Entangled Mean Estimation in High Dimensions

Ilias Diakonikolas
University of Wisconsin-Madison
Madison, USA
ilias.diakonikolas@gmail.com

Daniel M. Kane
University of California at San Diego
San Diego, USA
dakane@ucsd.edu

Sihan Liu
University of California at San Diego
San Diego, USA
sil046@ucsd.edu

Thanasis Pittas
University of Wisconsin-Madison
Madison, USA
pittas@wisc.edu

Abstract

We study the task of high-dimensional entangled mean estimation in the subset-of-signals model. Specifically, given N independent random points $x_1, \dots, x_N$ in $\mathbb{R}^D$ and a parameter $\alpha \in (0, 1)$ such that each x_i is drawn from a Gaussian with mean μ and unknown covariance, and an unknown α-fraction of the points have *identity-bounded* covariances, the goal is to estimate the common mean μ. The one-dimensional version of this task has received significant attention in theoretical computer science and statistics over the past decades. Recent work has given near-optimal upper and lower bounds for the one-dimensional setting. On the other hand, our understanding of even the information-theoretic aspects of the multivariate setting has remained limited.

In this work, we design a computationally efficient algorithm achieving an information-theoretically near-optimal error. Specifically, we show that the optimal error (up to polylogarithmic factors) is $f(\alpha, N) + \sqrt{D/(\alpha N)}$, where the term $f(\alpha, N)$ is the error of the one-dimensional problem and the second term is the sub-Gaussian error rate. Our algorithmic approach employs an iterative refinement strategy, whereby we progressively learn more accurate approximations $\widehat{\mu}$ to μ. This is achieved via a novel rejection sampling procedure that removes points significantly deviating from $\widehat{\mu}$, as an attempt to filter out unusually noisy samples. A complication that arises is that rejection sampling introduces bias in the distribution of the remaining points. To address this issue, we perform a careful analysis of the bias, develop an iterative dimension-reduction strategy, and employ a novel subroutine inspired by list-decodable learning that leverages the one-dimensional result.

CCS Concepts

• **Theory of computation → Unsupervised learning and clustering**.

Keywords

Machine Learning Theory, High Dimensional Statistics

STOC '25, Prague, Czechia

ACM ISBN 979-8-4007-1510-5/25/06
https://doi.org/10.1145/3717823.3718162

ACM Reference Format:
Ilias Diakonikolas, Daniel M. Kane, Sihan Liu, and Thanasis Pittas. 2025. Entangled Mean Estimation in High Dimensions. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing (STOC '25), June 23–27, 2025, Prague, Czechia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3717823.3718162

1 Introduction

Classical statistics has traditionally focused on the idealized scenario where the input dataset consists of independent and identically distributed samples drawn from a fixed but unknown distribution. In a wide range of modern data analysis applications, there is an increasing need to move beyond this assumption since datasets are often collected from heterogeneous sources [20, 22, 23, 25, 56, 62]. A natural formalization of heterogeneity in the context of mean estimation (the focus of this work) involves having each datapoint drawn independently from a potentially different distribution within a (known) family that shares a *common* mean parameter. Distributions with this property are referred to as *entangled*, and the setting is also known as sample *heterogeneity* or *heteroskedasticity*.

The task of estimating the mean of entangled distributions has gained significant attention in recent years for a number of reasons. First, from a practical viewpoint, entangled distributions intuitively capture the idea of collecting samples from diverse sources. One of the early works that studied this task [8] illustrates this with the following crowdsourcing example. Suppose that multiple users rate a product with some true value μ. Each user i has their own level of knowledge about the product, captured by a standard deviation parameter σ_i. The rating from user i is assumed to be sampled from a Gaussian distribution with mean μ and covariance σ_i^2, and the goal is to estimate μ in small absolute error using these samples. Other practical examples include datasets collected from sensors under varying environmental conditions; see, e.g., [34].

From a theoretical viewpoint, statistical estimation given access to non-identically distributed, heterogeneous data is a natural and fundamental task, whose roots trace back several decades in the statistics literature. Early work [29, 52, 53, 55, 57, 59] studied the asymptotic properties of such distributions. Specifically, [5, 32] studied maximum likelihood estimators and [27, 28, 45, 47] analyzed the median estimator for non-identically distributed samples. Heterogeneity has also been studied for moments of distributions [26] and linear regression [21, 36]. Mean estimation for entangled distributions, including the Gaussian setting considered in [8], is also related to the classical task of parameter learning for mixture models—albeit in a regime that is qualitatively different than the one commonly studied. While in the canonical setting—see [1, 2, 11, 35] for classic references and [3, 4, 7, 16, 16, 18, 19, 30, 37, 39, 43, 46] for more recent work—one typically assumes a small (constant) number of components $k \ll N$ with different means, in the entangled setting each sample comes from its own component ($k = N$). Importantly, the shared mean assumption allows for meaningful results despite the high number of components.

Prior Work We now summarize prior work for mean estimation of entangled Gaussians, starting with the (now well-understood) one-dimensional case. In this setting, we have access to samples $x_i \sim \mathcal{N}(\mu, \sigma_i^2)$ with unknown σ_i values. For a concrete and simple configuration for the σ_i's, we consider the so-called *subset-of-signals model*, introduced in [42]. In this model, it is assumed that at least an α-fraction of the samples have $\sigma_i \leq 1$, while the remaining can have arbitrary variances. The goal is to estimate μ in absolute error that is as small as possible in terms of the number of samples N and the rate α. A series of works [8, 10, 12, 42, 48–50, 60, 61] has established upper and lower bounds for this task. Specifically, the recent work [10] gave an estimator with error matching (up to polylogarithmic factors) the lower bound of [42] in the subset-of-signals model (for a very wide regime of α values). Entangled Gaussian mean estimation in the subset-of-signals model is thus essentially resolved in one dimension. Additional discussion on related work is provided in Section 1.3.

Entangled Mean Estimation in High Dimensions In contrast, the multivariate version of this problem is much less understood. Some of the prior work [8, 10, 48, 50] only tangentially considered higher dimensions, focusing on the rather restricted setting that the covariance matrices are spherical, i.e., of the form $\Sigma_i = \sigma_i^2 \mathbf{I}$. For this specific special case, it turns out that the problem becomes *easier* in higher dimensions—as each coordinate provides more information about the scalar parameter σ_i. A more general formulation would be to replace the sphericity assumption on the Σ_i's by a boundedness assumption. This leads to the following high-dimensional formalization of the subset-of-signals model.

Definition 1.1. Let $\mu \in \mathbb{R}^D$ be a target vector and $\alpha \in (0,1)$ be a parameter. A set of N points in $\mathbb{R}^D$ is generated as follows: First, an adversary chooses N positive semidefinite (PSD) matrices $\Sigma_1, \ldots, \Sigma_N \in \mathbb{R}^{D \times D}$ under the constraint that $\sum_{i=1}^N \mathbb{1}(\Sigma_i \preceq \mathbf{I}) \geq \alpha N$. Then, for each $i = 1, \ldots, N$, the sample x_i is drawn independently from $\mathcal{N}(\mu, \Sigma_i)$. The final dataset $\{x_1, \ldots, x_N\}$ is the input provided to the learning algorithm. We call μ the *common mean* and α the *signal-to-noise rate* of the model.

This natural definition was suggested by Jerry Li [41] at the TTIC Workshop on New Frontiers in Robust Statistics, where the complexity of the problem was posed as an open question.

We emphasize that our understanding of entangled mean estimation in the aforementioned setting is fairly limited—even information-theoretically. The results in [8, 10, 42] already imply that any estimator for the arbitrary covariance setting must incur error that is larger, by at least a polynomial factor, than the error achievable in the spherical covariance case (see Section 1.3.2 for more details). This suggests that the bounded covariance setting is more challenging than the spherical case and requires new ideas. Specifically, prior to this work, the optimal rate for the bounded covariance case was open—even ignoring computational considerations.

A standard attempt to obtain a (potentially tight) upper bound on the error involves using the one-dimensional estimator along an exponentially large cover of the unit ball in $\mathbb{R}^D$, and combining these estimates into a vector via a linear program (ala Tukey median) [58]. Unfortunately, this approach may fail in our setting, due to the following issue. Establishing correctness of the approach requires that the failure probability of the one-dimensional estimator is exponentially small in D. However, the currently best known error guarantees [10] hold only with probability $1 - \mathrm{poly}(N)$.[1] Moreover, even if this obstacle could be circumvented, we would still end up with an exponential-time estimator. Finally, we note that a simple and natural computationally efficient approach involves applying a one-dimensional estimator for each axis of the space. Unfortunately, the error incurred by this approach is $\sqrt{D}$ times that of the one-dimensional estimator, which turns out to be significantly suboptimal.

