# Sub-Gaussian mean estimators

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, Roberto I. Oliveira

# Sub-Gaussian mean estimators

Luc Devroye Matthieu Lerasle Gábor Lugosi Roberto I. Oliveira

September 24, 2015

**Abstract**

We discuss the possibilities and limitations of estimating the mean of a real-valued random variable from independent and identically distributed observations from a non-asymptotic 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

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

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,

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

where $\mu_{\mathrm{P}}$ and $\sigma_{\mathrm{P}}^2 > 0$ are the mean and variance of P (respectively) and $\Phi$ 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 non-asymptotic version of the mean estimation problem. We are interested in large, non-parametric classes of distributions, such as

$$\begin{aligned} \mathcal{P}_2 &:= \{\text{all distributions over } \mathbb{R} \text{ with finite second moment}\} && (2)\\ \mathcal{P}_2^{\sigma^2} &:= \{\text{all distributions } \mathrm{P} \in \mathcal{P}_2 \text{ with variance } \sigma_{\mathrm{P}}^2 = \sigma^2\} \ \ (\sigma^2 > 0) && (3)\\ \mathcal{P}_{\mathrm{krt}\le\kappa} &:= \{\text{all } \mathrm{P} \in \mathcal{P}_2 \text{ with kurtosis} \le \kappa\} \ \ (\kappa > 1), && (4) \end{aligned}$$

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 $\mathrm{P} \in \mathcal{P}$ and produce an estimate $\widehat{E}_n(X_1^n)$ of $\mu_{\mathrm{P}}$ that satisfies

$$\mathbb{P}\left( |\widehat{E}_n(X_1^n) - \mu_{\mathrm{P}}| > L\,\sigma_{\mathrm{P}} \sqrt{\frac{(1+\ln(1/\delta))}{n}} \right) \le \delta \quad \text{for all } \delta \in [\delta_{\min}, 1) \tag{5}$$

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 $\delta$ fixed, (5) is a weaker form of (1) since $\Phi^{-1}(1-\delta/2) \le \sqrt{2\ln(2/\delta)}$. The point is that (5) should hold non-asymptotically, for extremely small $\delta$, and uniformly over $\mathrm{P} \in \mathcal{P}$, 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 $\delta_{min}$ is quite large (cf. Section 2.3.1), so designing sub-Gaussian estimators with the kind of guarantee we look for is a non-trivial 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 $\delta$ 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 superexponentially small in $n$.) In the specific case of bounded-kurtosis distributions (cf. (4) above), we achieve $L \le \sqrt{2} + \epsilon$ for $\delta_{\min} \approx e^{-o\left((n/\kappa)^{2/3}\right)}$. 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 *$\delta$-dependent estimator*, that is, an estimator $\widehat{E}_n = \widehat{E}_{n,\delta}$ that may also depend on the confidence parameter $\delta$. By contrast, the estimators that we introduce here are called *multiple-$\delta$ 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 $\delta$-dependent estimators in the paper. In particular, we show that the distinction is substantial. For instance, there are no multiple-$\delta$ sub-Gaussian estimators for the full class $\mathcal{P}_2$ for any nontrivial range of $\delta_{\min}$. Interestingly, multiple-$\delta$ estimators do exist (with $\delta_{\min} \approx e^{-c\,n}$) 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 $\sigma_P^2$. 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., $\mathbb{E}\left[|X - \mathbb{E}X|^{1+\alpha}\right] = M$ 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 $\sigma_P$ 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 $\delta$, the estimator is at least $M^{1/(1+\alpha)} \left(\frac{\ln(1/\delta)}{n}\right)^{\alpha/(1+\alpha)}$ away from the target $\mu_{\mathrm{P}}$.

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 $\mathcal{P}^{\mathrm{Ber}}$ 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-$\delta$ 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 $P \in \mathcal{P}^{\mathrm{Ber}}$,

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

Of course, this bound has a sub-Gaussian flavor as it resembles (5) except that the confidence bounds do not scale by $\sigma_{\mathrm{P}}(\log(1/\delta)/n)^{1/2}$ 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 $\mathrm{P} \in \mathcal{P}$,

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

for some constant $L > 0$. $\delta$-dependent and multiple-$\delta$ 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 $\sup_{\mathrm{P}\in\mathcal{P}} \sigma_{\mathrm{P}} < \infty$, 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 $\delta$-dependent and multiple-$\delta$ estimators, and our construction of multiple-$\delta$ sub-Gaussian estimators for exponentially small $\delta$, are all new. On the other hand, constructions of $\delta$-dependent estimators are implicit in older work on stochastic optimization of Nemirovsky and Yudin [14] (see also Levin [12] and Hsu [6]), sampling from large discrete structures by Jerrum, Valiant, and Vazirani [8], and 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 [7], Brownlees, Joly, and Lugosi [3], Lerasle and Oliveira [11], Minsker [13], Audibert and Catoni [2], Bubeck, Cesa-Bianchi, and Lugosi [4]. Most of these papers use $\delta$-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 $\delta$-dependent sub-Gaussian estimators with nearly optimal $L = \sqrt{2} + o(1)$ for a wide range of $\delta$ and the classes $\mathcal{P}_2^{\sigma^2}$ and $\mathcal{P}_{\mathrm{krt}\le\kappa}$ defined in (3). The $\delta$-dependent sub-Gaussian estimator introduced by [5] may be converted into a multiple-$\delta$ estimators with subexponential (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-$\delta$ estimators for $\mathcal{P}_2$ with subexponential 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 $\delta$-dependent (Theorem 4.3), and the Poisson family for multiple-$\delta$ 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 $\delta$-dependent into multiple-$\delta$ estimators, one *can* build multiple-$\delta$ 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 $\mu_{\mathrm{P}}$ with "sub-Gaussian length", one may combine these intervals to produce a single multiple-$\delta$ estimator. This general construction is presented in Section 4.2 and is related at a high level to Lepskii's adaptation method [9, 10].

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 $\mathbb{N} = \{0, 1, 2, \dots\}$. For a positive integer $n$, denote $[n] = \{1, \dots, n\}$. $|A|$ denotes the cardinality of the finite set $A$.

