Multivariate mean estimation with direction-dependent accuracy [*†]

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

October 22, 2020

Abstract

We consider the problem of estimating the mean of a random vector based on N independent, identically distributed observations. We prove the existence of an estimator that has a near-optimal error in all directions in which the variance of the one dimensional marginal of the random vector is not too small: with probability $1-\delta$, the procedure returns $\widehat{\mu}_N$ which satisfies that for every direction $u \in S^{d-1}$,

$$\langle \widehat{\mu}_N - \mu, u \rangle \leq \frac{C}{\sqrt{N}} \left(\sigma(u) \sqrt{\log(1/\delta)} + \left(\mathbb{E} \|X - \mathbb{E}X\|_2^2 \right)^{1/2} \right),$$

where $\sigma^2(u) = \mathrm{Var}(\langle X, u \rangle)$ and C is a constant. To achieve this, we require only slightly more than the existence of the covariance matrix, in the form of a certain moment-equivalence assumption.

The proof relies on novel bounds for the ratio of empirical and true probabilities that hold uniformly over certain classes of random variables.

[*]Gábor Lugosi was supported by the Spanish Ministry of Economy and Competitiveness, Grant MTM2015-67304-P and FEDER, EU, and by "Google Focused Award Algorithms and Learning for AI".

[†]Shahar Mendelson would like to thank Jungo Connectivity for its support.

[‡]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, shahar.mendelson@anu.edu.au

1 Introduction

The problem of estimating the mean of a high-dimensional random vector with a possibly heavy-tailed distribution is a classical question that has been studied extensively over the years. Recently it has received the attention of mathematical statisticians and theoretical computer scientists. We refer to Lugosi and Mendelson [18] for a recent survey and to Bahmani [1], Dalalyan and Minasyan [5], Diakonikolas, Kane, Pensia [8], Minsker and Mathieu [26], Minsker and Ndaoud [27] for a sample of even more recent references.

To formulate the problem, let $X_1, \dots, X_N$ be independent, identically distributed random vectors taking values in $\mathbb{R}^d$ with mean $\mathbb{E}X = \mu \in \mathbb{R}^d$ (where X is distributed as X_1.) One is interested in constructing a measurable function $\widehat{\mu}_N : (\mathbb{R}^d)^N \to \mathbb{R}^d$ such that $\widehat{\mu}_N = \widehat{\mu}_N(X_1, \dots, X_N)$ is close, in some sense, to the mean μ. A possible meaningful goal is to find an estimator such that, given a confidence parameter δ, satisfies that with probability at least $1 - \delta$, the Euclidean distance $\|\widehat{\mu}_N - \mu\|$ is as small as possible.

Constructing an optimal $\widehat{\mu}_N$ is not that obvious even when $d = 1$. Without any further assumptions on the distribution of the X_i it is impossible to give meaningful performance guarantees even in that case, let alone for a random vector in $\mathbb{R}^d$. However, under minimal conditions, it is possible to construct estimators with remarkably strong properties. The most common assumption is that the X_i have finite second moment. When $d = 1$ one can find $\widehat{\mu}_N(\delta)$ such that, with probability at least $1 - \delta$,

$$|\widehat{\mu}_N(\delta) - \mu| \leq c\sigma \sqrt{\frac{\log(1/\delta)}{N}}\,, \tag{1.1}$$

where σ^2 is the variance of X and c is a universal constant (see, e.g., [18]).

The meaning of (1.1) is that even if X is heavy-tailed, $\widehat{\mu}_N$ performs as if X were a Gaussian random variable and the estimator were the empirical mean $N^{-1}\sum_{i=1}^N X_i$. Obviously, the empirical mean does not exhibit such a behavior unless X is actually Gaussian (or sub-Gaussian), which indicates that $\widehat{\mu}_N$ has to be rather carefully chosen when X is an arbitrary random variable.

The problem for $d > 1$ is significantly more complex, though it does have a satisfying answer. Let $X \in \mathbb{R}^d$ and assume (as we do throughout this article) that the covariance matrix of the X_i, denoted by $\Sigma = \mathbb{E}(X - \mu)(X - \mu)^T$, exists. Quite remarkably, there are mean estimators that, under this minimal condition, achieve "sub-Gaussian" performance. In particular, for any $\delta \in (0,1)$, there exists an estimator $\widehat{\mu}_N = \widehat{\mu}_N(\delta)$ such that, with probability at least $1 - \delta$,

$$\|\widehat{\mu}_N - \mu\| \leq C\left(\sqrt{\frac{\lambda_1 \log(1/\delta)}{N}} + \sqrt{\frac{\sum_{i=1}^d \lambda_i}{N}}\right), \tag{1.2}$$

where C is a universal constant and $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_d \geq 0$ are the eigenvalues of the covariance matrix Σ, see Lugosi and Mendelson [19, 20], Hopkins [12], Cherapanamjeri, Flammarion, and Bartlett [4], Depersin and Lecué [7], Lei, Luh, Venkat, and Zhang [16], Diakonikolas, Kane, Pensia [8].

The reason for the term "sub-Gaussian" is, again, the comparison to what happens in the Gaussian case and the estimator being the empirical mean. Indeed, when the X_i have a multivariate normal distribution and $\widetilde{\mu}_N = (1/N)\sum_{i=1}^N X_i$ is the empirical mean, then (1.2) holds with probability at least $1-\delta$. This bound is of the correct order, since the expected value $\mathbb{E}\|\widetilde{\mu}_N - \mu\|$ is proportional to $\sqrt{\mathrm{Tr}(\Sigma)/N} = \sqrt{(1/N)\sum_{i=1}^d \lambda_i}$, and the term $\sqrt{(1/N)\lambda_1 \log(1/\delta)}$ bounds the fluctuations, using the Gaussian concentration inequality. We refer to the two terms on the right-hand side of (1.2) as the *weak* and *strong* terms, respectively: the strong term is simply the L_2 norm of $\|X - \mathbb{E}X\|_2$ and the weak term corresponds to the largest variance of a one dimensional marginal of X, that is, to $\sup_{u\in S^{d-1}} \sigma(u)$. Here $S^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ denotes the Euclidean unit sphere and $\langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^d$.

Remark. Strong-weak inequalities are an important notion in high-dimensional probability (see, e.g., Latała and Wojtaszczyk [14]). We explain the connection between our results and these inequalities in Appendix B.

Clearly, an equivalent way of writing the inequality (1.2) is as follows:

$$\forall u \in S^{d-1} : \langle \widehat{\mu}_N - \mu, u \rangle \leq C\left(\sqrt{\frac{\lambda_1 \log(1/\delta)}{N}} + \sqrt{\frac{\sum_{i=1}^d \lambda_i}{N}} \right). \tag{1.3}$$

It is reasonable to expect that (1.3) is the best "directional formulation" that one can hope for. Firstly, the error must have a global component, which is the strong term: directional information corresponds only to one-dimensional marginals of X, while higher dimensional marginals impact the ability of approximating the mean. At the same time, any standard way of controlling fluctuations is based on estimates on the worst direction, leading to the weak term which involves λ_1. With that in mind, a more fine-grained version of (1.3) seems unlikely.

Perhaps contrary to intuition, our main result does precisely that: under a mild additional assumption on X we construct an estimator that, up to the optimal strong term, preforms in every direction as if it were an optimal estimator of the one dimensional marginal:

Theorem 1. *Let $X_1, \dots, X_N$ be i.i.d. random vectors, taking values in $\mathbb{R}^d$, with mean μ and covariance matrix Σ whose eigenvalues are $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_d \geq 0$. Suppose that there exists $q > 2$ and a constant κ such that, for all $u \in S^{d-1}$,*

$$\left(\mathbb{E}\langle X - \mu, u\rangle^q\right)^{1/q} \leq \kappa \left(\mathbb{E}\langle X - \mu, u\rangle^2\right)^{1/2} . \tag{1.4}$$

Then for every $\delta \in (0,1)$ there exists a mean estimator $\widehat{\mu}_N$ and constants $0 < c, c', C < \infty$ (depending on κ and q only) such that, if $\delta \geq e^{-c'N}$, then, with probability, at least $1 - \delta$,

$$\forall u \in S^{d-1} : \langle \widehat{\mu}_N - \mu, u\rangle \leq C \left(\sqrt{\frac{\sigma^2(u)\log(1/\delta)}{N}} + \sqrt{\frac{\sum_{i=c\log(1/\delta)}^{d} \lambda_i}{N}} \right). \tag{1.5}$$

It should be stressed that a proof of such a result, that is sensitive to directions, calls for the development of a completely new machinery. In (1.3) one is allowed fluctuations at scale $\sqrt{\frac{\lambda_1 \log(1/\delta)}{N}}$ in all directions, and that plays a crucial component in the proof of [19, 20]. In contrast, inequality (1.5) calls for (uniform) control over all nontrivial directions but the allowed scales of fluctuations in each direction can be much smaller than $\sqrt{\frac{\lambda_1 \log(1/\delta)}{N}}$. That difference renders useless the methods used in the proof of (1.3).

The main technical novelty of this article is the machinery that leads to the necessary directional control. It is presented in Section 3. This machinery consists of bounds for *ratios* of empirical and true probabilities that hold uniformly in a class of functions. Informally put, we are able to control

$$\sup_{\{f \in \mathcal{F}, \|f\|_{L_2} \geq r\}} \sup_{t : \mathbb{P}\{f(X) > t\} \geq \Delta} \left| \frac{N^{-1} \sum_{i=1}^{N} \mathbb{1}_{f(X_i) > t}}{\mathbb{P}\{f(X) > t\}} - 1 \right|$$

for appropriate values of r and Δ.

In other words, we show that, under minimal assumptions on the class $\mathcal{F}$, the empirical frequencies of level sets of every $f \in \mathcal{F}$ are close, in a multiplicative sense, to their true probabilities—as long as $\|f\|_{L_2}$ and $\mathbb{P}\{f(X) > t\}$ are large enough. Estimates of this flavor have been derived before, but only in a limited scope. Examples include the classical inequalities of Vapnik-Chervonenkis in vc theory, dealing with small classes of binary-valued functions (see also, Giné and Koltchinskii [10] for some results for real-valued classes). Existing ratio estimates are often based on the highly restrictive assumption that the collection of level sets, say of the form $\{\{x : f(x) > t\} : f \in \mathcal{F},\ t \geq t_0\}$, is small in the vc sense. Unfor-

tunately, in the general context that is required here, such an assumption need not hold.

The new method we develop here is based on a completely different argument that builds on the so-called *small-ball method*. The relevant ratio estimate can be found in Theorem 3.

While the main thrust in (1.5) is the directional dependence, it is worth noting that the strong term is better than $(\mathbb{E}\|X - \mathbb{E}X\|_2^2)^{1/2}$. To put this in perspective, consider again the example when the distribution of the data is Gaussian and the estimation procedure is the empirical mean $\widetilde{\mu}_N = (1/N)\sum_{i=1}^N X_i$. The next observation shows that even in this well-behaved example, the strong term needs to be at least of the order

$$\sqrt{\frac{\sum_{i>k}\lambda_i}{N}}$$

where k is proportional to $\log(1/\delta)$.

Proposition 1. *Let $\widetilde{\mu}_N = (1/N)\sum_{i=1}^N X_i$ where the X_i are independent Gaussian vectors with mean μ and covariance matrix Σ. Suppose that there exists a constant C such that, for all δ, N, μ, and Σ, with probability at least $1-\delta$,*

$$\forall u \in S^{d-1} : \langle \widetilde{\mu}_N - \mu, u\rangle \le C\sqrt{\frac{\sigma^2(u)\log(1/\delta)}{N}} + S\ . \tag{1.6}$$

Then there exists a constant C' depending on C only, such that the "strong term" S has to satisfy

$$S \ge C'\sqrt{\frac{\sum_{i>k_0}\lambda_i}{N}}$$

where $k_0 = 1 + (2C+\sqrt{2})^2\log(1/\delta)$.

The proof is given in Section A.1 of the Appendix.

Remark. An easy modification of the proof of Proposition 1 reveals that the lower bound is tight in the sense that, if the X_i are independent Gaussians, then the empirical mean $\widetilde{\mu}_N$ satisfies that, with probability, at least $1-\delta$,

$$\forall u \in S^{d-1} : \langle \widetilde{\mu}_N - \mu, u\rangle \le C\left(\sqrt{\frac{\sigma^2(u)\log(1/\delta)}{N}} + \sqrt{\frac{\sum_{i>k_1}\lambda_i}{N}}\right)$$

where $k_1 = c\log(1/\delta)$, for some constants c and C.

The article is organized as follows. In the next two sections we develop the technical machinery we require. First, in Section 2 we recall — and appropriately modify — the solution of Mendelson [23] of the following moment estimation

problem: given a real random variable Z, find an almost isometric, data dependent estimator of $\mathbb{E}|Z|^p$. In other words, find $\widehat{Z}_p$ such that, $(1-\varepsilon)\mathbb{E}|Z|^p \le \widehat{Z}_p \le (1+\varepsilon)\mathbb{E}|Z|^p$, with probability $1-\delta$.

The analysis of the moment estimation problem reveals the importance of uniform control of ratios of empirical and true probabilities, and that is the subject of Section 3 where the new general inequality is proven.

The next component in the proof of Theorem 1 is a covariance estimator (i.e., an estimator of all the directional variances $\sigma^2(u)$), which we introduce and analyze in Section 4. *Covariance estimation* has received quite a lot of attention lately, see, for example, Catoni [3], Giulini [11]. Koltchinskii and Lounici [13], Lounici [17], Mendelson [22], Mendelson and Zhivotovksiy [21], Minsker [25], Minsker and Wei [28].

The notion of estimation needed here is rather weak: we only require that the variances are approximated within a multiplicative constant factor. Our construction is based on the moment estimators developed in Section 2 and in Section 3: we construct an estimator $\psi_N(u)$ of the directional variances such that, with probability at least $1-\delta$, simultaneously for all u such that $\sigma^2(u) \ge \sum_{i>c\log(1/\delta)} \lambda_i$, all estimated variances $\psi_N(u)$ satisfy

$$\frac{1}{4} \le \frac{\psi_N(u)}{\sigma^2(u)} \le 2\,.$$

Moreover, for all other u, we have $\psi_N(u) \le C\sum_{i>c\log(1/\delta)} \lambda_i$.