In summary, none of the known approaches yields error better than $\mathrm{poly}(D) f(\alpha, N)$, where $f(\alpha, N)$ is the error of the optimal one-dimensional estimator, leaving even the information-theoretically tight bound wide open. This leads to the core question of our work:

What is the optimal error rate for high-dimensional entangled mean estimation, both
(i) from an information-theoretic perspective, and (ii) for computationally efficient algorithms?

In this work, we resolve both aspects of this question (up to polylogarithmic factors) for a wide range of the parameters N, D, α.

1.1 Main Result

Before we formally state our contributions, we recall the error guarantee of the 1-d estimator given in [10]. In particular, if we denote by N the number of samples and α the signal-to-noise rate (fraction of points with variances bounded from above by one), then their estimator $\widehat{\mu} \in \mathbb{R}$ satisfies $|\widehat{\mu} - \mu| \leq f(\alpha, N)$ with high probability, where $f(\alpha, N)$ is defined as

$$(\log(N/\alpha))^{O(1)} \cdot \begin{cases} \alpha^{-2} N^{-3/2}, & \Omega\left(\frac{\log N}{N}\right) \leq \alpha \leq N^{-3/4} \\ \alpha^{-2/3} N^{-1/2}, & N^{-3/4} < \alpha < 1 \\ \infty, & \text{otherwise} . \end{cases} \tag{1}$$

The above error upper bound had been previously shown [42] to be best possible up to polylogarithmic factors in the regime $\Omega(\log N/N) \leq \alpha \leq O(N^{1-\epsilon})$ for any arbitrarily small constant $\epsilon > 0$.

Roughly speaking, the error of our high-dimensional estimator is equal, up to polylogarithmic factors, to the sum of the above 1-d error and the statistical error for mean estimation of isotropic Gaussians. Specifically, our main result is the following:

Theorem 1.2 (High-Dimensional Entangled Mean Estimation). *EntangledMeanEstimation(N) in Algorithm 1 satisfies the following guarantee: The algorithm draws N samples in $\mathbb{R}^D$ from the*

[1] Though one could amplify the success probability of [10] in a black-box manner using the standard "median trick", we remark that such a strategy will lead to a factor of D loss in the error guarantee if the goal is to achieve success probability $1 - \exp(-\Theta(D))$.

subset-of-signals model of Definition 1.1 with common mean $\mu \in \mathbb{R}^D$ and signal-to-noise rate $\alpha \in (0,1)$. If $N \geq \frac{D}{\alpha}\log^C(\frac{D}{\alpha})$, where C is a sufficiently large absolute constant, the output $\widehat{\mu} \in \mathbb{R}^D$ of the algorithm satisfies the following with probability at least 0.99*:*

$$\|\widehat{\mu} - \mu\|_2 \leq \log^{O(1)}(N)\left(\sqrt{\frac{D}{\alpha N}} + f(\alpha, N)\right),$$

where $f(\cdot)$ is the function defined in Equation (1). Moreover, the algorithm runs in time $\mathrm{poly}(D, N)$*.*

We remark that the error bound achieved by our algorithm is optimal up to poly-logarithmic factors in the subset-of-signals model, provided that $N \geq \widetilde{\Omega}(D/\alpha)$. To show that the second error term, $f(\alpha, N)$, is necessary, we can simply embed the 1-d hard instance of [42] in the D-dimensional space. Specifically, we can set the mean and variance of the first coordinate according to the 1-d hard instance, and set the remaining coordinates to be deterministically 0. The second term $\sqrt{D/(\alpha N)}$ is the statistical error rate of estimating the mean of isotropic Gaussians. This term is also necessary, as can be seen by embedding the standard hard instance of D-dimensional isotropic Gaussian mean estimation into the αN many samples with bounded covariances, and setting the covariances of the rest of the samples to be sufficiently large so that they reveal almost no information.

Finally, the algorithm succeeds whenever $N > D/\alpha$ (times polylogarithmic factors). We remark that this is necessary for any estimator to achieve errors smaller than a constant, i.e., $\epsilon < 1/2$, even when the identities of the samples with bounded covariances are revealed to the algorithm. Extending the result to any $N \in \mathbb{Z}_+$ is an interesting open question that we leave for future work.

1.2 Brief Overview of Techniques

In this section, we summarize our approach for obtaining an estimator achieving the guarantees of Theorem 1.2. Towards this end, we will start by explaining how to obtain an initial rough estimate $\widetilde{\mu}$ such that $\|\widetilde{\mu} - \mu\|_2 \lesssim \sqrt{D}$. We note that the main novelty (and bulk) of our technical work will be on developing a recursive procedure that iteratively improves upon $\widetilde{\mu}$.

We now provide an efficient method to achieve the warm start. Specifically, provided that we are in the regime where $f(\alpha, N) = O(1)$, such an estimate $\widetilde{\mu}$ can be easily obtained by running the 1-d estimator from [10] along each axis. For the other regime, we design a sophisticated tournament procedure that outputs an estimate within $O(\sqrt{D})$ from the true mean (see the full version [17] for more details).

We next describe how to achieve improved estimation accuracy. Naïve approaches such as sample means are destined to fail in the subset-of-signals model, due to the fact that no assumptions are made on the covariances of $(1-\alpha)$-fraction of the samples. Specifically, these matrices can have arbitrarily large operator norms, which can cause the average of the samples to suffer from arbitrarily high statistical errors in ℓ_2 distance. Our approach to overcome this issue is to use our initial estimator $\widetilde{\mu}$ (warm start) to detect and reject samples that are too far from the true mean μ in Euclidean norm. Algorithmically, the rejection sampling procedure is to accept each sample x with probability $\exp(-\|x - \widetilde{\mu}\|_2^2/D)$. On the one hand, most of the "good" samples (the αN points with bounded covariance) survive the rejection sampling: this is because a Gaussian sample with covariance bounded above by $\mathbf{I}$ is $O(\sqrt{D})$-far from μ (and hence from $\widetilde{\mu}$) with high probability, which causes its acceptance with high probability. Regarding the remaining points (i.e., the ones with covariances Σ_i such that $\mathrm{tr}(\Sigma_i) \gg D$), the rejection sampling step ensures that the probability of acceptance is small enough so that the average of the covariance matrices of the surviving points (inliers and outliers), which we denote by $\widetilde{\Sigma}_{\mathrm{avg}}$, will have its trace bounded from above by $O(D)$. We show that this essentially allows for accurate estimation of the population mean of the surviving points; roughly speaking, this follows from the fact that the standard error for mean estimation of a bounded covariance random variable X is $O(\sqrt{\mathrm{tr}(\mathrm{Cov}(X))/N})$.

Unfortunately, the aforementioned approach does not quite work for the following reason: the rejection sampling procedure will also cause the population mean of the surviving samples to be biased. In particular, since the acceptance probability is given by the exponential of some quadratic in the input point, the resulting distribution of each point x_i conditioned on its acceptance will be some new Gaussian distribution whose means and covariances are functions of the true mean μ, the covariance Σ_i of x_i, and the center $\widetilde{\mu}$ used in the rejection sampling. After a careful calculation, one can show that the bias of the new population mean of a set of samples, conditioned on their acceptance, will be given by $\widetilde{\Sigma}_{\mathrm{avg}}\ (\mu - \widetilde{\mu})/D$, where $\widetilde{\Sigma}_{\mathrm{avg}}$ is the average of the conditional covariance of the surviving points. Since the operator norm of $\widetilde{\Sigma}_{\mathrm{avg}}$ could be as large as its trace, which could itself be as large as D, the bias caused by the rejection sampling over the *entire space* could hence be prohibitively large. That being said, if we can find some *low-variance* subspace $\mathcal{V}$ such that $v^\top \widetilde{\Sigma}_{\mathrm{avg}} v$ is small for any $v \in \mathcal{V}$, the magnitude of the bias within $\mathcal{V}$ will be only a small constant multiple of $\|\mu - \widetilde{\mu}\|_2$. Fortunately, since the trace of $\widetilde{\Sigma}_{\mathrm{avg}}$ is at most $O(D)$, by a simple averaging argument, $\widetilde{\Sigma}_{\mathrm{avg}}$ must have at least $D/2$ many eigenvalues that are at most $O(1)$. Thus, the subspace spanned by the corresponding eigenvectors gives the desired *low-variance* subspace. Moreover, we show that this subspace can be approximately computed from the samples (more precisely, we can compute a subspace of *similarly* low variance, up to polylogarithmic factors). It then remains to estimate the mean μ within the complement high-variance subspace $\mathcal{V}^\perp$. To achieve this goal, the idea is to just project the datapoints onto $\mathcal{V}^\perp$, and recursively run the same algorithm in that lower dimensional (of dimension $D/2$) subspace.[2] When the recursive procedure reaches a subspace with dimension $\mathrm{polylog}(D)$, we can simply run the 1-d estimator along each axis to finish the recursion. Since $\mathcal{V}^\perp$ now has dimension only $D/2$, the recursion terminates after $\log D$ many iterations.