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

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 $\theta$, 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. **$\delta$-dependent sub-Gaussian estimation:** *a $\delta$-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 $\mathrm{P} \in \mathcal{P}$, $\delta \in [\delta_{\min}, 1)$, and $X_1^n = (X_1, \ldots, X_n)$ is a sample of i.i.d. random variables distributed as P, then*

$$\mathbb{P}\left( |\widehat{E}_n(X_1^n, \delta) - \mu_{\mathrm{P}}| > L\, \sigma_{\mathrm{P}} \sqrt{\frac{(1 + \ln(1/\delta))}{n}} \right) \leq \delta \ . \tag{6}$$

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

2. **multiple-$\delta$ sub-Gaussian estimation:** *a multiple-$\delta$ $L$-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 $\delta \in [\delta_{\min}, 1)$, $\mathrm{P} \in \mathcal{P}$ and i.i.d. sample $X_1^n = (X_1, \ldots, X_n)$ distributed as P,*

$$\mathbb{P}\left( |\widehat{E}_n(X_1^n) - \mu_{\mathrm{P}}| > L\, \sigma_{\mathrm{P}} \sqrt{\frac{(1 + \ln(1/\delta))}{n}} \right) \leq \delta \ . \tag{7}$$

It transpires from these definitions that multiple-$\delta$ 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 $\delta$ in advance means that $\delta$-dependent estimators may be too pessimistic when a small $\delta$ 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-$\delta$ $L$-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*

$$\mathbb{P}\left(\widehat{E}_{n,\delta}(X_1^n) - \mu_{\mathrm{P}} > \frac{r\,\sigma_{\mathrm{P}}}{\sqrt{n}}\right) \leq \delta \text{ whenever } \mathrm{P} \in \mathcal{P} \ ,$$

*then $r \geq \Phi^{-1}(1-\delta)$. The same result holds for the lower tail. Since $\Phi^{-1}(1-\delta) \sim \sqrt{2\ln(1/\delta)}$ for small $\delta$, 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 $\delta$ is possible. In particular, sub-Gaussian estimators are optimal up to constants, and estimators with $L \leq \sqrt{2} + o(1)$ 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

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

is *not* a $L$-sub-Gaussian estimator for the class $\mathcal{P}_2^{\sigma^2}$ of all distibutions with variance $\sigma^2$. 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 ([15, Theorem 14, p. 125]) implies that, for large $n$, $\widehat{\mathsf{emp}}_n$ is a multiple-$\delta$ $\left(\sqrt{2}+\epsilon\right)$-sub-Gaussian estimator for $(\mathcal{P}_{3,\eta}, n, \delta_{\min,n})$, where

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

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 $\kappa$ 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 $\widehat{\mathsf{emp}}_n$ is $\left(\sqrt{2}+\epsilon\right)$-sub-Gaussian for the triple $(\mathcal{P}_{\infty,\eta}, n, \delta_{\min})$, with $\delta_{\min} = e^{-\epsilon^2 n/\eta^2}$ and

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

### 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 $\delta$-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 [14], Levin [12], Jerrum, Valiant, and Vazirani [8], 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 $\delta$-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-$\delta$ estimator, however with subexponential instead of sub-Gaussian tails (i.e., the $\sqrt{\ln(1/\delta)}$ term in (7) appears squared). Both estimators work for exponentially small $\delta$, although the constant in the exponent for $\mathcal{P}_{\mathrm{krt}\leq\kappa}$ depends on $\kappa$.

# 3 Main results

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

## 3.1 On the non-existence 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., $\mathbb{E}\left[|X - \mathbb{E}X|^{1+\alpha}\right] = M$). We start by pointing out that when $\alpha < 1$, no sub-Gaussian estimators exist (even if one allows $\delta$-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 $\widehat{E}_n$,*

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

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 $\delta$, there exists an estimator $\widehat{E}_n(X_1^n, \delta)$ such that

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

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-$\delta$ 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 $\sigma_1^2$ and $\sigma_2^2$:

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

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 proven 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-$\delta$ $L^{(1)}$-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-$\delta$ $L$-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 $\delta_{\min}^{(3)} = e^{1-5L^2 n}$, there is no $\delta$-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-$\delta$ $L$-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-$\delta$ $L$-sub-Gaussian estimators exist for any sequence $\delta = \delta_{\min,n} \to 0$. Finally, the third item says that even when $R_n \equiv 1$, $\delta$-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

Theorem 3.2 shows that finite, but completely unknown variance is too weak an assumption for multiple-$\delta$ sub-Gaussian estimation. The following shows that what we call *regularity conditions* can substitute for knowledge of the variance.

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

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

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

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

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

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

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

$$\mathcal{P}_{\alpha,\eta} = \{\mathrm{P} \in \mathcal{P}_2 \,:\, \mathrm{P}|X - \mu_{\mathrm{P}}|^\alpha \leq (\eta\,\sigma_{\mathrm{P}})^\alpha\}. \tag{8}$$

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

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

Our main result about $k$-regular classes states that sub-Gaussian multiple-$\delta$ estimators exist for $\mathcal{P}_{2,k-\mathrm{reg}}$ 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* $L_* = 4\sqrt{2\,(1+2\ln 2)\,(1+62\ln(3))}\,e^{\frac{5}{2}}$*. Then there exists a* $L_*$*-sub-Gaussian multiple-*$\delta$ *estimator for* $(\mathcal{P}_{2,k-\mathrm{reg}}, n, \delta_{\min,n,k})$*.*

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 $\delta$ for these two smaller classes. The next result, for symmetric distributions, is proven in Section 6.2.

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

1. *the estimator obtained in Theorem 3.3 for $k = 1$ is a $L_*$-sub-Gaussian multiple-$\delta$ estimator for $(\mathcal{P}_{2,\mathrm{sym}}, n, \delta_{\min,n,1})$ when $n \geq (3 + \ln 2)\, 124$;*
2. *on the other hand, for any $L \geq \sqrt{2}$, no $\delta$-dependent $L$-sub-Gaussian estimator can exist for $(\mathcal{P}_{2,\mathrm{sym}}, n, e^{1-5L^2 n})$.*

We also have an analogue 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 $\alpha$ 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-$\delta$ estimator for $(\mathcal{P}_{\alpha,\eta}, n, \delta_{\min,n,k_\alpha})$ when $n \geq (3 + \ln 4)\, 124 k_\alpha$;*
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-$\delta$ $L$-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 $\delta$-dependent $L$ sub-Gaussian estimator for $(\mathcal{P}_{\alpha,\eta}, n, e^{1-5L^2 n})$.*

### 3.4 Bounded kurtosis and nearly optimal constants

This section shows that multiple-$\delta$ sub-Gaussian estimation with nearly optimal constants can be proved when the kurtosis

$$\kappa_{\mathrm{P}} = \frac{\mathbb{E}(X - \mu_{\mathrm{P}})^4}{\sigma_{\mathrm{P}}^4}$$

(when $X =_d \mathrm{P}$) is uniformly bounded in the class. (For completeness, we set $\kappa_{\mathrm{P}} = 1$ when $\sigma_{\mathrm{P}}^2 = 0$.) More specifically, we will consider the class $\mathcal{P}_{\mathrm{krt} \leq \kappa}$ of all distributions $\mathrm{P} \in \mathcal{P}_2$ with $\kappa_{\mathrm{P}} \leq \kappa$.

