Near-Optimal Mean Estimation with Unknown, Heteroskedastic Variances

Spencer Compton
Stanford University
Stanford, USA
comptons@stanford.edu

Gregory Valiant
Stanford University
Stanford, USA
gvaliant@stanford.edu

ABSTRACT

Given data drawn from a collection of Gaussian variables with a common mean but different and unknown variances, what is the best algorithm for estimating their common mean? We present an intuitive and efficient algorithm for this task. As different closed-form guarantees can be hard to compare, the Subset-of-Signals model serves as a benchmark for "heteroskedastic" mean estimation: given n Gaussian variables with an unknown subset of m variables having variance bounded by 1, what is the optimal estimation error as a function of n and m? Our algorithm resolves this open question up to logarithmic factors, improving upon the previous best known estimation error by polynomial factors when $m = n^c$ for all $0 < c < 1$. Of particular note, we obtain error $o(1)$ with $m = \tilde{O}(n^{1/4})$ variance-bounded samples, whereas previous work required $m = \tilde{\Omega}(n^{1/2})$. Finally, we show that in the multi-dimensional setting, even for $d = 2$, our techniques enable rates comparable to knowing the variance of each sample.

CCS CONCEPTS

• **Theory of computation** → **Design and analysis of algorithms**; • **Mathematics of computing** → *Probability and statistics.*

KEYWORDS

heteroskedastic statistics, mean estimation

ACM Reference Format:
Spencer Compton and Gregory Valiant. 2024. Near-Optimal Mean Estimation with Unknown, Heteroskedastic Variances. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC '24), June 24–28, 2024, Vancouver, BC, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3618260.3649754

1 INTRODUCTION

Over the past decade, there has been a significant effort from the theoretical computer science and machine learning communities to reexamine fundamental learning and statistical estimation problems in non-i.i.d. settings. Many of these efforts have focused on relaxing the independence assumption. This includes the large body of work on *robust statistics*, where a portion of the data are assumed to be drawn i.i.d. from a fixed distribution and no assumptions are made about the remainder of the data. On the TCS side, work in robust statistics began by considering the problem of mean estimation in the Gaussian setting [8, 18], and then built up to considering more complex problems of learning or optimization (e.g [2, 10]).

Here, we instead consider the heterogeneous data setting, where samples are drawn independently, but from non-identical distributions. Even for some of the most fundamental problems, such as the problem of mean estimation with Gaussian data that we consider, much is still unknown about both the information theoretic and computational landscapes in this heterogeneous but independent setting. This is despite the practical importance of accurately extracting information from datasets whose contents have been gathered from heterogeneous sources (e.g. sourced from different workers, contributed by different hospitals or doctors, scraped from different websites, etc.).

Concretely, we consider the setting where we observe n independent heteroskedastic (meaning having different variances) Gaussian random variables that have a common mean: $X_1 \sim N(\mu, \sigma_1^2), \dots, X_n \sim N(\mu, \sigma_n^2)$, and our goal is to estimate their common mean, μ. Crucially, the variances σ_i^2 are *unknown*. This problem was explored in both the $d = 1$ and higher dimensional settings in the work of Chierichetti, Dasgupta, Kumar, and Lattanzi [4]. In the case where the variances are known, the unbiased estimator that weights X_i proportionally to $1/\sigma_i^2$ is easily shown to achieve optimal error $\Theta(1/\sqrt{\sum 1/\sigma_i^2})$ [15][1]. When the variances are unknown, however, both the problem and the optimal rates seems to change fundamentally.

In an effort to expose the core challenges of this problem, Liang and Yuan [20] introduced the Subset-of-Signals variant, parameterized by two numbers, m, n: as above, one observes n independent Gaussian random variables with a common mean, $X_1, \dots, X_n$, with the assumption that m have variance at most 1, and one makes no assumptions about the variances of the remaining $n - m$. Our results address the more general formulation, though are easier to interpret in this Subset-of-Signals setting, for which our approach achieves the known lower bounds, up to logarithmic factors.

1.1 Related Work

As mentioned above, this problem of heteroskedastic mean estimation was considered by Chierichetti, Dasgupta, Kumar, and Lattanzi in the $d = 1$ dimensional and (isotropic) high dimensional setting where $X_i \sim N(\mu, \sigma_i^2 I)$ [4]. Note that in this formulation, mean estimation becomes easier as d becomes larger, as there is more

STOC '24, June 24–28, 2024, Vancouver, BC, Canada

ACM ISBN 979-8-4007-0383-6/24/06
https://doi.org/10.1145/3618260.3649754

[1] Theorem 3.1 of [4] also contains a short proof of this.

information with which to infer the values of σ_i. Thus, while independent, the observations in different dimensions are often called "entangled." When $d = \Omega(\log(n))$, [4] attain estimation error of $\Theta\left(\sqrt{\frac{1}{\sum_{i=2}^{n} \frac{1}{\sigma_i^2}}}\right)$ for each dimension with high probability. Note that this is nearly identical to the classical known-variance rate, other than missing the dependence on σ_1. These results prompted subsequent work to focus on the more challenging small dimensional or one-dimensional settings for which it is more difficult or impossible to accurately recover the σ_i's.

In the one-dimensional setting, [4] attains a guarantee with respect to the $O(\log(n))$ smallest σ_i, giving an algorithm with expected error $\mathbb{E}[|\mu - \hat{\mu}|] = \min_{2 \le k \le \log(n)} \tilde{O}(n^{1/2(1+1/(k-1))}\sigma_k)$. Moreover, they showed lower bounds that demonstrated how the known-variance rates can be polynomially better than an optimal estimator that does not know the variances.

Subsequent works, [6, 7, 20, 22, 23, 26], which we discuss below, improve upon this in various regimes: their upper and lower bounds in the case of the Subset-of-Signals setting, together with our results, are depicted in Figure 1.

