Near-optimal mean estimators with respect to general norms *

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

June 16, 2018

Abstract

We study the problem of estimating the mean of a random vector in $\mathbb{R}^d$ based on an i.i.d. sample, when the accuracy of the estimator is measured by a general norm on $\mathbb{R}^d$. We construct an estimator (that depends on the norm) that achieves an essentially optimal accuracy/confidence tradeoff under the only assumption that the random vector has a well-defined covariance matrix. The estimator is based on the construction of a uniform median-of-means estimator in a class of real valued functions that may be of independent interest.

1 Introduction

In this note we explore the problem of multivariate mean estimation with respect to an arbitrary norm. To formulate the question, let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ and let X be a random vector in $\mathbb{R}^d$. One only assumes that X has a mean $\mu = \mathbb{E}X$ and a well-defined covariance matrix $\Sigma = \mathbb{E}(X-\mu)\otimes(X-\mu)$. The statistical problem we consider is estimating the mean vector μ from a sample $(X_i)_{i=1}^N$ of N independent copies of X. We do not assume any knowledge on the distribution. The goal is to

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

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

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

[§]Barcelona Graduate School of Economics

[¶]Mathematical Sciences Institute, The Australian National University and Department of Mathematics, Technion, I.I.T, shahar.mendelson@anu.edu.au

approximate the mean μ by finding some *mean estimator* $\widehat{\mu}_N = \widehat{\mu}_N(X_1, \ldots, X_N) \in \mathbb{R}^d$ such that $\|\widehat{\mu}_N - \mu\|$ is as small as possible.

Formally, the problem studied in this note is as follows:

> Given a norm $\|\cdot\|$, a confidence parameter $\delta \in (0,1)$ and an i.i.d. sample of cardinality N, find an estimator $\widehat{\mu}_N$ and the best possible accuracy ε for which
>
> $$\|\widehat{\mu}_N - \mu\| \leq \epsilon \quad \text{with probability at least } 1 - \delta\,.$$

Various versions of this question have been studied extensively in recent years, but it was far from resolved. In fact, even the correct order of the best accuracy ϵ was not clear, except in special situations. While there are some results for specific choices of norms, the only estimate that is known to be optimal was obtained in Lugosi and Mendelson [12] for the Euclidean norm, see also Joly, Lugosi, and Oliveira [9] and Catoni and Giulini [5]. In addition, there are also several partial results (see Minsker [15], Catoni and Giulini [5]) for other special norms (mainly in the context of the matrix operator norm) and which are suboptimal, as we will see below.

We start by discussing what kind of accuracy ϵ one should be aiming for. To this end, first consider the case when X is a real-valued random variable with finite mean μ and variance σ^2. Since the real-valued case is well-understood, it will eventually lead us to the possible identity of ϵ in the vector-valued scenario.

The first observation (see, e.g., Catoni [4]) is that if X is a Gaussian random variable then the best mean estimate that one can hope for is such that, with probability $1 - \delta$,

$$|\widehat{\mu}_N - \mu| \leq c\sigma \sqrt{\frac{\log(2/\delta)}{N}}\,. \tag{1.1}$$

Here c is an absolute constant. (In this article we focus on optimal orders of magnitude and ignore the–important–problem of optimizing constants.) If X is indeed Gaussian, then the choice of $\widehat{\mu}_N$ is simple: the empirical mean

$$\frac{1}{N}\sum_{i=1}^{N} X_i$$

has the desired accuracy at all confidence levels δ.

The empirical mean also yields (1.1) when X is L-sub-Gaussian, that is, if for every $p \geq 2$, $\|\overline{X}\|_{L_p} \leq L\sqrt{p}\|\overline{X}\|_{L_2}$, where $\overline{X} = X - \mu$ and $\|\overline{X}\|_{L_p} = \left(\mathbb{E}|\overline{X}|^p\right)^{1/p}$, see, for example, [3].

Unfortunately, this is as far as the empirical mean takes us. As soon as one leaves the sub-Gaussian realm, the empirical mean becomes a poor choice

and its performance deteriorates for 'heavy-tailed' distributions of X. In fact, for all δ there are distributions in which the estimate that follows from Chebyshev's inequality, that

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| \geq \frac{\sigma}{\sqrt{\delta N}}\right) \leq \delta, \tag{1.2}$$

is sharp. In other words, while the expected value

$$\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right|$$

is of the right order of magnitude ($\sim \sigma/\sqrt{N}$), the empirical mean exhibits rather poor concentration around μ.

Thus, the empirical mean has a performance comparable to the Gaussian case only in two situations:

- For an arbitrary distribution of X if one is only interested in constant confidence level (say $\delta = 0.1$), in which case the resulting accuracy is $\mathbb{E}|N^{-1}\sum_{i=1} X_i - \mu|$;
- If X is L-sub-Gaussian and one is interested in any confidence level, in which case the error is determined by estimating the probability $\mathbb{P}(|N^{-1}\sum_{i=1}^{N} X_i - \mu| \geq \eta)$.

