Optimal Sub-Gaussian Mean Estimation in Very High Dimensions

Jasper C. H. Lee
University of Wisconsin-Madison, WI, USA

Paul Valiant
Purdue University, West Lafayette, IN, USA

Abstract

We address the problem of mean estimation in very high dimensions, in the high probability regime parameterized by failure probability δ. For a distribution with covariance Σ, let its "effective dimension" be $d_{\text{eff}} = \frac{\text{Tr}(\Sigma)}{\lambda_{\max}(\Sigma)}$. For the regime where $d_{\text{eff}} = \omega(\log^2 \frac{1}{\delta})$, we show the first algorithm whose sample complexity is optimal to within $1 + o(1)$ factor. The algorithm has a surprisingly simple structure: 1) re-center the samples using a known sub-Gaussian estimator, 2) carefully choose an easy-to-compute positive integer t and then remove the t samples farthest from the origin and 3) return the sample mean of the remaining samples. The core of the analysis relies on a novel vector Bernstein-type tail bound, showing that under general conditions, the sample mean of a bounded high-dimensional distribution is highly concentrated around a spherical shell.

2012 ACM Subject Classification Mathematics of computing → Nonparametric statistics; Mathematics of computing → Multivariate statistics; Theory of computation → Sample complexity and generalization bounds; Theory of computation → Streaming, sublinear and near linear time algorithms

Keywords and phrases High-dimensional mean estimation

Digital Object Identifier 10.4230/LIPIcs.ITCS.2022.98

Funding *Jasper C. H. Lee*: Supported in part by the generous funding of a Croucher Fellowship for Postdoctoral Research and by NSF award DMS-2023239. Part of this work was done during Jasper's visit at the Simons Institute for the Theory of Computing.
Paul Valiant: Supported in part by NSF award CCF-2127806 and IIS-1562657. Part of this work was done at the Institute for Advanced Study, partially supported by NSF award DMS-1926686, and indirectly supported by NSF award CCF-1900460.

Acknowledgements We thank Avi Wigderson for insightful discussions on geometric intuitions for high-dimensional inequalities.

1 Introduction

Estimating the mean of a distribution over $\mathbb{R}^d$ is one of the most fundamental problems in statistics. In the past decade, there has been great interest and progress in settling the sample complexity question, under the minimal (and essentially necessary, c.f. [5, 14]) assumption that the (co)variance of the underlying distribution exists, without being known to the algorithm. For the $d = 1$ case, a line of work started by Catoni [2] and ending with a very recent result by Lee and Valiant [14] settled the sample complexity to within a $1 + o(1)$ multiplicative factor. In high dimensions, Lugosi and Mendelson [17] proposed the first estimator with sample complexity optimal to within a constant factor, but the estimator is not known to be efficiently computable. Motivated by these results, Hopkins [12] started a line of work, extended by others [4, 15], using sum-of-squares and spectral methods to propose estimators with polynomial running time, but still with sample complexity that is sub-optimal by constant factors. It remains an open problem to construct a high-dimensional mean estimator with sample complexity optimal to within a $1 + o(1)$ factor.

In this work, we take a step towards resolving this open problem: we propose an estimator with $1 + o(1)$-factor optimal sample complexity, in the regime where the dimensionality of the distribution is much larger than a function of the allowed failure probability δ. Concretely, define the *effective dimension* (also known as *stable rank*) of a distribution D with covariance Σ and maximum variance $\sigma^2_{\max} \equiv \lambda_{\max}(\Sigma)$ to be $d_{\text{eff}} = \text{Tr}(\Sigma)/\sigma^2_{\max}$. Our estimator has optimal sample complexity when $d_{\text{eff}} \geq \omega(\log^2 \frac{1}{\delta})$ (see Theorem 1 for the precise statement). In other words, our algorithm is optimal whenever the effective dimension of the distribution is large compared to the order of magnitude of the desired tail bound, δ, that we seek.

The estimator has a surprisingly simple structure, and is computable in linear time: 1) re-center the samples using a preliminary mean estimate, 2) remove some number of samples farthest away from the origin and 3) return the sample mean of the remaining samples. The structure of our estimator is natural enough that several estimators with similar structure have appeared in previous work, including: the "trimmed mean" is a standard tool in 1-dimension [21], often used in the context of robust statistics, but introduces complications when generalizing to high dimensions [18]; and the Catoni-Giulini estimator [3], which is easy to compute but does not have big-O optimal sample complexity.

The key ingredient in our analysis is a novel Bernstein-type vector concentration inequality (see Section 4) applied in a natural parameter regime where, surprisingly, standard vector tail bounds give vacuous probability bounds greater than 1. Intuitively, this result takes the standard notion that "high dimensional Gaussian distributions look like spherical shells" and shows how this essentially holds true for all sums of bounded independent random vectors. The tail bound is a general statement that we hope will find many more applications outside of the mean estimation context.

Our estimator uses the standard coordinate-wise median of means estimate (Estimator 2) as the preliminary mean estimate, which, like the rest of our algorithm, is computable in linear time. The standard analysis of the error of the coordinate-wise median of means estimate actually introduced a spurious dependence on the ambient dimension d. We give a natural new analysis of this classic algorithm, eliminating this d dependence, in Section 6.

1.1 The Model and Main Result

Given a set of n i.i.d. samples $X = \{x_1, \ldots, x_n\}$ from an $\mathbb{R}^d$-valued distribution D with unknown mean μ, the task is to estimate μ to within ℓ_2 distance ϵ, failing with probability at most δ. That is, we want to find an estimator $\hat{\mu}(X)$ such that

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

The goal is to find the optimal trade-off between the three problem parameters: sample size n, ℓ_2 error ϵ and failure probability δ. Fixing any two parameters and optimizing the third yields three equivalent formulations of the problem. For the rest of the paper, we will focus on the formulation of optimizing the error ϵ as a function of sample size n and failure probability δ. In this formulation the error $\epsilon(n, \delta)$, as a function of n and δ, is sometimes also known as the *statistical rate* of an estimator, and the statistical rate minimized over all estimators is known as the *minimax rate* of the problem.

Even in the case when the distribution D is a multivariate Gaussian, the minimax rate ϵ is lower bounded by $\Omega\left(\sqrt{\frac{\text{Tr}(\Sigma)}{n}}\right)$, even for a constant failure probability δ – this is a folklore result. Since this paper is emphasizing for the first time the *exact* multiplicative factor in statistical rate, we show in Section 5 a tight minimax rate lower bound of $\sqrt{\frac{\text{Tr}(\Sigma)}{n}}$. The intuition behind this quantity is that $\frac{\text{Tr}(\Sigma)}{n}$ is exactly the expected squared ℓ_2 error of the sample mean for *any* distribution with finite covariance.

Our main result shows that, in the regime where the effective dimension d_{eff} is very high, namely bigger than $\text{poly}\log\frac{1}{\delta}$, then it is possible to attain this optimal estimation error $\sqrt{\frac{\text{Tr}(\Sigma)}{n}}$, up to a $1+o(1)$ factor. Estimator 1 takes as input 1) a failure probability δ, 2) an approximation parameter γ (think of γ as a $o(1)$ term) and 3) n i.i.d. samples from a distribution, and outputs a mean estimate satisfying Theorem 1.

▶ **Theorem 1.** *Let $x_1,\ldots,x_n \overset{i.i.d.}{\leftarrow} D$ for some distribution D with finite but unknown mean μ and covariance Σ, let $\delta\in(0,0.01]$ be an arbitrary failure probability and let $\gamma\in(0,0.5]$ be an arbitrary approximation parameter. If $n\geq\frac{710000}{\gamma^{6.5}}\log^3\frac{20}{\delta}$ and $d_{\mathit{eff}}\geq\frac{16000}{\gamma^7}\log^2\frac{20}{\delta}$, then Estimator 1, on input $\delta,\gamma,x_1,\ldots,x_n$, will output a mean estimate $\hat{\mu}$ such that*

$$\mathop{\mathbb{P}}_{x_1,\ldots,x_n}\left(||\hat{\mu}-\mu||_2\geq(1+O(\gamma))\sqrt{\frac{\text{Tr}(\Sigma)}{n}}\right)\leq\delta$$

No effort has been made to optimize the numerical constants in the above theorem and in Estimator 1.

Beyond the question of statistical rate, previous work draws a distinction between a *single-δ* and *multiple-δ* estimator, that is, whether the estimator takes δ as an input [5]. The goal for a multiple-δ estimator is to *simultaneously* guarantee the optimality of the mean estimate for a wide range of δ. Devroye et al. [5], in the 1-dimensional setting, demonstrated how to construct a multiple-δ estimator with the correct constant in the statistical rate, as long as the kurtosis (normalized fourth moment) exists and that the algorithm knows an upper bound on the kurtosis. As a straightforward corollary of Theorem 1, surprisingly, we can get a result of a similar flavor essentially for free (Corollary 2): if we have a *lower bound* $\hat{d}_{\text{eff}}$ on the effective dimension, we can calculate a small $\hat{\delta}$ value from $\hat{d}_{\text{eff}}$ and use it as input to the main estimator to get a multiple-δ estimator, with optimal concentration guarantees across the entire range of δ satisfying $\hat{d}_{\text{eff}}\geq\text{poly}\log\frac{1}{\delta}$.

▶ **Corollary 2.** *Let $x_1,\ldots,x_n \overset{i.i.d.}{\leftarrow} D$ for some distribution D with finite but unknown mean μ and covariance Σ, let $\gamma\in(0,0.5]$ be an arbitrary approximation parameter, and suppose $\hat{d}_{\mathit{eff}}$ is a known lower bound of d_{eff}. If $n\geq\frac{\gamma^4}{3}\hat{d}_{\mathit{eff}}^{\frac{3}{2}}$, then Estimator 1, on input γ, $\delta=20e^{-\gamma^{3.5}\cdot\sqrt{\frac{\hat{d}_{\mathit{eff}}}{16000}}}$ and samples $x_1,\ldots,x_n$, will output a mean estimate $\hat{\mu}$ such that, for all $\delta\in[20e^{-\gamma^{3.5}\cdot\sqrt{\frac{\hat{d}_{\mathit{eff}}}{16000}}},1)$, we have*

$$\mathop{\mathbb{P}}_{x_1,\ldots,x_n}\left(||\hat{\mu}-\mu||_2\geq(1+O(\gamma))\sqrt{\frac{\text{Tr}(\Sigma)}{n}}\right)\leq\delta$$

1.2 Our Approach: a Vector Bernstein-Type Concentration Bound

The main tool we use in the analysis, and indeed the motivation for the simple algorithm we propose, is a novel vector tail bound (Theorem 3). The bound is a Bernstein-type bound, in the sense that it utilizes both covariance information as well as an absolute bound on the support of the distribution. See Section 4 for full details, including the explicit polynomial terms in the probability bound.

