, it is straightforward to verify that under minimal assumptions and for a small value of m, the $Z_f(j)$ do satisfy a sufficiently strong small-ball condition. This is an outcome of a Berry-Esseen type argument [3]: if there is some $q > 2$ such that $\|f\|_{L_q} \leq L\|f\|_{L_2}$ then for $m = c(q, L)$, $\sqrt{m}(Z_f(j) - \mathbb{E}|f|)$ is 'close enough' to a Gaussian variable and it follows that

$$\mathbb{P}(|Z_f(j)| \leq c_1\|f\|_{L_2}) \leq 0.05 \, .$$

5 Median-of-means tournaments in regression problems

The problem of regression function estimation essentially amounts to estimating *conditional* expectations and as such, it is a natural candidate for extending ideas of mean estimation discussed in this paper. In this section we explore some of the recent progress made in the study of regression problems driven by uniform median-of-means estimators.

The standard setup for regression function estimation may be formulated as follows. Let (X, Y) be a pair of random variables such that X takes its values in some set $\mathcal{X}$ while Y is real valued. Given a class $\mathcal{F}$ of real-valued functions

[3] The case $q = 3$ is the standard Berry-Esseen theorem while for $2 < q < 3$ one may use generalized Berry-Esseen bounds, see [71].

defined on $\mathcal{X}$, one's goal is to find $f \in \mathcal{F}$ for which $f(X)$ is a good prediction of Y. The performance of a predictor $f \in \mathcal{F}$ is measured by the *mean-squared error* $\mathbb{E}(f(X)-Y)^2$, also known as the *risk*. The best performance in the class is achieved by the risk minimizer

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}(f(X)-Y)^2 .$$

We assume in what follows that the minimum is attained and $f^* \in \mathcal{F}$ exists and is unique.

The difficulty stems from the fact that the joint distribution of (X, Y) is not known. Instead, one is given an i.i.d. sample $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$ distributed according to the joint distribution of X and Y. Given a sample size n, a *learning procedure* is a map that assigns to each sample $\mathcal{D}_n$ a (random) function in $\mathcal{F}$, which we denote by $\widehat{f}$.

The success of $\widehat{f}$ is measured by the tradeoff between the accuracy ϵ and the confidence δ with which $\widehat{f}$ attains that accuracy, that is, one would like to find $\widehat{f}$ which satisfies that

$$\mathbb{P}\left(\mathbb{E}\left(\left(\widehat{f}(X)-Y\right)^2 | \mathcal{D}_n\right) \leq \inf_{f \in \mathcal{F}} \mathbb{E}(f(X)-Y)^2 + \epsilon\right) \geq 1-\delta$$

for values of ϵ and δ as small as possible. [4]. The question of this accuracy/confidence tradeoff has been the subject of extensive study, see, for example, the books [82, 21, 29, 81, 3, 80, 60, 43, 76, 13] for a sample of the large body of work devoted to this question.

The most standard and natural way of choosing $\widehat{f}$ is by *least squares regression*, also known as *empirical risk minimization*:

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} (f(X_i)-Y_i)^2 .$$

A sample of the rich literature on the analysis of empirical risk minimization includes Györfi, Kohler, Krzyzak, Walk [29], van de Geer [80], Bartlett, Bousquet, and Mendelson [7], Koltchinskii [43], Massart [60].

The simple idea behind empirical risk minimization is that, for each $f \in \mathcal{F}$, the empirical risk $(1/n)\sum_{i=1}^n (f(X_i)-Y_i)^2$ is a good estimate of the risk $\mathbb{E}(f(X)-Y)^2$ and the minimizer of the empirical risk should nearly match that of the "true" risk. Naturally, when the empirical risks are not reliable estimates of their population counterparts, empirical risk minimization stands on shaky ground. It should not come as a surprise that the performance of empirical risk minimization changes

[4] Note that one has the freedom to select a function $\widehat{f}$ that does not belong to $\mathcal{F}$.