Perhaps surprisingly, the error one incurs in these two special and restrictive situations can be attained in full generality (though obviously the estimator one uses is not the empirical mean). One estimator that attains a "sub-Gaussian" performance (i.e., an accuracy bounded by $cN^{-1/2}\sigma\sqrt{\log(2/\delta)}$ for an absolute constant c) for any X with finite mean and variance is the *median-of-means estimator*. To compute this estimator, first the sample $X_1, \ldots, X_N$ is split into n blocks I_j, each one of the same cardinality m (here we assume without loss of generality that n divides N). For each block I_j, let

$$a_j = \frac{1}{m}\sum_{i \in I_j} X_i,$$

and put $\widehat{\mu}_N$ to be a median of $\{a_1, \ldots, a_n\}$. Setting $n \sim \log(2/\delta)$, it is straightforward to verify that this choice of $\widehat{\mu}_N$ satisfies (1.1). This estimator was introduced independently by Nemirovsky and Yudin [16]; Jerrum, Valiant, and Vazirani [8]; and Alon, Matias, and Szegedy [1]. Another, quite different, sub-Gaussian estimator was constructed by Catoni [4].

Note that unlike the empirical mean, here the procedure changes with the desired confidence. This is indeed necessary. As it is shown by Devroye, Lerasle,

Lugosi, and Oliveira [6], there is no single procedure that attains (1.1) for all confidence levels and for all distributions with finite second moment.

While the one-dimensional picture was well understood, in higher dimensions the situation was far less clear. Unfortunately, establishing the 'right' notion of error in higher dimensions and with respect to a general norm can be difficult, as parameters that are totally different in the multi-dimensional setup may 'collapse' to the same object in dimension one. However, one may still learn a lesson from the real-valued case and conclude the following:

- An estimator with accuracy of optimal order should depend on the prescribed confidence level and on the norm in question.
- A reasonable notion of error is dictated by what happens in the two 'trivial' situations—in both of which the empirical mean is essentially optimal—as in dimension one. For a real-valued random variable, when only a constant confidence is required, the error is of the order of $\mathbb{E}|N^{-1}\sum_{i=1}^N X_i - \mu|$. For small values of δ, the optimal accuracy is of the order of η for which $\mathbb{P}(|N^{-1}\sum_{i=1}^N X_i - \mu| \geq \eta) \leq \delta$ when X is L-sub-Gaussian. The analogous objects for a random vector in $(\mathbb{R}^d, \|\cdot\|)$ are the expectation of the norm

$$\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N X_i - \mu\right\| \tag{1.3}$$

  and the value η such that

$$\mathbb{P}\left(\left\|\frac{1}{N}\sum_{i=1}^N X_i - \mu\right\| \geq \eta\right) \leq \delta \tag{1.4}$$

  when X is an L-sub-Gaussian random vector[1].

We put (1.3) in a form more convenient for us. To this end, set

$$Y_N = \frac{1}{\sqrt{N}}\sum_{i=1}^N \varepsilon_i(X_i - \mu)$$

where $(\varepsilon_i)_{i=1}^N$ are independent, symmetric, $\{-1,1\}$-valued random variables that are also independent of $(X_i)_{i=1}^N$. A standard symmetrization argument shows that

$$\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N X_i - \mu\right\| \leq \frac{2}{\sqrt{N}}\mathbb{E}\|Y_N\| \ .$$

[1]Recall that X is L-sub-Gaussian if for every $t \in \mathbb{R}^d$ and every $p \geq 2$, $\|\langle X-\mu, t\rangle\|_{L_p} \leq L\sqrt{p}\|\langle X-\mu, t\rangle\|_{L_2}$.

(Also observe that by the central limit theorem, Y_N tends, in distribution, to the centred Gaussian random vector G that has the same covariance as X).

As for (1.4), if X is L-sub-Gaussian, then by a standard chaining argument combined with the majorizing measures theorem, one has that, with probability at least $1-\delta$,

$$\left\| \frac{1}{N}\sum_{i=1}^N X_i - \mu \right\| \leq \frac{c(L)}{\sqrt{N}} \left(\mathbb{E}\|G\| + \sqrt{\log(1/\delta)} \sup_{x^* \in \mathcal{B}^\circ} \left(\mathbb{E}(x^*(X-\mu))^2 \right)^{1/2} \right), \tag{1.5}$$

where again, G is the centred Gaussian vector that has the same covariance as X, $\mathcal{B}^\circ$ is the unit ball of the dual space[2] to $(\mathbb{R}^d, \|\ \|)$, and $c(L)$ is a constant that depends on L only.

Thus, if one believes that (1.3) and (1.4) should govern the error for a general mean estimation problem in $(\mathbb{R}^d, \|\cdot\|)$, one arrives to the following question:

Question 1. *Let $\|\cdot\|$, N and δ be as above. Does there exist an estimator $\widehat{\mu}_N$ (which may depend on δ and on the norm $\|\cdot\|$), such that, for all distributions whose covariance matrix exists, with probability at least $1-\delta$,*

$$\|\widehat{\mu}_N - \mu\| \leq \frac{c}{\sqrt{N}} \left(\max\left\{ \mathbb{E}\|Y_N\|,\ \ \mathbb{E}\|G\| + R\sqrt{\log(2/\delta)} \right\} \right), \tag{1.6}$$

where c is an absolute constant and

$$R = \sup_{x^* \in \mathcal{B}^\circ} \left(\mathbb{E}(x^*(X-\mu))^2 \right)^{1/2} \ ?$$

