Sub-Gaussian Confidence Intervals for Heavy-Tailed Data: Characterizing the Limits of Inference

Ilyes Hammouda
Department of Statistics
CREST ENSAE
ilyes.hammouda@ensae.fr

Stanislav Minsker
Department of Mathematics
University of Southern California
minsker@usc.edu

Mohamed Ndaoud
Information and Data Systems
ESSEC Business School
ndaoud@essec.edu

Abstract

We investigate the problem of uncertainty quantification in the mean estimation framework. Given an i.i.d. sample from a distribution with unknown mean μ and variance σ^2, we want to construct an estimator $\widehat{\mu}_n$ and provide a computable and size-optimal upper bound for the error $|\widehat{\mu}_n - \mu|$. While estimators possessing strong deviation guarantees are well known, no data-dependent non-asymptotic upper bounds for their performance exist in general. We show that this gap can be bridged by introducing a parameter that characterizes the "effective sample size" available for uncertainty quantification. This parameter captures the difficulty of the problem, interpolating between "easy" cases where sub-Gaussian confidence intervals exist and "hard" cases where their construction is impossible. Using this characterization, we design confidence intervals of optimal length that are fully adaptive to the unknown variance. Numerical experiments confirm that our approach maintains nominal coverage even in asymmetric and heavy-tailed regimes where other existing methods fail.

1 Introduction

Estimation and uncertainty quantification problems related to the mean of a population are among the oldest and most studied in mathematical statistics, going back to the works of W. Gosset [Student, 1908], Fisher [1925], and Neyman [1937], among others. Over the past 15 years, particular attention has been drawn to the range of questions dealing with robust mean estimation – a problem of mean estimation in the presence of contamination, often modeled via the outliers, or "atypical" observations. The research has largely centered around construction of optimal estimators while the problem of robust inference received less attention. In this paper, we ask the following question: given a sequence of i.i.d. copies $X_1, \dots, X_n$ of a random variable X with mean μ and finite variance σ^2 where both parameters are unknown, when is it possible to construct a (non-asymptotic) confidence interval $\widehat{C}_n(X_1, \dots, X_n)$ of finite length and such that $\mathbb{P}\Big(\widehat{C}_n \ni \mu\Big) \geq 0.95$? Moreover, we are interested in characterizing the optimal size of such intervals.

Distinctions between the problems of rate-optimal adaptive estimation and adaptive inference are well known in the literature: there exist scenarios where the first task admits a solution while the second one does not. Classical examples include the problem of constructing adaptive confidence

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

bands for the probability density functions of unknown smoothness [Low, 1997] as well as the problem of sparse linear regression with unknown sparsity level [Nickl and van de Geer, 2013].

As the statistical research paradigm shifts from an asymptotic to a non-asymptotic perspective, new challenges in the basic problem of mean estimation have emerged. It has been shown by Catoni [2012], Devroye et al. [2016] that for distributions with finite variance[1] there exist sub-Gaussian mean estimators, that is, measurable functions $\widehat{\mu}_n(\delta) := \widehat{\mu}_n(X_1, \ldots, X_n)(\delta)$ such that

$$\mathbb{P}\left(|\widehat{\mu}_n(\delta) - \mu| \geq C\sigma\sqrt{\frac{\log(1/\delta)}{n}}\right) \leq \delta$$

for all $e^{-cn} \leq \delta \leq e^{-1}$ where $c, C > 0$ are absolute constants. Moreover, it is impossible to improve this bound uniformly, aside from the absolute constants [Catoni, 2012]. Note that $\widehat{\mu}_n(\delta)$ does not require σ to be known. When X has normal distribution, the empirical mean $\bar{X}_n$ satisfies the inequality

$$\mathbb{P}\left(|\bar{X}_n - \mu| \geq Cs_n\sqrt{\frac{\log(1/\delta)}{n}}\right) \leq \delta, \tag{1}$$

for $e^{-cn} \leq \delta \leq e^{-1}$ where $s_n^2 = \frac{1}{n-1}\sum_{j=1}^n (X_j - \bar{X}_n)^2$ is the empirical variance (see Lemma B.1 below for a version with explicit constants). The inequality (1) can be equivalently rewritten as

$$\mathbb{P}\left(\mu \in \left[\bar{X}_n - Cs_n\sqrt{\frac{\log(1/\delta)}{n}}, \bar{X}_n + Cs_n\sqrt{\frac{\log(1/\delta)}{n}}\right]\right) \geq 1 - \delta.$$

Therefore, one may expect that, like in the mean estimation problem, this result for confidence intervals may be extended to all distributions in the domain of attraction of the normal law. However, such "universal" non-asymptotic confidence intervals do not exist: for any measurable $\widehat{L}_n(X_1, \ldots, X_n)$, $\widehat{U}_n(X_1, \ldots, X_n)$ such that $-\infty < \widehat{L}_n \leq \widehat{U}_n < \infty$ almost surely,

$$\inf_{X \sim P \in \mathcal{P}^{(2)}} \mathbb{P}\Big([\widehat{L}_n, \widehat{U}_n] \ni \mu\Big) = 0$$

where $\mathcal{P}^{(2)}$ is the class of all distributions with finite variance. This result follows from Corollary 3 in [Bahadur and Savage, 1956].

The present paper provides further refinement of this impossibility result: we construct a nested sequence of subclasses of $\mathcal{P}^{(2)}$ and characterize the optimal performance of confidence intervals over these subclasses. We also show that it is impossible to "adapt" to the scale defining these subclasses in a data-dependent way. Our "effective sample size" framework extends to distributions with finite $(1+\beta)$-moments, for $\beta \in (0, 1]$. The corresponding generalized effective-mass parameter, optimal confidence-interval rate, and matching lower bound are stated in Section 2 and proved in Appendix A.

Let us mention that another classical framework for modeling outliers is the Huber's contamination model [Huber, 1964] where the distribution of X is a mixture $(1-\varepsilon)P + \varepsilon Q$ of a distribution P with mean μ (a parameter of interest), and an arbitrary distribution Q. When ε is known and P has variance bounded by a known constant σ_0^2, Wang and Ramdas [2023] showed that it is possible to construct confidence intervals of length scaling as $\sigma_0\sqrt{\varepsilon}$ (also see [Bateni and Dalalyan, 2020, Maurer and Pontil, 2009, Shekhar and Ramdas, 2023] for results related to a class of compactly supported distributions). However, when ε is unknown, the price for adaptation is very high: recently, Luo and Gao [2024] showed that if P is the normal distribution with variance σ^2, then the confidence interval has to have length at least $\sigma(\log(n))^{-1/2}$ even if there are no outliers in the data ($\varepsilon = 0$) but potential contamination is suspected. If moreover σ^2 is unknown, finite-length confidence intervals do not exist (even though the "inliers" are normally distributed).

1.1 Notation

We use c, c_1, C, C_1, α, etc. to denote absolute positive constants and c_n, α_n, etc. to denote constants depending on n. For two sequences $(a_n)_n$ and $(b_n)_n$ we say that $a_n = \Omega(b_n)$ (resp. $a_n = \mathcal{O}(b_n)$) if

[1] Samples from such distributions exhibit naturally occurring outliers, hence they present a natural model for contaminated data.

$a_n \geq Cb_n$ (resp. $a_n \leq Cb_n$) for some $C > 0$ and large values of n. We say $a_n \simeq b_n$ if $a_n = \Omega(b_n)$ and $a_n = \mathcal{O}(b_n)$. We also say that $a_n = o(b_n)$ iff $a_n/b_n \to 0$. For a positive integer N, $[N]$ will denote the set $\{1, \dots, N\}$. For a given set A, $|A|$ will denote the cardinality of the latter. For a sequence of K elements $(u_n)_{n\in[K]}$, $\text{med}(u_1 \cdots u_K)$ will refer to the median of the sequence, i.e., if we order the K elements in ascending order, such that $u_{(1)} \leq u_{(2)} \cdots u_{(K)}$, then

$$\text{med}(u_1, \dots, u_K) = \begin{cases} u_{\left(\frac{K+1}{2}\right)}, & \text{if } K \text{ is odd} \\ \frac{1}{2}\left[u_{\left(\frac{K}{2}\right)} + u_{\left(\frac{K}{2}+1\right)}\right], & \text{if } K \text{ is even} \end{cases}.$$

We will also use notations $\left(\xrightarrow{d}\right)$ to denote convergence in distribution, $\left(\xrightarrow{a.s.}\right)$ for almost sure convergence and $\left(\xrightarrow{\mathbf{P}}\right)$ for convergence in probability. Finally, given a random variable ξ and a probability distribution P, we write $\xi \sim P$ to say that ξ is distributed according to P. Additional notation will be introduced on demand.

2 When do confidence intervals exist?

While there are no non-trivial adaptive confidence intervals for the class $\mathcal{P}^{(2)}$, they do exist for large sub-classes $\tilde{\mathcal{P}}$ of $\mathcal{P}^{(2)}$. A notable construction of such intervals is due to Devroye et al. [2016, Lemma 6.1] where it is shown that it suffices to require that the sample averages have "approximately symmetric" distributions for all sufficiently large samples. Specifically, the sufficient condition is that $\forall j \geq \lfloor \frac{n}{\log(1/\delta)} \rfloor$ and all $P \in \tilde{\mathcal{P}}$,

$$\min(p_-(j), p_+(j)) \geq 1/3, \tag{2}$$

where $p_-(j) = \mathbb{P}\Big(\frac{1}{j}\sum_{i=1}^{j} X_i \leq \mu\Big)$ and $p_+(j) = \mathbb{P}\Big(\frac{1}{j}\sum_{i=1}^{j} X_i \geq \mu\Big)$[2]. Of course, assumption (2) holds for every fixed $P \in \mathcal{P}^{(2)}$ and all $j \geq j(P)$ whenever $j(P)$ is sufficiently large in view of the Central Limit Theorem, but it does not hold uniformly over $\mathcal{P}^{(2)}$ since $\sup_{P\in\mathcal{P}^{(2)}} j(P) = \infty$. Natural classes of distributions $\tilde{\mathcal{P}}$ that satisfy assumption (2) are the ones that admit uniform bounds in the central limit theorem, for example, such that for some $p > 2$,

$$\sup_{P\in\tilde{\mathcal{P}}} \frac{\mathbf{E}^{1/p}|X-\mu|^p}{\sigma} \leq C\left(\tilde{\mathcal{P}}\right) < \infty.$$

In this case, condition (2) is implied by Berry-Esseen type inequalities [Petrov, 2012].

