Privately Learning High-Dimensional Distributions

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

Jerry Li JERRL@MICROSOFT.COM
Microsoft Research AI

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

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We present novel, computationally efficient, and differentially private algorithms for two fundamental high-dimensional learning problems: learning a multivariate Gaussian and learning a product distribution over the Boolean hypercube in total variation distance. The sample complexity of our algorithms nearly matches the sample complexity of the optimal non-private learners for these tasks in a wide range of parameters, showing that privacy comes essentially *for free* for these problems. In particular, in contrast to previous approaches, our algorithm for learning Gaussians does not require strong *a priori* bounds on the range of the parameters. Our algorithms introduce a novel technical approach to reducing the sensitivity of the estimation procedure that we call *recursive private preconditioning*.

Keywords: Privacy, learning, Gaussian, product distribution[1]

1. Introduction

A central problem in machine learning and statistics is to learn (estimate) the parameters of an unknown distribution using samples. However, in many applications, these samples consist of highly sensitive information belonging to individuals, and the output of the learning algorithm may inadvertently reveal this information. While releasing only the estimated parameters of a distribution may seem harmless, when there are enough parameters—that is, when the data is *high-dimensional*—these statistics can reveal a lot of individual-specific information (see e.g. Dinur and Nissim (2003); Homer et al. (2008); Bun et al. (2014); Dwork et al. (2015); Shokri et al. (2017), and the survey Dwork et al. (2017)). For example, the influential attack of Homer et al. (2008) showed how to use very simple statistical information released in the course of genome wide association studies to detect the presence of individuals in those studies, which implies these individuals have a particular medical condition. Thus it is crucial to design learning algorithms that ensure the privacy of the individuals in the dataset.

The most widely accepted solution to this problem is *differential privacy* (Dwork et al., 2006), which provides a strong individual privacy guarantee by ensuring that no individual sample has a significant influence on the learned parameters. A large body of literature now shows how to

1. A full version of this paper is available as Kamath et al. (2018).

implement nearly every statistical algorithm privately, and differential privacy is now being deployed by Apple (Differential Privacy Team, Apple, 2017), Google (Erlingsson et al., 2014), and the US Census Bureau (Dajani et al., 2017).

Differential privacy is typically achieved by adding noise to some non-private estimator, where the magnitude of the noise is calibrated to mask the effect of the single sample. However, straightforward methods for adding this noise require strong *a priori* bounds on the distribution to provide meaningful accuracy guarantees.

> *Example:* Suppose we want to estimate the mean $\mu \in (-R, R)$ of a Gaussian random variable with known variance 1, using the empirical mean of the data. The empirical mean itself has variance $1/n$. However, the naïve strategy for adding noise to the empirical mean would increase the variance by $O(R^2/n^2)$. Thus, the variance of the naïve private algorithm dominates the variance in the data unless we have $\Omega(R^2)$ samples, forcing the user to have strong *a priori* knowledge of the mean.

This problem is pervasive in applications of differential privacy, and considerable effort has been made to cope with this need for the parameters to lie in a small *range*, including multiple systems that have been built to help elicit this information from the user (Mohan et al., 2012; Gaboardi et al., 2016a).

When the distribution is *low-dimensional*, there are many more effective algorithms that avoid the polynomial dependence on the size of the range. For example, in the setting above of estimating a single univariate Gaussian, Karwa and Vadhan (2018) showed how to estimate the mean with *essentially no* dependence on R. More generally, there are also a plethora of general algorithmic techniques that can be used to address this general problem, such as *smooth sensitivity* (Nissim et al., 2007), *subsample-and-aggregate* (Nissim et al., 2007; Smith, 2011), *propose-test-release* (Dwork and Lei, 2009). Unfortunately, none of these methods extend well to *high-dimensional* problems. When they exist, natural extensions either incur a costly dependence on the dimension in the sample complexity, or have running time exponential in the dimension.

In this work we show how to privately learn two fundamental families of high-dimensional distributions with comparable costs to the corresponding optimal non-private learning algorithms:

- We give a computationally efficient algorithm for learning a multivariate Gaussian with unknown mean and covariance in total variation distance. This algorithm requires only weak *a priori* bounds on the mean and covariance, and in a wide range of parameters its sample complexity matches the optimal non-private algorithm up to lower-order terms.

- We give a computationally efficient algorithm for learning a product distribution over the Boolean hypercube in total variation distance, which requires adding noise to each coordinate proportional to its variance, despite not knowing the variance *a priori*. Again, for many parameter regimes, the sample complexity of this algorithm is similar to that of the optimal non-private algorithm.

Our results show that it is possible to obtain privacy nearly *for free* when learning these important classes of high-dimensional distribution. We obtain these results using a novel approach to reduce the sensitivity of the estimation procedure, which we call *private recursive preconditioning*.

1.1. Our Results

1.1.1. PRIVATELY LEARNING GAUSSIANS

The most fundamental class of high-dimensional distributions is the *multivariate Gaussian* in $\mathbb{R}^d$. Our first result is an algorithm that takes samples from a distribution $\mathcal{N}(\mu, \Sigma)$ with unknown mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$ and estimates parameters $\widehat{\mu}, \widehat{\Sigma}$ such that $\mathcal{N}(\widehat{\mu}, \widehat{\Sigma})$ is close to the true distribution in total variation distance (TV distance). Without privacy, $n = \Theta(\frac{d^2}{\alpha^2})$ samples suffice to guarantee total variation distance at most α (this is folklore, but see e.g. Diakonikolas et al. (2016)).

Despite the simplicity of this problem, it was only recently that Karwa and Vadhan (2018) gave an optimal algorithm for learning a univariate Gaussian. Specifically, they showed that just $n = \tilde{O}(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon} + \frac{\log(R\log\kappa)}{\varepsilon})$ samples are sufficient to learn a univariate Gaussian $\mathcal{N}(\mu, \sigma^2)$ with $|\mu| \leq R$ and $1 \leq \sigma^2 \leq \kappa$, up to α in total variation distance subject to ε-differential privacy. In contrast to naïve approaches, their result has two important features: (1) The sample complexity has only mild dependence on the range parameters R and κ, and (2) the sample complexity is only larger than that of the non-private estimator by a small multiplicative factor and an additive factor that is a lower order term for a wide range of parameters. When the covariance is unknown, a naïve application of their algorithm would preserve neither of these features.

We show that it is possible to privately estimate a multivariate normal while preserving both of these features. Our algorithms satisfy the strong notion of *concentrated differential privacy (zCDP)* (Dwork and Rothblum, 2016; Bun and Steinke, 2016), which is formally defined in Section B. To avoid confusion we remark that these definitions are on different scales so that $\frac{\varepsilon^2}{2}$-zCDP is comparable to ε-DP and (ε, δ)-DP.

Theorem 1 (Gaussian Estimation) *There is a polynomial time $\frac{\varepsilon^2}{2}$-zCDP algorithm that takes*

$$n = \tilde{O}\left(\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon} + \frac{d^{3/2}\log^{1/2}\kappa + d^{1/2}\log^{1/2} R}{\varepsilon}\right)$$

samples from a Gaussian $\mathcal{N}(\mu, \Sigma)$ with unknown mean $\mu \in \mathbb{R}^d$ such that $\|\mu\|_2 \leq R$ and unknown covariance $\Sigma \in \mathbb{R}^{d\times d}$ such that $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$, and outputs estimates $\widehat{\mu}, \widehat{\Sigma}$ such that, with high probability $\mathrm{d_{TV}}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\widehat{\mu}, \widehat{\Sigma})) \leq \alpha$. Here, $\tilde{O}(\cdot)$ hides polylogarithmic factors of d, $\frac{1}{\alpha}$, $\frac{1}{\varepsilon}$, $\log\kappa$, and $\log R$. The same algorithm satisfies $(\varepsilon\sqrt{\log(1/\delta)}, \delta)$-differential privacy for every $\delta > 0$.

Theorem 1 will follow by combining Theorem 9 for covariance estimation with Theorem 12 for mean estimation. Observe that, since the sample complexity without privacy is $\Theta(\frac{d^2}{\alpha^2})$, Theorem 1 shows that privacy comes almost for free unless $\frac{1}{\varepsilon}$, κ, or R are quite large.

The main difficulty that arises when trying to extend the results of Karwa and Vadhan (2018) to the multivariate case is that the covariance matrix of the Gaussian might be almost completely *unknown*. The main technically novel part of our algorithm is a method for learning a matrix A approximating the inverse of the covariance matrix so that $\mathbb{I} \preceq A\Sigma A \preceq 1000\mathbb{I}$. This matrix can be used to transform the Gaussian to be nearly spherical, making it possible to apply the methods of Karwa and Vadhan (2018).

Theorem 2 (Private Preconditioning) *There is an $\frac{\varepsilon^2}{2}$-zCDP algorithm that takes $n = \tilde{O}\left(\frac{d^{3/2}\log^{1/2}\kappa}{\varepsilon}\right)$ samples from an unknown Gaussian $\mathcal{N}(0, \Sigma)$ over $\mathbb{R}^d$ with $\Sigma \in \mathbb{R}^{d\times d}$ such that $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$, and*

outputs a symmetric matrix A such that $\mathbb{I} \preceq A\Sigma A \preceq 1000\mathbb{I}$. Here, $\tilde{O}(\cdot)$ hides polylogarithmic factors of d, $\frac{1}{\varepsilon}$, and $\log \kappa$.

We describe and analyze our algorithm for private covariance estimation in Section 2, and in Section 3 we combine it with the algorithms of Karwa and Vadhan (2018) to obtain Theorem 1.

1.1.2. PRIVATELY LEARNING PRODUCT DISTRIBUTIONS

The simplest family of high-dimensional discrete distributions are product distributions over $\{0,1\}^d$. Without privacy, $\Theta(\frac{d}{\alpha^2})$ are necessary and sufficient to learn up to α in total variation distance. The standard approach to achieving DP by perturbing each coordinate independently requires $\tilde{\Theta}(\frac{d}{\alpha^2} + \frac{d^{3/2}}{\alpha\varepsilon})$ samples. We give an improved algorithm for this problem that avoids this blowup in sample complexity. While our algorithm for learning product distributions is quite different to our algorithm for estimating Gaussian covariance, it uses a similar recursive preconditioning technique, highlighting the versatility of this approach.

Theorem 3 *There is a polynomial time $\frac{\varepsilon^2}{2}$-zCDP algorithm that takes $n = \tilde{O}\big(\frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon}\big)$ samples from an unknown product distribution $\mathcal{P}$ over $\{0,1\}^d$ and outputs a product distribution $\mathcal{Q}$ such that, with high probability, $\mathrm{d_{TV}}(\mathcal{P},\mathcal{Q}) \le \alpha$. Here, $\tilde{O}(\cdot)$ hides polylogarithmic factors of d, $\frac{1}{\alpha}$, and $\frac{1}{\varepsilon}$. The same algorithm satisfies $(\varepsilon\sqrt{\log(1/\delta)}, \delta)$-differential privacy for every $\delta > 0$.*

We describe and analyze our algorithm in Section D.

1.1.3. LOWER BOUNDS

We prove lower bounds for the problems we consider in this paper, demonstrating that for many problems, our sample complexity is optimal up to polylogarithmic factors. One example statement is the following lower bound for private mean estimation of a product distribution, for the more permissive notion of (ε,δ)-differential privacy (compared to our upper bounds, which are in terms of zCDP):

Theorem 4 *Any $(\varepsilon, \frac{1}{64n})$-differentially private algorithm that takes samples from an arbitrary unknown product distribution $\mathcal{P}$ over $\{0,1\}^d$ and outputs a product distribution $\mathcal{Q}$ such that $\mathrm{d_{TV}}(\mathcal{P},\mathcal{Q}) \le \alpha$ with probability $\ge 9/10$ requires $n = \Omega(\frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon\log d})$ samples.*

We also prove a qualitatively similar lower bound for privately estimating the mean of a Gaussian distribution.

In addition, we prove lower bounds for privately estimating a Gaussian with unknown covariance. These are qualitatively weaker, as they are only for ε-differential privacy, and we consider it an interesting open question to prove lower bounds for covariance estimation under concentrated or approximate differential privacy.

Theorem 5 *Any ε-differentially private algorithm that takes samples from an arbitrary unknown Gaussian distribution $\mathcal{P}$ and outputs a Gaussian distribution $\mathcal{Q}$ such that $\mathrm{d_{TV}}(\mathcal{P},\mathcal{Q}) \le \alpha$ with probability $\ge 9/10$ requires $n = \Omega(\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon})$ samples.*

The last question to address is the dependence on the parameters R and κ. It is well-known that, under zCDP, the existence of a 0.1-packing (in total variation distance) $\mathcal{P}$ results in a lower bound of $n = \Omega(\frac{1}{\varepsilon}\log^{1/2}|\mathcal{P}|)$ (Hardt and Talwar, 2010; Beimel et al., 2014; Bun and Steinke, 2016). Since, for a set of identity covariance Gaussians, there exists such a packing of size $R^{\Omega(d)}$, this implies that the dependence of Theorem 1 on R is optimal up to logarithmic factors. As for the dependence on κ, it can be shown that for $d = 2$, there exists a packing of zero-mean Gaussians of size $\mathrm{poly}(\kappa)$ (Kamath, 2019), in contrast to $\mathrm{poly}\log\kappa$ in one dimension. That is, though our algorithm's dependence on κ is exponentially greater than that of Karwa and Vadhan (2018), this is necessary for any $d \geq 2$.

All our lower bounds are presented in Section E.

1.1.4. COMPARISON TO LOWER BOUNDS FOR HIGH-DIMENSIONAL DP

Readers familiar with differential privacy may wonder why our results do not contradict known lower bounds for high-dimensional estimation in differential privacy (Bun et al., 2014; Steinke and Ullman, 2015; Dwork et al., 2015). The two key differences are (1) lower bounds showing that privacy is costly are for the relatively weak ℓ_∞ estimation guarantee whereas we want a rather stringent estimation guarantee, and (2) we exploit the structure of Gaussians and product distributions to obtain guarantees that are not possible for arbitrary distributions.

To understand the first issue, most lower bounds in differential privacy apply to estimating the mean of the distribution up to α in ℓ_∞ distance. This guarantee can be achieved with $\Theta(\frac{\log d}{\alpha^2})$ samples non-privately but requires $\Theta(\frac{\log d}{\alpha^2} + \frac{\sqrt{d}}{\alpha})$ samples with differential privacy. Thus, for ℓ_∞ estimation, privacy is costly in high dimensions. However, if we consider the more stringent ℓ_2 metric, then the cost of privacy goes away, and $\Theta(\frac{d}{\alpha^2})$ samples are sufficient with or without privacy.[2] Thus, for this stronger guarantee, privacy is not costly in high-dimensions. This phenomenon is fairly general, and is not specific to Gaussians or product distributions.

To understand the second issue, in order to learn Gaussians or product distributions in total variation distance, we need to learn in metrics that are related to ℓ_2, but take into account the variance of the distribution. In the case of Gaussians we learn in the Mahalanobis distance $\|\cdot\|_\Sigma$ and for product distributions our guarantees are closely related to χ^2-divergence. For example, Theorem 1 can actually be rephrased as saying that, for our algorithm,

$$n = \tilde{O}\left(\frac{d^{3/2}\log^{1/2}\kappa}{\varepsilon}\right) \implies \|\Sigma - \widehat{\Sigma}\|_\Sigma = O\left(\sqrt{\frac{d^2}{n}} + \frac{d^2}{\varepsilon n}\right).$$

This sort of guarantee where the error in the Σ-norm does not depend on the range parameter κ cannot be achieved for arbitrary distributions, and thus for this part of the guarantee we crucially use the fact that the data is i.i.d. from a Gaussian. A similar phenomenon arises for product distributions, where, as we show, learning the mean of a product distribution in the right metric can be done with $\tilde{O}(d)$ samples for product distributions but would require $\Omega(d^{3/2})$ samples for arbitrary distributions.

2. One way to see this is that estimation up to $\alpha/\sqrt{d}$ in ℓ_∞ implies estimation up to α in ℓ_2, so the non-private term and the private term in the ℓ_∞ bounds have roughly the same dependence on the dimension in this case.

1.2. Techniques

1.2.1. PRIVATELY LEARNING GAUSSIANS

We now give an overview of the main ideas that go into estimating multivariate Gaussians (Theorem 1). We make two simplifications to ease the presentation. First, we assume the distribution has mean zero, and focus only on covariance estimation. Second, we elide the accuracy and privacy parameters, α and ε, since they do not play a central role in the discussion.

Suppose we are given samples $X_1, \ldots, X_n \sim \mathcal{N}(0, \Sigma)$ and want to output $\widehat{\Sigma}$ such that $\mathcal{N}(0, \widehat{\Sigma})$ is close to the true distribution in TV distance. More precisely, we want to guarantee

$$\|\widehat{\Sigma} - \Sigma\|_\Sigma = \|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - \mathbb{I}\|_F \lesssim \frac{1}{100}, \tag{1}$$

which implies closeness in TV distance. Then the standard solution is to use the empirical covariance $\widetilde{\Sigma} = \frac{1}{n}\sum_i X_i X_i^T$, guaranteeing $\|\widetilde{\Sigma} - \Sigma\|_\Sigma \lesssim \sqrt{d^2/n}$, satisfying (1) when $n \gtrsim d^2$.

