SUB-GAUSSIAN MEAN ESTIMATORS

BY LUC DEVROYE[1,*], MATTHIEU LERASLE[†],
GABOR LUGOSI[2,3,‡] AND ROBERTO I. OLIVEIRA[2,4,5,§]

McGill University, CNRS—Université Nice Sophia Antipolis[†],
ICREA and Universitat Pompeu Fabra[‡] and IMPA[§]*

We discuss the possibilities and limitations of estimating the mean of a real-valued random variable from independent and identically distributed observations from a nonasymptotic point of view. In particular, we define estimators with a sub-Gaussian behavior even for certain heavy-tailed distributions. We also prove various impossibility results for mean estimators.

1. Introduction. Estimating the mean of a probability distribution P on the real line based on a sample $X_1^n = (X_1, \ldots, X_n)$ of n independent and identically distributed random variables is arguably the most basic problem of statistics. While the standard empirical mean

$$\widehat{\mathrm{emp}}_n(X_1^n) = \frac{1}{n} \sum_{i=1}^n X_i$$

is the most natural choice, its finite-sample performance is far from optimal when the distribution has a heavy tail.

The central limit theorem guarantees that if the X_i have a finite second moment, this estimator has Gaussian tails, asymptotically, when $n \to \infty$. Indeed,

$$(1) \qquad \mathbb{P}\left(\left| \widehat{\mathrm{emp}}_n(X_1^n) - \mu_P \right| > \frac{\sigma_P \Phi^{-1}(1 - \delta/2)}{\sqrt{n}} \right) \to \delta,$$

where μ_P and $\sigma_P^2 > 0$ are the mean and variance of P (resp.) and Φ is the cumulative distribution function of the standard normal distribution. This result is essentially optimal: no estimator can have better-than-Gaussian tails for all distributions in any "reasonable class" (cf. Remark 1 below).

This paper is concerned with a nonasymptotic version of the mean estimation problem. We are interested in large, nonparametric classes of distributions, such as

(2) $\mathcal{P}_2 := \{\text{all distributions over } \mathbb{R} \text{ with finite second moment}\}$,

(3) $\mathcal{P}_2^{\sigma^2} := \{\text{all distributions P} \in \mathcal{P}_2 \text{ with variance } \sigma_P^2 = \sigma^2\}$ $(\sigma^2 > 0)$,

(4) $\mathcal{P}_{\mathrm{krt} \leq \kappa} := \{\text{all P} \in \mathcal{P}_2 \text{ with kurtosis} \leq \kappa\}$ $(\kappa > 1)$,

as well as some other classes introduced in Section 3. Given such a class $\mathcal{P}$, we would like to construct *sub-Gaussian estimators*. These should take an i.i.d. sample X_1^n from some unknown P $\in \mathcal{P}$ and produce an estimate $\widehat{E}_n(X_1^n)$ of μ_P that satisfies

(5) $\mathbb{P}\left(|\widehat{E}_n(X_1^n) - \mu_P| > L\sigma_P\sqrt{\dfrac{(1 + \ln(1/\delta))}{n}}\right) \leq \delta$ for all $\delta \in [\delta_{\min}, 1)$

for some constant $L > 0$ that depends only on $\mathcal{P}$. One would like to keep $\delta_{\min}$ as small as possible (say exponentially small in n).

Of course, when $n \to \infty$ with δ fixed, (5) is a weaker form of (1) since $\Phi^{-1}(1 - \delta/2) \leq \sqrt{2\ln(2/\delta)}$. The point is that (5) should hold nonasymptotically, for extremely small δ, and uniformly over P $\in \mathcal{P}$, even for classes $\mathcal{P}$ containing distributions with heavy tails. The empirical mean cannot satisfy this property unless either $\mathcal{P}$ contains only sub-Gaussian distributions or δ_{min} is quite large (cf. Section 2.3.1), so designing sub-Gaussian estimators with the kind of guarantee we look for is a nontrivial task.

In this paper, we prove that, for most (but not all) classes $\mathcal{P} \subset \mathcal{P}_2$ we consider, there do exist estimators that achieve (5) for all large n, with $\delta_{\min} \approx e^{-c_{\mathcal{P}}n}$ and a value of L that does not depend on δ or n. In each case, $c_{\mathcal{P}} > 0$ is a constant that depends on the class $\mathcal{P}$ under consideration, and we also obtain nearly tight bounds on how $c_{\mathcal{P}}$ must depend on $\mathcal{P}$. (In particular, $\delta_{\min}$ cannot be super-exponentially small in n.) In the specific case of bounded-kurtosis distributions [cf. (4) above], we achieve $L \leq \sqrt{2} + \varepsilon$ for $\delta_{\min} \approx e^{-o((n/\kappa)^{2/3})}$. This value of L is nearly optimal by Remark 1 below.

Before this paper, it was known that (5) could be achieved for the whole class $\mathcal{P}_2$ of distributions with finite second moments, with a weaker notion of estimator that we call δ-*dependent estimator*, that is, an estimator $\widehat{E}_n = \widehat{E}_{n,\delta}$ that may also depend on the confidence parameter δ. By contrast, the estimators that we introduce here are called *multiple-δ estimators*: a single estimator works for the whole range of $\delta \in [\delta_{\min}, 1)$. This distinction is made formal in Definition 1 below. By way of comparison, we also prove some results on δ-dependent estimators in the paper. In particular, we show that the distinction is substantial. For instance, there are no multiple-δ sub-Gaussian estimators for the full class $\mathcal{P}_2$ for any nontrivial range of $\delta_{\min}$. Interestingly, multiple-δ estimators do exist (with $\delta_{\min} \approx e^{-cn}$)

for the class $\mathcal{P}_2^{\sigma^2}$ (corresponding to fixed variance). In fact, this is true when the variance is "known up to constants", but not otherwise.

Why finite variance? In all examples mentioned above, we assume that all distributions $P \in \mathcal{P}$ have a finite variance σ_P^2. In fact, our definition (5) implicitly requires that the variance exists for all $P \in \mathcal{P}$. A natural question is if this condition can be weakened. For example, for any $\alpha \in (0, 1]$ and $M > 0$, one may consider the class $\mathcal{P}_{1+\alpha}^M$ of all distributions whose $(1 + \alpha)$th central moment equals M (i.e., $\mathbb{E}[|X - \mathbb{E}X|^{1+\alpha}] = M$ if X is distributed according to any $P \in \mathcal{P}_{1+\alpha}^M$). It is natural to ask whether there exist estimators of the mean satisfying (5) with σ_P replaced by some constant depending on P. In Theorem 3.1, we prove that for every sample size n, $\delta < 1/2$, $\alpha \in (0, 1]$, and for any mean estimator $\widehat{E}_{n,\delta}$, there exists a distribution $P \in \mathcal{P}_{1+\alpha}^M$ such that with probability at least δ, the estimator is at least $M^{1/(1+\alpha)}(\frac{\ln(1/\delta)}{n})^{\alpha/(1+\alpha)}$ away from the target μ_P.

This result not only shows that one cannot expect sub-Gaussian confidence intervals for classes that contain distributions of infinite variance but also that in such cases it is impossible to have confidence intervals whose length scales as $n^{-1/2}$.

Weakly sub-Gaussian estimators. Consider the class $\mathcal{P}^{\text{Ber}}$ of all Bernoulli distributions, that is, the class that contains all distributions P of the form

$$P(\{1\}) = 1 - P(\{0\}) = p, \qquad p \in [0, 1].$$

Perhaps surprisingly, no multiple-δ estimator exists for this class of distributions, even when $\delta_{\min}$ is a constant. (We do not explicitly prove this here but it is easy to deduce it using the techniques of Sections 4.3 and 4.5.) On the other hand, by standard tail bounds for the binomial distribution (e.g., by Hoeffding's inequality), the standard empirical mean satisfies, for all $\delta > 0$ and $P \in \mathcal{P}^{\text{Ber}}$,

$$\mathbb{P}\left(\left|\widehat{\text{emp}}_n(X_1^n) - \mu_P\right| > \sqrt{\frac{\ln(2/\delta)}{2n}}\right) \leq \delta.$$

Of course, this bound has a sub-Gaussian flavor as it resembles (5) except that the confidence bounds do not scale by $\sigma_P(\log(1/\delta)/n)^{1/2}$ but rather by a distribution-free constant times $(\log(1/\delta)/n)^{1/2}$.

In general, we may call an estimate weakly sub-Gaussian with respect to the class $\mathcal{P}$ if there exists a constant $\overline{\sigma}_{\mathcal{P}}$ such that for all $P \in \mathcal{P}$,

$$\mathbb{P}\left(\left|\widehat{E}_n(X_1^n) - \mu_P\right| > L\overline{\sigma}_{\mathcal{P}}\sqrt{\frac{(1 + \ln(1/\delta))}{n}}\right) \leq \delta \qquad \text{for all } \delta \in [\delta_{\min}, 1)$$

for some constant $L > 0$. δ-dependent and multiple-δ versions of this definition may be given in analogy to those of sub-Gaussian estimators.

Note that if a class $\mathcal{P}$ is such that $\sup_{P \in \mathcal{P}} \sigma_P < \infty$, then any sub-Gaussian estimator is weakly sub-Gaussian. However, for classes of distributions without uniformly bounded variance, this is not necessarily the case and the two notions are incomparable.

In this paper, we focus on the notion of sub-Gaussian estimators and we do not pursue further the characterization of the existence of weakly sub-Gaussian estimators.

1.1. *Related work.* To our knowledge, the explicit distinction between δ-dependent and multiple-δ estimators, and our construction of multiple-δ sub-Gaussian estimators for exponentially small δ, are all new. On the other hand, constructions of δ-dependent estimators are implicit in older work on stochastic optimization of Nemirovsky and Yudin [17] (see also Levin [15] and Hsu [7]), sampling from large discrete structures by Jerrum, Valiant and Vazirani [10], and for sketching algorithms, see Alon, Matias and Szegedy [1]. Recently, there has been a surge of interest in sub-Gaussian estimators, their generalizations to multivariate settings, and their applications in a variety of statistical learning problems where heavy-tailed distributions may be present; see, for example, Catoni [5], Hsu and Sabato [9], Brownlees, Joly and Lugosi [3], Lerasle and Oliveira [14], Minsker [16], Audibert and Catoni [2], Bubeck, Cesa-Bianchi and Lugosi [4]. Most of these papers use δ-dependent sub-Gaussian estimators. Catoni's paper [5] is close in spirit to ours, as it focuses on sub-Gaussian mean estimation as a fundamental problem. That paper presents δ-dependent sub-Gaussian estimators with nearly optimal $L = \sqrt{2} + o(1)$ for a wide range of δ and the classes $\mathcal{P}_2^{\sigma^2}$ and $\mathcal{P}_{\mathrm{krt}\leq\kappa}$ defined in (3). The δ-dependent sub-Gaussian estimator introduced by [5] may be converted into a multiple-δ estimators with sub-exponential (instead of sub-Gaussian) tails for $\mathcal{P}_2^{\sigma^2}$ by choosing the single parameter of the estimator appropriately. Loosely speaking, this corresponds to squaring the term $\ln(1/\delta)$ in (5). Catoni also obtains multiple-δ estimators for $\mathcal{P}_2$ with sub-exponential tails. These ideas are strongly related to Audibert and Catoni's paper on robust least-squares linear regression [2].

1.2. *Main proof ideas.* The *negative results* we prove in this paper are minimax lower bounds for simple families of distributions such as scaled Bernoulli distributions (Theorem 3.1), Laplace distributions with fixed scale parameter for δ-dependent (Theorem 4.3), and the Poisson family for multiple-δ estimators (Theorem 4.4). The main point about the latter choices is that it is easy to compare the probabilities of events when one changes the values of the parameter. Interestingly, Catoni's lower bounds in [5] also follow from a one- dimensional family (in that case, Gaussians with fixed variance $\sigma^2 > 0$).

Our *constructions of estimators* use two main ideas. The first one is that, while one cannot turn δ-dependent into multiple-δ estimators, one *can* build multiple-δ estimators from the slightly stronger concept of *sub-Gaussian confidence intervals*. That is, if for each $\delta > 0$ one can find an empirical confidence interval for μ_{P} with "sub-Gaussian length," one may combine these intervals to produce a single multiple-δ estimator. This general construction is presented in Section 4.2 and is related at a high level to Lepskii's adaptation method [12, 13].

Although general, this method of confidence intervals loses constant factors. Our second idea for building estimators, which is specific to the bounded kurtosis case (see Theorem 3.6 below), is to use a data-driven truncation mechanism to make the empirical mean better behaved. By using preliminary estimators of the mean and variance, we truncate the random variables in the sample and obtain a Bennett-type concentration inequality with sharp constant $L = \sqrt{2} + o(1)$. A crucial point in this analysis is to show that our truncation mechanism is fairly insensitive to the preliminary estimators being used.

1.3. *Organization.* The remainder of the paper is organized as follows. Section 2 fixes notation, formally defines our problem and discusses previous work in light of our definition. Section 3 states our main results. Several general methods that we use throughout the paper are collected in Section 4. Proofs of the main results are given in Sections 5 to 7. Section 8 discusses several open problems.

2. Preliminaries.

2.1. *Notation.* We write $\mathbb{N} = \{0, 1, 2, \ldots\}$. For a positive integer n, denote $[n] = \{1, \ldots, n\}$. $|A|$ or $\#A$ denote the cardinality of the finite set A.