In our description so far, one full execution of the recursive algorithm yields some $\widehat{\mu}$ with error $\|\widehat{\mu} - \mu\|_2 \leq \|\widetilde{\mu} - \mu\|_2/2 + \sqrt{D/(\alpha N)} + f(\alpha, N)$ (up to polylogarithmic factors). The first term corresponds to the bias caused by the rejection sampling and the second term

[2] Here we assume that the algorithm can take multiple datasets, where each of them is generated by Definition 1.1. We argue that this can be easily simulated with a single dataset generated by Definition 1.1; see the full version of the paper [17] for more details.

corresponds to the statistical error of D-dimensional mean estimation of the "good" αN samples of bounded covariance. Finally, the last term corresponds to the error of the 1-d estimator used in the base case of the recursion. Thus, the execution of this recursive algorithm improves the estimation error by a constant factor, provided that $\|\widetilde{\mu} - \mu\|_2$ is still significantly larger than the error bound stated in Theorem 1.2. By iteratively repeating this process for $\log(ND)$ many iterations, the estimation error can be brought down to the error bound of Theorem 1.2.

1.3 Related Work

1.3.1 Additional Related Work on Entangled Mean Estimation. In this section we discuss the works that are most closely related to this paper. We refer the reader to [48–50] and [12] for additional references and in-depth discussion of earlier work in the statistics literature.

The work of [8] studied entangled mean estimation in one dimension. Instead of assuming that a subset of the samples have bounded variances, like in Definition 1.1, the N samples are $x_i \sim \mathcal{N}(\mu, \sigma_i^2)$ with $\sigma_1 \le \sigma_2 \le \cdots \le \sigma_N$ and the error guarantee is stated directly as a function of the σ_i's. They show that the best possible estimation error is on the order of $1/\sqrt{\sum_{i=1}^N \sigma_i^{-2}}$ when the variances are known a priori. Otherwise, they show an error of $\widetilde{O}(\sqrt{N}\sigma_{\log n})$ is achievable in the absence of such knowledge. [8] also studies the high-dimensional setting where the samples follow spherical Gaussian distributions, i.e., with covariances equal to $\sigma_i^2\mathbf{I}$. As already mentioned, the task with the spherical covariances becomes easier in higher dimensions, as it is possible to estimate the covariance scale parameter σ_i. Using this, they achieve an error bound on the order of $1/\sqrt{\sum_{i=2}^N \sigma_i^{-2}}$ in ℓ_∞ distance when $D \gg \log N$. Notably, this almost recovers the error bound when the covariances are known a priori, except for missing the dependency on σ_1.

Subsequent works [48–50] explore the more challenging non-Gaussian setting under only the assumptions of unimodality and radial (spherical) symmetry. Specifically, they give a hybrid estimator that achieves an error rate of $\sqrt{D}\sqrt{N}^{1/D}\sigma$ provided that at least $\Omega(\log N/N)$-fraction of samples have marginal variance bound σ. One could see that this result recovers that of [8] by setting $D = 1$. The authors also consider the more general settings where the distributions are only assumed to be *centrally* symmetric, i.e., the density function $\rho : \mathbb{R}^d \mapsto \mathbb{R}_+$ satisfies $\rho(x) = \rho(-x)$, and achieve an error of $O(\sqrt{N})$. This setting covers our setup of non-spherical Gaussians. Yet, as pointed out in [10], the error bound given in [48] is sub-optimal under the subset-of-signals model even in the 1-d case.

The work of [12] also uses symmetry in place of Gaussianity. Their algorithm is fully adaptive, i.e., requiring no parameter tuning for specific distribution families, and is made possible by the techniques of intersecting confidence intervals, which has later inspired the work of [10] that leads to (nearly) optimal 1-d estimators in the subset-of-signals model.

The work [42, 61] introduced the subset-of- signals model, provided a nearly optimal lower bound within the model, and showed theoretical guarantees for *the iterative trimming algorithm*—a widely used heuristic for entangled mean estimation. Notably, the algorithm works by iteratively searching for a mean parameter that minimizes the square distance to a subset of samples, and then searching for a subset of samples that minimize the squared distance to a given estimate of the mean parameter. Our algorithm bears some similarity to these techniques as we are also using estimates from past iterations to perform rejection sampling on samples, and then use the surviving samples to construct new estimates. We refer the readers to Section 2 for a detailed outline of our techniques.

Finally, Theorem 1 in [10] gives a nearly optimal 1-d estimator for the subset-of-signals model. Similar to [8], they show that the result can be easily applied in the multivariate spherical Gaussians setting to nearly recover the error bound achievable with prior knowledge on variance scales. Moreover, as an improvement, they only require the dimension to be at least 2 for the multivariate bound to be effective.

1.3.2 Comparison of Optimal Error in Spherical vs Arbitrary Gaussians. From the results of [8, 10] and [42], it can be seen that there is a polynomial gap between the the errors in the cases of spherical Gaussians and those with arbitrary covariances in the subset-of-signals model. The first observation is that in the arbitrary covariance matrix setting, any estimator must have an error of $f(\alpha, N)$ (up to polylogarithmic factors) as shown in [42] (this is because by allowing arbitrary covariances one can encode the hard one-dimensional instance in one of the axes). This means that the error is at least $f(\alpha, N) \ge \widetilde{\Omega}(1/(\alpha^{2/3}\sqrt{N}))$. On the other hand, for *spherical* Gaussians, [8] and [10] give estimators with errors $O(1/\sqrt{\sum_{i=1}^N \sigma_i^{-2}})$. Using $\sigma_1 \le \ldots \le \sigma_{\alpha N} \le 1$ to translate that to the subset-of-signals model, the estimator for the spherical Gaussians can achieve an error $O(1/\sqrt{\alpha N})$. This shows that there is a $\mathrm{poly}(1/\alpha)$ gap between the optimal errors in the two settings.

1.3.3 Further Related Work.

Robust Statistics. From a robustness perspective, samples with arbitrarily large covariances in Definition 1.1 can be viewed as outliers. Robust statistics typically considers stronger outlier models, such as the *Huber contamination model* [31], where each outlier point is sampled from an unknown and potentially arbitrary distribution; or the *strong contamination model* where outliers can be adversarial and even violate the independence between samples. Unlike the results of this paper, consistency in these more challenging models is impossible, with the error bounded from below by some positive function of the fraction of outliers. Efficient high-dimensional mean estimation in the aforementioned corruption models was first achieved in [13, 40], where the outlier fraction is a constant smaller than 1/2. When this fraction is more than 1/2, it is no longer possible to produce a single estimate with worst-case guarantee. This setting is known as 'list-decodable' mean estimation, first studied in [7]. We refer to the recent book [14] for an overview of algorithmic robust statistics.

Other models of semi-oblivious adversaries. Similar to the subset-of-signals model, there are other frameworks for modeling less adversarial outliers that are assumed to be independent and have additional (Gaussian) structure. These models are commonly referred as semi-oblivious noise models. One such model assumes that inlier points are distributed as $x_i \sim \mathcal{N}(\mu, 1)$ and an $\alpha < 1/2$ fraction of outlier points are sampled as $x_i \sim \mathcal{N}(z_i, 1)$ where z_i's are arbitrary centers. Recent work [38] has characterized the error for estimating μ in this model, improving upon a line of work [6, 9, 33]. [9] studies the multivariate extension of the model. Finally, beyond mean estimation, the idea of modeling outliers via mean shifts has also been explored in the context of regression in [24, 44, 51, 54].

