

Optimal Sub-Gaussian Mean Estimation in \mathbb{R}

Jasper C.H. Lee

Paul Valiant

Brown University
jasperchlee@brown.edu

IAS & Purdue University
pvaliant@gmail.com

November 18, 2020

Abstract

We revisit the problem of estimating the mean of a real-valued distribution, presenting a novel estimator with sub-Gaussian convergence: intuitively, “our estimator, on *any* distribution, is as accurate as the sample mean is for the Gaussian distribution of matching variance.” Crucially, in contrast to prior works, our estimator does not require prior knowledge of the variance, and works across the entire gamut of distributions with finite variance, including those without any higher moments. Parameterized by the sample size n , the failure probability δ , and the variance σ^2 , our estimator is accurate to within $\sigma \cdot (1 + o(1)) \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$, tight up to the $1 + o(1)$ factor. Our estimator construction and analysis gives a framework generalizable to other problems, tightly analyzing a sum of dependent random variables by viewing the sum implicitly as a 2-parameter ψ -estimator, and constructing bounds using mathematical programming and duality techniques.

1 Introduction

We revisit one of the most fundamental problems in statistics: estimating the mean of a real-valued distribution, using as few independent samples from it as possible. Our proposed estimator has convergence that is optimal not only in a big-O sense (i.e. “up to multiplicative constants”), but tight to a $1 + o(1)$ factor, under the minimal (and essentially necessary, see below) assumption of the finiteness of the variance. Previous works, discussed further in Section 2, are either only big-O tight [12, 20, 1], or require additional strong assumptions such as the variance being known to the estimator [4] or assumptions that allow for accurate estimates of the variance, such as the kurtosis (fourth moment) being finite [4, 7].

1.1 The Model and Main Result

Given a set of i.i.d. samples from a real-valued distribution, the goal is to return, with extremely high probability, an accurate estimate of the distribution’s mean. Specifically, given a sample set X of size n consisting of independent draws from a real-value distribution D , an (ϵ, δ) -estimator of the mean is a function $\hat{\mu} : \mathbb{R}^n \rightarrow \mathbb{R}$ such that, except with failure probability $\leq \delta$, the estimate $\hat{\mu}(X)$ is within ϵ of the true mean $\mu(D)$. Namely,

$$\mathbb{P}(|\hat{\mu}(X) - \mu(D)| \leq \epsilon) \geq 1 - \delta \tag{1}$$

The goal is to find the optimal tradeoff between the sample size n , and the error parameters ϵ and δ , for the distribution D . Fixing any two of the three parameters and minimizing the third

yields essentially equivalent reformulations of the problem: we can fix ϵ, δ and minimize the *sample complexity* n ; we can fix δ, n and minimize *error* ϵ ; or we can fix ϵ, n and minimize the *failure probability* δ (maximizing the *robustness* $1 - \delta$).

Perhaps the most standard and well-behaved setting for mean estimation is when the distribution D is a Gaussian. The sample mean (the empirical mean) is a provably optimal estimator in our sense when D is Gaussian: for any $\epsilon, \delta > 0$, the sample mean $\mu(X)$ is an (ϵ, δ) -estimator when given a sample set of size $n = (2 + o(1)) \frac{\sigma^2(D) \cdot \log \frac{1}{\delta}}{\epsilon^2}$ (all logarithms will be base e); and there is *no* (ϵ, δ) -estimator for Gaussians if the constant 2 in the previous expression for the sample size is changed to any smaller number.

The main result of this paper is an estimator that performs as well, on *any* distribution with finite variance, as the sample mean does on a Gaussian, without knowledge of the distribution or its variance:

Theorem 1. *Estimator 1, given $\delta, n > 0$, defines a function $\hat{\mu}$ such that with probability at least $1 - \delta$, given a sample set X of size n , yields an estimate with error*

$$|\hat{\mu}(X) - \mu(D)| \leq \sigma(D) \cdot (1 + o(1)) \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$$

Here, the $o(1)$ term tends to 0 as $(\frac{\log \frac{1}{\delta}}{n}, \delta) \rightarrow (0, 0)$. Furthermore, as evidenced by the Gaussian case, there is no estimator which, under the same settings, produces an error that improves on our guarantees by more than a $1 + o(1)$ multiplicative factor.

We have parameterized the above theorem in terms of fixing the sample size n and the robustness parameter δ and asking for the minimum error ϵ ; however, because of the simple functional form of the bounds of Theorem 1, we can equivalently rephrase it as saying that, for any ϵ, δ , (a reparameterized) Algorithm 1 is an (ϵ, δ) estimator using $(2 + o(1)) \frac{\sigma(D)^2}{\epsilon^2} \log \frac{1}{\delta}$ samples; or for any n, ϵ , Algorithm 1 gives an estimate that is $\delta = \exp(-\frac{n\epsilon^2}{(2+o(1)) \cdot \sigma^2(D)})$ -robust. For each of these formulations, the performance is optimal up to the $1 + o(1)$ factor, as evidenced by the well-known Gaussian case, as explained above.

We make the following observations regarding the main (minimal) assumption in the theorem, namely the finiteness of the variance of the unknown distribution. First, imposing further assumptions about the finiteness of higher moments will not yield any improvements to the result, since matching lower bounds are provided by Gaussians, for which all moments are finite. Second, as shown by Devroye et al. [7], relaxing the finite variance assumption by only assuming, say, the finiteness of the $(1 + \beta)^{\text{th}}$ moment for some $\beta < 1$ will yield strictly worse sample complexity. In particular, the sample complexity will have an ϵ -dependence that is $\omega(1/\epsilon^2)$. Thus, our result shows that mean estimation can be performed at a sub-Gaussian rate, with the optimal multiplicative constant of 2 in the sample complexity, if and only if the variance of the underlying distribution is finite.

We also contrast with previous works that attain optimal sub-Gaussian convergence but make additional assumptions such as the finiteness of the kurtosis (4th moment) [4, 7]. The gap in assumptions between those works and this work is not only theoretical, but also of practical consequence: power law distributions are known to be good models of certain real-world phenomena, and for exponents in the range $(3, 5]$, the variance exists, but not the kurtosis.

1.2 Our Approach

We briefly describe the main features of our estimator, as a setting for what follows, and to distinguish it from prior work. At the highest level: in order to return a δ -robust estimate of the mean, our estimator “throws out the $\frac{1}{3} \log \frac{1}{\delta}$ most extreme points in the samples”, and returns the mean of what remains. More specifically, outliers are thrown out in a *weighted* manner, where we throw out a *fraction* of each data point, with the fraction proportional to the square of its distance from a median-of-means initial guess for the mean, where the fraction is capped at 1, and the proportionality constant is chosen so that the total weight thrown out equals exactly $\frac{1}{3} \log \frac{1}{\delta}$. See Estimator 1 for full details, but we stress here that the estimator is simple to implement—it may be computed in linear time—and therefore applicable in practice.

The above description is rather different from the typical M-estimator/ ψ -estimator approach of Catoni [4] and other works in this area. However, as we see in Section 3, our estimator can be reinterpreted as a 2-parameter ψ -estimator, and the proof of our main result will crucially rely on this reformulation.

1.3 Motivation: 3rd-order corrections of the empirical mean

Perhaps the most non-obvious part of our estimator is throwing out exactly $\frac{1}{3} \log \frac{1}{\delta}$ many samples. We motivate this quantity in this section, by considering the special case of estimating the mean of asymmetric—very biased—Bernoulli distributions, which is in some sense an extremal case for our setting.

Example 1. *Consider the mean estimation problem, given n samples from a Bernoulli distribution supported on 0 and 1, where the probability of drawing 1 equals some parameter p . Thus the number of 1s observed is distributed as the Binomial distribution $\text{Bin}(n, p)$, of mean np and variance $np(1-p)$. The interesting regime for us is when p is very small, and thus $1-p \approx 1$, and the Binomial distribution is essentially the Poisson distribution $\text{Poi}(np)$ of mean and variance $\lambda = np$. In this setting, the mean estimation problem becomes: given a sample k from $\text{Poi}(np)$, and the parameters n and δ , return an estimate that, except with failure probability δ , is as close as possible to p (or equivalently np). Given a Poisson sample $k \leftarrow \text{Poi}(np)$, returning simply k is a natural estimate of np ; however, since Poisson distributions are slightly skewed, it turns out that one should instead return the correction $k - \frac{1}{3} \log \frac{1}{\delta}$.*

Explicitly, the Poisson distribution has pmf $\text{poi}(\lambda, k) = \frac{\lambda^k e^{-\lambda}}{k!}$, whose logarithm, using Stirling’s approximation for the factorial, expanding to 3rd order in k , and dropping lower-order terms in λ is $-\frac{(k-\lambda)^2}{2\lambda} + \frac{(k-\lambda)^3}{6\lambda}$. The 2nd-order term here corresponds to a Gaussian centered at $k = \lambda$ of variance λ , which is a standard approximation for the Poisson distribution. However, crucially, the 3rd order term, corresponding to the positive skewness of the Poisson distribution, increases the pmf to the right of $k = \lambda$ and decreases it by an essentially symmetric factor to the left.

Seeking a δ -robust estimation of λ from a single sample of k , we are concerned, essentially, with the interval where the Poisson pmf is greater than δ , or equivalently, where the log pmf is greater than $\log \delta$. The quadratic $-\frac{(k-\lambda)^2}{2\lambda}$ in the first term of the above approximation equals $\log \delta$ when $k = \lambda \pm \sqrt{2\lambda \log \frac{1}{\delta}}$, and this interval is centered at the λ . However, crucially, when we take into account the 3rd-order term, the interval where $\text{poi}(\lambda, k) \geq \delta$ essentially shifts to become $k = \frac{1}{3} \log \frac{1}{\delta} + \lambda \pm \sqrt{2\lambda \log \frac{1}{\delta}}$. Thus, given a single sample, one can δ -robustly estimate the mean of a Poisson distribution similarly well as the Gaussian of same mean and variance, but only if one returns the sample minus $\frac{1}{3} \log \frac{1}{\delta}$.

Thus, the $\frac{1}{3} \log \frac{1}{\delta}$ term in our estimator arises essentially from a 3rd order correction to the sample mean, at least in the special case of Bernoulli distributions. For additional intuition and motivation about the “3rd order correction” in our estimator, please refer to Appendix A.

1.4 Key Contributions in Our Construction and Analysis

In addition to settling the fundamental sample complexity question of mean estimation, we point out that the estimator construction and analysis may also be of independent interest. In particular, the analysis framework—as described below—is generalizable to other problem settings and estimator constructions.