We treat $\mathbb{R}$ and $\mathbb{R}^n$ as measurable spaces with the respective Borel σ-fields kept implicit. Elements of $\mathbb{R}^n$ are denoted by $x_1^n = (x_1, \ldots, x_n)$ with $x_1, \ldots, x_n \in \mathbb{R}$.

Probability distributions over $\mathbb{R}$ are denoted P. Given a (suitably measurable) function $f = f(X, \theta)$ of a real-valued random variable X distributed according to P and some other parameter θ, we let

$$\mathrm{P}f = \mathrm{P}f(X, \theta) = \int_{\mathbb{R}} f(x, \theta)\mathrm{P}(dx)$$

denote the integral of f with respect to X. Assuming $\mathrm{P}X^2 < \infty$, we use the symbols $\mu_{\mathrm{P}} = \mathrm{P}X$ and $\sigma_{\mathrm{P}}^2 = \mathrm{P}X^2 - \mu_{\mathrm{P}}^2$ for the mean and variance of P.

$Z =_d \mathrm{P}$ means that Z is a random object (taking values in some measurable space) and P is the distribution of this object. $X_1^n =_d \mathrm{P}^{\otimes n}$ means that $X_1^n = (X_1, \ldots, X_n)$ is a random vector in $\mathbb{R}^n$ with the product distribution corresponding to P. Moreover, given such a random vector X_1^n and a nonempty set $B \subset [n]$, $\widehat{\mathrm{P}}_B$ is the empirical measure of X_i, $i \in B$:

$$\widehat{\mathrm{P}}_B = \frac{1}{|B|} \sum_{i \in B} \delta_{X_i}.$$

We write $\widehat{\mathrm{P}}_n$ instead of $\widehat{\mathrm{P}}_{[n]}$ for simplicity.

2.2. *The sub-Gaussian mean estimation problem.* In this section, we begin a more formal discussion of the main problem in this paper. We start with the definition of a sub-Gaussian estimator of the mean.

DEFINITION 1. Let n be a positive integer, $L > 0$, $\delta_{\min} \in (0, 1)$. Let $\mathcal{P}$ be a family of probability distributions over $\mathbb{R}$ with finite second moments.

1. δ-dependent sub-Gaussian estimation: A δ-dependent L-sub-Gaussian estimator for $(\mathcal{P}, n, \delta_{\min})$ is a measurable mapping $\widehat{E}_n : \mathbb{R}^n \times [\delta_{\min}, 1) \to \mathbb{R}$ such that if $P \in \mathcal{P}$, $\delta \in [\delta_{\min}, 1)$, and $X_1^n = (X_1, \ldots, X_n)$ is a sample of i.i.d. random variables distributed as P, then

$$(6) \qquad \mathbb{P}\left(\left| \widehat{E}_n(X_1^n, \delta) - \mu_P \right| > L\sigma_P \sqrt{\frac{(1 + \ln(1/\delta))}{n}} \right) \le \delta.$$

We also write $\widehat{E}_{n,\delta}(\cdot)$ for $\widehat{E}_n(\cdot, \delta)$.

2. Multiple-δ sub-Gaussian estimation: A multiple-δ L-sub-Gaussian estimator for $(\mathcal{P}, n, \delta_{\min})$ is a measurable mapping $\widehat{E}_n : \mathbb{R}^n \to \mathbb{R}$ such that, for each $\delta \in [\delta_{\min}, 1)$, $P \in \mathcal{P}$ and i.i.d. sample $X_1^n = (X_1, \ldots, X_n)$ distributed as P,

$$(7) \qquad \mathbb{P}\left(\left| \widehat{E}_n(X_1^n) - \mu_P \right| > L\sigma_P \sqrt{\frac{(1 + \ln(1/\delta))}{n}} \right) \le \delta.$$

It transpires from these definitions that multiple-δ estimators are preferable whenever they are available, because they combine good typical behavior with nearly optimal bounds under extremely rare events. By contrast, the need to commit to a δ in advance means that δ-dependent estimators may be too pessimistic when a small δ is desired. The main problem addressed in this paper is the following:

Given a family $\mathcal{P}$ (or more generally a sequence of families $\mathcal{P}_n$), find the smallest possible sequence $\delta_{\min} = \delta_{\min,n}$ such that multiple-δ L-sub-Gaussian estimators for $(\mathcal{P}, n, \delta_{\min,n})$ [resp., $(\mathcal{P}_n, n, \delta_{\min,n})$] exist for all large n, and with a constant L that does not depend on n.

REMARK 1 (Optimality of sub-Gaussian estimators). Call a class $\mathcal{P}$ "reasonable" when it contains all Gaussian distributions with a given variance $\sigma^2 > 0$. Catoni [5], Proposition 6.1, shows that, if $\delta \in (0, 1)$, $\mathcal{P}$ is reasonable and some estimator $\widehat{E}_{n,\delta}$ achieves

$$\mathbb{P}\left(\widehat{E}_{n,\delta}(X_1^n) - \mu_P > \frac{r\sigma_P}{\sqrt{n}} \right) \le \delta \qquad \text{whenever } P \in \mathcal{P},$$

then $r \ge \Phi^{-1}(1 - \delta)$. The same result holds for the lower tail. Since $\Phi^{-1}(1 - \delta) \sim \sqrt{2\ln(1/\delta)}$ for small δ, this means that, for any reasonable class $\mathcal{P}$, no constant $L < \sqrt{2}$ is achievable for small $\delta_{\min}$, and no better dependence on n or δ is possible. In particular, sub-Gaussian estimators are optimal up to constants, and estimators with $L \le \sqrt{2} + o(1)$ are "nearly optimal."

2.3. *Known examples from previous work.* In what follows, we present some known estimators of the mean and discuss their sub-Gaussian properties (or lack thereof).

2.3.1. *Empirical mean as a sub-Gaussian estimator.* For large n, $\sigma^2 > 0$ fixed and $\delta_{\min} \to 0$, the empirical mean

$$\widehat{\mathrm{emp}}_n(x_1^n) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

is *not* a L-sub-Gaussian estimator for the class $\mathcal{P}_2^{\sigma^2}$ of all distributions with variance σ^2. This is a consequence of [5], Proposition 6.2, which shows that the deviation bound obtained from Chebyshev's inequality is essentially sharp.

Things change under slightly stronger assumptions. For example, a nonuniform version of the Berry–Esséen theorem ([18], Theorem 14, page 125) implies that, for large n, $\widehat{\mathrm{emp}}_n$ is a multiple-δ $(\sqrt{2} + \varepsilon)$-sub-Gaussian estimator for $(\mathcal{P}_{3,\eta}, n, \delta_{\min,n})$, where

$$\mathcal{P}_{3,\eta} = \left\{ \mathrm{P} \in \mathcal{P}_2 : \mathrm{P}|X - \mu_{\mathrm{P}}|^3 \leq (\eta\sigma)^3 \right\}$$

for some $\eta > 1$ and $\delta_{\min,n} \gg n^{-1/2}(\log n)^{-3/2}$. Similar results (with worse constants) hold for the class $\mathcal{P}_{\mathrm{krt} \leq \kappa}$ [cf. (4)] when $\delta_{\min} \gg 1/n$ and κ is bounded [5], Proposition 5.1. Catoni ([5], Proposition 6.3) shows that the sub-Gaussian property breaks down when $\delta_{\min} = o(1/n)$. Exponentially small $\delta_{\min}$ can be achieved under much stronger assumptions. For example, Bennett's inequality implies that $\widehat{\mathrm{emp}}_n$ is $(\sqrt{2} + \varepsilon)$-sub-Gaussian for the triple $(\mathcal{P}_{\infty,\eta}, n, \delta_{\min})$, with $\delta_{\min} = e^{-\varepsilon^2 n/\eta^2}$ and

$$\mathcal{P}_{\infty,\eta} := \left\{ \mathrm{P} \in \mathcal{P}_2 : |X - \mu_{\mathrm{P}}| \leq \eta\sigma_{\mathrm{P}} \text{ a.s.} \right\}.$$

2.3.2. *Median of means.* Quite remarkably, as it has been known for some time, one can do much better than the empirical mean in the δ-dependent setting. The so-called *median of means* construction gives L-sub-Gaussian estimators $\widehat{E}_{n,\delta}$ (with L some constant) for any triple $(\mathcal{P}_2, n, e^{1-n/2})$ where $n \geq 6$. The basic idea is to partition the data into disjoint blocks, calculate the empirical mean within each block, and finally take the median of them. This construction with a basic performance bound is reviewed in Section 4.1, as it provides a building block and an inspiration for the new constructions in this paper. We emphasize that, as pointed out in the Introduction, variants of this result have been known for a long time; see Nemirovsky and Yudin [17], Levin [15], Jerrum, Valiant and Vazirani [10], and Alon, Matias and Szegedy [1]. Note that this estimator has good performance even for distributions with infinite variance (see the remark following Theorem 3.1 below).

2.3.3. *Catoni's estimators.* The constant L obtained by the median-of-means estimator is larger than the optimal value $\sqrt{2}$ (see Remark 1). Catoni [5] designs δ-dependent sub-Gaussian estimators with nearly optimal $L = \sqrt{2} + o(1)$ for the classes $\mathcal{P}_2^{\sigma^2}$ (known variance) and $\mathcal{P}_{\mathrm{krt} \leq \kappa}$ (bounded kurtosis). A variant of Catoni's estimator is a multiple-δ estimator, however, with sub-exponential instead of sub-Gaussian tails [i.e., the $\sqrt{\ln(1/\delta)}$ term in (7) appears squared]. Both estimators work for exponentially small δ, although the constant in the exponent for $\mathcal{P}_{\mathrm{krt} \leq \kappa}$ depends on κ.

3. Main results. Here, we present the main results of the paper. Proofs are deferred to Sections 4–7.

3.1. *On the nonexistence of sub-Gaussian mean estimators.* Recall that for any $\alpha, M > 0$, $\mathcal{P}_{1+\alpha}^M$ denotes the class of all distributions on $\mathbb{R}$ whose $(1 + \alpha)$th central moment equals M (i.e., $\mathbb{E}[|X - \mathbb{E}X|^{1+\alpha}] = M$). We start by pointing out that when $\alpha < 1$, no sub-Gaussian estimators exist (even if one allows δ-dependent estimators).

THEOREM 3.1. *Let $n > 5$ be a positive integer, $M > 0$, $\alpha \in (0, 1]$, and $\delta \in (2e^{-n/4}, 1/2)$. Then for any mean estimator $\widehat{E}_n$,*

$$\sup_{P \in \mathcal{P}_{1+\alpha}^M} \mathbb{P}\left(|\widehat{E}_n(X_1^n, \delta) - \mu_P| > \left(\frac{M^{1/\alpha} \ln(2/\delta)}{n} \right)^{\alpha/(1+\alpha)} \right) \geq \delta.$$

The proof is given in Section 4.3. The bound of the theorem is essentially tight. It is shown in Bubeck, Cesa-Bianchi and Lugosi [4] that for each $M > 0$, $\alpha \in (0, 1]$, and δ, there exists an estimator $\widehat{E}_n(X_1^n, \delta)$ such that

$$\sup_{P \in \mathcal{P}_{1+\alpha}^M} \mathbb{P}\left(|\widehat{E}_n(X_1^n, \delta) - \mu_P| > \left(8\frac{(12M)^{1/\alpha} \ln(1/\delta)}{n} \right)^{\alpha/(1+\alpha)} \right) \leq \delta.$$

The estimator $\widehat{E}_n(X_1^n, \delta)$ satisfying this bound is the median-of-means estimator with appropriately chosen parameters.

It is an interesting question whether multiple-δ estimators exist with similar performance. Since our primary goal in this paper is the study of sub-Gaussian estimators, we do not pursue the case of infinite variance further.

3.2. *The value of knowing the variance.* Given $0 < \sigma_1 \leq \sigma_2 < \infty$, define the class of distributions with variance between σ_1^2 and σ_2^2:

$$\mathcal{P}_2^{[\sigma_1^2, \sigma_2^2]} = \{P \in \mathcal{P}_2 : \sigma_1^2 \leq \sigma_P^2 \leq \sigma_2^2\}.$$

This class interpolates between the classes of distributions with fixed variance $\mathcal{P}_2^{\sigma^2}$ and with completely unknown variance $\mathcal{P}_2$. The next theorem is proved in Section 5.

THEOREM 3.2. *Let $0 < \sigma_1 < \sigma_2 < \infty$ and define $R = \sigma_2/\sigma_1$.*

1. *Letting $L^{(1)} = (4e\sqrt{2 + 4\ln 2})R$ and $\delta_{\min}^{(1)} = 4e^{1-n/2}$, for every $n \geq 6$ there exists a multiple-δ $L^{(1)}$-sub-Gaussian estimator for $(\mathcal{P}_2^{[\sigma_1^2, \sigma_2^2]}, n, \delta_{\min}^{(1)})$.*

2. *For any $L \geq \sqrt{2}$, there exist $\phi^{(2)} > 0$ and $\delta_{\min}^{(2)} > 0$ such that, when $R > \phi^{(2)}$, there is no multiple-δ L-sub-Gaussian estimator for $(\mathcal{P}_2^{[\sigma_1^2, \sigma_2^2]}, n, \delta_{\min}^{(2)})$ for any n.*

3. *For any value of $R \geq 1$ and $L \geq \sqrt{2}$, if we let $\delta_{\min}^{(3)} = e^{1-9L^2 n}$, there is no δ-dependent L-sub-Gaussian estimator for $(\mathcal{P}_2^{[\sigma_1^2, \sigma_2^2]}, n, \delta_{\min}^{(3)})$ for any n.*

It is instructive to consider this result when n grows and $R = R_n$ may change with n. The theorem says that, when $\sup_n R_n < \infty$, there are multiple-δ L-sub-Gaussian estimators for all large n, with exponentially small $\delta_{\min}$ and a constant L. On the other hand, if $R_n \to \infty$, for any constant L and all large n, no multiple-δ L-sub-Gaussian estimators exist for any sequence $\delta = \delta_{\min,n} \to 0$. Finally, the third item says that even when $R_n \equiv 1$, δ-dependent estimators are limited to $\delta_{\min} = e^{-O(n)}$, so the median-of-means estimator is optimal in this sense.