2 The Dimensionality Reduction Algorithm and Proof Roadmap

This section describes our main algorithm in tandem with a detailed sketch of its correctness proof. The pseudocode is provided in Algorithms 1 to 4. Algorithm 1 describes our main algorithm that computes a rough initial estimate, and then leverages a dimensionality reduction routine to iteratively improve the estimation error. Algorithm 4 contains the main subroutine RecursiveEstimate. Aided by two subroutines, FindSubSpace and PartialEstimate (cf. Algorithms 2 and 3) whose functionalities will be explained later on, RecursiveEstimate takes as input an estimate $\widehat{\mu} \in \mathbb{R}^D$ and a batch of $\Theta\left(N/\left(\log N \log D\right)\right)$ independent samples from the data-generating model of Definition 1.1, and produces a more refined estimate $\widehat{\mu}'$ satisfying the following guarantee: if $\|\widehat{\mu} - \mu\|_2$ is significantly larger than the error bound specified in Theorem 1.2, then the new estimate $\widehat{\mu}'$ satisfies $\|\widehat{\mu}' - \mu\|_2 \leq \|\widehat{\mu} - \mu\|_2/2$. This is summarized in the following statement (see the full version [17] for the formal statement). [3]

Algorithm 1 Entangled Mean Estimation in High Dimension

```
function EntangledMeanEstimation($N$)
    Input: Total number of samples $N$ to use, and noise-to-signal ratio $\alpha \in (0, 1)$.
    Output: $\mu' \in \mathbb{R}^D$.
    $m \leftarrow \log_2 D, r \leftarrow \lceil \log_2 N \rceil$   ▷ max depth of recursion and number of outer loop iterations
    Set $\delta$ to be some sufficiently large constant. [4]
    $\tau \leftarrow N^{-\delta}/r, n \leftarrow \frac{N}{2+m(3r+1)}$   ▷ Probability of failure $\tau$ and sample budget $n$ per iteration.
    $\widehat{\mu} \leftarrow$ TournamentImprove($\mathbf{0}, n, \tau$).   ▷ Rough estimate $\widehat{\mu}$ with $O(\sqrt{D})$ error
    If $f_\delta(\alpha, n) \geq \sqrt{D}$ then return $\widehat{\mu}$   ▷ where $f_\delta(\cdot)$ is defined in Equation (1)
    for $i = 1, \dots, r$ do   ▷ each iteration improves estimation error
        $\widehat{\mu}' \leftarrow$ RecursiveEstimate($\mathbf{I}, n, \widehat{\mu}, \tau$),   ▷ Iterative improvement (cf. Algorithm 4)
        $\widehat{\mu} \leftarrow \widehat{\mu}'$
    end for
    return $\widehat{\mu}$
end function
```

Lemma 2.1 (Iterative Refinement (Informal; see the full version [17])). *There exists an algorithm* RecursiveEstimate *that takes as input $\widehat{\mu} \in \mathbb{R}^D$ satisfying $\|\widehat{\mu} - \mu\|_2 \lesssim \sqrt{D}$, uses a dataset of $n = \Theta\left(N/\left(\log N \log D\right)\right)$ independent samples from the model of Definition 1.1, and produces some $\widehat{\mu}'$ such that the following holds with high probability (where $f(\alpha, n)$ is defined in Equation (1)):*

$$\left\|\widehat{\mu}' - \mu\right\|_2 \leq \frac{1}{2}\|\widehat{\mu} - \mu\|_2 + \operatorname{polylog}(ND)\left(\sqrt{\frac{D}{\alpha N}} + f(\alpha, n)\right).$$

Suppose that we have a rough estimate $\widehat{\mu}$ satisfying $\|\widehat{\mu} - \mu\|_2 \lesssim \sqrt{D}$. It is easy to see that running RecursiveEstimate for at most $\Theta(\log(N))$ many iterations[5] will bring this error down to the one presented in Theorem 1.2. This iterative reduction is accomplished using the for loop in Algorithm 1.

We now summarize the idea behind Lemma 2.1. Let $\widehat{\mu}$ be the rough estimate provided as input. The routine RecursiveEstimate is a recursive function that performs divide-and-concur on the dimension D. In particular, in each step, it receives a batch of $n = \Theta\left(N/(\log(N)\log(D))\right)$ many samples, and uses the subroutine FindSubSpace to partition the current subspace further into a low-variance and a high-variance subspace[6], each with dimension at most half of the original subspace. Then the algorithm invokes PartialEstimate to compute an estimate within the low-variance subspace that has improved error; while for the high-variance subspace it makes a recursive call. The recursion continues until we reach a subspace of dimension $d \lesssim \operatorname{polylog}(ND)$. In that case, we can simply use the 1-d estimator from [10] along each axis in that low-dimensional subspace; the corresponding error guarantee will be at most $\sqrt{d} = \operatorname{polylog}(ND)$ factor larger than that of the error $f(\alpha, N)$ of the 1-d base estimator (see [17]).

Suppose that the dimension of the current subspace is d, i.e., assume that the samples are now vectors in $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$. The analysis consists of the following inductive claim: as long as the dimension is not too small, i.e., $d \gg \operatorname{polylog}(ND)$, a single step of RecursiveEstimate produces a subspace $\mathcal{V} \subset \mathbb{R}^d$ of dimension $\dim(\mathcal{V}) \leq d/2$ (this is the low-variance subspace identified by FindSubSpace) together with an estimate $\mu_{\text{low}} \in \mathcal{V}$ such that with high probability

$$\|\mu_{\text{low}} - \mathbf{\Pi}_{\mathcal{V}}\mu\|_2 \leq \|\widehat{\mu} - \mu\|_2/(2\log_2 D) + \widetilde{O}\left(\sqrt{\dim(\mathcal{V})/(\alpha N)}\right), \tag{2}$$

where $\mathbf{\Pi}_{\mathcal{V}}$ is the orthonormal projector onto $\mathcal{V}$. If D denotes the original dimension of the space for which RecursiveEstimate is first called, after the end of all recursive calls, the algorithm will have partitioned the entire space $\mathbb{R}^D$ into at most $\log_2 D$ orthogonal low-variance subspaces and will have produced an estimate for each of these subspaces with error as in Equation (2). The final estimate $\widehat{\mu}'$ returned is a combination of the estimate for each subspace as well

[3] As will be explained later on, the routine is designed to operate in any subspace $\mathcal{V}$ provided in the input. For this reason, the formal statement in the full version [17] uses a matrix $\mathbf{P}$ whose rows are the orthonormal vectors that span that subspace ($\mathbf{P}^\top\mathbf{P}$ is the orthogonal projection matrix). What we present here corresponds to the simplified case $\mathbf{P} = \mathbf{I}$ and $d = D$.

[4] More formally, with n input samples, the 1-d estimator from [10] with probability $1 - n^{-\delta}$ achieves error $f_\delta(\alpha, n)$ (cf. Equation (1)). For the purpose of our algorithm, we require its success probability to be a sufficiently large polynomial in n. Hence, we set δ to be some sufficiently large constant.

[5] For the purpose of this introductory section, we will pretend as if the algorithm can draw an independent dataset from Definition 1.1 in each iteration. This type of access can be simulated by randomly splitting a large dataset into smaller ones and is presented formally in the full version [17].

[6] It will be explained later on in this introductory section in what sense we call the subspace "low-variance".

as the base case estimator (which has error $\mathrm{polylog}(ND) f(\alpha, N)$). By summing all the errors together, it can be seen that the error of $\widehat{\mu}'$ is the one provided in Lemma 2.1. The rest of this section provides a sketch of the correctness proof for Equation (2).

2.1 Warm-Start Estimate via Tournament

RecursiveEstimate from Lemma 2.1 requires some $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_2 \lesssim \sqrt{D}$. If the 1-d estimator achieves accuracy $\epsilon = O(1)$, we can simply use it along each axis to obtain such a $\widehat{\mu}$. In the more challenging regime when $\epsilon = \omega(1)$, we will exploit the fact that a sample with covariance bounded by $\mathbf{I}$ is within distance of $\sqrt{D}$ from μ in expectation. While accurately identifying such a sample is hard, by taking roughly $1/\alpha$ many samples $\mu_1, \dots, \mu_{1/\alpha}$, one μ_i of them will satisfy $\|\mu_i - \mu\|_2 \lesssim \sqrt{D}$ with high probability. We then design a *tournament* procedure to choose a vector $\widetilde{\mu}$ from $\mu_1, \dots, \mu_{1/\alpha}$ which is almost as good as that μ_i.