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 $b_{\max} \ll (n/\kappa)^{2/3}$, $\xi = o(1)$. 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-$\delta$ $L$-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 $b_{\max} \ll (n/\kappa_n)^{2/3}$ and obtain multiple-$\delta$ $\big(\sqrt{2} + o\,(1)\big)$-sub-Gaussian estimators $(\mathcal{P}_{\mathrm{krt}\leq\kappa}, n, \delta_{\min}^{(4)})$ for $\delta_{\min}^{(4)} \approx e^{-b_{\max}}$. Catoni [5] obtained $\delta$-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 $\delta$ with a multiple-$\delta$ estimator, albeit with worse constant $L$.

# 4 General methods

We collect here some ideas that recur in the remainder of the paper.

1. Section 4.1 presents an analysis of the median-of-means estimator mentioned in Section 2.3.2 above. We present a proof based on Hsu's argument [6].

2. Section 4.2 presents a "black-box method" of deriving multiple-$\delta$ estimators from confidence intervals. The point is that confidence intervals are "$\delta$-dependent objects", and thus easier to design and analyze.

3. In Section 4.3 we use scaled Bernoulli distributions to prove the impossibility of designing (weakly) sub-Gaussian estimators for classes with distributions with unbounded variance.

4. Section 4.4 uses the family of Laplace distributions to lower bound $\delta_{\min}$ for $\delta$-dependent estimators.

5. Section 4.5 uses the Poisson family to derive lower bounds on $\delta_{\min}$ for multiple-$\delta$ estimators.

A combination of the above results will allow us to derive the sharp range for $\ln(1/\delta_{\min})$ for all families of distributions we consider.

### 4.1 Median of means

The next result is a well known performance bound for the median-of-means estimator. We include the proof for completeness.

**Theorem 4.1** *For any $n \geq 4$ and $L = 2\sqrt{2}\,e$ there exists a $\delta$-dependent $L$-sub-Gaussian estimators for $(\mathcal{P}_2, n, e^{1-n/2})$.*

*Proof:* We follow the argument of Hsu [6]. 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,

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

(If several $i$ fit the above description, we take the smallest one.) We need the following Lemma (proven subsequently):

**Lemma 4.1** *Let $Y_1^b = (Y_1, \ldots, Y_b) \in \mathbb{R}^b$ be independent random variables with the same mean $\mu$ and variances bounded by $\sigma^2$. Assume $L_0 > 1$ is given and $M_b = q_{1/2}(Y_1^b)$. Then $\mathbb{P}\left(|M_b - \mu| > 2\,L_0\,\sigma\right) \leq L_0^{-b}$.*

In our case we set $L_0 = e = L/2\sqrt{2}$. 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 $b \leq n/2$.

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

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

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

We now show that $\widehat{E}_{n,\delta}$ is a sub-Gaussian estimator for the class $\mathcal{P}_2$. Let $X_1^n =_d \mathrm{P}^{\otimes n}$ for a distribution $\mathrm{P} \in \mathcal{P}_2$. $\widehat{E}_{n,\delta}(X_1^n)$ is the median of random variables

$$Y_i = \frac{1}{|B_i|}\sum_{j \in B_i} X_j = \widehat{\mathrm{P}}_{B_i} X \qquad i \in [b]\ .$$

Each $Y_i$ has mean $\mu_{\mathrm{P}}$ and variance $\sigma_{\mathrm{P}}^2/\#B_i \leq \sigma_{\mathrm{P}}^2/k$. Then, using our choice of $b$, Lemma 4.1 implies

$$\mathbb{P}\left(|\widehat{E}_{n,\delta}(X_1^n) - \mu_{\mathrm{P}}| > \frac{2\,L_0\sigma_{\mathrm{P}}}{\sqrt{k}}\right) \leq L_0^{-b} \leq \delta\ .$$

Now, because $b = \lceil \ln(1/\delta) \rceil \leq n/2$

$$k = \left\lfloor \frac{n}{b} \right\rfloor \geq \frac{n}{b} - 1 \geq \frac{n}{2b} \geq \frac{n}{2(1 + \ln(1/\delta))},$$

and

$$\frac{2\,L_0}{\sqrt{k}} \leq \frac{2\,L_0\sqrt{2}\,\sqrt{1+\ln(1/\delta)}}{\sqrt{n}} = L\sqrt{\frac{1+\ln(1/\delta)}{n}}.$$

Therefore,

$$\mathbb{P}\left( |\widehat{E}_{n,\delta}(X_1^n) - \mu_{\mathrm{P}}| > L\,\sigma_{\mathrm{P}}\,\sqrt{\frac{\ln(1/\delta)}{n}} \right) \leq \delta\ ,$$

and since this works for any $\mathrm{P} \in \mathcal{P}_2$, the proof is complete. □

*Proof of Lemma 4.1:* Let $I = [\mu - 2\,L_0\,\sigma, \mu + 2\,L_0\sigma]$. Clearly,

$$M_b \notin I \Rightarrow \#\{j \in [b]\,:\,Y_j \notin I\} \geq \frac{b}{2} \Rightarrow \sum_{j=1}^{b} 1\,\{Y_j \notin I\} \geq \frac{b}{2}.$$

The indicators variables on the right-hand side are all independent, and by Chebyshev's inequality, for all $j \in [b]$,

$$\mathbb{P}\,(Y_j \notin I) \leq \frac{\mathbb{E}\left[(Y_j - \mu)^2\right]}{4\,L_0^2\sigma^2} \leq \frac{1}{4L_0^2}.$$

We deduce that $\sum_{j=1}^{b} 1\,\{Y_j \notin I\}$ is stochastically dominated by a binomial random variable $\mathrm{Bin}(b, (2L_0)^{-2})$ and therefore,

$$\begin{aligned}\mathbb{P}\,(M_b \notin I) &\leq \mathbb{P}\left(\mathrm{Bin}(b,(2L^0)^{-2}) \geq \frac{b}{2}\right) = \sum_{k=\lceil b/2\rceil}^{b} \binom{b}{k} \left(\frac{1}{(2L_0)^2}\right)^k \left(1 - \frac{1}{(2L_0)^2}\right)^{b-k} \\ &\leq \left(\frac{1}{(2L_0)^2}\right)^{\lceil b/2\rceil} \sum_{k=\lceil b/2\rceil}^{b} \binom{b}{k} \leq L_0^{-b}\end{aligned}$$

since $\sum_{k=\lceil b/2\rceil}^{b} \binom{b}{k} \leq \sum_{k=0}^{b} \binom{b}{k} = 2^b$.