3.3. *Regularity, symmetry and higher moments.* The following shows that what we call *regularity conditions* can substitute for knowledge of the variance.

DEFINITION 2. For $P \in \mathcal{P}_2$ and $j \in \mathbb{N} \setminus \{0\}$, let $X_1, \ldots, X_j$ be i.i.d. random variables with distribution P. Define

$$p_-(P, j) = \mathbb{P}\left(\sum_{i=1}^j X_i \leq j\mu_P\right) \quad \text{and} \quad p_+(P, j) = \mathbb{P}\left(\sum_{i=1}^j X_i \geq j\mu_P\right).$$

Given $k \in \mathbb{N}$, we define the *k-regular class* as follows:

$$\mathcal{P}_{2,k\text{-reg}} = \{P \in \mathcal{P}_2 : \forall j \geq k, \min(p_+(P, j), p_-(P, j)) \geq 1/3\}.$$

Note that this family of distributions is increasing in k. Also note that $\bigcup_{k \in \mathbb{N}} \mathcal{P}_{2,k\text{-reg}} = \mathcal{P}_2$, because the central limit theorem implies $p_+(P, j) \to 1/2$ and $p_-(P, j) \to 1/2$. Here are two important examples of large families of distributions in this class.

EXAMPLE 3.1. We say that a distribution $P \in \mathcal{P}_2$ is *symmetric around the mean* if, given $X =_d P$, $2\mu_P - X =_d P$ as well. Clearly, if P has this property, $p_+(P, j) = p_-(P, j) = 1/2$ for all j and thus $P \in \mathcal{P}_{2,1\text{-reg}}$. In other words, $\mathcal{P}_{2,\text{sym}} \subset \mathcal{P}_{2,1\text{-reg}}$ where $\mathcal{P}_{2,\text{sym}}$ is the class of all $P \in \mathcal{P}_2$ that are symmetric around the mean.

EXAMPLE 3.2. Given $\eta \geq 1$ and $\alpha \in (2, 3]$, set

$$(8) \qquad \mathcal{P}_{\alpha,\eta} = \{P \in \mathcal{P}_2 : P|X - \mu_P|^\alpha \leq (\eta \sigma_P)^\alpha\}.$$

We show in Lemma 6.2 that, for P in this family, $\min(p_+(P, j), p_-(P, j)) \geq 1/3$ once $j \geq (C_\alpha \eta)^{2\alpha/(\alpha-2)}$ for a constant C_α depending only on α. We deduce

$$\mathcal{P}_{\alpha,\eta} \subset \mathcal{P}_{2,k\text{-reg}} \qquad \text{if } k \geq (C_\alpha \eta)^{2\alpha/(\alpha-2)}.$$

Our main result about k-regular classes states that sub-Gaussian multiple-δ estimators exist for $\mathcal{P}_{2,k\text{-reg}}$ in the sense of the following theorem, proven in Section 6.1.

THEOREM 3.3. *Let n, k be positive integers with $n \geq (3 + \ln 4)124k$. Set $\delta_{\min,n,k} = 4e^{3-n/(124k)}$ and $L_* = 4\sqrt{2(1 + 2\ln 2)(1 + 62\ln(3))}e^{5/2}$. Then there exists a L_*-sub-Gaussian multiple-δ estimator for $(\mathcal{P}_{2,k\text{-reg}}, n, \delta_{\min,n,k})$.*

We also show that the range of $\delta_{\min} = e^{-O(n/k)}$ in this result is optimal. This follows directly from stronger results that we prove for Examples 3.1 and 3.2. In other words, the general family of estimators designed for k-regular classes has nearly optimal range of δ for these two smaller classes. The next result, for symmetric distributions, is proven in Section 6.2.

THEOREM 3.4. *Consider the class $\mathcal{P}_{2,\text{sym}}$ defined in Example 3.1. Then*:

1. *the estimator obtained in Theorem 3.3 for $k = 1$ is a L_*-sub-Gaussian multiple-δ estimator for $(\mathcal{P}_{2,\text{sym}}, n, \delta_{\min,n,1})$ when $n \geq (3 + \ln 2)124$;*

2. *on the other hand, for any $L \geq \sqrt{2}$, no δ-dependent L-sub-Gaussian estimator can exist for $(\mathcal{P}_{2,\text{sym}}, n, e^{1-9L^2 n})$.*

We also have an analogous result for the class $\mathcal{P}_{\alpha,\eta}$. The proof may be found in Section 6.3.

THEOREM 3.5. *Fix $\alpha \in (2, 3]$ and assume $\eta \geq 3^{1/3}2^{1/6}$. Consider the class $\mathcal{P}_{\alpha,\eta}$ defined in Example 3.2. Then there exists some $C_\alpha > 0$ depending only on α such that if $k_\alpha = \lceil C_\alpha \eta^{(2\alpha)/(\alpha-2)} \rceil$:*

1. *the estimator obtained in Theorem 3.3 for $k = k_\alpha$ is a L_*-sub-Gaussian multiple-δ estimator for $(\mathcal{P}_{\alpha,\eta}, n, \delta_{\min,n,k_\alpha})$ when $n \geq (3 + \ln 4)124k_\alpha$;*

2. *on the other hand, for any $L \geq \sqrt{2}$, there exist $n_{0,\alpha,L} \in \mathbb{N}$ and $c_{\alpha,L} > 0$ such that no multiple-δ L-sub-Gaussian estimator can exist for $(\mathcal{P}_{\alpha,\eta}, n, e^{1-c_{\alpha,L}n/k_\alpha})$ when $n \geq n_{0,\alpha,L}$ is large enough;*

3. *finally, for $L \geq \sqrt{2}$ there is no δ-dependent L sub-Gaussian estimator for $(\mathcal{P}_{\alpha,\eta}, n, e^{1-9L^2 n})$.*

3.4. *Bounded kurtosis and nearly optimal constants.* This section shows that multiple-δ sub-Gaussian estimation with nearly optimal constants can be proved when the kurtosis

$$\kappa_P = \frac{P(X - \mu_P)^4}{\sigma_P^4}$$

is uniformly bounded in the class. (We set $\kappa_P = 1$ when $\sigma_P^2 = 0$.) More specifically, we consider the class $\mathcal{P}_{krt \leq \kappa}$ of all distributions $P \in \mathcal{P}_2$ with $\kappa_P \leq \kappa$.

To state the result, let b_{max} be a positive integer to be specified below. Also define

$$\xi = 2\sqrt{2}\kappa \frac{b_{max}^{3/2}}{n} + 36\sqrt{\frac{\kappa b_{max}}{n}} + 1120\sqrt{\kappa}\frac{b_{max}}{n}.$$

Note that when $b_{max} = o((n/\kappa)^{2/3})$, $\xi = o(1)$. The main result for classes of distributions with bounded kurtosis is the following. For the proof, see Section 7.

THEOREM 3.6. *Let $n \geq 4$, $L = \sqrt{2}(1 + \xi)$, $\delta_{min}^{(4)} = \frac{4e}{e-2}e^{-b_{max}}$. There exists an absolute constant C such that, if $\kappa b_{max}/n \leq C$, then there exists a multiple-δ L-sub-Gaussian estimator for $(\mathcal{P}_4^\kappa, n, \delta_{min}^{(4)})$.*

This result is most interesting in the regime where $n \to \infty$, $\kappa = \kappa_n$ possibly depends on n and $n/\kappa_n \to \infty$. In this case, we may take $b_{max} = o((n/\kappa_n)^{2/3})$ and obtain multiple-δ $(\sqrt{2} + o(1))$-sub-Gaussian estimators $(\mathcal{P}_{krt \leq \kappa}, n, \delta_{min}^{(4)})$ for $\delta_{min}^{(4)} \approx e^{-b_{max}}$. Catoni [5] obtained δ-dependent $\sqrt{2} + o(1)$-estimators for a smaller value $\delta_{min}^{(5)} \approx e^{-n/\kappa}$. In Remark 2, we show how one can obtain a similar range of δ with a multiple-δ estimator, albeit with worse constant L.

4. General methods. We collect here some ideas that recur in the remainder of the paper. Section 4.1 presents an analysis of median-of-means, as described in Section 2.3.2 above. Section 4.2 presents a "black-box method" of deriving multiple-δ estimators from confidence intervals; the latter being δ-dependent objects that are easier to construct.

We then present several lower bounds. In Section 4.3, we use scaled Bernoulli distributions to prove the impossibility of designing (weakly) sub-Gaussian estimators for families of distributions with unbounded variance. Section 4.4 uses the family of Laplace distributions to lower bound δ_{min} for δ-dependent estimators. Finally, Section 4.5 uses the Poisson family to derive lower bounds on δ_{min} for multiple-δ estimators. A combination of the methods in this section will allow us to derive the sharp range for $\ln(1/\delta_{min})$ for almost all families of distributions we consider.

4.1. *Median of means.* Given a positive integer b and a vector $x_1^b \in \mathbb{R}^b$, we let $q_{1/2}$ denote the median of the numbers $x_1, x_2, \ldots, x_b$, that is,

$$q_{1/2}(x_1^b) = x_i \qquad \text{where } \#\{k \in [b] : x_k \le x_i\} \ge \frac{b}{2} \text{ and } \#\{k \in [b] : x_k \ge x_i\} \ge \frac{b}{2}.$$

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

To build our estimator for a given $\delta \in [e^{1-n/2}, 1)$, we first choose

$$b = \lceil \ln(1/\delta) \rceil$$

and note that $b \le n/2$.

Now divide $[n]$ into b *blocks* (i.e., disjoint subsets) B_i, $1 \le i \le b$, each of size $|B_i| \ge k = \lfloor n/b \rfloor \ge 2$. Given $x_1^n \in \mathbb{R}^n$, we define

$$y_{n,\delta}(x_1^n) = (y_{n,\delta,i}(x_1^n))_{i=1}^b \in \mathbb{R}^b \quad \text{with coordinates } y_{n,\delta,i}(x_1^n) = \frac{1}{|B_i|} \sum_{j \in B_i} x_j$$

and define the median-of-means estimator by $\widehat{E}_{n,\delta}(x_1^n) = q_{1/2}(y_{n,\delta}(x_1^n))$.

The next result is a well-known performance bound for the median-of-means estimator; see, for example, Hsu [7].

THEOREM 4.1. *For any $n \ge 4$ and $L = 2\sqrt{2}e$ the median-of-means estimator by $\widehat{E}_{n,\delta}(x_1^n) = q_{1/2}(y_{n,\delta}(x_1^n))$ is a δ-dependent L-sub-Gaussian estimator for $(\mathcal{P}_2, n, e^{1-n/2})$.*

4.2. *The method of confidence intervals for multiple-δ estimators.* In this section, we detail how sub-Gaussian confidence intervals may be combined to produce multiple-δ estimators. This will be our main tool in defining all multiple-δ estimators whose existence is claimed in Theorems 3.2 and 3.3. First, we need a definition.

DEFINITION 3. Let n be a positive integer, $\delta \in (0, 1)$ and let $\mathcal{P}$ be a class of probability distributions over $\mathbb{R}$. A measurable closed interval $\widehat{I}_{n,\delta}(\cdot) = [\hat{a}_{n,\delta}(\cdot), \hat{b}_{n,\delta}(\cdot)]$ consists of a pair of measurable functions $\hat{a}_{n,\delta}, \hat{b}_{n,\delta} : \mathbb{R}^n \to \mathbb{R}$ with $\hat{a}_{n,\delta} \le \hat{b}_{n,\delta}$. We let $\hat{\ell}_{n,\delta} = \hat{b}_{n,\delta} - \hat{a}_{n,\delta}$ denote the length of the interval. We say $\{\widehat{I}_{n,\delta}\}_{\delta \in [\delta_{\min}, 1)}$ is a collection L-sub-Gaussian confidence intervals for $(n, \mathcal{P}, \delta_{\min})$ if for any $P \in \mathcal{P}$, if $X_1^n =_d P^{\otimes n}$, then for all $\delta \in [\delta_{\min}, 1)$,

$$\mathbb{P}\left(\mu_P \in \widehat{I}_{n,\delta}(X_1^n) \text{ and } \hat{\ell}_{n,\delta}(X_1^n) \le L \sigma_P \sqrt{\frac{1 + \ln(1/\delta)}{n}} \right) \ge 1 - \delta.$$

The next theorem shows how one can combine sub-Gaussian confidence intervals to obtain a multiple-δ sub-Gaussian mean estimator.