Remark. It should be noted that covariance estimators, quite different in nature, but with performance bounds of the same spirit, were defined by Catoni [3] and Giulini [11], under certain fourth-moment assumptions.

Finally, in Section 5 we define the main multivariate mean estimation procedure and prove Theorem 1. Some of the proofs are relegated to the Appendix.

Remark. (NONATOMIC DISTRIBUTIONS.) To avoid unimportant but somewhat tedious technicalities, we assume throughout that the distribution of the X_i is absolutely continuous with respect to the Lebesgue measure. This implies that the distribution of $\langle X, u\rangle$ is nonatomic for all $u \in S^{d-1}$ and we do not need to worry about multiple points taking the same value—which makes the definition of trimming and quantiles simpler. This assumption is not restrictive because one may always add a tiny random perturbation to each data point, converting the distribution absolutely continuous, and without changing the mean vector too much.

Remark. (HILBERT SPACES.) For convenience, we present our results for random variables taking values in $\mathbb{R}^d$. However, the main results remain true without modification when X takes values in any separable Hilbert space $\mathcal{H}$. The only condition that needs to be modified is absolute continuity. It suffices that $\langle u, X\rangle$ has a continuous distribution for all $u \in \mathcal{H}\setminus\{0\}$.

Remark. (COMPUTATION.) Our definition of a mean estimator as a (measurable) function of the data ignores important computational issues. An important branch of research has focused on computational issues of robust statistical estimation. In particular, mean estimators achieving sub-Gaussian performance of the type (1.2) that can be computed in polynomial time have been proposed, see Hopkins [12], Cherapanamjeri, Flammarion, and Bartlett [4], Depersin and Lecué [7], Lei, Luh, Venkat, and Zhang [16], Diakonikolas, Kane, Pensia [8]. It is an interesting open problem whether there exists an efficiently computable estimator achieving a performance like the one announced in Theorem 1.

2 Empirical tail integration and moment estimation

In this section we consider the problem of estimating the raw moments of a real random variable. The estimators studied here are trimmed estimators, that is, based on the empirical mean after discarding the smallest and largest values of the sample. Such estimators have been studied extensively (see Lugosi and Mendelson [18] for pointers to the literature). The properties presented in this section have been mostly developed by Mendelson [23] and the exposition below is a slight modification of the arguments in [23]. For completeness, we present the proofs in Section 2.1.

Note that the results in this section are for moments of any order $p \geq 1$, although we only need the special cases $p = 1$ and $p = 2$.

Let $p \geq 1$ and let Z be a real-valued random variable with continuous distribution such that $\mathbb{E}|Z|^p < \infty$. Let $Z_1^N = (Z_1, \ldots, Z_N)$ be a sample of N independent copies of Z. For fixed $1/N < \theta \leq 1/2$ let J_+ be the set of indices of the θN largest values of $(Z_i)_{i=1}^N$ and denote by J_- the set of indices of the θN smallest values. (Since Z is assumed to have a continuous distribution, J_+ and J_- are uniquely defined, with probability one.) Writing $[N] = \{1, \ldots, N\}$, our goal is to show that with high probability,

$$\Psi_{p,\theta}(Z_1^N) = \frac{1}{N} \sum_{i \in [N] \setminus (J_+ \cup J_-)} |Z_i|^p$$

is a good estimator of $\mathbb{E}|Z|^p$ for an appropriately chosen value of θ.

Denote the empirical measure of Z by $\mathbb{P}_N$, that is, for all measurable sets $A \subset \mathbb{R}$, we write

$$\mathbb{P}_N\{Z \in A\} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{Z_i \in A} \, .$$

We show that, in order to ensure that $\Psi_{p,\theta}$ is a good estimator, it suffices to control

the ratios $\mathbb{P}_N\{Z \in I\}/\mathbb{P}\{Z \in I\}$, for intervals[1] I. Moreover, we show that, under similar assumptions,

$$\Phi_\theta(Z_1^N) = \frac{1}{N} \sum_{i \in [N] \setminus (J_+ \cup J_-)} Z_i$$

is a good estimator of the mean $\mathbb{E}Z$.

Our starting point is the following definition that describes some properties that ensure that $\Psi_{p,\theta}$ and Φ_θ are well-behaved, in a sense made precise below.

Definition 1. *Set $0 < \Delta, \theta < 1/100$. Let $\mathcal{A}$ be the event on which the following properties hold:*

(1) *for all nonnegative integers j and $t \geq 0$ such that $2^{-j}\mathbb{P}\{Z > t\} \geq \Delta$, we have*

$$\left| \frac{\mathbb{P}_N\{Z > t\}}{\mathbb{P}\{Z > t\}} - 1 \right| \leq 2^{-j/2-1}$$

and if $2^{-j}\mathbb{P}\{Z < -t\} \geq \Delta$, then

$$\left| \frac{\mathbb{P}_N\{Z < -t\}}{\mathbb{P}\{Z < -t\}} - 1 \right| \leq 2^{-j/2-1} \ ;$$

(2) *for any interval $I \subset \mathbb{R}$,*

$$\mathbb{P}_N\{Z \in I\} \leq \frac{3}{2}\mathbb{P}\{Z \in I\} + 2\Delta \ ;$$

(3) $\mathbb{P}\{Z > 0\} \geq \eta$ *and* $\mathbb{P}\{Z < 0\} \geq \eta$ *for* $\eta = 4\theta + 16\Delta$.

Note that property (1) requires that the relative error of the empirical measure becomes smaller for larger sets. The idea behind the conditions of Definition 1 is that one may write

$$\mathbb{E}|Z|^p = \int_0^\infty p t^{p-1} \left(\mathbb{P}\{Z > t\} + \mathbb{P}\{-Z > t\}\right) dt \ ,$$

and with sufficient control of the ratios $\mathbb{P}_N\{Z > t\}/\mathbb{P}\{Z > t\}$ and $\mathbb{P}_N\{-Z > t\}/\mathbb{P}\{-Z > t\}$, the integral can be well approximated by an empirical functional like $\Psi_{p,\theta}$. That approximation can be valid even when the distortion is relatively large for sets $\{Z > t\}$ or $\{-Z > t\}$ whose measure is small, but a small distortion is essential for sets of relatively large measure, as such sets have a much higher impact on the integral.

[1] By an *interval* we mean open, closed, half-open intervals in $\mathbb{R}$, including rays.

The main technical fact we use is the following estimate on the positive an negative parts of Z, denoted throughout by Z_+ and Z_-, respectively. For any $\alpha \in (0,1)$, we denote by Q_α the α-quantile of the random variable Z, that is, Q_α is the unique value such that $\mathbb{P}\{Z \le Q_\alpha\} = \alpha$.

Theorem 2. *Let $p \ge 1$ and let Z be a real random variable with continuous distribution such that $\mathbb{E}|Z|^p < \infty$. Let $0 < \Delta, \theta < 1/100$ be such that θN is an integer, $\theta \ge 7\Delta$, and let $\mathcal{A}$ be the event defined in Definition 1. Set*

$$\theta_1 = 2\theta + 8\Delta\,, \qquad \text{and} \qquad \theta_2 = (2\theta - 8\Delta)/3\,.$$

Then, on the event $\mathcal{A}$, we have

(a) For every $i \in J_+$, $Z_i \ge 0$ and for every $i \in J_-$, $Z_i \le 0$. In particular, $Z_i = (Z_i)_+$ for all $i \in J_+$, and $Z_i = (Z_i)_-$ for all $i \in J_-$.

(b)

$$\frac{1}{N}\sum_{j\in[N]\setminus J_+} (Z_i)_+^p \le \mathbb{E}Z_+^p + 2\sqrt{\Delta}\int_0^{Q_{1-\theta_2}} pt^{p-1}\sqrt{\mathbb{P}\{Z>t\}}dt\,,$$

and

$$\frac{1}{N}\sum_{i\in[N]\setminus J_+} (Z_i)_+^p \ge \mathbb{E}Z_+^p - 3\mathbb{E}\left[Z^p \mathbb{1}_{\{Z\ge Q_{1-\theta_1}\}}\right] - 2\sqrt{\Delta}\int_0^{Q_{1-\theta_1}} 2t\sqrt{\mathbb{P}\{Z>t\}}dt\,.$$

(c) The analogous claim to (b) holds for Z_-, the negative part of Z.

The proof of Theorem 2 is given in Section 2.1. We remark here that, as it is shown in the proof, $Q_{1-\theta_1}, Q_{1-\theta_2} > 0$ on the event $\mathcal{A}$.

In order to estimate the p-th moment of a random variable Z, we simply write

$$\begin{aligned}\frac{1}{N}\sum_{i\in[N]\setminus(J_+\cup J_-)} |Z_i|^p &= \frac{1}{N}\sum_{i\in[N]\setminus(J_+\cup J_-)} (Z_i)_+^p + \frac{1}{N}\sum_{i\in[N]\setminus(J_+\cup J_-)} (Z_i)_-^p \\ &= \frac{1}{N}\sum_{i\in[N]\setminus J_+} (Z_i)_+^p + \frac{1}{N}\sum_{i\in[N]\setminus J_-} (Z_i)_-^p\,,\end{aligned}$$

where the last equality holds on the event $\mathcal{A}$ because of part (a) of Theorem 2. Now part (b) may be used to show that the two terms on the right-hand side are close to their means $\mathbb{E}Z_+^p$ and $\mathbb{E}Z_-^p$.

We show in Section 3 that properties (1) – (2) in Definition 1 are satisfied with high probability. Property (3) means that the random variable Z is sufficiently "balanced".

Next we derive a corollary of Theorem 2 that is more convenient to use. To this end, set

$$\mathcal{E}_{T,p} = 2\sqrt{\Delta}\int_0^T pt^{p-1}\sqrt{\mathbb{P}\{|Z|>t\}}dt\ .$$

The following lemma was established by Mendelson [23] for nonnegative-valued, absolutely continuous random variables. The extension to the case below (allowing an atom at zero) is straightforward; the proof is omitted.

Lemma 1. *There is an absolute constant c for which the following holds. Let $p \geq 1$ and $\kappa \in (0,1)$. Let Z be nonnegative random variable whose distribution is a mixture of a an absolutely continuous component and an atom at 0, such that $\mathbb{E}Z^{2p} < \infty$. Then*

$$\mathbb{E}\Big[Z^p \mathbb{1}_{\{Z>Q_{1-\kappa}\}}\Big] \leq \sqrt{\kappa}\Big(\mathbb{E}Z^{2p}\Big)^{1/2}\ ,$$

and

$$\mathcal{E}_{Q_{1-\kappa},p} \leq c\sqrt{\Delta}\sqrt{\log\Big(\frac{1}{\kappa}\Big)}\Big(\mathbb{E}Z^{2p}\Big)^{1/2}\ .$$

Moreover, if $\mathbb{E}Z^q < \infty$ for some $q > 2p$, then

$$\mathcal{E}_{Q_{1-\kappa},p} \leq c_{q,p}\sqrt{\Delta}\,(\mathbb{E}Z^q)^{p/q}\ ,$$

where $c_{q,p} = cp/(q-2p)$ for a numerical constant $c > 0$.

Combining Lemma 1 with Theorem 2 leads to the following corollary:

Corollary 1. *There are absolute constants $c_1,\ldots,c_4$ for which the following holds. Assume the conditions of Theorem 2. Set $c_1\frac{\log N}{N} \leq \Delta < 100$, and $\theta = c_2\Delta$ with $c_2 > 7$. Then*

$$\left|\frac{1}{N}\sum_{i\in[N]\setminus J_+}(Z_i)_+^p - \mathbb{E}Z_+^p\right| \leq c_3\sqrt{\Delta\log\Big(\frac{1}{\Delta}\Big)}\Big(\mathbb{E}Z^{2p}\Big)^{1/2}\ ,$$

and if $Z \in L_q$ for $q > 2p$ then

$$\left|\frac{1}{N}\sum_{i\in[N]\setminus J_+} f_+^p(X_i) - \mathbb{E}f_+^p\right| \leq c_{q,p}\sqrt{\Delta}\,(\mathbb{E}Z^q)^{p/q}$$

where $c_{q,p} = c_4 p/(q-2p)$.

The analogous inequalities hold for Z_- with J_- replacing J_+.

2.1 Proof of Theorem 2

The proof is a minor modification of some arguments of Mendelson [23].

The proof of Theorem 2 requires a few preliminary steps. First, observe that $\theta_2 \geq 2\Delta$ and therefore

$$\text{for all } t \leq Q_{1-\theta_2}\,, \quad \mathbb{P}\{Z > t\} \geq 2\Delta\,. \tag{2.1}$$

(2.1) implies that all the level sets $\{Z > t\}$ for $0 < t \leq Q_{1-\theta_2}$ satisfy property (1) of Definition 1 with $j = 1$, a fact used frequently in what follows.

Another useful observation is that by Property (3),

$$\mathbb{P}\{Z > 0\} \geq \gamma = 4\theta + 16\Delta = 2\theta_1\,,$$

and therefore, the $(1-\theta_1)$-quantile of Z_+ coincides with the $(1-\theta_1)$-quantile of Z and the $(1-\theta_1)$-quantile of Z_- equals $-Q_{\theta_1}$.

Moreover, by property (1) (with $j = 0$) and the choice of θ we have

$$\mathbb{P}_N\{Z > 0\} \geq \frac{1}{2}\mathbb{P}\{Z > 0\} \geq \frac{\theta_1}{2} > \theta$$

and an identical argument shows that $\mathbb{P}_N\{Z < 0\} > \theta$. Now let $(Z_i^\sharp)_{i=1}^N$ be the monotone nonincreasing rearrangement of $(Z_i)_{i=1}^N$. Set $\widehat{Q}_+ = Z^\sharp_{\theta N}$ and $\widehat{Q}_- = Z^\sharp_{(1-\theta)N}$. Clearly, $\widehat{Q}_+ > 0$ and for $i \in J_+$, $Z_i = (Z_i)_+$. The analogous statement holds for $\widehat{Q}_-$ and J_-, proving part (a) of Theorem 2.