### 4.2 The method of confidence intervals for multiple-$\delta$ estimators

In this section we detail how sub-Gaussian confidence intervals may be combined to produce multiple-$\delta$ estimators. This will be our main tool in defining all multiple-$\delta$ 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 $\hat{a}_{n,\delta} \leq \hat{b}_{n,\delta}$. 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 a collection $L$-sub-Gaussian confidence intervals for $(n, \mathcal{P}, \delta_{\min})$ if for any $\mathrm{P} \in \mathcal{P}$, if $X_1^n =_d \mathrm{P}^{\otimes n}$, then for all $\delta \in [\delta_{\min}, 1)$,*

$$\mathbb{P}\left(\mu_{\mathrm{P}} \in \widehat{I}_{n,\delta}(X_1^n) \text{ and } \hat{\ell}_{n,\delta}(X_1^n) \leq L\,\sigma_{\mathrm{P}} \sqrt{\frac{1+\ln(1/\delta)}{n}}\right) \geq 1-\delta.$$

The next theorem shows how one can combine sub-Gaussian confidence intervals to obtain a multiple-$\delta$ 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-$\delta$ 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 $\mathrm{P} \in \mathcal{P}$ and $X_1^n =_d \mathrm{P}^{\otimes n}$, the event

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

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

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

One can easily check that

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

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 $X_1^n =_d \mathrm{P}^{\otimes n}$ with $\mathrm{P} \in \mathcal{P}$. Then

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

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

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

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

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

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

Putting it all together, we conclude

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

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

### 4.3 Scaled Bernoulli distributions and single-$\delta$ estimators

In this subsection we prove Theorem 3.1. In order to do so, we derive a simple minimax lower bound for single-$\delta$ 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

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

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 $\delta \geq 2e^{-n/4}$ and $p = (2/n)\log(2/\delta)$, then (using $1 - p \geq \exp(-p/(1-p))$),

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

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

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

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

$$\begin{aligned}
&\max\left(\mathbb{P}\left\{\left|\widehat{E}_{n,\delta}(X_1^n)-\mu_{P_+}\right|>\left(\frac{M^{1/\alpha}}{n}\log\frac{2}{\delta}\right)^{\alpha/(1+\alpha)}\right\},\right.\\
&\left.\mathbb{P}\left\{\left|\widehat{E}_{n,\delta}(Y_1^n)-\mu_{P_-}\right|>\left(\frac{M^{1/\alpha}}{n}\log\frac{2}{\delta}\right)^{\alpha/(1+\alpha)}\right\}\right)\geq\delta\ .
\end{aligned}$$

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

### 4.4 Laplace distributions and single-$\delta$ estimators

This section focuses on the class of *all Laplace distibutions with scale parameter equal to* 1. To define such a distribution, let $\lambda\in\mathbb{R}$ and let $\mathsf{La}_\lambda$ 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 $\delta$-dependent $L$-sub-Gaussian estimators are limited to exponentially small $\delta$ 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 $\delta$-dependent $L$-sub-Gaussian estimators for $(\mathcal{P}_{\mathsf{La}},n,e^{1-5L^2n})$.*

*Proof:* We proceed by contradiction, assuming that there exist $L$-sub-Gaussian $\delta$-dependent estimators $\widehat{E}_{n,\delta}$ for $(\mathcal{P}_{\mathsf{La}},n,\delta)$ where $\delta=e^{1-5L^2n}$ 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,

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

Using the definition of $\lambda$ 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

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

If we use again the definition of $\lambda$, we see that

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

or

$$e^{-2\sqrt{5}\,L^2\,n} \leq 2e^{1-5L^2 n} \Rightarrow n \leq \frac{1+\ln 2}{L^2\,(5-2\sqrt{5})}\ .$$

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

### 4.5 Poisson distributions and multiple-$\delta$ estimators

We use the family of Poisson distributions for bounding the range of confidence values of multiple-$\delta$ estimators. Denote by $\mathsf{Po}_\lambda$ the Poisson distribution with parameter $\lambda > 0$. Given $0 < \lambda_1 \leq \lambda_2 < \infty$, 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 $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ such that the following holds. Assume $L \geq \sqrt{2}$ and $n > 0$ are given. Then there exists no multiple-$\delta$ $L$-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 $c \geq c_0$, $L \geq \sqrt{2}$ and $C = \lceil s\, (L^2 \ln L) \rceil$, there is no multiple-$\delta$ sub-Gaussian estimator for

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

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

We proceed by contradiction. Assume

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

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

**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** $\mathbb{P}\left(S_Y = (1+2C)\, c\right) \geq 1/4\sqrt{(1+2C)\, c}$ 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 $c \geq c_0$, $\mathbb{P}\left(S_X = (1+2C)\, c\right) \geq e^{-h(C)\, c}$. 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 $\approx 1/L\sqrt{s \ln L\, c}$, which is much larger than the minimum confidence parameter $e^{1 - C^2\, c/L^2}$ allowed by $(\star)$ (at least if $c \geq c_0$ with a large enough $c_0$). Recalling **F0**, we obtain

$$\mathbb{P}\left( n\widehat{E}_n(Y_1^n) < (1+2C)\, c - L\sqrt{(1+2C)\, c(1 + \ln(8\sqrt{(1+2C)\, c}))} \right) \leq \frac{1}{4\sqrt{(1+C)\, c}}\ .$$

Therefore, by **F3**,

$$\mathbb{P}\left(n\widehat{E}_n(Y_1^n) < (1+2C)\,c - L\sqrt{(1+2C)\,c(1+\ln(8\sqrt{(1+2C)\,c}))} \mid S_Y = (1+C)\,c\right) \le 1/2.$$