▶ **Theorem 3.** *Consider a distribution $S=\sum_{i=1}^n Y_i$ in d dimensions, where each Y_i is an independent random variable, of mean 0, supported on the radius r ball. Let Σ_S be the covariance of S, $\sigma^2_{\max}$ be the maximum covariance $\lambda_{\max}(\Sigma_S)$ and d_{eff} be the effective dimension $\text{Tr}(\Sigma_S)/\sigma^2_{\max}$. Then, the probability $\mathbb{P}_{y\leftarrow S}(||y||\geq(1+\gamma)\sqrt{\text{Tr}(\Sigma_S)})$ is bounded by*

- *for* $\gamma \le 1$*:*

$$\sqrt{\pi}\left(1+\frac{3}{\sqrt{2}}\sqrt{\min\left(d_{\mathit{eff}},\frac{\sqrt{\mathrm{Tr}(\Sigma_S)}}{r}\right)}\right)\left(2+\frac{\mathrm{Tr}(\Sigma_S)}{r^2}\right)e^{-\frac{2\gamma^2}{\frac{2}{d_{\mathit{eff}}}+\frac{r}{\sqrt{\mathrm{Tr}(\Sigma_S)}}}\frac{0.7+0.15\gamma}{0.7+\gamma}}$$

- *for* $\gamma > 1$*:*

$$\sqrt{\pi}\left(1+\gamma\sqrt{d_{\mathit{eff}}}\right)\left(2+\frac{\mathrm{Tr}(\Sigma_S)}{r^2}\right)e^{-\min\left(\frac{1}{3}\gamma^2 d_{\mathit{eff}},\,0.35\gamma\frac{\sqrt{\mathrm{Tr}(\Sigma_S)}}{r}\right)}$$

As intuition, recall the standard notion that "high dimensional (spherical) Gaussian distributions look like spherical shells", specifically, that a sample from a Gaussian of covariance Σ is extremely likely to have distance from its mean of almost exactly $\sqrt{\mathrm{Tr}(\Sigma)}$, with concentration getting sharper in higher dimensions. Our new tail bound extends this result to *arbitrary* bounded distributions S, with arbitrary covariance Σ_S, bounding the probability that a sample differs from its mean by more than $(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma_S)}$ across our choice of γ. Specifically, the first term of the $\gamma \le 1$ case in Theorem 3 is essentially the probability $e^{-\gamma^2\frac{2\sqrt{\mathrm{Tr}(\Sigma_S)}}{r}}$, describing how the tail probability is exponential in (negative) the ratio between the root mean square length of a sample, $\sqrt{\mathrm{Tr}(\Sigma_S)}$, and the bound on the support radius r; and the probability decays exponentially in the square of the factor γ^2. Each of these dependencies is essentially tight in certain regimes. The second term of the min describes how, when the support bound r becomes tiny, these bounds blend into the corresponding tail bounds on Gaussian distributions, with the exponent now proportional to the effective dimension, $d_{\mathrm{eff}} = \frac{\mathrm{Tr}(\Sigma_S)}{\sigma^2_{\max}}$.

2 Related Work

The mean estimation problem has attracted much attention recently. Perhaps surprisingly, some basic questions were only resolved in the past few years, and others still remain open. Here, we briefly outline the line of work that this paper fits in. A survey by Lugosi and Mendelson [16] gives a comprehensive and technically in-depth overview of the state of the art circa 2019.

In 1-dimension, the sample complexity and minimax rate question is mostly settled, under the sole assumption that the underlying distribution has finite but unknown mean and variance. For decades, the median-of-means algorithm [13, 20, 1] has been known to have a minimax rate that is within a constant factor of optimal. Catoni [2], in a seminal work, demonstrated how to get the statistical rate to within a $1+o(1)$ factor of optimal, when the variance of the underlying distribution is known to a $1+o(1)$ factor or when the 4^{th} moment of the distribution is finite and bounded (in which case the variance can be estimated to high accuracy). Devroye et al. [5] showed, under the same 4^{th} moment assumption, how to construct a multiple-δ estimator – an estimator which does not take the failure probability δ as an input – with convergence also tight to within $1+o(1)$ simultaneously for a wide range of δ, despite the estimator not taking the parameter as an input. Very recently, Lee and Valiant [14] settled the 1-dimensional mean estimation question by constructing an efficiently computable estimator with the optimal convergence (up to a $1+o(1)$ factor) without requiring any knowledge of the variance or finiteness of any additional moments beyond the variance. This estimator is a single-δ estimator, meaning that it does require δ as an input, but this is necessarily so because Devroye et al. [5] showed that a multiple-δ estimator cannot exist under only the finite variance assumption.

In high dimensions, under the assumption of the finiteness of the covariance, prior works have constructed estimators with rate that is optimal up to constant multiplicative factors. Lugosi and Mendelson [17] proposed the first such estimator by generalizing the median-of-means approach to high dimensions. However, their estimator is not known to be computable in polynomial time. A recent line of work [12, 4, 15], started by Hopkins [12], improved the computation time first to polynomial time, then down to quadratic time.

It is still an open problem to construct a high dimensional mean estimator with rate optimal up to a $1 + o(1)$ factor. This paper achieves such an optimal rate by focusing only on the regime where the effective dimension is large with respect to $\operatorname{poly}\log \frac{1}{\delta}$.

There has also been great interest in the mean estimation problem in the *robust statistics* setting [8, 9, 7, 11, 10], where a small fraction of the the input samples can be adversarially corrupted. Lugosi and Mendelson [18] used a high dimensional trimmed mean approach to yield optimal robustness, but the estimator is not known to be efficiently computable. Recent work by Diakonikolas et al. [11] improves on this result by giving a computationally efficient estimator achieving the same optimal robustness guarantee, and furthermore, has an optimal sub-Gaussian convergence rate up to a multiplicative constant.

3 The Estimator

In this section, we present and analyze our estimator. Our estimator is enabled and inspired by the new tail bounds of Theorem 3. Recall that optimal mean estimation for Gaussian distributions is straightforward: as shown by Corollary 6, the sample mean is the optimal estimator, and Proposition 9 shows that the error in this case is at least $\sqrt{\frac{\operatorname{Tr}(\Sigma)}{n}}$. At a high level our new tail bound, Theorem 3, reveals the surprising fact that "the mean of n samples from *any* bounded high-dimensional distribution D behaves like the sample mean of the Gaussian of same mean and covariance", which suggests the following intuition to emulate Gaussian performance for any distribution: 1) somehow make D bounded without modifying its mean too much; and then 2) return the sample mean.

The most natural approach to transform a distribution D into a distribution of bounded support is to choose a center, and truncate the distribution beyond some radius. We choose a center by running a preliminary simple mean estimation algorithm. Choosing the radius of truncation must be done with some care, since picking too large a radius will make the tail bounds of Theorem 3 too weak, while picking too small a radius will truncate too many elements and thus shift the mean too much. We also point out that there are many natural other ways to truncate or trim the distribution that have appeared in the literature with similar goals, though which do not achieve our results (see for example the Catoni-Giulini estimator [3] and the multidimensional trimmed mean estimator [18]).

We point out that one of the main philosophical reasons for the $1 + o(1)$ tightness of our result comes from the intuition that "a high dimensional Gaussian distribution looks like a spherical shell". A standard bound for the radius of the δ-probability tail of a Gaussian of covariance Σ (see, e.g., Equation 1.1 in [18]) is $\sqrt{\operatorname{Tr}(\Sigma)} + \sqrt{2\sigma_{\max}^2 \log \frac{1}{\delta}}$, which we interpret as saying that the radius of the δ-probability tail remains *unchanged* up to $1 + o(1)$ factor for any δ ranging from a constant down to $e^{-o(d_{\text{eff}})} = e^{-o\left(\frac{\operatorname{Tr}(\Sigma)}{\sigma_{\max}^2}\right)}$. Theorem 3 essentially extends this intuition to all bounded distributions. This fact is crucially helpful for attaining the $1 + o(1)$ factor tightness in our error bounds, since it means that even if our analysis gives up huge factors in the failure probability δ, the resulting error bounds will still be $1 + o(1)$-tight. Thus our new tail bound, Theorem 3 enables a $1 + o(1)$-factor tight analysis, without requiring any miraculously tight analysis of the algorithm itself.

Our estimator uses the coordinate-wise median-of-means estimator as a preliminary mean estimate, which we analyze in Section 6. In particular, we note that previous analyses of the coordinate-wise median-of-means estimator were not optimal, introducing a spurious dependence on the ambient dimension d, and Section 6 shows a tighter analysis.

We now present the estimator.

Algorithm 1 Optimal Very High Dimensional Mean Estimator.

Inputs:

- n independent samples $\{x_i\}$ from the unknown underlying distribution D over $\mathbb{R}^d$ (guaranteed to have finite but unknown covariance Σ)
- Confidence parameter $\delta \in (0, 0.01]$, approximation parameter $\gamma \in (0, 0.5]$

1. Compute a preliminary mean estimate κ using the coordinate-wise median of means estimator (Estimator 2 in Section 6) on a γ fraction of the samples.
2. Let $t = \frac{2000}{\gamma^5} \log^2 \frac{20}{\delta}$. Compute $\mathbb{1}_t^\kappa(i)$, which is 0 if x_i is one of the t farthest samples from κ (breaking ties arbitrarily) or if x_i was part of the γ fraction used in the previous step, and 1 otherwise.
3. Return $\hat{\mu} = \kappa + \frac{1}{n} \sum_i (x_i - \kappa) \mathbb{1}_t^\kappa(i)$.

To analyze Estimator 1, we need 1) the concentration guarantee for the coordinate-wise median-of-means estimator shown in Section 6 and 2) the vector tail bound in Section 4. We state both results here.

▶ **Proposition 4.** *On input n samples from a d-dimensional distribution D of unknown mean and covariance μ and Σ respectively, and a probability bound $\delta \le e^{-2}$, Estimator 2 outputs a mean estimate $\hat{\mu}$ such that*

$$\mathbb{P}\left(||\hat{\mu} - \mu|| \le \sqrt{\left(60 + 24 \log \frac{1}{\delta}\right) \frac{\mathrm{Tr}(\Sigma)}{n}} \right) \ge 1 - \delta$$

▶ **Theorem 3.** *Consider a distribution $S = \sum_{i=1}^n Y_i$ in d dimensions, where each Y_i is an independent random variable, of mean 0, supported on the radius r ball. Let Σ_S be the covariance of S, $\sigma^2_{\max}$ be the maximum covariance $\lambda_{\max}(\Sigma_S)$ and d_{eff} be the effective dimension $\mathrm{Tr}(\Sigma_S)/\sigma^2_{\max}$. Then, the probability $\mathbb{P}_{y \leftarrow S}(||y|| \ge (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma_S)})$ is bounded by*

- *for $\gamma \le 1$:*

$$\sqrt{\pi}\left(1 + \frac{3}{\sqrt{2}} \sqrt{\min\left(d_{eff}, \frac{\sqrt{\mathrm{Tr}(\Sigma_S)}}{r}\right)}\right)\left(2 + \frac{\mathrm{Tr}(\Sigma_S)}{r^2}\right) e^{-\frac{2\gamma^2}{\frac{2}{d_{eff}} + \frac{r}{\sqrt{\mathrm{Tr}(\Sigma_S)}}} \frac{0.7 + 0.15\gamma}{0.7 + \gamma}}$$

- *for $\gamma > 1$:*

$$\sqrt{\pi}\left(1 + \gamma\sqrt{d_{eff}}\right)\left(2 + \frac{\mathrm{Tr}(\Sigma_S)}{r^2}\right) e^{-\min\left(\frac{1}{3}\gamma^2 d_{eff}, 0.35\gamma \frac{\sqrt{\mathrm{Tr}(\Sigma_S)}}{r}\right)}$$

We are now ready to prove Theorem 1, the main result of this paper, describing the concentration of Estimator 1.

▶ **Theorem 1.** *Let $x_1, \ldots, x_n \overset{i.i.d.}{\leftarrow} D$ for some distribution D with finite but unknown mean μ and covariance Σ, let $\delta \in (0, 0.01]$ be an arbitrary failure probability and let $\gamma \in (0, 0.5]$ be an arbitrary approximation parameter. If $n \ge \frac{710000}{\gamma^{6.5}} \log^3 \frac{20}{\delta}$ and $d_{eff} \ge \frac{16000}{\gamma^7} \log^2 \frac{20}{\delta}$, then Estimator 1, on input $\delta, \gamma, x_1, \ldots, x_n$, will output a mean estimate $\hat{\mu}$ such that*

$$\mathbb{P}_{x_1,\ldots,x_n}\left(||\hat{\mu} - \mu||_2 \ge (1 + O(\gamma)) \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} \right) \le \delta$$

Proof. After "recentering", subtracting the mean computed in Step 1 of Estimator 1, Proposition 4 guarantees that, except with $\frac{\delta}{4}$ probability we have $(1-\gamma)n$ samples from a distribution whose mean μ is within $\sqrt{(60+24\log\frac{4}{\delta})\frac{\mathrm{Tr}(\Sigma)}{\gamma n}} \le \sqrt{40\log\frac{4}{\delta}\frac{\mathrm{Tr}(\Sigma)}{\gamma n}}$ of the origin.