THEOREM 4.2. *Let n be a positive integer and let $\mathcal{P}$ be a class of probability distributions over $\mathbb{R}$. Assume that there exists a collection of L-sub-Gaussian confidence intervals for $(n, \mathcal{P}, \delta_{\min})$. Then there exists a multiple-δ estimator $\widehat{E}_n : \mathbb{R}^n \to \mathbb{R}$ that is L'-sub-Gaussian for $(n, \mathcal{P}, 2^{-m})$, where $L' = L\sqrt{1 + 2\ln 2}$ and $m = \lfloor \log_2(1/\delta_{\min}) \rfloor - 1 \geq \log_2(1/\delta_{\min}) - 2$ (in particular, $2^{-m} \leq 4\delta_{\min}$).*

PROOF. Our choice of m implies that, for each $k = 1, 2, 3, \ldots, m + 1$ there exists a measurable closed interval $\widehat{I}_k(\cdot) = [\hat{a}_k(\cdot), \hat{b}_k(\cdot)]$ with length $\hat{\ell}_k(\cdot)$, with the property that, if $P \in \mathcal{P}$ and $X_1^n =_d P^{\otimes n}$, the event

(9)
$$\mathcal{G}_k := \left\{ \mu_P \in \widehat{I}_k(X_1^n) \text{ and } \hat{\ell}_k(X_1^n) \leq L\sigma_P \sqrt{\frac{1 + k\ln 2}{n}} \right\}$$

has probability $\mathbb{P}(\mathcal{G}_k) \geq 1 - 2^{-k}$. To define our estimator, define, for $x_1^n \in \mathbb{R}^n$,

$$\hat{k}_n(x_1^n) = \min\left\{ k \in [m+1] : \bigcap_{j=k}^{m+1} \widehat{I}_j(x_1^n) \neq \varnothing \right\}.$$

One can easily check that

$$\bigcap_{j=\hat{k}_n(x_1^n)}^{m+1} \widehat{I}_j(x_1^n) \text{ is always a nonempty closed interval,}$$

so it makes sense to define the estimator $\widehat{E}_n(x_1^n)$ as its midpoint.

We claim that $\widehat{E}_n$ is the sub-Gaussian estimator we are looking for. To prove this, we let $2^{-m} \leq \delta \leq 1$ and choose the smallest $k \in \{1, 2, \ldots, m+1\}$ with $2^{1-k} \leq \delta$. Assume $X_1^n =_d P^{\otimes n}$ with $P \in \mathcal{P}$. Then:

1. $\mathbb{P}(\bigcap_{j=k}^{m+1} \mathcal{G}_j) \geq 1 - 2^{-k} - 2^{-k-1} \geq \cdots \geq 1 - 2^{1-k} \geq 1 - \delta$ by (9) and the choice of k.

2. When $\bigcap_{j=k}^{m+1} \mathcal{G}_j$ holds, $\mu_P \in \widehat{I}_j(X_1^n)$ for all $k \leq j \leq m + 1$, so $\mu_P \in \bigcap_{j=k}^{m+1} \widehat{I}_j(X_1^n)$. In particular, $\bigcap_{j=k}^{m+1} \widehat{I}_j(X_1^n) \neq \varnothing$ and $\hat{k}_n(X_1^n) \leq k$.

3. Now when $\hat{k}_n(X_1^n) \leq k$, $\widehat{E}_n(X_1^n) \in \bigcap_{j=k}^{m+1} \widehat{I}_j(X_1^n)$ as well, so both $\widehat{E}_n(X_1^n)$ and μ_P belong to $\widehat{I}_k(X_1^n)$. It follows that $|\widehat{E}_n(X_1^n) - \mu_P| \leq \hat{\ell}_k(X_1^n)$.

4. Finally, our choice of k implies $2^{1-k} \leq \delta \leq 2^{2-k}$, so, under $\bigcap_{j=k}^{m+1} \mathcal{G}_j$ we have

$$\hat{\ell}_k(X_1^n) \leq L\sigma_P \sqrt{\frac{1 + \ln(2^k)}{n}} \leq L\sigma_P \sqrt{\frac{1 + 2\ln 2 + \ln(1/\delta)}{n}} \leq L'\sigma_P \sqrt{\frac{1 + \ln(1/\delta)}{n}}$$

with $L' = L\sqrt{1 + 2\ln 2}$ as in the statement of the theorem.

Putting it all together, we conclude

$$\mathbb{P}\bigg(|\widehat{E}_n(X_1^n) - \mu_\mathrm{P}| \le L'\sigma_\mathrm{P}\sqrt{\frac{1 + \ln(1/\delta)}{n}}\bigg) \ge \mathbb{P}\bigg(\bigcap_{j=k}^{m+1}\mathcal{G}_j\bigg) \ge 1 - \delta,$$

and since this holds for all $\mathrm{P} \in \mathcal{P}$ and all $2^{-m} \le \delta \le 1/2$, the proof is complete. $\qquad\square$

4.3. *Scaled Bernoulli distributions and single-δ estimators.* In this subsection, we prove Theorem 3.1. In order to do so, we derive a simple minimax lower bound for single-δ estimators for the class $\mathcal{P}_{c,p} = \{P_+, P_-\}$ of distributions that contains two discrete distributions defined by

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

where $p \in [0, 1]$ and $c > 0$. Note that $\mu_{P_+} = pc$, $\mu_{P_-} = -pc$ and that for any $\alpha > 0$, the $(1 + \alpha)$th central moment of both distributions equals

$$(10) \qquad\qquad M = c^{1+\alpha} p(1 - p)(p^\alpha + (1 - p)^\alpha).$$

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

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

Note that $X_i \overset{\mathcal{L}}{\sim} P_+$ and $Y_i \overset{\mathcal{L}}{\sim} P_-$. Let $\delta \in (0, 1/2)$. If $\delta \ge 2e^{-n/4}$ and $p = (2/n)\log(2/\delta)$, then [using $1 - p \ge \exp(-p/(1 - p))$],

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

Let $\widehat{E}_{n,\delta}$ be any mean estimator, possibly depending on δ. Then

$$\begin{aligned}
\max\big(&\mathbb{P}\{|\widehat{E}_{n,\delta}(X_1^n) - \mu_{P_+}| > cp\}, \mathbb{P}\{|\widehat{E}_{n,\delta}(Y_1^n) - \mu_{P_-}| > cp\}\big) \\
&\ge \tfrac{1}{2}\mathbb{P}\{|\widehat{E}_{n,\delta}(X_1^n) - \mu_{P_+}| > cp \text{ or } |\widehat{E}_{n,\delta}(Y_1^n) - \mu_{P_-}| > cp\} \\
&\ge \tfrac{1}{2}\mathbb{P}\{\widehat{E}_{n,\delta}(X_1^n) = \widehat{E}_{n,\delta}(Y_1^n)\} \\
&\ge \tfrac{1}{2}\mathbb{P}\{X_1^n = Y_1^n\} \ge \delta.
\end{aligned}$$

From (10), we have that $cp \ge M^{1/(1+\alpha)}(p/2)^{\alpha/(1+\alpha)}$ and, therefore,

$$\max\bigg(\mathbb{P}\bigg\{|\widehat{E}_{n,\delta}(X_1^n) - \mu_{P_+}| > \bigg(\frac{M^{1/\alpha}}{n}\log\frac{2}{\delta}\bigg)^{\alpha/(1+\alpha)}\bigg\},$$

$$\mathbb{P}\bigg\{|\widehat{E}_{n,\delta}(Y_1^n) - \mu_{P_-}| > \bigg(\frac{M^{1/\alpha}}{n}\log\frac{2}{\delta}\bigg)^{\alpha/(1+\alpha)}\bigg\}\bigg) \ge \delta.$$

Theorem 3.1 simply follows by noting that $\mathcal{P}_{c,p} \subset \mathcal{P}_{1+\alpha}^M$.

4.4. *Laplace distributions and single-δ estimators.* This section focuses on the class of *all Laplace distributions with scale parameter equal to* 1. To define such a distribution, let $\lambda \in \mathbb{R}$ and let La_λ be the probability measure on $\mathbb{R}$ with density

$$\frac{d\,\mathsf{La}_\lambda}{dx}(x) = \frac{e^{-|x-\lambda|}}{2}.$$

Denote by $\mathcal{P}_{\mathsf{La}} = \{\mathsf{La}_\lambda : \lambda \in \mathbb{R}\}$ the class of all such distributions.

A simple calculation reveals that for all $\lambda \in \mathbb{R}$, the mean, variance and central third moment are $\mu_{\mathsf{La}_\lambda} = \lambda$, $\sigma^2_{\mathsf{La}_\lambda} = 2$ and $\mathsf{La}_\lambda\,|X - \lambda|^3 = 6 \leq (\eta \sigma_{\mathsf{La}_\lambda})^3$ with $\eta = 3^{1/3} 2^{1/6}$.

The next result proves that δ-dependent L-sub-Gaussian estimators are limited to exponentially small δ even over the one-dimensional family $\mathcal{P}_{\mathsf{La}}$.

THEOREM 4.3. *If $n \geq 3$ then, for any constant $L \geq \sqrt{2}$, there are no δ-dependent L-sub-Gaussian estimators for $(\mathcal{P}_{\mathsf{La}}, n, e^{1-9L^2 n})$.*

PROOF. We proceed by contradiction, assuming that there exist L-sub-Gaussian δ-dependent estimators $\widehat{E}_{n,\delta}$ for $(\mathcal{P}_{\mathsf{La}}, n, \delta)$ where $\delta = e^{1-9L^2 n}$ and arbitrarily large n. We set

$$\lambda = 2L\sqrt{2(1 + \ln(1/\delta))/n}$$

and consider $X_1^n =_d \mathsf{La}_0^{\otimes n}$ and $Y_1^n =_d \mathsf{La}_\lambda^{\otimes n}$. The triangle inequality applied to the exponents of $d\,\mathsf{La}_\lambda\,/dx$ and $d\,\mathsf{La}_0\,/dx$ shows that the densities of the two product measures satisfy, for all $x_1^n \in \mathbb{R}^n$,

$$\frac{d\,\mathsf{La}_0}{dx_1^n}(x_1^n) \geq e^{-\eta n}\frac{d\,\mathsf{La}_\lambda}{dx_1^n}(x_1^n)$$

and, therefore,

(11) $$\mathbb{P}\left(\widehat{E}_{n,\delta}(X_1^n) \geq \frac{\lambda}{2}\right) \geq e^{-\lambda n}\mathbb{P}\left(\widehat{E}_{n,\delta}(Y_1^n) \geq \frac{\lambda}{2}\right).$$

Using the definition of λ and the fact that $\mu_{\mathsf{La}_\lambda} = \lambda$ and $\sigma^2_{\mathsf{La}_\lambda} = 2$, we see that the right-hand side above is simply

$$e^{-\lambda n}\mathbb{P}\left(\widehat{E}_{n,\delta}(Y_1^n) \geq \mu_{\mathsf{La}_\lambda} - L\sigma_{\mathsf{La}_\lambda}\sqrt{\frac{1 + \ln(1/\delta)}{n}}\right) \geq e^{-\lambda n}(1 - \delta).$$

On the other hand, the left-hand side in (11) is

$$\mathbb{P}\left(\widehat{E}_{n,\delta}(X_1^n) \geq \mu_{\mathsf{La}_0} + L\sigma_{\mathsf{La}_0}\sqrt{\frac{1 + \ln(1/\delta)}{n}}\right) \leq \delta.$$

We deduce

$$e^{-\lambda n} \le \frac{\delta}{1-\delta} \le 2\delta.$$

If we use again the definition of λ, we see that

$$e^{-2L\sqrt{2n(1+\ln(1/\delta))}} \le 2\delta,$$

or

$$e^{-2\sqrt{18}L^2 n} \le 2e^{1-9L^2 n} \Rightarrow n \le \frac{1+\ln 2}{L^2(9-2\sqrt{18})}.$$

For $L \ge \sqrt{2}$, some simple estimates show that this leads to a contradiction when $n \ge 3$. □

4.5. *Poisson distributions and multiple-δ estimators.* We use the family of Poisson distributions for bounding the range of confidence values of multiple-δ estimators. Denote by Po_λ the Poisson distribution with parameter $\lambda > 0$. Given $0 < \lambda_1 \le \lambda_2 < \infty$, define

$$\mathcal{P}_{\mathsf{Po}}^{[\lambda_1, \lambda_2]} = \{\mathsf{Po}_\lambda : \lambda \in [\lambda_1, \lambda_2]\}.$$

THEOREM 4.4. *There exist positive constants c_0, s_0 and a function ϕ : $\mathbb{R}_+ \to \mathbb{R}_+$ such that the following holds. Assume $L \ge \sqrt{2}$ and $n > 0$ are given. Then there exists no multiple-δ L-sub-Gaussian estimator for $(\mathcal{P}_{\mathsf{Po}}^{[c/n, \phi(L)c/n]}, n, e^{1-s_0(L \ln L)^2 c})$.*

PROOF. We prove the following stronger result: there exist constants $c_0, s > 0$ such that, when $c \ge c_0$, $L \ge \sqrt{2}$ and $C = \lceil s(L^2 \ln L) \rceil$, there is no multiple-$\delta$ sub-Gaussian estimator for

$$(\star) = (\mathcal{P}_{\mathsf{Po}}^{[c/n, ((1+2C)c)/n]}, n, e^{1-C^2 c/L^2}).$$

The theorem then follows by taking $2C = \phi(L) - 1 = sL^2 \ln L$ and $s_0 = s^2$.