Main contribution. In this paper, we construct a family $\{\mathcal{P}_\alpha(n,\delta)\}_{0\leq\alpha\leq 1}$ of nested subclasses of $\mathcal{P}^{(2)}$, formally defined below, such that

$$\mathcal{P}^{(2)} = \mathcal{P}_0(n,\delta) \supseteq \mathcal{P}_{\alpha_1}(n,\delta) \supseteq \mathcal{P}_{\alpha_2}(n,\delta), \qquad \alpha_1 \leq \alpha_2,$$

and such that for each $\mathcal{P}_\alpha(n,\delta)$ there exist optimal-length sub-Gaussian confidence intervals.

Definition. *Let n be an integer, $0 < \delta < 1$. Let X be a random variable with mean μ and variance $\sigma^2 > 0$, and set $\xi := \xi(X) = \frac{X-\mu}{\sigma}$. Define*

$$\alpha_{n,\delta}(\xi) := \sup\left\{\gamma \geq 0\,, \mathbf{E}\left(\xi^2 \mathbf{1}\left(\xi^2 \leq \frac{n\gamma}{\log(1/\delta)}\right)\right) \geq \gamma\right\},$$

and the associated class $\mathcal{P}_\alpha(n,\delta) := \left\{P \in \mathcal{P}^{(2)} :\ X \sim P \text{ and } \alpha_{n,\delta}(\xi) \geq \alpha\right\}$.

Notice that $\alpha_{n,\delta}(\xi)$ is always finite. Indeed, $\gamma = 0$ belongs to the set

$$\mathcal{A} = \left\{\gamma \geq 0\,, \mathbf{E}\left(\xi^2 \mathbf{1}\left(\xi^2 \leq \frac{n\gamma}{\log(1/\delta)}\right)\right) \geq \gamma\right\},$$

and moreover $\mathcal{A} \subseteq [0,1]$. Hence, $0 \leq \alpha_{n,\delta}(\xi) \leq 1$. Observe also that $\alpha_{n,\delta}(\xi) \neq 0$ implies that $n \geq \log(1/\delta)$, a known requirement necessary (up to absolute constants) for existence of sub-Gaussian mean estimators. A lower bound on α excludes distributions with a "dominant" mass

[2]The constant $1/3$ in (2) can be replaced by any $0 < C < 1/2$.

at 0 from the class $\mathcal{P}_\alpha(n,\delta)$, as such distributions often lead to various impossibility results. For example, it is easy to see that the inequality

$$\mathbb{P}(1/2 \le \xi^2 \le 4) \ge 1/2$$

implies that $\alpha_{n,\delta}(\xi) \ge 1/4$ (of course, the constants in this relation can be modified). As we show below, $\alpha_{n,\delta}(\xi)$ quantifies the proportion of a sample of size n that can be used to estimate σ^2, and hence we can think of the product $n\alpha_{n,\delta}(\xi)$ as the *"effective sample size"*.

Let us also mention that for any $\varepsilon > 0$, there exists $\alpha(\varepsilon)$ small enough and a random variable X such that its law $P_X \in \mathcal{P}_\alpha$ and $\min(p_-(j), p_+(j)) \le \varepsilon$, as shown formally in Lemma B.2. Therefore, the approach of Devroye et al. [2016] is limited to the case $\alpha_{n,\delta}$ is a constant bounded away from 0. We are ready to state our main result.

Theorem 2.1. ***(i)*** *Let $0 < \delta < 1$ and assume that $\xi(X)$ is distributed according to $P \in \mathcal{P}^{(2)}$. There exists a sequence of confidence intervals $\widehat{C}_n(X_1,\dots,X_n) = [\widehat{L}_n(X_1,\dots,X_n),\ \widehat{U}_n(X_1,\dots,X_n]$ such that for all $\alpha_n \in (0,1]$,*

$$\inf_{P\in\mathcal{P}_{\alpha_n}(n,\delta)} \mathbb{P}\Big(\widehat{C}_n \ni \mu\Big) \ge 1-\delta \quad \text{and}$$

$$\inf_{P\in\mathcal{P}^{(2)},\sigma=\sigma_*} \mathbb{P}\left(|\widehat{C}_n| \le K\sigma_*\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}\right) \ge 1-\delta,$$

as long as $n\alpha_n \ge C\log(1/\delta)$, where $|\widehat{C}_n| = |\widehat{U}_n - \widehat{L}_n|$ and $C, K > 0$ are absolute constants. Moreover, $\widehat{L}_n, \widehat{U}_n$ are adaptive to P and $\alpha_{n,\delta}(\xi)$, meaning that they do not require any of these parameters to be known in advance, except for the lower bound α_n specifying the class $\mathcal{P}_{\alpha_n}(n,\delta)$.
(ii) *Let $\alpha_n, \gamma_n, \delta \in (0,1)$ be such that $\gamma_n > \alpha_n$, and let $c > 0$ be a sufficiently small absolute constant. Then there is no sequence of adaptive confidence intervals $\widetilde{C}_n(X_1,\dots,X_n)$ that simultaneously satisfy the conditions*

$$\inf_{P\in\mathcal{P}_{\alpha_n}(n,\delta)} \mathbb{P}\Big(\widetilde{C}_n \ni \mu\Big) \ge 1-\delta \quad \text{and}$$

$$\inf_{P\in\mathcal{P}_{\gamma_n}(n,\delta),\sigma=\sigma_*} \mathbb{P}\left(|\widetilde{C}_n| \le c\sigma_*\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}\right) \ge 1-\delta.$$

Remark 2.1. *(a) This theorem does not contradict the Central Limit Theorem as for every fixed distribution $P \in \mathcal{P}^{(2)}$, and $X \sim P$, $\alpha_{n,\delta}(\xi(X)) \to 1$ when $n \to \infty$.*

(b) Theorem 2.1 establishes that the optimal length for adaptive confidence intervals over the class $\mathcal{P}_{\alpha_n}(n,\delta)$ is of order $\sigma\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}$.

(c) While part (i) guarantees existence of such confidence intervals, part (ii) states that it is essentially impossible to adapt to the value of α_n: even if the data come from a distribution belonging to $\mathcal{P}_{\gamma_n}(n,\delta)$ with $\gamma_n > \alpha_n$, any confidence interval that guarantees uniform coverage of order δ over the class $\mathcal{P}_{\alpha_n}(n,\delta)$ has to have length at least $\tilde{C}\sigma\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}$.

(d) Theorem 2.1 characterizes the interpolation between the "easy" regime, where α_n is bounded away from zero by a constant, and the "hard" regime, where $\alpha_n \to 0$. Here, the product $n\alpha_n$ serves as the effective sample size, as the length of the adaptive sub-Gaussian confidence interval scales precisely with $(n\alpha_n)^{-1/2}$.

Interpretation of the hard regime. The case when $\alpha_{n,\delta}$ is close to 0 should not be interpreted as simply an artificial theoretical construct. Rather, it describes distributions with a "variance gap": observations of small or moderate magnitude account for only a small fraction of the population variance, while observations with high values have very low probability of occurrence but still contribute most to the variance. Such a regime naturally arises in imbalanced mixture distributions. For

instance, in the construction underlying our lower bounds, the minority component may have weight of order $\log(1/\delta)/n$. For $n = 200$ and $\delta = 0.05$, this corresponds to about 1.5% of the population, or roughly three observations on average. These observations may look like outliers relative to the dominant component, but they belong to the target distribution and removing them changes the estimand. Thus, a typical sample may appear approximately Gaussian with small variance, while a few legitimate observations determine both its centering and a substantial fraction of its population variance. There are natural examples where such rare "outliers" carry most of the useful information in the sample: financial returns with occasional large losses or gains, insurance portfolios with extremely rare catastrophic claims that can lead to insurer's financial losses if left unaccounted for, errors in measurement systems such as GPS positioning where one wants to account for the possible loss of satellite connection.

Note that within the class of distributions where $\alpha_n \asymp n^{-1}\log(1/\delta)$, all optimal adaptive intervals have length of order σ. This implies that in such cases, one cannot expect the confidence intervals to shrink around the true mean as the sample size n increases.

A deviation inequality whose radius depends on the unknown scale σ does not by itself provide a computable confidence interval. The construction of Devroye et al. [2016] requires the approximate-symmetry condition (2), whereas Lemma B.2 shows that this condition may fail within $\mathcal{P}_{\alpha_n}(n,\delta)$ when α_n is small.

Provided that $n\alpha_n = \Omega\left(\log(1/\delta)\right)$, the resulting interval is guaranteed to maintain the prescribed coverage. In table 1 below, we include a summary of existing results showing how our findings interpolate findings from the prior work.

Table 1: Comparison of optimal adaptive confidence interval lengths for the mean μ under two moments.

Case	**Optimal Interval Length**	**Reference**
σ known	$\sigma\sqrt{\frac{\log(1/\delta)}{n}}$	[Nemirovskii and Yudin, 1983]
σ unknown + symmetry condition	$\sigma\sqrt{\frac{\log(1/\delta)}{n}}$	Devroye et al. [2016]
σ unknown and $P_X \in \mathcal{P}_{\alpha_n}$	$\sigma\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}$	**Theorem 2.1**
σ unknown	∞ (Does not exist)	Bahadur and Savage [1956]

Extension to lower moments. Our results extend beyond the finite-variance setting. Assume that $\mathbf{E}|X-\mu|^{1+\beta} < \infty$ for some $\beta \in (0,1]$, and let $v_{1+\beta} := \left(\mathbf{E}|X-\mu|^{1+\beta}\right)^{1/(1+\beta)}$. Defining the corresponding effective-mass parameter $\alpha_{n,\delta}^{(\beta)}$ through the truncated $(1+\beta)$-moment, the optimal confidence interval length becomes

$$v_{1+\beta}\left(\frac{\log(1/\delta)}{n\alpha_{n,\delta}^{(\beta)}}\right)^{\frac{\beta}{1+\beta}}.$$

Both the upper bound and the matching lower bound extend to this setting: the proof follows the same scheme after replacing the truncated second-moment condition and the variance estimator by their $(1+\beta)$-moment analogues. The case $\beta = 1$ recovers the finite-variance results developed throughout the paper. We provide the formal definition of $\alpha_{n,\delta}^{(\beta)}$ and the corresponding proof details in Appendix A.