Consider the distribution of distance from the origin of these samples. The algorithm will throw out the farthest t samples, which is equivalent to finding the radius r of the t^{th} farthest sample, and then throwing out all samples of radius $\ge r$ (assuming without loss of generality there are no ties). If we let $r_{t/n}$ denote the radius such that a $\frac{t}{n}$ fraction of the probability mass of D has larger radius, then the probability that $r > r_{\frac{t}{2n}}$ is at most $\mathbb{P}(\mathrm{Bin}((1-\gamma)n), \frac{t}{2n}) \ge t)$. By the standard multiplicative Chernoff bound, we upper bound the probability by $(e/4)^{(1-\gamma)t/2}$. Similarly, we upper bound the probability that $r < r_{\frac{2t}{n}}$ by $(e^{-1/4}/(3/4)^{3/4})^{2(1-\gamma)t}$. For $\gamma \le 0.5$ and for our choice of t, both probabilities are upper bounded by $\frac{\delta}{16}$. The rest of the proof thus conditions on $r \in [r_{\frac{2t}{n}}, r_{\frac{t}{2n}}]$.

The high-level proof strategy is to first use Theorem 3 to show that, (A) except with $\frac{3\delta}{4}$ probability, the algorithm's output is at most distance $(1+O(\gamma))\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}$ from the mean of the distribution truncated at some radius $\tilde{r} \in [r_{\frac{2t}{n}}, r_{\frac{t}{2n}}]$ around the origin. Then, we complete the proof by observing that (B) the mean of the truncated distribution cannot be more than $\gamma\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}$ from the true mean for any $\tilde{r} \in [r_{\frac{2t}{n}}, r_{\frac{t}{2n}}]$.

For part (A), we begin by considering a mesh of $m = 9et+1$ radii, corresponding to evenly spaced probabilities in the interval $[\frac{t}{2n}, \frac{2t}{n}]$ (ignoring immaterial issues of ties, which can be broken arbitrarily). Define μ_i to be the mean of the distribution truncated at radius r_i from the origin, for each r_i in the mesh. We will show that, (A.1) except with $\frac{\delta}{2}$ probability, for all r_i in the mesh, the empirical mean of the samples truncated at radius r_i is within distance $(1+\gamma)\sqrt{\frac{\mathrm{Tr}(\Sigma)}{(1-\gamma)n}}$ of μ_i. Assuming (A.1), since the algorithm computes radius $r \in [r_{\frac{2t}{n}}, r_{\frac{t}{2n}}]$ and truncates at radius r, we consider the largest mesh point $r_i \le r$, and show that (A.2) except with probability $\frac{\delta}{8}$, there will be at most $\gamma\sqrt{t}$ many samples between distance r_i and r from the origin. This in turn lets us show that (A.3) the algorithm's output is at most $2\gamma\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}$ from the empirical mean of the samples truncated at radius r_i. Combining (A.1) and (A.3) means that except with probability $\frac{3\delta}{4}$, the algorithm's output is within distance $(1+O(\gamma))\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}$ from μ_i, concluding (A).

Showing (A.1) is a relatively straightforward application of Theorem 3 and the union bound, although requiring some calculations before the theorem can be applied. In order to use the theorem, we need to upper bound the distance between μ_i and any support element within radius r_i of the origin. By the triangle inequality, this distance is upper bounded by the sum of $||\mu_i - \mu||$, $||\mu||$ and $r_i \le r_{\frac{t}{2n}}$, which we bound separately. We already know from the very beginning of the proof that $||\mu|| \le \sqrt{40\log\frac{4}{\delta}\frac{\mathrm{Tr}(\Sigma)}{\gamma n}}$.

To (A.1.1) bound $||\mu_i - \mu||$, note that μ_i is the mean of a distribution after truncating some probability mass $p_i \le \frac{2t}{n}$ farthest away from the origin. Consider projecting the distribution D in the direction of $\mu_i - \mu$. The mean of this 1-dimensional distribution is at most $||\mu||$ from the origin. The variance of this 1-dimensional distribution in is at most $\sigma^2_{\max} = \frac{\mathrm{Tr}(\Sigma)}{d_{\text{eff}}}$. We observe that, for any 1-dimensional distribution with mean a and variance b, the contribution to the mean by any portion of the distribution with mass at most p is at most $ap + \sqrt{bp}$: consider q_y being probability mass at point y, with $\sum_y q_y = p$, and by the variance assumption, $\sum_y q_y(y-a)^2 \le b$. Then $|\sum_y q_y(y-a)| \le \sqrt{\sum_y q_y(y-a)^2}\sqrt{\sum_y q_y 1^2} \le \sqrt{bp}$

by Cauchy-Schwarz. Substituting the bounds $a \le \sqrt{40 \log \frac{4}{\delta} \frac{\mathrm{Tr}(\Sigma)}{\gamma n}}$, $b \le \frac{\mathrm{Tr}(\Sigma)}{d_{\mathrm{eff}}}$ and $p \le \frac{2t}{n}$, we have $||\mu_i - \mu|| \le ap + \sqrt{bp} \le \frac{2t}{n}\sqrt{40 \log \frac{4}{\delta} \frac{\mathrm{Tr}(\Sigma)}{\gamma n}} + \sqrt{\frac{2t \cdot \mathrm{Tr}(\Sigma)}{d_{\mathrm{eff}} n}} \le \gamma \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}$ since $t \le \frac{\gamma\sqrt{\gamma} n}{4\sqrt{40 \log \frac{4}{\delta}}}$ and $t \le \frac{\gamma^2}{8} d_{\mathrm{eff}}$.

To (A.1.2) bound $r_{\frac{t}{2n}}$, we observe that probability mass p at distance $\ge r$ from the origin is at least distance $r - |\mu|$ from its mean and thus contributes $\ge p(r - |\mu|)^2$ to $\mathrm{Tr}(\Sigma)$. Since by definition there is $\frac{t}{2n}$ probability mass at distance $\ge r_{\frac{t}{2n}}$ from the origin, then, letting $p = \frac{t}{2n}$ we have $\frac{t}{2n}(r_{\frac{t}{2n}} - |\mu|)^2 \le \mathrm{Tr}(\Sigma)$ and thus $r_{\frac{t}{2n}} \le |\mu| + \sqrt{\frac{2n}{t}\mathrm{Tr}(\Sigma)}$. We have $|\mu| \le \sqrt{40 \log \frac{4}{\delta} \frac{\mathrm{Tr}(\Sigma)}{\gamma n}} \le \frac{1}{5}\sqrt{\frac{n}{t}\mathrm{Tr}(\Sigma)}$ since $1000\, t \log \frac{4}{\delta} \le n$ and $\gamma n \ge 1$ by our choice of t in Estimator 1 and the theorem assumptions. Thus $r_{\frac{t}{2n}} \le 1.7\sqrt{\frac{n}{t}\mathrm{Tr}(\Sigma)}$.

Summarizing, to apply Theorem 3 on the distribution truncated at radius r_i from the origin, we use the upper bound $\gamma\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} + \sqrt{40 \log \frac{4}{\delta} \frac{\mathrm{Tr}(\Sigma)}{\gamma n}} + 1.7\sqrt{\frac{n}{t}\mathrm{Tr}(\Sigma)} \le 2\sqrt{\frac{n}{t}\mathrm{Tr}(\Sigma)}$ as "r" in the theorem. Thus, Theorem 3 recentered at μ_i and with $\Sigma_S = n\Sigma$ yields that the empirical mean of $(1-\gamma)n$ samples from this truncated distribution will be within distance $(1+\gamma)\sqrt{\frac{\mathrm{Tr}(\Sigma)}{(1-\gamma)n}} = (1 + O(\gamma))\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}$ from μ_i, except with probability $\sqrt{\pi}\left(1 + \frac{3}{\sqrt{2}}\sqrt{\min\left(d_{\mathrm{eff}}, \frac{\sqrt{t}}{2}\right)}\right)\left(2 + \frac{t}{4}\right) e^{-\frac{\gamma^2}{\frac{1}{d_{\mathrm{eff}}} + \frac{1}{\sqrt{t}}} \frac{0.7+0.15\gamma}{0.7+\gamma}} \le \sqrt{\pi}\left(1 + \frac{3}{2}t^{\frac{1}{4}}\right)\left(2 + \frac{t}{4}\right) e^{-\frac{\gamma^2}{\frac{1}{d_{\mathrm{eff}}} + \frac{1}{\sqrt{t}}} \frac{0.7+0.15\gamma}{0.7+\gamma}}$. A union bound over all $m = 9et + 1 \le 10et$ mesh points yields a failure probability (for A.1) of $10et \cdot \sqrt{\pi}\left(1 + \frac{3}{2}t^{\frac{1}{4}}\right)\left(2 + \frac{t}{4}\right) e^{-\frac{\gamma^2}{\frac{1}{d_{\mathrm{eff}}} + \frac{1}{\sqrt{t}}} \frac{0.7+0.15\gamma}{0.7+\gamma}}$. Noting that $d_{\mathrm{eff}} \ge \sqrt{t}$, $t \ge 10$ and that $\frac{0.7+0.15\gamma}{0.7+\gamma} \ge \frac{1}{2}$, we upper bound the above failure probability by $10t^3 \cdot e^{-\frac{\gamma^2\sqrt{t}}{4}}$. Since $t = \frac{2000}{\gamma^5}\log^2 \frac{20}{\delta}$, we have $10t^3 \cdot e^{-\frac{\gamma^2\sqrt{t}}{4}} = 10 \cdot \frac{2000^3}{\gamma^{15}} \log^6 \frac{20}{\delta} e^{-\frac{\sqrt{2000}}{4\sqrt{\gamma}} \log \frac{20}{\delta}} \le \frac{\delta}{2}$. Thus the A.1 failure probability is at most $\frac{\delta}{2}$.

Moving on to (A.2), we show a slightly stronger statement that except with probability $\frac{\delta}{8}$, for all r_i, r_{i+1} in the mesh, the number of samples with distance $[r_i, r_{i+1})$ from the origin is at most $\gamma\sqrt{t}$. By construction, there is at most $\frac{1}{m-1} \cdot \frac{3t}{2n}$ probability mass between radii r_i and r_{i+1}. Thus, the A.2 failure probability is upper bounded by $(m-1)\,\mathbb{P}\left(\mathrm{Bin}(n, \frac{1}{m-1} \cdot \frac{3t}{2n}) > \gamma\sqrt{t}\right)$. Recalling that $m - 1 = 9et$, using our choice of t and that $\gamma\sqrt{t} \gg 1$, standard Chernoff bounds bound this by $9et \cdot 2^{-\gamma\sqrt{t}} \le \frac{\delta}{8}$.

To complete the proof of part (A), we note that (A.3) is just combining (A.2) with the fact that all the elements between radius r_i and r are at distance at most $r_{\frac{t}{2n}} \le 1.7\sqrt{\frac{n}{t}\mathrm{Tr}(\Sigma)} \le 2\sqrt{\frac{n}{t}\mathrm{Tr}(\Sigma)}$ from the origin.

Finally, we note that condition (B) is implied by the bound of (A.1.1), and hence we conclude the desired result. ◀

4 Vector Bernstein-Type Concentration Inequality

In this section we discuss the main tail bound, Theorem 3. See Section 1.2 for a brief introduction.

As motivation for what follows, we recall the standard 1-dimensional Bernstein bound.

