Private Mean Estimation
with Person-Level Differential Privacy[*]

Sushant Agarwal[†]   Gautam Kamath[‡]   Mahbod Majid[§]   Argyris Mouzakis[¶]

Rose Silver[||]   Jonathan Ullman[**]

Abstract

We study person-level differentially private (DP) mean estimation in the case where each person holds multiple samples. DP here requires the usual notion of distributional stability when *all* of a person's datapoints can be modified. Informally, if n people each have m samples from an unknown d-dimensional distribution with bounded k-th moments, we show that

$$n = \tilde{\Theta}\left(\frac{d}{\alpha^2 m} + \frac{d}{\alpha m^{1/2}\varepsilon} + \frac{d}{\alpha^{k/(k-1)}m\varepsilon} + \frac{d}{\varepsilon}\right),$$

people are necessary and sufficient to estimate the mean up to distance α in ℓ_2-norm under ε-differential privacy (and its common relaxations). In the multivariate setting, we give computationally efficient algorithms under approximate DP and computationally inefficient algorithms under pure DP, and our nearly matching lower bounds hold for the most permissive case of approximate DP. Our computationally efficient estimators are based on the standard clip-and-noise framework, but the analysis for our setting requires both new algorithmic techniques and new analyses. In particular, our new bounds on the tails of sums of independent, vector-valued, bounded-moments random variables may be of interest.

Contents

1 Introduction

Mean estimation is the following classical statistical task: given a dataset $X_1, \ldots, X_n$ of samples drawn i.i.d. from some unknown distribution $\mathcal{D}$, output an estimate of its mean $\mathbb{E}_{X \sim \mathcal{D}}[X]$. Without special care, standard procedures for mean estimation can leak sensitive information about the underlying dataset [HSR+08, DSS+15], so there is now a line of recent work developing methods for mean estimation under the constraint of *differential privacy* (DP) [DMNS06].

Informally speaking, a DP algorithm is insensitive to any one person's data, but how we formalize the concept of one person's data can depend on the application.[1] In many applications, we assume that each person's data corresponds to a single element of the dataset, and thus we formalize DP as requiring that the algorithm be insensitive to changes in one sample X_i. This modeling protects one sample in the dataset, regardless of whether or not that captures all of one person's data, and is often called *item-level DP*.

However, there are many cases where a person can possess *multiple* samples, and we want to protect privacy of all the samples belonging to a single person. In these cases, we want to define datasets to be neighbors if they differ in exactly one person's set of samples. For example, systems for *federated learning* [MMR+17, KMA+21] adopt this modeling of privacy. More precisely, we study a setting where n people each have m samples[2] drawn i.i.d. from an unknown distribution $\mathcal{D}$. In this case, we formalize DP as requiring that the algorithm be insensitive to changing all m samples belonging to one person; we call this notion of privacy *person-level DP*. Note that this generalizes the initial setting, which corresponds to the case $m = 1$, and also that there are now nm total samples available.

Our goal is to design an estimator for $\mathbb{E}_{X \sim \mathcal{D}}[X]$, with small error in ℓ_2 distance, subject to person-level differential privacy. Our work focuses on two primary questions: What algorithms can we use to achieve this goal, and how many people and samples are required?

1.1 Results and Techniques

Our main results are upper and lower bounds on the sample complexity of mean estimation with person-level differential privacy, for both univariate and multivariate distributions with *bounded moments*. Bounded moments are a standard approach for capturing and quantifying what it means for data to be well behaved, and interpolates between distributions that have heavy tails and distributions that are well concentrated (e.g. Gaussians). Specifically, a distribution $\mathcal{D}$ over $\mathbb{R}^d$ with mean μ has bounded k-th moment if for all unit vectors $v \in \mathbb{R}^d$,

$$\mathbb{E}_{X \sim \mathcal{D}}[|\langle v, X - \mu \rangle|^k] \leqslant 1.[3]$$

Note that when $d = 1$, this is the standard k-th central moment condition, and when $d > 1$, this condition says that we can project the distribution onto any line and obtain a distribution with bounded k-th central moment.

In all cases, our lower bounds on sample complexity hold for the most permissive notion of *approximate differential privacy*. Our upper bounds on sample complexity hold even for the most restrictive notion of *pure differential privacy*. We complement our upper bounds on the sample complexity with a computationally efficient estimator in low-dimensions, or a computationally efficient estimator in high-dimensions satisfying approximate differential privacy.

As a starting point, we begin by presenting our estimator for the univariate case so we can introduce and discuss techniques and new phenomena that arise in the person-level setting.

THEOREM 1.1. (INFORMAL, SEE COROLLARY 3.4) *There is a computationally efficient person-level ε-DP estimator such that for every distribution $\mathcal{D}$ over $\mathbb{R}$ with mean μ and bounded k-th moment, the estimator takes m samples per person from*

$$n = \tilde{O}\left(\frac{1}{\alpha^2 m} + \frac{1}{\alpha m^{1/2} \varepsilon} + \frac{1}{\alpha^{k/(k-1)} m \varepsilon} + \frac{1}{\varepsilon} \right)$$

---

[1] Throughout, we use "person" to denote an entity for whom their entire dataset must have privacy preserved as the base level of granularity in the DP definition, which is often called *user-level privacy* in the literature. Depending on the application, the entity could naturally be called a person, individual, user, organization, silo, etc. without changing the underlying definition of privacy.

[2] We assume both m and n are publicly known. For simplicity, we focus on the case where each person has exactly m samples, though our general approach applies to heterogeneous cases where individuals have varying amounts of data.

[3] The choice of 1 on the right-hand side is arbitrary, and can be replaced with any other value by appropriate scaling.

people[4] and, with probability at least $2/3,$[5] *outputs an estimate* $\hat{\mu}$ *such that* $|\hat{\mu} - \mu| \leq \alpha$.

Observe that the required number n of people consists of four terms. As the total number of samples across all people is nm, the first term corresponds to the classical sample complexity of the problem, sans privacy constraints. The second and third terms are the most interesting. Loosely speaking, the second term corresponds to the optimal sample complexity of estimating the mean of a Gaussian in our setting. Since a Gaussian satisfies bounded k-th moments for every k, this number of samples is a lower bound in our setting. The third term can be arranged to say that the total number of samples we need is at least $1/\alpha^{k/(k-1)}\varepsilon$, which is the optimal sample complexity for estimating a k-th moment distribution under item-level privacy [BD14, KSU20], and is also a lower bound for our setting. The final term comes from a *coarse estimation step* in the algorithm, whose discussion we defer to the technical sections. This term is also necessary as any non-trivial DP algorithm requires data from at least $1/\varepsilon$ people.

Next, we consider the multivariate case, where we give an efficient estimator that satisfies approximate differential privacy.[6]

THEOREM 1.2. (INFORMAL, SEE THEOREM 4.3) *There is a computationally efficient* (ε, δ)-*DP estimator such that for every distribution* $\mathcal{D}$ *over* $\mathbb{R}^d$ *with mean* μ *and bounded* k-th *moment, the estimator takes* m *samples per person from*

$$n = \tilde{O}\left(\frac{d}{\alpha^2 m} + \frac{d \log^{1/2}(1/\delta)}{\alpha m^{1/2}\varepsilon} + \frac{d \log^{1/2}(1/\delta)}{\alpha^{k/(k-1)}m\varepsilon} + \frac{d \log^{1/2}(1/\delta) + d^{1/2}\log^{3/2}(1/\delta)}{\varepsilon}\right)$$

people and, with probability at least $2/3$, *outputs an estimate* $\hat{\mu}$ *such that* $\|\hat{\mu} - \mu\|_2 \leq \alpha$.

Observe that, in the first three terms, the sample complexity matches the sample complexity of the univariate case, scaled up by a factor of d, and, as we show, all three of these terms are optimal up to polylogarithmic factors. The final term does not decrease with m, meaning we will require $\gtrsim d/\varepsilon$ people no matter how many samples each person has. However, this final term does not depend on α and becomes dominated when α is sufficiently small, and in this case we provably obtain the optimal sample complexity.

In Section 1.2 we give an overview of the algorithms behind Theorems 1.1 and 1.2 and highlight the main new technical ideas.

Next, we give an algorithm with improved sample complexity and a guarantee of pure DP, at the cost of computational efficiency.

THEOREM 1.3. (INFORMAL, SEE THEOREM 5.1) *There is a computationally inefficient* ε-*DP estimator such that for every distribution* $\mathcal{D}$ *over* $\mathbb{R}^d$ *with mean* $\|\mu\| \leq R$ *and bounded* k-th *moment, the estimator takes* m-*samples per person from*

$$n = \tilde{O}\left(\frac{d}{\alpha^2 m} + \frac{d}{\alpha m^{1/2}\varepsilon} + \frac{d}{\alpha^{k/(k-1)}m\varepsilon} + \frac{d \log(R)}{\varepsilon}\right)$$

people and, with probability at least $2/3$, *outputs an estimate* $\hat{\mu}$ *such that* $\|\hat{\mu} - \mu\|_2 \leq \alpha$.

The first three terms in the sample complexity match those in the approximate DP case (Theorem 1.2). The fourth term is again the cost of coarse estimation, and is optimal for pure DP via standard packing lower bounds (see Lemma C.3).

Our pure-DP algorithm for the high-dimensional case employs the framework of Kamath, Singhal, and Ullman [KSU20], which reduces multivariate mean estimation to a collection of univariate mean testing problems. We slightly simplify their framework by viewing the set of testing problems as between a candidate mean and a "local cover" of its neighborhood, bringing it more into line with other works on private estimation [AAK21]. This reduction allows us to appeal to the techniques we developed for proving Theorem 1.1. We comment that,

---

[4]For simplicity, throughout this introduction we elide any assumptions that $|\mu| \leq R$ and any dependence on this parameter R, which is necessary for certain variants of DP but not others.

[5]The constant $2/3$ is arbitrary and can be amplified to any constant via standard techniques, at a cost of a constant factor blowup in sample complexity.

[6]We note that a similar algorithm can be designed under the stronger notion of *concentrated DP* [DR16, BS16], differing primarily in the coarse estimation step (and the corresponding term in the sample complexity).

inheriting the deficiencies of [KSU20], this algorithm is *not* computationally efficient. Indeed, most algorithms for pure-DP estimation in multivariate settings are computationally inefficient, barring a few recent works which rely on semidefinite programming and sum-of-squares optimization [HKM22, HKMN23]. However, the amount of computation and data required by these methods scales poorly as the order of the moment they employ increases. Since, in general, our algorithms employ higher-order moment information, it is not obvious how to make them computationally efficient, and we leave this open as an interesting open question for future work.[7]

Finally, we prove lower bounds on the sample complexity of private mean estimation under (ε, δ)-DP. Since both pure and concentrated DP imply (ε, δ)-DP, these lower bounds also apply to algorithms satisfying these variants.

THEOREM 1.4. (INFORMAL, SEE THEOREM 6.1) *Suppose there is an (ε, δ)-DP estimator such that for every distribution $\mathcal{D}$ over $\mathbb{R}^d$ with mean μ and bounded k-th moment, the estimator takes m samples per person from n people and, with probability at least $2/3$, outputs $\|\hat{\mu} - \mu\|_2 \leqslant \alpha$. Then we have that*

$$n = \tilde{\Omega}\left(\frac{d}{\alpha^2 m} + \frac{d}{\alpha m^{1/2}\varepsilon} + \frac{d}{\alpha^{k/(k-1)}m\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right).$$

This lower bound matches our one-dimensional upper bound (Theorem 1.1) in all four terms. It also matches our multivariate upper bounds (Theorems 1.2 and 1.3) in all terms except the last. As previously mentioned, we speculate the former upper bound (for approximate DP) is loose in this fourth term. The latter upper bound for pure DP is necessarily greater in the last term, due to the stronger privacy notion.

The first and third terms follow from a simple reduction from the item-level ($m = 1$) case: imagine starting with nm samples, splitting them into n batches of m samples, and feeding them into a person-level DP algorithm. If this algorithm required too few people, it would violate known lower bounds for the $m = 1$ case. The second term employs a conversion of Levy et al. [LSA$^+$21], which also reduces from the item-level case to the person-level case for Gaussian data, based on sufficient statistics. We also give an alternate, direct proof of this term using the fingerprinting technique [BUV14, KMS22a]. The fourth term is again required for any non-trivial DP algorithm.

1.2 Technical Overview of Theorems 1.1 and 1.2 Our efficient estimators use the following general framework, following [KSU20]:

1. **Reduce to item-level privacy.** For each person $i = 1, \ldots, n$, average their m samples to obtain $Y^{(i)}$. From now on, we will design an algorithm that takes $Y^{(1)}, \ldots, Y^{(n)}$, satisfies item-level privacy, and solves mean estimation for the distribution of Y. Notice that these random variables now come from the class of *averages of m independent random variables, each with bounded k-th moment*, which is more structured than the class of distributions with bounded k-th moments.

2. **Coarse Estimation.** Start by obtaining a *coarse estimate* μ_0 of the mean satisfying $\|\mu_0 - \mu\|_2 \leqslant \gamma$ for some suitable γ to narrow the search space.

3. **Clip-and-Noise.** Clip each value $Y^{(i)}$ to lie in an ℓ_2-ball of radius ρ centered at μ_0 for some suitable ρ. Then average these values and add Gaussian noise to this average. Notice that the noise we have to add to ensure DP will be proportional to the sensitivity of the average, which is bounded by $2\rho/n$ because of clipping.

Although the framework is well known, in the rest of this section we will describe the technical challenges that had to be overcome to apply this general framework in the setting of person-level privacy with data satisfying bounded-moments conditions.

For now we will ignore the coarse estimation step and suppress exactly what we need from the coarse estimate μ', so we can focus on the interplay between the first and third steps.

In the Gaussian case, this method is deceptively simple to analyze. In the first step, each value Y is itself a Gaussian with smaller variance (i.e., scaled down by a factor of m). Then, in the third step, since Gaussians are very tightly concentrated around their mean, we can clip to a small ball of radius $\approx d^{1/2}/m^{1/2}$ and introduce essentially no bias because there is only a very small probability that any individual point gets clipped. Formalizing the rest

---

[7]In particular, even for *item-level DP* where $m = 1$, computationally efficient algorithms with nearly optimal sample complexity are only known for the cases of $k = 2$ or for Gaussian data.

of this analysis is relatively straightforward. The case of bounded k-th moments is more challenging because the data can have heavy tails, posing a dilemma: either we clip to a relatively small ball and introduce significant bias, or we clip to a relatively large ball and have to add more noise for privacy, leading to an unavoidable bias-variance-tradeoff [KMR$^+$23].

To quantify this tradeoff and find the optimal balance of bias and variance, we need to understand the bias introduced by clipping the random variables Y, as a function of the clipping radius ρ. The standard approach to bounding the bias in one-dimension is more or less to integrate the function $f(t) = \mathbb{P}[|Y - \mu| \geq t]$ over all values of $t \geq \rho$, and thus we can reduce bounding the bias to computing tail bounds. In [KSU20], which is the case where $m = 1$ and Y is an arbitrary random variable with bounded k-th moment, the tightest possible bound on the tails is roughly $1/t^k$. However, things become more complicated in the person-level setting with $m > 1$.

Technical Challenge 1: Bounding the Bias in One-Dimension. Using the fact that Y is an average of m independent random variables with k-th moment bounded by 1, we can deduce that Y also has k-th moment bounded by roughly $1/m^{k/2}$, which yields a tail bound of roughly $1/m^{k/2}t^k$. However, this tail bound is actually suboptimal in its dependence on m. To see why, observe that as $m \to \infty$, Y converges to a Gaussian by the Central Limit Theorem, which has very thin tails of roughly $\exp(-mt^2)$. So for large m we'd expect a stronger tail bound than just what we get from assuming a bounded k-th moment. To make this argument rigorous, we apply a strong form of the Central Limit Theorem—*non-uniform Berry-Esseen* [Mic76]—that gives an optimal tail bound for this class of random variables (see Corollary 3.1). Plugging this tail bound into the standard bias arguments and then solving for the optimal tradeoff between bias and variance yields Theorem 1.1.

Technical Challenge 2: Tail Bounds in High-Dimensions. The natural way to generalize this approach to the high-dimensional case involves proving an analogous bound on $\mathbb{P}[\|Y - \mu\|_2 \geq t]$ for $t \geq \rho$. However, as far as we are aware, the non-uniform Berry-Esseen theorem does not have a generalization to high-dimensional random variables. To address this challenge, we use a somewhat-modified proof strategy that succeeds in high-dimensions and also gives an alternative proof of the optimal tail bound for the one-dimensional case. Roughly, our novel tail bound states that if Y is an average of m independent random variables with bounded k-th moment in d dimensions, then

$$\forall t \gtrsim \sqrt{\frac{d}{m}} \quad \mathbb{P}[\|Y - \mu\|_2 \geq t] \lesssim \exp\left(\frac{-mt^2}{d}\right) + \frac{d^{k/2}}{m^{k-1}t^k}.$$

(See Theorem 4.1 for a more precise statement.) Intuitively, the result says that the probability of the event $\|Y - \mu\|_2 \geq t$ is, roughly, at most the sum of two different probabilities: (1) The probability that a draw from $N(0, \frac{1}{m}\mathbb{I})$ has norm larger than t, and (2) the probability that a single random variable X with bounded k-th moment has norm larger than mt. Note that (1) is a lower bound on the tails because Gaussians have bounded k-th moments for every constant k, so $Y - \mu$ might be Gaussian. Also, (2) is a lower bound on the tails because Y is an average of m random variables with bounded k-th moments, and if a single one of these random variables has norm at least tm then there can be a constant probability that the norm of the average is at least mt. The theorem says that these are the two most likely ways the norm can be large. We believe that bounds on the tails of averages of independent random variables with bounded moments are quite natural, and we expect this result will find further applications.

Technical Challenge 3: Better Bias Bounds in High-Dimensions. The natural approach of integrating the function $f(t) = \mathbb{P}[\|Y - \mu\|_2 \geq t]$ for $t \geq \rho$ will yield some bound on the bias introduced by clipping to a ball of radius ρ. In particular, the integration would give a bound of the form $d^{k/2}/(m\rho)^{k-1}$. However, this bound is not strong enough to prove the sample-complexity bounds in Theorem 1.2. The looseness is because the bound only uses the fact that the random variables have a bound on the k-th moments of their norms, which, after suitable rescaling, is a *weaker* condition than assuming that the k-th moment is bounded in every individual direction. We give a novel, tighter bound on the bias introduced by clipping that takes into account the fact that we have control over our random variables in every direction. An interesting feature of our bound is that it depends heavily on how close the center of the ball used for clipping is to the true mean of the distribution. Specifically, if Y is a random variable satisfying our assumptions with mean μ, and we clip to a ball of radius ρ around a coarse estimate μ_0 such that $\|\mu_0 - \mu\|_2 \leq \gamma \ll \rho$, then the bias is at most

$$\frac{d^{(k-1)/2}}{(m\rho)^{k-1}}\left(1 + \frac{\gamma d^{1/2}}{\rho}\right).$$

The bound is a bit tricky to interpret. As a start, assume $\gamma = 0$, so we are clipping to a ball centered around the true mean. In this case, the bias improves over the straightforward bound by a factor of $d^{1/2}$, which is tight and sufficient for our purposes. We get the same conclusion as long as the ball is only off-center by $\gamma \lesssim \rho/d^{1/2}$. However, if the ball is off-center by more than $\rho/d^{1/2}$ we only get a weaker bias bound, and we can argue that this is inherent. Understanding the bias introduced by clipping is fundamental, and so we expect this bias bound will find further applications.

Technical Challenge 4: Iteratively Improving the Coarse Estimate. As the discussion above makes clear, the accuracy of the coarse estimate plays a significant role in determining the bias introduced by clipping. In order to say that the coarse estimate is accurate enough, we need it to have sufficiently small error $\gamma \lesssim \rho/d^{1/2}$. In some regimes, we will want to set $\rho \approx d^{1/2}/m^{1/2}$, so we want the coarse estimate to have error roughly $\gamma \approx 1/m^{1/2}$ to be safe. However, the straightforward way of obtaining a coarse estimate using the approach of [KSU20] will not produce such an accurate coarse estimate, and the result will be suboptimal sample complexity that is $\omega(d)$ overall.

To solve this problem, we use a novel iterative approach, inspired by the private preconditioning technique from [KLSU19, BDKU20, CKM$^+$20]. First, we use the simple strategy for producing a coarse estimate with somewhat large error γ_0. We then run the clip-and-noise algorithm with $O(d)$ people's data to obtain some new coarse estimate μ_1. While μ_1 will not itself have the error α that we desire, we can show that μ_1 is a better estimate than μ_0. Thus, we can restart the clip-and-noise algorithm with a smaller value γ_1 and get an even better estimate μ_2. We can analyze this process and show that by iterating this procedure a small number of times we will obtain μ_t such that $\|\mu_t - \mu\| \leqslant \gamma_t \approx 1/m^{1/2}$. As we've discussed above, such a coarse estimate μ_t is always accurate enough to get a tight bound on the bias of clipping, meaning we can run clip-and-noise a final time to get our last and most accurate estimate. Combining all these steps yields Theorem 1.2.

1.3 Related Work

Mean estimation under differential privacy has been studied extensively, including works which investigate mean estimation for Gaussians [KV18, BKSW19, KLSU19, BDKU20, DFM$^+$20, CWZ21, HLY21, HKMN23, BKS22, BDBC$^+$23, ALNP23, AKT$^+$23, AUZ23], distributions with bounded moments [BD14, KSU20, WXDX20, BGS$^+$21, KLZ22, HKM22, KDH23, BHS23], or both [BS19, LKKO21, LKO22, KMR$^+$23]. Several works have also focused on the related problem of privately estimating higher-order moments [AAK21, KMS$^+$22b, AL22, KMV22, KMS22a, Nar23, PH24]. Other works study private mean estimation of arbitrary (bounded) distributions [BUV14, SU15, DSS$^+$15]. However, all of these works focus on differential privacy in the item-level setting (i.e., when neighboring datasets differ in exactly one datapoint). This is a special case of the more general person-level privacy setting we consider, which corresponds to $m = 1$.

Recently, there have been several works focusing on estimation under person-level privacy. Some focus on mean estimation, but for distributions which are bounded or satisfy certain strong concentration properties [LSA$^+$21, GDD22]. These encompass (sub-)Gaussian distributions, but not those with heavier tails that we consider in our work. [NME22] also study mean estimation under a bounded-moment condition, but a qualitatively different type of condition than what we consider, and only with a bound on the second moment. Consequently, their work does not capture the rich tradeoff that we do as one varies the number m of samples per person and the number k of moments bounded. [CFMT22] focuses on mean estimation in a slightly different setting, where rather than people receiving samples from the distribution of interest, they instead have a latent sample from said distribution and receive several samples from the Bernoulli with that parameter. [GRST24] study mean estimation of Bernoullis under continual observation [HCSS10, DNPR10]. A line of works focuses on generic transformations from algorithms for item-level privacy to algorithms for person-level privacy [BGH$^+$23, GKK$^+$23b]. Even in the one-dimensional case, their results (combined with item-level algorithms of [KSU20]) are unable to recover our results for a couple reasons. First, their transformation only saves a factor of $1/\sqrt{m}$ in the private sample complexity, whereas in several parameter regimes, we save a factor of $1/m$. Based on details of our analysis, we believe a black-box technique that recovers our result from the item-level case seems unlikely. Furthermore, in general, their transformations are not computationally efficient. Finally, other works study adjacent learning tasks, such as learning discrete distributions [LSY$^+$20, CS23, ALS23], PAC learning [GKM21], and convex optimization [GKK$^+$23a, BS23, LA24].

Numerous other statistical estimation tasks have been studied under item-level differential privacy, including learning mixtures of Gaussians [KSSU19, AAL21, AAL23, AAL24], discrete distributions [DHS15], graphical models [ZKKW20], random graph models [BCS15, BCSZ18, SU19, CDd$^+$24], and median estimation [AMB19,

TVGZ20]. For more coverage of the literature on private statistics, see [KU20].

1.3.1 Independent work of Zhao et al. In May 2024, an initial version of this paper and a simultaneous and independent work of Zhao et al. [ZLS$^+$24] were posted on arXiv. Both works study mean estimation under person-level privacy, though considering slightly different notions of what it means for moments to be bounded. We consider a *directional* bound (Definition 2.4), which roughly says that the moment is bounded in every univariate projection. On the other hand, they consider a *non-directional* bound, which bounds the moment of the ℓ_2-norm of the random vector, i.e., $\mathbb{E}_{X \sim \mathcal{D}}\left[\|X - \mu\|_2^k\right]$.

It is not hard to see that a non-directional bound of 1 on the k-th moment implies a directional bound of 1 on the k-th moment. Similarly, one can show that a directional bound of 1 implies a non-directional bound of $d^{k/2}$ on the k-th moment. Thus, the two moment conditions are polynomially related to each other, and, by rescaling the data, upper and lower bounds for each case have implications for the other.

Using the latter implication that a directional bound implies a non-directional bound, one can employ algorithms for the non-directional setting in the directional setting. If one uses the algorithm in Theorem 3 of [ZLS$^+$24], and looks at the implications for the setting with directional moment bounds, one achieves a sample complexity comparable to Theorem 4.1 in the original May 2024 version of this paper. This sample complexity is roughly that of Theorem 1.2, albeit with an extra dimension-dependent factor in the third term. However, by working directly with the directional moment bound as we do, one can achieve the improved sample complexity in Theorem 1.2, which was done subsequent to the initial posting of our two papers. While our algorithms under our directional bound imply algorithms under their non-directional bound, the resulting sample complexities are loose by dimension-dependent factors.

Their work only considers algorithms, and not lower bounds. There may be implications of our lower bound arguments to show lower bounds for their non-directional moment bound setting.

2 Preliminaries

2.1 Privacy Preliminaries We say that two datasets X and X' are *neighboring datasets* if they differ in at most one datapoint, i.e. $d_H(X, X') \leqslant 1$ where d_H denotes the Hamming distance. We introduce the definition of differential privacy (which is also referred to as item-level differential privacy).

DEFINITION 2.1. (ITEM-LEVEL DIFFERENTIAL PRIVACY [DMNS06]) *Let* $\mathsf{A} : \mathcal{X}^n \to \mathcal{Y}$. *We say that* A *satisfies item-level* (ε, δ)-*differential privacy (DP) if, for all neighborhing datasets* $X, X' \in \mathcal{X}^n$ *and for all* $Y \in \mathcal{Y}$,

$$\mathbb{P}[\mathsf{A}(X) \in Y] \leqslant e^\varepsilon \, \mathbb{P}[\mathsf{A}(X') \in Y] + \delta.$$

It is reasonable to consider $\varepsilon, \delta \in [0, 1]$. We remark that $(\varepsilon, 0)$-DP is commonly referred to as *pure differential privacy*, and (ε, δ)-DP is commonly referred to as *approximate differential privacy*. In this paper, we focus on a generalized definition of differential privacy known as *person-level differential privacy*. For this definition, we consider datasets in $\mathcal{U}^n$ where $\mathcal{U} = \mathcal{X}^m$. We call each $X^{(i)} = (X_1^{(i)}, \ldots, X_m^{(i)}) \in \mathcal{U}$ a *person* and call each $X_j^{(i)} \in \mathcal{X}$ a *sample* (since in all our results these points will be chosen as independent samples from a probability distribution). In the definition of person-level differential privacy, we say that two datasets $X, X' \in \mathcal{U}^n$ are *neighboring datasets* if they differ in at most one one person's samples–that is, if $d_H(X, X') = \sum_{i=1}^n \mathbf{1}(X^{(i)} \neq X'^{(i)}) \leqslant 1$.

DEFINITION 2.2. (PERSON-LEVEL DIFFERENTIAL PRIVACY) *Let* $\mathcal{U} = \mathcal{X}^m$, *and let* $\mathsf{A} : \mathcal{U}^n \to \mathcal{Y}$. *We say that* A *satisfies person-level* (ε, δ)-*differential privacy (DP) if, for all neighboring datasets* $X, X' \in \mathcal{U}^n$ *and for all* $Y \in \mathcal{Y}$,

$$\mathbb{P}[\mathsf{A}(X) \in Y] \leqslant e^\varepsilon \, \mathbb{P}[\mathsf{A}(X') \in Y] + \delta.$$

We now introduce some standard lemmas for differentially private algorithms.[8] We first introduce a lemma surrounding a property known as *post-processing*.

LEMMA 2.1. (POST-PROCESSING) *Let* $\mathsf{A} : \mathcal{X}^n \to \mathcal{Y}$ *be a person-level* (ε, δ)-*DP algorithm. Let* $\mathsf{B} : \mathcal{Y} \to \mathcal{Z}$. *The mechanism* $\mathsf{B} \circ \mathsf{A}$ *satisfies person-level* (ε, δ)-*DP.*

---

[8]The specific formulations of some of these lemmas are inspired by those of [KSU20].

The next two lemmas deal with a property called *composition*. For these lemmas, we define $A_1, \ldots, A_k$ to be a sequence of mechanisms: $A_1 : \mathcal{X}^n \to \mathcal{Y}_1$ and, for all $t \in \{2, \ldots, k\}$, $A_t : \mathcal{X}^n \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{t-1} \to \mathcal{Y}_t$. For proofs of these lemmas, see, e.g. [DR14].

LEMMA 2.2. (BASIC COMPOSITION) *Suppose A_1 satisfies person-level $(\varepsilon_1, \delta_1)$-DP. Similarly, for each $t \in \{2, \ldots, k\}$, suppose that, under every fixed $y_1 \in \mathcal{Y}_1, \ldots, y_{t-1} \in \mathcal{Y}_{t-1}$, $A_t(\cdot, y_1, \ldots, y_{t-1})$ satisfies person-level $(\varepsilon_t, \delta_t)$-DP. It follows that the mechanism A_k which runs each of $A_1, \ldots, A_{k-1}$ in sequence satisfies person-level $\left(\sum_{i=1}^k \varepsilon_i, \sum_{i=1}^k \delta_i \right)$-DP.*

LEMMA 2.3. (ADVANCED COMPOSITION) *Suppose $A_1, \ldots, A_k$ are person-level $(\varepsilon_0, \delta_1), \ldots, (\varepsilon_0, \delta_k)$-DP respectively for some $\varepsilon_0 \leqslant 1$. Then, for all $\delta_0 > 0$, the mechanism A_k which runs each of $A_1, \ldots, A_{k-1}$ in sequence satisfies person-level (ε, δ)-DP for $\varepsilon = \varepsilon_0 \sqrt{6k \log(1/\delta_0)}$ and $\delta = \delta_0 + \sum_{t=1}^k \delta_t$.*

We now state a few definitions and lemmas related to standard private mechanisms.

DEFINITION 2.3. *Let $f : \mathcal{U}^n \to \mathbb{R}^d$ be a function, where $\mathcal{U} = \mathcal{X}^m$. The person-level ℓ_2-sensitivity (denoted $\Delta_{f,2}$) is defined as*

$$\Delta_{f,2} := \max_{\substack{X, X' \in \mathcal{U}^n \\ d_H(X, X') \leqslant 1}} \|f(X) - f(X')\|_2.$$

LEMMA 2.4. (GAUSSIAN MECHANISM) *Let $f : \mathcal{U}^n \to \mathbb{R}^d$ be a function with ℓ_2-sensitivity $\Delta_{f,2}$, where $\mathcal{U} = \mathcal{X}^m$. Let $S \in \mathcal{U}^n$, and let*

$$W \sim \mathcal{N}\left(0, \left(\frac{\Delta_{f,2}\sqrt{2 \ln(2/\delta)}}{\varepsilon} \right)^2 \cdot \mathbb{I}_{d \times d} \right).$$

Then the mechanism

$$M(X) = f(X) + W$$

satisfies person-level (ε, δ)-DP.

For a proof of Lemma 2.4, see e.g. [DR14].

