Sequential 1-bit Mean Estimation with Near-Optimal Sample Complexity

Ivan Lau
National University of Singapore

Jonathan Scarlett
National University of Singapore

Abstract

In this paper, we study the problem of distributed mean estimation with 1-bit communication constraints. We propose a mean estimator that is based on (randomized and sequentially-chosen) interval queries, whose 1-bit outcome indicates whether the given sample lies in the specified interval. Our estimator is (ϵ, δ)-PAC for all distributions with bounded mean ($-\lambda \leqslant \mathbb{E}(X) \leqslant \lambda$) and variance ($\mathrm{Var}(X) \leqslant \sigma^2$) for some known parameters λ and σ. We derive a sample complexity bound $\widetilde{O}\big(\frac{\sigma^2}{\epsilon^2}\log\frac{1}{\delta} + \log\frac{\lambda}{\sigma}\big)$, which matches the minimax lower bound for the unquantized setting up to logarithmic factors and the additional $\log\frac{\lambda}{\sigma}$ term that we show to be unavoidable. We also establish an adaptivity gap for interval-query based estimators: the best non-adaptive mean estimator is considerably worse than our adaptive mean estimator for large $\frac{\lambda}{\sigma}$. Finally, we give tightened sample complexity bounds for distributions with stronger tail decay, and present additional variants that (i) handle an unknown sampling budget (ii) adapt to the unknown true variance given (possibly loose) upper and lower bounds on the variance, and (iii) use only two stages of adaptivity at the expense of more complicated (non-interval) queries.

1 INTRODUCTION

Mean estimation is one of the simplest yet most ubiquitous tasks in statistics, machine learning, and theoretical computer science. In modern applications such as those arising in large-scale and decentralized systems, the learner often has limited access to the true data samples. A common limitation is communication constraints, which require each data sample to be compressed to a small number of bits, before being communicated to the learner. In this paper, we address the extreme case of this setting where the learner receives only one bit of feedback per sample. This raises a fundamental theoretical question:

> *How does 1-bit quantization affect the sample complexity of mean estimation?*

Our main contribution is a 1-bit mean estimator whose sample complexity nearly matches the minimax lower bound for the unquantized setting. To the best of our knowledge, analogous results were previously established only for Gaussian random variables (Cai and Wei, 2022a, 2024) and other known location-scale families such as Laplace and logistic (Kipnis and Duchi, 2022; Kumar and Vatedka, 2025), leaving the fundamental limits of 1-bit mean estimation for broader, non-parametric distribution classes (such as those defined solely by bounded variance) largely unresolved.

1.1 Problem Setup

Distributional assumption. Let X be a real-valued random variable[1] with unknown distribution D. We assume that D belongs to a (non-parametric) family $\mathcal{D} = \mathcal{D}(\lambda, \sigma)$, defined by known parameters $\lambda \geqslant \sigma > 0$; a distribution D is in this family if the following conditions hold:

1. Bounded mean: $\mu(D) \in [-\lambda, \lambda]$,[2]
2. Bounded variance: $\mathrm{Var}(X) \leqslant \sigma^2 \leqslant \lambda^2$,

where both λ and σ are known to the learner. Note that the support of D may be unbounded.

1-bit communication protocol. The learner is interested in estimating the population mean $\mu = \mu(D) = \mathbb{E}[X]$ from n independent and identically distributed (i.i.d.) samples $X_1, \ldots, X_n \sim D$, subject to a 1-bit

[1] Our results also have implications for certain multivariate settings; see Section 4.6 for details.

[2] Without loss of generality, we set the search range to be symmetric. Note that a dependence on the search range λ is unavoidable in the 1-bit setting (see Theorem 6), but a crude upper bound can be used due to the mild logarithmic dependence in the sample complexity (see Theorem 4).

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

communication constraint per sample. The estimation proceeds through an interactive protocol between a learner and a single memoryless agent[3] that observes i.i.d. samples and sends 1-bit feedback to the learner. Specifically, for $t = 1, \ldots, n$:

1. The learner sends a 1-bit quantization function $Q_t \colon \mathbb{R} \to \{0, 1\}$ to an agent;
2. The agent observes a fresh sample $X_t \sim D$ and sends a 1-bit message $Y_t = Q_t(X_t)$ to the learner.

After n rounds, the learner forms an estimate $\hat{\mu}$ based on the entire interaction history $\big(Q_1, Y_1, \ldots, Q_n, Y_n\big)$. This (and similar) setting was also adopted in previous communication-constrained learning works, e.g., (Hanna et al., 2022; Mayekar et al., 2023; Lau and Scarlett, 2025).

The learner's algorithm in this protocol is formally defined as follows:

Definition 1 (1-bit mean estimator)**.** A 1-bit mean estimator is an algorithm for the learner that operates within the above communication protocol. It consists of

1. A (potentially randomized) query strategy for selecting the quantization functions $Q_1, \ldots, Q_n$, where the choice of Q_t can depend adaptively on the history of interactions $(Q_1, Y_1, \ldots, Q_{t-1}, Y_{t-1})$.
2. An estimation rule that maps the full transcript $(Q_1, Y_1, \ldots, Q_n, Y_n)$ to a final estimate $\hat{\mu} \in \mathbb{R}$.

We say that an estimator is *non-adaptive* if the query strategy selects all quantization functions in advance, without access to any of the outcomes $Y_1, \ldots, Y_n$.

Interval query model. In the problem formulation, we placed no restriction on the choice of quantization function Q_t. However, motivated by the desire for "simple" choices in practice, we focus primarily on *interval queries*, which take the form "Is $X_t \in I_t$?" for some interval $I_t = [a_t, b_t]$ (possibly with $a_t = -\infty$ or $b_t = \infty$). The resulting 1-bit feedback Y_t is the corresponding binary answer $\mathbf{1}\{X_t \in I_t\}$. Our main estimator will only use such queries, though we will also present a variant that uses general 1-bit queries.

Learner's goal. The learner's goal is to design a 1-bit mean estimator that returns an accurate estimate with high probability, while using as few samples as possible. We formalize this notion as follows:

Definition 2 ((ϵ, δ)-PAC)**.** A mean estimator is (ϵ, δ)-PAC for distribution family $\mathcal{D}$ with sample complexity $n(\epsilon, \delta)$ if, for each distribution $D \in \mathcal{D}$, it returns an ϵ-correct estimate $\hat{\mu}$ with probability at least $1 - \delta$, i.e.,

$$\text{for each } D \in \mathcal{D}, \quad \Pr\left(|\hat{\mu} - \mu(D)| \leqslant \epsilon\right) \geqslant 1 - \delta$$

and the number of samples required is bounded by $n(\epsilon, \delta)$. The probability is taken over the samples $X_1, \ldots, X_n$ and any internal randomness of the estimator.

1.2 Summary of Contributions

With the problem setup now in place, we summarize our main contributions as follows:

- We propose a novel adaptive 1-bit mean estimator (see Section 2.1) that only makes use of interval queries.
- We show that the mean estimator is (ϵ, δ)-PAC for distribution family $\mathcal{D}(\lambda, \sigma)$, with a sample complexity that matches the minimax lower bound $\Omega(\sigma^2/\epsilon^2 \cdot \log(\delta^{-1}))$ for the unquantized setting up to logarithmic factors and an additional $\log(\lambda/\sigma)$ term (see Theorem 4). Our sample complexity bound scales logarithmically with λ/σ, which contrasts with existing bounds for communication-constrained non-parametric mean estimators scaling at least linearly in λ.
- We derive a worst-case lower bound, showing that the additional $\log(\lambda/\sigma)$ term is unavoidable (see Theorem 6). For the interval-query model, we establish an "adaptivity gap" by showing a worst-case lower bound $\Omega(\lambda\sigma/\epsilon^2 \cdot \log(\delta^{-1}))$ for non-adaptive estimators.
- We provide several extensions including improved logarithmic factors under stronger tail decay, handling partially unknown parameters (ϵ, σ), and a two-stage variant under general 1-bit queries.

1.3 Related Work

The related work on distributed mean estimation is extensive, we only provide a brief outline here, emphasizing the most closely related works.

Classical mean estimation. Mean estimation (in the unquantized setting) is a fundamental and well-studied problem in statistics, e.g., see (Lee and Valiant, 2022; Cherapanamjeri et al., 2022; Minsker, 2023; Dang et al., 2023; Gupta et al., 2024) and the references therein. The state-of-the-art (ϵ, δ)-PAC estimator by (Lee and Valiant, 2022) achieves a tight sample complexity $n = (2 + o(1)) \cdot (\sigma^2/\epsilon^2) \cdot \log(1/\delta)$ for all distributions with finite variance σ. These results serve as a natural benchmark for mean estimation problems under communication constraints.

Distributed estimation and learning. Early work in distributed estimation, learning, and optimization was motivated by the applications of wireless sensor networks (see (Xiao et al., 2006; Varshney, 2012; Veeravalli and

[3] Equivalently, this can be viewed as a sequence of memoryless agents where the agent in each round may be different. In particular, the agent in round t only has access to X_t and not to the previous samples $X_1, \ldots, X_{t-1}$.

Varshney, 2012; He et al., 2020) and the references therein), with a recent resurgence driven by the rise of large-scale machine learning systems. This has led to the characterization of the sample complexity or minimax risk/error for various distributed estimation problems (Zhang et al., 2013; Garg et al., 2014; Shamir, 2014; Braverman et al., 2016; Xu and Raginsky, 2017; Han et al., 2018a,b; Barnes et al., 2019, 2020; Acharya et al., 2020a,b, 2021a,b,d, 2023; Shah et al., 2025).

While abundant, most of the existing literature differs in major aspects including the estimation goal itself, the use of parametric models, and/or imposing significantly stronger assumptions. To our knowledge, none of the existing work on non-parametric distributed estimation captures our problem setup. For example, distributed non-parametric density estimation (Barnes et al., 2020; Acharya et al., 2021c) is an inherently harder problem, and accordingly the authors impose certain regularity conditions on the density function (e.g., belonging to Sobolev space). Similarly, distributed non-parametric function estimation problems in (Zhu and Lafferty, 2018; Szabó and van Zanten, 2018, 2020; Cai and Wei, 2022b; Zaman and Szabó, 2022) assume certain tail bounds on the likelihood ratio (e.g., Gaussian white noise model).

Distributed mean estimation (DME). Several works study variants of mean estimation under communication constraints directly. A large body of work focuses on parametric settings, often assuming a known location-scale family (Kipnis and Duchi, 2022; Kumar and Vatedka, 2025) with a particular emphasis on Gaussians (Ribeiro and Giannakis, 2006a; Cai and Wei, 2022a, 2024). These estimators crucially rely on CDF inversion, which are highly dependent on exact knowledge of the parametric family, and are not suitable for our non-parametric setting. The non-parametric mean estimators in (Luo, 2005; Ribeiro and Giannakis, 2006b) can handle broader distributional families but require additional assumptions such as bounded support and/or smooth density functions. Furthermore, some of these estimators require more than 1 bit of feedback (per coordinate) per sample. A recent independent work on non-adaptive 1-bit mean estimation (Abdalla and Chen, 2026) partially circumvents these restrictive assumptions. However, their estimator adopts a fixed quantization range whose width scales as $\Omega(\sigma^2/\epsilon)$ in the worst-case,[4] and this translates to a sample complexity of $\Omega(\sigma^4/\epsilon^4)$. In contrast, our adaptive 1-bit mean estimator achieves near-optimal $\widetilde{O}(\sigma^2/\epsilon^2)$ rates for all distributions whose first two moments lie within known bounds.

Empirical vs. population mean estimation. A closely related line of work focuses on distributed *empirical* mean estimation of a fixed dataset, which is a key primitive in federated learning (Suresh et al., 2017; Konečný and Richtárik, 2018; Davies et al., 2021; Vargaftik et al., 2021; Mayekar et al., 2021; Vargaftik et al., 2022; Ben-Basat et al., 2024; Babu et al., 2025). These estimators typically achieve a minimax optimal mean squared error (MSE) that scale as $\mathbb{E}[(\hat{\mu} - \mu_{\text{emp}})^2] = O(\lambda^2/n)$. By using Markov's inequality and the median-of-means method, they can be converted to (ϵ, δ)-PAC *population* mean estimator with a sample complexity of $n = \widetilde{O}(\lambda^2/\epsilon^2 \cdot \log(1/\delta))$. In contrast, our mean estimator achieves a sample complexity of $\widetilde{O}(\sigma^2/\epsilon^2 \cdot \log(1/\delta) + \log(\lambda/\sigma))$, which is considerably smaller when $\sigma^2 \ll \lambda^2$. Although some empirical mean estimators achieve MSE that depends on empirical deviation/variance σ_{emp} of the fixed dataset (Ribeiro and Giannakis, 2006b; Suresh et al., 2022), they require a bounded support. Furthermore, their MSE scale at least linearly with λ, e.g., the one in (Suresh et al., 2022) scales as $\mathbb{E}[(\hat{\mu} - \mu_{\text{emp}})^2] = O(\sigma_{\text{emp}}\lambda/n + \lambda^2/n^2)$. Consequently, converting them to (ϵ, δ)-PAC *population* mean estimator using standard techniques would result in a sample complexity bound that scales at least linearly with λ.

2 ESTIMATOR AND UPPER BOUND

In this section, we introduce our 1-bit mean estimator and provide its performance guarantee. Our estimator is designed as a target-accuracy driven procedure that takes parameters $(\lambda, \sigma, \epsilon, \delta)$ as input. It operates to ensure the desired accuracy ϵ is attained with probability at least $1 - \delta$ while minimizing the sample complexity n, and hence does not have an explicit pre-specified sample budget. However, the estimator can readily be applied to the fixed-budget setting where the sampling budget is given and the goal is to minimize the estimation error ϵ. In Section 4.3, we address a harder variant of this, where n is fixed but unknown to the learner.

2.1 Description of the Estimator

Our estimator first localizes an interval I of length $O(\sigma)$ containing the mean μ with high probability (see Step 1). Using the mid-point of I as the "centre", it partitions $\mathbb{R}$ into symmetric regions with width growing exponentially (see Step 2). The "outer regions" have low probability mass, from which we can infer that suffices to only estimate the contributions of "significant" regions (See Step 3). Finally, the estimator forms an estimate of the mean contribution from the significant regions to within an additive error of $\epsilon/2$ (see Steps 4–5). This high-level strategy of performing "localization" (coarse estimation) and "refinement" (finer estimation) has also appeared in prior works such as (Cai and Wei, 2022a), but with very different details, particularly for refinement (see Remark 3).

[4] To bound the worst-case truncation bias by $O(\epsilon)$ under only a finite-variance assumption, it can be shown that one must set the range to be $\Omega(\sigma^2/\epsilon)$ due to the worst-case tightness of Cantelli's inequality (a one-sided version of Chebyshev's inequality).

In more detail, our mean estimator is outlined as follows, with any omitted details deferred to Appendix A:

1. Using existing median estimation techniques, we localize a high probability confidence interval $[L, U]$ containing the median M using

$$n_{\text{loc}}(\delta, \lambda, \sigma) = \Theta\left(\log \frac{\lambda}{\sigma} + \log \frac{1}{\delta}\right) \tag{1}$$

1-bit threshold queries (which is a special case of interval queries). Using the well-known property $|\mathbb{E}[X] - M| \leqslant \sigma$, we have the high probability confidence interval $[L - \sigma, U + \sigma]$ containing the mean. It can then be verified that $|(U+\sigma)-(L-\sigma)| \leqslant 6\sigma$. Without loss of generality, we may assume that the interval $[L - \sigma, U + \sigma]$ is of length 6σ with the midpoint being exactly 0, i.e., $L + U = 0$.