First Attempt: The Gaussian Mechanism. The standard way to privately estimate a function f, is to use the *Gaussian mechanism*. In this case we are interested in the matrix-valued empirical covariance $\widetilde{\Sigma}(X_1, \ldots, X_n) = \frac{1}{n}\sum_i X_i X_i^T$, so the Gaussian mechanism would output

$$\widehat{\Sigma} = \widetilde{\Sigma}(X_1, \ldots, X_n) + Z \text{ where } Z \sim \mathcal{N}(0, O(\Delta_{\widetilde{\Sigma}}^2))^{d \times d}$$

is a random Gaussian *noise matrix* and $\Delta_{\widetilde{\Sigma}} = \max_{X \sim X'} \|\widetilde{\Sigma}(X) - \widetilde{\Sigma}(X')\|_F$ where $X \sim X'$ denotes that X, X' differ on at most one sample and $\Delta_{\widetilde{\Sigma}}$ is called the *global sensitivity*. Note that it is only a coincidence that the true distribution and the noise distribution are both Gaussian.

Unfortunately, the empirical covariance has *infinite sensitivity*, since X_i is an arbitrary vector, and thus changing a single sample can change the empirical covariance arbitrarily. The simplest way to address this problem is to assume that we have some *prior information* about Σ, namely that $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$. In this case we can *clamp* every sample so that $\|X_i\|_2^2 \le \tilde{O}(\kappa d)$. Once we do this, the sensitivity of the clamped empirical covariance is $\tilde{O}(\kappa d/n)$. However, if the data really came from a Gaussian, then the clamping will not have any effect, as long as there are no significant outliers. Thus, when we apply the Gaussian mechanism we obtain the guarantee

$$\left\|\widehat{\Sigma} - \Sigma\right\|_\Sigma \le \left\|\widehat{\Sigma} - \Sigma\right\|_F = \tilde{O}\left(\sqrt{\frac{d^2}{n}} + \frac{\kappa d^2}{n}\right), \tag{2}$$

where the first inequality follows from the assumption that $\Sigma \succeq \mathbb{I}$ and the second uses a standard bound on the Frobenius norm of a random Gaussian matrix. Unfortunately, (2) has a linear dependence on κ, meaning the number of samples has to be at least $\Omega(\kappa d^2)$ to achieve (1), and this analysis is tight when, say, $\Sigma = \mathbb{I}$.

Before moving on, we highlight the fact that, due to truncation, this algorithm ensures privacy *for any dataset*, even one not drawn from a Gaussian. This is a critical feature in private estimation that our algorithms will preserve.

Recursive Private Preconditioning. Looking at (2), we can see that the Gaussian mechanism actually has excellent accuracy when $\kappa = O(1)$, specifically the term corresponding to privacy vanishes faster than the term corresponding to sampling error. Thus, our approach to improve over the Gaussian mechanism is to private find a symmetric *preconditioner* A so that $\mathbb{I} \preceq A\Sigma A \preceq$

$O(1)\mathbb{I}$. Given such a matrix, we can apply the Gaussian mechanism to the data $AX_1, \ldots, AX_n \sim \mathcal{N}(0, A\Sigma A)$, and learn this transformed distribution using just $\tilde{O}(d^2)$ samples. Of course, to be useful, we need to find the private preconditioner using $\tilde{O}(d^2)$ samples.

In order to obtain such a matrix A, we essentially need a good multiplicative estimate of the covariance matrix Σ along every direction. However, the Gaussian mechanism makes this difficult because it adds an i.i.d. Gaussian matrix Z, which ignores the shape of Σ. Specifically, if we use the Gaussian mechanism and add the noise matrix Z, then unless we draw $\Omega(\kappa d^{3/2})$ samples, the directions of low variance will be completely overwhelmed by noise.

Our main observation is that when we use the Gaussian mechanism, even with just $n = \tilde{O}(d^{3/2})$ samples, we can still obtain a good enough estimate of Σ to make some progress. Specifically, we can find a matrix A such that $\mathbb{I} \preceq A\Sigma A \preceq \frac{7}{10}\kappa\mathbb{I}$. Given such a procedure, we can iterate $O(\log \kappa)$ times until the condition number is a constant.

To see how we make progress, we argue that if we draw just $\tilde{O}(d^{3/2})$ samples, and compute the noisy empirical covariance $\widehat{\Sigma} = \widetilde{\Sigma} + Z$, then the directions of $\widetilde{\Sigma}$ with *large variance* are approximately preserved, even though the directions with *small variance* will be overwhelmed by noise. We can leverage this fact in the following way. Suppose we see a direction of $\widehat{\Sigma}$ with variance at least $\kappa/2$. Then this direction cannot only appear to have large variance due to noise, it must also be large in Σ, meaning we have a good multiplicative approximation. On the other hand, suppose we see a direction of $\widehat{\Sigma}$ with variance at most $\kappa/2$. Then this direction may have almost no variance in Σ, but it cannot possibly have variance much larger than $\kappa/2$. Thus, we have discovered that this direction has less variance than our bound κ, meaning we can clamp the samples more aggressively in that direction to reduce the sensitivity of the estimator! More precisely, we compute the eigenvectors and eigenvalues of $\widehat{\Sigma}$ and let A be the matrix that partially projects out the eigenvectors with large eigenvalues. We can show that this matrix reduces the maximum variance by more than it reduces the minimum variance, thus (after some rescaling) $\mathbb{I} \preceq A\Sigma A \preceq \frac{7}{10}\kappa\mathbb{I}$, as desired.

Connection to Average Sensitivity. At a high level, the problem with the Gaussian mechanism is that it adds noise proportional to the *global sensitivity*, which is the maximum that changing one sample can change the estimate in the worst case. Intuitively, we'd like to add noise proportional to the *average sensitivity*, which is the amount that replacing one sample from the true distribution with an independent random sample from the true distribution changes the estimate. Simply adding noise proportional to average sensitivity is not private, however there are several techniques (e.g. Nissim et al. (2007); Dwork and Lei (2009)) that make it possible to add noise roughly proportional to average sensitivity for *univariate* statistics. Unfortunately, these techniques either do not apply, or are computationally inefficient for *multivariate* statistics. Our recursive private preconditioning technique can be viewed as a new technique for adding noise proportional to average sensitivity that is computationally efficient in high dimensions. Currently the technique is specific to Gaussian (or subgaussian) covariance estimation, but it would be interesting to understand how generally this technique can be applied.

1.2.2. PRIVATELY LEARNING PRODUCT DISTRIBUTIONS

Although learning Boolean product distributions is quite different from learning Gaussians, our algorithm uses a similar approach of recursively reducing the sensitivity of the natural estimator. Suppose we have a product distribution $\mathcal{P}$ over $\{0,1\}^d$ with mean $p = \mathbb{E}[\mathcal{P}]$ and want to output a product distribution $\widehat{\mathcal{P}}$ with mean $\widehat{p}$ that is close in TV distance. Bounding the TV distance between

$\mathcal{P}, \widehat{\mathcal{P}}$ requires some care, so to simplify this high-level discussion, we assume that $p \succeq 1/d$, in which case

$$\mathrm{d_{TV}}(\mathcal{P}, \widehat{\mathcal{P}}) \leq O\left(\sqrt{\sum_j \frac{(p_j - \widehat{p}_j)^2}{p_j}}\right) \tag{3}$$

is a reasonably tight bound. Without privacy, it would suffice to draw samples $X_1, \ldots, X_n \sim \mathcal{P}$ and compute the empirical mean $\widetilde{p} = \frac{1}{n}\sum_i X_i$ and output the corresponding product distribution, ensuring $\mathrm{d_{TV}}(\mathcal{P}, \widehat{\mathcal{P}}) \lesssim \sqrt{d/n}$.

First Attempt: The Gaussian Mechanism. As with Gaussians, the natural approach is to use the Gaussian mechanism, and compute $\widehat{p} = \widetilde{p} + Z$ where $Z \sim \mathcal{N}(0, (\sqrt{d}/n)^2)^d$ is a vector with i.i.d. Gaussian entries and $\sqrt{d}/n$ is the global sensitivity of the empirical mean. The problem with this mechanism is that if the mean of $\mathcal{P}$ is roughly $1/d$ in every coordinate, then we cannot upper bound (3) unless $\|Z\|_2 \lesssim 1/\sqrt{d}$, which requires $n = \Omega(d^{3/2})$.

Recursive Private Preconditioning. The key to improving the sample complexity is to recognize that the sensitivity analysis and the accuracy analysis cannot both be tight at the same time. That is, if $p = (\frac{1}{2}, \ldots, \frac{1}{2})$ then it suffices to have $\|Z\|_2 \lesssim 1$, which requires only $n = O(d)$ samples. On the other hand, if $p = (\frac{1}{d}, \ldots, \frac{1}{d})$, then we really need $\|Z\|_2 \lesssim 1/\sqrt{d}$. However, in this case each sample from $\mathcal{P}$ will satisfy $\|X_i\|_2 = O(\log d)$, so we reduce the sensitivity to just $O(\log d/n)$ by clamping the samples to have this norm, in which case we can obtain $\|Z\|_2 = 1/\sqrt{d}$ using just $n = \tilde{O}(d)$ samples.

The challenge is that p may have some coordinates that are roughly balanced and some that are very biased. If we could partition the coordinates into groups based on their bias, then we could apply an argument like the above on each group separately, but the challenge is to do this partitioning privately.

Similar to what we did for Gaussians, we can achieve this partitioning by starting with the Gaussian mechanism. Suppose we use the Gaussian mechanism to obtain $\widehat{p} = \widetilde{p} + Z$. Then if we draw $n = \tilde{O}(d)$ samples, we will have $\|Z\|_\infty = \tilde{O}(1/\sqrt{d})$. Now, consider two cases: for coordinates j such that $\widehat{p}_j \geq 1/4$, then we know that $\widetilde{p}_j$ is at least, say, $1/8$, so we have $(\widehat{p}_j - \widetilde{p}_j)^2/\widetilde{p}_j = O(1/d)$ which is a good enough estimate for that coordinate. Thus, we can lock in our estimate of these coordinates and move on. Now, the coordinates j we have left satisfy $\widehat{p}_j \leq 1/4$, and thus we know $\widetilde{p}_j$ is at most, say, $3/8$. Thus, if we restrict the distribution to just these coordinates, then we can clamp the norm of the samples and estimate again using less noise. Every time we iterate this process, we can reduce the upper bound on the bias by a constant factor until we get down to the case where all coordinates have bias at most $O(1/d)$, which we can handle separately. Iterating this approach $O(\log d)$ times requires us to add up estimation error and privacy loss across the different rounds, but this only incurs additional polylogarithmic factors.

1.3. Organization

In the body of this extended abstract, we go into more detail about our upper bounds for Gaussian estimation. We defer preliminaries (which include fairly standard definitions and properties of differential privacy), proofs of our upper bounds for Gaussian estimation, and our upper bounds for product distributions and lower bounds to the appendix.

2. Private Covariance Estimation for Gaussians

In this section we present our algorithm for privately estimating the covariance of an unknown Gaussian. Suppose we are given i.i.d. samples $X_1, \ldots, X_n \sim \mathcal{N}(0, \Sigma)$ where $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$. Our goal is to privately output $\widehat{\Sigma}$ so that $\|\Sigma - \widehat{\Sigma}\|_\Sigma \leq O(\alpha)$, where $\|A\|_\Sigma = \|\Sigma^{-1/2} A \Sigma^{-1/2}\|_F$. Here the matrix square root denotes any possible square root; it is trivial to check that all such choices are equivalent. By Lemma 21, this condition ensures $d_{\mathrm{TV}}(\mathcal{N}(0,\Sigma), \mathcal{N}(0,\widehat{\Sigma})) \leq O(\alpha)$. We cover some useful concentration inequalities in Section C.1, and some deterministic regularity conditions of our dataset that we condition on in Section C.2.

2.1. A Simple Algorithm for Well Conditioned Gaussians

We first consider the following simple algorithm: remove all points whose norm exceeds a certain threshold, then compute the empirical covariance of the resulting data set, and perturb the empirical covariance with noise to preserve privacy. This algorithm will have nearly-optimal dependence on most parameters, however, it will have a polynomial dependence on the condition number. Pseudocode for this algorithm is given in Algorithm 1.

Algorithm 1: Naive Private Gaussian Covariance Estimation $\textsc{NaivePCE}_{\rho,\beta,\kappa}(X)$
Input: A set of n samples $X_1, \ldots, X_n$ from an unknown Gaussian. Parameters $\rho, \beta, \kappa > 0$
Output: A covariance matrix M.

Let $S \leftarrow \left\{ i \in [n] : \|X_i\|_2^2 \leq O(d\kappa \log(n/\beta)) \right\}$
Let $\sigma \leftarrow \Theta\left(\frac{d\kappa \log(\frac{n}{\beta})}{n\rho^{1/2}} \right)$
Let $M' \leftarrow \frac{1}{n} \sum_{i \in S} X_i X_i^\top + N$ where $N \sim \mathrm{GUE}(\sigma^2)$
Let M be the projection of M' into the set of PSD matrices.
Return M

The following is an immediate consequence of Lemma 27 (in Section C.3), seen by noting that $\|\Sigma - M\|_\Sigma = \|\Sigma^{-1/2} N \Sigma^{-1/2} + N'\|_F \leq \|N\|_\Sigma + \|N'\|_F$.

Theorem 6 *For every $\rho, \beta, \kappa > 0$, the algorithm $\textsc{NaivePCE}_{\rho,\beta,\kappa}$ is ρ-zCDP and, when given $n = O\left(\frac{d^2 + \log(\frac{1}{\beta})}{\alpha^2} + \frac{\kappa d^2 \mathrm{polylog}(\frac{\kappa d}{\alpha\beta\rho})}{\alpha\rho^{1/2}} \right)$, samples from $\mathcal{N}(0,\Sigma)$ satisfying $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$, with probability at least $1 - O(\beta)$, it returns M such that $\|\Sigma - M\|_\Sigma \leq O(\alpha)$.*

2.2. A Private Recursive Preconditioner

When κ is a constant, Theorem 6 says that NAIVEPCE privately estimates the covariance of a Gaussian with little overhead compared to non-private estimation. In this section we will show how to nearly eliminate the dependence on the covariance by privately learning a *preconditioner* A such that $\mathbb{I} \preceq A\Sigma A \preceq 1000\mathbb{I}$. Once we have this preconditioner, we can reduce the condition number of the distribution to a constant. In this state, we can apply NAIVEPCE to estimate the covariance at no cost in κ.

2.2.1. REDUCING THE CONDITION NUMBER BY A CONSTANT

Our preconditioner works recursively. The main ingredient in the recursive construction is an algorithm, WEAKPPC (Algorithm 2) that privately estimates a matrix A such that the condition number of $A\Sigma A$ improves over that of Σ by a constant factor. Once we have this primitive we can apply it recursively in a straightforward way. Note that in Algorithm 2, when we apply NAIVEPCE to obtain a weak estimate of Σ, we use too few samples for NAIVEPCE to obtain a good estimate of Σ on its own.

Algorithm 2: Private Preconditioning $\text{WEAKPPC}_{\rho,\beta,\kappa,K}(X)$
Input: A set of n samples $X_1, \ldots, X_n$ from an unknown Gaussian. Parameters $\rho, \beta, \kappa, K > 0$.
Output: A symmetric matrix A.
Let $Z \leftarrow \text{NAIVEPCE}_{\rho,\beta,\kappa}(X_1, \ldots, X_n)$
Let $(\lambda_1, v_1), \ldots, (\lambda_d, v_d)$ be the eigenvalues and the corresponding eigenvectors of Z
Let $V \leftarrow \text{span}\left(\left\{v_i : \lambda_i \geq \frac{\kappa}{2}\right\}\right) \subseteq \mathbb{R}^d$
Return the pair (V, A) where $A = \frac{1}{\sqrt{K}}\Pi_V + \Pi_{V^\perp}$

The guarantee of Algorithm 2 is captured in the following theorem. In Section C.4, we state a more precise and complete version as Theorem 28, and prove it.

Theorem 7 *For every $\rho, \beta, \kappa, K > 0$, $\text{WEAKPPC}_{\rho,\beta,\kappa,K}(X)$ satisfies ρ-zCDP and, if $X_1, \ldots, X_n$ are sampled i.i.d. from $\mathcal{N}(0, \Sigma)$ for $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$ where $\kappa > 1000$, K is an appropriate constant, and $n = O\left(\frac{d^{3/2}\text{polylog}(\frac{d}{\rho\beta})}{\rho^{1/2}}\right)$, then with probability at least $1 - O(\beta)$ it outputs A such that $\mathbb{I} \preceq (1.1A)\Sigma(1.1A) \preceq \frac{7}{10}\kappa\mathbb{I}$.*

2.2.2. RECURSIVE PRECONDITIONING

Once we have WEAKPPC, we can apply it recursively to obtain a private preconditioner, PPC (Algorithm 3) that reduces the condition number down to a constant.