dramatically according to the tail behaviour of the functions involved in the given learning problem. One may show (see, e.g., [49]) that if $\mathcal{F}$ is convex and the random variables $\{f(X) : f \in \mathcal{F}\}$ and the target Y have well-behaved sub-Gaussian tails, empirical risk minimization performed in $\mathcal{F}$ yields good results: it essentially attains the optimal accuracy/confidence tradeoff for a certain range of the parameters. However, the situation deteriorates considerably when either members of $\mathcal{F}$ or Y is heavy-tailed in some sense. In such cases, the performance of empirical risk minimization may be greatly improved by employing more sophisticated mean estimation techniques. For the analysis of least squares regression for some heavy-tailed situations, see Han and Wellner [31].

A growing body of recent work has addressed the problem of constructing regression function estimators that work well even when some of the $f(X)$ and Y may be heavy tailed, see Audibert and Catoni [4], Brownlees, Joly, and Lugosi [11], Catoni and Giulini [15]. Chichignoud and Lederer [19], Fan, Li, and Wang [28], Hsu and Sabato [36], Lecué and Lerasle [46, 47], Lecué, Lerasle, and Mathieu [48], Lerasle and Oliveira [53], Lugosi and Mendelson [59, 58], Mendelson [62], and Minsker [65].

In this section we limit ourselves to sketching how median-of-means tournaments may be used in regression function estimation. Median-of-means tournaments were introduced in [59] for the study of such regression problems when $\mathcal{F}$ is a convex set. It was shown that one may attain the optimal accuracy/confidence tradeoff in prediction problems in convex classes. Similar methods were used in [58] and [47] to study the regularization framework. In these papers the convexity of the underlying class $\mathcal{F}$ played a central role in the analysis. In fact, it is convexity that allows one to define an optimal $\widehat{f}$ that takes values in $\mathcal{F}$. In the general case, when $\mathcal{F}$ need not be convex, selecting $\widehat{f} \in \mathcal{F}$ can be a poor choice (see, e.g. the discussion in [61]), and one has to adopt a totally different approach for naming an estimator.

An optimal choice of $\widehat{f}$ for an arbitrary class $\mathcal{F}$ was introduced by Mendelson [62], and that choice is also based on median-of-means tournament, though a different tournament than the one defined in [59].

Finally, we mention the general framework of ρ-estimators introduced by Baraud, Birgé, and Sart [6] and Baraud and Birgé [5]. The construction of their estimators bears certain similarities with the tournament procedures described here.

For the sake of simplicity, we will only consider the problem of regression in a closed and convex class $\mathcal{F}$. We set

$$f^* = \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}(f(X) - Y)^2$$

to be the minimizer in $\mathcal{F}$ of the risk, and since $\mathcal{F}$ is convex and closed, such a

minimizer exists and is unique. The excess risk of $f \in \mathcal{F}$ is defined to be

$$\mathbb{E}\mathcal{L}_f = \mathbb{E}(f(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2$$

and the aim is to ensure that $\mathbb{E}(\mathcal{L}_{\widehat{f}}|D) \leq \epsilon$ with probability at least $1 - \delta$.

As one may expect from a median-of-means estimator, we select $k \leq n$ wisely, split the given sample $(X_i, Y_i)_{i=1}^n$ to k blocks, each of cardinality $m = n/k$, and compare the statistical performance of every pair of functions on each block. Just as before, the belief is that because $\mathbb{E}(f^*(X) - Y)^2$ is smaller than $\mathbb{E}(f(X) - Y)^2$ this fact is exhibited by a median-of-means estimate, allowing us to prefer f^* over f. Hence, if we can find a uniform median-of-means estimator, such a comparison would lead us to a function that has almost the same risk as f^*.

With that in mind, the natural choice of a "match" in the tournament between two candidate functions f and h is counting the number of blocks on which $\frac{1}{m}\sum_{i\in B_j}(f(X_i) - Y_i)^2$ is larger than $\frac{1}{m}\sum_{i\in B_j}(h(X_i) - Y_i)^2$. The function that exhibits a superior performance (i.e., has a smaller empirical mean) on the majority of the blocks is the winner of the match.