2. We partition $\mathbb{R}$ into non-overlapping symmetric regions $R_1, R_{-1}, R_2, R_{-2}, \ldots$ with width growing exponentially as follows:

$$\frac{R_i}{\sigma} = \begin{cases} [m_{i-1}, m_i) & \text{if } i \geqslant 1 \\ -R_i/\sigma & \text{if } i \leqslant -1, \end{cases} \tag{2}$$

where[5]

$$m_i = \begin{cases} 0 & \text{if } i = 0 \\ 2^i & \text{if } 1 \leqslant i \leqslant 4 \\ 2(m_{i-1} - 3) & \text{if } i \geqslant 5. \end{cases} \tag{3}$$

Note that $m_i = \Theta(2^i)$ increases exponentially. Since the sum of all $\mu_i = \mathbb{E}\left[X \cdot \mathbf{1}\left(X \in R_i\right)\right]$ is μ, we can consider estimating each μ_i separately.

3. We identify a threshold $i_{\max} = \Theta\left(\log\left(\sigma/\epsilon\right)\right)$ such that the sum of μ_i for all i satisfying $|i| > i_{\max}$ has at most $\epsilon/2$ contribution to μ:

$$\left|\sum_{|i|>i_{\max}} \mu_i\right| \leqslant \frac{\epsilon}{2},$$

so that they can be estimated as being 0, i.e., $\hat{\mu}_i = 0$ for $|i| > i_{\max}$. It remains to form an estimate $\hat{\mu} = \sum_{|i|\leqslant i_{\max}} \hat{\mu}_i$ satisfying

$$\left|\sum_{|i|\leqslant i_{\max}} \mu_i - \sum_{|i|\leqslant i_{\max}} \hat{\mu}_i\right| \leqslant \frac{\epsilon}{2}$$

as this implies

$$\begin{aligned} |\mu - \hat{\mu}| &= \left|\sum_{|i|\leqslant i_{\max}} \mu_i + \sum_{|i|>i_{\max}} \mu_i - \sum_{|i|\leqslant i_{\max}} \hat{\mu}_i\right| \\ &\leqslant \left|\sum_{|i|\leqslant i_{\max}} \mu_i - \sum_{|i|\leqslant i_{\max}} \hat{\mu}_i\right| + \left|\sum_{|i|>i_{\max}} \mu_i\right| \\ &\leqslant \frac{\epsilon}{2} + \frac{\epsilon}{2} \\ &= \epsilon. \end{aligned}$$

4. Let $R_i = [a_i, b_i)$. and define the random variable $T_i \sim \text{Unif}(a_i, b_i)$. For ease of notation, we write $p_{a_i} := \Pr\left(X \in [a_i, T_i]\right)$ and $p_{b_i} := \Pr\left(X \in [T_i, b_i]\right)$. It can be verified that

$$\mu_i = \mathbb{E}\left[X \cdot \mathbf{1}\left(X \in R_i\right)\right] = a_i \cdot p_{a_i} + b_i \cdot p_{b_i}. \tag{4}$$

In Appendix B, we show that p_{a_i} (resp. p_{b_i}) is equivalent to the probability of X being in R_i and getting rounded down to a_i (resp. rounded up to b_i) by a binary stochastic quantizer for R_i.

5. It is therefore sufficient for the learner to form good estimates $\hat{p}_{a_i}$ of p_{a_i} and $\hat{p}_{b_i}$ of p_{b_i} such that

$$\hat{\mu}_i = a_i \cdot \hat{p}_{a_i} + b_i \cdot \hat{p}_{b_i}$$

satisfies

$$\begin{aligned} &\left|\sum_{|i|\leqslant i_{\max}} \mu_i - \sum_{|i|\leqslant i_{\max}} \hat{\mu}_i\right| \\ &= \left|\sum_{|i|\leqslant i_{\max}} a_i \cdot (p_{a_i} - \hat{p}_{a_i}) + b_i \cdot (p_{b_i} - \hat{p}_{b_i})\right| \\ &\leqslant \frac{\epsilon}{2}. \end{aligned}$$

The learner estimates them separately using randomized interval queries of the form $\mathbf{1}\{X \in [a_i, T_i]\}$ and $\mathbf{1}\{X \in [T_i, b_i]\}$, and empirical averages of the 1-bit feedback sent by agent. Using standard concentration inequalities, the number of 1-bit observations n_i needed to form "accurate" estimates can be bounded by $n_i = \widetilde{O}\left(\left(\frac{\sigma^2}{\epsilon^2} + \frac{2^i\sigma}{\epsilon}\right) \cdot \log\left(\frac{1}{\delta}\right)\right)$. Summing over all i satisfying $|i| \leqslant i_{\max} = \Theta\left(\log\left(\sigma/\epsilon\right)\right)$ gives

$$n_{\text{ref}}(\epsilon, \delta, \sigma) = \sum_{i:|i|\leqslant i_{\max}} n_i = \widetilde{O}\left(\frac{\sigma^2}{\epsilon^2} \cdot \log\left(\frac{1}{\delta}\right)\right).$$

Combining this with the $n_{\text{loc}}(\delta, \lambda, \sigma) = O\left(\log \frac{\lambda}{\sigma} + \log \frac{1}{\delta}\right)$ samples used in Step 1 for localization, we obtain a total sample complexity of

$$n := n(\epsilon, \delta, \lambda, \sigma) = \widetilde{O}\left(\frac{\sigma^2}{\epsilon^2} \cdot \log\left(\frac{1}{\delta}\right) + \log \frac{\lambda}{\sigma}\right).$$

[5] We choose to define m_i as in (3) for the convenience of analysis later on. Any exponential/geometric growth rate (e.g., $m_i = \Theta(a^i)$ for $a > 1$) would be sufficient to achieve the sample complexity in Theorem 4.

Remark 3. Standard refinement components such as stochastic quantization or CDF inversion fail to achieve the near-optimal variance-dependent rate in the general non-parametric setting. To achieve the near-optimal rate, we introduce specific design choices in the refinement stage that interact delicately:

- an exponential interval partitioning scheme that concentrates sampling on regions with "significant" probability mass;
- a carefully chosen truncation index that balances (a) bias from tail truncation and (b) sample complexity blow up from covering too many "insignificant" regions;
- a calibrated region-wise sample allocation to ensure the near-optimality of the "global" sample complexity.

Eliminating or simplifying any of these components is likely to "break" the claimed sample complexity given below.

2.2 Upper bound

We now formally state the main result of this paper, which is the performance guarantee of our mean estimator in Section 2.1. The proof is deferred to Appendix A, where we also provide the omitted details in the above outline.

Theorem 4. The mean estimator given in Section 2.1 is (ϵ, δ)-PAC for distribution family $\mathcal{D}(\lambda, \sigma)$, with sample complexity

$$n = O\left(\frac{\sigma^2}{\epsilon^2} \cdot \log^3\left(\frac{\sigma}{\epsilon}\right) \cdot \log\left(\frac{\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right) + \log\frac{\lambda}{\sigma}\right) \quad (5)$$

$$= \widetilde{O}\left(\frac{\sigma^2}{\epsilon^2}\log\frac{1}{\delta} + \log\frac{\lambda}{\sigma}\right). \quad (6)$$

Thus, we match the unquantized scaling up to logarithmic factors (see Section 1.3) and an additional $O(\log \lambda/\sigma)$ term. In Theorem 6 below, we show that this $\log(\lambda/\sigma)$ term is unavoidable. In Sections 4.1 and 4.2, we provide improved upper bounds for distributions with stronger tail decay. We also study variants where (ϵ, σ) are not prespecified in Sections 4.3 and 4.4 and a variant that uses only two rounds/stages of adaptivity in Section 4.5.

Remark 5 (Computational Complexity)**.** While our primary focus is sample complexity, the proposed estimator is also computationally efficient. The localization step (Step 1) relies on noisy binary search subroutine, requiring $O(\log^2(\lambda/\sigma))$ time (Gretta and Price, 2024, Section 6). The refinement step (Steps 2–5) operates in time linear in the sample complexity, assuming constant-time arithmetic operations and random threshold generation.

3 LOWER BOUND AND ADAPTIVITY GAP

In this section, we provide two lower bounds on the sample complexity. We first provide, in Theorem 6, a near-matching worst-case lower bound to the upper bound in Theorem 4. In particular, we show that the $\log(\lambda/\sigma)$ term is unavoidable. Perhaps more interestingly, we show in Theorem 7 that the best non-adaptive mean estimator is strictly worse than our adaptive mean estimator, at least under the interval query model. This shows that there is an "adaptivity gap" between the performance of adaptive and non-adaptive interval query based mean estimators. The proofs are given in Appendix C.

Theorem 6. For any (ϵ, δ)-PAC 1-bit mean estimator, and any $\epsilon < \sigma/2$, there exists a distribution $D \in \mathcal{D}(\lambda, \sigma)$ such that the number of samples n must satisfy

$$n = \Omega\left(\frac{\sigma^2}{\epsilon^2} \cdot \log\left(\frac{1}{\delta}\right) + \log\left(\frac{\lambda}{\sigma}\right)\right).$$

Theorem 7. For any non-adaptive (ϵ, δ)-PAC estimator that only makes interval queries, and any $\epsilon < \sigma/2$, there exists a distribution $D \in \mathcal{D}(\lambda, \sigma)$ such that the number of samples n must satisfy

$$n = \Omega\left(\frac{\lambda\sigma}{\epsilon^2} \cdot \log\left(\frac{1}{\delta}\right)\right).$$

Both proofs are based on constructing a (finite) "hard subset" of distributions that capture two sources of difficulty: (i) "coarsely" identifying the distribution's location in $[-\lambda, \lambda]$ among $\Theta(\lambda/\sigma)$ possibilities, and (ii) "finely" estimating the mean by distinguishing between two possibilities at that location whose means differ by 2ϵ. The fine estimation step inherently requires $\Omega(\sigma^2/\epsilon^2 \cdot \log(1/\delta))$ samples, based on standard hypothesis testing lower bound. However, the dependency on λ/σ arising from the coarse identification step differs in adaptive vs. non-adaptive settings:

- In Theorem 6 (adaptive setting), we can simply interpret the additive logarithmic dependence as the number of bits needed to identify the correct location among the $\Theta(\lambda/\sigma)$ possibilities, with each query giving at most 1 bit of information.
- In Theorem 7 (non-adaptive setting), the *multiplicative* dependence arises because the estimator needs to allocate enough queries in *every* one of the $\Theta(\lambda/\sigma)$ locations, as it does not know the correct location in advance.

We note that the distributed Gaussian mean estimator in (Cai and Wei, 2024) is non-adaptive and achieves an order-optimal MSE. However, their estimator is specific to Gaussian distributions, and their quantization functions

are not based on interval queries. We will build on their localization strategy in our two-stage variant (Section 4.5), but we avoid their refinement strategy which is much more Gaussian-specific (CDF inversion).

4 VARIATIONS AND REFINEMENTS

4.1 Bounded Higher-Order Moments

Suppose further that the random variable X has a finite k-th central moment bounded by σ^k for some $k > 2$, i.e.,

$$\mathbb{E}\big[|X-\mu|^k\big] \leqslant \sigma^k. \tag{7}$$

By Lyapunov's inequality, we have

$$\left(\mathbb{E}\big[|X-\mu|^2\big]\right)^{1/2} \leqslant \left(\mathbb{E}\big[|X-\mu|^k\big]\right)^{1/k} = \sigma,$$

which implies that the variance is bounded by σ^2. The condition (7) imposes a stronger tail decay that is imposed by variance alone,[6] and in this case we can tighten the $\log(\sigma/\epsilon)$ factors in Theorem 4 to $\log_{k/2}\log(\sigma/\epsilon)$.

Theorem 8. Suppose that random variable X satisfies (7) for some $\sigma > 0$ and $k > 2$. Then there exists an (ϵ, δ)-PAC 1-bit mean estimator with sample complexity

$$n = O\left(\frac{\sigma^2}{\epsilon^2}\cdot\left(\log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right)\right)^3\cdot\log\left(\frac{\log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right) + \log\frac{\lambda}{\sigma}\right).$$

The protocol and proof of its guarantee are similar to those given in Section 2.1, with the main difference being that we change m_i in (3) to a choice that scales doubly exponentially (see (65) in Appendix D.1). Consequently, i_{max} scales as $\log_{k/2}\log(\sigma/\epsilon)$. The details are given in Appendix D.1.

4.2 Sub-Gaussian Random Variables

Now we suppose that $X-\mu$ is sub-Gaussian with known parameter σ^2, i.e.,

$$\Pr(|X-\mu| \geqslant t) \leqslant 2\exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{8}$$

Note that we have $\mathrm{Var}(X) \leqslant \sigma^2$ and X has a finite k-th central moment for every k. In this case, we can tighten the $\log_{k/2}\log(\sigma/\epsilon)$ factors in Theorem 8 to $\log^*(\sigma/\epsilon)$, where the "iterated logarithm" $\log^*(\cdot)$ is the number of times the logarithm function must be iteratively applied before the result is less than or equal to 1.

[6] Note that if the learner knows $\mathrm{Var}(X) \leqslant \gamma^2$ for some known $\gamma \ll \sigma$, then using our main estimator in Section 2.1 would still lead to a better sample complexity. That is, this refinement is primarily of interest when γ and σ are comparable.

Theorem 9. Suppose that $X-\mu$ is sub-Gaussian with known parameter σ^2. Then there exists an (ϵ, δ)-PAC 1-bit mean estimator with sample complexity

$$n = O\left(\frac{\sigma^2}{\epsilon^2}\cdot\left(\log^*\left(\frac{\sigma}{\epsilon}\right)\right)^3\cdot\log\left(\frac{\log^*\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right) + \log\frac{\lambda}{\sigma}\right).$$

The protocol and proof of its guarantee are again similar to those of Theorem 4, with the main difference being that we change m_i in (3) to a choice that scales according to a tower of exponentials of height i. Consequently, i_{max} scales as $\log^*(\sigma/\epsilon)$. The details are given in Appendix D.2.

4.3 Unknown Target Accuracy

By inverting the ϵ term in (5), we obtain a performance guarantee on the target accuracy of our main algorithm in terms of parameters n, δ, σ, and λ, where n is the pre-specified sampling budget. In other words, by running our mean estimator with parameters

$$(\epsilon, \delta, \lambda, \sigma) = (\epsilon(n, \delta, \lambda, \sigma),\ \delta, \lambda, \sigma)$$

we obtain a mean estimate that is ϵ-accurate with error probability at most δ, where $\epsilon = \epsilon(n, \delta, \lambda, \sigma)$ is computed by inverting the ϵ term in (5). Furthermore, the number of samples used

$$n(\epsilon, \delta, \lambda, \sigma) = n_{\text{loc}}(\delta, \lambda, \sigma) + n_{\text{ref}}(\epsilon, \delta, \lambda, \sigma)$$

trivially satisfy the pre-specified sampling budget. Here we define

$$n_{\text{loc}}(\delta, \lambda, \sigma) = \Theta\left(\log\frac{\lambda}{\sigma} + \log\frac{1}{\delta}\right)$$

as the number of samples used in the localization step of our mean estimator (see Step 1 of Section 2.1), and

$$n_{\text{ref}}(\epsilon, \delta, \sigma) = \Theta\left(\frac{\sigma^2}{\epsilon^2}\cdot\log^3\left(\frac{\sigma}{\epsilon}\right)\cdot\log\left(\frac{\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right) \tag{9}$$

as the number of samples used in Steps 3-5 (refinement).

We now consider the scenario where the true sample budget n_{true} is not pre-specified in advance, but the other parameters σ, λ, and δ are still known. In this case, we can no longer run the algorithm using an ϵ inverted as before, since n is not pre-specified. A naive way is to guess some ϵ_{guess} and run the estimator with parameters $(\epsilon_{\text{guess}}, \delta, \lambda, \sigma)$. Moreover, if we guess some ϵ_{guess} which ends up being too big, i.e., $\epsilon_{\text{guess}} \gg \epsilon(n_{\text{true}}, \delta, \lambda, \sigma)$, then the resulting estimator is inaccurate given the sampling budget. Conversely, if the guess is too small, i.e., $\epsilon_{\text{guess}} \ll \epsilon(n_{\text{true}}, \delta, \lambda, \sigma)$, then the sampling budget may not be sufficient to guarantee an ϵ_{guess}-accurate estimate.