Algorithm 3: Privately estimating covariance $\text{PPC}_{\rho,\beta,\kappa}(X)$
Input: A set of n samples $X_1, \ldots, X_n$ from an unknown Gaussian $\mathcal{N}(0, \Sigma)$. Parameters $\rho, \beta, \kappa > 0$
Output: A symmetric matrix A
Let $T \leftarrow O(\log \kappa)$ $\quad \rho' \leftarrow \rho/T$ $\quad \beta' \leftarrow \beta/T$
Let K be the constant from Theorem 28
Let $\kappa^{(1)} \leftarrow \kappa$ and let $X_i^{(1)} \leftarrow X_i$ for $i = 1, \ldots, n$
For $t = 1, \ldots, T$
  Let $\tilde{A}^{(t)} \leftarrow \text{WEAKPPC}_{\rho',\beta',\kappa^{(t)},K}(X_1^{(t)}, \ldots, X_n^{(t)})$, and let $A^{(t)} \leftarrow 1.1\tilde{A}^{(t)}$
  Let $\kappa^{(t+1)} \leftarrow 0.7\kappa^{(t)}$
  Let $X_i^{(t+1)} \leftarrow A^{(t)}X_i^{(t)}$ for $i = 1, \ldots, n$
Return *The matrix* $A = \prod_{t=1}^{T} A^{(t)}$.

Theorem 8 *For every $\rho, \alpha, \beta, \kappa > 0$, the algorithm $\text{PPC}_{\rho,\beta,\kappa}$ satisfies ρ-zCDP, and when given $n = O\left(\frac{d^{3/2}\log^{1/2}(\kappa)\text{polylog}(\frac{d\log\kappa}{\rho\beta})}{\rho^{1/2}}\right)$, samples $X_1, \ldots, X_n \sim \mathcal{N}(0, \Sigma)$ for $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$, with probability $1 - O(\beta)$ it outputs a symmetric matrix A such that $\mathbb{I} \preceq A\Sigma A \preceq 1000\mathbb{I}$.*

2.3. Putting It All Together

We can now combine our private preconditioning algorithm with the naïve algorithm for covariance estimation to obtain a complete algorithm for covariance estimation.

Algorithm 4: Private Covariance Estimator $\mathrm{PGCE}_{\rho,\beta,\kappa}(X)$
Input: Samples $X_1,\dots,X_n$ from an unknown Gaussian $\mathcal{N}(0,\Sigma)$. Parameters $\rho,\beta,\kappa>0$.
Output: A matrix $\widehat{\Sigma}$ such that $\|\Sigma-\widehat{\Sigma}\|_\Sigma\le\alpha$.
Let $\rho'\leftarrow\rho/2$ and $\beta'\leftarrow\beta/2$
Let $A\leftarrow\mathrm{PPC}_{\rho',\beta',\kappa}(X_1,\dots,X_n)$ be the private preconditioner
Let $Y_i\leftarrow AX_i$ for $i=1,\dots,n$
Let $\widetilde{\Sigma}\leftarrow\textsc{NaivePCE}_{\rho',\beta',1000}(Y_1,\dots,Y_n)$
Return $\widehat{\Sigma}=A^{-1}\widetilde{\Sigma}A^{-1}$

This algorithm has the following guarantee.

Theorem 9 *For every $\rho,\beta,\kappa>0$, the algorithm $\mathrm{PGCE}_{\rho.\beta,\kappa}(X)$ is ρ-zCDP and, when given $n=O\left(\frac{d^2+\log(\frac{1}{\beta})}{\alpha^2}+\frac{d^2\mathrm{polylog}(\frac{d}{\alpha\beta\rho})}{\alpha\rho^{1/2}}+\frac{d^{3/2}\log^{1/2}(\kappa)\mathrm{polylog}(\frac{d\log\kappa}{\rho\beta})}{\rho^{1/2}}\right)$, $X_1,\dots,X_n\sim\mathcal{N}(0,\Sigma)$ for $\mathbb{I}\preceq\Sigma\preceq\kappa\mathbb{I}$, with probability $1-O(\beta)$, it outputs $\widehat{\Sigma}$ such that $\|\Sigma-\widehat{\Sigma}\|_\Sigma\le O(\alpha)$.*

3. Private Mean Estimation for Gaussians

Suppose we are given i.i.d. samples $X_1,\dots,X_n$, such that $X_i\sim\mathcal{N}(\mu,\Sigma)$ where $\|\mu\|_2\le R$ is an unknown mean and $I\preceq\Sigma\preceq\kappa I$ is an unknown covariance matrix. Our goal is to find an estimate $\widehat{\mu}$ such that $\|\mu-\widehat{\mu}\|_\Sigma\le O(\alpha)$ where $\|v\|_\Sigma=\|\Sigma^{-1/2}v\|_2$ is the Mahalanobis distance with respect to the covariance Σ. This guarantee ensures $\mathrm{d_{TV}}(\mathcal{N}(\mu,\Sigma),\mathcal{N}(\widehat{\mu},\Sigma))=O(\alpha)$.

When κ is a constant, we can obtain such a guarantee in a relatively straightforward way by applying the mean-estimation procedure for univariate Gaussians due to Karwa and Vadhan Karwa and Vadhan (2018) to each coordinate. To handle large values of κ, we combine their procedure with our procedure for privately learning a strong approximation to the covariance matrix.

3.1. Mean Estimation for Well-Conditioned Gaussians

We start with the following algorithm for learning the mean of a univariate Gaussian, which is a trivial variant of the algorithm of Karwa and Vadhan Karwa and Vadhan (2018) to the definition of zCDP.

Theorem 10 (Variant of Karwa and Vadhan (2018)) *For every $\varepsilon,\delta,\alpha,\beta,R,\sigma>0$, there is an $\frac{\varepsilon^2}{2}$-zCDP algorithm $\mathrm{KVMean}_{\varepsilon,\alpha,\beta,R,\kappa}(X)$ and an $n=O\left(\frac{\log(\frac{1}{\beta})}{\alpha^2}+\frac{\log(\frac{\log R}{\alpha\beta\varepsilon})}{\alpha\varepsilon}+\frac{\log^{1/2}(\frac{R}{\beta})}{\varepsilon}\right)$, such that if $X=(X_1,\dots,X_n)$ are i.i.d. samples from $\mathcal{N}(\mu,\sigma^2)$ for $|\mu|\le R$ and $1\le\sigma^2\le\kappa$ then, with probability at least $1-\beta$, KVMean outputs $\widehat{\mu}$ such that $|\mu-\widehat{\mu}|\le\alpha\kappa$. This algorithm also satisfies ρ-zCDP for $\rho=\frac{\varepsilon^2}{2}$ in which case we denote the algorithm $\mathrm{KVMean}_{\rho,\alpha,\beta,R,\kappa}(X)$.*

Note that the algorithm only needs an *upper bound* κ on the true variance σ^2 as a parameter. However, since the error guarantees depend on this upper bound, the upper bound needs to be reasonably tight in order to get a useful estimate of the mean.

We will describe our naïve algorithm for the case of ρ-zCDP since the parameters are cleaner. We could also obtain an (ε, δ)-DP version using the (ε, δ)-DP version of NAIVEPME and setting parameters appropriately.

Algorithm 5: Naïve Private Mean Estimator $\text{NAIVEPME}_{\rho,\alpha,\beta,R,\kappa}(X)$
Input: Samples $X_1, \ldots, X_n \in \mathbb{R}^d$ from a d-variate Gaussian. Parameters $\rho, \alpha, \beta, R, \kappa > 0$.
Output: A vector $\widehat{\mu}$ such that $\|\mu - \widehat{\mu}\|_\Sigma \leq \alpha$.
Let $\rho' \leftarrow \rho/d \quad \alpha' \leftarrow \alpha/\kappa\sqrt{d} \quad \beta' \leftarrow \beta/d$
For $j = 1, \ldots, d$
  Let $\widehat{\mu}_j \leftarrow \text{KVMEAN}_{\rho',\alpha',\beta',R,\kappa}(X_{1,j}, \ldots, X_{n,j})$
Return $\widehat{\mu} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_d)$

Theorem 11 *For every $\rho, \alpha, \beta, R, \kappa > 0$, the algorithm* $\text{NAIVEPME}_{\rho,\alpha,\beta,R,\kappa}$ *is ρ-zCDP and there is an $n = O\left(\frac{\kappa^2 d \log(\frac{d}{\beta})}{\alpha^2} + \frac{\kappa d \log(\frac{\kappa d \log R}{\alpha\beta\rho})}{\alpha\rho^{1/2}} + \frac{\sqrt{d}\log^{1/2}(\frac{Rd}{\beta})}{\rho^{1/2}}\right)$, such that if $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \Sigma)$ for $\|\mu\|_2 \leq R$ and $I \preceq \Sigma \preceq \kappa I$ then, with probability at least $1 - \beta$,* NAIVEPME *outputs $\widehat{\mu}$ such that $\|\mu - \widehat{\mu}\|_\Sigma \leq \alpha$.*

3.2. An Algorithm for General Gaussians

If Σ were known, then we could easily perform mean estimation without dependence on κ simply by applying $\Sigma^{-1/2}$ to each sample and running NAIVEPME. Specifically, if $X_i \sim \mathcal{N}(\mu, \Sigma)$ then $\Sigma^{-1/2}X_i \sim \mathcal{N}(\Sigma^{-1/2}\mu, I)$. Then applying NAIVEPME we would obtain $\widehat{\mu}$ such that $\|\mu - \widehat{\mu}\|_\Sigma = \|\Sigma^{-1/2}(\mu - \widehat{\mu})\|_2 \leq \alpha$ and the sample complexity would be independent of κ. Using the private preconditioner from the previous section, we can obtain a good enough approximation to $\Sigma^{-1/2}$ to carry out this reduction.

Algorithm 6: Private Mean Estimator $\text{PME}_{\rho,\alpha,\beta,R,\kappa}(X)$
Input: Samples $X_1, \ldots, X_{3n} \in \mathbb{R}^d$ from a d-variate Gaussian $\mathcal{N}(\mu, \Sigma)$ with unknown mean and covariance. Parameters $\rho, \alpha, \beta, R, \kappa > 0$.
Output: A vector $\widehat{\mu}$ such that $\|\mu - \widehat{\mu}\|_\Sigma \leq \alpha$.
For $i = 1, \ldots, n$, let $Z_i = \frac{1}{\sqrt{2}}(X_{2i} - X_{2i-1})$
Let $A \leftarrow \text{PPC}_{\rho,\beta,\kappa}(Z_1, \ldots, Z_n)$
For $i = 1, \ldots, n$, let $Y_i = AX_{2n+i}$
Let $\widetilde{\mu} \leftarrow \text{NAIVEPME}_{\rho,\alpha,\beta,1000R,1000}(Y_1, \ldots, Y_n)$
Return $\widehat{\mu} \leftarrow A^{-1}\widetilde{\mu}$

We capture the properties of PME in the following theorem

Theorem 12 *For every $\rho, \alpha, \beta, R, \kappa > 0$, the algorithm* $\text{PME}_{\rho,\alpha,\beta,R,\kappa}$ *is 2ρ-zCDP and there is an $n = O\left(\frac{d\log(\frac{d}{\beta})}{\alpha^2} + \frac{d\log(\frac{d\log R}{\alpha\beta\rho})}{\alpha\rho^{1/2}} + \frac{\sqrt{d}\log^{1/2}(\frac{Rd}{\beta})}{\rho^{1/2}} + n_{\text{PPC}}\right)$ such that if $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \Sigma)$ for $\|\mu\|_2 \leq R$ and $I \preceq \Sigma \preceq \kappa I$ then, with probability at least $1 - 2\beta$,* PME *outputs $\widehat{\mu}$ such that $\|\mu - \widehat{\mu}\|_\Sigma \leq \alpha$. In the above, n_{PPC} is the sample complexity required by* $\text{PPC}_{\rho,\beta,\kappa}$ *(Theorem 8).*

Acknowledgments

Work done when GK was a graduate student at MIT, supported by NSF Award CCF-1617730, CCF-1650733, CCF-1741137, and ONR N00014-12-1-0999, and when a Microsoft Research Fellow, as part of the Simons-Berkeley Research Fellowship program. Work done when JL was a graduate student at MIT, supported by NSF Award CCF-1453261 (CAREER), CCF-1565235, a Google Faculty Research Award, and an NSF Graduate Research Fellowship, and when a VMware Research Fellow, as part of the Simons-Berkeley Research Fellowship program. Part of this work was also performed while JL was an intern at Google. VS supported by NSF award CCF-1750640. JU supported by NSF awards CCF-1750640 and NSF awards CCF-1718088 and CNS-1816028, and a Google Faculty Research Award.

References

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., 2018a.

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., 2018b.

Maryam Aliakbarpour, Ilias Diakonikolas, and Ronitt Rubinfeld. Differentially private identity and closeness testing of discrete distributions. In *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pages 169–178. JMLR, Inc., 2018.

Mitali Bafna and Jonathan Ullman. The price of selection in differential privacy. In *Proceedings of the 30th Annual Conference on Learning Theory*, COLT '17, pages 151–168, 2017.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 464–473, Washington, DC, USA, 2014. IEEE Computer Society.

Amos Beimel, Hai Brenner, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. *Machine Learning*, 94(3):401–437, 2014.

Andrew C. Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941.

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, 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, Thomas Steinke, and Jonathan Ullman. Make up your mind: The price of online queries in differential privacy. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, pages 1306–1325, Philadelphia, PA, USA, 2017. SIAM.

Bryan Cai, Constantinos Daskalakis, and Gautam Kamath. Priv'it: Private and sample efficient identity testing. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 635–644. JMLR, Inc., 2017.

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.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM Symposium on the Theory of Computing*, STOC '17, pages 47–60, New York, NY, USA, 2017. ACM.

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.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The minimax learning rate of normal and Ising undirected graphical models. *arXiv preprint arXiv:1806.06887*, 2018.

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.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a Gaussian: Getting optimal error, efficiently. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, Philadelphia, PA, USA, 2018. SIAM.

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

Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, pages 202–210, New York, NY, USA, 2003. ACM.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st Annual ACM Symposium on the Theory of Computing*, STOC '09, pages 371–380, New York, NY, USA, 2009. ACM.

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

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, Moni Naor, Omer Reingold, Guy N. Rothblum, and Salil Vadhan. On the complexity of differentially private data release: Efficient algorithms and hardness results. In *Proceedings of the 41st Annual ACM Symposium on the Theory of Computing*, STOC '09, pages 381–390, New York, NY, USA, 2009. ACM.

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, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the 46th Annual ACM Symposium on the Theory of Computing*, STOC '14, pages 11–20, New York, NY, USA, 2014. ACM.

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.

Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4(1):61–84, 2017.

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.

Carl-Gustaf Esseen. On the Liapounoff limit of error in the theory of probability. *Arkiv för matematik, astronomi och fysik*, 28A(2):1–19, 1942.

Marco Gaboardi and Ryan Rogers. Local private hypothesis testing: Chi-square tests. In *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pages 1626–1635. JMLR, Inc., 2018.

Marco Gaboardi, James Honaker, Gary King, Jack Murtagh, Kobbi Nissim, Jonathan Ullman, and Salil Vadhan. Psi (ψ): A private data sharing interface. *arXiv preprint arXiv:1609.04340*, 2016a.

Marco Gaboardi, Hyun-Woo Lim, Ryan M. Rogers, and Salil P. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *Proceedings of the 33rd International Conference on Machine Learning*, ICML '16, pages 1395–1403. JMLR, Inc., 2016b.

Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the 42nd Annual ACM Symposium on the Theory of Computing*, STOC '10, pages 705–714, New York, NY, USA, 2010. ACM.

Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 454–463, Washington, DC, USA, 2014. IEEE Computer Society.

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.

Kazuya Kakizaki, Jun Sakuma, and Kazuto Fukuchi. Differentially private chi-squared test by unit circle mechanism. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 1761–1770. JMLR, Inc., 2017.

Gautam Kamath. A lower bound for private covariance estimation. Note, available upon request, 2019.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. *arXiv preprint arXiv:1805.00216*, 2018.

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.

Daniel Kifer and Ryan M. Rogers. A new class of private chi-square tests. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, AISTATS '17, pages 991–1000. JMLR, Inc., 2017.

Lucas Kowalczyk, Tal Malkin, Jonathan Ullman, and Mark Zhandry. Strong hardness of privacy from weak traitor tracing. In *Proceedings of the 14th Conference on Theory of Cryptography*, TCC '16-B, pages 659–689, Berlin, Heidelberg, 2016. Springer.

Lucas Kowalczyk, Tal Malkin, Jonathan Ullman, and Daniel Wichs. Hardness of non-interactive differential privacy from one-way functions. In *Proceedings of the 38th Annual International Cryptology Conference*, CRYPTO '18, Berlin, Heidelberg, 2018. Springer.

Kevin A. Lai, Anup B. Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 665–674, Washington, DC, USA, 2016. IEEE Computer Society.

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

Prashanth Mohan, Abhradeep Thakurta, Elaine Shi, Dawn Song, and David Culler. GUPT: Privacy preserving data analysis made easy. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 349–360, New York, NY, USA, 2012. ACM.

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.

Or Sheffet. Differentially private ordinary least squares. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 3105–3114. JMLR, Inc., 2017.

Or Sheffet. Locally private hypothesis testing. In *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pages 4605–4614. JMLR, Inc., 2018.

