Private Mean Estimation of Heavy-Tailed Distributions

Gautam Kamath G@CSAIL.MIT.EDU
Cheriton School of Computer Science, University of Waterloo

Vikrant Singhal SINGHAL.VI@NORTHEASTERN.EDU
Jonathan Ullman JULLMAN@CCS.NEU.EDU
Khoury College of Computer Sciences, Northeastern University

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We give new upper and lower bounds on the minimax sample complexity of differentially private mean estimation of distributions with bounded k-th moments. Roughly speaking, in the univariate case, we show that

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

samples are necessary and sufficient to estimate the mean to α-accuracy under ε-differential privacy, or any of its common relaxations. This result demonstrates a qualitatively different behavior compared to estimation absent privacy constraints, for which the sample complexity is identical for all $k \geq 2$. We also give algorithms for the multivariate setting whose sample complexity is a factor of $O(d)$ larger than the univariate case.

Keywords: Differential Privacy; Mean Estimation; Heavy-Tailed Distributions

1. Introduction

Given samples $X_1, \ldots, X_n$ from a distribution $\mathcal{D}$, can we estimate the mean of $\mathcal{D}$? This is the problem of *mean estimation* which is, alongside hypothesis testing, one of the most fundamental questions in statistics. As a result, answers to this problem are known in fairly general settings. For instance, the empirical mean is known to be an optimal estimate of a distribution's true mean under minimal assumptions.

That said, statistics like the empirical mean put aside any concerns related to the sensitivity, and might vary significantly based on the addition of a single datapoint in the dataset. While this is not an inherently negative feature, it becomes a problem when the dataset contains personal information, and large shifts based on a single datapoint could potentially violate the corresponding individual's *privacy*. In order to assuage these concerns, we consider the problem of mean estimation under the constraint of *differential privacy* (DP) Dwork et al. (2006), considered by many to be the gold standard of data privacy. Informally, an algorithm is said to be differentially private if its distribution over outputs is insensitive to the addition or removal of a single datapoint from the dataset. Differential privacy has enjoyed widespread adoption, including deployment in by Apple Differential Privacy Team, Apple (2017), Google Erlingsson et al. (2014), Microsoft Ding et al. (2017), and the US Census Bureau for the 2020 Census Dajani et al. (2017).

In this vein, a recent line of work Karwa and Vadhan (2018); Kamath et al. (2019a); Bun et al. (2019) gives nearly optimal differentially private algorithms for mean estimation of sub-Gaussian random variables. Roughly speaking, to achieve accuracy α under ε-differential privacy in a d-dimensional setting, one requires $n = \tilde{O}(\frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon})$ samples, a mild cost of privacy over the non-private sample complexity of $O(\frac{d}{\alpha^2})$, except when ε is very small (corresponding to a very high level of privacy). However, these results all depend on the strong assumption that the underlying distribution being sub-Gaussian. Indeed, many sources of data in the real world are known to be heavy-tailed in nature, and thus we require algorithms which are effective even under these looser restrictions. Thus, the core question of this work is

What is the cost of privacy when estimating the mean of heavy-tailed distributions?

We make progress on this question by giving both algorithms and lower bounds for differentially private mean estimation on distributions with bounded k-th moments, for $k \geq 2$. In particular, for univariate distributions, we show that the optimal worst-case sample complexity depends critically on the choice of k, which is qualitatively different from the non-private case.

1.1. Results, Techniques, and Discussion

In this section, we will assume familiarity with some of the most common notions of differential privacy: pure ε-differential privacy, ρ-zero-concentrated differential privacy, and approximate (ε, δ)-differential privacy. In particular, one should know that these are in (strictly) decreasing order of strength, formal definitions appear in Section B.

We first focus on the univariate setting, proving tight upper and lower bounds for estimation subject to bounds on every possible moment.

Theorem 1 (Theorems 10 and 12) *For every $k \geq 2$, $0 < \varepsilon, \alpha < 1$, and $R > 1$, there is an ε-DP algorithm that takes*

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

samples from an arbitrary distribution $\mathcal{D}$ with mean μ such that $\mu \in (-R, R)$ and $\mathbb{E}\left[|\mathcal{D} - \mu|^k\right] \leq 1$ and returns $\hat{\mu}$ such that, with high probability, $|\hat{\mu} - \mu| \leq \alpha$. Moreover, any such ε-DP algorithm requires $n = \Omega(\frac{1}{\alpha^2} + \frac{1}{\varepsilon\alpha^{\frac{k}{k-1}}} + \frac{\log(R)}{\varepsilon})$ samples in the worst case.[1]

Remark 2 *In Theorem 10 we chose to reduce the number of parameters by making only the assumption that $\mathbb{E}\left[|\mathcal{D} - \mu|^k\right] \leq 1$, and bounding the absolute error $|\hat{\mu} - \mu|$. More generally, we can consider a setting where the variance is $\sigma^2 = \mathbb{E}\left[(\mathcal{D} - \mu)^2\right]$ and the k-th moment satisfies $\mathbb{E}\left[|\mathcal{D} - \mu|^k\right] \leq M^k\sigma^k$ for some $M \geq 1$, and we want to bound the normalized error $|\hat{\mu} - \mu|/\sigma$. It is without loss of generality to solve the simplified problem. For example, if the standard deviation is known, then we can renormalize the data by a factor of $M\sigma$, after which the distribution has standard deviation $1/M$ and k-th moment at most 1. Now we can apply our theorem with accuracy $\alpha' = \alpha/M$, in which case we get a sample complexity of $O(M^{k/(k-1)}/\varepsilon\alpha^{k/(k-1)})$.*

Note that, absent privacy constraints, the sample complexity of mean estimation with bounded k-th moments is $n = O(1/\alpha^2)$ samples, for any $k \geq 2$. However, if we require the algorithm to be differentially private, there is a qualitatively different picture in which the cost of privacy

[1] Analogous tight bounds hold for zCDP and (ε, δ)-DP, and these bounds differ only in the dependence on R in the final term. In particular, $\Omega(1/\alpha^2 + 1/\varepsilon\alpha^{k/(k-1)})$ samples are necessary for any of the variants of differential privacy.

decays as we have stronger bounds on the moments of the distribution. Our upper bounds follow a noised and truncated-empirical-mean approach. While this is similar to prior work on private mean estimation Karwa and Vadhan (2018); Kamath et al. (2019a); Cai et al. (2019); Bun and Steinke (2019), we must be more aggressive with our truncation than before. In particular, for the Gaussian case, strong tail bounds allow one to truncate in a rather loose window and not remove any points if the data was actually sampled from a Gaussian. Since we consider distributions with much heavier tails, trying to not discard any points would result in a very wide truncation window, necessitating excessive amounts of noise. Instead, we truncate in a way that balances the two sources of error: bias due to valid points being discarded, and the magnitude of the noise due to the width of the truncation window. To be a bit more precise, our setting of parameters for truncation can be viewed in two different ways: either we truncate so that (in expectation) $1/\varepsilon$ points are removed, and we require n to be large enough to guarantee accuracy, or we truncate so that $\alpha^{k/(k-1)}$ probability mass is removed, and we require n to be large enough to guarantee privacy. These two perspectives on truncation are equivalent when n is at the critical value that makes up our sample complexity.

Our lower bound is proved via hypothesis testing. We demonstrate that two distributions that satisfy the conditions and are indistinguishable with fewer than the prescribed number of samples. Due to an equivalence between pure and approximate differential privacy in this setting, our lower bounds hold for the most permissive privacy notion of (ε, δ)-DP, even for rather large values of δ.

Turning to the multivariate setting, we provide separate algorithms for concentrated and pure differential privacy, both of which come at a multiplicative cost of $O(d)$ in comparison to the univariate setting. We state the concentrated DP result first.

Theorem 3 (Theorem 38) *For every d, $k \geq 2$, $\varepsilon, \alpha > 0$, and $R > 1$, there is a polynomial-time $\frac{\varepsilon^2}{2}$-zCDP algorithm that takes*

$$n \geq O\left(\frac{d}{\alpha^2} + \frac{d}{\varepsilon \alpha^{\frac{k}{k-1}}} + \frac{\sqrt{d \log(R)} \log(d)}{\varepsilon}\right)$$

samples from an arbitrary distribution $\mathcal{D}$ on $\mathbb{R}^d$ with mean vector μ such that $\|\mu\|_2 \leq R$ and bounded k-th moments $\sup_{v \in \mathbb{S}^{d-1}} \mathbb{E}\left[|\langle v, \mathcal{D} - \mu \rangle|^k\right] \leq 1$ and returns $\hat{\mu}$ such that, with high probability, $\|\hat{\mu} - \mu\|_2 \leq \alpha$.

Similar to the univariate case, we rely upon a noised and truncated empirical mean (with truncation to an ℓ_2 ball). The computations required to bound the bias of the truncated estimator are somewhat more involved and technical than the univariate case.

Our pure-DP multivariate mean estimator has the following guarantees.

Theorem 4 (Theorem 20) *For every d, $k \geq 2$, $\varepsilon, \alpha > 0$, and $R > 1$, there is a (possibly exponential time) pure ε-DP algorithm that takes*

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

samples from an arbitrary distribution $\mathcal{D}$ on $\mathbb{R}^d$ with mean vector μ such that $\|\mu\|_2 \leq R$ and bounded k-th moments $\sup_{v \in \mathbb{S}^{d-1}} \mathbb{E}\left[|\langle v, \mathcal{D} - \mu \rangle|^k\right] \leq 1$ and returns $\hat{\mu}$ such that, with high probability, $\|\hat{\mu} - \mu\|_2 \leq \alpha$.

We discuss the similarities and differences between Theorems 3 and 4. First, we note that the first two terms in the sample complexity are identical, similar to the multivariate Gaussian case, where distribution estimation under pure and concentrated DP share the same sample complexity Kamath et al. (2019a); Bun et al. (2019). This is contrary to certain problems in private mean estimation, where an $O(\sqrt{d})$ factor often separates the two complexities Bun et al. (2014); Steinke and Ullman (2015); Dwork et al. (2015). It appears that these qualitative gaps may or may not arise depending on the choice of norm and the assumptions we put on the underlying distribution (see Section 1.1.4 of Kamath et al. (2019a) and Remark 6.4 of Bun et al. (2019)) for more discussion. We point out that the estimator of Theorem 4 is not computationally efficient, while the estimator of Theorem 3 is. However, even for the well structured Gaussian case, no computationally-efficient algorithm is known under pure DP Kamath et al. (2019a); Bun et al. (2019).

Technically, our multivariate-pure-DP algorithm is quite different from our other algorithms. It bears significant resemblance to approaches based on applying the "Scheffé estimator" to a cover for the family of distributions Yatracos (1985); Devroye and Lugosi (1996, 1997, 2001). These approaches reduce an estimation problem to a series of pairwise comparisons (i.e., hypothesis tests) between elements of the cover. However, outside of density estimation, we are not aware of any other problems in this space which are solved by applying pairwise comparisons to elements of a cover, as they instead often appeal to uniform convergence arguments. In particular, we believe our algorithm is the first to use this approach for the problem of mean estimation. We cover the space of candidate means, and perform a series of tests of the form "Which of these two candidates is a better fit for the distribution's mean?" As mentioned before, there are often gaps between our understanding of multivariate estimation under pure and concentrated DP, and the primary reason is that the Laplace and Gaussian mechanisms have sensitivities based on the ℓ_1 and ℓ_2 norms, respectively. We avoid paying the extra $O(\sqrt{d})$ which often arises in the multivariate setting by reducing to a series of *univariate* problems—given two candidate means, we can project the problem onto the line which connects the two. By choosing whichever candidate wins all of its comparisons, we can get an accurate estimate for the mean overall. Crucially, using techniques from Bun et al. (2019), we only pay logarithmically in the size of the cover.