The work of Pensia, Jog, and Loh (preliminarily [22] and later [23]) develops machinery for analyzing the performance of classic estimators in this setting: the modal estimator, k-closest estimator, and the median. Using this, they show guarantees for a hybrid estimator and give complementary lower bounds that illustrate how under some conditions on $\sigma_1, \ldots, \sigma_n$ their estimator is near-optimal.[2] They also investigate the setting of heteroskedastic linear regression, as well as showing guarantees for their algorithm in $d > 1$ dimensions. Moreover, their results generalize from Gaussian distributions to radially symmetric and unimodal distributions.

The work of Devroye, Lattanzi, Lugosi, and Zhivotovskiy (preliminarily [6] and later [7]) also develops tools for sharp analysis of the sample median and modal estimator. In order to provide an adaptive algorithm requiring no parameter tuning, they employ subroutines that yield confidence intervals which they eventually intersect. Our algorithm will utilize a similar paradigm of intersecting confidence intervals obtained by (different) subroutines.

The works of Liang and Yuan [20, 26] provide estimation guarantees for the iterative trimming algorithm (a widely used heuristic). Importantly, they also introduce the Subset-of-Signals model, where m samples have variance bounded by 1, and it is desired to know the optimal estimation guarantee as a function of n and m. This framing is particularly helpful because the closed-form guarantees of various related work can otherwise be difficult to directly compare. In Fig. 1, we show the guarantees of related work in terms of the Subset-of-Signals model. Finally, Liang and Yuan show lower bounds for the optimal estimation error in this model.

We also highlight an alternative avenue (that we do not employ) for approaching heteroskedastic mean estimation. Consider the similar (but slightly different) task of receiving n i.i.d. samples from a mixture of n Gaussians, each with mean μ and variance σ_i^2. It appears plausible, for example, to analyze the Fisher information of this distribution, which would classically imply a lower bound and asymptotic guarantees. However, we are not aware of tools that would readily enable the desired finite-sample guarantees for this setting. For example, recent works of [11–13] obtain guarantees in terms of a distribution's *smoothed Fisher information*, yet such smoothing would result in suboptimal guarantees for this setting.

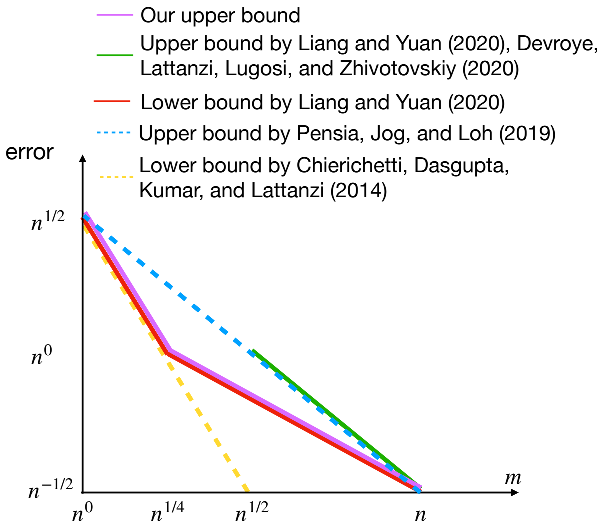

Figure 1: Guarantees of our upper bound and those of prior work for mean estimation in the Subset-of-Signals model, where one observes n independent Gaussian random variables with a common mean, and an unknown subset of $m \le n$ samples have variance at most 1, with no assumptions on the variance of the remaining $n - m$. The x-axis denotes m, the number of samples with variance bounded by 1, and the y-axis denotes the estimation error. Our upper bound matches the known lower bound up to logarithmic factors, and improves the estimation error by polynomial factors when $m = n^c$ for $0 < c < 1$. (Figure based on plot from [20]).

Related Work Beyond Heteroskedastic Mean Estimation: The previously mentioned work of [23] studied heteroskedastic linear regression where the variance of each observation's noise is arbitrary but independent of the covariates. Heteroskedastic linear regression has also been studied in the fundamentally different setting where noise variance is a rank 1 quadratic function of the covariates [5]. There have been several lines of work exploring property testing, estimation, and learning in settings with independent, but non-identical samples. These models span a large spectrum in terms of how much heterogeneity is present, relative to the sample size. On one extreme, there is a large volume of work on learning *mixture models* (of Gaussians, linear regressions, etc., see e.g. [1, 9, 14, 17, 21]). Typically, in these settings there are a small number (often just a constant number) of distributions, and each datapoint is drawn i.i.d. from one of these. Comparatively fewer works explore the other extreme, where a single sample (or small batch of samples) is drawn from each distribution—typically too little to learn the distribution—and the goal is to estimate some property of the set of distributions.

[2] We later observe in Fig. 1 that its guarantees can be polynomially suboptimal in a natural setting.

This includes the property testing work of Levi et al. [19], and work on estimating properties of populations of parameters, such as estimating the multiset of coin biases given a small number of tosses of each coin (e.g. [24, 25]).

1.2 Our Contributions

In our work, we design new algorithms for heteroskedastic mean estimation with polynomially-better error guarantees than prior work, explicitly answering the open problem of [20] (see Fig. 1):

Given samples of n independent Gaussians with a common mean, and with an unknown subset of m samples having variance bounded by 1, what is the best possible estimation error?