Lemma 2.2 (Tournament (Informal; see [17])). *There exists an algorithm that takes as input a list $L = \{\mu_1, \cdots, \mu_k\} \subset \mathbb{R}^D$, draws a dataset of size n from the model of Definition 3.5 from [17], and outputs an estimate $\mu_j \in L$ such that $\|\mu_j - \mu\|_2 \lesssim \min_{i \in [k]} \|\mu_i - \mu\|_2 + f_\delta(\alpha, n)$ with high probability.*

The algorithm establishing the above lemma is inspired by ideas developed in the context of list-decodable mean estimation [15]. In particular, for every pair of candidate estimates μ_j, μ_ℓ, the algorithm compares $\mu_j - \mu$ and $\mu_\ell - \mu$ *projected* along the direction of $\mu_j - \mu_\ell$ to decide wether or not to remove j from the list. This kind of comparisons, essentially reduce the problem into one-dimensional tasks, where we can again just use the 1-d estimator to learn the projection of μ along the direction of $\mu_\ell - \mu_j$. We remark that while the goal of a tournament procedure is to prune the candidate list down to a shorter list *containing* a good estimate in the list-decodable setting, our tournament procedure needs to produce a *single* estimate that has nearly optimal error. To achieve this, we require a novel analysis leveraging the structure of our setting. The formal statement and proof can be found in **??**.

2.2 Rejection Sampling

Recall that our procedure RecursiveEstimate relies critically on finding a low-variance subspace. Unfortunately, a low-variance subspace may not exist in general—as even a single noisy sample can cause the variance to be arbitrarily large. Our key idea is to use the warm-start estimate $\widetilde{\mu}$ obtained from the tournament procedure as a rejection sampling center to filter out unusually noisy samples: If $x_i \sim \mathcal{N}(\mu, \Sigma_i)$ denote the original samples, for each i we sample a decision variable $b_i \sim \mathrm{Ber}\left(\exp\left(-\|x_i - \widetilde{\mu}\|_2^2/d\right)\right)$ independently to decide whether or not to accept the sample.

Distribution of Accepted Samples. To analyze the effect of rejection sampling, denote by $\mathcal{A}_i$ the distribution of the i-th sample conditioned on its acceptance, i.e., the distribution of $x_i \sim \mathcal{N}(\mu, \Sigma_i)$ conditioned on $b_i = 1$. A convenient fact is that $\mathcal{A}_i$ is simply another Gaussian $\mathcal{N}(\widetilde{\mu}_i, \widetilde{\Sigma}_i)$, for some $\widetilde{\mu}_i, \widetilde{\Sigma}_i$ defined as functions of $\mu, \widetilde{\mu}, \Sigma_i$ that are provided in [17]. Denote by $S_{\mathcal{A}} = \{i \in [n] : b_i = 1\}$ the set of all accepted indices (which is a random set). By independence between samples and the aforementioned fact, if we condition on a set $S_{\mathcal{A}} = S$, then the conditional joint distribution of $\{x_i\}_{i \in S}$ is equal to the product distribution of the Gaussians $\mathcal{N}(\widetilde{\mu}_i, \widetilde{\Sigma}_i)$ for $i \in S$ (with $\widetilde{\mu}_i, \widetilde{\Sigma}_i$ defined in equation (5) in the full version [17]). We can thus adopt the following more convenient view of the random process that generates the *accepted* samples.

Definition 2.3 (Generation of accepted samples—alternative view).

(1) $S_{\mathcal{A}}$ is generated by independently including $i \in [n]$ with probability $\mathbf{E}_{x_i \sim \mathcal{N}(\mu, \Sigma_i)}[e^{-\|x_i - \widetilde{\mu}\|_2^2/d}]$.
(2) For each $i \in S_{\mathcal{A}}$, x_i is drawn independently from $\mathcal{A}_i := \mathcal{N}(\widetilde{\mu}_i, \widetilde{\Sigma}_i)$.

Thus, if we condition on a particular $S_{\mathcal{A}}$ of accepted indices, we could view the accepted vectors simply as independent samples from the distributions $\{\mathcal{A}_i = \mathcal{N}(\widetilde{\mu}_i, \widetilde{\Sigma}_i)\}$. This property makes it manageable to analyze the statistical properties of subroutines that work on the accepted samples.

Rejection of Noisy Samples. Let $x_1, \dots, x_k \in \mathbb{R}^d$ be the accepted samples (by renaming $S_{\mathcal{A}}$ to $[k]$). On the one hand, by definition of our acceptance rule, we will rarely see samples that are at a distance more than $\sqrt{d}$ from the rejection center $\widetilde{\mu}$ (where d is the dimension of the subspace we are currently working in). This ensures that the averaged covariance matrix of the accepted samples, $\frac{1}{k}\sum_{i=1}^k \widetilde{\Sigma}_i$, must have its trace appropriately bounded. In particular, we show in the full version [17] that with high probability over the randomness of the accepted indices $S_{\mathcal{A}}$, the distributions $\mathcal{A}_i$ for $i \in S_{\mathcal{A}}$ satisfy that $\mathbf{E}_{z \sim \mathcal{A}_i}\left[\|z_i - \widetilde{\mu}\|_2^2\right] \lesssim d \log(nd)$. Via a linear algebraic argument, this immediately implies that $\mathrm{tr}\left(\frac{1}{k}\sum_{i=1}^k \widetilde{\Sigma}_i\right) \lesssim d \log(nd)$. By a simple averaging argument, we therefore know that there must exist a subspace of dimension at least $d/2$ such that the eigenvalues of $\frac{1}{k}\sum_{i=1}^k \widetilde{\Sigma}_i$ within the subspace is at most $O(\log(nd))$. This therefore guarantees the existence of a low-variance subspace of the average population covariance matrix. Finally, note that $\widetilde{\Sigma}_i$ are unknown population covariances; thus, even though we showed that a low-variance subspace exists, we yet have to provide a way to compute such a subspace from samples. This will be done in Section 2.3.

Survival of Samples with Bounded Covariance. On the other hand, by standard Gaussian norm concentration, a sample $x_i \in \mathbb{R}^d$ with covariance bounded by $\mathbf{I}$ is rarely at a distance more than $\sqrt{d}$ from the mean μ. Provided that $\|\widetilde{\mu} - \mu\| \lesssim \sqrt{d}$, we thus have that such samples will pass the rejection sampling procedure with at least constant probability. A slight issue is that as we go deeper into the recursion tree of the algorithm, each call of the recursive routine projects everything onto a subspace with half the dimension d of the parent call. Thus, the requirement $\|\widetilde{\mu} - \mu\| \lesssim \sqrt{d}$, although true in the beginning when $d = D$, may no longer hold in later steps of the recursion. Fortunately, at each call of the routine, we can take more fresh samples, project them onto the current subspace $\mathcal{V}$, and invoke the tournament procedure from Lemma 2.2 to produce some new estimate $\widetilde{\mu} \in \mathbb{R}^d$ that is guaranteed to be within distance $\sqrt{d}$ from the projected mean (cf. Algorithm 4 in Algorithm 4).

Bias of Rejection Sampling. A more significant issue concerns the bias in the mean of the accepted samples. Let us examine the averaged mean over the accepted samples. A straightforward computation (see [17]) shows that this expectation is:

$$\frac{1}{k}\sum_{i=1}^{k}\widetilde{\mu}_i - \mu = \frac{2}{d}\widetilde{\Sigma}_{\text{avg}}\left(\widetilde{\mu}-\mu\right)\ , \tag{3}$$

where we define $\widetilde{\Sigma}_{\text{avg}} := \frac{1}{k}\sum_{i=1}^{k}\widetilde{\Sigma}_i$ for convenience. The question then is what is the best upper bound that we can use for the operator norm of $\widetilde{\Sigma}_{\text{avg}}$. A basic fact is that, conditioned on the accepted indices, each accepted sample x_i follows the Gaussian $\mathcal{A}_i := \mathcal{N}(\widetilde{\mu}_i, \widetilde{\Sigma}_i)$ with $\widetilde{\mu}_i = \widetilde{\Sigma}_i\left(2\widetilde{\mu}/d + \Sigma_i^{-1}\mu\right)$ and $\widetilde{\Sigma}_i = \left(\Sigma_i^{-1} + (2/d)\mathbf{I}\right)^{-1}$ (cf. [17]). This immediately implies that $\|\widetilde{\Sigma}_i\|_2 \le d/2$. However, if the operator norm of the averaged covariance $\widetilde{\Sigma}_{\text{avg}}$ could be as large as $d/2$, then the bias of the accepted samples would not be any better than the error of $\widetilde{\mu}$ that we started with! Fortunately, as discussed earlier, the fact that $\widetilde{\Sigma}_{\text{avg}}$ has bounded trace implies that there must exist a subspace $\mathcal{V}$ of dimension at least $d/2$ such that the eigenvalues of $\widetilde{\Sigma}_{\text{avg}}$ constrained to $\mathcal{V}$ are at most $\log(nd)$. Hence, the empirical mean over the accepted samples would constitute a good estimator within this low-variance subspace $\mathcal{V}$—if we could successfully *identify* it. After that, we can recurse on the orthogonal complement of $\mathcal{V}$, and that would complete the analysis of RecursiveEstimate. The remaining task is therefore to design a procedure that can identify this subspace.