Remark 5 *In Theorems 38 and 20 we use the standard formulation of bounded moments for distributions on $\mathbb{R}^d$, which means that for every direction v, the univariate distribution obtained by projecting onto v has bounded k-th moment. Although this is the standard definition of bounded moments for multivariate distributions, one could potentially consider other classes of heavy-tailed distributions, for example, one which bounds $\mathbb{E}\left[\|\mathcal{D}-\mu\|_2^k\right]$.*

Finally, we prove some lower bounds for multivariate private mean estimation.

Theorem 6 (Theorem 40) *Any pure ε-DP algorithm that takes samples from an arbitrary distribution on $\mathbb{R}^d$ with bounded 2nd moments and returns $\hat{\mu}$ such that $\|\hat{\mu}-\mu\|_2 \leq \alpha$ requires $n = \Omega\left(\frac{d}{\varepsilon\alpha^2}\right)$ samples from $\mathcal{D}$ in the worst case.*

In addition to showing that Theorem 4 is optimal for the case of $k = 2$, it specifically shows a qualitative difference between distributions with bounded 2nd moment and bounded k-th moment for $k > 2$. In the latter case the additional sample complexity due to privacy can be of lower order than the sample complexity without privacy, whereas for $k = 2$ it cannot be unless ε is a constant.

1.2. Paper Organization

Additional discussion of related work appears in Appendix A. A review of standard definitions in differential privacy is in Appendix B. In Section 2 we present upper and lower bounds for univariate estimation with pure DP. Variants for zCDP and approximate DP appear in Appendix D. In Section 3 we present our algorithm for multivariate estimation with pure differential privacy. Due to space restrictions, our computationally efficient algorithm with zCDP appears in Appendix F. Our lower bound for estimation in high dimensions appears in Appendix G. Relevant concentration inequalities appear in Appendix H.

2. Estimating in One Dimension

In this section, we discuss estimating the mean of a distribution whilst ensuring pure DP. Obtaining CDP and approximate DP algorithms for this is trivial, once we have the algorithm for pure DP, as shall be discussed towards the end of the upper bounds section. Finally, we show that our upper bounds are optimal.

2.1. Technical Lemmata

Here, we lay out the two main technical lemmata that we would use to prove our main results for the section. We defer their proofs to Appendix C for space. The first lemma says that if we truncate the distribution to within a large interval that is centered close to the mean, then the mean of this truncated distribution will be close to the original mean.

Lemma 7 *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean μ, and k^{th} moment bounded by 1. Let $\rho \in \mathbb{R}$, $0 < \tau < \frac{1}{16}$, and $\xi = \frac{C}{\tau^{\frac{1}{k-1}}}$ for a constant $C \geq 6$. Let $X \sim \mathcal{D}$, and Z be the following random variable.*

$$Z = \begin{cases} \rho - \xi & \text{if } X < \rho - \xi \\ X & \text{if } \rho - \xi \leq X \leq \rho + \xi \\ \rho + \xi & \text{if } X > \rho + \xi \end{cases}$$

If $|\mu - \rho| \leq \frac{\xi}{2}$, then $|\mu - \mathbb{E}[Z]| \leq \tau$.

The next lemma says that if we take a large number of samples from any distribution over $\mathbb{R}$, whose k^{th} moment is bounded by 1, then with high probability, the empirical mean of the samples lies close to the mean of the distribution.

Lemma 8 *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean μ and k^{th} moment bounded by 1. Suppose $(X_1, \dots, X_n)$ are samples from $\mathcal{D}$, where $n \geq O\left(\frac{1}{\alpha^2}\right)$. Then with probability at least 0.9,*

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| \leq \alpha.$$

2.2. The Algorithm

Here, we give an ε-DP algorithm to estimate the mean. The main algorithm consists of two parts: limiting the data to a reasonable range so as to achieve privacy, and to limit the amount of noise added for it to get optimal accuracy; and mean estimation in a differentially private way. We analyze the two separately.

2.2.1. PRIVATE RANGE ESTIMATION

Here, we explore the first part of the algorithm, that is, limiting the range of the data privately. We do that in a way similar to that of Karwa and Vadhan (2018). To summarize, we use differentially private histograms (Lemma 29) to find the bucket with the largest number of points. Due to certain moments of the distribution being bounded, the points tend to concentrate around the mean. Therefore, the above bucket would be the one closest to the mean, and by extending the size of the bucket a little, we could get an interval that contains a large number of points along with the mean. We show that the range is large enough that the mean of the distribution truncated to that interval will not be too far from the original mean.

Algorithm 1: Pure DP Range Estimator $\mathrm{PDPRE}_{\varepsilon,\alpha,R}(X)$

```
Input: Samples $X_1, \ldots, X_n \in \mathbb{R}$. Parameters $\varepsilon, \alpha, R > 1$.
Output: $[a, b] \in \mathbb{R}$.

Set parameters: $r \leftarrow 10/\alpha^{\frac{1}{k-1}}$
// Estimate range
Divide $[-R-2r, R+2r]$ into buckets: $[-R-2r, -R), \ldots, [-2r, 0), [0, 2r), \ldots, [R, R+2r]$
Run Pure DP Histogram for $X$ over the above buckets
Let $[a, b]$ be the bucket that has the maximum number of points
Let $I \leftarrow [a-2r, b+2r]$
return $I$
```

Theorem 9 *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean $\mu \in [-R, R]$ and k^{th} moment bounded by 1. Then for all $\varepsilon > 0$ and $0 < \alpha < \frac{1}{16}$, there exists an ε-DP algorithm that takes*

$$n \geq O\left(\frac{1}{\alpha} + \frac{\log(R\alpha)}{\varepsilon}\right)$$

samples from $\mathcal{D}$, and outputs $I = [a, b] \subset \mathbb{R}$, such that with probability at least 0.9, the following conditions all hold:

1. $b - a \in \Theta\left(\frac{1}{\alpha^{\frac{1}{k-1}}}\right)$.
2. *At most αn samples lie outside I.*
3. *$\mu \in I$ and $b - \mu, \mu - a \geq \frac{10}{\alpha^{\frac{1}{k-1}}}$.*

Proof We separate the privacy and accuracy proofs for Algorithm 1 for clarity.
Privacy:
Privacy follows from Lemma 29 and post-processing of the private output of private histograms (Lemma 24).
Accuracy:
The first part follows because the intervals are constructed to have length $6r \in \Theta\left(\frac{1}{\alpha^{\frac{1}{k-1}}}\right)$.

Let $(X_1, \dots, X_n)$ be independent samples from $\mathcal{D}$. We know from Lemma 42 that,

$$\underset{X \sim \mathcal{D}}{\mathbb{P}}\left[|X-\mu| > \frac{10}{\alpha^{\frac{1}{k-1}}}\right] \leq \underset{X \sim \mathcal{D}}{\mathbb{P}}\left[|X-\mu| > \frac{10}{\alpha^{\frac{1}{k}}}\right] \leq \frac{\alpha}{10^k}.$$

Using Lemma 44, we have, $\mathbb{P}[|\{i : X_i \notin [\mu - r, \mu + r]\}| > \alpha n] < 0.05$, because $n \geq O(1/\alpha)$. Therefore, there has to be a bucket that contains at least $0.5(1-\alpha)n \geq \frac{n}{4}$ points from the dataset, which implies that the bucket containing the maximum number of points has to have at least $\frac{n}{4}$ points. Now, from Lemma 29, we know that the noise added to any bucket cannot exceed $\frac{n}{16}$. Therefore, the noisy value for the largest bucket has to be at least $\frac{3n}{16}$. Since, all these points lie in a single bucket, and include points that are not in the tail of the distribution, the mean lies in either the same bucket, or in an adjacent bucket because the distance from the mean is at most r. Hence, the constructed interval of length $6r$ contains the mean and at least $1-\alpha$ fraction of the points.
From the above, since the mean is at most r far from at least one of a and b, the end points of I must be at least r far from μ. ■

2.2.2. Private Mean Estimation

Now, we detail the second part of the algorithm, that is, private mean estimation. In the previous step, we ensured that the range of the data is large enough that the mean of the truncated distribution would not be too far from the original mean. Here, we show that this range is small enough, that the noise we add to guarantee privacy is not too large. This would imply that the whole algorithm, whilst being differentially private, would also be accurate without adding a large overhead in the sample complexity.

Theorem 10 *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean $\mu \in [-R, R]$ and k^{th} moment bounded by 1. Then for all $\varepsilon, \alpha, \beta > 0$, there exists an ε-DP algorithm that takes*

$$n \geq O\left(\frac{\log(1/\beta)}{\alpha^2} + \frac{\log(1/\beta)}{\varepsilon \alpha^{\frac{k}{k-1}}} + \frac{\log(R)\log(1/\beta)}{\varepsilon}\right)$$

samples from $\mathcal{D}$, and outputs $\widehat{\mu} \in \mathbb{R}$, such that with probability at least $1-\beta$, $|\mu - \widehat{\mu}| \leq \alpha$.

Proof We first prove the privacy guarantee of Algorithm 2, then move on to accuracy.
Privacy:
The step of finding a good interval I_i is $\frac{\varepsilon}{2}$-DP by Theorem 9. Then the step of estimating the mean by adding Laplace noise is $\frac{\varepsilon}{2}$-DP by Lemma 27. Therefore, by Lemma 25, each iteration is ε-DP. Since, we use each disjoint part of the dataset only once in the entire loop, ε-DP still holds. Finally, we operate on private outputs to find the median, therefore, by Lemma 24, the algorithm is ε-DP.
Accuracy:
We fix an iteration i, and discuss the accuracy of that step. The accuracy of the rest of the iterations would be guaranteed in the same way, since all iterations are independent.