Theorem 1.1 (Optimal Subset-of-Signals). *Consider observing n Gaussian samples with a common mean $X_i \sim N(\mu, \sigma_i^2)$, where $\sigma_1 \leq \cdots \leq \sigma_m \leq 1$, the variances are unknown to the algorithm, and samples are presented in an arbitrary order. For any constant δ, there exists a constant C such that with probability at least $1 - \frac{1}{n^\delta}$, Algorithm 2 attains:*

- $\tilde{O}\left(\frac{n}{m^4}\right)^{1/2}$ *error if* $C\log(n) \leq m \leq n^{1/4}$
- $\tilde{O}\left(\frac{n}{m^4}\right)^{1/6}$ *error if* $n^{1/4} \leq m \leq n$

As our algorithm is scale-invariant and translation-invariant, this also enables the closed-form:

Corollary 1.2. *Consider observing n Gaussian samples with a common mean $X_i \sim N(\mu, \sigma_i^2)$, where $\sigma_1 \leq \cdots \leq \sigma_n$, the variances are unknown to the algorithm, and samples are presented in an arbitrary order. For any constant δ, there exists a constant C such that with probability at least $1 - \frac{1}{n^\delta}$, Algorithm 2 attains error*

$$\tilde{O}\left(\min\left(\min_{C\log(n)\leq i\leq n^{1/4}} \sigma_i \cdot \left(\frac{n}{i^4}\right)^{1/2}, \min_{n^{1/4}\leq i\leq n} \sigma_i \cdot \left(\frac{n}{i^4}\right)^{1/6}\right)\right)$$

Our techniques also naturally extend to the $d > 1$ dimensional setting, resolving the implicit open problem of [4]: *How large does the dimension d need to be to nearly attain the error rate that would be achievable if the variances were known?* We show that even when $d = 2$, this known-variance rate can nearly be attained, improving upon the prior guarantee of [4] that required $d = \Omega(\log(n))$:

Theorem 1.3. *Consider observing n 2-dimensional Gaussian samples with a common mean $X_i \sim N(\mu, \sigma_i^2 I)$, where $\sigma_1 \leq \cdots \leq \sigma_n$, the variances are unknown to the algorithm, and samples are presented in an arbitrary order. There exits an algorithm that attains error $\tilde{O}\left(\sqrt{\frac{1}{\sum_{i=2}^n \frac{1}{\sigma_i^2}}}\right)$ with probability $1 - o(1)$.*

1.3 Preliminaries

Let $\rho(l, r)$ denote the random variable corresponding to the number of samples with value $\in [l, r]$. $f_{\mathcal{D}}(\cdot)$ is the density function of distribution $\mathcal{D}$. For $d = 1$, in instances where we must refer to the samples in order of realized value, we refer to them by $Y_1 \leq \cdots \leq Y_n$. Meaning, we realize $X_1, \ldots, X_n$ with $X_i \sim N(\mu, \sigma_i^2)$, and observe $Y_1 \leq \cdots \leq Y_n$ where the Y_i's are the X_i's sorted in non-decreasing order. We use $\tilde{O}(\cdot)$ to suppress logarithmic factors in n.

2 OVERVIEW OF OUR TECHNIQUES

In this section, we provide the high-level intuition for our approach and results, and describe the key lemmas that facilitate our analysis. Finally, we discuss how our approach and analysis can be furthered to attain results for multi-dimensional heteroskedastic mean estimation.

2.1 Intuition and Existing Estimators

As discussed earlier, mean estimation and even heteroskedastic mean estimation has been studied by a variety of prior works that leverage different algorithmic ideas. Here, we provide a brief overview to give intuition into the challenges of the problem, and motivate our main algorithmic ideas.

The two most basic estimators are the *empirical mean* and the *empirical median*. Neither of these, however, adequately leverage the heterogeneity in the quality of samples in settings where some variances are significantly larger than others. In the case of returning the empirical mean, $\frac{X_1+\ldots X_n}{n}$, even if all but one sample has variance 1 and a single sample has arbitrarily large variance, the empirical mean also will have large variance. While the *median* of the X_i's has some robustness to such settings, it also fails to leverage heterogeneity—this is especially easy to see in the fact that the median is blind to settings where $\ll \sqrt{n}$ samples have significantly smaller variance than the rest. For example, suppose $X_1, \ldots, X_{n^{1/2-\varepsilon}} \sim N(\mu, 1)$ and $X_{n^{1/2-\varepsilon}}, \ldots, X_n \sim N(\mu, \infty)$. The median will incur unbounded expected error, while alternative algorithms, such as one that looks for the tightest cluster of $n^{1/2-\varepsilon}$ points and then takes the average of the cluster, would incur expected error of $\Theta(\frac{1}{\sqrt{n^{1/2-\varepsilon}}})$.

These settings where there are a small number of very good samples motivate creating estimators that search for tightly-clustered sets of samples, and return a statistic of the samples in the cluster. This intuitively reflects that if there are few low-variance samples, we would prefer our estimate to rely almost purely on those good samples *if we could identify them.* The k**-closest estimator**, and the **"modal"** estimator are two estimators that leverage this intuition. The k-closest estimator looks at the k-closest points and returns their midpoint. The "modal" estimator returns the value $\hat{\mu}$ containing the most samples within $[\hat{\mu} - w, \hat{\mu} + w]$. The parameters k and w are chosen so as to isolate an appropriate scale that focuses on the high-quality samples. As one might expect, these estimators are quite similar, and there is nearly a bijection between the k-closest estimator and the modal estimator with parameter $w = \arg\min_w(\max_{\hat{\mu}} \rho(\hat{\mu} - w, \hat{\mu} + w) \geq k)$. These estimators have been at the core of the previously-best guarantees for heteroskedastic mean estimation. Despite this, their shortcomings are illustrated even in the homoskedastic case where all samples have equal variance: when all samples $X_1, \ldots, X_n \sim N(\mu, 1)$ there is no choice of k or w for which the modal or k-closest estimators yield expected error better than $\Theta(n^{-1/3})$ [3, 16], despite expected error $O(1/\sqrt{n})$ being achievable by the mean or median.[3] Prior works have obtained guarantees demonstrated in Fig. 1 by leveraging

[3] For variants of the k-closest estimator that return the mean or median of the k-closest points, rather than their midpoint, this can behave similarly to the mean or median for sufficiently large k, although they are still suboptimal in the heteroskedastic case.

hybrid estimators (e.g. [4] uses the k-closest estimator consistent with the confidence interval quantiles of the empirical median).