To put Question 1 is some perspective, let us consider the case of the Euclidean norm $\|\cdot\| = \|\cdot\|_2$ in $\mathbb{R}^d$. Let $\mathrm{Tr}(\Sigma)$ be the trace of the covariance matrix of X and set λ_1 to be the largest eigenvalue of Σ. Observe that

$$\mathbb{E}\|Y_N\|_2 \leq (\mathbb{E}\|Y_N\|_2^2)^{1/2} \leq (\mathbb{E}\|X-\mu\|_2^2)^{1/2} = \sqrt{\mathrm{Tr}(\Sigma)}\,,$$

and a similar bound holds for $\mathbb{E}\|G\|_2$, since Y_N and G share the same covariance matrix. Also, because the Euclidean norm is self-dual, $\mathcal{B}^\circ = B_2^d$, the Euclidean unit ball. Therefore,

$$R = \sup_{t \in B_2^d} \left(\mathbb{E}\langle t, X-\mu\rangle^2 \right)^{1/2} \leq \sqrt{\lambda_1}\,.$$

Hence, if Question 1 has an affirmative answer, the resulting mean estimation error

[2]Here and in what follows we identify linear functionals on $\mathbb{R}^d$ with points in $\mathbb{R}^d$, and the action of $t \in \mathbb{R}^d$ is given by $x^*(x) = \langle t, x\rangle$, that is, the standard inner product with t.

for the Euclidean norm would satisfy

$$\|\widehat{\mu}_N - \mu\|_2 \leq \frac{c}{\sqrt{N}}\left(\sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{\lambda_1 \log(2/\delta)}\right) \tag{1.7}$$

and with probability $1-\delta$. This coincides with the performance of the empirical mean if X is Gaussian (see [9]).

As it happens, (1.7) was established in [12] for an arbitrary random vector X (that has a well-defined mean and covariance) using the notion of median-of-means tournaments.

In Section 4 we argue that (1.6) is not far from the best (uniform) estimate one can ever hope for. For now simply observe that the term $N^{-1/2}R\sqrt{\log(2/\delta)}$ is truly required. Indeed, let X be a Gaussian random vector with mean μ. Observe that for any estimator $\widehat{\psi}_N$ and any $x^* \in \mathcal{B}^\circ$,

$$\|\widehat{\psi}_N - \mu\| \geq |x^*(\widehat{\psi}_N) - x^*(\mu)| \,.$$

Now fix $x^* \in \mathcal{B}^\circ$ and consider the random variable $x^*(X)$, which is a real-valued Gaussian whose mean is $x^*(\mu)$. If $\widehat{\psi}_N$ performs with accuracy ϵ with probability $1-\delta$ given $X_1, \ldots, X_N$, then the real-valued estimator $x^*(\widehat{\psi}_N)$ would perform with at least as good accuracy and confidence for the real-valued Gaussian variable $x^*(X)$. However, the results of [4] imply that the best possible accuracy for any mean estimator for a real valued Gaussian is $\sim N^{-1/2}\sigma\sqrt{\log(2/\delta)}$, and in our case, $\sigma^2 = \mathbb{E}(x^*(X-\mu))^2$. Taking the 'worst choice' of $x^* \in \mathcal{B}^\circ$ shows that

$$\epsilon \gtrsim \sup_{x^* \in \mathcal{B}^\circ} \left(\mathbb{E}(x^*(X-\mu))^2\right)^{1/2} \sqrt{\frac{\log(2/\delta)}{N}} = R\sqrt{\frac{\log(2/\delta)}{N}} \,.$$

Our main result is an affirmative answer to Question 1, and the mean estimator that achieves the desired accuracy is defined as follows. The estimator depends on the desired confidence $\delta \in (0,1)$ and also on an "accuracy parameter" $\epsilon > 0$. We show below that the procedure achieves accuracy ϵ whenever it is at least as large as the expression on the right-hand side on (1.6). For simplicity of presentation we assume that $n = \log(2/\delta)$ is an integer and that N is divisible by n. (Otherwise an obvious modification only effects the value of the unspecified constants so we do not lose any generality.)

- Set $\epsilon > 0$.
- Let $n = \log(2/\delta)$ and split the sample $(X_i)_{i=1}^N$ to n blocks I_j, each of cardinality N/n. Set $Z_j = \frac{1}{m}\sum_{i\in I_j} X_i$.
- Let T be the set of extreme points of the dual unit ball $\mathcal{B}^\circ$. For every $x^* \in T$ set

$$S_{x^*} = \left\{ y \in \mathbb{R}^d : |x^*(Z_j) - x^*(y)| \le \epsilon \right\} \text{ for more than } \frac{n}{2} \text{ blocks.}$$

- Set $\mathbb{S}(\epsilon) = \bigcap_{x^*\in T} S_{x^*}$ and select $\widehat{\mu}_N(\epsilon,\delta)$ to be any point in $\mathbb{S}(\epsilon)$.

Note that S_{x^*} is a union of intersections of shifts of the same 'slab' in $\mathbb{R}^d$, defined by the linear functional x^* and of 'width' ϵ. Thus, each intersection is just a (data dependent) slab, making S_{x^*} to be the union of slabs defined by x^*. As a result, $\mathbb{S}(\epsilon)$ is an intersection of unions of slabs generated by the extreme points of the dual unit ball of the given norm.

Our main result is the following—formulated using the notation introduced previously.

Theorem 1. *There exist absolute constants c, c' such that the following holds. Given a norm $\|\ \|$, confidence parameter $\delta \in (0,1)$ and sample size N, if*

$$\epsilon \ge \frac{c}{\sqrt{N}}\left(\max\left\{\mathbb{E}\|Y_N\|,\ \ \mathbb{E}\|G\| + R\sqrt{\log(2/\delta)}\right\}\right), \tag{1.8}$$

then the estimator $\widehat{\mu}_N(\epsilon,\delta)$ defined above satisfies that, with probability at least $1 - c'\delta$, $\mathbb{S}(\epsilon)$ is nonempty, and

$$\|\widehat{\mu}_N(\epsilon,\delta) - \mu\| \le \epsilon\ .$$

Remark. Observe that Theorem 1 also implies that if ϵ is as in (1.8) then the set $\mathbb{S}(\epsilon)$ is bounded. Moreover, the set is also closed because each S_{x^*} is closed.