Proposition 11 *Fix $1 \leq i \leq m$. Then in iteration i, if*

$$|Y^i| \geq O\left(\frac{1}{\alpha^2} + \frac{1}{\varepsilon \alpha^{\frac{k}{k-1}}} + \frac{\log(R)}{\varepsilon}\right),$$

then with probability at least 0.7, $|\mu_i - \mu| \leq \alpha$.

```
Algorithm 2: Pure DP 1-Dimensional Mean Estimator PDPODME$_{\varepsilon,\alpha,R}(X)$
Input: Samples $X_1, \dots, X_{2n} \in \mathbb{R}$. Parameters $\varepsilon, \alpha, R > 0$.
Output: $\widehat{\mu} \in \mathbb{R}$.

Set parameters: $m \leftarrow 200\log(2/\beta)$    $Z \leftarrow (X_1, \dots, X_n)$    $W \leftarrow (X_{n+1}, \dots, X_{2n})$

// Partition the dataset, and estimate mean on each subset
for $i \leftarrow 1, \dots, m$ do
    Let $Y^i \leftarrow (W_{(i-1)\cdot\frac{n}{k}+1}, \dots, W_{i\cdot\frac{n}{k}-1})$ and $Z^i \leftarrow (Z_{(i-1)\cdot\frac{n}{k}+1}, \dots, Z_{i\cdot\frac{n}{k}-1})$

    // Find small interval containing the mean and large fraction
       of points
    $I_i \leftarrow \mathrm{PDPRE}_{\varepsilon,\alpha,\mathbb{R}}(Z_i)$   and   $r_i \leftarrow |I_i|$

    // Truncate to within the small interval above
    for $y \in Y^i$ do
        if $x \notin I_i$ then
            Set $x$ to be the nearest end-point of $I_i$
        end
    end

    // Estimate the mean
    $\widehat{\mu_i} \leftarrow \frac{1}{n} \sum_{y \in Y^i} y + \mathrm{Lap}\left(\frac{mr_i}{\varepsilon n}\right)$
end

// Median of means to select a good mean with high probability
$\widehat{\mu} \leftarrow \mathrm{Median}(\mu_1, \dots, \mu_m)$

return $\widehat{\mu}$
```

Proof We know from Theorem 9 that with probability at least 0.9, $\mu \in I_i$, such that if $I_i = [a, b]$, then $\mu - a, b - \mu \in \Omega\left(\frac{1}{\alpha^{\frac{1}{k-1}}}\right)$. Therefore, from Lemma 7, the mean of the truncated distribution (let's call it μ_i') will be at most $\frac{\alpha}{3}$ from μ. But from Lemma 8, we know that $|\mu_i - \mu_i'| \leq \frac{\alpha}{3}$ with probability at least 0.9. Finally, from Lemma 46, with probability at least 0.9, the Laplace noise added is at most $\frac{\alpha}{3}$ because we have at least $O\left(\frac{1}{\varepsilon\alpha^{\frac{k}{k-1}}}\right)$ samples. Therefore, by triangle inequality, $|\mu_i - \mu| \leq \alpha$, and by the union bound, this happens with probability at least 0.7. ■

Now, by the claim above, and using Lemma 45, more than $\frac{m}{2}$ iterations should yield μ_i that are α close to μ, which happens with probability at least $1 - \beta$ (because $m \geq O(\log(1/\beta))$). Therefore, the median, that is, $\widehat{\mu}$ is at most α far from μ with probability at least $1 - \beta$. ■

In Appendix C, we prove the following matching lower bound.

Theorem 12 *Let $\mathcal{D}$ be a distribution with mean $\mu \in (-1,1)$ and k^{th} moment bounded by 1. Then given $\varepsilon, \delta, \alpha > 0$, any (ε,δ)-DP algorithm takes $n \geq \Omega\left(\frac{1}{\varepsilon\alpha^{\frac{k}{k-1}}}\right)$ samples to estimate μ to within α absolute error with constant probability.*

3. Estimating in High Dimensions with Pure DP

We prove an upper bound for mean estimation in case of high-dimensional distributions with bounded k^{th} moment, whilst having pure DP guarantee for our algorithm. It involves creating a cover over $[R,R]^d$, and using the Exponential Mechanism (Lemma 30) for selecting a point that would be a good estimate for the mean with high probability. Note that while this algorithm achieves sample complexity that is linear in the dimension, it is computationally inefficient.

3.1. Technical Lemma

We start by stating two lemmata that would be used in the proof of accuracy of our proposed algorithm, but prove them in Appendix E. Unlike Lemma 7, the first lemma says that if we truncate a one-dimensional distribution with bounded k^{th} moment around a point that is far from the mean, then the mean of this truncated distribution would be far from the said point.

Lemma 13 *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean μ, and k^{th} moment bounded by 1. Let $\rho \in \mathbb{R}$, $0 < \tau < \frac{1}{16}$, and $\xi = \frac{C}{\tau^{\frac{1}{k-1}}}$ for a universal constant C. Let $X \sim \mathcal{D}$, and Z be the following random variable.*

$$Z = \begin{cases} \rho - \xi & \text{if } X < \rho - \xi \\ X & \text{if } \rho - \xi \leq X \leq \rho + \xi \\ \rho + \xi & \text{if } X > \rho + \xi \end{cases}$$

If $\rho > \mu + \frac{\xi}{2}$, then the following holds.

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

If $\rho < \mu - \frac{\xi}{2}$, then the following holds.

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

The guarantees of the next lemma are similar to Lemma 8, except that this one promises a much higher correctness probability. It is just the well-known median of means estimator, which is adapted for the case of distributions with bounded k^{th} moment.

Lemma 14 (Median of Means) *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean μ and k^{th} moment bounded by 1. For $0 < \beta < 1$, and $\alpha > 0$, suppose $(X_1, \ldots, X_n)$ are independent samples from $\mathcal{D}$, such that*

$$n \geq O\left(\frac{\log(1/\beta)}{\alpha^2}\right).$$

Let $m \geq 200\log(2/\beta)$. For $i = 1, \dots, m$, suppose $Y^i = \left(X_{(i-1)m+1}, \dots, X_{(i-1)m+\frac{n}{m}}\right)$, and μ_i is the empirical mean of Y^i. Define $\overline{\mu} = \mathrm{Median}(\mu_1, \dots, \mu_m)$. Then with probability at least $1-\beta$,

$$|\overline{\mu} - \mu| \leq \alpha.$$

3.2. The Algorithm

Our algorithm for estimating the mean with pure differential privacy is a so-called, "cover-based algorithm." It creates a net of points, some of which could be used to get a good approximation for the mean, then privately choses one of those points with high probability. This is done by assigning a "score" to each point in the net, which depends on the dataset, such that it ensures that the points which are good, have significantly higher scores than the bad points. Privacy comes in for the part of selecting a point because the score is directly linked with the dataset. So we use the Exponential Mechanism (Lemma 30) for this purpose.

This framework is reminiscent of the classic approach of density estimation via Scheffé estimators (see, e.g., Devroye and Lugosi (2001)), and its private analogue in Bun et al. (2019). In particular, in a similar way, we choose from the cover by setting up several pairwise comparisons, and privatize using the exponential mechanism in the same way as Bun et al. (2019). This is where the similarities end: the method for performing a comparison between two elements is quite different, and the application is novel (mean estimation versus density estimation). We adopt this style of pairwise comparisons to reduce the problem from d dimensions to 2 dimensions: indeed, certain aspects of pure differential privacy are not well understood in high-dimensional settings, so this provides a new tool to get around this roadblock. To the best of our knowledge, we are the first to use pairwise comparisons for a statistical estimation task besides general density estimation. Most cover-based arguments for other tasks instead appeal to uniform convergence, which is not clear how to apply in this setting when we must preserve privacy.

The first task is to come up with a good SCORE function. It means that it should satisfy two properties. First: the points $O(\alpha)$ close to μ should have very high scores, but the ones further than that must have very low scores. Second: the function should have low sensitivity so that we don't end up selecting a point with low utility (SCORE). We create a required function, which is based on games or "matches" between pairs of points, as defined below.

Definition 15 (Match between Two Points) *Let $X^1, \dots, X^m \in \mathbb{R}^{n \times d}$ be datasets, X be their concatenation, and p, q be points in $\mathbb{R}^d$, and $\xi > 0$. Suppose $Y^1, \dots, Y^m$ are the respective datasets after projecting their points on to the line $p-q$ and truncating to within $B_\xi(p)$, and $\mu_1, \dots, \mu_m$ are the respective empirical means of Y_i's. Let $\mu' = \mathrm{Median}(\mu_1, \dots, \mu_m)$. Then we define the function $\mathrm{Match}_{X,\xi} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as follows.*

$$\mathrm{Match}_{X,\xi}(p,q) = \begin{cases} \mathsf{Tie} & \text{if } \|p-q\|_2 \leq 20\alpha \\ \mathsf{Win} & \text{if } \|p-\mu'\|_2 < \|q-\mu'\|_2 \\ \mathsf{Lose} & \text{if } \|p-\mu'\|_2 \geq \|q-\mu'\|_2 \end{cases}$$

Note that the above definition is not symmetric.

Definition 16 (SCORE of a Point) *Let $X^1, \dots, X^m \in \mathbb{R}^{n \times d}$ be datasets, X be their concatenation, and p be a point in $\mathbb{R}^d$, and $\alpha, \xi > 0$. We define $\mathrm{SCORE}_{X,\xi,D,\alpha}(p)$ with respect to a domain*

$D \subset \mathbb{R}^d$ of a point p to be the minimum number of points of X that need to be changed to get a dataset $\overline{X}$ so that there exists $q \in D$, such that $\mathrm{Match}_{\overline{X},\xi}(p,q) = \mathsf{Lose}$. If for all $q \in D \setminus \{p\}$ and all $Y \in \mathbb{R}^{nm \times d}$, $\mathrm{Match}_{Y,\xi}(p,q) \neq \mathsf{Lose}$, then we define $\textsc{Score}_{Y,\xi,D,\alpha}(p) = n\alpha$.
If the context is clear, we abbreviate the quantity to $\textsc{Score}_X(p)$.

By the above definition, if there already exists a $q \in D$, such that $\mathrm{Match}_{X,\xi}(p,q) = \mathsf{Lose}$, then $\textsc{Score}_X(p) = 0$. Let S be the set as defined in Algorithm 2. We state two properties of the score function, and prove that it satisfies them in Appendix E. We start by showing that the points in S that are close to μ have a high score with high probability.

Algorithm 3: Pure DP High-Dimensional Mean Estimator $\mathrm{PDPHDME}_{\varepsilon,\alpha,R}(X)$

```
Input: Samples $X_1, \dots, X_{2n} \in \mathbb{R}^d$. Parameters $\varepsilon, \alpha, R > 0$.
Output: $\widehat{\mu} \in \mathbb{R}^d$.

Set parameters: $\xi \leftarrow \frac{60}{\alpha^{\frac{1}{k-1}}}$    $Y \leftarrow (X_1, \dots, X_n)$    $Z \leftarrow (X_{n+1}, \dots, X_{2n})$

// Reduce search space
for $i \leftarrow 1, \dots, d$ do
    $c_i \leftarrow \mathrm{PDPODME}_{\frac{\varepsilon}{d},\alpha,R}(Y^i)$
    $I_i \leftarrow [c_i - \alpha, c_i + \alpha]$
    $J_i \leftarrow \{c_i - \alpha, c_i - \alpha + \frac{\alpha}{\sqrt{d}}, \dots, c_i + \alpha - \frac{\alpha}{\sqrt{d}}, c_i + \alpha\}$
end

// Find a good estimate
Let $S \leftarrow J_1 \times \cdots \times J_d$
Compute $\textsc{Score}_{X,\xi,S,\alpha}(p)$ with respect to $Z$ for every $p \in S$
Run Exponential Mechanism w.r.t. $\textsc{Score}$ (sensitivity 1, privacy budget $\varepsilon$) to output $\widehat{\mu} \in S$
return $\widehat{\mu}$
```

Lemma 17 *Let $m \geq 400d\log(8\sqrt{d}/\beta)$ be the same quantity as in Definition 15, and let $S_{\leq} \subset S$, such that for all $p \in S_{\leq}$, $\|p - \mu\|_2 \leq 5\alpha$. If*

$$n \geq O\Big(\frac{m}{\alpha^2}\Big),$$

then

$$\mathbb{P}\Big[\exists p \in S_{\leq},\ st.\ \textsc{Score}_{X,\xi,S,\alpha}(p) < \tfrac{4n\alpha}{5\xi}\Big] \leq \beta.$$

Now, we show that the points in S that are far from μ have a very low score.