2.2 A "Balanced" Modal Estimator

At its core, our estimator behaves similarly to a modal estimator, that returns the estimate $\hat{\mu}$ which maximizes the number of samples in the range $[\hat{\mu}-w, \hat{\mu}+w]$, with the additional condition that this range be "balanced" in the sense that the number of samples in the interval $[\hat{\mu}-w, \hat{\mu}]$ is approximately the same as the number of samples in the interval $[\hat{\mu}, \hat{\mu}+w]$.

Before discussing how w is chosen, we describe the intuition for this balanced condition. Returning to the homoskedastic case where all variances are 1, suppose we are trying to decide whether to return the true mean, μ, versus a slightly offset version of it, $\mu+\Delta$. The standard modal estimator with parameter $w=1$ is trying to decide whether there is more probability mass in the interval $[\mu-1, \mu+1]$ versus the interval $[\mu-1+\Delta, \mu+1+\Delta]$. This depends on the difference between the mass in the intervals $[\mu-1, \mu-1+\Delta]$ and $[\mu+1, \mu+1+\Delta]$. The difference in expectation is roughly the derivative of the probability density function of the standard Gaussian, evaluated at 1 times the *square* of Δ, namely $O(\Delta^2 n)$, while the standard deviation of the difference is roughly $O(\sqrt{\Delta n})$. The signal of the true mean overpowers the variance when $\Delta \gg n^{-1/3}$, matching classical guarantees for the modal estimator. In contrast, when evaluating the balance condition at $\mu+\Delta$, the relevant quantity is the difference between the densities in the intervals $[\mu+\Delta-1, \mu+\Delta]$ and $[\mu+\Delta, \mu+\Delta+1]$. The difference in expectation is roughly the difference in the standard Gaussian density in the interval $[\mu, \mu+\Delta]$ and the interval $[\mu+1, \mu+1+\Delta]$. In particular, this quantity is *linear* in the offset Δ, as opposed to quadratic. We obtain a difference in expectation that is $O(\Delta n)$, while the standard deviation is $O(\sqrt{n})$. Hence, we can detect imbalance when $\Delta \gg n^{-1/2}$, yielding more accurate estimates that match the best guarantees for homoskedastic estimation.

This "balanced" modal estimator attains nearly-optimal error for homoskedastic mean estimation in a way that seems amenable to zooming into scales that would leverage heteroskedasticity, unlike the median or mean. We will see that (perhaps surprisingly), this balanced modal estimator can also provide a near-optimal estimator from *heteroskedastic* observations if the perfect width w to use was known. To address this caveat that we do not know which width, w, to use, we propose a similarly-intuited approach we call the *balance-finding algorithm*. Oversimplifying, this algorithm will enable us to accomplish something similar to looking for the information of the balanced modal estimator at multiple scales of w simultaneously.

Balance finding. Our primary algorithmic technique is to search for the phenomenon of a particular kind of *balance* that implies a high-probability confidence interval for the mean. We will look for such balance at many scales (similar to trying many values of w) and intersect our obtained confidence intervals to determine our final estimate. To illustrate this phenomenon, consider counting the number of samples that are slightly less than μ, and the number of samples slightly larger than μ. If we use "slightly" to mean within an interval of size w, we are considering $\rho(\mu-w, \mu)$ and $\rho(\mu, \mu+w)$ respectively (recall that $\rho(l, r)$ denotes the number of samples within $[l, r]$). Naturally, as our density is symmetric, we expect $\rho(\mu-w, \mu) \approx \rho(\mu, \mu+w)$, meaning these terms are $\tilde{\Theta}(\sqrt{\rho(\mu-w, \mu+w)})$ apart. For appropriately chosen w and any estimate $\hat{\mu}$, an observation that $\rho(\hat{\mu}-w, \hat{\mu}) \approx \rho(\hat{\mu}, \hat{\mu}+w)$ can be roughly interpreted as evidence that either $|\mu-\hat{\mu}|$ is small, or that $[\hat{\mu}-w, \hat{\mu}+w]$ corresponds to a relatively flat region of the density curve.

This illuminates the desire to distinguish between estimates near μ and estimates far from μ but in flat regions of the density curve. Intuitively, in the case that our estimate is merely in a flat region, we expect to still see this balance if we perturb our estimate. More concretely, suppose we perturb our flat-region $\hat{\mu}$ by a term Δ, we still expect to see $\rho((\hat{\mu}+\Delta)-w, \hat{\mu}+\Delta) \approx \rho(\hat{\mu}+\Delta, (\hat{\mu}+\Delta)+w)$. On the other hand, we do not expect to see this balance when $\hat{\mu}$ is near μ. If we move our estimate Δ to the left then we expect to see many more samples to its right, or $\rho((\hat{\mu}-\Delta)-w, \hat{\mu}-\Delta) \ll \rho(\hat{\mu}-\Delta, (\hat{\mu}-\Delta)+w)$. Similarly, if we move the estimate Δ to the right we expect $\rho((\hat{\mu}+\Delta)-w, \hat{\mu}+\Delta) \gg \rho(\hat{\mu}+\Delta, (\hat{\mu}+\Delta)+w)$. This motivates searching for a meaningful type of balance, where the balance is not observed for the perturbed estimates, and thus resembling the case where $|\mu-\hat{\mu}|$ is small. Observing imbalance in the correct directions with both perturbations is actually sufficient, so our algorithm does not need to test that there is balance centered at $\hat{\mu}$.