Lemma 2. *On the event $\mathcal{A}$ defined in Definition 1,*

$$Q_{1-\theta_1} < \widehat{Q}_+ < Q_{1-\theta_2} \;\text{ and }\; -Q_{\theta_1} < \widehat{Q}_- < -Q_{\theta_2}\,.$$

Proof. We present a proof of the first claim. The proof of the second one is identical and is omitted. By definition, $\mathbb{P}_N\{Z \geq \widehat{Q}_+\} = \theta$. Applying property (2) in Definition 1 for $I = [\widehat{Q}_+, \infty)$,

$$\theta \leq \mathbb{P}_N\{Z \geq \widehat{Q}_+\} \leq \frac{3}{2}\mathbb{P}\{Z \geq \widehat{Q}\} + 2\Delta\,.$$

In particular,

$$\mathbb{P}\{Z > \widehat{Q}_+\} = \mathbb{P}\{Z \geq \widehat{Q}_+\} \geq \frac{2}{3}(\theta - 2\Delta) \geq \Delta$$

provided that $\theta \geq 7\Delta/2$, as was assumed. Hence, we may use property (1) of Definition 1 with $j = 1$ and $t = \widehat{Q}_+$, to get

$$\left|\frac{\mathbb{P}_N\{Z > \widehat{Q}_+\}}{\mathbb{P}\{Z > \widehat{Q}_+\}} - 1\right| \leq \frac{1}{2}\,.$$

This implies

$$\mathbb{P}\{Z > \widehat{Q}_+\} \le 2\mathbb{P}_N\{Z > \widehat{Q}_+\} \le 2\theta < \theta_1$$

and

$$\mathbb{P}\{Z > \widehat{Q}_+\} \ge \frac{2\mathbb{P}_N\{Z > \widehat{Q}_+\}}{3} = \frac{2(\theta - 1/N)}{3} \ge \theta_2 \,,$$

and therefore

$$Q_{1-\theta_1} < \widehat{Q}_+ < Q_{1-\theta_2} \,,$$

as claimed. ∎

Lemma 3. *Consider the conditions of Theorem 2. Then, on the event $\mathcal{A}$,*

$$\int_0^{Q_{1-\theta_1}} pt^{p-1}\mathbb{P}_N\{Z > t\}dt - \theta\widehat{Q}_+^p \le \frac{1}{N}\sum_{j\in[N]\setminus J_+}(Z_i)_+^p \le \int_0^{Q_{1-\theta_2}} pt^{p-1}\mathbb{P}_N\{Z > t\}dt \,.$$

A similar estimate holds for Z_-.

Proof. Recall that $\widehat{Q}_+ = Z^\sharp_{\theta N}$, and therefore,

$$\frac{1}{N}\sum_{i=1}^N (Z_i)_+^p \mathbb{1}_{\{(Z_i)_+\le\widehat{Q}_+\}} - \theta\widehat{Q}_+^p \le \frac{1}{N}\sum_{i\in[N]\setminus J_+}(Z_i)_+^p \le \frac{1}{N}\sum_{i=1}^N (Z_i)_+^p \mathbb{1}_{\{(Z_i)_+\le\widehat{Q}_+\}} \,.$$

By tail integration,

$$\frac{1}{N}\sum_{i=1}^N (Z_i)_+^p \mathbb{1}_{\{(Z_i)_+\le\widehat{Q}_+\}} = \int_0^\infty pt^{p-1}\mathbb{P}_N\left\{(Z)_+^p \mathbb{1}_{\{(Z)_+\le\widehat{Q}_+\}} > t\right\}dt = \int_0^{\widehat{Q}_+} pt^{p-1}\mathbb{P}_N\{Z > t\}dt \,.$$

Since by Lemma 2, on the event $\mathcal{A}$, $Q_{1-\theta_1} < \widehat{Q}_+ < Q_{1-\theta_2}$, the claimed inequalities follow. ∎

With Lemma 3 in mind, next we may use a general estimate of Mendelson [23] for $\int_0^T pt^{p-1}\mathbb{P}_N\{Z > t\}dt$ that holds as long as T is such that $\mathbb{P}\{Z > t\}$ is large enough. To formulate the claim, recall that

$$\mathcal{E}_{T,p} = 2\sqrt{\Delta}\int_0^T pt^{p-1}\sqrt{\mathbb{P}\{|Z| > t\}}dt \,.$$

Lemma 4. *(Mendelson [23].) Let T be such that $\mathbb{P}\{Z > T\} \ge \Delta$. On the event $\mathcal{A}$ of Definition 1, we have*

$$\mathbb{E}\left[Z_+^p \mathbb{1}_{\{Z_+\le T\}}\right] - \mathcal{E}_{T,p} \le \int_0^T pt^{p-1}\mathbb{P}_N\{Z > t\}dt \le \mathbb{E}Z_+^p + \mathcal{E}_{T,p} \,.$$

Now we are ready to prove Theorem 2.

Proof of Theorem 2. Assume that the event $\mathcal{A}$ holds. Apply Lemma 4 with $T = Q_{1-\theta_1}$ and $T = Q_{1-\theta_2}$. Both valid choices, as

$$\mathbb{P}\{Z > Q_{1-\theta_1}\} \geq \mathbb{P}\{Z > Q_{1-\theta_2}\} \geq \Delta \, .$$

Thus,

$$\int_0^{Q_{1-\theta_2}} pt^{p-1}\mathbb{P}_N\{Z > t\}dt \leq \mathbb{E}Z_+^p + \mathcal{E}_{Q_{1-\theta_2},p} \, ,$$

and

$$\begin{aligned}\int_0^{Q_{1-\theta_1}} pt^{p-1}\mathbb{P}_N\{Z > t\}dt &\geq \mathbb{E}\left[Z_+^p \mathbb{1}_{\{Z\leq Q_{1-\theta_1}\}}\right] - \mathcal{E}_{Q_{1-\theta_1},p} \\ &= \mathbb{E}Z_+^p - \left(\mathbb{E}\left[Z^p \mathbb{1}_{\{Z>Q_{1-\theta_1}\}}\right] + \mathcal{E}_{Q_{1-\theta_1},p}\right) .\end{aligned}$$

It remains to show that

$$\widehat{Q}_+^p\theta \leq 2\mathbb{E}\left[Z^p \mathbb{1}_{\{Z\geq Q_{1-\theta_1}\}}\right] , \tag{2.2}$$

which, by Lemma 3 completes the proof. To that end, recall that $\mathbb{P}\{Z > Q_{1-\theta_2}\} \geq \Delta$ and that, by Lemma 2, $Q_{1-\theta_1} < \widehat{Q}_+ < Q_{1-\theta_2}$. Thus,

$$\mathbb{E}\left[Z^p \mathbb{1}_{\{Z\geq \widehat{Q}_+\}}|Z_1,\ldots,Z_N\right] \leq \mathbb{E}\left[Z_+^p \mathbb{1}_{\{Z\geq Q_{1-\theta_1}\}}\right].$$

Also, since $\mathbb{P}_N\{Z \geq \widehat{Q}_+\} \geq \theta$ and

$$\mathbb{P}\{Z \geq \widehat{Q}_+|Z_1,\ldots,Z_N\} \geq \mathbb{P}\{Z \geq Q_{1-\theta_2}\} \geq \Delta \, ,$$

it follows from property (1) in Definition 1 (taking $j = 0$) that

$$\mathbb{E}\left[Z^p \mathbb{1}_{\{Z\geq \widehat{Q}_+\}}|Z_1,\ldots,Z_N\right] \geq \widehat{Q}_+^p\mathbb{P}\{Z \geq \widehat{Q}_+|Z_1,\ldots,Z_N\} \geq \widehat{Q}_+^p 2\mathbb{P}_N\{Z \geq \widehat{Q}_+\} = 2\widehat{Q}_+^p\theta \, ,$$

proving (2.2).

A similar estimate holds for Z_-. ∎

3 Uniform relative deviations of empirical measures

In this section we present inequalities for uniform relative deviations of empirical measures. These are the main technical novelties of the article that allow us to construct covariance and mean estimators with the desired properties. More precisely, we prove that properties (1) and (2) of Definition 1 hold *uniformly* over

a class of random variables (with high probability) above a certain critical level depending on the class.

To properly set up the main result of this section, we need a few definitions. Since the results may be of independent interest, we present it in a greater generality than what is needed for our purposes in this article.

Let $\mathcal{X}$ be a measurable set and let $X, X_1, \dots, X_N$ be independent, identically distributed random variables taking values in $\mathcal{X}$.

Consider a class $\mathcal{F}$ of real-valued functions defined on $\mathcal{X}$. We assume that, for all $f \in \mathcal{F}$, the random variable $f(X)$ has mean zero. We denote by L_2 the set of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}f(X)^2 < \infty$. We write $\|f\|_{L_2} = \left(\mathbb{E}f(X)^2\right)^{1/2}$ for all $f \in L_2$ and denote the unit sphere and the unit ball in L_2 by $S(L_2) = \{f \in L_2 : \|f\|_{L_2} = 1\}$ and $D = \{f \in L_2 : \|f\|_{L_2} \le 1\}$, respectively.

Recall that for an indicator function $h(x) = \mathbb{1}_{x \in A}$ for some $A \subset \mathcal{X}$, we abbreviate

$$\mathbb{P}\{h\} = \mathbb{P}\{X \in A\} \quad \text{and} \quad \mathbb{P}_N\{h\} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{X_i \in A} \, .$$

We assume that $0 \in \mathcal{F}$ (i.e., $\mathcal{F}$ contains the function that equals zero everywhere) and that $\mathcal{F}$ is star-shaped around 0, that is, if $f \in \mathcal{F}$ then $cf \in \mathcal{F}$ for all $c \in [0,1]$.

For $\epsilon > 0$, denote by $\mathcal{M}(\mathcal{F}, \epsilon)$ the packing number of $\mathcal{F}$, that is, the size of the largest ϵ-net in $\mathcal{F}$ (i.e., a subset whose elements have pairwise distance at least ϵ).

Definition 2. *For $\Delta, c, L > 0$, let $\overline{\rho}_N = \overline{\rho}_N(c, \Delta, L)$ denote the infimum of all those values $r > 0$ such that*

$$(1) \quad \log \mathcal{M}\left(\mathcal{F} \cap rS(L_2), \frac{\Delta^{3/2} r}{1000 \cdot L}\right) \le cN\Delta \, , \quad \textit{and}$$

$$(2) \quad \mathbb{E} \sup_{u \in (\mathcal{F}-\mathcal{F}) \cap \Delta^{3/2} rD} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i u(X_i) \right| \le \frac{\Delta^2 r}{40000 \cdot L} \, ,$$

where $\epsilon_1, \dots, \epsilon_N$ are independent symmetric Bernoulli random variables with $\mathbb{P}\{\epsilon_i = 1\} = \mathbb{P}\{\epsilon_i = -1\} = 1/2$ and $\mathcal{F} - \mathcal{F} = \{f = g - h : g, h \in \mathcal{F}\}$. We call $\overline{\rho}_N$ the critical level *of the class $\mathcal{F}$.*

Note that since $\mathcal{F}$ is star-shaped, inequalities (1) and (2) in Definition 2 are satisfied for all $r > \overline{\rho}_N$. The constants 1/1000 and 1/40000 do not have any special role. Their values have not been optimized and they are chosen by convenience.

The main result of this section is that there exists a positive numerical constant c such that, above the critical level $\overline{\rho}_N(c, \Delta, L)$, functions in $\mathcal{F}$ simultaneously satisfy properties (1) and (2) of Definition 1, given a certain condition involving the constant L.

To formulate the theorem, for $r > 0$, define the classes U_r and V_r of indicator functions by

$$U_r = \left\{ \mathbb{1}_{\{f \in I\}} : I \text{ is an interval, } f \in \mathcal{F},\ \|f\|_{L_2} \geq r \right\}$$

and

$$V_r = \left\{ \mathbb{1}_{\{h > t\}} : t > 0,\ h \in \mathcal{F} \cup (-\mathcal{F}),\ \|h\|_{L_2} \geq r \right\} .$$

The main result of this section is the following:

Theorem 3. *Let $\mathcal{F} \subset L_2$ be a class of real-valued functions that is star-shaped about 0, such that $\mathbb{E}f(X) = 0$ for all $f \in \mathcal{F}$. There exist positive numerical constants c, c_0, c_1 such that the following holds. Let $c_0 \frac{\log N}{N} \leq \Delta \leq \frac{1}{2}$ and assume that functions in $\mathcal{F}$ satisfy the following* small-ball condition *with constants $L > 0$ and $\gamma = c_1 \Delta$: for any interval $I \subset \mathbb{R}$,*

$$\mathbb{P}\{f(X) \in I\} \leq \max\left\{ \frac{L|I|}{\|f\|_{L_2}}, \gamma \right\} . \tag{3.1}$$

Suppose that $r > \overline{\rho}_N(c, \Delta, L)$. Then, with probability at least $1 - 2\exp(-c_2(L)\Delta N)$, for all $u \in U_r$ and $v \in V_r$,

(a) for any integer $j \geq 0$, if $2^{-j}\mathbb{P}\{v\} \geq \Delta$, then

$$\left| \frac{\mathbb{P}_N\{v\}}{\mathbb{P}\{v\}} - 1 \right| \leq 2^{-j/2-1} ;$$

(b) $\mathbb{P}_N\{u\} \leq \frac{3}{2}\mathbb{P}\{u\} + 2\Delta$,

where $c_2(L)$ is a positive constant depending on L only.

Theorem 3 resembles classical results in empirical processes theory. In fact, in certain special situations, the classes U_r and V_r have a well-behaved vc dimension and then uniform ratio estimates are known (see, for example, [23] for a recent application). However, in the general case we study here, the vc dimensions of U_r and V_r can be very large or even infinite—even when the class $\mathcal{F}$ itself is relatively well-behaved. As a result, the proof of Theorem 3 calls for a different way of showing that U_r and V_r are "small". The key feature of $\mathcal{F}$ used in the proof of Theorem 3 is that functions in $\mathcal{F}$ satisfy the small-ball condition (3.1).