Lemma 18 *Let $m \geq 400d\log(8\sqrt{d}/\beta)$ be as in Definition 15, and let $S_{>} \subset S$, such that for all $p \in S_{>}$, $\|p - \mu\|_2 > 20\alpha$. If*

$$n \geq O\Big(\frac{m}{\alpha^2}\Big),$$

then

$$\mathbb{P}[\exists p \in S_{>},\ st.\ \textsc{Score}_{X,\xi,S,\alpha}(p) > 0] \leq \beta.$$

Finally, we state that the SCORE function has low sensitivity, but prove it in Appendix E.

Lemma 19 *The* SCORE *function satisfies the following:*

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

We can now move on to the main theorem of the section.

Theorem 20 *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean $\mu \in B_R(\mathbf{0})$ and k^{th} moment bounded by* 1. *Then for all $\varepsilon, \alpha, \beta > 0$, there exists an ε-DP algorithm that takes*

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

samples from $\mathcal{D}$, and outputs $\widehat{\mu} \in \mathbb{R}^d$, such that with probability at least $1-\beta$,

$$\|\mu - \widehat{\mu}\|_2 \leq \alpha.$$

Proof We again separate the proofs of privacy and accuracy of Algorithm 3.
Privacy:
The first step is ε-DP from Lemma 10, and from Lemma 25. The second step is ε-DP from Lemmata 19 and 30. Therefore, the algorithm is 2ε-DP.
Accuracy:
The first step is meant to reduce the size of the search space. From Lemma 10, we have that for each i, the distance between c_i and the mean along the i^{th} axis is at most α. So, I_i contains the mean with high probability, and is of length 2α by construction.

We know from Lemma 30 that with high probability, the point returned by Exponential Mechanism has a high SCORE. So, it will be enough to argue that with high probability, only the points in S, which are $O(\alpha)$ close to μ, have a high quality score, while the rest have SCORE close to 0. This exactly what we have from Lemmata 17 and 18. Let $\text{OPT}_{\text{SCORE}}(Z)$ be the maximum score of any point in S. Then we know that $\text{OPT}_{\text{SCORE}}(Z) \geq \frac{4n\alpha}{5\xi}$, and that the points that have this score have to be at most 20α far from μ. From Lemma 30, we know that with probability at least $1-\beta$,

$$\begin{aligned}
\text{SCORE}(X, \widehat{\mu}) &\geq \text{OPT}_{\text{SCORE}}(Z) - \frac{2\Delta_{\text{SCORE},1}}{\varepsilon}(\log(|S|) + \log(1/\beta)) \\
&\geq \frac{4n\alpha}{5\xi} - \frac{2}{\varepsilon}\left(d\log\left(4\sqrt{d}\right) + \log(1/\beta)\right) \\
&\geq O\left(n\alpha^{\frac{k}{k-1}}\right). && \text{(Because of our bounds on } n \text{ and } \xi)
\end{aligned}$$

Therefore, we get a point that is at most 20α far from μ. Rescaling α and β by constants, we get the required result. ■

Acknowledgments

GK is supported by a University of Waterloo startup grant. Part of this work was done while GK was supported as a Microsoft Research Fellow, as part of the Simons-Berkeley Research Fellowship program at the Simons Institute for the Theory of Computing, and while visiting Microsoft Research Redmond. VS and JU are supported by NSF grants CCF-1718088, CCF-1750640, CNS-1816028, and CNS-1916020. This work was initiated while all authors were visiting the Simons Institute for the Theory of Computing.

References

Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Sorting with adversarial comparators and application to density estimation. In *Proceedings of the 2014 IEEE International Symposium on Information Theory*, ISIT '14, pages 1682–1686, Washington, DC, USA, 2014. IEEE Computer Society.

Jayadev Acharya, Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Maximum selection and sorting with adversarial comparators. *Journal of Machine Learning Research*, 19(1):2427–2457, 2018a.

Jayadev Acharya, Gautam Kamath, Ziteng Sun, and Huanyu Zhang. Inspectre: Privately estimating the unseen. In *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pages 30–39. JMLR, Inc., 2018b.

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private testing of identity and closeness of discrete distributions. In *Advances in Neural Information Processing Systems 31*, NeurIPS '18, pages 6878–6891. Curran Associates, Inc., 2018c.

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

Victor Balcer and Salil Vadhan. Differential privacy on finite computers. *Journal of Privacy and Confidentiality*, 9(2), Sep. 2019. doi: 10.29012/jpc.679. URL https://journalprivacyconfidentiality.org/index.php/jpc/article/view/679.

Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. *arXiv preprint arXiv:2006.06618*, 2020.

Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: The SuLQ framework. In *Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '05, pages 128–138, New York, NY, USA, 2005. ACM.

Olivier Bousquet, Daniel M. Kane, and Shay Moran. The optimal approximation factor in density estimation. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pages 318–341, 2019.

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.

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.

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, New York, NY, USA, 2014. ACM.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 634–649, Washington, DC, USA, 2015. IEEE Computer Society.

Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. In *Proceedings of the 7th Conference on Innovations in Theoretical Computer Science*, ITCS '16, pages 369–380, New York, NY, USA, 2016. ACM.

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.

T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*, 2019.

Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. Fast mean estimation with sub-Gaussian rates. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pages 786–806, 2019.

Aref N. Dajani, Amy D. Lauger, Phyllis E. Singer, Daniel Kifer, Jerome P. Reiter, Ashwin Machanavajjhala, Simson L. Garfinkel, Scot A. Dahl, Matthew Graham, Vishesh Karwa, Hang Kim, Philip Lelerc, Ian M. Schmutte, William N. Sexton, Lars Vilhuber, and John M. Abowd. The modernization of statistical disclosure limitation at the U.S. census bureau, 2017. Presented at the September 2017 meeting of the Census Scientific Advisory Committee.

Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians. In *Proceedings of the 27th Annual Conference on Learning Theory*, COLT '14, pages 1183–1213, 2014.

Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning Poisson binomial distributions. In *Proceedings of the 44th Annual ACM Symposium on the Theory of Computing*, STOC '12, pages 709–728, New York, NY, USA, 2012. ACM.

Jules Depersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*, 2019.

Luc Devroye and Gábor Lugosi. A universally acceptable smoothing factor for kernel density estimation. *The Annals of Statistics*, 24(6):2499–2512, 1996.

Luc Devroye and Gábor Lugosi. Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes. *The Annals of Statistics*, 25(6):2626–2637, 1997.

Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer, 2001.

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.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 655–664, Washington, DC, USA, 2016. IEEE Computer Society.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 999–1008. JMLR, Inc., 2017.

Differential Privacy Team, Apple. Learning with privacy at scale. https://machinelearning.apple.com/docs/learning-with-privacy-at-scale/appledifferentialprivacysystem.pdf, December 2017.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems 30*, NIPS '17, pages 3571–3580. Curran Associates, Inc., 2017.

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

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '13, pages 429–438, Washington, DC, USA, 2013. IEEE Computer Society.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 2017.

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.

Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '10, pages 51–60, Washington, DC, USA, 2010. IEEE Computer Society.

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, Washington, DC, USA, 2015. IEEE Computer Society.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM Conference on Computer and Communications Security*, CCS '14, pages 1054–1067, New York, NY, USA, 2014. ACM.

Marco Gaboardi, Ryan Rogers, and Or Sheffet. Locally private confidence intervals: Z-test and tight confidence intervals. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, AISTATS '19, pages 2545–2554. JMLR, Inc., 2019.

Samuel B. Hopkins. Sub-Gaussian mean estimation in polynomial time. *arXiv preprint arXiv:1809.07425*, 2018.

Matthew Joseph, Janardhan Kulkarni, Jieming Mao, and Zhiwei Steven Wu. Locally private Gaussian estimation. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 2980–2989. Curran Associates, Inc., 2019.

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

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, 2019a.

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., 2019b.

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.

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

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019b.

Satyaki Mahalanabis and Daniel Stefankovic. Density estimation in linear time. In *Proceedings of the 21st Annual Conference on Learning Theory*, COLT '08, pages 503–512, 2008.

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, Washington, DC, USA, 2007. IEEE Computer Society.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th Annual ACM Symposium on the Theory of Computing*, STOC '07, pages 75–84, New York, NY, USA, 2007. ACM.

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing*, STOC '11, pages 813–822, New York, NY, USA, 2011. ACM.

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.

Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *The Journal of Privacy and Confidentiality*, 7(2):3–22, 2017.

Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical Gaussian mixtures. In *Advances in Neural Information Processing Systems 27*, NIPS '14, pages 1395–1403. Curran Associates, Inc., 2014.

Salil Vadhan. The complexity of differential privacy. In Yehuda Lindell, editor, *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*, chapter 7, pages 347–450. Springer International Publishing AG, Cham, Switzerland, 2017.

Yannis G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13(2):768–774, 1985.

Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics. *arXiv preprint arXiv:1909.08755*, 2019.

Appendix A. Related Work

The most closely related works to ours are Blum et al. (2005); Bun et al. (2014); Steinke and Ullman (2015); Dwork et al. (2015); Steinke and Ullman (2017); Karwa and Vadhan (2018); Kamath et al. (2019a); Cai et al. (2019); Bun and Steinke (2019); Avella-Medina and Brunel (2019); Du et al. (2020); Biswas et al. (2020), which study differentially private estimation of the mean of a distribution. Some of these focus on restricted cases, such as product distributions or sub-Gaussians, which we generalize by making weaker moment-based assumptions. Some instead study more general cases, including unrestricted distributions over the hypercube – by making assumptions on the moments of the generating distributions, we are able to get better sample complexities. The work of Bun and Steinke Bun and Steinke (2019) explicitly studies mean estimation of distributions with bounded second moment, but their sample complexity can be roughly stated as $O(1/\alpha^2\varepsilon^2)$, whereas we prove a tight bound of $\Theta(1/\alpha^2\varepsilon)$. Furthermore, we go beyond second-moment assumptions, and show a hierarchy of sample complexities based on the number of moments which are bounded. Some very recent works Du et al. (2020); Biswas et al. (2020) focus on designing practical tools for private estimation of mean and covariance, in univariate and multivariate settings.

Among other problems, Duchi, Jordan, and Wainwright Duchi et al. (2013, 2017) study univariate mean estimation with moment bounds under the stricter constraint of *local* differential privacy. Our univariate results and techniques are similar to theirs: morally the same algorithm and lower-bound construction works in both the local and central model. Translating their results to compare to ours, they show that the sample complexity of mean estimation in the local model is $O(1/\alpha^{\frac{2k}{k-1}}\varepsilon^2)$, the square of the "second term" in our sample complexity for the central model. However, their investigation is limited to the univariate setting, while we provide new algorithms and lower bounds for the multivariate setting as well. There has also been some work on locally private mean estimation in the Gaussian case Gaboardi et al. (2019); Joseph et al. (2019).

This is just a small sample of work in differentially private distribution estimation, and there has been much study into learning distributions beyond mean estimation. These are sometimes (but not always) equivalent problems – for instance, learning the mean of a Gaussian distribution with known covariance is equivalent to learning the distribution in total variation distance. Diakonikolas, Hardt, and Schmidt Diakonikolas et al. (2015) gave algorithms for learning structured univariate distributions. Privately learning mixtures of Gaussians was considered in Nissim et al. (2007); Kamath et al. (2019b). Bun, Nissim, Stemmer, and Vadhan Bun et al. (2015) give an algorithm for learning distributions in Kolmogorov distance. Acharya, Kamath, Sun, and Zhang Acharya et al. (2018b) focus on estimating properties of a distribution, such as the entropy or support size. Smith Smith (2011) gives an algorithm which allows one to estimate asymptotically normal statistics with optimal convergence rates, but no finite sample complexity guarantees. Bun, Kamath, Steinke, and Wu Bun et al. (2019) give general tools for private hypothesis selection and apply this to learning many distribution classes of interest. For further coverage of differentially private statistics, see Kamath and Ullman (2020).