To overcome this, we can use a standard halving trick on ϵ_{guess} (along with careful consideration of δ) to "anytime-ify" the mean estimator. We run the localization step and

the partitioning step (see Steps 1-2 of Section 2.1) once, which requires knowing only the parameters $(\delta, \lambda, \sigma)$ and uses $n_{\text{loc}}(\delta, \lambda, \sigma)$ samples. Then for each round $\tau = 1, 2, \ldots,$ we run Steps 3-5 of our mean estimator with parameters

$$(\epsilon_\tau, \delta_\tau, \sigma) \quad \text{where} \quad \epsilon_\tau = \frac{\sigma}{2^\tau} \quad \text{and}^7 \quad \delta_\tau = \frac{6\delta}{\pi^2 \tau^2}.$$

Note that this process would use $n_{\text{ref}}(\epsilon_\tau, \delta_\tau, \sigma)$ samples in each round τ. It follows that round τ will complete as long as the true sampling budget n_{true} satisfies

$$n_{\text{loc}}(\delta, \lambda, \sigma) + \sum_{s=1}^{\tau} n_{\text{ref}}(\epsilon_s, \delta_s, \sigma) \leqslant n_{\text{true}}.$$

When the real sampling budget n_{true} is exhausted, we stop and output the last estimate we fully computed, i.e., we output $\hat{\mu}_T$, where

$$T = \max_{\tau \geqslant 1} \left\{ \sum_{s=1}^{\tau} n_{\text{ref}}(\epsilon_s, \delta_s, \sigma) \leqslant n_{\text{true}} - n_{\text{loc}}(\delta, \lambda, \sigma) \right\} \tag{10}$$

is the last round where the subroutine is completed. By the union bound and the guarantee of the subroutine for each round τ, we have with probability at least $1 - \delta$ that every estimate $\hat{\mu}_\tau$ formed is ϵ_τ-accurate. In particular, under this high-probability event, the final output $\hat{\mu}_T$ satisfies

$$|\hat{\mu}_T - \mu| \leqslant \epsilon_T = \frac{\sigma}{2^T}.$$

Ideally, we would like to compare ϵ_T against the "oracle accuracy" ϵ^*, which satisfies

$$n_{\text{ref}}(\epsilon^*, \delta, \sigma) = n_{\text{true}} - n_{\text{loc}}(\delta, \lambda, \sigma),$$

i.e, ϵ^* is the optimal target accuracy that could be achieved (with high probability) had we known the unknown sampling budget n_{true} in advance. Indeed, under a mild assumption of n not being too small, we show that ϵ_T obtained from the doubling trick matches the "oracle" value to within a constant factor.

Theorem 10. Under the preceding setup, assuming $n_{\text{true}} \geqslant n_{\text{loc}}(\delta, \lambda, \sigma)$, we have $\epsilon_T = O(\epsilon^*)$.

The proof is given in Appendix E.1.

4.4 Adapting to Unknown Variance

The sample complexity of our mean estimator, as stated in Theorem 4, scales quadratically with σ/ϵ, where σ^2 is a known upper bound on the true variance $\sigma^2_{\text{true}} = \text{Var}(X)$. This scaling is not ideal when the upper bound is loose. This is in contrast to the unquantized setting, where there exist mean estimators whose sample complexity scales quadratically with $\sigma_{\text{true}}/\epsilon$ without any knowledge of σ (Lee and Valiant, 2022).

Under the 1-bit communication constraint, it may be difficult to learn the true variance (and estimate mean at the same time). We consider the case where both target accuracy ϵ and true variance σ_{true} are unknown, but we know that

$$\sigma_{\text{true}} \in [\sigma_{\min}, \sigma_{\max}] \quad \text{and} \quad \epsilon = r\sigma_{\text{true}}$$

for some known r. That is, we seek accuracy to within r multiples of standard deviation, even though we do not know standard deviation σ_{true}.

In this case, we construct a mean estimator that uses the mean estimator in Section 2.1 as subroutine. Set

$$T = \lceil \log_2 (\sigma_{\max}/\sigma_{\min}) \rceil. \tag{11}$$

For $i = 0, \ldots, T$, we define

$$\sigma_i = \sigma_{\min} \cdot 2^i \iff \sigma_i = \begin{cases} \sigma_{\min} & \text{if } i = 0 \\ 2\sigma_{i-1} & \text{if } 1 \leqslant i \leqslant T \end{cases}, \tag{12}$$

and run the mean estimator in Section 2.1 with parameters

$$(\epsilon_i, \delta_i, \lambda, \sigma_i) = \left(\frac{r\sigma_i}{5}, \quad \frac{\delta}{T+1}, \quad \lambda, \quad \sigma_i, \right) \tag{13}$$

to obtain an estimate $\hat{\mu}^{(i)}$. For each i, we define a confidence interval

$$I_i = [\hat{\mu}^{(i)} \pm \epsilon_i] = \left[\hat{\mu}^{(i)} - \frac{r\sigma_i}{5}, \hat{\mu}^{(i)} + \frac{r\sigma_i}{5} \right] \tag{14}$$

of length $2\epsilon_i$, and we say that σ_i is *feasible* if I_i overlaps with all confidence intervals of higher indices:

$$I_i \cap I_j \neq \varnothing \quad \text{for all } j > i. \tag{15}$$

Note that σ_T is trivially feasible. We return the mean estimate corresponding to the smallest feasible σ_i, i.e., we return the estimate $\hat{\mu}^{(i^*)}$ where i^* is the smallest i that satisfies condition (15). The resulting mean estimator has a sample complexity that scales quadratically with $\sigma_{\text{true}}/\epsilon = 1/r$, but pays an extra multiplicative factor $\log(\sigma_{\max}/\sigma_{\min})$.

Theorem 11. The mean estimator above is (ϵ, δ)-PAC with sample complexity

$$n = \tilde{O}\left(\log\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right) \left(\frac{1}{r^2} \log\left(\frac{1}{\delta}\right) + \log \frac{\lambda}{\sqrt{\sigma_{\min}\sigma_{\max}}} \right) \right).$$

The proof is given in Appendix E.2.

Remark 12. Intuitively, the feasibility condition (15) tells us whether an interval I_i is consistent with the intervals obtained using larger/more conservative σ-values.

[7] Alternatively, we could pick any suitable ϵ_0 as the initial target accuracy and let $\epsilon_\tau = \epsilon_0/2^{\tau-1}$.

In particular, if $\sigma_i \geqslant \sigma_{\text{true}}$ then σ_i is feasible (see Appendix E.2), but the converse may not hold. In practice, we can start with the largest σ-value and sequentially half it (i.e., $\sigma_i = \sigma_{\max}/2^i$), until we find the first i where σ_i is infeasible, and return $\hat{\mu}^{(i+1)}$. Although this may not lead to an improvement in the upper bound (e.g., the loop may not terminate even when $\sigma_i < \sigma_{\text{true}}$), it can help avoid using all T loops when it is unnecessary to do so.

4.5 Two-Stage Variant

Our mean estimator in Section 2.1 uses $O\left(\log \frac{\lambda}{\sigma} + \log \frac{1}{\delta}\right)$ rounds of adaptivity. Specifically, the localization step (Step 1 of Section 2.1), which performs median estimation through noisy binary search, requires $O\left(\log \frac{\lambda}{\sigma} + \log \frac{1}{\delta}\right)$ rounds of adaptivity; while the refinement step can be done in just one additional round after we have localized an interval of length $O(\sigma)$ containing the mean. In this section, we provide an alternative localization procedure that is non-adaptive, with the remaining steps unchanged. This gives us an alternative mean estimator that requires only two rounds of adaptivity – one for localization and one for refinement. However, this comes at the cost of using general 1-bit queries in the first round, as opposed to only using interval queries.

Our alternative localization step is adapted from the localization step of the non-adaptive Gaussian mean estimator in (Cai and Wei, 2024), which is presented therein for Gaussian distributions but also noted to extend to the general sub-Gaussian case (unlike their refinement stage). We modify their localization step so that it works on all distributions with mean and variance lying within known bounds (namely, $[-\lambda, \lambda]$ and $[0, \sigma^2]$ respectively), with the following performance guarantee:

Theorem 13. There exists a 1-bit non-adaptive localization protocol taking $(\delta, \lambda, \sigma)$ as input such that for each $D \in \mathcal{D}$, it returns an interval I containing μ with probability at least $1 - \delta/2$. Furthermore, the number of samples used is $\Theta\left(\log\left(\frac{\lambda}{\sigma}\right) \cdot \log \frac{\log(\lambda/\sigma)}{\delta}\right)$ and $|I| = O(\sigma)$.

We describe the high-level idea here. The learner partitions the interval $[-\lambda, \lambda]$ into 2^K subintervals $\{I_0, I_1, \ldots, I_{2^K-1}\}$ of same length for some $K = \Theta(\log(\lambda/\sigma))$, and the learner tries to estimate all K bits of the Gray code representation of the subinterval containing μ. Each of these K bits is estimated reliably by taking a majority vote over $J = \Theta\left(\log \frac{K}{\delta}\right)$ samples. The details are given in Appendix F.

By replacing the localization step of our main estimator (Step 1 of Section 2.1) with the alternative localization step above, we have a mean estimator with the following performance guarantee.

Corollary 14. The alternative mean estimator described above is (ϵ, δ)-PAC for distribution family $\mathcal{D}(\lambda, \sigma)$, with sample complexity

$$n = \widetilde{O}\left(\frac{\sigma^2}{\epsilon^2} \log \frac{1}{\delta} + \log\left(\frac{\lambda}{\sigma}\right) \cdot \log\log\left(\frac{\lambda}{\sigma}\right)\right).$$

Furthermore, it uses only two rounds of adaptivity, the first of which uses general (non-interval) 1-bit queries.

4.6 Multivariate Mean Estimation

The multivariate case (i.e., $X \in \mathbb{R}^d$ with $d > 1$) is naturally of significant interest. We have focused on the univariate case since it is the natural starting point and is already challenging. However, our results turn out to also provide some preliminary findings for multivariate settings.

Specifically, suppose that X takes values in $\mathbb{R}^d$ and has entries $X_1, \ldots, X_d$ satisfying our earlier assumptions individually for each coordinate $i = 1, \ldots, d$. By applying our univariate techniques coordinate-wise with parameters $\epsilon/\sqrt{d}$ and δ/d, we obtain an overall estimate that is ϵ-accurate in ℓ_2 norm with probability at least $1 - \delta$. In accordance with Theorem 4, the sample complexity is

$$\widetilde{O}\left(\frac{d^2\sigma^2}{\epsilon^2} \log \frac{1}{\delta} + d \log \frac{\lambda}{\sigma}\right),$$

where the d^2 factor arises from (i) using the scaled accuracy parameter $\epsilon/\sqrt{d}$, and (ii) running the univariate subroutine d times. This may seem potentially loose on first glance, due to the correct scaling being $\frac{\sigma^2}{\epsilon^2} \cdot (d + \log(1/\delta))$ in the absence of a communication constraint (Lugosi and Mendelson, 2019). However, under 1-bit feedback, the $d^2\sigma^2/\epsilon^2$ dependence in fact unavoidable even in the special case of Gaussian random variables; see (Cai and Wei, 2024, Theorem 8) with the parameter m' therein equating to n/d in our notation under 1-bit feedback.[8] Moreover, if the communication bottleneck is relaxed to allow d bits of feedback per sample (i.e., one bit *per coordinate*), applying our univariate estimator coordinate-wise yields a sample complexity of $\widetilde{O}\big(\frac{d\sigma^2}{\epsilon^2} \log(1/\delta) + d \log(\lambda/\sigma)\big)$. In the constant error probability regime ($\delta = \Theta(1)$), this matches the unconstrained rate up to logarithmic factors. In the regime $\delta = o(1)$ there remains a significant gap due to the fact that $d \log \frac{1}{\delta} \gg d + \log \frac{1}{\delta}$, but this gap is inherent to any approach that controls each coordinate's error to $O\big(\frac{\epsilon}{\sqrt{d}}\big)$ separately.

Beyond the issue of joint (d, δ) dependence, another limitation of the coordinate-wise approach is that it does not capture the dependence on off-diagonal terms in the covariance matrix Σ. Doing so may be significantly more difficult, particularly when Σ is not known exactly and so "whitening" techniques cannot readily be used. We leave such considerations for future work.

[8] To give slightly more detail, the parameters m and $b_i = 1$ therein equate respectively to the number of samples n and number of bits allowed per feedback.

5 CONCLUSION

In this paper, we studied the problem of estimating the mean of a distribution under the extreme constraint of a single bit of communication per sample. We proposed an adaptive estimator that is (ϵ, δ)-PAC for all distributions with bounded mean and variance, which achieves near-optimal sample complexity. This result demonstrates that the statistical efficiency of mean estimation is largely preserved under 1-bit communication constraints. We also established an adaptivity gap for the interval query model, showing that non-adaptive strategies are strictly suboptimal. Several directions remain for future research, including tightening the polylogarithmic factors, adapting to unknown variance and target accuracy with as few assumptions as possible, and extending to multivariate settings beyond the coordinate-wise approach.

Acknowledgment

This work was supported by the Singapore National Research Foundation (NRF) under its AI Visiting Professorship programme.

References

Abdalla, P. and Chen, J. (2026). Robust mean estimation under quantization. *arXiv preprint arXiv:2601.07074*.

Acharya, J., Canonne, C., Liu, Y., Sun, Z., and Tyagi, H. (2021a). Distributed estimation with multiple samples per user: Sharp rates and phase transition. In *Advances in Neural Information Processing Systems*, volume 34, pages 18920–18931.

Acharya, J., Canonne, C. L., Liu, Y., Sun, Z., and Tyagi, H. (2021b). Interactive inference under information constraints. *IEEE Transactions on Information Theory*, 68(1):502–516.

Acharya, J., Canonne, C. L., Singh, A. V., and Tyagi, H. (2021c). Optimal rates for nonparametric density estimation under communication constraints. *CoRR*, abs/2107.10078.

Acharya, J., Canonne, C. L., Sun, Z., and Tyagi, H. (2023). Unified lower bounds for interactive high-dimensional estimation under information constraints. *Advances in Neural Information Processing Systems*, 36:51133–51165.

Acharya, J., Canonne, C. L., and Tyagi, H. (2020a). Inference under information constraints I: Lower bounds from chi-square contraction. *IEEE Transactions on Information Theory*, 66(12):7835–7855.

Acharya, J., Canonne, C. L., and Tyagi, H. (2020b). Inference under information constraints II: communication constraints and shared randomness. *IEEE Transactions on Information Theory*, 66(12):7856–7877.

Acharya, J., Kairouz, P., Liu, Y., and Sun, Z. (2021d). Estimating sparse discrete distributions under privacy and communication constraints. In *Algorithmic Learning Theory (ALT)*, pages 79–98. PMLR.

Babu, N. S., Kumar, R., and Vatedka, S. (2025). Unbiased quantization of the l_1 ball for communication-efficient distributed mean estimation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258, pages 1270–1278. PMLR.

Barnes, L. P., Han, Y., and Özgür, A. (2019). Fisher information for distributed estimation under a blackboard communication protocol. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2704–2708. IEEE.

Barnes, L. P., Han, Y., and Özgür, A. (2020). Lower bounds for learning distributions under communication constraints via Fisher information. *Journal of Machine Learning Research*, 21:1–30.