In a perfect world, we would choose a function that won all of its matches. However, the world is far from perfect and the outcomes of matches between functions that are 'too close' are not reliable. To address this issue, the tournament requires an additional component: a *distance oracle*, similar to the one presented in the previous section. Thanks to the distance oracle one may verify in a data-dependent way when two functions are too close, and in such cases disregard the outcome of the match between them.

Let us describe some technical facts that are at the heart of the results in [59, 58]. Define the "quadratic" and "multiplier" processes

$$\mathbb{Q}_{f,h}(j) = \frac{1}{m}\sum_{i\in B_j}(f(X_i) - h(X_i))^2, \quad \mathbb{M}_{f,h}(j) = \frac{2}{m}\sum_{i\in B_j}(f(X_i) - h(X_i))(h(X_i) - Y_i)$$

and put

$$\mathbb{B}_{f,h}(j) \equiv \frac{1}{m}\sum_{i\in B_j}(f(X_i) - Y_i)^2 - \frac{1}{m}\sum_{i\in B_j}(h(X_i) - Y_i)^2 = \mathbb{Q}_{f,h}(j) + \mathbb{M}_{f,h}(j)\,.$$

Note that $\mathbb{E}\mathbb{B}_{f,h}(j) = \mathbb{E}(f(X) - Y)^2 - \mathbb{E}(h(X) - Y)^2$. Therefore, at least intuitively, if $\mathbb{B}_{f,h}(j) > 0$ for a majority of indices $1 \leq j \leq k$, one would expect that $\mathbb{E}(f(X) - Y)^2 > \mathbb{E}(h(X) - Y)^2$, making h a better candidate to be a risk minimizer than f.

When one is given a sample $(X_i, Y_i)_{i=1}^{3n}$, the choice of $\widehat{f}$ is carried out as follows:

Step 1:

- Fix $r > 0$, corresponding to the desired accuracy parameter $\epsilon \sim r^2$.
- Let $\widehat{\Phi}_n$ be a distance oracle in $\mathcal{F}$ similar to the one described in the previous section, which uses as data the first part of the sample $(X_i)_{i=1}^n$. Thus, for the right choice of parameters α and β and with high probability the following holds: if $f, h \in \mathcal{F}$ and $\widehat{\Phi}_n(f,h) \geq \beta r$ then $\|f-h\|_{L_2} \sim_{\alpha,\beta} \widehat{\Phi}_n(f,h)$, and if $\widehat{\Phi}_n(f,h) \leq \beta r$ then $\|f-h\|_{L_2} \leq (\beta/\alpha) r$.

  Define $\mathcal{DO}(f,h) = 1$ if $\widehat{\Phi}_n(f,h) \geq \beta r$ and $\mathcal{DO}(f,h) = 0$ otherwise.

The binary valued functional $\mathcal{DO}$ serves as the 'referee' of the tournament. Its role is to decide when a match between two functions is allowed to take place. In a more mathematical language, when $\mathcal{DO}(f,h) = 1$ one has a good reason to expect that f and h are far enough to ensure that $(\mathbb{B}_{f,h}(j))_{j=1}^k$ reflects the true value $\mathbb{E}(f(X)-Y)^2 - \mathbb{E}(h(X)-Y)^2$.

Step 2:

- This round of the tournament consists of statistical matches between class members which are preformed using the second part of the sample $(X_i, Y_i)_{i=n+1}^{2n}$. A match is allowed to proceed only if $\mathcal{DO}(f,h) = 1$; otherwise, the match is drawn. If a match does take place then h defeats f if $\mathbb{B}_{f,h}(j) > 0$ for a majority of indices j, and f defeats h if the reverse inequality holds for a majority of the blocks.
- A function f qualifies from this round if it has has won or drawn all of its matches.

