The Annals of Statistics
2021, Vol. 49, No. 1, 393–410
https://doi.org/10.1214/20-AOS1961
© Institute of Mathematical Statistics, 2021

ROBUST MULTIVARIATE MEAN ESTIMATION: THE OPTIMALITY OF TRIMMED MEAN

BY GÁBOR LUGOSI[1] AND SHAHAR MENDELSON[2]

[1]*ICREA; Department of Economics and Business, Pompeu Fabra University; Barcelona Graduate School of Economics*
gabor.lugosi@upf.edu

[2]*Mathematical Sciences Institute, Australian National University, shahar.mendelson@gmail.com*

We consider the problem of estimating the mean of a random vector based on i.i.d. observations and adversarial contamination. We introduce a multivariate extension of the trimmed-mean estimator and show its optimal performance under minimal conditions.

1. Introduction. Estimating the mean of a random vector based on independent and identically distributed samples is one of the most basic statistical problems. In the last few years, the problem has attracted a lot of attention and important advances have been made both in terms of statistical performance and computational methodology.

In the simplest form of the mean estimation problem, one wishes to estimate the expectation $\mu = \mathbb{E}X$ of a random vector X taking values in $\mathbb{R}^d$, based on a sample $X_1, \dots, X_N$ consisting of independent copies of X. An *estimator* is a (measurable) function of the data

$$\widehat{\mu} = \widehat{\mu}(X_1, \dots, X_N) \in \mathbb{R}^d.$$

We measure the quality of an estimator by the distribution of its Euclidean distance to the mean vector μ. More precisely, for a given $\delta > 0$—the *confidence parameter*—one would like to ensure that

$$\|\widehat{\mu} - \mu\| \le \varepsilon(N, \delta) \quad \text{with probability at least } 1 - \delta$$

with $\varepsilon(N, \delta)$ as small as possible. Here and in the entire article, $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$.

The obvious choice of $\widehat{\mu}$ is the empirical mean $N^{-1}\sum_{i=1}^N X_i$, which, apart from its computational simplicity, has good statistical properties when the distribution is sufficiently well behaved. However, it is well known that, even when X is real valued, the empirical mean behaves suboptimally and much better mean estimators are available.[1] The reason for the suboptimal performance of the empirical mean is the damaging effect of *outliers* that are inevitably present when the distribution is heavy-tailed.

Informally put, outliers are sample points that are, in some sense, atypical; as a result they cause a significant distortion to the empirical mean. The crucial fact is that when X is a heavy-tailed random variable, a typical sample contains a significant number of outliers, implying the empirical mean is likely to be distorted.

To exhibit the devastating effect that outliers cause, let $\varepsilon > 0$ and note that there is a square integrable (univariate) random variable X such that

$$\left|\frac{1}{N}\sum_{i=1}^N X_i - \mu\right| \ge \varepsilon \quad \text{with probability at least } c\frac{\sigma_X^2}{\varepsilon^2 N}$$

Received July 2019; revised February 2020.
MSC2020 subject classifications. Primary 62J02, 62G08; secondary 60G25.
Key words and phrases. Mean estimation, robust estimation.

[1]We refer the reader to the recent survey [20] for an extensive discussion.

for a positive absolute constant c; σ_X^2 is the variance of X. In other words, the best possible error $\varepsilon(N,\delta)$ that can be guaranteed by the empirical mean (when only finite variance is assumed) is of the order of $\sigma_X/\sqrt{\delta N}$. On the other hand, it is well known (see, e.g., the survey [20]) that there are estimators of the mean $\widehat{\mu}$ such that for all square-integrable random variables X,

$$|\widehat{\mu}-\mu| \le c\sigma_X\sqrt{\frac{\log(2/\delta)}{N}} \quad \text{with probabilty } 1-\delta, \tag{1.1}$$

where c is a suitable absolute constant. An estimator that performs with an error $\varepsilon(N,\delta)$ of the order of $\sigma_X\sqrt{\log(2/\delta)/N}$ is called a sub-Gaussian estimator. Such estimators are optimal in the sense that no estimator can perform with a better error $\varepsilon(N,\delta)$ even if X is known to be a Gaussian random variable.

Because the empirical mean is such a simple estimator and seeing that outliers are the probable cause of its suboptimality, for real-valued random variables, a 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 γN largest and smallest points for some parameter $\gamma \in (0,1)$, 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 [28], Huber and Ronchetti [15], Bickel [1] and Stigler [26] for early work on the theoretical properties of the trimmed-mean estimator. However, the nonasymptotic sub-Gaussian property of the trimmed mean was established only recently, by Oliveira and Orenstein in [23]. They proved that if $\gamma = \kappa\log(1/\delta)/N$ for a constant κ, then the trimmed mean estimator $\widehat{\mu}$ satisfies (1.1) for all distributions with a finite variance σ_X and with a constant c that depends on κ only.

An added value of the trimmed mean is that it seems to be robust to malicious noise, at least intuitively. Indeed, assume that an adversary can corrupt ηN of the N points for some $\eta < 1$. The trimmed-mean estimator can withstand at least one sort of contamination: the adversary making the corrupted points either very large or very small. This does not rule out other damaging changes to the sample, but at least it gives the trimmed mean another potential edge over other estimators. And, in fact, as we prove in this article, the performance of the trimmed-mean estimator is as good as one can hope for under both heavy-tailed distributions and adversarial corruption. We show that—a simple variant of—the trimmed-mean estimator achieves

$$|\widehat{\mu}-\mu| \le c\sigma_X\left(\sqrt{\eta}+\sqrt{\frac{\log(1/\delta)}{N}}\right) \tag{1.2}$$

with probability $1-\delta$, for an absolute constant c (see Theorem 1 for the detailed statement). The bound (1.2) holds for all univariate distributions with a finite variance, and is minimax optimal in that class of distributions. For distributions with lighter tail, the dependence on the contamination level η can be improved. For example, for sub-Gaussian distributions $\sqrt{\eta}$ may be replaced by $\eta\sqrt{\log(1/\eta)}$ and the trimmed-mean estimator achieves that. As we explain in what follows, the parameter γ that determines the level of trimming depends on the confidence parameter δ and contamination level η only.

The problem of mean estimation in the multivariate case (i.e., when X takes values in $\mathbb{R}^d$ for some $d>1$) is considerably more complex. For i.i.d. data without contamination, the best possible statistical performance for square-integrable random vectors is well understood: if $\Sigma = \mathbb{E}[(X-\mu)(X-\mu)^T]$ is the covariance matrix of X whose largest eigenvalue and trace are denoted by λ_1 and $\operatorname{Tr}(\Sigma)$, respectively, then for every $\delta>0$, there exists a mean estimator $\widehat{\mu}$ such that, regardless of the distribution, with probability at least $1-\delta$,

$$\|\widehat{\mu}-\mu\| \le c\left(\sqrt{\frac{\operatorname{Tr}(\Sigma)}{N}}+\sqrt{\frac{\lambda_1\log(1/\delta)}{N}}\right) \tag{1.3}$$

for some absolute constant c. This bound is optimal in the sense that one cannot improve it even when the distribution is known to be Gaussian. The existence of such a "sub-Gaussian" estimator was established by Lugosi and Mendelson [21]. Computationally efficient versions have been subsequently constructed by Hopkins [14] and by Cherapanamjeri, Flammarion and Bartlett [5]; see also Depersin and Lecué [7]. Once again, we refer to the survey [20] for related results.

A natural question is how well one can estimate the mean of a random vector in the presence of adversarial contamination. In particular, one may ask the following.

Let X be a random vector in $\mathbb{R}^d$ whose mean and covariance matrix exist. Let $X_1, \dots, X_N$ be i.i.d. copies of X. Then the adversary, maliciously (and knowing in advance of statistician's intentions), is free to change at most ηN of the sample points. How accurately can $\mu = \mathbb{E}X$ be estimated with respect to the Euclidean norm? In particular, given δ and η, does there exist an estimator and an absolute constant c such that, regardless of the distribution of X, with probability at least $1 - \delta$,

$$\|\widehat{\mu} - \mu\| \le c\left(\sqrt{\frac{\operatorname{Tr}(\Sigma)}{N}} + \sqrt{\frac{\lambda_1 \log(1/\delta)}{N}} + \sqrt{\lambda_1 \eta}\right)? \tag{1.4}$$

The main result of this article, Theorem 2, answers this question in the affirmative. To that end, we construct a procedure, based on the one-dimensional trimmed-mean estimator, that has the desired performance guarantees.