Our overall analysis framework may be viewed as a Chernoff bound—showing exponentially small probability of estimation error via bounds on a moment generation function (expectation of an exponentiated real-valued random variable). However, since we seek to analyze our estimator to sub-constant accuracy, many standard approaches fail to yield the required resolution. We point out three crucial components of our approach.

First, our estimator (Estimator 1) is *not* a sum of independent terms, which is fundamental to standard Chernoff bound approaches, and thus we instead reformulate our estimator as a 2-parameter ψ -estimator (see Definition 2). This technique rewrites our estimate $\hat{\mu}$ as the first coordinate of the root $(\hat{\mu}, \hat{\alpha})$ of a system of 2 equations $\psi_{\mu}(\hat{\mu}, \hat{\alpha}) = 0$ and $\psi_{\alpha}(\hat{\mu}, \hat{\alpha}) = 0$, where the functions $\psi_{\mu}(\hat{\mu}, \hat{\alpha}) = \sum_i \psi_{\mu}(x_i, \hat{\mu}, \hat{\alpha})$ and $\psi_{\alpha}(\hat{\mu}, \hat{\alpha}) = \sum_i \psi_{\alpha}(x_i, \hat{\mu}, \hat{\alpha})$ are explicitly sums of a corresponding function applied to each of the n independent data points in the sample set. Thus we have bought independence at the price of making the estimator an implicit function, introducing two new variables. One-dimensional estimators of this form are standard: for example, Catoni’s [4] mean estimator in the case of known variance is a (1 parameter) ψ -estimator for which he proves finite sample concentration. However, adding another dimension— $\hat{\alpha}$, a new implicit variable whose value the estimator will ultimately discard—is less standard, without standard analysis techniques, yet significantly increases the expressive power of such estimators [24]. Our high-level approach is to find carefully chosen linear combinations of the functions ψ_{μ} and ψ_{α} , each of which is now a sum of independent terms, and prove Chernoff bounds about these linear combinations.

Second, even after identifying these linear combinations of ψ functions, the corresponding Chernoff bound analysis is difficult to directly tackle. The Chernoff bound analysis, as it turns out, is essentially equivalent to bounding a max-min optimization problem where the maximization is over the set of real-valued probability measures with mean 0 and variance 1. In other words, the max-min optimization problem can be interpreted as having uncountably infinitely many variables. In order to drastically simplify the problem and make it amenable to analysis, we use convex-concave programming and linear programming duality techniques to reduce the problem to a pure minimization problem with a small finite number of variables, which we can analyze tightly.

We believe that the above two ideas—1) reformulating an estimator as a multi-parameter ψ -estimator, so as to find a proxy of the estimator that is a sum of independent variables, and 2) viewing the corresponding Chernoff bound analysis as an optimization problems and applying relevant duality techniques—form a general analysis framework which expands the space of possible estimators that are amenable to *tight* analysis.

2 Related Work

There is a long history of work on real-valued mean estimation in a variety of models. In the problem setting we adopt, where the sole assumption is on the finiteness of the second moment, the median-of-means algorithm [12, 20, 1] has long been known to have sample complexity tight

to within constant multiplicative factors, albeit with a sub-optimal constant. Catoni [4] improved this sample complexity to essentially optimal (tight up to a $1 + o(1)$ factor), by focusing on the special cases where the variance of the underlying distribution is known or the 4th moment is finite and bounded (in which case the second moment can be accurately estimated). We stress however that the finiteness of the 4th moment is nonetheless a much stronger assumption than our minimal assumption on the finiteness of the variance (see the discussion at the end of Section 1.1).

Moving beyond the original problem formulation, Devroye et al. [7] drew the distinction between a *single- δ estimator*, which takes in the robustness parameter δ as input, versus a *multiple- δ estimator*, which does not take any δ as input, but still provides guarantees across a wide range of δ values. In their work, making the same finite kurtosis assumption as Catoni, they achieved a multiple- δ estimator with essentially optimal sample complexity, for a wide range of δ values. It is thus natural and prudent to ask whether a multiple- δ estimator can exist for the entire class of distributions with finite variance, for a meaningful range of δ values. Unfortunately, Devroye et al. [7] showed strong lower bounds answering the question in the negative. Hence, in this work, our proposed estimator is (and must be) a single- δ estimator, taking δ as input.

Many applications have arisen from the success of sub-Gaussian mean estimation, showing how to leverage or extend Catoni-style estimators to new settings, achieving sub-Gaussian performance on problems such as regression, empirical risk minimization, and online learning (bandit settings): for example see [18, 2, 5, 3].

A separate but closely related line of work is on *high dimensional* mean estimation. While estimators generalizing the “median-of-means” construction were found to have statistical convergence tight to multiplicative constants, until recently, such estimators took super-polynomial time to compute [17]. A recent line of work [10, 6, 15], started by Hopkins [10], thus focuses on the computational aspect, and brought the computation time first down to polynomial time, with subsequent work bringing it further down to quadratic time using spectral methods.

A recent comprehensive survey by Lugosi and Mendelson [16] explains much of the above works in greater detail.

Other works have focused on mean estimation in restrictive settings, for example, with differential privacy constraints. For example, Kamath et al. [14] studied the differentially private mean estimation problem in the constant probability regime, and showed strong sample complexity separations from our unrestricted setting. Duchi, Jordan and Wainright [8, 9] also studied the problem under the stricter constraint of *local* differential privacy. See the work of Kamath et al. [14] for a more comprehensive literature review on differentially private mean estimation.

Part of our tight analysis relies on insights from mathematical programming and duality; see [22] for a detailed discussion of prior works that use such mathematical programming and duality tools to either design or analyze statistical estimators [21, 19, 26, 27, 13, 25].

3 Our Estimator

In this section, we present our estimator (Estimator 1), as well as its reformulation as a 2-parameter ψ -estimator. We then present some perspective and basic structural properties of the estimator that will serve as a foundation for the analysis to follow.

3.1 Meaning of the Estimator

Consider the expression in Step 3 for the final returned value of the estimator, $\hat{\mu} = \kappa + \frac{1}{n} \sum_i (x_i - \kappa)(1 - \min(\alpha(x_i - \kappa)^2, 1))$. Without the final min expression, the expression $\kappa + \frac{1}{n} \sum_i (x_i - \kappa) \cdot 1$ computes exactly the sample mean. The factor $(1 - \min(\alpha(x_i - \kappa)^2, 1))$ may be thought of as a weight

Estimator 1 The Main Estimator

Inputs:

- n independent samples $\{x_i\}$ from the unknown underlying distribution D (guaranteed to have finite variance)
 - Confidence parameter δ
1. Compute the median-of-means estimate κ : evenly partition the data into $\log \frac{1}{\delta}$ groups and let κ be the median of the set of means of the groups.
 2. Find the solution α to the monotonic, piecewise-linear equation $\sum_i \min(\alpha(x_i - \kappa)^2, 1) = \frac{1}{3} \log \frac{1}{\delta}$
 3. Output: $\hat{\mu} = \kappa + \frac{1}{n} \sum_i (x_i - \kappa)(1 - \min(\alpha(x_i - \kappa)^2, 1))$
-

on the i^{th} element, between 0 and 1, where a weight of 1 leaves that element as is, but a weight towards 0 essentially throws out part of the sample x_i and instead defaults to the median-of-means estimate κ . Thus, rather than either keeping or discarding each entry, the weight $\min(\alpha(x_i - \kappa)^2, 1)$ specifies what *fraction* of the i^{th} sample to discard.

The condition in Step 2 of Estimator 1 picks α so that the total, weighted, number of discarded samples equals $\frac{1}{3} \log \frac{1}{\delta}$. The expression $\min(\alpha(x_i - \kappa)^2, 1)$ specifying what fraction of each x_i to discard says, essentially, that this fraction should be proportional to the square of the deviation of x_i from the mean estimate κ , capped at 1 so that we do not discard “more than 100% of” any sample x_i .

3.2 Structural Properties of the Estimator

We point out three basic structural properties of Estimator 1 that both shed light on the estimator itself, and will be crucial to its analysis. We formally state and prove these properties in Appendix B.

First, the estimator is “affine invariant” in the sense that, if its input samples $\{x_i\}$ undergo an affine map $x \rightarrow ax + b$ then its output will be mapped correspondingly. Second, as is well known, the median-of-means estimate κ of Step 1, while not as accurate as what we will eventually return, is robust in the sense that, with probability at least $1 - \delta/2$, the median-of-means estimate has additive error from the true mean that is at most $O(\sigma \cdot \sqrt{\frac{\log \frac{1}{\delta}}{n}})$ —proportional to the eventual guarantees of our estimator, but with somewhat worse proportionality constant. Third, if we temporarily ignore Step 1, treating κ as a free parameter, we show that the final output of the algorithm, $\hat{\mu}$, varies very little with κ . Combined with the accuracy guarantees of the median-of-means estimate, the difference in the final estimate between using the median-of-means as κ versus using the *true* mean as κ is inconsequential (a $o(1)$ factor) compared to the total additive error we aim for. Therefore, for the purposes of *analysis*, it suffices to assume that κ takes the value of the true mean (though an algorithm could not do this in practice, as the true mean is unknown).

These structural properties allow us to drastically simplify the analysis: the affine invariance means it is sufficient to show our estimator works for the special case of distributions with mean 0 and variance 1; the second and third properties mean that errors in κ effectively do not matter, and, for distributions with mean 0, it is sufficient to omit Step 1 and instead just analyze the case where $\kappa = 0$.

We point out that Estimator 1 when modified to set $\kappa = 0$ (independently of the samples) is

no longer affine invariant, nor is its reformulation as a ψ -estimator in Section 3.3. The structural properties in this section show that, instead of analyzing the actual estimator (Estimator 1 which is affine invariant), it suffices to analyze this artificially simplified, although no longer affine invariant, estimator which sets $\kappa = 0$, on distributions with mean 0 and variance 1. Explicitly, in the rest of the paper we will show Proposition 3 (Section 4), which analyzes the mean-0 variance-1 case of the ψ -estimator defined below in Definition 2; the discussion of this section—made formal in Appendix B—shows that this proposition implies our main result, Theorem 1.

3.3 Representing a Special Case of Estimator 1 as a ψ -Estimator

As discussed in Section 1.4, our estimator, even its simplified version with $\kappa = 0$, is not a sum of independent terms, making it difficult to tightly bound its moment generating function, and hence also difficult to prove its concentration around the true mean using a Chernoff-style bound. Our solution is to reformulate Estimator 1, with the simplifying assumption that $\kappa = 0$, as a 2-parameter ψ -estimator, as defined in Definition 2. This reformulation defines our estimate $\hat{\mu}$ implicitly in terms of two new functions ψ_μ and ψ_α that are indeed sums of n independent terms, each term depending on a single x_i . We will use this representation crucially for the concentration analysis of the estimator.