The crucial fact behind Step 2 is that, with high probability, the risk minimizer f^* qualifies for the next round: if $\mathcal{DO}(h, f^*) = 1$ then h and f^* are far enough to ensure that $(\mathbb{B}_{h,f^*}(j))_{j=1}^k$ reflects the true value $\mathbb{E}(h(X)-Y)^2 - \mathbb{E}(f^*(X)-Y)^2$. Since f^* is the unique minimizer of the risk, the majority of values are positive.

Moreover, the same argument implies that if h is a qualifier from Step 2, then $\|h-f^*\|_{L_2} \leq \beta r$. Indeed, the match between h and f^* (or between any two qualifiers) must have been drawn by the referee's decision; thus h must be 'close' to f^*.

Step 2 is not enough to identify a function with a small excess risk. Indeed, all the qualifiers are close to f^*, but the fact that $\|f-f^*\|_{L_2} \leq \beta r$ does not imply that $\mathbb{E}(f(X)-Y)^2 - \mathbb{E}(f^*(X)-Y)^2 \lesssim r^2$. Therefore, the tournament has an additional step: the *Champions League* round, in which all the qualifiers play each other in a

different type of match.

To find a function that does have an almost optimal risk one uses the third part of the sample $(X_i, Y_i)_{i=2n+1}^{3n}$ to define a 'home and away' style matches:

Step 3:

- Let $\Psi_{h,f} = (h(X) - f(X))(f(X) - Y)$ and set $\Psi_{h,f}(j) = \frac{1}{m}\sum_{i \in B_j} \Psi_{h,f}(X_i, Y_i)$. Let α, β and r be as above and put $r_1 = 2(\beta/\alpha)r$.
- f wins its home match against h if $\Psi_{h,f}(j) \geq -r_1^2/10$ for a majority of the indices j.
- A winner of the tournament is any qualifier that wins all of its home matches. We set $\widehat{f}$ to be any such winner.

To see the reason behind this choice of matches, recall that all the qualifiers h satisfy that $\|h - f^*\|_{L_2} \leq \beta r$. At the same time, the excess risk of h is

$$\mathbb{E}(h(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2 = \|h - f^*\|_{L_2}^2 + 2\mathbb{E}(h(X) - f^*(X)) \cdot (f^*(X) - Y) \, .$$

Since $\|h - f^*\|_{L_2}^2$ is of the order of r^2 it is evident that if $\mathbb{E}(h(X) - f^*(X)) \cdot (f^*(X) - Y) \lesssim r^2$, then the excess risk of h is also of the order of r^2.

Observe that $\mathbb{E}\Psi_{h,f^*} = \mathbb{E}(h(X) - f^*(X)) \cdot (f^*(X) - Y)$ and that by the convexity of $\mathcal{F}$, $\mathbb{E}\Psi_{h,f^*} \geq 0$ (this follows from the characterization of the nearest point map onto a closed, convex subset of a Hilbert space). Moreover,

$$\mathbb{E}\Psi_{h,f^*} = -\|h - f^*\|_{L_2}^2 - \mathbb{E}\Psi_{f^*,h} \, . \tag{5.1}$$

One shows that $\Psi_{h,f^*}(j) \gtrsim -r^2$ for a majority of indices j. This follows because the median of $(\Psi_{h,f^*}(j))_{j=1}^k$ happens to be a uniform median-of-means estimator of the true mean $\mathbb{E}\Psi_{h,f^*}$. As a consequence, f^* wins all of its home matches. Also, if h wins a home match against f^*, (i.e., $\Psi_{f^*,h}(j) \gtrsim -r^2$ for a majority of indices j), then $\mathbb{E}\Psi_{f^*,h} \gtrsim -r^2$ and by (5.1), $\mathbb{E}\Psi_{h,f^*} \lesssim r^2$. That implies that every function that wins all of its home matches must have a small excess risk.

To conclude, all three components of the tournament procedure from [59] are derived using uniform median-of-means estimators (of different functionals) in the class $\mathcal{F}$.