In the non-private setting, there has recently been significant work in mean estimation of distributions with bounded second moments, in the "high probability" regime. That is, we wish to estimate the mean of a distribution with probability $1-\beta$, where $\beta > 0$ might be very small. While the empirical mean is effective in the sub-Gaussian case, more advanced techniques are necessary to achieve the right dependence on $1/\beta$ when we only have a bound on the second moment. A recent series of papers has focused on identifying effective methods and making them computationally efficient Lugosi and Mendelson (2019a); Hopkins (2018); Cherapanamjeri et al. (2019); Depersin and Lecué (2019); Lugosi and Mendelson (2019b). This high-probability consideration is not the focus of the present work, though we note that, at worst, our estimators incur a multiplicative factor of $\log(1/\beta)$ in achieving this guarantee. We consider determining the correct dependence on the failure probability with privacy constraints an interesting direction for future study.

Our work bears a significant resemblance to a line on hypothesis selection, reducing to pairwise comparisons using the Scheffé estimator. This style of approach was pioneered by Yatracos Yatracos (1985), and refined in subsequent work by Devroye and Lugosi Devroye and Lugosi (1996, 1997, 2001). After this, additional considerations have been taken into account, such as computation, approximation factor, robustness, and more Mahalanabis and Stefankovic (2008); Daskalakis et al. (2012); Daskalakis and Kamath (2014); Suresh et al. (2014); Acharya et al. (2014); Diakonikolas et al. (2016); Acharya et al. (2018a); Bousquet et al. (2019); Bun et al. (2019). As mentioned before, to the best of our knowledge, we are the first to apply this pairwise-comparison-based approach combined with a net-based argument for a problem besides density estimation.

Appendix B. Preliminaries

We formally state what it means for a distribution to have its k^{th} moment bounded.

Definition 21 *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ. We say that for $k \geq 2$, the k^{th} moment of $\mathcal{D}$ is bounded by M, if for every unit vector $v \in \mathbb{S}^{d-1}$,*

$$\mathbb{E}\left[|\langle X-\mu, v\rangle|^k\right] \leq M.$$

Also, we define $B_r(p) \subset \mathbb{R}^d$ to be the ball of radius $r > 0$ centered at $p \in \mathbb{R}^d$.

B.1. Privacy Preliminaries

Definition 22 (Differential Privacy (DP) Dwork et al. (2006)) *A randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies (ε, δ)-differential privacy ((ε, δ)-DP) if for every pair of neighboring datasets $X, X' \in \mathcal{X}^n$ (i.e., datasets that differ in exactly one entry),*

$$\forall Y \subseteq \mathcal{Y} \quad \mathbb{P}[M(X) \in Y] \leq e^{\varepsilon} \cdot \mathbb{P}\big[M(X') \in Y\big] + \delta.$$

When $\delta = 0$, we say that M satisfies ε-differential privacy or pure differential privacy.

Definition 23 (Concentrated Differential Privacy (zCDP) Bun and Steinke (2016)) *A randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies ρ-zCDP if for every pair of neighboring datasets $X, X' \in \mathcal{X}^n$,*

$$\forall \alpha \in (1, \infty) \quad D_\alpha\big(M(X)||M(X')\big) \le \rho\alpha,$$

where $D_\alpha(M(X)||M(X'))$ is the α-Rényi divergence between $M(X)$ and $M(X')$.[2]

Note that zCDP and DP are on different scales, but are otherwise can be ordered from most-to-least restrictive. Specifically, $(\varepsilon, 0)$-DP implies $\frac{\rho^2}{2}$-zCDP, which implies $(\varepsilon\sqrt{\log(1/\delta)}, \delta)$-DP for every $\delta > 0$ Bun and Steinke (2016).

Both these definitions are closed under post-processing and can be composed with graceful degradation of the privacy parameters.

Lemma 24 (Post Processing Dwork et al. (2006); Bun and Steinke (2016)) *If $M : \mathcal{X}^n \to \mathcal{Y}$ is (ε, δ)-DP, and $P : \mathcal{Y} \to \mathcal{Z}$ is any randomized function, then the algorithm $P \circ M$ is (ε, δ)-DP. Similarly if M is ρ-zCDP then the algorithm $P \circ M$ is ρ-zCDP.*

Lemma 25 (Composition of DP Dwork et al. (2006, 2010); Bun and Steinke (2016)) *If M is an adaptive composition of differentially private algorithms $M_1, \dots, M_T$, then the following all hold:*

1. *If $M_1, \dots, M_T$ are $(\varepsilon_1, \delta_1), \dots, (\varepsilon_T, \delta_T)$-DP then M is (ε, δ)-DP for*

$$\varepsilon = \sum_t \varepsilon_t \quad \textit{and} \quad \delta = \sum_t \delta_t.$$

2. *If $M_1, \dots, M_T$ are $(\varepsilon_0, \delta_1), \dots, (\varepsilon_0, \delta_T)$-DP for some $\varepsilon_0 \le 1$, then for every $\delta_0 > 0$, M is (ε, δ)-DP for*

$$\varepsilon = \varepsilon_0\sqrt{6T\log(1/\delta_0)} \quad \textit{and} \quad \delta = \delta_0 + \sum_t \delta_t$$

3. *If $M_1, \dots, M_T$ are $\rho_1, \dots, \rho_T$-zCDP then M is ρ-zCDP for $\rho = \sum_t \rho_t$.*

B.2. Basic Differentially Private Mechanisms.

We first state standard results on achieving privacy via noise addition proportional to sensitivity Dwork et al. (2006).

Definition 26 (Sensitivity) *Let $f : \mathcal{X}^n \to \mathbb{R}^d$ be a function, its ℓ_1-*sensitivity *and ℓ_2-*sensitivity *are*

$$\Delta_{f,1} = \max_{X \sim X' \in \mathcal{X}^n} \|f(X) - f(X')\|_1 \quad \textit{and} \quad \Delta_{f,2} = \max_{X \sim X' \in \mathcal{X}^n} \|f(X) - f(X')\|_2,$$

respectively. Here, $X \sim X'$ denotes that X and X' are neighboring datasets (i.e., those that differ in exactly one entry).

For functions with bounded ℓ_1-sensitivity, we can achieve ε-DP by adding noise from a Laplace distribution proportional to ℓ_1-sensitivity. For functions taking values in $\mathbb{R}^d$ for large d it is more useful to add noise from a Gaussian distribution proportional to the ℓ_2-sensitivity, to get (ε, δ)-DP and ρ-zCDP.

[2] Given two probability distributions P, Q over Ω, $D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\big(\sum_x P(x)^\alpha Q(x)^{1-\alpha}\big)$.

Lemma 27 (Laplace Mechanism) *Let $f : \mathcal{X}^n \to \mathbb{R}^d$ be a function with ℓ_1-sensitivity $\Delta_{f,1}$. Then the Laplace mechanism*

$$M(X) = f(X) + \mathrm{Lap}\left(\frac{\Delta_{f,1}}{\varepsilon}\right)^{\otimes d}$$

satisfies ε-DP.

Lemma 28 (Gaussian Mechanism) *Let $f : \mathcal{X}^n \to \mathbb{R}^d$ be a function with ℓ_2-sensitivity $\Delta_{f,2}$. Then the Gaussian mechanism*

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

satisfies (ε, δ)-DP. Similarly, the Gaussian mechanism

$$M_f(X) = f(X) + \mathcal{N}\left(0, \left(\frac{\Delta_{f,2}}{\sqrt{2\rho}}\right)^2 \cdot \mathbb{I}_{d\times d}\right)$$

satisfies ρ-zCDP.

Lemma 29 (Private Histograms) *Let $(X_1, \dots, X_n)$ be samples in some data universe U, and let $\Omega = \{h_u\}_{u\subset U}$ be a collection of disjoint histogram buckets over U. Then we have ε-DP, ρ-zCDP, and (ε, δ)-DP histogram algorithms with the following guarantees.*

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

Part 1 follows from Balcer and Vadhan (2019). Part 2 follows trivially by using the Gaussian Mechanism (Lemma 28) instead of the Laplace Mechanism (Lemma 27) in Part 1. Part 3 holds due to Bun et al. (2016); Vadhan (2017).

Finally, we recall the widely used *exponential mechanism*.

Lemma 30 (Exponential Mechanism McSherry and Talwar (2007)) *The exponential mechanism $\mathcal{M}_{\varepsilon,S,\textsc{Score}}(X)$ takes a dataset $X \in \mathcal{X}^n$, computes a score ($\textsc{Score} : \mathcal{X}^n \times S \to \mathbb{R}$) for each $p \in S$ with respect to X, and outputs $p \in S$ with probability proportional to $\exp\left(\frac{\varepsilon\cdot\textsc{Score}(X,p)}{2\cdot\Delta_{\textsc{Score},1}}\right)$, where*

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

It satisfies the following.

1. *$\mathcal{M}$ is ε-differentially private.*
2. *Let $\mathrm{OPT}_{\textsc{Score}}(X) = \max_{p\in S}\{\textsc{Score}(X,p)\}$. Then*

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

Appendix C. Missing Proofs from Section 2

C.1. Proof of Lemma 7

Without loss of generality, we assume that $\rho \geq \mu$, since the argument for the other case is symmetric. Let $a = \rho - \xi$ and $b = \rho + \xi$.

$$|\mu - \mathbb{E}[Z]| \leq |\mathbb{E}[(X-a)\mathbb{1}_{X<a}]| + |\mathbb{E}[(X-b)\mathbb{1}_{X>b}]| \tag{1}$$

Now, we compute the first term on the right hand side. The second term would follow by an identical argument.

$$\begin{aligned}
|\mathbb{E}[(X-a)\mathbb{1}_{X<a}]| &= |\mathbb{E}[(X-\mu-(a-\mu))\mathbb{1}_{X<a}]| \\
&\leq \mathbb{E}[|X-\mu|\mathbb{1}_{X<a}] + (|a-\mu|)\mathbb{E}[\mathbb{1}_{X<a}] \\
&\leq \left(\mathbb{E}\left[|X-\mu|^k\right]\right)^{\frac{1}{k}} (\mathbb{P}[X<a])^{\frac{k-1}{k}} + (|a-\mu|)\mathbb{P}[X<a] \\
&\leq \left(\frac{2}{C}\right)^{k-1}\tau + C\left(\frac{2}{C}\right)^{k}\tau \\
&= 3\left(\frac{2}{C}\right)^{k-1}\tau
\end{aligned}$$

In the above, the first inequality follows from linearity of expectations, triangle inequality, and Lemma 49. The second inequality follows from Lemma 48, and the third inequality follows from the fact that

$$\mathbb{P}[X<a] \leq \mathbb{P}\left[X < \mu - \tfrac{\xi}{2}\right] \leq \left(\frac{2}{C}\right)^{k}\tau^{\frac{k}{k-1}},$$

which holds due to Lemma 42. Similarly, we can bound the second term in (1) as follows:

$$|\mathbb{E}[(X-b)\mathbb{1}_{X>b}]| \leq 4\left(\frac{2}{C}\right)^{k-1}\tau.$$

Substituting these two values in Inequality 1, we get that

$$|\mu - \mathbb{E}[Z]| \leq 7\left(\frac{2}{C}\right)^{k-1}\tau \leq \tau.$$

C.2. Proof of Lemma 8

Using Lemma 49, we know that

$$\mathbb{E}\left[|X-\mu|^2\right] \leq \mathbb{E}\left[|X-\mu|^k\right]^{\frac{2}{k}} \leq 1.$$

Let $Z = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$. Then we have the following.