Roadmap of the proof of Theorem 2.1. We now outline the key steps and ideas of the proof.

1. **Construct a preliminary robust mean estimator.** We start from a sub-Gaussian estimator $\widehat{\mu}$ satisfying
$$|\widehat{\mu}-\mu| \lesssim \sigma\sqrt{\frac{\log(1/\delta)}{n}}$$
under the finite variance assumption. The same observations are then reused for scale estimation. *No sample splitting is required.*

2. **Estimate the scale through blockwise truncated mass.** We partition the sample into $K \asymp \log(1/\delta)$ blocks and compute blockwise empirical second moments around $\widehat{\mu}$. The definition of $\alpha_{n,\delta}$ guarantees that each block contains enough moderate observations to provide a lower bound of order $\alpha_n \sigma^2$, while the finite variance assumption gives a constant-order upper bound. Median aggregation amplifies these fixed-probability blockwise controls, yielding Proposition 4.1: $c\,\alpha_n \sigma^2 \le \widehat{\sigma}^2_{\mathrm{MoM}} \le C\,\sigma^2$ with probability at least $1-\delta$.
3. **Build the confidence interval.** Combining the robust mean estimator with the scale estimate gives an interval with radius $\widehat{\sigma}_{\mathrm{MoM}}\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}$, and hence length of order $\sigma\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}$. This proves the upper bound in Theorem 2.1.
4. **Design an α_n-sensitive lower bound pair.** The lower bound requires more than a classical two-mass-point construction. We must simultaneously ensure that the null belongs to the "difficult" class $\mathcal{P}_{\alpha_n}(n,\delta)$, that the alternative belongs to an "easier" class $\mathcal{P}_{\gamma_n}(n,\delta)$, and that their means differ at the target scale $\sigma_*\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}$. Writing $L = \log(1/\delta)$, we choose

$$\tau = \frac{L}{n}, \qquad A \asymp \sqrt{\frac{L}{n\alpha_n}},$$

and construct two mixtures sharing the same dominant Gaussian component:

$$P_0 = (1-\tau)\mathcal{N}(A,1) + \tau\delta_{z_0}, \qquad P_1 = (1-\tau)\mathcal{N}(A,1) + \tau\delta_0,$$

where z_0 is chosen so that P_0 is centered. The rare component under P_0 is genuine and population-relevant: it centers the distribution and carries the variance needed to place the standardized law in the hard class. Replacing this component by an atom at zero produces an easier alternative whose mean is shifted at the desired scale.
5. **Exploit the common-mixture event.** If none of the n observations come from the rare component, the two samples are generated from exactly the same Gaussian distribution. This event has probability

$$(1-\tau)^n = \left(1 - \frac{L}{n}\right)^n \gtrsim e^{-L} = \delta.$$

Therefore, the product measures retain an overlap of order δ, and every test must incur total error at least of this order. A Neyman–Pearson argument yields the separation lower bound of order $\rho_n \gtrsim \sigma_*\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}$.
6. **Transfer testing limits to confidence intervals.** Test/confidence set duality converts the testing lower bound into the optimal length and non-adaptivity statements of Theorem 2.1. The same mixture principle also gives Proposition 4.2 by producing nearly indistinguishable models whose variances differ by a factor of order α_n.

3 A hypothesis testing perspective

In order to get a better understanding of the problem related to adaptive confidence intervals, we will reduce it to a hypothesis testing problem. We still assume that $X = \mu + \sigma\xi$, where $\mu \in \mathbb{R}, \sigma > 0$ and ξ a centered random variable such that $\mathbf{E}(\xi^2) = 1$. We will also denote via $P_\xi \in \mathcal{P}^{(2)}$ the distribution of ξ. Consider the following testing problem:

$$\mathbf{H}_0 : \mu = 0 \quad \text{vs} \quad \mathbf{H}_1 : |\mu| \ge \rho_n,$$

where $\rho_n > 0$. Given n i.i.d. observations $X_1, \dots, X_n$, we define the minimax risk $\mathcal{R}^*(n, \rho_n, \sigma_*)$ via

$$\mathcal{R}^*(n,\rho_n,\sigma_*) := \inf_{\hat{T}} \left(\sup_{\mu=0,\sigma=\sigma_*,P_\xi\in\mathcal{P}^{(2)}} \mathbb{P}\Big(\hat{T}=1\Big) + \sup_{|\mu|\ge\rho_n,\sigma=\sigma_*,P_\xi\in\mathcal{P}^{(2)}} \mathbb{P}\Big(\hat{T}=0\Big) \right),$$

where the infimum is taken over all measurable test functions $\hat{T}(X_1,\dots,X_n)$. We are interested in necessary and sufficient "separation threshold" $\rho_n^*(\delta)$ that yields minimax risk of order δ.

Under just two finite moments, it is well known that in order to achieve minimax risk of order δ, the necessary and sufficient separation $\rho_n^*(\delta)$ is of order $\rho_n^*(\delta) := \sigma_* \sqrt{\log(1/\delta)/n}$ as long as $n = \Omega(\log(1/\delta))$. An optimal testing procedure is given by

$$\hat{T} = \mathbf{1}\{|\mathrm{MoM}(X_1, \ldots, X_n)| \geq 4\sigma_* \sqrt{\log(1/\delta)/n}\},$$

where $\mathrm{MoM}(X_1, \ldots, X_n)$ is the median-of-means estimator [Nemirovskii and Yudin, 1983]. *This fact shows that testing the mean under two finite moments is not more difficult than testing under Gaussian noise*. Unfortunately, this testing procedure requires prior knowledge of σ_*. Our next goal is to understand the required separation for testing when σ_* is unknown.

In order to characterize adaptive testing, we need to define the risk differently. The following definition of the minimax risk is inspired by the work Luo and Gao [2024]. Specifically, let

$$\mathcal{R}(n, \rho_n, \sigma_*) := \inf_{\hat{T}} \left(\sup_{\mu=0, \sigma>0, P_\xi \in \mathcal{P}^{(2)}} \mathbb{P}\Big(\hat{T} = 1\Big) + \sup_{|\mu| \geq \rho_n, \sigma=\sigma_*, P_\xi \in \mathcal{P}^{(2)}} \mathbb{P}\Big(\hat{T} = 0\Big) \right),$$

Here, the Type I error is uniform over $\sigma > 0$ in order to allow for adaptation, while the Type II error is given for a fixed value $\sigma = \sigma_*$. This definition of the risk allows us to capture the optimal adaptive separation ρ_n as a function of σ_*.

The asymmetric treatment of the scale in this definition is intentional. Type-I error is controlled uniformly over $\sigma > 0$, since validity must hold without prior knowledge of the scale, whereas Type-II error is evaluated at a fixed $\sigma = \sigma_*$, since the target interval length itself depends on the true scale. Up to immaterial constants in δ, this corresponds to requiring a level δ test with power at least $1 - \delta$, and is precisely the testing criterion obtained from the desired properties of adaptive confidence intervals through test/confidence-set duality.

Proposition 3.1. *For all values of $\rho_n \geq 0$, we have $\mathcal{R}(n, \rho_n, \sigma_*) = 1$.*

Proposition 3.1 implies that we can't test the mean when σ is unknown, i.e. $\rho_n^*(\delta) = \infty$. The proof of this result is based on a simple construction of two distributions both of which have a large "mass" at 0 so that they cannot be distinguished based on a sample of size n. This "all-or-nothing" risk reflects the impossibility of testing when σ is entirely unknown. However, by restricting to the class $\mathcal{P}_\alpha$ we recover meaningful results. Indeed, if we define for each δ and $\gamma_n > \alpha_n$

$$\begin{aligned}
&\mathcal{R}(n, \rho_n, \sigma_*, \alpha_n, \gamma_n, \delta) \\
&\qquad := \inf_{\hat{T}} \left(\sup_{\mu=0, \sigma>0, P_\xi \in \mathcal{P}_{\alpha_n}(n,\delta)} \mathbb{P}\Big(\hat{T} = 1\Big) + \sup_{|\mu| \geq \rho_n, \sigma=\sigma_*, P_\xi \in \mathcal{P}_{\gamma_n}(n,\delta)} \mathbb{P}\Big(\hat{T} = 0\Big) \right),
\end{aligned}$$

then we get the following statement.

Proposition 3.2. *For each value of δ such that $n\alpha_n = \Omega(\log(1/\delta))$ and $\gamma_n > \alpha_n$, we have that*

$$\mathcal{R}(n, \rho_n, \sigma_*, \alpha_n, \gamma_n, \delta) \leq \delta \text{ iff } \rho_n = \Omega\left(\sigma_* \sqrt{\frac{\log(1/\delta)}{\alpha_n n}} \right).$$

In that case we show (via explicitly constructing a test function) that in order to achieve a testing error of order δ, the optimal separation must be of order $\rho_n^* := \sigma_* \sqrt{\frac{\log(1/\delta)}{n\alpha_n}}$. Note that the bound is uniform over $\sigma > 0$. Observe also that we recover the impossibility result as $n\alpha_n \underset{n\to\infty}{\to} 0$. In other words, in order to get shrinking confidence intervals, we need to ensure that $n\alpha_n \underset{n\to\infty}{\to} \infty$. Finally, to understand the role of parameter γ_n, we refer the reader to the previously stated Remark 2.1 (c).

Duality between existence of tests and confidence sets immediately implies the following lemma.