Without going into technical details, at the heart of the analysis of Steps 2 and 3 of the tournament is the following fact: given a convex class $\mathcal{F}$ that satisfies some minimal conditions, for the right choice of k and r (the choice of r depends on the geometry of the class $\mathcal{F}$ and on the parameters γ_1 and γ_2 appearing below), and for an absolute constant c_1, we have that, with probability $1 - 2\exp(-c_1 k)$,

(1) for every $f \in \mathcal{F}$ such that $\|f - f^*\|_{L_2} \geq r$, one has

$$\mathbb{B}_{f,f^*}(j) \geq \gamma_1 \|f - f^*\|_{L_2}^2$$

for $0.99k$ of the blocks;

(2) for every $f \in \mathcal{F}$ such that $\|f - f^*\|_{L_2} < r$, one has

$$|\mathbb{M}_{f,f^*}(j) - \mathbb{E}\mathbb{M}_{f,f^*}(j)| \leq \gamma_2 r^2$$

for $0.99k$ of the blocks.

These facts suffice for proving the validity of steps (2) and (3) in the tournament procedure. A general bound for the performance of the procedure defined above was proven by Lugosi and Mendelson [59]. The achievable accuracy depends on the interaction between the geometry of the class $\mathcal{F}$ and the distribution of (X, Y). Instead of recalling the technical details in their full generality, we simply illustrate the performance on the canonical example of linear regression.

Let $\mathcal{F} = \{\langle t, \cdot \rangle : t \in \mathbb{R}^d\}$ be the class of linear functionals on $\mathbb{R}^d$. Let X be an isotropic random vector in $\mathbb{R}^d$ (i.e., $\mathbb{E}\langle t, X\rangle^2 = 1$ for every t in the Euclidean unit sphere) and assume that the distribution of X is such that there are $q > 2$ and $L > 1$ for which, for every $t \in \mathbb{R}^d$, $\|\langle X, t\rangle\|_{L_q} \leq L\|\langle X, t\rangle\|_{L_2}$.

Assume that one is given n noisy measurements of $\langle t_0, \cdot \rangle$ for a fixed but unknown $t_0 \in \mathbb{R}^d$, that is, assume that $Y = \langle t_0, X\rangle + W$ for some symmetric random variable W that is independent of X and has variance σ^2. One observes the "noisy" data $(X_i, Y_i)_{i=1}^n$ and the aim is to approximate t_0 with a small error (accuracy) and with high probability (confidence).

Invoking standard methods as in [50], the best that one can guarantee using empirical risk minimization is a choice of $\widehat{t} \in \mathbb{R}^d$, for which

$$\|\widehat{t} - t_0\|_2 \leq \frac{C}{\delta}\sigma\sqrt{\frac{d}{n}} \quad \text{with probability } 1 - \delta - 2\exp(-c_1 d)$$

for some constant C that depends on q and L. Therefore, if one wishes for an error that is proportional to $\sigma\sqrt{d/n}$, the best that one can hope for is a constant confidence δ.

The median-of-means tournament procedure, when applied to this example, selects $\widehat{t}$ for which

$$\|\widehat{t} - t_0\|_2 \leq C\sigma\sqrt{\frac{d}{n}} \quad \text{with probability } 1 - 2\exp(-cd)$$

for some numerical constants $c, C > 0$. As it is argued in [59] that this is the optimal confidence at any level that is proportional to $\sqrt{d/n}$. In fact, the median-of-means tournament procedure gives the optimal confidence for any accuracy $r \geq c'\sigma\sqrt{d/n}$. Standard empirical risk minimization can only achieve such accuracy/confidence tradeoff for sub-Gaussian distributions.

Acknowledgements. We thank Sam Hopkins, Stanislav Minsker, and Roberto Imbuzeiro Oliveira for illuminating discussions on the subject. We also thank two referees for their thorough reports and insightful comments.

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] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[4] J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39:2766–2794, 2011.

[5] Y. Baraud and L. Birgé. Rho-estimators revisited: General theory and applications. *The Annals of Statistics*, 46(6B):3767–3804, 2018.