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

$$
\begin{aligned}
&\le \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}|X_i-\mu|^2\right]\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[|X_i-\mu|^2\right]\\
&\le \frac{1}{n}.
\end{aligned}
$$

Then using Lemma 42, we have

$$\mathbb{P}[|Z-\mu|>\alpha] \le \frac{1}{\sqrt{n}\alpha} \le 0.9.$$

C.3. Proof of Theorem 12

We construct two dstributions that are "close", and show that any (ε,δ)-DP algorithm that distinguishes between them requires a large number of samples.

Proposition 31 *Let $\varepsilon,\delta,\alpha>0$, and $\mathcal{D}_1$, $\mathcal{D}_2$ be two distributions on $\mathbb{R}$ defined as follows.*

$$\mathcal{D}_1 \equiv \mathop{\mathbb{P}}_{X\sim\mathcal{D}_1}[X=0]=1$$

$$\mathcal{D}_2 \equiv \begin{cases} X=0 & \textit{with probability } 1-p\\ X=\tau & \textit{with probability } p\end{cases}$$

Where in the above, $\tau>0$, $p\tau=\alpha$ and $p\le\frac{1}{\alpha^{\frac{k}{k-1}}}$. Then the following holds.

1. $\mathop{\mathbb{E}}_{X\sim\mathcal{D}_2}\left[|X-p\tau|^k\right]\le 1$
2. *Any (ε,δ)-DP algorithm that can distinguish between $\mathcal{D}_1$ and $\mathcal{D}_2$ with constant probability requires at least $\frac{1}{\varepsilon\alpha^{\frac{k}{k-1}}}$ samples.*

Proof For the first part, note that $\mathop{\mathbb{E}}_{X\sim\mathcal{D}_2}[X]=p\tau$. Then we have the following.

$$
\begin{aligned}
\mathop{\mathbb{E}}_{X\sim\mathcal{D}_2}\left[|X-p\tau|^k\right] &= p|\tau-p\tau|^k+(1-p)|p\tau|^k\\
&= p(1-p)\tau^k\left((1-p)^{k-1}+p^k\right)\\
&\le p\tau^k\\
&\le 1. && \text{(Using our restrictions on } p,\tau)
\end{aligned}
$$

Now, for the second part, we know that

$$\left|\mathop{\mathbb{E}}_{X\sim\mathcal{D}_1}[X]-\mathop{\mathbb{E}}_{X\sim\mathcal{D}_2}[X]\right|=\alpha.$$

Suppose we take n samples each from $\mathcal{D}_1$ and $\mathcal{D}_2$. Then by Theorem 11 of Acharya et al. (2018c), we get that

$$\begin{aligned} pn &\in \Omega\left(\frac{1}{\varepsilon}\right) \\ \implies n &\in \Omega\left(\frac{1}{\varepsilon\alpha^{\frac{k}{k-1}}}\right). \end{aligned}$$

We conclude by using the equivalence of pure and approximate DP for testing problems (e.g., Lemma 5 of Acharya et al. (2018c)). ■

Finally, since being able to learn to within α absolute error implies distinguishing two distributions that are at least 2α apart, from the above claim, the lemma holds.

Appendix D. One-Dimensional Range and Mean with CDP and Approximate DP

The CDP equivalent of Algorithm 1 for range-estimation (that we call, ”CDPRE”) could be created by using the CDP version of private histograms as mentioned in Lemma 29. Its approximate DP version (which we call, ”ADPRE”) could be obtained via approximate differentially private histograms as mentioned in the same lemma.

Theorem 32 *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean $\mu \in [-R, R]$ and k^{th} moment bounded by 1. Then for all $\varepsilon, \delta, \rho > 0$ and $0 < \alpha < \frac{1}{16}$, there exist (ε, δ)-DP and ρ-zCDP algorithms that take*

$$n_{(\varepsilon,\delta)} \geq O\left(\frac{1}{\alpha} + \frac{\log(1/\delta)}{\varepsilon}\right)$$

and

$$n_\rho \geq O\left(\frac{1}{\alpha} + \sqrt{\frac{\log(R\alpha)}{\rho}}\right)$$

samples from $\mathcal{D}$ respectively, and output $I = [a, b] \subset \mathbb{R}$, such that with probability at least 0.9, the following hold.

1. *$b - a \in \Theta\left(\frac{1}{\alpha^{\frac{1}{k-1}}}\right)$.*
2. *At most αn samples lie outside I.*
3. *$\mu \in I$ and $b - \mu, \mu - a \geq \frac{10}{\alpha^{\frac{1}{k-1}}}$.*

For mean-estimation, Algorithm 2 could be used to get CDP guarantees by using CDPRE instead of Algorithm 1, and using the Gaussian Mechanism (Lemma 28) instead of the Laplace Mechanism (Lemma 27). We call this algorithm CDPODME. To get approximate DP guarantees, Algorithm 2, with the exception of using ADPRE instead, could be used, and we call this modified algorithm ADPODME.

Theorem 33 *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean $\mu \in [-R, R]$ and k^{th} moment bounded by 1. Then for all $\varepsilon, \delta, \rho, \alpha, \beta > 0$, there exist (ε, δ)-DP and ρ-zCDP algorithms that take*

$$n_{(\varepsilon,\delta)} O\left(\frac{\log(1/\beta)}{\alpha^2} + \frac{\log(1/\beta)}{\varepsilon\alpha^{\frac{k}{k-1}}} + \frac{\log(1/\delta)\log(1/\beta)}{\varepsilon}\right)$$

and

$$n_\rho \geq O\left(\frac{\log(1/\beta)}{\alpha^2} + \frac{\log(1/\beta)}{\sqrt{\rho}\alpha^{\frac{k}{k-1}}} + \frac{\sqrt{\log(R)}\log(1/\beta)}{\sqrt{\rho}}\right)$$

samples from $\mathcal{D}$ respectively, and output $\widehat{\mu} \in \mathbb{R}$, such that with probability at least $1 - \beta$,

$$|\mu - \widehat{\mu}| \leq \alpha.$$

Appendix E. Missing Proofs from Section 3

E.1. Proof of Lemma 13

Without loss of generality, we assume that $\rho \geq \mu$,i since the argument for the other case is symmetric. Let $a = \max\left\{\mu + \frac{\xi}{16}, \rho - \xi\right\}$, $b = \rho + \xi$, and $q = \mathbb{P}[|X - \mu| > |a - \mu|]$. Then the highest value $\mathbb{E}[Z]$ can achieve is when $1 - q$ probability mass is at a and q of the mass is at b. This is because the mass that lies beyond a is q, and b is the highest value that X can take because of truncation. We get the following:

$$\begin{aligned} \mathbb{E}[Z] &\leq (1-q)a + qb \\ &= a + q(b - a). \end{aligned} \tag{2}$$

Now, there are two cases. First, when $a = \mu + \frac{\xi}{16}$, and second, when $a = \rho - \xi$. In the first case, since $\mu + \frac{\xi}{16} \geq \rho - \xi$, it must be the case that $\rho - \mu \leq \frac{17\xi}{16}$. So, we have the following from (2).

$$\begin{aligned} \mathbb{E}[Z] &\leq \mu + \frac{\xi}{16} + q\left(\rho + \xi - \mu - \frac{\xi}{16}\right) \\ &= \mu + \frac{\xi}{16} + q(\rho - \mu) + \frac{15q\xi}{16} \\ &\leq \mu + \frac{\xi}{16} + 2q\xi. \end{aligned}$$

From Lemma 42, we know that

$$q = \mathbb{P}\left[|X - \mu| > \frac{\xi}{16}\right] \leq \left(\frac{16}{\xi}\right)^k = \left(\frac{16}{C}\right)^k \tau^{\frac{k}{k-1}}.$$

This, along with our restrictions on C and τ, gives us,

$$\mathbb{E}[Z] \leq \mu + \frac{\xi}{8}.$$

We now have to show that $\mathbb{E}[Z] \geq \mu - \frac{\xi}{8}$. We have the following:

$$\mathbb{E}[Z] \geq (1-q)(\mu - \frac{\xi}{16}) + q(\rho - \xi)$$

$$
\begin{aligned}
&= \mu - \frac{\xi}{16} + q\left(\frac{\xi}{16} - \xi + \rho - \mu\right)\\
&= \mu - \frac{\xi}{16} - \frac{15q\xi}{16} + q(\rho - \mu)\\
&\geq \mu - \frac{\xi}{8}.
\end{aligned}
$$

For the second case, we have the following from (2):

$$
\begin{aligned}
\mathbb{E}[Z] &\leq \rho - \xi + q(\rho + \xi - \rho + \xi)\\
&= \rho - \xi + 2q\xi.
\end{aligned}
$$

From Lemma 42, we know that

$$
q = \mathbb{P}[|X-\mu| > |\rho - \xi - \mu|] \leq \mathbb{P}\left[|X-\mu| > \frac{\xi}{16}\right] \leq \left(\frac{16}{\xi}\right)^k = \left(\frac{16}{C}\right)^k \tau^{\frac{k}{k-1}}.
$$

This gives us,

$$
\mathbb{E}[Z] \leq \rho - \frac{15\xi}{16}.
$$

Now it is trivial to see that $\mathbb{E}[Z] \geq \rho - \xi$ because the minimum value Z can take is $\rho - \xi$.

E.2. Proof of Lemma 14

Since $\frac{n}{m} \geq n_k$, by Lemma 8, with probability at least 0.9, for a fixed i, $|\mu_i - \mu| \leq \alpha$. Now, because $m \geq 200\log(2/\beta)$, using Lemma 45, we have that with probability at least $1-\beta$, at least $0.8m$ of the μ_i's are at most α far from μ. Therefore, their median has to be at most α far from μ.

E.3. Proof of Lemma 17

Fix a $p \in S_{\leq}$. First, note that p cannot lose to any other point that is at most 5α far from μ. So, it can only lose or win against points that are at least 15α far from μ.

Now, let $q \in S$ be any point that is at least 20α away from p, and let ℓ_q be the line $p - q$. If we project μ on to ℓ_q, the projected mean μ_0 will be at most 5α away from p as projection cannot increase the distance between the projected point any of the points on the line. This implies that μ_0 will be at least 15α far from q.

From Lemma 7, we know that the mean of the distribution truncated around p for a sufficiently large ξ (which we call μ_p) is at most α far from μ_0. So, by Lemma 14, we know that with probability at least $1 - \frac{\beta}{|S|^2}$, the median of means (μ') is at most α far from μ_p. This implies that μ' is more than 14α far from q, and at most 6α far from p. Therefore, to lose to q, μ' will have to be moved by at least 4α towards q.

Since moving a point in X can move μ' by at most $\frac{2m\xi}{n}$, it means that we need to change at least $\frac{2n\alpha}{m\xi}$ points each from at least $0.4m$ of the sub-datasets in X to make p lose to q (from the proof of Lemma 14), since we want to have the means of at least half of the sub-datasets to be closer to q. Taking the union bound over all pairs of points in S, we get the desired error probability bound. This proves the claim.