We begin by showing that any fixed function f satisfies Properties (a) and

(b), with high probability. This observation, while interesting on its own right, is needed in the proof of Theorem 3. As it happens, the required bounds for a single function do follow from vc theory.

Definition 3. *Let H be a class of $\{0,1\}$-valued functions on $\mathcal{X}$. A set $\{x_1,\ldots,x_n\}$ is shattered by H if for every $I \subset [n]$ there is some $h_I \in H$ for which $h_I(x_i) = 1$ if $i \in I$ and $h_I(x_i) = 0$ otherwise. The* vc *dimension of H is the maximal cardinality of a subset of $\mathcal{X}$ that is shattered by H. It is denoted by* $\mathrm{VC}(H)$.

We refer the reader to van der Vaart and Wellner [31] for basic facts on vc classes and on the vc-dimension.

The connection between the vc-dimension and properties (a) and (b) for a single function is that for any function f, the class of indicator functions

$$U_f = \left\{ \mathbb{1}_{\{f \in I\}} : I \text{ is an interval} \right\}$$

satisfies that $VC(U_f) \leq 2$.

For classes of sets with finite vc dimension, the following analogue of Theorem 3 was established by Mendelson [23]. Again, it should be stressed that in such cases, uniform ratio estimates are far simpler than in the general scenario that is needed in what follows.

Theorem 4. *([23].) There are absolute constants c_0 and c_1 for which the following holds. Let U be a class of functions on $\mathcal{X}$ taking values in $\{0,1\}$, such that $VC(U) \leq d$. Then for*

$$c_0 \frac{d}{N} \log\left(\frac{eN}{d}\right) \leq \Delta \leq \frac{1}{2} ,$$

with probability at least $1 - 2\exp(-c_1 \Delta N)$, for every $u \in U$ and nonnegative integer j,

(a') if $\mathbb{P}\{u\} \geq 2^j \Delta$, then

$$\left| \frac{\mathbb{P}_N\{u\}}{\mathbb{P}\{u\}} - 1 \right| \leq 2^{-j/2-2} ;$$

(b') $\mathbb{P}_N\{u\} \leq \frac{3}{2}\mathbb{P}\{u\} + 2\Delta$.

In particular, when applied to the class U_f whose VC dimension is at most 2, Theorem 4 implies that any function f satisfies properties (a) and (b) of Theorem 3. Indeed, $VC(U_f) \leq 2$ and therefore, Theorem 4 may be applied when $\Delta \geq c_0 \frac{\log N}{N}$ for a numerical constant $c_0 > 0$.

3.1 Proof of Theorem 3

Let us now turn to the proof of Theorem 3. It is important to note once again that there is no reason to expect that the corresponding classes of indicator functions U_r and V_r have a well-behaved vc dimension.

We begin with the following straightforward observation.

Lemma 5. *For any f, h, any $t \in \mathbb{R}$ and $\delta > 0$,*

$$\left| \mathbb{1}_{\{f>t\}} - \mathbb{1}_{\{h>t+\delta\}} \right| \le \mathbb{1}_{\{|f-h|>\delta\}} + \mathbb{1}_{\{h\in(t-\delta,t+\delta]\}}. \tag{3.2}$$

Proof. If $\mathbb{1}_{\{f>t\}}(x) - \mathbb{1}_{\{h>t+\delta\}}(x) \neq 0$ and $|h-f|(x) \le \delta$ then one of the two alternatives holds: either $f(x) > t$ and $h(x) \le t+\delta$, or $f(x) \le t$ and $h(x) > t+\delta$. The former implies that $h(x) \in (t-\delta, t+\delta]$, while the latter is impossible. ∎

Thanks to Lemma 5, it is possible to estimate

$$|\mathbb{P}\{f > t\} - \mathbb{P}\{h > t+\delta\}| \quad \text{and} \quad |\mathbb{P}_N\{f > t\} - \mathbb{P}_N\{h > t+\delta\}|$$

by a combination of a tail estimate for $|f-h|$ and a small-ball estimate for $|h|$ — first with respect to the underlying measure $\mathbb{P}$ and then with respect to the empirical measure $\mathbb{P}_N$.

Fix $\Delta > c_0 \log N/N$ for the numerical constant c_0 mentioned in the paragraph following Theorem 4. Let $c > 0$ be a constant to be specified later and let $r > \overline{\rho}_N(c, \Delta, L)$ be above the critical level for $\mathcal{F}$. Let j be a nonnegative integer. By property (1) in Definition 2, a maximal $\Delta^{3/2} r/(1000L)$-net of the subset of $\mathcal{F}$ consisting of functions with $\|f\|_{L_2} = r$ has cardinality at most $\exp(cN\Delta)$. Let H_r be such an $\Delta^{3/2} r/(1000L)$-net.

If $c \le c_1/2$ for the constant c_1 appearing in Theorem 4, then by the ratio estimate for a single function (Theorem 4) and the union bound, with probability at least $1 - 2\exp(-c'N\Delta)$,

$$\sup_{h\in H_r} \sup_{t:\mathbb{P}\{h>t\}\ge 2^{j-1}\Delta} \left| \frac{\mathbb{P}_N\{h > t\}}{\mathbb{P}\{h > t\}} - 1 \right| \le \frac{2^{-(j-1)/2}}{4}, \tag{3.3}$$

where we may take $c' = c_1/2$.

For $f \in \mathcal{F} \cap rS(L_2)$, let $\pi f \in H_r$ be the best approximation to f in the net H_r with respect to the L_2 norm. Then, for any $t \in \mathbb{R}$ and $\delta > 0$, on the same event where (3.3) holds, we have

$$\left| \frac{\mathbb{P}_N\{f > t\}}{\mathbb{P}\{f > t\}} - 1 \right| \le \left| \frac{\mathbb{P}_N\{f > t\}}{\mathbb{P}\{f > t\}} - \frac{\mathbb{P}_N\{\pi f > t+\delta\}}{\mathbb{P}\{\pi f > t+\delta\}} \right| + \frac{2^{-(j-1)/2}}{4}, \tag{3.4}$$

provided that $\mathbb{P}\{\pi f > t + \delta\} \geq 2^{(j-1)}\Delta$.

Note that in the inequality above, we may choose the value of δ at will, even depending on f. For each $f \in \mathcal{F}$, define $\delta_f = \Delta\|f\|_{L_2}/(100 \cdot L)$

The next lemma shows that with this choice of δ_f, one indeed has $\mathbb{P}\{\pi f > t + \delta\} \geq 2^{(j-1)}\Delta$ whenever $\mathbb{P}\{f > t\} \geq 2^j\Delta$.

Lemma 6. *Assume the small-ball condition* (3.1) *where* $\gamma \leq \Delta/18$. *Then, for every* $f \in \mathcal{F} \cap rS(L_2)$,

$$\left|\mathbb{P}\{f > t\} - \mathbb{P}\{\pi f > t + \delta_f\}\right| \leq \frac{2\Delta}{9}\,.$$

Proof. Fix $f \in \mathcal{F} \cap rS(L_2)$. By the small-ball condition, if $2L\delta_f/r \geq \gamma$ then

$$\mathbb{P}\{\pi f \in [t - \delta_f, t + \delta_f]\} \leq \frac{2L\delta_f}{\|\pi f\|_{L_2}} = \frac{2L\delta_f}{r}\,.$$

Therefore,

$$\begin{aligned}
\mathbb{P}\{|f - \pi f| \geq \delta_f\} + \mathbb{P}\{\pi f \in [t - \delta_f, t + \delta_f]\} &\leq \frac{\|f - \pi f\|_{L_2}^2}{\delta_f^2} + \frac{2L\delta_f}{\|\pi f\|_{L_2}} \\
&\leq \frac{\left(\frac{\Delta^{3/2} r}{1000 \cdot L}\right)^2}{\left(\frac{\Delta r}{100 \cdot L}\right)^2} + \frac{2L\Delta}{100 \cdot L} \\
&= \frac{3\Delta}{100}\,.
\end{aligned}$$

The stated inequality now follows from Lemma 5. ∎

Hence, under the small-ball condition, (3.4) indeed holds whenever $\mathbb{P}\{f > t\} \geq 2^j\Delta$. Consider such an a function $f \in \mathcal{F} \cap rS(L_2)$. Using

$$\left|\frac{\mathbb{P}_N\{\pi f > t + \delta_f\}}{\mathbb{P}\{\pi f > t + \delta_f\}} - 1\right| \leq \frac{2^{-(j-1)/2}}{4} < \frac{1}{2}\,,$$

the first term on the right-hand side of (3.4) may be bounded as

$$\begin{aligned}
&\left|\frac{\mathbb{P}_N\{f>t\}}{\mathbb{P}\{f>t\}}-\frac{\mathbb{P}_N\{\pi f>t+\delta_f\}}{\mathbb{P}\{\pi f>t+\delta_f\}}\right| \\
&\quad\le \left|\frac{\mathbb{P}_N\{f>t\}-\mathbb{P}_N\{\pi f>t+\delta_f\}}{\mathbb{P}\{f>t\}}\right|+\frac{\mathbb{P}_N\{\pi f>t+\delta_f\}}{\mathbb{P}\{\pi f>t+\delta_f\}}\left|\frac{\mathbb{P}\{f>t\}-\mathbb{P}(\pi f>t+\delta_f)}{\mathbb{P}\{f>t\}}\right| \\
&\quad\le \left|\frac{\mathbb{P}_N\{f>t\}-\mathbb{P}_N\{\pi f>t+\delta_f\}}{\mathbb{P}\{f>t\}}\right|+2\left|\frac{\mathbb{P}\{f>t\}-\mathbb{P}\{\pi>t+\delta_f\}}{\mathbb{P}\{f>t\}}\right| \\
&\quad\le \frac{\mathbb{P}_N\{|f-\pi f|\ge\delta_f\}}{\mathbb{P}\{f>t\}}+\frac{\mathbb{P}_N\{\pi f\in[t-\delta_f,t+\delta_f]\}}{\mathbb{P}\{f>t\}} \\
&\qquad +2\frac{\mathbb{P}\{|f-\pi f|\ge\delta_f\}}{\mathbb{P}\{f>t\}}+2\frac{\mathbb{P}\{\pi f\in[t-\delta_f,t+\delta_f]\}}{\mathbb{P}\{f>t\}} \\
&\qquad \text{(using Lemma 5 twice)} \\
&\quad\overset{\text{def.}}{=} (I)+(II)+(III)+(IV)\,. \qquad (3.5)
\end{aligned}$$

Hence, we need to bound the four terms on the right-hand side. The last two terms may be bounded without further work, as we have already seen in the proof of Lemma 6 that

$$(III)+(IV)\le\frac{3\Delta/100}{\mathbb{P}\{f>t\}}\le\frac{3\cdot 2^{-j}}{100}\,.$$

In the remaining part of the proof we bound the empirical counterpart, that is, the terms (I) and (II).

Since $\pi f\in H_r$, for term (II), we may simply invoke part (b') of Theorem 4 that implies that, with probability at least $1-e^{-c\Delta N}$, for all $f\in\mathcal{F}\cap rS(L_2)$, we have

$$(II)\le\frac{3}{4}\,(IV)+\frac{\Delta/10}{\mathbb{P}\{f>t\}}\le 2^{-j-3}\,.$$

It remains to bound

$$\sup_{f\in\mathcal{F}\cap rS(L_2)}\mathbb{P}_N\{|f-\pi f|>\delta_f\}=\sup_{f\in\mathcal{F}\cap rS(L_2)}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{\{|f-\pi f|>\delta_f\}}(X_i)\,. \qquad (3.6)$$

This is done in the next lemma that implies that, with probability at least $1-e^{-c\Delta N}$, for all $f\in\mathcal{F}\cap rS(L_2)$ with $\mathbb{P}\{f>t\}\ge 2^j\Delta$,

$$(I)\le\frac{2^{-j}}{10}\,.$$

Putting everything together, we get that there exists a universal constant $c>0$ such that, for every nonnegative integer j, with probability at least $1-e^{-c\Delta N}$, we have

$$\left|\frac{\mathbb{P}_N\{f>t\}}{\mathbb{P}\{f>t\}}-1\right|\le 2^{-j}\left(\frac{3}{100}+\frac{1}{8}+\frac{1}{10}\right)+\frac{2^{-(j-1)/2}}{4}\le 2^{-(j/2-1)}\,.$$

Since there are at most $\log_2 N$ relevant values of j, the union bound implies part (a) of Theorem 3 for functions $f \in \mathcal{F} \cap rS(L_2)$. To deal with functions that satisfy $\|f\|_{L_2} \geq r$, fix such a function and $t \in \mathbb{R}$ for which $\mathbb{P}\{f > t\} \geq 2^j\Delta$. Put $t_r = rt/\|f\|_{L_2}$ and $f_r = rf/\|f\|_{L_2} \in F \cap rS(L_2)$, and note that

$$\{f > t\} = \left\{ \frac{rf}{\|f\|_{L_2}} \cdot \frac{\|f\|_{L_2}}{r} > t \right\} = \{f_r > t_r\} \ .$$

Thus, $\mathbb{P}\{f_r > t_r\} \geq 2^j\Delta$,

$$\frac{\mathbb{P}_N\{f_r > t_r\}}{\mathbb{P}\{f_r > t_r\}} = \frac{\mathbb{P}_N\{f > t\}}{\mathbb{P}\{f > t\}},$$

and the claim follows from the bound for $\mathcal{F} \cap rS(L_2)$ and by the star-shaped property of $\mathcal{F}$.