We proceed by contradiction. Assume

$$X_1^n =_d \mathsf{Po}_{c/n}^{\otimes n}, \qquad Y_1^n =_d \mathsf{Po}_{(1+2C)c/n}^{\otimes n}$$

and that there exists an L-sub-Gaussian estimator $\widehat{E}_n : \mathbb{R}^n \to \mathbb{R}$ for $(\star)$ above. We use the following well-known facts about Poisson distributions:

F0: $\mu_{\mathsf{Po}_{c/n}} = \sigma_{\mathsf{Po}_{c/n}}^2 = c/n$ and $\mu_{\mathsf{Po}_{(1+2C)c/n}} = \sigma_{\mathsf{Po}_{(1+2C)c/n}}^2 = (1+2C)c/n$.

F1: $S_X = X_1 + X_2 + \cdots + X_n =_d \mathsf{Po}_c$ and $S_Y = Y_1 + Y_2 + \cdots + Y_n =_d \mathsf{Po}_{(1+2C)c}$.

F2: Given any $k \in \mathbb{N}$, the distribution of X_1^n conditioned on $S_X = k$ is *the same* as the distribution of Y_1^n conditioned on $S_Y = k$.

F3: $\mathbb{P}(S_Y = (1 + 2C)c) \geq 1/4\sqrt{(1 + 2C)c}$ if $C > 0$ and $c \geq c_0$ for some c_0. [This follows from the fact that $\mathsf{Po}_m(\{m\}) = e^{-m}m^m/m!$ is asymptotic to $1/\sqrt{2\pi m}$ when $m \to \infty$, by Stirling's formula.]

F4: There exists a function h with $0 < h(C) \approx (1 + C)\ln(1 + C)$ such that, for all $c \geq c_0$, $\mathbb{P}(S_X = (1 + 2C)c) \geq e^{-h(C)c}$. This follows from another asymptotic estimate proven by Stirling's formula: as $c \to \infty$

$$\mathsf{Po}_c(\{(1 + 2C)c\}) = e^{-c}\frac{c^{(1+2C)c}}{[(1 + 2C)c]!} \sim \frac{e^{-[(1+2C)\ln(1+2C)-2C]c}}{\sqrt{2\pi(1 + 2C)c}}.$$

We apply the sub-Gaussian property for the triple $(\star)$ to $\delta = 1/4\sqrt{(1 + 2C)c}$. This is possible because, for $C = \lceil s(L^2 \ln L) \rceil$ with a large enough s, this value is $\approx 1/L\sqrt{s \ln Lc}$, which is much larger than the minimum confidence parameter e^{1-C^2c/L^2} allowed by $(\star)$ (at least if $c \geq c_0$ with a large enough c_0). Recalling F0, we obtain

$$\mathbb{P}\big(n\widehat{E}_n(Y_1^n) < (1 + 2C)c - L\sqrt{(1 + 2C)c\big(1 + \ln(8\sqrt{(1 + 2C)c})\big)}\big)$$

$$\leq \frac{1}{4\sqrt{(1 + C)c}}.$$

Therefore, by F3,

$$\mathbb{P}\big(n\widehat{E}_n(Y_1^n) < (1 + 2C)c - L\sqrt{(1 + 2C)c\big(1 + \ln(8\sqrt{(1 + 2C)c})\big)}\big|S_Y = (1 + C)c\big)$$

$$\leq 1/2.$$

Now F1 implies that the left-hand side is the same if we switch from Y to X. In particular, by looking at the complementary event we obtain

$$\mathbb{P}\big(n\widehat{E}_n(X_1^n) \geq (1 + 2C)c - L\sqrt{(1 + 2C)c\big(1 + \ln(8\sqrt{(1 + 2C)c})\big)}\big|$$

(12) $\qquad S_X = (1 + 2C)c)$

$$\geq 1/2.$$

Since we are taking $c \geq c_0$ and $C \geq sL^2 \ln L$, a calculation reveals

$$L\sqrt{(1 + 2C)c\big(1 + \ln(8\sqrt{(1 + 2C)c})\big)} = O\left(\sqrt{\frac{C^2c(\ln C + \ln c)}{\ln C}}\right) = O(C\sqrt{c \ln c}).$$

Therefore, by taking a large enough c_0 we can ensure that

$$L\sqrt{(1 + 2C)c\big(1 + \ln(8\sqrt{(1 + 2C)c})\big)} \leq Cc.$$

So (12) gives

$$\mathbb{P}\big(n\widehat{E}_n(X_1^n) \geq (1 + C)c|S_X = (1 + 2C)c\big) \geq 1/2.$$

We may combine this with F4 to deduce

$$(13) \qquad \mathbb{P}\big(n\widehat{E}_n(X_1^n) \geq (1+C)c\big) \geq \frac{e^{-h(C)c}}{2}.$$

We now use F0 to rewrite the previous probability as

$$\mathbb{P}\big(n\widehat{E}_n(X_1^n) \geq (1+C)c\big) = \mathbb{P}\left(\widehat{E}_n(X_1^n) - \mu_P \geq L\sigma_P \frac{\sqrt{1 + \ln(1/\delta_0)}}{\sqrt{n}}\right),$$

where

$$\delta_0 = e^{1 - C^2 c/L^2}.$$

Since we assumed $\widehat{E}_n$ is L-sub-Gaussian for the triple $(\star)$, we obtain

$$\frac{e^{-h(C)c}}{2} \leq \mathbb{P}\big(n\widehat{E}(X_1^n) \geq (1+C)c\big) \leq e^{1 - C^2 c/(4L^2)}.$$

Comparing the left- and right-hand sides, and recalling $c \geq c_0$, we obtain $h(C) \geq C^2/4L^2 - 1 - (\ln 2/c_0)$. This is a contradiction if $C \gg L^2 \ln L$ because $h(C)$ grows like $C \ln C$ (cf. F4). This contradiction shows that there does not exist a L-sub-Gaussian estimator for $(\star)$, as desired. $\square$

5. Degrees of knowledge about the variance. In this section, we present the proof of Theorem 3.2. This is mostly a matter of combining the main results in the previous section. Recall that we consider the class

$$\mathcal{P}_2^{[\sigma_1^2, \sigma_2^2]} = \{P \in \mathcal{P}_2 : \sigma_1^2 \leq \sigma_P^2 \leq \sigma_2^2\}.$$

and that $R = \sigma_2/\sigma_1$. The three parts of the theorem are proven separately.

Part 1. (Existence of a multiple-δ estimator with constant depending on R.) Theorem 4.1 ensures that, irrespective of σ_1 or σ_2, for all $\delta \in (e^{1-n/2}, 1)$ there exists a δ-dependent estimator $\widehat{E}_{n,\delta} : \mathbb{R}^n \to \mathbb{R}$ with

$$(14) \qquad \mathbb{P}\left(|\widehat{E}_{n,\delta}(X_1^n) - \mu_P| > 2\sqrt{2}e\sigma_P\sqrt{\frac{1 + \ln(1/\delta)}{n}}\right) \leq \delta$$

whenever $X_1^n = P^{\otimes n}$ for some $P \in \mathcal{P}_2$. We define a confidence interval for each δ via

$$\widehat{I}_{n,\delta}(x_1^n) = \left[\widehat{E}_{n,k}(x_1^n) - 2\sqrt{2}e\sigma_2\sqrt{\frac{1 + \ln(1/\delta)}{n}},\right.$$

$$\left.\widehat{E}_{n,k}(x_1^n) + 2\sqrt{2}e\sigma_2\sqrt{\frac{1 + \ln(1/\delta)}{n}}\right].$$

Clearly, (14) and the fact that $\sigma_2 \leq R\sigma_P$ for all $\mathcal{P}_2^{[\sigma_1^2,\sigma_2^2]}$ imply that $\{\widehat{I}_{n,\delta}\}_{\delta \in [e^{1-n/2},1)}$ is a $4\sqrt{2}eR$-sub-Gaussian confidence interval for $(\mathcal{P}_2^{[\sigma_1^2,\sigma_2^2]}, n, e^{1-n/2})$. Applying Theorem 4.2 gives the desired result.

Part 2. [Nonexistence of multiple-δ estimators when $R > \phi^{(2)}(L)$.] We use Theorem 4.4. By rescaling, we may assume $\sigma_1^2 = c_0/n$, where c_0 is the constant appearing in Theorem 4.4. We also set $\phi^{(2)}(L) := \sqrt{\phi(L)}$ for $\phi(L)$ as in Theorem 4.4. The assumption on R ensures that $\mathcal{P}_{\mathsf{Po}}^{[c_0/n,\phi(L)c_0/n]} \subset \mathcal{P}_2^{[\sigma_1^2,\sigma_2^2]}$, so there cannot be a L-sub-Gaussian estimator when $\delta_{\min}^{(2)}(L) = e^{1-s(L \ln L)^2 c_0}$.

Part 3. (Nonexistence of δ-dependent estimators when $\delta_{\min} = e^{1-9L^2 n}$.) By rescaling, we may assume $\sigma_1^2 = 2$. Then the class $\mathcal{P}_{\mathsf{La}}$ in Theorem 4.3 is contained in $\mathcal{P}_2^{[\sigma_1^2,\sigma_2^2]}$, and the theorem implies the desired result directly.

6. The regularity condition, symmetry and higher moments. In this section, we prove the results described in Section 3.3.

6.1. *An estimator under k-regularity.* We start with Theorem 3.3, the general positive result on k-regular classes.

PROOF OF THEOREM 3.3. By Theorem 4.2, it suffices to build a $4\sqrt{2(1 + 62\ln(3))}e^{5/2}$-sub-Gaussian confidence interval for $(\mathcal{P}_{2,k\text{-reg}}, n, e^{3-n/(124)k})$.

To build these intervals, we use an idea related to the proof of Theorem 4.1. Just like in the case of the median-of-means estimator, we divide the data into blocks, but instead of taking the median of the means, we look at the $1/4$ and $3/4$-quantiles to build an interval.

To make this precise, given $\alpha \in (0, 1)$, we define the α-quantile $q_\alpha(y_1^b)$ of a vector $y_1^b \in \mathbb{R}^b$ as the smallest index $i \in [b]$ with

$$\#\{j \in [b] : y_j \leq y_i\} \geq \alpha b \quad \text{and} \quad \#\{\ell \in [b] : y_\ell \geq y_i\} \geq (1-\alpha)b.$$

The next result is proven subsequently.

LEMMA 6.1. *Let* $Y_1^b = (Y_1, \ldots, Y_b) \in \mathbb{R}^b$ *be a vector of independent random variables with the same mean* μ *and variances bounded by* σ^2. *Assume further that* $\mathbb{P}(Y_i \leq \mu) \geq 1/3$ *and* $\mathbb{P}(Y_i \geq \mu) \geq 1/3$ *for each* $i \in [b]$. *Then*

$$\mathbb{P}(\mu \in [q_{1/4}(Y_1^b), q_{3/4}(Y_1^b)] \text{ and } q_{3/4}(Y_1^b) - q_{1/4}(Y_1^b) \leq 2L_0\sigma) \geq 1 - 3e^{-db},$$

where d *is the numerical constant*

$$d = \tfrac{1}{4}\ln(\tfrac{3}{4}) + \tfrac{3}{4}\ln(\tfrac{9}{8}) \approx 0.0164 > \tfrac{1}{62}$$

and $L_0 = 2e^{2d+1/2} \leq 2e^{5/2}$.

Now fix $\delta \in [e^{3-n/(124k)}, 1)$. We define a confidence interval $\widehat{I}_{n,\delta}(\cdot)$ as follows. First, set $b = \lceil 62 \ln(3/\delta) \rceil$ and note that

$$(15) \qquad b \leq 62 \ln(3/e^{3-n/(124k)}) + 1 \leq \frac{n}{2k} \leq n/2.$$

Partition

$$[n] = B_1 \cup B_2 \cup \cdots \cup B_b$$

into disjoint blocks of sizes $|B_i| \geq \lfloor n/b \rfloor$. For each $i \in [b]$ and $x_1^n \in \mathbb{R}^n$, we define

$$y_1^b(x_1^n) = (y_1(x_1^n), \ldots, y_b(x_1^n)) \qquad \text{where } y_i(x_1^n) = \frac{1}{\#B_i} \sum_{j \in B_i} x_j$$

and set, for $x_1^n \in \mathbb{R}^n$,

$$\widehat{I}_{n,\delta}(x_1^n) = [q_{1/4}(y_1^b(x_1^n)), q_{3/4}(y_1^b(x_1^n))].$$

CLAIM 1. $\{\widehat{I}_{n,\delta}(\cdot)\}_{\delta \in [e^{3-n/(124k)}, 1)}$ *is a* $4\sqrt{2(1 + 62 \ln(3))}e^{5/2}$*-sub-Gaussian collection of confidence intervals for* $(\mathcal{P}_{2,k\text{-reg}}, n, e^{3-n/(124k)})$.

To see this, we take a distribution P in this family and assume $X_1^n =_d P^{\otimes n}$. Set $s = \lfloor n/b \rfloor$. Because the blocks B_i are disjoint and have at least s elements each, the random variables

$$Y_i = y_i(X_1^n) = \widehat{P}_{B_i} X,$$

all have mean μ_P and variance $\leq \sigma_P^2/s$. Moreover, using (15),

$$s = \left\lfloor \frac{n}{b} \right\rfloor \geq \frac{n}{b} - 1 \geq 2k - 1 \geq k,$$

so the k-regularity property implies that for all $i \in [b]$,

$$\mathbb{P}(Y_i \leq \mu) \geq \tfrac{1}{3}, \quad \mathbb{P}(Y_i \geq \mu) \geq \tfrac{1}{3}.$$

Lemma 6.1 implies

$$(16) \quad \mathbb{P}\left(\mu_P \in \widehat{I}_{n,\delta}(X_1^n) \text{ and length of } \widehat{I}_{n,\delta}(X_1^n) \leq 2L_0 \frac{\sigma}{\sqrt{s}} \right) \geq 1 - 3e^{-db} \geq 1 - \delta$$

by the choice of b and the fact that $d \geq 1/62$. To finish, we use (15) and the definition of b to obtain

$$\frac{1}{s} = \frac{1}{\lfloor n/b \rfloor} \leq \frac{1}{(n/b) - 1} \leq \frac{2b}{n} \leq \frac{2(\lceil 62 \ln(3) + 62 \ln(1/\delta) \rceil)}{n}$$

$$\leq 2(1 + 62 \ln 3) \frac{1 + \ln(1/\delta)}{n}.$$

Plugging this into (16) and recalling $L_0 \leq 2e^{5/2}$ implies the desired result. $\quad \square$

PROOF OF LEMMA 6.1. Define $J = [\mu - L_0\sigma, \mu + L_0\sigma]$. Assume the following three properties hold:

1. $q_{1/4}(Y_1^b) \leq \mu$.
2. $q_{3/4}(Y_1^b) \geq \mu$.
3. The number of indices $i \in [b]$ with $Y_i \in J$ is at least $3b/4$.

Then clearly $\mu \in [q_{1/4}(Y_1^b), q_{3/4}(Y_1^b)]$. Moreover, item 3 implies that $q_{1/4}(Y_1^b)$, $q_{3/4}(Y_1^b) \in J$, so that

$$q_{3/4}(Y_1^b) - q_{1/4}(Y_1^b) \leq \text{(length of } J) = 2L_0\sigma.$$

It follows that

$$(17) \quad \begin{aligned} &\mathbb{P}(\mu \notin [q_{1/4}(Y_1^b), q_{3/4}(Y_1^b)] \text{ or } q_{3/4}(Y_1^b) - q_{1/4}(Y_1^b) > 2L_0\sigma) \\ &\leq \mathbb{P}(q_{1/4}(Y_1^b) > \mu) + \mathbb{P}(q_{3/4}(Y_1^b) < \mu) + \mathbb{P}(\#\{i \in [b] : Y_i \notin J\} > b/4). \end{aligned}$$

We bound the three terms by e^{-bd} separately. By assumption, $\mathbb{P}(Y_i \leq \mu) \geq 1/3$ for each $i \in [b]$. Since there events are also independent, we have that $\sum_{i=1}^{b} \mathbb{1}\{Y_i \leq \mu\}$ stochastically dominates a binomial random variable $\text{Bin}(b, 1/3)$. Thus,

$$\mathbb{P}(q_{1/4}(Y_1^b) > \mu) = \mathbb{P}\left(\sum_{i=1}^{b} \mathbb{1}\{Y_i \leq \mu\} < b/4\right) \leq \mathbb{P}(\text{Bin}(b, 1/3) < b/4) \leq e^{-db}$$

by the relative entropy version of the Chernoff bound and the fact that d is the relative entropy between two Bernoulli distributions with parameters $1/4$ and $1/3$. A similar reasoning shows that $\mathbb{P}(q_{3/4}(Y_1^b) > \mu) \leq e^{-db}$ as well.

It remains to bound $\mathbb{P}(\#\{i \in [b] : Y_i \notin J\} > b/4)$. To this end note that for all $i \in [b]$,

$$(18) \quad \mathbb{P}(Y_i \notin J) = \mathbb{P}(|Y_i - \mu| \geq L_0\sigma) \leq \frac{1}{L_0^2},$$

and these events are independent. It follows that

$$\mathbb{P}(\#\{i \in [b] : Y_i \notin J\} > b/4) \leq \mathbb{P}\left(\bigcup_{A \subset [b], |A| = \lceil b/4 \rceil} \bigcap_{i \in A} \{Y_i \notin J\}\right),$$

$$\text{(union bound)} \leq \binom{b}{\lceil b/4 \rceil} \max_{A \subset [b], |A| = \lceil b/4 \rceil} \mathbb{P}\left(\bigcap_{i \in A} \{Y_i \notin J\}\right),$$

$$\text{[independence of } Y_i + (18)] \leq \binom{b}{\lceil b/4 \rceil} \left(\frac{1}{L_0^2}\right)^{-\lceil b/4 \rceil},$$

$$(\binom{b}{k} \leq (eb/k)^k \text{ for all } 1 \leq k \leq b) \leq \left(\frac{eb}{L_0^2 \lceil b/4 \rceil}\right)^{\lceil b/4 \rceil},$$

$$(b \leq 4\lceil b/4 \rceil \text{ and } L_0^2 = 4e^{4d+1}) \leq e^{-4d\lceil b/4 \rceil} \leq e^{-bd}. \qquad \square$$

6.2. *Symmetric distributions.* To prove Theorem 3.4, notice that the existence of the multiple-δ sub-Gaussian estimator follows from Theorem 3.3. The second part is a simple consequence of Theorem 4.3 and the fact that Laplace distributions are symmetric around their means.

6.3. *Higher moments.* In this section, we first prove that $\mathcal{P}_{\alpha,\eta} \subset \mathcal{P}_{2,k\text{-reg}}$ for large enough k, and then prove Theorem 3.5. We recall the definition of $\min(p_+(\mathrm{P}, j))$ and $p_-(\mathrm{P}, j)$ from Definition 2.

LEMMA 6.2. *For all* $\alpha \in (2, 3]$, *there exists* $C = C_\alpha$ *such that, if* $j \geq (C_\alpha \eta)^{2\alpha/(\alpha-2)}$, *then* $\min(p_+(\mathrm{P}, j), p_-(\mathrm{P}, j)) \geq 1/3$.

PROOF. We only consider $p_+(\mathrm{P}, j)$, as p_- is similar. Let N be standard normal and $X_1^j =_d \mathrm{P}^{\otimes j}$ for some $\mathrm{P} \in \mathcal{P}_{\alpha,\eta}$ with positive variance. Define

$$M_j := \frac{1}{\sigma_\mathrm{P}\sqrt{j}} \sum_{i=1}^{j} (X_i - \mu_\mathrm{P}).$$

Let $h : \mathbb{R} \to \mathbb{R}$ be Lipschitz function with $0 \leq h(x) \leq \mathbf{1}_{[0,+\infty)}$ and $\mathbb{E}[h(N)] \geq 5/12 > 1/3$. We have $p_+(\mathrm{P}, j) = \mathbb{P}(M_j \geq 0) \geq \mathbb{E}[h(M_j)] \geq 5/12 - |\mathbb{E}[h(N)] - \mathbb{E}[h(M_j)]|$. Theorem 3.2 in [6] shows that there exists a universal $c > 0$ (depending on h) with

$$|\mathbb{E}[h(N)] - \mathbb{E}[h(M_j)]| \leq cj\mathbb{E}\left[\left(\frac{X_1}{\sigma_\mathrm{P}\sqrt{j}}\right)^2 \vee \left(\frac{X_1}{\sigma_\mathrm{P}\sqrt{j}}\right)^3\right] \leq \frac{\mathbb{E}[|X_1|^\alpha]}{j^{\alpha/2}} \leq c\frac{\eta^\alpha}{j^{(\alpha-2)/2}}.$$

Therefore, $\mathbb{E}[h(M_j)] \geq 1/3$ for all $j \geq (C_\alpha \eta)^{2\alpha/(\alpha-2)}$. $\square$

PROOF OF THEOREM 3.5. The positive result follows directly from Theorem 3.3 plus Lemma 6.2, which guarantees $p_\pm(\mathrm{P}, j) \geq 1/3$ for $j \geq k_\alpha$. For the second part, we first assume $\eta > \eta_0$ for a sufficiently large constant η_0. We use the Poisson family of distributions from Section 4.5. For $\lambda = o(1)$ and $\alpha \in (2, 3]$, we have that

$$\mathrm{Po}_\lambda |X - \lambda|^\alpha = (1 + o(1))\lambda = (1 + o(1))\sigma_{\mathrm{Po}_\lambda}^\alpha \lambda^{-(\alpha-2)/2}.$$

If we compare this to Example 3.2, we see that $\mathrm{Po}_\lambda \in \mathcal{P}_{\alpha,\eta}$ if $\lambda \geq h/\eta^{2\alpha/(\alpha-2)}$ for some constant $h = h_\alpha > 0$ (recall we are assuming that $\eta \geq \eta_0$ is at least a large constant). Now take $c > 0$ such that $c/n = h/\eta^{2\alpha/(\alpha-2)}$. If $c > c_0$ for the constant c_0 in the statement of Theorem 4.4, we can apply the theorem to deduce that there is no multiple-δ estimator for $(\mathcal{P}_{\mathrm{Po}}^{[c/n,\phi(L)c/n]}, n, e^{-c})$. Noting that c is of the order n/k_α completes the proof in this case.

Now assume $\eta \le \eta_0$. In this case, we use the Laplace distributions in Section 4.4. Since $2 < \alpha \le 3$, we may apply the fact that the central third moment of a Laplace distribution satisfies $\mathsf{La}_\lambda |X - \lambda|^3 = 6 \le (3^{1/3}2^{1/6}\sigma_{\mathsf{La}_\lambda})^3$ to obtain

$$\mathsf{La}_\lambda |X - \lambda|^\alpha \le (\mathsf{La}_\lambda |X - \lambda|^3)^{\alpha/3} \le (3^{1/3}2^{1/6}\sigma_{\mathsf{La}_\lambda})^\alpha.$$

Our assumption on η implies that $\mathcal{P}_{\mathsf{La}} \subset \mathcal{P}_{\alpha,\eta}$. Thus, Theorem 4.3 implies that there is no δ-dependent or multiple-δ sub-Gaussian estimator for $(\mathcal{P}_{\alpha,\eta}, n, e^{1-9L^2n})$. This is the desired result since k_α is bounded when $\eta \le \eta_0$.

Finally, the third part of the theorem follows from the same reasoning as in the previous paragraph. $\square$

7. Bounded kurtosis and nearly optimal constants. In this section, we prove Theorem 3.6. Throughout the proof, we assume $X =_d \mathsf{P}$ and $X_1^n =_d \mathsf{P}^{\otimes n}$ for some $\mathsf{P} \in \mathcal{P}_{\mathrm{krt} \le \kappa}$, and let $b_{\max}, C, \xi$ be as in Section 3.4. Our proof is divided into four steps:

1. *Preliminary estimates for mean and variance.* Median-of-means gives preliminary estimates for μ_P and σ_P^2. With high probability, they are not too far off.

2. *Truncation at the ideal point.* We introduce a two-parameter family of truncation-based estimators for μ_P, and analyze their behavior under knowledge of μ_P and σ_P.

3. *Truncated estimators are insensitive.* We use a chaining argument to show that this two-parameter family is insensitive to the choice of parameters.

4. *Wrap up.* The insensitivity property means that the preliminary estimates from step 1 are good enough to "make everything work".

Remark 2 at the end of the section shows how one could obtain a broader range of $\delta_{\min}$ with a worse constant L.

Step 1. (Preliminary estimates via median of means.) First, we need the following elementary fact, also used in the proof of Theorem 4.1. The proof is omitted.

LEMMA 7.1. *Let $Y_1^b = (Y_1, \ldots, Y_b) \in \mathbb{R}^b$ be independent random variables with the same mean μ and variances bounded by σ^2. Assume $L_0 > 1$ is given and $M_b = q_{1/2}(Y_1^b)$. Then $\mathbb{P}(|M_b - \mu| > 2L_0\sigma) \le L_0^{-b}$.*

Denote by $\widehat{\mu}_{b_{\max}} = \widehat{\mu}_{b_{\max}}(X_1^n)$ the estimator given by Theorem 4.1 with $\delta = e^{-b_{\max}}$, which is possible if $C \ge 6/(1 - \log 2)$. The next lemma provides an estimator of the variance.