Finding balance can be defined with respect to the estimate $\hat{\mu}$, the perturbation Δ, the width w, and a confidence parameter that determines thresholds for $\ll, \gg$ as used above. In this section, assume the confidence parameter is defined such that the probability of ever finding a false-positive meaningful balance is inverse-polynomially small. We will then more precisely describe a balance as a $(w, \Delta, \hat{\mu})$-balance. We claim that, with high probability, there will be no $(\cdot, \Delta, \hat{\mu})$-balance where $|\mu-\hat{\mu}| > \Delta$: yielding a confidence interval of $\mu \in [\hat{\mu}-\Delta, \hat{\mu}+\Delta]$. Accordingly, our strategy is to test many carefully-chosen tuples of $(w, \Delta, \hat{\mu})$-balance and intersect the confidence intervals we obtain. In Algorithm 1, we outline our subroutine for testing a $(w, \Delta, \hat{\mu})$-balance.

What remains is to design an algorithm that tests the correct balances that yield sufficiently small and correct confidence intervals. Algorithmically, we remark that for a given w and Δ, we can use a sweep-line method to find all ranges of $\hat{\mu}$ where there exists $(w, \Delta, \hat{\mu})$-balance in $\tilde{O}(n)$ time. Thus, we may obtain an $\tilde{O}(n)$ time algorithm if we can select $\tilde{O}(1)$ pairs of (w, Δ) to consider, and can show that testing just balances with these parameters will obtain our desired estimation error. While we do not fully motivate it until later, we provide our approach in Algorithm 2.

2.3 Analyzing Estimation Error

Near-optimal guarantees for simplified Subset-of-Signals. We will now informally show that finding balance is sufficient for obtaining near-optimal guarantees in a simplified version of the Subset-of-Signals model where at least m samples have $\sigma_i \leq 1$, *and the remaining samples all have the same value of $\sigma_i = \sigma^*$ (this additional assumption is only to permit a cleaner explanation here).* More sophisticated techniques will later enable us to show the same guarantees for (unsimplified) Subset-of-Signals, and results for more general settings.

Algorithm 1 Testing $(w, \Delta, \hat{\mu})$-balance

Input: width w, shift Δ, and potential mean $\hat{\mu}$
Output: PASS (it likely holds that the true mean $\mu \in [\hat{\mu} - \Delta, \hat{\mu} + \Delta]$), or FAIL (insufficient evidence or evidence against $\mu \in [\hat{\mu} - \Delta, \hat{\mu} + \Delta]$)
Description: This test will PASS if the number of samples in the intervals $[\hat{\mu} - w, \hat{\mu}]$ and $[\hat{\mu}, \hat{\mu} + w]$ are approximately equal, yet after shifting these intervals by ±Δ the halves become significantly unbalanced (evidencing a higher density of samples near $\hat{\mu}$ versus $\hat{\mu} \pm w$).

```
procedure Test($w, \Delta, \hat{\mu}$):
    $L_{\text{shift-right}} \leftarrow \rho(\hat{\mu} + \Delta - w, \hat{\mu} + \Delta)$   ▷ Count samples within $[\hat{\mu} + \Delta - w, \hat{\mu} + \Delta]$.
    $R_{\text{shift-right}} \leftarrow \rho(\hat{\mu} + \Delta, \hat{\mu} + \Delta + w)$   ▷ Count samples within $[\hat{\mu} + \Delta, \hat{\mu} + \Delta + w]$.
    $T_{\text{shift-right}} \leftarrow \rho(\hat{\mu} + \Delta - w, \hat{\mu} + \Delta + w)$   ▷ Count samples within $[\hat{\mu} + \Delta - w, \hat{\mu} + \Delta + w]$.
    if $L_{\text{shift-right}} - R_{\text{shift-right}} \leq \sqrt{C_{\delta_{\text{false-pos}}} \log(n) T_{\text{shift-right}}}$ or $T_{\text{shift-right}} < C_{\delta_{\text{false-pos}}} \log(n)$ then
        return FAIL
    end if
    $L_{\text{shift-left}} \leftarrow \rho(\hat{\mu} - \Delta - w, \hat{\mu} - \Delta)$   ▷ Count samples within $[\hat{\mu} - \Delta - w, \hat{\mu} - \Delta]$.
    $R_{\text{shift-left}} \leftarrow \rho(\hat{\mu} - \Delta, \hat{\mu} - \Delta + w)$   ▷ Count samples within $[\hat{\mu} - \Delta, \hat{\mu} - \Delta + w]$.
    $T_{\text{shift-left}} \leftarrow \rho(\hat{\mu} - \Delta - w, \hat{\mu} - \Delta + w)$   ▷ Count samples within $[\hat{\mu} - \Delta - w, \hat{\mu} - \Delta + w]$.
    if $R_{\text{shift-left}} - L_{\text{shift-left}} \leq \sqrt{C_{\delta_{\text{false-pos}}} \log(n) T_{\text{shift-left}}}$ or $T_{\text{shift-left}} < C_{\delta_{\text{false-pos}}} \log(n)$ then
        return FAIL
    end if
    return PASS
end procedure
```

The existence of (w, Δ, μ)-balance will typically imply that our algorithm obtains $O(\Delta)$ error with high probability. This will follow from showing that: (i) with high probability there is no $(\cdot, \Delta', \hat{\mu})$-balance where $|\mu - \hat{\mu}| > \Delta'$, and (ii) our algorithm will test sufficiently similar tuples that find a $(\cdot, \Delta', \cdot)$-balance with $\Delta' = O(\Delta)$. Accordingly, if there exists a (w, Δ, μ)-balance, then we expect our algorithm to find a balance yielding a correct confidence interval of width $O(\Delta)$ containing μ. This motivates our focus on studying the conditions under which (w, Δ, μ)-balance exists:

Informal Claim 2.1. *(w, Δ, μ)-balance will exist with high probability if* $\mathbb{E}[\rho(\mu, \mu+\Delta) - \rho(\mu+w, \mu+w+\Delta)]^2 \geq C_1 \log(n) \cdot \mathbb{E}[\rho(\mu, \mu+w)]$.