The estimator $\widehat{\mu}_N(\epsilon,\delta)$ has the disadvantage that it requires the knowledge of the accuracy level ϵ. However, the achievable optimal accuracy depends on the distribution and it is generally unknown to the statistician. Luckily, it is easy to use the theorem above to construct an estimator that does not depend on such previous knowledge and yet achieves the same performance bound. We may simply define our estimator $\widetilde{\mu}_N = \widetilde{\mu}_N(\delta)$ as follows. Let $\epsilon_0 = \inf\{\epsilon > 0 : \mathbb{S}(\epsilon) \ne \emptyset\}$. The sets $\mathbb{S}(\epsilon)$ for $\epsilon > \epsilon_0$ are nested and compact and therefore $\cap_{\epsilon>\epsilon_0}\mathbb{S}(\epsilon) \ne \emptyset$. We define $\widetilde{\mu}_N$ to be an arbitrary element of $\cap_{\epsilon>\epsilon_0}\mathbb{S}(\epsilon)$. It follows from Theorem 1 that for ϵ satisfying (1.8), with probability at least $1-\delta$, $\mathbb{S}(\epsilon) \ne \emptyset$, and in particular, $\widetilde{\mu}_N \in \mathbb{S}(\epsilon)$. Hence, we obtain the following.

Corollary 1. *There exist absolute constants c, c' such that the following holds. Given a norm, confidence parameter $\delta \in (0,1)$ and sample size N, if*

$$\epsilon \geq \frac{c}{\sqrt{N}} \left(\max \left\{ \mathbb{E}\|Y_N\|, \ \ \mathbb{E}\|G\| + R\sqrt{\log(2/\delta)} \right\} \right) , \tag{1.9}$$

then the estimator $\widetilde{\mu}_N$ satisfies that, with probability at least $1 - c'\delta$,

$$\|\widetilde{\mu}_N - \mu\| \leq \epsilon \ .$$

Theorem 1 is established using a general fact that is of independent interest: we construct an effective *uniform* median-of-means estimator in a class of real valued functions, as described in the next section.

Related work

The multivariate median-of-means estimators that behave well under heavy-tailed distributions have been the subject of intensive study. Minsker [14] and Hsu and Sabato [7] defined and analyzed multivariate extensions of the median-of-means estimator, see also Lerasle and Oliveira [11]. The first truly sub-Gaussian estimator (under the Euclidean norm) was shown to exist by Lugosi and Mendelson [12]. See Joly, Lugosi, and Oliveira [9] for an earlier attempt and Catoni and Giulini [5] for a different estimator.

Minsker [15] and Catoni and Giulini [5] consider estimating the mean of random matrices based on an i.i.d. sample under the spectral norm and the Hilbert-Schmidt norm. They both prove sub-Gaussian performance bounds but the bounds of these papers fall short, in various aspects, of the optimal order of magnitude achieved by the estimator of Theorem 1 above. As far as we know, estimators achieving the accuracy/tradeoff of Theorem 1 have only been known for the Euclidean norm.

2 Uniform median-of-means estimators

In this section we explore the next problem:

> Let F be a class of functions on a probability space (Ω, ν) and let $\delta \in (0,1)$. Given an independent sample $(X_1, \ldots, X_N)$ distributed according to ν^N, find an estimator $\widehat{\Phi}_N$, such that, with probability at least $1 - \delta$, for every $f \in F$, $|\widehat{\Phi}_N(f) - \mathbb{E}f|$ is small.

The obvious choice of $\widehat{\Phi}_N$ is simply the standard median-of-means estimator we use for a single random variable. However, expecting $\widehat{\Phi}_N$ to have the

'individual' sub-Gaussian error is too optimistic. The best uniformly achievable accuracy must depend on some appropriate notion of the 'size' of the class F.

To address the problem above, fix integers n and m and let $N = mn$. As before, we split the given sample to n blocks, each one of cardinality n, while keeping in mind that the natural choice is $n \sim \log(2/\delta)$. Our goal is to find the smallest possible value of r such that $\sup_{f\in F}|\widehat{\Phi}_N(f) - \mathbb{E}f| \leq r$ with probability at least $1-\delta$.

Recall that if one would like to ensure that the median-of-means estimator performs with an error of at most r for a *single function* $f \in F$, then it suffices that

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} f(X_i) - \mathbb{E}f\right| \geq r\right) \leq \frac{1}{2} - \theta \tag{2.1}$$

for some $\theta > 0$. Indeed, if (2.1) holds then with probability at least $1-2\exp(-c\theta^2 n)$,

$$\mathbb{E}f - r \leq \frac{1}{m}\sum_{i\in I_j} f(X_i) \leq \mathbb{E}f + r$$

for more than $n/2$ of the blocks I_j, where c is an absolute constant. However, a uniform result calls for a little more flexibility. Firstly, there is a need to have a larger number of 'good' blocks I_j. It suffices that for any fixed function one controls $0.9n$ of them. Clearly, that may be achieved if (2.1) holds for $1/2 - \theta \leq 0.05$. With that in mind, let

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

From here on we write at times p_m instead of $p_m(\eta)$. We set D to be the unit ball in $L_2(\nu)$ and let $\mathcal{M}(F, rD)$ be the maximal cardinality of a subset of F that is r-separated with respect to the $L_2(\nu)$ norm. We also denote $F - F = \{f_1 - f_2 : f_1, f_2 \in F\}$.