▶ **Fact 1.** *Consider a distribution $S = \sum_{i=1}^{n} Y_i$ in 1 dimension where each Y_i is an independent random variable, of mean 0, supported on the radius r interval $[-r, r]$. Let σ_S^2 be the variance of S. Then*

$$\mathop{\mathbb{P}}_{y \leftarrow S}(y \ge t) \le e^{-\frac{\frac{1}{2}t^2}{\sigma_S^2 + \frac{1}{3}rt}}$$

When the radius bound r is sufficiently small ($r \approx \sigma_S$), this bound becomes essentially the standard Gaussian tail bound $e^{-\frac{t^2}{2\sigma_S^2}}$. By contrast, when the radius r becomes large enough that the term $\frac{1}{3}rt$ in the denominator of the exponent is much larger than the remaining term σ_S^2, then the tail bound becomes $e^{-\frac{3}{2}\cdot\frac{t}{r}}$. This bound is exponential in $\frac{t}{r}$, roughly capturing the tail of a Poisson distribution: if each Y_i takes value r with extremely small probability, then the number of times r is sampled in total, $\frac{S}{r}$ behaves like a Poisson distribution, and its chance of exceeding $\frac{t}{r}$ is roughly exponential in $-\frac{t}{r}$ without much dependence on the Poisson parameter, provided it is sufficiently small. In short, as r varies, the 1-dimensional Bernstein bound smoothly blends between the "Gaussian regime" with exponent proportional to $-t^2$ and the "Poisson regime" with exponent proportional to $-t$.

Our vector extension of the Bernstein bound, Theorem 3, maintains the same flavor. In Section 7, we prove the crucial $\gamma \le 1$ case, and defer the details of the similar proof of the $\gamma > 1$ case to the full paper on arXiv.

▶ **Theorem 3.** *Consider a distribution $S = \sum_{i=1}^{n} Y_i$ in d dimensions, where each Y_i is an independent random variable, of mean 0, supported on the radius r ball. Let Σ_S be the covariance of S, $\sigma_{\max}^2$ be the maximum covariance $\lambda_{\max}(\Sigma_S)$ and d_{eff} be the effective dimension $\mathrm{Tr}(\Sigma_S)/\sigma_{\max}^2$. Then, the probability $\mathbb{P}_{y\leftarrow S}(||y|| \ge (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma_S)})$ is bounded by*

- *for $\gamma \le 1$:*

$$\sqrt{\pi}\left(1+\frac{3}{\sqrt{2}}\sqrt{\min\left(d_{\mathit{eff}}, \frac{\sqrt{\mathrm{Tr}(\Sigma_S)}}{r}\right)}\right)\left(2+\frac{\mathrm{Tr}(\Sigma_S)}{r^2}\right)e^{-\frac{2\gamma^2}{\frac{2}{d_{\mathit{eff}}}+\frac{r}{\sqrt{\mathrm{Tr}(\Sigma_S)}}}\frac{0.7+0.15\gamma}{0.7+\gamma}}$$

- *for $\gamma > 1$:*

$$\sqrt{\pi}\left(1+\gamma\sqrt{d_{\mathit{eff}}}\right)\left(2+\frac{\mathrm{Tr}(\Sigma_S)}{r^2}\right)e^{-\min\left(\frac{1}{3}\gamma^2 d_{\mathit{eff}}, 0.35\gamma\frac{\sqrt{\mathrm{Tr}(\Sigma_S)}}{r}\right)}$$

In particular, using the notation of the theorem, for $\gamma = 1$, we point out that when the radius bound r is small, the exponent of the bound is $-d_{\mathrm{eff}}$, which matches up to constant factor the exponent in standard Gaussian tail bounds. (For example, the above-mentioned bound on the radius of the δ-probability tail, from Equation 1.1 in [18], is $\sqrt{\mathrm{Tr}(\Sigma_S)} + \sqrt{2\sigma_{\max}^2 \log\frac{1}{\delta}}$; and substituting $\delta = e^{-d_{\mathrm{eff}}}$ yields the desired radius up to a constant factor.)

Meanwhile, again for the $\gamma = 1$ case, as $r \to \infty$, the exponent in the bound of Theorem 3 is $-\frac{2\sqrt{\mathrm{Tr}(\Sigma_S)}}{r}$, which is proportional to the ratio of the deviation – $(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma_S)}$ – to the radius bound r, exactly as in the 1-dimensional case. This captures the fact that a distribution S with significant probability mass at a single point at radius r will have a Poisson-like tail bound from the contribution of this single support point, justifying an exponent proportional to the ratio of the desired deviation to the radius r, as in the 1-dimensional case.

Generally, for the $\gamma \le 1$ case, the exponent of the bound smoothly interpolates between the two regimes $-\gamma^2 d_{\mathrm{eff}}$ and $-2\frac{\gamma^2\sqrt{\mathrm{Tr}(\Sigma_S)}}{r}$, depending on which is larger; the multiplicative term $\frac{0.7+0.15\gamma}{0.7+\gamma}$ in the exponent monotonically varies from 1 to $\frac{1}{2}$ as γ goes from 0 to 1, and thus for small γ can be essentially ignored.

Finally, we point out that the exponents are proportional to γ^2 for $\gamma \le 1$, which matches the Gaussian case – as may be verified by solving for δ as a function of the above tail bound $\sqrt{\mathrm{Tr}(\Sigma_S)} + \sqrt{2\sigma_{\max}^2 \log\frac{1}{\delta}}$.

Thus the bounds of Theorem 3 are tight in several natural regimes. For the purposes of this paper only the $\gamma \leq 1$ regime is relevant, though the results extend naturally to the $\gamma > 1$ case.

We briefly point out that all previous results we are aware of are essentially vacuous for small γ, and even for sufficiently small *constant* γ. The most influential result of this nature is the matrix Bernstein bound of Tropp [22], which, when restricted to vectors in our setting gives

$$\mathbb{P}_{y \leftarrow S}(||y|| \geq t) \leq (d+1) \cdot e^{-\frac{\frac{1}{2}t^2}{\mathrm{Tr}(\Sigma_S)+\frac{1}{3}rt}}$$

Substituting in $t = (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma_S)}$ yields

$$\mathbb{P}_{y \leftarrow S}(||y|| \geq (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma_S)}) \leq (d+1) \cdot e^{-\frac{\frac{1}{2}(1+\gamma)^2\mathrm{Tr}(\Sigma_S)}{\mathrm{Tr}(\Sigma_S)+\frac{1}{3}r(1+\gamma)\sqrt{\mathrm{Tr}(\Sigma_S)}}}$$

where the above bound is at least $(d+1) \cdot e^{-\frac{\frac{1}{2}(1+\gamma)^2\mathrm{Tr}(\Sigma_S)}{\mathrm{Tr}(\Sigma_S)}} = (d+1)e^{-\frac{1}{2}(1+\gamma)^2}$, which is vacuous when $\gamma \leq 1$ and $d \geq 7$. To further differentiate this line of work from our result, we recall the intuition that Gaussians (and, now, sums of independent bounded random vectors) "are distributed very close to a spherical shell", and in particular, for any constant γ, the tail bound should decay *exponentially* with the effective dimension $d_{\text{eff}} = \frac{\mathrm{Tr}(\Sigma_S)}{\sigma^2_{\max}}$, instead of increasing linearly with the ambient dimension d. Other variants of vector Bernstein inequalities have similarly weak bounds in the constant γ regime [19, 23].

5 Minimax Rate Lower Bound

In this section, we show that no mean estimator can achieve a rate better than $\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}$ even for the well-behaved case of multivariate Gaussians with *known* covariance Σ, and even with constant failure probability. We show that for Gaussians, the sample mean is optimal, and show that the sample mean has a constant probability of having error at least $\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}$.

While similar bounds exist in the literature and folklore, most focus on the mean squared ℓ_2 error of estimators instead of our "ϵ, δ" regime where we ask for estimators that achieve ℓ_2 error ϵ with probability $1-\delta$. Since the ultimate goal of this line of research is to resolve the high probability mean estimation problem to within $1+o(1)$ factors, we derive here the corresponding lower bound for this high probability setting, with the desired tightness, which can be used as a benchmark for this and for future work.

We start with an observation about the indistinguishability of combinations of Gaussian random variables.

▶ **Fact 2.** *Let $X \sim \mathcal{N}(0, t^2 I)$, and $Y, Z \sim \mathcal{N}(0, \Sigma)$. Then $d_{\mathrm{TV}}((X+Y, Y), (X+Y, Z)) = O_\Sigma(\frac{1}{t})$. Here, $O_\Sigma(\frac{1}{t})$ means that the term is $O(\frac{1}{t})$ for a fixed Σ.*

Since the ordered pairs $(X+Y, Y)$ and $(X+Y, Z)$ are themselves multivariate Gaussians, this result easily follows from bounds on the total variation distance of multivariate Gaussians, such as [6].

Using Fact 2, we show that the optimal estimator using a *single* Gaussian sample is just the sample itself.

▶ **Lemma 5.** *Given a covariance matrix Σ in d dimensions, and a radius $r > 0$, consider the problem of estimating a vector μ to within distance r with as high probability as possible, given 1 sample from the Gaussian distribution $x \leftarrow \mathcal{N}(\mu, \Sigma)$. We claim that the minimax risk is attained by the estimator that simply returns the sample x.*

Proof. Consider an arbitrary estimator that returns $f(x)$ on input a sample x. Fix a large radius $t > 0$. Choosing $\mu \leftarrow \mathcal{N}(0, t^2 I)$ and sampling $x \leftarrow \mathcal{N}(\mu, \Sigma)$, Fact 2 implies that the joint distribution of $x, x - \mu$ is within $O_\Sigma(\frac{1}{t})$ of the joint distribution of $\mathcal{N}(0, \Sigma + t^2 I), \mathcal{N}(0, \Sigma)$; thus for the remainder of the proof we instead take $x, \mu - x$ to be drawn independently from these latter two distributions, and add $O_\Sigma(\frac{1}{t})$ to the resulting probability bounds.

We wish to bound the probability that the estimator is accurate to within radius r, namely, the probability that $||f(x) - \mu|| \le r$, when x, μ are sampled as above. Fixing x, by the previous paragraph the posterior distribution of μ is $\mathcal{N}(x, \Sigma)$. We claim that the probability that μ is in the ball $B(f(x), r)$ is maximized when $f(x) = x$: in 1 dimension this is the trivial claim that, for a 1-dimensional Gaussian distribution centered at $y \in \mathbb{R}$, the interval of radius r with maximum probability is the interval $[y - r, y + r]$; in d dimensions, diagonalizing so that the Gaussian can be represented as independent Gaussians along each dimension, we see that shifting a ball of radius r, coordinate-by-coordinate, until it is centered at x will monotonically increase the probability, since the intersection of a ball with any line parallel to the i^{th} axis will be an interval, whose contribution will increase if we center the interval at i^{th} coordinate x_i.

Thus for each x, $\mathbb{P}_{\mu \leftarrow \mathcal{N}(x,\Sigma)} (||\mu - f(x)|| \le r) \ge \mathbb{P}_{\mu \leftarrow \mathcal{N}(x,\Sigma)} (||\mu - x|| \le r)$. And thus, averaged over $x \leftarrow \mathcal{N}(0, \Sigma + t^2 I)$ the same relation holds.

From the total variation bound, we thus have that, for all estimators f,

$$\mathop{\mathbb{P}}_{\substack{\mu \leftarrow \mathcal{N}(0, t^2 I) \\ x \leftarrow \mathcal{N}(\mu, \Sigma)}} (||\mu - f(x)|| \le r) \ge \mathop{\mathbb{P}}_{\substack{\mu \leftarrow \mathcal{N}(0, t^2 I) \\ x \leftarrow \mathcal{N}(\mu, \Sigma)}} (||\mu - x|| \le r) - O_\Sigma(\frac{1}{t})$$

Thus there is no estimator that beats the trivial estimator when averaged over μ drawn from a sufficiently large Gaussian, or by extension, when μ is drawn from any sufficiently smooth distribution; consequently, the minimax risk is exactly that of the trivial estimator that returns x. ◀

Combining the previous lemma with the standard fact that the sample mean is a sufficient statistic for i.i.d multivariate Gaussian samples (Fact 3, which has a straightforward proof from the Fisher-Neyman factorization theorem), we show that the sample mean is an optimal mean estimator for Gaussians.