Related work. The model of estimation under adversarial contamination has been extensively addressed in the literature of computational learning theory. Its origins may be traced back to the malicious noise model of Valiant [29] and Kearns and Li [16]. In the context of mean estimation, it has been investigated by Diakonikolas, Kamath, Kane, Li, Moitra and Stewart [9–11], Steinhardt, Charikar and Valiant [25] and Minsker [22]. In particular, in [10] it is shown that when $N = \Omega((d/\eta)\log d)$ and λ_1 is the largest eigenvalue of the covariance matrix Σ of X, then there exists a computationally efficient estimator of the mean that satisfies

$$\|\widehat{\mu} - \mu\| \le c\sqrt{\lambda_1 \eta}$$

with probability at least $9/10$ for all distributions. Although this bound is suboptimal in terms of the conditions and does not recover the sub-Gaussian bounds, the goal in [10], and in other articles in this direction as well, was mainly on computational efficiency. In contrast, our aim is to construct an estimator with optimal statistical performance, and the multivariate estimator we propose is not computationally feasible—at least in its naive implementation—in the sense that computing the estimator takes time that is exponential in the dimension. It is an intriguing problem to find computationally efficient mean estimators that have optimal statistical performance under the weakest possible assumptions: although such estimators are available for i.i.d. data from the results of Hopkins [14] and Cherapanamjeri, Flammarion and Bartlett [5], these estimators are not expected to perform well under adversarial contamination.

The sub-Gaussian estimators achieving the bound (1.3) are based on median-of-means estimators. Such estimators have been studied under a (somewhat more restrictive) adversarial contamination model by Lecué and Lerasle [17] and by Minsker [22]; see also see Rodriguez and Valdora [24]. In particular, Minsker [22] studies estimators that cleverly combine Huber's robust M-estimators with the median-of-means technique. His results imply a performance bound exactly of the form of (1.4). A disadvantage of Minsker's estimator is that it assumes that the trace and operator norm of the covariance matrix are known up to a constant factor.

In a recent manuscript, Depersin and Lecué [7] study the problem of robust mean estimation a slightly more restrictive model of contamination. Their main result is a computationally efficient multivariate mean estimator that achieves a performance similar to (1.4), though only when η is at most a small constant times $\log(1/\delta)/N$; thus, it is only able to handle low levels of contamination.

Chen, Gao and Ren [3] develop a general theory of minimax bounds under Huber's contamination model (i.e., when the contamination is i.i.d.) for parametric families of distributions. In [4] the same authors study robust estimation of the mean vector and covariance matrix under Huber's contamination model and derive sharp minimax bounds for Gaussian, and more generally elliptical, distributions. In particular, they show that if the uncontaminated data is Gaussian with identity covariance matrix, then Tukey's median $\widehat{\mu}$ satisfies that, with probability at least $1-\delta$,

$$\|\widehat{\mu}-\mu\| \le c\biggl(\sqrt{\frac{d}{N}}+\sqrt{\frac{\log(1/\delta)}{N}}+\eta\biggr).$$

Moreover, they prove that this estimator is minimax optimal up to constant factors. Note that (1.4) has a similar form except that the term η is replaced by the weaker $\sqrt{\eta}$. It is remarkable that this is the only (necessary) price one has to pay for moving from Gaussian distributions to arbitrary ones whose covariance matrix exists and from Huber's contamination to adversarial one. Moreover, as we argue below, for sub-Gaussian distributions the term $\sqrt{\eta}$ may be improved to $\eta\sqrt{\log(1/\eta)}$. We also refer to Dalalyan and Thompson [6] for recent related work.

The rest of the article is organized as follows. In Section 2, we discuss the univariate case and establish a performance bound for a version of the trimmed-mean estimator in Theorem 1. We argue that this bound is best possible up to the value of the absolute constant. In Section 3, we extend the discussion to the multivariate case, and construct a new estimator. The proof of the performance bound of the multivariate estimator is given in Section 4.

2. The real-valued case. Let X be a real-valued random variable that has finite variance σ_X^2. Set $\mu=\mathbb{E}X$ and define $\overline{X}=X-\mu$. In what follows, c, C denote positive absolute constants whose value may change at each appearance. For $0<p<1$, define the quantile

$$Q_p(\overline{X}) = \sup\bigl\{M\in\mathbb{R}:\mathbb{P}(\overline{X}\ge M)\ge 1-p\bigr\}. \tag{2.1}$$

For simplicity of presentation, we assume throughout the article that X has an absolutely continuous distribution. Under this assumption, it follows that $\mathbb{P}(\overline{X}\ge Q_p(\overline{X}))=1-p$. However, we emphasize that this assumption is not restrictive; one may easily adjust the proof to include all distributions with a finite second moment. Another solution is that the statistician can always add a small independent Gaussian noise to the sample points, thus ensuring that the distribution has a density and without affecting statistical performance.

For reasons of comparison, our starting point is a simple lower bound that limits the performance of every mean estimator. Similar arguments appear in [10] and [22].

While the adversary has total freedom to change at most ηN of the sample points, consider first a rather trivial action: changing the i.i.d. sample $(\mathcal{X}_i)_{i=1}^N$ to $(\widetilde{X}_i)_{i=1}^N$ defined by

$$\widetilde{X}_i = \min\bigl\{X_i,\mu+Q_{1-\eta/2}(\overline{X})\bigr\}. \tag{2.2}$$

Since

$$\mathbb{P}\bigl(\overline{X}\ge Q_{1-\eta/2}(\overline{X})\bigr)=\frac{\eta}{2},$$

by a binomial tail bound, with probability at least $1-2\exp(-c\eta N)$,

$$\bigl|\bigl\{i:X_i-\mu\ge Q_{1-\eta/2}(\overline{X})\bigr\}\bigr|\le\frac{3}{4}\eta N.$$

In particular, on this event, the adversary can change all sample points X_i that are bigger than $\mu + Q_{1-\eta/2}(\overline{X})$. As a result, there is no way one can determine whether $(\widetilde{X}_i)_{i=1}^N$ is a corrupted sample, originally selected according to X and then changed as in (2.2), or an uncorrupted sample selected according to the random variable

$$Z = \min\{X, \mu + Q_{1-\eta/2}(\overline{X})\}.$$

Therefore, on this event, no procedure can distinguish between $\mathbb{E}X$ and $\mathbb{E}Z$, which means that the error caused by this action is at least $|\mathbb{E}Z - \mu|$. Note that for $M = Q_{1-\eta/2}(\overline{X})$ one has that

$$|\mathbb{E}Z - \mu| = \mathbb{E}\big[(\overline{X} - M)\mathbb{1}_{\overline{X} \geq M}\big].$$

Since the adversary can target the lower tail of X in exactly the same way, it follows that, with probability at least $1 - 2\exp(-c\eta N)$, no estimator can perform with accuracy better than

$$\overline{\mathcal{E}}(\eta, X) \stackrel{\text{def.}}{=} \max\big\{\mathbb{E}\big[|\overline{X} - Q_{\eta/2}(\overline{X})|\mathbb{1}_{\overline{X} \leq Q_{\eta/2}(\overline{X})}\big], \mathbb{E}\big[|\overline{X} - Q_{1-\eta/2}(\overline{X})|\mathbb{1}_{\overline{X} \geq Q_{1-\eta/2}(\overline{X})}\big]\big\}.$$

Of course, the adversary has a second trivial action: do nothing. That is a better corruption strategy (in the minimax sense) when

$$\overline{\mathcal{E}}(\eta, X) \leq C\sigma_X\sqrt{\frac{\log(2/\delta)}{N}}.$$

Therefore, if one wishes to find a procedure that performs with probability at least $1 - \delta - 2\exp(-c\eta N)$, the best error one can hope for is

$$\overline{\mathcal{E}}(\eta, X) + C\sigma_X\sqrt{\frac{\log(2/\delta)}{N}}, \tag{2.3}$$

where c and C are absolute constants.