Let us describe the performance of the uniform median-of-means estimator:

Theorem 2. *There exist absolute constants $c_0, \ldots, c_4$ 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) $p_m(\eta_0) \leq 0.05$;

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

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

where $W = (F-F)\cap\eta_1 D$ and $\overline{W} = \{w - \mathbb{E}w : w \in W\}$.

Let $r = \eta_0 + \eta_2$. Then with probability at least $1 - 2\exp(-c_4 n)$, for any $f \in F$ one has that

$$\left| \frac{1}{m} \sum_{i \in I_j} f(X_i) - \mathbb{E}f \right| \leq r \quad \text{for at least } 0.6n \text{ blocks } I_j \, .$$

To put Theorem 2 in some perspective, note that η_0 captures the worst individual error caused by a function in F. Moreover, as noted previously, the standard median-of-means estimator would perform with accuracy η_0 and confidence $1 - \delta$ if

$$\mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^{m} f(X_i) - \mathbb{E}f \right| \geq \eta_0 \right) \leq 0.05,$$

and by Chebyshev's inequality, one may set

$$\eta_0 \gtrsim \left(\mathbb{E}(f(X) - \mathbb{E}f)^2 \right)^{1/2} \cdot \frac{1}{\sqrt{m}} \sim \left(\mathbb{E}(f(X) - \mathbb{E}f)^2 \right)^{1/2} \cdot \sqrt{\frac{\log(2/\delta)}{N}},$$

as one would expect from a sub-Gaussian estimate.

In contrast, the role of η_2 is to calibrate the impact of the 'size' of F.

Proof. Fix $f \in F$ and let δ_j be the indicator of the event $\left| \frac{1}{m} \sum_{i \in I_j} f(X_i) - \mathbb{E}f \right| \geq \eta_0$. By a standard binomial tail estimate, for $k \geq 0.06n$,

$$\mathbb{P}\Big(\big| \{ j : \delta_j = 1 \} \big| \leq k \Big) \geq 1 - 2\exp(-c_0 k \log(ek/p_m n)) \, .$$

In particular, $|\{j : \delta_j = 1\}| \leq 0.1n$ with probability at least $1 - 2\exp(-c_1 n \log(e/p_m))$.

The importance of the high-probability estimate is seen in the next step of the proof: one may control all the elements of an η_1-net of F (with respect to the $L_2(\nu)$ norm) as long as its cardinality is at most $\exp(c_2 n \log(e/p_m))$. Indeed, by the union bound, with probability at least $1 - 2\exp(-c_3 n \log(e/p_m))$, for every h in the net there are at least $0.9n$ blocks I_j such that

$$\left| \frac{1}{m} \sum_{i \in I_j} h(X_i) - \mathbb{E}h \right| \leq \eta_0 \, .$$

The final and crucial step in the proof is passing from the net to the entire class: for every $f \in F$ set πf to be the best approximation to f in the net. Thus, $\|f - \pi f\|_{L_2} \leq \eta_1$. We show that for every $f \in F$ there are at most $0.2n$ blocks I_j such that

$$\left| \frac{1}{m} \sum_{i \in I_j} (f(X_i) - \mathbb{E}f) - \frac{1}{m} \sum_{i \in I_j} (\pi f - \mathbb{E}\pi f)(X_i) \right| \leq \eta_2. \tag{2.2}$$

If that is indeed the case then for every $f \in F$ there are at least $0.7n$ blocks for which

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i\in I_j} f(X_i) - \mathbb{E}f\right| \\
&\quad\leq \left|\frac{1}{m}\sum_{i\in I_j}(\pi f)(X_i) - \mathbb{E}\pi f\right| + \left|\frac{1}{m}\sum_{i\in I_j}(f-\mathbb{E}f)(X_i) - \frac{1}{m}\sum_{i\in I_j}(\pi f - \mathbb{E}\pi f)(X_i)\right| \\
&\quad\leq \eta_0 + \eta_2 \,,
\end{aligned}$$

as required.

It remains to prove (2.2). To this end, note that $f - \pi f \in (F-F)\cap \eta_1 D = W$, and thus $f - \mathbb{E}f - (\pi f - \mathbb{E}\pi f) \in \overline{W}$ were $\overline{W} = \{w - \mathbb{E}w : w \in W\}$. Hence, the proof is completed once it is established that, with probability at least $1 - 2e^{-c_4 n}$,

$$S \stackrel{\text{def.}}{=} \sup_{w\in\overline{W}} \left|\left\{j : \left|\frac{1}{m}\sum_{i\in I_j} w(X_i)\right| \geq \eta_2\right\}\right| \leq 0.2n \,.$$

To control S, note that by the bounded differences inequality (see, e.g., [3]) there is an absolute constant c_1 such that

$$\mathbb{P}(S \geq \mathbb{E}S + 0.1n) \leq 2\exp(-c_1 n) \,.$$

Thus, all that remains is to show that $\mathbb{E}S \leq 0.1n$. Observe that for any $(a_j)_{j=1}^n$,

$$|\{j : |a_j| \geq \eta\}| = \sum_{j=1}^n \mathbb{1}_{\{|a_j|\geq\eta\}} \leq \frac{1}{\eta}\sum_{j=1}^n |a_j| \,.$$