▶ **Fact 3.** *If $Y_1, \ldots, Y_n$ are independent samples from a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ in $\mathbb{R}^d$, for fixed Σ, then $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n Y_i$ is a sufficient statistic for the mean, μ.*

▶ **Corollary 6.** *Given a covariance matrix Σ in d dimensions, and a radius $r > 0$, consider the problem of estimating a vector μ to within distance r with as high probability as possible, given n samples from the Gaussian distribution $x_1, \ldots, x_n \leftarrow \mathcal{N}(\mu, \Sigma)$. In this situation, the minimax risk is attained by the estimator that simply returns the sample mean $\frac{1}{n} \sum_{i=1}^n x_i$.*

Proof. By Fact 3, the sample mean is a sufficient statistic for i.i.d. Gaussian samples. The sample mean is distributed as a single sample from $\mathcal{N}(\mu, \frac{\Sigma}{n})$ and thus we apply Lemma 5 to yield the result. ◀

We point out that the above results essentially show that one cannot beat the sample mean estimator, *even if one knows the covariance matrix*, Σ; clearly this also implies that one cannot beat the sample mean when one does *not* know Σ.

Lastly, we lower bound the statistical rate of the sample mean for estimating Gaussian means (Lemma 8). We show the lower bound via analyzing the squared ℓ_2 error of the sample mean, which is a 1-dimensional distribution that is a sum of scaled χ^2 independent random variables, showing that there is at least constant probability in exceeding its mean.

To prove Lemma 8, we use the fact that the χ^2 distribution has a finite 4^{th} moment, and apply the following anti-concentration lemma (Lemma 7).

▶ **Lemma 7.** *A 1-dimensional distribution D with variance v and 4^{th} (central) moment m has at least $(2\sqrt{3}-3)\frac{v^2}{m}$ fraction of its probability mass on both sides of its mean.*

Proof. Without loss of generality, assume the mean is 0. Let D^+ and D^- respectively denote the distribution restricted to support respectively above and below 0 (without normalizing, so they are interpreted as measures, not distributions). And let D_p^+, D_p^- respectively denote their p^{th} (central) moments. Hölder's inequality yields $D_4^+ \geq \frac{(D_2^+)^2}{D_0^+}$, and $D_1^+ \leq \sqrt{D_0^+ D_2^+}$, and $D_4^- \geq (D_2^-)^3 (D_1^-)^{-2}$. Rearranging, using the fact that $D_1^- = -D_1^+$ (since the overall mean of D is 0), and that $D_2^- = v - D_2^+$ (since the total variance is v), we bound the total 4^{th} moment as $m = D_4^+ + D_4^- \geq \frac{1}{D_0^+}\left((D_2^+)^2 + \frac{(v-D_2^+)^3}{D_2^+}\right)$. It is straightforward to compute the minimum over D_2^+ of this expression as $(2\sqrt{3}-3)\frac{v^2}{D_0^+}$. Rearranging yields our desired bound of $D_0^+ \geq (2\sqrt{3}-3)\frac{v^2}{m}$. Symmetry yields the same bound for D_0^-. ◀

▶ **Lemma 8.** *Given n samples from a Gaussian distribution, $x_1, \ldots, x_n \leftarrow \mathcal{N}(\mu, \Sigma)$, the probability that the empirical mean has error at least $\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}$ is at least 0.03:*

$$\mathop{\mathbb{P}}_{x_1,\ldots,x_n \leftarrow \mathcal{N}(\mu,\Sigma)} \left(\left\| \mu - \frac{1}{n}\sum_{i=1}^n x_i \right\| \geq \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} \right) \geq 0.03$$

Proof. Diagonalizing Σ so that $\sigma_1^2, \ldots, \sigma_n^2$ represent the covariances along the n principal axes, the distribution of the squared error $\left\|\mu - \frac{1}{n}\sum_{i=1}^n x_i\right\|^2$ can be expressed as $\sum_{i=1}^n y_i^2 \sigma_i^2$ where $y_1, \ldots, y_n \sim \mathcal{N}(0,1)$. We compute that the random variable $\sum_{i=1}^n y_i^2\sigma_i^2$ has mean $\frac{1}{n}\sum_{i=1}^n \sigma_i^2 = \frac{\mathrm{Tr}(\Sigma)}{n}$, it has 2^{nd} moment (variance times n) of $2\sum_{i=1}^n \sigma_i^4$, and it has excess kurtosis $12(\sum_{i=1}^n \sigma_i^4)^{-2}\sum_{i=1}^n \sigma_i^8$

Thus the squared error has 4^{th} moment $12(\sum_{i=1}^n \sigma_i^4)^2 + 48\sum_{i=1}^n \sigma_i^8 \leq 60(\sum_{i=1}^n \sigma_i^4)^2$, and by Lemma 7, the probability of exceeding its mean is at least

$$\frac{(2\sqrt{3}-3)4(\sum_{i=1}^n \sigma_i^4)^2}{60(\sum_{i=1}^n \sigma_i^4)^2} = \frac{2\sqrt{3}-3}{15} \geq 0.03$$ ◀

▶ **Proposition 9.** *Given a covariance matrix Σ in d dimensions, and a probability $\delta > 0$, consider the problem of finding the best estimate to μ given n samples from the Gaussian distribution $x_1, \ldots, x_n \leftarrow \mathcal{N}(\mu, \Sigma)$, in the sense that we desire the smallest possible error radius r subject to a success probability $\geq 1-\delta$. If $\delta < 0.03$ then the error r must be at least $\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}$.*

Proof. Assume for the sake of contradiction that there is an estimator with failure probability $\delta < 0.03$ and error $r < \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}$. Corollary 6 implies that the sample mean estimator must attain this same accuracy $\leq r$ with at least as small a failure probability. However, this

contradicts Lemma 8 which says that the sample mean has error $\geq \sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}} > r$ with probability at least $0.03 > \delta$. Thus the assumption must be false and the error must be at least $\sqrt{\frac{\mathrm{Tr}(\Sigma)}{n}}$ as claimed in the proposition. ◀

6 New Analysis of Coordinate-wise Median of Means

In this section, we provide a new analysis of the coordinate-wise median of means estimator (Estimator 2), which is used as a sub-routine of the main estimator.

Algorithm 2 Coordinate-wise Median of Means.

Inputs:

- n independent samples $\{x_i\}$ from the unknown underlying distribution D over $\mathbb{R}^d$ (guaranteed to have finite but unknown covariance Σ)
- Confidence parameter $\delta \in (0, 1/e^2]$

1. Divide the samples into $t = 3\log\frac{1}{\delta}$ blocks, compute the sample mean of each block.
2. For each of the d coordinates, return the median of the means from the previous step at that particular coordinate.

The standard analysis [16] performs the 1-dimensional median-of-means analysis on each coordinate with failure probability δ/d, and takes a union bound over all the d dimensions. The resulting ℓ_2 error bound is thus $O\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)\log\frac{d}{\delta}}{n}}\right)$. We show that there is in fact no dependence on d, and it is possible to circumvent the union bound to improve the ℓ_2 error to $O\left(\sqrt{\frac{\mathrm{Tr}(\Sigma)\log\frac{1}{\delta}}{n}}\right)$ (Proposition 4). We bound the tail of the (squared) ℓ_2 error as a sum instead of only bounding the tails of the errors in the individual coordinates separately. Even though the coordinates are in general *dependent*, thus preventing us from using Chernoff-type tail bounds, here it suffices to use a generalized version of Markov's inequality (Lemma 10) that shifts the distribution before applying the standard Markov inequality. This lets us use Markov's inequality only far into the tail of the distribution where the contribution to the expectation is small, giving a small failure probability. Combined with the fact that in each coordinate, median-of-means has error with sub-Gaussian tail, this yields our tighter analysis of the coordinate-wise median-of-means estimator. Here and below, we use brackets around an inequality, as in $[a > b]$ to denote "1" if the inequality is true and "0" otherwise.

▶ **Lemma 10.** *Given random variables $x_1, \ldots, x_n$ over the reals – that are not necessarily independent, nor non-negative – and bounds $\epsilon_1, \ldots, \epsilon_n \geq 0$, define $\alpha_i = \mathbb{E}[x_i \cdot [x_i > \epsilon_i]]$. Then for any $c > 0$ we have $\mathbb{P}(\sum_{i=1}^n x_i \geq \sum_{i=1}^n \epsilon_i + \frac{\alpha_i}{c}) \leq c$.*

Proof. Let $y_i = x_i - \epsilon_i$ and $y_i^+ = \max(0, y_i)$. Noting that y_i^+ is non-negative, by Markov's inequality we have $\mathbb{P}(\sum_i^n y_i^+ \geq \frac{1}{c}\sum_i^n \mathbb{E}[y_i^+]) \leq c$. Since $y_i \leq y_i^+$, we also have $\mathbb{P}(\sum_i^n y_i \geq \frac{1}{c}\sum_i^n \mathbb{E}[y_i^+]) \leq c$. Now observe that $\mathbb{E}[y_i^+] = \mathbb{E}[y_i \cdot [y_i > 0]] = \mathbb{E}[(x_i - \epsilon_i) \cdot [x_i > \epsilon_i]] \leq \alpha_i$, since $\epsilon_i \geq 0$. Thus, $\mathbb{P}(\sum_i^n x_i \geq \sum_i^n \epsilon_i + \frac{\alpha_i}{c}) \leq \mathbb{P}(\sum_i^n y_i \geq \sum_i^n \frac{\mathbb{E}[y_i^+]}{c}) \leq c$ ◀

A standard argument involving swapping the order of integration yields the following:

▶ **Fact 4.** *Given a non-negative random variable x and a bound s, we have*

$$\mathbb{E}[x^2[x \geq s]] = s^2 \cdot \mathbb{P}(x \geq s) + \int_s^\infty 2y\,\mathbb{P}(x \geq y)\,dy$$

With the above fact, we can use the sub-Gaussian tail of median-of-means in each coordinate to bound the expectation required for applying Lemma 10.

▶ **Lemma 11.** *For a 1-dimensional distribution D with mean μ and variance σ^2, if we take n samples from D, divide them into $t \geq 6$ blocks, and let $\hat{\mu}$ be the random variable computed as the median of the t means of each block, then*

$$\mathbb{E}\left[(\hat{\mu}-\mu)^2\left[|\hat{\mu}-\mu| \geq \sqrt{\frac{8t\sigma^2}{n}}\right]\right] \leq 60\frac{\sigma^2 e^{-\frac{t}{3}}}{n}$$

Proof. Denote the error bound as $s = \sqrt{\frac{8t\sigma^2}{n}}$. The probability that the error of the mean of a single block exceeds s is bounded, by Chebyshev's inequality, by $\frac{t\sigma^2}{ns^2}$. The median cannot have error s unless at least $\frac{t}{2}$ of the blocks have error at least s, which happens with probability at most $\mathbb{P}(\mathrm{Bin}(t, \frac{t\sigma^2}{ns^2}) \geq \frac{t}{2})$. The standard Chernoff bound for this probability is $e^{-t\cdot D_{\mathrm{KL}}(\frac{1}{2}||\frac{t\sigma^2}{ns^2})}$, where $D_{\mathrm{KL}}(p||q)$ is a shorthand for the KL-divergence between Bernoulli coins of probabilities p and q, namely $D_{\mathrm{KL}}(p||q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$. Using Fact 4, the expectation in the lemma statement is thus bounded by $s^2 e^{-t\cdot D_{\mathrm{KL}}(\frac{1}{2}||\frac{t\sigma^2}{ns^2})} + \int_s^\infty 2ye^{-t\cdot D_{\mathrm{KL}}(\frac{1}{2}||\frac{t\sigma^2}{ny^2})}\,dy$.