This follows from how the imbalance after shifting will be much larger than the standard deviation of the difference between correctly balanced halves centered at μ. We will use the simple condition of Claim 2.1 to obtain desired estimation error. As seen in Fig. 1, the optimal rate for Subset-of-Signals undergoes a phase transition at $m = n^{1/4}$. We obtain this rate up to logarithmic factors:

Algorithm 2 Estimation-Algorithm

Input: $Y_1 \leq \cdots \leq Y_n$
Output: Range C_{conf} (can choose any arbitrary value in this range as the estimate $\hat{\mu}$)

```
procedure Sweep-Test($w, \Delta$):
    return $S_{w,\Delta}$   ▷ Returns set $S_{w,\Delta}$ of $O(n)$ intervals of $\hat{\mu}$ that PASS Test($w, \Delta, \hat{\mu}$)
end procedure
procedure Generate-Tests($Y_1 \leq \cdots \leq Y_n$):
    $S_{\text{params}} \leftarrow \{\infty\}$
    for $i \in [\lfloor \log(n) \rfloor]$ do
        $r_{2^i} \leftarrow \min_j Y_{j+2^i} - Y_j$   ▷ $r_{2^i}$ is the gap between the closest $2^i$ samples
        for $j \in \{-\lceil C_{\delta_{\text{param}}} \log(n) \rceil, \ldots, \lceil C_{\delta_{\text{param}}} \log(n) \rceil$ do
            $S_{\text{params}} \leftarrow S_{\text{params}} \cup r_{2^i} \cdot 2^j$   ▷ Approximating $\sigma_{2^i}$ by powers of 2 near $r_{2^i}$.
        end for
    end for
    return $S_{\text{params}}$   ▷ Returns $S_{\text{params}}$, including $\infty$ and approximations of $\sigma_{2^i}$
end procedure
procedure Estimation-Algorithm($Y_1 \leq \cdots \leq Y_n$):
    $C_{\text{conf}} \leftarrow [-\infty, \infty]$   ▷ Interval we are confident $\mu$ is within
    $S_{\text{params}} \leftarrow$ Generate-Tests(Y)   ▷ Determine values of $w, \Delta$
    for $w, \Delta \in S_{\text{params}}$ do
        $S_{w,\Delta} \leftarrow$ Sweep-Test(w, Δ)   ▷ Values of $\hat{\mu}$ that Pass Test($w, \Delta, \hat{\mu}$).
        if $S_{w,\Delta} \neq \emptyset$ then
            $C_{\text{conf}} \leftarrow C_{\text{conf}} \cap \min_{\hat{\mu} \in S_{w,\Delta}} [\hat{\mu} - \Delta, \hat{\mu} + \Delta]$
            $C_{\text{conf}} \leftarrow C_{\text{conf}} \cap \max_{\hat{\mu} \in S_{w,\Delta}} [\hat{\mu} - \Delta, \hat{\mu} + \Delta]$   ▷ Intersect confidence intervals.
        end if
    end for
    return $C_{\text{conf}}$   ▷ Can estimate $\hat{\mu}$ as any arbitrary value in $C_{\text{conf}}$.
end procedure
```

Lemma 2.2. *When $m \in [n^{1/4}, n]$, with high probability there exists a (w, Δ, μ)-balance with $\Delta = \tilde{O}\left(\frac{n}{m^4}\right)^{1/6}$.*

Proof. We will consider evaluating two types of balance, and conclude that at least one of these balances must exist with the desired Δ.

By Claim 2.1, we can find $(1, \Delta, \mu)$-balance if $(m \cdot \Delta)^2 \geq O(1) \cdot C_1 \log(n) \cdot \mathbb{E}[\rho(\mu-1, \mu+1)]$. Meaning, if we do not find such balance, $\Delta \leq O(1) \cdot \sqrt{\frac{C_1 \log(n) \mathbb{E}[\rho(\mu-1,\mu+1)]}{m^2}}$.

Intuitively, if this is an undesirable bound on Δ, then $\mathbb{E}[\rho(\mu - 1, \mu + 1)]$ must be large, meaning many of the $n - m$ samples of standard deviation σ^* must be realized in $[-1, +1]$, and thus σ^* must not be too large. In other words, either we are able to find balance from our m "good" points, or our remaining $n - m$ "bad" points must not actually be too bad. For our other type of balance, we will notice how (∞, Δ, μ)-balance behaves similarly to classical high-probability guarantees for the median. We will find such a balance if $\mathbb{E}[\rho(\mu-\Delta,\mu+\Delta)] \geq O(1)\cdot\sqrt{C_1\log(n)n}$.

Combining both restrictions, if we cannot find either balance then $\Delta \leq O(1)\cdot\sqrt{\frac{C_1\log(n)\mathbb{E}[\rho(\mu-1,\mu+1)]}{m^2}} \leq O(1)\cdot \sqrt{\frac{C_1\log(n)\mathbb{E}[\rho(\mu-\Delta,\mu+\Delta)]}{\Delta m^2}} \leq O(1)\cdot\sqrt{\frac{C_1\log(n)\sqrt{C_1\log(n)n}}{\Delta m^2}}$. This implies $\Delta \leq O(1)\cdot(C_1^{3/2}\log^{3/2}(n))^{1/3}\cdot(\frac{n}{m^4})^{1/6} = O(\sqrt{\log(n)}\cdot(\frac{n}{m^4})^{1/6}) = \tilde{O}((\frac{n}{m^4})^{1/6}))$.

□

Lemma 2.3. *When $m \in [C'\log(n), n^{1/4}]$, with high probability there exists a (w,Δ,μ)-balance with $\Delta = \tilde{O}\left(\frac{n}{m^4}\right)^{1/2}$.*

Proof. We will again consider evaluating a pair of balances, and conclude that at least one of these values must exhibit balance with the desired Δ.