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

$$\mathbb{P}\left(n\widehat{E}_n(X_1^n) \ge (1+2C)\,c - L\sqrt{(1+2C)\,c(1+\ln(8\sqrt{(1+2C)\,c}))} \mid S_X = (1+2C)\,c\right) \ge 1/2. \tag{12}$$

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

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

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

$$L\sqrt{(1+2C)\,c(1+\ln(8\sqrt{(1+2C)\,c}))} \le C\,c\ .$$

So (12) gives

$$\mathbb{P}\left(n\widehat{E}_n(X_1^n) \ge (1+C)\,c \mid S_X = (1+2C)\,c\right) \ge 1/2.$$

We may combine this with **F4** to deduce:

$$\mathbb{P}\left(n\widehat{E}_n(X_1^n) \ge (1+C)\,c\right) \ge \frac{e^{-h(C)\,c}}{2}. \tag{13}$$

We now use **F0** to rewrite the previous probability as

$$\mathbb{P}\left(n\widehat{E}_n(X_1^n) \ge (1+C)\,c\right) = \mathbb{P}\left(\widehat{E}_n(X_1^n) - \mu_{\mathrm{P}} \ge L\,\sigma_{\mathrm{P}}\frac{\sqrt{1+\ln(1/\delta_0)}}{\sqrt{n}}\right),$$

where

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

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

$$\frac{e^{-h(C)\,c}}{2} \le \mathbb{P}\left(n\widehat{E}(X_1^n) \ge (1+C)\,c\right) \le e^{1-\frac{C^2 c}{4L^2}}.$$

Comparing the left and right hand sides, and recalling $c \ge c_0$, we obtain $h(C) \ge C^2/4L^2 - 1 - (\ln 2/c_0)$. 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

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

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

**Part 1:** (Existence of a multiple-$\delta$ estimator with constant depending on $R$.) Theorem 4.1 ensures that, irrespective of $\sigma_1$ or $\sigma_2$, for all $\delta \in (e^{1-n/2}, 1)$ there exists a $\delta$-dependent estimator $\widehat{E}_{n,\delta} : \mathbb{R}^n \to \mathbb{R}$ with

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

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

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

Clearly, (14) and the fact that $\sigma_2 \leq R\,\sigma_{\mathrm{P}}$ 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 $4\sqrt{2}\,e\,R$-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:** (Non-existence of multiple-$\delta$ 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:** (Non-existence of $\delta$-dependent estimators when $\delta_{\min} = e^{1-5L^2n}$.) 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 $4\sqrt{2\,(1+62\ln(3))}\,e^{\frac{5}{2}}$-sub-Gaussian confidence interval for $(\mathcal{P}_{2,k-\mathrm{reg}}, n, e^{3-n/(124)k})$.

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 $\alpha$-quantile $q_\alpha(y_1^b)$ of a vector $y_1^b \in \mathbb{R}^b$ as the smallest index $i \in [b]$ with

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

The next result (proven subsequently) is an analogue of Lemma 4.1.

**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 $\mu$ and variances bounded by $\sigma^2$. Assume further that $\mathbb{P}\,(Y_i \le \mu) \ge 1/3$ and $\mathbb{P}\,(Y_i \ge \mu) \ge 1/3$ for each $i \in [b]$. Then*

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

*where $d$ is the numerical constant*

$$d = \frac{1}{4}\,\ln\left(\frac{3}{4}\right) + \frac{3}{4}\,\ln\left(\frac{9}{8}\right) \approx 0.0164 > \frac{1}{62}$$

*and $L_0 = 2\,e^{2d+\frac{1}{2}} \le 2\,e^{\frac{5}{2}}$.*

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

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

Partition

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

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

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

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

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

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

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

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

all have mean $\mu_{\mathrm{P}}$ and variance $\leq \sigma_{\mathrm{P}}^2/s$. 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 $i \in [b]$,

$$\mathbb{P}\left(Y_i \leq \mu\right) \geq \frac{1}{3}, \mathbb{P}\left(Y_i \geq \mu\right) \geq \frac{1}{3}.$$

Lemma 6.1 implies

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

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

*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

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

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 $\sum_{i=1}^b 1\{Y_i \leq \mu\}$ stochastically dominates a binomial random variable $\mathrm{Bin}(b, 1/3)$. Thus,

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

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 $\mathbb{P}\left(q_{3/4}(Y_1^b) > \mu\right) \leq e^{-db}$ as well.

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

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

and these events are independent. It follows that

$$\begin{aligned}
\mathbb{P}\left(\#\{i \in [b] \,:\, Y_i \notin J\} > b/4\right) &\leq \mathbb{P}\left(\bigcup_{A \subset [b],\, |A| = \lceil b/4 \rceil} \bigcap_{i \in A} \{Y_i \notin J\}\right) \\
\text{(union bound)} &\leq \binom{b}{\lceil b/4 \rceil} \max_{A \subset [b],\, |A| = \lceil b/4 \rceil} \mathbb{P}\left(\bigcap_{i \in A} \{Y_i \notin J\}\right) \\
\text{(independence of } Y_i + (18)) &\leq \binom{b}{\lceil b/4 \rceil} \left(\frac{1}{L_0^2}\right)^{-\lceil \frac{b}{4} \rceil} \\
\left(\tbinom{b}{k} \leq (eb/k)^k \text{ for all } 1 \leq k \leq b\right) &\leq \left(\frac{e\,b}{L_0^2\,\lceil b/4 \rceil}\right)^{\lceil \frac{b}{4} \rceil} \\
\left(b \leq 4\lceil b/4 \rceil \text{ and } L_0^2 = 4e^{4d+1}\right) &\leq e^{-4d\lceil \frac{b}{4} \rceil} \leq e^{-bd}.
\end{aligned}$$

## 6.2 Symmetric distributions

To prove Theorem 3.4, notice that the existence of the multiple-$\delta$ 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 $\mathcal{P}_{\alpha,\eta} \subset \mathcal{P}_{2,k-\mathrm{reg}}$ 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 $j \geq (C_\alpha \eta)^{\frac{2\alpha}{\alpha-2}}$, then* $\min(p_+(\mathrm{P}, j), p_-(\mathrm{P}, j)) \geq 1/3$.

*Proof:* We only prove that $p_+(\mathrm{P}, j) \geq 1/3$, as the other proof is analogous.