Definition 2. Consider Estimator 1 but with Step 1 replaced with “ $\kappa = 0$ ”. The estimator can be equivalently expressed as follows:

1. Input: n independent samples $X = x_1, \dots, x_n$
2. Solve for the (unique) pair $(\hat{\mu}, \hat{\alpha})$ satisfying $\psi_\mu = 0$ and $\psi_\alpha = 0$, where the functions are defined as follows:

$$\begin{aligned}\psi_\mu(X, \hat{\mu}, \hat{\alpha}) &= \sum_{i=1}^n (\hat{\mu} - x_i (1 - \min(\hat{\alpha}x_i^2, 1))) \\ \psi_\alpha(X, \hat{\mu}, \hat{\alpha}) &= \sum_{i=1}^n \left(\min(\hat{\alpha}x_i^2, 1) - \frac{1}{3n} \log \frac{1}{\delta} \right)\end{aligned}$$

(Note that $\hat{\alpha} > 0$ always)

3. Output: $\hat{\mu}$ from the previous step

We will sometimes omit $\hat{\mu}$ from the arguments of ψ_α since $\hat{\mu}$ is not used in the definition of the function. We will often refer to the pair (ψ_μ, ψ_α) as a 2-element vector ψ .

For convenience in the rest of the paper, we define $\hat{v} \equiv \frac{\log(1/\delta)}{3n\hat{\alpha}}$, which we refer to as the “truncated empirical variance”; this is because, if we modify the $\psi_\alpha = 0$ condition by removing the “truncation” of taking the min with 1, then the resulting condition, when expressed in terms of $\hat{v} = \frac{\log(1/\delta)}{3n\hat{\alpha}}$ and rearranged, is exactly the condition that \hat{v} is the empirical variance: $\frac{1}{n} \sum_{i=1}^n x_i^2$. Thus $\hat{\alpha}$ may be thought of as a proxy for the empirical variance, as $\hat{v} = \frac{\log(1/\delta)}{3n\hat{\alpha}}$ equals the empirical variance, except in cases when samples are far enough from 0 that they are “truncated” by the “min”.

Interestingly, in the case that none of the samples are “truncated”, (and $\kappa = 0$), the overall output of the estimator becomes $\frac{1}{n} \sum_i x_i - \alpha x_i^3 = \frac{1}{n} \sum_i x_i - \frac{\log(1/\delta)}{3n\hat{v}} x_i^3$, namely, $\hat{\mu}$ is “the empirical mean, corrected by subtracting $\frac{1}{3n} \log \frac{1}{\delta}$ times the ratio of the empirical 3rd moment over the empirical 2nd moment.”

Proof that Definition 2 is equivalent to Estimator 1 when κ is set to 0. Fix a set of samples $X = \{x_i\}$. We observe that Estimator 1, with the additional simplifying assumption that $\kappa = 0$, can be represented by the following 2 equations.

$$\begin{aligned} \sum_i \min(\alpha x_i^2, 1) &= \frac{1}{3} \log \frac{1}{\delta} \\ \hat{\mu} &= \frac{1}{n} \sum_i x_i (1 - \min(\alpha x_i^2, 1)) \end{aligned} \tag{2}$$

Estimator 1 solves for α in the first line, and uses this α value to compute the estimate $\hat{\mu}$ in the second line. The two conditions of Equation 2 are equivalent to the two conditions $\psi_\alpha = 0$, $\psi_\mu = 0$ respectively, and thus the two estimators are equivalent. \square

4 Analyzing our estimator

In this section, we outline the proof of our main theorem, restated as follows.

Theorem 1. *Estimator 1, given $\delta, n > 0$, and a sample set X of n independent samples from distribution D , will, with probability at least $1 - \delta$ over the sampling process, yield an estimate $\hat{\mu}$ with error at most $|\hat{\mu}(X) - \mu(D)| \leq \sigma(D) \cdot (1 + o(1)) \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$. Here, the $o(1)$ term tends to 0 as $(\frac{\log \frac{1}{\delta}}{n}, \delta) \rightarrow (0, 0)$.*

The discussion of the structural properties of Estimator 1 in Section 3.2 shows that it is sufficient to instead show that, for any distribution of mean 0 and variance 1, the ψ -estimator of Definition 2 will return an estimate $\hat{\mu}$ that is close to 0, except with tiny probability. (See Appendix B for the formal statements of the claims of Section 3.2.) Recall also that, since the ψ -estimator solves for $(\hat{\mu}, \hat{\alpha})$ such that $\psi(X, \hat{\mu}, \hat{\alpha}) = 0$ (where X is the sample set) and returns $\hat{\mu}$, the claim that the returned estimate will be close to 0 is equivalent to saying that *every* $(\hat{\mu}, \hat{\alpha})$ pair with $\hat{\mu}$ far from 0 must violate the equation, namely $\psi(X, \hat{\mu}, \hat{\alpha}) \neq 0$. We thus prove the following proposition (Proposition 3), to yield Theorem 1. Note that the failure probability in Proposition 3 is $\delta/2$ (instead of δ , as in Theorem 1), accounting for an additional $\delta/2$ probability that the median-of-means estimate in Step 1 of Estimator 1 fails.

Proposition 3. *There exists a universal constant $c > 0$ such that, fixing $\epsilon' = \left(1 + \frac{c \log \log \frac{1}{\delta}}{\log \frac{1}{\delta}}\right) \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$, we have that for all distributions D with mean 0 and variance 1, with probability at least $1 - \frac{\delta}{2}$ over the set of samples X , for all $\hat{\mu}, \hat{\alpha}$ where $|\hat{\mu}| > \epsilon'$ and $\hat{\alpha} > 0$, the vector $\psi(X, \hat{\mu}, \hat{\alpha}) \neq 0$.*

Proposition 3 asks us to show that, with high probability, $\psi(X, \hat{\mu}, \hat{\alpha})$ is not at the origin for any choice of $|\hat{\mu}| > \epsilon'$, $\hat{\alpha}$; instead, as a proof strategy, we choose a finite bounded mesh of $\hat{\mu}, \hat{\alpha}$ and show that the function $\psi(X, \hat{\mu}, \hat{\alpha})$ is 1) not just nonzero, but far from the origin on this set, 2) Lipschitz in between mesh elements, and 3) monotonic (in an appropriate sense) outside the mesh bounds. Step 1), discussed below, contains the most noteworthy part of the proof, a mathematical programming-inspired bound to help complete a delicate Chernoff bound argument.

For simplicity, we reparameterize to work with $\hat{v} \equiv \frac{\log(1/\delta)}{3n\hat{\alpha}}$ (the “truncated empirical variance”) instead of $\hat{\alpha}$: the mesh we analyze, covering the most delicate region for analysis, will span the interval $\hat{v} \in [0.05, 55.5]$, namely, where the truncated empirical variance \hat{v} is within a constant factor of the true variance of 1. Note that this should *not* be taken to imply that $\hat{v} \in [0.05, 55.5]$ with high probability—the truncated empirical variance is not designed to be a good estimate of the variance,

merely as a step in robustly estimating the mean; and further, accurate estimates of the variance are simply impossible in general without further assumptions such as bounds on the distribution's 3rd or 4th moments. We also want to distinguish our estimator from Catoni's [4]: Catoni's estimator relies on having a high-precision estimate of the variance (to within a $1 + o(1)$ factor) in order to achieve the desired performance. By contrast, our estimator is robust against wild inaccuracies of the (truncated) empirical variance \hat{v} compared to the true variance of 1. In short, the approach of our estimator should be viewed as distinct from Catoni's, since, while Catoni's estimator relies on an initial good guess at the variance, ours thrives in the inevitable situations where \hat{v} is far from 1.

We return to describing our strategy for analyzing the performance of our estimator. For each $\hat{\mu}, \hat{v} = \frac{\log(1/\delta)}{3n\hat{\alpha}}$ that we analyze (from the finite mesh): instead of directly showing that, with $\geq 1 - \frac{\delta}{2}$ probability, $\psi(X, \hat{\mu}, \hat{\alpha})$ is far from the origin in some direction, we instead *linearize* this claim; we prove the stronger claim that there exists a specific direction $\mathbf{d}(\hat{v})$ such that with $\geq 1 - \frac{\delta}{2}$ probability, $\psi(X, \hat{\mu}, \hat{\alpha})$ is more than $\frac{1}{\log(1/\delta)}$ distance from the origin in direction \mathbf{d} (specifically we lower bound the dot product $\mathbf{d}(\hat{v}) \cdot \psi(\hat{\mu}, \hat{\alpha})$, while we upper bound each coordinate of \mathbf{d} inversely with the Lipschitz coefficients of ψ). The crucial advantage of this reformulation is that, since each of ψ_μ, ψ_α is a sum of n terms, that are each a function of an independent sample x_i from D , the dot product $\mathbf{d}(\hat{v}) \cdot \psi(X, \hat{\mu}, \hat{\alpha})$ is thus also a sum of n independent terms, and thus we finish the proof with a Chernoff bound, Lemma 4. The Chernoff bound argument itself is standard; however, to bound the resulting expression requires an extremely delicate analysis that we pull out into a separate 4-variable inequality expressed as Lemma 7—see the discussion around the lemma for more details and for motivation of the analysis from a mathematical programming perspective.

We state the crucial Chernoff bound (Lemma 4) and the Lipschitz bounds (Lemma 5), and then use them to prove Proposition 3. We prove Lemmas 4 and 5 in the next section, along with the statement and proof of the delicate component that is Lemma 7.

Lemma 4. *Consider an arbitrary distribution D with mean 0 and variance 1. There exists a universal constant c where the following claim is true. Fixing $\hat{\mu} = \epsilon' = \left(1 + \frac{c \log \log \frac{1}{\delta}}{\log \frac{1}{\delta}}\right) \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$, then for all δ smaller than some universal constant, and for all $\hat{v} \in [0.05, 55.5]$, there exists a vector $\mathbf{d}(\hat{v})$ where $d_\mu \geq 0$, and both $\sqrt{\frac{n}{\log(1/\delta)}} |d_\mu|, |d_\alpha|$ are bounded by a universal constant, such that*

$$\mathbb{P}_{X \leftarrow D^n} \left(\mathbf{d}(\hat{v}) \cdot \psi \left(X, \hat{\mu} = \epsilon', \hat{\alpha} = \frac{\log(1/\delta)}{3n\hat{v}} \right) > \frac{1}{\log \frac{1}{\delta}} \right) \geq 1 - \frac{\delta}{\log^4 \frac{1}{\delta}}$$

Furthermore, for $\hat{v} = 0.05$ we have $d_\mu = \sqrt{3.75 \frac{\log(1/\delta)}{n}}$, $d_\alpha = \sqrt{3}$; and for $\hat{v} = 55.5$ we have $d_\mu = 0$, $d_\alpha < 0$.