Ben-Basat, R., Vargaftik, S., Portnoy, A., Einziger, G., Ben-Itzhak, Y., and Mitzenmacher, M. (2024). Accelerating federated learning with quick distributed mean estimation. In *International Conference on Machine Learning (ICML)*, volume 235, pages 3410–3442. PMLR.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Braverman, M., Garg, A., Ma, T., Nguyen, H. L., and Woodruff, D. P. (2016). Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Symposium on Theory of Computing Conference, STOC'16*, pages 1011–1020. ACM.

Cai, T. T. and Wei, H. (2022a). Distributed adaptive Gaussian mean estimation with unknown variance: Interactive protocol helps adaptation. *The Annals of Statistics*, 50(4):1992–2020.

Cai, T. T. and Wei, H. (2022b). Distributed nonparametric function estimation: Optimal rate of convergence and cost of adaptation. *The Annals of Statistics*, 50(2):698–725.

Cai, T. T. and Wei, H. (2024). Distributed Gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms. *Journal of Machine Learning Research*, 25(37):1–63.

Cherapanamjeri, Y., Tripuraneni, N., Bartlett, P., and Jordan, M. (2022). Optimal mean estimation without a variance. In *Conference on Learning Theory*, pages 356–357. PMLR.

Dang, T., Lee, J., Song, M., and Valiant, P. (2023). Optimality in mean estimation: Beyond worst-case, beyond sub-Gaussian, and beyond $1 + \alpha$ moments. In *Advances in Neural Information Processing Systems*, volume 36, pages 4150–4176.

Davies, P., Gurunanthan, V., Moshrefi, N., Ashkboos, S., and Alistarh, D. (2021). New bounds for distributed mean estimation and variance reduction. In *International Conference on Learning Representations, (ICLR)*.

Garg, A., Ma, T., and Nguyen, H. L. (2014). On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems 27*, pages 2726–2734.

Gretta, L. and Price, E. (2024). Sharp Noisy Binary Search with Monotonic Probabilities. In *51st International Colloquium on Automata, Languages, and Programming (ICALP)*, volume 297, pages 75:1–75:19.

Gupta, S., Hopkins, S., and Price, E. (2024). Beyond catoni: Sharper rates for heavy-tailed and robust mean estimation. In *Conference on Learning Theory (COLT)*, pages 2232–2269. PMLR.

Han, Y., Mukherjee, P., Özgür, A., and Weissman, T. (2018a). Distributed statistical estimation of high-dimensional and non-parametric distributions. In *IEEE International Symposium on Information Theory (ISIT)*, pages 506–510.

Han, Y., Özgür, A., and Weissman, T. (2018b). Geometric lower bounds for distributed parameter estimation under communication constraints. In *Conference on Learning Theory (COLT)*, volume 75, pages 3163–3188. PMLR.

Hanna, O. A., Yang, L., and Fragouli, C. (2022). Solving Multi-Arm Bandit Using a Few Bits of Communication. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 11215–11236.

He, S., Shin, H.-S., Xu, S., and Tsourdos, A. (2020). Distributed estimation over a low-cost sensor network: A review of state-of-the-art. *Information Fusion*, 54:21–43.

Kipnis, A. and Duchi, J. C. (2022). Mean estimation from one-bit measurements. *IEEE Transactions on Information Theory*, 68(9):6276–6296.

Konečnỳ, J. and Richtárik, P. (2018). Randomized distributed mean estimation: Accuracy vs. communication. *Frontiers in Applied Mathematics and Statistics*, 4:62.

Kumar, R. and Vatedka, S. (2025). One-bit distributed mean estimation with unknown variance. *arXiv preprint arXiv:2501.18502*.

Lau, I. and Scarlett, J. (2025). Quantile multi-armed bandits with 1-bit feedback. In *Algorithmic Learning Theory*, pages 664–699. PMLR.

Lee, J. (2020). CSCI 1951-W Sublinear Algorithms for Big Data, Lecture 11. `https://cs.brown.edu/courses/csci1951-w/lec/lec%2011%20notes.pdf`. Accessed: 2025-07-01.

Lee, J. C. and Valiant, P. (2022). Optimal sub-Gaussian mean estimation in $\mathbb{R}$. In *IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 672–683. IEEE.

Lugosi, G. and Mendelson, S. (2019). Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783 – 794.

Luo, Z.-Q. (2005). Universal decentralized estimation in a bandwidth constrained sensor network. *IEEE Transactions on Information Theory*, 51(6):2210–2219.

Mayekar, P., Scarlett, J., and Tan, V. Y. (2023). Communication-constrained bandits under additive Gaussian noise. In *International Conference on Machine Learning (ICML)*, pages 24236–24250.

Mayekar, P., Suresh, A. T., and Tyagi, H. (2021). Wyner-ziv estimators: Efficient distributed mean estimation with side-information. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3502–3510. PMLR.

Minsker, S. (2023). Efficient median of means estimator. In *Conference on Learning Theory (COLT)*, pages 5925–5933. PMLR.

Polyanskiy, Y. and Wu, Y. (2025). *Information Theory: From Coding to Learning*. Cambridge University Press.

Ribeiro, A. and Giannakis, G. B. (2006a). Bandwidth-constrained distributed estimation for wireless sensor networks-part i: Gaussian case. *IEEE Transactions on Signal Processing*, 54(3):1131–1143.

Ribeiro, A. and Giannakis, G. B. (2006b). Bandwidth-constrained distributed estimation for wireless sensor networks-part ii: Unknown probability density function. *IEEE Transactions on Signal Processing*, 54(7):2784–2796.

Shah, J., Cardone, M., Rush, C., and Dytso, A. (2025). Generalized linear models with 1-bit measurements: Asymptotics of the maximum likelihood estimator. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Shamir, O. (2014). Fundamental limits of online and distributed algorithms for statistical learning and estimation. In *Advances in Neural Information Processing Systems 27*, pages 163–171.

Suresh, A. T., Felix, X. Y., Kumar, S., and McMahan, H. B. (2017). Distributed mean estimation with limited communication. In *International Conference on Machine Learning (ICML)*, pages 3329–3337. PMLR.

Suresh, A. T., Sun, Z., Ro, J., and Yu, F. (2022). Correlated quantization for distributed mean estimation

and optimization. In *International Conference on Machine Learning*, pages 20856–20876. PMLR.

Szabó, B. and van Zanten, H. (2018). Adaptive distributed methods under communication constraints. *The Annals of Statistics*.

Szabó, B. and van Zanten, H. (2020). Distributed function estimation: Adaptation using minimal communication. *Mathematical Statistics and Learning*.

Vargaftik, S., Basat, R. B., Portnoy, A., Mendelson, G., Itzhak, Y. B., and Mitzenmacher, M. (2022). EDEN: Communication-efficient and robust distributed mean estimation for federated learning. In *International Conference on Machine Learning (ICML)*, volume 162, pages 21984–22014. PMLR.

Vargaftik, S., Ben-Basat, R., Portnoy, A., Mendelson, G., Ben-Itzhak, Y., and Mitzenmacher, M. (2021). Drive: One-bit distributed mean estimation. In *Advances in Neural Information Processing Systems*, volume 34, pages 362–377.

Varshney, P. K. (2012). *Distributed detection and data fusion*. Springer Science & Business Media.

Veeravalli, V. V. and Varshney, P. K. (2012). Distributed inference in wireless sensor networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1958):100–117.

Vershynin, R. (2026). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2 edition.

Xiao, J.-J., Ribeiro, A., Luo, Z.-Q., and Giannakis, G. B. (2006). Distributed compression-estimation using wireless sensor networks. *IEEE Signal Processing Magazine*, 23(4):27–41.

Xu, A. and Raginsky, M. (2017). Information-theoretic lower bounds on Bayes risk in decentralized estimation. *IEEE Transactions on Information Theory*, 63(3):1580–1600.

Zaman, A. and Szabó, B. (2022). Distributed nonparametric estimation under communication constraints. *arXiv preprint arXiv:2204.10373*.

Zhang, Y., Duchi, J., Jordan, M. I., and Wainwright, M. J. (2013). Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems 26*, pages 2328–2336.

Zhu, Y. and Lafferty, J. (2018). Distributed nonparametric regression under communication constraints. In *International Conference on Machine Learning (ICML)*, pages 6009–6017. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material

A Proof of Theorem 4 (Performance Guarantee of 1-bit Mean Estimator)

We proceed in several steps as we outlined in Section 2.1.

Step 1 (Narrowing Down the Mean via the Median): We discretize the interval $[-\lambda, \lambda]$ containing $\mathbb{E}[X]$ into a discrete set of points with uniform spacing of σ:[9]

$$\{-\lambda, -\lambda+\sigma, \ldots, -\sigma, 0, \sigma, \ldots, \lambda-\sigma, \lambda\}.$$

We then form estimates $L, U \in \{-\lambda, -\lambda+\sigma, \ldots, \lambda-\sigma, \lambda\}$ using noisy binary search (Gretta and Price, 2024) that satisfy

$$\Pr\left([F(L), F(L+\sigma)] \cap (0.49, 0.5) \text{ is non-empty}\right) \geqslant 1-\delta \tag{16}$$

and

$$\Pr\left([F(U-\sigma), F(U)] \cap (0.5, 0.51) \text{ is non-empty}\right) \geqslant 1-\delta. \tag{17}$$

The algorithm in (Gretta and Price, 2024) achieves this using at most $O\big(\log \frac{\lambda}{\sigma\delta}\big)$ 1-bit queries. Under these high-probability events, the median M satisfies $L \leqslant M \leqslant U$. Since $|\mu - M| \leqslant \sigma$ (e.g., see (Boucheron et al., 2013, Exercise 2.1), we have

$$\mu \in [L-\sigma, U+\sigma].$$

We would like to bound the length of the interval, $(U+\sigma)-(L-\sigma)$. To do so, we consider two different cases: (i) $L+\sigma \geqslant U-\sigma$ and (ii) $L+\sigma < U-\sigma$. In case (i), the interval length is trivially at most 4σ. In case (ii), we claim that the interval length is at most 6σ. Seeking contradiction, suppose the length of interval $(U+\sigma)-(L-\sigma) \geqslant 7\sigma$. Then we must have either

$$\mu-(L-\sigma) \geqslant 3.5\sigma \quad \text{or} \quad (U+\sigma)-\mu \geqslant 3.5\sigma.$$

We will show that $\mu-(L-\sigma) \geqslant 3.5\sigma$ (which implies $\mu - 1.5\sigma \geqslant L+\sigma$) will lead to a contradiction; the case $(U+\sigma)-\mu \geqslant 3.5\sigma$ is similar. Using (16), we have

$$\Pr\left(X \leqslant \mu - 1.5\sigma\right) \geqslant \Pr(X \leqslant L+\sigma) = F_X(L+\sigma) > 0.49.$$

On the other hand, by Chebyshev's inequality, we have

$$\Pr\left(X \leqslant \mu - 1.5\sigma\right) \leqslant \Pr\left(|X-\mu| \geqslant 1.5\sigma\right) \leqslant \frac{1}{1.5^2} < 0.49,$$

which is a contradiction.

Step 2 (Partitioning into Regions): Define $\mu_i := \mathbb{E}\left[X \cdot \mathbf{1}(X \in R_i)\right]$, with the regions R_i defined in (2) and (3). By the linearity of expectation, we have

$$\sum_i \mu_i = \mathbb{E}\left[X \cdot \sum_i \mathbf{1}(X \in R_i)\right] = \mathbb{E}\left[X \cdot \mathbf{1}\left(X \in \bigcup_i R_i\right)\right] = \mathbb{E}[X]. \tag{18}$$

Therefore, it is sufficient to estimate each μ_i.

Step 3 (Ignoring Insignificant Regions): For $i \geqslant 1$, we have $\max(R_i) \leqslant m_i\sigma$ and $\min(R_i) \geqslant m_{i-1}\sigma$, where m_i is as defined in (3). Using $\max(R_i) \leqslant m_i\sigma$, we have

$$\mu_i = \mathbb{E}\left[X \cdot \mathbf{1}(X \in R_i)\right] \leqslant \max(X \in R_i) \cdot \mathbb{E}\left[\mathbf{1}(X \in R_i)\right] \leqslant m_i\sigma \cdot \Pr\left[X \in R_i\right]. \tag{19}$$

[9] For ease of analysis, we assume that λ is an integer multiple of σ.

We now bound $\Pr[X \in R_i]$. First, recall that $\mu \leqslant 3\sigma$ by our "centering" step in Step 1. Using this and $\min(X \in R_i) \geqslant m_{i-1}\sigma$, we have

$$\Pr[X \in R_i] \leqslant \Pr[X \geqslant \min(X \in R_i)] \leqslant \Pr[X \geqslant m_{i-1}\sigma] \leqslant \Pr[X - \mu \geqslant (m_{i-1} - 3)\sigma]. \tag{20}$$

For $i \geqslant 5$, using (20), Chebyshev's inequality, and the definition of m_i (see (3)) gives

$$\Pr[X \in R_i] \leqslant \Pr[X - \mu \geqslant (m_{i-1} - 3)\sigma] \leqslant \Pr[|X - \mu| \geqslant (m_{i-1} - 3)\sigma] \leqslant \frac{1}{(m_{i-1} - 3)^2} = \frac{4}{m_i^2}. \tag{21}$$

Combining (19) and (21), we have for $i \geqslant 5$ that

$$0 \leqslant \mu_i \leqslant \frac{4m_i}{m_i^2}\sigma = 4\sigma m_i^{-1}. \tag{22}$$

By a symmetric argument, we have an analogous bound for $i \leqslant -5$. Combining these, we have

$$|\mu_i| \leqslant 4\sigma m_i^{-1} \quad \text{for } |i| \geqslant 5. \tag{23}$$

Consider the "tail sum" $\sum_{i:|i|>i_{\max}} \mu_i$, where

$$i_{\max} = \min_{i \geqslant 5} \left\{ i : 2^{-i} \leqslant \frac{5\epsilon}{128\sigma} \right\} = \Theta\left(\log\left(\frac{\sigma}{\epsilon}\right)\right). \tag{24}$$

Note that since $\frac{5}{8} \cdot 2^i \leqslant \frac{5}{8} \cdot 2^i + 6 \leqslant m_i \leqslant 2^i$ (which can be verified using (3) and induction), we have

$$m_i^{-1} \leqslant \frac{8}{5} \cdot 2^{-i} \quad \text{and} \quad \sum_{i=1}^{i_{\max}} m_i = \Theta(2^{i_{\max}}) = \Theta(m_{i_{\max}}) = O\left(\frac{\sigma}{\epsilon}\right). \tag{25}$$

Using triangle inequality and (23)–(25), the tail sum can be bounded by

$$\left| \sum_{i:|i|>i_{\max}} \mu_i \right| \leqslant \sum_{i<-i_{\max}} |\mu_i| + \sum_{i>i_{\max}} |\mu_i| \leqslant 8\sigma \sum_{i>i_{\max}} m_i^{-1} \leqslant \frac{64\sigma}{5} \cdot \sum_{i>i_{\max}} 2^{-i} = \frac{64\sigma}{5} \cdot 2^{-i_{\max}} \leqslant \frac{\epsilon}{2}.$$

It follows that

$$\left| \mathbb{E}[X] - \sum_{i:|i|\leqslant i_{\max}} \mu_i \right| = \left| \sum_i \mu_i - \sum_{i:|i|\leqslant i_{\max}} \mu_i \right| = \left| \sum_{i:|i|>i_{\max}} \mu_i \right| \leqslant \frac{\epsilon}{2}, \tag{26}$$

and so it is sufficient to estimate μ_i for $|i| < i_{\max}$; the rest can be estimated as being 0 while only contributing at most $\epsilon/2$ to the error.