A rather surprising fact is that in the real-valued case, the two trivial actions cause the largest possible damage. Indeed, we show that there is an estimator that is a simple modification of trimmed mean that attains what is almost the optimal error—with $\overline{\mathcal{E}}(\eta, X)$ replaced by

$$\mathcal{E}(\eta, X) \stackrel{\text{def.}}{=} \max\big\{\mathbb{E}\big[|\overline{X}|\mathbb{1}_{\overline{X} \leq Q_{\eta/2}(\overline{X})}\big], \mathbb{E}\big[|\overline{X}|\mathbb{1}_{\overline{X} \geq Q_{1-\eta/2}(\overline{X})}\big]\big\}.$$

REMARK. It is straightforward to construct a random variable X for which $\overline{\mathcal{E}}(\eta, X) \geq c_1\sqrt{\eta}\sigma_X$. (Take, e.g., X that takes value 0 with probability $1 - \eta$ and values $\pm\sigma_X/\sqrt{\eta}$ with probability $\eta/2$ each.) Thus, in terms of η, σ_X, δ and N, the best minimax error rate that is possible in the corrupted mean estimation problem for real-valued random variables is

$$c\sigma_X \max\left\{\sqrt{\eta}, \sqrt{\frac{\log(2/\delta)}{N}}\right\}$$

for a suitable absolute constant c.

Next, let us define the modified trimmed-estimator. The estimator splits the data into two equal parts. Half of the data points are used to determine the truncation at the appropriate level. The points from the other half are averaged as is, except for the data points that fall outside of the estimated quantiles, which are truncated prior to averaging. For convenience, assume that the data consists of $2N$ independent copies of the random variable

X, denoted by $X_1, \dots, X_N, Y_1, \dots, Y_N$. The statistician has access to the corrupted sample $\widetilde{X}_1, \dots, \widetilde{X}_N, \widetilde{Y}_1, \dots, \widetilde{Y}_N$, where at most $2\eta N$ of the sample points have been changed by an adversary.

For $\alpha \le \beta$, let

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

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

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

Univariate mean estimator.

(1) Consider the corrupted sample $\widetilde{X}_1, \dots, \widetilde{X}_N, \widetilde{Y}_1, \dots, \widetilde{Y}_N$ as input.
(2) Given the corruption parameter η and confidence level δ, set

$$\varepsilon = 8\eta + 12\frac{\log(4/\delta)}{N}.$$

(3) Let $\alpha = \widetilde{Y}^*_{\varepsilon N}$ and $\beta = \widetilde{Y}^*_{(1-\varepsilon)N}$ and set

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \phi_{\alpha,\beta}(\widetilde{X}_i).$$

THEOREM 1. *Let $\delta \in (0,1)$ be such that $\delta \ge e^{-N}/4$. Then, with probability at least $1-\delta$,*

$$|\widehat{\mu} - \mu| \le 3\mathcal{E}(4\varepsilon, X) + 2\sigma_X\sqrt{\frac{\log(4/\delta)}{N}}.$$

Moreover, with probability at least $1 - 4\exp(-\varepsilon N/12)$,

$$|\widehat{\mu} - \mu| \le 10\sqrt{\varepsilon}\sigma_X.$$

REMARK. The necessity of prior knowledge of the confidence parameter δ was pointed out (even in the contamination-free case) by Devroye, Lerasle, Lugosi and Oliveira [8]; see [20] for further discussion. The contamination level need not be known exactly. If an upper bound $\overline{\eta} \ge \eta$ is available and one uses the estimator with parameter $\overline{\eta}$ instead of η, then the same bound holds with η replaced by $\overline{\eta}$.

To explain the meaning of Theorem 1, observe that for $M = Q_{1-\varepsilon/2}(\overline{X})$, one has

$$\frac{\varepsilon}{2} = \mathbb{P}(\overline{X} \ge M) \le \frac{\sigma_X^2}{M^2},$$

and in particular,

$$Q_{1-\varepsilon/2}(\overline{X}) \le \frac{\sigma_X\sqrt{2}}{\sqrt{\varepsilon}}. \tag{2.4}$$

Also,

$$\begin{aligned} \mathbb{E}\big[(\overline{X} - M)\mathbb{1}_{\overline{X}\ge M}\big] &\le \mathbb{E}\big[|\overline{X}|\mathbb{1}_{\overline{X}\ge M}\big] + \mathbb{E}[M\mathbb{1}_{\overline{X}\ge M}] \\ &\le \sigma_X\mathbb{P}^{1/2}(\overline{X} \ge M) + |M|\mathbb{P}(\overline{X} \ge M) \\ &\le \sigma_X\sqrt{8\varepsilon}, \end{aligned} \tag{2.5}$$

implying that for every X,

$$\mathcal{E}(\varepsilon, X) \le \sigma_X \sqrt{8\varepsilon}. \tag{2.6}$$

Hence, Theorem 1 shows that the estimator attains the minimax rate of the corrupted mean-estimation problem, noted previously.

Of course, Theorem 1 actually implies sharper individual bounds: if $\eta N \le \log(2/\delta)$, then $\varepsilon \sim N^{-1}\log(2/\delta)$ and the assertion of Theorem 1 is that, with probability at least $1-\delta$,

$$|\widehat{\mu} - \mu| \le C\sigma_X \sqrt{\frac{\log(2/\delta)}{N}},$$

which matches the optimal sub-Gaussian error rate. If, on the other hand, $\eta N > \log(2/\delta)$, then with probability at least $1-\delta$,

$$|\widehat{\mu} - \mu| \le C\mathcal{E}(c\eta, X),$$

essentially matching the lower bound (2.3).

REMARK. Observe that the upper bound on $\mathcal{E}(\varepsilon, X)$ in (2.6) is based only on σ_X and, therefore, on the fact that X is square-integrable. Under stronger moment assumptions on X, an improved bound can be easily established. For example, if X is sub-Gaussian, that is, if for every $p \ge 2$, $(\mathbb{E}|\overline{X}|^p)^{1/p} \le c\sqrt{p}\sigma_X$, the same argument used in (2.6) for $p = \log(1/\varepsilon)$ shows that

$$2\mathcal{E}(4\varepsilon, X) + \frac{\varepsilon}{2}\max\{|Q_{\varepsilon/2}(\overline{X})|, |Q_{1-\varepsilon/2}(\overline{X})|\} \le c\varepsilon\sqrt{\log(1/\varepsilon)}\sigma_X.$$

One may wonder if $\eta\sqrt{\log(1/\eta)}$ is the correct order of dependence on the contamination level for sub-Gaussian distributions. As it is proved by Chen, Gao and Ren [4], if X is Gaussian and the contamination comes from Huber's model, the correct dependence on the contamination level is proportional to η, suggesting a possible slight improvement. At the same time, as we discuss it above, $\overline{\mathcal{E}}(\eta, X)$ is a lower bound for any estimator. One may easily check that, if X is Gaussian, $\overline{\mathcal{E}}(\eta, X)$ is of the order of $\eta/\sqrt{\log(1/\eta)}$ so this lower bound is loose in this case. Interestingly, however, there exist sub-Gaussian distributions under which $\overline{\mathcal{E}}(\eta, X)$ is of the order of $\eta\sqrt{\log(1/\eta)}$. (As an example, one may take $X = \mathbb{1}_{|G|\le Q}\min(1, |G|) + \mathbb{1}_{|G|>Q}|G|$ where G is a standard Gaussian random variable and Q is its $1-\eta/2$ quantile.) This means that for sub-Gaussian distributions, the upper bound of Theorem 1 is indeed tight, up to constant factors. Note that our lower bound uses the adversarial nature of the contamination, so it might be the case that under Huber's model, even for sub-Gaussian distributions, η is the correct order.

2.1. *Proof of Theorem 1.* Recall that one is given the corrupted sample $\widetilde{X}_1, \dots, \widetilde{X}_N, \widetilde{Y}_1, \dots, \widetilde{Y}_N$, out of which at most $2\eta N$ of the sample points have been corrupted. Also, $(z_i^*)_{i=1}^N$ denotes a *nondecreasing* rearrangement of the sequence $(z_i)_{i=1}^N$.

The first step of the estimation procedure determines the truncation level, which is done using the first half of the corrupted sample.