Lemma 5. *Consider an arbitrary set of n samples X . Consider the expressions $\psi_\mu(X, \hat{\mu}, \hat{\alpha}), \psi_\alpha(X, \hat{\alpha})$, reparameterized in terms of $\hat{v} \equiv \frac{\log(1/\delta)}{3n\hat{\alpha}}$ in place of $\hat{\alpha}$. Suppose the equation $\psi_\alpha(X, \hat{\alpha}) = 0$ has a solution in the range $\hat{v} \in [0.05, 55.5]$. Then the functions $\sqrt{\frac{\log(1/\delta)}{n}} \psi_\mu(X, \hat{\mu}, \hat{\alpha})$ and $\psi_\alpha(X, \hat{\alpha})$ are Lipschitz with respect to \hat{v} on the entire interval $\hat{v} \in [0.05, 55.5]$, with Lipschitz constant $c \log \frac{1}{\delta}$ for some universal constant c .*

We now prove Proposition 3, which per our previous discussion, implies our main result, Theorem 1.

Proof of Proposition 3. As in Lemma 4, we fix $\epsilon' = \left(1 + \frac{c \log \log \frac{1}{\delta}}{\log \frac{1}{\delta}}\right) \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$, where c is some universal constant.

By symmetry, instead of considering positive and negative $\hat{\mu}$, it suffices to consider the case $\hat{\mu} > \epsilon'$ (as opposed to $\hat{\mu} < -\epsilon'$) and show that this case succeeds with probability at least $1 - \frac{\delta}{4}$.

To prove the claim, we first prove a stronger statement on a restricted domain, that with probability at least $1 - \frac{\delta}{4}$ over the randomness of the sample set X , for each $\hat{v} \in [0.05, 55.5]$ there exists a vector $\mathbf{d} = (d_\mu, d_\alpha)$ such that $\mathbf{d} \cdot \boldsymbol{\psi}(X, \epsilon', \hat{\alpha} \equiv \frac{\log(1/\delta)}{3n\hat{v}}) > 0$, with $d_\mu \geq 0$ throughout, and, for $\hat{v} = 0.05$ we have $d_\mu = \sqrt{3.75 \frac{\log(1/\delta)}{n}}$, $d_\alpha = \sqrt{3}$; and for $\hat{v} = 55.5$ we have $d_\mu = 0$, $d_\alpha < 0$.

We will first apply Lemma 4 to each \hat{v} in a discrete mesh: let M consist of evenly spaced points between 0.05 and 55.5 with spacing $1/\log^3 \frac{1}{\delta}$ (thus with $\Theta(\log^3 \frac{1}{\delta})$ many points).

By Lemma 4 and a union bound over these $\Theta(\log^3 \frac{1}{\delta})$ points, we have that with probability at least $1 - \frac{\delta}{\Theta(\log^3 \frac{1}{\delta})}$ (which is at least $1 - \frac{\delta}{4}$ for δ smaller than some universal constant) over the set of n samples X , for all $\hat{v} \in M$, there exists a vector $\mathbf{d}(\hat{v})$ such that $\mathbf{d}(\hat{v}) \cdot \boldsymbol{\psi}(X, \hat{\mu} = \epsilon', \hat{\alpha} \equiv \frac{\log(1/\delta)}{3n\hat{v}}) > 1/\log \frac{1}{\delta}$, where \mathbf{d} further satisfies the desired positivity and boundary conditions, and where both $\sqrt{\frac{n}{\log(1/\delta)}}|d_\mu|, |d_\alpha|$ are bounded by a universal constant. For the rest of the proof, we will only consider sets of samples X satisfying the above condition.

Now consider an arbitrary $\hat{v}' \in [0.05, 55.5] \setminus M$ and consider the vector $\boldsymbol{\psi}$ evaluated at $\hat{\alpha}' = \frac{\log(1/\delta)}{3n\hat{v}'}$. We wish to extend the dot product inequality to hold also for \hat{v}' . If $\psi_\alpha \neq 0$ then there is nothing to prove: set $d_\mu = 0$ and $d_\alpha = \text{sign}(\psi_\alpha)$; otherwise, $\psi_\alpha = 0$ means we may apply Lemma 5 to conclude that both $\sqrt{\frac{\log(1/\delta)}{n}}\psi_\mu(X, \hat{\mu}, \hat{\alpha}')$ and $\psi_\alpha(X, \hat{\mu}, \hat{\alpha}')$ are Lipschitz with respect to \hat{v}' on the interval $\hat{v}' \in [0.05, 55.5]$, with Lipschitz constant $c \log \frac{1}{\delta}$ for some universal constant c .

Consider the closest $\hat{v} \in M$ to \hat{v}' , which by definition of M is at most $1/\log^3 \frac{1}{\delta}$ away. By assumption on X , there exists a vector \mathbf{d} such that $\mathbf{d} \cdot \boldsymbol{\psi}(X, \hat{\mu} = \epsilon', \hat{\alpha} = \frac{\log(1/\delta)}{3n\hat{v}}) > 1/\log \frac{1}{\delta}$, with $d_\mu \geq 0$ and both $\sqrt{\frac{n}{\log(1/\delta)}}|d_\mu|, |d_\alpha|$ are bounded by a universal constant. Because of the Lipschitz bounds on $\boldsymbol{\psi}$, combined with the bounds on the size of the d_μ, d_α , we conclude that the Lipschitz constant of the dot product (treating the vector \mathbf{d} as fixed) is $O(\log \frac{1}{\delta})$. Thus, the large positive dot product at \hat{v} implies at least a positive dot product nearby at \hat{v}' : $\mathbf{d} \cdot \boldsymbol{\psi}(X, \hat{\mu} = \epsilon', \hat{v}') > \frac{1}{\log \frac{1}{\delta}} - O(\log \frac{1}{\delta}) \frac{1}{\log^3 \frac{1}{\delta}} > 0$, for sufficiently small δ as given in the proposition statement.

Having shown the stronger version of the claim for the restriction $\hat{\mu} = \epsilon'$ and $\hat{v} \in [0.05, 55.5]$ we now extend to the entire domain via three monotonicity arguments. Explicitly, assume the set of samples X satisfies the dot product inequality above with the vector function $\mathbf{d}(\hat{v})$, where $\mathbf{d}(\hat{v})$ satisfies the boundary conditions at $\hat{v} = 0.05$ and 55.5 specified in Lemma 4. From this assumption, we will show that $\boldsymbol{\psi} \neq 0$ for *any* positive $\hat{v} = \frac{\log(1/\delta)}{3n\hat{\alpha}}$, and for *any* $\hat{\mu} \geq \epsilon'$.

First consider $\hat{v} > 55.5$ (still fixing $\hat{\mu} = \epsilon'$). The function $\psi_\alpha = \sum_{i=1}^n (\min(\hat{\alpha}x_i^2, 1) - \frac{1}{3n} \log \frac{1}{\delta})$ is an increasing function of $\hat{\alpha}$, and thus a decreasing function of $\hat{v} \equiv \frac{\log(1/\delta)}{3n\hat{\alpha}}$. Since for $\hat{v} = 55.5$, the dot product $\mathbf{d} \cdot \boldsymbol{\psi} > 0$ with $d_\mu = 0, d_\alpha < 0$, the dot product will thus remain positive for this same choice of \mathbf{d} as we increase \hat{v} from 55.5.

Next, for $\hat{v} < 0.05$ (again still fixing $\hat{\mu} = \epsilon'$), we analogously show that the dot product of $\boldsymbol{\psi}(X, \epsilon', \hat{\alpha} \equiv \frac{\log(1/\delta)}{3n\hat{v}})$ with the fixed vector $\mathbf{d}(0.05)$ will increase as we decrease \hat{v} . The i^{th} term in the sums defining ψ_μ or ψ_α depends on $\hat{\alpha}$ (and thus \hat{v}) only in the factor $\min(\hat{\alpha}x_i^2, 1)$. Further, there is no dependence unless the first term attains the min, namely $|x_i| \leq \sqrt{1/\hat{\alpha}}$, which in turn is upper bounded by $\sqrt{0.15 \frac{n}{\log(1/\delta)}}$ because of our assumption that $\hat{v} < 0.05$. Thus, the only i^{th} terms in the dot product which have $\hat{\alpha}$ dependent are simply equal to $d_\mu \hat{\alpha}x_i^3 + d_\alpha \hat{\alpha}x_i^2 = \hat{\alpha}x_i^2(d_\alpha + x_i d_\mu)$. By our choice of $d_\mu(0.05) = \sqrt{3.75 \frac{\log(1/\delta)}{n}}$ and $d_\alpha(0.05) = \sqrt{3}$ from Lemma 4, the expression $(d_\alpha + x_i d_\mu) \geq \sqrt{3} - \sqrt{0.15} \sqrt{3.75}$ is thus always non-negative, and thus the overall dot product

cannot decrease as we send $\hat{\alpha}$ to ∞ —equivalently, sending \hat{v} to 0—as desired.

We have thus shown that, for all non-negative $\hat{\alpha} = \frac{\log(1/\delta)}{3n\hat{v}}$, there is a vector \mathbf{d} with $d_\mu \geq 0$ whose dot product with $\boldsymbol{\psi}(X, \epsilon', \hat{\alpha})$ is greater than 0. We complete the proof by noting that the only dependence on $\hat{\mu}$ in $\boldsymbol{\psi}$ is that ψ_μ is (trivially) increasing in $\hat{\mu}$. Since $d_\mu \geq 0$, increasing $\hat{\mu}$ from ϵ' will only increase the dot product, and thus the dot product remains strictly greater than 0, implying that $\boldsymbol{\psi}(X, \hat{\mu}, \hat{\alpha}) \neq 0$ as desired. \square

5 Proofs of Lemmas 4 and 5

The main purpose of this section is to present and motivate the proof of Lemma 4—since our results are tight across such a wide parameter space, the resulting inequalities are somewhat subtle. After, we also present the short proof of Lemma 5.