Step 4 (Studying Region-Wise Randomized Interval Queries): For each i, write $R_i = [a_i, b_i)$ and let $T_i \sim \mathrm{Unif}(a_i, b_i)$. Using the law of total expectation, we have

$$p_{a_i} := \Pr(X \in [a_i, T_i]) = \mathbb{E}[\mathbf{1}(X \in [a_i, T_i])] = \mathbb{E}[\mathbb{E}[\mathbf{1}(X \in [a_i, T_i]) \mid X]] = \mathbb{E}[\Pr(X \in [a_i, T_i] \mid X)]. \tag{27}$$

Using the CDF of the uniform distribution $T_i \sim \mathrm{Unif}(a_i, b_i)$, we have

$$\Pr(X \in [a_i, T_i] \mid X = x) = \begin{cases} \Pr(T_i \geqslant x) = \dfrac{b_i - x}{b_i - a_i} & \text{if } x \in [a_i, b_i) \\ 0 & \text{otherwise} \end{cases},$$

which can be rewritten as

$$\Pr(X \in [a_i, T_i] \mid X) = \frac{(b_i - X) \cdot \mathbf{1}(X \in R_i)}{b_i - a_i}. \tag{28}$$

Combining (27)–(28) gives

$$p_{a_i} = \mathbb{E}\left[\frac{(b_i - X) \cdot \mathbf{1}(X \in R_i)}{b_i - a_i}\right]. \tag{29}$$

Likewise, similar steps give

$$p_{b_i} := \Pr\left(X \in [T_i, b_i]\right) = \mathbb{E}\left[\frac{(X - a_i) \cdot \mathbf{1}\left(X \in R_i\right)}{b_i - a_i}\right]. \tag{30}$$

Using (29) and (30), linearity of expectation, and basic algebraic manipulations, we can verify that

$$a_i \cdot p_{a_i} + b_i \cdot p_{b_i} = \mathbb{E}\left[X \cdot \mathbf{1}(X \in R_i)\right] = \mu_i. \tag{31}$$

It follows that, to estimate μ_i, it is sufficient to estimate p_{a_i} and p_{b_i}. We denote the estimates as $\hat{p}_{a_i}$ and $\hat{p}_{b_i}$ respectively, and we form them using empirical averages of (randomized) interval queries in the next step.

Step 5 (Estimating p_{a_i} and p_{b_i}): Using the identity $p_{a_i} = \mathbb{E}\left[\mathbf{1}\left(X \in [a_i, T_i]\right)\right]$ in (27), the learner can form an estimate $\hat{p}_{a_i}$ of p_{a_i} as follows:

1. Generate random variables $T_{i,j} \sim \mathrm{Unif}(a_i, b_i)$ for $j = 1, \ldots, n_i$ for some n_i that will be determined later;
2. Ask the agent n_i randomized interval queries "Is $X_{i,j} \in [a_i, T_{i,j}]$?";
3. Compute the empirical averages based on the 1-bit feedback.

The learner can also form an estimate $\hat{p}_{b_i}$ of p_{b_i} using a similar procedure but with queries "Is $X_{i,j} \in [T_{i,j}, b_i]$?". We summarize the estimates as follows:

$$\hat{p}_{a_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{1}\left(X_{i,j} \in [a_i, T_{i,j}]\right) \quad \text{and} \quad \hat{p}_{b_i} = \frac{1}{n_i}\sum_{j=n_i+1}^{2n_i} \mathbf{1}\left(X_{i,j} \in [T_{i,j}, b_i]\right). \tag{32}$$

The number of samples used to form each pair $(\hat{p}_{a_i}, \hat{p}_{b_i})$ is $2n_i$, and the procedure to obtain all pairs $\{(\hat{p}_{a_i}, \hat{p}_{b_i})\}_i$ can be done in a non-adaptive manner. Observe that if the estimates $\hat{p}_{a_i}$ of p_{a_i} and $\hat{p}_{b_i}$ of p_{b_i} satisfy

$$|\hat{p}_{a_i} - p_{a_i}| \leqslant \frac{\epsilon}{(2 \cdot i_{\text{max}}) \cdot (|a_i| + |b_i|)} \quad \text{and} \quad |\hat{p}_{b_i} - p_{b_i}| \leqslant \frac{\epsilon}{(2 \cdot i_{\text{max}}) \cdot (|a_i| + |b_i|)}, \tag{33}$$

then we have

$$|\mu_i - (a_i\hat{p}_{a_i} + b_i\hat{p}_{b_i})| = |(a_i p_{a_i} + b_i p_{b_i}) - (a_i\hat{p}_{a_i} + b_i\hat{p}_{b_i})| = |a_i \cdot (p_{a_i} - \hat{p}_{a_i}) + b_i \cdot (p_{b_i} - \hat{p}_{b_i})| \leqslant \frac{\epsilon}{2 \cdot i_{\text{max}}}, \tag{34}$$

from which it follows that

$$\left|\sum_{i:|i| \leqslant i_{\max}} \mu_i - \sum_{i:|i| \leqslant i_{\max}} (a_i\hat{p}_{a_i} + b_i\hat{p}_{b_i})\right| \leqslant \sum_{i:|i| \leqslant i_{\max}} |(\mu_i - (a_i\hat{p}_{a_i} + b_i\hat{p}_{b_i}))| \leqslant \frac{\epsilon}{2}. \tag{35}$$

Towards establishing (33), we set

$$\delta_i = \frac{\delta}{4 \cdot i_{\text{max}}} = \frac{\delta}{\Theta\left(\log(\sigma/\epsilon)\right)} \quad \text{and} \quad \epsilon_i := \frac{\epsilon}{(2 \cdot i_{\text{max}}) \cdot (|a_i| + |b_i|)} = \frac{\epsilon}{\Theta\left(\log(\sigma/\epsilon) \cdot 2^i \cdot \sigma\right)}, \tag{36}$$

where we recall i_{max} from (24) as well as $\{|a_i|, |b_i|\} = \{m_{i-1}, m_i\}$ and $m_i = \Theta(2^i)$ from (2) and (3).

For $|i| \leqslant 4$, we take

$$n_i = \left\lceil \frac{1}{2\epsilon_i^2} \log\left(\frac{2}{\delta_i}\right)\right\rceil. \tag{37}$$

Recalling p_{a_i} and $\hat{p}_{a_i}$ from (27) and (32), applying Hoeffding's inequality for each $|i| \leqslant 4$ gives:

$$\Pr\left(|p_{a_i} - \hat{p}_{a_i}| > \epsilon_i\right) = \Pr\left(\left|\mathbb{E}\left[\mathbf{1}\left(X \in [a_i, T_i]\right)\right] - \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{1}\left(X_{i,j} \in [a_i, T_{i,j}]\right)\right| > \epsilon_i\right) \leqslant 2\exp\left(-2n_i\epsilon_i^2\right) \leqslant \delta_i. \tag{38}$$

For $|i| \geqslant 5$, we take

$$n_i = \left\lceil \left(\frac{8}{m_i^2}\frac{1}{\epsilon_i^2} + \frac{2}{3}\frac{1}{\epsilon_i}\right) \cdot \log\left(\frac{2}{\delta_i}\right)\right\rceil \geqslant \left\lceil \left(\frac{2\Pr(X \in R_i)}{\epsilon_i^2} + \frac{2}{3}\frac{1}{\epsilon_i}\right) \cdot \log\left(\frac{2}{\delta_i}\right)\right\rceil, \tag{39}$$

where the inequality follows from (21). Applying Bernstein's inequality (Vershynin, 2026)[Theorem 2.9.5] to the i.i.d. mean zero bounded random variables

$$Y_{i,j} := \mathbf{1}\left(X_{i,j} \in [a_i, T_{i,j}]\right) - \mathbb{E}\left[\mathbf{1}\left(X \in [a_i, T_i]\right)\right],$$

we obtain:

$$\begin{aligned}
\Pr\left(|p_{a_i} - \hat{p}_{a_i}| > \epsilon_i\right) &= \Pr\left(\left|\mathbb{E}\left[\mathbf{1}\left(X \in [a_i, T_i]\right)\right] - \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{1}\left(X_{i,j} \in [a_i, T_{i,j}]\right)\right| > \epsilon_i\right) && (40)\\
&= \Pr\left(\left|\frac{1}{n_i}\sum_{j=1}^{n_i} Y_{i,j}\right| > \epsilon_i\right) && (41)\\
&\leqslant 2\exp\left(-\frac{n_i\epsilon_i^2}{2\,\mathbb{E}\left[Y_{ij}^2\right] + \frac{2}{3}\epsilon_i}\right) && (42)\\
&= 2\exp\left(-\frac{n_i\epsilon_i^2}{2\,\mathrm{Var}\left(\mathbf{1}\left(X \in [a_i, T_i]\right)\right) + \frac{2}{3}\epsilon_i}\right) && (43)\\
&\leqslant 2\exp\left(-\frac{n_i}{2\Pr(X \in R_i)\frac{1}{\epsilon_i^2} + \frac{2}{3}\frac{1}{\epsilon_i}}\right) && (44)\\
&\leqslant \delta_i, && (45)
\end{aligned}$$

where in (43) we use $X_{i,j} \overset{d}{=} X$ and $T_{i,j} \overset{d}{=} T_i$ to derive

$$\mathbb{E}\left[Y_{ij}^2\right] = \mathbb{E}\left[\left(\mathbf{1}\left(X_{i,j} \in [a_i, T_{i,j}]\right) - \mathbb{E}\left[\mathbf{1}\left(X \in [a_i, T_i]\right)\right]\right)^2\right] = \mathrm{Var}\left(\mathbf{1}\left(X \in [a_i, T_i]\right)\right),$$

and in (44) we use $\mathrm{Var}(\mathrm{Ber}(p)) = p(1-p) \leqslant p$ and $T_i \sim \mathrm{Unif}(a_i, b_i)$ to derive

$$\mathrm{Var}\left(\mathbf{1}\left(X \in [a_i, T_i]\right)\right) \leqslant \Pr\left(X \in [a_i, T_i]\right) \leqslant \Pr\left(X \in [a_i, b_i]\right) = \Pr(X \in R_i).$$

Likewise, we have $\Pr\left(|p_{b_i} - \hat{p}_{b_i}| > \epsilon_i\right) \leqslant \delta_i$.

We now substitute δ_i and ϵ_i from (36) into n_i. For $|i| \leqslant 4$, substituting these into n_i from (37) gives

$$n_i = O\left(\underbrace{2^{2i}}_{\text{bounded by } 2^8} \cdot \frac{\sigma^2}{\epsilon^2}\log^2\left(\frac{\sigma}{\epsilon}\right)\log\left(\frac{\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right) = O\left(\frac{\sigma^2}{\epsilon^2}\log^2\left(\frac{\sigma}{\epsilon}\right)\log\left(\frac{\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right). \tag{46}$$

For $4 \leqslant |i| \leqslant i_{\max}$, substituting into n_i from (39) and recalling $m_i = \Theta(2^i)$ from (3) gives

$$\begin{aligned}
n_i &= O\left(\left(\frac{1}{2^{2i}}\frac{2^{2i}\sigma^2}{\epsilon^2} \cdot \log^2\left(\frac{\sigma}{\epsilon}\right) + \frac{2^i\sigma}{\epsilon} \cdot \log\left(\frac{\sigma}{\epsilon}\right)\right) \cdot \log\left(\frac{\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right)\\
&= O\left(\left(\frac{\sigma^2}{\epsilon^2} \cdot \log\left(\frac{\sigma}{\epsilon}\right) + \frac{2^i\sigma}{\epsilon}\right) \cdot \log\left(\frac{\sigma}{\epsilon}\right) \cdot \log\left(\frac{\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right).
\end{aligned} \tag{47}$$

Summing up all n_i, we obtain

$$\begin{aligned}\sum_{i:|i|\leqslant i_{\max}} n_i &= 2\left(\sum_{1\leqslant i\leqslant 4} n_i + \sum_{5\leqslant i\leqslant i_{\max}} n_i\right)\\ &= O\left(\left(i_{\max}\frac{\sigma^2}{\epsilon^2}\cdot\log\left(\frac{\sigma}{\epsilon}\right) + \sum_{i\leqslant i_{\max}}\frac{2^i\sigma}{\epsilon}\right)\cdot\log\left(\frac{\sigma}{\epsilon}\right)\cdot\log\left(\frac{\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right)\\ &= O\left(\left(\frac{\sigma^2}{\epsilon^2}\cdot\log^2\left(\frac{\sigma}{\epsilon}\right) + \frac{2^{i_{\max}}\sigma}{\epsilon}\right)\cdot\log\left(\frac{\sigma}{\epsilon}\right)\cdot\log\left(\frac{\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right)\\ &= O\left(\left(\frac{\sigma^2}{\epsilon^2}\cdot\log^2\left(\frac{\sigma}{\epsilon}\right) + \frac{\sigma^2}{\epsilon^2}\right)\cdot\log\left(\frac{\sigma}{\epsilon}\right)\cdot\log\left(\frac{\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right)\\ &= O\left(\frac{\sigma^2}{\epsilon^2}\cdot\log^3\left(\frac{\sigma}{\epsilon}\right)\cdot\log\left(\frac{\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right),\end{aligned} \tag{48}$$

where the second last step follows since $2^{i_{\max}} = O\left(\frac{\sigma}{\epsilon}\right)$ (see (25)).

B Equivalence of Randomized Interval Queries and Stochastic Rounding in Step 4

In this appendix, we show that our randomized interval queries from Step 4 of Section 2.1 can be interpreted as performing a form of binary stochastic quantization. Note that this connection is presented purely for the sake of intuition, and it is not needed in the proof of Theorem 4.

For each i, let $R_i = [a_i, b_i)$ as before, and define the stochastic quantizer $\mathrm{SQ}_i(\cdot)$ as follows:

$$\mathrm{SQ}_i(x) = \begin{cases} 0 & \text{if } x \notin R_i \\ a_i & \text{with probability } \frac{b_i - x}{b_i - a_i} \text{ if } x \in R_i \\ b_i & \text{with probability } \frac{x - a_i}{b_i - a_i} \text{ if } x \in R_i. \end{cases} \tag{49}$$

As before, we write $p_{a_i} := \Pr\left(X \in [a_i, T_i]\right)$ and $p_{b_i} := \Pr\left(X \in [T_i, b_i]\right)$. We now show that p_{a_i} (resp. p_{b_i}) is equivalent to the probability of X being in R_i and getting rounded down to a_i (resp. rounded up to b_i) by SQ_i, i.e.,

$$p_{a_i} = \Pr(X \in R_i \cap \mathrm{SQ}_i(X) = a_i) \quad \text{and} \quad p_{b_i} = \Pr(X \in R_i \cap \mathrm{SQ}_i(X) = b_i).$$

Using (49) as well as standard properties of conditional probability, indicator functions, Bernoulli random variables, and linearity of expectation, we have

$$\begin{aligned}\Pr(X \in R_i \cap \mathrm{SQ}_i(X) = a_i) &= \Pr(X \in R_i)\cdot\Pr\left(\mathrm{SQ}_i(X) = a_i \mid X \in R_i\right)\\ &= \Pr(X \in R_i)\cdot\mathbb{E}\left[\mathbf{1}\left(\mathrm{SQ}_i(X) = a_i\right) \mid X \in R_i\right]\\ &= \Pr(X \in R_i)\cdot\mathbb{E}\left[\frac{b_i - X}{b_i - a_i}\,\middle|\, X \in R_i\right]\\ &= \frac{b_i\cdot\Pr(X \in R_i) - \mathbb{E}\left[X \mid X \in R_i\right]\cdot\Pr(X \in R_i)}{b_i - a_i}.\end{aligned} \tag{50}$$

Moreover, using $p_{a_i} = \mathbb{E}\left[(b_i - X)/(b_i - a_i)\cdot\mathbf{1}\left(X \in R_i\right)\right]$ (see (29)) and (50) as well as linearity of expectation and law of total expectation, we have

$$\begin{aligned}p_{a_i} &= \mathbb{E}\left[\frac{(b_i - X)\cdot\mathbf{1}\left(X \in R_i\right)}{b_i - a_i}\right]\\ &= \frac{b_i\cdot\mathbb{E}\left[\mathbf{1}\left(X \in R_i\right)\right] - \mathbb{E}\left[X\cdot\mathbf{1}\left(X \in R_i\right)\right]}{b_i - a_i}\\ &= \frac{b_i\cdot\Pr(X \in R_i) - \mathbb{E}\left[X \mid X \in R_i\right]\cdot\Pr(X \in R_i)}{b_i - a_i}\\ &= \Pr(X \in R_i \cap \mathrm{SQ}_i(X) = a_i)\end{aligned} \tag{51}$$

as desired. Analogous steps give $p_{b_i} = \Pr(X \in R_i \cap \mathrm{SQ}_i(X) = b_i)$.