LEMMA 2.5. (PRIVATE HISTOGRAMS) *Let $(S_1, \ldots, S_n)$ be samples in some data universe Ω, and let $U = \{h_u\}_{u \subset \Omega}$ be a collection of disjoint histogram buckets over Ω. Then, we have person-level ε-DP and (ε, δ)-DP histogram algorithms with the following guarantees.*

- *ε-DP: ℓ_∞ error $O\left(\frac{\log(|U|/\beta)}{\varepsilon} \right)$ with probability at least $1 - \beta$; run time $\mathrm{poly}(n, \log(|U|/(\varepsilon\beta)))$,*

- *(ε, δ)-DP: ℓ_∞ error $O\left(\frac{\log(1/\delta\beta)}{\varepsilon} \right)$ with probability at least $1 - \beta$; run time $\mathrm{poly}(n, \log(|U|/(\varepsilon\beta)))$.*

LEMMA 2.6. (EXPONENTIAL MECHANISM [MT07]) *The exponential mechanism $M_{\varepsilon, S, \mathrm{Score}}(X)$ takes a dataset $X \in \mathcal{X}^n$, computes a score ($\mathrm{Score} : \mathcal{X}^n \times S \to \mathbb{R}$) for each $p \in S$ with respect to X, and outputs $p \in S$ with probability proportional to $\exp(\frac{\varepsilon \cdot \mathrm{Score}(X,p)}{2 \cdot \Delta_{\mathrm{Score},1}})$, where:*

$$\Delta_{\mathrm{Score},1} := \max_{p \in S} \max_{X \sim X' \in \mathcal{X}^n} |\mathrm{Score}(X, p) - \mathrm{Score}(X', p)|.$$

It satisfies the following properties:

1. *M is person-level ε-differentially private.*

2. *Let $\mathrm{OPT}_{\mathrm{Score}}(X) = \max_{p \in S}\{\mathrm{Score}(X, p)\}$. Then*

$$\mathbb{P}\left[\mathrm{Score}(X, M_{\varepsilon, S, \mathrm{Score}}(X)) \leqslant \mathrm{OPT}_{\mathrm{Score}}(X) - \frac{2\Delta_{\mathrm{Score},1}}{\varepsilon} (\ln(|S|) + t) \right] \leqslant e^{-t}.$$

2.2 Bounded Moments

DEFINITION 2.4. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean μ. The k-th moment $\sigma_k(\mathcal{D})$ of $\mathcal{D}$ is defined as*

$$\sigma_k(\mathcal{D}) := \mathop{\mathbb{E}}_{X \sim \mathcal{D}}\left[|X - \mu|^k\right]^{1/k}.$$

For $\mathcal{D}$ over $\mathbb{R}^d$ with mean μ, $\sigma_k(\mathcal{D})$ is defined as

$$\sigma_k(\mathcal{D}) := \sup_{v \in \mathbb{S}^{d-1}} \mathop{\mathbb{E}}_{X \sim \mathcal{D}}\left[|\langle X - \mu, v \rangle|^k\right]^{1/k}.$$

LEMMA 2.7. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean μ and $\sigma_k(\mathcal{D}) \leqslant 1$. Let $X_1, \ldots, X_m \sim \mathcal{D}$. Let $X = \frac{1}{m}(X_1 + \cdots + X_m)$. It follows that*

$$\mathbb{E}[|X - \mu|^k]^{1/k} = O\left(\frac{\sigma_k(\mathcal{D})}{\sqrt{m}}\right).$$

Throughout this paper, we will treat k as a constant. We define the following notation to denote that the inequality holds up to multiplicative factors that depend only on k.

DEFINITION 2.5. (HIDING k: $\leqslant_k$) *We write $f(x) \leqslant_k g(x)$, if and only if there exists a function $h(k) \geqslant 0$ such that $f(x) \leqslant h(k)g(x)$.*

3 Private Mean Estimation in One Dimension

In this section, we describe an efficient algorithm for estimating means of univariate distributions with bounded k-th moments under person-level differential privacy. The algorithm is based on the *Clip-and-Noise* framework and consists of two phases:

1. Coarse Estimation: Find a coarse estimate of the mean with only a few samples.

2. Fine Estimation: Given a coarse estimate of the mean, produce an estimate to the mean with error α.

The fine-estimation step is where we see *clip-and-noise* come into play: First, for each person $i \in [n]$, compute the mean $\bar{S}^{(i)}$ of their m samples. Then, clip each $\bar{S}^{(i)}$ to a radius around the coarse estimate to obtain some value $Z^{(i)}$. Finally, output the average of $Z^{(1)}, \ldots, Z^{(n)}$ with additional noise for privacy.

In Section 3.2, we prove Theorem 3.1 which gives a statement on the final error α of applying the clip-and-noise paradigm to general distributions. In Section 3.3, we prove Theorem 3.2 which analyzes the error and sample complexity of the coarse-estimation subroutine when the samples come from a distribution with bounded k-th moment. Finally, in Section 3.4, we piece together Theorem 3.1 and Theorem 3.2 in order to achieve Theorem 1.1.

3.1 Preliminaries We will make use of the Laplace mechanism which uses additive noise. We use the following tail bound on the Laplace distribution to bound the noise term.

LEMMA 3.1. (LAPLACE DISTRIBUTION TAIL BOUND) *For a Laplace distribution $\mathcal{L}$ with parameter b, and for any $t > 0$,*

$$\mathbb{P}[|\mathcal{L}(b)| \geqslant tb] \leqslant \exp(-t).$$

In order to reduce sensitivity we use a truncation operation defined as follows.

DEFINITION 3.1. (CLIPPING OPERATION $\text{Trunc}_{\ell,r}$) *For two numbers $\ell \leqslant r \in \mathbb{R}$ we define the clipping operation to be*

$$\text{Trunc}_{\ell,r}(x) := \begin{cases} r & x \geqslant r \\ l & x \leqslant l \\ x & o.w. \end{cases}$$

In the analysis of the algorithm, we use the following fact about truncated random variables. (See Appendix A for the proof.)

LEMMA 3.2. *Suppose X is a random variable in $\mathbb{R}$ and $Z = \mathrm{Trunc}_{\ell,r}(X)$, where $\mathrm{Trunc}_{\ell,r}$ is the truncation operation. Then*

$$\mathbb{E}\left[|Z - \mathbb{E}[Z]|^2\right] \leq \mathbb{E}\left[|X - \mathbb{E}[X]|^2\right].$$

Finally, in our analysis we will use the non-uniform Berry-Esseen tail bound for the average of m i.i.d. samples from a distribution with bounded k-th moments to bound the bias caused by truncation. See Appendix D for discussion and proof.

COROLLARY 3.1. (TAIL BOUND FOR AVERAGES OF BOUNDED-MOMENTS VARIABLES [MIC76]) *Let $k \geq 3$. Assume X is a distribution with mean 0 and k-th moment bounded by 1, and X_i's are m i.i.d. copies of X. Then,*

$$\forall t \geq \sqrt{\frac{(k-1)\log m}{m}}, \quad \mathbb{P}\left[\frac{1}{m}\sum_{i=1}^{m} X_i \geq t\right] \leq O\left(\frac{1}{m^{k-1}t^k}\right).$$

3.2 Generic Clip and Noise Theorem In this section, we create a generic theorem that applies the clip-and-noise framework for private mean estimation, given access to tail bounds for a univariate distribution, a coarse estimate of its mean, and an upper bound on its variance. The algorithm is as follows: take n i.i.d. samples from the distribution, take their average, clip the average, and add Laplace noise in order to output an ε-DP mean estimation. Lemma 3.3 bounds the error caused by the clipping operation. Lemma 3.4 bounds the error caused by the addition of Laplace noise and the sampling error from the truncated distribution. Finally, putting together the clipping bias, sampling bias, and noise term guarantees, Theorem 3.1 gives a guarantee on the accuracy of the clip-and-noise framework, given distributions with arbitrary tail bounds.

LEMMA 3.3. (CLIPPING BIAS) *Let $X \in \mathbb{R}$ be a random variable with finite mean. For two numbers $l \leq r$, let $Z := \mathrm{Trunc}_{l,r}(X)$. Then, we have:*

$$|\mathbb{E}[X] - \mathbb{E}[Z]| \leq \int_{-\infty}^{\ell} \mathrm{LeftTail}(x)\,dx + \int_{r}^{+\infty} \mathrm{RightTail}(x)\,dx,$$

where $\mathrm{RightTail}(x) := \mathbb{P}[X \geq x]$ and $\mathrm{LeftTail}(x) := \mathbb{P}[X \leq x]$.

Proof. We observe that $Z = \ell\mathbb{1}\{X < \ell\} + X\mathbb{1}\{\ell \leq X \leq r\} + r\mathbb{1}\{X > r\}$. Using the above, as well as a combination of the triangle inequality and Jensen's inequality for the function $\phi(x) = |x|$, the clipping bias can be upper-bounded by:

$$\begin{aligned}|\mathbb{E}[Z] - \mathbb{E}[X]| &= |\mathbb{E}[(\ell - X)\mathbb{1}\{X \leq \ell\}] + \mathbb{E}[(r - X)\mathbb{1}\{X \geq r\}]| \\ &\leq \mathbb{E}[(\ell - X)\mathbb{1}\{X \leq \ell\}] + \mathbb{E}[(X - r)\mathbb{1}\{X \geq r\}].\end{aligned}$$

We note now that the individual terms in the above sum involve non-negative random variables, so we have:

$$\begin{aligned}|\mathbb{E}[Z] - \mathbb{E}[X]| &\leq \int_{0}^{+\infty} \mathbb{P}[(\ell - X)\mathbb{1}\{X \leq \ell\} \geq x]\,dx + \int_{0}^{+\infty} \mathbb{P}[(X - r)\mathbb{1}\{X \geq r\} \geq x]\,dx \\ &= \int_{0}^{+\infty} \mathbb{P}[X \leq \ell - x]\,dx + \int_{0}^{+\infty} \mathbb{P}[X \geq r + x]\,dx \\ &= \int_{-\infty}^{\ell} \mathbb{P}[X \leq x]\,dx + \int_{r}^{+\infty} \mathbb{P}[X \geq x]\,dx,\end{aligned}$$

where the last equality relies on the changes of variables $x \to \ell - x$, and $x \to x - r$, respectively. $\quad\square$

Often the form of the tail bounds we have on distributions are of the form $\mathbb{P}[X \leqslant \mu - t] \leqslant \cdots$, or $\mathbb{P}[X \geqslant \mu + t] \leqslant \cdots$, where μ is the mean of X. The following corollary adapts Lemma 3.3 to this setting where we have access to a coarse estimate of the mean.

COROLLARY 3.2. (CLIPPING BIAS GIVEN COARSE ESTIMATE) *In Lemma 3.3 suppose that we are given access to a coarse estimate μ_{coarse} such that $|\mu_{coarse} - \mu| \leqslant u$, and we set $\ell = \mu_{coarse} - \rho, r = \mu_{coarse} + \rho$. Then*

$$|\mathbb{E}[X] - \mathbb{E}[Z]| \leqslant \int_{\rho - u}^{+\infty} \mathbb{P}[|X - \mu| \geqslant t]\, dt.$$

Proof. For the first term we have

$$\int_{-\infty}^{\mu_{coarse} - \rho} \text{LeftTail}(x)\, dx = \int_{-\infty}^{\mu_{coarse} - \rho} \mathbb{P}[X \leqslant x]\, dx \leqslant \int_{-\infty}^{\mu - \rho + u} \mathbb{P}[X \leqslant x]\, dx = \int_{0}^{+\infty} \mathbb{P}[X \leqslant \mu - \rho + u - t]\, dt$$

$$= \int_{0}^{+\infty} \mathbb{P}[X - \mu \leqslant -(\rho - u + t)]\, dt.$$

Similarly,

$$\int_{\mu_{coarse} + \rho}^{+\infty} \text{RightTail}(x)\, dx \leqslant \int_{0}^{+\infty} \mathbb{P}[X - \mu \geqslant \rho - u + t]\, dt.$$

Putting these together finishes the proof. $\quad\Box$

Next, let's analyze the sampling bias for the truncated distribution and the error induced by the additive Laplace noise.

LEMMA 3.4. (SAMPLING BIAS AND NOISE TERM) *Suppose $\varepsilon > 0$, and a random variable Z supported on $[\ell, r]$ is given, with variance $\mathbb{V}[Z]$. Suppose n i.i.d. copies of Z, Z_i's are given. Let $2\rho = r - \ell$. and $\hat{\mu} = \frac{Z_1 + \cdots + Z_n}{n} + \mathcal{L}(\frac{2\rho}{n\varepsilon})$, where $\mathcal{L}$ is the Laplace distribution. Then*

$$\mathbb{P}\left[|\hat{\mu} - \mathbb{E}[Z]| \geqslant \Omega\left(\sqrt{\frac{\mathbb{V}[Z]}{n\beta}} + \frac{\rho \log(1/\beta)}{n\varepsilon}\right)\right] \leqslant \beta.$$

Proof. Note that $\hat{\mu} - \mathbb{E}[Z] = \frac{Z_1 + \cdots + Z_n}{n} - \mathbb{E}[Z] + \mathcal{L}(\frac{\rho}{n\varepsilon})$. From Chebyshev's inequality we know

$$\forall t \geqslant 0: \quad \mathbb{P}\left[\left|\frac{Z_1 + \ldots Z_n}{n} - \mathbb{E}[Z]\right| \geqslant t\right] \leqslant \frac{\mathbb{V}[Z]}{nt^2}.$$

From tail bounds for the Laplace distribution (Lemma 3.1)

$$\forall t \geqslant 0: \quad \mathbb{P}\left[\left|\mathcal{L}\left(\frac{2\rho}{n\varepsilon}\right)\right| \geqslant t\right] \leqslant \exp\left(-\frac{tn\varepsilon}{2\rho}\right).$$

Therefore,

$$\mathbb{P}[|\hat{\mu} - \mathbb{E}[Z]| \geqslant t] \leqslant \mathbb{P}\left[\left|\frac{Z_1 + \cdots + Z_n}{n} - \mathbb{E}[Z]\right| \geqslant t/2\right] + \mathbb{P}\left[\left|\mathcal{L}\left(\frac{2\rho}{n\varepsilon}\right)\right| \geqslant t/2\right]$$

$$\leqslant \frac{4\mathbb{V}[Z]}{nt^2} + \exp(-\frac{tn\varepsilon}{4\rho}).$$

2830

Therefore,

$$\mathbb{P}\left[|\hat{\mu} - \mathbb{E}[Z]| \geqslant \sqrt{\frac{8\,\mathbb{V}[Z]}{n\beta}} + \frac{4\rho\log(2/\beta)}{n\varepsilon}\right] \leqslant \beta \,.$$

□

Together Lemma 3.4 and Corollary 3.2 give us a recipe for estimating the mean for arbitrary distributions with known tail bounds, and variances.

THEOREM 3.1. (COARSE TO FINE ESTIMATE) *Suppose X is a distribution with variance $\mathbb{V}[X]$. Suppose a coarse estimate of $\mathbb{E}[X]$, the true mean is given with accuracy u. Then truncation to a radius of $\rho/2$ around this coarse estimate gives an ε-DP estimate of μ with success probability $1 - \beta$, and accuracy*

$$\alpha = O\left(\sqrt{\frac{\mathbb{V}[X]}{n\beta}} + \frac{\rho\log(1/\beta)}{n\varepsilon}\right) + \int_{\rho-u}^{+\infty} \mathbb{P}[|X - \mu| \geqslant t]\,\mathrm{d}t.$$

Proof. Putting together Lemma 3.4, and Corollary 3.2, and noting that variance decreases after truncation by Lemma 3.2. □

3.3 Coarse Estimation In this section we give an algorithm for obtaining a coarse estimate of the mean up to accuracy $O(1/\sqrt{m})$, under both pure and approximate differential privacy, given access to n people, each taking m samples.

Algorithm 1 One-Dimensional Range Estimator

Input: Samples $X = \left\{X_j^{(i)}\right\}_{i\in[n], j\in[m]}$ where each $X_j^{(i)} \in \mathbb{R}$. Parameters $\varepsilon, \delta, r, R$.

Output: $\mu_{\texttt{coarse}} \in \mathbb{R}$.

1: **procedure** RANGEESTIMATOR($X; \varepsilon, \delta, r, R$)
2:     For each $i \in [n]$, do $S_i = \frac{1}{m}(X_1^{(i)} + \cdots + X_m^{(i)})$.
3:     Divide $[-R - 2r, R + 2r)$ into buckets of width r: $[-R - 2r, -R - r), \ldots, [-r, 0), [0, r), \ldots, [R + r, R + 2r)$.
4:     If $\delta = 0$, run ε-DP Histogram, and if $\delta > 0$, run (ε, δ)-DP Histogram from Lemma 2.5 using samples $(S_1, \ldots, S_n)$ over the buckets defined above.
5:     Return $\mu_{\texttt{coarse}} = \frac{a+b}{2}$, where $[a, b]$ is the bucket with the largest count.
6: **end procedure**

THEOREM 3.2. (PERSON-LEVEL COARSE ESTIMATION) *For all people $i \in [n]$ and number $j \in [m]$ of samples per person, let $X = \{X_j^{(i)}\}_{i,j}$. Let $r > 0$. Then for all $\varepsilon > 0$, $\delta \in [0, 1)$, the algorithm RangeEstimator($X; \cdot, \cdot, \cdot, \cdot$) satisfies person-level (ε, δ)-DP. In addition, let P be a distribution over $\mathbb{R}$ with mean $|\mu| \leqslant R$ for some $R > 0$ and k-th moment bounded by 1. Suppose each $X_j^{(i)} \sim P$ i.i.d.. For all $\varepsilon > 0$, $\beta \in (0, 1)$, and $16^{1/k}/\sqrt{m} < r < R$, there exists*

$$n_0 = O\left(\frac{\log(1/\beta)}{\log(\sqrt{m}r)} + \frac{\log(R/(r\beta))}{\varepsilon}\right)$$

such that, for all $n \geqslant n_0$, $\mu_{coarse} \leftarrow$ RangeEstimator($X; \varepsilon, 0, r, R$) satisfies

$$|\mu - \mu_{coarse}| < 2r$$

with probability at least $1 - \beta$. Moreover, for all $\varepsilon > 0$, $\delta \in (0, 1)$, $\beta \in (0, 1)$, and $16^{1/k}/\sqrt{m} < r < R$, there exists

$$n_1 = O\left(\frac{\log(1/\beta)}{\log(\sqrt{m}r)} + \frac{\log(1/(\delta\beta))}{\varepsilon}\right)$$

such that, for all $n \geqslant n_1$, $\mu_{coarse} \leftarrow RangeEstimator(X; \varepsilon, \delta, r, R)$ satisfies

$$|\mu - \mu_{coarse}| < 2r$$

with probability at least $1 - \beta$.

In the proof of Theorem 3.2, we use the following Chernoff bound:

THEOREM 3.3. (CHERNOFF) *Let $Y_1, \ldots, Y_n$ be 0-1 random variables. Let $Y = Y_1 + \ldots + Y_n$, and let $\mu = \mathbb{E}[Y]$. Then for all $t > 1$,*

$$\mathbb{P}[Y > t\mu] = \frac{1}{t^{\Omega(t\mu)}}.$$

We now prove Theorem 3.2.

Proof. Privacy follows by running the (ε, δ)-DP Histogram (Lemma 2.5) and by post-processing (Lemma 2.1). We focus on analyzing the accuracy.

Let $[a, b]$ be the bucket such that $\mu_{\text{coarse}} \in [a, b]$ (that is, $[a, b]$ is the bucket with the largest noisy count). To show that $|\mu - \mu_{\text{coarse}}| < 2r$, it suffices to show that $\max(|a - \mu|, |b - \mu|) < 2r$, or equivalently, the bucket with the largest count resides within $\mu \pm 2r$. In proving this, we show that the each of the following two events hold with probability at least $1 - \beta/2$: (1) A large fraction of the samples resides within $\mu \pm 2r$, and (2) no bucket will have too large a magnitude of additive Laplace or Gaussian noise.

We start by showing (1). In particular, we show that, with probability at least $1 - \beta/2$, a $(1 - \frac{1}{16})$-fraction of the samples will lie within $\mu \pm r$. For all $i \in [n]$, let $Y_i = \mathbb{I}(|X_i - \mu| > r)$, let $p = \mathbb{P}[Y_i = 1]$, and let $Y = Y_1 + \ldots + Y_n$. Note that $\mathbb{E}[Y] = np$ and

$$p = \mathbb{P}[|X_i - \mu| > r] = \mathbb{P}[|X_i - \mu|^k > r^k] \leqslant \left(\frac{1}{\sqrt{m}r}\right)^k$$

where the inequality follows from Markov's inequality and Lemma 2.7. To show (1), we want to bound $\mathbb{P}\left[Y > \frac{n}{16}\right]$. To do this, we can assume that p is as maximal as possible (i.e. when $p = \left(1/(\sqrt{m}r)\right)^k$) as this maximizes $\mathbb{P}\left[Y > \frac{n}{16}\right]$. Thus,

$$\mathbb{P}\left[Y > \frac{n}{16}\right] = \mathbb{P}\left[Y > \frac{(\sqrt{m}r)^k \mathbb{E}[Y]}{16}\right]$$

(by Theorem 3.3) $$= \left(\frac{16}{(\sqrt{m}r)^k}\right)^{\Omega\left((\sqrt{m}r)^k \mathbb{E}[Y]/16\right)}$$

$$= \left(\frac{16}{(\sqrt{m}r)^k}\right)^{\Omega(n/16)}$$

$$= \left(\frac{16}{(\sqrt{m}r)^k}\right)^{\Omega\left(\log(\beta^{-1})/\log((\sqrt{m}r)^k/16)\right)}$$

$$= \left(\frac{1}{2}\right)^{\Omega\left(\log \beta^{-1}\right)}$$

$$\leqslant \beta/2.$$

Note that there are at most 3 buckets which satisfy $\max(|a - \mu|, |b - \mu|) < 2r$. Thus, the largest of these 3 buckets must have at least $(15n/16)/3 = 5n/16$ points with probability at least $1 - \beta/2$.

To show (2), we can directly apply Lemma 2.5. If $n_0 = O\left(\log(R/(r\beta))/\varepsilon\right)$, then the ℓ_∞-error induced by the Laplace noise is at most $n/16$ with probability at least $1 - \beta/2$. Similarly, If $n_1 = O\left(\log(1/\beta\delta)/\varepsilon\right)$, then the ℓ_∞-error induced by the Gaussian noise is at most $n/16$ with probability at least $1 - \beta/2$. Thus, by a union bound,

$$\mathbb{P}[|\mu - \mu_{\text{coarse}}| > 2r] \leqslant \beta/2 + \beta/2 \leqslant \beta.$$

$\square$

3.4 Applying Clip and Noise to Distributions with Bounded k-th Moments In this section, we use non-uniform Berry-Esseen and a coarse estimate of the mean to find a fine estimate of the mean. We assume $\overline{S}_m$ is a random variable corresponding to the mean of m samples, and apply Berry-Esseen to obtain tail bounds for this distribution and then use Theorem 3.1.

LEMMA 3.5. (COARSE TO FINE FOR BOUNDED k-TH MOMENTS) *Assume $\rho \geqslant u + \sqrt{\frac{(k-1)\log m}{m}}$. Suppose X is a distribution with mean μ and k-th moment bounded by 1. Let $\overline{S}_m$ be the random variable corresponding to the mean of m i.i.d. samples from X. Suppose we have access to a coarse estimate of the mean with accuracy u. Then there exists an ε-DP estimator of the mean that takes samples from the sample mean distribution, truncates them within a radius of ρ around the coarse estimate, and outputs an estimate of the mean with success probability $1 - \beta$ and accuracy*

$$\alpha = O\left(\sqrt{\frac{1}{mn\beta}} + \frac{\rho \log(1/\beta)}{n\varepsilon} + m^{-k+1} \cdot (\rho - u)^{-k+1}\right).$$

Proof. We aim to apply Theorem 3.1, and note that by Corollary 3.1 we have that for all $t \geqslant \sqrt{\frac{(k-1)\log m}{m}}$,

$$\mathbb{P}\left[\left|\overline{S}_m - \mu\right| \geqslant t\right] \leqslant_k m^{-k+1} t^{-k}.$$

Therefore, since $\rho - u \geqslant \sqrt{\frac{(k-1)\log m}{m}}$, we have that

$$
\int_{\rho-u}^{+\infty} \mathbb{P}\left[\left|\overline{S}_m - \mu\right| \geqslant t\right] dt \leqslant_k \int_{\rho-u}^{+\infty} m^{-k+1} t^{-k} \, dt
$$

$$
= m^{-k+1} \int_{\rho-u}^{+\infty} t^{-k} \, dt
$$

$$
= m^{-k+1} \left. \frac{t^{-k+1}}{-k+1} \right|_{\rho-u}^{+\infty}
$$

$$
= m^{-k+1} \frac{(\rho - u)^{-k+1}}{k - 1}.
$$

Finally, note that since the k-th moment of X is bounded by 1, its variance is bounded by 1 as well. Therefore, $\mathbb{V}[\overline{S}_m] \leqslant 1/m$, and hence

$$\alpha = O_k\left(\sqrt{\frac{1}{mn\beta}} + \frac{\rho \log(1/\beta)}{n\varepsilon} + m^{-k+1} \cdot (\rho - u)^{-k+1}\right).$$

□

Before presenting the main theorem of the section, we state a corollary that is related to the contribution of the bias term in the error rate given in the previous lemma.

COROLLARY 3.3. *Suppose X is a distribution with mean μ and k-th moment bounded by 1. Let $\overline{S}_m$ be the random variable corresponding to the mean of m i.i.d. samples from X. Suppose that we are given access to a coarse estimate μ_{coarse} such that $|\mu_{coarse} - \mu| \leqslant u$, and that Z is the result of truncating $\overline{S}_m$ in the interval $[\mu_{coarse} - \rho, \mu_{coarse} + \rho]$, where $\rho - u \geqslant \sqrt{\frac{(k-1)\log m}{m}}$. Then, the bias induced by this truncation operation satisfies:*

$$|\mathbb{E}[Z] - \mathbb{E}[X]| \leqslant m^{-k+1} \frac{(\rho - u)^{-k+1}}{k - 1}.$$

Thus, if we want the bias to be at most α, we need to take $\rho \geqslant_k u + \frac{1}{\alpha^{\frac{1}{k-1}} m}$.

We now conclude by presenting the main result of this section.

THEOREM 3.4. (MEAN ESTIMATION FOR BOUNDED k-TH MOMENT DISTRIBUTIONS) *Suppose X is a distribution over $\mathbb{R}$ with mean $|\mu| \leqslant R$ for some $R > 0$ and k-th moment bounded by 1. Then there exists a person-level ε-DP mean estimation algorithm that takes*

$$\tilde{O}_k\left(\frac{1}{\alpha^2 m \beta} + \frac{\log(1/\beta)}{\alpha \sqrt{m} \varepsilon} + \frac{\log(1/\beta)}{\alpha^{\frac{k}{k-1}} m \varepsilon} + \frac{\log(Rm/\beta)}{\varepsilon}\right)$$

and outputs an estimate of the mean with failure probability β and accuracy α. Moreover $\tilde{O}_k$ hides multiplicative factors that only depend on k and lower order logarithmic factors in m.

Proof. Suppose we have access to a coarse estimate of the mean with accuracy $16\sqrt{1/m}$. Then by Lemma 3.5 we have that there exists an ε-DP estimator of the mean that takes samples from the the sample mean distribution truncates them within a radius of ρ around the coarse estimate, and outputs an estimate of the mean with success probability β and accuracy

$$\alpha \leqslant_k \sqrt{\frac{1}{mn\beta}} + \frac{\rho \log(1/\beta)}{n\varepsilon} + m^{-k+1} \cdot (\rho - u)^{-k+1},$$

where $u = 16\sqrt{1/m}$. Take $\rho = \Theta\left(\sqrt{\frac{(k-1)\log m}{m}} + \left(\frac{n\varepsilon}{\log(1/\beta)}\right)^{1/k} \cdot \frac{1}{m^{1-1/k}}\right)$, hiding a sufficiently large constant to ensure that $\rho - u \geqslant \sqrt{\frac{(k-1)\log m}{m}}$. Therefore, we have that

$$\alpha \leqslant_k \sqrt{\frac{1}{mn\beta}} + \frac{\log(1/\beta)\sqrt{\log m}}{\sqrt{m}n\varepsilon} + \left(\frac{\log(1/\beta)}{mn\varepsilon}\right)^{1-1/k}.$$

Rearranging the terms and using the coarse estimation sample complexity from Theorem 3.2, with $r = 16\sqrt{1/m}$ we conclude that there exists a person-level ε-DP mean-estimation algorithm that takes

$$n = \tilde{O}_k\left(\frac{1}{\alpha^2 m \beta} + \frac{\log(1/\beta)}{\alpha \sqrt{m} \varepsilon} + \frac{\log(1/\beta)}{\alpha^{\frac{k}{k-1}} m \varepsilon} + \frac{\log(Rm/\beta)}{\varepsilon}\right),$$

many samples and outputs an estimate of the mean with success probability β and accuracy α. Moreover $\tilde{O}_k$ hides multiplicative factors that only depend on k and lower order logarithmic factors in m. □

4 Mean Estimation in High Dimensions with Approximate-DP

In this section, we introduce Algorithm 4, an algorithm for privately and efficiently estimating the mean of distributions with bounded k-th moments.

Using Clip-and-Noise in d dimensions. As in the univariate case, our algorithm for computing an estimate of the mean of a multivariate distribution is as follows: (1) compute a coarse estimate to the mean via private histograms, and (2) apply the clip-and-noise framework, clipping around the coarse estimate. As in the univariate case, our overall error is proportional to the bias induced by clipping and also the noise added to ensure privacy. To ensure overall error α, we set the clipping radius ρ to be as small as possible while still ensuring that the bias is $\leqslant \alpha$. Under this fixed ρ, we can guarantee that the noise is also $\leqslant \alpha$ so long as the number n of people is sufficiently large.

Main technical obstacles in d dimensions. In one dimension, we used a corollary of the Non-Uniform Berry-Esseen Theorem (Corollary 3.1) from [Mic76] in order to bound the bias due to clipping. Unfortunately, this bound does not apply to distributions over $\mathbb{R}^d$ for $d > 1$, and moreover the proof introduced by [Mic76] does not readily extend to higher dimensions. In this section, we prove an analogous concentration bound in high dimensions.

THEOREM 4.1. *Let $k > 2$, and let $\mathcal{D}$ have mean 0 and $\sigma_k(\mathcal{D}) \leqslant 1$. There exists $t_1 = O\left(\sqrt{\frac{d \log m}{m}}\right)$ such that, for all $t \geqslant t_1$,*

$$\mathbb{P}[\|X\|_2 \geqslant t] = \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}t^k} + e^{-mt^2/d}\right).$$

We use this concentration inequality in order to bound the bias due to Clip-and-Noise. As discussed in Section 1, the techniques used to bound the bias in the univariate case do not give tight results in the multivariate case. In this section, we introduce Theorem 4.2 whose proof circumvents this issue.

THEOREM 4.2. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ and $\sigma_k(\mathcal{D}) \leqslant 1$. Suppose that $X_j \sim \mathcal{D}$ for all $j \in [m]$, and let $X = \frac{1}{m}\sum_{j=1}^m X_j$. Let $u \in \mathbb{R}^d$, $\rho \geqslant 0$, and $\gamma = \|u - \mu\|_2$. If $\rho \geqslant t_1$ for some $t_1 = O\left(\sqrt{\frac{d \log m}{m}}\right)$, and if $\gamma \leqslant \sqrt{\frac{d}{m}}$, then,*

$$(4.1) \qquad \left\|\mu - \mathbb{E}\left[\mathrm{clip}_{\rho,u}(X)\right]\right\|_2 = \tilde{O}\left(\frac{d^{\frac{k-1}{2}}}{m^{k-1}\rho^{k-1}}\left(1 + \gamma \cdot \frac{d^{\frac{1}{2}}}{\rho}\right)\right).$$

Main algorithmic obstacle in d dimensions. Even equipped with a high-dimensional tail bound (Theorem 4.1) and bound on the bias (Theorem 4.2), there is an inherent algorithmic obstacle that exists in the multivariate case that does not exist in the univariate case.