Lemma 4. *Consider an arbitrary distribution D with mean 0 and variance 1. There exists a universal constant c where the following claim is true. Fixing $\hat{\mu} = \epsilon' = \left(1 + \frac{c \log \log \frac{1}{\delta}}{\log \frac{1}{\delta}}\right) \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$, then for all δ smaller than some universal constant, and for all $\hat{v} \in [0.05, 55.5]$, there exists a vector $\mathbf{d}(\hat{v})$ where $d_\mu \geq 0$, and both $\sqrt{\frac{n}{\log(1/\delta)}}|d_\mu|, |d_\alpha|$ are bounded by a universal constant, such that*

$$\mathbb{P}_{X \leftarrow D^n} \left(\mathbf{d}(\hat{v}) \cdot \boldsymbol{\psi} \left(X, \hat{\mu} = \epsilon', \hat{\alpha} = \frac{\log(1/\delta)}{3n\hat{v}} \right) > \frac{1}{\log \frac{1}{\delta}} \right) \geq 1 - \frac{\delta}{\log^4 \frac{1}{\delta}} \quad (3)$$

Furthermore, for $\hat{v} = 0.05$ we have $d_\mu = \sqrt{3.75 \frac{\log(1/\delta)}{n}}$, $d_\alpha = \sqrt{3}$; and for $\hat{v} = 55.5$ we have $d_\mu = 0$, $d_\alpha < 0$.

We start the analysis via standard Chernoff bounds on the complement of the probability in Equation 3 via Lemma 6, before pausing to discuss how mathematical programming and duality insights lead to the formulation of the crucial Lemma 7; we then complete the proof.

Lemma 6. *Consider an arbitrary distribution D with mean 0 and variance 1. For all sufficiently small δ , for any $\hat{\mu}, \hat{\alpha}$ and vector $\mathbf{d} = (d_\mu, d_\alpha)$, we have*

$$\mathbb{P}_{X \leftarrow D^n} \left(\mathbf{d} \cdot \boldsymbol{\psi}(X, \hat{\mu}, \hat{\alpha}) \leq \frac{1}{\log \frac{1}{\delta}} \right) \leq 2 \left(e^{-d_\mu \hat{\mu} + d_\alpha \frac{1}{3n} \log \frac{1}{\delta}} \mathbb{E}_{x \leftarrow D} (e^{d_\mu x(1 - \min(\hat{\alpha}x^2, 1)) - d_\alpha \min(\hat{\alpha}x^2, 1)} \right)^n$$

Proof. We upper-bound the probability by exponentiating the negation of both sides of the expression inside the probability, and then using Markov's inequality:

$$\begin{aligned} & \mathbb{P}_{X \leftarrow D^n} \left(\mathbf{d}(\hat{v}) \cdot \boldsymbol{\psi}(X, \hat{\mu}, \hat{\alpha}) \leq \frac{1}{\log \frac{1}{\delta}} \right) \\ &= \mathbb{P}_{X \leftarrow D^n} \left(e^{-\mathbf{d}(\hat{v}) \cdot \boldsymbol{\psi}(X, \hat{\mu}, \hat{\alpha})} \geq e^{-\frac{1}{\log \frac{1}{\delta}}} \right) \\ &\leq 2 \mathbb{E}_{X \leftarrow D^n} \left(e^{-\mathbf{d}(\hat{v}) \cdot \boldsymbol{\psi}(X, \hat{\mu}, \hat{\alpha})} \right) \quad \text{by Markov's inequality; and } e^{\frac{1}{\log(1/\delta)}} \leq 2 \text{ for sufficiently small } \delta \\ &= 2 \mathbb{E}_{x \leftarrow D} (e^{-\mathbf{d}(\hat{v}) \cdot \boldsymbol{\psi}(x, \hat{\mu}, \hat{\alpha})})^n \quad \text{by independence} \\ &= 2 \left(e^{-d_\mu \hat{\mu} + d_\alpha \frac{1}{3n} \log \frac{1}{\delta}} \mathbb{E}_{x \leftarrow D} (e^{d_\mu x(1 - \min(\hat{\alpha}x^2, 1)) - d_\alpha \min(\hat{\alpha}x^2, 1)} \right)^n \quad \text{substituting the definition of } \boldsymbol{\psi} \end{aligned} \quad (4)$$

\square

5.1 Mathematical Programming and Duality Analysis

In order to show Lemma 4, we aim to find bounds on the failure probability that are as strong as possible. Appealing to Lemma 6 that we have just proven, recall that, as in the standard Chernoff bound methodology, we are still free to choose the parameters d_μ, d_α , which we do so as to minimize the resulting bound on the failure probability. Phrased abstractly, the goal is, for the $\hat{\mu}, \hat{\alpha}$ of Lemma 4, to show that, for any distribution D of mean 0 and variance 1, there is a choice $\mathbf{d} = (d_\mu, d_\alpha)$ that makes Equation 4 sufficiently small. Phrased as an optimization problem, our goal is to evaluate (or tightly bound):

$$\max_D \min_{\mathbf{d}=(d_\mu, d_\alpha)} e^{-d_\mu \hat{\mu} + d_\alpha \frac{1}{3n} \log \frac{1}{\delta}} \mathbb{E}_{x \leftarrow D} (e^{d_\mu x(1-\min(\hat{\alpha}x^2, 1)) - d_\alpha \min(\hat{\alpha}x^2, 1)}) \quad (5)$$

where D ranges over distributions of mean 0 and variance 1.

We will use convex-concave programming and linear programming duality to significantly simplify the max-min program in Equation 5 before we dive into the part of analysis that is ad hoc for this problem. We wish to emphasize here again that the steps of 1) writing an estimator as a multi-parameter ψ -estimator and finding an analogous lemma to our Lemma 4, then 2) using mathematical programming duality to simplify the Chernoff bound analysis, are a framework generalizable for tightly analyzing other estimators.

For simplicity of exposition, assume that we restrict the support of D to some sufficiently fine-grained finite set, meaning that the maximization in Equation 5 is now finite-dimensional, albeit an arbitrarily large finite number. For each support element x , let D_x be a variable representing the probability of choosing x under distribution D . The expectation component of Equation 5 may now be expressed as sum that is a linear function in the variables D_x :

$$\max_D \min_{\mathbf{d}=(d_\mu, d_\alpha)} e^{-d_\mu \hat{\mu} + d_\alpha \frac{1}{3n} \log \frac{1}{\delta}} \sum_x D_x \cdot e^{d_\mu x(1-\min(\hat{\alpha}x^2, 1)) - d_\alpha \min(\hat{\alpha}x^2, 1)} \quad (6)$$

Using the standard max-min inequality (a form of weak duality in optimization), we have that Equation 6 is upper bounded by swapping the maximization and minimization (Equation 7), meaning that the vector \mathbf{d} no longer depends on the distribution D .

$$\min_{\mathbf{d}=(d_\mu, d_\alpha)} \max_D e^{-d_\mu \hat{\mu} + d_\alpha \frac{1}{3n} \log \frac{1}{\delta}} \sum_x D_x \cdot e^{d_\mu x(1-\min(\hat{\alpha}x^2, 1)) - d_\alpha \min(\hat{\alpha}x^2, 1)} \quad (7)$$

Crucially, however, Equation 7 is not just an upper bound on Equation 6, but is in fact *equal* to it, due to Sion's minimax theorem [23]. To apply Sion's minimax theorem, it suffices to check that 1) both \mathbf{d} and D are constrained to be in convex sets, at least one of which is compact, 2) the objective is convex in \mathbf{d} and 3) concave in the variables D_x . For the first condition, we note that the set of distributions on a finite domain is compact. The objective is convex in \mathbf{d} since the objective is the sum of exponentials that are each linear in \mathbf{d} . And the objective is concave in D_x because it is in fact a linear function of D .

The guarantee of Sion's minimax theorem means that we may work with Equation 7 instead of Equation 6 without sacrificing tightness in our analysis. This justifies why we are free to choose $\mathbf{d} = (d_\mu, d_\alpha)$ in Lemma 4 that does not depend on the distribution D .

To further simplify the problem in Equation 7, we note again that both the objective and the constraints on D are linear in the variables D_x , meaning that the inner maximization is in fact a linear program. We can then apply linear programming (strong) duality to yield the following equivalent optimization (Equation 8). We note that, as above, for the purposes of upper bounding

Equation 5, it suffices to only use weak duality. Strong duality however guarantees that this step does not introduce slack into the analysis.

The three variables V, M, S in the inner minimization below are the dual variables corresponding to the three constraints on distribution D originally: that D has variance 1, mean 0, and total probability mass 1.

$$\min_{\mathbf{d}=(d_\mu, d_\alpha)} \min_{V, M, S} V + S \quad (8)$$

for all x : $Vx^2 + Mx + S \geq e^{-d_\mu \hat{\mu} + d_\alpha \frac{1}{3n} \log \frac{1}{\delta} + d_\mu x(1 - \min(\hat{\alpha}x^2, 1)) - d_\alpha \min(\hat{\alpha}x^2, 1)}$

We have thus reduced the infinite-dimensional optimization problem of Equation 5 to the five-dimensional problem of Equation 8 (or six dimensions, if we include the universal quantification for $x \in \mathbb{R}$), a significant simplification. We bound Equation 8 by explicitly choosing values for $\mathbf{d} = (d_\mu, d_\alpha), V, M, S$ as functions of $\hat{\alpha}, n, \log \frac{1}{\delta}$, and showing that they jointly satisfy the constraint of Equation 8, for all x . We factor out the terms in the exponential that do not depend on x ; we make the variable substitutions $y \equiv \sqrt{\hat{\alpha}}x$ and $\hat{v} \equiv \frac{\log(1/\delta)}{3n\hat{\alpha}}$ to replace dependence on $\hat{\alpha}, n, \log \frac{1}{\delta}$ with dependence on the single variable \hat{v} ; taking the log of both sides (and swapping sides) yields an expression that is recognizable in the following lemma, where the multipliers of $1, y, y^2$ respectively on the right hand side are essentially our choices of S, M, V :

Lemma 7. *For all $\hat{v} \in [0.05, 55.5]$, there exist $a > 0$ and b such that*

$$\forall y \in \mathbb{R} : ay(1 - \min(y^2, 1)) - b \cdot \min(y^2, 1) \leq \log \left(1 + ay + y^2 \hat{v} \left(-3 + \frac{a\sqrt{6}}{\sqrt{\hat{v}}} - b \right) \right)$$

where $a \in [C, C']$ and $b \in [-C', C']$ for positive constants C, C' . Further, for $\hat{v} = 0.05$, the pair $a = 0.75, b = \sqrt{3}$ works.