C Lower Bound and Adaptivity Gap

C.1 Proof of Theorem 6 (General Lower Bound)

Even if the (ϵ, δ)-PAC estimator has no 1-bit constraint, the lower bound $n = \Omega\big(\frac{\sigma^2}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\big)$ is well known. For instance, this can be derived via a reduction to distinguishing two Bernoulli distributions (Lee, 2020, Section 4). Therefore, it is sufficient for us to establish that $n = \Omega\left(\log \frac{\lambda}{\sigma}\right)$.

We create $N = \Theta(\lambda/\sigma)$ instances of "hard-to-distinguish" distribution pairs, which we will reuse in the proof of Theorem 7 in Appendix C.2. Divide $[-\lambda, \lambda]$ into a grid of $N = \lambda/\sigma - 1$ "center-points" spaced 2σ apart,[10] i.e., the center-points are

$$c_j = -\lambda + 2j\sigma \quad \text{for each } j = 1, 2 \ldots, N. \tag{52}$$

For each instance j, we define two probability distributions $D_{j,-}$ and $D_{j,+}$, each with a two-point support set $\{c_j - \sigma/2, c_j + \sigma/2\}$, as follows:

$$\begin{aligned} D_{j,-}\colon \Pr\left(X = c_j + \frac{\sigma}{2}\right) &= \frac{1}{2} - \frac{\epsilon}{\sigma} = 1 - \Pr\left(X = c_j - \frac{\sigma}{2}\right) \implies \mathbb{E}[X] = c_j - \epsilon \\ D_{j,+}\colon \Pr\left(X = c_j + \frac{\sigma}{2}\right) &= \frac{1}{2} + \frac{\epsilon}{\sigma} = 1 - \Pr\left(X = c_j - \frac{\sigma}{2}\right) \implies \mathbb{E}[X] = c_j + \epsilon. \end{aligned} \tag{53}$$

We readily observe the following:

- By the assumption $\epsilon < \frac{\sigma}{2}$, each each of these $2N$ distributions has their mean in $[-\lambda, \lambda]$;
- Since a distribution on $[a, b]$ as variance at most $\frac{(b-a)^2}{4}$, each of these $2N$ distributions has variance at most σ^2.

Therefore, when the distributions are restricted to only these $2N$ distributions, the task of being able to form an ϵ-good estimation of the true mean of each unknown underlying distribution is at least as hard as being able to distinguish the distributions from each other.[11] We proceed to establish a lower bound for this goal of *identification*, also known as *multiple hypothesis testing*.

Let Θ be a uniform random variable over the $2N$ distributions, which implies

$$H(\Theta) = \log(2N), \tag{54}$$

where $H(X) := -\sum_{x \in \mathcal{X}} p(x) \log p(x)$ is the entropy function. Fix an adaptive mean estimator that makes n queries, and let $Y^n = (Y_1, \ldots, Y_n)$ be the resulting binary responses. Using the chain rule for mutual information (see e.g. (Polyanskiy and Wu, 2025, Theorem 3.7)) and the fact that each query yields at most 1 bit of information, we have

$$I(\Theta; Y^n) = \sum_{k=1}^{n} I\big(\Theta; Y_k \mid Y^{k-1}\big) \leqslant \sum_{k=1}^{n} H\big(Y_k \mid Y^{k-1}\big) \leqslant \sum_{k=1}^{n} H(Y_k) \leqslant \sum_{k=1}^{n} 1 = n. \tag{55}$$

Moreover, Fano's inequality (see (Polyanskiy and Wu, 2025, Theorem 3.12)) gives:

$$H(\Theta \mid Y^n) \leqslant H_2(\delta) + \delta \log(2N - 1) \leqslant 1 + \delta \log(2N), \tag{56}$$

where δ is the error probability and $H_2(p) = -p \log_p -(1-p)\log(1-p)$ is the binary entropy function. Using (54)–(56) and the definition of mutual information, we obtain

$$n \geqslant I(\Theta; Y^n) = H(\Theta) - H(\Theta \mid Y^n) \geqslant \log(2N) - 1 - \delta \log(2N) = (1-\delta)\log(2N) - 1. \tag{57}$$

Combining this with $N = \Theta(\lambda/\sigma)$, we have

$$n = \Omega((1-\delta)\log N) = \Omega\left(\log \frac{\lambda}{\sigma}\right)$$

as desired.

[10] For convenience, we assume that λ is an integer multiple of 2σ. This is justified by a simple rounding argument and the fact that when $\lambda = \Theta(\sigma)$ the $\Omega\big(\log \frac{\lambda}{\sigma}\big)$ lower bound is trivial.

[11] Strictly speaking this is true when the algorithm is required to attain accuracy *strictly smaller* than ϵ, rather than *smaller or equal*, but this distinction clearly has no impact on the final result stated using $O(\cdot)$ notation, and by ignoring it we can avoid cumbersome notation.

C.2 Proof of Theorem 7 (Adaptivity Gap)

We consider the same instance as that of Section C.1, and accordingly re-use the notation therein. Before proving Theorem 7, we first introduce the idea of an interval query being "informative" or "uninformative" for distinguishing between the distributions $D_{j,-}$ and $D_{j,+}$.

Definition 15 (Informative Interval Queries)**.** For a fixed interval query $Q =$ "Is $X \in [a,b]$?", we say that Q is informative for the j-th pair of distributions $(D_{j,-}, D_{j,+})$ if its binary feedback $B = \mathbf{1}\{X \in [a,b]\}$ satisfies

$$\Pr_{X\sim D_{j,-}}(B=1) \neq \Pr_{X\sim D_{j,+}}(B=1).$$

Otherwise, Q is said to be uninformative.

The following lemma shows that each interval query can be simultaneously informative for at most two different pairs.

Lemma 16. An interval query $Q =$ "Is $X \in [a,b]$?"can be simultaneously informative for at most two different $(D_{j,-}, D_{j,+})$ pairs, i.e., at most two different values of j.

Proof of Lemma 16. The claim follows from the following two facts:

1. For a fixed distribution pair (indexed by j), an interval query $Q =$ "Is $X \in [a,b]$?" is informative for distinguishing between $D_{j,-}$ and $D_{j,+}$ only if $[a,b]$ contains exactly one of the two support points $\{c_j \pm \sigma/2\}$, i.e., $\big|[a,b] \cap \{c_j \pm \sigma/2\}\big| = 1$.

2. There are at most two indices j for which $\big|[a,b] \cap \{c_j \pm \sigma/2\}\big| = 1$.

Fact 1 can be verified by analyzing the binary feedback $B = \mathbf{1}\{X \in [a,b]\}$ for all cases of $[a,b] \cap \{c_j \pm \sigma/2\}$:

$$\big|\,[a,b] \cap \{c_j \pm \sigma/2\}\big| \in \{0,2\} \implies \Pr_{X\sim D_{j,-}}(B=1) = \Pr_{X\sim D_{j,+}}(B=1) \implies Q \text{ is uninformative,}$$

and

$$\big|\,[a,b] \cap \{c_j \pm \sigma/2\}\big| = 1 \implies \big|\Pr_{X\sim D_{j,-}}(B=1) - \Pr_{X\sim D_{j,+}}(B=1)\big| = \frac{2\epsilon}{\sigma} \implies Q \text{ is informative.} \tag{58}$$

For Fact 2, we first observe from (52) that the support points of all $2N$ distributions satisfy

$$c_1 - \frac{\sigma}{2} < c_1 + \frac{\sigma}{2} < c_2 - \frac{\sigma}{2} < \cdots < c_N - \frac{\sigma}{2} < c_N + \frac{\sigma}{2},$$

with each pair j having a unique disjoint interval $(c_j - \sigma/2, c_j + \sigma/2)$ between its support points. An interval $[a,b]$ satisfies $\big|[a,b] \cap \{c_j \pm \sigma/2\}\big| = 1$ if and only if exactly one endpoint of $[a,b]$ lies in the interval $(c_j - \sigma/2, c_j + \sigma/2)$. Since the gaps are disjoint and $[a,b]$ has only two endpoints, it follows that at most two indices j satisfy $\big|[a,b] \cap \{c_j \pm \sigma/2\}\big| = 1$. □

Proof of Theorem 7. Consider an arbitrary algorithm that makes n non-adaptive interval queries. Recall the set of $2N$ distributions $\{D_{j,-}, D_{j,+}\}_{j=1}^N \subseteq \mathcal{D}(\lambda,\sigma)$ constructed in the proof of Theorem 6, where $N = \lambda/\sigma - 1$. We will again establish a lower bound for this "hard subset" of distributions, but with different details to exploit the assumption of non-adaptive interval queries.

Recall from Section C.1 that the means of the $2N$ distributions are pairwise separated by 2ϵ or more, and thus, attaining ϵ-accuracy implies being able to identify the underlying distribution from the hard subset. We proceed to establish a lower bound for this goal of identification (multiple hypothesis testing).

Suppose that the true distribution is drawn uniformly at random from the $2N$ distributions in the hard subset. By Yao's minimax principle, the worst-case error probability is lower bounded by the average-case error probability of the best *deterministic* strategy, so we may assume that the algorithm is deterministic (in the choice of queries and the procedure for forming the final estimate).

Letting $(\hat{j}, \hat{s})$ be the estimated index (in $\{1, \ldots, N\}$) and sign (in $\{1, -1\}$), the average-case error probability is given by

$$\Pr(\text{error}) = \frac{1}{2N} \sum_{j=1}^{N} \sum_{s \in \{+1,-1\}} \Pr_{j,s}((\hat{j}, \hat{s}) \neq (j, s)) \tag{59}$$

$$\geqslant \frac{1}{N} \sum_{j=1}^{N} \Bigg(\underbrace{\frac{1}{2}\Pr_{j,+}\big(\hat{s} \neq 1\big) + \frac{1}{2}\Pr_{j,-}\big(\hat{s} \neq -1\big)}_{=:\Pr_j(\text{error})} \Bigg), \tag{60}$$

where $\Pr_{j,s}$ denotes probability when the underlying distribution is $D_{j,s}$.

For each $j = 1, \ldots, N$, we define n_j to be the algorithm's total number of interval queries that are informative (in the sense of Definition 15) for distinguishing between $D_{j,-}$ and $D_{j,+}$. Since the algorithm is deterministic and the n queries are assumed to be non-adaptive (i.e., they must all be chosen in advance), it follows that the values $\{n_j\}_{j=1}^{N}$ are also deterministic.

Recall from (58) that each informative query provides binary feedback that follows either $\mathrm{Bern}(p_+)$ or $\mathrm{Bern}(p_-)$, where $p_+ = 1/2 + \epsilon/\sigma$ and $p_- = 1/2 - \epsilon/\sigma = 1 - p_+$. Distinguishing between these two cases is a *binary hypothesis testing* problem, and the associated error probability $\Pr_j(\text{error})$ is given by the j-th summand in (60).

Using standard binary hypothesis testing lower bounds (Lee, 2020, Theorem 11.9), we have[12]

$$\Pr_j(\text{error}) > \exp\left(-c' \cdot n_j \cdot d_H^2(p_+, p_-)\right) \tag{61}$$

for some constant c', where $d_H^2(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2$ is the Squared Hellinger distance. For $\mathrm{Bern}(p_+)$ and $\mathrm{Bern}(p_-)$, we have the following standard calculation:

$$d_H^2(p_+, p_-) = \left(\sqrt{p_+} - \sqrt{p_-}\right)^2 = \left(\frac{p_+ - p_-}{\sqrt{p_+} + \sqrt{p_-}}\right)^2 = \frac{|p_+ - p_-|^2}{\left(1 + 2\sqrt{p_+ p_-}\right)^2} = \Theta\left(|p_+ - p_-|^2\right) = \Theta\left(\frac{\epsilon^2}{\sigma^2}\right), \tag{62}$$

where the equalities follow from the facts that $p_+ + p_- = 1$ and $p_+ p_- \in [0, 1/4]$. Combining (61) and (62), we obtain

$$\Pr_j(\text{error}) > \exp\left(-c'' \cdot \frac{n_j\, \epsilon^2}{\sigma^2}\right) \tag{63}$$

for some constant $c'' > 0$. Applying Jensen's inequality (since exp is convex) and using $\sum_{j=1}^{N} n_j \leqslant 2n$ (see Lemma 16), it follows that

$$\frac{1}{N}\sum_{j=1}^{N} \Pr_j(\text{error}) > \frac{1}{N}\sum_{j=1}^{N} \exp\left(-c'' \cdot \frac{n_j\, \epsilon^2}{\sigma^2}\right) \geqslant \exp\left(-c'' \cdot \frac{\epsilon^2}{\sigma^2} \cdot \frac{1}{N}\sum_{j=1}^{N} n_j\right) \geqslant \exp\left(-c'' \cdot \frac{\epsilon^2}{\sigma^2} \cdot \frac{2n}{N}\right).$$

It follows that if

$$n < \frac{1}{4c''} \cdot \frac{\lambda\sigma}{\epsilon^2} \log\left(\frac{1}{\delta}\right) = \frac{1}{4c''} \cdot \frac{\lambda}{\sigma} \cdot \frac{\sigma^2}{\epsilon^2} \cdot \log\left(\frac{1}{\delta}\right) = \frac{1}{4c''} \cdot (N+1) \cdot \frac{\sigma^2}{\epsilon^2} \cdot \log\left(\frac{1}{\delta}\right) \leqslant \frac{N}{2c''}\frac{\sigma^2}{\epsilon^2} \log\left(\frac{1}{\delta}\right),$$

then the average error probability is lower bounded by

$$\frac{1}{N}\sum_{j=1}^{N} \Pr_j(\text{error}) > \exp\left(-c'' \cdot \frac{\epsilon^2}{\sigma^2} \cdot \frac{2n}{N}\right) \geqslant \exp\left(\log\left(\frac{1}{\delta}\right)\right) = \delta.$$

Therefore, to attain an error probability no higher than δ, we must have

$$n = \Omega\left(\frac{\lambda\sigma}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right)$$

as desired. □

[12] We have re-arranged their result to express other quantities in term of the error probability.

D Improvements for Random Variables with Stronger Tail Decay

D.1 Proof of Theorem 8 (Improvement with Finite Higher-order Central Moments)

The main difference compared to the case with only bounded variance is that we now have a better tail bound through the higher-moment Chebyshev's inequality:

$$\Pr(|X-\mu| \geqslant t) \leqslant \frac{\mathbb{E}|X-\mu|^k}{t^k} \leqslant \frac{\sigma^k}{t^k}. \tag{64}$$

Since the proof mostly follows that of Theorem 4, we focus our attention on the steps that are different.

Modified Step 2: We let the width of the regions R_i grow doubly exponentially instead of exponentially. Specifically, we still let R_i have the form in (2), but we modify m_i in (3) as follows:

$$m_i = \begin{cases} 0 & \text{if } i = 0 \\ 2^{(k/2)^i} & \text{if } 1 \leqslant i \leqslant 4 \\ (m_{i-1}-3)^{k/2} & \text{if } i \geqslant 5 \end{cases} \tag{65}$$

Note that the last case can be expanded as $\underbrace{\left(\left(\left(2^{(k/2)^4}-3\right)^{k/2}-3\right)^{k/2}\cdots-3\right)^{k/2}}_{i-4 \text{ times}}$, from which we can verify by induction that m_i scales doubly exponentially according to $\Theta\big(2^{(k/2)^i}\big)$.