Lemma 7. *Let $r > \overline{\rho}_N(c, \Delta, L)$. Then, for some constant $c_1 > 0$, with probability at least $1 - e^{c_1\Delta N}$,*

$$\sup_{f \in \mathcal{F} \cap rS(L_2)} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{|f - \pi f| > \delta_f\}}(X_i) \leq \frac{\Delta}{10} \ .$$

Proof. Recall that, by definition, $r > \overline{\rho}_N(c, \Delta, L)$ implies that

$$\mathbb{E} \sup_{f \in \mathcal{F} \cap rS(L_2)} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i (f - \pi f)(X_i) \right| \leq \frac{\Delta^2 r}{40000 \cdot L} \ . \tag{3.7}$$

Define

$$Z \stackrel{\text{def.}}{=} \sup_{f \in \mathcal{F} \cap rS(L_2)} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{|f - \pi f| > \delta_f\}}(X_i) \ .$$

The proof is based on Talagrand's concentration inequality for the supremum of empirical processes [30] which implies that for $x > 0$, with probability at least $1 - \exp(-x)$,

$$Z \leq \frac{3}{2}\mathbb{E}Z + 4\sigma_{\mathcal{F}} \sqrt{\frac{x}{N}} + \frac{6x}{N}$$

(see, e.g., Theorems 11.8 and 12.2 in [2]) where $\sigma_{\mathcal{F}} = \sup_{f \in \mathcal{F} \cap rS(L_2)} \mathbb{P}^{1/2}\{|f - \pi f| > \delta_f\}$.

Here, recalling that $\delta_f = \Delta\|f\|_{L_2}$ and the fact that δ_f is the same for every $f \in \mathcal{F} \cap rS(L_2)$, by Chebyshev's inequality,

$$\sigma_{\mathcal{F}} \leq \sup_{f \in \mathcal{F} \cap rS(L_2)} \frac{\|f - \pi f\|_{L_2}}{\delta_f} \leq \frac{\Delta^{3/2} r}{\Delta r} = \frac{\sqrt{\Delta}}{10} \ .$$

Next, let

$$\phi_\delta(t) = \begin{cases} 1 & \text{if } t \geq \delta\,, \\ \frac{2}{\delta}\left(t - \frac{\delta}{2}\right) & \text{if } t \in \left[\frac{\delta}{2}, \delta\right], \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $\phi_\delta(t) \geq \mathbb{1}_{t>\delta}$ is a Lipschitz function with Lipschitz constant $2/\delta$ that satisfies $\phi_\delta(0) = 0$. By the Giné-Zinn symmetrization theorem [9] followed by the contraction inequality for Bernoulli processes [15],

$$\begin{aligned} \mathbb{E}Z &\leq \mathbb{E} \sup_{f \in \mathcal{F} \cap rS(L_2)} \frac{1}{N} \sum_{i=1}^N \phi_\delta(|f - \pi f|(X_i)) \\ &\leq 2\mathbb{E} \sup_{f \in \mathcal{F} \cap rS(L_2)} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i \phi_\delta(|f - \pi f|(X_i)) \right| + \sup_{f \in \mathcal{F} \cap rS(L_2)} \mathbb{E}\phi_\delta(|f - \pi f|(X_i)) \\ &\leq \frac{4}{\delta}\mathbb{E} \sup_{f \in \mathcal{F} \cap rS(L_2)} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i (f - \pi f)(X_i) \right| + \frac{\Delta}{25}\,, \end{aligned}$$

where the last inequality holds because

$$\mathbb{E}\phi_\delta(|f - \pi f|(X_i)) \leq \mathbb{P}(|f - \pi f|(X) \geq \delta_f/2) \leq \frac{4\|f - \pi f\|_{L_2}^2}{\delta_f^2} \leq \frac{\Delta}{25}\,.$$

Using (3.7), we have

$$\frac{4}{\delta_f}\mathbb{E} \sup_{f \in \mathcal{F} \cap rS(L_2)} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i (f - \pi f)(X_i) \right| \leq \frac{\Delta}{100}\,,$$

and therefore, with probability at least $1 - 2\exp(-x)$,

$$\sup_{f \in \mathcal{F} \cap rS(L_2)} \mathbb{P}_N(|f - \pi f| > \delta_f) \leq \frac{3\Delta}{40} + 4\Delta^{1/2}\sqrt{\frac{x}{N}} + \frac{6x}{N}\,.$$

The claim follows by setting $x = c_0 N\Delta$ for a sufficiently small value of c_0. ∎

In order to complete the proof of Theorem 3, it only remains to prove part (b). The proof is completely analogous with part (a). First we consider a net of $\mathcal{F} \cap rS(L_2)$ and use part (b') of Theorem 4. Then one may extend the inequality to all $\mathcal{F} \cap rS(L_2)$ in a similar fashion. In order to avoid repetition of the same ideas, we omit the details.

4 Covariance estimation with trimmed means

In this section we present the first main component of the mean estimation procedure. As explained in the introduction, in this first step we need to estimate the directional variances $\sigma^2(u) = \mathrm{Var}(\langle X, u\rangle)$ in all directions where $\sigma^2(u)$ is "not too small." This estimator does not need to be very accurate. It is sufficient for our purposes that the estimator is correct up to a constant factor.

For this purpose, we require a bit more than the existence of the covariance matrix Σ. The key assumption we use is "L_q-L_2 norm equivalence" for some $q > 2$. More precisely, we assume that there exist $q > 2$ and $\kappa > 0$ such that, for all $u \in S^{d-1}$,

$$\left(\mathbb{E}\langle X - \mu, u\rangle^q\right)^{1/q} \le \kappa\left(\mathbb{E}\langle X - \mu, u\rangle^2\right)^{1/2} .$$

In other words, writing $\overline{X} = X - \mathbb{E}X$, we assume that for all $u \in S^{d-1}$,

$$\left\|\left\langle \overline{X}, u\right\rangle\right\|_{L_q} \le \kappa \left\|\left\langle \overline{X}, u\right\rangle\right\|_{L_2} .$$

The proposed estimator is quite natural. For each direction $u \in S^{d-1}$, we compute an appropriately trimmed empirical variance. Unlike standard trimmed mean estimators, where the truncation occurs at a pre-set level, here we trim by removing a fixed number of the largest and smallest values of $\langle X_i, u\rangle$ corresponding to each direction $u \in S^{d-1}$. To show that this estimator satisfies the desired properties simultaneously for all directions, we make use of the tools developed in Sections 2 and 3. The main ingredient of the analysis is Theorem 3. This theorem requires that the "small-ball" condition (3.1) is satisfied. To guarantee this property, we form blocks of a fixed size of the given sample, and take the empirical average within each block. In Section 4.1 we show that under L_q-L_2 norm equivalence, it suffices to form blocks of constant size (depending in κ and q).

Let us describe the covariance estimation procedure. In order to estimate variances without knowing the expected values, we use the standard trick that, if X' is an independent copy of X, then $\mathrm{Var}(\langle X, u\rangle) = (1/2)\mathbb{E}\langle X - X', u\rangle^2$. It is easy to see that if X satisfies $L_q - L_2$ norm equivalence with constant κ then $\widetilde{X} = X - X'$ satisfies L_q-L_2 norm equivalence with constant $\sqrt{2}\kappa$.

Thus, we split the data in two halves to form independent pairs of observations. For the sake of simpler notation, assume that we are given $2N$ independent copies, $X_1, \ldots, X_{2N}$ and for $i \in [N]$, define $\widetilde{X}_i = X_i - X_{N+i}$.

The proposed covariance estimator has two tuning parameters, $\gamma, \theta \in (0, 1)$.

For a positive integer m, define $Z = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \widetilde{X}_i$. By Lemma 8, there is a constant $c_1(\kappa, q)$ such that if $m = \left\lceil \frac{c_1(\kappa,q)}{\gamma^2} \right\rceil$, then for any $u \in S^{d-1}$, and for all intervals

$I \subset \mathbb{R}$,

$$\mathbb{P}\{\langle u, Z\rangle \in I\} \le \max\left\{L\frac{|I|}{\sigma(u)}, \gamma\right\},$$

where $L > 0$ is a numerical constant. This implies that every function in the class $\mathcal{F} = \left\{\langle u, Z\rangle : u \in B_2^d\right\}$ satisfies the key "small-ball" assumption in Theorem 3.

Once the value of m is set, the sample $\widetilde{X}_1, \dots, \widetilde{X}_N$ is divided into $n = N/m$ blocks, each one of cardinality m. (We may assume, without loss of generality, that m divides N.) For each block $j \in [n]$, we may compute

$$Z_j = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \widetilde{X}_{m(j-1)+i} .$$

For every $u \in S^{d-1}$, denote by $J_+(u)$ the set of indices of the θn largest values of $\left\langle Z_j, u\right\rangle$ and define

$$\psi_N(u) = \frac{1}{2n} \sum_{j \in [n] \setminus J_+(u)} \left(\left\langle Z_j, u\right\rangle\right)^2 .$$

The following theorem summarizes the main performance guarantees of the estimator $\psi_N(u)$. It is a crucial ingredient of the mean estimator introduced in the next section.

Proposition 2. *Assume the condition of Theorem 1. There are constants $\gamma, \theta \in (0,1)$ and $c_0, c' > 0$ depending on κ and q for which the following holds. Set $m \ge \frac{c_1(\kappa,q)}{\gamma^2}$ and*

$$r^2 = \frac{c_0}{n} \sum_{i \ge c_0 N} \lambda_i .$$

Then, with probability at least $1 - 2\exp(-c'N)$, the estimator ψ_N satisfies

(i) If $u \in S^{d-1}$ is such that $\sigma(u) \ge r$, then

$$\frac{1}{4}\sigma^2(u) \le \psi_N(u) \le 2\sigma^2(u) .$$

(ii) If $\sigma(u) \le r$ then $\psi_N(u) \le Cr^2$ for an absolute constant C.

Proof. Fix $\gamma, \theta \in (0,1)$ and consider the resulting estimator ψ_N. (Recall that in the definition of ψ_N, the block size m depends on γ, as well as on the constants κ and q of the norm equivalence condition). We show that γ and θ may be chosen so that the inequalities of the theorem hold. In particular, let c, c_1 be the constants appearing in Theorem 3. We show that it is sufficient if the parameters $\gamma, \theta \in (0,1)$ satisfy $2\theta + 8\gamma/c_1 < (5\kappa^2)^{-q/(q-2)}$ and $\theta > 7\gamma/c_1$.

The proof consists of three parts. First we show that (i) holds above such a the critical level r. This is based on Theorems 2 and 3. Second, we show that the announced value of r satisfies $r > \overline{\rho}_n(c, \Delta, L)$ and therefore it is a valid choice. Finally, we prove part (ii) of the theorem, based on the *small-ball method*. Before the proof, we establish some consequences of the $L_q - L_2$ norm equivalence condition (1.4).

4.1 $L_q - L_2$ norm equivalence implies a small ball property

Let Y be an absolutely continuous real-valued random variable that satisfies $\|Y - \mathbb{E}Y\|_{L_q} \le \kappa \|Y - \mathbb{E}Y\|_{L_2}$ for some $q > 2$ and $\kappa > 0$. Let $Y_1, \dots, Y_m$ be independent copies of Y. To ease notation, set $\overline{Y} = Y - \mathbb{E}Y$.

Lemma 8. *Define* $Z_m = \frac{1}{\sqrt{m}} \sum_{i=1}^m \overline{Y}_i$.

- *There exists a constant* $m_0(q, \kappa)$ *such that if* $m \ge m_0(q, \kappa)$, *then*

$$\mathbb{P}\{Z_m \ge 0\} \ge \frac{1}{4} \quad \textit{and} \quad \mathbb{P}\{Z_m \le 0\} \ge \frac{1}{4}\,. \tag{4.1}$$

- *Let* $0 < \xi < 1/2$. *Then*

$$\mathbb{E}\overline{Y}^2 \mathbb{1}_{\{|\overline{Y}| \ge Q_{1-\alpha}(|\overline{Y}|)\}} \le \xi \mathbb{E}\overline{Y}^2\,, \tag{4.2}$$

  where $\alpha = (\xi/\kappa^2)^{q/(q-2)}$.

- $$\|Z_m\|_{L_q} \le (4(q-1))^{1/2} \kappa \|Z_m\|_{L_2}\,. \tag{4.3}$$

- *There exists a numerical constant* L *and a constant* $c_1 = c_1(\kappa, q)$ *such that, for any* $\gamma \in (0,1)$, *if* $m \ge c_1/\gamma^2$, *then for all intervals* $I \subset \mathbb{R}$,

$$\mathbb{P}\{Z_m \in I\} \le \max\left\{ L \frac{|I|}{\left\|\overline{Y}\right\|_{L_2}}, \gamma \right\}. \tag{4.4}$$

Proof. (4.1) follows from a generalization of the Berry-Esseen theorem (see, e.g., [29]),

To prove (4.2), note that by Hölder's inequality for $\beta = q/2$ and the $L_q - L_2$ norm equivalence,

$$\mathbb{E}\left[\overline{Y}^2 \mathbb{1}_{\{|\overline{Y}| \ge Q_{1-\alpha}(\overline{Y})\}}\right] \le \left\|\overline{Y}\right\|_{L_q}^2 \mathbb{P}\left\{\overline{Y} \ge Q_{1-\alpha}(|\overline{Y}|)\right\}^{1-\frac{2}{q}} \le \kappa^2 \mathbb{E}\overline{Y}^2 \cdot \alpha^{1-\frac{2}{q}}\,.$$

For the proof of (4.3), observe that, if $\varepsilon_1, \ldots, \varepsilon_m$ are independent symmetric Bernoulli random variables with $\mathbb{P}\{\epsilon_i = 1\} = \mathbb{P}\{\epsilon_i = -1\} = 1/2$, then by symmetrization and Khintchine's inequality (see, e.g, [6, p.21],

$$\mathbb{E}|Z_m|^q \le 2^q \mathbb{E}\left|\frac{1}{\sqrt{m}}\sum_{i=1}^m \varepsilon_i \overline{Y}_i\right|^q \le (4(q-1))^{q/2}\mathbb{E}\left|\frac{1}{m}\sum_{i=1}^m \overline{Y}_i^2\right|^{\frac{q}{2}} \le (4(q-1)\kappa^2)^{q/2}\mathbb{E}\overline{Y}^2 ,$$

where the last inequality follows from the norm-equivalence condition (1.4) because, by the convexity of $\phi(t) = |t|^{q/2}$,

$$\left|\frac{1}{m}\sum_{i=1}^m \overline{Y}_i^2\right|^{\frac{q}{2}} \le \frac{1}{m}\sum_{i=1}^m \left|\overline{Y}_i\right|^q .$$

It remains to prove the small-ball bound of (4.4). Setting $\xi = 1/50$, it follows from (4.2) that $\mathbb{E}\left[\overline{Y}^2 \mathbb{1}_{\{|\overline{Y}| \ge Q_{1-\alpha}(|\overline{Y}|)\}}\right] \le (1/50)\mathbb{E}\overline{Y}^2$ where $\alpha = (1/(50\kappa^2)^{q/(q-2)}$. Moreover, by Chebychev's inequality, $Q_{1-\alpha}(|\overline{Y}|) \le \left\|\overline{Y}\right\|_{L_2}/\sqrt{\alpha}$, and therefore

$$\mathbb{E}\left[\overline{Y}^2 \mathbb{1}_{\{|\overline{Y}| \ge \left\|\overline{Y}\right\|_{L_2}/\sqrt{\alpha}\}}\right] \le \frac{\mathbb{E}\left[\overline{Y}^2\right]}{50} .$$

By the second part of Theorem 3.2 in Mendelson [24], there exists a constant c_1 depending on α (and hence on κ and q) such that, for any $\gamma \in (0,1)$, if $m \ge c_1/\gamma^2$, then

$$\sup_{x \in R} \mathbb{P}\{|Z_m - x| \le c_2 \gamma \left\|\overline{Y}\right\|_{L_2}\} \le \gamma ,$$

where c_2 is an absolute constant. Hence, for any interval $I \subset \mathbb{R}$,

$$\mathbb{P}\{Z_m \in I\} \le \max\left\{\frac{|I|}{c_2\left\|\overline{Y}\right\|_{L_2}}, \gamma\right\} .$$

■

Above the critical level

We start by proving part (i) of Proposition 2. The proof is based on applying Theorem 3 for the class $\mathcal{F} = \left\{\langle u, \cdot\rangle : u \in B_2^d\right\}$.