We emphasize that the application of Lemma 7 in the proof of Lemma 4 below is straightforward, though finding the particular form of Lemma 7 is not. Further, one would not seek a result of the form of Lemma 7 without the guarantees of this section, derived via duality and mathematical programming, showing that “results of the form of Lemma 7 encompass the full power of the Chernoff bounds of Equation 4.” See the end of Section 5.2 for the proof of Lemma 7.

5.2 Proof of Lemma 4

We now prove Lemma 4 by combining the Chernoff bound analysis of Lemma 6 with the inequality from Lemma 7. We point out that the proof below is direct, without any reference to duality or mathematical programming; however, the discussion of Section 5.1 was crucial to discovering the right formulation for Lemma 7. We prove Lemma 7 at the end of the section.

Lemma 4. *Consider an arbitrary distribution D with mean 0 and variance 1. There exists a universal constant c where the following claim is true. Fixing $\hat{\mu} = \epsilon' = \left(1 + \frac{c \log \log \frac{1}{\delta}}{\log \frac{1}{\delta}}\right) \sqrt{\frac{2 \log \frac{1}{\delta}}{n}}$, then for all δ smaller than some universal constant, and for all $\hat{v} \in [0.05, 55.5]$, there exists a vector $\mathbf{d}(\hat{v})$ where $d_\mu \geq 0$, and both $\sqrt{\frac{n}{\log(1/\delta)}} |d_\mu|, |d_\alpha|$ are bounded by a universal constant, such that*

$$\mathbb{P}_{X \leftarrow D^n} \left(\mathbf{d}(\hat{v}) \cdot \psi \left(X, \hat{\mu} = \epsilon', \hat{\alpha} = \frac{\log(1/\delta)}{3n\hat{v}} \right) > \frac{1}{\log \frac{1}{\delta}} \right) \geq 1 - \frac{\delta}{\log^4 \frac{1}{\delta}}$$

Furthermore, for $\hat{v} = 0.05$ we have $d_\mu = \sqrt{3.75 \frac{\log(1/\delta)}{n}}$, $d_\alpha = \sqrt{3}$; and for $\hat{v} = 55.5$ we have $d_\mu = 0$, $d_\alpha < 0$.

Proof. Start with the bound on the probability of failure given by Lemma 6:

$$2 \left(e^{-d_\mu \hat{\mu} + d_\alpha \frac{1}{3n} \log \frac{1}{\delta}} \mathbb{E}_{x \leftarrow D} (e^{d_\mu x(1 - \min(\hat{\alpha}x^2, 1)) - d_\alpha \min(\hat{\alpha}x^2, 1)} \right)^n$$

For $\hat{v} \in [0.05, 55.5)$ we bound the exponential inside the expectation via the exponential of Lemma 7; we also use Lemma 7 to choose d_μ, d_α for us (the $\hat{v} = 55.5$ case is covered at the end). Namely, in Lemma 7 use \hat{v} as given, substitute $x\sqrt{\hat{\alpha}} \equiv y$ (where $\hat{\alpha} \equiv \frac{\log(1/\delta)}{3n\hat{v}}$ as always), and choose $d_\mu \equiv a\sqrt{\hat{\alpha}}$, and $d_\alpha \equiv b$ —in particular, for $\hat{v} = 0.05$ this gives $d_\mu(0.05) = 0.75\sqrt{\hat{\alpha}} = 0.75\sqrt{\frac{\log(1/\delta)}{3n\hat{v}}} = \sqrt{3.75 \frac{\log(1/\delta)}{n}}$. Thus the failure probability is bounded by

$$\begin{aligned} & 2 \left(e^{-d_\mu \hat{\mu} + d_\alpha \frac{1}{3n} \log \frac{1}{\delta}} \mathbb{E}_{\substack{x \leftarrow D \\ y = x\sqrt{\hat{\alpha}}}} \left(1 + ay + y^2 \hat{v} \left(-3 + \frac{a\sqrt{6}}{\sqrt{\hat{v}}} - b \right) \right) \right)^n \\ &= 2 \left(e^{-d_\mu \hat{\mu} + d_\alpha \frac{1}{3n} \log \frac{1}{\delta}} \left(1 + \frac{\log \frac{1}{\delta}}{3n} \left(-3 + 3d_\mu \sqrt{\frac{2n}{\log(1/\delta)}} - d_\alpha \right) \right) \right)^n \quad \text{since } D \text{ has mean 0, variance 1} \\ &\leq 2 \left(e^{-d_\mu \hat{\mu} + d_\alpha \frac{1}{3n} \log \frac{1}{\delta} + \frac{\log \frac{1}{\delta}}{3n} \left(-3 + 3d_\mu \sqrt{\frac{2n}{\log(1/\delta)}} - d_\alpha \right)} \right)^n \quad \text{since } 1 + z \leq e^z \text{ for any } z \\ &\leq 2e^{-d_\mu \sqrt{\frac{2n}{\log(1/\delta)}} c \log \log \frac{1}{\delta} - \log \frac{1}{\delta}} \quad \text{substituting } \hat{\mu} = \epsilon' = \left(1 + \frac{c \log \log \frac{1}{\delta}}{\log \frac{1}{\delta}} \right) \sqrt{\frac{2 \log \frac{1}{\delta}}{n}} \\ &\leq \frac{\delta}{\log^4 \frac{1}{\delta}} \quad \text{as desired, for large enough } c, \text{ since } d_\mu \sqrt{\frac{n}{\log(1/\delta)}} = \frac{a}{\sqrt{3\hat{v}}} \text{ is greater than some positive constant.} \end{aligned}$$

We prove the $\hat{v} = 55.5$ case now. We choose $d_\mu = 0$ and $d_\alpha = -4$, substituting into the bound of Equation 4 to yield

$$\begin{aligned} 2 \left(e^{-\frac{4}{3n} \log \frac{1}{\delta}} \mathbb{E}_{x \leftarrow D} (e^{4 \min(\hat{\alpha}x^2, 1)}) \right)^n &\leq 2\delta^{4/3} \mathbb{E}_{x \leftarrow D} (1 + 54\hat{\alpha}x^2)^n \quad \text{since for } y \in [0, 1], e^{4y} \leq 1 + 54y \\ &= 2\delta^{4/3} (1 + 55.5\hat{\alpha})^n \quad \text{since } D \text{ has variance 1} \\ &\leq 2\delta^{4/3} e^{n \cdot 54 \frac{\log(1/\delta)}{3 \cdot 55.5n}} \quad \text{since } 1 + z \leq e^z, \text{ substituting def. of } \hat{\alpha} \\ &= 2\delta^{4/3} \delta^{-\frac{54}{3 \cdot 55.5}} \leq 2\delta^{1.009} \end{aligned}$$

which is bounded as desired for small enough δ . □

We now prove Lemma 7.

Lemma 7. *For all $\hat{v} \in [0.05, 55.5]$, there exist $a > 0$ and b such that*

$$\forall y \in \mathbb{R} : ay(1 - \min(y^2, 1)) - b \cdot \min(y^2, 1) \leq \log \left(1 + ay + y^2 \hat{v} \left(-3 + \frac{a\sqrt{6}}{\sqrt{\hat{v}}} - b \right) \right) \quad (9)$$

where $a \in [C, C']$ and $b \in [-C', C']$ for positive constants C, C' . Further, for $\hat{v} = 0.05$, the pair $a = 0.75, b = \sqrt{3}$ works.

Proof. We first prove the special case of 1) $\hat{v} = 0.05$, before moving to the general case of 2) $\hat{v} \in (0.05, 55.5]$. We note that our choice of $a(\hat{v}), b(\hat{v})$ is *not* continuous in \hat{v} at 0.05, but the usage of the lemma does not require any continuity. We choose a, b at the edge case $\hat{v} = 0.05$ for convenience.

1) For $\hat{v} = 0.05$, we choose $a = 0.75, b = \sqrt{3}$. This special case of Equation 9 simplifies to:

$$\forall y \in \mathbb{R} : 0.75y (1 - \min(y^2, 1)) - \sqrt{3} \cdot \min(y^2, 1) \leq \log(1 + 0.75y + 0.174y^2)$$

(where 0.174 is a lower bound on $\hat{v}(-3 + \frac{a\sqrt{6}}{\sqrt{\hat{v}}} - b)$). This is a 1-dimensional bound and can be easily analyzed in many ways. For the range $y \in [-1, 1]$: the right hand side is at least $\log(1 + 0.75y)$, which in this range is at least $.75y - .75y^2$, which is easily shown to be greater than the polynomial expression that the left hand side reduces to in this range, $0.75y - \sqrt{3}y^2 - 0.75y^3$. For the remaining range, $y \notin [-1, 1]$, the left hand side is the constant $-\sqrt{3}$, and it is easy to check that the quadratic in the argument of the right hand side, $1 + 0.75y + 0.174y^2$, always exceeds $e^{-\sqrt{3}}$.

2) To show Equation 9 for the rest of the range of $\hat{v} \in (0.05, 55.5]$, we choose a to be the positive root of the quadratic equation $\sqrt{\hat{v}}(a^2 - 12) + \sqrt{6}a = 0$ and let $b = 3 - a^2/2$ —we will see the motivation for this choice shortly. For now, note that the definition of a implies $a \leq \sqrt{12}$, for otherwise $\sqrt{\hat{v}}(a^2 - 12) + \sqrt{6}a$ would be greater than 0.

Our proof will analyze the sign of the derivative with respect to y of the difference between the right and left hand sides of Equation 9. For the critical region $|y| \leq 1$ this derivative equals:

$$\frac{a + 2y\hat{v}(-3 + \frac{a\sqrt{6}}{\sqrt{\hat{v}}} - b)}{1 + ay + y^2\hat{v}(-3 + \frac{a\sqrt{6}}{\sqrt{\hat{v}}} - b)} - a + 3ay^2 + 2by \quad (10)$$

The crucial step is to choose a to be the positive root of the quadratic equation $\sqrt{\hat{v}}(a^2 - 12) + \sqrt{6}a = 0$ and let $b = 3 - a^2/2$, after which Equation 10 miraculously factors as

$$\frac{1}{3a^2} \cdot \frac{y(y + \frac{2}{a})(y + \frac{2}{a} - \frac{a}{3})^2}{y^2 + (\frac{4}{a} - \frac{a}{3})y + (\frac{4}{a^2} - \frac{1}{3})}$$

From this expression for the derivative, it is straightforward to read off its sign. The discriminant of the quadratic in the denominator is $\frac{1}{9}(a^2 - 12) > 0$, meaning the denominator is always positive. The squared term in the numerator cannot affect the overall sign. Thus the sign of the derivative equals the sign of $y(y + \frac{2}{a})$, meaning that the difference between the right and left side of Equation 9 is monotonically increasing for $y > 0$, and unimodal for $y < 0$, having non-positive derivative for $y \in [\frac{2}{a}, 0]$ and nonnegative derivative for smaller y . Thus to show the inequality holds for all $y \in [-1, 1]$ it suffices to check it at $y = 0$ and $y = -1$.