Substituting the definition of $s = \sqrt{\frac{8t\sigma^2}{n}}$, the first term becomes $\frac{8t\sigma^2}{n}e^{-t\cdot D_{\mathrm{KL}}(\frac{1}{2}||\frac{1}{8})} \leq 37\frac{\sigma^2}{n}e^{-\frac{t}{3}}$. For the second (integral) term, we note that the 2nd argument of the KL-divergence is always at most $\frac{t\sigma^2}{ns^2} = \frac{1}{8}$, and thus we use the bound that $D_{\mathrm{KL}}(\frac{1}{2}||z) \leq -0.42 - 0.4\log z$ when $z \leq \frac{1}{8}$. Thus the integral is bounded as

$$\begin{aligned}\int_s^\infty 2ye^{-t\cdot D_{\mathrm{KL}}(\frac{1}{2}||\frac{t\sigma^2}{ny^2})}\,dy &= \frac{t\sigma^2}{n}\int_8^\infty e^{-t\cdot D_{\mathrm{KL}}(\frac{1}{2}||\frac{1}{z})}\,dz \\ &\leq \frac{t\sigma^2}{n}\int_8^\infty e^{-t\cdot(-0.42+0.4\log z)}\,dz = \frac{t\sigma^2}{n}e^{0.42t}\frac{8^{1-0.4t}}{0.4t-1}\end{aligned}$$

This final expression is easily verified to be at most $22\frac{\sigma^2}{n}e^{-\frac{t}{3}}$ when $t \geq 6$. Summing our two bounds yields the desired result. ◀

Proposition 4 combines Lemmas 10 and 11 to yield our tighter analysis of the coordinate-wise median of means estimator.

▶ **Proposition 4.** *On input n samples from a d-dimensional distribution D of unknown mean and covariance μ and Σ respectively, and a probability bound $\delta \leq e^{-2}$, Estimator 2 outputs a mean estimate $\hat{\mu}$ such that*

$$\mathbb{P}\left(||\hat{\mu}-\mu|| \leq \sqrt{\left(60 + 24\log\frac{1}{\delta}\right)\frac{\mathrm{Tr}(\Sigma)}{n}}\right) \geq 1-\delta$$

Proof. The i^{th} coordinate of samples from D will have variance $\Sigma_{i,i}$, the i^{th} diagonal entry of the covariance matrix. Towards applying Lemma 10, let random variable x_i denote the squared error of the estimator in the i^{th} coordinate, and let $\epsilon_i = \frac{8t\Sigma_{i,i}}{n}$. In the notation of Lemma 10, α_i is bounded from Lemma 11 by $60\frac{\Sigma_{i,i}}{n}e^{-\frac{t}{3}} = 60\frac{\Sigma_{i,i}}{n}\delta$. Thus, letting c in Lemma 10 equal δ, we bound the expression $\sum_{i=1}^n \epsilon_i + \frac{\alpha_i}{c} \leq \frac{8t}{n}\mathrm{Tr}(\Sigma) + \frac{60}{n}\mathrm{Tr}(\Sigma) = (60 + 24\log\frac{1}{\delta})\frac{\mathrm{Tr}(\Sigma)}{n}$, meaning that Lemma 10 yields the desired bound on the squared error, that $\mathbb{P}(\sum_{i=1}^n x_i \geq (60 + 24\log\frac{1}{\delta})\frac{\mathrm{Tr}(\Sigma)}{n}) \leq \delta$. ◀

7 Proof of Theorem 3

In this section, we prove Theorem 3, except for details of the $\gamma > 1$ case, which we include in the arXiv version of the paper.

▶ **Theorem 3.** *Consider a distribution $S = \sum_{i=1}^{n} Y_i$ in d dimensions, where each Y_i is an independent random variable, of mean 0, supported on the radius r ball. Let Σ_S be the covariance of S, $\sigma^2_{\max}$ be the maximum covariance $\lambda_{\max}(\Sigma_S)$ and d_{eff} be the effective dimension $\mathrm{Tr}(\Sigma_S)/\sigma^2_{\max}$. Then, the probability $\mathbb{P}_{y \leftarrow S}(||y|| \geq (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma_S)})$ is bounded by*

- *for $\gamma \leq 1$:*

$$\sqrt{\pi}\left(1 + \frac{3}{\sqrt{2}}\sqrt{\min\left(d_{\mathit{eff}}, \frac{\sqrt{\mathrm{Tr}(\Sigma_S)}}{r}\right)}\right)\left(2 + \frac{\mathrm{Tr}(\Sigma_S)}{r^2}\right) e^{-\frac{2\gamma^2}{\frac{2}{d_{\mathit{eff}}} + \frac{r}{\sqrt{\mathrm{Tr}(\Sigma_S)}}} \frac{0.7+0.15\gamma}{0.7+\gamma}}$$

- *for $\gamma > 1$:*

$$\sqrt{\pi}\left(1 + \gamma\sqrt{d_{\mathit{eff}}}\right)\left(2 + \frac{\mathrm{Tr}(\Sigma_S)}{r^2}\right) e^{-\min\left(\frac{1}{3}\gamma^2 d_{\mathit{eff}}, 0.35\gamma\frac{\sqrt{\mathrm{Tr}(\Sigma_S)}}{r}\right)}$$

Two of the main features of the proof, both appearing in Equation 1 in Lemma 12, are that we analyze the high-dimensional case by instead proving 1-dimensional Chernoff bounds in every direction; however, before applying the Chernoff bound, we first carefully pull out the contribution from support elements sufficiently far from the origin. Thus we apply Chernoff bounds only to elements near the origin (when projected to the current 1-dimensional line being considered), which enables robust tools such as polynomial expansions of the exponential terms, bounded by information about the first few moments of the distribution.

▶ **Lemma 12.** *Consider the setting of Theorem 3. Let $\sigma_x^2 = x^\top \Sigma_S x$, namely the variance in the direction of x scaled by $||x||^2$. Also represent each distribution Y_i as having probability $q_{i,j}$ of drawing the vector $w_{i,j}$. Then, the probability $\mathbb{P}_{y \leftarrow S}(||y|| \geq (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma_S)})$ is bounded by*

$$\min_{\beta>0,\tilde{t}>0,p\in[0,1]} \sqrt{\pi}\, e^{\frac{\beta^2 d}{2(1+\gamma)^2\mathrm{Tr}(\Sigma_S)}}\left(1 + \frac{\beta}{1+\gamma}\sqrt{\frac{d}{2\mathrm{Tr}(\Sigma_S)}}\right)$$
$$\cdot \mathop{\mathbb{E}}_{x \leftarrow \mathcal{N}(0,\frac{1}{d})^d}\left[\left(\min_{t\in[0,\tilde{t}]} e^{-\beta t + \sigma_x^2 \frac{t^2}{2}(1+\frac{p}{2})}\right) + \sum_{i,j} q_{i,j}\left[x \cdot w_{i,j} > \frac{p}{\tilde{t}}\right]\right]$$

Proof. We show the desired tail bound by analyzing a directional tail bound, with threshold $\beta > 0$ and parameters $p \in (0,1]$ and $\tilde{t}$ chosen later. In the following, let $S_{||x \leq \frac{p}{\tilde{t}}}$ denote the distribution S, but where we move to the origin all support elements whose projections to direction x exceed $\frac{p}{\tilde{t}}$.

$$\mathop{\mathbb{P}}_{y\leftarrow S}[||y|| \geq (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma_S)}] \leq \mathrm{erfc}\left(\frac{\beta\sqrt{d}}{(1+\gamma)\sqrt{2\mathrm{Tr}(\Sigma_S)}}\right)^{-1} \mathop{\mathbb{E}}_{x\leftarrow\mathcal{N}(0,\frac{1}{d})^d}\left[\mathop{\mathbb{P}}_{y\leftarrow S}(x\cdot y \geq \beta)\right]$$

$$\leq \sqrt{\pi}e^{\frac{\beta^2 d}{2(1+\gamma)^2\mathrm{Tr}(\Sigma_S)}}\left(1+\frac{\beta}{1+\gamma}\sqrt{\frac{d}{2\mathrm{Tr}(\Sigma_S)}}\right)$$

$$\cdot \mathop{\mathbb{E}}_{x\leftarrow\mathcal{N}(0,\frac{1}{d})^d}\left[\mathop{\mathbb{P}}_{y\leftarrow S_{||x\leq\frac{p}{t}}}(x\cdot y\geq\beta) + \sum_{i=1}^{n}\sum_j q_{i,j}\left[x\cdot w_{i,j} > \frac{p}{\tilde{t}}\right]\right]$$

$$\leq \sqrt{\pi}e^{\frac{\beta^2 d}{2(1+\gamma)^2\mathrm{Tr}(\Sigma_S)}}\left(1+\frac{\beta}{1+\gamma}\sqrt{\frac{d}{2\mathrm{Tr}(\Sigma_S)}}\right)$$

$$\cdot \mathop{\mathbb{E}}_{x\leftarrow\mathcal{N}(0,\frac{1}{d})^d}\left[\left(\min_{t\geq 0} e^{-\beta t}\prod_{i=1}^{n}\sum_j q_{i,j}e^{tx\cdot w_{i,j}[x\cdot w_{i,j}\leq\frac{p}{t}]}\right) + \sum_{i=1}^{n}\sum_j q_{i,j}\left[x\cdot w_{i,j} > \frac{p}{\tilde{t}}\right]\right] \quad (1)$$

where the first inequality is seen by switching the order of the expectation and probability, since $\mathrm{erfc}\left(\frac{\beta\sqrt{d}}{(1+\gamma)\sqrt{2\mathrm{Tr}(\Sigma_S)}}\right)$ is the probability that a vector y with $||y|| = (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma_S)}$ has dot product with $x \leftarrow \mathcal{N}(0,\frac{1}{d})^d$ that exceeds β; the second inequality is the bound $\mathrm{erfc}(x) \geq \frac{1}{\sqrt{\pi}}\cdot\frac{e^{-x^2}}{1+x}$ and considering moving to the origin all support elements S whose projections to direction x exceed $\frac{p}{t}$ (as the distribution $S_{||x\leq\frac{p}{t}}$), where we add up the probability masses $q_{i,j}$ of all the moved elements since these upper bound the total variation distance between $S_{||x\leq\frac{p}{t}}$ and S and hence the change in probability of exceeding β; the third inequality is a standard 1-dimensional Chernoff bound on the distribution $S_{||x\leq\frac{p}{t}}$ projected to direction x, choosing Chernoff parameter $t(x)$.

For the moment, using $c_{i,j}$ as shorthand for $tx\cdot w_{i,j}[x\cdot w_{i,j}\leq\frac{p}{t}]$, and noting that, for $t\leq\tilde{t}$ we always have $c_{i,j}\leq p\leq 1$, thus

$$\begin{aligned}
\min_{t\geq 0} e^{-\beta t}\prod_{i=1}^{n}\sum_j q_{i,j}e^{c_{i,j}} &\leq \min_{t\in[0,\tilde{t}]} e^{-\beta t}\prod_{i=1}^{n}\sum_j q_{i,j}e^{c_{i,j}} \\
&\leq \min_{t\in[0,\tilde{t}]} e^{-\beta t}\prod_{i=1}^{n}\sum_j q_{i,j}(1+c_{i,j}+\frac{1}{2}c_{i,j}^2+\frac{1}{4}\max(0,c_{i,j})^3) \\
&\leq \min_{t\in[0,\tilde{t}]} e^{-\beta t}\prod_{i=1}^{n} e^{\sum_j q_{i,j}(c_{i,j}+\frac{1}{2}c_{i,j}^2+\frac{1}{4}\max(0,c_{i,j})^3)} \\
&= \min_{t\in[0,\tilde{t}]} e^{-\beta t+\sum_{i,j} q_{i,j}(c_{i,j}+\frac{1}{2}c_{i,j}^2+\frac{1}{4}\max(0,c_{i,j})^3)}
\end{aligned}$$

Let σ_x^2 be shorthand for $x^\top\Sigma_S x$, the variance in direction x, scaled by $||x||^2$ when $||x||\neq 1$. We bound this last expression by noting that $\sum_{i,j} q_{i,j}c_{i,j}$ is at most t times the first moment of S in direction x, namely at most 0; also, $\sum_{i,j} q_{i,j}\frac{1}{2}c_{i,j}^2$ is at most $\frac{1}{2}t^2$ times the second moment of S in direction x, namely at most $\sigma_x^2\frac{t^2}{2}$; finally, $\sum_{i,j} q_{i,j}\frac{1}{4}\max(0,c_{i,j})^3$ is at most $\frac{p}{2}$ times the previous expression since $\max(0,c_{i,j})\leq p$. Thus the above equation is bounded as $\min_{t\in[0,\tilde{t}]} e^{-\beta t+\sigma_x^2\frac{t^2}{2}(1+\frac{p}{2})}$, yielding the lemma. ◀

Having parameterized the argument by the crucial choices of $\beta,\tilde{t}$, and p, the next step of the proof refines the bound from Lemma 12 so that it is independent of the underlying distributions Y_i.