Let $\Delta = \gamma/c_1$ and let $\overline{\rho}_n(c, \Delta, L)$ be the critical level of the class $\mathcal{F}$, as defined in Definition 2. Let $r > \overline{\rho}_n(c, \Delta, L)$ be arbitrary.

Let $D = \{u \in \mathbb{R}^d : \sigma(u) \le 1\}$. Clearly, $\sqrt{2}\sigma(u) = \|\langle u, Z\rangle\|_{L_2}$ and for each u, $\langle u, Z\rangle$ is a symmetric random variable.

The class $\mathcal{F} = \left\{\langle u, \cdot\rangle : u \in B_2^d\right\}$ is star-shaped around 0 and therefore it satisfies the conditions of Theorem 3. By Theorem 3, there is a numerical constant $c_2 > 0$ and an event $\mathcal{A}$ of probability at least $1 - 2\exp(-c_2\gamma n)$, on which the following holds: for every $u \in B_2^d$ such that $\left\|\left\langle u, \overline{X}\right\rangle\right\|_{L_2} \ge r$, the random variable $\langle u, Z\rangle$ satisfies properties (1)-(2) of Definition 1. Moreover, property (3) holds trivially for every $\langle u, Z\rangle$ by the symmetry of these random variables.

Suppose that the event $\mathcal{A}$ occurs and let $u \in B_2^d$ be such that $\sigma(u) \ge r$.

Let $\left(\left\langle u, Z_j\right\rangle^\sharp\right)_{j=1}^n$ be the monotone nonincreasing rearrangement of $\langle u, Z_1\rangle, \dots, \langle u, Z_n\rangle$. Define $\widehat{Q}(u) = \langle u, Z_{\theta n}\rangle^\sharp$ and denote by $Q_q(u)$ the q-quantile of the random variable $\langle u, Z\rangle$. By Lemma 2 and the symmetry of the random variables, it follows that

$$Q_{1-(2\theta+8\Delta)}(u) \le \widehat{Q}(u) \le Q_{1-(2\theta-8\Delta)/3}(u)\,.$$

Setting

$$Q_3 \stackrel{\text{def.}}{=} Q_{1-(2\theta+8\Delta)}(u) \quad \text{and} \quad Q_4 \stackrel{\text{def.}}{=} Q_{1-(2\theta-8\Delta)/3}(u)\,,$$

just as in the proof of Lemma 3, we have

$$2\psi_N(u) \ge \int_0^{Q_3} 2t\mathbb{P}_n\{|\langle u, Z\rangle| > t\}\,dt - \theta\widehat{Q}^2(u)$$

and

$$2\psi_N(u) \le \int_0^{Q_4} 2t\mathbb{P}_n\{|\langle u, Z\rangle| > t\}\,dt\,.$$

Since $\theta > 7\Delta$, we have $(2\theta - 8\Delta)/3 \ge \Delta$ and therefore if $0 \le t \le Q_4$ then $\mathbb{P}\{|\langle u, Z\rangle| > t\} \ge \Delta$. Thus, by Theorem 3, for $t \in [0, Q_4]$,

$$\frac{1}{2}\mathbb{P}\{|\langle u, Z\rangle| > t\} \le \mathbb{P}_n\{|\langle u, Z\rangle| > t\} \le \frac{3}{2}\mathbb{P}\{|\langle u, Z\rangle| > t\}\,.$$

Also, just as in (2.2),

$$\theta\widehat{Q}^2 \le 2\mathbb{E}\left[\langle u, Z\rangle^2\,\mathbb{1}_{\{|\langle u,Z\rangle| \ge Q_3\}}\right].$$

Hence,

$$2\psi_N(u) \ge \frac{1}{2}\mathbb{E}\langle u, Z\rangle^2 - \frac{5}{2}\mathbb{E}\left[\langle u, Z\rangle^2\,\mathbb{1}_{\{|\langle u,Z\rangle| \ge Q_3\}}\right] \ge \frac{1}{4}\mathbb{E}\langle u, Z\rangle^2\,,$$

provided that $5\mathbb{E}\langle u, Z\rangle^2\,\mathbb{1}_{\{|\langle u,Z\rangle| \ge Q_3\}} \le \mathbb{E}\langle u, Z\rangle^2$, which holds by (4.2) of Lemma 8 whenever $2\theta + 8\Delta \le (5\kappa^2)^{-q/(q-2)}$.

In the reverse direction,

$$2\psi_N(u) \le \int_0^{Q_4} 2t\mathbb{P}_n\{|\langle u, Z\rangle| > t\}\,dt \le 2\int_0^{\infty} 2t\mathbb{P}\{|\langle u, Z\rangle| > t\}\,dt = 2\mathbb{E}\langle Z, u\rangle^2 ,$$

and combining the two inequalities we have that for every $v \in \mathbb{R}^d$,

$$\frac{1}{4}\sigma^2(u) \le \psi(v) \le 2\sigma^2(u)^2 ,$$

as claimed.

The critical level

Next we show that there exists a constant c_0 such that, for the class of functions $\mathcal{F} = \left\{\langle u, \cdot\rangle : u \in B_2^d\right\}$,

$$r = \sqrt{\frac{c_0}{n}\sum_{i \ge c_0 n} \lambda_i} \ge \overline{\rho}_n(c, \Delta, L)$$

for the values of c, Δ, L introduced the first part of the proof above. (Recall Definition 2 where $\overline{\rho}_n(c, \Delta, L)$ was introduced; also note that c is an absolute constant while Δ and L depend on the constants κ and q of the norm-equivalence condition (1.4)).

Without loss of generality, we may assume that the covariance matrix Σ is positive definite. We may write the random vector $Z = X - X'$ as $Z = TW$ where the random vector W has identity covariance matrix and $T : \mathbb{R}^d \to \mathbb{R}^d$ is a positive definite linear transformation. Since $Z/\sqrt{2}$ has the same covariance matrix Σ as X, the eigenvalues of T are $2\lambda_1, \ldots, 2\lambda_d$.

It is straightforward to verify that $D = \{u \in \mathbb{R}^d : \sigma(u) \le 1\} = T^{-1}B_2^d$ and therefore the packing numbers of Definition 2 satisfy

$$\mathcal{M}(B_2^d \cap rD, \Delta^{3/2} r/(1000 \cdot L)) = \mathcal{M}\Big(TB_2^d \cap rB_2^d, \Delta^{3/2} r/(1000 \cdot L)\Big) .$$

Also, $TB_2^d \cap rB_2^d \subset \mathcal{E}$ for an ellipsoid $\mathcal{E}$ whose principal axes are of lengths proportional to the values $\min\{\sqrt{\lambda_i}, r\}$. By Sudakov's inequality (see, e.g. [15]), there is a constant c_3 (depending on L only) such that

$$c_3\Delta^{3/2} r \log^{\frac{1}{2}} \mathcal{M}\Big(TB_2^d \cap rB_2^d, \Delta^{3/2}/(1000 \cdot L)\Big) \le \mathbb{E}\sup_{x \in \mathcal{E}}\langle G, x\rangle ,$$

where G is the standard Gaussian random vector in $\mathbb{R}^d$. A straightforward computation shows that

$$\mathbb{E}\sup_{x \in \mathcal{E}}\langle G, x\rangle \le c_4\left(\sum_{i=1}^{d} \min\left\{\lambda_i, r^2\right\}\right)^{\frac{1}{2}}$$

for a numerical constant $c_4 > 0$ and, in particular, inequality (1) in Definition 2 holds if

$$\left(\sum_{i=1}^{d} \min\left\{\lambda_i, r^2\right\}\right)^{\frac{1}{2}} \leq c_5 \Delta^2 \sqrt{n} r \,, \tag{4.5}$$

where the constant c_5 depends on L.

Turning to inequality (2) in Definition 2, observe that

$$\begin{aligned}
\mathbb{E} \sup_{u \in 2B_2^d \cap \Delta^{3/2} r D} \left| \sum_{i=1}^{n} \varepsilon_i \langle Z_i, u \rangle \right| &\leq 2\mathbb{E} \sup_{u \in T^{-1}(TB_2^d \cap \Delta^{3/2} r B_2^d)} \left| \sum_{i=1}^{n} \varepsilon_i \langle W_i, Tu \rangle \right| \\
&= 2\mathbb{E} \sup_{v \in TB_2^d \cap \Delta^{3/2} r B_2^d} \left| \sum_{i=1}^{n} \varepsilon_i \langle W_i, v \rangle \right| \\
&\leq c_6 \mathbb{E} \sup_{v \in \mathcal{E}'} \left| \sum_{i=1}^{n} \varepsilon_i \langle W_i, v \rangle \right|
\end{aligned}$$

for an ellipsoid $\mathcal{E}'$ that may be written as $\mathcal{E}' = QB_2^d$ for a linear transformation Q whose eigenvalues are proportional to $\min\{\sqrt{\lambda_i}, \Delta^{3/2} r\}$. Thus,

$$\mathbb{E} \sup_{v \in \mathcal{E}'} \left| \sum_{i=1}^{n} \varepsilon_i \langle W_i, v \rangle \right| = \mathbb{E} \sup_{v \in B_2^d} \left| \sum_{i=1}^{n} \varepsilon_i \langle QW_i, v \rangle \right| \leq \sqrt{n} \left(\mathbb{E} \|QW\|_2^2 \right)^{1/2} .$$

Since W is isotropic,

$$\mathbb{E}\|QW\|_2^2 = \sum_{i=1}^{d} \mathbb{E}\langle W, Q^* e_i \rangle^2 = \sum_{i=1}^{d} \|Q^* e_i\|_2^2 \leq c_7 \sum_{i=1}^{d} \min\{\lambda_i, \Delta^3 r^2\} \,.$$

Therefore, inequality (2) in Definition 2 is verified once

$$\left(\sum_{i=1}^{d} \min\{\lambda_i, \Delta^3 r^2\}\right)^{\frac{1}{2}} \leq c_8 \Delta^2 \sqrt{n} r \tag{4.6}$$

for a constant c_8 (that depends on κ and q). Recalling that $\Delta < 1$, it is evident that both (4.5) and (4.6) are satisfied when

$$\left(\sum_{i=1}^{d} \min\{\lambda_i, r^2\}\right)^{\frac{1}{2}} \leq c_9 \sqrt{n} r \tag{4.7}$$

for a constant $c_9 = c_9(\kappa, q)$. Using that $\min\{\lambda_i, r^2\} \leq r^2$ for $i \leq \frac{1}{2} c_9$, it suffices that

$$r^2 \geq c_0 \frac{1}{n} \sum_{i \geq c_0 n} \lambda_i \,,$$

for a constant depending on κ and q only, as claimed.

Below the critical level

Finally, we prove part (ii) of Proposition 2. Let the parameters $\theta, \gamma \in (0,1)$ be constant as specified in the proof of part (i) above.

Fix $r_0 > 0$ that satisfies (4.7) and set $\mathcal{F}_{r_0} = \{\langle u, \cdot \rangle : u \in B_2^d \cap r_0 D\}$. Our goal is to show that, with high probability, for every $u \in B_2^d \cap r_0 D$,

$$(\langle u, Z \rangle_{\theta n}))^\sharp \leq c_0 \frac{r_0}{\sqrt{\theta}}$$

for a constant c_0, where $\left(\langle u, Z \rangle_j)^\sharp\right)_{j=1}^n$ is the monotone nonincreasing rearrangement of $\langle u, Z_1 \rangle, \ldots, \langle u, Z_n \rangle$. On this event, for all $u \in B_2^d \cap r_0 D$, we have

$$\psi_N(u) \leq \frac{c_0^2}{\theta} r_0^2 \,,$$

as required.