E.4. Proof of Lemma 18

Fix a $p \in S_{>}$. We have to deal with two cases here. First, when $\|\mu - p\|_2 \leq \frac{\xi}{2}$, and when $\|\mu - p\|_2 > \frac{\xi}{2}$.

For the first case, let z be the point in S that is nearest to μ, and let ℓ_z be the line $p - z$. Suppose μ_1 is the projection of μ on to ℓ_z. Then μ_1 will be at most α from z. By Lemma 7, the mean of the distribution truncated around p (which we call μ_p) will be at most α far from μ_1. Then by Lemma 14, with probability at least $1 - \frac{\beta}{|S|^2}$, the median of means (μ') will be at most α far from μ_p, hence, at most 3α far from z. This implies that μ' will be at least 17α far from p. Therefore, the score of p will be 0, since it has already lost to z.

In the second case, let ℓ_μ be the line $p - \mu$. Suppose μ_1 is the mean of the distribution projected on to ℓ_μ and truncated around p, and let z be the point in S that is closest to μ_1. If $\mu_1 = z$, then we're done because by Lemma 13, μ_1 is at least $\frac{15\xi}{16}$ far from p, and then by Lemma 14, with probability at least $1 - \frac{\beta}{|S|^2}$, the median of means lies α close to μ_1, and is closer to μ_1 than it is to p.

If not, then we have to do some more work. Now, let ℓ_z be the line $p - z$, let μ_z be the projection of μ on to ℓ_z, and let μ_2 be the projection of μ_1 on to ℓ_z. Using basic geometry, we have the following.

$$\begin{aligned}
& \frac{\|p - \mu\|_2}{\|p - \mu_z\|_2} = \frac{\|p - \mu_1\|_2}{\|p - \mu_2\|_2} \\
\iff & \|p - \mu_z\|_2 = \frac{\|p - \mu\|_2 \|p - \mu_2\|_2}{\|p - \mu_1\|_2} \\
\implies & \|p - \mu_z\|_2 \geq \frac{\|p - \mu\|_2(\|p - \mu_1\|_2 - \alpha)}{\|p - i\mu_1\|_2} && \text{(Triangle Inequality)} \\
& = \|p - \mu\|_2 \left(1 - \frac{\alpha}{\|p - \mu_1\|_2}\right) \\
& \geq \|p - \mu\|_2 \left(1 - \frac{16\alpha}{15\xi}\right) \\
& = \|p - \mu\|_2 \left(1 - \frac{16\alpha^{\frac{k}{k-1}}}{15C}\right) \\
& \geq \frac{15\|p - \mu\|_2}{16} && \text{(Due to our restrictions on } C \text{ and } \alpha)
\end{aligned}$$

If $\|p - \mu_z\|_2 \leq \frac{\xi}{2}$, then by a similar argument as in the first case, z wins against p because the mean of the truncated distribution is close to μ_z, and the empirical median of means is close to that mean with probability at least $1 - \frac{\beta}{|S|^2}$, hence, closer to z than to p. If not, then by Lemma 13, the mean of the distribution projected on to ℓ_z, and truncated around p will be at most $\frac{\xi}{16}$ far from $p - \xi$, so by Lemma 14, the median of means (μ') will be at most α far from that mean with probability at least $1 - \frac{\beta}{|S|^2}$. This implies that it will be at most $\frac{\xi}{16} + 2\alpha$ far from z, but will be at least $\frac{15\xi}{16} - 2\alpha$ from p, which means that p will lose to z by default.

Taking the union bound over all sources of error, and all pairs of points in S, we get the error probabiity of 4β, which we can rescale to get the required bounds.

E.5. Proof of Lemma 19

Let X be any dataset, and $p \in \mathbb{R}^d$ be a point in the domain in question, and let $\textsc{Score}_X(p)$ be the score of p. Suppose X' is a neighbouring dataset of X. Then by changing a point x in X to x' (to get X'), we can only change the score of p by 1. Let the median of means of projected, truncated X be μ_1, and that of X' be μ_2.

Suppose p was already losing to a point q, that is, its score was 0. Then switching from X to X' can either imply that μ_2 is further from p than μ_1 was from p, or it could go further. In the first case, p would still lose to q. In the second case, if μ_2 is closer to p than it is to q, then the score of p would increase at most by 1 because we can switch back to X from X' by switching one point; or p could still be losing to q, in which case, the score wouldn't change at all.

Now, suppose p was winning against all points that are more than 30α away from p with respect to X. Let q be a point that determined the score of p. If switching to X' made μ_2 closer to p than μ_1 was, then the score can only increase by 1 because we can always switch back to X, and get the original score. If it moved μ_2 closer to q than μ_1 was, then the score can only decrease by 1. This is because q determined $\textsc{Score}_X(p)$ via some optimal strategy, and changing a point of X cannot do better than that. Therefore, the sensitivity is 1, as required.

Appendix F. Estimating in High Dimensions with CDP

In this section, we give a computationally efficient, ρ-zCDP algorithm for estimaing the mean of a distribution with bounded k^{th} moment. The analogous (ε, δ)-DP algorithm would be the same, with the same analysis, and we state the theorem for it at the end.

Remark 34 *Throughout the section, we assume that the dimension d is greater than some absolute constant ($32\ln(4)$). If it is less than that, then we can just use our one-dimensional estimator from Theorem 33 multiple times to individually estimate each coordinate with constant multiplicative overhead in sample complexity.*

F.1. Technical Lemmata

Similar to our one-dimensional distribution estimator, the idea is to aggressively truncate the distribution around a point, and compute the noisy empirical mean. We first have to define what truncation in high dimensions means. The definition essentially says that if a point lies outside the specified range (in this case, a sphere around a point), then we project the point on to the surface of the sphere in the direction towards the centre.

Definition 35 *Let $\rho, x \in \mathbb{R}^d$, and $r > 0$. Then we define $\mathsf{trunc}(\rho, r, x)$ as follows.*

$$\mathsf{trunc}(\rho, r, x) = \begin{cases} x & \text{if } \|\rho - x\|_2 \le r \\ y \text{ st. } \|y - \rho\|_2 = r \text{ and } y = \rho + \gamma \cdot (\rho - x) \text{ for some } \gamma \in \mathbb{R} & \text{if } \|\rho - x\|_2 > r \end{cases}$$

Similarly, for a dataset $S = (X_1, \dots, X_n) \in \mathbb{R}^{n \times d}$, we define $\mathsf{trunc}(\rho, r, S)$ as the dataset $S' = (X'_1, \dots, X'_n)$, where for each $1 \le i \le n$, $X'_i = \mathsf{trunc}(\rho, r, X_i)$.

Now, we show that the mean of the distribution, truncated to within a large-enough ball centered close to the mean, does not move too far from the original mean. Our lemmas are somewhat similar to techniques in Diakonikolas et al. (2017) (see Lemma A.18), but their results are specific to bounded second moments ($k = 2$), while we focus on the case of general k.

Lemma 36 *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ, and k^{th} moment bounded by 1, where $k \geq 2$. Let $\rho \in \mathbb{R}^d$, $0 < \tau < \frac{1}{16}$, and $\xi = \frac{C\sqrt{d}}{\tau^{\frac{1}{k-1}}}$ for a constant $C > 2$. Let $X \sim \mathcal{D}$, and Z be the following random variable.*

$$Z = \mathsf{trunc}(\rho, \xi, X)$$

If $\|\mu - \rho\|_2 \leq \frac{\xi}{2}$, then $\|\mu - \mathbb{E}[Z]\|_2 \leq \tau$.

Proof By self-duality of the Euclidean norm, it is sufficient to prove that for each unit vector $v \in \mathbb{R}^d$, $|\langle \mu - \mathbb{E}[Z], v\rangle| \leq \tau$. Let $\gamma = \mathbb{E}[Z]$. Then we have the following.

$$\begin{aligned}
|\langle \mu - \mathbb{E}[Z], v\rangle| &= \left|\mathbb{E}\left[\langle X - \gamma, v\rangle \mathbb{1}_{X \notin B_\xi(\rho)}\right]\right| \\
&\leq \left(\mathbb{E}\left[|\langle X - \gamma, v\rangle|^k\right]\right)^{\frac{1}{k}} \left(\mathbb{E}\left[\mathbb{1}_{X \notin B_\xi(\rho)}\right]\right)^{\frac{k-1}{k}} && \text{(Lemmata 49 and 48)} \\
&\leq 1 \cdot \left(\mathbb{P}\left[\|X - \mu\|_2 > \frac{\xi}{2}\right]\right)^{\frac{k-1}{k}} \\
&\leq \tau. && \text{(Lemma 43)}
\end{aligned}$$

■

Our last lemma shows that the empirical mean of a set of samples from a high-dimensional distribution with bounded k^{th} moment is close to the mean of the distribution.

Lemma 37 *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ and k^{th} moment bounded by 1. Suppose for $\tau > 0$, $(X_1, \dots, X_n)$ are samples from $\mathcal{D}$, where*

$$n \geq O\left(\frac{d}{\tau^2}\right).$$

Then with probability at least 0.9,

$$\left\|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right\|_2 \leq \tau.$$

Proof Suppose Σ is the covariance matrix of $\mathcal{D}$. We know that $\|\Sigma\|_2 = 1$. Let $\overline{\mu} = \sum\limits_{i=1}^{n} X_i$. Then

$$\mathbb{E}[\overline{\mu}] = \mu \quad \text{and} \quad \mathbb{E}\left[(\overline{\mu} - \mu)^T(\overline{\mu} - \mu)\right] = \frac{1}{n}\Sigma.$$

Using Lemma 43, we have the following.

$$\begin{aligned}
\mathbb{P}[\|\overline{\mu} - \mu\|_2 > \tau] &\leq \left(\sqrt{\frac{d}{n\tau^2}}\right) \\
&\leq 0.9.
\end{aligned}$$

■

F.2. The Algorithm

We finally state the main theorem of the section here. Algorithm 4 first computes a rough estimate of the mean using the one-dimensional mean estimator from Theorem 33 that lies at most $\sqrt{d}$ from μ. Then it truncates the distribution to within a small ball around the estimate, and uses the Gaussian Mechanism (Lemma 28) to output a private empirical mean.

```
Algorithm 4: CDP High-Dimensional Mean Estimator $\mathrm{CDPHDME}_{\rho,\alpha,R}(X)$
Input: Samples $X_1, \ldots, X_{2n} \in \mathbb{R}^d$. Parameters $\rho, \alpha, R > 0$.
Output: $\widehat{\mu} \in \mathbb{R}^d$.

Set parameters: $Y \leftarrow (X_1, \ldots, X_n)$    $Z \leftarrow (X_{n+1}, \ldots, X_{2n})$    $r \leftarrow \frac{4\sqrt{d}}{\alpha^{\frac{1}{k-1}}}$

// Obtain a rough estimate of the mean via coordinate-wise
   estimation
for $i \leftarrow 1, \ldots, d$ do
    $c_i \leftarrow \mathrm{CDPODME}_{\frac{\rho}{d},1,\frac{0.1}{d},R}(Y^i)$
end
Let $\boldsymbol{c} \leftarrow (c_1, \ldots, c_d)$

// Truncate to within a small ball around the mean
Let $Z' \leftarrow \mathsf{trunc}(\boldsymbol{c}, r, Z)$

// Estimate the mean
$\widehat{\mu} \leftarrow \frac{1}{n} \sum_{z \in Z'} z + \mathcal{N}\left(\mathbf{0}, \frac{2r^2}{\rho n^2} \mathbb{I}_{d \times d}\right)$
return $\widehat{\mu}$
```

Theorem 38 *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean $\mu \in B_R(\mathbf{0})$ and k^{th} moment bounded by* 1. *Then for all $\rho, \alpha > 0$, there exists a polynomial-time, ρ-zCDP algorithm that takes*

$$n \geq O\left(\frac{d}{\alpha^2} + \frac{d}{\sqrt{\rho}\alpha^{\frac{k}{k-1}}} + \frac{\sqrt{d \log(R)} \log(d)}{\sqrt{\rho}}\right)$$

samples from $\mathcal{D}$, and outputs $\widehat{\mu} \in \mathbb{R}^d$, such that with probability at least 0.7,

$$\|\mu - \widehat{\mu}\|_2 \leq \alpha.$$

Proof The proofs of privacy and accuracy (for Algorithm 4) are separated again as follows.
Privacy:
Privacy follows from Lemmata 33, 28, and 25 (since the ℓ_2-sensitivity of the estimation step is $\frac{2r}{n}$).
Accuracy:
The first step finds a centre c_i for each coordinate i, such that the i^{th} coordinate of the mean is at most 1 far from c_i in absolute distance. Therefore, $\boldsymbol{c}$ is at most $\sqrt{d}$ away from μ. By Theorem 33, this happens with probability at least 0.9.