By Claim 2.1, we can find $(1,\frac{1}{2},\mu)$-balance if $m^2 \geq O(1)\cdot C_1\log(n)\cdot\mathbb{E}[\rho(\mu-1,\mu+1)]$. Since our guarantees for Δ in this lemma are super-constant, finding this balance would be sufficient. If we do not find such balance, then $m^2 \leq O(1)\cdot C_1\log(n)\cdot\mathbb{E}[\rho(\mu-1,\mu+1)] \leq O(1)\cdot C\log(n)\frac{n}{\sigma^*} \implies \sigma^* \leq O(1)\cdot\frac{Cn\log(n)}{m^2}$.

Similar to Lemma 2.2, our inability to find balance from the m samples implies σ^* cannot be too large. We will then find the median-like balance of (∞,Δ,μ)-balance if $\mathbb{E}[\rho(\mu-\Delta,\mu+\Delta)] \geq \sqrt{C_1\log(n)n}$. Finally, this implies we find (∞,Δ,μ)-balance for a $\Delta \leq O(1)\cdot\frac{\sqrt{C_1\log(n)n}}{\mathbb{E}[\rho(\mu-1,\mu+1)]} \leq O(1)\cdot\frac{\sigma^*\sqrt{\log(n)}}{\sqrt{n}} \leq O(1)\cdot\frac{\sqrt{n}\log^{1.5}(n)}{m^2} = \tilde{O}((\frac{n}{m^4})^{1/2})$.

□

Accordingly, one may obtain desired rates for simplified Subset-of-Signals by just testing the collection of tuples we discussed in the proofs of Lemmas 2.2 and 2.3.

Additional considerations. We will need additional non-trivial considerations for proving our unsimplified results. Some include:

(Unsimplified) Subset-of-Signals. If the $n-m$ remaining samples are allowed to have any value of σ_i, then checking just the tuples of balances in Lemmas 2.2 and 2.3 will not be sufficient to find the desired balance. This is roughly because there may be groups of σ_i that interfere with balance at the scale of 1, while still not helping produce a good median. With some nuance, we later show (i) there still must exist some scale at which to find desired balance, and (ii) we can choose a set of $\tilde{O}(1)$ tuples which will test something sufficiently close to discover said desired balance.

Choosing testing tuples. The previous point touches on how we require some way of testing the correct collection of balances. Moreover, it would be desirable if our estimator was scale-invariant so that if m samples have $\sigma_i \leq v$, then we could attain the analogous Subset-of-Signals guarantee scaled by v. One may expect that if we are looking for balance driven by k good samples, the correct Δ and w to test may be within a polynomial factor of the distance between the k-closest points (r_k). Later, we will show it is sufficient to consider pairs of w and Δ that are powers of 2 and polynomially-close to a r_{2^i} for $i \in [1,\log(n)]$, giving $\tilde{O}(1)$ tuples to test in a scale-invariant manner.

2.4 Multi-Dimensional Estimation

In this section, we focus on estimation with d-dimensional observations. Each $X_i \sim N(\mu,\Sigma_i)$, where μ is a d-dimensional vector and Σ_i is a $d\times d$ covariance matrix. If each Σ_i can be an arbitrary diagonal covariance matrix, then observations in different dimensions are unrelated and thus there is nothing possible beyond considering d independent instances of 1-dimensional estimation. However, if $\Sigma_i = \sigma_i^2 I$, then each sample has the same variance in every dimension, and high dimensional observations are extremely helpful. [4] initiated the study of this problem and obtained (in Theorem 5.2) an algorithm that with probability $1-\Theta(1/n)$, it holds that $\mathbb{E}[|\hat{\mu}_i-\mu_i|] = O\left(\sqrt{\frac{1}{\sum_{j=2}^n\frac{1}{\sigma_j^2}}}\right)$ when $d=\Omega(\log(n))$. Note how this quantity is exactly the error for estimation with known-variances, other than the removal of the term depending on σ_1. The crux of their approach leverages that with $d=\Omega(\log(n))$ dimensions, one can approximate $\sigma_i^2+\sigma_j^2$ well for every pair of $i\neq j$.

Interestingly, we will obtain similar guarantees while only requiring $d\geq 2$. We provide a high-level overview focusing on the most interesting case of $d=2$. Let us denote the known-variance error ignoring σ_1 as $R(\sigma) \triangleq \sqrt{\frac{1}{\sum_{i=2}^n\frac{1}{\sigma_i^2}}}$. We note its relation to a simpler closed-form:

Lemma 2.4. $\min_{2\leq i\leq n}\frac{\sigma_i}{\sqrt{i}} \leq \tilde{O}(R(\sigma))$.

Establishing this simpler closed-form as our goal, we sketch an approach based on balance-testing that may hope to obtain error near $\frac{\sigma_i}{\sqrt{i}}$:

- Consider a guess for the mean $\hat{\mu}=\hat{\mu}_1,\hat{\mu}_2$.
- Filter all X_j whose observation in the first dimension is farther than σ_i from $\hat{\mu}_1$.
- With the filtered points in the second dimension, perform balance testing around $\hat{\mu}_2$.

Informally, consider how often a sample X_j with large σ_j would "interfere" with a balance test at the scale of σ_i in the 1-dimensional setting: it would land in $[\mu-\sigma_i,\mu+\sigma_i]$ with probability $\Theta(\frac{\sigma_i}{\sigma_j})$. However, in the 2-dimensional setting, this probability is much smaller given our filtering, and is accordingly $\Theta\left(\left(\frac{\sigma_i}{\sigma_j}\right)^2\right)$. This difference will be enough to obtain known-variance rates. Algorithmically, we will try all $O(n^2)$ possible filterings, each creating an instance of 1-dimensional estimation, and we will intersect all the confidence intervals yielded from each instance to obtain an estimate.