I.G. Shevtsova. An improvement of convergence rate estimates in the Lyapunov theorem. *Doklady Mathematics*, 82(3):862–864, 2010.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Proceedings of the 38th IEEE Symposium on Security and Privacy*, SP '17, pages 3–18, Washington, DC, USA, 2017. IEEE Computer Society.

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.

Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pages 45:1–45:21, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

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, 2017a.

Thomas Steinke and Jonathan Ullman. Tight lower bounds for differentially private selection. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '17, pages 552–563, Washington, DC, USA, 2017b. IEEE Computer Society.

Terence Tao. *Topics in random matrix theory*, volume 132 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

Jonathan Ullman. Answering $n^{2+o(1)}$ counting queries with differential privacy is hard. *SIAM Journal on Computing*, 45(2):473–496, 2016.

Jonathan Ullman and Salil Vadhan. PCPs and the hardness of generating private synthetic data. In *Proceedings of the 8th Conference on Theory of Cryptography*, TCC '11, pages 400–416, Berlin, Heidelberg, 2011. Springer.

Yue Wang, Jaewoo Lee, and Daniel Kifer. Revisiting differentially private hypothesis tests for categorical data. *arXiv preprint arXiv:1511.03376*, 2015.

Appendix A. Additional Related Work

Differentially Private Learning and Statistics. The most directly comparable papers to ours are recent results on learning *low-dimensional* statistics. In addition to the aforementioned work of Karwa and Vadhan (2018), Bun et al. (2015) showed how to private learn an arbitrary distribution in Kolmogorov distance, which is weaker than TV distance, with almost no increase in sample complexity. Diakonikolas et al. (2015) extended that work to give a practical algorithm for learning structured one-dimensional distributions in TV distance.

An elegant work of Smith (2011) showed how to estimate arbitrary *asymptotically normal* statistics with only a small increase in sample complexity compared to non-private estimation. Technically, this work doesn't apply to covariance estimation or estimating sparse product distributions, for which the asymptotic distribution is not normal. More fundamentally, this algorithm learns a high-dimensional distribution one coordinate at a time, which is quite costly for the distributions we consider here.

Subsequent to our work, Cai et al. (2019) studied mean and covariance estimation of subgaussian distributions (as well as sparse mean estimation, which we don't consider in this work) subject to differential privacy, but in a setting with strong *a priori* bounds on the parameters. In particular, they prove a lower bound for mean estimation of subgaussian distributions that is incomparable to our Theorems 52 and 56. It is quantitatively larger by a factor of $\log^{1/2}(1/\delta)$, but it only holds for the more general class of subgaussian non-product distributions. In particular they use a reduction from Steinke and Ullman (2017a) to boost one of Theorem 52 or 56 in a way that fails to preserve the property of being Gaussian or being a product distribution.

Covariance Estimation. For covariance estimation, the works closest to ours are that of Dwork et al. (2014) and Sheffet (2017). Their algorithms require that the norm of the data be bounded, and the sample complexity depends polynomially on this bound. In contrast, our algorithms have either mild or no dependence on the norm of the data.

Robust Statistical Estimation on High-Dimensional Data. Recently, there has been significant interest in the computer science community in robustly estimating distributions (Diakonikolas et al., 2016; Lai et al., 2016; Charikar et al., 2017; Diakonikolas et al., 2017, 2018; Steinhardt et al., 2018), where the goal is to estimate some distribution from samples even when a constant fraction of the samples may be corrupted by an adversary. As observed by Dwork and Lei (2009), differentially private estimation and robust estimation both seek to minimize the influence of outliers, and thus there is a natural conceptual connection between these two problems. Technically, the two problems are incomparable. Differential privacy seeks to limit the influence of outliers in a very strong sense, and without making any assumptions on the data, but only when up to $O(1/\varepsilon)$ samples are corrupted. In contrast, robust estimation limits the influence of outliers in a weaker sense, and only when the remaining samples are chosen from a nice distribution, but tolerates up to $\Omega(n)$ corruptions.

Differentially Private Testing. There have also been a number of works on differentially private *hypothesis testing*. For example, Wang et al. (2015); Gaboardi et al. (2016b); Kifer and Rogers (2017); Cai et al. (2017); Kakizaki et al. (2017) gave private algorithms for goodness-of-fit testing,

closeness, and independence testing. Recently, Acharya et al. (2018b) and Aliakbarpour et al. (2018) have given essentially optimal algorithms for goodness-of-fit and closeness testing of arbitrary distributions. Acharya et al. (2018a) designed nearly optimal algorithms for estimating properties like support size and entropy. Gaboardi and Rogers (2018); Sheffet (2018) study hypothesis testing in the local differential privacy setting. All these works consider testing of *arbitrary* distributions, and so they necessarily have sample complexity growing exponentially in the dimension.

Privacy Attacks and Lower Bounds. A complementary line of work has established limits on the accuracy of private algorithms for high-dimensional learning. For example, Dwork et al. (2015) (building on Bun et al. (2014); Hardt and Ullman (2014); Steinke and Ullman (2015, 2017a)) designed a *robust tracing attack* that can infer sensitive information about individuals in a dataset using highly noisy statistical information about the dataset. These attacks apply to nice distributions like product distributions and Gaussians, but require that the dataset be too small to learn the underlying distribution in total variation distance, and thus do not contradict our results. These attacks apply to a number of learning problems, such as PCA (Dwork et al., 2014), ERM (Bassily et al., 2014), and variable selection (Bafna and Ullman, 2017; Steinke and Ullman, 2017b). Similar attacks lead to computational hardness results for differentially private algorithms for high-dimensional data (Dwork et al., 2009; Ullman and Vadhan, 2011; Ullman, 2016; Kowalczyk et al., 2016, 2018), albeit for learning problems that encode certain cryptographic functionalities.

Appendix B. Preliminaries

A *dataset* $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$ is a collection of elements from some *universe*. We say that two datasets $X, X' \in \mathcal{X}^n$ are *neighboring* if they differ on at most a single entry, and denote this by $X \sim X'$. Informally, differential privacy requires that for every pair of datasets $X, X' \in \mathcal{X}^n$ that differ on at most a single entry, the distributions $M(X)$ and $M(X')$ are close. In our work we consider a few different variants of differential privacy. The first is the standard variant of differential privacy.

Definition 13 (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 $X, X' \in \mathcal{X}^n$,*

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

The second variant is so-called *concentrated differential privacy* (Dwork and Rothblum, 2016), specifically the refinement *zero-mean concentrated differential privacy* (Bun and Steinke, 2016).

Definition 14 (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 $X, X' \in \mathcal{X}^n$,*

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

where $D_\alpha(M(X)||M(X'))$ is the α-Rényi divergence between $M(X)$ and $M(X')$.[3]

Both of these definitions are closed under post-processing

3. 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 15 (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.*

Qualitatively, DP with $\delta = 0$ is stronger than zCDP, which is stronger than DP with $\delta > 0$. These relationships are quantified in the following lemma.

Lemma 16 (Relationships Between Variants of DP (Bun and Steinke, 2016)) *For every $\varepsilon \geq 0$,*

1. *If M satisfies $(\varepsilon, 0)$-DP, then M is $\frac{\varepsilon^2}{2}$-zCDP.*
2. *If M satisfies $\frac{\varepsilon^2}{2}$-zCDP, then M satisfies $(\frac{\varepsilon^2}{2} + \varepsilon\sqrt{2\log(\frac{1}{\delta})}, \delta)$-DP for every $\delta > 0$.*

Note that the parameters for DP and zCDP are on different scales, with (ε, δ)-DP roughly commensurate with $\frac{\varepsilon^2}{2}$-zCDP.

Composition. A crucial property of all the variants of differential privacy is that they can be composed adaptively. By adaptive composition, we mean a sequence of algorithms $M_1(X), \ldots, M_T(X)$ where the algorithm $M_t(X)$ may also depend on the outcomes of the algorithms $M_1(X), \ldots, M_{t-1}(X)$.

Lemma 17 (Composition of DP (Dwork et al., 2006, 2010; Bun and Steinke, 2016)) *If M is an adaptive composition of differentially private algorithms $M_1, \ldots, M_T$, then the following all hold:*

1. *If $M_1, \ldots, M_T$ are $(\varepsilon_1, \delta_1), \ldots, (\varepsilon_T, \delta_T)$-DP then M is (ε, δ)-DP for*
$$\varepsilon = \sum_t \varepsilon_t \quad \text{and} \quad \delta = \sum_t \delta_t$$
2. *If $M_1, \ldots, M_T$ are $(\varepsilon_0, \delta_1), \ldots, (\varepsilon_0, \delta_T)$-DP for some $\varepsilon_0 \leq 1$, then for every $\delta_0 > 0$, M is (ε, δ)-DP for*
$$\varepsilon = \varepsilon_0\sqrt{6T\log(1/\delta_0)} \quad \text{and} \quad \delta = \delta_0 + \sum_t \delta_t$$
3. *If $M_1, \ldots, M_T$ are $\rho_1, \ldots, \rho_T$-zCDP then M is ρ-zCDP for $\rho = \sum_t \rho_t$.*

Note that the first and the third properties say that (ε, δ)-DP and ρ-zCDP compose linearly—the parameters simply add up. The second property says that (ε, δ)-DP actually composes sublinearly—the parameter ε grows roughly with the square root of the number of steps in the composition, provided we allow a small increase in δ.

The Gaussian Mechanism. Our algorithms will extensively use the well known and standard Gaussian mechanism to ensure differential privacy.

Definition 18 (ℓ_2-Sensitivity) *Let $f : \mathcal{X}^n \to \mathbb{R}^d$ be a function, its ℓ_2-*sensitivity *is*
$$\Delta_f = \max_{X \sim X' \in \mathcal{X}^n} \|f(X) - f(X')\|_2$$

Lemma 19 (Gaussian Mechanism) *Let $f : \mathcal{X}^n \to \mathbb{R}^d$ be a function with ℓ_2-sensitivity Δ_f. Then the Gaussian mechanism*

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

satisfies ρ-zCDP.

In order to prove accuracy, we will use the following standard tail bounds for Gaussian random variables.

Lemma 20 *If $Z \sim N(0, \sigma^2)$ then for every $t > 0$, $\mathbb{P}[|Z| > t\sigma] \leq 2e^{-t^2/2}$.*

Parameter Estimation to Distribution Estimation. In this work, our goal is to estimate some underlying distribution in total variation distance. For both Gaussian and product distributions, we will achieve this by estimating the parameters of the distribution, and we argue that a distribution from the class with said parameters will be accurate in total variation distance. For product distributions, we require an estimate of the parameters, which is accurate in terms of a type of chi-squared distance; this is shown in the proof of Theorem 32. For Gaussian distributions, the parameter estimate we require is slightly more difficult to describe. For a vector x, define $\|x\|_\Sigma = \|\Sigma^{-1/2}x\|_2$. Similarly, for a matrix X, define $\|X\|_\Sigma = \|\Sigma^{-1/2}X\Sigma^{-1/2}\|_F$. With these two norms in place, we have the following lemma, which is a combination of Corollaries 2.13 and 2.14 of Diakonikolas et al. (2016).

Lemma 21 *Let $\alpha \geq 0$ be smaller than some absolute constant. Suppose that $\|\mu - \hat{\mu}\|_\Sigma \leq \alpha$, and $\|\Sigma - \hat{\Sigma}\|_\Sigma \leq \alpha$, where $\mathcal{N}(\mu, \Sigma)$ is a Gaussian distribution in $\mathbb{R}^d$, $\hat{\mu} \in \mathbb{R}^d$, and $\Sigma \in \mathbb{R}^{d\times d}$ is a PSD matrix. Then $d_{\mathrm{TV}}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\hat{\mu}, \hat{\Sigma})) \leq O(\alpha)$.*

Appendix C. Additional Details for Gaussian Estimation

C.1. Useful Concentration Inequalities

We will need several facts about Gaussians and Gaussian matrices. Throughout this section, let $\mathrm{GUE}(\sigma^2)$ denote the distribution over $d \times d$ symmetric matrices M where for all $i \leq j$, we have $M_{ij} \sim \mathcal{N}(0, \sigma^2)$ i.i.d.. From basic random matrix theory, we have the following guarantee.

Theorem 22 (see e.g. Tao (2012) Corollary 2.3.6) *For d sufficiently large, there exist absolute constants $C, c > 0$ such that*

$$\mathbb{P}_{M\sim\mathrm{GUE}(\sigma^2)}\left[\|M\|_2 > A\sigma\sqrt{d}\right] \leq C\exp(-cAd)$$

for all $A \geq C$.

We also require the following, well known tail bound on quadratic forms on Gaussians.

Theorem 23 (Hanson-Wright Inequality (see e.g. Laurent and Massart (2000))) *Let $X \sim \mathcal{N}(0, \mathbb{I})$ and let A be a $d \times d$ matrix. Then, for all $t > 0$, the following two bounds hold:*

$$\mathbb{P}\left[X^\top AX - \mathrm{tr}(A) \geq 2\|A\|_F\sqrt{t} + 2\|A\|_2 t\right] \leq \exp(-t) \tag{4}$$

$$\mathbb{P}\left[X^\top AX - \mathrm{tr}(A) \leq -2\|A\|_F\sqrt{t}\right] \leq \exp(-t) \tag{5}$$

As a special case of the above inequality, we also have

Fact 24 (Laurent and Massart (2000)) *Fix $\beta > 0$, and let $X_1, \dots, X_m \sim \mathcal{N}(0, \sigma^2)$ be independent. Then*

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} X_i^2 - \sigma^2\right| > 4\sigma^2\left(\sqrt{\frac{\log(1/\beta)}{m}} + \frac{2\log(1/\beta)}{m}\right)\right] \le \beta$$

C.2. Deterministic Regularity Conditions

We will rely on certain regularity properties of i.i.d. samples from a Gaussian. These are standard concentration inequalities, and a reference for these facts is Section 4 of Diakonikolas et al. (2016).

Fact 25 *Let $X_1, \dots, X_n \sim \mathcal{N}(0, \Sigma)$ i.i.d. for $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$. Let $Y_i = \Sigma^{-1/2}X_i$ and let*

$$\widehat{\Sigma}_Y = \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^\top$$

Then for every $\beta > 0$, the following conditions hold except with probability $1 - O(\beta)$.

$$\forall i \in [n] \quad \|Y_i\|_2^2 \le O(d\log(n/\beta)) \tag{6}$$

$$\left(1 - O\left(\sqrt{\frac{d+\log(1/\beta)}{n}}\right)\right)\cdot\mathbb{I} \preceq \widehat{\Sigma}_Y \preceq \left(1 + O\left(\sqrt{\frac{d+\log(1/\beta)}{n}}\right)\right)\cdot\mathbb{I} \tag{7}$$

$$\left\|\mathbb{I} - \widehat{\Sigma}_Y\right\|_F \le O\left(\sqrt{\frac{d^2+\log(1/\beta)}{n}}\right) \tag{8}$$

We now note some simple consequences of these conditions. These inequalities follow from simple linear algebra and we omit their proof for conciseness.

Lemma 26 *Let $Y_1, \dots, Y_n$ satisfy (6)–(8). Fix $M \succ 0$, and for all $i = 1, \dots, n$, let $Z_i = M^{1/2}Y_i$, and let $\widehat{\Sigma}_Z = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^\top$. Let κ' be the top eigenvalue of M. Then*

$$\forall i \in [n] \quad \|Z_i\|_2^2 \le O\big(\kappa' d\log(n/\beta)\big)$$

$$\left(1 - O\left(\sqrt{\frac{d+\log(1/\beta)}{n}}\right)\right)\cdot M \preceq \widehat{\Sigma}_Z \preceq \left(1 + O\left(\sqrt{\frac{d+\log(1/\beta)}{n}}\right)\right)\cdot M$$

$$\left\|M - \widehat{\Sigma}_Z\right\|_M \le O\left(\sqrt{\frac{d^2+\log(1/\beta)}{n}}\right)$$

C.3. Analysis of Algorithm 1

Lemma 27 (Analysis of NAIVEPCE) *For every ρ, β, κ, n, $\text{NAIVEPCE}_{\rho,\beta,\kappa}(X)$ satisfies ρ-zCDP, and if $X_1, \dots, X_n$ are sampled i.i.d. from $\mathcal{N}(0, \Sigma)$ for $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$ and satisfy (6)–(8), then with probability at least $1 - O(\beta)$, it outputs M so that $M = \Sigma^{1/2}(I + N')\Sigma^{1/2} + N$ where*

$\Sigma^{1/2}(I+N')\Sigma^{1/2} \succeq 0$, *and*

$$\|N'\|_F \le O\left(\sqrt{\frac{d^2+\log(1/\beta)}{n}}\right) \text{ and } \|N\|_\Sigma \le O\left(\frac{d^2\kappa\log(n/\beta)\log^{1/2}(1/\beta)}{n\rho^{1/2}}\right), \text{ and} \tag{9}$$

$$\|N'\|_2 \le O\left(\sqrt{\frac{d+\log(1/\beta)}{n}}\right) \text{ and } \|N\|_2 \le O\left(\frac{d^{3/2}\kappa\log(n/\beta)\log(1/\beta)}{n\rho^{1/2}}\right). \tag{10}$$

Proof We first prove the privacy guarantee. Given two neighboring data sets X, X' of size n which differ in that one contains X_i and the other contains X_i', the truncated empirical covariance of these two data sets can change in Frobenius norm by at most

$$\left\|\frac{1}{n}\left(X_iX_i^\top - X_i'(X_i')^\top\right)\right\|_F \le \frac{1}{n}\|X_i\|_2^2 + \frac{1}{n}\|X_i'\|_2^2 \le O\left(\frac{d\kappa\log(n/\beta)}{n}\right).$$

Thus the privacy guarantee follows immediately from Lemma 19.