Lemma 3.1. *For any ρ_n such that that $\mathcal{R}(n, \rho_n, \sigma_*, \alpha_n, \gamma_n, \delta) \leq \delta$, there exists a confidence interval with coverage probability at least $1 - \delta$ and length at most ρ_n, also with probability at least $1 - \delta$. Alternatively, if we can find an adaptive confidence interval of level δ such that its length is uniformly less than ρ_n with probability at least $1 - \delta$, then $\mathcal{R}(n, \rho_n, \sigma_*, \alpha_n, \gamma_n, \delta) \leq \delta$.*

4 On the construction of optimal adaptive confidence intervals

For this section, we assume that $n\alpha_n = \Omega(\log(1/\delta))$. Notably, this condition is not significantly more restrictive than $n = \Omega(\log(1/\delta))$, as emphasized earlier, since $n\alpha_n$ can be interpreted as the "effective sample size" derived from the data.

4.1 An estimator of the variance

To construct an estimator of the variance, we need an initial estimator of the mean. To this end we choose an estimator $\hat{\mu}$ of μ such that $|\hat{\mu}-\mu| \leq 10\sigma\sqrt{\frac{\log(1/\delta)}{n}}$ with probability $1-\delta$. While there are several viable candidates in the literature, we choose the most well-known and arguably the simplest median of means (MoM) estimator [Nemirovskii and Yudin, 1983]. In this section, we assume a sample of size n is provided to construct $\hat{\mu}$ and $\hat{\sigma}$. By avoiding sample splitting, we utilize the full sample for both estimators and maintain maximum efficiency. Next, we consider the estimator of σ^2 constructed as follows.

Definition. *Let $X_1, \ldots, X_n$ be a sequence of i.i.d copies of a random variable X with mean μ and finite variance σ^2. We partition the index set $[n]$ into K disjoint subsets $(A_i^n)_{i\in[K]}$, where each subset A_i^n contains exactly m elements, i.e.,*

$$(A_i^n)_{i\in[K]} \subset \{J \subset [n] \mid |J| = m\}.$$

The parameter m satisfies $m \leq \lfloor n/K \rfloor$, and K is an integer such that $K \leq \lfloor n/2 \rfloor$. For each subset A_i^n, we define $\hat{\sigma}_i^2$ to be the usual empirical variance estimator of σ^2 for a given estimator $\hat{\mu}$ of the mean μ, that is,

$$\forall i \in [K], \quad \hat{\sigma}_i^2 := \frac{1}{m}\sum_{j\in A_i^n}(X_j - \hat{\mu})^2.$$

Finally, we define the median-of-means estimator via

$$(\hat{\sigma}_{\mathrm{MoM}}^K)^2 := \mathrm{med}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_K^2).$$

Proposition 4.1. *Assume that $n\alpha_n = \Omega(\log(1/\delta))$ and that $P_\xi \in \mathcal{P}_{\alpha_n}(n,\delta)$. Let $\hat{\mu}$ be an estimator of μ satisfying $|\hat{\mu}-\mu| \leq 10\sigma\sqrt{\frac{\log(1/\delta)}{n}}$ with probability at least $1-\delta$. Then the event*

$$\mathbf{O} := \left\{\frac{\alpha_n}{4} \leq \frac{(\hat{\sigma}_{\mathrm{MoM}}^K)^2}{\sigma^2} \leq 10\right\}$$

occurs with probability at least $1-\delta/2$ whenever $K = C\log(1/\delta)$ for an absolute constant $C > 0$.

Remark 4.1. *Note that only the lower bound in the inequality depends on α_n. It means that we can construct a* lower *bound for σ^2 that is valid uniformly over $\mathcal{P}_2$. However, the upper bound requires the lower bound for α_n.*

The estimator of the variance constructed above is actually minimax optimal in the sense that we cannot get better estimates of the variance within the class $\mathcal{P}_{\alpha_n}$. The next result proposes a matching lower bound.

Proposition 4.2. *Assume that $n\alpha_n = \Omega(\log(1/\delta))$. Then there exists an absolute constant $c > 0$ such that no estimator of the variance $\tilde{\sigma}^2$ can satisfy the inequality*

$$\inf_{\sigma>0,\mu\in\mathbb{R},P_\xi\in\mathcal{P}_{\alpha_n}(n,\delta)} \mathbb{P}\left(c\alpha_n \leq \tilde{\sigma}^2/\sigma^2 \leq 1\right) \geq 1-\delta.$$

Combining both results provides minimax optimal confidence intervals for estimation of the variance σ^2. Notice that the optimal estimator $(\hat{\sigma}_{\mathrm{MoM}}^K)^2$ can be employed beyond scenarios where α_n is a constant bounded away from 0.

4.2 An optimal confidence interval

For $\delta > 0$ and $K = C\log(1/\delta)$, let $\hat{\mu}(\delta)$ and $\hat{\sigma}_{\mathrm{MoM}}^K$ be the estimators defined above. Consider the confidence interval

$$\hat{I}(\delta,\alpha_n) = \left[\hat{\mu}(\delta) - 5\hat{\sigma}_{\mathrm{MoM}}^K\sqrt{\frac{\log(1/\delta)}{\alpha_n n}}, \hat{\mu}(\delta) + 5\hat{\sigma}_{\mathrm{MoM}}^K\sqrt{\frac{\log(1/\delta)}{\alpha_n n}}\right].$$

Then $\hat{I}(\delta, \alpha_n)$ leads to a minimax optimal test with respect to the risk $\mathcal{R}(n, \rho_n, \sigma_*, \alpha_n, \gamma_n, \delta)$: indeed, $\hat{T}$ defined via

$$\hat{T} = \mathbf{1}\{0 \notin \hat{I}(\delta)\},$$

achieves desired optimality. It is also clear that $|\hat{I}(\delta, \alpha_n)| \leq C\sigma\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}$ with probability at least $1 - \delta$. Unlike the problem of variance estimation, constructing $\hat{I}(\delta, \alpha_n)$ requires prior knowledge of the lower bound α_n. On the one hand, our result remains adaptive to the choice of $\alpha_{n,\delta}(\xi)$ (conditionally on having access to its lower bound). On the other hand, we showed that constructing adaptive confidence intervals is impossible without a specified lower bound on α_n. This clarifies the cost of adaptation: while we do not need to know the exact distribution P or the variance σ^2, the "effective sample size" $n\alpha_n$ dictates the fundamental limits of uncertainty quantification.

5 Numerical experiments

In this section, we provide a numerical illustration comparing our adaptive approach to the construction proposed by Devroye et al. [2016] that we shall denote "DLLO". We focus on a scenario where the "approximate symmetry" condition (2) required by their method fails, as formally established in Lemma B.2. To this end, we utilize the mixture distribution constructed in the proof of that lemma for different values of α, which serves as a stress test for uncertainty quantification under heavy tails and asymmetry. For a given confidence level $1 - \delta$, both procedures are calibrated by setting the number of blocks K of order $\log(1/\delta)$ to ensure the nominal coverage is met in the well-behaved regime. In the experiments that follow, we have fixed $\alpha_n = 0.25$ for the construction of our adaptive procedure, representing a robust middle-ground for distributions that deviate from symmetry but retain sufficient "effective" mass for variance estimation.

Calibration and baselines. All practical constants are calibrated once in the well-behaved regime $\alpha = 1$ and are subsequently kept fixed across all experiments. Specifically, we use $K = \lceil 7\log(3/\delta) \rceil$ blocks and the numerical constant $C = 3.7$. The lower bound α_n supplied to our procedure is also fixed once and for all at $\alpha_n = 0.25$. We then vary only the data-generating distribution, considering the imbalanced mixtures from Lemma B.2 with $\alpha \in \{1, 0.25, 0.1\}$. We compare our method with the DLLO procedure of Devroye et al. and with the Student-t interval as an asymptotic benchmark.

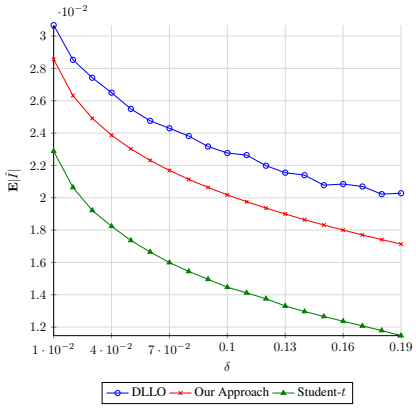
The right panels report empirical miscoverage probabilities; hence the dashed line $y = \delta$ represents the nominal boundary. In the well-behaved regime $\alpha = 1$, the robust procedures achieve nominal coverage. As α decreases, DLLO substantially exceeds the nominal miscoverage level and the Student-t interval also exhibits undercoverage, whereas our procedure remains below the nominal boundary. The same calibration constants are used throughout, so the improved coverage in the difficult regimes is not the result of regime-specific tuning.

Figures 1a and 1b illustrate the regime where $\alpha \approx 1$. In this case, both the mean interval length and the empirical coverage of our method are comparable to those achieved by the approach of DLLO. This suggests that our refinement does not incur a significant "robustness price" when the distribution is well-behaved. However, as we move away from the easy regime, the performance gap becomes evident.

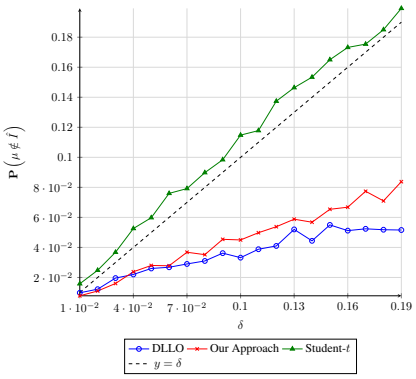
- **Moderate α:** for $\alpha = 0.25$ (Figures 1c and 1d), the method in Devroye et al. [2016] begins to exhibit a degradation in reliability.
- **Small α:** for $\alpha = 0.1$ (Figures 1e and 1f), the coverage of the standard approach drops significantly below the nominal level.

While our intervals are naturally wider to account for the increased difficulty of the problem, they consistently maintain the prescribed $1 - \delta$ coverage guarantee. This demonstrates that our method maintains nominal empirical coverage across the considered regimes, whereas non-adaptive methods fail to remain valid when the distribution's concentration properties are poor.