Let $N$ be a standard normal random variable. Take some smooth function $\Psi : \mathbb{R} \to \mathbb{R}$ with bounded second and third derivatives, such that $\Psi(x) = 0$ for $x \in (-\infty, 0]$, $0 \leq \Psi(x) \leq 1$ for $x > 0$ and $\mathbb{E}[\Psi(N)] \geq 1/\sqrt{6}$. (It is easy to see that such a $\Psi$ exists.) Also let $X_1^j =_d \mathrm{P}^{\otimes j}$ and assume, without loss of generality, that $\sigma_{\mathrm{P}} > 0$. Then

$$p_+(\mathrm{P}, j) = \mathbb{P}\left(\frac{1}{\sigma_{\mathrm{P}}^2 \sqrt{j}} \sum_{i=1}^{j} (X_i - \mu_{\mathrm{P}}) \geq 0\right) \geq \mathbb{E}\left[\Psi\left(\frac{1}{\sigma_{\mathrm{P}}^2 \sqrt{j}} \sum_{i=1}^{j} (X_i - \mu_{\mathrm{P}})\right)\right].$$

Lindberg's proof of the central limit theorem (see [16]), specialized to the case where $X_1^j$ are i.i.d., gives

$$\mathbb{E}\left[\Psi\left(\frac{1}{\sigma_{\mathrm{P}}^2 \sqrt{j}} \sum_{i=1}^{j} (X_i - \mu_{\mathrm{P}})\right)\right] \geq \mathbb{E}[\Psi(N)] - C_0\, j\, \mathrm{P}\left[\phi\left(\frac{X - \mu_{\mathrm{P}}}{\sigma_{\mathrm{P}} \sqrt{j}}\right)\right],$$

where $\phi(t) = t^2 \wedge t^3$ and $C_0 > 0$ is a universal constant. Since $\mathbb{E}[\Psi(N)] \geq 1/\sqrt{6} > 1/3$ and $\phi(t) \leq t^\alpha$, we obtain

$$\mathbb{E}\left[\Psi\left(\frac{1}{\sigma_{\mathrm{P}}^2 \sqrt{j}} \sum_{i=1}^{j} (X_i - \mu_{\mathrm{P}})\right)\right] \geq \frac{1}{\sqrt{6}} - C_0\, j^{\frac{\alpha}{2}-1}\, \eta^\alpha\ .$$

The right-hand side is $\geq 1/3$ when $j \geq (C\eta)^{\frac{2\alpha}{\alpha-2}}$ for some universal $C = C_\alpha$. $\square$

*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 $\eta_0$. We use the Poisson family of distributions from Section 4.5. For $\lambda = o(1)$ and $\alpha \in (2, 3]$, we have that

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

If we compare this to Example 3.2, we see that $\mathsf{Po}_\lambda \in \mathcal{P}_{\alpha,\eta}$ 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-$\delta$ estimator for $(\mathcal{P}_{\mathsf{Po}}^{[c/n,\phi(L)\,c/n]}, n, e^{-c})$. Noting that $c$ is of the order $n/k_\alpha$ finishes 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

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

Our assumption on $\eta$ implies that $\mathcal{P}_{\mathsf{La}} \subset \mathcal{P}_{\alpha,\eta}$. Thus Theorem 4.3 implies that there is no $\delta$-dependent or multiple-$\delta$ sub-Gaussian estimator for $(\mathcal{P}_{\alpha,\eta}, n.e^{1-5L^2n})$. This is the desired result since $k_\alpha$ is bounded when $\eta \le \eta_0$.

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

# 7 Bounded kurtosis and nearly optimal constants

In this section we prove Theorem 3.6. Throughout the proof we assume $X =_d \mathrm{P}$ and $X_1^n =_d \mathrm{P}^{\otimes n}$ for some $\mathrm{P} \in \mathcal{P}_{\mathrm{krt}\le\kappa}$, 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.* We use the median-of-means technology to obtain preliminary estimates for the mean and variance of P. These estimates are not good enough to satisfy the claimed properties, but with extremely high probability they are reasonably close to the true values.

2. *Truncation at the ideal point.* We introduce a two-parameter family of truncation-based estimators for $\mu_{\mathrm{P}}$, and analyze the behavior of one such estimator, chosen under knowledge of $\mu_{\mathrm{P}}$ and $\sigma_{\mathrm{P}}$.

3. *Truncated estimators are insensitive.* Finally, 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."

We conclude the section by a remark on how to obtain a broader range of $\delta_{\min}$ with a worse constant $L$.

**Step 1.** (Preliminary estimates via median of means.) 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 $C \geq 6/(1-\log 2)$. The next lemma provides an estimator of the variance.

**Lemma 7.1** *Let $B_1, \dots, B_{b_{\max}}$ denote a partition of $[n]$ into blocks of size $|B_i| \geq k = \lfloor n/b_{\max} \rfloor \geq 2$. For each block $B_i$ with $i \in [b_{\max}]$, define*

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

*Then*

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

*In particular, if*

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

*then*

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

*Proof:* Compute

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}_i^4\right] = {} & \frac{1}{|B_i|^2(|B_i|-1)^2} \sum_{(j,k) \in B_i^{(2)}} \mathbb{E}\left[(X_j - X_k)^4\right] \\
& + \frac{6}{|B_i|^2(|B_i|-1)^2} \sum_{(j,k,l) \in B_i^{(3)}} \mathbb{E}\left[(X_j - X_k)^2 (X_j - X_l)^2\right] \\
& + \frac{1}{|B_i|^2(|B_i|-1)^2} \sum_{(j,k,l,m) \in B_i^{(4)}} \mathbb{E}\left[(X_j - X_k)^2 (X_l - X_m)^2\right]
\end{aligned}$$

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

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

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

Therefore,

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

Lemma 4.1 with $L_0 = e$ gives then

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

In particular, we get

$$\mathbb{P}\left(\frac{1}{2}\sigma^2_{\mathrm{P}} \leq \widehat{\nu}^2_{b_{\max}} \leq \frac{3}{2}\sigma^2_{\mathrm{P}}\right) \geq 1 - e^{-b_{\max}} \ .$$

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

**Step 2:** (Two-parameter family of estimators at the ideal point.) Given $\mu$ 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 7.2** *Assume $b_{\max} \geq t$, $R = \sigma_{\mathrm{P}}\sqrt{n/b_{\max}}$ and $\mu = \mu_{\mathrm{P}}$. Then, with probability at least $1 - 2e^{-t}$,*