The $y = 0$ case is trivial as both sides of Equation 9 equal 0.

For $y = -1$, Equation 9, after expressing both $\sqrt{\hat{v}}$ and b in terms of a becomes

$$\frac{a^2}{2} - 3 \leq \log\left(-2 + a - \frac{36}{a^2 - 12}\right) \quad (11)$$

For $a \in [0, \sqrt{12})$, the inverse of the rational expression inside the log is bounded by its linear approximation, $\frac{\sqrt{12}-a}{\sqrt{12}}$. Calling this a new variable $z = \frac{\sqrt{12}-a}{\sqrt{12}}$, which is between 0 and 1, Equation 11 becomes the claim that $6(1-z)^2 - 3 \leq -\log z$, which is easily verified for $z \in (0, 1]$.

Lastly, we show Equation 9 for $|y| > 1$. Reexpressing b and $\sqrt{\hat{v}}$ in terms of a , the left hand side of the inequality is the constant value $-b = -3 + \frac{a^2}{2}$ (independent of y), while the right hand side

is $\log(1 + ay + \frac{3a^2}{12-a^2}y^2)$. Analyzing the quadratic inside the log shows that the right hand side has a minimum of $\frac{a^2}{12}$, attained at $y = -\frac{12-a^2}{6a}$.

When the location of this minimum, $y = -\frac{12-a^2}{6a}$, is inside the interval $[-1, 1]$, then because this quadratic is monotonic to either side of the minimum, the fact that we have already proven Equation 9 for $y = \pm 1$ implies the inequality holds for all y further from 0.

The remaining case is when the minimum is not in $[-1, 1]$, namely, $-\frac{12-a^2}{6a} < -1$, meaning $a < 1.59$; since a is monotonic in \hat{v} , a is at least its value when $\hat{v} = 0.05$, namely $a \geq 1.003$. Equation 9 thus reduces to showing that, for $a \in [1.003, 1.59]$ we have $\frac{a^2}{2} - 3 \leq \log \frac{a^2}{12}$, which is trivially implied, substituting $z = \frac{a^2}{12}$, by the inequality $6z - 3 \leq \log z$ for $z \in [0.083, 0.22]$, yielding the claim. \square

5.3 Proof of Lemma 5

Lemma 5. *Consider an arbitrary set of n samples X . Consider the expressions $\psi_\mu(X, \hat{\mu}, \hat{\alpha}), \psi_\alpha(X, \hat{\alpha})$, reparameterized in terms of $\hat{v} \equiv \frac{\log(1/\delta)}{3n\hat{\alpha}}$ in place of $\hat{\alpha}$. Suppose the equation $\psi_\alpha(X, \hat{\alpha}) = 0$ has a solution in the range $\hat{v} \in [0.05, 55.5]$. Then the functions $\sqrt{\frac{\log(1/\delta)}{n}}\psi_\mu(X, \hat{\mu}, \hat{\alpha})$ and $\psi_\alpha(X, \hat{\alpha})$ are Lipschitz with respect to \hat{v} on the entire interval $\hat{v} \in [0.05, 55.5]$, with Lipschitz constant $c \log \frac{1}{\delta}$ for some universal constant c .*

Proof. Consider the \hat{v} derivative of $\psi_\alpha(X, \hat{\mu}, \hat{\alpha} \equiv \frac{\log(1/\delta)}{3n\hat{v}}) = \sum_{i=1}^n \left(\min \left(\frac{\log(1/\delta)}{3n\hat{v}} x_i^2, 1 \right) - \frac{1}{3n} \log \frac{1}{\delta} \right)$. The \hat{v} derivative of $\min \left(\frac{\log(1/\delta)}{3n\hat{v}} x_i^2, 1 \right)$ is either $-\frac{\log(1/\delta)}{3n\hat{v}^2} x_i^2 = -\frac{1}{\hat{v}} \hat{\alpha} x_i^2$ or 0, depending on which term in the min is the smallest, and in either case has magnitude at most $\frac{1}{\hat{v}} \min(\hat{\alpha} x_i^2, 1)$. Thus the overall \hat{v} derivative of $\psi_\alpha(X, \hat{\mu}, \hat{\alpha})$ has magnitude at most $\frac{1}{\hat{v}} \sum_i \min(\hat{\alpha} x_i^2, 1)$. Since, we are guaranteed that $\sum_{i=1}^n \min(\hat{\alpha} x_i^2, 1) = \frac{1}{3} \log \frac{1}{\delta}$ for some $\hat{v} \in [0.05, 55.5]$, we thus have that the derivative is within a constant factor of this across the entire range, as desired.

Similarly, consider the \hat{v} derivative of $\psi_\mu(X, \hat{\mu}, \hat{\alpha}) = \sum_{i=1}^n (\hat{\mu} - x_i (1 - \min(\hat{\alpha} x_i^2, 1)))$. The i^{th} term of this is the \hat{v} derivative of $\min(\hat{\alpha} x_i^2, 1)$, which is either $-\frac{1}{\hat{v}} \hat{\alpha} x_i^2$ or 0 depending on whether $x_i \leq \sqrt{1/\hat{\alpha}}$, and thus the magnitude of this derivative may be bounded by $\frac{1}{\hat{v}\sqrt{\hat{\alpha}}} \sum_{i=1}^n \min(\hat{\alpha} x_i^2, 1)$. Since $\sum_{i=1}^n \min(\hat{\alpha} x_i^2, 1)$ is bounded by a constant times $\log \frac{1}{\delta}$ (as in the last paragraph), and $\frac{1}{\hat{v}\sqrt{\hat{\alpha}}}$ is bounded by a constant times $\frac{1}{\sqrt{\hat{v}\hat{\alpha}}} = \sqrt{\frac{3n}{\log(1/\delta)}}$, the magnitude of the derivative of $\sqrt{\frac{\log(1/\delta)}{n}}\psi_\mu(X, \hat{\mu}, \hat{\alpha})$ is bounded by a constant times $\log \frac{1}{\delta}$, as desired. \square

References

- [1] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.
- [2] Christian Brownlees, Emilien Joly, Gábor Lugosi, et al. Empirical risk minimization for heavy-tailed losses. *Ann. Stat.*, 43(6):2507–2536, 2015.
- [3] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Trans. Inf. Theory*, 59(11):7711–7717, 2013.
- [4] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. I. H. Poincaré-PR*, 48(4):1148–1185, 2012.
- [5] Olivier Catoni and Ilaria Giulini. Dimension-free pac-bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv:1712.02747*, 2017.
- [6] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L Bartlett. Fast mean estimation with sub-gaussian rates. In *Proc. COLT '20*, pages 786–806, 2019.
- [7] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-gaussian mean estimators. *Ann. Stat.*, 44(6):2695–2725, 2016.
- [8] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *Proc. FOCS '13*, pages 429–438, 2013.
- [9] John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *J. Am. Stat. Assoc.*, 113(521):182–201, 2018.
- [10] Samuel B Hopkins et al. Mean estimation with sub-gaussian rates in polynomial time. *Ann. Stat.*, 48(2):1193–1213, 2020.
- [11] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.*, 17(1):543–582, 2016.
- [12] Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theor. Comput. Sci.*, 43:169–188, 1986.
- [13] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory*, 61(5):2835–2885, 2015.
- [14] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. volume 125 of *Proc. COLT '20*, pages 2204–2235. PMLR, 09–12 Jul 2020.
- [15] Zhixian Lei, Kyle Luh, Prayaag Venkat, and Fred Zhang. A fast spectral algorithm for mean estimation with sub-gaussian rates. In *Proc. COLT '20*, pages 2598–2612, 2020.
- [16] Gabor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions—a survey, 2019.
- [17] Gábor Lugosi, Shahar Mendelson, et al. Sub-Gaussian estimators of the mean of a random vector. *The annals of statistics*, 47(2):783–794, 2019.
- [18] Stanislav Minsker. Uniform bounds for robust mean estimators. *arXiv:1812.03523*, 2018.

- [19] Ankur Moitra and Michael Saks. A polynomial time algorithm for lossy population recovery. In *Proc. FOCS'13*, pages 110–116, 2013.
- [20] A.S. Nemirovsky and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.
- [21] Yury Polyanskiy, Ananda Theertha Suresh, and Yihong Wu. Sample complexity of population recovery. In *Proc. COLT'17*, volume 65, 2017.
- [22] Yury Polyanskiy and Yihong Wu. Dualizing Le Cam’s method, with applications to estimating the unseens. *arXiv:1902.05616*, 2019.
- [23] Maurice Sion. On general minimax theorems. *Pac. J. Math*, 8(1):171–176, 1958.
- [24] Leonard A. Stefanski and Dennis D. Boos. The calculus of M-estimation. *The American Statistician*, 56(1):29–38, 2002.
- [25] Gregory Valiant and Paul Valiant. The power of linear estimators. In *Proc. FOCS'11*, pages 403–412, 2011.
- [26] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inf. Theory*, 62(6):3702–3720, 2016.
- [27] Yihong Wu, Pengkun Yang, et al. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Ann. Stat*, 47(2):857–883, 2019.

A Additional “3rd Order” Motivation for Our Estimator

In this appendix, we give additional motivation of our estimator as a “3rd order correction” to the sample mean.

Suppose (for this section only), as in [4], that one knows the variance $\sigma^2(D)$ of the distribution in question, or has a good estimate of it.