We now prove correctness. Recall M' is the original noised covariance before projection back to the SDP cone. We first prove that M' satisfies this form, with $\Sigma^{1/2}(I+N')\Sigma^{1/2} = \frac{1}{n}\sum_{i=1}^n X_iX_i^\top$. Clearly this implies that $\Sigma^{1/2}(I+N')\Sigma^{1/2} \succeq 0$. Since (6) holds, we have $S = [n]$. The first inequality in (9) now follows from Lemma 26, and the second follows from Fact 24 and since $\|N\|_\Sigma \le \|N\|_F$, as $\Sigma \succeq I$. By a similar logic, (10) holds since we can apply Lemma 26 and Theorem 22. Finally, to argue about M, simply observe that $\Sigma^{1/2}(I+N')\Sigma^{1/2}$ is PSD, and projection onto the PSD cone can only decrease distance (in either spectral norm or Frobenius norm) to any element in the PSD cone. ■

C.4. Analysis of Algorithm 2

Theorem 28 *For every $\rho, \beta, \kappa, K > 0$, $\textsc{WeakPPC}_{\rho,\beta,\kappa,K}(X)$ satisfies ρ-zCDP and, if $X_1, \dots, X_n$ are sampled i.i.d. from $\mathcal{N}(0,\Sigma)$ for $\mathbb{I} \preceq \Sigma \preceq \kappa\mathbb{I}$ and satisfy (6)–(8), then with probability at least $1 - O(\beta)$ it outputs (V, A) such that*

$$(1-\psi)^2(1-\Gamma)\mathbb{I} \preceq A\Sigma A \preceq (1+\psi)\cdot\kappa\left(\max\left(\frac{1}{K}, \frac{1}{2}\right) + \varphi\right)\mathbb{I} \tag{11}$$

where

$$\varphi = O\left(\frac{d^{3/2}\log(n/\beta)\log(1/\beta)}{n\rho^{1/2}}\right),$$

$$\psi = O\left(\sqrt{\frac{d+\log(1/\beta)}{n}}\right), \text{ and}$$

$$\Gamma = \max\left\{\frac{2K}{(1/2-\varphi)\kappa}, \frac{16K\varphi^2}{(1/2-\varphi)^2}\right\}$$

In particular, if $\kappa > 1000$, and K is an appropriate constant, and

$$n \ge O\left(\frac{d^{3/2}\text{polylog}(\frac{d}{\rho\beta})}{\rho^{1/2}}\right)$$

then $1.1A$ is such that $\mathbb{I} \preceq (1.1A)\Sigma(1.1A) \preceq \frac{7}{10}\kappa\mathbb{I}$.

Proof Privacy follows since we are simply post-processing the output of Algorithm 1 (Lemma 15). Thus it suffices to prove correctness. We assume that (6)–(8) hold simultaneously. By Lemma 27, (9)–(10) hold simultaneously for the matrix Z except with probability $O(\beta)$. We will condition on these events throughout the remainder of the proof. Observe that (7) implies that $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top$ is non-singular.

We will prove the upper bound and lower bound in (11) in two separate lemmata.

Lemma 29 *Let V, A be as in Algorithm 2. Then, conditioned on (6)–(10), with probability $1 - O(\beta)$, we have*

$$\|A\Sigma A\|_2 \leq (1+\psi)\cdot\kappa\left(\max\left(\frac{1}{K},\frac{1}{2}\right)+\varphi\right) .$$

Lemma 30 *Let V, A be as in Algorithm 2. Then, conditioned on (6)–(10), with probability $1 - O(\beta)$, we have*

$$A\Sigma A \succeq (1-\psi)^2(1-\Gamma)\mathbb{I} . \tag{12}$$

These two lemmata therefore together imply Theorem 28. We now turn our attention to the proofs of these lemmata. Let N be the Gaussian noise added to the empirical covariance in NAIVEPCE, so that $Z = \widehat{\Sigma} + N$.

Proof [Proof of Lemma 29] By Lemma 26 (with $M = \Sigma$), it suffices to show that

$$\|A\widehat{\Sigma}A\|_2 \leq \kappa\left(\max\left(\frac{1}{K},\frac{1}{2}\right)+\varphi\right) .$$

But with probability $1-\beta$, we have

$$\begin{aligned}\left\|A\widehat{\Sigma}A\right\|_2 &\leq \|AZA\|_2 + \|ANA\|_2 \\ &\overset{(a)}{\leq} \|AZA\|_2 + \kappa\varphi ,\end{aligned}$$

where (a) follows since $\|A\|_2 \leq 1$ and Theorem 22. We now observe that since V is a span of eigenvectors of Z, we have

$$AZA = \frac{1}{K}\Pi_V Z \Pi_V + \Pi_{V^\perp} Z \Pi_{V^\perp} ,$$

and so by our choice of V, we have $\|AZA\|_2 \leq \kappa\cdot\max(1/K, 1/2)$. This completes our proof. ■

We now prove the lower bound in Theorem 28:

Proof [Proof of Lemma 30] As before, by Lemma 26, it suffices to prove that

$$A\widehat{\Sigma}A \succeq (1-\psi)(1-\Gamma)\mathbb{I} .$$

This is equivalent to showing that for all unit vectors u, we have

$$u^T A\widehat{\Sigma}Au \geq (1-\psi)(1-\Gamma) .$$

Fix any such u. Expanding, we have

$$u^T A\widehat{\Sigma}Au = \frac{1}{K}u^T\Pi_V\widehat{\Sigma}\Pi_V u + \frac{1}{K^{1/2}}u^T\Pi_V\widehat{\Sigma}\Pi_{V^\perp}u + \frac{1}{K^{1/2}}u^T\Pi_{V^\perp}\widehat{\Sigma}\Pi_V u + u^T\Pi_{V^\perp}\widehat{\Sigma}\Pi_{V^\perp}u . \tag{13}$$

The first and last terms are non-negative since $\widehat{\Sigma}$ is PSD, but the other terms may be negative, so we need to control their magnitude. Note that

$$\begin{aligned} u^T \Pi_V \widehat{\Sigma} \Pi_V u &= u^T \Pi_V Z \Pi_V u - u^T \Pi_V N \Pi_V u \\ &\geq \frac{\kappa}{2} \|\Pi_V u\|_2^2 - \kappa\varphi \|\Pi_V u\|_2^2 = \kappa\left(\frac{1}{2} - \varphi\right)\|\Pi_V u\|_2^2 \,. \end{aligned}$$

where the inequality follows from our choice of V (the "large" directions of Z), and Theorem 22 (bounding the spectral norm of N). On the other hand, we have

$$\begin{aligned} \left|\frac{1}{K^{1/2}} u^T \Pi_V \widehat{\Sigma} \Pi_{V^\perp} u\right| &= \left|\frac{1}{K^{1/2}} u^T \Pi_V (Z - N) \Pi_{V^\perp} u\right| \\ &\overset{(a)}{=} \left|\frac{1}{K^{1/2}} u^T \Pi_V N \Pi_{V^\perp} u\right| \\ &\overset{(b)}{\leq} \frac{\kappa}{K^{1/2}} \varphi \|\Pi_V u\|_2 \|\Pi_{V^\perp} u\|_2 \\ &\leq \frac{\kappa}{K^{1/2}} \varphi \|\Pi_V u\|_2 \,, \end{aligned}$$

where (a) follows since $\Pi_V Z \Pi_{V^\perp} = 0$, and (b) follows from Theorem 22. Similarly, we have

$$\left|\frac{1}{K^{1/2}} u^T \Pi_{V^\perp} \widehat{\Sigma} \Pi_V u\right| \leq \frac{\kappa}{K^{1/2}} \varphi \|\Pi_V u\|_2 \,.$$

Thus, if we have $\|\Pi_V u\|_2^2 \geq \Gamma$, by our choice of Γ, we have

$$\begin{aligned} &\frac{1}{K} u^T \Pi_V \widehat{\Sigma} \Pi_V u + \frac{1}{K^{1/2}} u^T \Pi_V \widehat{\Sigma} \Pi_{V^\perp} u + \frac{1}{K^{1/2}} u^T \Pi_{V^\perp} \widehat{\Sigma} \Pi_V u \\ \geq\ & \frac{\kappa}{K}\left(\frac{1}{2} - \varphi\right)\|\Pi_V u\|_2^2 - 2\frac{\kappa}{K^{1/2}} \varphi \|\Pi_V u\|_2 \\ \geq\ & \frac{\kappa}{2K}\left(\frac{1}{2} - \varphi\right)\|\Pi_V u\|_2^2 \\ \geq\ & 1 \,. \end{aligned}$$

Thus in this case the claim follows since the final term in (13) is nonnegative since $\widehat{\Sigma}$ is PSD.

Now consider the case where

$$\|\Pi_V u\|_2 < \Gamma \,,$$

or equivalently, since by the Pythagorean theorem we have $\|\Pi_V u\|_2^2 + \|\Pi_{V^\perp} u\|_2^2 = 1$,

$$\|\Pi_{V^\perp} u\|_2^2 > 1 - \Gamma \,.$$

Then, since we have $\widehat{\Sigma} \succeq (1 - \psi)\mathbb{I}$ (Fact 25), we have

$$\begin{aligned} u^T A \widehat{\Sigma} A u &\geq (1 - \psi) u^T \left(\frac{1}{K} \Pi_V \Pi_V + \Pi_{V^\perp} \Pi_{V^\perp}\right) u \\ &\geq (1 - \psi)\|\Pi_{V^\perp} u\|_2^2 \\ &\geq (1 - \psi)(1 - \Gamma) \,, \end{aligned}$$

as claimed. ∎

Combining Lemma 29 and Lemma 30 yield the desired conclusion. ∎

C.5. Analysis of Algorithm 3

We present the proof of Theorem 8.

Privacy is immediate from Theorem 28 and composition of ρ-zCDP (Lemma 17).

By Fact 25, (6)–(8) hold for the sample $X_1, \ldots, X_n$ except with probability $O(\beta)$. Define $\Sigma^{(1)} = \Sigma$ and recursively define $\Sigma^{(t)} = A^{(t-1)}\Sigma^{(t-1)}A^{(t-1)}$ to be the covariance after the t-th round of preconditioning. By the guarantee of WEAKPPC (Theorem 28), a union bound, and our choice of n, we have that for every t, we obtain a matrix $A^{(t)}$ such that

$$\mathbb{I} \preceq A^{(t)}\Sigma^{(t)}A^{(t)} \preceq 0.7\kappa^{(t)}\mathbb{I}.$$

The theorem now follows by induction on t.

C.6. Analysis of Algorithm 4

We present the proof of Theorem 9.

Privacy follows from Theorem 28 and composition of ρ-zCDP (Lemma 17).

By construction, the samples $Y_1, \ldots, Y_n$ are i.i.d. from $\mathcal{N}(0, A\Sigma A)$. By Theorem 8, and our choice of n, we have that except with probability $O(\beta)$, A is such that $\mathbb{I} \preceq A\Sigma A \preceq 1000\mathbb{I}$. Therefore, combining the guarantees of NAIVEPCE with our choice of n we obtain that, except with probability $O(\beta)$, $\widetilde{\Sigma}$ satisfies $\|A\Sigma A - \widetilde{\Sigma}\|_{A\Sigma A} \leq O(\alpha)$. The theorem now follows because $\|\Sigma - \widehat{\Sigma}\|_\Sigma = \|\Sigma - A^{-1}\widetilde{\Sigma}A^{-1}\|_\Sigma = \|A\Sigma A - \widetilde{\Sigma}\|_{A\Sigma A} \leq O(\alpha)$.

C.7. Analysis of Algorithm 5

We present the proof of Theorem 11.

The fact that the algorithm satisfies ρ-zCDP follows immediately from the assumed privacy of KVMEAN and the composition property for zCDP (Lemma 17).

Next we argue that with probability at least $1 - \beta$, for every coordinate $j = 1, \ldots, d$, we have $|\mu_j - \widehat{\mu}_j| \leq \alpha/\sqrt{d}$. Observe that, since $X_1, \ldots, X_n$ are distributed as $\mathcal{N}(\mu, \Sigma)$, the j-th coordinates $X_{1,j}, \ldots, X_{n,j}$ are distributed as $\mathcal{N}(\mu_j, \Sigma_{jj})$ and, by assumption $|\mu_j| \leq R$ and $1 \leq \Sigma_{jj} \leq \kappa$. Thus, by Theorem 10, we have $|\mu_j - \widehat{\mu}_j| \leq \alpha'\kappa = \alpha/\sqrt{d}$ except with probability at most $\beta' = \beta/d$. The statement now follows by a union bound.

Assuming that every coordinate-wise estimate is correct up to $\alpha/\sqrt{d}$, we have

$$\|\mu - \widehat{\mu}\|_\Sigma = \|\Sigma^{-1/2}(\mu - \widehat{\mu})\|_2 \leq \|\Sigma^{-1/2}\|_2 \cdot \|\mu - \widehat{\mu}\|_2 \leq \alpha$$

where the final equality uses the coordinate-wise bound on $\mu - \widehat{\mu}$ and the fact that $I \preceq \Sigma$. To complete the proof, we can plug our choices of ρ', α', β' into the sample complexity bound for KVMEAN from Theorem 10.

The proof and algorithm for the final statement of the theorem regarding (ε, δ)-DP are completely analogous. This completes the proof of the theorem.

C.8. Analysis of Algorithm 6

We present the proof of Theorem 12.

Privacy will follow immediately from the composition property of 2ρ-zCDP and the assumed privacy of PPC and NAIVEPME. The sample complexity bound will also follow immediately

from the sample complexity bounds for PCE and NAIVEPME. Thus, we focus on proving that $\|\mu - \widehat{\mu}\|_\Sigma \leq \alpha$.

Since $X_1, \ldots, X_{2n}$ are i.i.d. from $\mathcal{N}(\mu, \Sigma)$, the values $Z_1, \ldots, Z_n$ are i.i.d. from $\mathcal{N}(0, \Sigma)$. Therefore, with probability at least $1 - \beta$, $\text{PPC}(Z_1, \ldots, Z_n)$ returns a matrix A such that $I \preceq A\Sigma A \preceq 1000I$. Note that since $I \preceq \Sigma$ and $A\Sigma A \preceq 1000$ we have $\|A\|_2 \leq 1000$.

Now, since the samples $X_{2n+1}, \ldots, X_{3n}$ are i.i.d. from $\mathcal{N}(\mu, \Sigma)$, the values $Y_1, \ldots, Y_n$ are i.i.d. from $\mathcal{N}(A\mu, A\Sigma A)$. Note that $\|A\mu\|_2 \leq \|A\|_2\|\mu\|_2 \leq 1000R$ and, by assumption, $I \preceq A\Sigma A \preceq 1000I$. When we apply $\text{NAIVEPME}_{\rho,\alpha,\beta,1000R,1000}$ to $Y_1, \ldots, Y_n$, with probability at least $1 - \beta$ we will obtain $\widetilde{\mu}$ such that $\|A\mu - \widetilde{\mu}\|_{A\Sigma A} \leq \alpha$. Finally, we can write $\|\mu - \widehat{\mu}\|_\Sigma = \|A\mu - A\widehat{\mu}\|_{A\Sigma A} = \|A\mu - \widetilde{\mu}\|_{A\Sigma A} \leq \alpha$. The theorem now follows by a union bound over the two possible failure events.

Appendix D. Privately Learning Product Distributions

In this section we introduce and analyze our algorithm for learning a product distribution P over $\{0,1\}^d$ in total variation distance, thereby proving Theorem 3 in the introduction. The pseudocode appears in Algorithm 7. For simplicity of presentation, we assume that the product distribution has mean that is bounded coordinate-wise by $\frac{1}{2}$ (i.e. $\mathbb{E}[P] \preceq \frac{1}{2}$), although we emphasize that this assumption is essentially without loss of generality, and can easily be removed while paying only a constant factor in the sample complexity.

D.1. A Private Product-Distribution Estimator

To describe the algorithm, we need to introduce notation for the *truncated mean*. Given a dataset element (a vector) $X_i \in \{0,1\}^d$ and $B \geq 0$, we use

$$\mathsf{trunc}_B(X_i) = \begin{cases} X_i & \text{if } \|X_i\|_2 \leq B \\ \frac{B}{\|X_i\|_2} \cdot X_i & \text{if } \|X_i\|_2 > B \end{cases}$$

to denote the truncation of x to an ℓ_2-ball of radius B. Given a dataset $X = (X_1, \ldots, X_m) \in \{0,1\}^{m \times d}$ and $B > 0$, we use

$$\mathsf{tmean}_B(X) = \frac{1}{m} \sum_{i=1}^{m} \mathsf{trunc}_B(X_i)$$

to denote the mean of the truncated vectors. Observe that the ℓ_2-sensitivity of tmean_B is $\frac{B}{m}$, while the ℓ_2-sensitivity of the untruncated mean is infinite. Note that $\mathsf{tmean}_B(X) = \frac{1}{m}\sum_{i=1}^m X_i$ unless $\|X_i\|_2 > B$ for some i. If one of the inputs to tmean_B does not satisfy the norm bound then we will say, "truncation occurred," as a shorthand.