2.3 Searching for a Low-Variance Subspace

We devise a procedure FindSubSpace (cf. Algorithm 2) for identifying a low-variance subspace with respect to the accepted samples.

Algorithm 2 Function to find a subspace in which the accepted samples have low variance

```
function FindSubSpace($\widetilde{\mu}, x_1, \cdots, x_k$)
    Input: Rough mean estimate $\widetilde{\mu} \in \mathbb{R}^d$, and samples $x_1, \cdots, x_k \in \mathbb{R}^d$.
    Output: Matrices $\mathbf{P}_{\text{low}}, \mathbf{P}_{\text{high}} \in \mathbb{R}^{d/2 \times d}$ whose rows form an orthonormal basis of $\mathbb{R}^d$.
    Let $\widetilde{\mathbf{M}}_{(\text{emp})} = \frac{1}{k}\sum_{i\in[k]}(x_i-\widetilde{\mu})(x_i-\widetilde{\mu})^\top$.
    Let $v_1, v_2, \cdots, v_d$ be the $d$ eigenvectors of $\widetilde{\mathbf{M}}_{(\text{emp})}$ in descending order of their eigenvalues.
    Let $\mathbf{P}_{\text{high}} = [v_1, \ldots, v_{\lfloor d/2 \rfloor}]^\top$ and $\mathbf{P}_{\text{low}} = [v_{\lfloor d/2 \rfloor+1}, \ldots, v_d]^\top$.
    return $\mathbf{P}_{\text{low}}, \mathbf{P}_{\text{high}}, \mu_{\text{low}}$
end function
```

Lemma 2.4 (Low-Variance Subspace Identification (Informal; see full version [17])). *Let $x_i \sim \mathcal{N}(\widetilde{\mu}_i, \widetilde{\Sigma}_i) \in \mathbb{R}^d$ for $i \in [k]$ be the set of accepted samples. Assume that $k \gg d\,\mathrm{polylog}(nd)$. The procedure FindSubSpace takes the samples as input, and produces a subspace $\mathcal{V}$ such that the following holds with high probability: $v^\top \frac{1}{k}\sum_{i\in[k]}\widetilde{\Sigma}_i v \lesssim \log(nd)$ for all unit vectors $v \in \mathcal{V}$.*

To search for the low-variance subspace, we take the following natural approach. Given accepted samples $x_1, \cdots, x_k \in \mathbb{R}^d$ that pass through the rejection sampling centered at $\widetilde{\mu}$, we compute the second moment matrix $\widetilde{\mathbf{M}}_{(\text{emp})} := \frac{1}{k}\sum_{i=1}^{k}(x_i-\widetilde{\mu})(x_i-\widetilde{\mu})^\top$, and take the subspace spanned by the bottom $d/2$ eigenvectors. The challenging part is to show that the averaged covariance matrix $\widetilde{\Sigma}_{\text{avg}}$ will indeed have small eigenvalues within this subspace with high probability. The proof strategy comprises of showing that the following two claims hold with high probability:

(1) There is a subspace $\mathcal{V}$ in which the empirical second moment $\widetilde{\mathbf{M}}_{(\text{emp})}$ of accepted samples has small operator norm.
(2) The averaged covariance matrix $\widetilde{\Sigma}_{\text{avg}}$ can be bounded from above (in Lowner order) by the sample second moment matrix $\widetilde{\mathbf{M}}_{(\text{emp})}$ (up to a constant factor).

It is not hard to see that combining the two statements above yields that, with high probability, $\widetilde{\Sigma}_{\text{avg}}$ will have bounded eigenvalues within the subspace $\mathcal{V}$. The formal statement of Item 1 is given in [17]. Its proof follows mostly from the properties of rejection sampling and the details can be found in the full version of the paper [17]. The formal statement of Item 2 also is given in [17]. The proof idea is to use the "truncation" technique. In particular, define $\mathbf{X}_i = (x_i - \widetilde{\mu})(x_i - \widetilde{\mu})^\top$. We will decompose $\mathbf{X}_i$ into

$$\mathbf{Y}_i = \mathbb{1}\{\|x_i-\widetilde{\mu}\|_2 \le \widetilde{\Theta}(\sqrt{d})\}\mathbf{X}_i \text{ and } \mathbf{Z}_i = \mathbb{1}\{\|x_i-\widetilde{\mu}\|_2 > \widetilde{\Theta}(\sqrt{d})\}.$$

By the property of the rejection sampling procedure, the $\mathbf{Z}_i$ are almost always $\mathbf{0}$. On the other hand, the $\mathbf{Y}_i$ now have bounded operator norm almost surely. We could then apply the Matrix Bernstein Inequality to argue about the spectral concentration of $\sum_{i\in[k]}\mathbf{Y}_i$. The details are provided in [17].

2.4 Improvement within the Low-Variance Subspace

Building on top of FindSubSpace, we give another procedure PartialEstimate (cf. Algorithm 4) that simultaneously identifies a low variance subspace $\mathcal{V} \subset \mathbb{R}^d$ and an estimate μ_{low} such that they satisfy Equation (2) with high probability. See [17] for the formal statement. Notably, the routine uses the same batch of accepted samples to search for the low-variance subspace, and to compute the empirical mean estimator μ_{low}. To avoid adaptive conditioning (i.e., conditioning on concentration of the sample mean within the subspace $\mathcal{V}$ defined by the samples), we instead show that with high probability the deviation of the sample mean from the population mean along *every* direction is bounded from above by some quantity proportional to the population variance along that direction. That is, conditioned on that $\{1, \ldots, k\}$ are the indices of the accepted samples, then with probability at least $1-\tau$, the following holds for all $v \in \mathbb{R}^d$

$$v^\top\left(\frac{1}{k}\sum_{i=1}^{k}x_i - \frac{1}{k}\sum_{i=1}^{k}\widetilde{\mu}_i\right) \lesssim \sqrt{\frac{1}{d}}\sqrt{\frac{1}{k}\sum_{i=1}^{k}v^\top\widetilde{\Sigma}_i v}. \tag{4}$$

This is because, conditioned on the acceptance set being $\{1, \ldots, k\}$, the accepted samples are just independent Gaussians conditioned on a specific set S of accepted indices; thus, standard Gaussian concentration can be applied. The detailed proof is given in [17].

Algorithm 3 Function to estimate the mean within the Low-Variance Subspace

```
function PartialEstimate($\widetilde{\mu}, x_1, \cdots, x_n$)
    Input: Rough mean estimate $\widetilde{\mu} \in \mathbb{R}^d$, and samples $x_1, \cdots, x_n \in \mathbb{R}^d$.
    Output: Matrix $\mathbf{P}_{\mathrm{high}} \in \mathbb{R}^{d/2 \times d}$ and an estimate $\mu_{\mathrm{low}} \in \mathbb{R}^d$.

    For each $i \in [n]$ draw $b_i \sim \mathrm{Ber}(e^{-\|x_i - \widetilde{\mu}\|_2^2/d})$.    ▷ Rejection sampling
    $\mathbf{P}_{\mathrm{low}}, \mathbf{P}_{\mathrm{high}} \leftarrow$ FindSubSpace($\widetilde{\mu}, \{x_i : b_i = 1\}$).    ▷ Split space (cf. Section 2.3)
    $\mu_{\mathrm{low}} \leftarrow \frac{1}{\sum_{i \in [n]} b_i} \sum_{i \in [n]} b_i \mathbf{P}_{\mathrm{low}}^\top \mathbf{P}_{\mathrm{low}} x_i$    ▷ Empirical mean of surviving samples after projection
    return $\mathbf{P}_{\mathrm{high}}, \mu_{\mathrm{low}}$
end function
```