Modified Step 3: Because $m_i = \Theta\big(2^{(k/2)^i}\big)$, we expect $i_{\max}$ to have $\log_{k/2}\log(\sigma/\epsilon)$ scaling instead of $\log(\sigma/\epsilon)$. We proceed to show this. For $|i| \geqslant 5$, using steps similar to those in (19)–(23), but with higher-moment Chebyshev's inequality (64) and the modified definition of m_i gives

$$\Pr[X \in R_i] \leqslant \Pr[X-\mu \geqslant (m_{i-1}-3)\sigma] \leqslant \frac{1}{(m_{i-1}-3)^k} = \frac{1}{m_i^2}, \tag{66}$$

which implies

$$|\mu_i| \leqslant \sigma m_i^{-1} \text{ for } |i| \geqslant 5. \tag{67}$$

Consider the "tail sum" $\sum_{i:|i|>i_{\max}} \mu_i$, where

$$i_{\max} = \min\left\{i : m_{i+1}^{-1} \leqslant \frac{\epsilon}{8\sigma}\right\} = \min\left\{i : m_{i+1} \geqslant \frac{8\sigma}{\epsilon}\right\} = \Theta\left(\log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right)\right). \tag{68}$$

Note that due to the "super-geometric" growth of m_i, we have

$$m_{i+1} \geqslant \frac{m_i}{2} \quad \text{and} \quad \sum_{i=1}^{i_{\max}} m_i = \Theta(m_{i_{\max}}) = O\left(\frac{\sigma}{\epsilon}\right). \tag{69}$$

Using (67)–(69), the tail sum can be bounded by

$$\left|\sum_{i<-i_{\max}} \mu_i + \sum_{i>i_{\max}} \mu_i\right| \leqslant \sum_{i<-i_{\max}} |\mu_i| + \sum_{i>i_{\max}} |\mu_i| \leqslant 2\sigma \sum_{i>i_{\max}} m_i^{-1} \leqslant 2\sigma\left(\frac{\epsilon}{8\sigma} + \frac{\epsilon}{16\sigma} + \frac{\epsilon}{32\sigma} + \cdots\right) \leqslant \frac{\epsilon}{2}.$$

It follows that

$$\left|\mathbb{E}[X] - \sum_{i:|i|\leqslant i_{\max}} \mu_i\right| = \left|\sum_i \mu_i - \sum_{i:|i|\leqslant i_{\max}} \mu_i\right| = \left|\sum_{i:|i|>i_{\max}} \mu_i\right| \leqslant \frac{\epsilon}{2}, \tag{70}$$

and so it is sufficient to estimate μ_i for $|i| < i_{\max}$.

Modified Step 5: We adjust δ_i and ϵ_i according to the new m_i and $i_{\max}$, which gives us a smaller n_i and $\sum_{i:|i|\leqslant i_{\max}} n_i$. Specficially, we set

$$\delta_i = \frac{\delta}{4 \cdot i_{\max}} = \frac{\delta}{\Theta\left(\log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right)\right)} \quad \text{and} \quad \epsilon_i := \frac{\epsilon}{(2 \cdot i_{\max}) \cdot (|a_i| + |b_i|)} = \frac{\epsilon}{\Theta\left(\log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right) \cdot m_i \cdot \sigma\right)}. \tag{71}$$

For $|i| \leqslant 4$, we take $n_i = \left\lceil \frac{1}{2\epsilon_i^2}\log\left(\frac{2}{\delta_i}\right)\right\rceil$, and for $|i| \geqslant 5$, we take

$$n_i = \left\lceil \left(\frac{2}{m_i^2}\frac{1}{\epsilon_i^2} + \frac{2}{3}\frac{1}{\epsilon_i}\right) \cdot \log\left(\frac{2}{\delta_i}\right)\right\rceil \geqslant \left\lceil \left(\frac{2\Pr(X \in R_i)}{\epsilon_i^2} + \frac{2}{3}\frac{1}{\epsilon_i}\right) \cdot \log\left(\frac{2}{\delta_i}\right)\right\rceil, \tag{72}$$

where the inequality follows from (66). Applying Hoeffding's inequality for each $|i| \leqslant 4$ as in (38) and Bernstein's inequality for each $|i| \geqslant 5$ as in (40)–(45), we obtain:

$$\Pr\left(|p_{a_i} - \hat{p}_{a_i}| > \epsilon_i\right) \leqslant \delta_i \quad \text{and} \quad \Pr\left(|p_{b_i} - \hat{p}_{b_i}| > \epsilon_i\right) \leqslant \delta_i. \tag{73}$$

To substitute δ_i and ϵ_i from (71) into n_i, we use steps similar to (46) and (47), which gives:

$$n_i = O\left(\frac{\sigma^2}{\epsilon^2}\left(\log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right)\right)^2 \log\left(\frac{\log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right) \quad \text{for } |i| \leqslant 4 \tag{74}$$

and

$$n_i = O\left(\left(\frac{\sigma^2}{\epsilon^2} \cdot \log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right) + \frac{m_i\sigma}{\epsilon}\right) \cdot \log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right) \cdot \log\left(\frac{\log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right) \quad \text{for } 5 \leqslant |i| \leqslant i_{\max}. \tag{75}$$

Summing up all n_i as in (48), we obtain

$$\begin{aligned}
\sum_{i:|i|\leqslant i_{\max}} n_i &= O\left(\left(i_{\max}\frac{\sigma^2}{\epsilon^2} \cdot \log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right) + \sum_{i\leqslant i_{\max}} \frac{m_i\sigma}{\epsilon}\right) \cdot \log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right) \cdot \log\left(\frac{\log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right) \\
&= O\left(\left(\frac{\sigma^2}{\epsilon^2} \cdot \left(\log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right)\right)^2 + \frac{\sigma}{\epsilon} \cdot \frac{\sigma}{\epsilon}\right) \cdot \log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right) \cdot \log\left(\frac{\log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right) \\
&= O\left(\frac{\sigma^2}{\epsilon^2} \cdot \left(\log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right)\right)^3 \cdot \log\left(\frac{\log_{k/2}\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right),
\end{aligned} \tag{76}$$

where the second step follows from (69).

D.2 Proof of Theorem 9 (Improvement for Sub-Gaussian Random Variables)

The main difference is that we now have an even faster tail decay through the sub-Gaussian tail bound (8).

Modified Step 2: Due to the strong tail decay of sub-Gaussian random variables, we can let the width of regions R_i grow much more rapidly. Specifically, we keep R_i as in (2) but modify m_i in (3) as follows:

$$m_i = \begin{cases} 0 & \text{if } i = 0 \\ \exp\left(\frac{m_{i-1}^2}{2}\right) & \text{if } 1 \leqslant i \leqslant 4 \\ \exp\left(\frac{(m_{i-1}-3)^2}{4}\right) & \text{if } i \geqslant 5. \end{cases} \tag{77}$$

Note that m_i scales according to a tower of exponentials of height i, which can be verified by induction:

$$m_i = \Theta\left(\underbrace{\exp\left(\exp\left(\cdots\exp\left(\Theta(1)\right)\right)\right)}_{i \text{ times}}\right). \tag{78}$$

Modified Step 3: Because m_i scales according to a tower of exponentials, we expect $i_{\max}$ to have $\log^*(\sigma/\epsilon)$ scaling. Because the arguments are almost identical to those in modified Step 3 of Appendix D.1 (improvement for random variables with finite k-th central moment), we will omit most of the details. The main difference is that we use the sub-Gaussian bound (8) and the modified definition of m_i (see (77)) in obtaining

$$\Pr[X \in R_i] \leqslant \Pr[X - \mu \geqslant (m_{i-1} - 3)\,\sigma] \leqslant \exp\left(-\frac{(m_{i-1}-3)^2}{2}\right) = \frac{1}{m_i^2}. \tag{79}$$

Consequently, we have

$$i_{\max} = \min\left\{i : \frac{\sigma}{m_{i+1}} \leqslant \frac{\epsilon}{8}\right\} = \Theta\left(\log^*\left(\frac{\sigma}{\epsilon}\right)\right) \quad \text{and} \quad \sum_{i=1}^{i_{\max}} m_i = \Theta(m_{i_{\max}}) = O\left(\frac{\sigma}{\epsilon}\right). \tag{80}$$

Modified Step 5: We adjust δ_i and ϵ_i according to the new m_i and $i_{\max}$, which gives us a smaller n_i and $\sum_{i:|i|\leqslant i_{\max}} n_i$. As the steps are almost identical to those in modified Step 5 of Appendix D.1, we will omit most of the details for brevity. We set

$$\delta_i = \frac{\delta}{4 \cdot i_{\max}} = \frac{\epsilon}{\Theta\left(\log^*\left(\frac{\sigma}{\epsilon}\right)\right)} \quad \text{and} \quad \epsilon_i := \frac{\epsilon}{(2 \cdot i_{\max}) \cdot (|a_i| + |b_i|)} = \frac{\epsilon}{\Theta\left(\log^*\left(\frac{\sigma}{\epsilon}\right) \cdot m_i \cdot \sigma\right)}, \tag{81}$$

and take

$$n_i = \begin{cases} \left\lceil \frac{1}{2\epsilon_i^2} \log\left(\frac{2}{\delta_i}\right) \right\rceil = O\left(\frac{\sigma^2}{\epsilon^2}\left(\log^*\left(\frac{\sigma}{\epsilon}\right)\right)^2 \log\left(\frac{\log^*\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right) & \text{if } |i| \leqslant 4 \\ \left\lceil \left(\frac{2}{m_i^2}\frac{1}{\epsilon_i^2} + \frac{2}{3}\frac{1}{\epsilon_i}\right) \cdot \log\left(\frac{2}{\delta_i}\right) \right\rceil = O\left(\left(\frac{\sigma^2}{\epsilon^2} \cdot \log^*\left(\frac{\sigma}{\epsilon}\right) + \frac{m_i \sigma}{\epsilon}\right) \cdot \log^*\left(\frac{\sigma}{\epsilon}\right) \cdot \log\left(\frac{\log^*\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right) & \text{if } |i| \geqslant 5. \end{cases} \tag{82}$$

Applying Hoeffding's inequality for each $|i| \leqslant 4$ as in (38) and Bernstein's inequality for each $|i| \geqslant 5$ as in (40)–(45) gives

$$\Pr\left(|p_{a_i} - \hat{p}_{a_i}| > \epsilon_i\right) \leqslant \delta_i \quad \text{and} \quad \Pr\left(|p_{b_i} - \hat{p}_{b_i}| > \epsilon_i\right) \leqslant \delta_i \tag{83}$$

Summing up all n_i, we obtain

$$\begin{aligned} \sum_{i:|i|\leqslant i_{\max}} n_i &= O\left(\left(i_{\max}\frac{\sigma^2}{\epsilon^2} \cdot \log^*\left(\frac{\sigma}{\epsilon}\right) + \sum_{i \leqslant i_{\max}} \frac{m_i \sigma}{\epsilon}\right) \cdot \log^*\left(\frac{\sigma}{\epsilon}\right) \cdot \log\left(\frac{\log^*\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right) \\ &= O\left(\left(\frac{\sigma^2}{\epsilon^2} \cdot \left(\log^*\left(\frac{\sigma}{\epsilon}\right)\right)^2 + \frac{\sigma}{\epsilon} \cdot \frac{\sigma}{\epsilon}\right) \cdot \log^*\left(\frac{\sigma}{\epsilon}\right) \cdot \log\left(\frac{\log\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right) \\ &= O\left(\frac{\sigma^2}{\epsilon^2} \cdot \left(\log^*\left(\frac{\sigma}{\epsilon}\right)\right)^3 \cdot \log\left(\frac{\log^*\left(\frac{\sigma}{\epsilon}\right)}{\delta}\right)\right), \end{aligned} \tag{84}$$

where the second step follows from (80).

E Unknown Parameters

E.1 Proof of Theorem 10 (Unknown Target Accuracy)

To establish that $\epsilon_T = O(\epsilon^*)$, we will compare the last round T (see (10)) and $\tau^* := \log_2(\epsilon_0/\epsilon^*) = \log_2(\sigma/2\epsilon^*)$. By the definition of τ^* and the definition of n_{ref} (see (9)), we have

$$\log(\sigma/\epsilon^*) = \Theta(\tau^*) \quad \text{and} \quad \frac{\sigma}{\epsilon^*} = \Theta\big(2^{\tau^*}\big) \implies n_{\text{ref}}(\epsilon^*, \delta, \sigma) = \Theta\left(4^{(\tau^*)} \cdot (\tau^*)^3 \cdot \log\left(\frac{\tau^*}{\delta}\right)\right). \tag{85}$$

By using $\epsilon_s = \sigma/2^s$, $\delta_s = \frac{6\delta}{\pi^2 s^2}$, and the fact that a sum of exponentially increasing terms is dominated by its last term, we have for any $\tau \geqslant 1$ that

$$S(\tau) := \sum_{s=1}^{\tau} n_{\text{ref}}(\epsilon_s, \delta_s, \sigma) = \sum_{s=1}^{\tau} \Theta\left(4^s \cdot s^3 \cdot \log\left(\frac{s}{\delta}\right)\right) = \Theta\left(4^\tau \cdot \tau^3 \cdot \log\left(\frac{\tau}{\delta}\right)\right).$$

Using the definition of T in (10), we have

$$S(T) \leqslant n_{\text{true}} - n_{\text{loc}}(\delta, \lambda, \sigma) < S(T+1) = \Theta\left(4^{T+1} \cdot (T+1)^3 \cdot \log\left(\frac{T+1}{\delta}\right)\right), \tag{86}$$

Combining (85) and (86), and recalling that ϵ^* is defined such that $n_{\text{true}} - n_{\text{loc}}(\delta, \lambda, \sigma) = n_{\text{ref}}(\epsilon^*, \delta, \sigma)$, we have

$$4^{(\tau^*)} \cdot (\tau^*)^3 \cdot \log\left(\frac{\tau^*}{\delta}\right) = O\left(4^{T+1} \cdot (T+1)^3 \cdot \log\left(\frac{T+1}{\delta}\right)\right)$$

which implies $T \geqslant \tau^* - O(1)$. It follows that

$$\epsilon_T = \frac{\sigma}{2^T} \leqslant \frac{\sigma}{2^{\tau^* - O(1)}} = 2^{O(1)} \cdot \frac{\sigma}{2^{\tau^*}} = O(\epsilon^*)$$

as desired.

E.2 Proof of Theorem 11 (Adapting to Unknown Variance)

Recall that our proposed method for this result was given in Section 4.4. We first bound the sample complexity n. Recalling our choices of problem parameters in terms of T (see (13)), we have

$$n = \sum_{i=0}^{T} n\left(\epsilon_i, \delta_i, \lambda, \sigma_i\right) = \sum_{i=0}^{T} n\left(\frac{r\sigma_i}{5}, \frac{\delta}{T+1}, \lambda, \sigma_i\right) = \sum_{i=0}^{T} O\left(\frac{1}{r^2} \cdot \log^3\left(\frac{1}{r}\right) \cdot \log\left(\frac{T \log\left(\frac{1}{r}\right)}{\delta}\right) + \log\frac{\lambda}{\sigma_i}\right), \tag{87}$$

where the last step substitutes the sample complexity from Theorem 4. Recalling that $T = \lceil \log_2\left(\sigma_{\text{max}}/\sigma_{\text{min}}\right) \rceil$ and $\sigma_i = \sigma_{\min} \cdot 2^i$ (see (11) and (13)), we have

$$\sum_{i=0}^{T} \log_2 \frac{\lambda}{\sigma_i} = (T+1)\log_2 \frac{\lambda}{\sigma_{\min}} - \sum_{i=0}^{T} i = (T+1) \cdot \log_2 \frac{\lambda}{\sigma_{\min}} - \frac{T(T+1)}{2} = \Theta\left(T \log_2 \frac{\lambda}{\sqrt{\sigma_{\min}\sigma_{\max}}}\right).$$

Combining the above two findings gives

$$n = O\left(\log\left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\right) \cdot \frac{1}{r^2} \cdot \log^3\left(\frac{1}{r}\right) \cdot \log\left(\frac{\log\left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\right) \cdot \log\left(\frac{1}{r}\right)}{\delta}\right) + \log\left(\frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}\right) \cdot \log\frac{\lambda}{\sqrt{\sigma_{\min}\sigma_{\max}}}\right)$$

as desired.