Example 2. *Given samples x_1, \dots, x_n from a distribution of mean 0 and variance 1 and bounded higher moments, suppose our goal is to construct a slight variant of the empirical mean that will robustly return an estimate that is close to 0, the true mean; we consider estimates of the form $\frac{1}{n} \sum_{i=1}^n (x_i + c(x_i))$ for some function $c: \mathbb{R} \rightarrow \mathbb{R}$. Explicitly, given a bound b , we want our estimate to be between $\pm b$, with as high probability as possible. For simplicity we will consider the positive case, namely, bounding $\mathbb{P}_{x_1, \dots, x_n}(\frac{1}{n} \sum_i (x_i + c(x_i)) \geq b)$. With a view towards deriving a Chernoff bound, we rearrange, multiply by an arbitrary positive constant α , and exponentiate inside the probability to yield that this probability equals $\mathbb{P}_{x_1, \dots, x_n}(\exp(\alpha \sum_i (x_i + c(x_i) - b)) \geq 1)$; by Markov’s inequality, this probability is at most $\mathbb{E}_{x_1, \dots, x_n}(\exp(\alpha \sum_i (x_i + c(x_i) - b)))$, for our choice of $\alpha > 0$. We set $\alpha = b$. Since each x_i is independent, this probability becomes $\mathbb{E}_x(\exp(b(x + c(x) - b)))^n$.*

Considering the empirical estimator, where $c(x_i) = 0$, we thus have that the probability the empirical mean estimate exceeds b is at most the n^{th} power of $\mathbb{E}_x(\exp(b(x - b^2)))$, where this expression can be expanded to 3rd order as

$$e^{-b^2} \left(1 + b \mathbb{E}(x) + \frac{1}{2} b^2 \mathbb{E}(x^2) + \frac{1}{6} b^3 \mathbb{E}(x^3) + O(x^4) \right)$$

As we assumed the data distribution has mean 0 and variance 1, we can simplify the above expression to

$$e^{-b^2} \left(1 + \frac{1}{2} b^2 + \frac{1}{6} b^3 \mathbb{E}(x^3) + O(b^4) \right)$$

Ignoring, for the moment, the 3rd or higher-order terms, this expression is $e^{-b^2}(1 + \frac{1}{2}b^2) \approx e^{-b^2/2}$, whose n^{th} power equals $e^{-b^2 n/2}$, which is exactly the bound one would expect for the standard Gaussian case, of the probability that the empirical mean of n samples is more than b from the true value. However, the 3rd order term is a crucial obstacle here, as the third moment $\mathbb{E}(x^3)$ could be of either sign, skewing either the left tail or right tail to have substantially more mass than in our benchmark of the Gaussian case.

We thus choose a correction function $c(x_i)$ so as to cancel out this 3rd-order term and improve the estimate in this regime: to cancel out the term $\frac{1}{6}b^3 \mathbb{E}(x^3)$ in the 3rd-order expansion of our Chernoff bound $\mathbb{E}_x(\exp(b(x - b)))$, we replace x by $x - \frac{1}{6}x^3b^2$, yielding a bound on the failure probability of the n^{th} power of

$$e^{-b^2} \left(1 + \frac{1}{2} b^2 + O(b^4) \right) = e^{-b^2/2 + O(b^4)}$$

as desired.

For the sake of clarity, we can change variables, letting the leading term of our probability bound $e^{-b^2 n/2}$ equal δ , and thus the correction $-\frac{1}{6}x^3b^2$ becomes $c(x) \equiv -\frac{1}{n}x^3\frac{1}{3} \log \frac{1}{\delta}$, meaning the correction amounts essentially to a 3rd moment correction, split n ways and scaled by the same $\frac{1}{3} \log \frac{1}{\delta}$ of our main algorithm.

We explicitly relate this estimator to Estimator 1 by pointing out that, when none of the samples x_i are “truncated” by Estimator 1 (namely, $\hat{\alpha}x_i^2 \leq 1$ always), then $\hat{\alpha} \equiv \frac{\log(1/\delta)}{3n\bar{v}}$ may be expressed in

terms of the empirical variance \hat{v} ; taking $\kappa = 0$ for simplicity, the returned estimate will be $\frac{1}{n} \sum_i x_i - \alpha x_i^3 = \frac{1}{n} \sum_i x_i - \frac{1}{\hat{v}n} x_i^3 \frac{1}{3} \log \frac{1}{\delta}$, which equals the above-derived “3rd-order corrected estimator” when the empirical variance is the true variance, 1.

In the above example we showed that Chernoff bounds for the empirical mean deteriorate for distributions with large 3rd moments (skew), and that adding a 3rd-order correction to the empirical mean corrects for this, leaving essentially “Gaussian-like” performance. These calculations motivate several features of Estimator 1—including the $\frac{1}{3} \log \frac{1}{\delta}$ parameter, and the 3rd-order terms in the expression for $\hat{\mu}$ —even though the overall form of Estimator 1 is rather different, as it must work in all regimes and not just in the cartoon asymptotic regime considered in this example.

B Proposition 3 implies Theorem 1

For completeness’ sake, we explicitly state and prove the properties described in Section 3.2, which combine to show that Proposition 3 implies Theorem 1.

Lemma 8. *Suppose X is a set of samples in \mathbb{R} . Then for any $\delta > 0$ and any scale $a > 0$ and shift b ,*

$$\hat{\mu}(aX + b, \delta) = a \hat{\mu}(X, \delta) + b$$

where $\hat{\mu}$ denotes the output of Estimator 1.

The above lemma follows trivially from the fact that the median-of-means estimate also respects shift and scale in the input samples, and that α is chosen in Step 2 of Estimator 1 so that $\min(\alpha(x_i - \kappa)^2, 1)$ does not depend on the affine parameters a, b .

Fact 9 ([11]). *For any distribution D with mean μ and standard deviation σ , the median-of-means estimate κ , on input n samples, satisfies*

$$\mathbb{P} \left(|\kappa - \mu| > O \left(\sigma \sqrt{\frac{\log \frac{1}{\delta}}{n}} \right) \right) \leq \delta$$

Lemma 10. *Consider a fixed sample set X of size n , and a confidence parameter δ . Let $e(X, \delta, \kappa)$ denote Estimator 1 but where Step 1 is omitted and κ is instead considered as an input. Then,*

$$\left| \frac{d e(X, \delta, \kappa)}{d \kappa} \right| = O \left(\sqrt{\frac{\log \frac{1}{\delta}}{n}} \right)$$

Fact 9 shows that, except with δ probability, the median-of-means estimate is within $O \left(\sigma \sqrt{\frac{\log \frac{1}{\delta}}{n}} \right)$ of the true mean, and multiplying this by the Lipschitz constant $O \left(\sqrt{\frac{\log \frac{1}{\delta}}{n}} \right)$ from Lemma 10 shows that the change in output of Algorithm 1, between using the median-of-means versus setting $\kappa = 0$, has magnitude $O \left(\sigma \sqrt{\frac{\log \frac{1}{\delta}}{n}} \right)^2 = o \left(\sigma \sqrt{\frac{\log \frac{1}{\delta}}{n}} \right)$. This discrepancy is therefore a $o(1)$ fraction of the additive error guaranteed by Theorem 1.

We now prove Lemma 10.

Proof. We compute the derivatives with respect to α and κ of the $\hat{\mu}$ (computed in Step 3 of Estimator 1), and the expression on the left hand side of Step 2, which we denote $\nu \equiv \sum_i \min(\alpha(x_i - \kappa)^2, 1)$. We note that for terms where $\min(\alpha(x_i - \kappa)^2, 1) = 1$, all derivatives are 0, so we adopt the notation “ $\sum_{<}$ ” to denote summing only over those indices i for which $\alpha(x_i - \kappa)^2 < 1$. Thus we have

$$\begin{aligned}\frac{d\nu}{d\kappa} &= 2 \sum_{<} \alpha(x_i - \kappa) \\ \frac{d\nu}{d\alpha} &= \sum_{<} (x_i - \kappa)^2 \\ \frac{d\hat{\mu}}{d\kappa} &= 1 + \frac{1}{n} \sum_{<} (-1 + 3\alpha(x_i - \kappa)^2) \\ \frac{d\hat{\mu}}{d\alpha} &= -\frac{1}{n} \sum_{<} (x_i - \kappa)^3\end{aligned}$$

Recall that α is defined implicitly so as to make the expression $\nu = \frac{1}{3} \log \frac{1}{\delta}$; thus in Estimator 1, if we change κ at a rate of 1, then α also changes at rate $-\frac{d\nu}{d\kappa} / \frac{d\nu}{d\alpha}$ to keep ν unchanged. Thus, the overall derivative of the estimate with respect to changing κ equals $\frac{d\hat{\mu}}{d\kappa} - \frac{d\hat{\mu}}{d\alpha} \frac{d\nu}{d\kappa} / \frac{d\nu}{d\alpha}$. We bound this from the derivatives computed above.

To bound $\frac{d\hat{\mu}}{d\kappa}$, we note that the number of indices *not* in the sum “ $\sum_{<}$ ” is at most $\frac{1}{3} \log \frac{1}{\delta}$ because each such i contributes 1 to the left hand side of the condition in Step 2 of Estimator 1 and the right hand side equals $\frac{1}{3} \log \frac{1}{\delta}$. Thus the initial terms of $\frac{d\hat{\mu}}{d\kappa}$ are bounded as $1 + \frac{1}{n} \sum_{<} (-1) \leq \frac{1}{3n} \log \frac{1}{\delta}$. The remaining part of $\frac{d\hat{\mu}}{d\kappa}$, namely $\frac{1}{n} \sum_{<} 3\alpha(x_i - \kappa)^2$ is $\frac{3}{n}$ times the corresponding terms in $\nu \leq \frac{1}{3} \log \frac{1}{\delta}$ itself, and thus is at most $\frac{1}{n} \log \frac{1}{\delta}$. Thus $\frac{d\hat{\mu}}{d\kappa} = O(\frac{1}{n} \log \frac{1}{\delta})$.

We now bound the remaining term $-\frac{d\hat{\mu}}{d\alpha} \frac{d\nu}{d\kappa} / \frac{d\nu}{d\alpha}$. Since for each index i in “ $\sum_{<}$ ” we have $|x_i - \kappa| \leq \frac{1}{\sqrt{\alpha}}$, we may bound $\frac{d\hat{\mu}}{d\alpha}$, involving a 3rd moment term, by the simpler $|\frac{d\hat{\mu}}{d\alpha}| \leq \frac{1}{n} \sum_{<} |x_i - \kappa|^2 / \sqrt{\alpha}$. Combining this, with the other derivatives and the bound $\alpha \leq \frac{1}{3} \log \frac{1}{\delta} / \sum_{<} (x_i - \kappa)^2$ from the previous paragraph yields:

$$\left| \frac{d\hat{\mu}}{d\alpha} \frac{d\nu}{d\kappa} / \frac{d\nu}{d\alpha} \right| \leq \frac{2\sqrt{\alpha}}{n} \left| \frac{\sum_{<} (x_i - \kappa) \sum_{<} (x_i - \kappa)^2}{\sum_{<} (x_i - \kappa)^2} \right| \leq \frac{2\sqrt{\frac{1}{3} \log \frac{1}{\delta}}}{n} \left| \frac{\sum_{<} (x_i - \kappa)}{\sqrt{\sum_{<} (x_i - \kappa)^2}} \right| \leq \sqrt{\frac{4 \log \frac{1}{\delta}}{3n}}$$

where the last inequality is Cauchy-Schwarz applied to the sequence $(x_i - \kappa)$ and the all-1s sequence. \square