For some intuition regarding why we obtain known-variance rates, consider the case where $i^* = \arg\min_{C'\log^2(n)\leq i\leq n}\frac{\sigma_i}{\sqrt{i}}$. We claim that (after some calculation) the conditions of Claim 2.1 under which we expect to find balance are satisfied when $\Delta \geq \frac{\log(n)\sigma_{i^*}}{\sqrt{i^*}}$. Accordingly, there exists a C' such that if $i^*\geq C'\log^2(n)$ then we obtain error $\tilde{O}(\frac{\sigma_{i^*}}{\sqrt{i^*}})$ with high probability. Handling other cases where $i^* < C'\log^2(n)$ involve other considerations that ultimately yield:

Theorem 1.3. *Consider observing n 2-dimensional Gaussian samples with a common mean $X_i \sim N(\mu,\sigma_i^2 I)$, where $\sigma_1 \leq \cdots \leq \sigma_n$,* *the variances are unknown to the algorithm, and samples are presented in an arbitrary order. There exits an algorithm that attains error* $\tilde{O}\left(\sqrt{\frac{1}{\sum_{i=2}^{n}\frac{1}{\sigma_i^2}}}\right)$ *with probability* $1 - o(1)$.

ACKNOWLEDGEMENTS

This work was supported by the National Defense Science & Engineering Graduate (NDSEG) Fellowship Program, Tselil Schramm's NSF CAREER Grant no. 2143246, and Gregory Valiant's Simons Foundation Investigator Award.

REFERENCES

[1] Mikhail Belkin and Kaushik Sinha. 2010. Polynomial learning of distribution families. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, 103–112.

[2] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. 2017. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. 47–60.

[3] Herman Chernoff. 1964. Estimation of the mode. *Annals of the Institute of Statistical Mathematics* 16, 1 (1964), 31–41.

[4] Flavio Chierichetti, Anirban Dasgupta, Ravi Kumar, and Silvio Lattanzi. 2014. Learning entangled single-sample Gaussians. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 511–522.

[5] Aniket Das, Dheeraj M Nagaraj, Praneeth Netrapalli, and Dheeraj Baby. 2023. Near optimal heteroscedastic regression with symbiotic learning. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 3696–3757.

[6] Luc Devroye, Silvio Lattanzi, Gabor Lugosi, and Nikita Zhivotovskiy. 2020. On Mean Estimation for Heteroscedastic Random Variables. *arXiv preprint arXiv:2010.11537* (2020).

[7] Luc Devroye, Silvio Lattanzi, Gábor Lugosi, and Nikita Zhivotovskiy. 2023. On mean estimation for heteroscedastic random variables. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, Vol. 59. Institut Henri Poincaré, 1–20.

[8] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2019. Robust estimators in high-dimensions without the computational intractability. *SIAM J. Comput.* 48, 2 (2019), 742–864.

[9] Ilias Diakonikolas, Daniel M Kane, Thanasis Pittas, and Nikos Zarifis. 2023. SQ Lower Bounds for Learning Mixtures of Separated and Bounded Covariance Gaussians. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 2319–2349.

[10] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. 2019. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2745–2754.

[11] Shivam Gupta, Jasper Lee, Eric Price, and Paul Valiant. 2022. Finite-sample maximum likelihood estimation of location. *Advances in Neural Information Processing Systems* 35 (2022), 30139–30149.

[12] Shivam Gupta, Jasper CH Lee, and Eric Price. 2023. Finite-sample symmetric mean estimation with fisher information rate. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 4777–4830.

[13] Shivam Gupta, Jasper CH Lee, and Eric Price. 2023. High-dimensional location estimation via norm concentration for subgamma vectors. In *International Conference on Machine Learning*. PMLR, 12132–12164.

[14] Samuel B Hopkins and Jerry Li. 2018. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. 1021–1034.

[15] IA Ibragimov, RZ Has' minskii, et al. 1981. Statistical Estimation: Asymptotic Theory. *Springer Book Archive-Mathematics* (1981).

[16] Jeankyung Kim and David Pollard. 1990. Cube root asymptotics. *The Annals of Statistics* (1990), 191–219.

[17] Weihao Kong, Raghav Somani, Zhao Song, Sham Kakade, and Sewoong Oh. 2020. Meta-learning for Mixed Linear Regression. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 5394–5404.

[18] Kevin A Lai, Anup B Rao, and Santosh Vempala. 2016. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 665–674.

[19] Reut Levi, Dana Ron, and Ronitt Rubinfeld. 2013. Testing properties of collections of distributions. *Theory of Computing* 9, 1 (2013), 295–347.

[20] Yingyu Liang and Hui Yuan. 2020. Learning entangled single-sample Gaussians in the subset-of-signals model. In *Conference on Learning Theory*. PMLR, 2712–2737.

[21] Ankur Moitra and Gregory Valiant. 2010. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, 93–102.

[22] Ankit Pensia, Varun Jog, and Po-Ling Loh. 2019. Estimating location parameters in entangled single-sample distributions. *arXiv preprint arXiv:1907.03087* (2019).

[23] Ankit Pensia, Varun Jog, and Po-Ling Loh. 2022. Estimating location parameters in sample-heterogeneous distributions. *Information and Inference: A Journal of the IMA* 11, 3 (2022), 959–1036.

[24] Kevin Tian, Weihao Kong, and Gregory Valiant. 2017. Learning populations of parameters. *Advances in neural information processing systems* 30 (2017).

[25] Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, and Sham Kakade. 2019. Maximum likelihood estimation for learning populations of parameters. In *International Conference on Machine Learning*. PMLR, 6448–6457.

[26] Hui Yuan and Yingyu Liang. 2020. Learning entangled single-sample distributions via iterative trimming. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2666–2676.

Received 13-NOV-2023; accepted 2024-02-11