Consider the corruption-free sample $Y_1, \dots, Y_N$ and let $U = \mathbb{1}_{\overline{X} \ge Q_{1-2\varepsilon}(\overline{X})}$. Since X is absolutely continuous, we have that $\mathbb{P}(\overline{X} \ge Q_{1-2\varepsilon}(\overline{X})) = 2\varepsilon$ and

$$\sigma_U \le \mathbb{P}^{1/2}\big(\overline{X} \ge Q_{1-2\varepsilon}(\overline{X})\big) = (2\varepsilon)^{1/2}.$$

A straightforward application of Bernstein's inequality shows that, with probability at least $1 - \exp(-\varepsilon N/12)$,

$$\big|\{i : Y_i \ge \mu + Q_{1-2\varepsilon}(\overline{X})\}\big| \ge \frac{3}{2}\varepsilon N. \tag{2.7}$$

A similar argument for $U = \mathbb{1}_{\overline{X} > Q_{1-\varepsilon/2}(\overline{X})}$ implies that, with probability at least $1 - \exp(-\varepsilon N/12)$,

$$\big|\{i : Y_i \le \mu + Q_{1-\varepsilon/2}(\overline{X})\}\big| \ge \big(1 - (3/4)\varepsilon\big)N. \tag{2.8}$$

Similarly, with probability at least $1 - 2\exp(-\varepsilon N/12)$,

$$\big|\{i : Y_i \le \mu + Q_{2\varepsilon}(\overline{X})\}\big| \ge \frac{3}{2}\varepsilon N, \tag{2.9}$$

and, with probability at least $1 - 2\exp(-\varepsilon N/12)$,

$$\big|\{i : Y_i \ge \mu + Q_{\varepsilon/2}(\overline{X})\}\big| \ge \big(1 - (3/4)\varepsilon\big)N. \tag{2.10}$$

Thus, with probability at least $1 - 4\exp(-\varepsilon N/12) \ge 1 - \delta/2$, (2.7)–(2.10) hold simultaneously on an event we denote by E. Importantly, the event E only depends on the uncorrupted sample $Y_1, \dots, Y_N$.

Since $\eta \le \epsilon/8$, following any corruption of at most $2\eta N$ points, on the event E,

$$\big|\{i : \widetilde{Y}_i \ge \mu + Q_{1-2\varepsilon}(\overline{X})\}\big| \ge \big((3/2)\varepsilon - 2\eta\big)N \ge \varepsilon N$$

and

$$\big|\{i : \widehat{Y}_i \le \mu + Q_{1-\varepsilon/2}(\overline{X})\}\big| \ge \big(1 - (3/4)\varepsilon - 2\eta\big)N \ge (1-\varepsilon)N;$$

in other words,

$$Q_{1-2\varepsilon}(\overline{X}) \le \widetilde{Y}^*_{(1-\varepsilon)N} - \mu \le Q_{1-\varepsilon/2}(\overline{X}). \tag{2.11}$$

Similarly, on the event E, we also have

$$Q_{\varepsilon/2}(\overline{X}) \le \widetilde{Y}^*_{\varepsilon N} - \mu \le Q_{2\varepsilon}(\overline{X}). \tag{2.12}$$

Recall that the truncation levels are

$$\alpha = \widetilde{Y}^*_{\varepsilon N} \quad \text{and} \quad \beta = \widetilde{Y}^*_{(1-\varepsilon)N}.$$

To prove Theorem 1, first we show that $(1/N)\sum_{i=1}^N \phi_{\alpha,\beta}(X_i)$ satisfies an inequality of the wanted form, and then we prove that corruption does not change the empirical mean of $\phi_{\alpha,\beta}$ by too much; that is, that

$$\left|\frac{1}{N}\sum_{i=1}^N \phi_{\alpha,\beta}(X_i) - \frac{1}{N}\sum_{i=1}^N \phi_{\alpha,\beta}(\widetilde{X}_i)\right|$$

is also small enough.

For the first step, note that on the event E,

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \phi_{\alpha,\beta}(X_i) &\le \frac{1}{N}\sum_{i=1}^N \phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X_i) \\
&= \mathbb{E}\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X) \\
&\quad + \frac{1}{N}\sum_{i=1}^N \big(\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X_i) \\
&\quad - \mathbb{E}\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X)\big).
\end{aligned} \tag{2.13}$$

The first term on the right-hand side of (2.13) is bounded by

$$\begin{aligned}
\mathbb{E}\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X) &\le \mu + \mathbb{E}[\overline{X}\mathbb{1}_{\overline{X}\ge Q_{1-\varepsilon/2}(\overline{X})}] \\
&\le \mu + \mathcal{E}(\varepsilon, X).
\end{aligned}$$

On the other hand, since

$$\begin{aligned}\mathbb{E}\phi_{\mu+Q_{2\epsilon}(\overline{X}),\mu+Q_{1-\epsilon/2}(\overline{X})}(X) &\geq \mu - \mathbb{E}[\overline{X}\mathbb{1}_{\overline{X}\leq Q_{2\varepsilon}(\overline{X})}] \\ &\geq \mu - \mathcal{E}(4\varepsilon, X),\end{aligned}$$

the second term on the right-hand side of (2.13) is a sum of centered i.i.d. random variables (independent of E) that are upper bounded by $Q_{1-\varepsilon/2}(\overline{X}) + \mathcal{E}(4\varepsilon, X)$ and whose variance is at most σ_X^2. Therefore, by Bernstein's inequality, conditioned on $Y_1, \ldots, Y_n$, with probability at least $1 - \delta/4$,

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\phi_{\alpha,\beta}(X_i) \\ &\quad\leq \mu + \mathcal{E}(\varepsilon, X) + \sigma_X\sqrt{\frac{2\log(4/\delta)}{N}} + \frac{Q_{1-\varepsilon/2}(\overline{X})\log(4/\delta)}{N} + \frac{\mathcal{E}(4\varepsilon, X)\log(4/\delta)}{N} \\ &\quad\leq \mu + 2\mathcal{E}(4\varepsilon, X) + 2\sigma_X\sqrt{\frac{\log(4/\delta)}{N}},\end{aligned}$$

where we used the fact that by (2.4), $Q_{1-\varepsilon/2}(\overline{X})\log(4/\delta)/N \leq \sigma_X\sqrt{\frac{\log(4/\delta)}{6N}}$ and that $\mathcal{E}(4\varepsilon, X)\log(4/\delta)/N \leq \mathcal{E}(4\varepsilon, X)$ by the assumption that $\delta \geq e^{-N}/4$.

An identical argument for the lower tail shows that, on the event E, with probability at least $1 - \delta/2$,

$$\left|\frac{1}{N}\sum_{i=1}^{N}\phi_{\alpha,\beta}(X_i) - \mu\right| \leq 2\mathcal{E}(4\varepsilon, X) + 2\sigma_X\sqrt{\frac{\log(4/\delta)}{N}}.$$

It remains to show that, on the event E,

$$\left|\frac{1}{N}\sum_{i=1}^{N}\phi_{\alpha,\beta}(X_i) - \frac{1}{N}\sum_{i=1}^{N}\phi_{\alpha,\beta}(\widetilde{X}_i)\right|$$

is small. Since $\phi_{\alpha,\beta}(X_i) \neq \phi_{\alpha,\beta}(\widetilde{X}_i)$ for at most $2\eta N$ indices, and for such points that maximal gap is

$$|\phi_{\alpha,\beta}(X_i) - \phi_{\alpha,\beta}(\widetilde{X}_i)| \leq |Q_{\varepsilon/2}(\overline{X})| + |Q_{1-\varepsilon/2}(\overline{X})|,$$

it follows that

$$\begin{aligned}\left|\frac{1}{N}\sum_{i=1}^{N}\phi_{\alpha,\beta}(X_i) - \frac{1}{N}\sum_{i=1}^{N}\phi_{\alpha,\beta}(\widetilde{X}_i)\right| &\leq 2\eta\big(|Q_{\varepsilon/2}(\overline{X})| + |Q_{1-\varepsilon/2}(\overline{X})|\big) \\ &\leq \frac{\varepsilon}{2}\max\{|Q_{\varepsilon/2}(\overline{X})|, |Q_{1-\varepsilon/2}(\overline{X})|\},\end{aligned}$$

since $\eta \leq \varepsilon/8$. Finally, note that

$$\frac{\varepsilon}{2}Q_{1-\varepsilon/2}(\overline{X}) = \mathbb{E}\big[Q_{1-\varepsilon/2}(\overline{X})\mathbb{1}_{\overline{X}\geq Q_{1-\varepsilon/2}(\overline{X})}\big] \leq \mathbb{E}[\overline{X}\mathbb{1}_{\overline{X}\geq Q_{1-\varepsilon/2}(\overline{X})}],$$

and, therefore, on the event E, we have

$$\left|\frac{1}{N}\sum_{i=1}^{N}\phi_{\alpha,\beta}(X_i) - \frac{1}{N}\sum_{i=1}^{N}\phi_{\alpha,\beta}(\widetilde{X}_i)\right| \leq \mathcal{E}(\varepsilon, X).$$

The second statement of the theorem now follows by (2.6).