The proof uses a net argument. Let U_{r_0} be a maximal $\sqrt{\theta} r_0$-separated subset of $B_2^d \cap r_0 D$. A standard argument using the norm equivalence condition, the union bound, and Markov's inequality shows that, for every $u \in U_{r_0}$, with probability at least $1 - \exp(-c_1 k \log(n/(c_2 k)))$,

$$(\langle u, Z \rangle)_k^\sharp \leq \| \langle u, Z \rangle \|_{L_2} \sqrt{\frac{n}{k}} \,,$$

where $c_1 = q/2 - 1$ and $c_2 = (e\kappa^q)^{2/(q-2)}$. Hence, by setting $k = \theta n/2$ and if

$$|U_{r_0}| \leq (c_1/4) n\theta \log\left(\frac{2c_2}{\theta}\right) ,$$

that is, if

$$\log \mathcal{M}(B_2^d \cap r_0 D, \sqrt{\theta} r_0) \leq (c_1/4) n\theta \log\left(\frac{2c_2}{\theta}\right) , \tag{4.8}$$

it follows that, with probability at least $1 - \exp\left(-(c_1/4) n\theta \log\left(\frac{2c_2}{\theta}\right)\right)$, for every $u \in U_{r_0}$,

$$(\langle u, Z \rangle)_{\theta n/2}^\sharp \leq \frac{r_0}{\sqrt{\theta/2}} \,.$$

Denote by πu the best approximation to u in U_{r_0} with respect to the $L_2(Z)$ norm. In particular, by the choice of U_{r_0}, $\|u - \pi u\|_{L_2} \leq \sqrt{\theta} r_0$. To complete the proof it suffices to show that, with high probability,

$$\Gamma \stackrel{\text{def.}}{=} \sup_{v \in B_2^d \cap r_0 D} \left| \left\{ j : \left| \left\langle u - \pi u, Z_j \right\rangle \right| \geq \frac{8 r_0}{\sqrt{\theta}} \right\} \right| \leq \frac{\theta n}{2} \,.$$

This follows from what is, by now, a standard argument:

By the bounded differences inequality we have that, with probability at least $1 - 2\exp(-\theta^2 n/8)$, $\Gamma \le \frac{\theta n}{2}$, provided that

$$\mathbb{E}\Gamma \le \frac{\theta n}{4} .$$

By symmetrization and contraction,

$$\begin{aligned}
\mathbb{E}\Gamma &\le \frac{\sqrt{\theta}}{8r_0}\mathbb{E}\sup_{u\in B_2^d\cap r_0 D}\sum_{j=1}^{n}|\langle u-\pi u, Z_i\rangle| \\
&\le \frac{\sqrt{\theta}}{8r_0}\left(\mathbb{E}\sup_{u\in B_2^d\cap r_0 D}\sum_{j=1}^{n}(|\langle u-\pi u, Z_i\rangle| - \mathbb{E}|\langle u-\pi u, Z_i\rangle|) + \sqrt{\theta} r_0 n\right) \\
&\le \frac{\sqrt{\theta}}{8r_0}\left(2\mathbb{E}\sup_{u\in B_2^d\cap r_0 D}\left|\sum_{j=1}^{n}\varepsilon_j\langle u-\pi u, Z_i\rangle\right| + \sqrt{\theta} r_0 n\right) \\
&\le \frac{\sqrt{\theta}}{8r_0}\left(4\mathbb{E}\sup_{u\in B_2^d\cap \sqrt{\theta} r_0 D}\left|\sum_{j=1}^{n}\varepsilon_j\langle u, Z_i\rangle\right| + \sqrt{\theta} r_0 n\right) .
\end{aligned}$$

Hence, $\mathbb{E}\Gamma \le \theta n/4$ provided that

$$\mathbb{E}\sup_{u\in B_2^d\cap \sqrt{\theta} r_0 D}\left|\sum_{j=1}^{n}\varepsilon_j\langle u, Z_i\rangle\right| \le \frac{r_0\sqrt{\theta} n}{4} .$$

This may be proved by the same argument used in the second part of the proof above. In particular, the inequality holds once $r_0 \ge \sqrt{(c_0/n)\sum_{i\ge c_0 n}\lambda_i}$ for an appropriate constant (depending on κ and q), as required. ∎

5 Multivariate mean estimator and its performance

Now we are prepared to define the mean estimator announced in Theorem 1 and prove its performance bound. The estimator receives, as input, $3N$ independent, identically distributed random vectors $X_1, \dots, X_{3N}$, the parameters $\kappa > 0$ and $q > 2$, and the confidence parameter $\delta \in (0,1)$. (The sample size is set to be $3N$ for convenience as the proposed estimator splits the data into three equal parts.)

The data $X_{N+1}, \dots, X_{3N}$ are used to estimate the variances $\sigma^2(u) = \mathrm{Var}(\langle X, u\rangle)$ for $u \in S^{d-1}$. Using the estimator ψ_N, we have that, on an event $\mathcal{A}$ of probability at

least $1 - e^{-cN}$,

$$\begin{array}{ll} \frac{1}{4}\sigma^2(u) \le \psi_N(u) \le 2\sigma^2(u) & \text{for all } u \in S^{d-1} \text{ such that } \sigma(u) \ge r \\ \psi_N(u) \le Cr^2 & \text{otherwise} \end{array} \tag{5.1}$$

Here c, C are constants depending on κ and q only and

$$r = \sqrt{\frac{c_0}{N} \sum_{i \ge c_0 N} \lambda_i}$$

for another constant $c_0 > 0$ depending on κ and q. (Recall that the variance estimator ψ_N has two parameters θ and γ, both may be determined by κ and q.)

The data $X_1, \ldots, X_N$ are used to estimate the mean $\mathbb{E}\langle X, u\rangle = \langle \mu, u\rangle$ for all $u \in S^{d-1}$. One would like to construct an estimator such that it is approximately correct simultaneously for all directions $u \in S^{d-1}$. To this end, similarly to the covariance estimation procedure of the previous section, we divide the sample $X_1, \ldots, X_N$ into $n = N/m$ blocks, each of size m, and compute, for $j \in [n]$,

$$Y_j = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} X_{m(j-1)+i} \, .$$

The recommended value of m is specified below. It is not necessarily the same as the block size in the covariance estimation procedure defined in Section 4. However, the role of this blocking procedure is the same as in the covariance estimation procedure of Section 4: by an appropriate choice of m, the random vectors Y_j satisfy the small-ball condition that allows us to apply Theorem 3. The estimator is a simple trimmed-mean estimator defined as

$$\widehat{\nu}_N(u) = \frac{1}{\sqrt{m}} \frac{1}{n - 2\theta n} \sum_{j \in [n] \setminus J_+(u) \cup J_-(u)} Y_j \, ,$$

where $\theta \in (0, 1/2)$ is a parameter of the estimator and the sets $J_+(u)$ and $J_-(u)$ correspond to the indices of the θn smallest and θn largest values of $\left\langle Y_j, u \right\rangle$. (We may assume that θn is an integer and the value of θ is specified below.) The key property of the marginal mean estimator $\widehat{\nu}_N(u)$ is summarized in the next proposition.

Proposition 3. *Assume the condition of Theorem 1. There exist choices of the parameters m and θ of the estimator $\widehat{\nu}_N(u)$ that depend only on κ and q and there exist constants $c, C' > 0$ depending on κ, q such that, with probability at least $1 - \delta$, for all $u \in S^{d-1}$,*

$$\left|\widehat{\nu}_N(u) - \langle \mu, u\rangle\right| \le C' \left(\sqrt{\frac{\sigma^2(u) \log(1/\delta)}{N}} + \sqrt{\frac{\sum_{i=c\log(1/\delta)}^{d} \lambda_i}{N}} \right) . \tag{5.2}$$

The proof of the proposition follows from Theorems 2 and 3, similarly to the arguments presented in the previous section for covariance estimation. In order to avoid repetitions, we defer the details to Section A.2 in the Appendix.

Equipped with Proposition 3, it is now easy to define the mean estimator announced in the introduction and prove Theorem 1.

Let $\psi_N(u)$ and $\widehat{\nu}_N(u)$ be the variance and marginal mean estimators defined above. For a parameter $\rho > 0$, and for each $u \in S^{d-1}$, define the slabs

$$E_{u,\rho} = \left\{ v \in \mathbb{R}^d : |\widehat{\nu}_N(u) - \langle v, u \rangle| \le \rho + 2C' \sqrt{\frac{\psi_N(u) \log(1/\delta)}{N}} \right\}$$

and let

$$S_\rho = \bigcap_{u \in S^{d-1}} E_{u,\rho} \, .$$

Note that for every $\rho > 0$, the set S_ρ is a compact set and that S_ρ is nonempty for a sufficiently large ρ. Therefore, the set

$$S = \bigcap_{\rho > 0 : S_\rho \neq \emptyset} S_\rho$$

is not empty. We define the mean estimator as any element $\mu_N \in S$.

Theorem 1 now follows easily.

Proof of Theorem 1. By Propositions 2 and 3, with probability at least $1 - \delta$, both (5.1) and (5.2) hold, where $\delta \in (0,1)$ is such that $c \log(1/\delta) \le c_0 N$. Denote this event by $\mathcal{A}$. (Here c_0 is the constant appearing in the definition of r and c is as in (5.2).)

On the event $\mathcal{A}$, if $u \in S^{d-1}$ is such that $\sigma(u) \ge r$, then $\sigma^2(u) \le 4\psi_N(u)$ and therefore, (5.2) implies that

$$\left| \widehat{\nu}_N(u) - \langle \mu, u \rangle \right| \le \rho + 2C' \sqrt{\frac{\psi_N(u) \log(1/\delta)}{N}}$$

whenever $\rho \ge C' \sqrt{\frac{\sum_{i=c\log(1/\delta)}^{d} \lambda_i}{N}}$. On the other hand, if $\sigma(u) \ge r$, then $\psi_N(u) \le Cr^2$,

and therefore, by bounding the right-hand side of (5.2) further, we get

$$\begin{aligned}\left|\widehat{\nu}_N(u) - \langle \mu, u\rangle\right| &\leq C'\left(\sqrt{\frac{Cr^2\log(1/\delta)}{N}} + \sqrt{\frac{\sum_{i=c\log(1/\delta)}^{d}\lambda_i}{N}}\right)\\ &\leq C'\sqrt{\frac{Cc_0}{c}}r + C'\sqrt{\frac{\sum_{i=c\log(1/\delta)}^{d}\lambda_i}{N}}\\ &\quad \text{(since } c\log(1/\delta)\leq c_0 N\text{)}\\ &\leq C'\left(c_0\sqrt{\frac{C}{c}}+1\right)\sqrt{\frac{\sum_{i=c\log(1/\delta)}^{d}\lambda_i}{N}}\,,\end{aligned}$$

where at the last step again we used the fact that $c\log(1/\delta)\leq c_0 N$.

Hence, on the event $\mathcal{A}$, we have that $\mu\in S_\rho$ when

$$\rho = C_1\sqrt{\frac{\sum_{i=c\log(1/\delta)}^{d}\lambda_i}{N}}\,,$$

where $C_1 = C'\left(c_0\sqrt{\frac{C}{c}}+1\right)$. Thus, S_ρ is nonempty and, by definition, $\widehat{\mu}_N\in S_\rho$ for this choice of ρ. This means that, for all $u\in S_\rho$,

$$\left|\langle\widehat{\mu}-\mu,u\rangle\right| \leq C_1\sqrt{\frac{\sum_{i=c\log(1/\delta)}^{d}\lambda_i}{N}} + 2C'\sqrt{\frac{\psi_N(u)\log(1/\delta)}{N}}\,.$$

The theorem is now proved by noticing that $\psi_N(u)\leq 2\sigma^2(u)$ when $\sigma(u)\geq r$ and $\psi_N(u)\leq Cr^2$ otherwise. ∎

A Appendix: additional proofs

A.1 Proof of Proposition 1

We may assume, without loss of generality, that Σ is diagonal such that the eigenvalues $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_d$ have the canonical basis vectors $e_1,\ldots,e_d$ as corresponding eigenvectors. Denote by $Y=\sqrt{N}\,(\widetilde{\mu}_N-\mu)$ and note that Y is a zero-mean Gaussian vector with covariance matrix Σ.

Actually, we prove the lower bound

$$S\geq C'\left(\sqrt{\frac{\sum_{i>k_0}\lambda_i}{N}}+\sqrt{\frac{\lambda_{k_0+1}\log(1/\delta)}{N}}\right), \tag{A.1}$$

which is seemingly stronger than the announced inequality. However, note that the second term in the expression of the lower bound of the strong term above satisfies

$$\sqrt{\frac{\lambda_{k_0+1}\log(1/\delta)}{N}} \sim \sqrt{\frac{k_0\lambda_{k_0+1}}{N}} \le \sqrt{\frac{2\sum_{i>k_0/2}\lambda_i}{N}}\,.$$

Hence, we do not lose much by ignoring the second term.

Suppose that (1.6) holds on an event Ω_δ such that $\mathbb{P}\{\Omega_\delta\} \ge 1-\delta$. Let k be the unique value such that $S \in (C\sqrt{\lambda_{k+1}\log(1/\delta)/N}, C\sqrt{\lambda_k\log(1/\delta)/N}]$.