We now show that the mean estimator is (ϵ, δ)-PAC, i.e.,

$$\Pr\left(|\hat{\mu}_{i^*} - \mu| \leqslant \epsilon\right) \geqslant 1 - \delta. \tag{88}$$

Let k be the smallest index satisfying $\sigma_k \geqslant \sigma_{\text{true}}$:

$$k = \operatorname*{argmin}_{i \geqslant 0}\{\sigma_i \geqslant \sigma_{\text{true}}\}. \tag{89}$$

For each $i \geqslant k$, the event

$$\mathcal{E}_i = \{\mu \in I_i\} \quad \text{where} \quad I_i = [\hat{\mu}^{(i)} \pm \epsilon_i] \text{ is as defined as in (14)}$$

occurs with probability at least

$$\Pr\left(\mathcal{E}_i\right) = \Pr\left(\mu \in I_i\right) = \Pr\left(|\hat{\mu}^{(i)} - \mu| \leqslant \epsilon_i\right) \geqslant 1 - \delta_i$$

by the subroutine's guarantee. By the union bound, the "good event" $\mathcal{E} = \bigcap_{i \geqslant k} \mathcal{E}_i$ happens with probability at least

$$\Pr\left(\mathcal{E}\right) = \Pr\left(\bigcap_{i \geqslant k} \mathcal{E}_i\right) = 1 - \Pr\left(\bigcup_{i \geqslant k} \neg\mathcal{E}_i\right) \geqslant 1 - \sum_{i \geqslant k} \Pr\left(\neg\mathcal{E}_i\right) \geqslant 1 - \sum_{i \geqslant k} \delta_i \geqslant 1 - \sum_{i=0}^{T} \delta_i \geqslant 1 - \delta.$$

We now condition on event $\mathcal{E}$. Observe that we have

$$\sigma_k = \begin{cases} \sigma_{\min} & \text{if } k = 0 \\ 2\sigma_{k-1} & \text{otherwise} \end{cases} \implies \sigma_k \leqslant 2\sigma_{\text{true}} = \frac{2\epsilon}{r} \tag{90}$$

due to (12), the definition of k (see (89)) and the assumption that $\sigma_{\text{true}} \geqslant \sigma_{\min}$. Based on (90), it is sufficient to show that

$$|\hat{\mu}_{i*} - \mu| \leqslant \frac{r\sigma_k}{2} \tag{91}$$

whenever the good event $\mathcal{E}$ holds.

Recall that i^* is the smallest index i satisfying the feasibility condition in (15). Towards showing (91), we first establish that $i^* \leqslant k$, i.e., σ_k is feasible. Under event $\mathcal{E}$, we have $\mu \in I_k$ and also $\mu \in I_j$ for all $j > k$. It follows that k satisfies (15) and so $i^* \leqslant k$ by definition. If $i^* = k$, then

$$|\hat{\mu}_{i*} - \mu| = |\hat{\mu}_k - \mu| \leqslant \epsilon_k = \frac{r\sigma_k}{5} < \frac{r\sigma_k}{2}$$

as desired. On the other hand, if $i^* < k$, then by definition (see (12)), we have

$$\sigma_k \geqslant 2\sigma_{i*}. \tag{92}$$

Furthermore, the two confidence intervals I_{i*} and I_k must overlap by the feasibility of i^* and the fact that $i^* < k$. Therefore, there is a common point z such that $z \in I_{i*}$ and $z \in I_k$. By the definition of the intervals (see (14)), we have $|\hat{\mu}_{i*} - z| \leqslant \epsilon_{i*}$ and $|\hat{\mu}_k - z| \leqslant \epsilon_k$, which implies

$$|\hat{\mu}_{i*} - \hat{\mu}_k| \leqslant |\hat{\mu}_{i*} - z| + |z - \hat{\mu}_k| \leqslant \epsilon_{i*} + \epsilon_k \tag{93}$$

by the triangle inequality. Using the triangle inequality a second time along with (93), event $\mathcal{E}$, the choice of ϵ_i in (13), and (92), we have

$$|\hat{\mu}_{i*} - \mu| \leqslant |\hat{\mu}_{i*} - \hat{\mu}_k| + |\hat{\mu}_k - \mu| \leqslant \epsilon_{i*} + 2\epsilon_k = \frac{r\sigma_{i*}}{5} + \frac{2r\sigma_k}{5} \leqslant \frac{r\sigma_k}{2},$$

thus giving the desired sufficient condition (91).

F Details of Two-stage Mean Estimator

Here we provide the technical details for the non-adaptive localization protocol described in Section 4.5. The goal of this localization protocol is to identify an interval I of length $O(\sigma)$ that contains the mean μ with high probability. The core idea is adapted from (Cai and Wei, 2024), whose focus is on Gaussian distributions. We modify their approach to handle our general non-parametric family $\mathcal{D}(\lambda, \sigma)$. The protocol works by encoding the location of the mean using a binary Gray code of length $K = \Theta(\log(\lambda/\sigma))$, and estimating each of these K bits by aggregating responses from suitably chosen non-adaptive queries. We now formalize the necessary definitions and describe the procedure.

Definition 17 (Gray function)**.** For integers $k \geqslant 0$, we let $g_k \colon [0,1] \to \{0,1\}$ be the k-th Gray function, defined by

$$g_k(x) \coloneqq \begin{cases} 0 & \text{if } \lfloor 2^k \cdot x \rfloor \bmod 4 \in \{0,3\} \\ 1 & \text{if } \lfloor 2^k \cdot x \rfloor \bmod 4 \in \{1,2\} \end{cases}.$$

Definition 18 (Change points set)**.** The set G_k of change points for g_k is defined as the collection of points $x \in [0,1]$ where $g_k(x)$ changes its value from 0 to 1 or from 1 to 0. Formally, we define

$$G_k = \left\{(2j-1) \cdot 2^{-k} : 1 \leqslant j \leqslant 2^{k-1}\right\} = \left\{x \in [0,1] : \lim_{y \to x^-} g_k(y) \neq \lim_{y \to x^+} g_k(y)\right\}.$$

Note that the G_k are pairwise disjoint, i.e., $G_k \cap G_{k'} = \varnothing$ for $k \neq k'$.

Definition 19 (Decoding)**.** For any $K \geqslant 1$, we let $\mathrm{Dec}_K \colon \{0,1\}^K \to 2^{[0,1]}$ be the decoding function defined by

$$\mathrm{Dec}_K(y_1, \ldots, y_K) \coloneqq \{x \in [0,1] : g_k(x) = y_k \quad \text{for } 1 \leqslant k \leqslant K\}$$

This is a dyadic interval of length 2^{-K} that is consistent with the gray code bits $y_1, y_2, \ldots, y_K$, and so we can express it as follows for some $x_0 \in [0,1]$:

$$\mathrm{Dec}_K(y_1, \ldots, y_K) = \left[x_0, x_0 + 2^{-K}\right] \subset [0,1].$$

With these definitions in mind, we now describe the localization procedure.

1. We first rescale
$$X_i' = \frac{X_i + \lambda}{2\lambda} \in [0,1] \quad \text{and} \quad \mu' = \frac{\mu + \lambda}{2\lambda} \in [0,1],$$
and note that the resulting variance scales as follows:
$$\mathbb{E}\big[|X_i' - \mu'|^2\big] \leqslant \left(\frac{\sigma}{2\lambda}\right)^2. \tag{94}$$

2. We view the samples as being collected in groups. Let the number of groups to be
$$K = \left\lfloor \log_2\left(\frac{2\lambda}{\sigma}\right) - 3 \right\rfloor, \tag{95}$$
with each group having the fixed number of samples $J = \left\lceil 8 \log \frac{3K}{\delta} \right\rceil$. Thus, the total number of samples used (for localization) is
$$KJ = \Theta\left(\log\left(\frac{\lambda}{\sigma}\right) \cdot \log \frac{\log(\lambda/\sigma)}{\delta}\right).$$

3. For sample j in group k, the agent sends the single bit
$$Z_{k,j} = g_k(X_{k,j}'),$$
where $X_{k,j}'$ is the unquantized transform sample.

4. For each group $k = 1, \ldots, K$, the learner computes the majority bit
$$\hat{z}_k = \mathrm{Maj}\{Z_{k,1}, \ldots, Z_{k,J}\} = \begin{cases} 1 & \text{if } \sum_j Z_{k,j} \geqslant J/2, \\ 0 & \text{otherwise.} \end{cases}$$

5. The learner first computes the interval $\left[x_0, x_0 + 2^{-K}\right] = \mathrm{Dec}_K(\hat{z}_1, \ldots, \hat{z}_K)$, and then widens it by shifting the left end and right end by $2^{-(K+2)}$:
$$I' = \left[x_0 - 2^{-(K+2)},\ x_0 + 2^{-K} + 2^{-(K+2)}\right] \cap [0,1]. \tag{96}$$
Finally, it scales and shifts the interval $I' = [L', U']$ by using the transformation
$$I = 2\lambda I' - \lambda = \left[2\lambda L' - \lambda,\ 2\lambda U' - \lambda\right]$$
and returns this as the final interval. Note that the length satisfies
$$|I| = 2\lambda \cdot (U' - L') \leqslant 2\lambda \cdot \left(2^{-K} + 2 \cdot 2^{-(K+2)}\right) = 2^{-K} \cdot 3\lambda = O(\sigma), \tag{97}$$
where the last step follows from the choice of K in (95).

Before proving Theorem 13, we first state three useful lemmas below. Lemma 20 is a restatement of (Cai and Wei, 2024)[Lemma 17] (whose proof is elementary and straightforward), while the other two lemmas bound the encoding and decoding error probability.

Lemma 20. ((Cai and Wei, 2024)[Lemma 17]) Let I' be the widened interval as stated in (96). If each $k \in \{1, \ldots, K\}$ satisfies the condition
$$\underbrace{\inf_{y \in G_k} |\mu' - y|}_{=:d_k} < 2^{-(K+2)} \quad \text{or} \quad \hat{z}_k = g_k(\mu'), \tag{98}$$
then it holds that $\mu' \in I'$. Note that there is at most one k satisfying the condition $d_k < 2^{-(K+2)}$.

Lemma 21. For each $k = 1, \dots K$ and each $j = 1, \dots, J$, we have

$$\Pr\left(g_k(X'_{k,j}) \neq g_k(\mu')\right) \leqslant \left(\frac{\sigma}{2\lambda d_k}\right)^2,$$

where $d_k = \inf_{y \in G_k} |\mu' - y|$ is the distance from the transformed mean to the set G_k from Definition 18.

Proof. We first claim that

$$\Pr\left(g_k(X'_{k,j}) \neq g_k(\mu')\right) \leqslant \Pr\left(|X'_{k,j} - \mu'| \geqslant d_k\right). \tag{99}$$

Before proving this, we note that given that it holds, Chebyshev's inequality (with the variance bound in (94)) gives the desired bound:

$$\Pr\left(g_k(X'_{k,j}) \neq g_k(\mu')\right) \leqslant \Pr\left(|X'_{k,j} - \mu'| \geqslant d_k\right) \leqslant \left(\frac{\sigma}{2\lambda d_k}\right)^2.$$

It remains to establish (99), or equivalently

$$\Pr\left(|X'_{k,j} - \mu'| < d_k\right) \leqslant \Pr\left(g_k(X'_{k,j}) = g_k(\mu')\right). \tag{100}$$

This follows from the event implication

$$\left\{|X'_{k,j} - \mu'| < d_k\right\} \implies \left\{g_k(X'_{k,j}) = g_k(\mu')\right\},$$

which follows immediately from the definition of d_k. □

Lemma 22 (Majority-vote reliability)**.** Fix a group $k \in \{1, \dots, K\}$. Suppose that each i.i.d. sample $X'_{k,j}$ with $j \in \{1, \dots, J\}$ satisfies

$$\Pr\left(g_k(X'_{k,j}) \neq g_k(\mu')\right) \leqslant \frac{1}{4}.$$

Then, under the choice $J = \lceil 8 \log \frac{3K}{\delta} \rceil$, the majority vote $\hat{z}_k = \mathrm{Maj}\{g_k(X'_{k,1}, \dots, g_k(X'_{k,J})\}$ satisfies

$$\Pr\left(\hat{z}_k \neq g_k(\mu')\right) \leqslant \exp(-J/8) \leqslant \frac{\delta}{3K}.$$

Proof. Let $B_j := \mathbf{1}\{g_k(X'_{k,J}) \neq g_j(\mu')\}$, which gives $B_j \sim \mathrm{Bern}(p_j)$ with $p_j \leqslant 1/4$. Let $S = \sum_{j=1}^{J} B_j$ count the number of errors in the group. The majority vote is incorrect only when at least half are wrong:

$$\Pr\left(\hat{z}_k \neq g_k(\mu')\right) = \Pr\left(S \geqslant \frac{J}{2}\right) = \Pr\left(S - \mathbb{E}[S] \geqslant \frac{J}{2} - \mathbb{E}[S]\right).$$

Since $\mathbb{E}[S] \leqslant J/4$, applying Hoeffding inequality yields

$$\Pr\left(S - \mathbb{E}[S] \geqslant \frac{J}{2} - \mathbb{E}[S]\right) \leqslant \Pr\left(S - \mathbb{E}[S] \geqslant \frac{J}{4}\right) \leqslant \exp\left(-\frac{J}{8}\right) \leqslant \frac{\delta}{3K}$$

as desired. □

Proof of Theorem 13. Given (97), it remains to show with probability at least $1 - \delta/2$ that $\mu \in I$, or equivalently, the scaled mean $\mu' = (\mu + \lambda)/(2\lambda)$ lies in I'. In view of Lemma 20, we define the "good events"

$$E_k = \left\{d_k < 2^{-(K+2)} \text{ or } \hat{z}_k = g_k(\mu')\right\}$$

and show that

$$\Pr\left(\bigcup_{k=1}^{K} E_k\right) \geqslant 1 - \frac{\delta}{2}.$$

By the union bound, it is sufficient to show that each "bad event" $\bar{E}_k$ happens with probability at most

$$\Pr(\bar{E}_k) = \Pr\left(d_k \geqslant 2^{-(K+2)} \text{ and } \hat{z}_k \neq g_k(\mu')\right) \leqslant \frac{\delta}{2K}.$$

Fix an arbitrary $k \in \{1, \ldots, K\}$. If $d_k < 2^{-(K+2)}$, then $\Pr(\bar{E}_k) = 0$. Therefore, we may assume without loss of generality that

$$d_k \geqslant 2^{-(K+2)} \iff \frac{1}{d_k} \leqslant 2^{K+2} = 4 \cdot 2^K.$$

Using this assumption, the choice of K (see (95)), and Lemma 21, we have

$$\Pr\left(g_k(X'_{k,j}) \neq g_k(\mu')\right) \leqslant \left(\frac{\sigma}{2\lambda d_k}\right)^2 \leqslant \left(\frac{\sigma}{2\lambda} \cdot 4 \cdot \frac{2\lambda}{\sigma} \cdot \frac{1}{8}\right)^2 = \frac{1}{4}.$$

It then follows from Lemma 22 that

$$\Pr\left(\hat{z}_k \neq g_k(\mu')\right) \leqslant \frac{\delta}{3K} \implies \Pr(\bar{E}_k) \leqslant \frac{\delta}{3K} < \frac{\delta}{2K}$$

as desired. □