3. Robust multivariate mean estimation. In this section, we present the main findings of the article; we construct a multivariate version of the robust mean estimator and establish the corresponding performance bound announced in the introduction.

As one may expect, the procedure in the multidimensional case is significantly more involved than in dimension one. In what follows, X is a random vector taking values in $\mathbb{R}^d$ with mean $\mu = \mathbb{E}X$ and covariance matrix of Σ. As before, we write $\overline{X} = X - \mu$, λ_1 denotes the largest eigenvalue of Σ, and $\mathrm{Tr}(\Sigma) = \mathbb{E}\|\overline{X}\|^2$ is its trace.

Recall that a mean estimator receives as data a sample $(\widetilde{X}_i)_{i=1}^N$ that an adversary fabricates by corrupting at most ηN points of a sample $X_1, \dots, X_N$ of independent, identically distributed copies of the random vector X. As in the univariate case, the estimator requires knowledge of the contamination level η and the confidence parameter δ. Once again, for clarity of the presentation, we assume that X has an absolutely continuous distribution with respect to the Lebesgue measure.

THEOREM 2. *Assume that X is a random vector in $\mathbb{R}^d$ that has a mean and covariance matrix. There exists a mean estimator $\widehat{\mu}$ that takes the parameters $\delta \in (0,1)$, $\eta \in [0,1)$ and the contaminated data $(\widetilde{X}_i)_{i=1}^N$ as input, and satisfies that, with probability at least $1-\delta$,*

$$\|\widehat{\mu} - \mu\| \le c\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{\lambda_1 \log(1/\delta)}{N}} + \sqrt{\lambda_1 \eta}\right),$$

where $c > 0$ is a numerical constant.

A value of the numerical constant is explicitly given in the proof. However, no attempt has been made to optimize its value.

The same remark as in the univariate case on the previous knowledge of η and δ, mentioned after Theorem 1, applies here as well.

As it is pointed out in the Introduction, the bound of Theorem 2 coincides with the best possible bound in the corruption-free case up to the term $\sqrt{\lambda_1 \eta}$ that is the price one has to pay for adversarial corruption. The fact that the term $\sqrt{\lambda_1 \eta}$ is inevitable in the upper bound follows from the fact that for any upper bound for the norm of difference $\|\widehat{\mu} - \mu\|$, the same upper bound holds for any one-dimensional marginal. Hence, the necessity of this term follows from our arguments in the univariate case. At the same time, similar to the univariate case, under higher moment assumptions, the term $\sqrt{\lambda_1 \eta}$ may be improved. For instance, if the distribution is sub-Gaussian (in the sense that all one-dimensional projections are sub-Gaussian), then this term may be replaced by $\eta\sqrt{\log(1/\eta)}\sqrt{\lambda_1}$. This may be seen by a straightforward modification of the proof.

Remarkably, the malicious sample corruption affects only the "weak" term of the bound, that is, it scales with the square root of the operator norm of the covariance matrix. Indeed, if the corruption parameter η is such that $\eta N \le \log(2/\delta)$, then, with probability at least $1-\delta$, $\widehat{\mu}$ satisfies

$$\|\widehat{\mu} - \mu\| \le c\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\lambda_1}\sqrt{\frac{\log(1/\delta)}{N}}\right), \tag{3.1}$$

matching the optimal bound for multivariate mean estimation bound from [21] for the corruption-free case. If, on the other hand, the corruption parameter is larger, then Theorem 2 implies that with probability at least $1 - 2\exp(-\eta N/c)$,

$$\|\widehat{\mu} - \mu\| \le c\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\eta}\sqrt{\lambda_1}\right) \tag{3.2}$$

for a numerical constant $c > 0$.

In what follows we describe the construction of the mean estimator $\widehat{\mu}$ that satisfies the announced performance bound.

3.1. *The multivariate mean estimator.* The main component is a mean estimation procedure that, in order to perform well, requires information on $\mathrm{Tr}(\Sigma)$ and λ_1. Since such information is not assumed to be available, we produce an estimator depending on a tuning parameter Q. Then we use a simple mechanism of choosing the appropriate value of Q.

Just like in the univariate case, for simplicity of notation, assume that the estimator receives $2N$ data points $\widetilde{X}_1, \dots, \widetilde{X}_N, \widetilde{Y}_1, \dots, \widetilde{Y}_N$, and that at most $2\eta N$ points of the original independent sample $X_1, \dots, X_N, Y_1, \dots, Y_N$ have been changed by the adversary. The procedure computes, for each unit vector v and tuning parameter $Q > 0$, the trimmed mean estimate of the expectation of the projection of X to the line spanned by v with a minor difference: the truncation level is widened depending on the parameter Q. Each one of these estimators defines a slab in $\mathbb{R}^d$. The details are as follows.

Multivariate mean estimator.

(1) Set

$$\varepsilon = \max\biggl(10\eta, 2560\frac{\log(2/\delta)}{N}\biggr).$$

(2) Let S^{d-1} be the Euclidean unit sphere in $\mathbb{R}^d$ and for every $v \in S^{d-1}$ define

$$\alpha_v = \bigl(\langle \widetilde{Y}_i, v\rangle\bigr)^*_{(\varepsilon/2)N} \quad \text{and} \quad \beta_v = \bigl(\langle \widetilde{Y}_i, v\rangle\bigr)^*_{(1-\varepsilon/2)N}.$$

(3) For every $v \in S^{d-1}$ and $Q > 0$, set

$$U_Q(v) = \frac{1}{N}\sum_{i=1}^{N} \phi_{\alpha_v - Q, \beta_v + Q}\bigl(\langle \widetilde{X}_i, v\rangle\bigr),$$

and let

$$\Gamma(v, Q) = \bigl\{x \in \mathbb{R}^d : \bigl|\langle x, v\rangle - U_Q(v)\bigr| \le 2\varepsilon Q\bigr\}.$$

(4) For each $Q > 0$, set

$$\Gamma(Q) = \bigcap_{v \in S^{d-1}} \Gamma(v, Q).$$

(5) Let $i^* \in \mathbb{Z}$ be the smallest such that $\bigcap_{i \ge i^*} \Gamma(2^i) \neq \varnothing$. Define $\widehat{\mu}$ to be any point in

$$\bigcap_{i \in \mathbb{Z}: i \ge i^*} \Gamma\bigl(2^i\bigr).$$

Each set $\Gamma(Q)$ is an intersection of random slabs, one for each direction in the sphere S^{d-1}. The "center" of the slab associated with the direction v is $U_Q(v)$ and its width is proportional to εQ. As we show in what follows, there is some $i_0 \in \mathbb{Z}$ such that with probability at least $1 - \delta$, the sets $\Gamma(2^i), i \ge i_0$ are nested, implying that $\widehat{\mu}$ is well defined. Note that the last step of selecting the value of Q is reminiscent of Lepski's method [19] or the related method "intersection of confidence intervals" by Goldenshluger and Nemirovski [13].

4. Proof of Theorem 2. The heart of the proof of Theorem 2 is the following proposition that describes the performance of an estimator with the correct tuning parameter Q.

The role of Q is to incorporate the "global complexity" of S^{d-1}. In particular, if Q is selected properly, that is enough to ensure that $\Gamma(Q)$ is nonempty and contains a good estimator of μ. This is formalized in the next proposition.