▶ **Lemma 13.** *Consider the setting of Theorem 3 and the notation introduced in the statement of Lemma 12. Then, for all $y > 0$ and $p \in [0,1]$ such that $r \leq \frac{p}{2\sqrt{2}}\sqrt{y\mathrm{Tr}(\Sigma_S)d_{eff}(1+\frac{p}{2})}$, the probability $\mathbb{P}_{y\leftarrow S}(||y|| \geq (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma_S)})$ is bounded by*

$$\sqrt{\pi}\left(1+\frac{\sqrt{d_{eff}(2+p)y}}{2(1+\gamma)}\right)e^{\frac{y(2+p)d_{eff}}{4(1+\gamma)^2}} \cdot \left(\left(e^{-1+\sqrt{1+4y}}\cdot\frac{-1+\sqrt{1+4y}}{2y}\right)^{-\frac{d_{eff}}{2}} + \left(1+\frac{\mathrm{Tr}(\Sigma_S)}{r^2}\right)e^{-\left(\frac{p^2(2+p)yd_{eff}\mathrm{Tr}(\Sigma_S)}{4r^2}\right)^{\frac{1}{3}}}\right)$$

Proof. We start from the resulting bound of Lemma 12:

$$\min_{\beta>0,\tilde{t}>0,p\in[0,1]} \sqrt{\pi}\, e^{\frac{\beta^2 d}{2(1+\gamma)^2\mathrm{Tr}(\Sigma_S)}}\left(1+\frac{\beta}{1+\gamma}\sqrt{\frac{d}{2\mathrm{Tr}(\Sigma_S)}}\right) \cdot \mathop{\mathbb{E}}_{x\leftarrow\mathcal{N}(0,\frac{1}{d})^d}\left[\left(\min_{t\in[0,\tilde{t}]} e^{-\beta t+\sigma_x^2\frac{t^2}{2}(1+\frac{p}{2})}\right)+\sum_{i,j}q_{i,j}\left[x\cdot w_{i,j}>\frac{p}{\tilde{t}}\right]\right]$$

The second term in the expectation, for a given i,j, has expectation $q_{i,j}\ \mathrm{erfc}(\frac{p\sqrt{d}}{\sqrt{2}||w_{i,j}||\tilde{t}}) \leq q_{i,j}\, e^{-\frac{p^2 d}{2||w_{i,j}||^2\tilde{t}^2}}$. Thus, since $\mathrm{Tr}(\Sigma_S) = \sum_{i,j} q_{i,j}||w_{i,j}||^2$, the expectation of the entire second term can be bounded by

$$\mathrm{Tr}(\Sigma_S)\sum_{i,j}\frac{q_{i,j}||w_{i,j}||^2}{\mathrm{Tr}(\Sigma_S)}\frac{1}{||w_{i,j}||^2}e^{-\frac{p^2 d}{2||w_{i,j}||^2\tilde{t}^2}} \leq \mathrm{Tr}(\Sigma_S)\sum_{i,j}\frac{q_{i,j}||w_{i,j}||^2}{\mathrm{Tr}(\Sigma_S)}\frac{1}{r^2}e^{-\frac{p^2 d}{2r^2\tilde{t}^2}} = \frac{\mathrm{Tr}(\Sigma_S)}{r^2}e^{-\frac{p^2 d}{2r^2\tilde{t}^2}} \tag{2}$$

provided that $r \leq \sqrt{\frac{p^2 d}{2\tilde{t}^2}}$, since $\frac{1}{r^2}e^{-\alpha/r^2}$ increases in r as long as $r^2 \leq \alpha$. We adopt the condition $r \leq \sqrt{\frac{p^2 d}{2\tilde{t}^2}}$ as it is equivalently expressed in the lemma statement, given our below choices for $\tilde{t}$ and β.

We bound the first term in the expectation as

$$\min_{t\in[0,\tilde{t}]} e^{-\beta t+\sigma_x^2\frac{t^2}{2}(1+\frac{p}{2})} \leq e^{\max\left(-\frac{\beta^2}{(2+p)\sigma_x^2},-\frac{\beta\tilde{t}}{2}\right)} \leq e^{-\frac{\beta^2}{(2+p)\sigma_x^2}}+e^{-\frac{\beta\tilde{t}}{2}} \tag{3}$$

where the first inequality follows from the fact that the quadratic $-\beta t+\sigma_x^2\frac{t^2}{2}(1+\frac{p}{2})$ has a non-negative t^* at its minimum: thus either $t^* \in [0,\tilde{t}]$, in which case the minimum has value $-\frac{\beta^2}{(2+p)\sigma_x^2}$, or $t^* \geq \tilde{t}$; in the latter case, the value of the quadratic is upper bounded by $-\frac{\beta\tilde{t}}{2}$.

Our overall bound is thus the sum of the last two equations (times the factors outside the expectation in the expression of Lemma 12), which we bound – by setting $\tilde{t} = \left(\frac{p^2 d}{\beta r^2}\right)^{\frac{1}{3}}$ to make the exponentials in last two terms below both equal to $e^{-\left(\frac{p^2\beta^2 d}{2r^2}\right)^{\frac{1}{3}}}$ – as

$$e^{-\frac{\beta^2}{(2+p)\sigma_x^2}}+e^{-\frac{\beta\tilde{t}}{2}}+\frac{\mathrm{Tr}(\Sigma_S)}{r^2}e^{-\frac{p^2 d}{2r^2\tilde{t}^2}} \leq e^{-\frac{\beta^2}{(2+p)\sigma_x^2}}+\left(1+\frac{\mathrm{Tr}(\Sigma_S)}{r^2}\right)e^{-\left(\frac{p^2\beta^2 d}{2r^2}\right)^{\frac{1}{3}}}$$

We now further bound the first term of this. We note that, by considering the tangent line of the concave function $-\frac{\beta^2}{(2+p)z}$ at $z=\frac{1}{b}$ for any $b>0$, we have $-\frac{\beta^2}{(2+p)\sigma_x^2} \leq -2b\frac{\beta^2}{2+p}+$

$b^2\frac{\beta^2}{2+p}\sigma_x^2$. Thus, letting σ_i^2 represent the axial variances after diagonalizing the covariance matrix Σ_S,

$$\mathop{\mathbb{E}}_{x\leftarrow\mathcal{N}(0,\frac{1}{d})^d}\left[e^{-\frac{\beta^2}{(2+p)\sigma_x^2}}\right] \le e^{-2b\frac{\beta^2}{2+p}} \mathop{\mathbb{E}}_{x\leftarrow\mathcal{N}(0,\frac{1}{d})^d}\left[e^{b^2\frac{\beta^2}{2+p}\sigma_x^2}\right] = e^{-2b\frac{\beta^2}{2+p}}\prod_{i=1}^{d}\frac{1}{\sqrt{1-2b^2\frac{\beta^2}{(2+p)d}\sigma_i^2}}$$

We bound this last expression under a variance bound $\sigma_x^2 \le \sigma_{\max}^2$, where since the logarithm of $\frac{1}{\sqrt{1-x}}$ is convex, our bound is maximized when each σ_i^2 equals either 0 or $\sigma_{\max}^2$. Thus our bound equals

$$\min_{b>0} e^{-2b\frac{\beta^2}{2+p}}\left(1-2b^2\frac{\beta^2}{(2+p)d}\sigma_{\max}^2\right)^{-\frac{\mathrm{Tr}(\Sigma_S)}{2\sigma_{\max}^2}}$$

Recall that the effective dimension is defined as $d_{\text{eff}} = \frac{\mathrm{Tr}(\Sigma_S)}{\sigma_{\max}^2}$, and introducing the variable substitutions $y = \frac{2d\beta^2\sigma_{\max}^2}{(2+p)\mathrm{Tr}(\Sigma_S)^2}$ and $b' = b\frac{\mathrm{Tr}(\Sigma_S)}{d}$, this expression becomes simply $\left(\min_{b'} e^{2b'y}(1-b'^2y)\right)^{\frac{d_{\text{eff}}}{2}}$. The min is at $b' = \frac{-1+\sqrt{1+4y}}{2y}$ leading to a value of $\left(e^{-1+\sqrt{1+4y}}\cdot\frac{-1+\sqrt{1+4y}}{2y}\right)^{-\frac{k}{2}}$.

Thus, reexpressing in terms of $y = \frac{2d\beta^2\sigma_{\max}^2}{(2+p)\mathrm{Tr}(\Sigma_S)^2}$ and d_{eff} we have the desired bound of

$$\min_{y>0,p\in[0,1]} \sqrt{\pi}\left(1+\frac{\sqrt{d_{\text{eff}}(2+p)y}}{2(1+\gamma)}\right)e^{\frac{y(2+p)d_{\text{eff}}}{4(1+\gamma)^2}} \cdot\left(\left(e^{-1+\sqrt{1+4y}}\cdot\frac{-1+\sqrt{1+4y}}{2y}\right)^{-\frac{d_{\text{eff}}}{2}} + \left(1+\frac{\mathrm{Tr}(\Sigma_S)}{r^2}\right)e^{-\left(\frac{p^2(2+p)yd_{\text{eff}}\mathrm{Tr}(\Sigma_S)}{4r^2}\right)^{\frac{1}{3}}}\right)$$

Lastly, we check that for all pairs of (y,p) satisfying the lemma conditions, the above proof goes through. The above analysis assumed $r \le \sqrt{\frac{p^2d}{2\tilde{t}^2}}$; substituting in our choice $\tilde{t} = \left(\frac{p^2d}{\beta r^2}\right)^{\frac{1}{3}}$ yields $r \le \frac{1}{\sqrt{2}}p^{\frac{1}{3}}d^{\frac{1}{6}}\beta^{\frac{1}{3}}r^{\frac{2}{3}}$, which is equivalent to $r \le \frac{1}{2\sqrt{2}}p\sqrt{d}\beta$. Substituting $y = \frac{2d\beta^2\sigma_{\max}^2}{(2+p)\mathrm{Tr}(\Sigma_S)^2}$ yields $r \le \frac{p}{2\sqrt{2}}\sqrt{y\mathrm{Tr}(\Sigma_S)d_{\text{eff}}(1+\frac{p}{2})}$. Summarizing, for every pair of (y,p) satisfying the lemma conditions, there exists a corresponding triple of $(\beta,\tilde{t},p)$ in the domain of the minimization for the bound in Lemma 12. ◀

We complete the proof of Theorem 3 for the $\gamma \le 1$ case by picking the parameters y and p in the result of Lemma 13, and simplifying the expression.

▶ **Lemma 14.** *Consider the setting of Theorem 3. If $\gamma \le 1$, then the probability $\mathbb{P}_{y\leftarrow S}(||y|| \ge (1+\gamma)\sqrt{\mathrm{Tr}(\Sigma_S)})$ is bounded by*

$$\sqrt{\pi}\left(1+\frac{3}{\sqrt{2}}\sqrt{\min\left(d_{\mathit{eff}},\frac{\sqrt{\mathrm{Tr}(\Sigma_S)}}{r}\right)}\right)\left(2+\frac{\mathrm{Tr}(\Sigma_S)}{r^2}\right)e^{-\frac{2\gamma^2}{\frac{2}{d_{\mathit{eff}}}+\frac{r}{\sqrt{\mathrm{Tr}(\Sigma_S)}}}\frac{0.7+0.15\gamma}{0.7+\gamma}}$$

Proof. In the following proof, we first assume that our final choices of y and p satisfy the preconditions of Lemma 13. At the end of the proof, we show that if the preconditions are not satisfied, then the lemma statement is vacuously true.