[6] Y. Baraud, L. Birgé, and M. Sart. A new method for estimation and model selection: ρ-estimation. *Inventiones Mathematicae*, 207(2):425–517, 2017.

[7] P.L. Bartlett, O. Bousquet, and S. Mendelson. Localized Rademacher complexities. *Annals of Statistics*, 33:1497–1537, 2005.

[8] P.J. Bickel. On some robust estimates of location. *The Annals of Mathematical Statistics*, 36:847–858, 1965.

[9] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36:929–965, 1989.

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

[11] C. Brownlees, E. Joly, and G. Lugosi. Empirical risk minimization for heavy-tailed losses. *Annals of Statistics*, 43:2507–2536, 2015.

[12] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59:7711–7717, 2013.

[13] P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[14] 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.

[15] O. Catoni and I. Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017.

[16] O. Catoni and I. Giulini. Dimension-free pac-bayesian bounds for the estimation of the mean of a random vector. *arXiv preprint arXiv:1802.04308*, 2018.

[17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60. ACM, 2017.

[18] Y. Cherapanamjeri, N. Flammarion, and P. Bartlett. Fast mean estimation with sub-gaussian rates. *arXiv preprint arXiv:1902.01998*, 2019.

[19] M. Chichignoud and J. Lederer. A robust, adaptive m-estimator for pointwise estimation in heteroscedastic regression. *Bernoulli*, 20(3):1560–1599, 2014.

[20] 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.

[21] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

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

[23] I. Diakonikolas, G. Kamath, D.M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 655–664. IEEE, 2016.

[24] I. Diakonikolas, G. Kamath, D.M. Kane, J. Li, A. Moitra, and A. Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, 2017.

[25] I. Diakonikolas, G. Kamath, D.M. Kane, J. Li, A. Moitra, and A. Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. Society for Industrial and Applied Mathematics, 2018.

[26] I. Diakonikolas, D.M. Kane, and A. Stewart. Efficient robust proper learning of log-concave distributions. *arXiv preprint arXiv:1606.03077*, 2016.

[27] I. Diakonikolas, W. Kong, and A. Stewart. Efficient algorithms and lower bounds for robust linear regression. *arXiv preprint arXiv:1806.00040*, 2018.

[28] J. Fan, Q. Li, and Y. Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):247–265, 2017.

[29] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer-Verlag, New York, 2002.

[30] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 1986.

[31] Q. Han and J.A. Wellner. A sharp multiplier inequality with applications to heavy-tailed regression problems. *arXiv preprint arXiv:1706.02410*, 2017.

[32] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[33] S.B. Hopkins. Sub-gaussian mean estimation in polynomial time. *Annals of Statistics*, 2019, to apear.

[34] S.B. Hopkins and J. Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034. ACM, 2018.

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

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

[37] M. Huber. An optimal (ϵ, δ)-randomized approximation scheme for the mean of random variables with bounded relative variance. *Random Structures & Algorithms*, 2019.

[38] P.J. Huber. Robust estimation of a location parameter. *The annals of mathematical statistics*, 35(1):73–101, 1964.

[39] P.J. Huber and E.M. Ronchetti. *Robust statistics*. Wiley, New York, 2009. Second edition.

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

[41] E. Joly, G. Lugosi, and R. I. Oliveira. On the estimation of the mean of a random vector. *Electronic Journal of Statistics*, 11:440–451, 2017.

[42] A. Klivans, P.K. Kothari, and R. Meka. Efficient algorithms for outlier-robust regression. In *Proceedings of the 31st Annual Conference of Learning Theory (COLT 2018)*, 2018.

[43] V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

[44] P.K. Kothari, J. Steinhardt, and D. Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046. ACM, 2018.

[45] Kevin A. Lai, Anup B. Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 665–674. IEEE, 2016.

[46] G. Lecué and M. Lerasle. Learning from mom's principles: Le cam's approach. *arXiv preprint arXiv:1701.01961*, 2017.