Now, by Lemma 36, the mean of the truncated distribution around $\boldsymbol{c}$ (that we call μ') is at most α far from μ is ℓ_2 distance. Therefore, by Lemma 37, the empirical mean of the truncated distribution ($\overline{\mu}$) will be at most α far from μ' with probability at least 0.9.

Finally, let $z = (z_1, \dots, z_d)$ be the noise vector added in the estimation step. and let $S_z = \sum\limits_{i \in [d]} z_i$. Then since $d \geq 32 \ln(4)$, by Lemma 47, we have the following.

$$\mathbb{P}\left[\left|\sum_{i=1}^{d} z_i^2 - \frac{2dr^2}{\rho n^2}\right| \geq 0.5 \times \frac{2dr^2}{\rho n^2}\right] \leq 0.1.$$

Therefore, it is enough to have the following.

$$\begin{aligned}
\frac{3dr^2}{\rho n^2} &\leq \alpha^2 \\
\iff n &\geq \sqrt{\frac{48d^2}{\rho \alpha^{\frac{2k}{k-1}}}} \\
&= \frac{4\sqrt{3}d}{\sqrt{\rho}\alpha^{\frac{k}{k-1}}}.
\end{aligned}$$

This is what we required in our sample complexity. Hence, by the union bound and rescaling α by a constant, we get the required result. ■

To get the analogous (ε, δ)-DP algorithm, we just use ADPODME instead of CDPODME in the first step, and keep the rest the same.

Theorem 39 *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean $\mu \in B_R(\mathbf{0})$ and k^{th} moment bounded by* 1. *Then for all $\varepsilon, \delta, \alpha > 0$, there exists a polynomial-time, (ε, δ)-DP algorithm that takes*

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

samples from $\mathcal{D}$, and outputs $\widehat{\mu} \in \mathbb{R}^d$, such that with probability at least 0.7,

$$\|\mu - \widehat{\mu}\|_2 \leq \alpha.$$

Appendix G. Lower Bounds for Estimating High-Dimensional Distributions

Theorem 40 *Suppose $\mathcal{A}$ is an $(\varepsilon, 0)$-DP algorithm and $n \in \mathbb{N}$ is a number is such that, for every product distribution P on $\mathbb{R}^d$ such that $\mathbb{E}[P] = \mu$ and $\sup_{v:\|v\|_2=1} \mathbb{E}\left[\langle v, P - \mu\rangle^2\right] \leq 1$,*

$$\mathop{\mathbb{E}}_{X_1,\dots,X_n \sim P, \mathcal{A}}\left[\|\mathcal{A}(X) - \mu\|_2^2\right] \leq \alpha^2.$$

Then $n = \Omega\left(\frac{d}{\alpha^2 \varepsilon}\right)$.

The proof uses a standard *packing argument*, which we encapsulate in the following lemma.

Lemma 41 *Let $\mathcal{P} = \{P_1, P_2, \dots\}$ be a family of distributions such that, for every $P_i, P_j \in \mathcal{P}$, $\mathrm{d_{TV}}(P_i, P_j) \le \tau$. Suppose $\mathcal{A}$ is an $(\varepsilon, 0)$-DP algorithm and $n \in \mathbb{N}$ is a number such that, for every $P_i \in \mathcal{P}$,*

$$\underset{X_1,\dots,X_n \sim P_i, \mathcal{A}}{\mathbb{P}}[\mathcal{A}(X) = i] \ge 2/3,$$

then $n = \Omega\Big(\frac{\log|\mathcal{P}|}{\tau\varepsilon}\Big)$.

Proof We will define a packing as follows. As a shorthand, define

$$\begin{cases} 0 & \text{w.p. } 1 - \frac{\alpha^2}{d} \\ \frac{\sqrt{d}}{\alpha} & \text{w.p. } \frac{\alpha^2}{d} \end{cases}$$

For $c \in \{0,1\}^d$, let

$$P_c = \bigotimes_{j=1}^{d} Q_{c_j}$$

be the product of the distributions Q_0 and Q_1 where we choose each coordinate of the product based on the corresponding coordinate of c.

Note that $\mathrm{d_{TV}}(Q_0, Q_1) \le \alpha^2/d$, and therefore, for every $c, c' \in \{0,1\}^d$, $\mathrm{d_{TV}}(P_c, P_{c'}) \le \alpha^2$. Let $\mathcal{C} \subseteq \{0,1\}^d$ be a code of relative distance $1/4$. That is, every distinct $c, c' \in \mathcal{C}$ differ on at least $d/4$ coordinates. By standard information-theoretic arguments, there exists such a code such that $|\mathcal{C}| = 2^{\Omega(d)}$. We will define the packing to be $\mathcal{P} = \{\mathcal{P}_c\}_{c \in \mathcal{C}}$. By Lemma 41, if there is an $(\varepsilon, 0)$-DP algorithm $\mathcal{A}$ that takes n samples from an arbitrary one of the distribution $P_c \in \mathcal{P}$ and correctly identifies P_c with probability at least $2/3$, then $n = \Omega\big(\frac{d}{\alpha^2\varepsilon}\big)$.

We make two more observation about the distributions in $\mathcal{P}$. First, since this is a product distribution, its 2nd moment is bounded by the maximum 2nd moment of any coordinate, so

$$\sup_{v : \|v\|_2 = 1} \mathbb{E}\big[\langle v, P - \mu\rangle^2\big] \le \max\{\mathrm{Var}[Q_0], \mathrm{Var}[Q_1]\} \le 1.$$

Second, since any distinct c, c' differ on $d/4$ coordinates, and $\mathbb{E}[Q_1 - Q_0] = \alpha/\sqrt{d}$, we have that for every distinct c, c',

$$\|\mathbb{E}[P_c - P_{c'}]\|_2 \ge \sqrt{\frac{d}{4}} \cdot \frac{\alpha}{\sqrt{d}} = \frac{\alpha}{2}.$$

By a standard packing argument, any $(\varepsilon, 0)$-DP algorithm that takes n samples from P_c for an arbitrary $c \in \mathcal{C}$, and correctly identifies c, must satisfy $n = \Omega(\frac{d}{\alpha^2\varepsilon})$. Therefore, if we can estimate the mean to within ℓ_2^2 error $< \alpha^2/64$, we can identify c uniquely. Moreover, if $\mathcal{A}$ satisfies

$$\underset{X_1,\dots,X_n \sim P, \mathcal{A}}{\mathbb{E}}\big[\|\mathcal{A}(X) - \mathbb{E}[P]\|_2^2\big] < \alpha^2/192$$

for every distribution P with bounded $2nd$ moment, then by Markov's inequality, we have

$$\underset{X_1,\dots,X_n \sim P, \mathcal{A}}{\mathbb{P}}\big[\|\mathcal{A}(X) - \mathbb{E}[P]\|_2^2 < \alpha^2/64\big] \ge 2/3.$$

Therefore, any $(\varepsilon, 0)$-DP algorithm $\mathcal{A}$ with low expected ℓ_2^2 error must have $n = \Omega(\frac{d}{\alpha^2\varepsilon})$. The theorem now follows by a change-of-variables for α. ■

Appendix H. Useful Inequalities

The following standard concentration inequalities are used frequently in this document.

Lemma 42 (Chebyshev's Inequality) *Let $\mathcal{D}$ be a distribution over $\mathbb{R}$ with mean μ, and k^{th} moment bounded by M. Then the following holds for any $a > 1$.*

$$\underset{X\sim\mathcal{D}}{\mathbb{P}}\left[|X-\mu| > aM^{\frac{1}{k}}\right] \leq \frac{1}{a^k}$$

Lemma 43 (Concentration in High Dimensions Zhu et al. (2019)) *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean $\mathbf{0}$, and k^{th} moment bounded by M. Then the following holds for any $t > 0$.*

$$\underset{X\sim\mathcal{D}}{\mathbb{P}}[\|X\|_2 > t] \leq M\left(\frac{\sqrt{d}}{t}\right)^k$$

Lemma 44 (Multiplicative Chernoff) *Let $X_1, \dots, X_m$ be independent Bernoulli random variables taking values in $\{0,1\}$. Let X denote their sum and let $p = \mathbb{E}[X_i]$. Then for $m \geq \frac{12}{p}\ln(2/\beta)$,*

$$\mathbb{P}\left[X \notin \left[\frac{mp}{2}, \frac{3mp}{2}\right]\right] \leq 2e^{-mp/12} \leq \beta.$$

Lemma 45 (Bernstein's Inequality) *Let $X_1, \dots, X_m$ be independent Bernoulli random variables taking values in $\{0,1\}$. Let $p = \mathbb{E}[X_i]$. Then for $m \geq \frac{5p}{2\varepsilon^2}\ln(2/\beta)$ and $\varepsilon \leq p/4$,*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum X_i - p\right| \geq \varepsilon\right] \leq 2e^{-\varepsilon^2 m/2(p+\varepsilon)} \leq \beta.$$

Lemma 46 (Laplace Concentration) *Let $Z \sim \mathrm{Lap}(t)$. Then $\mathbb{P}[|Z| > t \cdot \ln(1/\beta)] \leq \beta$.*

Lemma 47 (Gaussian Empirical Variance Concentration) *Let $(X_1, \dots, X_m) \sim \mathcal{N}(0, \sigma^2)$ be independent. If $m \geq \frac{8}{\tau^2}\ln(2/\beta)$, for $\tau \in (0,1)$, then*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} X_i^2 - \sigma^2\right| > \tau\sigma^2\right] \leq \beta.$$

We also mention two well-known and useful inequalities.

Lemma 48 (Hölder's Inequality) *Let X, Y be random variables over $\mathbb{R}$, and let $k > 1$. Then,*

$$\mathbb{E}[|XY|] \leq \left(\mathbb{E}\left[|X|^k\right]\right)^{\frac{1}{k}}\left(\mathbb{E}\left[|Y|^{\frac{k}{k-1}}\right]\right)^{\frac{k-1}{k}}.$$

Lemma 49 (Jensen's Inequality) *Let X be an integrable, real-valued random variable, and ψ be a convex function. Then*

$$\psi(\mathbb{E}[X]) \leq \mathbb{E}[\psi(X)].$$