We also use the following notational conventions: Given a dataset element $X_i \in \{0,1\}^d$, we will use the array notation $X_i[j]$ to refer to its j-th coordinate, and the notation $X_i[S] = (X_i[j])_{j \in S}$ to refer to the vector X restricted to the subset of coordinates $S \subseteq [d]$. Given a dataset $X = (X_1, \ldots, X_m)$, we use the notation $X[S] = (X_1[S], \ldots, X_m[S])$ to refer to the dataset consisting of each X_i restricted to the subset of coordinates $S \subseteq [d]$.

The privacy analysis of Algorithm 7 is straightforward, based on privacy of the Gaussian mechanism and bounded sensitivity of the truncated mean.

Algorithm 7: Private Product-Distribution Estimator $\mathrm{PPDE}_{\rho,\alpha,\beta}(X)$

Input: Samples $X_1,\ldots,X_n \in \{0,1\}^d$ from an unknown product distribution P satisfying $\mathbb{E}[P] \preceq \frac{1}{2}$. Parameters $\rho,\alpha,\beta > 0$.

Output: A product distribution Q over $\{0,1\}^d$ such that $\mathrm{d_{TV}}(P,Q) \leq \alpha$.

Set parameters: $R \leftarrow \log_2(d/2)$ $c \leftarrow 128\log^{5/4}(d/\alpha\beta(2\rho)^{1/2})$ $c' \leftarrow 128\log^3(dR/\beta)$ $m \leftarrow \frac{c'd}{\alpha^2} + \frac{cd}{\alpha(2\rho)^{1/2}}$

Split X into $R+1$ blocks of m samples each, denoted $X^r = (X_1^r,\ldots,X_m^r)$

(Halt and output $\perp$ if n is too small.)

Let $q[j] \leftarrow 0$ for every $j \in [d]$, and let $S_1 = [d]$, $u_1 \leftarrow \frac{1}{2}$, $\tau_1 \leftarrow \frac{3}{16}$, and $r \leftarrow 1$

`// Partitioning Rounds`

While $u_r|S_r| \geq 1$

- Let $S_{r+1} \leftarrow \emptyset$
- Let $B_r \leftarrow \sqrt{6u_r|S_r|\log(mR/\beta)}$
- Let $q_r[S_r] \leftarrow_{\mathrm{R}} \mathsf{tmean}_{B_r}(X^r[S_r]) + \mathcal{N}\left(0, \frac{B_r^2}{2\rho m^2}\cdot\mathbb{I}\right)$ **For** $j \in S_r$
  - **If** $q_r[j] < \tau_r$
    - Add j to S_{r+1}
  - **Else**
    - Set $q[j] \leftarrow q_r[j]$
- Let $u_{r+1} \leftarrow \frac{1}{2}u_r$, $\tau_{r+1} \leftarrow \frac{1}{2}\tau_r$, and $r \leftarrow r+1$

`// Final Round`

If $|S_r| \geq 1$

- Let $B_r \leftarrow \sqrt{6\log(m/\beta)}$
- Let $q[S_r] \leftarrow_{\mathrm{R}} \mathsf{tmean}_{B_r}(X^r[S_r]) + \mathcal{N}\left(0, \frac{B_r^2}{2\rho m^2}\cdot\mathbb{I}\right)$

Return $Q = \mathrm{Ber}(q[1]) \otimes \cdots \otimes \mathrm{Ber}(q[d])$

Theorem 31 *For every $\rho,\alpha,\beta > 0$, $\mathrm{PPDE}_{\rho,\alpha,\beta}(X)$ satisfies ρ-zCDP.*

Proof Since each individual's data is used only to compute $\mathsf{tmean}_{B_r}(X^r)$ for a single round r, privacy follows immediately from Lemma 19 and from observing that the ℓ_2-sensitivity of tmean_B is $\frac{B}{n}$. Note that, since a disjoint set of samples X^r is used for each round r, each sample only affects a single one of the rounds, so we do not need to apply composition. ■

D.2. Accuracy Analysis for PPDE

In this section we prove the following theorem bounding the sample complexity required by PPDE to learn a product distribution up to α in total variation distance.

Theorem 32 *For every $d \in \mathbb{N}$, every product distribution P over $\{0,1\}^d$, and every $\rho,\alpha,\beta > 0$, if $X = (X_1,\ldots,X_n)$ are independent samples from P for*

$$n = \tilde{O}\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\sqrt{\rho}}\right),$$

then with probability at least $1 - O(\beta)$, $\mathrm{PPDE}_{\rho,\alpha,\beta}(X)$ outputs Q, such that $\mathrm{d_{TV}}(P,Q) \leq O(\alpha)$. The notation $\tilde{O}(\cdot)$ hides polylogarithmic factors in d, $\frac{1}{\alpha}$, $\frac{1}{\beta}$, and $\frac{1}{\rho}$.

Before proving the theorem, we will introduce or recall a few useful tools and inequalities.

Distances Between Distributions. We use several notions of distance between distributions.

Definition 33 *If P, Q are distributions, then*

- *the* statistical distance *is* $\mathrm{d_{TV}}(P,Q) = \frac{1}{2}\sum_x |P(x) - Q(x)|$,
- *the* χ^2-divergence *is* $\mathrm{d}_{\chi^2}(P\|Q) = \sum_x \frac{(P(x)-Q(x))^2}{Q(x)}$, *and*
- *the* KL-divergence *is* $\mathrm{d_{KL}}(P\|Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)}$.

For product distributions $P = P_1 \otimes \cdots \otimes P_k$ and $Q = Q_1 \otimes \cdots \otimes Q_k$, the χ^2 and $\mathrm{d_{KL}}$ divergences are additive, and the statistical distance is subadditive. Specifically,

Lemma 34 *Let $P = P_1 \otimes \cdots \otimes P_k$ and $Q = Q_1 \otimes \cdots \otimes Q_k$ be two product distributions. Then*

- $\mathrm{d_{TV}}(P,Q) \le \sum_{j=1}^d \mathrm{d_{TV}}(P_j, Q_j)$,
- $\mathrm{d}_{\chi^2}(P\|Q) \le \sum_{j=1}^d \mathrm{d}_{\chi^2}(P_j\|Q_j)$, *and*
- $\mathrm{d_{KL}}(P\|Q) = \sum_{j=1}^d \mathrm{d_{KL}}(P_j\|Q_j)$.

The three definitions also satisfy some useful relationships.

Lemma 35 *For any two distributions P, Q we have $2\cdot\mathrm{d_{TV}}(P,Q)^2 \le \mathrm{d_{KL}}(P\|Q) \le \mathrm{d}_{\chi^2}(P\|Q)$.*

Tail Bounds. We need a couple of useful tail bounds for sums of independent Bernoulli random variables. The first lemma is a useful form of the Chernoff bound.

Lemma 36 *For every $p > 0$, if $X_1, \ldots, X_m$ are i.i.d. samples from $\mathrm{Ber}(p)$ then for every $\varepsilon > 0$*

$$\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^m X_i \ge p+\varepsilon\right] \le e^{-\mathrm{d_{KL}}(p+\varepsilon\|p)\cdot m} \quad \textit{and} \quad \mathbb{P}\left[\frac{1}{m}\sum_{i=1}^m X_i \le p-\varepsilon\right] \le e^{-\mathrm{d_{KL}}(p-\varepsilon\|p)\cdot m}$$

The next lemma follows easily from a Chernoff bound.

Lemma 37 *Suppose $X_1, \ldots, X_k$ are sampled i.i.d. from a product distribution P over $\{0,1\}^t$, where the mean of each coordinate is upper bounded by p (i.e. $\mathbb{E}[P] \preceq p$). Then*

1. *if $pt > 1$, then for each i, $\mathbb{P}\Big[\|X_i\|_2^2 \ge pt\Big(1 + 3\log(\frac{k}{\beta})\Big)\Big] \le \frac{\beta}{k}$, and*
2. *if $pt \le 1$, then for each i, $\mathbb{P}\Big[\|X_i\|_2^2 \ge 6\log(\frac{k}{\beta})\Big] \le \frac{\beta}{k}$.*

D.2.1. ANALYSIS OF THE PARTITIONING ROUNDS

In this section we analyze the progress made during the partitioning rounds. We show two properties: (1) any coordinate j such that $q[j]$ was set during the partitioning rounds has small error and (2) any coordinate j such that $q[j]$ was not set in the partitioning rounds has a small mean. We capture the properties of the partitioning rounds that will be necessary for the proof of Theorem 32 in the following lemma.

Lemma 38 *If $X^1, \ldots, X^R$ each contain at least*

$$m = \frac{128d\log^3(dR/\beta)}{\alpha^2} + \frac{128d\log^{5/4}(d/\alpha\beta(2\rho)^{1/2})}{\alpha\rho^{1/2}}$$

i.i.d. samples from P, then, with probability at least $1 - O(\beta)$, in every partitioning round $r = 1, \ldots, R$:

1. *If a coordinate j does not go to the next round (i.e. $j \in S_r$ but $j \notin S_{r+1}$) then $q[j]$ has small χ^2-divergence with $p[j]$,*
$$\frac{4(p[j]-q[j])^2}{q[j]} \leq \frac{\alpha^2}{d}.$$
Thus, if S_A consists of all coordinates such that $q[j]$ is set in one of the partitioning rounds, $\mathrm{d_{TV}}(P[S_A], Q[S_A]) \leq \alpha$.

2. *If a coordinate j does go to the next round (i.e. $j \in S_r, S_{r+1}$), then $p[j]$ is small,*
$$p[j] \leq u_{r+1} = \frac{u_r}{2}.$$

Proof We will prove the lemma by induction on r (taking a union bound over the events that one of the two conditions fails in a given round r). Therefore, we will assume that in every round r, $p[j] \leq u_r$ for every $j \in S_r$ and prove that if this bound holds then the two conditions in the lemma hold. For the base of the induction, observe that, by assumption, $p[j] \leq u_1 = \frac{1}{2}$ for every $j \in S_1 = [d]$. In what follows we fix an arbitrary round $r \in [R]$. Throughout the proof, we will use the notation $\tilde{p}_r = \frac{1}{m}\sum_{i=1}^m X_i^r$ to denote the empirical mean of the r-th block of samples.

Claim 39 *If $\tilde{p}_r[j] = \frac{1}{m}\sum_{i=1}^m X_i^r[j]$ and $p_j > \frac{1}{d}$, then with probability at least $1 - \frac{2\beta}{R}$,*

$$\forall j \in S_r \;\; |p[j] - \tilde{p}_r[j]| \leq \sqrt{\frac{4p[j]\log\left(\frac{dR}{\beta}\right)}{m}}$$

Proof [Proof of Claim 39] We use a Chernoff Bound (Theorem 36), and facts that

$$\forall \gamma > 0 \quad \mathrm{d_{KL}}(p+\gamma||p) \geq \frac{\gamma^2}{2(p+\gamma)} \quad \text{and} \quad \mathrm{d_{KL}}(p-\gamma||p) \geq \frac{\gamma^2}{2p},$$

and set

$$\gamma = \sqrt{\frac{4p[j]\log\left(\frac{dR}{\beta}\right)}{m}}.$$

Note that when $p[j] > \frac{1}{d}$, due to our choice of parameters, $\gamma \leq p[j]$. Therefore, $2(p[j]+\gamma) \leq 4p[j]$. Finally, taking a union bound over the cases when $\tilde{p}_r[j] \leq p[j]-\gamma$ and when $\tilde{p}_r[j] \geq p[j]+\gamma$, we prove the claim. ■

Claim 40 *With probability at least* $1-\frac{\beta}{R}$*, for every* $X_i^r \in X^r$*,* $\|X_i^r\|_2 \leq B_r$*, so no rows of* X^r *are truncated in the computation of* $\mathsf{tmean}_{B_r}(X^r)$*.*

Proof [Proof of Claim 40] By assumption, all marginals specified by S_r are upper bounded by u_r. Now, the expected value of $\|X_i^r\|_2^2$ is at most $u_r|S_r|$. Since $B_r = \sqrt{u_r|S_r|6\log(mR/\beta)}$, we know that $B_r \geq \sqrt{u_r|S_r|(1+3\log(mR/\beta))}$. The claim now follows from a Chernoff Bound (Lemma 37) and a union bound over the entries of X^r. ■

Claim 41 *With probability at least* $1-\frac{2\beta}{R}$*,*

$$\forall j \in S_r \;\; |\tilde{p}_r[j]-q_r[j]| \leq \sqrt{\frac{6u_r|S_r|\log\left(\frac{mR}{\beta}\right)\log\left(\frac{2dR}{\beta}\right)}{\rho m^2}}$$

Proof [Proof of Claim 41] We assume that all marginals specified by S_r are upper bounded by u_r. From Claim 40, we know that, with probability at least $1-\beta/R$, there is no truncation, so $\mathsf{tmean}_{B_r}(X^r[S_r]) = \frac{1}{m}\sum_{X_i^r \in X^r} X^r[S_r] = \tilde{p}_r[S_r]$. So, the Gaussian noise is added to $\tilde{p}_r[j]$ for each $j \in S_r$. Therefore, the only source of error here is the Gaussian noise. Using the standard tail bound for zero-mean Gaussians (Lemma 20), with the following parameters,

$$\sigma = \sqrt{\frac{3u_r|S_r|\log\left(\frac{mR}{\beta}\right)}{\rho m^2}} \quad \text{and} \quad t = \sqrt{2\log\left(\frac{2dR}{\beta}\right)},$$

and taking a union bound over all coordinates in S_r, and the event of truncation, we obtain the claim. ■

Plugging our choice of m into Claims 39 and 41, applying the triangle inequality, and simplifying, we get that (with high probability),

$$|p[j]-q_r[j]| \leq \frac{\alpha}{\log^{1/4}\left(\frac{d}{\beta}\right)}\sqrt{\frac{u_r}{d}}.$$

To simplify our calculations, we will define

$$e_r = \frac{\alpha}{\log^{1/4}\left(\frac{d}{\beta}\right)}\sqrt{\frac{u_r}{d}}$$

to denote the above bound on $|p[j]-q[j]|$ in round r.

Claim 42 *For all $j \in S_r$, with probability, at least, $1 - \frac{4\beta}{R}$,*

$$\chi^2(p[j], q_r[j]) \leq \frac{4(p[j] - q_r[j])^2}{q_r[j]}.$$

Proof For every r, and d more than some absolute constant, $|e_r| \leq \frac{1}{4},$. Also, by assumption, $p[j] \leq \frac{1}{2}$, for all $j \in [d]$. Therefore, for every r, and every $j \in S_r$,

$$\begin{aligned}\chi^2(p[j], q_r[j]) &= \frac{(p[j] - q_r[j])^2}{q_r[j]} + \frac{(p[j] - q_r[j])^2}{1 - q_r[j]} \\ &= \frac{(p[j] - q_r[j])^2}{q_r[j](1 - q_r[j])} \\ &\leq \frac{4(p[j] - q_r[j])^2}{q_r[j]}\end{aligned}$$

■

Claim 43 *With probability at least $1 - \frac{4\beta}{R}$, for every $j \in S_r$,*

$$q_r[j] \geq \tau_r \implies \frac{4(p[j] - q[j])^2}{q[j]} \leq \frac{\alpha^2}{d}.$$

Proof [Proof of Claim 43] We want to show the following inequality.

$$\frac{4(p_j - q_r[j])^2}{q_r[j]} \leq \frac{\alpha^2}{d}$$

We know that $|p_j - q_r[j]| \leq e_r$ with high probability. Thus, we need to show that if $q_r[j] \geq \tau_r$, the following inequality holds:

$$\frac{4e_r^2}{q_r[j]} \leq \frac{\alpha^2}{d} \iff \frac{4de_r^2}{\alpha^2} \leq q_r[j].$$

We now show that the left-hand side is at most τ_r, which completes the proof. In the algorithm, we have $\tau_r = \frac{3}{4}u_{r+1}$:

$$\frac{4de_r^2}{\alpha^2} \leq \frac{3}{4}u_{r+1} \iff \frac{4u_r}{\log^{1/2}\left(\frac{d}{\beta}\right)} \leq \frac{3}{4}u_{r+1} \iff \frac{16}{3\log^{1/2}\left(\frac{d}{\beta}\right)} \leq \frac{1}{2}.$$

Note that the final inequality is satisfied as long as d is larger than some absolute constant. ■

Claim 44 *With probability at least $1 - \frac{4\beta}{R}$, for every $j \in S_r$,*

$$q_r[j] < \tau_r \implies p[j] \leq u_{r+1} = \frac{u_r}{2}.$$

Proof [Proof of Claim 44] We know that with high probability, $p_j \le q_r[j] + e_r$. But since $q_r[j] < \tau_i$, we know that $p_j < \tau_r + e_r$. Also, $\tau_r = \frac{3}{4}u_{r+1}$. Then it is sufficient to show the following.

$$\begin{aligned} & e_r \le \frac{u_{r+1}}{4} \\ \iff & \frac{\alpha}{\log^{1/4}\left(\frac{d}{\beta}\right)}\sqrt{\frac{u_r}{d}} \le \frac{u_{r+1}}{4} \\ \iff & \frac{16\alpha^2}{du_1\log^{1/2}\left(\frac{d}{\beta}\right)} \le \left(\frac{1}{2}\right)^{r+1} \\ \iff & \left(\frac{16\alpha^2}{du_1\log^{1/2}\left(\frac{d}{\beta}\right)}\right)^{\frac{1}{r+1}} \le \frac{1}{2}. \end{aligned} \tag{14}$$

Now we have

$$\begin{aligned} \left(\frac{16\alpha^2}{du_1\log^{1/2}\left(\frac{d}{\beta}\right)}\right)^{\frac{1}{r+1}} &= \left(\frac{1}{d}\right)^{\frac{1}{r+1}} \cdot \left(\frac{16\alpha^2}{u_1\log^{1/2}\left(\frac{d}{\beta}\right)}\right)^{\frac{1}{r+1}} \\ &\le \frac{1}{2}\cdot\left(\frac{16\alpha^2}{u_1\log^{1/2}\left(\frac{d}{\beta}\right)}\right)^{\frac{1}{r+1}} && (r \le R = \log_2(d/2)) \\ &\le \frac{1}{2} \end{aligned}$$

where the last inequality holds for d larger than some absolute constant. Therefore, (14) is satisfied for all $1 \le r \le R$. ■

Claim 44 completes the inductive step of the proof. It establishes that at the beginning of round $r+1$, $p_j < u_{r+1}$ for all $j \in S_{r+1}$. Now we can take a union bound over all the failure events in each round and over each of the R rounds so that the conclusions of the Lemma hold with probability $1 - O(\beta)$. This completes the proof of Lemma 38. ■

D.2.2. ANALYSIS OF THE FINAL ROUND

In this section we show that the error of the coordinates j such that $q[j]$ was set in the final round r is small.