Hence, by standard methods of empirical processes, via an analogous argument to

that in [12], one has

$$\begin{aligned}
&\mathbb{E}\sup_{w\in\overline{W}}\left|\left\{j:\left|\frac{1}{m}\sum_{i\in I_j}w(X_i)\right|\geq\eta_2\right\}\right| \\
&\leq \frac{1}{\eta_2}\mathbb{E}\sup_{w\in\overline{W}}\sum_{j=1}^{n}\left|\frac{1}{m}\sum_{i\in I_j}w(X_i)\right| \\
&\leq \frac{1}{\eta_2}\mathbb{E}\sup_{w\in\overline{W}}\sum_{j=1}^{n}\left(\left|\frac{1}{m}\sum_{i\in I_j}w(X_i)\right|-\mathbb{E}\left|\frac{1}{m}\sum_{i\in I_j}w(X_i)\right|\right)+\frac{n}{\eta_2}\sup_{w\in\overline{W}}\mathbb{E}\left|\frac{1}{m}\sum_{i\in I_j}w(X_i)\right| \\
&\leq \frac{2}{\eta_2}\mathbb{E}\sup_{w\in\overline{W}}\left|\sum_{j=1}^{n}\varepsilon_j\left(\frac{1}{m}\sum_{i\in I_j}w(X_i)\right)\right|+\frac{n}{\eta_2}\cdot\sup_{w\in\overline{W}}\frac{\|w\|_{L_2}}{\sqrt{m}} \\
&\leq \frac{4n}{\eta_2}\left(\mathbb{E}\sup_{w\in\overline{W}}\left|\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i w(X_i)\right|+\frac{\eta_1}{\sqrt{m}}\right).
\end{aligned}$$

In particular, $\mathbb{E}S\leq 0.1n$ provided that

$$\mathbb{E}\sup_{w\in\overline{W}}\left|\sum_{i=1}^{N}\varepsilon_i w(X_i)\right|\leq c_2\eta_2 N \quad\text{and}\quad \frac{\eta_1}{\sqrt{m}}\leq c_3\eta_2\,,$$

as we assumed. ∎

Remark. Note that if F is a finite class and $\log|F|\leq c_2 n\log(e/p_m(\eta_0))$ then $\widehat{\Phi}_N$ performs with accuracy η_0. The proof follows from the standard bound on the performance of the median-of-means estimator for each real random variable $f(X)$ and a straightforward application of the union bound.

3 Estimation with respect to a general norm

In this section we establish Theorem 1 by invoking Theorem 2.

Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ and let $\mathcal{B}^\circ$ be the unit ball of the dual norm. Recall that for any $v\in\mathbb{R}^d$,

$$\|v\|=\sup_{x^*\in\mathrm{ext}(\mathcal{B}^\circ)}x^*(v)\,,$$

where $\mathrm{ext}(\mathcal{B}^\circ)$ denotes the set of extreme points in $\mathcal{B}^\circ$, and that the empirical average within block I_j, for $1\leq j\leq n$, is denoted by

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

Let $r > 0$ be as in Theorem 2 for the class of functions $F = \{x^*(\cdot) : x^* \in \mathrm{ext}(\mathcal{B}^\circ)\}$ and with the respect to the measure ν endowed by $X - \mu$. Finally, let $\mathcal{A}$ be the event for which the assertion of Theorem 2 holds.

Consider the sets

$$S_{x^*} = \left\{ y \in \mathbb{R}^d : \left| x^*(Z_j) - x^*(y) \right| \le r \ \text{ for more than } \frac{n}{2} \text{ indices } j \right\}$$

and put $\widehat{\mu}_N(\epsilon, \delta)$ to be any point that belongs to the set

$$\mathbb{S}(\epsilon) = \bigcap_{x^* \in \mathrm{ext}(\mathcal{B}^\circ)} S_{x^*} \,. \tag{3.1}$$

To show that selecting $\widehat{\mu}_N(\epsilon, \delta) \in \mathbb{S}(\epsilon)$ has the desired properties, fix a sample $(X_i)_{i=1}^N \in \mathcal{A}$. First, observe that $\mathbb{S}(\epsilon)$ is nonempty as it contains μ. Indeed, setting $f(x) = x^*(x)$, it is evident that

$$\mathbb{E} f(X - \mu) = 0 \quad \text{and} \quad \frac{1}{m} \sum_{i \in I_j} f(X_i - \mu) = x^*(Z_j) - x^*(\mu) \,.$$

By Theorem 2 it follows that

$$|x^*(Z_j) - x^*(\mu)| \le r$$

for a majority of the indices j, which means that $\mu \in S_{x^*}$ for every $x^* \in \mathrm{ext}(\mathcal{B}^\circ)$.

Next, one has to show that if $y \in \mathbb{S}(\epsilon)$, then $\|y - \mu\|$ is 'small'. To that end, observe that for every $x^* \in \mathrm{ext}(\mathcal{B}^\circ)$ there is some index j such that

$$\left| x^*(Z_j) - x^*(y) \right| \le r \quad \text{and} \quad \left| x^*(Z_j) - x^*(\mu) \right| \le r \,,$$

because both conditions hold for more than half of the indices j. Thus,

$$\left| x^*(y) - x^*(\mu) \right| \le \left| x^*(Z_j) - x^*(y) \right| + \left| x^*(Z_j) - x^*(\mu) \right| \le 2r \,.$$

Finally, recalling that $\|v\| = \sup_{x^* \in \mathrm{ext}(\mathcal{B}^\circ)} x^*(v)$, one has that

$$\|y - \mu\| = \sup_{x^* \in \mathrm{ext}(\mathcal{B}^\circ)} |x^*(y) - x^*(\mu)| \le 2r \,,$$

as claimed.