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

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

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

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

$$\mathbb{E}\left[\left(\Psi_{\mu,R}(X) - \mu_{\mathrm{P}}\right)^2\right] \leq \sigma^2_{\mathrm{P}}$$

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

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

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

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

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

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

By Chernoff's bound,

$$\mathbb{P}\left(\widehat{\mathrm{P}}_n\Psi_{\mu,R} - \mu_{\mathrm{P}} > 2\sqrt{2}\kappa_{\mathrm{P}}\sigma_{\mathrm{P}}\left(\frac{b_{\max}}{n}\right)^{3/2} + \frac{s}{n}\sigma_{\mathrm{P}}^2 + \frac{s^2}{6n^2}\sqrt{\kappa_{\mathrm{P}}}\sigma_{\mathrm{P}}^3 + \frac{5s^3}{96n^3}\kappa_{\mathrm{P}}\sigma_{\mathrm{P}}^4\right) \leq e^{-\frac{s^2\sigma_{\mathrm{P}}^2}{2n}}$$

or, equivalently,

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

Repeat the same computations with $s = -\sqrt{2nt}/\sigma_{\mathrm{P}}$ to prove the lower bound. □

**Step 3:** (Insensitivity of the estimators.) Given $\epsilon_\mu, \epsilon_R \in (0, 1/2)$, define

$$\mathcal{R} = \left\{(\mu, R) \ : \ |\mu - \mu_{\mathrm{P}}| \leq \epsilon_\mu\sigma_{\mathrm{P}}, \quad \left|R - \sigma_{\mathrm{P}}\sqrt{n/(2b_{\max})}\right| \leq \epsilon_R\sigma_{\mathrm{P}}\right\} \ ,$$

$$\Delta_{\mu,R} = \widehat{\mathrm{P}}_n\left(\Psi_{\mu,R} - \Psi_{\mu_{\mathrm{P}},\sigma_{\mathrm{P}}\sqrt{n/(2b_{\max})}}\right) \ .$$

**Lemma 7.3** *Assume $\sqrt{n/(2b_{\max})} \geq 2(\epsilon_\mu + \epsilon_R)$ then for any $t > 0$, with probability at least $1 - e^{-t}$, for all $(\mu, R) \in \mathcal{R}$,*

$$|\Delta_{\mu,R}| \leq (\epsilon_\mu + \epsilon_R)\sigma_{\mathrm{P}}\left(\frac{56b_{\max}}{n} + \frac{4\sqrt{b_{\max}t}}{n} + \frac{2t}{3n}\right) \ .$$

*Proof:* Start with the trivial bound

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

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

$$|x - \mu| \leq |x - \mu_P| + \epsilon_\mu \sigma_{\mathrm{P}} \leq \sigma_{\mathrm{P}}\left(2\sqrt{n/(8b_{\max})} - \epsilon_R\right) \leq R \ .$$

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

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

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

$$\left|\Psi_{\mu,R}(x) - \Psi_{\mu',R'}(x)\right| \leq \left(\left|\mu - \mu'\right| + \left|R - R'\right|\right) 1\left\{|x - \mu_{\mathrm{P}}| > \sigma_{\mathrm{P}}\sqrt{n/(8b_{\max})}\right\} \ .$$

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

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

By Bennett's inequality,

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

To apply a chaining argument, consider the sequence $(D_j)_{j\geq 0}$ of points of $\mathcal{R}$ obtained by the following construction. $D_0 = (\mu_{\mathrm{P}}, \sigma_{\mathrm{P}}\sqrt{n/(2b_{\max})}$ and, 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 $\ell_1$-distance between $(\mu, R)$ and $\pi_j(\mu, R)$ is upper-bounded by $2^{-j}(\epsilon_\mu + \epsilon_R)\sigma_{\mathrm{P}}$. Therefore,

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

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

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

Summing up these inequalities gives the desired bound. □

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

$$\left|\widehat{\mathrm{P}}_n \Psi_{\mu,R} - \mu_{\mathrm{P}}\right| \leq \frac{\sigma_{\mathrm{P}}}{\sqrt{n}} \left(\sqrt{2t}(1+\xi_1) + \xi_2\right) \ ,$$

*where*

$$\xi_1 = \frac{1}{3\sqrt{2}}\sqrt{\frac{\kappa_{\mathrm{P}} t}{n}} + \frac{5}{48}\frac{\kappa_{\mathrm{P}} t}{n} + 2(\epsilon_\mu + \epsilon_R)\left(\sqrt{\frac{2b_{\max}}{n}} + \frac{1}{3}\sqrt{\frac{t}{n}}\right) \ ,$$

$$\xi_2 = 2\sqrt{2}\kappa_{\mathrm{P}}\frac{b_{\max}^{3/2}}{n} + 56(\epsilon_\mu + \epsilon_R)\frac{b_{\max}}{n} \ .$$

**Step 4:** (Wrap-up.) Define now

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

From Lemma 7.1, with probability at least $1 - 2e^{-b_{\max}}$,

$$|\widehat{\mu}_n - \mu| \leq 2\sqrt{2}\, e\sigma_{\mathrm{P}}\sqrt{\frac{b_{\max}}{n}}, \quad \left|\widehat{\nu}^2_{b_{\max}} - \sigma^2_{\mathrm{P}}\right| \leq 2e\sqrt{6(\kappa+3)}\sigma^2_{\mathrm{P}}\sqrt{\frac{b_{\max}}{n}} \ .$$

The second inequality gives

$$\sqrt{1 - 2e\sqrt{6(\kappa_{\mathrm{P}}+3)}\sqrt{\frac{b_{\max}}{n}}} \leq \frac{\widehat{\nu}_{b_{\max}}}{\sigma_{\mathrm{P}}} \leq \sqrt{1 + 2e\sqrt{6(\kappa_{\mathrm{P}}+3)}\sqrt{\frac{b_{\max}}{n}}}$$

Since we can assume that

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

we deduce that

$$|\widehat{\nu}_{b_{\max}} - \sigma_{\mathrm{P}}| \leq e\sqrt{12(\kappa_{\mathrm{P}}+3)}\sigma_{\mathrm{P}}\sqrt{\frac{2b_{\max}}{n}} \ .$$