Lemma 45 *Let r be the final round for which $u_r|S_r| \le 1$. If $p[j] \le u_r$ for every $j \in S_r$, and X^r contains at least*

$$m = \frac{128d\log^3(dR/\beta)}{\alpha^2} + \frac{128d\log^{5/4}(d/\alpha\beta(2\rho)^{1/2})}{\alpha\rho^{1/2}}$$

i.i.d. samples from P, then with probability at least $1 - O(\beta)$, then $\mathrm{d_{TV}}(P[S_r], Q[S_r]) \le O(\alpha)$

Proof Again, we use the notation, $\tilde{p}_r = \frac{1}{m}\sum_{i=1}^m X_i^r$, for the rest of this proof. First we have, two claims that bound the difference between $p[j]$ and $\tilde{p}[j]$.

Claim 46 *For each $j \in S_r$, such that $p_j > \frac{1}{d}$, with probability at least $1 - 2\beta/d$, we have,*

$$\forall j \in S_r \;\; |p[j] - \tilde{p}_r[j]| \leq \sqrt{\frac{4p[j]\log\left(\frac{d}{\beta}\right)}{m}}$$

Proof [Proof of Claim 46] The proof is identical to that of Claim 39. ■

Claim 47 *For each $j \in S_r$, such that $p_j \leq \frac{1}{d}$, with probability at least $1 - 4\beta/d$, we have,*

$$\forall j \in S_r \;\; |p[j] - \tilde{p}_r[j]| \leq \frac{\alpha}{4d\log\left(\frac{d}{\beta}\right)}$$

Proof [Proof of Claim 47] We use a Chernoff Bound (Theorem 36), and facts that

$$\forall \gamma > 0 \quad \mathrm{d_{KL}}(p+\gamma||p) \geq \frac{\gamma^2}{2(p+\gamma)} \quad \text{and} \quad \mathrm{d_{KL}}(p-\gamma||p) \geq \frac{\gamma^2}{2p}.$$

There are two cases to analyze.

- $p_j > \frac{\alpha^2}{16d\ln^2\left(\frac{dR}{\beta}\right)}$: In this case, setting $\gamma = \sqrt{\frac{4p[j]\log\left(\frac{d}{\beta}\right)}{m}}$, we get $\gamma \leq p[j]$. Then we apply Theorem 36 with appropriate parameters.
- $p_j \leq \frac{\alpha^2}{16d\ln^2\left(\frac{dR}{\beta}\right)}$: In this case, setting $\gamma = \frac{4\log\left(\frac{d}{\beta}\right)}{m}$, we get $\gamma \geq p[j]$. Then we apply Theorem 36 with the corresponding parameters.

Since $p[j] \leq \frac{1}{d}$, if m satisfies the assumption of the lemma, then

$$\max\left\{\sqrt{\frac{4p[j]\log\left(\frac{d}{\beta}\right)}{m}}, \frac{4\log\left(\frac{d}{\beta}\right)}{m}\right\} \leq \frac{\alpha}{4d\log\left(\frac{d}{\beta}\right)}.$$

Therefore, with high probability, the maximum error is $\frac{\alpha}{4d\log\left(\frac{d}{\beta}\right)}$. ■

Claim 48 *With probability at least $1 - \beta$, for every $X_i^r \in X^r$, $\|X_i^r\|_2 \leq B_r$, so no rows of X^r are truncated in the computation of* $\mathsf{tmean}_{B_r}(X^r)$.

Proof [Proof of Claim 48] We assume that all marginals specified by S_r are upper bounded by u_r. Now, the expected value of $\|X_i^r\|_2^2$ is upper bounded by 1. Also, $B_r = \sqrt{6\log(m/\beta)}$. With this, we use a Chernoff Bound (Fact 37) and get the required result. ■

Claim 49 *With probability at least* $1 - 2\beta$,

$$\forall j \in S_r \;\; |\tilde{p}[j] - q_r[j]| \leq \sqrt{\frac{6\log\left(\frac{m}{\beta}\right)\log\left(\frac{2d}{\beta}\right)}{\rho m^2}}$$

Proof [Proof of Claim 49] We assume that all marginals specified by S_r are upper bounded by u_r. From Claim 48, we know that with high probability, there is no truncation, so, the Gaussian noise is added to $\tilde{p}_j$ for each $j \in S_r$. Therefore, the only source of error here is the Gaussian noise. Using the standard tail bound for zero-mean Gaussians (Lemma 20), with the following parameters,

$$\sigma = \sqrt{\frac{3\log\left(\frac{m}{\beta}\right)}{\rho m^2}} \quad \text{and} \quad t = \sqrt{2\log\left(\frac{2d}{\beta}\right)},$$

and taking the union bound over all columns of the dataset in that round and the event of truncation, we obtain the claim. ■

By the above claim, the magnitude of Gaussian noise added to each coordinate in the final round is less than,

$$\frac{\alpha}{2d\log^{1/4}\left(d/\alpha\beta\sqrt{2\rho}\right)}.$$

We partition the set S_r into $S_{r,L} = | \in \mathcal{S}_\nabla : \sqrt{}[|] \leq \frac{\infty}{\lceil}$ and $S_{r,H} = | \in \mathcal{S}_\nabla : \sqrt{}[|] > \frac{\infty}{\lceil}$.

Claim 50 $\mathrm{d_{TV}}(P[S_{r,H}], Q[S_{r,H}]) \leq \alpha$

Proof [Proof of Claim 50] For every coordinate $j \in S_{r,H}$, due Claim 46, and from the upper bound on the Gaussian noise added, we know that,

$$|p[j] - q[j]| \leq \frac{\alpha}{\log^{1/4}\left(\frac{d}{\beta}\right)}\sqrt{\frac{p[j]}{d}} = e_{r,H}.$$

Now, we know that $p[j] > \frac{1}{d}$, and $e_{r,H} \leq \frac{p[j]}{2}$, when d is greater than some absolute constant. So, we can bound the χ^2-divergence between such $p[j]$ and $q[j]$ by,

$$\begin{aligned}\frac{4(p[j]-q[j])^2}{q[j]} &\leq \frac{4e_{r,H}^2}{p[j]-e_{r,H}} \\ &\leq \frac{8\alpha^2}{d\log^{1/2}\left(\frac{d}{\beta}\right)} \\ &\leq \frac{\alpha^2}{d}.\end{aligned}$$

Thus, we have $\mathrm{d}_{\chi^2}(P[S_{r,H}] \| Q[S_{r,H}]) \leq \alpha^2$, which implies $\mathrm{d_{TV}}(P[S_{r,H}], Q[S_{r,H}]) \leq \alpha$. ■

Claim 51 $\mathrm{d_{TV}}(P[S_{r,L}], Q[S_{r,L}]) \leq \alpha$

Proof [Proof of Claim 51] By Claim 47, and from the upper bound on the Gaussian noise added, for every coordinate $j \in S_{r,L}$, we have,

$$|p[j] - q[j]| \leq 2 \cdot \max\left\{ \frac{\alpha}{4d\log\left(\frac{d}{\beta}\right)}, \frac{\alpha}{2d\log^{1/4}\left(d/\alpha\beta\sqrt{2\rho}\right)} \right\}$$
$$\leq \frac{\alpha}{d\log^{1/4}(d/\beta)}.$$

Thus, by Lemma 34, we have $\mathrm{d_{TV}}(P[S_{r,L}], Q[S_{r,L}]) \leq \alpha$. ■

Now, combining Claims 50 and 51, and applying Lemma 34 completes the proof. ■

D.2.3. PUTTING IT TOGETHER

In this section we combine Lemmas 38 and 45 to prove Theorem 32. First, by Lemma 38, with probability at least $1 - O(\beta)$, if S_A is the set of coordinates j such that $q[j]$ was set in any of the partitioning rounds, then

1. $\mathrm{d_{TV}}(P[S_A], Q[S_A]) \leq O(\alpha)$ and
2. if $j \notin S_A$ and r is the final round, then $p[j] \leq u_r$.

Next, by the second condition, we can apply Lemma 45 to obtain that if S_F consists of all coordinates set in the final round, then with probability at least $1 - O(\beta)$, $\mathrm{d_{TV}}(P[S_F], Q[S_F]) \leq O(\alpha)$. Finally, we use a union bound and Lemma 34 to conclude that, with probability at least $1 - O(\beta)$,

$$\mathrm{d_{TV}}(P, Q) \leq \mathrm{d_{TV}}(P[S_A], Q[S_A]) + \mathrm{d_{TV}}(P[S_F], Q[S_F]) = O(\alpha).$$

This completes the proof of Theorem 32.

Appendix E. Lower Bounds for Private Distribution Estimation

In this section we prove a number of lower bounds for private distribution estimation, matching our upper bounds up to polylogarithmic factors. For estimating the mean of product or Gaussian distributions, we prove lower bounds for the weaker notion of (ε, δ)-differential privacy, but still show that they nearly match our upper bounds which are under the stronger notion of $\frac{\varepsilon^2}{2}$-zCDP. For estimating the covariance of Gaussian distributions, our lower bound is for ε-DP, a stronger notion than our upper bound, which is $\frac{\varepsilon^2}{2}$-zCDP. Proving lower bounds for covariance estimation with stronger privacy (i.e., concentrated or approximate differential privacy) is an interesting open question.

Our proofs generally consist of two parts. First, we prove a lower bound on the sample complexity required for private parameter estimation. For our lower bounds on mean estimation, we use modifications of the "fingerprinting" method. Then, we show that if two distributions are distance in parameter distance (either ℓ_2-distance between their means, or Frobenius distance between their covariances), then they will be far in statistical distance. Though we consider questions of the latter sort to be very natural, we were surprised to find they have not been studied as sigificantly as we

expected. For example, while a lower bound on the statistical distance between Gaussian distributions in terms of the ℓ_2-distance between their means is folklore, a bound in terms of the Frobenius distance between their covariance matrices is fairly recent Devroye et al. (2018). Furthermore, to the best of our knowledge, our lower bound on the statistical distance between binary product distributions in terms of the ℓ_2-distance between their means is entirely novel.

In Section E.1, we describe our lower bounds for learning product distributions. In Section E.2, we describe our lower bounds for learning Gaussian distributions with known covariance. Finally, in Section E.3, we describe our lower bounds for learning the covariance of Gaussian distributions.

E.1. Privately Learning Product Distributions

In this section we prove that our algorithm for learning product distributions has optimal sample complexity up to polylogarithmic factors. Our proof actually shows that our algorithm is optimal even if we only require the learner to work for *somewhat balanced* product distributions (i.e. those whose marginals are bounded away from 0 and 1) and allow the learner to satisfy the weaker variant of (ε, δ)-DP. The lower bound has two steps: (1) a proof that estimating the mean of a somewhat balanced product distribution up to α in ℓ_2 distance (Lemma 53) and (2) a proof that estimating a somewhat balanced product distribution in total-variation distance implies estimating its mean in ℓ_2 distance (Lemma 55). Putting these two lemmata together immediately implies the following theorem:

Theorem 52 *For any $\alpha \leq 1$ smaller than some absolute constant, any $\left(\varepsilon, \frac{3}{64n}\right)$-DP mechanism which estimates a product distribution to accuracy $\leq \alpha$ in total variation distance with probability $\geq 2/3$ requires $n = \Omega\left(\frac{d}{\alpha\varepsilon \log d}\right)$ samples.*

Proof We will show that no algorithm can estimate the mean of a product distribution up to accuracy α with probability $2/3$ with fewer than $O\left(\frac{d}{\alpha\varepsilon \log d}\right)$ samples (for an appropriate choice of the constant in the big-Oh notation). By Lemma 55, this would imply an algorithm with the same sample complexity which estimates the distribution in total variation distance up to accuracy $C\alpha$. The theorem statement follows after a rescaling of α.

Suppose that such an algorithm existed. By repeating the algorithm $O(\log d)$ times, the success probability could be boosted by a standard argument[4] to $\geq 1 - 1/d^2$, with the overall algorithm requiring $O\left(\frac{d}{\alpha\varepsilon}\right)$ samples. Since the domain is bounded, any answer will be, at worst, an $O(\sqrt{d})$-accurate estimate in ℓ_2-distance. This implies that the expected accuracy of the resulting algorithm is at most $O(\alpha)$, which is precluded by Lemma 53, for an appropriate choice of constant in the big-Oh notation. ■

Lemma 53 *If $M : \{\pm 1\}^{n \times d} \to [-\frac{1}{3}, \frac{1}{3}]^d$ is $\left(\varepsilon, \frac{3}{64n}\right)$-DP, and for every product distribution P over $\{\pm 1\}^d$ such that $-\frac{1}{3} \preceq \mathbb{E}[P] \preceq \frac{1}{3}$,*

$$\mathop{\mathbb{E}}_{X \sim P^{\otimes n}}\left[\|M(X) - \mathbb{E}[P]\|_2^2\right] \leq \alpha^2 \leq \frac{d}{54}$$

4. Specifically, repeat the algorithm $O(\log d)$ times, and choose any output which is close to at least half the outputs. This is correct with high probability by using the Chernoff bound and the fact that the original algorithm was accurate with probability $\geq 2/3$.

then $n \geq \frac{d}{72\alpha\varepsilon}$. *Equivalently, if* M *is* $\left(\varepsilon, \frac{3}{64n}\right)$*-DP and is such that for every product distribution* P *over* $\{0,1\}^d$ *such that* $\frac{1}{3} \preceq \mathbb{E}[P] \preceq \frac{2}{3}$,

$$\mathop{\mathbb{E}}_{X\sim P^{\otimes n}}\left[\|M(X) - \mathbb{E}[P]\|_2^2\right] \leq \alpha^2 \leq \frac{d}{216}$$

then $n \geq \frac{d}{144\alpha\varepsilon}$.

Proof We will only prove the first part of the theorem for estimation over $\{\pm 1\}^d$, and the second part will follow immediately by a change of variables.

Let $P^1, \ldots, P^d \sim [-\frac{1}{3}, \frac{1}{3}]$ be chosen uniformly and independently from $[-\frac{1}{3}, \frac{1}{3}]$. Let $P = \mathrm{Ber}(P^1)\otimes\cdots\otimes\mathrm{Ber}(P^d)$ be the product distribution with mean $\overline{P} = (P^1, \ldots, P^d)$. Let $X_1, \ldots, X_n \sim P$ be independent samples from this product distribution. Define:

$$Z_i^j = \left(\frac{\frac{1}{9} - (P^j)^2}{1 - (P^j)^2}\right)(M^j(X) - P^j)(X_i^j - P^j)$$
$$Z_i = \sum_{j=1}^{d} Z_i^j$$

where Z_i is a measure of the correlation between the estimate $M(X)$ and the i-th sample X_i. We will use the following key lemma, which is an extension of a similar statement in Steinke and Ullman (2015) for the uniform distribution over $[-1, 1]$.