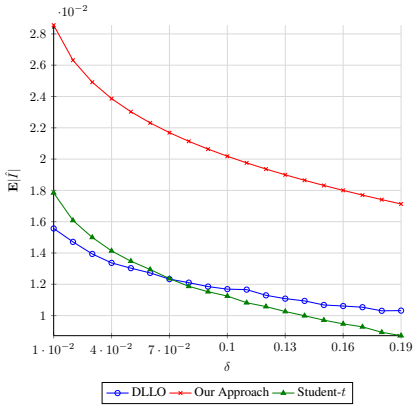
Remark 5.1. *Numerical experiments were performed using a standard laptop computer and did not require significant computational resources.*



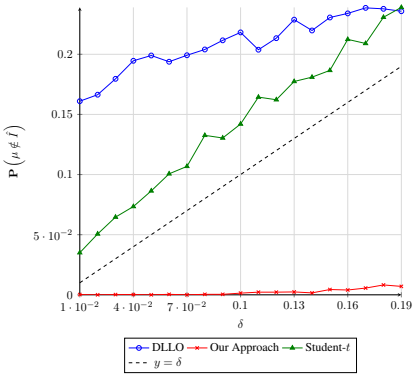
(a) Mean length: $\alpha = 1$.



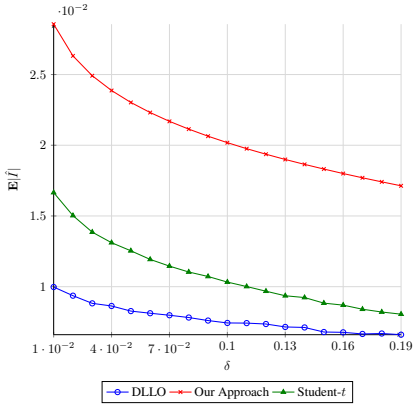
(b) Miscoverage: $\alpha = 1$.



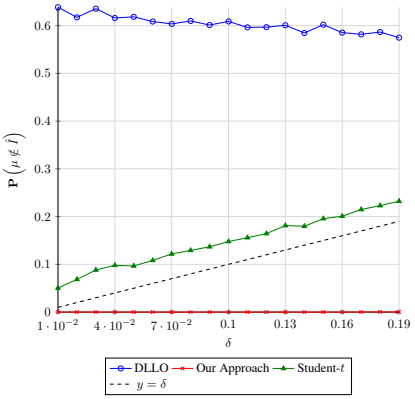
(c) Mean length: $\alpha = 0.25$.



(d) Miscoverage: $\alpha = 0.25$.



(e) Mean length: $\alpha = 0.1$.



(f) Miscoverage: $\alpha = 0.1$.

Figure 1: Comparison of mean interval length and empirical miscoverage for DLLO, our proposed adaptive approach, and the Student-t interval under the considered mixture distributions.

Final remarks and future directions

We showed that the difficulty of the problem of constructing sub-Gaussian non-asymptotic confidence intervals for the mean is driven by the "effective sample size". This parameter offers natural interpolation between the "easy" and "hard" cases of the problem. Moreover, we established that it is not possible to learn this effective sample size from the data alone, and that a known lower bound is required for constructing valid confidence intervals. It would be interesting to extend our results to the case of distributions and confidence sets in $\mathbb{R}^d$, and find the correct extension of the notion of "effective sample size" to this more general framework.

Acknowledgments and Disclosure of Funding

The work of M.N. was supported by France 2030. The work of S.M. was supported by the National Science Foundation grant DMS CAREER-2045068. The authors declare no competing interests.

References

Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

Ronald Fisher. Applications of "Student's"' distribution. *Metron*, 5:90–104, 1925.

Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380, 1937.

Mark G Low. On nonparametric confidence intervals. *The Annals of Statistics*, 25(6):2547–2554, 1997.

Richard Nickl and Sara van de Geer. Confidence sets in sparse regression. *The Annals of Statistics*, 41(6):2852–2876, 2013.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185. Institut Henri Poincaré, 2012.

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

Raghu R Bahadur and Leonard J Savage. The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics*, 27(4):1115–1122, 1956.

P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964.

Hongjian Wang and Aaditya Ramdas. Huber-robust confidence sequences. In *International Conference on Artificial Intelligence and Statistics*, pages 9662–9679. PMLR, 2023.

Amir-Hossein Bateni and Arnak S Dalalyan. Confidence regions and minimax rates in outlier-robust estimation on the probability simplex. *Electronic Journal of Statistics*, 14:2653–2677, 2020.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Shubhanshu Shekhar and Aaditya Ramdas. On the near-optimality of betting confidence sets for bounded means. *arXiv preprint arXiv:2310.01547*, 2023.

Yuetian Luo and Chao Gao. Adaptive robust confidence intervals. *arXiv preprint arXiv:2410.22647*, 2024.

Valentin V Petrov. *Sums of independent random variables*. Springer Science & Business Media, 2012.

Arkadij Nemirovskii and David Yudin. Problem complexity and method efficiency in optimization. 1983.

Evarist Giné, Friedrich Götze, and David M Mason. When is the student t-statistic asymptotically standard normal? *The Annals of Probability*, 25(3):1514–1531, 1997.

A Proofs of the main results

Everywhere below, $\mathbb{V}(X)$ denotes the variance of a random variable X.

A.1 Proof of Proposition 3.1

To show that the minimax risk $\mathcal{R}(n, \rho_n, \sigma_*) = 1$, it suffices to construct two distributions, P_0 and P_1, that are statistically indistinguishable based on a sample of size n, yet have different means. To this end, let $\tau \in (0, 1)$ be a small constant to be chosen later. Without loss of generality, we set σ_* to be 1 (the general case can be deduced via rescaling). Consider first the distribution P_0 (under $\mathbf{H}_0$) defined by

$$\begin{cases} P_0(X = -A) = 1 - \tau, \\ P_0(X = A/\tau - A) = \tau. \end{cases}$$

For $X \sim P_0$, the mean is $\mathbb{E}[X] = 0$. The variance is given by $\sigma_0^2 = A^2(1-\tau)/\tau$, which is finite for any $A > 0$. Therefore, $P_0 \in \mathcal{P}^{(2)}$. Now, consider the distribution P_1 defined via

$$\begin{cases} P_1(X = -A) = 1 - \tau, \\ P_1(X = 1/\sqrt{\tau} - A) = \tau/2, \\ P_1(X = -1/\sqrt{\tau} - A) = \tau/2. \end{cases}$$

It is easy to see that $\mathbb{E}[X] = -A$ and $\mathbb{V}(X) = 1$. Next, note that

$$\mathbb{P}_{X_1 \sim P_0}\left(X_1 = \cdots = X_n = -A\right) = \mathbb{P}_{X_1 \sim P_1}\left(X_1 = \cdots = X_n = -A\right) = (1-\tau)^n.$$

Therefore, for any test $\hat{T}(X_1, \ldots, X_n)$, using the Neyman-Pearson lemma, we get

$$\begin{aligned} &\mathbb{P}_{X_1 \sim P_0}\left(\hat{T} = 1\right) + \mathbb{P}_{X_1 \sim P_1}\left(\hat{T} = 0\right) \geq \int \min\left(dP_0^{\otimes n}, dP_1^{\otimes n}\right) \\ &\quad \geq \min\left(\mathbb{P}_{X_1 \sim P_0}\left(X_1 = \cdots = X_n = -A\right), \mathbb{P}_{X_1 \sim P_1}\left(X_1 = \cdots = X_n = -A\right)\right) = (1-\tau)^n. \end{aligned}$$

Choosing τ to be small enough, we can make the right-hand side arbitrarily close to 1.

Note that this simple construction demonstrates that without further restrictions on the distribution class, the presence of a dominant mass at the mean of a distribution makes adaptive uncertainty quantification impossible.

A.2 Proofs of Proposition 3.2 and Theorem 2.1

1. Proof of the lower bound. Assume that $\frac{\log(1/\delta)}{n} \leq \alpha_n \leq \frac{1}{8}$. To establish the lower bound, we will construct two distributions, P_0 and P_1, that are statistically indistinguishable under n observations but have sufficiently different means. Let $A = \frac{1}{2}\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}$ and define P_0 via

$$P_0 = \left(1 - \frac{\log(1/\delta)}{2n}\right)\mathcal{N}(A, 1) + \frac{\log(1/\delta)}{2n}\delta_{z_0}, \quad \text{where } z_0 = -\frac{1 - \frac{\log(1/\delta)}{2n}}{\frac{\log(1/\delta)}{2n}}A.$$

Let $X \sim P_0$. By construction, $\mathbb{E}(X) = 0$. A direct calculation of the variance yields that

$$\mathbb{V}(X) = \left(1 - \frac{\log(1/\delta)}{2n}\right)(A^2 + 1) + \frac{\log(1/\delta)}{2n}z_0^2 \simeq \frac{1}{\alpha_n}.$$

Let $\xi = X/\sqrt{\mathbb{V}(X)}$ be the standardized version. To show that $P_0 \in \mathcal{P}_{\alpha_n}$, we verify the effective mass condition. For $n\alpha_n \gg \log(1/\delta)$, we have that

$$\begin{aligned} \mathbb{E}\left[\xi^2 \mathbf{1}\left\{\xi^2 \leq \frac{n\alpha_n}{\log(1/\delta)}\right\}\right] &\geq \frac{1 - \frac{\log(1/\delta)}{2n}}{\mathbb{V}(X)}\mathbb{E}\left[Z^2 \mathbf{1}\left\{Z^2 \leq \frac{n\alpha_n \mathbb{V}(X)}{\log(1/\delta)}\right\}\right] \\ &\geq \frac{2\alpha_n}{1 + 2\alpha_n}\left(\mathbb{E}[Z^2] - \frac{\log(1/\delta)\mathbb{E}[Z^4]}{n\alpha_n \mathbb{V}(X)}\right) \geq \frac{2\alpha_n}{1 + 2\alpha_n}\left(1 - \frac{16\log(1/\delta)}{n}\right) \geq \frac{6}{5}\alpha_n \geq \alpha_n. \end{aligned}$$

where $Z \sim \mathcal{N}(A, 1)$ and c is a sufficiently small universal constant. Thus, $P_0 \in \mathcal{P}_{\alpha_n}$.

Next, consider the alternative distribution

$$P_1 = \left(1 - \frac{\log(1/\delta)}{2n}\right)\mathcal{N}(A, 1) + \frac{\log(1/\delta)}{2n}\delta_0.$$

For $X \sim P_1$, the mean satisfies $\mathbb{E}(X) = cA = c\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}$ for some $c > 0$, and since P_1 has finite variance, $P_1 \in \mathcal{P}^{(2)}$. We standardize X by setting $\zeta = (X - \mathbb{E}(X))/\sqrt{\mathbb{V}(X)}$, and it is straightforward to verify that ζ satisfies the effective mass condition $\alpha_{n,\delta}(\zeta) \geq c_2$ for some constant $c_2 > 0$ that does not depend on n and δ. Since $\alpha_{n,\delta}(\zeta) \geq c_2$ then $P_1 \in \mathcal{P}_{\gamma_n}$ for any $\gamma_n \leq c_2$ and in particular for $\gamma_n > \alpha_n$. Using the Neyman-Pearson lemma, the minimax risk is lower-bounded by the total variation distance between the product measures :

$$\mathcal{R}(n, \rho_n, \sigma_*, \alpha_n, \delta) \geq \int \min\left(dP_0^{\otimes n}, dP_1^{\otimes n}\right) \geq \mathbb{P}_{P_1}\left(\forall i \in [n], X_i \neq 0\right)$$
$$= \left(1 - \frac{\log(1/\delta)}{2n}\right)^n \geq e^{-\log(1/\delta)} = \delta.$$

This confirms that no test can distinguish P_0 and P_1 with risk better than δ, establishing that the separation condition $\rho_n = \Omega\left(\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}\right)$ is necessary.

2. Proof of the upper bound. Consider the test $\hat{T} := \mathbf{1}\left\{\left|\frac{\hat{\mu}}{\hat{\sigma}}\right| \geq 5\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}\right\}$. We recall that $|\hat{\mu}-\mu| \leq 10\sigma\sqrt{\frac{\log(1/\delta)}{n}}$ with probability at least $1-\delta/2$, and that $\hat{\sigma}$ satisfies the bound of Proposition 4.1.

Type I Error: Under $\mathbf{H}_0$ ($\mu = 0$), using the union bound and the concentration properties of $\hat{\sigma}$, we deduce that

$$\mathbb{P}_0(\hat{T} = 1) \leq \mathbb{P}_0\left(\left|\frac{\hat{\mu}}{\sigma}\right| \geq 10\sqrt{\frac{\log(1/\delta)}{n}}\right) + \mathbb{P}_0\left(\frac{\hat{\sigma}^2}{\sigma^2} \leq \frac{\alpha_n}{4}\right) \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

Type II Error: Under $\mathbf{H}_1$, assume that $|\mu| \geq 30\sigma\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}$. By triangle inequality,

$$\begin{aligned}
\mathbb{P}_1(\hat{T} = 0) &= \mathbb{P}_1\left(|\hat{\mu}| < 5\hat{\sigma}\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}\right) \\
&\leq \mathbb{P}_1\left(|\mu| - |\hat{\mu} - \mu| < 5\hat{\sigma}\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}\right) \\
&\leq \mathbb{P}_1\left(|\hat{\mu} - \mu| > 30\sigma\sqrt{\frac{\log(1/\delta)}{n\alpha_n}} - 5\hat{\sigma}\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}\right).
\end{aligned}$$

On the event $\left\{\hat{\sigma}^2 \leq 10\sigma^2\right\}$, we have that $5\hat{\sigma} \leq 5\sqrt{10}\sigma \leq 16\sigma$. Therefore,

$$\mathbb{P}_1(\hat{T} = 0) \leq \mathbb{P}\left(|\hat{\mu} - \mu| > 14\sigma\sqrt{\frac{\log(1/\delta)}{n\alpha_n}}\right) + \mathbb{P}\left(\hat{\sigma}^2 > 10\sigma^2\right) \leq \delta.$$

The result follows. Observe that as $\alpha_n \to 0$, we recover the impossibility result of Proposition 3.1.

A.3 Proof of Proposition 4.1

Let $\mathcal{A}$ be the event $\{|\hat{\mu} - \mu| \leq 10\sigma\sqrt{\frac{\log(1/\delta)}{n}}\}$, which holds with probability at least $1 - \delta$ by assumption. On $\mathcal{A}$, we have that

$$(\hat{\mu} - \mu)^2 \leq 100\sigma^2\frac{\log(1/\delta)}{n}.$$

We aim to bound the probability that the median-of-means estimator $\hat{\sigma}^2_{\mathrm{MoM}}$ significantly deviates from the true variance σ^2. By the definition of the median, this occurs only if more than half of the block estimators $\hat{\sigma}^2_k$ turn out to be "bad". Specifically, we wish to show that

$$\mathbb{P}\left(\frac{\alpha_n}{4} \leq \frac{\hat{\sigma}^2_{\mathrm{MoM}}}{\sigma^2} \leq 10\right) \geq 1 - 2\delta.$$

Recall that for any block $k \in [K]$, the estimator is $\hat{\sigma}^2_k = \frac{1}{m}\sum_{j\in A_k}(X_j - \hat{\mu})^2$. Using the identity $(X_j - \hat{\mu})^2 \leq 2(X_j - \mu)^2 + 2(\hat{\mu} - \mu)^2$ (and similar lower bounds), we can relate the block estimators to the oracle centered sums $S_k := \frac{1}{m}\sum_{j\in A_k}\xi_j^2$, where $\xi_j = (X_j - \mu)/\sigma$. On event $\mathcal{A}$, the condition $\frac{\alpha_n}{4}\sigma^2 \leq \hat{\sigma}^2_k \leq 10\sigma^2$ is satisfied if

$$\frac{\alpha_n}{2} + 200\frac{\log(1/\delta)}{n} \leq S_k \leq 5 - 100\frac{\log(1/\delta)}{n}.$$

Under the assumption $n\alpha_n \geq C\log(1/\delta)$, we can simplify these thresholds to $S_k \in [\frac{2}{3}\alpha_n, 4]$. By the union bound, the probability that the median fails is bounded from above by

$$\mathbb{P}(\mathcal{A}^c) + \mathbb{P}\left(\sum_{k=1}^K \mathbf{1}\{S_k > 4\} \geq \frac{K}{4}\right) + \mathbb{P}\left(\sum_{k=1}^K \mathbf{1}\{S_k < \frac{2}{3}\alpha_n\} \geq \frac{K}{4}\right).$$

1. Upper bound (concentration of the mean). Using Markov's inequality and the fact that $\mathbb{E}[S_k] = 1$, we see that $\mathbb{P}(S_k > 4) \leq 1/4$. However, to deduce a tighter bound, we assume n is large enough such that $\mathbb{P}(S_k > 4) \leq 1/8$. Applying Hoeffding's inequality to the sum of indicators, we obtain that

$$\mathbb{P}\left(\sum_{k=1}^K \mathbf{1}_{\{S_k>4\}} \geq \frac{K}{4}\right) \leq \exp\left(-\frac{K}{150}\right) \leq \delta,$$

provided that $K \geq 150\log(1/\delta)$.

2. Lower bound. To control the lower tail, we consider the truncated variables $\bar{\xi}_j^2 = \xi_j^2\mathbf{1}_{\{\xi_j^2 \leq n\alpha_n/\log(1/\delta)\}}$. By the definition of α_n, we have $\mathbb{E}[\bar{\xi}_j^2] = \alpha_n$. We need to show that $\mathbb{P}(S_k < \frac{2}{3}\alpha_n) \leq 1/8$. Note that $\mathbb{V}(\bar{\xi}_j^2) \leq \mathbb{E}[\bar{\xi}_j^4] \leq \frac{n\alpha_n}{\log(1/\delta)}\mathbb{E}[\bar{\xi}_j^2]$. Applying Bernstein's inequality to the sum of m i.i.d. truncated variables, we see that

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{m}\sum_{j=1}^m \bar{\xi}j^2 < \frac{2}{3}\alpha_n\right) &\leq \mathbb{P}\left(\frac{1}{m}\sum_{j=1}^m \bar{\xi}_j^2 - \mathbb{E}[\bar{\xi}_j^2] < -\frac{\alpha_n}{3}\right) \\
&\leq \exp\left(-\frac{m\alpha_n^2/18}{\mathbb{V}(\bar{\xi}_j^2) + \frac{n\alpha_n}{3\log(1/\delta)}\frac{\alpha_n}{3}}\right) \\
&\leq \exp\left(-\frac{m\alpha_n^2}{C'\frac{n\alpha_n^2}{\log(1/\delta)}}\right) = \exp\left(-\frac{m\log(1/\delta)}{C'n}\right).
\end{aligned}$$

Since $m = n/K$, this probability is bounded by a constant (e.g., $1/8$) as long as K is of order $\log(1/\delta)$. Finally, applying Hoeffding's inequality again for the K blocks ensures that the number of blocks failing the lower threshold is less than $K/4$ with probability $1-\delta$. Combining these results yields the claim.

A.4 Proof of Proposition 4.2

Assume that $\frac{\log(1/\delta)}{n} \leq \alpha_n \leq \frac{1}{4}$. We will establish the lower bound by constructing two triples (μ_0, σ_0, P_0) and (μ_1, σ_1, P_1) such that the resulting data distributions are nearly indistinguishable, yet their variances differ by a factor of α.

1. Construction of P_0. Let $A := \sqrt{\frac{\log(1/\delta)}{n}}$ and define P_0 as the distribution of ξ via the mixture $(1-\tau)\mathcal{N}(A, \alpha_n) + \tau\delta_{z_0}$, where $\tau = \frac{\log(1/\delta)}{n}$ and z_0 is chosen such that $\mathbb{E}(\xi) = 0$. Simple algebra

shows that $z_0 = -\frac{1-\frac{\log(1/\delta)}{n}}{\frac{\log(1/\delta)}{n}} A$. A direct calculation yields that $\mathbb{V}(\xi) = (1-\tau)(1-\tau+A^2+\alpha_n)$. Hence, $\mathbb{V}(\xi)$ is of order 1. To verify the effective mass condition, let $Z \sim \mathcal{N}(A, \alpha_n)$. Then

$$\mathbb{E}[\xi^2 \mathbf{1}_{\{\xi^2 \le \frac{n\alpha_n}{\log(1/\delta)}\}}] \ge (1-\tau)\mathbb{E}[Z^2 \mathbf{1}_{\{Z^2 \le \frac{n\alpha_n}{\log(1/\delta)}\}}] \ge \alpha_n - C\frac{\log(1/\delta)}{n} \ge \frac{\alpha_n}{2},$$

where we used the fact that $A^2 = \log(1/\delta)/n$. Thus, $\xi \in \mathcal{P}_{c_1\alpha_n}$ for some constant c_1.

2. Construction of P_1. Let P_1 be the mixture $\sim (1-\tau)\mathcal{N}(A, \alpha_n) + \tau\delta_0$. For this distribution, the mean is $\mathbb{E}(X) = (1-\tau)A$ and the variance $\mathbb{V}(X)$ is of order α_n. We standardize X by setting $\zeta = (X - \mathbb{E}(X))/\sqrt{\mathbb{V}(X)}$, and it is straightforward to verify that ζ satisfies the effective mass condition $\alpha_{n,\delta}(\zeta) \ge c_2$ for some constant $c_2 > 0$.

3. Indistinguishability argument. We will now compare the product measures $Q_0 = P_0^{\otimes n}$ (with $\sigma^2 = 1$) and $Q_1 = P_1^{\otimes n}$ (with σ^2 of order α). Any estimator $\tilde{\sigma}^2$ must satisfy the coverage requirement for both. Let $c_3 = \min(c_1, c_2)$. The minimax error is bounded below by the testing error between Q_0 and Q_1, namely

$$\sup_{P \in \mathcal{P}_{P\alpha_n}} \mathbb{P}\left(\left(c_3\alpha_n \le \frac{\tilde{\sigma}^2}{\sigma^2} \le 1\right)^c\right) \ge \frac{1}{2}\left[\mathbb{P}_{Q_0}\left(\tilde{\sigma}^2 < c_3\alpha_n \text{ or } \tilde{\sigma}^2 > 1\right) + \mathbb{P}_{Q_1}\left(\frac{\tilde{\sigma}^2}{\alpha_n} < c_3\alpha_n \text{ or } \frac{\tilde{\sigma}^2}{\alpha_n} > 1\right)\right].$$

Note that the intervals $[c_3\alpha_n, 1]$ and $[c_3\alpha - n^2, \alpha_n]$ are disjoint for sufficiently small α_n. Consequently, for any $\tilde{\sigma}^2$, the probability of error is at least the probability that the two distributions cannot be distinguished. Applying the same total variation argument as in Proposition 3.1, we see that

$$\mathcal{E} \;\ge\; \frac{1}{2}\int \min(dQ_0, dQ_1) \;\ge\; \frac{1}{2}\mathbb{P}Q_1(X_1, \ldots, X_n \;\neq\; 0) \;=\; \frac{1}{2}\left(1 - \frac{\log(1/\delta)}{n}\right)^n \;\ge\; \frac{\delta}{2}.$$

This implies that no estimator can consistently distinguish between a variance of order 1 and a variance of order α_n across the class $\mathcal{P}_{\alpha_n}$, completing the proof.

A.5 Extension to finite moment of order $1 < p < 2$

Fix $\beta \in (0, 1]$ and write $p = 1 + \beta$. Given a random variable X with mean μ and p-th central moment

$$v_p := \left(\mathbb{E}|X - \mu|^p\right)^{1/p} \in (0, \infty),$$

define $\xi = \xi(X) := \frac{X-\mu}{v_p}$ and

$$\alpha_{n,\delta}^{(\beta)}(\xi) := \sup\left\{a \in [0,1] : \mathbb{E}\left[|\xi|^p \mathbf{1}\left\{|\xi|^p \le \frac{na}{\log(1/\delta)}\right\}\right] \ge a^\beta\right\}.$$

For $\alpha \in (0, 1]$, let $\mathcal{P}_\alpha^{(\beta)}(n, \delta)$ denote the class of distributions satisfying $\alpha_{n,\delta}^{(\beta)}(\xi) \ge \alpha$. For $\beta = 1$, this definition coincides with the effective-mass parameter used in the main text.

Generalized result. Up to constants depending only on β, the conclusions of Theorem 2.1 remain valid with σ replaced by v_p and with the rate

$$v_p\left(\frac{\log(1/\delta)}{n\alpha}\right)^{\frac{\beta}{1+\beta}}.$$

More precisely, if $n\alpha \ge C_\beta \log(1/\delta)$, there exists a confidence interval $\widehat{C}_n^{(\beta)}$ such that

$$\inf_{P \in \mathcal{P}_\alpha^{(\beta)}(n,\delta)} \mathbb{P}\Big(\mu \in \widehat{C}_n^{(\beta)}\Big) \ge 1 - \delta$$

and, with probability at least $1 - \delta$,

$$\left|\widehat{C}_n^{(\beta)}\right| \le C_\beta v_p\left(\frac{\log(1/\delta)}{n\alpha}\right)^{\frac{\beta}{1+\beta}}.$$

Conversely, this rate cannot be improved, up to constants depending only on β. The same lower-bound construction also yields the analogue of the non-adaptivity statement in Theorem 2.1.

Proof: upper bound. Set $L = \log(1/\delta)$ and let $K \asymp_\beta L$. Partition the observations into K blocks of size $m \asymp n/K$. A standard median-of-means estimator $\widehat{\mu}$ satisfies

$$|\widehat{\mu} - \mu| \le C_\beta v_p \left(\frac{L}{n}\right)^{\beta/p}$$

with probability at least $1 - \delta/3$. Indeed, if $\overline{X}_B$ denotes the empirical mean on a block, the standard L_p inequality for sums of independent centered random variables, valid for $1 < p \le 2$, gives

$$\mathbb{E}|\overline{X}_B - \mu|^p \le C_\beta m^{1-p} \mathbb{E}|X - \mu|^p = C_\beta v_p^p m^{-\beta}.$$

Markov's inequality gives a fixed-probability bound on each block, and median aggregation amplifies it to probability $1 - \delta/3$.

We next estimate the p-th central moment. For every block B_k, set

$$\widehat{v}_k^p := \frac{1}{m} \sum_{j \in B_k} |X_j - \widehat{\mu}|^p, \qquad \widehat{v}_{\mathrm{MoM}}^p := \mathrm{med}\big(\widehat{v}_1^p, \dots, \widehat{v}_K^p\big).$$

Let

$$S_k := \frac{1}{m} \sum_{j \in B_k} |\xi_j|^p, \qquad e := \frac{\widehat{\mu} - \mu}{v_p}.$$

Since $p \in (1, 2]$,

$$2^{1-p} S_k - |e|^p \le \frac{\widehat{v}_k^p}{v_p^p} \le 2^{p-1}\big(S_k + |e|^p\big). \tag{A.1}$$

The upper tail of S_k is controlled directly by Markov's inequality, because $\mathbb{E}|\xi|^p = 1$.

For the lower tail, take a witness $a_\star \ge \alpha/2$ in the definition of $\alpha_{n,\delta}^{(\beta)}(\xi)$ and define

$$Y := |\xi|^p \mathbf{1}\left\{|\xi|^p \le \frac{n a_\star}{L}\right\}.$$

Then

$$q := \mathbb{E}Y \ge a_\star^\beta, \qquad 0 \le Y \le \frac{n a_\star}{L},$$

and hence

$$\mathbb{V}(Y) \le \mathbb{E}Y^2 \le \frac{n a_\star}{L} \mathbb{E}Y.$$

Bernstein's inequality therefore gives

$$\begin{aligned} \mathbb{P}\left(\frac{1}{m} \sum_{j \in B_k} Y_j < \frac{q}{2}\right) &\le \exp\left(-c_\beta \frac{mq}{n a_\star / L}\right) \\ &\le \exp\left(-c_\beta \frac{mL}{n} a_\star^{\beta - 1}\right) \le \exp\left(-c_\beta \frac{mL}{n}\right), \end{aligned} \tag{A.2}$$

where the last inequality uses $\beta \le 1$ and $a_\star \le 1$. Choosing the constants in $K \asymp_\beta L$ so that the blockwise success probability is bounded away from $1/2$, and applying the usual median amplification argument, yields

$$c_\beta \alpha^\beta \le \frac{\widehat{v}_{\mathrm{MoM}}^p}{v_p^p} \le C_\beta \tag{A.3}$$

with probability at least $1 - \delta/3$. To use (A.1), note that on the event controlling $\widehat{\mu}$,

$$|e|^p \le C_\beta \left(\frac{L}{n}\right)^\beta \le c_\beta \alpha^\beta$$

whenever $n\alpha \ge C_\beta L$.

Combining the mean and scale bounds, define

$$\widehat{C}_n^{(\beta)} := \left[\widehat{\mu} \pm C_\beta \widehat{v}_{\mathrm{MoM}} \left(\frac{L}{n\alpha}\right)^{\beta/p}\right].$$

The lower bound in (A.3) implies $v_p \le C_\beta \alpha^{-\beta/p} \widehat{v}_{\mathrm{MoM}}$, so the preceding deviation inequality for $\widehat{\mu}$ gives coverage at least $1 - \delta$. The upper bound in (A.3) gives the stated bound on the interval length.