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

$$\epsilon_\mu = 2\sqrt{2}\, e\sqrt{\frac{b_{\max}}{n}} \leq \sqrt{\kappa_{\mathrm{P}}}, \quad \epsilon_R = 2e\sqrt{3(\kappa_{\mathrm{P}}+3)} \leq 19\sqrt{\kappa_{\mathrm{P}}} \ .$$

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

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

where

$$\xi_1 = 36\sqrt{\frac{\kappa_{\mathrm{P}}\, b_{\max}}{n}}, \quad \xi_2 = 2\sqrt{2}\kappa_{\mathrm{P}}\frac{b_{\max}^{3/2}}{n} + 1120\sqrt{\kappa_{\mathrm{P}}}\frac{b_{\max}}{n}\ .$$

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

$$\mathbb{P}\left(\left|\widehat{\mathrm{P}}_n\Psi_{\widehat{\mu}_n,\widehat{R}_n} - \mu_{\mathrm{P}}\right| \le \frac{\sigma_{\mathrm{P}}}{\sqrt{n}}\left(\sqrt{2(1+\ln(1/\delta))}(1+\xi_1)+\xi_2\right)\right) \\ \ge 1 - \left(\frac{2}{e} + \frac{4}{4e/(e-2)}\right)\delta = 1-\delta\ .$$

**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 $\delta$-dependent estimators for $\mu_{\mathrm{P}}$ via median-of-means, as in (14), but then use the value $2\widehat{\sigma}_b(X_1^n)$ from Lemma 7.1 instead of the value $\sigma_2^2$ when building the confidence interval, with a choice of $b \approx \ln(1/\delta)$. Then one obtains an empirical confidence interval that contains $\mu_{\mathrm{P}}$ and has the appropriate length with probability $\ge 1-2\delta$ whenever $\ln(1/\delta) \le c\, n/\kappa$ for some constant $c > 0$. Using Theorem 4.2 as in Section 5 then gives a multiple-$\delta$ $L$-sub-Gaussian estimator for $(\mathcal{P}_{\mathrm{krt}\le\kappa}, n, e^{1-cn/\kappa})$ for large enough values of $n/\kappa$, where $L$ does not depend on $n$ or $\kappa$. 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 the paper by a partial list of problems related to our results that seem especially interesting.

**Sharper constants and truly sub-Gaussian estimators.** For what families $\mathcal{P}$ of distributions and what values of $\delta_{\min}$ can one find multiple-$\delta$ estimators with sharp constant $L = \sqrt{2} + o\,(1)$? One may even sharpen our definition of a sub-Gaussian estimator and ask for estimators that satisfy

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

for all $\mathrm{P} \in \mathcal{P}$ and $\delta \in [\delta_{\min}, 1)$?

**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-$\delta$ estimators?

**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 distribution P. In a prototypical risk minimization problem, one wishes to find an approximate minimum $\widehat{\theta}_n(X_1^n)$ of a functional $\ell(\theta) := \mathrm{P}\, f(\theta, X)$ over choices of $\theta \in \Theta$. The usual way to do this is via empirical risk minimization, which consists of minimizing the empirical risk $\hat{\ell}_n(\theta) := \widehat{\mathrm{P}}\, f(\theta, X)$ instead. Under strong assumptions on the family $F := \{f(\theta, \cdot)\}_{\theta \in \Theta}$ (such as uniform boundedness), the fluctuations of the empirical process $\{(\widehat{\mathrm{P}}_n - \mathrm{P})\, f(\theta, X)\}_{\theta \in \Theta}$ can be bounded in terms of geometric or combinatorial properties of $F$, and this leads to results on empirical risk minimization. However, the strong sub-Gaussian concentration results one may obtain are only available when $F$ has very light tails.

A natural way to obtain strong sub-Gaussian concentration for heavier-tailed $F$ would be to replace the usual empirical estimates $\widehat{\mathrm{P}}\, f(\theta, X)$ by one of our multiple-$\delta$ sub-Gaussian estimates. This, however, is not straightforward. The usual chaining technique for controlling empirical processes rely on linearity, and our estimators are nonlinear in the sample. Although there are (artificial) ways around this, we do not know of any *efficient method* for doing the analogue of empirical risk minimization with our estimators in any nontrivial setting. These difficulties were overcome by Brownlees et al. [3] via Catoni's multiple-$\delta$ subexponential estimator, at the cost of obtaining weaker concentration. Can one do something similar and achieve truly sub-Gaussian results at low computational cost?

## 9 Acknowledgements

Luc Devroye was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada. Gábor Lugosi and Roberto Imbuzeiro Oliveira gratefully acknowledge support from CNPq, Brazil via the *Ciência sem Fronteiras* grant # 401572/2014-5. Gábor Lugosi was supported by the Spanish Ministry of Science and Technology grant MTM2012-37195. Roberto Imbuzeiro Oliveira's work was supported by a *Bolsa de Produtividade em Pesquisa* from CNPq. His work in this article is part of the activities of FAPESP Center for Neuromathematics (grant# 2013/ 07699-0 , FAPESP - S.Paulo Research Foundation).

## References

[1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pages 20–29, New York, NY, USA, 1996. ACM.

[2] J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794, 10 2011.

[3] C. Brownlees, E. Joly, and G. Lugosi. Empirical risk minimization for heavy-tailed losses. *Annals of Statistics*, to appear, 2015.

[4] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *Information Theory, IEEE Transactions on*, 59(11):7711–7717, Nov 2013.

[5] O. Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. H. Poincaré Probab. Statist.*, 48(4):1148–1185, 11 2012.

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

[7] D. Hsu and S. Sabato. Heavy-tailed regression with a generalized median-of-means. In Tony Jebara and Eric P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 37–45. JMLR Workshop and Conference Proceedings, 2014.

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

[9] O. V. Lepskii. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and its Applications*, 36:454–466, 1990.

[10] O. V. Lepskii. Asymptotically minimax adaptive estimation I: Upper bounds. optimally adaptive estimates. *Theory of Probability and its Applications*, 36:682–697, 1991.

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

[12] L. A. Levin. Notes for miscellaneous lectures. *CoRR*, abs/cs/0503039, 2005.

[13] S. Minsker. Geometric median and robust estimation in Banach spaces. *arXiv preprint*, 2013.

[14] A. Nemirovsky and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.

[15] V.V. Petrov. *Sums of Independent Random Variables*. Springer-Verlag, Berlin, 1975.

[16] D. W. Stroock. *Probability theory: an analytic view*. Cambridge University Press, 2003.