[47] G. Lecué and M. Lerasle. Robust machine learning by median-of-means: theory and practice. *Annals of Stastistics*, 2019, to appear.

[48] G. Lecué, M. Lerasle, and T. Mathieu. Robust classification via mom minimization. *arXiv preprint arXiv:1808.03106*, 2018.

[49] G. Lecué and S. Mendelson. Learning subgaussian classes: Upper and minimax bounds. In S. Boucheron and N. Vayatis, editors, *Topics in Learning Theory*. Societe Mathematique de France, 2016.

[50] G. Lecué and S. Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016.

[51] M. Ledoux. *The concentration of measure phenomenon*. American Mathematical Society, Providence, RI, 2001.

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

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

[54] Po-Ling Loh and Xin Lu Tan. High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electronic Journal of Statistics*, 12(1):1429–1467, 2018.

[55] G. Lugosi and S. Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *manuscript*, 2019.

[56] G. Lugosi and S. Mendelson. Sub-Gaussian estimators of the mean of a random vector. *Annals of Statistics*, 47:783–794, 2019.

[57] G. Lugosi and S. Mendelson. Near-optimal mean estimators with respect to general norms. *Probability Theory and Related Fields*, 2019, to appear.

[58] G. Lugosi and S. Mendelson. Regularization, sparse recovery, and median-of-means tournaments. *Bernoulli*, 2019, to appear.

[59] G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 2019, to appear.

[60] P. Massart. *Concentration inequalities and model selection*. Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics. Springer, 2006.

[61] S. Mendelson. Learning without concentration. *Journal of the ACM*, 62:21, 2015.

[62] S. Mendelson. An optimal unrestricted learning procedure. *arXiv preprint arXiv:1707.05342*, 2017.

[63] S. Mendelson. Learning without concentration for general loss functions. *Probability Theory and Related Fields*, 171(1-2):459–502, 2018.

[64] S. Mendelson and N. Zhivotovskiy. Robust covariance estimation under $L_4 - L_2$ norm equivalence. *arXiv preprint arXiv:1809.10462*, 2018.

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

[66] Stanislav Minsker. Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903, 2018.

[67] Stanislav Minsker. Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523*, 2018.

[68] Stanislav Minsker and Nate Strawn. Distributed statistical estimation and rates of convergence in normal approximation. *arXiv preprint arXiv:1704.02658*, 2017.

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

[70] Roberto I. Oliveira and Paulo Orenstein. The sub-gaussian property of trimmed means estimators. Technical report, IMPA, 2019.

[71] Valentin V Petrov. Limit theorems of probability theory: sequences of independent random variables. Technical report, Oxford, New York, 1995.

[72] IG Shevtsova. On the absolute constants in the Berry-Esseen-type inequalities. In *Doklady Mathematics*, volume 89, pages 378–381. Springer, 2014.

[73] C.G. Small. A survey of multidimensional medians. *International Statistical Review*, pages 263–277, 1990.

[74] S.M. Stigler. The asymptotic distribution of the trimmed mean. *The Annals of Statistics*, 1:472–477, 1973.

[75] B.S. Tsirelson, I.A. Ibragimov, and V.N. Sudakov. Norm of Gaussian sample function. In *Proceedings of the 3rd Japan-U.S.S.R. Symposium on Probability Theory*, volume 550 of *Lecture Notes in Mathematics*, pages 20–41. Springer-Verlag, Berlin, 1976.

[76] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

[77] J.W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.

[78] J.W. Tukey and D.H. McLaughlin. Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/winsorization 1. *Sankhyā: The Indian Journal of Statistics, Series A*, 25:331–352, 1963.

[79] L.G. Valiant. A theory of the learnable. *Communications of the ACM*, 27:1134–1142, 1984.

[80] S. van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.

[81] A.W. van der Waart and J.A. Wellner. *Weak convergence and empirical processes*. Springer, 1996.

[82] V.N. Vapnik and A.Ya. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974. (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979.

[83] R. Vershynin. Lectures in geometric functional analysis. 2009.