Recall that the error of Clip-and-Noise is proportional to both the bias induced by the clipping and the noise added to ensure privacy. Thus, if the overall error is α, then each of the bias and the noise must be $\lesssim \alpha$. As seen in Theorem 4.2, the bias is proportional to the error γ of the coarse estimate. In order for the bias to be $\lesssim \alpha$, the clipping radius ρ must be large enough to offset the error γ. However, enlarging ρ also enlarges the sensitivity and thus the noise! The only way to then decrease the sensitivity is to increase the number of available people in the data set. Unfortunately, the γ produced by the coarse estimate is too large, ultimately requiring n to be larger than optimal.

To get around this problem, instead of forcing the optimal error α and settling for a sub-optimal n, we instead enforce the optimal n and observe a different error $\gamma' > \alpha$. In our parameter setting, while γ' is not smaller than α, it is *smaller* than γ; thus, we can actually interpret the clip-and-noise algorithm as boosting the error of the coarse estimate. The insight is that this process is repeatable. In particular, we will run a process called Threaded Clip-and-Noise which does exactly this, and is able to drive the error of the coarse estimate down to optimal.[9] Then, with the optimal coarse estimate, we can do one last application of Clip-and-Noise in order to achieve error α with the optimal sample complexity.

In this section, we work towards introducing Algorithm 4, an algorithm which composes the processes described above. The guarantees of Algorithm 4 are described in the following theorem.

THEOREM 4.3. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ, and let $\sigma_k(\mathcal{D}) \leqslant 1$ for some $k > 2$. For all $\varepsilon, \delta > 0$, Algorithm 4 satisfies person-level (ε, δ)-DP. Furthermore, Algorithm 4 is efficient, and for all $\varepsilon, \delta, \alpha \in (0,1)$, there exists*

$$(4.2) \qquad n_0 = \tilde{O}_k\left(\frac{d}{\alpha^2 m} + \frac{d\log^{1/2}(1/\delta)}{\alpha m^{1/2}\varepsilon} + \frac{d\log^{1/2}(1/\delta)}{\alpha^{k/(k-1)}m\varepsilon} + \frac{d\log^{1/2}(1/\delta)}{\varepsilon} + \frac{d^{1/2}\log^{3/2}(1/\delta)}{\varepsilon}\right)$$

such that, if $n \geqslant n_0$, then, with probability at least $2/3$, Algorithm 4 outputs $\hat{\mu} \in \mathbb{R}^d$ such that

$$\|\hat{\mu} - \mu\|_2 \leqslant \alpha.$$

Organization of Section 4. The remainder of this section is organized as follows. In Section 4.1, we prove Theorem 4.1 and Theorem 4.2. In Section 4.2, we discuss Threaded Clip-and-Noise. In Section 4.3 we give the full algorithm and prove Theorem 4.3.

4.1 Technical Lemmata: Bounding the Bias In this section, we prove a concentration bound for distributions in d dimensions with bounded kth moments. Using this concentration bound, we analyze the bias induced by Clip-and-Noise.

---

[9]*We can get the optimal coarse estimate for free!*

4.1.1 Theorem 4.1: Proof Overview and Core Lemmata Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$, let $X_1, \ldots, X_m \overset{\text{iid}}{\sim} \mathcal{D}$, and let $X = \frac{1}{m} \sum_{i=1}^m X_i$. We introduce the following theorem which is central in the analysis of the bias:

THEOREM 4.4. *Let $k > 2$, and let $\mathcal{D}$ have mean 0 and $\sigma_k(\mathcal{D}) \leqslant 1$. There exists $t_1 = O\left(\sqrt{\frac{d \log m}{m}}\right)$ such that, for all $t \geqslant t_1$,*

$$\mathbb{P}[\|X\|_2 \geqslant t] = \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}t^k} + e^{-mt^2/d}\right).$$

This theorem is a direct corollary of the following two theorems:

THEOREM 4.5. *Let $k > 2$, and let $\mathcal{D}$ have mean 0 and $\sigma_k(\mathcal{D}) \leqslant 1$. There exists $t_1 = O\left(\sqrt{\frac{d \log m}{m}}\right)$ and $t_2 = \Omega\left(\sqrt{d} \log^{\frac{-1}{k-2}} m\right)$ such that, for all $t \in [t_1, t_2]$,*

$$\mathbb{P}[\|X\|_2 \geqslant t] = \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}t^k} + e^{-mt^2/d}\right).$$

THEOREM 4.6. *Let $k > 2$, and let $\mathcal{D}$ have mean 0 and $\sigma_k(\mathcal{D}) \leqslant 1$. There exists $t_3 = O\left(\sqrt{d} \log^{1/k} m\right)$ such that, for all $t \geqslant t_3$,*

$$\mathbb{P}[\|X\|_2 \geqslant t] = \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}t^k}\right).$$

In this subsection, we give a proof overview of both Theorem 4.5 and Theorem 4.6 (whose combination yields Theorem 4.1). As we give the overview, we will also develop some high-level lemmas that will be used in both analyses. Building on these lemmas, we will give the full proofs of the theorems in the following subsections.

As notation, we say that a random variable $Z \in \mathbb{R}^d$ has a *t-deviation* if $\|Z\|_2 \geqslant t$. In Theorem 4.5, we analyze the probability that X has a t-deviation for all $\sqrt{d/m} \lesssim t \lesssim \sqrt{d}$ (we call all t in this range *small t*). In Theorem 4.6, we analyze the probability that X has a t-deviation for all $t \gtrsim \sqrt{d}$ (we call all t in this range *large t*). The analyses of the small-t event and the large-t event closely follow the same blueprint, which we describe in this subsection.

To start, let $r_1, r_2 \geqslant 0$ such that $r_1 < r_2$. We call a sample X_i *light* if $\|X_i\|_2 < r_1$. We call a sample X_i *moderate* if $\|X_i\|_2 \in [r_1, r_2)$. We call a sample X_i *heavy* if $\|X_i\|_2 \geqslant r_2$. We can group together the light, moderate, and heavy samples as follows:

$$A_1 = \frac{1}{m} \sum_{i=1}^m X_i \cdot \mathbb{I}(\|X_i\|_2 < r_1)$$

$$A_2 = \frac{1}{m} \sum_{i=1}^m X_i \cdot \mathbb{I}(r_1 \leqslant \|X_i\|_2 < r_2)$$

$$A_3 = \frac{1}{m} \sum_{i=1}^m X_i \cdot \mathbb{I}(\|X_i\|_2 \geqslant r_2)$$

We can use these groupings to expand the probability that X has a t-deviation:

(4.3) $$\mathbb{P}[\|X\|_2 \geqslant t] \leqslant \mathbb{P}\left[\|A_1\|_2 \geqslant \frac{t}{3}\right] + \mathbb{P}\left[\|A_2\|_2 \geqslant \frac{t}{3}\right] + \mathbb{P}\left[\|A_3\|_2 \geqslant \frac{t}{3}\right].$$

The proofs of Theorem 4.5 and Theorem 4.6 focus on bounding the probabilities that each of A_1, A_2, and A_3 have $(t/3)$-deviations.

In order to bound these events, we must take care to set r_1 and r_2 appropriately as functions of t. In particular, for both the small-t regime and the large-t regime, we will always set $r_2 = \frac{mt}{3\log m} \approx mt$. That is, we should think of r_2 as being a global variable for both the small-t regime and the large-t regime. On the other hand, we should think of r_1 as being a local variable which is set differently depending on whether t is small or large (this is actually why we divide t into small t and large t in the first place). In the small-t regime, we set $r_1 = d/t$. In the large-t regime, we set $r_1 = t/3$.

As it turns out, the proof bounding the deviation of the heavy samples is agnostic to whether t is small or large. On the other hand, the proofs bounding the deviation of the light samples will necessarily be different for small t and large t. The proofs bounding the deviation of the moderate samples are largely similar for small t and large t. We focus for now on proving bounds for the deviations of just the heavy and moderate samples. (In ensuing subsections, we bound the probability that A_1 has a $(t/3)$-deviation separately for small t and for large t.)

Throughout the proof, we make extensive use of the following lemma:

LEMMA 4.1. (TAIL BOUND FOR THE NORM OF BOUNDED k-TH MOMENT DISTRIBUTION [ZJS22]) *Assume p has its k-th moment bounded by* 1 *for $k \geqslant 2$. Then for all $t > 0$,*

$$\mathbb{P}[\|X\|_2 \geqslant t] \leqslant d^{k/2}t^{-k}.$$

Bounding the deviation of the heavy samples. We begin by introducing Lemma 4.2, which bounds the probability that A_3 has a $(t/3)$-deviation. This is the easiest step in the analysis and is used in both the small-t and large-t regimes. As mentioned earlier, we set $r_2 = \frac{mt}{3\log m}$.

LEMMA 4.2. (DEVIATION OF HEAVY SAMPLES) *Let $k \geqslant 2$, and let the distribution $\mathcal{D}$ over $\mathbb{R}^d$ have mean* 0 *and k-th moment bounded by* 1. *Then for all $t > 0$,*

$$(4.4) \qquad \mathbb{P}\left[\|A_3\|_2 \geqslant \frac{t}{3}\right] \leqslant \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}t^k}\right).$$

Proof. Note that

$$(4.5) \qquad \mathbb{P}\left[\|A_3\|_2 \geqslant \frac{t}{3}\right] = \mathbb{P}\left[\left\|\frac{1}{m}\sum_{\substack{i \\ \|X_i\|_2 \geqslant r_2}} X_i\right\|_2 \geqslant \frac{t}{3}\right]$$

$$\leqslant \mathbb{P}\left[\sum_{\substack{i \\ \|X_i\|_2 \geqslant r_2}} \|X_i\|_2 \geqslant \frac{mt}{3}\right]$$

$$\leqslant \mathbb{P}\left[\sum_{\substack{i \\ \|X_i\|_2 \geqslant r_2}} \|X_i\|_2 \geqslant r_2\right]$$

$$= \mathbb{P}[\exists i \text{ such that } \|X_i\|_2 \geqslant r_2]$$

$$(4.6) \qquad \leqslant m\,\mathbb{P}[\|X_1\|_2 \geqslant r_2],$$

where the last line follows from the Union Bound. We now use Lemma 4.1 to bound the probability of the norm of one sample having an r_2-deviation:

$$(4.7) \qquad \mathbb{P}[\|X_1\|_2 \geqslant r_2] \leqslant r_2^{-k}d^{k/2} = \tilde{O}\left(\frac{d^{k/2}}{m^k t^k}\right),$$

which combined with (4.6) completes the proof. $\square$

Bounding the deviation of the moderate samples. We give an overview of how to bound the probability of the moderate samples deviating for both small t and large t. This is the most interesting part of the analysis.

We begin by partitioning the moderate samples further. WLOG assume that r_2/r_1 is a power of two, and for all integers $\ell \in \{1, \ldots, \log(r_2/r_1)\}$ we define the interval

$$B_\ell = \left\{ i : 2^{-\ell} \cdot r_2 \leqslant \|X_i\|_2 < 2^{-\ell+1} \cdot r_2 \right\}.$$

Note that, as ℓ gets larger, the interval B_ℓ shrinks in width and also covers a smaller range of values. Consider the following remark about these intervals.

REMARK 4.1. *If $\log(r_2/r_1) \leqslant \log m$ and A_2 has a $(t/3)$-deviation, then there exists $\ell \in \{1, \ldots, \log(r_2/r_1)\}$ such that $|B_\ell| \geqslant 2^{\ell-1}$.*

Proof. Suppose for contradiction that $\forall \ell : |B_\ell| < 2^{\ell-1}$. Then,

$$\|A_2\|_2 = \left\| \frac{1}{m} \sum_{\substack{i \\ \|X_i\|_2 \in [r_1, r_2)}} X_i \right\|_2$$

$$\leqslant \frac{1}{m} \sum_{\substack{i \\ \|X_i\|_2 \in [r_1, r_2)}} \|X_i\|_2$$

$$= \frac{1}{m} \sum_{\ell=1}^{\log(r_2/r_1)} \sum_{i \in B_\ell} \|X_i\|_2$$

$$< \frac{1}{m} \sum_{\ell=1}^{\log(r_2/r_1)} 2^{\ell-1} \cdot (2^{-\ell+1} \cdot r_2)$$

(4.8) $$= \frac{\log(r_2/r_1) r_2}{m}.$$

By assumption on $\log(r_2/r_1) \leqslant \log m$, equation (4.8) becomes

$$\|A_2\|_2 < \frac{\log(r_2/r_1) r_2}{m} \leqslant \frac{\log m}{m} \cdot \frac{mt}{3 \log m} \leqslant \frac{t}{3},$$

which is a contradiction. $\quad\square$

We will assume $\log(r_2/r_1) \leqslant \log m$ and use Remark 4.1 to expand the probability that A_2 has a $(t/3)$-deviation:

(by Remark 4.1) $$\mathbb{P}\left[\|A_2\|_2 \geqslant \frac{t}{3} \right] \leqslant \sum_{\ell=1}^{\log(r_2/r_1)} \mathbb{P}\left[|B_\ell| \geqslant 2^{\ell-1} \right]$$

(by Union Bound) $$\leqslant \sum_{\ell=1}^{\log(r_2/r_1)} \binom{m}{2^{\ell-1}} \mathbb{P}\left[\|X_i\|_2 \geqslant 2^{-\ell} \cdot r_2 \right]^{2^{\ell-1}}$$

(by Lemma 4.1 and Lemma B.1) $$\leqslant \sum_{\ell=1}^{\log(r_2/r_1)} \left(\frac{em}{2^{\ell-1}} \right)^{2^{\ell-1}} \left(\left(\frac{2^\ell \cdot \sqrt{d}}{r_2} \right)^k \right)^{2^{\ell-1}}$$

$$\leqslant \sum_{\ell=1}^{\log(r_2/r_1)} \left(\frac{e \cdot m \cdot d^{k/2} \cdot 2^{\ell k-\ell+1}}{r_2^k} \right)^{2^{\ell-1}}$$

(4.9) $$\leqslant \log(r_2/r_1) \cdot \max_{\ell \in \{1, \ldots, \log(r_2/r_1)\}} \left(\frac{e \cdot m \cdot d^{k/2} \cdot 2^{\ell k-\ell+1}}{r_2^k} \right)^{2^{\ell-1}}$$

Since $\log(r_2/r_1) = \tilde{O}(1)$ and $r_2 = \frac{mt}{3\log m}$, then the rest of the proof will reduce to showing

$$(4.10) \qquad \max_\ell \left(\frac{2 \cdot 3^k e \cdot d^{k/2} \cdot 2^{\ell(k-1)} \log^k m}{m^{k-1} t^k} \right)^{2^{\ell-1}} = \tilde{O}\left(\frac{d^{k/2}}{m^{k-1} t^k} \right).$$

Showing (4.10) will follow from algebraic calculations. To lighten the notation, we will define the function

$$f(\ell) = \Theta\left(\frac{d^{k/2} \cdot 2^{\ell(k-1)} \log^k m}{m^{k-1} t^k} \right)^{2^{\ell-1}},$$

where the Θ-notation hides the $2 \cdot 3^k e$ term in (4.10). To bound $\max_\ell f(\ell)$, it helps to observe Lemma 4.3 which states that $f(\ell)$ is convex with respect to ℓ (see Appendix B.1 for the proof).

LEMMA 4.3. *Fix any $m, d, t > 0$ and $k \geqslant 1$. Then the function $f(\ell)$ is convex with respect to ℓ.*

By Lemma 4.3, it follows that

$$(4.11) \qquad \max_\ell f(\ell) \in \{f(1), f(\log(r_2/r_1)\},$$

and so it suffices to show that $f(1)$ and $f(\log(r_2/r_1))$ are bounded. Note that, for both small t and large t,

$$(4.12) \qquad f(1) = \Theta\left(\frac{d^{k/2} \cdot 2^{k-1} \log^k m}{m^{k-1} t^k} \right) = \tilde{O}\left(\frac{d^{k/2}}{m^{k-1} t^k} \right).$$

On the other hand, we must bound $f(\log(r_2/r_1))$ separately for small t and large t, as in each analysis we will set r_1 differently. We leave this subsection and bookmark our progress with the following lemma:

LEMMA 4.4. *If $\log(r_2/r_1) \leqslant \log m$ and $f(\log(r_2/r_1)) = \tilde{O}(d^{k/2} m^{-k+1} t^{-k})$, then*

$$\mathbb{P}\left[\|A_2\|_2 \geqslant \frac{t}{3} \right] = \tilde{O}\left(\frac{d^{k/2}}{m^{k-1} t^k} \right).$$

4.1.2 The Small-t Regime: Proving Theorem 4.5

In this section, we prove the following theorem:

THEOREM 4.7. *Let $k > 2$, and let $\mathcal{D}$ have mean 0 and $\sigma_k(\mathcal{D}) \leqslant 1$. There exists $t_1 = O\left(\sqrt{\frac{d \log m}{m}} \right)$ and $t_2 = \Omega\left(\sqrt{d} \log^{\frac{-1}{k-2}} m \right)$ such that, for all $t \in [t_1, t_2]$,*

$$\mathbb{P}[\|X\|_2 \geqslant t] = \tilde{O}\left(\frac{d^{k/2}}{m^{k-1} t^k} + e^{-mt^2/d} \right).$$

Let $r_1 = \frac{d}{t}$, $r_2 = \frac{mt}{3\log m}$, and let A_1, A_2, and A_3 be defined as follows:

$$A_1 = \frac{1}{m} \sum_{i=1}^m X_i \cdot \mathbb{I}(\|X_i\|_2 < r_1)$$

$$A_2 = \frac{1}{m} \sum_{i=1}^m X_i \cdot \mathbb{I}(r_1 \leqslant \|X_i\|_2 < r_2)$$

$$A_3 = \frac{1}{m} \sum_{i=1}^m X_i \cdot \mathbb{I}(\|X_i\|_2 \geqslant r_2)$$

As discussed earlier, we must bound the probabilities that A_1, A_2, and A_3 have $(t/3)$-deviations. We have already bounded the heavy samples by Lemma 4.2. We move on to showing the complete proof of the bound for the moderate samples in Lemma 4.5:

LEMMA 4.5. (DEVIATION OF MODERATE SAMPLES IN SMALL-t REGIME) *Let $k > 2$, $\mathcal{D}$ have mean 0 and k-th moment bounded by 1, and $t > 0$. Then, there exists $t_1 = O\left(\sqrt{\frac{d \log m}{m}}\right)$ and $t_2 = \Omega(\sqrt{d} \log^{\frac{-1}{k-2}} m)$ such that, for all $t \in [t_1, t_2]$,*

$$(4.13) \qquad \mathbb{P}\left[\|A_2\|_2 \geqslant \frac{t}{3}\right] \leqslant \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}t^k}\right).$$

Proof. We seek to apply Lemma 4.4 (as described in the previous subsection). Note that there exists $t_2 = \Omega(\sqrt{d} \log^{\frac{-1}{k-2}} m)$ such that, for all $t \leqslant t_2$,

$$\log(r_2/r_1) = \log\left(\frac{mt^2}{3d \log m}\right) \leqslant \log m.$$

Thus, all that remains is proving that $f(\log(r_2/r_1)) = \tilde{O}(d^{k/2} m^{-k+1} t^{-k})$. We defer the proof of this to Lemma B.6 in Appendix B.2. $\square$

We now introduce Lemma 4.6, which shows that the deviation of the light samples is bounded.

LEMMA 4.6. (DEVIATION OF LIGHT SAMPLES IN SMALL-t REGIME) *Let $k > 2$, and let the distribution $\mathcal{D}$ over $\mathbb{R}^d$ have mean 0 and k-th moment bounded by 1. Then, there exists $t_2 = \Omega(\sqrt{d})$ such that, for all positive $t \leqslant t_2$,*

$$(4.14) \qquad \mathbb{P}\left[\|A_1\|_2 \geqslant \frac{t}{3}\right] \leqslant \exp\left(-\Theta\left(\frac{mt^2}{d}\right)\right).$$

Lemma 4.6 critically uses a Bernstein Inequality for vectors (see, e.g., Lemma 18 of [KL17]):

LEMMA 4.7. (VECTOR BERNSTEIN INEQUALITY) *Let $Z_1, \ldots, Z_m$ be independent random vectors in $\mathbb{R}^d$ such that $\mathbb{E}[Z_i] = 0$, $\mathbb{E}[\|Z_i\|_2^2] \leqslant \sigma^2$ and $\|Z_i\|_2 \leqslant r$ for all i. Then for all $0 < t < \sigma^2/r$,*

$$\mathbb{P}\left[\left\|\sum_{i=1}^{m} Z_i\right\|_2 \geqslant mt\right] \leqslant \exp\left(\frac{-mt^2}{8\sigma^2} + \frac{1}{4}\right).$$

We now prove Lemma 4.6.

Proof. For each $i \in [m]$, set $Y_i = X_i \cdot \mathbb{I}(\|X_i\|_2 < r_1)$, and $Z_i = Y_i - \mathbb{E}[Y_i]$. We can rewrite the left-hand side of (4.14) in terms of Z_i and $\mathbb{E}[Y_i]$ as follows:

$$\mathbb{P}\left[\|A_1\|_2 \geqslant \frac{t}{3}\right] = \mathbb{P}\left[\left\|\frac{1}{m} \sum_{\substack{i \\ \|X_i\|_2 < r_1}} X_i\right\|_2 \geqslant \frac{t}{3}\right]$$

$$= \mathbb{P}\left[\left\|\sum_{\substack{i \\ \|X_i\|_2 < r_1}} X_i\right\|_2 \geqslant \frac{mt}{3}\right]$$

$$= \mathbb{P}\left[\left\|\sum_{i=1}^{m} Y_i\right\|_2 \geqslant \frac{mt}{3}\right]$$

$$= \mathbb{P}\left[\left\|\sum_{i=1}^{m} Y_i - \mathbb{E}[Y_i] + \mathbb{E}[Y_i]\right\|_2 \geqslant \frac{mt}{3}\right]$$

$$= \mathbb{P}\left[\left\|\sum_{i=1}^{m} Z_i + \mathbb{E}[Y_i]\right\|_2 \geq \frac{mt}{3}\right]$$

(by the Triangle Inequality)
$$\leq \mathbb{P}\left[\left\|\sum_{i=1}^{m} Z_i\right\|_2 \geq m\left(\frac{t}{3} - \|\mathbb{E}[Y_i]\|_2\right)\right].$$

We have that $\|\mathbb{E}[Y_i]\|_2 \leq t/6$ by Lemma B.3, which uses the fact that $\mathcal{D}$ has a bounded k-th moment (see Appendix B.2 for the proof). This gives us

$$\mathbb{P}\left[\|A_1\|_2 \geq \frac{t}{3}\right] \leq \mathbb{P}\left[\left\|\sum_{i=1}^{m} Z_i\right\|_2 \geq \frac{mt}{6}\right].$$

Applying Lemma 4.7 to the expression above—using the parameters $\sigma^2 = d$ and $r = 2r_1$—completes the proof. This is done in detail in Lemma B.4 and Lemma B.5 in Section B.2. In particular, Lemma B.4 checks the constraints on the Z_i's, and Lemma B.5 performs the necessary algebraic manipulation to complete the proof. $\quad\square$

We now prove Theorem 4.5.

Proof. We have that

$$\mathbb{P}[\|X\|_2 \geq t] \leq \mathbb{P}\left[\|A_1\|_2 \geq \frac{t}{3}\right] + \mathbb{P}\left[\|A_2\|_2 \geq \frac{t}{3}\right] + \mathbb{P}\left[\|A_3\|_2 \geq \frac{t}{3}\right]$$

(by Lemma 4.6, Lemma 4.5, and Lemma 4.2)
$$\leq \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}t^k} + e^{-mt^2/d}\right).$$

$\square$

4.1.3 The Large-t Regime: Proving Theorem 4.6

THEOREM 4.8. *Let $k > 2$, and let $\mathcal{D}$ have mean 0 and $\sigma_k(\mathcal{D}) \leq 1$. There exists $t_3 = O\left(\sqrt{d}\log^{1/k} m\right)$ such that, for all $t \geq t_3$,*

$$\mathbb{P}[\|X\|_2 \geq t] = \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}t^k}\right).$$

Proof. Let $r_1 = \frac{t}{3}$, $r_2 = \frac{mt}{3\log m}$ and let A_1, A_2, and A_3 be defined as follows:

$$A_1 = \frac{1}{m}\sum_{i=1}^{m} X_i \cdot \mathbb{I}(\|X_i\|_2 < r_1)$$

$$A_2 = \frac{1}{m}\sum_{i=1}^{m} X_i \cdot \mathbb{I}(r_1 \leq \|X_i\|_2 < r_2)$$

$$A_3 = \frac{1}{m}\sum_{i=1}^{m} X_i \cdot \mathbb{I}(\|X_i\|_2 \geq r_2).$$

We have that

$$\mathbb{P}[\|X\|_2 \geq t] \leq \mathbb{P}\left[\|A_1\|_2 \geq \frac{t}{3}\right] + \mathbb{P}\left[\|A_2\|_2 \geq \frac{t}{3}\right] + \mathbb{P}\left[\|A_3\|_2 \geq \frac{t}{3}\right]$$

$$= 0 + \mathbb{P}\left[\|A_2\|_2 \geq \frac{t}{3}\right] + \mathbb{P}\left[\|A_3\|_2 \geq \frac{t}{3}\right]$$

(by Lemma 4.2)
$$\leq \mathbb{P}\left[\|A_2\|_2 \geq \frac{t}{3}\right] + \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}t^k}\right).$$

To bound $\mathbb{P}\left[\|A_2\|_2 \geq \frac{t}{3}\right]$, we seek to apply Lemma 4.4. When $r_1 = t/3$,

$$\log(r_2/r_1) = \log\left(\frac{m}{\log m}\right) \leq \log m.$$

Thus, all that remains is proving that $f(\log(r_2/r_1)) = \tilde{O}(d^{k/2}m^{-k+1}t^{-k})$. Note that

$$f\left(\log\left(\frac{m}{\log m}\right)\right) = \Theta\left(\frac{md^{k/2}\left(\frac{m}{\log m}\right)^{k-1}\log^k m}{m^k t^k}\right)^{\frac{m}{2\log m}}$$

$$= \Theta\left(\frac{d^{k/2}\log m}{t^k}\right)^{\frac{m}{2\log m}}$$

(since $t \geq t_3$)
$$\leq \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}t^k}\right).$$

This completes the proof. $\square$

As an immediate corollary, we have:

COROLLARY 4.1. *Let $k > 2$, and let $\mathcal{D}$ have mean μ and $\sigma_k(\mathcal{D}) \leq 1$. There exists $t_1 = O\left(\sqrt{\frac{d\log m}{m}}\right)$ such that, for all $t \geq t_1$,*

$$(4.15) \qquad \mathbb{P}[\|X - \mu\|_2 \geq t] = \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}t^k}\right).$$

Proof. Without loss of generality, we can assume that $\mathcal{D}$ has arbitrary mean μ. Thus, to complete the proof, we show that the exponential term in Theorem 4.1 is dominated by the polynomial term in Theorem 4.1 for all $t \geq t_0\sqrt{\frac{kd\log m}{m}}$. Set $t = c\sqrt{\frac{kd\log m}{m}}$ for some $c \geq t_0$. Observe that, for all $c \geq t_0$,

$$\exp\left(-\frac{mt^2}{d}\right) = \exp\left(-c^2 k\log m\right) \leq \frac{1}{m^{kc^2}} \leq \frac{1}{m^k 2^{kc^2}} \leq \frac{1}{m^k 2^{k\log c}} \leq \frac{1}{m^k c^k}$$

$$\leq \frac{1}{m^{k/2-1}c^k}$$

$$= \frac{d^{k/2}}{m^{k-1}t^k}.$$

$\square$

4.1.4 Using Corollary 4.1 to Bound the Bias We now define the operation $\text{clip}_{\rho,u}$. The operation takes as input some $X \in \mathbb{R}^d$ and clips X to the ℓ_2 ball centered at $u \in \mathbb{R}^d$ with radius ρ.

DEFINITION 4.1. (CLIPPING OPERATION) *For $\rho \geq 0$ and $u \in \mathbb{R}^d$, the operation $\text{clip}_{\rho,u} : \mathbb{R}^d \to \mathbb{R}^d$ is defined as follows:*

$$\text{clip}_{\rho,u}(x) := \begin{cases} x & \|x-u\|_2 \leq \rho \\ u + \rho \cdot \frac{x-u}{\|x-u\|_2} & \text{otherwise} \end{cases}.$$

We now introduce the following theorem, which describes the bias of the statistical estimator that averages m samples from some distribution $\mathcal{D}$ with bounded k-th moments and applies $\text{clip}_{\rho,u}$ to the average.

THEOREM 4.9. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ and $\sigma_k(\mathcal{D}) \leq 1$. Suppose that $X_j \sim \mathcal{D}$ for all $j \in [m]$, and let $X = \frac{1}{m} \sum_{j=1}^m X_j$. Let $u \in \mathbb{R}^d$, $\rho \geq 0$, and $\gamma = \|u - \mu\|_2$. If $\rho \geq t_1$ for some $t_1 = O\left(\sqrt{\frac{d \log m}{m}}\right)$, and if $\gamma \leq \sqrt{\frac{d}{m}}$, then,*

$$(4.1) \qquad \left\|\mu - \mathbb{E}\left[\mathrm{clip}_{\rho,u}(X)\right]\right\|_2 = \tilde{O}\left(\frac{d^{\frac{k-1}{2}}}{m^{k-1}\rho^{k-1}}\left(1 + \gamma \cdot \frac{d^{\frac{1}{2}}}{\rho}\right)\right).$$

Proof. Define $Z = \mathrm{clip}_{\rho,u}(X)$, and define

$$v = \frac{\mathbb{E}[X - Z]}{\|\mathbb{E}[X - Z]\|_2}.$$

We use v to expand the expression for the bias:

$$\|\mu - \mathbb{E}[Z]\|_2 = \|\mathbb{E}[X - Z]\|_2 = \mathbb{E}[\langle X - Z, v\rangle]$$

$$\leq \int_{\ell=0}^{\infty} \mathbb{P}[\langle X - Z, v\rangle \geq \ell]\, d\ell$$

$$(4.16) \qquad = \underbrace{\int_{\ell=0}^{\rho/\sqrt{d}+\gamma} \mathbb{P}[\langle X - Z, v\rangle \geq \ell]\, d\ell}_{T_1 :=} + \underbrace{\int_{\ell=\rho/\sqrt{d}+\gamma}^{\infty} \mathbb{P}[\langle X - Z, v\rangle \geq \ell]\, d\ell}_{T_2 :=}.$$

The rest of the proof is dedicated to bounding T_1 and T_2 in (4.16).

Bounding T_1. Our goal is to bound T_1 using Corollary 4.1. Be begin by bounding the projection of $X - Z$ onto v by the ℓ_2 norm of $X - Z$:

$$T_1 = \int_{\ell=0}^{\rho/\sqrt{d}+\gamma} \mathbb{P}[\langle X - Z, v\rangle \geq \ell]\, d\ell \leq \int_{\ell=0}^{\rho/\sqrt{d}+\gamma} \mathbb{P}[\|X - Z\|_2 \geq \ell]\, d\ell$$

For $\ell > 0$, one can check that $\|X - Z\|_2 \geq \ell$ implies $\|X - \mu\|_2 \geq \ell + \rho - \gamma$. (See Lemma B.7 in Appendix B.3 for details). Thus,

$$T_1 \leq \int_{\ell=0}^{\rho/\sqrt{d}+\gamma} \mathbb{P}[\|X - \mu\|_2 \geq \ell + \rho - \gamma]\, d\ell$$

(by assumption on γ and ρ) $\qquad \leq \int_{\ell=0}^{\rho/\sqrt{d}+\gamma} \mathbb{P}[\|X - \mu\|_2 \geq \ell + \rho/2]\, d\ell$

(since $\ell \geq 0$) $\qquad \leq \int_{\ell=0}^{\rho/\sqrt{d}+\gamma} \mathbb{P}[\|X - \mu\|_2 \geq \rho/2]\, d\ell$

(by Corollary 4.1 and assumption on ρ) $\qquad = \lim_{R \to 0^+} \int_{\ell=R}^{\rho/\sqrt{d}+\gamma} \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}\rho^k}\right) d\ell$

$$(4.17) \qquad = \tilde{O}\left(\frac{d^{\frac{k-1}{2}}}{m^{k-1}\rho^{k-1}}\left(1 + \gamma \cdot \frac{d^{\frac{1}{2}}}{\rho}\right)\right).$$

We now focus bounding T_2.

Bounding T_2. Our goal is to bound T_2 using Corollary 4.1 for univariate distributions. We start by expanding $\langle X - \mu, v\rangle$:

$$\langle X - \mu, v\rangle = \langle X - Z + Z - u + u - \mu, v\rangle$$

$$= \langle X - Z, v\rangle + \langle Z - u, v\rangle + \langle u - \mu, v\rangle$$

$$(4.18) \qquad \geq \langle X - Z, v\rangle + \langle Z - u, v\rangle - \gamma.$$

We use (4.18) to expand T_2:

$$T_2 = \int_{\ell=\rho/\sqrt{d}+\gamma}^{\infty} \mathbb{P}[\langle X - Z, v \rangle \geqslant \ell]\, \mathrm{d}\ell$$

(by (4.18))
$$\leqslant \int_{\ell=\rho/\sqrt{d}+\gamma}^{\infty} \mathbb{P}[\langle X - \mu, v \rangle \geqslant \ell + \langle Z - u, v \rangle - \gamma]\, \mathrm{d}\ell$$

Note that $X - Z$ and $Z - u$ are non-negative scalar multiples of each other. Thus, the inequality $\langle X - Z, v \rangle \geqslant 0$ (implied by $\langle X - Z, v \rangle \geqslant \ell$) implies $\langle Z - u, v \rangle \geqslant 0$. Thus,

$$T_2 \leqslant \int_{\ell=\rho/\sqrt{d}+\gamma}^{\infty} \mathbb{P}[\langle X - \mu, v \rangle \geqslant \ell - \gamma]\, \mathrm{d}\ell$$

(by a change of variables)
$$\leqslant \int_{\ell'=\rho/\sqrt{d}}^{\infty} \mathbb{P}[\langle X - \mu, v \rangle \geqslant \ell']\, \mathrm{d}\ell'.$$

Let $Y = \langle X, v \rangle$. Note that $\mathbb{E}[Y] = \mathbb{E}[\langle X, v \rangle] = \langle \mu, v \rangle$ and also that

$$\sigma_k(Y) = \mathbb{E}\left[|Y - \mathbb{E}[Y]|^k\right]^{1/k} = \mathbb{E}\left[|\langle X - \mu, v \rangle|^k\right]^{1/k} \leqslant 1$$

by assumption on $\mathcal{D}$. Thus, we can use Corollary 4.1 in the case of univariate distributions:

$$T_2 \leqslant \int_{\ell'=\rho/\sqrt{d}}^{\infty} \mathbb{P}[\langle X - \mu, v \rangle \geqslant \ell']\, \mathrm{d}\ell'$$

$$= \int_{\ell'=\rho/\sqrt{d}}^{\infty} \mathbb{P}[Y - \mathbb{E}[Y] \geqslant \ell']\, \mathrm{d}\ell'$$

$$\leqslant \int_{\ell'=\rho/\sqrt{d}}^{\infty} \mathbb{P}[|Y - \mathbb{E}[Y]| \geqslant \ell']\, \mathrm{d}\ell'$$

(by Corollary 4.1 and assumption on ρ)
$$\leqslant \int_{\ell'=\rho/\sqrt{d}}^{\infty} \tilde{O}\left(\frac{1}{m^{k-1}(\ell')^k}\right)\, \mathrm{d}\ell'$$

(4.19)
$$\leqslant \tilde{O}\left(\frac{\sqrt{d}}{m\rho}\right)^{k-1}.$$

Combining (4.16), (4.17), and (4.19) completes the proof. $\quad\square$

4.2 Threaded Clip-and-Noise In this section, we introduce the algorithm Threaded Clip-and-Noise (TCN). As discussed in the introduction of Section 4, Threaded Clip-and-Noise takes as input an initial coarse estimate u_0 to the mean and outputs a new estimate u^*. The formal guarantees of Threaded Clip-and-Noise are described in Theorem 4.10.

THEOREM 4.10. *Let* $S = \{S_1, \ldots, S_n\}$, *where each* $S_i = \left\{X_1^{(i)}, \ldots, X_m^{(i)}\right\}$ *and each* $X_j^{(i)} \in \mathbb{R}^d$. *Let* $u_0 \in \mathbb{R}^d$. *For all* $\varepsilon, \delta \in (0, 1)$, $\mathrm{TCN}(S; \varepsilon, \delta, u_0)$ *satisfies person-level* (ε, δ)-*DP. Furthermore, let* $\mathcal{D}$ *be a distribution over* $\mathbb{R}^d$ *with mean* μ *and* $\sigma_k(\mathcal{D}) \leqslant 1$. *Suppose that* $X_j^{(i)} \sim \mathcal{D}$ *for all* $i \in [n]$ *and* $j \in [m]$, *and suppose that* $\|u_0 - \mu\|_2 \leqslant \frac{d^{1/2}}{m^{1/2}}$. *Let* u^* *be the output of* $\mathrm{TCN}(S; \varepsilon, \delta, u_0)$. *There exists* $n_0 = \tilde{O}\left(\frac{d\sqrt{\log(1/\delta)}}{\varepsilon}\right)$, *such that, if* $n \geqslant n_0$, *then* $\|u^* - \mu\|_2 = \tilde{O}\left(\frac{1}{m^{1/2}}\right)$ *with probability at least* 0.9.

Overview of Threaded Clip-and-Noise. We give a high-level overview of Threaded Clip-and-Noise (see Algorithm 3 for the full details). The algorithm is composed of a total of T layers. In each layer $t \in [T]$, the algorithm runs an application of the high-dimensional clip-and-noise algorithm ClipAndNoise (Algorithm 2) and

outputs $u_t \in \mathbb{R}^d$. The name Threaded Clip-and-Noise comes from the fact that the layers are threaded together: the output u_{t-1} is fed as input into layer t, and it is used as the center of the clipping ball in the application of ClipAndNoise.

To discuss Threaded Clip-and-Noise in more detail, we introduce the following notation. For each $t \in [T]$ and $i \in [n]$, define $Z_i(t) = \text{clip}_{\rho_t, u_{t-t}}\left(X^{(i)}\right)$ where each $X^{(i)} = \frac{1}{m}\sum_{j=1}^m X_j^{(i)}$ and where ρ_t is a value specified at each layer t. Let W_t be the noise added by ClipAndNoise in layer t. The output of layer t is

$$u_t = \frac{1}{n}\sum_{i=1}^n Z_i(t) + W_t.$$

As further notation, let $\gamma_t = \|u_t - \mu\|_2$. We refer to γ_t as *the error of u_t*.

Algorithm 2 Clip-and-Noise

Input: $S = \left\{X_j^{(i)}\right\}_{i \in [n], j \in [m]}$ where each $X_j^{(i)} \in \mathbb{R}^d$. Parameters $\varepsilon, \delta, \rho, u$.

Output: $u' \in \mathbb{R}^d$.

1: **procedure** CLIPANDNOISE($S;\varepsilon,\delta,\rho,u$)
2:     **for** $i \in [n]$ **do**
3:         $Z_i \leftarrow \text{clip}_{\rho,u}\left(\frac{1}{m}\sum_{j=1}^m X_j^{(i)}\right)$.
4:     **end for**
5:     Let $W \sim \mathcal{N}\left(0, \left(\frac{2\rho\sqrt{2\ln(4/\delta)}}{n\varepsilon}\right)^2 \cdot \mathbb{I}_{d \times d}\right)$.
6:     Return $u' = \frac{1}{n}\sum_{i=1}^n Z_i + W$.
7: **end procedure**

The iterative refinements of Threaded Clip-and-Noise. The layered/threaded design choices of Threaded Clip-and-Noise hinge on the following key idea: Each layer t produces an estimate u_t that is better than that of the previous layer; that is, for all $t \in [T]$, the algorithm produces γ_t such that $\gamma_t < \gamma_{t-1}$. In addition, by the time the algorithm terminates, the final estimate u_T will have error $\lesssim \sqrt{1/m}$ (so long as the initial estimate u_0 had error $\lesssim \sqrt{d/m}$).

To see why this is true, it is useful to understand how γ_t evolves with t. We can think of γ_t as being roughly equal to the sum of a bias term and a noise term, both of which come from unraveling the details of ClipAndNoise. In our parameter regime, the bias term will be

$$(4.20) \qquad\qquad\qquad\qquad \text{bias}(t) \approx \frac{d^{k/2}\gamma_{t-1}}{m^{k-1}\rho_t^k}$$

due to Theorem 4.2. When $n \gtrsim \frac{d\sqrt{\log(1/\delta)}}{\varepsilon}$, the noise term will be

$$\text{noise}(t) \approx \frac{\rho_t}{\sqrt{d}}.$$

Notice that both terms are functions of ρ_t, which is a parameter set by the algorithm at each layer. In order to minimize γ_t, the algorithm sets ρ_t to be the value that balances $\text{bias}(t)$ and $\text{noise}(t)$, i.e. setting

$$\rho_t \approx \frac{(\gamma_{t-1})^{\frac{1}{k+1}} d^{\frac{1}{2}}}{m^{\frac{k-1}{k+1}}}$$

(where we pretend for now that the algorithm knows the value of γ_{t-1}). Plugging this value of ρ_t into γ_t gives the following evolution of γ_t with respect to t:

$$\gamma_t \approx \frac{(\gamma_{t-1})^{\frac{1}{k+1}}}{m^{\frac{k-1}{k+1}}}.$$

If we assume that $\gamma_0 \leqslant \sqrt{d/m}$, then it is ideal to set

$$(4.21) \qquad \rho_t = \frac{d^{\frac{1}{2} + \frac{1}{2k(k+1)^{t-1}}}}{m^{1-1/k}}$$

at each round, yielding that $\gamma_t = \frac{d^{\frac{1}{2k(k+1)^{t-1}}}}{m^{1-1/k}}$. After $\log\log d$ layers, it follows that $\gamma_{\log\log d} \lesssim \sqrt{\frac{1}{m}}$.

The gross oversimplification above glosses over one important detail: We can only apply Theorem 4.2 in (4.20) if we assume that $\rho_t \gtrsim \sqrt{d/m}$. Thus, if the right-hand side of (4.21) ever dips below $\approx \sqrt{d/m}$, we must instead set $\rho_t \approx \sqrt{d/m}$. We then apply ClipAndNoise one final time and then terminate the algorithm. In this event, the completion of the terminating layer actually gives a final error of $\sqrt{1/m}$, in which case the algorithm is happy to terminate.

The rest of the section is dedicated to formalizing the intuition above.

Algorithm 3 Threaded Clip-and-Noise

Input: $S = \left\{X_j^{(i)}\right\}_{i\in[n],j\in[m]}$ where each $S_i = \left\{X_j^{(i)}\right\}_{j\in[m]}$ and $X_j^{(i)} \in \mathbb{R}^d$. Parameters ε, δ, u.

Output: $u^* \in \mathbb{R}^d$.

1: **procedure** TCN($S;\varepsilon,\delta,u$)
2:      $u_0 = u$
3:      $\varepsilon' = \frac{\varepsilon}{\sqrt{\log\log d \cdot \log((2\log\log d)/\delta)}}, \delta' = \frac{\delta}{2\log\log d}$
4:      $T = \max\left(0, \max\left(t \in \mathbb{Z} : \frac{d^{\frac{1}{2} + \frac{1}{2k(k+1)^{t-1}}}}{m^{1-1/k}} > \tilde{O}\left(\frac{d^{1/2}}{m^{1/2}}\right)\right)\right) + 1$
5:      **if** $T > 1$ **then**
6:          **for** $t = 1,\dots,T-1$ **do**
7:              Let $S(t) = \left\{S_{\lfloor\frac{n}{T}\rfloor(t-1)+1},\dots,S_{\lfloor\frac{n}{T}\rfloor t}\right\}$
8:              $u_t \leftarrow$ ClipAndNoise($S(t); \varepsilon', \delta', \rho_t, u_{t-1}$), where $\rho_t = \frac{d^{\frac{1}{2} + \frac{1}{2k(k+1)^{t-1}}}}{m^{1-1/k}}$
9:          **end for**
10:      **end if**
11:      $S(T) = \left\{S_{\lfloor\frac{n}{T}\rfloor(T-1)+1},\dots,S_n\right\}$
12:      $u^* \leftarrow$ ClipAndNoise($S(T); \varepsilon', \delta', \rho_T, u_{T-1}$), where $\rho_T = \tilde{O}\left(\frac{d^{1/2}}{m^{1/2}}\right)$
13:      Return u^*
14: **end procedure**

Full analysis of Threaded Clip-and-Noise. We decompose γ_t into *noise*, *sampling error*, and *bias* terms as follows:

$$\begin{aligned}
\gamma_t &= \|u_t - \mu\|_2 \\
&= \left\|W_t + \frac{1}{n}\sum_{i=1}^{n} Z_i(t) - \mathbb{E}[Z_1(t)] + \mathbb{E}[Z_1(t)] - \mu\right\|_2 \\
&\leqslant \|W_t\|_2 + \left\|\frac{1}{n}\sum_{i=1}^{n} Z_i(t) - \mathbb{E}[Z_1(t)]\right\|_2 + \|\mathbb{E}[Z_1(t)] - \mu\|_2 \\
&:= \text{noise}(t) + \text{sampling}(t) + \text{bias}(t).
\end{aligned}$$

We first introduce Lemma 4.8 which bounds noise(t) and sampling(t). Its proof is deferred to Appendix B.4

LEMMA 4.8. (NOISE AND SAMPLING) *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ and $\sigma_k(\mathcal{D}) \leqslant 1$. Suppose that, for all $i \in [n]$ and $j \in [m]$, $X_j^{(i)} \sim \mathcal{D}$. There exists $n_0 = \tilde{O}\left(\frac{d\sqrt{\log(1/\delta)}}{\varepsilon}\right)$ such that, for all $n \geqslant n_0$,*

$$\text{noise}(t) + \text{sampling}(t) = \tilde{O}\left(\frac{\rho_t}{\sqrt{d}}\right)$$

with probability at least $1 - 0.1/\log\log d$,

We can now give guarantees of each of these three terms at any given layer t of Threaded Clip-and-Noise.

LEMMA 4.9. (ERROR OF ITERATION t) *For all t, define $r_t = \frac{d^{\frac{1}{2} + \frac{1}{2k(k+1)^{t-1}}}}{m^{1-1/k}}$. For all $t \in [T]$, let ρ_t and u_t be as defined in Threaded Clip-and-Noise, and let $\gamma_t = \|u_t - \mu\|_2$. If $\gamma_{t-1} = \tilde{O}\left(r_{t-1}/\sqrt{d}\right)$, then there exists $n_0 = \tilde{O}\left(\frac{d\sqrt{\log(1/\delta)}}{\varepsilon}\right)$ such that for all $n \geqslant n_0$,*

$$\gamma_t = \tilde{O}\left(\rho_t/\sqrt{d}\right)$$

with probability at least $1 - 0.1/\log\log d$.

Proof. We've seen that

$$\gamma_t \leqslant \text{noise}(t) + \text{sampling}(t) + \text{bias}(t).$$

Note by Lemma 4.8 that the $\text{noise}(t) + \text{sampling}(t) = \tilde{O}\left(\rho_t/\sqrt{d}\right)$ with probability at least $1 - 0.1/\log\log d$. Thus, it suffices to show that $\text{bias}(t)$ is also $\tilde{O}\left(\rho_t/\sqrt{d}\right)$, which is shown in the following calculation:

$$\text{(from Theorem 4.2)} \qquad \text{bias}(t) = \tilde{O}\left(\frac{d^{k/2}\gamma_{t-1}}{m^{k-1}\rho_t^k} + \frac{d^{k/2}}{m^{k-1}\rho_t^{k-1}} \cdot \frac{1}{d^{1/2}}\right)$$

$$\text{(by assumption on } \gamma_{t-1}) \qquad = \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}\rho_t^k} \cdot \frac{r_{t-1}}{d^{1/2}} + \frac{d^{k/2}}{m^{k-1}\rho_t^{k-1}} \cdot \frac{1}{d^{1/2}}\right)$$

$$\text{(since } \rho_t \geqslant r_t) \qquad = \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}r_t^k} \cdot \frac{r_{t-1}}{d^{1/2}} + \frac{d^{k/2}}{m^{k-1}r_t^{k-1}} \cdot \frac{1}{d^{1/2}}\right)$$

$$\text{(since } r_{t-1} = r_t d^{\frac{1}{2(k+1)^{t-1}}}) \qquad = \tilde{O}\left(\frac{d^{k/2}d^{\frac{1}{2(k+1)^{t-1}}}}{m^{k-1}r_t^k} \cdot \frac{r_t}{d^{1/2}} + \frac{d^{k/2}}{m^{k-1}r_t^k} \cdot \frac{r_t}{d^{1/2}}\right)$$

$$= \tilde{O}\left(\frac{d^{k/2}d^{\frac{1}{2(k+1)^{t-1}}}}{m^{k-1}r_t^k} \cdot \frac{r_t}{d^{1/2}}\right)$$

$$= \tilde{O}\left(\frac{r_t}{\sqrt{d}}\right)$$

$$\text{(since } \rho_t \geqslant r_t) \qquad = \tilde{O}\left(\frac{\rho_t}{\sqrt{d}}\right).$$

$\square$

LEMMA 4.10. *Algorithm 2 satisfies person-level (ε, δ)-DP.*

Proof. For each user, let $S_i = \left\{X_1^{(i)}, \ldots, X_m^{(i)}\right\}$. Define each Z_i as in Algorithm 2. Consider the function $f(S_1, \ldots, S_n) = \frac{1}{n}\sum_{i=1}^n Z_i$. The ℓ_2-sensitivity of this function (with respect to person-level privacy) is $O\left(\frac{\rho}{n}\right)$. The proof then immediately follows by applying Lemma 2.4. $\square$

THEOREM 4.11. *Let $S = \{S_1, \ldots, S_n\}$, where each $S_i = \left\{X_1^{(i)}, \ldots, X_m^{(i)}\right\}$ and each $X_j^{(i)} \in \mathbb{R}^d$. Let $u_0 \in \mathbb{R}^d$. For all $\varepsilon, \delta \in (0, 1)$, $\mathrm{TCN}(S; \varepsilon, \delta, u_0)$ satisfies person-level (ε, δ)-DP. Furthermore, let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ and $\sigma_k(\mathcal{D}) \leqslant 1$. Suppose that $X_j^{(i)} \sim \mathcal{D}$ for all $i \in [n]$ and $j \in [m]$, and suppose that $\|u_0 - \mu\|_2 \leqslant \frac{d^{1/2}}{m^{1/2}}$. Let u^* be the output of $\mathrm{TCN}(S; \varepsilon, \delta, u_0)$. There exists $n_0 = \tilde{O}\left(\frac{d\sqrt{\log(1/\delta)}}{\varepsilon}\right)$, such that, if $n \geqslant n_0$, then $\|u^* - \mu\|_2 = \tilde{O}\left(\frac{1}{m^{1/2}}\right)$ with probability at least 0.9.*

Proof. We first prove the privacy guarantees of Theorem 4.10. Without loss of generality, we can assume that each call to ClipAndNoise receives as input the entire dataset X of size n and chooses to only compute on a fraction of this dataset. By Lemma 4.10, each of the T iterations of ClipAndNoise satisfies person-level (ε', δ')-DP. Thus, by Lemma 2.3 and the fact that $T \leqslant \log \log d$, Threaded Clip-and-Noise satisfies person-level (ε, δ)-DP.

We now prove the accuracy guarantees of u^*. Let $\gamma^* = \|u^* - \mu\|_2$. Our goal will be to show that $\gamma^* = \tilde{O}\left(\rho_T/\sqrt{d}\right)$ as this implies that $\gamma^* = \tilde{O}\left(1/m^{1/2}\right)$. We divide T into two cases: (1) when $T > 1$ and (2) when $T = 1$.

Case 1 ($T > 1$): For all $t \in [T]$, let E_t be the event that $\gamma_t = \tilde{O}\left(\rho_t/\sqrt{d}\right)$, and let $\bar{E}_t$ be the negation of this event. We have that
$$\mathbb{P}[E_T] \geqslant \mathbb{P}[E_0, E_1, \ldots, E_T] = 1 - \mathbb{P}\left[\exists t \in \{0, \ldots, T\} : \bar{E}_t\right]$$

We can make the following realization.

(4.22)
$$\mathbb{P}\left[\exists t \in \{0, \ldots, T\} : \bar{E}_t\right] = \sum_{t=1}^{T} \mathbb{P}\left[\bar{E}_t \mid E_{t-1}, \ldots, E_0\right] + \mathbb{P}[\bar{E}_0].$$

We briefly show that, when $T > 1$, $\mathbb{P}[\bar{E}_0] = 0$ (i.e. $\gamma_0 \leqslant r_0/\sqrt{d}$). By assumption, $\gamma_0 \leqslant \frac{d^{1/2}}{m^{1/2}}$. When $T > 1$, $\tilde{O}\left(\frac{d^{1/2}}{m^{1/2}}\right) \leqslant r_1 = r_0/\sqrt{d}$, and thus, $\gamma_0 \leqslant r_0/\sqrt{d}$.

Plugging in $\mathbb{P}[\bar{E}_0] = 0$, we finish bounding (4.22):

$$\mathbb{P}[E_T] \geqslant 1 - \sum_{t=1}^{T} \mathbb{P}\left[\bar{E}_t \mid E_{t-1}, \ldots, E_0\right]$$

$$= 1 - \sum_{t=1}^{T} \mathbb{P}\left[\bar{E}_t \mid E_{t-1}\right]$$

(since $\rho_t = r_t$ for all $t \leqslant T$)
$$= 1 - \sum_{t=1}^{T} \mathbb{P}\left[\bar{E}_t \mid \gamma_{t-1} = \tilde{O}\left(r_{t-1}/\sqrt{d}\right)\right]$$

(by Lemma 4.9)
$$\geqslant 1 - T \cdot \frac{0.01}{\log \log d}$$

$$\geqslant 0.09,$$

where the second-to-last line comes from the fact that

$$T = \max\left(0, \max\left(t \in \mathbb{Z} : \frac{d^{\frac{1}{2} + \frac{1}{2k(k+1)^{t-1}}}}{m^{1-1/k}} > O\left(\frac{d^{1/2} \log^{1/2} m}{m^{1/2}}\right)\right)\right) + 1 \leqslant \log \log d.$$

Case 2 ($T = 1$): When $T = 1$,

$$\gamma_1 \leqslant \mathrm{noise}(1) + \mathrm{sampling}(1) + \mathrm{bias}(1)$$

(by Lemma 4.8)
$$= \tilde{O}\left(\rho_1/\sqrt{d}\right) + \mathrm{bias}(1)$$

2848

where the last line holds with probability at least 0.9. Thus it remains to show that bias$(1) = \tilde{O}\left(\rho_1/\sqrt{d}\right)$, which we show in the following calculations:

(from Theorem 4.2)

$$\text{bias}(1) = \tilde{O}\left(\frac{d^{k/2}\gamma_0}{m^{k-1}\rho_1^k} + \frac{d^{k/2}}{m^{k-1}\rho_1^{k-1}} \cdot \frac{1}{d^{1/2}}\right)$$

(by assumption on γ_0)

$$= \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}\rho_1^k} \cdot \rho_1 + \frac{d^{k/2}}{m^{k-1}\rho_1^{k-1}} \cdot \frac{1}{d^{1/2}}\right)$$

$$= \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}\rho_1^k} \cdot \rho_1\right)$$

(since $\rho_1 \geqslant r_1$)

$$= \tilde{O}\left(\frac{d^{k/2}}{m^{k-1}r_1^k} \cdot \rho_1\right)$$

$$= \tilde{O}\left(\frac{\rho_1}{\sqrt{d}}\right).$$

☐

4.3 Full Algorithm We are now ready to introduce Algorithm 4. The algorithm first runs the high-dimensional private histogram subroutine CoarseEstimate, which is the natural extension of the one-dimensional coarse estimate. (The details are deferred to the Appendix B.5.) After running CoarseEstimate, the algorithm then runs Threaded Clip-and-Noise and finishes by running a final round of ClipAndNoise. In the remainder of this section,

---

Algorithm 4 Person-Level Approximate-DP High-Dimensional Mean Estimator

Input: $S = \left\{X_j^{(i)}\right\}_{i \in [3n], j \in [m]}$ where each $X_j^{(i)} \in \mathbb{R}^d$. Parameters $\varepsilon, \delta, k > 0$.

Output: $\hat{\mu} \in \mathbb{R}^d$.

1: **procedure** MEANESTIMATION$(X; \varepsilon, \delta, k)$
2:      Let $\rho^* = \max\left(\sqrt{\frac{d}{m}}, \frac{n^{1/k}\varepsilon^{1/k}d^{\frac{1}{2}-\frac{1}{k}}}{(\log 1/\delta)^{1/(2k)}m^{1-1/k}}\right)$.
3:      Let $Y := \left(Y^{(1)}, \ldots, Y^{(n)}\right) = \left(X^{(1)}, \ldots, X^{(n)}\right)$
4:      Let $P := \left(P^{(1)}, \ldots, P^{(n)}\right) = \left(X^{(n+1)}, \ldots, X^{(2n)}\right)$
5:      Let $V := \left(V^{(1)}, \ldots, V^{(n)}\right) = \left(X^{(2n+1)}, \ldots, X^{(3n)}\right)$
6:      $u_0 \leftarrow$ CoarseEstimate$(Y; \varepsilon/2, \delta/2, 16\sqrt{d/m})$.
7:      Let $u^* \leftarrow$ TCN$(P; \varepsilon/4, \delta/4)$.
8:      Let $\hat{\mu} \leftarrow$ ClipAndNoise$(V; \varepsilon/4, \delta/4, \rho^*, u^*)$.
9:      Return $\hat{\mu}$.
10: **end procedure**

---

we prove the following theorem:

THEOREM 4.12. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ, and let $\sigma_k(\mathcal{D}) \leqslant 1$ for some $k > 2$. For all $\varepsilon, \delta > 0$, Algorithm 4 satisfies person-level (ε, δ)-DP. Furthermore, Algorithm 4 is efficient, and for all $\varepsilon, \delta, \alpha \in (0, 1)$, there exists*

(4.2)
$$n_0 = \tilde{O}_k\left(\frac{d}{\alpha^2 m} + \frac{d \log^{1/2}(1/\delta)}{\alpha m^{1/2}\varepsilon} + \frac{d \log^{1/2}(1/\delta)}{\alpha^{k/(k-1)} m \varepsilon} + \frac{d \log^{1/2}(1/\delta)}{\varepsilon} + \frac{d^{1/2}\log^{3/2}(1/\delta)}{\varepsilon}\right)$$

such that, if $n \geqslant n_0$, then, with probability at least 2/3, Algorithm 4 outputs $\hat{\mu} \in \mathbb{R}^d$ such that

$$\|\hat{\mu} - \mu\|_2 \leqslant \alpha.$$

We first state the guarantees of CoarseEstimate.

THEOREM 4.13. (APPROXIMATE DP COARSE ESTIMATION IN HIGH DIMENSIONS) *Let n be the number of people, m be the number of samples per person, and $\varepsilon > 0, \delta \in (0,1)$. Assume P is a distribution over $\mathbb{R}^d$ with k-th moment bounded by 1. Then there exists an efficient person-level (ε, δ)-DP algorithm CoarseEstimate$(X; \varepsilon, \delta, r)$ that takes i.i.d. samples $\{X_j^{(i)}\}_{i,j}$ from P, and outputs an estimate of the mean up to accuracy $16^{1/k}\sqrt{d/m} < r$, with success probability $1 - \beta$, as long as $n \geqslant n_0$, for some*

$$n_0 = \tilde{O}\left(\frac{\log(1/\beta)}{\log(r\sqrt{m/d})} + \min\left\{\frac{d\log(1/(\delta\beta))}{\varepsilon}, \frac{\sqrt{d\log(1/\delta)}\log(1/\delta\beta)}{\varepsilon}\right\}\right),$$

where $\tilde{O}$ hides lower order logarithmic factors in d.

Finally, we prove the following lemma about the accuracy of the final application of ClipAndNoise.

LEMMA 4.11. *Let $d, m, n \geqslant 2$, let $\varepsilon, \delta \in (0,1)$, let $k > 2$, and let ρ^*, u^*, and $\hat{\mu}$ be as defined in Algorithm 4. Suppose $\mathcal{D}$ is a distribution over $\mathbb{R}^d$ with mean μ and $\sigma_k(\mathcal{D}) \leqslant 1$. If $V_1^{(i)}, \ldots, V_m^{(i)} \stackrel{iid}{\sim} \mathcal{D}$ for all $i \in [n]$, and if $\|u^* - \mu\|_2 \leqslant \frac{\rho^*}{2\sqrt{d}}$, then there exists*

$$n_0 = O\left(\frac{d}{m\alpha^2} + \frac{d\sqrt{\log 1/\delta}}{m\varepsilon\alpha^{\frac{k}{k-1}}} + \frac{d\sqrt{\log(1/\delta)}}{\sqrt{m}\alpha\varepsilon}\right)$$

such that if $n \geqslant n_0$, then with probability at least 0.9,

$$\|\hat{\mu} - \mu\|_2 \leqslant \alpha.$$

Proof. Recall that $\hat{\mu}$ is the output of an application of Algorithm 2. Define $V_i := \text{clip}_{\rho^*, u^*}\left(\frac{1}{m}\sum_{j=1}^m V_j^{(i)}\right)$ and define W such that

$$W \sim \mathcal{N}\left(0, \left(\frac{8\rho^*\sqrt{2\ln(16/\delta)}}{n\varepsilon}\right)^2 \cdot \mathbb{I}_{d\times d}\right).$$

We begin by expanding $\|\hat{\mu} - \mu\|_2$:

$$\|\hat{\mu} - \mu\|_2 = \left\|W + \frac{1}{n}\sum_{i=1}^n V_i - \mathbb{E}[V_1] + \mathbb{E}[V_1] - \mu\right\|_2$$

$$\leqslant \underbrace{\|W\|_2}_{T_1:=} + \underbrace{\left\|\frac{1}{n}\sum_{i=1}^n V_i - \mathbb{E}[V_1]\right\|_2}_{T_2:=} + \underbrace{\|\mathbb{E}[V_1] - \mu\|_2}_{T_3:=}.$$

By assumption on n, we can apply Lemma B.9 which says that, with probability at least 0.95,

$$T_2 = \left\|\frac{1}{n}\sum_{i=1}^n V_i - \mathbb{E}[V_1]\right\|_2 \leqslant \frac{\alpha}{2}.$$

By the properties of Gaussian random variables, it follows that with probability at least 0.95,

$$T_1 = \|W\|_2 = O\left(\frac{\rho^*\sqrt{d\log(1/\delta)}}{n\varepsilon}\right).$$

When $\rho^* = \sqrt{d/m}$,

(4.23) $$T_1 = O\left(\frac{d\sqrt{\log(1/\delta)}}{\sqrt{m}n\varepsilon}\right) \leqslant \frac{\alpha}{4},$$

by assumption on n and n_0. Likewise, when $\rho^* = \frac{n^{1/k} \varepsilon^{1/k} d^{\frac{1}{2} - \frac{1}{k}}}{(\log 1/\delta)^{1/(2k)} m^{1-1/k}}$,

$$(4.24) \qquad T_1 = O\left(\frac{d\sqrt{\log(1/\delta)}}{mn\varepsilon} \right)^{\frac{k-1}{k}} \leqslant \frac{\alpha}{4},$$

by assumption on n and n_0. By Theorem 4.2 and assumption on u^*,

$$\|\mathbb{E}[\mathsf{V}_1] - \mu\|_2 = \tilde{O}\left(\frac{\sqrt{d}}{m\rho^*} \right)^{k-1}.$$

When $\rho^* = \sqrt{d/m}$, T_3 is dominated by T_1. Otherwise, when $\rho^* = \frac{n^{1/k} \varepsilon^{1/k} d^{\frac{1}{2} - \frac{1}{k}}}{(\log 1/\delta)^{1/(2k)} m^{1-1/k}}$, T_3 and T_1 have the same asymptotics. Thus,

$$(4.25) \qquad \|\hat{\mu} - \mu\|_2 \leqslant T_1 + T_2 + T_3 \leqslant 2T_1 + T_2 \leqslant \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha,$$

which completes the proof. □

We can now prove Theorem 4.3

Proof. We break the proof into two pieces. First, we prove that Algorithm 4 satisfies person-level (ε, δ)-DP. Afterwords, we analyze the accuracy of Algorithm 4.

Analysis of privacy. Without loss of generality, we can assume that each of Line 6, Line 7, and Line 8 each receive as input the entire dataset X and choose to only compute on a third of the dataset. By Theorem 4.13, CoarseEstimate satisfies person-level $\left(\frac{\varepsilon}{2}, \frac{\delta}{2}\right)$-DP. By Theorem 4.10, Algorithm 3, is $\left(\frac{\varepsilon}{4}, \frac{\delta}{4}\right)$-DP. By Lemma 4.10, Line 8 is person-level $\left(\frac{\varepsilon}{4}, \frac{\delta}{4}\right)$-DP. Thus, by basic composition (Lemma 2.2), Algorithm 4 is person-level (ε, δ)-DP.

Analysis of accuracy. By Theorem 4.13, the estimate u_0 satisfies $\|u_0 - \mu\|_2 \leqslant \frac{\sqrt{d}}{2\sqrt{m}}$ with probability at least 0.9. Conditioned on this event, by Theorem 4.10, $\|u^* - \mu\|_2 \leqslant \frac{\rho^*}{2\sqrt{d}}$ with probability at least 0.9. Conditioned on this event, by Lemma 4.11, $\|\hat{\mu} - \mu\|_2 \leqslant \alpha$ with probability at least 0.9. Thus, by a union bound, $\|\hat{\mu} - \mu\|_2 \leqslant \alpha$ with probability at least 0.7. □

5   Mean Estimation in High Dimensions with Pure-DP

The section is split into two parts. The first focuses on fine estimation, whereas the second integrates coarse estimation to the overall process and presents the full algorithm.

5.1   Pure-DP Fine Estimation The high-level overview of our fine estimation algorithm is the following. We have a distribution $\mathcal{D}$ over $\mathbb{R}^d$ such that $\mu := \mathbb{E}_{X \sim \mathcal{D}}[X]$ and $\mathbb{E}_{X \sim \mathcal{D}}\left[|\langle X - \mu, v \rangle|^k \right] \leqslant 1, \forall \|v\|_2 = 1$. Since we're in the fine estimation setting, we assume that we have $\|\mu\|_\infty \leqslant \alpha$, where α is our target error. The idea is to cover the space of candidate means in a way that, for every candidate, there exists a point in the cover that is at ℓ_2-distance at most α from it. Then, around each point, we construct a "local cover". The goal of the local cover is to help us examine whether our point is a good candidate for an estimate of the true mean by performing binary comparisons. If the "central point" loses a comparison with an element of the local cover, this outcome is treated as a certificate that the true mean is far from that point. Conversely, if the central point wins all comparisons, this implies that the point is indeed close to the true mean.

Implementing this privately involves instantiating the exponential mechanism in a way that is tailored to the above setting. Given that each person is contributing a batch consisting of multiple samples, the score of a point is the minimum number of batches that need to be altered so that the candidate point loses at least one comparison with an element of its corresponding local cover. In the rest of this section, the analysis will first focus on establishing the necessary facts which describe how comparisons behave, and then the focus will shift to establishing the utility guarantees of the exponential mechanism.

We start by analyzing how binary comparisons are performed between the central point and the elements of its local cover. Our algorithm first calculates the empirical mean of each batch, and then projects the result on the line that connects the two points. Then, the projections are averaged, and, depending on which point the average is closer to, the outcome of the comparison is determined. Given that this will be used to instantiate the exponential mechanism, we need to ensure that the extent to which the outcome of the comparison is affected if one or multiple batches are altered is limited, thus guaranteeing bounded sensitivity of the score function. For that reason, the mean of each batch is truncated around the point that is the candidate for the true mean that is under consideration at each stage of the algorithm's execution.

Analyzing the effect of the truncation is the first step to obtaining our result. To do so, we will need the a variant of Lemma 5.1 from [KSU20]. The lemma in question quantifies the bias error induced by the truncation, when the truncation center is far from the true mean.

LEMMA 5.1. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean μ, and kth moment bounded by* 1. *Let $x_0, \rho \in \mathbb{R}$, with $\rho := \Theta\left(\sqrt{\frac{(k-1)\log m}{m}} + \frac{1}{m\alpha^{\frac{1}{k-1}}}\right)$. For $X := (X_1, \ldots, X_m) \sim \mathcal{D}^{\otimes m}$, we define Z be the following random variable:*

$$
Z := \begin{cases}
x_0 - \rho, & \text{if } \frac{1}{m}\sum_{i=1}^{m} X_i < x_0 - \rho \\
\frac{1}{m}\sum_{i=1}^{m} X_i, & \text{if } \left|\frac{1}{m}\sum_{i=1}^{m} X_i - x_0\right| \leq \rho \\
x_0 + \rho, & \text{if } \frac{1}{m}\sum_{i=1}^{m} X_i > x_0 + \rho
\end{cases}.
$$

Then, the following hold:

1. *If $x_0 > \mu + \frac{\rho}{2}$, we have:*

$$
\mathbb{E}[Z] \in \begin{cases}
\left[\mu - \frac{\rho}{8}, \mu + \frac{\rho}{8}\right], & \text{if } \frac{\rho}{2} < |x_0 - \mu| \leq \frac{17\rho}{16} \\
\left[x_0 - \rho, x_0 - \frac{15\rho}{16}\right], & \text{if } |x_0 - \mu| > \frac{17\rho}{16}
\end{cases}.
$$

2. *If $x_0 < \mu - \frac{\rho}{2}$, we have:*

$$
\mathbb{E}[Z] \in \begin{cases}
\left[\mu - \frac{\rho}{8}, \mu + \frac{\rho}{8}\right], & \text{if } \frac{\rho}{2} < |x_0 - \mu| \leq \frac{17\rho}{16} \\
\left[x_0 + \frac{15\rho}{16}, x_0 + \rho\right], & \text{if } |x_0 - \mu| > \frac{17\rho}{16}
\end{cases}.
$$

Proof. Without loss of generality, we assume that $x_0 > \mu$, since the other case is symmetric. We define $a := \max\{x_0 - \rho, \mu + \frac{\rho}{16}\}$ and $b := x_0 + \rho$. The previous implies that $\mu < a < b$. To help us bound the value of $\mathbb{E}[Z]$, we also consider the probability $q := \mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| > a - \mu\right]$. For the upper bound, we have:

$$
\mathbb{E}[Z] = \mathbb{E}\left[\mathbb{1}\left\{\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \leq a - \mu\right\} Z\right] + \mathbb{E}\left[\mathbb{1}\left\{\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| > a - \mu\right\} Z\right]
$$

$$
= (1 - q)\,\mathbb{E}\left[Z \,\middle|\, \left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \leq a - \mu\right] + q\,\mathbb{E}\left[Z \,\middle|\, \left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| > a - \mu\right]
$$

$$
\leq (1 - q)a + qb
$$

(5.26)
$$
= a + (b - a)q,
$$

where we used the property of conditional expectation that $\mathbb{E}[X\mathbb{1}\{X \in A\}] = \mathbb{P}[X \in A]\,\mathbb{E}[X|X \in A]$.

Indeed, when the event $\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| > a - \mu \Leftrightarrow \left(\frac{1}{m}\sum_{i=1}^{m} X_i > a\right) \vee \left(\frac{1}{m}\sum_{i=1}^{m} X_i < 2\mu - a\right)$ is realized, the value of Z is maximized when one assumes that $\frac{1}{m}\sum_{i=1}^{m} X_i \geq x_0 + \rho = b$ holds with probability q. Similarly, when

$\left|\frac{1}{m}\sum_{i=1}^{m}X_i - \mu\right| \le a - \mu \Leftrightarrow 2\mu - a \le \frac{1}{m}\sum_{i=1}^{m}X_i \le a$ is realized, the value of Z is maximized when $\frac{1}{m}\sum_{i=1}^{m}X_i = a$ happens with probability $1 - q$. Now, we take cases depending on the value of a. If $a = \mu + \frac{\rho}{16} \Leftrightarrow \mu + \frac{\rho}{16} \ge x_0 - \rho \Leftrightarrow x_0 - \mu \le \frac{17}{16}\rho$, (5.26) becomes:

$$(5.27) \qquad\qquad \mathbb{E}[Z] \le \mu + \frac{\rho}{16} + 2q\rho.$$

Conversely, if $a = x_0 - \rho \Leftrightarrow \mu + \frac{\rho}{16} < x_0 - \rho \Leftrightarrow x_0 - \mu > \frac{17}{16}\rho$, (5.26) yields:

$$(5.28) \qquad\qquad \mathbb{E}[Z] \le x_0 - \rho + 2\rho q.$$

We observe that, in either of the two cases, in order to establish the desired upper bounds, it suffices to establish that $q \le \frac{1}{32}$. Since $\rho = \Theta\left(\sqrt{\frac{(k-1)\log m}{m}} + \frac{1}{m\alpha^{\frac{1}{k-1}}}\right)$, Corollary 3.1 yields:

$$q = \mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}X_i - \mu\right| > \alpha - \mu\right] \le \mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}X_i - \mu\right| > \frac{\rho}{16}\right]$$

$$\le_k \frac{1}{m^{k-1}\left(\frac{\rho}{16}\right)^k}.$$

Due to our choice of ρ, we get that $q \le \frac{1}{32}$, so the desired upper bounds follow from (5.27) and (5.28).

We now turn our attention to establishing the lower bounds. The bound $\mathbb{E}[Z] \ge x_0 - \rho$ holds trivially because of the truncation interval considered. Thus, we focus on the case where $x_0 - \mu < \frac{17}{16}\rho$. Based on a similar reasoning as the one used in the upper bound, we get:

$$\mathbb{E}[Z] \ge (1-q)(2\mu - a) + q(x_0 - \rho) = (1-q)\left(\mu - \frac{\rho}{16}\right) + q(x_0 - \rho) = \mu - \frac{\rho}{16} + q(x_0 - \mu) - \frac{15q\rho}{16}$$

$$\ge \mu - \frac{\rho}{8},$$

where we used that $x_0 - \mu > 0$ and $q \le \frac{1}{32}$. $\square$

The next step in the analysis involves reasoning about the *sampling error*. In particular, we have to consider how much the truncated sample mean of a batch deviates from the true mean of the truncated distribution.

LEMMA 5.2. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean μ, and kth moment bounded by 1. Assume we are given $n \ge O\left(\frac{\log\left(\frac{1}{\beta}\right)}{m\alpha^2}\right)$ independently-drawn batches of size m, i.e., $X^{(i)} := \left(X_1^{(i)}, \dots, X_m^{(i)}\right) \sim \mathcal{D}^{\otimes m}, \forall i \in [n]$. For $x_0, \rho \in \mathbb{R}$, we define the random variables Z_i as in Lemma 5.1.[10] We denote the mean of the above random variables by μ_{tr}. Now, let $k = \Theta\left(\log\left(\frac{1}{\beta}\right)\right)$, so that $\frac{n}{k} \ge \frac{10}{m\alpha^2}$. For $i = 1, \dots, k$, we define $Y^i := \left(Z_{(i-1)\frac{n}{k}+1}, \dots Z_{i\frac{n}{k}}\right)$, and let $\mu_{\mathrm{tr},i}$ be the empirical mean of Y^i. Then, for $\bar{\mu} := \mathrm{Median}(\mu_{\mathrm{tr},1}, \dots, \mu_{\mathrm{tr},k})$, we will have:*

$$\mathbb{P}[|\bar{\mu} - \mu_{\mathrm{tr}}| > \alpha] \le \beta.$$

In particular, at least 90% of $(\mu_{\mathrm{tr},1}, \dots, \mu_{\mathrm{tr},k})$ will be at most α-far from $\bar{\mu}$, except with probability β.

Proof. First, we note that Jensen's inequality implies:

$$\mathbb{E}\left[\left|X_j^i - \mu\right|^2\right] \le \mathbb{E}\left[\left|X_j^i - \mu\right|^k\right]^{\frac{2}{k}} \le 1.$$

---

[10] x_0, ρ do not need to be the same as in that lemma.

Additionally, looking at the second moment of the sample mean of each batch, we get:

$$\mathbb{E}\left[\left\|\frac{1}{m}\sum_{j=1}^{m}X_j^{(i)} - \mu\right\|^2\right] = \frac{1}{m^2}\sum_{j=1}^{m}\mathbb{E}\left[\left|X_j^{(i)} - \mu\right|^2\right] \leqslant \frac{1}{m}.$$

By Lemma 3.2, we get that:

$$\mathbb{E}\left[|Z_i - \mathbb{E}[Z_i]|^2\right] \leqslant \mathbb{E}\left[\left\|\frac{1}{m}\sum_{j=1}^{m}X_j^{(i)} - \mu\right\|^2\right] \leqslant \frac{1}{m}.$$

By Chebyshev's inequality and our assumption that $\frac{n}{k} \geqslant \frac{10}{m\alpha^2}$, we have that:

$$\mathbb{P}[|\mu_{\mathrm{tr},i} - \mu_{\mathrm{tr}}| > \alpha] \leqslant 0.1.$$

Then, by a direct application of the *median trick*, we amplify the probability of success to get the desired high probability guarantee, leading to the bound $n \geqslant O\left(\frac{\log\left(\frac{1}{\beta}\right)}{m\alpha^2}\right)$. $\square$

Having established the above two lemmas which account for the effect of various sources of error, we now proceed to perform a step-by-step analysis of our algorithm. This involves working in a bottom-up fashion, meaning that we will start by focusing on the binary comparison between one individual candidate mean and an element of its local cover. We consider two cases based on whether we're truncating around a point that's close to the true mean or not. For the first case, we have a lemma that establishes that, if we're truncating around a point p that's "close" to the true mean, p will win the comparison against any point q that's far from the true mean with high probability, and there will be a lower bound on the number of batches that would have to be changed in order to alter the result of the comparison.

Algorithm 5 Binary Mean Comparison

Input: A dataset $X := \left(X^{(1)}, \ldots, X^{(n)}\right) \sim \left(\mathcal{D}^{\otimes m}\right)^{\otimes n}$, points $p, q \in \mathbb{R}^d$, target error α, target failure probability β.
Output: True, False.

1: **procedure** BinMeanComp$_{p,q,\alpha,\beta}(X)$

2:     Let $\rho = \Theta\left(\sqrt{\frac{(k-1)\log m}{m}} + \frac{1}{m\alpha^{\frac{1}{k-1}}}\right)$.

3:     **for** $i \in [n]$ **do**

4:         Consider the 1-d projection $\left\langle \frac{q-p}{\|q-p\|_2}, \frac{1}{m}\sum_{j=1}^{m}X_j^{(i)} \right\rangle$, and, for this projection, define Z_i as in Lemma 5.1 with

         truncation center $x_0 := p$ and truncation radius ρ.

5:     **end for**

6:     Let $p_0 := \left\langle \frac{q-p}{\|q-p\|_2}, p \right\rangle$, and $q_0 := \left\langle \frac{q-p}{\|q-p\|_2}, q \right\rangle$.

7:     **if** $\bar{\mu} \leqslant \frac{p_0+q_0}{2}$ **then**   ▷ $\bar{\mu}$ is the median of means, as in Lemma 5.2.

8:         **return** True.

9:     **else**

10:         **return** False.

11:     **end if**

12: **end procedure**

LEMMA 5.3. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ, and kth moment bounded by* 1*. Assume we are given* $n \geqslant O\left(\frac{\log\left(\frac{1}{\beta}\right)}{m\alpha^2}\right)$ *independently-drawn batches of size m, each denoted by* $X^{(i)} := \left(X_1^{(i)}, \ldots, X_m^{(i)}\right) \sim \mathcal{D}^{\otimes m}, \forall i \in [n]$. *Also,*

assume we are given a pair of points $p, q \in \mathbb{R}^d$ such that $\|p - \mu\|_2 \leqslant \frac{\alpha}{8}$ and $\|p - q\|_2 > \alpha$. Then, we have for Algorithm 5, with probability at least $1 - \beta$, the number of batches an adversary needs to change to make p lose the comparison with q is at least $\Theta\left(\frac{n\alpha}{\rho}\right)$.

Proof. Let $\mu_0 := \left\langle \frac{q-p}{\|q-p\|_2}, \mu \right\rangle$ be the projection of μ along the line connecting p and q. We have that $|\mu_0 - p_0| \leqslant \|\mu - p\|_2 \leqslant \frac{\alpha}{8}$. The mean $\mu_{0,\mathrm{tr}}$ of the samples Z_i satisfies $|\mu_{0,\mathrm{tr}} - \mu_0| \leqslant \frac{\alpha}{16}$ (by Corollary 3.3 and our choice of truncation radius), whereas we have $|\bar{\mu} - \mu_{0,\mathrm{tr}}| \leqslant \frac{\alpha}{16}$, except with probability β (by Lemma 5.2 for $\alpha \to \frac{\alpha}{16}$). By triangle inequality, we get overall that $|p_0 - \bar{\mu}| \leqslant \frac{\alpha}{8} + \frac{\alpha}{16} + \frac{\alpha}{16} = \frac{\alpha}{4}$. We note that $|p_0 - q_0| = \|p - q\|_2 > \alpha$. Thus, for q to win the comparison instead of p, we need to shift $\bar{\mu}$ by at least $\frac{\alpha}{4}$. By Lemma 5.2 we know that, for an adversary to achieve this, they need to corrupt at least $0.4k = \Theta\left(\log\left(\frac{1}{\beta}\right)\right)$ subsamples in order to shift their empirical mean by $\frac{\alpha}{4}$. Focusing on an individual subsample, corrupting a single batch can shift the mean by at most $\frac{2\rho k}{n}$. Thus, in each subsample, the adversary needs to corrupt at least $\frac{n\alpha}{8\rho k}$ batches, implying at least $\Theta\left(\frac{n\alpha}{\rho}\right)$ batches corrupted overall. $\square$

Our second lemma considers the case of truncating around a point p that's far from the true mean, whereas we have a point q that's close to the true mean. In this case, there are two possible outcomes, depending on how far the true mean μ is from p. If μ (and q with it) is very far from p, what might happen is that, due to the effect of the truncation, the truncated mean might end up being closer to p than q. If the previous doesn't happen, we show that p loses with high probability. Conversely, if it happens, we argue that there exists a set of points q' such that, if we run Algorithm 5 with input p and q', p will lose the comparison with high probability.

LEMMA 5.4. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ, and kth moment bounded by 1. For $\alpha \leqslant O\left(\frac{1}{m^{\frac{k-1}{k}}}\right)$, we are given $n \geqslant O\left(\frac{\log\left(\frac{1}{\beta}\right)}{m\alpha^2}\right)$ independently-drawn batches of size m, each denoted by $X^{(i)} := \left(X_1^{(i)}, \ldots, X_m^{(i)}\right) \sim \mathcal{D}^{\otimes m}, \forall i \in [n]$. Assume we are given a pair of points $p, q \in \mathbb{R}^d$ such that $\|p - \mu\|_2 > \frac{9\alpha}{8}$ and $\|q - \mu\|_2 \leqslant \frac{\alpha}{8}$. Then, at least one of the following occurs:*

- *p loses the comparison to q with probability at least $1 - \beta$.*

- *Let $\widetilde{\mu} := p + \frac{\rho}{2} \cdot \frac{\mu - p}{\|\mu - p\|_2}$, and let q' be any point in $\mathbb{R}^d$ such that $\|q' - \widetilde{\mu}\|_2 \leqslant \frac{\alpha}{8}$. Then, if we run Algorithm 5 with points p and q' as input, p will lose with probability at least $1 - \beta$.*

Proof. As in the proof of Lemma 5.3, we have $\mu_0 := \left\langle \frac{q-p}{\|q-p\|_2}, \mu \right\rangle$ be the projection of μ along the line connecting p and q. We start by noting that $\|p - q\|_2 > \alpha$ holds because of the triangle inequality. Additionally, we have $|q_0 - \mu_0| \leqslant \|q - \mu\|_2 \leqslant \frac{\alpha}{8}$. We now need to consider two cases, depending on how $|p_0 - \mu_0|$ compares with $\frac{\rho}{2}$.

- $|p_0 - \mu_0| \leqslant \frac{\rho}{2}$ (Figure 1). By Corollary 3.3 and our choice of truncation radius, we get that $|\mu_{0,\mathrm{tr}} - \mu_0| \leqslant \frac{\alpha}{16}$. Additionally, Lemma 5.2 yields that $|\bar{\mu} - \mu_{0,\mathrm{tr}}| \leqslant \frac{\alpha}{16}$ with probability at least $1 - \beta$. Thus, by triangle inequality we get $|\bar{\mu} - q_0| \leqslant \frac{\alpha}{4} < \frac{\alpha}{2} \leqslant \frac{\|p-q\|_2}{2}$. This implies that p will lose to q, yielding the desired result.

- $|p_0 - \mu_0| > \frac{\rho}{2}$. In this case, the triangle inequality implies that $|p_0 - q_0| > \frac{\rho}{2} - \frac{\alpha}{8} \geqslant \frac{3\rho}{8}$, where the last inequality follows from the assumption $\alpha \leqslant O\left(\frac{1}{m^{\frac{k-1}{k}}}\right)$. Now, we have to consider two cases depending on how $|p_0 - q_0|$ compares with $\frac{17\rho}{16}$.

  When $|p_0 - \mu_0| \leqslant \frac{17\rho}{16}$ (Figure 2), Lemma 5.1 implies $|\mu_0 - \mu_{0,\mathrm{tr}}| \leqslant \frac{\rho}{8}$. Additionally, we have by Lemma 5.2 that $|\bar{\mu} - \mu_{0,\mathrm{tr}}| \leqslant \frac{\alpha}{16}$ with probability at least $1 - \beta$. The triangle inequality yields that $|q_0 - \bar{\mu}| \leqslant \frac{\alpha}{8} + \frac{\rho}{8} + \frac{\alpha}{16} = \frac{\rho}{8} + \frac{3\alpha}{16} \leqslant \frac{3\rho}{16} < \frac{\|p-q\|_2}{2}$. This immediately yields that p loses the comparison to q.
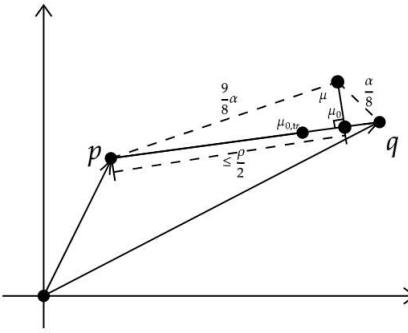
Figure 1: $\|p - \mu\|_2 > \frac{9\alpha}{8}$, $\|q - \mu\|_2 \leq \frac{\alpha}{8}$, and $|p_0 - \mu_0| \leq \frac{\rho}{2}$.
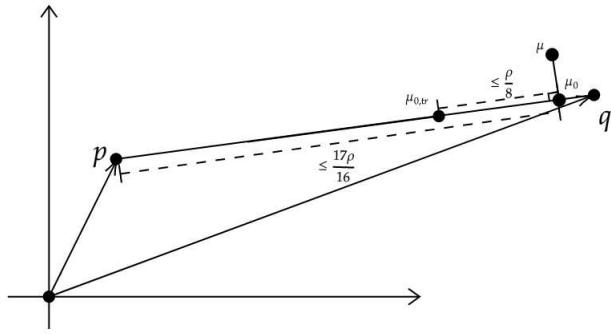
Figure 2: $|p_0 - \mu_0| > \frac{\rho}{2}$, and $|p_0 - \mu_0| \leq \frac{17\rho}{16}$.

When $|p_0 - \mu_0| > \frac{17\rho}{16}$, we get that $|p_0 - \mu_{0,\text{tr}}| \in \left(\frac{15\rho}{16}, \rho \right]$. We note that, by definition, we have $q_0 > p_0$, so $q_0 - p_0 = \|q - p\|_2 \geq \frac{3\rho}{8}$. Combining this with the fact that $|q_0 - \mu_0| \leq \frac{\alpha}{8}$, we get that $\mu_0 - p_0 \geq \frac{3\rho}{8} - \frac{\alpha}{8} > 0 \implies \mu_0 > p_0 \implies \mu_{0,\text{tr}} > p_0 + \frac{15\rho}{16}$. By Lemma 5.2, we have that $|\bar{\mu} - \mu_{0,\text{tr}}| \leq \frac{\alpha}{16}$ with probability at least $1 - \beta$, implying that $\bar{\mu} > p_0 + \frac{15}{16}(\rho - \alpha)$. If $|q_0 - \bar{\mu}| \leq \frac{15}{16}(\rho - \alpha)$, then q wins and p loses. However, if that does not happen, we have to work differently.

For the case where $|q_0 - \bar{\mu}| > \frac{15}{16}(\rho - \alpha)$, let q' be any point that is at ℓ_2-distance at most $\frac{\alpha}{8}$ from $\widetilde{\mu} = p + \frac{\rho}{2} \cdot \frac{\mu - p}{\|\mu - p\|_2}$ (Figure 3).
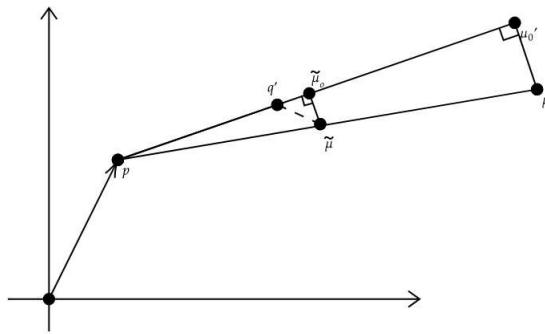
Figure 3: $|q_0 - \bar{\mu}| > \frac{15}{16}(\rho - \alpha)$.

We project everything along the line that connects p and q', so we get $p_0' := \left\langle \frac{q' - p}{\|q' - p\|_2}, p \right\rangle, q_0' :=$

$\left\langle \frac{q'-p}{\|q'-p\|_2}, q' \right\rangle$, $\mu'_0 := \left\langle \frac{q'-p}{\|q'-p\|_2}, \mu \right\rangle$, and $\widetilde{\mu}_0 := \left\langle \frac{q'-p}{\|q'-p\|_2}, \widetilde{\mu} \right\rangle$. By triangle inequality, we have $|q'_0 - p'_0| = \|q' - p\|_2 \geq \|\widetilde{\mu} - p\|_2 - \|q' - \widetilde{\mu}\|_2 \geq \frac{\rho}{2} - \frac{\alpha}{8} \geq \frac{3\rho}{8}$. Additionally, by triangle similarity, we have:

$$\frac{|\widetilde{\mu}_0 - p'_0|}{|\mu'_0 - p'_0|} = \frac{\|p - \widetilde{\mu}\|_2}{\|p - \mu\|_2}$$

$$\Leftrightarrow |\mu'_0 - p'_0| = |\widetilde{\mu}_0 - p'_0| \frac{\|p - \mu\|_2}{\|p - \widetilde{\mu}\|_2} = |\widetilde{\mu}_0 - p'_0| \frac{|p_0 - \mu_0|}{\|p - \widetilde{\mu}\|_2} = |\widetilde{\mu}_0 p - p'_0| \frac{\frac{17}{16}\rho}{\frac{\rho}{2}} = \frac{17}{32}|\widetilde{\mu}_0 - p'_0|,$$

where, by the Pythagorean theorem $|p'_0 - \widetilde{\mu}_0| = \sqrt{\|p - \widetilde{\mu}\|_2^2 - \left\| \widetilde{\mu} - \left(p + \widetilde{\mu}_0 \frac{q'-p}{\|q'-p\|_2} \right) \right\|_2^2} \geq \sqrt{\|p - \widetilde{\mu}\|_2^2 - \|\widetilde{\mu} - q'\|_2^2} \geq \sqrt{\|p - \widetilde{\mu}\|_2^2 - \left(\frac{\alpha}{8}\right)^2} = \sqrt{\left(\frac{\rho}{2}\right)^2 - \left(\frac{\alpha}{8}\right)^2}$. As a result of all the previous, we get that $|\mu'_0 - p'_0|$ is sufficiently large, so we can work similarly to the previous cases, but p is guaranteed to lose this time.

□

Assuming that $\|\mu\|_\infty \leq \alpha$, our goal is to use the above lemma to analyze the comparisons between a point p with $\|p\|_\infty \leq \alpha$ and the elements of a local cover of the ℓ_∞-ball of radius 2α that is centered around p. We will remove from the cover all the points that are at ℓ_2-distance at most α from p. Our goal is to plug the above into the exponential mechanism in order to determine whether a point p is close to the true mean of our distribution. For that reason, we will need to define a *score function*, which we do formally below:

DEFINITION 5.1. (SCORE OF A POINT) *Let* $X := \left(X^{(1)}, \dots, X^{(n)} \right)$ *be a dataset consisting of n batches $X^{(i)} \in \mathbb{R}^{m \times d}$, $\forall i \in [d]$, and p be a point in $\mathbb{R}^d$, and $\alpha > 0$. We define* $\text{Score}_{X,D,\alpha}(p)$ *of p with respect to a domain $D \subset \mathbb{R}^d$ to be the minimum number of batches of X that need to be changed to get a dataset $\overline{X}$ so that there exists $q \in D$, such that Algorithm 5 outputs* False. *If for all $q \in D \setminus \{p\}$ and all X, Algorithm 5 outputs* True, *then we define* $\text{Score}_{X,D,\alpha}(p) = n\alpha$. *If the context is clear, we abbreviate the quantity to* $\text{Score}_X(p)$.

At this point, we recall the following lemma from [KSU20]:

LEMMA 5.5. *The* Score *function satisfies the following:*

$$\Delta_{\text{Score},1} \leq 1.$$

We note that this lemma was established in the setting of *item-level privacy*, where the score is defined in terms of points that have to be changed for a point p to lose a comparison, but the result still holds if instead of individual points we consider batches.

Algorithm 6 implements the above reasoning. The algorithm outputs a value that corresponds to the score of p, as defined in Definition 5.1. Its performance guarantees will be analyzed by using Lemmas 5.3 and 5.4.

LEMMA 5.6. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean $\|\mu\|_\infty \leq \alpha \leq O\left(\frac{1}{m^{\frac{k-1}{k}}} \right)$, and kth moment bounded by 1.*

Assume we are given $n \geq O\left(\frac{d \log\left(\frac{d}{\beta}\right)}{m\alpha^2} \right)$ independently-drawn batches of size m, each denoted by $X^{(i)} := \left(X_1^{(i)}, \dots, X_m^{(i)} \right) \sim \mathcal{D}^{\otimes m}, \forall i \in [n]$. For any $p \in \mathbb{R}^d$ such that $\|p\|_\infty \leq \alpha$, we have:

- *If $\|p - \mu\|_2 \leq \frac{\alpha}{8}$, we have for Algorithm 6 that, with probability at least $1 - \beta$, it will output $\text{Score}_{X, J_p, \alpha}(X) \geq \Theta\left(\frac{n\alpha}{\rho} \right)$.*

- *If $\|p - \mu\|_2 > \frac{9\alpha}{8}$, we have for Algorithm 6 that, with probability at least $1 - \beta$, it will output $\text{Score}_{X, J_p, \alpha}(X) = 0$.*

Algorithm 6 Test Candidate

Input: A dataset $X := \left(X^{(1)}, \ldots, X^{(n)}\right) \sim \left(\mathcal{D}^{\otimes m}\right)^{\otimes n}$, a point $p \in \mathbb{R}^d$, target error α, target failure probability β.
Output: $s \in \mathbb{N}$.

1: **procedure** $\text{TestCan}_{p,\alpha,\beta}(X)$
2:     **for** $i \in [d]$ **do**
3:          Let $J_{p,i} := \left\{ p_i - 2\alpha, p_i - 2\alpha + \frac{\alpha}{4\sqrt{d}}, \ldots, p_i + 2\alpha - \frac{\alpha}{4\sqrt{d}}, p_i + 2\alpha \right\}$.
4:     **end for**
5:      Let $J_p := \left(\bigotimes_{i \in [d]} J_{p,i} \right) \setminus \left\{ x \in \mathbb{R}^d : \|x - p\|_2 \leqslant \alpha \right\}$.
6:     **return** $\text{Score}_{X,J_p,\alpha}(X)$.
7: **end procedure**

Proof. The proof follows almost immediately from Lemmas 5.3 and 5.4.

For the first case, the fact that Algorithm 6 constructs J_p in a way that ensures that points q considered are at least α-far from p suffices for us to be able to apply Lemma 5.3. Given that we want to apply the lemma for all points q, we need to set $\beta \to \frac{\beta}{\left(\Theta\left(\sqrt{d}\right)\right)^d}$. Indeed, by a union bound, the probability of getting the wrong result for at least one comparison is upper bounded by β, leading to the bound $n \geqslant O\left(\frac{d \log\left(\frac{d}{\beta}\right)}{m\alpha^2} \right)$.

For the second case, we note that J_p is an $\frac{\alpha}{8\sqrt{d}}$-cover with respect to the ℓ_∞-distance of the ℓ_∞-ball with radius 2α that's centered at p (modulo the ℓ_2-ball centered around p, which we've removed). We want to argue that there must exist a point $q \in J_p$ such that $\|q - \mu\|_\infty \leqslant \frac{\alpha}{8\sqrt{d}}$. This would follow immediately from the definition of J_p, had the points that lie at the ℓ_2-ball of radius α that's centered at p not been removed. Thus, we need to argue that any point q which satisfies $\|q - \mu\|_\infty \leqslant \frac{\alpha}{8\sqrt{d}}$ must also satisfy $\|p - q\|_2 > \alpha$. We have that $\|q - \mu\|_\infty \leqslant \frac{\alpha}{8\sqrt{d}} \implies \|q - \mu\|_2 \leqslant \frac{\alpha}{8}$. Since $\|p - \mu\|_2 > \frac{9\alpha}{8}$, the triangle inequality yields that $\|p - q\|_2 > \frac{9\alpha}{8} - \frac{\alpha}{8} = \alpha$, leading to the desired result. Additionally, the previous guarantee that the conditions of Lemma 5.4 are satisfied, and setting again $\beta \to \left(\Theta\left(\sqrt{d}\right)\right)^d$ completes the proof. $\square$

We now have all the necessary tools to present the complete algorithm for fine estimation. Algorithm 7 constructs an ℓ_∞-cover of the set $\{x \in \mathbb{R}^d : \|x\|_\infty \leqslant \alpha\}$ that has granularity $\frac{2\alpha}{\sqrt{d}}$, and then calculates the score of all points of the cover and samples from the exponential mechanism.

LEMMA 5.7. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean $\|\mu\|_\infty \leqslant \alpha \leqslant O\left(\frac{1}{m^{\frac{k-1}{k}}}\right)$, and kth moment bounded by 1. Assume we are given $n \geqslant O_k\left(\frac{d \log\left(\frac{d}{\beta}\right)}{m\alpha^2} + \frac{d \log\left(\frac{d}{\beta}\right)\sqrt{\log(m)}}{\sqrt{m}\alpha\varepsilon} + \frac{d \log\left(\frac{d}{\beta}\right)}{m\alpha^{\frac{k}{k-1}}\varepsilon} \right)$ independently-drawn batches of size m, each denoted by $X^{(i)} := \left(X_1^{(i)}, \ldots, X_m^{(i)}\right) \sim \mathcal{D}^{\otimes m}, \forall i \in [n]$. Then, for $\varepsilon, \beta > 0$, there exists an ε-DP mechanism (Algorithm 7) which, given $\left(X^{(1)}, \ldots X^{(n)}\right)$ as input, outputs a point $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_2 \leqslant \alpha$ with probability at least $1 - \beta$.*

Proof. The privacy guarantee is an immediate consequence of the privacy guarantee of the exponential mechanism (Lemma 2.6), so we focus on the accuracy guarantee. By the definition of the set J, we have that there must exist a point $p \in J$ such that $\|p - \mu\|_\infty \leqslant \frac{\alpha}{\sqrt{d}} \implies \|p - \mu\|_2 \leqslant \alpha$. Thus, what we must do is argue that, thanks to our choice of parameters, and the guarantees of the exponential mechanism imply that a point with $\|p - \mu\|_2 \leqslant \alpha$ will be chosen with high probability. First, we note that Algorithm 7 uses Algorithm 6 with target error $\frac{8\alpha}{9}$, and probability $\frac{\beta}{2\left(\Theta\left(\sqrt{d}\right)\right)^d}$. Due to the number of batches n that we have, as well as the guarantees of Lemma 5.6 we

---

Algorithm 7 Fine Estimation

Input: A dataset $X := \left(X^{(1)}, \ldots, X^{(n)}\right) \sim (\mathcal{D}^{\otimes m})^{\otimes n}$, target error α, target failure probability β, privacy parameter ε.

Output: $\widehat{\mu} \in \mathbb{R}^d$.

1: **procedure** $\mathrm{FineEst}_{\alpha, \beta, \varepsilon}(X)$
2:     **for** $i \in [d]$ **do**
3:         Let $J_i := \left\{ -\alpha, -\alpha + \frac{2\alpha}{\sqrt{d}}, \ldots, \alpha - \frac{2\alpha}{\sqrt{d}}, \alpha \right\}$.
4:     **end for**
5:     Let $J := \bigotimes\limits_{i \in [d]} J_i$.
6:     **for** $p \in J$ **do**
7:         Run $\mathrm{TestCan}_{p, \frac{8\alpha}{9}, \frac{\beta}{2\left(\Theta\left(\sqrt{d}\right)\right)^d}}(X)$ to calculate $\mathrm{Score}_{X, J_p, \alpha}(X)$.
8:     **end for**
9:     Run the Exponential Mechanism with score function $\{\mathrm{Score}_{X, J_p, \alpha}(X)\}_{p \in J}$ and privacy budget ε, to get a point $\widehat{\mu}$.
10:     **return** $\widehat{\mu}$.
11: **end procedure**

---

have that, for each point $p \in J$, except with probability $\frac{\beta}{2}$, the score of a point p with $\|p - \mu\|_2 \leqslant \frac{\alpha}{9}$ will be $\Theta\left(\frac{n\alpha}{\rho}\right)$, whereas the score of a point p with $\|p - \mu\|_2 > \alpha$ will be 0. We have by Lemma 5.5 that the sensitivity of our score function is at most 1. Thus, thanks to the guarantees of the exponential mechanism (Lemma 2.6), we have that, except with probability at least $\frac{\beta}{2}$, we have:

$$\mathrm{Score}(X, \widehat{\mu}) \geqslant \mathrm{OPT}_{\mathrm{Score}}(X) - \frac{2\Delta_{\mathrm{Score}, 1}}{\varepsilon}(\ln(|S|) + t)$$

$$\geqslant \Theta\left(\frac{n\alpha}{\rho}\right) - \frac{2}{\varepsilon}\left(\Theta\left(d \log(d)\right) + \log\left(\frac{2}{\beta}\right)\right)$$

$$\geqslant \Theta\left(\frac{n\alpha}{\sqrt{\frac{(k-1)\log m}{m}} + \frac{1}{m\alpha^{\frac{1}{k-1}}}}\right) - \frac{2}{\varepsilon}\left(\Theta\left(d \log(d)\right) + \log\left(\frac{2}{\beta}\right)\right)$$

We now need to consider two cases, depending on which term dominates in the denominator. If $\sqrt{\frac{(k-1)\log m}{m}} \leqslant \frac{1}{m\alpha^{\frac{1}{k-1}}}$, the previous can be lower-bounded by:

$$\Theta\left(n\alpha^{\frac{k}{k-1}} m\right) - \frac{2}{\varepsilon}\left(\Theta\left(d \log(d)\right) + \log\left(\frac{2}{\beta}\right)\right)$$

which, by assumption, is greater than 0.

Conversely, if $\sqrt{\frac{(k-1)\log m}{m}} > \frac{1}{m\alpha^{\frac{1}{k-1}}}$, we get the lower bound:

$$\Theta\left(n\alpha\sqrt{\frac{m}{(k-1)\log(m)}}\right) - \frac{2}{\varepsilon}\left(\Theta\left(d \log(d)\right) + \log\left(\frac{2}{\beta}\right)\right),$$

which, by assumption again, is greater than 0.

Thus, except with probability β, the exponential mechanism will output a point with score greater than 0, implying that the point will be at distance at most α from the true mean. $\quad\square$

5.2 Coarse Estimation and the Full Algorithm After focusing on fine estimation in the previous section, we can reason about coarse estimation, and tie everything together. Our coarse estimator is not a new algorithm: it consists of a component-wise application of the single-dimensional mean estimator that is implied by Theorem 3.4. In particular, given sufficiently many samples from a distribution with kth moment bounded by 1 that has mean μ, the estimator of Theorem 3.4 outputs a $\widehat{\mu}$ such that $|\widehat{\mu} - \mu| \leqslant \alpha$. Now, given samples from a distribution in d dimensions with mean μ and kth moment bounded by 1, we can apply this algorithm independently for each coordinate, and obtain an estimate $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\|_\infty \leqslant \alpha$. This allows us to reduce to the case where $\|\mu\|_\infty \leqslant \alpha$, which was addressed in Section 5.1. Algorithm 8 presents the full pseudocode that handles both coarse and fine estimation.

Algorithm 8 Person-Level Pure-DP High-Dimensional Mean Estimation

---

Input: A dataset $X := \left(X^{(1)}, \ldots, X^{(2n)}\right) \sim (\mathcal{D}^{\otimes m})^{\otimes 2n}$, target error α, target failure probability β, privacy parameter ε.
    Output: $\widehat{\mu} \in \mathbb{R}^d$.
1: **procedure** PersonLevelMeanEst$_{\alpha,\beta,\varepsilon}(X)$
2:     **for** $k \in [d]$ **do**
3:         Let $Y_i := \left(\left(X_1^{(i)}\right)_k, \ldots, \left(X_m^{(i)}\right)_k\right) \in \mathbb{R}^n, \forall i \in [n]$ be the batch consisting of the k-th component of each element of the i-th batch.
4:         Run the algorithm of Theorem 3.4 over $(Y_1, \ldots, Y_n)$ with target error α, target failure probability $\frac{\beta}{2d}$, and privacy parameter $\frac{\varepsilon}{d}$, to obtain $\mu_{\texttt{coarse},i}$.
5:     **end for**
6:     Let $\mu_{\texttt{coarse}} := (\mu_{\texttt{coarse},1}, \ldots, \mu_{\texttt{coarse},d})$.
7:     **for** $i \in \{n+1, \ldots, 2n\}$ **do**
8:         Let Z_{i-n} be the batch that is obtained by subtracting $\mu_{\texttt{coarse}}$ from each element of $X^{(i)}$.
9:     **end for**
10:    Let $Z := (Z_1, \ldots, Z_n)$
11:    Let $\widehat{\mu} := \text{FineEst}_{\alpha, \frac{\beta}{2}, \varepsilon}(Z) + \mu_{\texttt{coarse}}$.
12:    **return** $\widehat{\mu}$.
13: **end procedure**

THEOREM 5.1. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ such that $\|\mu\|_2 \leqslant R$, and kth moment bounded by 1. Assume we are given $n \geqslant O_k\left(\frac{d\log(d)}{m\alpha^2\beta} + \frac{d\log\left(\frac{d}{\beta}\right)\sqrt{\log(m)}}{\sqrt{m}\alpha\varepsilon} + \frac{d\log\left(\frac{d}{\beta}\right)}{m\alpha^{\frac{k}{k-1}}\varepsilon} + \frac{d\log\left(\frac{dRm}{\beta}\right)}{\varepsilon}\right)$ independently-drawn batches of size m, each denoted by $X^{(i)} := \left(X_1^{(i)}, \ldots, X_m^{(i)}\right) \sim \mathcal{D}^{\otimes m}, \forall i \in [n]$. Then, for $\varepsilon, \alpha, \beta > 0$ with $\alpha \leqslant O\left(\frac{1}{m^{\frac{k-1}{k}}}\right)$, there exists an ε-DP mechanism (Algorithm 7) which, given $\left(X^{(1)}, \ldots X^{(n)}\right)$ as input, outputs a point $\widehat{\mu}$ such that $\left\|\widehat{\mu} - \mu\right\|_2 \leqslant \alpha$ with probability at least $1 - \beta$.*

Proof. We start by establishing the privacy guarantee first. By the privacy guarantee of Theorem 3.4, and basic composition (Lemma 2.2), we get that we have ε-DP over the batches $(X^{(1)}, \ldots, X^{(n)})$. The guarantee is not affected when we later construct $\mu_{\texttt{coarse}}$ and subtract it from the other datapoints, due to closure under post-processing (Lemma 2.1). By the privacy guarantee of Lemma 5.7, we are guaranteed ε-DP for the batches $X^{(n+1)}, \ldots X^{(2n)}$. The overall privacy guarantee then follows from parallel composition.

    Now, it remains to establish the accuracy guarantee. By the guarantees of Theorem 3.4, we get that, except with probability $\frac{\beta}{2}$, we will have $\|\mu_{\texttt{coarse}} - \mu\|_\infty \leqslant \alpha$. Thus, by the guarantees of Lemma 5.7, we get that, except with probability $\frac{\beta}{2}$, we will get a $\widehat{\mu}$ that satisfies $\left\|\widehat{\mu} - (\mu - \mu_{\texttt{coarse}})\right\|_2 = \left\|(\widehat{\mu} + \mu_{\texttt{coarse}}) - \mu\right\|_2 \leqslant \alpha$. The desired accuracy guarantee follows directly from the last inequality and a union bound over failure events. $\quad\square$

6   Approximate-DP Lower Bounds for Bounded k-th Moments

In this section, we establish lower bounds for private mean estimation of distributions with bounded kth moments in the person-level setting, under the constraint of (ε, δ)-DP. Our bounds nearly match the upper bounds of Section 4. We start by stating the main theorem of the section.[11]

THEOREM 6.1.   *Let $C_1 > 1$ be a sufficiently large absolute constant. Fix $k \geq 2$, and let $C(k) := 2^k C^{(0)}(k) + 1$, where $C^{(0)}(k)$ is a constant that is sufficiently large so that $\mathbb{E}\left[\left|\mathcal{N}(\mu, \sigma^2)\right|^k\right] \leq C^{(0)}(k)(|\mu| + \sigma)^k$ is satisfied.[12] Suppose $\alpha, \varepsilon \leq 1$,*

and $\delta \leq \min\left\{\widetilde{\Omega}\left(\frac{\alpha\varepsilon}{d}\right)^{C_1}, \Omega\left(\min\left\{\frac{1}{nm}, \frac{\sqrt{d}}{nm\sqrt{\log\left(\frac{nm}{\sqrt{d}}\right)}}\right\}\right)\right\}$. Suppose that M is an (ε, δ)-DP mechanism that takes as input

$X := \left(X^{(1)}, \ldots, X^{(n)}\right)$ with $X^{(i)} := \left(X_1^{(i)}, \ldots, X_m^{(i)}\right) \in \mathbb{R}^{m \times d}, \forall i \in [n]$. Let $\mathcal{D}$ be any distribution over $\mathbb{R}^d$ such that $\mathbb{E}\left[|\langle\mathcal{D} - \mathbb{E}[\mathcal{D}], v\rangle|^k\right] \leq C(k), v \in \mathbb{S}^{d-1}$, and assume that $X_j^{(i)} \sim_{i.i.d.} \mathcal{D}, \forall i \in [n], j \in [m]$. If, for any such $\mathcal{D}$, we have that $\mathbb{P}[\|M(X) - \mathbb{E}[\mathcal{D}]\|_2 \leq \alpha] \geq \frac{2}{3}$, it must hold that $n \geq \widetilde{\Omega}\left(\frac{d}{m\alpha^2} + \frac{d}{\sqrt{m}\alpha\varepsilon} + \frac{d}{m\alpha^{k/k-1}\varepsilon}\right)$.

   The proof of Theorem 6.1 can be split into two parts. The first part involves establishing the first and the last term of the sample complexity, whereas the second part focuses on the middle term. The former is significantly simpler since, as we will see, it comes as a direct consequence of a reduction from the item-level setting. In particular, we recall the following result from [Nar23]:

PROPOSITION 6.1.   *[Theorem 5.1 from [Nar23]]. Let $C_1 > 1$ be a sufficiently large absolute constant. Fix $k \geq 2$, and let $C(k) := 2^k C^{(0)}(k) + 1$, where $C^{(0)}(k)$ is a constant that is sufficiently large so that $\mathbb{E}\left[\left|\mathcal{N}(\mu, \sigma^2)\right|^k\right] \leq C^{(0)}(k)(|\mu| + \sigma)^k$ is satisfied. Suppose $\alpha, \varepsilon \leq 1$, and $\delta \leq \widetilde{\Omega}\left(\frac{\alpha\varepsilon}{d}\right)^{C_1}$. Suppose that M is an (ε, δ)-DP mechanism that takes as input $X := (X_1, \ldots, X_n)$ with $X_i \in \mathbb{R}^d, \forall i \in [n]$. Let $\mathcal{D}$ be any distribution over $\mathbb{R}^d$ such that $\mathbb{E}\left[|\langle\mathcal{D} - \mathbb{E}[\mathcal{D}], v\rangle|^k\right] \leq C(k), \forall v \in \mathbb{S}^{d-1}$, and assume that $X_i \sim_{i.i.d.} \mathcal{D}, \forall i \in [n]$. If, for any such $\mathcal{D}$, we have that $\mathbb{P}[\|M(X) - \mathbb{E}[\mathcal{D}]\|_2 \leq \alpha] \geq \frac{2}{3}$, it must hold that $n \geq \widetilde{\Omega}\left(\frac{d}{\alpha^2} + \frac{d}{\alpha^{\frac{k}{k-1}}\varepsilon}\right)$.*

   To understand why Proposition 6.1 implies a lower bound for the batch setting, we need the following lemma. The lemma describes how, given oracle access to a mechanism for the person-level setting, we can construct a mechanism for the item-level setting. This allows us to reduce instances of the item-level problem to the person-level. Hence, lower bounds on the number of items in the item-level setting imply lower bounds on the number of people in the person-level setting.

LEMMA 6.1.   *Let $\mathcal{D}$ be any distribution over $\mathbb{R}^d$. Let $\varepsilon, \delta, \alpha \geq 0$. Assume that any (ε, δ)-DP mechanism $M: \mathbb{R}^{n' \times d} \to \mathbb{R}^d$ which, given $X := (X_1, \ldots, X_{n'}) \sim \mathcal{D}^{\otimes n'}$, outputs $M(X)$ such that $\mathbb{P}[\|M(X) - \mathbb{E}[\mathcal{D}]\|_2 \leq \alpha] \geq \frac{2}{3}$ requires at least $n' \geq n_0$ samples. Then, for any (ε, δ)-DP mechanism $\bar{M}: \left(\mathbb{R}^{m \times d}\right)^n \to \mathbb{R}^d$ which, given $Y := \left(Y^{(1)}, \ldots, Y^{(n)}\right) \sim (\mathcal{D}^{\otimes m})^{\otimes n}$, outputs $\bar{M}(Y)$ such that $\mathbb{P}[\|M(Y) - \mathbb{E}[\mathcal{D}]\|_2 \leq \alpha] \geq \frac{2}{3}$, it must hold that $n \geq \frac{n_0}{m}$.*

Proof. Let $\bar{M}$ be an (ε, δ)-DP mechanism for the person-level setting that takes n batches of size m as input, and satisfies the desired accuracy guarantee. Given oracle access to this mechanism, we will show how to construct a mechanism M for the item-level setting which shares the same accuracy guarantees. Assume that we have a dataset X of size $n' := nm$ that has been drawn i.i.d. from $\mathcal{D}$. We partition the dataset into n batches of size m. This results in a dataset Y of size n, where each datapoint is an i.i.d. sample from $\mathcal{D}^{\otimes m}$. As a result, we can feed Y into $\bar{M}$, and the resulting mechanism will be M. M inherits the accuracy and privacy guarantees of $\bar{M}$. Consequently, any lower bound on n_0 on n' also implies a lower bound on n, yielding the desired result.   □

---

[11]We note that this differs from Theorem 1.4 in that it omits the $\log(1/\delta)/\varepsilon$ term. This is a known lower bound for the set of all point-mass distributions in the item-level case, and observe that the problem is unchanged in the person-level setting (regardless of the value of m).

[12]It is a standard fact that such a $C^{(0)}(k)$ exists.

As a direct consequence of Proposition 6.1 and Lemma 6.1, we get the following corollary which accounts for two out of three terms that appear in the lower bound of Theorem 6.1.

COROLLARY 6.1. *Let $C_1 > 1$ be a sufficiently large absolute constant. Fix $k \geq 2$, and let $C(k) := 2^k C^{(0)}(k) + 1$, where $C^{(0)}(k)$ is a constant that is sufficiently large so that $\mathbb{E}\left[\left|\mathcal{N}(\mu, \sigma^2)\right|^k\right] \leq C^{(0)}(k)(|\mu| + \sigma)^k$ is satisfied. Suppose $\alpha, \varepsilon \leq 1$, and $\delta \leq \widetilde{\Omega}\left(\frac{\alpha \varepsilon}{d}\right)^{C_1}$. Suppose that M is an (ε, δ)-DP mechanism that takes as input $X := \left(X^{(1)}, \ldots, X^{(n)}\right)$ with $X^{(i)} := \left(X_1^{(i)}, \ldots, X_m^{(i)}\right) \in \mathbb{R}^{m \times d}, \forall i \in [n]$. Let $\mathcal{D}$ be any distribution over $\mathbb{R}^d$ such that $\mathbb{E}\left[\left|\langle \mathcal{D} - \mathbb{E}[\mathcal{D}], \forall v \rangle\right|^k\right] \leq C(k), v \in \mathbb{S}^{d-1}$, and assume that $X_j^{(i)} \sim_{i.i.d.} \mathcal{D}, \forall i \in [n], j \in [m]$. If, for any such $\mathcal{D}$, we have that $\mathbb{P}[\|M(X) - \mathbb{E}[\mathcal{D}]\|_2 \leq \alpha] \geq \frac{2}{3}$, it must hold that $n \geq \widetilde{\Omega}\left(\frac{d}{m\alpha^2} + \frac{d}{m^{\frac{k}{k-1}}\varepsilon}\right)$.*

It remains to argue about the term $\frac{d}{\sqrt{m}\alpha\varepsilon}$ that appears in Theorem 6.1. To do so, we need to invoke a result that is implicit in [LSA+21]. Theorem 6 in that work is concerned with proving a lower bound on the rates of Stochastic Convex Optimization with person-level privacy. Establishing the result involves reducing from Gaussian mean estimation under person-level privacy, for which [LSA+21] proves a lower bound in Appendix E.2. We explicitly state the result here.

PROPOSITION 6.2. *Given $X \sim (\mathcal{N}(\mu, \mathbb{I})^{\otimes m})^{\otimes n}$ with $\mu \in [\pm 1]^d$, for any $\alpha = O(\sqrt{d})$ and any (ε, δ)-DP mechanism $M \colon \left(\mathbb{R}^{m \times d}\right)^n \to [\pm 1]^d$ with $\varepsilon, \delta \in [0, 1]$, and $\delta \leq \Omega\left(\min\left\{\frac{1}{nm}, \frac{\sqrt{d}}{nm\sqrt{\log\left(\frac{nm}{\sqrt{d}}\right)}}\right\}\right)$ that satisfies $\mathbb{E}_{X,M}\left[\|M(X) - \mu\|_2^2\right] \leq \alpha^2$, it holds that $n = \Omega\left(\frac{d}{\sqrt{m}\alpha\varepsilon}\right)$.*

We note that the result in [LSA+21] comes from a reduction-based approach which works by leveraging the lower bound for Gaussian mean estimation in the item-level setting that was shown in [KLSU19]. We provide a direct proof of Proposition 6.2 in Appendix E, which might be of independent interest.

Having said all the above, Theorem 6.1 now follows directly by combining Corollary 6.1 and Proposition 6.2.

References

[AAK21] Ishaq Aden-Ali, Hassan Ashtiani, and Gautam Kamath. On the sample complexity of privately learning unbounded high-dimensional gaussians. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, ALT '21, pages 185–216. JMLR, Inc., 2021.

[AAL21] Ishaq Aden-Ali, Hassan Ashtiani, and Christopher Liaw. Privately learning mixtures of axis-aligned gaussians. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

[AAL23] Jamil Arbas, Hassan Ashtiani, and Christopher Liaw. Polynomial time and private learning of unbounded gaussian mixture models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML '23, pages 1018–1040. JMLR, Inc., 2023.

[AAL24] Mohammad Afzali, Hassan Ashtiani, and Christopher Liaw. Mixtures of gaussians are privately learnable with a polynomial number of samples. In *Proceedings of the 35th International Conference on Algorithmic Learning Theory*, ALT '24, pages 47–73. JMLR, Inc., 2024.

[AKT+23] Daniel Alabi, Pravesh K Kothari, Pranay Tankala, Prayaag Venkat, and Fred Zhang. Privately estimating a Gaussian: Efficient, robust and optimal. In *Proceedings of the 55th Annual ACM Symposium on the Theory of Computing*, STOC '23. ACM, 2023.

[AL22] Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning Gaussians and beyond. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 1075–1076, 2022.

[ALNP23] Martin Aumüller, Christian Janos Lebeda, Boel Nelson, and Rasmus Pagh. Plan: variance-aware private mean estimation. *arXiv preprint arXiv:2306.08745*, 2023.

[ALS23] Jayadev Acharya, Yuhan Liu, and Ziteng Sun. Discrete distribution estimation under user-level local differential privacy. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, AISTATS '23, pages 8561–8585. JMLR, Inc., 2023.

[AMB19] Marco Avella-Medina and Victor-Emmanuel Brunel. Differentially private sub-Gaussian location estimators. *arXiv preprint arXiv:1906.11923*, 2019.

[AUZ23] Hilal Asi, Jonathan Ullman, and Lydia Zakynthinou. From robustness to privacy and back. *arXiv preprint arXiv:2302.01855*, 2023.

[BCS15] Christian Borgs, Jennifer Chayes, and Adam Smith. Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 1369–1377. Curran Associates, Inc., 2015.

[BCSZ18] Christian Borgs, Jennifer Chayes, Adam Smith, and Ilias Zadik. Revealing network structure, confidentially: Improved rates for node-private graphon estimation. In *Proceedings of the 59th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '18, pages 533–543. IEEE Computer Society, 2018.

[BD14] Rina Foygel Barber and John C Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*, 2014.

[BDBC+23] Shai Ben-David, Alex Bie, Clément L. Canonne, Gautam Kamath, and Vikrant Singhal. Private distribution learning with public data: The view from sample compression. *arXiv preprint arXiv:2308.06239*, 2023.

[BDKU20] Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20, pages 14475–14485. Curran Associates, Inc., 2020.

[BGH+23] Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell. Stability is stable: Connections between replicability, privacy, and adaptive generalization. In *Proceedings of the 55th Annual ACM Symposium on the Theory of Computing*, STOC '23, pages 520–527. ACM, 2023.

[BGS+21] Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, and Lydia Zakynthinou. Covariance-aware private mean estimation without private covariance estimation. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

[BHS23] Gavin Brown, Samuel B Hopkins, and Adam Smith. Fast, sample-efficient, affine-invariant private mean and covariance estimation for subgaussian distributions. In *Proceedings of the 36th Annual Conference on Learning Theory*, COLT '23, pages 5578–5579, 2023.

[BKS22] Alex Bie, Gautam Kamath, and Vikrant Singhal. Private estimation with public data. In *Advances in Neural Information Processing Systems 35*, NeurIPS '22. Curran Associates, Inc., 2022.

[BKSW19] Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 156–167. Curran Associates, Inc., 2019.

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Proceedings of the 14th Conference on Theory of Cryptography*, TCC '16-B, pages 635–658, Berlin, Heidelberg, 2016. Springer.

[BS19] Mark Bun and Thomas Steinke. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 181–191. Curran Associates, Inc., 2019.

[BS23] Raef Bassily and Ziteng Sun. User-level private stochastic convex optimization with optimal rates. In *Proceedings of the 40th International Conference on Machine Learning*, ICML '23, pages 1838–1851. JMLR, Inc., 2023.

[BUV14] Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the 46th Annual ACM Symposium on the Theory of Computing*, STOC '14, pages 1–10. ACM, 2014.

[CDd+24] Hongjie Chen, Jingqiu Ding, Tommaso d'Orsi, Yiding Hua, Chih-Hung Liu, and David Steurer. Private graphon estimation via sum-of-squares. In *Proceedings of the 56th Annual ACM Symposium on the Theory of Computing*, STOC '24. ACM, 2024.

[CFMT22] Rachel Cummings, Vitaly Feldman, Audra McMillan, and Kunal Talwar. Mean estimation with user-level privacy under data heterogeneity. In *Advances in Neural Information Processing Systems 35*, NeurIPS '22, pages 29139–29151. Curran Associates, Inc., 2022.

[CKM+20] Clément L. Canonne, Gautam Kamath, Audra McMillan, Jonathan Ullman, and Lydia Zakynthinou. Private identity testing for high-dimensional distributions. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20, pages 10099–10111. Curran Associates, Inc., 2020.

[CS23] Julien Chhor and Flore Sentenac. Robust estimation of discrete distributions under local differential privacy. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory*, ALT '23, pages 411–446. JMLR, Inc., 2023.

[CWZ21] T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

[DFM+20] Wenxin Du, Canyon Foot, Monica Moniot, Andrew Bray, and Adam Groce. Differentially private confidence intervals. *arXiv preprint arXiv:2001.02285*, 2020.

[DHS15] Ilias Diakonikolas, Moritz Hardt, and Ludwig Schmidt. Differentially private learning of structured discrete distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 2566–2574. Curran Associates, Inc., 2015.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.

[DNPR10] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. Differential privacy under continual observation. In *Proceedings of the 42nd Annual ACM Symposium on the Theory of Computing*, STOC '10, pages 715–724. ACM, 2010.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[DR16] Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[DSS+15] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 650–669. IEEE Computer Society, 2015.

[GDD22] Antonious M Girgis, Deepesh Data, and Suhas Diggavi. Distributed user-level private mean estimation. In *Proceedings of the 2022 IEEE International Symposium on Information Theory*, ISIT '22, pages 2196–2201. IEEE Computer Society, 2022.

[GKK+23a] Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Raghu Meka, and Chiyuan Zhang. On user-level private convex optimization. In *Proceedings of the 40th International Conference on Machine Learning*, ICML '23, pages 11283–11299. JMLR, Inc., 2023.

[GKK+23b] Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Raghu Meka, and Chiyuan Zhang. User-level differential privacy with few examples per user. In *Advances in Neural Information Processing Systems 36*, NeurIPS '23, pages 19263–19290. Curran Associates, Inc., 2023.

[GKM21] Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. User-level differentially private learning via correlated sampling. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21, pages 20172–20184. Curran Associates, Inc., 2021.

[GRST24] Anand Jerry George, Lekshmi Ramesh, Aditya Vikram Singh, and Himanshu Tyagi. Continual mean estimation under user-level privacy. *IEEE Journal on Selected Areas in Information Theory*, 5:28–43, 2024.

[HCSS10] TH Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. In *Proceedings of the 37th International Colloquium on Automata, Languages, and Programming*, ICALP '10, pages 405–417, 2010.

[HKM22] Samuel B Hopkins, Gautam Kamath, and Mahbod Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *Proceedings of the 54th Annual ACM Symposium on the Theory of Computing*, STOC '22, pages 1406–1417. ACM, 2022.

[HKMN23] Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. In *Proceedings of the 55th Annual ACM Symposium on the Theory of Computing*, STOC '23. ACM, 2023.

[HLY21] Ziyue Huang, Yuting Liang, and Ke Yi. Instance-optimal mean estimation under differential privacy. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

[HSR+08] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8):1–9, 2008.

[KDH23] Rohith Kuditipudi, John Duchi, and Saminul Haque. A pretty fast algorithm for adaptive private mean estimation. In *Proceedings of the 36th Annual Conference on Learning Theory*, COLT '23, pages 2511–2551, 2023.

[KL17] Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. *arXiv preprint arXiv:1705.05933*, 2017.

[KLSU19] Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pages 1853–1902, 2019.

[KLZ22] Gautam Kamath, Xingtu Liu, and Huanyu Zhang. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. In *Proceedings of the 39th International Conference on Machine Learning*, ICML '22, pages 10633–10660. JMLR, Inc., 2022.

[KMA+21] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[KMR+23] Gautam Kamath, Argyris Mouzakis, Matthew Regehr, Vikrant Singhal, Thomas Steinke, and Jonathan Ullman. A bias-variance-privacy trilemma for statistical estimation. *arXiv preprint arXiv:2301.13334*, 2023.

[KMS22a] Gautam Kamath, Argyris Mouzakis, and Vikrant Singhal. New lower bounds for private estimation and a generalized fingerprinting lemma. In *Advances in Neural Information Processing Systems 35*, NeurIPS '22, pages 24405–24418. Curran Associates, Inc., 2022.

[KMS⁺22b] Gautam Kamath, Argyris Mouzakis, Vikrant Singhal, Thomas Steinke, and Jonathan Ullman. A private and computationally-efficient estimator for unbounded gaussians. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 544–572, 2022.

[KMV22] Pravesh K Kothari, Pasin Manurangsi, and Ameya Velingker. Private robust estimation by stabilizing convex relaxations. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 723–777, 2022.

[KSSU19] Gautam Kamath, Or Sheffet, Vikrant Singhal, and Jonathan Ullman. Differentially private algorithms for learning mixtures of separated Gaussians. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 168–180. Curran Associates, Inc., 2019.

[KSU20] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In *Proceedings of the 33rd Annual Conference on Learning Theory*, COLT '20, pages 2204–2235, 2020.

[KU20] Gautam Kamath and Jonathan Ullman. A primer on private statistics. *arXiv preprint arXiv:2005.00010*, 2020.

[KV18] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pages 44:1–44:9, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[LA24] Daogao Liu and Hilal Asi. User-level differentially private stochastic convex optimization: Efficient algorithms with optimal rates. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, AISTATS '24, pages 4240–4248. JMLR, Inc., 2024.

[LKKO21] Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

[LKO22] Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 1167–1246, 2022.

[LM00] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

[LSA⁺21] Daniel Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri, and Ananda Theertha Suresh. Learning with user-level privacy. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

[LSY⁺20] Yuhan Liu, Ananda Theertha Suresh, Felix Yu, Sanjiv Kumar, and Michael Riley. Learning discrete distributions: User vs item-level privacy. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20. Curran Associates, Inc., 2020.

[Mic76] R. Michel. Nonuniform Central Limit Bounds with Applications to Probabilities of Deviations. *The Annals of Probability*, 4(1):102 – 106, 1976.

[MMR⁺17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, AISTATS '17, pages 1273–1282. JMLR, Inc., 2017.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 94–103. IEEE Computer Society, 2007.

[Nar23] Shyam Narayanan. Better and simpler lower bounds for differentially private statistical estimation. *arXiv preprint arXiv:2310.06289*, 2023.

[NME22] Shyam Narayanan, Vahab Mirrokni, and Hossein Esfandiari. Tight and robust private mean estimation with few users. In *Proceedings of the 39th International Conference on Machine Learning*, ICML '22, pages 16383–16412. JMLR, Inc., 2022.

[PH24] Victor S Portella and Nick Harvey. Lower bounds for private estimation of gaussian covariance matrices under all reasonable parameter regimes. *arXiv preprint arXiv:2404.17714*, 2024.

[RW05] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 11 2005.

[SU15] Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Proceedings of the 28th Annual Conference on Learning Theory*, COLT '15, pages 1588–1628, 2015.

[SU19] Adam Sealfon and Jonathan Ullman. Efficiently estimating Erdos-Renyi graphs with node differential privacy. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 3765–3775. Curran Associates, Inc., 2019.

[TVGZ20] Christos Tzamos, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Ilias Zadik. Optimal private median estimation under minimal distributional assumptions. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20, pages 3301–3311. Curran Associates, Inc., 2020.

[Wu16] Yihong Wu. Lecture notes in information-theoretic methods in high-dimensional statistics, lecture 14: Packing, covering, and consequences on minimax risk, Spring 2016.

[WXDX20] Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *Proceedings of the 37th International Conference on Machine Learning*, ICML '20, pages 10081–10091. JMLR, Inc., 2020.

[ZJS22] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics. *The Annals of Statistics*,

50(4):2256 – 2283, 2022.

[ZKKW20] Huanyu Zhang, Gautam Kamath, Janardhan Kulkarni, and Zhiwei Steven Wu. Privately learning Markov random fields. In *Proceedings of the 37th International Conference on Machine Learning*, ICML '20, pages 11129–11140. JMLR, Inc., 2020.

[ZLS$^+$24] Puning Zhao, Lifeng Lai, Li Shen, Qingming Li, Jiafei Wu, and Zhe Liu. A Huber loss minimization approach to mean estimation under user-level differential privacy. *arXiv preprint arXiv:2405.13453*, 2024.

A   Missing Proofs from Section 3

LEMMA 3.2. *Suppose X is a random variable in $\mathbb{R}$ and $Z = \mathrm{Trunc}_{\ell,r}(X)$, where $\mathrm{Trunc}_{\ell,r}$ is the truncation operation. Then*

$$\mathbb{E}\left[|Z - \mathbb{E}[Z]|^2\right] \leqslant \mathbb{E}\left[|X - \mathbb{E}[X]|^2\right].$$

Proof. First note that for any random variable Y we have that:

$$\mathbb{E}\left[|Y - \mathbb{E}[Y]|^2\right] = \frac{1}{2}\,\mathbb{E}\left[|Y - Y'|^2\right],$$

where Y' is an independent copy of Y. Now,

$$\mathbb{E}\left[|Z - \mathbb{E}[Z]|^2\right] = \frac{1}{2}\,\mathbb{E}\left[|Z - Z'|^2\right] = \frac{1}{2}\,\mathbb{E}\left[|\mathrm{Trunc}_{\ell,r}(X) - \mathrm{Trunc}_{\ell,r}(X')|^2\right] \leqslant \frac{1}{2}\,\mathbb{E}\left[|X - X'|^2\right]$$
$$= \mathbb{E}\left[|X - \mathbb{E}[X]|^2\right],$$

where the last inequality relies on the fact that the distance between X and X' cannot increase post-truncation.
□

B   Missing Proofs from Section 4

B.1   Missing Proofs from Section 4.1.1

LEMMA B.1. *Let m, j be integers satisfying $1 \leqslant j \leqslant m$. It follows that*

$$\binom{m}{j} \leqslant \left(\frac{em}{j}\right)^j.$$

LEMMA B.2. *Let $a > 0$ and $b \geqslant 0$ be constants. The function $(ax^b)^x$ defined on $x > 0$ is convex.*

Proof. Note that

$$\frac{d^2}{dx^2}(ax^b)^x = (ax^b)^x\left(\left(\ln(ax^b) + b\right)^2 + \frac{b}{x}\right) \geqslant 0$$

when $a, x > 0$, $b \geqslant 0$, completing the proof.   □

For the next lemma, recall that the function $f(\ell)$ is defined as

$$f(\ell) = \Theta\left(\frac{d^{k/2} \cdot 2^{\ell(k-1)} \log^k m}{m^{k-1}t^k}\right)^{2^{\ell-1}}.$$

LEMMA 4.3. *Fix any $m, d, t > 0$ and $k \geqslant 1$. Then the function $f(\ell)$ is convex with respect to ℓ.*

Proof. Set $x = 2^\ell$, $a = \frac{m^{1/2}d^{k/4}\log^{k/2}m}{m^{k/2}t^{k/2}}$, and $b = \frac{k-1}{2}$. Observe that $a, b, x > 0$. Thus,

$$f(\ell) = \left(\frac{md^{k/2}2^{\ell(k-1)}\log^k m}{m^k t^k}\right)^{2^{\ell-1}}$$
$$= \left(\frac{m^{1/2}d^{k/4}\log^{k/2}m}{m^{k/2}t^{k/2}} \cdot 2^{\ell\left(\frac{k-1}{2}\right)}\right)^{2^\ell}$$
$$= (ax^b)^x,$$

which we know by Lemma B.2 is convex with respect to $x > 0$, and thus is convex with respect to ℓ.   □

B.2 Missing Proofs from Section 4.1.2

LEMMA B.3. *Let $k > 2$ and $t > 0$. Let X_i be a random vector in $\mathbb{R}^d$ with $\mathbb{E}[X_i] = 0$ and k-th moment bounded by 1. Let $r_1 = \frac{d}{t}$. Let $Y_i = X_i \cdot \mathbb{I}[\|X_i\|_2 < r_1]$. There exists $t_2 = \Omega(\sqrt{d})$ such that, for all positive $t \leqslant t_2$,*

$$\|\mathbb{E}[Y_i]\|_2 \leqslant \min\left(\frac{t}{6}, \frac{d}{t}\right).$$

Proof. Since $\mathbb{E}[X_i] = 0$, we have that

$$\mathbb{E}[X_i] = \mathbb{E}[X_i \cdot \mathbb{I}[\|X_i\|_2 < r_1]] + \mathbb{E}[X_i \cdot \mathbb{I}[\|X_i\|_2 \geqslant r_1]]$$

which implies that

$$
\begin{aligned}
\|\mathbb{E}[Y_i]\|_2 &= \|\mathbb{E}[X_i \cdot \mathbb{I}[\|X_i\|_2 < r_1]]\|_2 \\
&= \|\mathbb{E}[X_i] - \mathbb{E}[X_i \cdot \mathbb{I}[\|X_i\|_2 \geqslant r_1]]\|_2 \\
&= \|\mathbb{E}[X_i \cdot \mathbb{I}[\|X_i\|_2 \geqslant r_1]]\|_2
\end{aligned}
$$

So now,

$$
\begin{aligned}
\|\mathbb{E}&[X_i \cdot \mathbb{I}[\|X_i\|_2 \geqslant r_1]]\|_2 \\
&\leqslant \mathbb{E}[\|X_i\|_2 \cdot \mathbb{I}[\|X_i\|_2 \geqslant r_1]] \\
&= \int_{x=0}^{\infty} \mathbb{P}[\|X_i\|_2 \cdot \mathbb{I}[\|X_i\|_2 \geqslant r_1] > x]\,\mathrm{d}x \\
&= \int_{x=0}^{r_1} \mathbb{P}[\|X_i\|_2 \cdot \mathbb{I}[\|X_i\|_2 \geqslant r_1] > x]\,\mathrm{d}x \\
&\qquad + \int_{x=r_1}^{\infty} \mathbb{P}[\|X_i\|_2 \cdot \mathbb{I}[\|X_i\|_2 \geqslant r_1] > x]\,\mathrm{d}x \\
&= \int_{x=0}^{r_1} \mathbb{P}[\|X_i\|_2 \geqslant r_1]\,\mathrm{d}x + \int_{x=r_1}^{\infty} \mathbb{P}[\|X_i\|_2 > x]\,\mathrm{d}x
\end{aligned}
$$

$$
\begin{aligned}
\text{(by Lemma 4.1)} \qquad &= \int_{x=0}^{r_1} \left(\frac{d^{1/2}}{r_1}\right)^k \mathrm{d}x + \int_{x=r_1}^{\infty} \left(\frac{d^{1/2}}{x}\right)^k \mathrm{d}x \\
&= \frac{d^{k/2}}{r_1^{k-1}} + \frac{d^{k/2}}{r_1^{k-1}(k-1)} \\
\text{(since } k \geqslant 2\text{)} \qquad &\leqslant \frac{2d^{k/2}}{r_1^{k-1}} \\
&= \frac{2t^{k-1}}{d^{k/2-1}}.
\end{aligned}
$$

To finish the proof, note that

$$\frac{2t^{k-1}}{d^{k/2-1}} \leqslant 2t\left(\frac{t}{\sqrt{d}}\right)^{k-2} \leqslant \frac{t}{6},$$

which uses the assumption that $t \leqslant t_2$ for some $t_2 \leqslant 12^{\frac{-1}{k-2}}\sqrt{d}$. Also note that

$$\frac{2t^{k-1}}{d^{k/2-1}} \leqslant \frac{2t^k}{d^{k/2}} \cdot \frac{d}{t} \leqslant \frac{d}{t}$$

which uses the assumption that $t \leqslant t_2$ for some $t_2 \leqslant 2^{-1/k}\sqrt{d}$. $\qquad\square$

LEMMA B.4. *Let $k > 2$ and $t > 0$. Let X_i be a random vector in $\mathbb{R}^d$ with $\mathbb{E}[X_i] = 0$ and k-th moment bounded by 1. Let $r_1 = \frac{d}{t}$. Let $Y_i = X_i \cdot \mathbb{I}[\|X_i\|_2 < r_1]$. Let $Z_i = Y_i - \mathbb{E}[Y_i]$. There exists $t_2 = \Omega(\sqrt{d})$ such that, for all positive $t \leqslant t_2$,*

$$(\text{B.1}) \qquad\qquad \mathbb{E}[Z_i] = 0,$$

$$(\text{B.2}) \qquad\qquad \mathbb{E}\big[\|Z_i\|_2^2\big] \leqslant d, \text{ and}$$

$$(\text{B.3}) \qquad\qquad \|Z_i\|_2 \leqslant 2r_1.$$

Proof. Note that (B.1) is satisfied by design of the Z_i's. We now show that (B.2) holds. Note that

$$\mathbb{E}\big[\|Z_i\|_2^2\big] = \mathbb{E}\big[\|Y_i - \mathbb{E}[Y_i]\|_2^2\big] \leqslant \mathbb{E}\big[\|Y_i\|_2^2\big] \leqslant \mathbb{E}\big[\|X_i\|_2^2\big],$$

where the first inequality above comes from the fact that setting $y = \mathbb{E}[Y_i]$ minimizes $\mathbb{E}\big[\|Y_i - y\|_2^2\big]$. Observe that

$$(\text{B.4}) \qquad \mathbb{E}\big[\|X_i\|_2^2\big] = \mathbb{E}\left[\sum_{j=1}^d \langle X_i, e_j \rangle^2\right] = \sum_{j=1}^d \mathbb{E}\big[\langle X_i, e_j \rangle^2\big] \leqslant \sum_{j=1}^d \mathbb{E}\big[\langle X_i, e_j \rangle^k\big] \leqslant d,$$

where the last inequality comes from the assumption that, for every u with $\|u\|_2 = 1$, $\mathbb{E}\big[\langle X_i, u \rangle^k\big] \leqslant 1$. Thus, $\mathbb{E}\big[\|Z_i\|_2^2\big] \leqslant d$ and so (B.2) is satisfied.

Now we show that the last constraint, (B.3), also holds. We start by expanding $\|Z_i\|_2$ for some i.

$$
\begin{aligned}
\|Z_i\|_2 &= \|X_i \cdot \mathbb{I}[\|X_i\|_2 < r_1] - \mathbb{E}[Y_i]\|_2 \\
&\leqslant \|X_i \cdot \mathbb{I}[\|X_i\|_2 < r_1]\|_2 + \|\mathbb{E}[Y_i]\|_2 \\
&\leqslant r_1 + \|\mathbb{E}[Y_i]\|_2 \\
\text{(by Lemma B.3)} \qquad &\leqslant 2r_1.
\end{aligned}
$$

$\square$

LEMMA B.5. *Let $k > 2$ and $t > 0$. Let X_i be a random vector in $\mathbb{R}^d$ with $\mathbb{E}[X_i] = 0$ and k-th moment bounded by 1. Let $r_1 = \frac{d}{t}$. Let $Y_i = X_i \cdot \mathbb{I}[\|X_i\|_2 < r_1]$. Let $Z_i = Y_i - \mathbb{E}[Y_i]$. There exists $t_2 = \Omega(\sqrt{d})$ such that, for all positive $t \leqslant t_2$,*

$$\mathbb{P}\left[\left\|\sum_{i=1}^m Z_i\right\|_2 \geqslant m\left(\frac{t}{3} - \|\mathbb{E}[Y_i]\|_2\right)\right] \leqslant \exp\left(-\Theta\left(\frac{mt^2}{d}\right)\right)$$

Proof. We apply the Vectorized Bernstein's Inequality (i.e. Lemma 4.7) by setting $t' = t/3 - \|\mathbb{E}[Y_i]\|_2$, $r = 2r_1$, and $\sigma^2 = d$. One can check using Lemma B.3 that $0 < t' < \sigma^2/r$ is satisfied for all $0 < t < \sqrt{d}/12$. Thus,

$$
\begin{aligned}
\mathbb{P}\left[\left\|\sum_{i=1}^m Z_i\right\|_2 \geqslant m(t/3 - \|\mathbb{E}[Y_i]\|_2)\right] &= \mathbb{P}\left[\left\|\sum_{i=1}^m Z_i\right\|_2 \geqslant mt'\right] \\
\text{(by Lemma 4.7)} \qquad &\leqslant \exp\left(-m\frac{(t')^2}{8\sigma^2} + \frac{1}{4}\right) \\
&= \exp\left(-m\frac{(t/3 - \|\mathbb{E}[Y_i]\|_2)^2}{8d} + \frac{1}{4}\right) \\
\text{(by Lemma B.3)} \qquad &\leqslant \exp\left(-m\frac{(t/3 - t/6)^2}{8d} + \frac{1}{4}\right) \\
&\leqslant \exp\left(-\Theta\left(\frac{mt^2}{d}\right)\right).
\end{aligned}
$$

$\square$

For the next lemma, recall that the function $f(\ell)$ is defined as

$$f(\ell) = \Theta\left(\frac{d^{k/2} \cdot 2^{\ell(k-1)} \log^k m}{m^{k-1} t^k}\right)^{2^{\ell-1}}.$$

LEMMA B.6. *Let $m, d, t > 0$ and $k > 2$. Let $r_2 = mt/(3\log m)$ and let $r_1 = d/t$. There exists $t_1 = O\left(\sqrt{\frac{d\log m}{m}}\right)$ and $t_2 = \Omega\left(\sqrt{d}\log^{\frac{-1}{k-2}} m\right)$ such that, for all $t \in [t_1, t_2]$,*

$$f(\log(r_2/r_1)) = \tilde{O}\left(\frac{d^{k/2}}{m^{k-1} t^k}\right).$$

Proof. Define $g(t) = f\left(\log\left(\frac{mt^2}{3d\log m}\right)\right)$. Note that

$$\min_{t\in[t_1,t_2]} \tilde{O}\left(\frac{d^{k/2}}{m^{k-1} t^k}\right) = \tilde{O}\left(\frac{1}{m^{k-1}}\right).$$

Thus, it suffices to show that

(B.5) $$\max_{t\in[t_1,t_2]} g(t) \leqslant \tilde{O}\left(\frac{1}{m^{k-1}}\right).$$

We break the proof into two parts. First, we show that

(B.6) $$\max_{t\in[t_1,t_2]} g(t) \in \{g(t_1), g(t_2)\}.$$

Second, we show that

(B.7) $$g(t_1), g(t_2) \leqslant \tilde{O}\left(\frac{1}{m^{k-1}}\right).$$

To show (B.6), it suffices to show that $g(t)$ is convex with respect to t for $t > 0$. Note that

$$g(t) = \Theta\left(\frac{md^{k/2}\left(\frac{mt^2}{3d\log m}\right)^{k-1}\log^k m}{m^k t^k}\right)^{\frac{mt^2}{6d\log m}}$$

$$= \Theta\left(\frac{d^{-k/2+1} t^{k-2}\log m}{3^{k-1}}\right)^{\frac{mt^2}{6d\log m}}$$

$$= \Theta\left(\log m\left(\frac{t^2}{d}\right)^{\frac{k-2}{2}}\right)^{\frac{mt^2}{6d\log m}}.$$

Note that

$$\Theta\left(\log m\left(\frac{t^2}{d}\right)^{\frac{k-2}{2}}\right)^{\frac{mt^2}{6d\log m}} = \left(\left(\frac{\log m}{d^{\frac{k-2}{2}}}\right)^{\frac{m}{6d\log m}}\left(t^2\right)^{\frac{k-2}{2}\cdot\frac{m}{6d\log m}}\right)^{t^2} = (ax^b)^x,$$

for $x = t^2$, $a = \left(\frac{\log m}{d^{\frac{k-2}{2}}}\right)^{\frac{m}{6d\log m}}$ and $b = \frac{k-2}{2}\cdot\frac{m}{6d\log m}$. Since $a, b, x > 0$, then by Lemma B.2, $(ax^b)^x$ is convex with respect to x, and so g is also convex for $t > 0$. Thus, (B.6) holds.

We move on to showing (B.7). Let $t_1 = \sqrt{\frac{cd \log m}{m}}$ for some $c \geq \frac{12(k-1)}{k-2}$. When $t = t_1$,

$$
\begin{aligned}
g(t_1) &= \Theta\left(\log m \left(\frac{t_1^2}{d}\right)^{\frac{k-2}{2} \cdot \frac{mt_1^2}{6d \log m}}\right) \\
&= \Theta\left(\log m \left(\frac{\log m}{m}\right)^{\frac{k-2}{2} \cdot \frac{c}{6}}\right) \\
&= \tilde{O}\left(\frac{1}{m^{\left(\frac{k-2}{2} \cdot \frac{c}{6}\right)}}\right) \\
&= \tilde{O}\left(\frac{1}{m^{k-1}}\right).
\end{aligned}
$$

Let $t_2 = \frac{\sqrt{d} \log^{\frac{-1}{k-2}} m}{2}$. When $t = t_2$,

$$
\begin{aligned}
g(t_2) &= \Theta\left(\log m \left(\frac{t_2^2}{d}\right)^{\frac{k-2}{2} \cdot \frac{mt_2^2}{6d \log m}}\right) \\
&= O\left(\frac{1}{2}\right)^{O(m)} \\
&= \tilde{O}\left(\frac{1}{m^{k-1}}\right).
\end{aligned}
$$

$\square$

B.3 Missing Proofs from Section 4.1.4

LEMMA B.7. *Let $\ell > 0$. Let $\|u - \mu\|_2 < \rho$. It follows that*

$$
\|X - \mu\|_2 \geq \|X - \mathrm{clip}_{\rho,u}(X)\|_2 + \rho - \|\mu - u\|_2.
$$

Proof. Let $r = \|u - \mu\|_2$, and let $\mathbb{B}^d(u, r)$ be the ℓ_2 ball in d dimensions of radius r that is centered at u. Note that, by assumption, $r < \rho$, and so it follows that $\mu \in \mathbb{B}^d(u, r) \subseteq \mathbb{B}^d(u, \rho)$. Let

$$
\mu^* = \arg\min_{\mu' \in \mathbb{B}^d(u,r)} \|X - \mu'\|_2.
$$

One can check that μ^* is a point on the surface of $\mathbb{B}^d(u, r)$, and also that u, μ^*, $\mathrm{clip}_{\rho,u}(X)$, and X are collinear and arranged sequentially on the line in that order. We use these facts to relate the quantities $\|X - \mu\|_2$ and $\|X - \mathrm{clip}_{\rho,u}(X)\|_2$:

$$
\begin{aligned}
\|X - \mu\|_2 &\geq \|X - \mu^*\|_2 \\
&= \|X - u\|_2 - \|\mu^* - u\|_2 \\
&= \|X - \mathrm{clip}_{\rho,u}(X)\|_2 + \|\mathrm{clip}_{\rho,u}(X) - u\|_2 - \|\mu^* - u\|_2 \\
(\text{B.8}) \qquad &\geq \|X - \mathrm{clip}_{\rho,u}(X)\|_2 + \rho - \|\mu - u\|_2
\end{aligned}
$$

Note that $\|X - \mathrm{clip}_{\rho,u}(X)\|_2 \geq \ell$ thus implies $\|X - \mu\|_2 \geq \ell + \rho - \|\mu - u\|_2$, completing the proof. $\square$

B.4 Missing Proofs from Section 4.3 To prove Lemma 4.8, we use the following lemma about Gaussian random variables:

LEMMA B.8. (LAURENT AND MASSART [LM00]) *Let $d \in \mathbb{N}$. Consider the random vector $W \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$. Then, for every $s \geqslant 0$,*

$$\mathbb{P}\left[\|W\|_2^2 \geqslant \sigma^2\left(d + 2d\sqrt{s} + 2ds\right)\right] \leqslant \exp\left(-sd\right).$$

We also use Lemma B.9 which gives an analysis of the error incurred due to the sampling process.

LEMMA B.9. *Suppose $\mathcal{D}$ over $\mathbb{R}^d$ satisfies $\sigma_k(\mathcal{D}) \leqslant 1$. For all $i \in [n]$ and $j \in [m]$, let $X_j^{(i)} \overset{iid}{\sim} \mathcal{D}$, and for some $\rho \geqslant 0$, $u \in \mathbb{R}^d$ let*

$$Z_i = \text{clip}_{\rho,u}\left(\frac{1}{m}\sum_{j=1}^{m} X_j^{(i)}\right).$$

Let $\alpha, \beta > 0$. There exists $n_0 = O\left(\frac{d}{m\beta\alpha^2}\right)$ such that for all $n \geqslant n_0$ and for all ρ and u,

$$(\text{B.9}) \qquad \mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right]\right\|_2 \geqslant \alpha\right] \leqslant \beta.$$

Proof. Let $Y^{(i)} = \frac{1}{m}\sum_{j=1}^{m} X_j^{(i)} \in \mathbb{R}^d$, and let $Y^{(i)}(\ell) \in \mathbb{R}$ be the ℓth coordinate of $Y^{(i)}$. We begin by expanding the following expression of expectation:

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right]\right\|_2^2\right]$$

$$= \frac{1}{n^2}\,\mathbb{E}\left[\left\|\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])\right\|_2^2\right]$$

$$\leqslant \frac{1}{n^2}\,\mathbb{E}\left[\sum_{i=1}^{n}\|Z_i - \mathbb{E}[Z_i]\|_2^2\right]$$

$$= \frac{1}{n}\,\mathbb{E}\left[\|Z_1 - \mathbb{E}[Z_1]\|_2^2\right]$$

$$\leqslant \frac{1}{n}\,\mathbb{E}\left[\left\|\frac{1}{m}\sum_{j=1}^{m} X_j^{(1)} - \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} X_j^{(1)}\right]\right\|_2^2\right]$$

$$= \frac{1}{n}\,\mathbb{E}\left[\sum_{\ell=1}^{d}\left(Y^{(1)}(\ell) - \mathbb{E}\left[Y^{(1)}(\ell)\right]\right)^2\right]$$

$$= \frac{1}{n}\sum_{\ell=1}^{d}\mathbb{E}\left[\left(Y^{(1)}(\ell) - \mathbb{E}\left[Y^{(1)}(\ell)\right]\right)^2\right]$$

$$\leqslant \frac{d}{nm}.$$

Where the last line follows from the assumption that $\sigma_k(\mathcal{D}) \leqslant 1$. By Chebyshev's Inequality,

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right]\right\|_2 \geqslant \alpha\right] = O\left(\frac{d}{mn\alpha^2}\right) \leqslant \beta.$$

□

LEMMA 4.8. (NOISE AND SAMPLING) *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ and $\sigma_k(\mathcal{D}) \leqslant 1$. Suppose that, for all $i \in [n]$ and $j \in [m]$, $X_j^{(i)} \sim \mathcal{D}$. There exists $n_0 = \tilde{O}\left(\frac{d\sqrt{\log(1/\delta)}}{\varepsilon}\right)$ such that, for all $n \geqslant n_0$,*

$$\text{noise}(t) + \text{sampling}(t) = \tilde{O}\left(\frac{\rho_t}{\sqrt{d}}\right)$$

with probability at least $1 - 0.1/\log\log d$,

Proof. Recall from ClipAndNoise that $W_t \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ with $\sigma = \tilde{O}(\rho_t/(n\varepsilon'))$. Using Lemma B.8, let $s = \frac{\log(\log\log d/0.05)}{d}$. With probability at least $1 - 0.05/(\log\log d)$,

$$\text{noise}(t) = \|W_t\|_2 = \tilde{O}\left(\frac{\sqrt{d}\rho_t}{n\varepsilon'}\right).$$

Thus, there exists $n_0 = \tilde{O}\left(\frac{d\sqrt{\log(1/\delta)}}{\varepsilon}\right)$, such that, for $n \geqslant n_0$, it follows that with probability at least $1 - 0.05/(\log\log d)$,

$$\|W\|_2 \leqslant \frac{\rho_t}{4\sqrt{d}}.$$

In order to bound sampling(t), we apply Lemma B.9 directly, setting $\alpha = \frac{\rho_t}{2\sqrt{d}}$ and $\beta = 0.05/\log\log d$. In particular, if

$$n \geqslant \frac{d\log\log d}{0.05} \geqslant \frac{d\log\log d}{0.05m} \cdot \frac{4d}{\rho_t^2} \geqslant \frac{d}{m\beta\alpha},$$

then sampling$(t) > \frac{\rho_t}{2\sqrt{d}}$ with probability at most $\frac{0.05}{\log\log d}$. $\square$

B.5 Coarse Estimation In this subsection, we provide an algorithm to obtain a coarse estimate of the mean, up to accuracy $O(\sqrt{d/m})$ in high dimensions under approximate differential privacy, using 1-dimensional estimates (Theorem 3.2).

THEOREM B.1. (APPROXIMATE DP COARSE ESTIMATION IN HIGH DIMENSIONS) *Let n be the number of people, m be the number of samples per person, and $\varepsilon > 0$, $\delta \in (0, 1)$. Assume P is a distribution over $\mathbb{R}^d$ with k-th moment bounded by 1. Then there exists an efficient person-level (ε, δ)-DP algorithm CoarseEstimate$(X; \varepsilon, \delta, r)$ that takes i.i.d. samples $\{X_j^{(i)}\}_{i,j}$ from P, and outputs an estimate of the mean up to accuracy $16^{1/k}\sqrt{d/m} < r$, with success probability $1 - \beta$, as long as $n \geqslant n_0$, for some*

$$n_0 = \tilde{O}\left(\frac{\log(1/\beta)}{\log(r\sqrt{m/d})} + \min\left\{\frac{d\log(1/(\delta\beta))}{\varepsilon}, \frac{\sqrt{d\log(1/\delta)}\log(1/\delta\beta)}{\varepsilon}\right\}\right),$$

where $\tilde{O}$ hides lower order logarithmic factors in d.

Proof. Run d instances of Algorithm 1 with $r' = r/\sqrt{d}$, $\varepsilon' = \varepsilon/\sqrt{d\log(1/\delta)}$, $\delta' = \delta/d$, on each coordinate of the dataset X, and output the coordinate-wise estimate of the mean. By Theorem 3.2 (one dimensional coarse estimation), and Lemma 2.3 (advanced composition), we know that this output will be person-level $(O(\varepsilon), O(\delta))$-DP. Moreover, the output will have accuracy r, with probability $1 - \beta$, as long as $n \geqslant n_1$, for some

$$n_1 = O\left(\frac{\log(d/\beta)}{\log(r\sqrt{m/d})} + \frac{\sqrt{d}\sqrt{\log(1/\delta)}\log(d^2/\delta\beta)}{\varepsilon}\right).$$

Alternatively, using basic composition (Lemma 2.2), and applying Algorithm 1 with $r' = r/\sqrt{d}$, $\varepsilon' = \varepsilon/d$, $\delta' = \delta/d$, on each coordinate of X, and outputting the coordinate-wise estimate of the mean, we obtain a person-level (ε, δ)-DP algorithm that has accuracy r, with success probability $1 - \beta$, as long as $n \geqslant n_2$, for some

$$n_2 = O\left(\frac{\log(d/\beta)}{\log(r\sqrt{m/d})} + \frac{d\log(d^2/\delta\beta)}{\varepsilon} \right).$$

□

C Lower Bounds Under Pure Differential Privacy

Suppose $\mathcal{D}$ is a distribution over $\mathbb{R}^d$ with k-th moment bounded from above by 1. Assume n people each take m samples from the distribution $\mathcal{D}$. In this section we show lower bounds against the number of people required to estimate the mean of the distribution $\mathcal{D}$ up to accuracy α with success probability $1 - \beta$, under person-level ε-differential privacy.

We can describe the problem in an equivalent language: take n samples from the tensorized distribution $\mathcal{D}^{\otimes m}$. How many samples are required in order to estimate the mean of $\mathcal{D}$ up to accuracy α with success probability $1 - \beta$, under item-level ε differential privacy?

Technically, we apply packing lower bounds to distributions of the form $\mathcal{D}^{\otimes m}$. Specifically, we repeatedly apply the following theorem, that shows sample complexity lower bounds against learning a set of distributions that are all γ far from a central distribution and have disjoint parameters that can be learned with probability at least $1 - \beta$. Moreover, we use some facts about Kullback-Leibler divergence in order to analyze the tensorized distribution.

THEOREM C.1. (THEOREM 7.1 IN [HKM22]) *Let $\mathcal{P} = \{P_1, \ldots, P_m\}$ be a set of distributions, and P_O be a distribution such that for every $P_i \in \mathcal{P}$, $\|P_i - P_O\|_{\mathrm{TV}} \leqslant \gamma$. Let $\mathcal{G} = \{G_1, \ldots, G_m\}$ be a collection of disjoint subsets of some set $\mathcal{Y}$. If there is an ε-DP algorithm M such that $\mathbb{P}_{X \sim P_i^n}[M(X) \in G_i] \geqslant 1 - \beta$ for all $i \in [m]$, then*

$$n \geqslant \Omega\left(\frac{\log m + \log(1/\beta)}{\gamma(e^{2\varepsilon} - 1)} \right).$$

Note that for the usual regime $\varepsilon \leqslant 1$, we can replace the $e^{2\varepsilon} - 1$ in the denominator with ε.

We instantiate Theorem C.1 three times in order to prove each term of the lower bound. The first term is the non-private cost, the second term corresponds to the cost of estimating the mean of a Gaussian, the third term corresponds to the cost of estimating the mean of point distributions with bounded k-th moment, and the last term corresponds to the cost of estimating up to a coarse estimate.

THEOREM C.2. (PURE DP LOWER BOUND FOR BOUNDED k-TH MOMENTS) *Suppose $k \geqslant 2$, and $\mathcal{D}$ is a distribution with k-th moment bounded by 1 and mean μ. Moreover, suppose μ is in a ball of radius R, where $\alpha \leqslant \min\{\frac{6}{25}R, (1/25)^{(k-1)/k}\}$, and assume $\varepsilon \leqslant 1$. Any ε-DP algorithm that takes n samples from $\mathcal{D}^{\otimes m}$ and outputs $\hat{\mu}$ such that $\|\mu - \hat{\mu}\|_2 \leqslant \alpha$ with probability $1 - \beta$ requires*

$$n = \Omega_k\left(\frac{d + \log(1/\beta)}{\alpha^2 m} + \frac{d + \log(1/\beta)}{\alpha\sqrt{m}\varepsilon} + \frac{d + \log(1/\beta)}{\alpha^{\frac{k}{k-1}}m\varepsilon} + \frac{d\log(R/\alpha) + \log(1/\beta)}{\varepsilon} \right).$$

many samples, where Ω_k hides multiplicative factors that only depend on k.

Proof. The first term is the cost of estimating the mean of a distribution with nm samples without privacy. For the other terms putting together Corollary C.1, Lemma C.2, and Lemma C.3 finishes the proof. □

C.1 Preliminaries First, we state some facts about Kullback–Leibler divergence and total variation distance. In our analysis we intend to analyze the total variation distance of the tensorized distribution $\mathcal{D}^{\otimes m}$. However, analyzing the total variation distance of the tensorized distribution directly, is difficult, as the total variation distance does not behave well under tensorization. Thereofore, instead we use the Kullback-Leibler divergence to analyze the total variation distance of the tensorized distribution through Pinkser's inequality.

We want to analyze the tensorized distribution $\mathcal{D}^{\otimes m}$. The following fact states that the Kullback-Leibler divergence tensorizes well.

FACT C.1. (TENSORIZATION OF KL DIVERGENCE) *Assume $\{P_i\}_{i=1}^n$, and $\{Q_i\}_{i=1}^n$ are distributions over X and $P = \otimes_{i=1}^n P_i$ and $Q = \otimes_{i=1}^n Q_i$. Then*

$$\mathrm{KL}(P\|Q) = \sum_{i=1}^n \mathrm{KL}(P_i\|Q_i)\,.$$

We apply Pinsker's inequality to relate Kullback-Leibler divergence and total variation distance.

FACT C.2. (PINSKER'S INEQUALITY) *Let P and Q be two distributions over X. Then*

$$\mathrm{d}_{\mathrm{TV}}(P, Q) \leqslant \sqrt{\frac{\mathrm{KL}(P\|Q)}{2}}\,.$$

Next, for two arbitrary Gaussians their closed form KL divergence is known, which we use in our analysis.

FACT C.3. (KL DIVERGENCE OF GAUSSIANS [RW05]) *For two arbitrary Gaussians $\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)$ we have*

$$\mathrm{KL}(\mathcal{N}(\mu_1, \Sigma_1)\|\mathcal{N}(\mu_2, \Sigma_2)) = \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_1^{-1}\Sigma_2 - I\right) + (\mu_1 - \mu_2)^\top \Sigma_1^{-1}(\mu_1 - \mu_2) - \log\det\left(\Sigma_2\Sigma_1^{-1}\right)\right).$$

Specifically for identity covariance,

$$\mathrm{KL}(\mathcal{N}(\mu_1, I)\|\mathcal{N}(\mu_2, I)) = \frac{\|\mu_1 - \mu_2\|_2^2}{2}\,.$$

Next we state a basic fact about packing number $M(\Theta, \|\cdot\|, \varepsilon)$ and covering number $N(\Theta, \|\cdot\|, \varepsilon)$. This fact shows and upper and lower bound for the packing and covering number of a unit norm ball.

FACT C.4. (METRIC ENTROPY OF NORM BALLS [WU16]) *Let $B_2(1)$ be the unit ℓ_2 ball. Consider $N(B_2(1), \|\cdot\|_2, \varepsilon)$. When $\varepsilon \geqslant 1$, $N(B_2(1), \|\cdot\|_2, \varepsilon) = 1$. When $\varepsilon < 1$,*

$$\left(\frac{1}{\varepsilon}\right)^d \leqslant N(B_2(1), \|\cdot\|_2, \varepsilon) \leqslant M(B_2(1), \|\cdot\|_2, \varepsilon) \leqslant \left(1 + \frac{2}{\varepsilon}\right)^d \leqslant \left(\frac{3}{\varepsilon}\right)^d.$$

Hence $d\log\frac{1}{\varepsilon} \leqslant \log N(B_2(1), \|\cdot\|_2, \varepsilon) \leqslant \log M(B_2(1), \|\cdot\|_2, \varepsilon) \leqslant d\log\frac{3}{\varepsilon}$.

C.2 Proof of Theorem C.2 In this section we prove the lower bounds for each term in Theorem C.2. First we prove the second term which corresponds to learning a Gaussian.

LEMMA C.1. (PURE DP GAUSSIAN MEAN ESTIMATION LOWER BOUND) *Suppose μ is in a ball of radius R, where $\alpha \leqslant R/4$, and suppose $\varepsilon \leqslant 1$. Any ε-DP algorithm that takes n samples from $\mathcal{N}(\mu, I_d)^{\otimes m}$ and outputs $\hat{\mu}$ such that $\|\mu - \hat{\mu}\|_2 \leqslant \alpha$ with probability $1 - \beta$ requires*

$$n = \Omega\left(\frac{d + \log(1/\beta)}{\alpha\sqrt{m}\varepsilon}\right),$$

many samples.

Proof. We want to use Theorem C.1. Let $P_O = \mathcal{N}(0, I_d)^{\otimes m}$, $P_i = \mathcal{N}(\lambda v_i, I_d)^{\otimes m}$, for v_i in the unit ball, such that $\|v_i - v_j\|_2 \geqslant 2\alpha/\lambda$, and for λ to be determined later. We know that a set I such that $\log|I| \geqslant d\log(\lambda/2\alpha)$ of such $v_i's$ exists from Fact C.4. Let $G_i = \lambda v_i + \alpha B_d$, for such a set I and $i \neq j$, we will have that G_i's are disjoint. Now, from Fact C.3, and $\|v_i\|_2 \leqslant 1$ we know that $\mathrm{KL}(\mathcal{N}(0, I_d)\|\mathcal{N}(\lambda v_i, I_d)) \leqslant \lambda^2/2$. Tensorization of KL (Fact C.1) and Pinsker's inequality (Fact C.2) imply that

$$\|P_O - P_i\|_{\mathrm{TV}} \leqslant \sqrt{\frac{\mathrm{KL}(P_i\|P_O)}{2}} \leqslant \sqrt{\frac{m \cdot \mathrm{KL}(\mathcal{N}(\lambda v_i, I_d), \mathcal{N}(0, I_d))}{2}} \leqslant \sqrt{m}\lambda/2\,.$$

Taking $\lambda = 4\alpha$ we obtain

$$n = \Omega\left(\frac{d + \log(1/\beta)}{\alpha\sqrt{m}\varepsilon}\right).$$

□

However, note that the the k-th moment of $\mathcal{N}(\mu, I_d)$, is not bounded by 1, but instead it is bounded from above by $(k-1)!!$, where the !! denotes the double factorial that is the product of all numbers from 1 to $k-1$ with the same parity as $k-1$. We state the corollary below to extend the lemma above to our setting where the k-th moment is bounded by 1.

COROLLARY C.1. (PURE DP NORMALIZED GAUSSIAN MEAN ESTIMATION LOWER BOUND) *Let σ_k^k be the k-th moment of $\mathcal{N}(0,1)$. Suppose μ is in a ball of radius R, where $\alpha \leqslant R/4$, and suppose $\varepsilon \leqslant 1$. Any ε-DP algorithm that takes n samples from $\mathcal{N}(\mu, I_d/\sigma_k^2)^{\otimes m}$, and outputs $\hat{\mu}$ such that $\|\mu - \hat{\mu}\|_2 \leqslant \alpha$ with probability $1 - \beta$ requires*

$$n = \Omega_k\left(\frac{d + \log(1/\beta)}{\alpha\sqrt{m}\varepsilon}\right),$$

where Ω_k hides multiplcative factors that only depend on k. Moreover, $\mathcal{N}(\mu, I_d/\sigma_k^2)^{\otimes m}$ has k-th moment bounded by 1.

Proof. We apply Lemma C.1. Assume there exists an ε-DP algorithm that takes n samples from $\mathcal{N}(\mu, I_d/\sigma_k^2)^{\otimes m}$, and outputs an estimate of μ up to accuracy α, with probability $1 - \beta$. Now we can take the same algorithm and apply it to Y_i's where $Y_i = X_i/\sigma_k$ and $X_i \sim \mathcal{N}(\mu\sigma_k, I_d)^{\otimes m}$. Multiplying this estimate by σ_k would give an $\alpha\sigma_k$ estimate of the mean of $\mathcal{N}(\mu\sigma_k, I_d)$. However, we know that such an estimate requires at least $\Omega((d + \log(1/\beta))/(\alpha\sigma_k\sqrt{m}\varepsilon))$ many samples. Since σ_k is a constant that only depends on k we must have

$$n = \Omega_k\left(\frac{d + \log(1/\beta)}{\alpha\sqrt{m}\varepsilon}\right),$$

as desired. □

Next we prove the third term which we show through constructing point mass distributions that have bounded k-th moements.

LEMMA C.2. (PURE DP POINT DISTRIBUTION WITH BOUNDED MOMENTS LOWER BOUND) *Suppose $k \geqslant 2$, and $\mathcal{D}$ is a distribution with k-th moment bounded by 1 and mean μ. Moreover, suppose μ is in a ball of radius R, where $\alpha \leqslant \min\{\frac{6}{25}R, (1/25)^{(k-1)/k}\}$, and assume $\varepsilon \leqslant 1$. Any ε-DP algorithm that takes n samples from $\mathcal{D}^{\otimes m}$ and outputs $\hat{\mu}$ such that $\|\mu - \hat{\mu}\|_2 \leqslant \alpha$ with probability $1 - \beta$ requires*

$$n = \Omega\left(\frac{d + \log(1/\beta)}{\alpha^{\frac{k}{k-1}}m\varepsilon}\right)$$

many samples.

Proof. We want to use Theorem C.1. A set I, with $\log|I| \geqslant d\log(2)$ of indices and vectors v_i in the unit ball exist such that $\|v_i - v_i\|_2 \geqslant 1/2$.

Suppose Q_O is a distribution with all of the mass on the origin, and Q_i is equal to 0 with probability $1 - \lambda$, and is equal to $\frac{1}{6\alpha^{\frac{1}{k-1}}}v_i$, with probability λ, where $\lambda = 25\alpha^{\frac{k}{k-1}}$. Then $\mathbb{E}_{x \sim Q_i}[x] = \frac{25}{6}\alpha v_i$, and for all unit u, $\mathbb{E}_{x \sim Q_i}[|\langle x, u \rangle|^k] \leqslant \frac{25}{6^k} < 1$, therefore Q_i's satisfy the moment assumption. Moreover if we take $G_i = \frac{26}{6}\alpha v_i + \alpha B$, then G_i's are disjoint. Let $P_i = Q_i^{\otimes m}$, then

$$\|P_O - P_i\|_{\text{TV}} = \left\|Q_O^{\otimes m} - Q_i^{\otimes m}\right\|_{\text{TV}} \leqslant \underset{x \sim Q_i^{\otimes m}}{\mathbb{P}}[x \neq 0^{\otimes m}] = 1 - \underset{x \sim Q_i^{\otimes m}}{\mathbb{P}}[x = 0^{\otimes m}] \leqslant 1 - (1 - \lambda)^m \leqslant m\lambda,$$

where the last inequality comes from the fact that the function $f(y) = 1 - my + (1-x)^m$, $y \in [0,1]$ is an increasing function and is minimized at 0, and therefore $f(y) \geqslant 0, \forall y \in [0,1]$. Applying Theorem C.1 finishes the proof. □

Lastly, we prove the last term which corresponds to the cost of finding a coarse estimate. We show this by choosing a set of point mass distributions over the ball of radius R.

LEMMA C.3. (PURE DP LOWER BOUND RANGE PARAMETER) *Suppose $k \geqslant 2$, and $\mathcal{D}$ is a distribution with k-th moment bounded by 1 and mean μ. Moreover, suppose μ is in a ball of radius R, where $\alpha < R/2$, and assume $\varepsilon \leqslant 1$. Any ε-DP algorithm that takes n samples from $\mathcal{D}^{\otimes m}$ and outputs $\hat{\mu}$ such that $\|\mu - \hat{\mu}\|_2 \leqslant \alpha$ with probability $1 - \beta$ requires*

$$n = \Omega\left(\frac{d \log(R/\alpha) + \log(1/\beta)}{\varepsilon}\right)$$

many samples.

Proof. We want to use Theorem C.1. Fact C.4 implies that there exists a set I of indices such that $\log|I| \geqslant d \log(R/2\alpha)$ and v_i in the ball of radius R such that $\|v_i - v_j\|_2 \geqslant 2\alpha$, for every $i \neq j \in I$. Let Q_O be the distribution with all of the mass on the origin, and Q_i's be the distributions with all of the mass on v_i, then $\|Q_O - Q_i\|_{\mathrm{TV}} = 1$ for every $i \in I$. Let $P_i = Q_i^{\otimes m}$, then $\|P_O - P_i\|_{\mathrm{TV}} = 1$ for every $i \in I$. Let $G_i = v_i + \alpha B$, then G_i's are disjoint. Applying Theorem C.1 finishes the proof. $\square$

D Non Uniform Berry-Esseen

In this section we restate and prove the Non Uniform Berry-Esseen Theorem [Mic76].

THEOREM D.1. ([MIC76]) *Let $k \geqslant 3$. Assume X is a distribution with mean 0 and k-th moment bounded by 1 . Then for any $t \geqslant \sqrt{(k-1)\log m}$, we have*

$$\mathbb{P}\left[\sum X_i/\sqrt{m} \geqslant t\right] \leqslant_k m^{-k/2+1}t^{-2k} + m\,\mathbb{P}\left[X \geqslant rm^{1/2}t\right],$$

where $r = 1/(2(k-1)k)$.

Proof. Let Y_i be the truncation of X_i at $rm^{1/2}t$, i.e. $Y_i = \min(X_i, rm^{1/2}t)$, where $r = 1/(2(k-1)k)$. Then

$$\mathbb{P}\left[\sum X_i/\sqrt{m} \geqslant t\right] \leqslant \mathbb{P}\left[\sum Y_i/\sqrt{m} \geqslant t\right] + m\,\mathbb{P}\left[X \geqslant r\sqrt{m}t\right].$$

Therefore, we need to show that

$$\mathbb{P}\left[\sum Y_i/\sqrt{m} \geqslant t\right] \leqslant_k m^{-k/2+1}t^{-2k}.$$

Let $h = m^{-1/2}t^{-1} \cdot ((k-2)\log m + 2(k-1)k\log t)$, and $f(x) = \exp(hm^{1/2}x)$. f is a non-negative non-decreasing function. We apply Markov's inequality.

$$\begin{aligned}
\mathbb{P}\left[\sum Y_i/\sqrt{m} \geqslant t\right] &= \mathbb{P}\left[f\left(\sum Y_i/\sqrt{m}\right) \geqslant f(t)\right] \\
&\leqslant \frac{\mathbb{E}\left[f\left(\sum Y_i/\sqrt{m}\right)\right]}{f(t)} \\
&= \frac{\mathbb{E}\left[\prod f\left(Y_i/\sqrt{m}\right)\right]}{f(t)} \\
&= \frac{\mathbb{E}\left[f\left(Y_i/\sqrt{m}\right)\right]^m}{f(t)} \\
&= \frac{\mathbb{E}\left[\exp(hY_i)\right]^m}{\exp(hm^{1/2}t)} \\
&= \mathbb{E}\left[\exp(hY_i)\right]^m \cdot \exp(-hm^{1/2}t) \\
&\leqslant \mathbb{E}\left[\exp(hY_i)\right]^m \cdot m^{-k+2} \cdot t^{-2(k-1)k}
\end{aligned}$$

Now we have to bound $\mathbb{E}\left[\exp(hY_i)\right]$. We use the Taylor expansion of exp, and the fact that $Y_i \leqslant rm^{1/2}t$, to get

$$\exp(hY) \leqslant 1 + hY + \frac{h^2Y^2}{2} + \frac{h^3|Y|^3}{6} \cdot \max_{y' \leqslant rm^{1/2}t} \exp(hy') \leqslant 1 + hY + \frac{h^2Y^2}{2} + \frac{h^3|Y|^3}{6} \cdot \exp(hrm^{1/2}t).$$

Taking expectation, and using the fact that $\mathbb{E}[Y_i^2] \le \mathbb{E}[X_i^2] \le 1$,

$$\mathbb{E}\big[\exp(hY_i)\big] \le 1 + h|\mathbb{E}[Y_i]| + \frac{h^2\,\mathbb{E}\big[Y_i^2\big]}{2} + \frac{h^3\,\mathbb{E}\big[|Y_i|^3\big]}{6} \cdot \exp(hrm^{1/2}t)$$

$$\le 1 + h|\mathbb{E}[Y_i]| + \frac{h^2}{2} + \frac{h^3\,\mathbb{E}\big[|Y_i|^3\big]}{6} \cdot \exp(hrm^{1/2}t).$$

Now we bound $|\mathbb{E}[Y_i]|$ and $\mathbb{E}\big[|Y_i|^3\big]$. We bound $|\mathbb{E}[Y_i]|$ using Lemma 3.3.

$$|\mathbb{E}[Y_i]| \le \int_{rm^{1/2}t}^{\infty} \mathbb{P}[X \ge x]\,\mathrm{d}x$$

$$\le \int_{rm^{1/2}t}^{\infty} \mathbb{P}[|X| \ge x]\,\mathrm{d}x$$

$$\le \int_{rm^{1/2}t}^{\infty} \frac{\mathbb{E}\big[X^2\big]}{x^2}\,\mathrm{d}x$$

$$\le \frac{1}{rm^{1/2}t}.$$

In order to bound $\mathbb{E}\big[|Y_i|^3\big]$, note that $\mathbb{E}\big[|Y_i|^3\big] \le \mathbb{E}\big[|X_i|^3\big] \le 1$. Therefore, since $t \ge 1$, we have

$$\mathbb{E}\big[\exp(hY_i)\big] \le 1 + \frac{h}{rm^{1/2}t} + \frac{h^2}{2} + \frac{h^3}{6} \cdot \exp(hrm^{1/2}t)$$

$$= 1 + \frac{h}{rm^{1/2}t} + \frac{h^2}{2} + \frac{h^3}{6} \cdot m^{(k-2)/2(k-1)k}t$$

$$\le 1 + \frac{h^2}{2} + b_1 m^{-1},$$

where b_1 is a constant depending only on k. Therefore,

$$\mathbb{E}\big[\exp(hY_i)\big] \le 1 + \frac{h^2}{2} + b_1 m^{-1} \le \exp(h^2/2 + b_1 m^{-1}),$$

and

$$\mathbb{E}\big[\exp(hY_i)\big]^m \le \exp(h^2 m/2 + b_1).$$

Note that

$$h^2 m/2 \le \frac{((k-2)\log m + 2(k-1)k\log t)^2}{2t^2}$$

$$= \frac{(k-2)^2\log^2 m}{2t^2} + \frac{(k-2)\log m \cdot 2(k-1)k\log t}{t^2} + \frac{4(k-1)^2 k^2 \log^2 t}{2t^2}$$

$$= \frac{(k-2)\log m}{t^2} \cdot \frac{(k-2)\log m}{2} + \frac{(k-1)\log m}{t^2} \cdot 2(k-2)k\log t + \frac{\log^2 t}{t^2} \cdot 2(k-1)^2 k^2$$

$$\le \frac{k-2}{2}\log m + 2(k-1)k\log t + b_2,$$

where b_2 is a constant depending only on k. The last inequality follows from the assumption that $t \ge \sqrt{(k-1)\log m}$, and the fact that $t \ge \log t$. Therefore,

$$\mathbb{E}\big[\exp(hY_i)\big]^m \le \exp(\frac{k-2}{2}\log m + 2(k-2)k\log t + b_3),$$

where b_3 is a constant depending only on k. Therefore,

$$\mathbb{P}\Big[\sum Y_i/\sqrt{m} \ge t\Big] \le \exp(\frac{k-2}{2}\log m + 2(k-2)k\log t + b_3) \cdot m^{-k+2} \cdot t^{-2(k-1)k} \le_k m^{-k/2+1}t^{-2k},$$

and we are done. $\quad\square$

COROLLARY D.1. (NON-UNIFORM TAIL BOUND FOR k-TH MOMENT BOUNDED BY 1 - LARGE t) *Let $k \geqslant 3$. Assume X is a distribution with mean 0 and k-th moment bounded by 1. Then for any $t \geqslant \sqrt{(k-1)\log m}$, we have*

$$\mathbb{P}\left[\sum X_i / \sqrt{m} \geqslant t\right] \leqslant_k m^{-k/2+1} t^{-k},$$

and consequently, for any $t \geqslant \sqrt{\frac{(k-1)\log m}{m}}$, we have

$$\mathbb{P}\left[\sum X_i / m \geqslant t\right] \leqslant_k m^{-k+1} t^{-k}.$$

Proof. Follows from the previous theorem and the fact that $\mathbb{P}\left[X \geqslant r m^{1/2} t\right] \leqslant r^{-k} m^{-k/2} t^{-k}$. $\qquad\square$

E Direct Proof of Proposition 6.2

In this appendix, we focus on proving Proposition 6.2 directly using fingerprinting. To do that, we will use the general lemma for exponential families from [KMS22a], which we state here (in a slightly simplified form):

PROPOSITION E.1. *Let p_η be a distribution over $S \subseteq \mathbb{R}^d$ belonging to an exponential family $\mathcal{E}(T, h)$ with natural parameter vector $\eta \in \mathcal{H} \subseteq \mathbb{R}^k$. Also, let $\eta^{(1)}, \eta^{(2)} \in \mathcal{H}$ and let $I_j := \left[\eta_j^{(1)}, \eta_j^{(2)}\right], \forall j \in [k]$ be a collection of intervals and $R := \eta^{(2)} - \eta^{(1)}, m := \frac{\eta^{(1)}+\eta^{(2)}}{2}$ be the corresponding width and midpoint vectors, respectively. Assume that $\bigotimes_{j\in[k]} I_j \subseteq \mathcal{H}$ and $\eta \sim \mathcal{U}\left(\bigotimes_{j\in[k]} I_j\right)$. Moreover, assume that we have a dataset $X \sim \left(p_\eta^{\otimes n}\middle|\eta\right)$ and an independently drawn point $X_i' \sim \left(p_\eta\middle|\eta\right)$ and $X_{\sim i}$ denotes the dataset where X_i has been replaced with X_i'. Finally, let $M \colon S^n \to \bigotimes_{j\in[k]} \left[\pm\frac{R_j}{2}\right]$ be an (ε, δ)-DP mechanism with $\varepsilon \in [0, 1], \delta \geqslant 0$ and $\underset{X,M}{\mathbb{E}}\left[\|M(X) - (\eta - m)\|_2^2\right] \leqslant \alpha^2 \leqslant \frac{\|R\|_2^2}{24}$. Then, for any $T_0 > 0$, it holds that:*

$$n\left\{2\delta T_0 + 2\alpha\varepsilon\underset{\eta}{\mathbb{E}}\left[\sqrt{\|\Sigma_T\|_2}\right] + 2\underset{\eta}{\mathbb{E}}\left[\int_{T_0}^{\infty} \underset{X_i}{\mathbb{P}}\left[\|T(X_i) - \mu_T\|_2 > \frac{4t}{\|R\|_\infty^3 \sqrt{k}}\right] dt\right]\right\} \geqslant \frac{\|R\|_2^2}{24}.$$

We note now that $\mathcal{N}(\mu, \mathbb{I})^{\otimes m}$ is an exponential family. Indeed, the density of $\mathcal{N}(\mu, \mathbb{I})^{\otimes m}$ is:

$$\frac{1}{(2\pi)^{\frac{md}{2}}} e^{-\frac{1}{2}\sum_{i\in[m]}\|x_i-\mu\|_2^2},$$

which can be written in the form $h(x)e^{\langle\eta, T(x)\rangle - Z(\eta)}$ for $x := (x_1, \ldots, x_m) \in \left(\mathbb{R}^d\right)^m$ with:

$$h(x) = \frac{1}{(2\pi)^{\frac{md}{2}}} e^{-\frac{1}{2}\sum_{i\in[m]}\|x_i\|_2^2},$$

$$T(x) = \sum_{i=1}^m x_i,$$

$$\eta = \mu,$$

$$Z(\eta) = \frac{m\|\eta\|_2^2}{2}.$$

At this point, we recall the following lemma from [KMS22a].

LEMMA E.1. *[Lemma C.5 from [KMS22a]]. Let $X \sim \mathcal{N}(\mu, \mathbb{I})$. Assuming that $T_0 \geqslant 2d$, we have:*

$$\int_{T_0}^{\infty} \underset{X}{\mathbb{P}}\left[\|X - \mu\|_2 > \frac{t}{2\sqrt{d}}\right] dt \leqslant \sqrt{2\pi d}\, e^{-\frac{\left(\sqrt{T_0^2 - 2d^2} - \sqrt{2}d\right)^2}{8d}}.$$

Using this lemma, we prove the following result.

LEMMA E.2. *Let $X \sim \mathcal{N}(\mu, \mathbb{I})^{\otimes m}$. Assuming that $T_0 \geqslant 2md$, we have:*

$$\int_{T_0}^{\infty} \mathbb{P}_X \left[\|T(X) - \mu_T\|_2 > \frac{t}{2\sqrt{d}} \right] dt \leqslant \sqrt{2\pi d} e^{-\frac{\left(\sqrt{\left(\frac{T_0}{m}\right)^2 - 2d^2} - \sqrt{2}d\right)^2}{8d}}.$$

Proof. We note that we have:

$$\int_{T_0}^{\infty} \mathbb{P}_X \left[\|T(X) - \mu_T\|_2 > \frac{t}{2\sqrt{d}} \right] dt = \int_{T_0}^{\infty} \mathbb{P}_{X_1,\dots,X_m} \left[\left\| \sum_{i=1}^{m} X_i - m\mu \right\|_2 > \frac{t}{2\sqrt{d}} \right] dt$$

$$= \int_{T_0}^{\infty} \mathbb{P}_{X \sim \mathcal{N}(0,1)} \left[\|X - \mu\|_2 > \frac{t}{2m\sqrt{d}} \right] dt$$

$$= \int_{\frac{T_0}{m}}^{\infty} \mathbb{P}_{X \sim \mathcal{N}(0,1)} \left[\|X - \mu\|_2 > \frac{t}{2\sqrt{d}} \right] dt$$

$$\leqslant \sqrt{2\pi d} e^{-\frac{\left(\sqrt{\left(\frac{T_0}{m}\right)^2 - 2d^2} - \sqrt{2}d\right)^2}{8d}},$$

where the last inequality follows directly from Lemma E.1. $\square$

We now prove Proposition 6.2.

Proof. Assume that we have $\mu \sim \mathcal{U}([\pm 1]^d)$, and $X \sim \left((\mathcal{N}(\mu, \mathbb{I})^{\otimes m})^{\otimes n} | \mu \right)$. Also, let $M : \left(\mathbb{R}^{m \times d} \right)^n \to [\pm 1]^d$ be an (ε, δ)-DP mechanism with:

$$\underset{X,M}{E} \left[\|M(X) - \mu\|_2^2 \right] \leqslant \alpha^2 \leqslant \frac{d}{6} = \frac{\|R\|_2^2}{24}.$$

By Proposition E.1 for batches of Gaussian samples $\mathcal{N}(\mu, \mathbb{I})^{\otimes m}$ with $\mu_j \in [\pm 1], \forall j \in [d]$, we get:

$$n \left(2\delta T_0 + 2\sqrt{m}\alpha\varepsilon + 2\sqrt{2\pi d} e^{-\frac{\left(\sqrt{\left(\frac{T_0}{m}\right)^2 - 2d^2} - \sqrt{2}d\right)^2}{8d}} \right) \geqslant \frac{d}{6}.$$

The rest of the proof follows the same steps as the proof in Appendix C.2 of [KMS22a], so we do not repeat it here in its entirety. The core idea is to show that, for $T_0 = 2m\sqrt{d}\sqrt{\ln\left(\frac{1}{\delta}\right) + \left(\sqrt{\ln\left(\frac{1}{\delta}\right)} + \sqrt{d}\right)^2}$, and:

$$\delta \leqslant \min\left\{ \frac{1}{144\sqrt{2}nm}, \frac{\sqrt{d}}{288\sqrt{2}nm\sqrt{\ln\left(\frac{144\sqrt{2}nm}{\sqrt{d}}\right)}} \right\},$$

we have that:

$$\delta T_0 \geqslant 2\sqrt{2\pi d} e^{-\frac{\left(\sqrt{\left(\frac{T_0}{m}\right)^2 - 2d^2} - \sqrt{2}d\right)^2}{8d}} \text{ and } 3n\delta T_0 \leqslant \frac{d}{12}.$$

$\square$