Algorithm 4 Recursive Function for Mean Estimation

```
function RecursiveEstimate($\mathbf{P}, n, \widehat{\mu}, \tau$)
    Input: Row orthonormal matrix $\mathbf{P} \in \mathbb{R}^{d \times D}$, sample budget $n \in \mathbb{N}$, failure probability $\tau \in (0, 1)$, mean estimate $\widehat{\mu} \in \mathbb{R}^d$ from last iteration, and noise-to-signal ratio $\alpha \in (0, 1)$.
    Output: Better mean estimate $\widehat{\mu}' \in \mathbb{R}^d$.

    Let $C, \delta$ be sufficiently large absolute constants.
    Draw an independent batch of $n$ samples $y_1, \cdots, y_n \in \mathbb{R}^D$ from the model of Definition 1.1 (see Footnote 5).
    Form a projected set of samples: $S \leftarrow \{x_i = \mathbf{P} y_i \ \forall i \in [n]\} \subset \mathbb{R}^d$.
    $\widetilde{\mu} \leftarrow$ TournamentImprove($\widehat{\mu}, n, \tau$) of [17].    ▷ Ensures $\|\widetilde{\mu} - \mathbf{P}\mu\|_2 \lesssim \sqrt{d} + f_\delta(\alpha, n)$.
    if $d \le C \log(nd/\tau)(\log D)^2$ then
        Let $\widehat{\mu}' \in \mathbb{R}^d$ be the output of the estimator from [17] computed on the dataset $S$.
        return $\widehat{\mu}'$    ▷ End the recursion
    else if $\sqrt{d} \le f_\delta(\alpha, n)$ then    ▷ Recall that $f_\delta(\cdot)$ is defined in Equation (1)
        return $\widehat{\mu}' = \widetilde{\mu}$.    ▷ We already have $\|\widetilde{\mu} - \mathbf{P}\mu\|_2 \lesssim f_\delta(\alpha, n)$ in this case
    else
        $\mathbf{P}_{\mathrm{high}}, \mu_{\mathrm{low}} \leftarrow$ PartialEstimate($\widetilde{\mu}, x_1, \cdots, x_n$)    ▷ (cf. Algorithm 3) Returns an orthonormal matrix $\mathbf{P}_{\mathrm{high}} \in \mathbb{R}^{d/2 \times d}$ corresponding to the high-variance subspace and some mean estimate $\mu_{\mathrm{low}} \in \mathbb{R}^d$ in the low-variance subspace.
        $\mu_{\mathrm{high}} \leftarrow$ RecursiveEstimate($\mathbf{P}_{\mathrm{high}}\mathbf{P}, n, \mathbf{P}_{\mathrm{high}}\widehat{\mu}, \tau$)    ▷ Recurse on high-variance subspace
        $\widehat{\mu}' \leftarrow \mu_{\mathrm{low}} + \mathbf{P}_{\mathrm{high}}^\top \mu_{\mathrm{high}}$    ▷ Combination of the two estimates in $\mathbb{R}^d$
        return $\widehat{\mu}'$
    end if
end function
```

Putting everything together to show (2). We are now ready to put everything together to prove (2). Recall that the procedure RecursiveEstimate works in a d-dimensional subspace of $\mathbb{R}^D$ defined by some orthonormal matrix $\mathbf{P} \in \mathbb{R}^{d \times D}$ given as its input. After projecting the sample points into the subspace with $\mathbf{P}$, it runs the tournament procedure from Lemma 2.2 to obtain a rough estimate $\widetilde{\mu} \in \mathbb{R}^d$ within the subspace. It then performs rejection sampling as outlined in Section 2.2 centered at $\widetilde{\mu}$. Let $x_1, \cdots, x_k$ be the set of accepted samples. By Equation (3), the bias of the samples will be of the form $(2/d)\, \widetilde{\Sigma}_{\mathrm{avg}}\, (\widetilde{\mu} - \mu)$. With high probability, the routine FindSubSpace yields a subspace $\mathcal{V}$ such that $v^T \widetilde{\Sigma}_{\mathrm{avg}} v = O(1)$ (up to polylogarithmic factors) for all $v \in \mathcal{V}$. Via a linear algebraic argument (cf. **??**), one can show that $\left\| \mathbf{\Pi}_{\mathcal{V}} \left(\frac{1}{k} \sum_{i=1}^k \mathbf{E}[x_i] - \mu \right) \right\|_2 \ll \|\widetilde{\mu} - \mu\|_2$. On the other hand, as shown in Equation (4), the deviation of the empirical mean $\frac{1}{k} \sum_{i=1}^k x_i$ from its expected value will be at most $\sqrt{d/k}$ (up to polylogarithmic factors). Combining the above two observations with the triangle inequality then concludes the proof of Equation (2).

The detailed pseudocode of RecursiveEstimate is given in Algorithm 4. For the formal statement and proof, we refer the reader to [17].

References

[1] D. Achlioptas and F. McSherry. 2005. On spectral learning of mixtures of distributions. In *Proc. 18th Annual Conference on Learning Theory (COLT)*.

[2] S. Arora and R. Kannan. 2001. Learning mixtures of arbitrary Gaussians. In *Proc. 33rd Annual ACM Symposium on Theory of Computing (STOC)*.

[3] A. Bakshi, I. Diakonikolas, H. Jia, D. M. Kane, P. K. Kothari, and S. S. Vempala. 2022. Robustly Learning Mixtures of k Arbitrary Gaussians. In *Proc. 54th Annual ACM Symposium on Theory of Computing (STOC)*.

[4] Mikhail Belkin and Kaushik Sinha. 2015. Polynomial learning of distribution families. *SIAM J. Comput.* 44, 4 (2015), 889–911.

[5] Rudolf Beran. 1982. Robust estimation in models for independent non-identically distributed data. *The Annals of Statistics* (1982), 415–428.

[6] T. T. Cai and J. Jin. 2010. Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *The Annals of Statistics* 38, 1 (2010), 100–145.

[7] M. Charikar, J. Steinhardt, and G. Valiant. 2017. Learning from Untrusted Data. In *Proc. 49th Annual ACM Symposium on Theory of Computing (STOC)*. ACM Press, 47–60. doi:10.1145/3055399.3055491

[8] F. Chierichetti, A. Dasgupta, R. Kumar, and S. Lattanzi. 2014. Learning Entangled Single-Sample Gaussians. In *Proc. 25th Annual Symposium on Discrete Algorithms (SODA)*. SIAM, 511–522. doi:10.1137/1.9781611973402.38

[9] O. Collier and A. Dalalyan. 2019. Multidimensional linear functional estimation in sparse Gaussian models and robust estimation of the mean. *Electronic Journal of Statistics* 13 (2019), 2830–2864.

[10] S. Compton and G. Valiant. 2024. Near-Optimal Mean Estimation with Unknown, Heteroskedastic Variances. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. 194–200. doi:10.1145/3618260.3649754

[11] S. Dasgupta. 1999. Learning mixtures of Gaussians. In *Proc. 40th IEEE Symposium on Foundations of Computer Science (FOCS)*. doi:10.1109/sffcs.1999.814639

[12] L. Devroye, S. Lattanzi, G. Lugosi, and N. Zhivotovskiy. 2023. On mean estimation for heteroscedastic random variables. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 59, 1 (2023), 1–20. doi:10.1214/21-aihp1239

[13] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. 2016. Robust Estimators in High Dimensions without the Computational Intractability. In *Proc. 57th IEEE Symposium on Foundations of Computer Science (FOCS)*. doi:10.1109/FOCS.2016.85

[14] I. Diakonikolas and D. M. Kane. 2023. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press. doi:10.1017/9781108943161

[15] I. Diakonikolas, D. M. Kane, and D. Kongsgaard. 2020. List-decodable mean estimation via iterative multi-filtering. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*. doi:10.1109/focs46700.2020.00022

[16] I. Diakonikolas, D. M. Kane, D. Kongsgaard, J. Li, and K. Tian. 2022. Clustering Mixture Models in Almost-Linear Time via List-Decodable Mean Estimation. In *Proc. 54th Annual ACM Symposium on Theory of Computing (STOC)*. doi:10.1145/3519935.3520014

[17] Ilias Diakonikolas, Daniel M Kane, Sihan Liu, and Thanasis Pittas. 2025. Entangled Mean Estimation in High-Dimensions. *arXiv preprint arXiv:2501.05425* (2025).

[18] Ilias Diakonikolas, Daniel M. Kane, Thanasis Pittas, and Nikos Zarifis. 2023. SQ Lower Bounds for Learning Mixtures of Separated and Bounded Covariance Gaussians. In *Proceedings of Thirty Sixth Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 195)*. PMLR, 2319–2349. https://proceedings.mlr.press/v195/diakonikolas23b.html

[19] I. Diakonikolas, D. M. Kane, and A. Stewart. 2018. List-Decodable Robust Mean Estimation and Learning Mixtures of Spherical Gaussians. In *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*. doi:10.1145/3188745.3188758

[20] Murat Dundar, Balaji Krishnapuram, Jinbo Bi, and R Bharat Rao. 2007. Learning classifiers when the training data is not IID.. In *IJCAI*, Vol. 2007. Citeseer, 756–61.