LEMMA 7.2. *Let $B_1, \ldots, B_{b_{\max}}$ denote a partition of $[n]$ into blocks of size $|B_i| \ge k = \lfloor n/b_{\max} \rfloor \ge 2$. For each block B_i with $i \in [b_{\max}]$, define*

$$\widehat{\sigma}_i^2 = \frac{1}{|B_i|(|B_i| - 1)} \sum_{j \ne k \in B_i} (X_j - X_k)^2 \quad and \quad \widehat{v}_{b_{\max}}^2 = q_{1/2}(\widehat{\sigma}_1, \ldots, \widehat{\sigma}_{b_{\max}}).$$

Then

$$\mathbb{P}\left(|\widehat{v}_{b_{\max}}^2 - \sigma_{\mathrm{P}}^2| \le 2e\sqrt{6(\kappa+3)}\sigma_{\mathrm{P}}^2\sqrt{\frac{b_{\max}}{n}}\right) \ge 1 - e^{-b_{\max}}.$$

In particular, if

$$\frac{96e(\kappa+3)b_{\max}}{n} \le 1,$$

then

$$\mathbb{P}\left(|\widehat{\mu}_{b_{\max}} - \mu_{\mathrm{P}}| \le 2\sqrt{2}e\widehat{v}_{b_{\max}}\sqrt{\frac{b_{\max}}{n}} \text{ and } \widehat{v}_{b_{\max}}^2 \le \frac{3}{2}\sigma_{\mathrm{P}}^2\right) \ge 1 - 2e^{-b_{\max}}.$$

PROOF. Compute

$$\mathbb{E}[\widehat{\sigma}_i^4] = \frac{1}{|B_i|^2(|B_i|-1)^2}\sum_{(j,k)\in B_i^{(2)}} \mathbb{E}[(X_j - X_k)^4]$$

$$+ \frac{6}{|B_i|^2(|B_i|-1)^2}\sum_{(j,k,l)\in B_i^{(3)}} \mathbb{E}[(X_j - X_k)^2(X_j - X_l)^2]$$

$$+ \frac{1}{|B_i|^2(|B_i|-1)^2}\sum_{(j,k,l,m)\in B_i^{(4)}} \mathbb{E}[(X_j - X_k)^2(X_l - X_m)^2].$$

Expanding all the squares, using independence and noticing that $\mathbb{E}[X_j - \mu_{\mathrm{P}}] = 0$, we get

$$\mathbb{E}[(X_j - X_k)^4] = 2(\kappa_{\mathrm{P}} + 3)\sigma_{\mathrm{P}}^4,$$
$$\mathbb{E}[(X_j - X_k)^2(X_j - X_l)^2] = (\kappa_{\mathrm{P}} + 3)\sigma_{\mathrm{P}}^4,$$
$$\mathbb{E}[(X_j - X_k)^2(X_l - X_m)^2] = 4\sigma_{\mathrm{P}}^4.$$

Therefore,

$$\mathbb{E}[\widehat{\sigma}_i^4] \le \left(\frac{3(\kappa_{\mathrm{P}}+3)}{|B_i|} + 1\right)\sigma_{\mathrm{P}}^4 \le \mathbb{E}[\widehat{\sigma}_i^2]^2 + 6(\kappa+3)\sigma_{\mathrm{P}}^4\frac{b_{\max}}{n}.$$

Lemma 7.1 with $L_0 = e$ gives then

$$\mathbb{P}\left(|\widehat{v}_{b_{\max}}^2 - \sigma_{\mathrm{P}}^2| > 2e\sqrt{6(\kappa+3)}\sigma_{\mathrm{P}}^2\sqrt{\frac{b_{\max}}{n}}\right) \le e^{-b_{\max}}.$$

In particular, we get

$$\mathbb{P}\left(\frac{1}{2}\sigma_{\mathrm{P}}^2 \le \widehat{v}_{b_{\max}}^2 \le \frac{3}{2}\sigma_{\mathrm{P}}^2\right) \ge 1 - e^{-b_{\max}}.$$

The theorem follows by the definition of $\widehat{\mu}_{b_{\max}}$ and an application of Theorem 4.1. $\square$

Step 2. (Two-parameter family of estimators at the ideal point.) Given μ and R define, for all $x \in \mathbb{R}$,

$$\Psi_{\mu, R}(x) = \mu + \left(\frac{R}{|x - \mu|} \wedge 1 \right)(x - \mu).$$

LEMMA 7.3. *Assume* $b_{\max} \geq t$, $R = \sigma_{\mathrm{P}}\sqrt{n/b_{\max}}$ *and* $\mu = \mu_{\mathrm{P}}$. *Then, with probability at least* $1 - 2e^{-t}$,

$$|\widehat{\mathbb{P}}_n \Psi_{\mu, R} - \mu_{\mathrm{P}}| \leq 2\sqrt{2}\kappa_{\mathrm{P}}\sigma_{\mathrm{P}}\left(\frac{b_{\max}}{n} \right)^{3/2} + \sqrt{\frac{2t}{n}}\sigma_{\mathrm{P}}\left(1 + \frac{1}{3\sqrt{2}}\sqrt{\frac{\kappa_{\mathrm{P}}t}{n}} + \frac{5}{48}\frac{\kappa_{\mathrm{P}}t}{n} \right).$$

PROOF. The proof is a consequence of Bennett's inequality. It suffices to estimate the moments of $\Psi_{\mu, R}(X) - \mu_{\mathrm{P}}$. For the first moment,

$$
\begin{aligned}
\big|\mathbb{E}[\Psi_{\mu, R}(X) - \mu_{\mathrm{P}}]\big| &= \left|\mathbb{E}\left[\left(1 \wedge \frac{R}{|X - \mu_{\mathrm{P}}|} - 1 \right)(X - \mu_{\mathrm{P}}) \right]\right| \\
&\leq \mathbb{E}\left[\left(1 - \frac{R}{|X - \mu_{\mathrm{P}}|} \right)_{+} |X - \mu_{\mathrm{P}}| \right] \\
&\leq \mathbb{E}\big[|X - \mu_{\mathrm{P}}|\mathbb{1}\{|X - \mu_{\mathrm{P}}| > R\} \big] \\
&\leq \mathbb{E}\big[|X - \mu_{\mathrm{P}}|^4 \big]^{1/4}\mathbb{P}(|X - \mu_{\mathrm{P}}| > R)^{3/4} \leq \frac{\kappa_{\mathrm{P}}\sigma_{\mathrm{P}}^4}{R^3},
\end{aligned}
$$

where we used Hölder's inequality. On the other hand,

$$\mathbb{E}\big[(\Psi_{\mu, R}(X) - \mu_{\mathrm{P}})^2 \big] \leq \sigma_{\mathrm{P}}^2.$$

By the Cauchy–Schwarz inequality, and using the bounded kurtosis assumption,

$$\mathbb{E}\big[|\Psi_{\mu, R}(X) - \mu_{\mathrm{P}}|^3 \big] = \mathbb{E}\left[\left| 1 \wedge \frac{R}{|X - \mu_{\mathrm{P}}|} \right|^3 |X - \mu_{\mathrm{P}}|^3 \right] \leq \mathbb{E}\big[|X - \mu_{\mathrm{P}}|^3 \big] \leq \sqrt{\kappa_{\mathrm{P}}}\sigma_{\mathrm{P}}^3.$$

Finally, for any $p \geq 4$, since $|\Psi_{\mu, R}(X) - \mu_{\mathrm{P}}| \leq R$,

$$\mathbb{E}\big[|\Psi_{\mu, R}(X) - \mu_{\mathrm{P}}|^p \big] \leq R^{p-4}\mathbb{E}\big[|\Psi_{\mu, R}(X) - \mu_{\mathrm{P}}|^4 \big] = R^{p-4}\kappa_{\mathrm{P}}\sigma_{\mathrm{P}}^4.$$

For $s = \sqrt{2nt}/\sigma_{\mathrm{P}}$, we have $|sR|/n \leq 1$ and, therefore,

$$
\begin{aligned}
&\mathbb{E}\big[e^{(s/n)(\Psi_{\mu, R}(X) - \mu_{\mathrm{P}})} \big] \\
&\quad \leq 1 + \frac{s}{n}\frac{\kappa_{\mathrm{P}}\sigma_{\mathrm{P}}^4}{R^3} + \frac{s^2}{2n^2}\sigma_{\mathrm{P}}^2 + \frac{s^3}{6n^3}\sqrt{\kappa_{\mathrm{P}}}\sigma_{\mathrm{P}}^3 + \frac{s^4}{24n^4}\kappa_{\mathrm{P}}\sigma_{\mathrm{P}}^4\left(1 + \sum_{p \geq 5}\frac{4!}{p!} \right) \\
&\quad \leq \exp\left(2\sqrt{2}\frac{s}{n}\kappa_{\mathrm{P}}\sigma_{\mathrm{P}}\left(\frac{b_{\max}}{n} \right)^{3/2} + \frac{s^2}{2n^2}\sigma_{\mathrm{P}}^2 + \frac{s^3}{6n^3}\sqrt{\kappa_{\mathrm{P}}}\sigma_{\mathrm{P}}^3 + \frac{5s^4}{96n^4}\kappa_{\mathrm{P}}\sigma_{\mathrm{P}}^4 \right).
\end{aligned}
$$

By Chernoff's bound,

$$\mathbb{P}\left(\widehat{P}_n \Psi_{\mu,R} - \mu_P > 2\sqrt{2}\kappa_P \sigma_P \left(\frac{b_{\max}}{n}\right)^{3/2} + \frac{s}{n}\sigma_P^2 + \frac{s^2}{6n^2}\sqrt{\kappa_P}\sigma_P^3 + \frac{5s^3}{96n^3}\kappa_P \sigma_P^4\right)$$

$$\leq e^{-(s^2 \sigma_P^2)/(2n)}$$

or, equivalently,

$$\mathbb{P}\left(\widehat{P}_n \Psi_{\mu,R} - \mu_P > 2\sqrt{2}\kappa_P \sigma_P \left(\frac{b_{\max}}{n}\right)^{3/2} + \sqrt{\frac{2t}{n}}\sigma_P\left(1 + \frac{1}{3\sqrt{2}}\sqrt{\frac{\kappa_P t}{n}} + \frac{5}{48}\frac{\kappa_P t}{n}\right)\right)$$

$$\leq e^{-t}.$$

Repeat the same computations with $s = -\sqrt{2nt}/\sigma_P$ to prove the lower bound. $\quad\square$

Step 3. (Insensitivity of the estimators.) Given $\varepsilon_\mu, \varepsilon_R \in (0, 1/2)$, define

$$\mathcal{R} = \{(\mu, R) : |\mu - \mu_P| \leq \varepsilon_\mu \sigma_P, |R - \sigma_P\sqrt{n/(2b_{\max})}| \leq \varepsilon_R \sigma_P\},$$

$$\Delta_{\mu,R} = \widehat{P}_n(\Psi_{\mu,R} - \Psi_{\mu_P, \sigma_P\sqrt{n/(2b_{\max})}}).$$

LEMMA 7.4. *Assume* $\sqrt{n/(2b_{\max})} \geq 2(\varepsilon_\mu + \varepsilon_R)$ *then for any* $t > 0$, *with probability at least* $1 - e^{-t}$, *for all* $(\mu, R) \in \mathcal{R}$,

$$|\Delta_{\mu,R}| \leq (\varepsilon_\mu + \varepsilon_R)\sigma_P\left(\frac{56b_{\max}}{n} + \frac{4\sqrt{b_{\max}t}}{n} + \frac{2t}{3n}\right).$$

PROOF. Start with the trivial bound

$$|\Psi_{\mu,R}(x) - \Psi_{\mu',R'}(x)| \leq |\mu - \mu'| + |R - R'|$$

that holds for all $(\mu, R), (\mu', R') \in \mathcal{R}$ and $x \in \mathbb{R}$. Moreover, assume that $|x - \mu_P| \leq \sigma_P\sqrt{n/(8b_{\max})}$. Then

$$|x - \mu| \leq |x - \mu_P| + \varepsilon_\mu \sigma_P \leq \sigma_P\left(2\sqrt{n/(8b_{\max})} - \varepsilon_R\right) \leq R.$$

Hence, for any $x \in \mathbb{R}$ such that $|x - \mu_P| \leq \sigma_P\sqrt{n/(8b_{\max})}$ and for all $(\mu, R) \in \mathcal{R}$,

$$\Psi_{\mu,R}(x) = x.$$

Therefore, for any (μ, R) and (μ', R') in $\mathcal{R}$ and for any $x \in \mathbb{R}$,

$$|\Psi_{\mu,R}(x) - \Psi_{\mu',R'}(x)| \leq (|\mu - \mu'| + |R - R'|)\mathbb{1}\{|x - \mu_P| > \sigma_P\sqrt{n/(8b_{\max})}\}.$$

By Chebyshev's inequality, this implies that, for any positive integer p,

$$P|\Psi_{\mu,R} - \Psi_{\mu',R'}|^p \leq (|\mu - \mu'| + |R - R'|)^p \frac{8b_{\max}}{n}.$$

By Bennett's inequality,

$$\mathbb{P}\left(\frac{|\Delta_{\mu,R} - \Delta_{\mu',R'}|}{|\mu - \mu'| + |R - R'|} > \frac{8b_{\max}}{n} + \frac{4\sqrt{b_{\max}t}}{n} + \frac{t}{3n}\right) \leq 2e^{-t}.$$

To apply a chaining argument, consider the sequence $(D_j)_{j\geq 0}$ of subsets of $\mathcal{R}$ obtained by the following construction. $D_0 = \{(\mu_P, \sigma_P\sqrt{n/(2b_{\max})})\}$. For any $j \geq 1$, divide $\mathcal{R}$ into 4^j pieces by dividing each axis into 2^j pieces of equal sizes. Define then D_j as the set of lower left corners of the 4^j rectangles. Then $|D_j| = 4^j$ and, for any $(\mu, R) \in \mathcal{R}$, there exists a point $\pi_j(\mu, R) \in D_j$ such that the ℓ_1-distance between (μ, R) and $\pi_j(\mu, R)$ is upper-bounded by $2^{-j}(\varepsilon_\mu + \varepsilon_R)\sigma_P$. Therefore,

$$\sup_{(\mu,R)\in\mathcal{R}} |\Delta_{(\mu,R)}| \leq \sum_{j\geq 1} \sup_{(\mu,R)\in D_j} |\Delta_{(\mu,R)} - \Delta_{\pi_{j-1}(\mu,R)}|.$$

A union bound in Bennett's inequality gives that, with probability at least $1 - 2^{1-j}e^{-t}$, for any $(\mu, R) \in D_j$,

$$|\Delta_{\mu,R} - \Delta_{\pi_{j-1}(\mu,R)}|$$
$$\leq (\varepsilon_\mu + \varepsilon_R)\sigma_P\left(\frac{16b_{\max}}{2^j n} + \frac{8\sqrt{b_{\max}(t + j\log 8)}}{2^j n} + \frac{2t + 2j\log 8}{3n2^j}\right).$$