PROPOSITION 1. *Let*

$$Q_0 = \max\left(\frac{256}{\varepsilon}\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}}, 16\sqrt{\frac{\lambda_1}{\varepsilon}}\right) \tag{4.1}$$

and consider $Q \in [2Q_0, 4Q_0]$. *Then, with probability at least* $1 - 2\exp(-\varepsilon N/2560) \geq 1 - \delta$, $\Gamma(Q) \neq \varnothing$ *and for every* $z \in \Gamma(Q)$,

$$\|z - \mu\| \leq 4\varepsilon Q_0.$$

Observe that for every Q, the diameter of $\Gamma(Q)$ is at most $4\varepsilon Q$. Indeed, if $x_1, x_2 \in \Gamma(Q)$ then for every $v \in S^{d-1}$,

$$\big|\langle x_1 - x_2, v\rangle\big| = \big|\langle x_1, v\rangle - \langle x_2, v\rangle\big| \leq \big|\langle x_1, v\rangle - U_Q(v)\big| + \big|\langle x_2, v\rangle - U_Q(v)\big| \leq 4\varepsilon Q,$$

implying that $\|x_1 - x_2\| \leq 4\varepsilon Q$.

The key component in the proof of Proposition 1 is the next lemma.

LEMMA 1. *For each* $i \in \{1, \ldots, N\}$ *and* $v \in S^{N-1}$, *define* $\overline{Y}_i(v) = \langle Y_i - \mu, v\rangle$. *With probability at least* $1 - \exp(-\varepsilon N/2560) \geq 1 - \delta/2$,

$$\sup_{v \in S^{d-1}} \big|\{i : \overline{Y}_i(v) \geq Q_0\}\big| \leq \frac{\varepsilon}{8}N \quad \text{and} \quad \sup_{v \in S^{d-1}} \big|\{i : \overline{Y}_i(v) \leq -Q_0\}\big| \leq \frac{\varepsilon}{8}N. \tag{4.2}$$

Lemma 1 is a uniform version of the analogous claim used in the univariate case.

PROOF. Let us prove the first inequality; the second is proved by an identical argument and is omitted. Consider the function $\chi : \mathbb{R} \to \mathbb{R}$, defined by

$$\chi(x) = \begin{cases} 0 & \text{if } x \leq Q_0/2, \\ \dfrac{2x}{Q_0} - 1 & \text{if } x \in (Q_0/2, Q_0], \\ 1 & \text{if } x > Q_0. \end{cases}$$

Observe that $\mathbb{1}_{\{\overline{Y}(v) \geq Q_0\}} \leq \chi(\overline{Y}(v)) \leq \mathbb{1}_{\{\overline{Y}(v) \geq Q_0/2\}}$, and that χ is Lipschitz with constant $2/Q_0$. Therefore, if $\varepsilon_1, \ldots, \varepsilon_N$ are independent, symmetric $\{-1, 1\}$-valued random variables that are independent of the $(Y_i)_{i=1}^N$, then

$$\begin{aligned}
&\mathbb{E}\sup_{v \in S^{d-1}} \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\{\overline{Y}_i(v) \geq Q_0\}} \\
&\quad \leq \mathbb{E}\sup_{v \in S^{d-1}} \frac{1}{N}\sum_{i=1}^N \chi\big(\overline{Y}_i(v)\big) \\
&\quad \leq 2\mathbb{E}\sup_{v \in S^{d-1}} \frac{1}{N}\left|\sum_{i=1}^N \varepsilon_i \chi\big(\overline{Y}_i(v)\big)\right| + \sup_{v \in S^{d-1}} \mathbb{E}\chi\big(\overline{Y}(v)\big) \\
&\qquad \big(\text{by the Giné–Zinn symmetrization theorem [12]}\big) \\
&\quad \leq \frac{4}{Q_0}\mathbb{E}\sup_{v \in S^{d-1}} \frac{1}{N}\left|\sum_{i=1}^N \varepsilon_i \overline{Y}_i(v)\right| + \sup_{v \in S^{d-1}} \mathbb{E}\chi\big(\overline{Y}(v)\big) \\
&\quad \stackrel{\text{def.}}{=} (*),
\end{aligned}$$

where in the second step one uses the standard contraction lemma for Rademacher averages; see Ledoux and Talagrand [18].

To bound the second term on the right-hand side, recall that $Q_0 \geq 16\sqrt{\lambda_1/\varepsilon}$, and thus, for every $v \in S^{d-1}$,

$$\begin{aligned}\mathbb{E}\chi\big(\overline{Y}(v)\big) &\leq \mathbb{E}\mathbb{1}_{\{\overline{Y}(v)\geq Q_0/2\}} = \mathbb{P}\bigg(\langle \overline{X}, v\rangle \geq \frac{Q_0}{2}\bigg) \\ &\leq \frac{4\mathbb{E}\langle \overline{X}, v\rangle^2}{Q_0^2} \leq \frac{4\lambda_1}{Q_0^2} \leq \frac{\varepsilon}{64}.\end{aligned} \tag{4.3}$$

To bound the first term, note that

$$\mathbb{E}\sup_{v\in S^{d-1}}\bigg|\frac{1}{N}\sum_{i=1}^N \varepsilon_i \overline{Y}_i(v)\bigg| = \mathbb{E}\sup_{v\in S^{d-1}}\bigg|\frac{1}{N}\sum_{i=1}^N \varepsilon_i \langle X_i - \mu, v\rangle\bigg| \leq \sqrt{\frac{\operatorname{Tr}(\Sigma)}{N}}.$$

Hence, by the definition of Q_0,

$$(*) \leq \frac{\varepsilon}{32}.$$

By Talagrand's concentration inequality for empirical processes indexed by a class of uniformly bounded functions [27], with probability at least $1 - \exp(-x)$,

$$\frac{1}{N}\sup_{v\in S^{d-1}}\big|\{i : \overline{Y}_i(v) \geq Q_0\}\big| \leq \frac{\varepsilon}{16} + \sqrt{\frac{x}{N}}\cdot\frac{\sqrt{\varepsilon}}{128} + \frac{10x}{N}$$

(see [2], Exercise 12.15, for the value of the numerical constant).

With the choice of $x = \varepsilon N/2560$ one has that, with probability at least $1 - \exp(-\varepsilon N/2560)$,

$$\sup_{v\in S^{d-1}}\big|\{i : \overline{Y}_i(v) \geq Q_0\}\big| \leq \frac{\varepsilon}{8}N,$$

as required. □

Note that, when (4.2) holds, we have, for every $v \in S^{d-1}$,

$$\alpha_v - \langle \mu, v\rangle \geq -Q_0 \quad \text{and} \quad \beta_v - \langle \mu, v\rangle \leq Q_0.$$

Indeed, this follows from the fact that for every $v \in S^{d-1}$ there are at most $(\varepsilon/8)N$ of the $\overline{Y}_i(v)$ that are larger than Q_0. If, in addition, the adversary corrupts at most $(\varepsilon/8)N$ of the points Y_i, then there are still no more than $(\varepsilon/4)N$ values $\langle \widetilde{Y}_i, v\rangle$ that are larger than $\langle \mu, v\rangle + Q_0$, which suffices for our purposes. And, by the definition of ε, one has that $\varepsilon/8 \geq \eta$, as required.

Now consider some Q that satisfies $2Q_0 < Q \leq 4Q_0$, and from here we condition on an event E such that the inequalities (4.2) both hold. By Lemma 1, E occurs with probability at least $1 - \exp(-\varepsilon N/2560)$; importantly, this event only depends on $Y_1, \ldots, Y_N$, the first half of the uncontaminated sample.

In particular, on the event E, for every $v \in S^{d-1}$,

$$\beta_v - \langle \mu, v\rangle + Q \leq Q_0 + Q \leq 5Q_0$$

and

$$\beta_v - \langle \mu, v\rangle + Q \geq \alpha_v - \langle \mu, v\rangle + Q \geq -Q_0 + Q \geq Q_0.$$

By a similar argument, one may obtain lower and upper bounds for $\alpha_v - \langle \mu, v \rangle$. Hence, on E, for every $v \in S^{d-1}$,

$$-5Q_0 \le \alpha_v - \langle \mu, v \rangle - Q \le -Q_0 \quad \text{and} \quad Q_0 \le \beta_v - \langle \mu, v \rangle + Q \le 5Q_0. \tag{4.4}$$

Finally, recall that

$$U_Q(v) = \frac{1}{N} \sum_{i=1}^N \phi_{\alpha_v - Q, \beta_v + Q}\big(\langle \widetilde{X}_i, v \rangle\big),$$

and in order to complete the proof of Proposition 1, it suffices to show that $U_Q(v)$ is uniformly close to $\langle \mu, v \rangle$, with high probability. In particular, the next lemma implies Proposition 1.