Proof: matching lower bound. We adapt the mixture construction used in the proof of Theorem 2.1. Let

$$\tau := \frac{L}{2n}, \qquad A := \left(\frac{\tau}{\alpha}\right)^{\beta/p},$$

and let ε be a Rademacher random variable. Consider the common dominant component

$$Q := \mathcal{L}(A + \varepsilon),$$

where $\mathcal{L}$ denotes the law of a random variable. This construction is analogous to the common Gaussian component used in the proof of Theorem 2.1. Define

$$P_0 = (1-\tau)Q + \tau\delta_{z_0}, \qquad z_0 := -\frac{1-\tau}{\tau}A,$$

and

$$P_1 = (1-\tau)Q + \tau\delta_0.$$

Then $\mathbb{E}_{P_0} X = 0$, whereas

$$\mu_1 := \mathbb{E}_{P_1} X = (1-\tau)A \asymp A.$$

Let $v_{0,p}$ and $v_{1,p}$ denote the corresponding p-th central scales. Since $\tau \le \alpha$ under $n\alpha \ge C_\beta L$, we have $A \le 1$, and a direct calculation gives

$$v_{0,p}^p \asymp_\beta \tau|z_0|^p \asymp_\beta A^p \tau^{1-p} = \alpha^{-\beta}, \qquad v_{1,p}^p \asymp_\beta 1. \tag{A.4}$$

After standardization, the dominant component of P_0 contributes a p-th moment of order α^β. Taking $a_0 = c_\beta \alpha$ in the definition of $\alpha_{n,\delta}^{(\beta)}$, and choosing $c_\beta > 0$ small enough, the truncation threshold na_0/L contains the whole dominant component, while

$$\mathbb{E}_{P_0}\left[|\xi_0|^p \mathbf{1}\left\{|\xi_0|^p \le \frac{na_0}{L}\right\}\right] \ge c_\beta \alpha^\beta \ge a_0^\beta.$$

Hence $\alpha_{n,\delta}^{(\beta)}(\xi_0) \ge c_\beta \alpha$. For P_1, the scale is of constant order and the centered dominant component is bounded; the same argument with a fixed sufficiently small $a_1 > 0$ gives

$$\alpha_{n,\delta}^{(\beta)}(\xi_1) \ge a_1,$$

uniformly in n, δ, and α. Thus the null belongs to the hard class of order α, whereas the alternative belongs to an easier class bounded away from zero.

Finally, P_0 and P_1 share the sub-probability measure $(1-\tau)Q$. Therefore,

$$\int \min\{dP_0^{\otimes n}, dP_1^{\otimes n}\} \ge (1-\tau)^n.$$

Since $\tau = L/(2n)$ and $L \le n$, $\log(1-x) \ge -2x$ for $0 \le x \le 1/2$ yields

$$(1-\tau)^n \ge e^{-L} = \delta.$$

By the Neyman–Pearson lemma, every test between the two experiments has total error at least δ. On the other hand, by (A.4),

$$\frac{|\mu_1 - \mu_0|}{v_{1,p}} \asymp_\beta A = \left(\frac{L}{n\alpha}\right)^{\beta/(1+\beta)}.$$

Rescaling the construction gives any prescribed value of $v_{1,p}$. Consequently, no uniformly valid confidence interval can have length of smaller order than

$$v_p\left(\frac{\log(1/\delta)}{n\alpha}\right)^{\beta/(1+\beta)}.$$

Since the alternative has effective mass bounded away from zero, the same pair also gives the analogue of the non-adaptivity conclusion of Theorem 2.1. This completes the proof.

B Auxiliary results

It is well known that when X is Gaussian we can achieve the optimal deviations bounds for the confidence interval around the empirical mean $\bar{X}_n$ such that $\forall \delta > 0$, we have

$$\mathbf{P}\left(|\bar{X}_n - \mathbf{E}(X)| \geq s_n \sqrt{\frac{2(1+o(1))\log(1/\delta)}{n}}\right) = \delta,$$

where s_n denotes the empirical standard deviation.

Lemma B.1. *Let $X_1, \ldots, X_n$ be i.i.d. with normal distribution $\mathcal{N}(\mu, \sigma^2)$. Then for all $e^{-n/2} \leq \delta \leq e^{-1}$,*

$$\mathbb{P}\left(|\bar{X}_n - \mathbf{E}(X)| \geq s_n \sqrt{\frac{6\log(2/\delta)}{n}}\right) = \delta$$

Moreover, the bound holds with $\sqrt{6\log(2/\delta)}$ replaced by $\sqrt{2(1+o(1)\log(1/\delta)}$ where $o(1) \to 0$ as $n \to \infty$ for each $0 < \delta \leq 1$.

Proof. The probability density function of the Student's t-distribution with $n-1$ degrees of freedom is

$$p_n(x) = c_n \left(1 + \frac{x^2}{n-1}\right)^{-n/2}$$

where $c_n = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\Gamma\left(\frac{n-1}{2}\right)}$. Therefore, letting Z have t-distribution with $n-1$ d.f., we deduce that

$$\mathbb{P}(Z \geq t) = \int_t^\infty p_n(x)dx \leq \int_t^\infty \frac{x}{t} c_n \left(1 + \frac{x^2}{n-1}\right)^{-n/2} \leq \frac{c_n}{t}\frac{n-1}{n-2}\left(1 + \frac{t^2}{n-1}\right)^{-(n-2)/2},$$

where we used the fact that $\frac{x}{t} \geq 1$ in the region of integration. It remains to simplify the resulting expression using the inequality $\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \leq \sqrt{n/2}$ (Gautschi's inequality) so that

$$c_n \frac{n-1}{n-2} \leq \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{n-1}}\frac{n-1}{n-2} \leq \sqrt{\frac{16}{15\pi}}$$

for $n \geq 5$. Moreover, since $(1+z)^{-p} \leq e^{-pz/2}$ for all $p > 0$ and $0 \leq z \leq \frac{1}{2}$,

$$\left(1 + \frac{t^2}{n-1}\right)^{-(n-2)/2} \leq e^{-t^2/6}$$

for $n \geq 6$. The same calculation yields a bound for the lower tail, hence the result follows by setting $t = \sqrt{6\log(2/\delta)}$.

The second part of the claim stems from the well-known fact that the Student's t-statistic is asymptotically normally distributed [Giné et al., 1997]. □

Lemma B.2. *For α small enough, there exists a distribution $\xi \in \mathcal{P}_\alpha$ such that $\min(p_-(j), p_+(j)) \leq 1/4$.* [3]

Proof. Assume $\log(1/\delta)/n \leq \alpha \leq 1/4$. We aim to show that for sufficiently small α, there exists a distribution in $\mathcal{P}_\alpha$ that is highly asymmetric, such that the sample mean remains on one side of the true mean with high probability. Consider the distribution P_0 defined as in the proof of Proposition 4.2:

$$P_0 = (1-\tau)\mathcal{N}(A, \alpha) + \tau\delta_{z_0},$$

where $\tau = \frac{\log(1/\delta)}{n}$, $A = \sqrt{\frac{\log(1/\delta)}{n}}$, and $z_0 = -\frac{1-\tau}{\tau}A$. As established previously, $P_0 \in \mathcal{P}_\alpha$ and $\mathbb{E}(\xi) = 0$.

[3] It is clear from the proof that $\frac{1}{4}$ can be replaced by any positive constant.

Let $\xi_1, \ldots, \xi_m$ be i.i.d. copies of $\xi \sim P_0$, where $m = \lfloor \frac{n}{10\log(1/\delta)} \rfloor$. We can represent each ξ_i using an auxiliary Bernoulli variable $B_i \sim \text{Bern}(1-\tau)$:

$$\xi_i = B_i Z_i + (1 - B_i) z_0, \quad \text{where } Z_i \sim \mathcal{N}(A, \alpha).$$

Now, consider the event $\mathcal{E} = \{B_1 = \cdots = B_m = 1\}$, which corresponds to the case where no "outlier" from the δ_{z_0} component is sampled. The probability of this event is:

$$\mathbb{P}(\mathcal{E}) = (1-\tau)^m = \left(1 - \frac{\log(1/\delta)}{n}\right)^{\frac{n}{10\log(1/\delta)}} \geq \exp(-1/10) \approx 0.904.$$

On the event $\mathcal{E}$, the sample mean $\bar{\xi}_m$ is simply the empirical mean of the Gaussian components Z_i. Thus:

$$\begin{aligned} p_+(m) = \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m \xi_i \geq 0\right) &\geq \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m Z_i \geq 0 \,\middle|\, \mathcal{E}\right)\mathbb{P}(\mathcal{E}) \\ &\geq \mathbb{P}\left(A + \sqrt{\frac{\alpha}{m}}\mathcal{N}(0,1) \geq 0\right) e^{-1/10}. \end{aligned}$$

Since $A = \sqrt{\log(1/\delta)/n}$ and $m \approx n/(10\log(1/\delta))$, the term inside the probability is:

$$\mathbb{P}\left(\mathcal{N}(0,1) \geq -A\sqrt{m/\alpha}\right) \approx \mathbb{P}\left(\mathcal{N}(0,1) \geq -\sqrt{\frac{1}{10\alpha}}\right).$$

As $\alpha \to 0$, this probability tends to 1. Consequently, $p_+(m) \geq e^{-1/10}(1-\epsilon) > 3/4$ for small α. Since $p_-(m) = 1 - p_+(m)$, it follows that:

$$\min(p_-(m), p_+(m)) \leq 1 - e^{-1/10}(1-\epsilon) < 1/4.$$

This demonstrates that the approximate symmetry condition (2) is violated for $P \in \mathcal{P}_\alpha$, proving that the construction in Devroye et al. [2016] cannot be applied uniformly to this class.

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: the abstract explains the main contributions while sections 1 and 2 make the claims mathematically precise.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: yes, we discuss limitations and possible extensions of our results in section **"Final remarks and future directions"** at the end of the manuscript.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: while high-level intuition behind the proofs is outlined in the main text, the complete technical arguments are given in the supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: detailed overview of numerical experiments is provided in section 5 of the paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Zip file with anonymized code has been submitted with the supplementary materials.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: details are given in section 5 of the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification:

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: numerical experiments were produced using a standard laptop computer and do not require significant computational resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is theoretical in nature and focuses on mathematical statistics. It does not involve human subjects, sensitive data, or potential for immediate harmful applications.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: this is primarily a theoretical paper and it does not have any immediate societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.