To complete the proof of Theorem 1 let us bound η_0 and η_2. To that end, recall that

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

G is the centred Gaussian vector that has the same covariance as X, and

$$Y_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i (X_i - \mu) \, .$$

We show that the three conditions of Theorem 2 can be controlled when $F = \{x^*(\cdot) : x^* \in \mathrm{ext}(\mathcal{B}^\circ)\}$ and with respect to the measure ν endowed by $\overline{X} = X - \mu$.

To verify (1), fix $x^* \in \mathcal{B}^\circ$ and note that

$$\mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^{m} x^*(X_i - \mu) \right| \geq \eta_0 \right) \leq \frac{\mathbb{E}|x^*(X-\mu)|^2}{\eta_0^2 m} = \frac{\mathbb{E}|x^*(X-\mu)|^2 \log(2/\delta)}{\eta_0^2 N}$$

where we have used the fact that $n = \log(2/\delta)$. (Recall that we assume, without loss of generality, that $\log(2/\delta)$ is an integer that divides N.) Thus, to ensure that $p_m \leq 0.05$ it suffices that

$$\eta_0 \geq c_0 R \sqrt{\frac{\log(2/\delta)}{N}} \, .$$

Turning to (2), we identify $\mathcal{B}^\circ$ with the set $\{t \in \mathbb{R}^d : \sup_{x \in \mathcal{B}} \langle t, x \rangle \leq 1\}$, and the action of a functional x^* associated with t is given by $x*(x) = \langle t, x \rangle$. We also abuse notation and denote by D the unit ball of the $L_2(\overline{X})$ norm endowed on $\mathbb{R}^d$ by identifying each $t \in \mathbb{R}^d$ with a linear functional. By Sudakov's inequality (see [10]), there is an absolute constant c such that

$$\log \mathcal{M}(\mathcal{B}^\circ, \eta_1 D) \leq c \eta_1^{-2} (\mathbb{E} \sup_{x^* \in \mathcal{B}^\circ} x^*(G))^2 = c \left(\frac{\mathbb{E}\|G\|}{\eta_1} \right)^2 \, ,$$

implying that one may set

$$\eta_1 = c_1 \frac{\mathbb{E}\|G\|}{\sqrt{n}} \, .$$

In particular, this forces the constraint

$$\eta_2 \geq c_2 \frac{\mathbb{E}\|G\|}{\sqrt{N}} \, .$$

Finally, to control (3), observe that

$$W \subset \{\langle t, \cdot \rangle : t \in 2\mathcal{B}^\circ \cap \eta_1 D\} \, .$$

Therefore, one has to show that

$$\mathbb{E} \sup_{t \in 2\mathcal{B}^\circ \cap \eta_1 D} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \langle t, (X_i - \mu) \rangle \right| = \mathbb{E} \sup_{t \in 2\mathcal{B}^\circ \cap \eta_1 D} |\langle t, Y_N \rangle| \leq \eta_2 \sqrt{N} \, ,$$

where $Y_N = N^{-1/2} \sum_{i=1}^N \varepsilon_i (X_i - \mu)$. Clearly, it suffices that

$$\eta_2 \geq c_3 \frac{\mathbb{E}\|Y_N\|}{\sqrt{N}} \,, \tag{3.2}$$

and one may set

$$\eta_2 = \frac{c_4}{\sqrt{N}} \max\left\{\mathbb{E}\|Y_N\|, \mathbb{E}\|G\|\right\} \,.$$

Now Theorem 1 follows from Theorem 2. ∎

Remark. Note that the choices of η_0, η_1 and η_2 need not be optimal for each F and $\overline{X}$ as above. Indeed, η_1 was chosen via Sudakov's inequality which is not always sharp, and η_2 was determined after the 'localization' $2\mathcal{B}^\circ \cap \eta_1 D$ was replaced by $2\mathcal{B}^\circ$. Therefore, it stands to reason that there are cases in which the resulting estimate may be improved with more care. However, as we explain in the next section, Theorem 1 is likely to be the best uniform result that one can hope for.

4 Lower bounds

In this section we discuss the optimality of the upper bound of Theorem 1. In the introduction we already pointed out that the term $N^{-1/2} R \sqrt{\log(2/\delta)}$ is inevitable. Here we discuss the necessity of the term $N^{-1/2}\mathbb{E}\|G\|$ (and, equivalently, the term $N^{-1/2}\mathbb{E}\|Y_N\|$ since $\lim_{N\to\infty} \mathbb{E}\|Y_N\| = \mathbb{E}\|G\|$ by the central limit theorem).

Here we show that the order of magnitude of the bound of Theorem 1 is essentially un-improvable even if one only considers isotropic Gaussian distributions, with one minor caveat. Recall that the proof of Theorem 1 uses Sudakov's inequality to ensure that

$$\log \mathcal{M}(\mathcal{B}^\circ, \eta_1 D) \lesssim n \,, \tag{4.1}$$

and then the contribution to the error is $\sim \eta_1/\sqrt{m} = \eta_1 \sqrt{n/N}$. As noted previously, while it is convenient to use Sudakov's inequality, its application may be loose. A more accurate upper estimate on the error is $\sim \eta_1/\sqrt{m}$ where η_1 is the smallest value for which (4.1) holds.