Denote by $U_k \subset \mathbb{R}^d$ the vector space spanned by $e_1,\dots,e_k$. For all $u \in U_k \cap S^{d-1}$ we have $\sigma^2(u) \ge \lambda_k$, and for all such vectors $S \le C\sigma(u)\sqrt{\log(1/\delta)/N}$. Therefore, on the event Ω_δ,

$$\forall u \in U_k \cap S^{d-1} : \langle \widetilde{\mu}_N - \mu, u\rangle \le 2C\sqrt{\frac{\sigma^2(u)\log(1/\delta)}{N}}\,.$$

Equivalently,

$$\sup_{u\in U_k\cap S^{d-1}} \frac{\langle Y,u\rangle}{\sigma(u)} \le 2C\log(1/\delta)\,.$$

If $G=(G_1,\dots,G_d)$ is a standard normal vector in $\mathbb{R}^d$, then we may write $Y=\Sigma^{1/2}G$. Since $\sigma(u)=\|\Sigma^{1/2}u\|$, and Σ is a diagonal matrix,

$$\sup_{u\in U_k\cap S^{d-1}} \frac{\langle Y,u\rangle}{\sigma(u)} = \sup_{v\in U_k\cap S^{d-1}} \langle G,v\rangle = \|G^{(k)}\|\,,$$

where $G^{(k)}=(G_1,\dots,G_k)$ is a standard normal vector in $\mathbb{R}^k$. By the Gaussian concentration inequality, with probability at least $1-\delta$, we have

$$\|G^{(k)}\| \ge \mathbb{E}\|G^{(k)}\| - \sqrt{2\log(1/\delta)} \ge \sqrt{k-1} - \sqrt{2\log(1/\delta)}\,.$$

Comparing the upper and lower bounds for $\|G^{(k)}\|$, we conclude that

$$k \le 1 + (2C+\sqrt{2})^2\log(1/\delta) = k_0\,,$$

implying that

$$S \ge C\sqrt{\frac{\lambda_{k_0+1}\log(1/\delta)}{N}}\,. \tag{A.2}$$

Next we consider the orthogonal complement $U_k^\perp$ of U_k. Since $\sigma^2(u) \le \lambda_{k+1}$ for all $u \in U_k^\perp \cap S^{d-1}$, on the event Ω_δ, we have

$$\sup_{u\in U_k^\perp\cap S^{d-1}} \langle Y,u\rangle \le 2S\sqrt{N}\,.$$

Writing $Y = \Sigma^{1/2} G$ as before, and noting that $\sup_{u \in U_k^\perp \cap S^{d-1}} \left\langle \Sigma^{1/2} G, u \right\rangle$ is a Lipschitz function of G with constant $\sqrt{\lambda_{k+1}}$, the Gaussian concentration inequality implies that, with probability at least $1 - \delta$,

$$\sup_{u \in U_k^\perp \cap S^{d-1}} \langle Y, u \rangle \geq \mathbb{E} \sup_{u \in U_k^\perp \cap S^{d-1}} \langle Y, u \rangle - \sqrt{2\lambda_{k+1} \log(1/\delta)} = \mathbb{E} \|\Sigma^{1/2} G^{(k)\perp}\| - \sqrt{2\lambda_{k+1} \log(1/\delta)},$$

where $G^{(k)\perp} = (G_{k+1}, \ldots, G_d)$. By the Gaussian Poincaré inequality (see., e.g., [2, Theorem 3.20]),

$$\mathbb{E} \|\Sigma^{1/2} G^{(k)\perp}\| \geq \sqrt{\mathbb{E} \|\Sigma^{1/2} G^{(k)\perp}\|^2} - \sqrt{\lambda_{k+1}} = \sqrt{\sum_{i>k} \lambda_i} - \sqrt{\lambda_{k+1}},$$

and therefore

$$S \geq \frac{1}{2\sqrt{N}} \left(\sqrt{\sum_{i>k} \lambda_i} - \sqrt{\lambda_{k+1}} \left(1 + \sqrt{2 \log(1/\delta)} \right) \right).$$

If

$$\sqrt{\lambda_{k+1}} \left(1 + \sqrt{2 \log(1/\delta)} \right) \leq \frac{1}{2} \sqrt{\sum_{i>k} \lambda_i},$$

then

$$S \geq \frac{1}{4\sqrt{N}} \left(\sqrt{\sum_{i>k} \lambda_i} \right)$$

which, together with (A.2) and the fact that $k \leq k_0$, implies (A.1). On the other hand, if

$$\sqrt{\lambda_{k+1}} \left(1 + \sqrt{2 \log(1/\delta)} \right) > \frac{1}{2} \sqrt{\sum_{i>k} \lambda_i},$$

then (A.2) already implies inequality (A.1).

A.2 Proof of Proposition 3

For $j \in [n]$, we write $\overline{Y}_j = Y_j - \mathbb{E} Y_j = Y_j - \mu$. Then for all $u \in S^{d-1}$,

$$\left| \widehat{\nu}(u) - \langle \mu, u \rangle \right| = \frac{1}{\sqrt{m}} \left(\frac{1}{n - 2n\theta} \sum_{j \in [n] \setminus (J_+ \cup J_-)} \left\langle \overline{Y}_j, u \right\rangle \right),$$

and it suffices to obtain an upper estimate on

$$\left| \frac{1}{n - 2n\theta} \sum_{j \in [n] \setminus (J_+ \cup J_-)} \left\langle \overline{Y}_j, u \right\rangle \right|.$$

That is precisely the question addressed in Theorem 2 for $p = 1$ (and with the sample size being n rather than N). To apply Theorem 2 and the subsequent Corollary 1, one needs to ensure that the random variables $\left\langle \overline{Y}, u \right\rangle$ satisfy properties (1)-(3) of Definition 1.

Observe that $\left\| \left\langle \overline{Y, u} \right\rangle \right\|_{L_2} = \|u\|_{L_2} = \sigma(u)$ because the L_2 norm endowed by $\overline{Y}$ coincides with the one endowed by $\overline{X}$. (4.1) in Lemma 8 shows that property (3) holds for $\eta = 1/4$ if $m \geq m_0(q, \kappa)$ for a constant $m_0(q, \kappa)$.

Also, by (4.4), for any $\gamma \in (0, 1)$, if $m \geq c_1/\gamma^2$ for some constant c_1, then for any interval $I \subset \mathbb{R}$,

$$\mathbb{P}\left\{ \overline{Y} \in I \right\} \leq \max\left\{ L\frac{|I|}{\sigma(u)}, \gamma \right\} ,$$

and therefore, with such a choice of m, the class $\mathcal{F} = \left\{ \left\langle \overline{Y}, u \right\rangle : u \in B_2^d \right\}$ satisfies the assumptions of Theorem 3.

Now set $\Delta = \gamma/c_1$ (with c_1 as in the statement of Theorem 3) and choose $\theta \geq 7\Delta$. Let $\rho_1 \geq \overline{\rho}_n(c, \Delta, L)$ where $\overline{\rho}_n(c, \Delta, L)$ is the critical level of the class $\mathcal{F}$ (with c as in Theorem 3).

The theorem implies that there is an event $\mathcal{A}$ with probability at least $1 - 2\exp(-c_2\Delta n)$, such that for all $\|u\|_{L_2} \geq \rho_1$, the random variable $\left\langle v, \overline{Y} \right\rangle$ satisfies properties (1) and (2) of Definition 1. Therefore, Theorem 1 shows that on that event, if $\|u\|_{L_2} \geq \rho_1$, then

$$\left| \frac{1}{n - 2n\theta} \sum_{j \in [n] \setminus (J_+ \cup J_-)} \left\langle \overline{Y}_j, u \right\rangle \right| \leq c_3\sqrt{\Delta \log(1/\Delta)}\sigma(u) \leq c_4\sigma(u) .$$

In particular, there is a constant $c(\kappa, q)$ such that, if

$$m = c(\kappa, q)\frac{N}{\log(1/\delta)} ,$$

then $\mathbb{P}\{\mathcal{A}\} \geq 1 - \delta/2$ and on the event $\mathcal{A}$, for any $u \in S^{d-1}$ for which $\sigma(u) \geq \rho_1$,

$$\left| \widehat{\nu}_N(u) - \langle \mu, u \rangle \right| \leq c'(\kappa, q)\sigma(u)\sqrt{\frac{\log(e/\delta)}{N}} , \tag{A.3}$$

satisfying (5.2).

It remains to check that the inequality also holds for those $u \in S^{d-1}$ with $\sigma(u) < \rho_1$. Let $\rho_2 \geq \rho_1$ to be specified in what follows. Clearly, (A.3) holds when $\sigma(u) \geq \rho_2$. When $\sigma(u) < \rho_2$, one may repeat the argument used in the last part of the proof of Theorem 2 ("below the critical level"). It is evident that if

$$\rho_2 = c(\kappa, q)\sqrt{\frac{\sum_{i \geq c'(\kappa,q)n} \lambda_i}{n}} ,$$

for some constants $c(\kappa,q), c'(\kappa,q)$, then, with probability $1-2\exp(-c_1 n) \geq 1-\delta/2$,

$$\sup_{u\in B_2^d\cap\rho_2 D} \left|\left\langle \overline{Y}_j, u\right\rangle\right|_{\theta n}^{\sharp} \leq c(\kappa,q)\rho_2 \,.$$

Therefore, on this event,

$$\sup_{v\in B_2^d\cap\rho D} \left|\frac{1}{n-2n\theta}\sum_{j\in[n]\setminus(J_+\cup J_-)} \left\langle \overline{Y}_j, u\right\rangle\right| \leq c'(\kappa,q)\rho_2 \,,$$

and

$$\left|\widehat{\nu}_N(u) - \langle \mu, u\rangle\right| \leq \frac{1}{\sqrt{m}}\cdot c''(\kappa,q)\rho_2 \,.$$

The announced bound now follows for all u, on an event of probability at least $1-\delta$. ∎

B The connection with strong-weak norm inequalities

Strong-weak norm inequalities are a natural way of quantifying the tail behaviour of random vectors. Given a random vector X in $\mathbb{R}^d$ and a norm $\|\cdot\|$, we say that X satisfies a strong-weak inequality with constant C if for every $p \geq 1$,

$$(\mathbb{E}\|X-\mathbb{E}X\|^p)^{1/p} \leq C\left(\mathbb{E}\|X-\mathbb{E}X\| + \sup_{z^*\in B^*}(\mathbb{E}|x^*(X-\mathbb{E}X)|^p)^{1/p}\right) \,,$$

where B^* is the unit ball of the dual space to $(\mathbb{R}^d,\|\cdot\|)$. In other words, the way X concentrates around its mean with respect to the norm $\|\cdot\|$ is governed by the L_1 norm of $\|X-\mathbb{E}X\|$ and the largest L_p norm of all the one-dimensional marginals of the centred random vector $X-\mathbb{E}X$.

The fact that the L_p norm of $\|X-\mathbb{E}X\|$ can be controlled by such a combination and no information on the L_p norms of higher dimensional marginals is a rather powerful feature. The best type of a strong-weak inequality one can hope for is a *sub-Gaussian* one, that is, if the tails of one dimensional marginals of $X-\mathbb{E}X$ decay at least as fast as those of a Gaussian:

$$\sup_{z^*\in B^*}(\mathbb{E}|x^*(X-\mathbb{E}X)|^p)^{1/p} \leq \sqrt{p}\sup_{z^*\in B^*}\sigma(z^*) \,,$$

where $\sigma(z^*) = \left(\mathbb{E}(z^*(X-\mathbb{E}X))^2\right)^{1/2}$ is the variance of the one dimensional marginal defined by z^*. In such a case, an equivalent "in-probability" version of the strong-weak inequality is that for any $0<\delta<1/2$,

$$\mathbb{P}\left(\|X-\mathbb{E}X\| \geq C\left(\mathbb{E}\|X-\mathbb{E}X\| + \sqrt{\log(1/\delta)}\sup_{z^*\in B^*}\sigma(z^*)\right)\right) \leq \delta \,. \tag{B.1}$$

Clearly, if the one-dimensional marginals of $X - \mathbb{E}X$ do not exhibit a subgaussian tail decays, there is no hope that (B.1) can be true, even, when $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$, which is our main focus.

When the random vector $Y_N = N^{-1}\sum_{i=1}^N X_i$ satisfies a subgaussian "in probability" version of the strong-weak inequality, that implies that the empirical mean is an optimal mean estimation procedure. However, almost no random vectors satisfy that strong property. At the same time, (1.2) shows that by replacing the empirical mean with $\widehat{\mu}_N$, every random vector satisfies a version of a subgaussian strong-weak inequality. Moreover, Theorem 1 shows that under a minimal norm equivalence condition, the weak term can be replaced by the optimal directional-dependent term.

References

[1] Sohail Bahmani. Nearly optimal robust mean estimation via empirical characteristic function. *arXiv preprint arXiv:2004.02287*, 2020.

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

[3] Olivier Catoni. Pac-bayesian bounds for the gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*, 2016.

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

[5] Arnak S Dalalyan and Arshak Minasyan. All-in-one robust estimator of the gaussian mean. *arXiv preprint arXiv:2002.01432*, 2020.

[6] V.H. de la Peña and E. Giné. *Decoupling: from Dependence to Independence*. Springer, New York, 1999.

[7] J. Depersin and G. Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*, 2019.

[8] Ilias Diakonikolas, Daniel M Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. *arXiv preprint arXiv:2007.15618*, 2020.

[9] E. Giné and J. Zinn. Some limit theorems for empirical processes. *Annals of Probability*, 12:929–989, 1984.

[10] Evarist Giné and Vladimir Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.*, 34(3):1143–1216, 2006.

[11] Ilaria Giulini. Robust dimension-free Gram operator estimates. *Bernoulli*, 24(4B):3864–3923, 2018.

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

[13] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, February 2017.

[14] R. Latała and J. O. Wojtaszczyk. On the infimum convolution inequality. *Studia Math.*, 189(2):147–187, 2008.

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

[16] Zhixian Lei, Kyle Luh, Prayaag Venkat, and Fred Zhang. A fast spectral algorithm for mean estimation with sub-gaussian rates. In *Conference on Learning Theory*, pages 2598–2612, 2020.

[17] Karim Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

[18] G. Lugosi and S. Mendelson. Mean estimation and regression under heavy-tailed distributions—a survey. *Foundations of Computational Mathematics*, 2019.

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

[20] G. Lugosi and S. Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *Annals of Statistics*, to appear, 2020.

[21] S. Mendelson and N. Zhivotovskiy. Robust covariance estimation under $L_4 - L_2$ norm equivalence. *Annals of Statistics*, 48(3):16481664, 2020.

[22] Shahar Mendelson. Approximating the covariance ellipsoid. *arXiv preprint arXiv:1804.05402*, 2018.

[23] Shahar Mendelson. Approximating l_p unit balls via random sampling. *arXiv preprint arXiv:2008.08380*, 2020.

[24] Shahar Mendelson. Learning bounded subsets of l_p. *arXiv preprint arXiv:2002.01182*, 2020.

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

[26] Stanislav Minsker and Timothée Mathieu. Excess risk bounds in robust empirical risk minimization. *arXiv preprint arXiv:1910.07485*, 2019.

[27] Stanislav Minsker and Mohamed Ndaoud. Robust and efficient mean estimation: approach based on the properties of self-normalized sums. *arXiv preprint arXiv:2006.01986*, 2020.

[28] Stanislav Minsker and Xiaohan Wei. Robust modifications of u-statistics and applications to covariance estimation problems. *Bernoulli*, 26(1):694–727, 2020.

[29] E. Mossel, R. O'Donnell, and K. Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. *Annals of Mathematics*, 171:295–341, 2010.

[30] M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996.

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