LEMMA 2. *Let $2Q_0 \le Q \le 4Q_0$. Conditioned on the event E, with probability at least $1 - 2\exp(-\varepsilon N/2560)$,*

$$\sup_{v \in S^{d-1}} \big| U_Q(v) - \langle \mu, v \rangle \big| \le 2\varepsilon Q.$$

PROOF. We prove that

$$\sup_{v \in S^{d-1}} \big(U_Q(v) - \langle \mu, v \rangle \big) \le 2\varepsilon Q$$

holds with the wanted probability; the proof that

$$\sup_{v \in S^{d-1}} \big(\langle \mu, v \rangle - U_Q(v) \big) \le 2\varepsilon Q$$

follows an identical argument and is omitted.

As a first step, note that, in the expression of $U_Q(v)$, the corrupted samples $\widetilde{X}_i$ may be harmlessly replaced by their uncorrupted counterparts X_i. Indeed, by (4.4), on the event E, the range of the function $\phi_{\alpha_v - Q, \beta_v + Q}$ is an interval of length at most $10Q$ and, therefore, deterministically, for all $v \in S^{d-1}$,

$$\frac{1}{N} \sum_{i=1}^N \phi_{\alpha_v - Q, \beta_v + Q}\big(\langle \widetilde{X}_i, v \rangle\big) - \frac{1}{N} \sum_{i=1}^N \phi_{\alpha_v - Q, \beta_v + Q}\big(\langle X_i, v \rangle\big) \le \eta \cdot 10Q \le \varepsilon Q.$$

Once again, recalling that on E (4.4) holds, it follows that

$$\frac{1}{N} \sum_{i=1}^N \phi_{\alpha_v - Q, \beta_v + Q}\big(\langle X_i, v \rangle\big) \le \frac{1}{N} \sum_{i=1}^N \phi_{\langle \mu, v \rangle - Q_0, \langle \mu, v \rangle + 5Q_0}\big(\langle X_i, v \rangle\big).$$

Since the event E only depends on the uncorrupted sample $Y_1, \dots, Y_N$, the right-hand side of the above inequality is independent of E. Thus, writing

$$\overline{U}_Q(v) = \frac{1}{N} \sum_{i=1}^N \phi_{\langle \mu, v \rangle - Q_0, \langle \mu, v \rangle + 5Q_0}\big(\langle X_i, v \rangle\big) - \langle \mu, v \rangle = \frac{1}{N} \sum_{i=1}^N \phi_{-Q_0, 5Q_0}\big(\langle X_i - \mu, v \rangle\big),$$

it suffices to prove that, with probability at least $1 - 2e^{-\varepsilon N/2560}$,

$$\sup_{v \in S^{d-1}} \overline{U}_Q(v) \le \varepsilon Q.$$

To that end, consider the decomposition

$$\sup_{v \in S^{d-1}} \overline{U}_Q(v) \le \sup_{v \in S^{d-1}} \big(\overline{U}_Q(v) - \mathbb{E}\overline{U}_Q(v) \big) + \sup_{v \in S^{d-1}} \mathbb{E}\overline{U}_Q(v) \stackrel{\text{def.}}{=} (1) + (2).$$

First, let us bound the term (1) in several steps.

Set

$$\overline{W}_Q(v) = \frac{1}{N}\sum_{i=1}^{N} \phi_{-3Q,3Q}\big(\langle X_i - \mu, v\rangle\big),$$

and note that

$$\begin{aligned}\sup_{v\in S^{d-1}} \big(\overline{U}_Q(v) - \mathbb{E}\overline{U}_Q(v)\big) \;\le\; & \sup_{v\in S^{d-1}} \big(\overline{U}_Q(v) - \overline{W}_Q(v)\big) \\ & + \sup_{v\in S^{d-1}} \big(\overline{W}_Q(v) - \mathbb{E}\overline{W}_Q(v)\big) \\ & + \sup_{v\in S^{d-1}} \big(\mathbb{E}\overline{W}_Q(v) - \mathbb{E}\overline{U}_Q(v)\big) \\ \overset{\text{def.}}{=}\; & (a) + (b) + (c).\end{aligned}$$

To bound term (a), recall that $2Q_0 \le Q \le 4Q_0$, implying that $\phi_{-Q_0,5Q_0}(x) \ne \phi_{-3Q,3Q}(x)$ only if

$$\text{either} \quad x < -Q_0 \quad \text{or} \quad x > 5Q_0.$$

In both cases,

$$\big|\phi_{-Q_0,5Q_0}(x) - \phi_{-3Q,3Q}(x)\big| \le 3Q.$$

By Lemma 1, with probability at least $1 - \exp(-\varepsilon N/2560)$,

$$\sup_{v\in S^{d-1}} \big|\{i : \langle X_i - \mu, v\rangle > 5Q_0 \text{ or } \langle X_i - \mu, v\rangle < -Q_0\}\big| \le \frac{\varepsilon N}{4},$$

hence, on this event,

$$(a) \le \frac{3\varepsilon Q}{4}.$$

One may control term (c) similarly. For each $v \in S^{d-1}$,

$$\mathbb{E}\overline{W}_Q(v) - \mathbb{E}\overline{U}_Q(v) \le 3Q \cdot \mathbb{P}\big\{\big|\langle X - \mu, v\rangle\big| > Q_0\big\} \le \frac{3\varepsilon Q}{64}$$

by recalling (4.3).

The term (b) is controlled using Talagrand's concentration inequality for the supremum of empirical processes. Note that for every $v \in S^{d-1}$,

$$\big|\phi_{-3Q,3Q}\big(\langle \overline{X}, v\rangle\big)\big| \le 3Q \quad \text{and} \quad \mathbb{E}\big|\phi_{-3Q,3Q}\big(\langle \overline{X}, v\rangle\big)\big|^2 \le \mathbb{E}\big|\langle \overline{X}, v\rangle\big|^2 \le \lambda_1.$$

Also, since $\phi_{-3Q,3Q}(x)$ is a 1-Lipschitz function that passes through 0, by a contraction argument (see Ledoux and Talagrand [18]),

$$\mathbb{E}\sup_{v\in S^{d-1}} \big|\overline{W}_Q(v) - \mathbb{E}\overline{W}_Q(v)\big| \le 2\mathbb{E}\sup_{v\in S^{d-1}} \left|\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i \langle X_i - \mu, v\rangle\right| \le 2\sqrt{\frac{\operatorname{Tr}(\Sigma)}{N}}.$$

Hence, by Talagrand's inequality, with probability at least $1 - 2\exp(-x)$,

$$\sup_{v\in S^{d-1}} \big|\overline{W}_Q(v) - \mathbb{E}\overline{W}_Q(v)\big| \le 4\sqrt{\frac{\operatorname{Tr}(\Sigma)}{N}} + 2\sqrt{\lambda_1}\sqrt{\frac{x}{N}} + 20Q\frac{x}{N} \le \frac{\varepsilon Q}{64}$$

with the choice of $x = \varepsilon N/2560$, recalling the definition of Q_0, and using that $Q \geq 2Q_0$. This concludes the proof that (1) $\leq (1/2 + 1/32 + 1/400)\varepsilon Q$ with probability $1 - e^{-\varepsilon N/2560}$.