We show now that if X is a Gaussian measure whose covariance is the identity matrix, then this more accurate upper estimate is actually a lower bound as well.

To formulate the lower bound, let G be the standard Gaussian random vector in $\mathbb{R}^d$ and denote by ν the corresponding measure. We study the performance of an arbitrary mean estimation procedure with respect to a norm $\|\cdot\|$, for a collection of Gaussian measures endowed by $\{G + t : t \in T\}$ and where $T \subset \mathbb{R}^d$ is a well chosen set. Note that for any $X = G + t$ one has that $\overline{X} = G$, let $D \subset \mathbb{R}^d$ be the unit

ball endowed by the norm $L_2(\overline{X}) = L_2(\nu)$ (which in this case is simply B_2^d), and set $\mathcal{B}$ to be the unit ball of $(\mathbb{R}^d, \|\cdot\|)$.

Theorem 3. *There exist absolute constants c_1 and c_2 for which the following holds. Let $n \le N$ and assume that $\log \mathcal{M}(\mathcal{B}^\circ, \eta D) \ge c_1 n$. There is a set $T \subset \mathbb{R}^d$ such that any mean estimator $\widehat{\Psi}_N$ that performs with confidence $1/2$ with respect to all the Gaussian measures $\{G + t : t \in T\}$, cannot perform with higher accuracy than $c_2 \eta \sqrt{n/N}$.*

Remark. An immediate outcome of Theorem 3 is that when Sudakov's inequality is sharp at scale η, the lower bound on the accuracy that holds with constant confidence is indeed $\sim \mathbb{E}\|G\|/\sqrt{N}$.

Proof. To define the set T, observe that if $\log \mathcal{M}(\mathcal{B}^\circ, \eta D) \ge c_1 n$ then by the duality theorem of metric entropy [2], and since $D^\circ = B_2^d$, it follows that $\log \mathcal{M}(D, c_2 \eta \mathcal{B}) \ge c_3 n$. In other words, the set D contains a subset that is $c_2\eta$-separated with respect to the norm $\|\ \|$ and whose cardinality is $c_3 n$. Set $R = \sqrt{n/N} = 1/\sqrt{m}$ and $r = c_2\eta/\sqrt{m}$. Clearly,

$$\log \mathcal{M}(D, c_2\eta\mathcal{B}) = \log \mathcal{M}(D, (r/R)\mathcal{B}) = \log \mathcal{M}(RD, r\mathcal{B}) \ge c_3 n \,,$$

and let $T \subset RD$ be the r-separated set with respect to the norm $\|\ \|$.

Now, assume that there is a mean estimator that performs with confidence $1/2$ and accuracy $r/3$ for every one of the Gaussian random vector $G + t$, $t \in T$, and let as reach a contradiction.

If we denote by $\widehat{\Psi}_N$ such an estimator, it follows that for every $t \in T$ there is a set $\mathcal{A}_t \subset (\mathbb{R}^d)^N$ with $\nu^N(A_t) \ge 1/2$, such that for $\omega \in \mathcal{A}_t + (t, \dots, t)$ one has that $\|\widehat{\Psi}_N(\omega) - t\| \le r/3$. Moreover, since the set T is r-separated, it is evident that the sets $U_t = \mathcal{A}_t + (t, \dots, t)$ are disjoint. Indeed, if $\omega \in (\mathcal{A}_x + (x, \dots, t)) \cap (\mathcal{A}_y + (y, \dots, y))$ then $\widehat{\Psi}_N$ would have to be 'close' to both x and y, which is impossible.

Observe that ν^N is the standard Gaussian measure on $(\mathbb{R}^d)^N$, and $\|(t, \dots, t)\|_{L_2(\nu^N)} = \sqrt{N}\|t\|_2 \le \sqrt{N}R$. Using the same argument as in Talagrand's proof of the dual Sudakov inequality [10] (see also [13]) and recalling that $t \in RD = RB_2^d$,

$$\nu^N(U_t) \ge \nu^N(\mathcal{A}_t) \exp(-cN\|t\|_2^2) \ge \frac{1}{2}\exp(-cNR^2) \,,$$

for an absolute constant c.

On the other hand, the sets U_t are disjoint and thus

$$1 \ge \nu^N\left(\bigcup_{t\in T} U_t\right) = \sum_{t\in T} \nu^N(U_t) \ge c|T|\exp(-c'NR^2) \,.$$

In particular,

$$\log|T| \le c''NR^2 = c''n \,,$$

which is a contradiction to our choice of T. ∎

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] Shiri Artstein, Vitali Milman, and Stanisław J Szarek. Duality of metric entropy. *Annals of Mathematics*, pages 1313–1328, 2004.

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

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

[5] Olivier Catoni and Ilaria Giulini. Dimension-free pac-bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*, 2017.

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

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

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

[9] 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.

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

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

[12] G. Lugosi and S. Mendelson. Sub-Gaussian estimators of the mean of a random vector. *Annals of Statistics*, 2018, to appear.

[13] S. Mendelson. "Local" vs. "global" parameters—breaking the Gaussian complexity barrier. *Ann. Statist.*, 45(5):1835–1862, 2017.

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

[15] S. Minsker. Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *arXiv preprint arXiv:1605.07129*, 2016.

[16] A.S. Nemirovsky and D.B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, 1983.