Letting $f(y) = -1+\sqrt{1+4y}+\log\frac{-1+\sqrt{1+4y}}{2y}$ represent the logarithm of the inside of the first term in the main parenthesis, then choose $p = \frac{rd_{\text{eff}}}{\sqrt{\mathrm{Tr}(\Sigma_S)}}\frac{f(y)^{\frac{3}{2}}}{2\sqrt{y}}$ so as to make the

exponential portion of the second term slightly smaller than the first term. Namely, the exponent of the second term becomes

$$-\left(\frac{p^2(2+p)yd_{\text{eff}}\text{Tr}(\Sigma_S)}{4r^2}\right)^{\frac{1}{3}} \leq -\left(\frac{p^2(2)yd_{\text{eff}}\text{Tr}(\Sigma_S)}{4r^2}\right)^{\frac{1}{3}} = -\frac{d_{\text{eff}}}{2}f(y)$$

Thus, as we are not trying to optimize the polynomial terms, we simply bound the exponential part of the second term by the first term, and merge them, to yield the bound

$$\min_{y>0}\sqrt{\pi}\left(1+\frac{\sqrt{d_{\text{eff}}(2+p)y}}{2(1+\gamma)}\right)\left(2+\frac{\text{Tr}(\Sigma_S)}{r^2}\right)e^{\frac{y(2+p)d_{\text{eff}}}{4(1+\gamma)^2}-\frac{d_{\text{eff}}}{2}f(y)}$$

Considering, for simplicity, only the exponential portion of this bound, substituting in our choice of $p = \frac{rd_{\text{eff}}}{\sqrt{\text{Tr}(\Sigma_S)}}\frac{f(y)^{\frac{3}{2}}}{2\sqrt{y}}$, and defining $r' = \frac{rd_{\text{eff}}}{\sqrt{\text{Tr}(\Sigma_S)}}$ yields

$$\min_{y>0} e^{\frac{y(2+p)d_{\text{eff}}}{4(1+\gamma)^2}-\frac{d_{\text{eff}}}{2}f(y)} = \min_{y>0} e^{d_{\text{eff}}\left(\frac{2y+\frac{1}{2}r'\sqrt{y}f(y)^{\frac{3}{2}}}{4(1+\gamma)^2}-\frac{1}{2}f(y)\right)}$$

To simplify this expression we substitute $z = -1+\sqrt{1+4y}$, so that $f(y) = z-\log(1+\frac{z}{2})$, and $y = \frac{z^2}{4}+\frac{z}{2}$. Choose $z = \frac{8\gamma}{2+r'}$, so that $r' = \frac{8\gamma}{z}-2 = \frac{8\gamma-2z}{z}$. It is easy to check that, for all $z \geq 0$, $\frac{1}{2}\sqrt{y}f(y)^{\frac{3}{2}} = \frac{1}{2}\sqrt{\frac{z^2}{4}+\frac{z}{2}}\left(z-\log(1+\frac{z}{2})\right)^{\frac{3}{2}} \leq \frac{1}{8}z^2+\frac{5}{64}z^3$, bounding the above expression by

$$e^{d_{\text{eff}}z\left(\frac{1}{(1+\gamma)^2}\left(\frac{z}{8}+\frac{1}{4}\right)+\frac{8\gamma-2z}{4(1+\gamma)^2}\left(\frac{1}{8}+\frac{5}{64}z\right)-\frac{1}{2}+\frac{1}{2z}\log\left(1+\frac{z}{2}\right)\right)}$$

We claim the expression inside the parentheses in the exponent is at most $-\frac{\gamma}{4}\frac{0.7+0.15\gamma}{0.7+\gamma}$, for any $z \geq 0$ (thus for all $\gamma > 0, r' > 0$), which we prove by maximizing the parenthetical expression with respect to z. Specifically, we note that $\frac{1}{2z}\log(1+\frac{z}{2})$ is bounded by *both* its quadratic power series expansion $\frac{1}{4}-\frac{z}{16}+\frac{z^2}{48}$, and a second, more ad hoc quadratic, $\frac{1}{4}-\frac{7}{128}z+\frac{5}{512}z^2$. For $\gamma < 0.35$ we use the first quadratic bound on the logarithmic expression, making the parenthetical expression a quadratic in z, whose maximum we may explicitly evaluate (e.g., with Mathematica) as $\frac{56\gamma-75\gamma^2-180\gamma^3-76\gamma^4}{32(1+\gamma)^2(-7+16\gamma+8\gamma^2)}$; correspondingly, for $\gamma \geq 0.35$ we use the second bound, which similarly yields a maximum of the parenthetical expression of $\frac{1-467\gamma-605\gamma^2-209\gamma^3}{640(1+\gamma)^2(3+\gamma)}$. Both of these expressions are bounded, in the domains $\gamma < 0.35$ and $\gamma \geq 0.35$ respectively, by $-\frac{\gamma}{4}\frac{0.7+0.15\gamma}{0.7+\gamma}$, as desired. Thus, substituting back the definitions of z, r', our final probability bound (omitting the polynomial multipliers) is $e^{-d_{\text{eff}}z\frac{\gamma}{4}\frac{0.7+0.15\gamma}{0.7+\gamma}} = e^{-\frac{2\gamma^2}{\frac{2}{d_{\text{eff}}}+\frac{r}{\sqrt{\text{Tr}(\Sigma_S)}}}\frac{0.7+0.15\gamma}{0.7+\gamma}}$.

We now bound the polynomial multipliers. Consider the term $d_{\text{eff}} \cdot y$, which we will bound as $d_{\text{eff}} \cdot y \leq 6\min(d_{\text{eff}}, \frac{\sqrt{\text{Tr}(\Sigma_S)}}{r})$. Since $z = \frac{8\gamma}{2+r'} \leq 4$, and $y = \frac{z^2}{4}+\frac{z}{2} \leq 6$, we have $d_{\text{eff}} \cdot y \leq 6d_{\text{eff}}$. Suppose $d_{\text{eff}} \geq \frac{\sqrt{\text{Tr}(\Sigma_S)}}{r}$; equivalently, $r' \equiv \frac{rd_{\text{eff}}}{\sqrt{\text{Tr}(\Sigma_S)}} \geq 1$. Since $z = \frac{8\gamma}{2+r'} \leq \frac{8}{2+r'}$ and $y = \frac{z^2}{4}+\frac{z}{2}$, it is easy to verify that $y \leq \frac{9}{2}\frac{1}{r'}$. Thus $d_{\text{eff}} \cdot y \leq \frac{9}{2}\frac{\sqrt{\text{Tr}(\Sigma_S)}}{r}$. Thus in all cases, $d_{\text{eff}} \cdot y \leq 6\min(d_{\text{eff}}, \frac{\sqrt{\text{Tr}(\Sigma_S)}}{r})$. Thus, since $\gamma \geq 0$, $p \leq 1$, we bound the polynomial term by

$$\sqrt{\pi}\left(1+\frac{\sqrt{d_{\text{eff}}(2+p)y}}{2(1+\gamma)}\right)\left(2+\frac{\text{Tr}(\Sigma_S)}{r^2}\right) \leq \sqrt{\pi}\left(1+\frac{3}{\sqrt{2}}\sqrt{\min\left(d_{\text{eff}}, \frac{\sqrt{\text{Tr}(\Sigma_S)}}{r}\right)}\right)\left(2+\frac{\text{Tr}(\Sigma_S)}{r^2}\right)$$

which is the claimed multiplier. We note for below that this multiplier is at least 5, provided $\frac{\sqrt{\mathrm{Tr}(\Sigma_S)}}{r} \geq 1$ and otherwise the desired overall probability bound is vacuously > 1.

Finally, Lemma 13 assumes $r \leq \frac{p}{2\sqrt{2}}\sqrt{y\mathrm{Tr}(\Sigma_S)d_{\mathrm{eff}}(1+\frac{p}{2})}$, which we simplify to the sufficient condition $r \leq \frac{p}{2\sqrt{2}}\sqrt{y\mathrm{Tr}(\Sigma_S)d_{\mathrm{eff}}}$; we substitute our choice $p = \frac{rd_{\mathrm{eff}}}{\sqrt{\mathrm{Tr}(\Sigma_S)}}\frac{f(y)^{\frac{3}{2}}}{2\sqrt{y}}$ to yield $r \leq \frac{rd_{\mathrm{eff}}}{\sqrt{\mathrm{Tr}(\Sigma_S)}}\frac{f(y)^{\frac{3}{2}}}{4\sqrt{2y}}\sqrt{y\mathrm{Tr}(\Sigma_S)d_{\mathrm{eff}}}$. Simplifying yields $4\sqrt{2} \leq (d_{\mathrm{eff}}f(y))^{\frac{3}{2}}$, and thus $d_{\mathrm{eff}}f(y) \geq 2^{\frac{5}{3}}$. Since $f(y) = z - \log(1+\frac{z}{2}) \geq \frac{z}{2}$ it is sufficient for $d_{\mathrm{eff}}z \geq 2^{\frac{8}{3}}$. If this is violated, then the exponential term of the final probability bound is at least $e^{-d_{\mathrm{eff}}z\frac{\gamma}{4}\frac{0.7+0.15\gamma}{0.7+\gamma}} \geq e^{-d_{\mathrm{eff}}z\frac{1}{4}} \geq e^{-2^{\frac{8}{3}}\frac{1}{4}} \geq \frac{1}{5}$. Since the polynomial multiplier of the exponential is at least 5 when $\frac{\sqrt{\mathrm{Tr}(\Sigma_S)}}{r} \geq 1$, either the above proof holds or the desired inequality is vacuously true. ◀

References

1 Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci*, 58(1):137–147, 1999.

2 Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Ann. I. H. Poincaré -PR*, 48(4):1148–1185, 2012.

3 Olivier Catoni and Ilaria Giulini. Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector. *arXiv:1802.04308*, 2018.

4 Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-Gaussian rates. In *Proc. COLT '20*, pages 786–806, 2019.

5 Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *Ann. Stat*, 44(6):2695–2725, 2016.

6 Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians. *arXiv:1810.08693*, 2020.

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

8 Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proc. ICML'17*, pages 999–1008, 2017.

9 Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a Gaussian: Getting optimal error, efficiently. In *Proc. SODA'18*, pages 2683–2702, 2018.

10 Ilias Diakonikolas and Daniel Kane. Robust high-dimensional statistics. In Tim Roughgarden, editor, *Beyond the Worst-Case Analysis of Algorithms*, pages 382–402. Cambridge University Press, 2021.

11 Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. In *Proc. NeuRIPS'20*, pages 1830–1840, 2020.

12 Samuel B. Hopkins. Mean estimation with sub-Gaussian rates in polynomial time. *Ann. Stat.*, 48(2):1193–1213, 2020.

13 Mark R. Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theor. Comput. Sci*, 43:169–188, 1986.

14 Jasper C.H. Lee and Paul Valiant. Optimal sub-Gaussian mean estimation in $\mathbb{R}$. To appear in *Proc. FOCS'21*.

15 Zhixian Lei, Kyle Luh, Prayaag Venkat, and Fred Zhang. A fast spectral algorithm for mean estimation with sub-Gaussian rates. In *Proc. COLT '20*, pages 2598–2612, 2020.

16 Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions—a survey. *Found. Comput. Math.*, 19(5):1145–1190, 2019.

17 Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *Ann. Stat.*, 47(2):783–794, 2019.

18 Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *Ann. Stat.*, 49(1):393–410, 2021.

19 Stanislav Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Stat. Probab. Lett.*, 127:111–119, 2017.

20 A.S. Nemirovsky and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley, 1983.

21 Roberto I. Oliveira and Paulo Orenstein. The sub-Gaussian property of trimmed means estimators. *Technical Report, IMPA*, 2019.

22 Joel A. Tropp. An Introduction to Matrix Concentration Inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

23 V.V. Yurinskiĭ. Exponential inequalities for sums of random vectors. *J. Multivar. Anal.*, 6(4):473–499, 1976.