Finally, it remains to estimate term (2),

$$(2) = \sup_{v \in S^{d-1}} \mathbb{E}\overline{U}_Q(v) = \sup_{v \in S^{d-1}} \mathbb{E}\phi_{-Q_0,5Q_0}(\langle \overline{X}, v\rangle).$$

Clearly, $X_v = \langle \overline{X}, v\rangle$ is centered and $\phi_{-Q_0,5Q_0}(X_v) \neq X_v$ only when either $X_v \geq 5Q_0$ or $X_v \leq -Q_0$. Hence,

$$\begin{aligned}\mathbb{E}\phi_{-Q_0,5Q_0}(X_v) &= \mathbb{E}\big(\phi_{-Q_0,5Q_0}(X_v) - X_v\big)\\ &\leq \mathbb{E}|Q_0 + X_v|\mathbb{1}_{X_v \leq -Q_0}\\ &\leq \frac{\varepsilon Q}{64}\end{aligned}$$

by an argument analogous to (2.5) and using (4.3). □

With Proposition 1 proved, let us complete the proof of Theorem 2. Let i_0 be such that $Q \overset{\text{def.}}{=} 2^{i_0} \in [2Q_0, 4Q_0)$ and let E be the "good" event that both (4.2) and

$$\sup_{v \in S^{d-1}} \big|U_Q(v) - \langle \mu, v\rangle\big| \leq 2\varepsilon Q$$

hold. Recall that

$$U_Q(v) = \frac{1}{N}\sum_{i=1}^{N} \phi_{\alpha_v - Q, \beta_v + Q}\big(\langle \widetilde{X}_i, v\rangle\big);$$

E holds with probability at least $1 - \delta$; and on E, any point in $\Gamma(2^{i_0})$ is within distance $4\varepsilon Q_0$ of the mean μ. Hence, it suffices to show that on the event E, the sets $\Gamma(2^i)$ for $i \geq i_0$ are nested. Indeed, by the definition of i^*,

$$\varnothing \neq \bigcap_{i \geq i^*} \Gamma\big(2^i\big) \subset \Gamma\big(2^{i_0}\big),$$

and thus $\|\widehat{\mu} - \mu\| \leq 4\varepsilon Q_0$.

To see that $\Gamma(2^{i_0}) \subset \Gamma(2^{i_0+1})$, it is enough to show that, for all $v \in S^{d-1}$, $|\langle x, v\rangle - U_{2Q}(v)| \leq 4\varepsilon Q$. But if $x \in \Gamma(v, Q)$ for some $v \in S^{d-1}$, it follows that

$$\big|\langle x, v\rangle - U_{2Q}(v)\big| \leq \big|\langle x, v\rangle - U_Q(v)\big| + \big|U_Q(v) - U_{2Q}(v)\big| \leq 2\varepsilon Q + \big|U_Q(v) - U_{2Q}(v)\big|;$$

therefore, it suffices to show that $|U_Q(v) - U_{2Q}(v)| \leq 2\varepsilon Q$.

Note that on the event E, there are at most $\varepsilon N/4$ sample points $\widetilde{X}_i$ such that $\langle \widetilde{X}_i, v\rangle$ is above or below the levels $\alpha_v - 2^{i_0}$ and $\beta_v + 2^{i_0}$. Hence, the number of points for which $U_Q(v) \neq U_{2Q}(v)$ is at most $\varepsilon N/4$ and so the difference is at most $(2Q\varepsilon N/4)/N = \varepsilon Q/2$.

By induction, the same argument shows that, on the event E, $\Gamma(2^i) \subset \Gamma(2^{i+1})$ for every $i \geq i_0$, completing the proof of Theorem 2.

Acknowledgments. We thank the referees and the Associate Editor for insightful comments and pointing out relevant connections to previous work.

Gábor Lugosi was supported by the Spanish Ministry of Economy and Competitiveness, Grant PGC2018-101643-B-I00; "High-dimensional problems in structured probabilistic models—Ayudas Fundación BBVA a Equipos de Investigación Cientifica 2017"; and Google Focused Award "Algorithms and Learning for AI."

REFERENCES

[1] BICKEL, P. J. (1965). On some robust estimates of location. *Ann. Math. Stat.* **36** 847–858. MR0177484 https://doi.org/10.1214/aoms/1177700058

[2] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[3] CHEN, M., GAO, C. and REN, Z. (2016). A general decision theory for Huber's ϵ-contamination model. *Electron. J. Stat.* **10** 3752–3774. MR3579675 https://doi.org/10.1214/16-EJS1216

[4] CHEN, M., GAO, C. and REN, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* **46** 1932–1960. MR3845006 https://doi.org/10.1214/17-AOS1607

[5] CHERAPANAMJERI, Y., FLAMMARION, N. and BARTLETT, P. (2019). Fast mean estimation with sub-gaussian rates. Preprint. Available at arXiv:1902.01998.

[6] DALALYAN, A. and THOMPSON, P. (2019). Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's M-estimator. In *Advances in Neural Information Processing Systems* 13188–13198.

[7] DEPERSIN, J. and LECUÉ, G. (2019). Robust subgaussian estimation of a mean vector in nearly linear time. Preprint. Available at arXiv:1906.03058.

[8] DEVROYE, L., LERASLE, M., LUGOSI, G. and OLIVEIRA, R. I. (2016). Sub-Gaussian mean estimators. *Ann. Statist.* **44** 2695–2725. MR3576558 https://doi.org/10.1214/16-AOS1440

[9] DIAKONIKOLAS, I., KAMATH, G., KANE, D. M., LI, J., MOITRA, A. and STEWART, A. (2016). Robust estimators in high dimensions without the computational intractability. In 57*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2016 655–664. IEEE Computer Soc., Los Alamitos, CA. MR3631028 https://doi.org/10.1109/FOCS.2016.85

[10] DIAKONIKOLAS, I., KAMATH, G., KANE, D. M., LI, J., MOITRA, A. and STEWART, A. (2017). Being robust (in high dimensions) can be practical. In *Proceedings of the* 34*th International Conference on Machine Learning* (*ICML* 2017).

[11] DIAKONIKOLAS, I., KAMATH, G., KANE, D. M., LI, J., MOITRA, A. and STEWART, A. (2018). Robustly learning a Gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* 2683–2702. SIAM, Philadelphia, PA. MR3775959 https://doi.org/10.1137/1.9781611975031.171

[12] GINÉ, E. and ZINN, J. (1984). Some limit theorems for empirical processes. *Ann. Probab.* **12** 929–998. MR0757767

[13] GOLDENSHLUGER, A. and NEMIROVSKI, A. (1997). On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.* **6** 135–170. MR1466625

[14] HOPKINS, S. B. (2020). Sub-Gaussian mean estimation in polynomial time. *Ann. Statist.* To appear.

[15] HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2488795 https://doi.org/10.1002/9780470434697

[16] KEARNS, M. and LI, M. (1993). Learning in the presence of malicious errors. *SIAM J. Comput.* **22** 807–837. MR1227763 https://doi.org/10.1137/0222052

[17] LECUÉ, G. and LERASLE, M. (2020). Robust machine learning by median-of-means: Theory and practice. *Ann. Statist.* **48** 906–931.

[18] LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces*: *Isoperimetry and Processes*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) [*Results in Mathematics and Related Areas* (3)] **23**. Springer, Berlin. MR1102015 https://doi.org/10.1007/978-3-642-20212-4

[19] LEPSKIĬ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682–697.

[20] LUGOSI, G. and MENDELSON, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comput. Math.* **19** 1145–1190. MR4017683 https://doi.org/10.1007/s10208-019-09427-x

[21] LUGOSI, G. and MENDELSON, S. (2019). Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.* **47** 783–794. MR3909950 https://doi.org/10.1214/17-AOS1639

[22] MINSKER, S. (2018). Uniform bounds for robust mean estimators. Preprint. Available at arXiv:1812.03523.

[23] OLIVEIRA, R. I. and ORENSTEIN, P. (2019). The sub-gaussian property of trimmed means estimators. Technical report, IMPA.

[24] RODRIGUEZ, D. and VALDORA, M. (2019). The breakdown point of the median of means tournament. *Statist. Probab. Lett.* **153** 108–112. MR3962785 https://doi.org/10.1016/j.spl.2019.05.012

[25] STEINHARDT, J., CHARIKAR, M. and VALIANT, G. (2018). Resilience: A criterion for learning in the presence of arbitrary outliers. In 9*th Innovations in Theoretical Computer Science*. *LIPIcs*. *Leibniz Int. Proc. Inform.* **94** Art. No. 45, 21. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3761781

[26] STIGLER, S. M. (1973). The asymptotic distribution of the trimmed mean. *Ann. Statist.* **1** 472–477. MR0359134

[27] TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563. MR1419006 https://doi.org/10.1007/s002220050108

[28] TUKEY, J. W. and MCLAUGHLIN, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization. I. *Sankhyā Ser. A* **25** 331–352. MR0169354

[29] VALIANT, L. G. (1985). Learning disjunction of conjunctions. In *IJCAI* 560–566.