[21] Faouzi El Bantli and Marc Hallin. 1999. L1-estimation in linear models with heterogeneous white noise. *Statistics & probability letters* 45, 4 (1999), 305–315. doi:10.1016/s0167-7152(99)00072-3

[22] J. Fan, X. Han, and H. Liu. 2014. Challenges of Big Data Analysis. *National Science Review* 1, 2 (2014), 293–314. doi:doi.org/10.1093/nsr/nwt032

[23] Seth R Flaxman, Daniel B Neill, and Alexander J Smola. 2015. Gaussian processes for independence tests with non-iid data in causal inference. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7, 2 (2015), 1–23. doi:doi.org/10.1145/2806892

[24] I. Gannaz. 2007. Robust estimation and wavelet thresholding in partially linear models. *Statistics and Computing* 17 (2007), 293–310. doi:10.1007/s11222-007-9019-x

[25] D. Gao, Y. Yao, and Q. Yang. 2022. A survey on heterogeneous federated learning. *arXiv preprint arXiv:2210.04505* (2022).

[26] Yehoram Gordon, Alexander Litvak, Carsten Schütt, and Elisabeth Werner. 2006. On the minimum of several random variables. *Proc. Amer. Math. Soc.* 134, 12 (2006), 3665–3675. doi:10.1090/s0002-9939-06-08453-x

[27] M. Hallin and I. Mizera. 1997. Unimodality and the asymptotics of M-estimators. *Lecture Notes-Monograph Series* (1997), 47–56. doi:doi.org/10.1214/lnms/1215454126

[28] Marc Hallin and Ivan Mizera. 2001. Sample heterogeneity and M-estimation. *Journal of statistical planning and inference* 93, 1-2 (2001), 139–160. doi:10.1016/s0378-3758(00)00174-9

[29] Wassily Hoeffding. 1956. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics* (1956), 713–721. doi:10.1214/aoms/1177728178

[30] S. B. Hopkins and J. Li. 2018. Mixture Models, Robustness, and Sum of Squares Proofs. In *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*. doi:10.1145/3188745.3188748

[31] P. J. Huber. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* 35, 1 (March 1964), 73–101. doi:10.1214/aoms/1177703732

[32] IA Ibragimov and RZ Has' minskii. 1976. Local asymptotic normality for non-identically distributed observations. *Theory of Probability & Its Applications* 20, 2 (1976), 246–260. doi:10.1137/1120032

[33] J. Jin. 2008. Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70, 3 (2008), 461–493. doi:10.1111/j.1467-9868.2007.00645.x

[34] S. Kamm, S. S. Veekati, T. Müller, N. Jazdi, and M. Weyrich. 2023. A survey on machine learning based analysis of heterogeneous data in industrial automation. *Computers in Industry* 149 (2023), 103930. doi:doi.org/10.1016/j.compind.2023.103930

[35] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. 2008. The Spectral Method for General Mixture Models. *SIAM J. Comput.* 38, 3 (2008), 1141–1156. doi:10.1137/S0097539704445925

[36] Keith Knight. 1999. Asymptotics for L1-estimators of regression parameters under heteroscedasticityY. *Canadian Journal of Statistics* 27, 3 (1999), 497–507.

[37] Weihao Kong, Raghav Somani, Zhao Song, Sham Kakade, and Sewoong Oh. 2020. Meta-learning for mixed linear regression. In *International Conference on Machine Learning*. PMLR, 5394–5404.

[38] Subhodh Kotekal and Chao Gao. 2025. Optimal Estimation of the Null Distribution in Large-Scale Inference. *IEEE Transactions on Information Theory* 71, 3 (2025), 2075–2103. doi:10.1109/TIT.2025.3529457

[39] P. K. Kothari, J. Steinhardt, and D. Steurer. 2018. Robust Moment Estimation and Improved Clustering via Sum of Squares. In *Proc. 50th Annual ACM Symposium on Theory of Computing (STOC)*. doi:10.1145/3188745.3188970

[40] K. A. Lai, A. B. Rao, and S. Vempala. 2016. Agnostic Estimation of Mean and Covariance. In *Proc. 57th IEEE Symposium on Foundations of Computer Science (FOCS)*. doi:10.1109/FOCS.2016.76

[41] J. Li. 2024. Entangled Mean Estimation in High Dimensions. Open Problems Session at Workshop on New Frontiers in Robust Statistics, TTI-Chicago, June 2024. Personal Communication.

[42] Y. Liang and H. Yuan. 2020. Learning Entangled Single-Sample Gaussians in the Subset-of-Signals Model. In *Proc. 33rd Annual Conference on Learning Theory (COLT)*. doi:10.1137/1.9781611973402.38

[43] A. Liu and J. Li. 2022. Clustering mixtures with almost optimal separation in polynomial time. In *Proc. 54th Annual ACM Symposium on Theory of Computing (STOC)*. doi:10.1137/22m1538788

[44] L. McCann and R. E. Welsch. 2007. Robust variable selection using least angle regression and elemental set sampling. *Computational Statistics & Data Analysis* 52, 1 (2007), 249–257. doi:10.1016/j.csda.2007.01.012

[45] Ivan Mizera and Jon A Wellner. 1998. Necessary and sufficient conditions for weak consistency of the median of independent but not identically distributed random variables. *Annals of statistics* (1998), 672–691. doi:10.1214/aos/1028144854

[46] 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. doi:10.1109/focs.2010.15

[47] VB Nevzorov. 1984. Rate of convergence to the normal law of order statistics for nonidentically distributed random variables. *Journal of Soviet Mathematics* 27 (1984), 3263–3270. doi:10.1007/bf01850675

[48] A. Pensia, V. Jog, and P. Loh. 2019. Estimating Location Parameters in Entangled Single-Sample Distributions. *CoRR* abs/1907.03087 (2019).

[49] A. Pensia, V. Jog, and P. Loh. 2019. Mean Estimation for Entangled Single-Sample Distributions. In *Proc. 2019 IEEE International Symposium on Information Theory*. 3052–3056. doi:10.1109/ISIT.2019.8849279

[50] Ankit Pensia, Varun Jog, and Po-Ling Loh. 2021. Estimating location parameters in sample-heterogeneous distributions. *Information and Inference: A Journal of the IMA* 11 (June 2021), 959–1036. doi:10.1093/imaiai/iaab013

[51] S. Sardy, P. Tseng, and A. Bruce. 2001. Robust wavelet denoising. *IEEE transactions on signal processing* 49, 6 (2001), 1146–1152. doi:10.1109/78.923297

[52] Pranab Kumar Sen. 1968. Asymptotic normality of sample quantiles for m-dependent processes. *The annals of mathematical statistics* 39, 5 (1968), 1724–1730. doi:10.1214/aoms/1177698155

[53] Pranab Kumar Sen. 1970. A note on order statistics for heterogeneous distributions. *The Annals of Mathematical Statistics* 41, 6 (1970), 2137–2139. doi:10.1214/aoms/1177696715

[54] Y. She and A. B. Owen. 2011. Outlier detection using nonconvex penalized regression. *J. Amer. Statist. Assoc.* 106, 494 (2011), 626–639. doi:10.1198/jasa.2011.tm10390

[55] Galen R Shorack and Jon A Wellner. 2009. *Empirical processes with applications to statistics*. SIAM. doi:10.1137/1.9780898719017

[56] Ingo Steinwart and Andreas Christmann. 2009. Fast learning from non-iid observations. *Advances in neural information processing systems* 22 (2009). doi:10.1109/vtc2023-spring57618.2023.10200108

[57] Stephen M Stigler. 1976. The effect of sample heterogeneity on linear functions of order statistics, with applications to robust estimation. *J. Amer. Statist. Assoc.* 71, 356 (1976), 956–960. doi:10.2307/2286868

[58] J.W. Tukey. 1975. Mathematics and picturing of data. In *Proceedings of the International Congress of Mathematicians (ICM)*, Vol. 6. 523–531.

[59] Lionel Weiss. 1969. The Asymptotic Distribution of Quantiles from Mixed Samples. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 31, 3 (1969), 313–318. http://www.jstor.org/stable/25049590

[60] Dong Xia. 2019. Non-asymptotic bounds for percentiles of independent non-identical random variables. *Statistics & Probability Letters* 152 (2019), 111–120. doi:10.1016/j.spl.2019.04.018

[61] H. Yuan and Y. Liang. 2020. Learning entangled single-sample distributions via iterative trimming. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2666–2676. doi:10.1137/1.9781611973402.38

[62] Tianqing Zhu, Ping Xiong, Gang Li, and Wanlei Zhou. 2014. Correlated differential privacy: Hiding information in non-IID data set. *IEEE Transactions on Information Forensics and Security* 10, 2 (2014), 229–242. doi:10.1109/TIFS.2014.2368363

Received 2024-11-04; accepted 2025-02-01