Summing up these inequalities gives the desired bound. $\quad\square$

COROLLARY 7.1. *Assume $t \geq 1$, $\sqrt{n/(2b_{\max})} \geq 2(\varepsilon_\mu + \varepsilon_R)$. Then, with probability at least $1 - 2e^{-t} - 2e^{-b_{\max}}$, for all $(\mu, R) \in \mathcal{R}$,*

$$|\widehat{P}_n\Psi_{\mu,R} - \mu_P| \leq \frac{\sigma_P}{\sqrt{n}}(\sqrt{2t}(1 + \xi_1) + \xi_2),$$

where

$$\xi_1 = \frac{1}{3\sqrt{2}}\sqrt{\frac{\kappa_P t}{n}} + \frac{5}{48}\frac{\kappa_P t}{n} + 2(\varepsilon_\mu + \varepsilon_R)\left(\sqrt{\frac{2b_{\max}}{n}} + \frac{1}{3}\sqrt{\frac{t}{n}}\right),$$

$$\xi_2 = 2\sqrt{2}\kappa_P\frac{b_{\max}^{3/2}}{n} + 56(\varepsilon_\mu + \varepsilon_R)\frac{b_{\max}}{n}.$$

Step 4. (Wrap-up.) Define now

$$(\widehat{\mu}_n, \widehat{R}_n) = \left(\widehat{\mu}_{b_{\max}}, \widehat{v}_{b_{\max}}\sqrt{\frac{n}{2b_{\max}}}\right).$$

From Lemma 7.2, with probability at least $1 - 2e^{-b_{\max}}$,

$$|\widehat{\mu}_n - \mu| \leq 2\sqrt{2}e\sigma_P\sqrt{\frac{b_{\max}}{n}}, \qquad |\widehat{v}_{b_{\max}}^2 - \sigma_P^2| \leq 2e\sqrt{6(\kappa + 3)}\sigma_P^2\sqrt{\frac{b_{\max}}{n}}.$$

The second inequality gives

$$\sqrt{1 - 2e\sqrt{6(\kappa_P + 3)}\sqrt{\frac{b_{\max}}{n}}} \leq \frac{\widehat{\nu}_{b_{\max}}}{\sigma_P} \leq \sqrt{1 + 2e\sqrt{6(\kappa_P + 3)}\sqrt{\frac{b_{\max}}{n}}}.$$

Since we can assume that

$$2e\sqrt{6(\kappa_P + 3)}\sqrt{\frac{b_{\max}}{n}} \leq 1,$$

we deduce that

$$|\widehat{\nu}_{b_{\max}} - \sigma_P| \leq e\sqrt{12(\kappa_P + 3)}\sigma_P\sqrt{\frac{2b_{\max}}{n}}.$$

This means that, with probability at least $1 - 2e^{-b_{\max}}$, $(\widehat{\mu}_n, \widehat{R}_n)$ belongs to $\mathcal{R}$ if we define

$$\varepsilon_\mu = 2\sqrt{2}e\sqrt{\frac{b_{\max}}{n}} \leq \sqrt{\kappa_P}, \qquad \varepsilon_R = 2e\sqrt{3(\kappa_P + 3)} \leq 19\sqrt{\kappa_P}.$$

By an appropriate choice of the constant C, we can always assume that $\sqrt{n/(2b_{\max})\kappa_P}$ is at least some large constant, to ensure that $2(\varepsilon_\mu + \varepsilon_R) \leq \sqrt{n/(2b_{\max})}$. So Corollary 7.1 applies and gives

$$\mathbb{P}\left(|\widehat{\mathbb{P}}_n\Psi_{\widehat{\mu}_n, \widehat{R}_n} - \mu_P| \leq \frac{\sigma_P}{\sqrt{n}}(\sqrt{2t}(1 + \xi_1) + \xi_2)\right) \geq 1 - 2e^{-t} - 4e^{-b_{\max}},$$

where

$$\xi_1 = 36\sqrt{\frac{\kappa_P b_{\max}}{n}}, \qquad \xi_2 = 2\sqrt{2}\kappa_P\frac{b_{\max}^{3/2}}{n} + 1120\sqrt{\kappa_P}\frac{b_{\max}}{n}.$$

In particular, if $\delta > \frac{4e}{e-2}e^{-b_{\max}}$, we get

$$\mathbb{P}\left(|\widehat{\mathbb{P}}_n\Psi_{\widehat{\mu}_n, \widehat{R}_n} - \mu_P| \leq \frac{\sigma_P}{\sqrt{n}}\left(\sqrt{2(1 + \ln(1/\delta))}(1 + \xi_1) + \xi_2\right)\right)$$

$$\geq 1 - \left(\frac{2}{e} + \frac{4}{4e/(e-2)}\right)\delta = 1 - \delta.$$

REMARK 2. Let us quickly sketch how one may get a smaller value of $\delta_{\min}$ at the expense of a larger constant L. The idea is to redo the proof of part 1 of Theorem 3.2 (cf. Section 5). We build δ-dependent estimators for μ_P via median-of-means, as in (14), but then use the value $2\widehat{\sigma}_b(X_1^n)$ from Lemma 7.2 instead of the value σ_2^2 when building the confidence interval, with a choice of $b \approx \ln(1/\delta)$. Then one obtains an empirical confidence interval that contains μ_P and has the appropriate length with probability $\geq 1 - 2\delta$ whenever $\ln(1/\delta) \leq cn/\kappa$ for some

constant $c > 0$. Using Theorem 4.2 as in Section 5 then gives a multiple-δ L-sub-Gaussian estimator for $(\mathcal{P}_{\mathrm{krt} \leq \kappa}, n, e^{1-cn/\kappa})$ for large enough values of n/κ, where L does not depend on n or κ. It is an open question whether one can obtain a similar value of $\delta_{\min}$ with $L = \sqrt{2} + o(1)$.

8. Open problems.

We conclude with some especially interesting open problems.

Sharper estimators. For what families $\mathcal{P}$ and what values of $\delta_{\min}$ can one find multiple-δ estimators with sharp constant $L = \sqrt{2} + o(1)$? What about estimators that satisfy the stronger property

$$\mathbb{P}\left(\left| \widehat{E}_n(X_1^n) - \mu_{\mathrm{P}} \right| > \sigma_{\mathrm{P}} \frac{\Phi^{-1}(1 - \delta/2)}{\sqrt{n}} \right) \leq (1 + o(1))\delta?$$

Sub-Gaussian confidence intervals. The notion of sub-Gaussian confidence interval introduced in Section 4.2 seems interesting on its own right. For which classes of distributions $\mathcal{P}$ can one find sub-Gaussian confidence intervals? Can one reverse the implication in Theorem 4.2, and build sub-Gaussian confidence intervals from multiple-δ estimators?

Multivariate settings. There has been some work on sub-Gaussian δ-dependent estimates for vectors; see [9, 11, 16]. A natural question is whether one can obtain similar multiple-δ estimators. By way of comparison, let $X_1, \ldots, X_n \in \mathbb{R}^p$ be i.i.d. random Gaussian (column) vectors with law P, mean μ_{P} and covariance matrix $\Sigma_{\mathrm{P}} := \mathrm{P}(X - \mu_{\mathrm{P}})(X - \mu_{\mathrm{P}})^{\dagger}$. Then Gaussian concentration implies

$$\mathbb{P}\left(\left\| (\widehat{\mathrm{P}}_n - \mathrm{P})X \right\|_2 > \sqrt{\frac{\mathrm{Tr}(\Sigma_{\mathrm{P}})}{n}} + \sqrt{2 \frac{\|\Sigma_{\mathrm{P}}\|_{2 \to 2} \ln(1/\delta)}{n}} \right) \leq \delta,$$

where $\|\cdot\|_2$ is the Euclidean norm, Tr is the trace and $\|\cdot\|_{2 \to 2}$ is the corresponding operator norm. Slightly stronger bounds are also available for certain sub-Gaussian ensembles [8]. Can some multiple-δ estimator $\widehat{E}_n(X_1, \ldots, X_n)$ achieve similar properties under much heavier tails? One natural path toward these results would be to work with higher dimensional analogues of quantiles, kurtosis and k-regular classes $\mathcal{P}_{2,k-\mathrm{reg}}$. (We thank an anonymous referee for suggesting this problem.)

Empirical risk minimization. Suppose now that the X_i are i.i.d. random variables that live in an arbitrary measurable space and have common law P. In a prototypical risk minimization problem, one wishes to find an approximate minimum $\widehat{\theta}_n(X_1^n)$ over choices of $\theta \in \Theta$ of a loss functional $\ell(\theta) := \mathrm{P}f(X, \theta)$. The usual way to do this is via empirical risk minimization, which consists of minimizing the empirical risk $\widehat{\ell}_n(\theta) := \widehat{\mathrm{P}}_n f(X, \theta)$ instead. Under strong assumptions on the family $F := \{f(\cdot, \theta)\}_{\theta \in \Theta}$ (such as uniform boundedness), the fluctuations of the empirical process $\{(\widehat{\mathrm{P}}_n - \mathrm{P})f(X, \theta)\}_{\theta \in \Theta}$ can be bounded in terms of geometric or combinatorial properties of F.

One may try to extend these results to heavier-tailed functions by replacing $\widehat{\mathrm{P}}_n f(X, \theta)$ by one of our multiple-δ sub-Gaussian estimates. However, this is not

straightforward: our estimators are nonlinear in the sample, and usual chaining breaks down. There are (artificial) ways around this, we do not know of any *natural method* for doing empirical risk minimization with our estimators. These difficulties were overcome by Brownlees et al. [3] via Catoni's multiple-δ sub-exponential estimator, at the cost of obtaining weaker concentration. Can one do something similar and achieve truly sub-Gaussian results, preferably at low computational cost?

REFERENCES

[1] ALON, N., MATIAS, Y. and SZEGEDY, M. (1996). The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing* (*Philadelphia*, *PA*, 1996) 20–29. ACM, New York. MR1427494

[2] AUDIBERT, J.-Y. and CATONI, O. (2011). Robust linear least squares regression. *Ann*. *Statist*. **39** 2766–2794. MR2906886

[3] BROWNLEES, C., JOLY, E. and LUGOSI, G. (2015). Empirical risk minimization for heavy-tailed losses. *Ann*. *Statist*. **43** 2507–2536. MR3405602

[4] BUBECK, S., CESA-BIANCHI, N. and LUGOSI, G. (2013). Bandits with heavy tail. *IEEE Trans*. *Inform*. *Theory* **59** 7711–7717. MR3124669

[5] CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann*. *Inst*. *Henri Poincaré Probab*. *Stat*. **48** 1148–1185. MR3052407

[6] CHEN, L. H. Y., GOLDSTEIN, L. and SHAO, Q. (2011). *Normal Approximation by Stein's Method*. Springer, Berlin.

[7] HSU, D. (2010). Robust statistics. Available at http://www.inherentuncertainty.org/2010/12/robust-statistics.html.

[8] HSU, D., KAKADE, S. M. and ZHANG, T. (2012). A tail inequality for quadratic forms of sub-Gaussian random vectors. *Electron*. *Commun*. *Probab*. **17** 1–6. MR2994877

[9] HSU, D. and SABATO, S. (2014). Heavy-tailed regression with a generalized median-of-means. In *Proceedings of the* 31*st International Conference on Machine Learning* (*ICML*-14) (T. Jebara and E. P. Xing, eds.) 37–45. JMLR Workshop and Conference Proceedings.

[10] JERRUM, M. R., VALIANT, L. G. and VAZIRANI, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoret*. *Comput*. *Sci*. **43** 169–188. MR0855970

[11] JOLY, E. (2015). Estimation robuste de distributions à queue lourde. Ph.D. thesis, Mathematics at Univ. Paris-Saclay.

[12] LEPSKIĬ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab*. *Appl*. **35** 454–466.

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

[14] LERASLE, M. and OLIVEIRA, R. I. (2012). Robust empirical mean estimators. Available at arXiv:1112.3914.

[15] LEVIN, L. A. (2005). Notes for miscellaneous lectures. *CoRR*. Available at arXiv:cs/0503039.

[16] MINSKER, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468

[17] NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York. MR0702836

[18] PETROV, V. V. (1975). *Sums of Independent Random Variables*. Springer, Berlin. MR0388499

L. DEVROYE
SCHOOL OF COMPUTER SCIENCE
MCGILL UNIVERSITY
3480 UNIVERSITY STREET
MONTREAL H3A 0E9
CANADA
E-MAIL: lucdevroye@gmail.com

G. LUGOSI
DEPARTMENT OF ECONOMICS
UNIVERSITAT POMPEU FABRA
RAMON TRIAS FARGAS 25-27
08005 BARCELONA
SPAIN
E-MAIL: gabor.lugosi@gmail.com

M. LERASLE
LABORATOIRE J. A. DIEUDONNÉ
UNIVERSITÉ NICE SOPHIA ANTIPOLIS
06108 NICE CEDEX 02
FRANCE
E-MAIL: mlerasle@unice.fr

R. I. OLIVEIRA
IMPA
ESTRADA DA. CASTORINA, 110
RIO DE JANEIRO, RJ 22460-320
BRAZIL
E-MAIL: rimfo@impa.br