Lemma 54 (Fingerprinting Lemma) *For every* $f : \{\pm 1\}^n \to [-\frac{1}{3}, \frac{1}{3}]$*, we have*

$$\mathop{\mathbb{E}}_{P\sim[-\frac{1}{3},\frac{1}{3}],X_{1\ldots n}\sim P}\left[\left(\frac{\frac{1}{9} - P^2}{1 - P^2}\right)\cdot(f(X) - P)\cdot\sum_{i=1}^{n}(X_i - P) + (f(X) - P)^2\right] \geq \frac{1}{27}$$

Proof [Proof of Lemma 54] Define the function

$$g(p) = \mathop{\mathbb{E}}_{X_{1\ldots n}\sim p}[f(X)].$$

For brevity, we will write $\mathop{\mathbb{E}}_{P}[\cdot]$ to indicate that the expectation is being taken over P, where P is chosen uniformly from $[-\frac{1}{3}, \frac{1}{3}]$. By (Bun et al., 2017, Lemma A.1), for every fixed p,

$$h(p) := \mathop{\mathbb{E}}_{X_{1\ldots n}\sim p}\left[\left(\frac{\frac{1}{9} - p^2}{1 - p^2}\right)\cdot(f(X) - p)\cdot\sum_{i=1}^{n}(X_i - p)\right] = \left(\frac{1}{9} - p^2\right)g'(p). \tag{15}$$

Where we have defined the function $h(p)$ for brevity. Now we have,

$$\begin{aligned}\mathop{\mathbb{E}}_{P}[h(P)] = \mathop{\mathbb{E}}_{P}\left[\left(\frac{1}{9} - P^2\right)g'(P)\right] &= \frac{3}{2}\int_{-1/3}^{1/3}\left(\frac{1}{9} - p^2\right)g'(p)dp \\ &= 2\cdot\mathop{\mathbb{E}}_{P}[Pg(P)].\end{aligned} \tag{16}$$

Now, using the above identity, we have:

$$\begin{aligned}\mathop{\mathbb{E}}_{P,X_{1\ldots n}\sim P}\Big[(f(X)-P)^2\Big] &= \mathop{\mathbb{E}}_{P,X_{1\ldots n}\sim P}\big[f(X)^2\big] + \mathop{\mathbb{E}}_{P}\big[P^2\big] - 2\cdot\mathop{\mathbb{E}}_{P}[Pg(P)] \\ &\geq \mathop{\mathbb{E}}_{P}\big[P^2\big] - 2\cdot\mathop{\mathbb{E}}_{P}[Pg(P)] \\ &= \mathop{\mathbb{E}}_{P}\big[P^2\big] - \mathop{\mathbb{E}}_{P}[h(P)] \qquad \text{(Using (16))}\end{aligned}$$

Rearranging the above inequality gives:

$$\mathop{\mathbb{E}}_{P,X_{1\ldots n}\sim P}\Big[(f(X)-P)^2\Big] + \mathbb{E}[h(P)] \geq \mathop{\mathbb{E}}_{P}\big[P^2\big] = \frac{1}{27}.$$

■

Henceforth, all expectations are taken over $\overline{P}$, X, and M. We can now apply the lemma to the function $M^j(X)$ for every $j \in [d]$, use linearity of expectation, and the accuracy assumption to get the bound,

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}[Z_i] &= \sum_{j=1}^{d}\mathbb{E}\left[\sum_{i=1}^{n} Z_i^j\right] \\ &\geq \frac{d}{27} - \mathbb{E}\big[\|M(X)-\overline{P}\|_2^2\big] \\ &\geq \frac{d}{27} - \alpha^2 \\ &\geq \frac{d}{54},\end{aligned}$$

where the second inequality follows from the assumption $\mathbb{E}\big[\|M(X)-\overline{P}\|_2^2\big] \leq \alpha^2 \leq \frac{d}{54}$.

To complete the proof, we will give an upper bound on $\sum_{i=1}^{n}\mathbb{E}[Z_i]$ that contradicts the lower bound unless n is sufficiently large. Consider any $i \in [n]$. Define:

$$\begin{aligned}\widetilde{Z}_i^j &= \left(\frac{\frac{1}{9} - (P^j)^2}{1-(P^j)^2}\right)(M(X_{\sim i}) - P^j)(X_i^j - P^j) \\ \widetilde{Z}_i &= \sum_{j=1}^{d}\widetilde{Z}_i^j\end{aligned}$$

where $X_{\sim i}$ denotes X with the i-th sample replaced with an independent draw from P. Since $X_{\sim i}$ and X_i are conditionally independent conditioned on P, $\mathbb{E}\Big[\widetilde{Z}_i\Big] = 0$. Also, we have:

$$\mathbb{E}\Big[|\widetilde{Z}_i|\Big]^2 \leq \mathbb{E}\Big[\widetilde{Z}_i^2\Big] = \mathrm{Var}\Big[\widetilde{Z}_i\Big] \leq \frac{1}{9}\mathbb{E}\big[\|M(X)-\overline{P}\|_2^2\big] \leq \frac{1}{9}\alpha^2$$

where the first inequality is Jensen's. Furthermore, we have the following upper bound on the maximum value of $\widetilde{Z}_i$ and Z_i: $\|Z_i\|_\infty \leq 8d/81$ and $\|\widetilde{Z}_i\|_\infty \leq 8d/81$.

Now we can apply differential privacy to bound $\mathbb{E}[Z_i]$, using the fact that X and $X_{\sim i}$ differ on at most one sample. The approach is akin to Lemma 8 of Steinke and Ullman (2017b). The main

idea is to split Z_i into its positive and negative components $Z_{i,+}$ and $Z_{i,-}$, write each of them as $\mathbb{E}[Z_{i,*}] = \int_0^{\|Z_i\|_\infty} \mathbb{P}[Z_{i,*} \geq t]\, dt$, and apply the definition of (ε, δ)-approximate differential privacy to relate them to the similar quantities for $\widetilde{Z}_i$. Implementing this strategy gives the following:

$$\begin{aligned}
\mathbb{E}[Z_i] &\leq \mathbb{E}\Big[\widetilde{Z}_i\Big] + 2\varepsilon \cdot \mathbb{E}\Big[|\widetilde{Z}_i|\Big] + 2\delta \cdot \|Z_i\|_\infty \\
&\leq 0 + 2\varepsilon \cdot \frac{1}{3}\alpha + \frac{3}{32n} \cdot \frac{8d}{81} \\
&\leq \frac{2}{3}\alpha\varepsilon + \frac{d}{108n}.
\end{aligned}$$

Note that we used the upper bound $e^\varepsilon - 1 \leq 2\varepsilon$ for $\varepsilon \leq 1$. Thus, we have:

$$\sum_{i=1}^{n} \mathbb{E}[Z_i] \leq \frac{2}{3}\alpha\varepsilon n + \frac{d}{108}.$$

Combining the upper and lower bounds gives:

$$\frac{d}{54} \leq \frac{2}{3}\alpha\varepsilon n + \frac{d}{108} \iff n \geq \frac{d}{72\alpha\varepsilon}.$$

This completes the proof. ■

Lemma 55 *Let P and Q be two product distributions with mean vectors p and q respectively, such that $p_i \in [1/3, 2/3]$ for all $i \in [d]$. Suppose that*

$$\|\mathbb{E}[P] - \mathbb{E}[Q]\|_2 \geq \alpha,$$

for any $\alpha \leq \alpha_0$, where $0 < \alpha_0 \leq 1$ is some absolute constant. Then $\mathrm{d_{TV}}(P, Q) \geq C\alpha$, for some absolute constant, C.

Proof Consider the set, $A = \{x \mid \log(P(x)/Q(x)) > \alpha\}$. If we show that $P(A) = \Omega(1)$, then we would have the following.

$$\begin{aligned}
&\forall x \in A \quad \frac{P(x)}{Q(x)} > e^\alpha \geq 1 + \alpha \\
\implies &\forall x \in A \quad P(x) - Q(x) > \frac{\alpha}{1+\alpha} P(x) \geq \frac{\alpha}{2} P(x) \\
\implies &P(A) - Q(A) > \frac{\alpha}{2} P(A) \\
\implies &P(A) - Q(A) > \Omega(\alpha) \\
\implies &\mathrm{d_{TV}}(P, Q) > \Omega(\alpha).
\end{aligned}$$

To this end, let $x = (x_1, \ldots, x_d) \in \{0, 1\}^d$. Then,

$$P(x) = \prod_{i=1}^{d} p_i^{x_i}(1 - p_i)^{1 - x_i} \quad \text{and} \quad Q(x) = \prod_{i=1}^{d} q_i^{x_i}(1 - q_i)^{1 - x_i}.$$

Therefore,

$$\begin{aligned} Z(x) \coloneqq \log(P(x)/Q(x)) &= \log(P(x)) - \log(Q(x)) \\ &= \sum_{i=1}^{d} x_i \log \frac{p_i}{q_i} + \sum_{i=1}^{d} (1-x_i) \log \frac{1-p_i}{1-q_i} \\ &= -\sum_{i=1}^{d} x_i \log \frac{q_i}{p_i} - \sum_{i=1}^{d} (1-x_i) \log \frac{1-q_i}{1-p_i}. \end{aligned}$$

Now, we lower bound $Z(x)$ by some function of x, so that if that function takes a value larger than α with probability $\Omega(1)$ (measured with respect to P), then $Z(x) \geq \alpha$ with probability $\Omega(1)$. Noting that $\log(t) \leq t-1$ for all $t > 0$, we get the following:

$$\begin{aligned} \log(P(x)/Q(x)) &\geq \sum_{i=1}^{d} x_i \left(1 - \frac{q_i}{p_i}\right) + \sum_{i=1}^{d} (1-x_i)\left(1 - \frac{1-q_i}{1-p_i}\right) \\ &= \sum_{i=1}^{d} \frac{(x_i - p_i)(p_i - q_i)}{p_i(1-p_i)} \end{aligned}$$

Let $Y_i = \frac{(X_i - p_i)(p_i - q_i)}{p_i(1-p_i)}$ be a transformation of the random variable $X_i \sim P_i$. To be precise, it will have the following PMF:

$$Y_i = \begin{cases} \frac{p_i - q_i}{p_i} & \text{w.p.} \quad p_i \\ -\frac{p_i - q_i}{1-p_i} & \text{w.p.} \quad 1-p_i. \end{cases}$$

Y_i has the following properties:

$$\mathbb{E}[Y_i] = 0, \quad \sigma_i^2 = \mathbb{E}\big[Y_i^2\big] = \frac{(p_i - q_i)^2}{p_i(1-p_i)}, \quad \text{and} \quad \mathbb{E}\big[Y_i^3\big] = \frac{(p_i - q_i)^3\big[p_i^2 + (1-p_i)^2\big]}{p_i^2(1-p_i)^2}.$$

Let $\sigma^2 = \sum_{i \in [d]} \sigma_i^2$, and $Y = \frac{1}{\sigma} \sum_{i=1}^{d} Y_i$. Hence, $Z(x) \geq \sigma Y$ for all x. At this point, it suffices to show that $\mathbb{P}[Y > \alpha/\sigma] \geq \Omega(1)$. We will do this in two parts: we show that if Y was a Gaussian with the same mean and variance, then this inequality would hold, and we also show that Y is well-approximated by said Gaussian. We start with the latter.

We apply the Berry-Esseen theorem Berry (1941); Esseen (1942); Shevtsova (2010) to approximate the distribution of Y by the standard normal distribution. Let ψ be the actual CDF of Y, and ϕ be the CDF of the standard normal distribution:

$$\begin{aligned} |\psi(y) - \phi(y)| &\leq C_1 \sigma^{-1} \cdot \max_i \frac{\mathbb{E}\big[Y_i^3\big]}{\mathbb{E}\big[Y_i^2\big]} \\ &= C_1 \sigma^{-1} \cdot \max_i \frac{(p_i - q_i)\big[p_i^2 + (1-p_i)^2\big]}{p_i(1-p_i)} \\ &\leq \frac{5C_1}{2} \sigma^{-1} \cdot \max_i (p_i - q_i) \end{aligned}$$

Here, $C_1 = 0.56$ is a universal constant. Now, we can assume that $p_i - q_i \leq C_2\alpha$ (for some constant C_2 of our choosing), otherwise $\mathrm{d_{TV}}(P, Q) > C_2\alpha$ trivially. Note that, by our assumption on $p_i \in [1/3, 2/3]$, we have the following:

$$\sigma^2 = \sum_{i\in[d]} \frac{(p_i - q_i)^2}{p_i(1-p_i)} \geq \frac{9}{2}\sum_{i\in[d]}(p_i - q_i)^2 = \frac{9}{2}\alpha^2.$$

Therefore, $\sigma \geq 2\alpha$, and we get the following:

$$|\psi(y) - \phi(y)| \leq \frac{5C_1C_2}{4}.$$

We now use this to prove an $\Omega(1)$ lower bound on $\mathbb{P}[Y > \alpha/\sigma]$.

$$\begin{aligned}
1 - \psi(\alpha/\sigma) &> 1 - \phi(\alpha/\sigma) - \frac{5C_1C_2}{4} \\
&= 1 - \frac{1}{\sqrt{2\pi}}\int_0^{\alpha/\sigma} e^{-t^2}dt - \frac{5C_1C_2}{4} \\
&\geq 1/2 - \frac{1}{\sqrt{2\pi}}\frac{\alpha}{\sigma} - \frac{5C_1C_2}{4} \\
&\geq 1/2 - \frac{1}{2\sqrt{2\pi}} - \frac{5C_1C_2}{4} \\
&\geq 0.30 - \frac{5C_1C_2}{4}
\end{aligned}$$

We want the above quantity to be a constant greater than zero. We could pick any "small enough" constant, so we pick 0.1. Therefore, by choosing $C_2 < 0.16/C_1$ (say 0.25), we guarantee that $\mathbb{P}[Y > \alpha/\sigma] > 0.1$. Hence, we have $\mathrm{d_{TV}}(P, Q) > 0.05\alpha$, which completes the proof. ■

E.2. Privately Learning Gaussian Distributions with Known Covariance

In this section, we will show a lower bound on the number of samples required to estimate the mean of a Gaussian distribution when its covariance matrix is known. The approach is similar to the product distribution case (Section E.1), but with modifications required for the different structure and unbounded data.

Theorem 56 *For any $\alpha \leq 1$ smaller than some absolute constant, any (ε, δ)-DP mechanism (for $\delta \leq \tilde{O}\left(\frac{\sqrt{d}}{Rn}\right)$) which estimates a Gaussian distribution (with mean $\mu \in [-R, R]^d$ and known covariance $\sigma^2\mathbb{I}$) to accuracy $\leq \alpha$ in total variation distance with probability $\geq 2/3$ requires $n = \Omega\left(\frac{d}{\alpha\varepsilon\log(dR)}\right)$ samples.*

While the expression for δ might seem complex, one can note that if $R = 1$ and for $d \geq 1$, we have $\delta = O\left(\frac{1}{n\sqrt{\log n}}\right)$, very similar to the statement of Lemma 53. Our statement is stronger and more general for settings of d and R.

Proof The proof is very similar to that of Theorem 52, so we only sketch the differences. To estimate the Gaussian to total variation distance α, it is necessary to estimate the mean in ℓ_2-distance to accuracy $\alpha\sigma$, evidenced by the following folklore fact (see, e.g., Diakonikolas et al. (2018)):

Fact 57 *The total variation distance between $\mathcal{N}(\mu_1, \sigma^2\mathbb{I})$ and $\mathcal{N}(\mu_2, \sigma^2\mathbb{I})$ is at least $C\frac{\|\mu_1-\mu_2\|_2}{\sigma}$, for an appropriate constant C, and all μ_1, μ_2, σ such that $\frac{\|\mu_1-\mu_2\|_2}{\sigma}$ is smaller than some absolute constant.*

Similar to before, we can argue that the existence of such an algorithm implies the existence of an algorithm which is correct in expectation, at a multiplicative cost of $O(\log dR)$ in the sample complexity, as any estimate output by the algorithm is accurate up to $O(\sqrt{d}R)$ in ℓ_2-distance. Such an algorithm is precluded by Lemma 58 (noting that we must rescale α by a factor of σ), concluding the proof. ■

Lemma 58 *If $M : \mathbb{R}^{n\times d} \to [-R\sigma, R\sigma]^d$ is (ε, δ)-DP for $\delta \leq \frac{\sqrt{d}}{48\sqrt{2}Rn\sqrt{\log(48\sqrt{2}Rn/\sqrt{d})}}$, and for every Gaussian distribution P with known covariance matrix, $\sigma^2\mathbb{I}$, such that $-R\sigma \leq \mathbb{E}[P] \leq R\sigma$,*

$$\mathop{\mathbb{E}}_{X\sim P^{\otimes n}}\left[\|M(X) - \mathbb{E}[P]\|_2^2\right] \leq \alpha^2 \leq \frac{d\sigma^2R^2}{6},$$

then $n \geq \frac{d\sigma}{24\alpha\varepsilon}$.

Proof By a scaling argument, we will focus on the case where $\sigma = 1$. We prove the following statement: If $M : \mathbb{R}^{n\times d} \to [-R, R]^d$ is (ε, δ)-DP for $\delta \leq \frac{\sqrt{d}}{48\sqrt{2}Rn\sqrt{\log(48\sqrt{2}Rn/\sqrt{d})}}$, and for every Gaussian distribution P with known covariance matrix $\mathbb{I}$ such that $-R \leq \mathbb{E}[P] \leq R$,

$$\mathop{\mathbb{E}}_{X\sim P^{\otimes n}}\left[\|M(X) - \mathbb{E}[P]\|_2^2\right] \leq \alpha^2 \leq \frac{dR^2}{6},$$

implies $n \geq \frac{d}{24\alpha\varepsilon}$.

Let $\mu^1, \ldots, \mu^d$ be chosen independently and uniformly at random from the interval $[-R, +R]$. