Learning Entangled Single-Sample Gaussians in the Subset-of-Signals Model

Yingyu Liang   YLIANG@CS.WISC.EDU
University of Wisconsin-Madison

Hui Yuan[*]   YUANHUI@MAIL.USTC.EDU.CN
Univeristy of Science and Technology of China

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

In the setting of entangled single-sample distributions, the goal is to estimate some common parameter shared by a family of n distributions, given one single sample from each distribution. This paper studies mean estimation for entangled single-sample Gaussians that have a common mean but different unknown variances. We propose the subset-of-signals model where an unknown subset of m variances are bounded by 1 while there are no assumptions on the other variances. In this model, we analyze a simple and natural method based on iteratively averaging the truncated samples, and show that the method achieves error $O\left(\frac{\sqrt{n \ln n}}{m}\right)$ with high probability when $m = \Omega(\sqrt{n \ln n})$, slightly improving existing bounds for this range of m. We further prove lower bounds, showing that the error is $\Omega\left(\left(\frac{n}{m^4}\right)^{1/2}\right)$ when m is between $\Omega(\ln n)$ and $O(n^{1/4})$, and the error is $\Omega\left(\left(\frac{n}{m^4}\right)^{1/6}\right)$ when m is between $\Omega(n^{1/4})$ and $O(n^{1-\epsilon})$ for an arbitrarily small $\epsilon > 0$, improving existing lower bounds and extending to a wider range of m.
Keywords: Entangled Gaussians, Mean Estimation, Subset-of-Signals

1. Introduction

This work considers the novel parameter estimation setting called entangled single-sample distributions. In this setting, distributions are entangled in the sense that they share some common parameter and our goal is to estimate the common parameter based on one sample from each distributions obtained. We focus on the mean estimation problem in the subset-of-signals model when the distributions are Gaussians. In this problem, we have n independent Gaussians with a common mean with different unknown variances. Given one sample from each of the Gaussians, our goal is to estimate the mean.

There can be different configurations of the unknown variances. In this work, we propose a basic model called subset-of-signals, which assumes that an unknown subset of m variances are bounded by 1 while there are no assumptions on the other variances. Equivalently, $\sigma_{(m)} \leq 1$ where $\sigma_{(m)}$ is the m-th smallest value in $\{\sigma_i\}_{i=1}^n$. The subset-of-signals model gives a simple setting specifying the possible configurations of n unknown variances $\{\sigma_i\}_{i=1}^n$ for analysis. While even in this simple setting, the optimal rates of mean estimation for entangled single-sample Gaussians are still unknown (for most values of m).

---

[*] The work was done during the summer internship of H. Yuan at the University of Wisconsin-Madison.

The setting of entangled single-sample distributions is motivated for both theoretical and practical reasons. From the theoretical perspective, it goes beyond the typical i.i.d. setting and raises many interesting open questions in the most fundamental topics like mean estimation of Gaussians. It can also be viewed as a generalization of the traditional mixture modeling, since the number of distinct mixture components could grow with the number of samples and even be as large as the number of samples. From the practical perspective, traditional i.i.d. assumption can lead to a bad modeling of data in modern applications, where various forms of heterogeneity occur. In particular, entangled Gaussians capture heteroscedastic noises in various applications and thus can be a natural model for studying robustness.

Though theoretically interesting and practically important, few studies exist in this setting. Chierichetti et al. (2014) considered the mean estimation for entangled Gaussians and showed the existence of a gap between estimation error rates of the best possible estimator in this setting and the maximum likelihood estimator when the variances are known. It focused on the case where most samples are "high-noised" (i.e., most variances are large), and provided bounds in terms of $\sigma_{(m)}$ with small m like $\Theta(\ln n)$. Pensia et al. (2019) considered means estimation for symmetric, unimodal distributions with sharpened bounds, and provided extensive discussion on the performance of their estimators in different configurations of the variances. Many questions are still largely open. In particular, when instantiated in the subset-of-signals model, existing studies provide interesting upper bounds and lower bounds but a large gap remains. See the related work section and remarks after our theorems for more details.

This work thus proposes the subset-of-signals model and attempts to gain better understanding on the problem. For the upper bound, we aim to achieve a vanishing error bound (i.e., the error bound tends to 0 when $n \to +\infty$). We analyze a simple algorithm based on iteratively averaging the truncated samples: it keeps an iterate and each time it truncates the samples in an interval around the current iterate and then averages the truncated samples to compute the next iterate. We also prove lower bounds for a wide range of m, improving known bounds. Our main results are summarized below.

1.1. Main Results

Problem Setup. Suppose we have n independent samples $x_i \sim \mathcal{N}(\mu^\star, \sigma_i^2)$, where the distributions have a common mean $\mu^\star$ but different variances σ_i^2. The mean and variances are all unknown. We consider the subset-of-signal model, where an unknown subset of m variances are bounded by 1 while there are no assumptions on the other variances. That is, $\sigma_{(m)} \leq 1$ where $\sigma_{(m)}$ is the m-th smallest value in $\{\sigma_i\}_{i=1}^n$. Our goal is to estimate the common mean $\mu^\star$ from the samples $\{x_i\}_{i=1}^n$.

As usual, we use $f(n, m) = O(g(n, m))$ (or $f(n, m) \lesssim g(n, m)$) if there exist N, M and $C > 0$ such that when $n > N$ and $m > M$, $f(n, m) \leq Cg(n, m)$. $f = \tilde{O}(g)$ hides logarithmic terms. $f = \Omega(g)$ (or $f \gtrsim g$), $f = \Theta(g)$ (or $f \simeq g$), $f = o(g)$, and $f = \omega(g)$ are defined as usual.

Upper bound. We obtain the following result for an algorithm based on iteratively averaging truncated samples(see Algorithm 1 for the details).

Theorem 1 *If $\sigma_{(m)} \leq 1$ for $m = \Omega(\sqrt{n \ln n})$, then with probability at least $1 - 1/n$, the output $\hat{\mu}$ of Algorithm 1 satisfies*

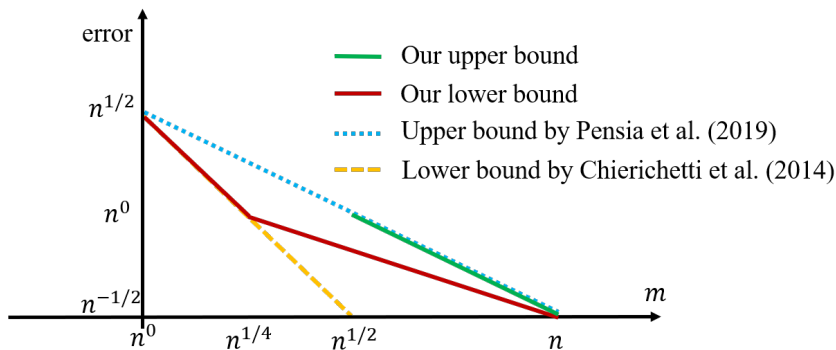$$|\hat{\mu} - \mu^\star| \lesssim \frac{\sqrt{n \ln n}}{m}.$$

Figure 1: Our bounds and those from previous works for mean estimation of entangled Gaussians in the subset-of-signals model. x-axis is the number of Gaussians with variances 1, y-axis is the error. See the text for the details of the bounds.

The result shows that the algorithm can achieve a vanishing error when $m = \omega(\sqrt{n \ln n})$. Therefore, we can achieve vanishing error with only an $\omega(\sqrt{\ln n/n})$ fraction of samples with bounded variances. This means even when the noisy samples dominates the data and the fraction of signals diminishes when $n \to +\infty$, we can still obtain accurate estimation. The result also shows that when there are only a constant fraction of "heavy-noised" data (i.e., $m = \Theta(n)$), the error rate is $O(\sqrt{\ln n/n})$, which matches the optimal error rate $O(1/\sqrt{n})$ up to a logarithmic factor. Our result matches the best bound known: the hybrid estimator proposed in Pensia et al. (2019) achieved $O(\sqrt{n} \ln n/m)$ in the subset-of-signals model, almost the same upper bound as ours, but for essentially all values of m. On the other hand, our algorithm is simple and natural. The iterative truncation is widely used in practice, so our analysis can be viewed as a justification for this heuristic. While the estimator in Pensia et al. (2019) combines two estimators and is not related to practical heuristics.

Our upper bound is in sharp contrast to the robust mean estimation in the commonly studied adversarial contamination model (Valiant, 1985; Huber, 2011; Diakonikolas et al., 2019), where an ϵ fraction of the data are adversarially modified and it has been shown that vanishing error is impossible when $\epsilon = \Omega(1)$. This means that the entangled distributions setting can be much more benign than the adversarial contamination model. For mean estimation for entangled Gaussians in the subset-of-signals model, one can view it as an adversary picking $n - m$ variances but having no control over the sampling process after specifying those variances. That is, it is a semi-adversarial model and can be much more benign than the fully adversarial contamination model.

Lower bound. We now turn to the lower bound. Note that an instance of our problem is specified by $\mu^\star$ and $\{\sigma_i\}_{i=1}^n$.

Theorem 2 *Suppose $\sigma_{(m)} \leq 1$.*

- *If $m = \Omega(\ln n)$ and $m = O(n^{1/4})$, then there exist a family of instances and a distribution over these instances such that any estimator has expected error $\Omega\left(\left(\frac{n}{m^4}\right)^{1/2}\right)$.*

3

- *For any arbitrarily small $\epsilon > 0$, if m is between $\Omega(n^{1/4})$ and $O(n^{1-\epsilon})$, then there exist a family of instances and a distribution over these instances such that any estimator has expected error $\Omega\left(\left(\frac{n}{m^4}\right)^{1/6}\right)$.*

The bound is for a distribution over the instances, which then implies the typical minimax bound. The result show that when $m = O(n^{1/4})$, it is impossible to obtain vanishing error. When m is as small as $\Theta(\ln n)$, the error is $\tilde{\Omega}(\sqrt{n})$, paying a factor of $\tilde{\Omega}(\sqrt{n})$ compared to the oracle bound $O(1/\sqrt{m})$ when the m bounded variance samples are known. When $m = \Omega(n^{1/4})$, the lower bound does not exclude the possibility of vanishing error. On the other hand, it shows that one needs to pay a factor of $\Omega\left(\left(\frac{n}{m}\right)^{1/6}\right)$, compared to the oracle bound $O(1/\sqrt{m})$ when the m bounded variance samples are known. It also shows that one needs to pay a factor of $\Omega\left(\left(\frac{n}{m}\right)^{2/3}\right)$, compared to the bound $O(1/\sqrt{n})$ when all samples have bounded variance 1.

Our result extends and improves the lower bound in Chierichetti et al. (2014). Their bound is $\Omega\left(\left(\frac{n}{m^4}\right)^{1/2}\right)$ for m between $\Omega(\ln n)$ and $o(\sqrt{n})$. Our result extends the range of m by including the values between $\Omega(n^{1/2})$ and $O(n^{1-\epsilon})$ (for any arbitrarily small $\epsilon > 0$). It also improves their bound in the range between $\Omega(n^{1/4})$ and $o(n^{1/2})$, by a factor of $\Omega\left(\left(\frac{m^4}{n}\right)^{1/3}\right)$.

As far as we know, our upper and lower bounds are the best existing bounds for mean estimation of entangled single-sample Gaussians in the subset-of-signals model. See Figure 1 for an illustration. There is still a gap between the known upper and lower bounds. A natural direction is to close the gap and obtain the optimal rates, which we left as future work.

2. Related Work

Entangled distributions. This setting is first studied by Chierichetti et al. (2014), which considered mean estimation for entangled Gaussians and presented a algorithm combining the k-median and the k-shortest gap algorithms. It also showed the existence of a gap between the error rates of the best possible estimator in this setting and the maximum likelihood estimator when the variances are known. Pensia et al. (2019) considered a more general class of distributions (unimodal and symmetric) and provided analysis on both individual estimator (r-modal interval, k-shortest gap, k-median estimators) and hybrid estimator, which combines Median estimator with Shortest Gap or Modal Interval estimator. They also discussed slight relaxation of the symmetry assumption and provided extensions to linear regression. Our work focuses on the subset-of-signals model that allows to study the minimax rate and helps a clearer understanding of the problem (but our results can also be used for some other configurations). The algorithm we analyzed is based on the natural iterative truncation heuristics frequently used in practice to handle heteroscedastic noises, and our bound for it matches the best known rates (obtained by the hybrid estimator in Pensia et al. (2019)) in the range $m = \Omega(\sqrt{n \ln n})$. We also extends (to a wider range of m) and improves the lower bound in Chierichetti et al. (2014).

Yuan and Liang (2020) considered mean estimation for entangled distributions, but the distributions are not assumed to be Gaussians (it only assumed the distributions have the same mean and their variances exist). Due to this generality, their upper bound is significantly worse than ours: it's only for $m \geq 4n/5$ (i.e., only a constant fraction of high noise points); it does not achieve a vanishing error when n tends to infinity. The paper doesn't provide lower bounds. Their algorithm is also

based on iterative truncation, but has the following important difference: it removes a fixed fraction of data points in each iteration, rather than doing adaptive truncation. In contrast, our algorithm uses adaptive truncation interval lengths. This is crucial to obtain our results, since intuitively the best bias-variance trade-off introduced by the truncation can only be achieved with adaptive truncation.

The entangled distributions setting is also closely related to robust estimation, which have been extensively studied in the literature of both classic statistics and machine learning theory.

Robust mean estimation. There are several classes of data distribution models for robust mean estimators. The most commonly addressed is adversarial contamination model, whose origin can be traced back to the malicious noise model by Valiant (1985) and the contamination model by Huber (2011). Under contamination, mean estimation has been investigated in Diakonikolas et al. (2017, 2019); Cheng et al. (2019). Another related model is the mixture of distributions. There has been steady progress in algorithms for leaning mixtures, in particular, leaning Gaussian mixtures. Starting from Dasgupta (1999), a rich collection of results are provided in many studies, such as Sanjeev and Kannan (2001); Achlioptas and McSherry (2005); Kannan et al. (2005); Belkin and Sinha (2010a,b); Kalai et al. (2010); Moitra and Valiant (2010); Diakonikolas et al. (2018).

Heteroscedastic models. The setting of entangled distributions is also closely related to heteroscedastic models, which have been a classic topic in statistics. For example, in heterogeneous linear regression (Munoz et al., 1986; Vicari and Vichi, 2013), the errors for different response variables may have different variances, and weighted least squares has been used for estimating the parameters in this setting. Another example is Principal Component Analysis for heteroscedastic data (Hong et al., 2018a,b; Zhang et al., 2018). The entangled Gaussians can be viewed as a model of mean estimation in the presence of heteroscedastic noises.

3. Upper Bound

The naïve method of averaging all samples cannot achieve a small error when some distributions have large variances. A natural idea is then to reduce the variances. Truncation is a frequently used heuristic, i.e., projecting the samples to an interval (around a current estimation) to get controlled variances. However, while averaging the original samples is consistent, truncation can lead to bias. So truncation introduces some form of bias-variance tradeoff and the width of the interval controls the tradeoff. Intuitively, the best width will depend on how aligned the interval is with the true mean; for intervals around estimations of different error, the width for the best tradeoff can be different. Therefore, we consider iterative truncation using adaptive widths for the interval.

Algorithm 1 describes the details of our method. Given an initial estimation μ_0, it averages the truncated data in an interval around the estimation iteratively. In particular, the algorithm has K stages, and each stage has T steps. In step t of stage k, given a current estimation $\mu_t^{(k)}$ and a width parameter $\delta_t^{(k)}$, the algorithm computes the new estimation $\mu_{t+1}^{(k)}$ by averaging the truncated data $\phi(x_i; \Delta_t^{(k)})$, where $\Delta_t^{(k)}$ is the interval around $\mu_t^{(k)}$ with radius $\delta_t^{(k)}$, and ϕ is defined as:

$$\phi(x; [a, b]) = \begin{cases} a, & \text{if } x < a, \\ x, & \text{if } a \leq x \leq b, \\ b, & \text{if } x > b. \end{cases} \tag{1}$$

For this algorithm, we prove the following guarantee.

Algorithm 1: Mean Estimation via Iterative Truncation

Input: $\{x_i\}_{i=1}^n$, initialization μ_0, and parameters B, m s.t. $B \geq 2|\mu_0 - \mu^*|, \sigma_{(m)} \leq 1$

Set $\delta^{(0)} = B, \mu_0^{(0)} = \mu_0, K = \lfloor \log_2 \delta^{(0)} \rfloor, T = \lceil 64n \ln n/m \rceil$

for $k = 0, 1, \ldots, K$ **do**

    for $t = 0, 1, \ldots, T$ **do**

        $\Delta_t^{(k)} = [\mu_t^{(k)} - \delta^{(k)}, \mu_t^{(k)} + \delta^{(k)}]$

        $\mu_{t+1}^{(k)} = \frac{1}{n} \sum_{i=1}^n \phi(x_i; \Delta_t^{(k)})$ // ϕ is defined in Eqn (1)

    end

    $\mu_0^{(k+1)} = \mu_{T+1}^{(k)}, \delta^{(k+1)} = \delta^{(k)}/2$

end

Output: $\hat{\mu} \leftarrow \mu_{T+1}^{(K)}$

Theorem 1 *If $\sigma_{(m)} \leq 1$ for $m = \Omega(\sqrt{n \ln n})$, then with probability at least $1 - 1/n$, the output $\hat{\mu}$ of Algorithm 1 satisfies*

$$|\hat{\mu} - \mu^\star| \lesssim \frac{\sqrt{n \ln n}}{m}.$$

Remark. The algorithm needs an initialization μ_0 and parameter B. There exist methods to achieve this, e.g., set μ_0 as the sample mean and B as two times the diameter of the sample points.

Remark. Our proof actually gives more general results. Let $m(\delta) = \max\{i : \sigma_{(i)} \leq \delta\}$ and let H_δ^σ be the harmonic mean of $\{\max(\sigma_i, \delta)\}_{i=1}^n$, i.e., $H_\delta^\sigma = n/(\sum_{i=1}^n \frac{1}{\max(\sigma_i, \delta)})$. Then our analysis shows that for any k in the algorithm, the estimation at the end of the k-th iteration satisfies $|\mu_{T+1}^{(k)} - \mu^\star| \lesssim H_{\delta^{(k)}}^\sigma \sqrt{\frac{\ln n}{n}}$. That is, with probability at least $1 - 1/n$, the algorithm can output an estimation $\hat{\mu}$ (by setting proper K and T) for any δ with $m(\delta) = \Omega(\sqrt{n \ln n})$, such that

$$|\hat{\mu} - \mu^\star| \lesssim H_\delta^\sigma \sqrt{\frac{\ln n}{n}}. \tag{2}$$

Since $H_\delta^\sigma \leq n\delta/m(\delta)$, the error is $\lesssim \frac{\delta\sqrt{n \ln n}}{m(\delta)}$. So for any $t \geq m = \Omega(\sqrt{n \ln n})$, by setting $\delta = \sigma_{(t)}$ (the t-th smallest variance), we can get with probability $1 - 1/n$,

$$|\hat{\mu} - \mu^\star| \lesssim \frac{\sigma_{(t)} \sqrt{n \ln n}}{t} \tag{3}$$

When $t = m$, we recover the bound in the theorem.

    The more general results are more adaptive. First, they can be applied to more general threshold values δ. For example, for the configuration of variances where $\sigma_{(t)}$ can increase with n, one can still get vanishing error when $\sigma_{(t)} = o(t/\sqrt{n \ln n})$. Second, (2) can be applied to different configurations of σ_i's and obtain better bounds. When $\sigma_{(i)}$'s for $i > m$ are benign, (2) shows that they can help the estimation and quantifies the provided information with the notion H_δ^σ.

Remark. We would also like to point out, the hybrid estimator proposed in Pensia et al. (2019) also achieved almost the same upper bound $O(\sqrt{n} \ln n / m)$ as ours in the subset-of-signals model, but for essentially all values of m. Their bound is obtained by combining two estimators, and depends on a notion r_k, the length of the smallest interval containing k samples. In comparison, our bound is for the iterative truncation method frequently used in practice, and depends on the notion H_δ^σ.

More details of the existing bounds are as follows. Chierichetti et al. (2014) achieved an error bound $\min_{2 \le k \le \log n} \tilde{O}\left(n^{1/2(1+1/(k-1))} \sigma_k\right)$. Among all estimators studied in Pensia et al. (2019), the superior performance is obtained by the hybrid estimators, which includes version (1): combining k_1-median with k_2-shorth and version (2): combining k_1-median with modal interval estimator. These two versions achieve similar guarantees. Version 1 of the hybrid estimator outputs $\hat{\mu}_{k_1,k_2}$ such that $|\hat{\mu}_{k_1,k_2} - \mu| \le \frac{4\sqrt{n}\log n}{k_2} r_{2k_2}$ with probability $1 - 2\exp(-c'k_2) - 2\exp(-c\log^2 n)$, where $k_1 = \sqrt{n}\log n$ and $k_2 \ge C \log n$. Here r_k is defined as $\inf\left\{r : \frac{1}{n}\sum_{i=1}^n \mathbb{P}(|x_i - \mu^\star| \le r) \ge \frac{k}{n}\right\}$. So the error bound varies with specific configurations of the variances. Furthermore, the modal interval estimator or the shorth estimator still work for small m's, so their bound holds also for $m = \tilde{O}(n^{1/2})$.

3.1. Proof of Theorem 1

To prove the theorem, we first focus on one stage and omit the superscript (k). Define

$$e_t := |\mu_t - \mu^\star|, \tag{4}$$

$$z_i := \phi(x_i; \Delta_t) - \mu^\star, \tag{5}$$

$$\bar{z}_i := z_i - \mathbb{E}z_i. \tag{6}$$

We have

$$\mu_{t+1} - \mu^\star = \frac{\sum_{i=1}^n (\phi(x_i; \Delta_t) - \mu^\star)}{n} = \frac{1}{n}\sum_{i=1}^n z_i. \tag{7}$$

To bound $|\sum_{i=1}^n z_i|$, we need to bound $\bar{z}_i$'s and $|\mathbb{E}z_i|$'s. Since z_i is 1-Lipschitz w.r.t. μ_t, a standard ϵ-net argument gives a uniform concentration bound of $\bar{z}_i$'s in Lemma 3. $|\mathbb{E}z_i|$ is bounded in Lemma 4. See Appendix A for their proofs.

Lemma 3 *Let $z_i(\mu) = \phi(x_i; [\mu - \delta, \mu + \delta]) - \mu^\star, \bar{z}_i(\mu) = z_i(\mu) - \mathbb{E}z_i(\mu)$. With probability at least $1 - 1/n^3$, for any μ satisfying $|\mu - \mu^\star| \le \delta_e$, we have*

$$\left|\sum_{i=1}^n \bar{z}_i(\mu)\right| \lesssim \delta\sqrt{n \ln n} + \frac{\delta_e}{n}.$$

Lemma 4 *Let $z_i(\mu) = \phi(x_i; [\mu - \delta, \mu + \delta]) - \mu^\star$ and $\delta_e = |\mu - \mu^\star|$. Then*

$$|\mathbb{E}z_i| \le \delta_e \left(1 - \frac{1}{5}\frac{\delta}{\max\{\delta_e, \delta\}}\frac{\delta}{\max\{\sigma_i, \delta\}}\right).$$

Using these two lemmas, we can analyze one iteration of the algorithm.

Lemma 5 *If $\delta \geq e_t$, then with probability at least $1 - 1/n^3$,*

$$e_{t+1} \leq C\delta\sqrt{\frac{\ln n}{n}} + e_t\left(1 - \frac{\delta}{5H_\delta^\sigma}\right)$$

where H_δ^σ is the harmonic mean of $\{\max(\sigma_i, \delta)\}_{i=1}^n$: $H_\delta^\sigma = n/\sum_{i=1}^n \frac{1}{\max(\sigma_i,\delta)}$.

Proof By Lemma 3 and Lemma 4, with probability at least $1 - 1/n^3$,

$$\left|\sum_{i=1}^n z_i\right| \leq \left|\sum_{i=1}^n \bar{z}_i\right| + \sum_{i=1}^n |\mathbb{E}z_i|$$

$$\leq C\delta\sqrt{n\ln n} + \frac{e_t}{n} + e_t\left(n - \frac{\delta}{5}\sum_{i=1}^n \frac{1}{\max(\sigma_i,\delta)}\right).$$

This leads to the final bound. ∎

Now we are ready to prove Theorem 1.

At stage $k = 0$, we have $\delta^{(k)} \geq 2|\mu_0^{(k)} - \mu^\star|$. Suppose this is true for stage $k < K$, we show that it is true for $k + 1$.

In stage k, we have $\delta^{(k)} \geq 2e_t$ for $t = 0$. Suppose this is true for a step $t \leq T$, we show that it is true for $t + 1$. Let $m(\delta) = \max\{i : \sigma_{(i)} \leq \delta\}$. We have

$$H_\delta^\sigma = \frac{n}{\sum_{i=1}^n \frac{1}{\max(\sigma_i,\delta)}} \leq \frac{n}{m(\delta)\frac{1}{\delta}} = \frac{n\delta}{m(\delta)}.$$

Then by Lemma 5,

$$e_{t+1} \leq C\delta^{(k)}\sqrt{\frac{\ln n}{n}} + e_t\left(1 - \frac{m(\delta^{(k)})}{5n}\right).$$

If $e_t \gtrsim \frac{\delta^{(k)}\sqrt{n\ln n}}{m(\delta^{(k)})}$, $e_{t+1} \leq e_t \leq \delta^{(k)}/2$. If $e_t \lesssim \frac{\delta^{(k)}\sqrt{n\ln n}}{m(\delta^{(k)})}$, we have $e_{t+1} \leq e_t + C\delta^{(k)}\sqrt{\frac{\ln n}{n}} \lesssim \delta^{(k)}\frac{\sqrt{n\ln n}}{m(\delta^{(k)})} + \delta^{(k)}\sqrt{\frac{\ln n}{n}} \lesssim \delta^{(k)}\frac{\sqrt{n\ln n}}{m(\delta^{(k)})} \leq \delta^{(k)}/4$. Therefore, we can always guarantee $e_t \leq \delta^{(k)}/2$ for $t \leq T$. Then Lemma 5 can be applied for all $t \leq T$, and thus after T iterations,

$$e_T \leq C\delta^{(k)}\sqrt{\frac{\ln n}{n}}\sum_{i=0}^{T-1}\left(1 - \frac{m(\delta^{(k)})}{5n}\right)^i + \left(1 - \frac{m(\delta^{(k)})}{5n}\right)^T e_0 \lesssim \frac{\delta^{(k)}\sqrt{n\ln n}}{m(\delta^{(k)})}.$$

Since $m(\delta^{(k)}) \geq m$, $e_T < \delta^{(k)}/4$, so $\delta^{(k+1)} = \delta^{(k)}/2 > 2e_t = 2|\mu_0^{(k+1)} - \mu^\star|$.

Therefore, $\delta^{(k)} \geq 2|\mu_0^{(k)} - \mu^\star|$ for all $k \leq K$. Since $1 \leq \delta^{(K)}$, at the end of stage K:

$$e_T \lesssim \frac{\delta^{(K)}\sqrt{n\ln n}}{m(\delta^{(K)})} \lesssim \frac{\sqrt{n\ln n}}{m}.$$

This is $|\hat{\mu} - \mu^\star| \lesssim \frac{\sqrt{n\ln n}}{m}$.

4. Lower Bound

To complement the upper bound, we also provide the following lower bound.

Theorem 2 *Suppose $\sigma_{(m)} \leq 1$.*

- *If $m = \Omega(\ln n)$ and $m = O(n^{1/4})$, then there exist a family of instances and a distribution over these instances such that any estimator has expected error $\Omega\left(\left(\frac{n}{m^4}\right)^{1/2}\right)$.*

- *For any arbitrarily small $\epsilon > 0$, if m is between $\Omega(n^{1/4})$ and $O(n^{1-\epsilon})$, then there exist a family of instances and a distribution over these instances such that any estimator has expected error $\Omega\left(\left(\frac{n}{m^4}\right)^{1/6}\right)$.*

Remark. The lower bound considers two ranges of m. In the first range, the bound is $\tilde{\Omega}(\sqrt{n})$ at one end point $m = \Theta(\ln n)$, and is $\Omega(1)$ at the other end point $m = \Theta(n^{1/4})$. It decreases at a rate of $1/m^2$ as m increases in this range. In the second range, the bound is $\Omega(1)$ at one end point $m = \Theta(n^{1/4})$, and is $\Omega(1/n^{1/2-2\epsilon/3})$ at the other end point $m = \Theta(n^{1-\epsilon})$ (for any arbitrarily small $\epsilon > 0$). It decreases at a rate of $1/m^{2/3}$ as m increases, which is slower than that in the first range. Roughly speaking, the bound excludes the possibility of vanishing error in the first range while still allows that in the second range, and the transition point is $m = \Theta(n^{1/4})$.

Our result extends and improves the lower bound in Chierichetti et al. (2014). Their bound is $\Omega\left(\left(\frac{n}{m^4}\right)^{1/2}\right)$ for m between $\Omega(\ln n)$ and $o(\sqrt{n})$. Our result extends the range of m by including the values between $\Omega(n^{1/2})$ and $O(n^{1-\epsilon})$ (for any arbitrarily small $\epsilon > 0$). It also improves their bound in the range between $\Omega(n^{1/4})$ and $o(n^{1/2})$, by a factor of $\Omega\left(\left(\frac{m^4}{n}\right)^{1/3}\right)$. The improvement is obtained by a tighten analysis in the second range of m, which is discussed below.

4.1. Proof of Theorem 2

Our proof follows the high-level idea of Chierichetti et al. (2014) but with a tightened analysis. We also consider the following distribution over a family of instances: σ_i's are i.i.d. sampled; with probability p, $\sigma_i = \sigma_p$, and with probability $q = 1 - p$, $\sigma_i = \sigma_q$; $\mu^\star$ is uniform over $\{+L, -L\}$. Here, $p = m/n$, $\sigma_p = 1$, while σ_q, L are parameters to be chosen.

The goal is then to choose σ_q, L (based on n, m), such that conditioned on $\mu^\star = +L$ or $\mu^\star = -L$, the other choice of mean has a higher likelihood with a constant probability. If this is true, then any estimator has an expected error $\Omega(L)$ over the above distribution on the instances and the randomness of the sample points. When m large enough, the probability that $\sigma_{(m/2)} > 1$ is exponentially small. Then on the distribution over the instances conditioned on $\sigma_{(m/2)} \leq 1$, the lower bound holds under the assumption $\sigma_{(m/2)} \leq 1$. By changing the variable m to $2m$, the theorem follows.

We improve over Chierichetti et al. (2014) by noting that, roughly speaking, the requirement on σ_q when $m = \Omega(n^{1/4})$ is more relaxed compared to that when $m = O(n^{1/4})$. This allows us to set σ_q differently to get improved results and also over a more general range of m, as detailed below.

Following the idea above, denote the likelihood of the mean being L as $\mathcal{L}_+$, and the likelihood of the mean being $-L$ as $\mathcal{L}_-$. We will show that the log-likelihood ratio has sufficiently large variances so can be negative or positive with constant probabilities. From now on, we condition on

the true mean is L (the proof for the case with $-L$ is symmetric). Let $S_p = \{i : \sigma_i = \sigma_p\}$ and $S_q = \{i : \sigma_i = \sigma_q\}$. Define

$$N_i = \frac{p/\sigma_p}{q/\sigma_q} \exp\left(-\frac{(x_i - L)^2}{2}\left(\frac{1}{\sigma_p^2} - \frac{1}{\sigma_q^2}\right)\right) \tag{8}$$

$$D_i = \frac{p/\sigma_p}{q/\sigma_q} \exp\left(-\frac{(x_i + L)^2}{2}\left(\frac{1}{\sigma_p^2} - \frac{1}{\sigma_q^2}\right)\right). \tag{9}$$

Then we have

$$\ln\frac{\mathcal{L}_+}{\mathcal{L}_-} = \sum_{i=1}^{n}\left(\ln\frac{1+N_i}{1+D_i} + \frac{2L}{\sigma_q^2}x_i\right) = \underbrace{\sum_{i\in S_p}\left(\ln\frac{1+N_i}{1+D_i} + \frac{2L}{\sigma_q^2}x_i\right)}_{X_p} + \underbrace{\sum_{i\in S_q}\left(\ln\frac{1+N_i}{1+D_i} + \frac{2L}{\sigma_q^2}x_i\right)}_{X_q}.$$

We next bound X_q and X_p respectively. The road map is to show that X_q has sufficiently large variances so can make the log-likelihood ratio negative with constant probability, shown via the Berry-Essen Theorem. This requires computing the moments, so we first approximate $\ln\frac{1+N_i}{1+D_i}$ via the Taylor expansion of the function $\ln(1+x)$, and then compute the moments of the approximation. When $m = \Omega(n^{1/4})$, the likelihood of $x_i \in S_q$ is comparable to that of $x_i \in S_q$, so their ratio (as in (8) or (9)) is in the same order as a constant. We thus use a tighter approximation for $\ln(1 + N_i)$ and $\ln(1 + D_i)$ in the log-likelihood ratio, and improve over Chierichetti et al. (2014).

Lemma 6 *Suppose the mean is L, and $q > C_q p$, $\sigma_q > C_\sigma \sigma_p$, $L < c_L \sigma_q$ for sufficiently large absolute constants C_q, C_σ and a sufficiently small absolute constant c_L.[1] Suppose $\frac{p/\sigma_p}{q/\sigma_q} < c_\alpha$ for a sufficiently small absolute constant $c_\alpha < 1$. Let t be a positive integer. Let $V_i = \sum_{j=1}^{2t-1}(-1)^{j+1}(N_i^j - D_i^j)/j$ and $Y_i = \frac{2L}{\sigma_q^2}x_i + V_i$. Then for $i \in S_q$,*

$$\mathbb{E}[Y_i] \lesssim \frac{L^2}{\sigma_q^2}, \quad \mathbb{E}[Y_i^2] \simeq \frac{p^2/\sigma_p}{q^2/\sigma_q}\min\left\{1, \frac{L^2}{\sigma_p^2}\right\} + \frac{L^2}{\sigma_q^2}.$$

And with probability at least $1 - n^{-\Theta(1)} - \exp\left(\Theta\left(U_q\right)\right)$, $\left|X_q - \sum_{i\in S_q}Y_i\right| \lesssim U_q := \left(\frac{p/\sigma_p}{q/\sigma_q}\right)^{2t}\frac{\sigma_p}{\sigma_q}n$. Also, with probability at least $1 - c$ for a sufficiently small absolute constant c, $\left|X_q - \sum_{i\in S_q}Y_i\right| \lesssim U_q' := \left(\frac{p/\sigma_p}{q/\sigma_q}\right)^{2t}\left(\frac{\sigma_p}{\sigma_q}n + \sqrt{n}\right)$.

Lemma 7 *Under the same conditions as in Lemma 6, for $i \in S_p$,*

$$\mathbb{E}[Y_i] \lesssim \frac{p/\sigma_p}{q/\sigma_q}\min\left\{1, \frac{L^2}{\sigma_p^2}\right\}, \quad \mathbb{E}[Y_i^2] \simeq \frac{L^2\sigma_p^2}{\sigma_q^4} + \frac{L^4}{\sigma_q^4} + \left(\frac{p/\sigma_p}{q/\sigma_q}\right)^2\min\left\{1, \frac{L^2}{\sigma_p^2}\right\} + \frac{p/\sigma_p}{q/\sigma_q}\frac{L^2}{\sigma_q^2}.$$

And with probability at least $1 - c$ for a sufficiently small absolute constant c, $\left|X_p - \sum_{i\in S_p}Y_i\right| \lesssim U_p := \left(\frac{p/\sigma_p}{q/\sigma_q}\right)^{2t}pn$.

---

1. C_q is a constant chosen for the inequality $q > C_q p$. It doesn't depend on the value of q. Similar for C_σ, c_L etc.

Now define

$$Z_i = Y_i - \mathbb{E}[Y_i], \quad Z = \frac{1}{\sqrt{M_2|S_q|}} \sum_{i \in S_q} Z_i.$$

To apply the Berry-Essen Theorem, we bound the first three moments of Z_i. Clearly, $\mathbb{E}[Z_i] = 0$.

Lemma 8 *Under the same conditions as in Lemma 6, for $i \in S_q$,*

$$M_2 := \mathbb{E}[Z_i^2] \simeq \frac{p^2/\sigma_p}{q^2/\sigma_q} \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} + \frac{L^2}{\sigma_q^2},$$

$$M_3 := \mathbb{E}[|Z_i|^3] \lesssim \frac{p^3/\sigma_p^2}{q^3/\sigma_q^2} \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} + \frac{p^2/\sigma_p^2}{q^2\sigma_q} L^2$$

$$+ \frac{p/\sigma_p}{q\sigma_q^4} L^2(\sigma_p^3 + \sigma_p^2 L + \sigma_p L^2) + \frac{L^3}{\sigma_q^3}.$$

By the Berry-Essen Theorem, conditioned on S_q, the CDF $F(t)$ of Z satisfies $|F(t) - \Phi(t)| \lesssim \frac{M_3}{\sqrt{M_2^3|S_q|}}$ where $\Phi(t)$ is the CDF of a standard normal distribution. By the Chernoff's bound, with probability $1 - n^{\Theta(1)}$, $|S_p| \simeq pn, |S_q| \simeq qn \simeq n$. Assume this is true in the rest of the proof.

Now we consider different cases for p and set σ_p, σ_q and L accordingly.

Case 1. Suppose $p \geq \Omega(\ln n/n)$ and $p \leq \frac{c_p}{n^{3/4}}$ for some sufficiently small constant $c_p > 0$. Then set $\sigma_p = 1$, $\sigma_q = C_\sigma/(p^2 n)$ and $L = c_L/(p^2 n^{3/2}) \simeq \sigma_q/\sqrt{n}$ for some sufficiently large constant $C_\sigma > 0$ and some sufficiently small constant $c_L > 0$. Set $t = 1$. Then

$$M_2 \simeq p^2\sigma_q + \frac{L^2}{\sigma_q^2} \simeq \frac{1}{n}.$$

$$M_3 \lesssim p^3\sigma_q^2 + p^2 \frac{L^2}{\sigma_q} + \frac{p}{\sigma_q^4} L^2(1 + L + L^2) + \frac{L^3}{\sigma_q^3} \simeq \frac{1}{pn^2} + \frac{1}{n^{3/2}}.$$

Then conditioned on $|S_q| > qn/2 > n/4$, we have $\frac{M_3}{\sqrt{M_2^3|S_q|}} \lesssim \frac{1}{pn} = o(1)$. Then we have for constants $C_Z > 0$ and $c_z > 0$, $\Pr[Z \leq -C_Z] = \Pr\left[\sum_{i \in S_q} Z_i \leq -C_Z\sqrt{M_2|S_q|}\right] \geq c_z$. So with a constant probability, $-\sum_{i \in S_q} Z_i \geq C_Z\sqrt{M_2 n} \simeq C_Z$. We also have with probability $1 - c$ for a sufficiently small absolute constant c,

$$\sum_{i \in S_q} \mathbb{E}[Y_i] \lesssim \frac{L^2}{\sigma_q^2} qn \simeq 1,$$

$$U_q = \left(\frac{p/\sigma_p}{q/\sigma_q}\right)^2 \frac{\sigma_p}{\sigma_q} n \simeq \frac{p^2/\sigma_p}{q^2/\sigma_q} n \simeq 1,$$

$$U_p = \left(\frac{p/\sigma_p}{q/\sigma_q}\right)^2 pn \simeq \frac{1}{pn} = o(1),$$

$$\sum_{i \in S_p} \mathbb{E}[Y_i] \lesssim pn \frac{p/\sigma_p}{q/\sigma_q} \simeq p^2\sigma_q n \simeq 1,$$

$$\sum_{i \in S_p} \mathbb{E}[Y_i^2] \lesssim pn\left(\frac{L^2\sigma_p^2}{\sigma_q^4} + \frac{L^4}{\sigma_q^4} + \left(\frac{p/\sigma_p}{q/\sigma_q}\right)^2 + \frac{p/\sigma_p}{q/\sigma_q}\frac{L^2}{\sigma_q^2}\right)$$

11

$$\lesssim pn\left(\frac{1}{\sigma_q^2 n} + \frac{1}{n^2} + p^2\sigma_q^2 + \frac{p\sigma_q}{n}\right) \lesssim \frac{1}{pn} = o(1).$$

Therefore, with a constant probability, $\ln\frac{\mathcal{L}_+}{\mathcal{L}_-}$ is negative. The expected error $\mathbb{E}|\hat{\mu} - \mu^*|$ of any estimator $\hat{\mu}$ is $\Omega(L) = \Omega(1/(p^2 n^{3/2})) = \Omega(\sqrt{n}/m^2)$.

Case 2. Suppose $p \geq \frac{C_p}{n^{3/4}}$ and $p < \frac{c_p}{n^{2/t}}$ for some sufficiently large absolute constant C_p and sufficiently small absolute constant c_p. Then set $\sigma_p = 1$, $\sigma_q = C_\sigma/p^{2/3}$ and $L = c_L/(p^{2/3} n^{1/2}) \simeq \sigma_q/\sqrt{n}$ for some sufficiently large constant $C_\sigma > 0$ and some sufficiently small constant $c_L > 0$. Then

$$M_2 \simeq p^2\sigma_q L^2 + \frac{L^2}{\sigma_q^2} \simeq \frac{1}{n}.$$

$$M_3 \lesssim p^3\sigma_q^2 L^2 + p^2\frac{L^2}{\sigma_q} + \frac{p}{\sigma_q^4}L^2 + \frac{L^3}{\sigma_q^3} \simeq \frac{p^{1/3}}{n} + \frac{1}{n^{3/2}}.$$

Then conditioned on $|S_q| > qn/2 > n/4$, we have $\frac{M_3}{\sqrt{M_2^3|S_q|}} = o(1)$. Then we have for constants $C_Z > 0$ and $c_z > 0$, $\Pr[Z \leq -C_Z] = \Pr\left[\sum_{i\in S_q} Z_i \leq -C_Z\sqrt{M_2|S_q|}\right] \geq c_z$. So with a constant probability, $-\sum_{i\in S_q} Z_i \geq C_Z\sqrt{M_2 n} \simeq C_Z$. We also have with probability $1 - c$ for a sufficiently small absolute constant c,

$$\sum_{i\in S_q}\mathbb{E}[Y_i] \lesssim \frac{L^2}{\sigma_q^2}qn \simeq 1,$$

$$U_q' = \left(\frac{p/\sigma_p}{q/\sigma_q}\right)^{2t}\left(\frac{\sigma_p}{\sigma_q}n + \sqrt{n}\right) \lesssim 1,$$

$$U_p = \left(\frac{p/\sigma_p}{q/\sigma_q}\right)^{2t} pn \lesssim 1,$$

$$\sum_{i\in S_p}\mathbb{E}[Y_i] \lesssim pn\frac{p/\sigma_p}{q/\sigma_q}L^2 \simeq 1,$$

$$\sum_{i\in S_p}\mathbb{E}[Y_i^2] \lesssim pn\left(\frac{L^2\sigma_p^2}{\sigma_q^4} + \frac{L^4}{\sigma_q^4} + \left(\frac{p/\sigma_p}{q/\sigma_q}\right)^2 L^2 + \frac{p/\sigma_p}{q/\sigma_q}\frac{L^2}{\sigma_q^2}\right)$$

$$\lesssim pn\left(\frac{1}{\sigma_q^2 n} + \frac{1}{n^2} + p^2\sigma_q^2 L^2 + \frac{p\sigma_q}{n}\right) \lesssim p^{1/3} = o(1).$$

Therefore, with a constant probability, $\ln\frac{\mathcal{L}_+}{\mathcal{L}_-}$ is negative. The expected error $\mathbb{E}|\hat{\mu} - \mu^*|$ of any estimator $\hat{\mu}$ is $\Omega(L) = \Omega(1/(p^{2/3} n^{1/2})) = \Omega(n^{1/6}/m^{2/3})$.

5. Conclusion

This work considered mean estimation in the setting of entangled single-sampled Gaussians where given one sample from each of n Gaussians with a common mean but different variances, the goal is to learn the mean. It studied the subset-of-signals model where an unknown subset of m variances are bounded, and proved upper and lower bounds, which are summarized in Figure 1. A natural future direction is to close the gap between the upper bound and the lower bound.

Acknowledgement

References

Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.

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

Mikhail Belkin and Kaushik Sinha. Toward learning gaussian mixtures with arbitrary separation. In *COLT*, pages 407–419. Citeseer, 2010b.

Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2755–2771. SIAM, 2019.

Flavio Chierichetti, Anirban Dasgupta, Ravi Kumar, and Silvio Lattanzi. Learning entangled single-sample gaussians. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 511–522. Society for Industrial and Applied Mathematics, 2014.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE, 1999.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 999–1008. JMLR. org, 2017.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2683–2702. Society for Industrial and Applied Mathematics, 2018.

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

David Hong, Laura Balzano, and Jeffrey A Fessler. Asymptotic performance of pca for high-dimensional heteroscedastic data. *Journal of multivariate analysis*, 167:435–452, 2018a.

David Hong, Jeffrey A Fessler, and Laura Balzano. Optimally weighted pca for high-dimensional heteroscedastic data. *arXiv preprint arXiv:1810.12862*, 2018b.

Peter J Huber. *Robust statistics*. Springer, 2011.

Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562. ACM, 2010.

Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. In *International Conference on Computational Learning Theory*, pages 444–457. Springer, 2005.

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

Alvaro Munoz, Bernard Rosner, and Vincent Carey. Regression analysis in the presence of heterogeneous intraclass correlations. *Biometrics*, pages 653–658, 1986.

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

Arora Sanjeev and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. ACM, 2001.

Leslie G Valiant. Learning disjunction of conjunctions. In *IJCAI*, pages 560–566. Citeseer, 1985.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

Donatella Vicari and Maurizio Vichi. Multivariate linear regression for heterogeneous data. *Journal of Applied Statistics*, 40(6):1209–1230, 2013.

Hui Yuan and Yingyu Liang. Learning entangled single-sample distributions via iterative trimming. In *The 23rd International Conference on Artificial Intelligence and Statistics*, 2020.

Anru Zhang, T Tony Cai, and Yihong Wu. Heteroskedastic pca: Algorithm, optimality, and applications. *arXiv preprint arXiv:1810.08316*, 2018.

Appendix A. Proofs for Upper Bound

A.1. Proof of Lemma 3

Note that $z_i(\mu)$ is 1-Lipschitz w.r.t. μ. So a standard ϵ-net argument over the interval $[\mu^\star - \delta_e, \mu^\star + \delta_e]$ gives the bound.

More precisely, let $\mathcal{X}$ be an ϵ-net over $[\mu^\star - \delta_e, \mu^\star + \delta_e]$, with $\epsilon = \delta_e / n^2$. A standard construction gives $|\mathcal{X}| < 2n^2$. For a fixed $\mu \in \mathcal{X}$, we have

$$\mu - \delta - \mu^\star \leq \mathbb{E}z_i(\mu) \leq \mu + \delta - \mu^\star, \tag{10}$$

$$-2\delta \leq \bar{z}_i(\mu) \leq 2\delta. \tag{11}$$

Since $\bar{z}_i(\mu)$ is bounded, we have by sub-Gaussian properties (see, e.g., Section 2.5 and 2.6 of Vershynin (2018)),

$$\|\bar{z}_i(\mu)\|_{\psi_2} \lesssim \delta, \tag{12}$$

and we have with probability at least $1 - 1/n^6$, for the fixed μ,

$$\left| \sum_{i=1}^{n} \bar{z}_i(\mu) \right| \lesssim \delta\sqrt{n \ln n}. \tag{13}$$

Taking a union bound over $\mathcal{X}$, we have with probability at least $1 - 1/n^3$, for all $\mu \in \mathcal{X}$,

$$\left| \sum_{i=1}^{n} \bar{z}_i(\mu) \right| \lesssim \delta\sqrt{n \ln n}. \tag{14}$$

For any $\mu' \notin \mathcal{X}$, there is $\mu \in \mathcal{X}$ satisfying $|\mu' - \mu| \leq \epsilon$. Therefore,

$$\left| \sum_{i=1}^{n} \bar{z}_i(\mu') \right| \leq \left| \sum_{i=1}^{n} \bar{z}_i(\mu) \right| + \left| \sum_{i=1}^{n} \bar{z}_i(\mu') - \bar{z}_i(\mu) \right| \tag{15}$$

$$\leq \left| \sum_{i=1}^{n} \bar{z}_i(\mu) \right| + \left| \sum_{i=1}^{n} z_i(\mu') - z_i(\mu) \right| + \left| \sum_{i=1}^{n} \mathbb{E}[z_i(\mu') - z_i(\mu)] \right| \tag{16}$$

$$\leq \left| \sum_{i=1}^{n} \bar{z}_i(\mu) \right| + \epsilon n + \epsilon n \tag{17}$$

$$\lesssim \delta\sqrt{n \ln n} + \delta_e/n. \tag{18}$$

This completes the proof.

A.2. Proof of Lemma 4

W.L.O.G., suppose $\mu \geq \mu^\star$, and let $\delta_e = |\mu - \mu^\star|$. Let z_i be a shorthand for $z_i(\mu)$. Let $g(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), a_i = \frac{\delta_e - \delta}{\sigma_i}, b_i = \frac{\delta_e + \delta}{\sigma_i}$. Then

$$|\mathbb{E}z_i| = \mathbb{E}z_i \tag{19}$$

$$= (\delta_e - \delta) \int_{-\infty}^{\delta_e - \delta} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{x^2}{2\sigma_i^2}\right) dx \tag{20}$$

$$+ \int_{\delta_e - \delta}^{\delta_e + \delta} x \cdot \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{x^2}{2\sigma_i^2}\right) dx \tag{21}$$

$$+ (\delta_e + \delta) \int_{\delta_e + \delta}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{x^2}{2\sigma_i^2}\right) dx \tag{22}$$

$$= (\delta_e - \delta) \int_{-\infty}^{a_i} g(x) dx \tag{23}$$

$$+ \int_{a_i}^{b_i} \sigma_i x g(x) dx \tag{24}$$

$$+ (\delta_e + \delta) \int_{b_i}^{+\infty} g(x) dx \tag{25}$$

$$= (\delta_e - \delta) \int_{-b_i}^{a_i} g(x) dx + (\delta_e - \delta) \int_{-\infty}^{-b_i} g(x) dx \tag{26}$$

$$+ \int_{-a_i}^{b_i} \sigma_i x g(x) dx \tag{27}$$

$$+ (\delta_e + \delta) \int_{b_i}^{+\infty} g(x) dx \tag{28}$$

$$= (\delta_e - \delta) \int_{-a_i}^{b_i} g(x) dx + (\delta_e - \delta) \int_{b_i}^{+\infty} g(x) dx \tag{29}$$

$$+ \int_{-a_i}^{b_i} \sigma_i x g(x) dx \tag{30}$$

$$+ (\delta_e + \delta) \int_{b_i}^{+\infty} g(x) dx \tag{31}$$

$$= \int_{-a_i}^{b_i} [\sigma_i x + (\delta_e - \delta)] g(x) dx + 2\delta_e \int_{b_i}^{+\infty} g(x) dx. \tag{32}$$

We consider two cases.

Case 1: $\delta \geq \delta_e$. Then $-a_i \geq 0$.

$$\int_{-a_i}^{b_i} [\sigma_i x + (\delta_e - \delta)] g(x) dx \tag{33}$$

$$= \int_{-a_i}^{b_i} \delta_e g(x) dx + \int_{-a_i}^{\frac{b_i - a_i}{2}} [\sigma_i x - \delta] g(x) dx + \int_{\frac{b_i - a_i}{2}}^{b_i} [\sigma_i x - \delta] g(x) dx \tag{34}$$

$$= \int_{-a_i}^{b_i} \delta_e g(x) dx - \int_{\frac{b_i - a_i}{2}}^{b_i} [\sigma_i y - \delta] g\left(\frac{2\delta}{\sigma_i} - y\right) dy + \int_{\frac{b_i - a_i}{2}}^{b_i} [\sigma_i x - \delta] g(x) dx \tag{35}$$

$$\leq \int_{-a_i}^{b_i} \delta_e g(x) dx. \tag{36}$$

Therefore,

$$|\mathbb{E}z_i| \le \delta_e \int_{-a_i}^{b_i} g(x)dx + 2\delta_e \int_{b_i}^{+\infty} g(x)dx \tag{37}$$

$$= \delta_e \left(1 - \int_{-b_i}^{-a_i} g(x)dx \right) \tag{38}$$

$$= \delta_e \left(1 - \int_0^{b_i} g(x)dx \right). \tag{39}$$

Case 2: $\delta < \delta_e$. Then $a_i > 0$.

$$\int_{-a_i}^{b_i} [\sigma_i x + (\delta_e - \delta)]g(x)dx \tag{40}$$

$$= \int_{-a_i}^{a_i} [\sigma_i x + (\delta_e - \delta)]g(x)dx + \int_{a_i}^{b_i} [\sigma_i x + (\delta_e - \delta)]g(x)dx \tag{41}$$

$$= \int_{-a_i}^{a_i} (\delta_e - \delta)g(x)dx + \int_{a_i}^{b_i} [\sigma_i x + (\delta_e - \delta)]g(x)dx. \tag{42}$$

Then second term can be bounded as in Case 1.

$$\int_{a_i}^{b_i} [\sigma_i x + (\delta_e - \delta)]g(x)dx \le \int_{a_i}^{b_i} \delta_e g(x)dx. \tag{43}$$

Therefore,

$$|\mathbb{E}z_i| \le \int_{-a_i}^{a_i} (\delta_e - \delta)g(x)dx + \delta_e \int_{a_i}^{b_i} g(x)dx + 2\delta_e \int_{b_i}^{+\infty} g(x)dx \tag{44}$$

$$= \int_{-a_i}^{a_i} (\delta_e - \delta)g(x)dx + \delta_e - \delta_e \int_{-b_i}^{a_i} g(x)dx \tag{45}$$

$$= -\delta \int_{-a_i}^{a_i} g(x)dx + \delta_e - \delta_e \int_{-b_i}^{-a_i} g(x)dx \tag{46}$$

$$\le \delta_e - \delta \int_{-b_i}^{a_i} g(x)dx \tag{47}$$

$$\le \delta_e \left(1 - \frac{\delta}{\delta_e} \int_0^{b_i} g(x)dx \right). \tag{48}$$

In summary, for both cases, we have

$$|\mathbb{E}z_i| \le \delta_e \left(1 - \frac{\delta}{\max\{\delta_e, \delta\}} \int_0^{b_i} g(x)dx \right). \tag{49}$$

To simplify the bound, we consider two cases. If $\sigma_i \le \delta + \delta_e$, then $b_i \ge 1$, and

$$\int_0^{b_i} g(x)dx \ge \int_0^1 g(x)dx \ge 1/2. \tag{50}$$

17

If $\sigma_i > \delta + \delta_e$, then

$$\int_0^{b_i} g(x)dx \geq b_i g(b_i) \tag{51}$$

$$= \frac{1}{\sqrt{2\pi}} \frac{\delta + \delta_e}{\sigma_i} \exp\left(-\frac{(\delta + \delta_e)^2}{2\sigma_i^2}\right) \tag{52}$$

$$\geq \frac{\delta + \delta_e}{5\sigma_i} \geq \frac{\delta}{5\sigma_i}. \tag{53}$$

Then for both cases, we have

$$|\mathbb{E}z_i| \leq \delta_e \left(1 - \frac{\delta}{\max\{\delta_e, \delta\}} \frac{\delta}{5\max\{\sigma_i, \delta\}}\right). \tag{54}$$

Appendix B. Proofs for Lower Bound

For convenience, define

$$\alpha = \frac{p/\sigma_p}{q/\sigma_q}, \tag{55}$$

$$\beta = \frac{2L}{\sigma_q^2}, \tag{56}$$

$$\gamma = \frac{\sigma_p}{\sigma_q}, \tag{57}$$

$$\frac{1}{\sigma_{pq}^2} = \frac{1}{\sigma_p^2} - \frac{1}{\sigma_q^2}. \tag{58}$$

Then we have

$$N_i = \alpha \exp\left(-\frac{(x_i - L)^2}{2\sigma_{pq}^2}\right) \tag{59}$$

$$D_i = \alpha \exp\left(-\frac{(x_i + L)^2}{2\sigma_{pq}^2}\right) \tag{60}$$

and

$$\ln \frac{\mathcal{L}_+}{\mathcal{L}_-} = \sum_{i=1}^n \ln \frac{1 + N_i}{1 + D_i} + \beta x_i \tag{61}$$

$$= \underbrace{\sum_{i \in S_p} \left(\ln \frac{1 + N_i}{1 + D_i} + \beta x_i\right)}_{X_p} + \underbrace{\sum_{i \in S_q} \left(\ln \frac{1 + N_i}{1 + D_i} + \beta x_i\right)}_{X_q}. \tag{62}$$

B.1. Proof of Lemma 6

Lemma 9 *Suppose the mean is L. For a positive integer j, N_i^j and D_i^j are sub-Gaussian with norms*

$$\|N_i^j\|_{\psi_2} \lesssim \left(\frac{p/\sigma_p}{q/\sigma_q}\right)^j, \quad \|D_i^j\|_{\psi_2} \lesssim \left(\frac{p/\sigma_p}{q/\sigma_q}\right)^j. \tag{63}$$

Proof Recall that if the moments of a random variable X satisfy $\|X\|_{L_p} = (\mathbb{E}|X|^p)^{1/p} \leq K\sqrt{p}$ for all $p \geq 1$, then $\|X\|_{\psi_2} \lesssim K$. The lemma then follows from Lemma 10. ∎

Since for any $x > 0$,

$$\sum_{j=1}^{2t} \frac{(-1)^{j+1}x^j}{j} \leq \ln(1+x) \leq \sum_{j=1}^{2t-1} \frac{(-1)^{j+1}x^j}{j}, \tag{64}$$

we have

$$V_i - \frac{N_i^{2t}}{2t} \leq \ln \frac{1+N_i}{1+D_i} \leq V_i + \frac{D_i^{2t}}{2t}, \tag{65}$$

and thus

$$\left| \sum_{i \in S_q} \ln \frac{1+N_i}{1+D_i} + \beta x_i - \sum_{i \in S_q} Y_i \right| \leq \sum_{i \in S_q} \max \left\{ \frac{N_i^{2t}}{2t}, \frac{D_i^{2t}}{2t} \right\}. \tag{66}$$

By Lemma 10 and 11, for $i \in S_q$,

$$\mathbb{E}[N_i^j] = \alpha^j \frac{\sigma_{pq}}{\sqrt{\sigma_{pq}^2 + j\sigma_q^2}}, \tag{67}$$

$$\mathbb{E}[D_i^j] = \alpha^j \frac{\sigma_{pq}}{\sqrt{\sigma_{pq}^2 + j\sigma_q^2}} \exp\left(-\frac{2jL^2}{\sigma_{pq}^2 + j\sigma_q^2}\right) \leq \mathbb{E}[N_i^j], \tag{68}$$

$$\mathbb{E}[N_i^j D_i^j] = \alpha^{2j} \frac{\sigma_{pq}}{\sqrt{\sigma_{pq}^2 + 2j\sigma_q^2}} \exp\left(-2jL^2 \frac{\sigma_{pq}^2 + j\sigma_q^2}{\sigma_{pq}^4 + 2j\sigma_{pq}^2\sigma_q^2}\right). \tag{69}$$

By the Chernoff's bound, with probability $1 - n^{\Theta(1)}$, $|S_q| \simeq qn \simeq n$. Conditioned on S_q, we have with probability at least $1 - e^{\Theta(\alpha^{2t}\gamma n)}$,

$$\max \left\{ \sum_{i \in S_q} \frac{N_i^{2t}}{2t}, \sum_{i \in S_q} \frac{D_i^{2t}}{2t} \right\} \leq 2 \sum_{i \in S_q} \mathbb{E}N_i^{2t} \simeq \alpha^{2t}\gamma n. \tag{70}$$

So with probability $1 - n^{\Theta(1)} - e^{\Theta(\alpha^{2t}\gamma n)}$,

$$\left| \sum_{i \in S_q} \ln \frac{1+N_i}{1+D_i} + \beta x_i - \sum_{i \in S_q} Y_i \right| \lesssim \alpha^{2t}\gamma n. \tag{71}$$

Now consider Y_i. Since p is sufficiently small compared to q and L is sufficiently small compared to σ_q, and $\alpha < c_\alpha$ for some sufficiently small absolute constant $c_\alpha < 1$, we have

$$\mathbb{E}[Y_i] \lesssim \beta L + \sum_{j=1}^{2t-1} \alpha^j \gamma \frac{L^2}{\sigma_q^2} \lesssim \frac{L^2}{\sigma_q^2}\left(1 + \sum_{j=1}^{2t-1} \alpha^j \gamma\right) \tag{72}$$

19

$$\lesssim \frac{L^2}{\sigma_q^2}. \tag{73}$$

Let $V_{ij} = (-1)^{j+1}(N_i^j - D_i^j)/j$, then $Y_i = \beta x_i + \sum_{j=1}^{2t-1} V_{ij}$. By Lemma 10, 11, 12, and that $\alpha < c_\alpha$ for some sufficiently small absolute constant $c_\alpha < 1$,

$$\mathbb{E}[Y_i^2] = \mathbb{E}\left[\beta^2 x_i^2 + \sum_{j=1}^{2t-1} V_{ij}^2 + 2\sum_{j=1}^{2t-1} \beta x_i V_{ij} + 2\sum_{j<k;j,k=1}^{2t-1} V_{ij}V_{ik}\right] \tag{74}$$

$$\simeq \beta^2 \sigma_q^2 + \sum_{j=1}^{2t-1} \alpha^{2j}\gamma \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} \tag{75}$$

$$+ \sum_{j=1}^{2t-1} (-1)^{j+1}\alpha^j \beta\gamma L + \sum_{j<k;j,k=1}^{2t-1} (-1)^{j+k}\alpha^{j+k}\gamma \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} \tag{76}$$

$$\simeq \beta^2 \sigma_q^2 + \alpha^2\gamma \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} + \alpha\beta\gamma L \tag{77}$$

$$\simeq \frac{L^2}{\sigma_q^2} + \frac{p^2/\sigma_p}{q^2/\sigma_q} \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} + \frac{p}{q}\frac{L^2}{\sigma_q^2} \tag{78}$$

$$\simeq \frac{p^2/\sigma_p}{q^2/\sigma_q} \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} + \frac{L^2}{\sigma_q^2} \tag{79}$$

where the last line follows from $p < q$.

B.2. Proof of Lemma 7

The proof is similar to that of Lemma 6.

Again, we have

$$\left|\sum_{i\in S_p} \ln\frac{1+N_i}{1+D_i} + \beta x_i - \sum_{i\in S_p} Y_i\right| \le \sum_{i\in S_p} \max\left\{\frac{N_i^{2t}}{2t}, \frac{D_i^{2t}}{2t}\right\}. \tag{80}$$

By Lemma 10 and 11, for $i \in S_p$,

$$\mathbb{E}[N_i^j] = \alpha^j \frac{\sigma_{pq}}{\sqrt{\sigma_{pq}^2 + j\sigma_p^2}}, \tag{81}$$

$$\mathbb{E}[D_i^j] = \alpha^j \frac{\sigma_{pq}}{\sqrt{\sigma_{pq}^2 + j\sigma_p^2}} \exp\left(-\frac{2jL^2}{\sigma_{pq}^2 + j\sigma_p^2}\right) \le \mathbb{E}[N_i^j], \tag{82}$$

$$\mathbb{E}[N_i^j D_i^j] = \alpha^{2j} \frac{\sigma_{pq}}{\sqrt{\sigma_{pq}^2 + 2j\sigma_p^2}} \exp\left(-2jL^2\frac{\sigma_{pq}^2 + j\sigma_p^2}{\sigma_{pq}^4 + 2j\sigma_{pq}^2\sigma_p^2}\right). \tag{83}$$

Conditioned on S_p, by Lemma 9, we have with probability at least $1 - \delta$,

$$\sum_{i\in S_p} \max\left\{\frac{N_i^{2t}}{2t}, \frac{D_i^{2t}}{2t}\right\} \lesssim \frac{\alpha^{2t}}{2t}|S_p| + \alpha^{2t}\sqrt{|S_p|\log\frac{1}{\delta}}. \tag{84}$$

By the Chernoff's bound, with probability $1 - n^{\Theta(1)}$, $|S_p| \simeq pn$. So with probability $1 - n^{\Theta(1)} - c$ for a sufficiently small absolute constant c,

$$\left| \sum_{i \in S_p} \ln \frac{1 + N_i}{1 + D_i} + \beta x_i - \sum_{i \in S_p} Y_i \right| \lesssim \alpha^{2t} pn + \alpha^{2t} \sqrt{pn}. \tag{85}$$

Now consider Y_i. Since p is sufficiently small compared to q and L is sufficiently small compared to σ_q, and $\alpha < c_\alpha$ for some sufficiently small absolute constant $c_\alpha < 1$, we have

$$\mathbb{E}[Y_i] \lesssim \beta L + \sum_{j=1}^{2t-1} \alpha^j \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} \lesssim \frac{L^2}{\sigma_q^2} + \alpha \min\left\{1, \frac{L^2}{\sigma_p^2}\right\}. \tag{86}$$

Let $V_{ij} = (-1)^{j+1}(N_i^j - D_i^j)/j$, then $Y_i = \beta x_i + \sum_{j=1}^{2t-1} V_{ij}$. By Lemma 10, 11, 12, and that $\alpha < c_\alpha$ for some sufficiently small absolute constant $c_\alpha < 1$,

$$\mathbb{E}[Y_i^2] = \mathbb{E}\left[\beta^2 x_i^2 + \sum_{j=1}^{2t-1} V_{ij}^2 + 2\sum_{j=1}^{2t-1} \beta x_i V_{ij} + 2\sum_{j<k; j,k=1}^{2t-1} V_{ij} V_{ik}\right] \tag{87}$$

$$\simeq \beta^2(\sigma_p^2 + L^2) + \sum_{j=1}^{2t-1} \alpha^{2j} \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} \tag{88}$$

$$+ \sum_{j=1}^{2t-1} (-1)^{j+1} \alpha^j \beta L + \sum_{j<k; j,k=1}^{2t-1} (-1)^{j+k} \alpha^{j+k} \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} \tag{89}$$

$$\simeq \beta^2(\sigma_p^2 + L^2) + \alpha^2 \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} + \alpha\beta L \tag{90}$$

$$\simeq \frac{L^2 \sigma_p^2}{\sigma_q^4} + \frac{L^4}{\sigma_q^4} + \left(\frac{p/\sigma_p}{q/\sigma_q}\right)^2 \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} + \frac{p/\sigma_p}{q/\sigma_q} \frac{L^2}{\sigma_q^2}. \tag{91}$$

where the last line follows from $p < q$.

B.3. Proof of Lemma 8

The second moment is

$$M_2 := \mathbb{E}[Z_i^2] = \mathbb{E}[Y_i^2] - \mathbb{E}^2[Y_i] \tag{92}$$

$$\simeq \frac{p^2/\sigma_p}{q^2/\sigma_q} \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} + \frac{L^2}{\sigma_q^2} - \left(\frac{L^2}{\sigma_q^2}\right)^2 \tag{93}$$

$$\simeq \frac{p^2/\sigma_p}{q^2/\sigma_q} \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} + \frac{L^2}{\sigma_q^2}. \tag{94}$$

where the last line follows from L is sufficiently small compared to σ_q.

To compute the third moment, let $R_i = \beta x_i - \beta\mathbb{E}[x_i] = \beta(x_i - L)$. Then

$$\mathbb{E}[|Z_i|^3] = \mathbb{E}[|Y_i - \mathbb{E}[Y_i]|^3] \tag{95}$$

21

$$= \mathbb{E}[|V_i - \mathbb{E}[V_i] + R_i|^3] \tag{96}$$

$$\leq \mathbb{E}[|V_i - \mathbb{E}[V_i]|^3] + \mathbb{E}[|R_i|^3] \tag{97}$$

$$+ 3\mathbb{E}[|V_i - \mathbb{E}[V_i]|^2 |R_i|] + 3\mathbb{E}[|V_i - \mathbb{E}[V_i]||R_i|^2]. \tag{98}$$

The terms can be bounded respectively.

$$\mathbb{E}[|V_i - \mathbb{E}[V_i]|^3] \leq \mathbb{E}[|V_i - \mathbb{E}[V_i]|^2] \max_{x_i} |V_i - \mathbb{E}[V_i]| \tag{99}$$

$$\lesssim \mathbb{E}[|V_i - \mathbb{E}[V_i]|^2] \max_{j,x_i} \left| (N_i^j - D_i^j) - \mathbb{E}(N_i^j - D_i^j) \right| \tag{100}$$

$$\lesssim \mathbb{E}[|V_i - \mathbb{E}[V_i]|^2] \max_{x_i} \{N_i, D_i\} \tag{101}$$

$$\lesssim \mathbb{E}[V_i^2] \max_{x_i} \{N_i, D_i\} \tag{102}$$

$$\lesssim \alpha^3 \gamma \min \left\{ 1, \frac{L^2}{\sigma_p^2} \right\}. \tag{103}$$

By Lemma 18 and Lemma 16,

$$\mathbb{E}[|V_i - \mathbb{E}[V_i]|^2 |R_i|] \lesssim \mathbb{E}[V_i^2 |R_i|] + \mathbb{E}^2[V_i]\mathbb{E}[|R_i|] \tag{104}$$

$$\lesssim \sum_{j=1}^{2t-1} \mathbb{E}\left[(N_i^j - D_i^j)^2 |R_i| \right] + \left(\sum_{j=1}^{2t-1} \mathbb{E}\left[N_i^j - D_i^j \right] \right)^2 \mathbb{E}[|R_i|] \tag{105}$$

$$\lesssim \sum_{j=1}^{2t-1} \alpha^{2j}\beta \frac{L}{\sigma_q} + \left(\sum_{j=1}^{2t-1} \alpha^j \gamma \frac{L^2}{\sigma_q^2} \right)^2 \beta\sigma_q \tag{106}$$

$$\lesssim \alpha^2 \beta \frac{L}{\sigma_q} + \alpha^2 \gamma^2 \frac{L^5}{\sigma_q^5} \tag{107}$$

$$\lesssim \alpha^2 \beta \frac{L}{\sigma_q}. \tag{108}$$

For $\mathbb{E}[|V_i - \mathbb{E}[V_i]||R_i|^2]$, let $V_{ij} = (-1)^{j+1}(N_i^j - D_i^j)/j$.

$$\mathbb{E}[|V_i - \mathbb{E}[V_i]||R_i|^2] \leq \mathbb{E}[|V_i||R_i|^2] + |\mathbb{E}[V_i]|\mathbb{E}[|R_i|^2] \tag{109}$$

$$\leq \sum_{j=1}^{2t-1} \mathbb{E}[|V_{ij}||R_i|^2] + \sum_{j=1}^{2t-1} |\mathbb{E}[V_{ij}]|\mathbb{E}[|R_i|^2]. \tag{110}$$

For the first part, by Lemma 19,

$$\mathbb{E}[|V_{ij}||R_i|^2] \lesssim \alpha^j \beta^2 \frac{\sigma_p^3}{\sigma_q} \mathrm{erf}(\Theta(L/\sigma_p)) \tag{111}$$

$$+ \alpha^j \beta^2 \frac{\sigma_p(\sigma_p^2 + L^2)}{\sigma_q} \mathrm{erf}(\Theta(L/\sigma_p)) \exp\left(-\Theta\left(L^2/\sigma_q^2\right)\right) \tag{112}$$

$$+ \alpha^j \beta^2 \frac{\sigma_p^2 L}{\sigma_q} \exp\left(-\frac{L^2}{2\sigma_p^2}\right) \tag{113}$$

$$\lesssim \alpha^j \beta^2 \frac{\sigma_p^3 + \sigma_p^2 L + \sigma_p L^2}{\sigma_q}. \tag{114}$$

For the second part, by Lemma 16,

$$\mathbb{E}[V_{ij}]\mathbb{E}[|R_i|^2] \lesssim \alpha^j \gamma \frac{L^2}{\sigma_q^2}\beta^2 \sigma_q^2. \tag{115}$$

Combining the two parts,

$$\mathbb{E}[|V_i - \mathbb{E}[V_i]||R_i|^2] \leq \mathbb{E}[|V_i||R_i|^2] + \mathbb{E}[V_i]\mathbb{E}[|R_i|^2] \tag{116}$$

$$\lesssim \sum_{j=1}^{2t-1} \alpha^j \beta^2 \frac{\sigma_p^3 + \sigma_p^2 L + \sigma_p L^2}{\sigma_q} + \sum_{j=1}^{2t-1} \alpha^j \gamma \frac{L^2}{\sigma_q^2}\beta^2 \sigma_q^2 \tag{117}$$

$$\lesssim \alpha\beta^2 \frac{\sigma_p^3 + \sigma_p^2 L + \sigma_p L^2}{\sigma_q} + \alpha\gamma \frac{L^2}{\sigma_q^2}\beta^2 \sigma_q^2 \tag{118}$$

$$\lesssim \alpha\beta^2 \frac{\sigma_p^3 + \sigma_p^2 L + \sigma_p L^2}{\sigma_q}. \tag{119}$$

where the last line follows from $\gamma = \sigma_p/\sigma_q$. Finally, also by Lemma 16,

$$\mathbb{E}[|R_i|^3] \lesssim (\beta\sigma_q)^3. \tag{120}$$

Combining all terms together gives

$$M_3 = \mathbb{E}[|Z_i|^3] \lesssim \alpha^3 \gamma \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} + \alpha^2 \beta \frac{L}{\sigma_q} \tag{121}$$

$$+ \alpha\beta^2 \frac{\sigma_p^3 + \sigma_p^2 L + \sigma_p L^2}{\sigma_q} + (\beta\sigma_q)^3 \tag{122}$$

$$\lesssim \frac{p^3/\sigma_p^2}{q^3/\sigma_q^2} \min\left\{1, \frac{L^2}{\sigma_p^2}\right\} + \frac{p^2/\sigma_p^2}{q^2\sigma_q}L^2 \tag{123}$$

$$+ \frac{p/\sigma_p}{q\sigma_q^4}L^2(\sigma_p^3 + \sigma_p^2 L + \sigma_p L^2) + \frac{L^3}{\sigma_q^3}. \tag{124}$$

B.4. Toolbox

The following properties of Gaussian distributions are useful for proving the lower bounds.

Lemma 10

$$\int_{\mathbb{R}} e^{-\frac{(x-L)^2}{2b^2}} \frac{1}{\sqrt{2\pi}c} e^{-\frac{(x-L)^2}{2c^2}} dx = \frac{b}{\sqrt{b^2 + c^2}}, \tag{125}$$

$$\int_{\mathbb{R}} e^{-\frac{(x+L)^2}{2b^2}} \frac{1}{\sqrt{2\pi}c} e^{-\frac{(x-L)^2}{2c^2}} dx = \frac{b}{\sqrt{b^2 + c^2}} e^{-\frac{2L^2}{b^2 + c^2}}. \tag{126}$$

Lemma 11

$$\int_{\mathbb{R}} e^{-\frac{(x-L)^2}{2a^2} - \frac{(x+L)^2}{2b^2}} \frac{1}{\sqrt{2\pi}c} e^{-\frac{(x-L)^2}{2c^2}} dx \tag{127}$$

$$= \frac{ab}{\sqrt{a^2b^2 + a^2c^2 + b^2c^2}} \exp\left(-2L^2 \frac{a^2 + c^2}{a^2b^2 + a^2c^2 + b^2c^2}\right). \tag{128}$$

Lemma 12

$$\int_{\mathbb{R}} x e^{-\frac{(x-L)^2}{2b^2}} \frac{1}{\sqrt{2\pi}c} e^{-\frac{(x-L)^2}{2c^2}} dx = \frac{b}{\sqrt{b^2 + c^2}} L, \tag{129}$$

$$\int_{\mathbb{R}} x e^{-\frac{(x+L)^2}{2b^2}} \frac{1}{\sqrt{2\pi}c} e^{-\frac{(x-L)^2}{2c^2}} dx = \frac{b}{\sqrt{b^2 + c^2}} e^{-\frac{2L^2}{b^2+c^2}} \frac{b^2 - c^2}{b^2 + c^2} L. \tag{130}$$

Lemma 13

$$\int x^2 e^{-\frac{x^2}{2b^2}} \frac{1}{\sqrt{2\pi}c} e^{-\frac{(x-L)^2}{2c^2}} dx \tag{131}$$

$$= \frac{1}{\sqrt{2\pi}c} \frac{b^2c^2}{2(b^2 + c^2)^{3/2}} \left(\sqrt{2\pi}bc \cdot \mathrm{erf}\left(\frac{x\sqrt{b^2 + c^2}}{\sqrt{2}bc}\right) \right. \tag{132}$$

$$\left. - 2x\sqrt{b^2 + c^2} \exp\left(-\frac{x^2}{2b^2} - \frac{x^2}{2c^2}\right)\right) + \mathrm{constant} \tag{133}$$

$$= \frac{b^3c^2}{2(b^2 + c^2)^{3/2}} \mathrm{erf}\left(\frac{x\sqrt{b^2 + c^2}}{\sqrt{2}bc}\right) \tag{134}$$

$$- \frac{b^2cx}{\sqrt{2\pi}(b^2 + c^2)^{3/2}} \exp\left(-\frac{x^2}{2b^2} - \frac{x^2}{2c^2}\right) + \mathrm{constant}. \tag{135}$$

Lemma 14

$$\int x^2 e^{-\frac{(x+M)^2}{2b^2}} \frac{1}{\sqrt{2\pi}c} e^{-\frac{(x-L)^2}{2c^2}} dx \tag{136}$$

$$= \frac{1}{\sqrt{2\pi}c} \frac{bc^2}{2(b^2 + c^2)^{5/2}} \exp\left(-\frac{(M+x)^2}{2b^2} - \frac{x^2}{2c^2}\right) \tag{137}$$

$$\left[\sqrt{2\pi}c(b^4 + b^2c^2 + c^2M^2) \cdot \mathrm{erf}\left(\frac{b^2x + c^2M + c^2x}{\sqrt{2}bc\sqrt{b^2 + c^2}}\right) \exp\left(\frac{(b^2x + c^2M + c^2x)^2}{2b^2c^2(b^2 + c^2)}\right) \right. \tag{138}$$

$$\left. - 2b\sqrt{b^2 + c^2}(b^2x + c^2(x - M))\right] + \mathrm{constant} \tag{139}$$

$$= \frac{bc^2(b^4 + b^2c^2 + c^2M^2)}{2(b^2 + c^2)^{5/2}} \mathrm{erf}\left(\frac{b^2x + c^2M + c^2x}{\sqrt{2}bc\sqrt{b^2 + c^2}}\right) \exp\left(-\frac{M^2}{2(b^2 + c^2)}\right) \tag{140}$$

$$- \frac{b^2c}{\sqrt{2\pi}(b^2 + c^2)^2} \exp\left(-\frac{(M+x)^2}{2b^2} - \frac{x^2}{2c^2}\right) (b^2x + c^2(x - M)) + \mathrm{constant}. \tag{141}$$

Lemma 15 *For any $\epsilon > 0$, $\mathrm{erf}(\epsilon) \leq \frac{2}{\sqrt{\pi}}\epsilon$.*

Lemma 16 *For any non-negative integer p,*

$$\int_{\mathbb{R}} |x|^p \frac{1}{\sqrt{2\pi}c} e^{-\frac{x^2}{2c^2}} dx = c^p (p-1)!! \cdot \begin{cases} \sqrt{\frac{2}{\pi}} & \text{if } p \text{ is odd} \\ 1 & \text{if } p \text{ is even} \end{cases} \tag{142}$$

Lemma 17

$$\int_{\mathbb{R}} \left| e^{-\frac{(x-L)^2}{2b^2}} - e^{-\frac{(x+L)^2}{2b^2}} \right| \frac{1}{\sqrt{2\pi}c} e^{-\frac{(x-L)^2}{2c^2}} dx \tag{143}$$

$$= \frac{b}{\sqrt{b^2+c^2}} \text{erf}\left(\sqrt{\frac{b^2+c^2}{2b^2c^2}} L \right) + \frac{b e^{-\frac{2L^2}{b^2+c^2}}}{\sqrt{b^2+c^2}} \text{erf}\left(\frac{c^2-b^2}{bc\sqrt{2b^2+2c^2}} L \right) \tag{144}$$

$$\le \frac{4}{\sqrt{\pi}} \frac{c^2}{c^2+b^2} \frac{L}{c}. \tag{145}$$

Lemma 18

$$\int_{\mathbb{R}} |x| \left(e^{-\frac{x^2}{2b^2}} - e^{-\frac{(x+M)^2}{2b^2}} \right)^2 \frac{1}{\sqrt{2\pi}c} e^{-\frac{x^2}{2c^2}} dx \tag{146}$$

$$= \frac{bc^2 M}{(b^2+2c^2)^{3/2}} \exp\left(-\frac{b^2+c^2}{2b^2(b^2+2c^2)} M^2 \right) \text{erf}\left(\frac{cM}{b\sqrt{2b^2+4c^2}} \right) \tag{147}$$

$$+ \frac{bc^2 M}{(b^2+2c^2)^{3/2}} \exp\left(-\frac{1}{b^2+2c^2} M^2 \right) \text{erf}\left(\frac{2cM}{b\sqrt{2b^2+4c^2}} \right) \tag{148}$$

$$\le \sqrt{\frac{8}{\pi}} \frac{c^3 M}{(b^2+2c^2)^2} \left(\exp\left(-\frac{(b^2+c^2)M^2}{2b^2(b^2+2c^2)} \right) + \exp\left(-\frac{M^2}{b^2+2c^2} \right) \right) \tag{149}$$

$$\le \sqrt{\frac{32}{\pi}} \frac{c^3 M}{(b^2+2c^2)^2}. \tag{150}$$

Lemma 19

$$\int_{\mathbb{R}} x^2 \left| e^{-\frac{x^2}{2b^2}} - e^{-\frac{(x+M)^2}{2b^2}} \right| \frac{1}{\sqrt{2\pi}c} e^{-\frac{x^2}{2c^2}} dx \tag{151}$$

$$= \frac{b^3 c^2}{(b^2+c^2)^{3/2}} \text{erf}\left(\frac{\sqrt{b^2+c^2}}{\sqrt{8}bc} M \right) \tag{152}$$

$$+ \frac{bc^2(b^4+b^2c^2+c^2M^2)}{(b^2+c^2)^{5/2}} \text{erf}\left(\frac{(c^2-b^2)M}{2bc\sqrt{2b^2+2c^2}} \right) \exp\left(-\frac{M^2}{2b^2+2c^2} \right) \tag{153}$$

$$\tag{154}$$

$$+ \frac{2}{\sqrt{2\pi}} \frac{b^2 c^3 M}{(b^2+c^2)^2} \exp\left(-\frac{M^2}{8b^2} - \frac{M^2}{8c^2} \right). \tag{155}$$

Proof Let $L = M/2$. Since $\frac{x^2}{2b^2} \le \frac{(x+M)^2}{2b^2}$ when $x \ge -L$, and $\frac{x^2}{2b^2} > \frac{(x+M)^2}{2b^2}$ otherwise, we have

$$\int_{\mathbb{R}} x^2 \left| e^{-\frac{x^2}{2b^2}} - e^{-\frac{(x+M)^2}{2b^2}} \right| \frac{1}{\sqrt{2\pi}c} e^{-\frac{x^2}{2c^2}} dx \tag{156}$$

25

$$= \left(\int_{-L}^{+\infty} - \int_{-\infty}^{-L} \right) x^2 \left(e^{-\frac{x^2}{2b^2}} - e^{-\frac{(x+M)^2}{2b^2}} \right) \frac{1}{\sqrt{2\pi}c} e^{-\frac{x^2}{2c^2}} \, dx. \tag{157}$$

By Lemma 13,

$$\left(\int_{-L}^{+\infty} - \int_{-\infty}^{-L} \right) x^2 e^{-\frac{x^2}{2b^2}} \frac{1}{\sqrt{2\pi}c} e^{-\frac{x^2}{2c^2}} \, dx \tag{158}$$

$$= \frac{b^3 c^2}{(b^2 + c^2)^{3/2}} \operatorname{erf} \left(\frac{\sqrt{b^2 + c^2}}{\sqrt{2}bc} L \right) - \frac{2b^2 cL}{\sqrt{2\pi}(b^2 + c^2)} \exp \left(-\frac{L^2}{2b^2} - \frac{L^2}{2c^2} \right). \tag{159}$$

By Lemma 14,

$$\left(\int_{-L}^{+\infty} - \int_{-\infty}^{-L} \right) x^2 e^{-\frac{(x+M)^2}{2b^2}} \frac{1}{\sqrt{2\pi}c} e^{-\frac{x^2}{2c^2}} \, dx \tag{160}$$

$$= - \frac{bc^2(b^4 + b^2 c^2 + c^2 M^2)}{(b^2 + c^2)^{5/2}} \operatorname{erf} \left(\frac{c^2 - b^2}{\sqrt{2}bc\sqrt{b^2 + c^2}} L \right) \exp \left(-\frac{4L^2}{2(b^2 + c^2)} \right) \tag{161}$$

$$- \frac{2b^2 c}{\sqrt{2\pi}(b^2 + c^2)^2} \exp \left(-\frac{L^2}{2b^2} - \frac{L^2}{2c^2} \right) (b^2 + 3c^2)L. \tag{162}$$

Combining the terms completes the proof. ∎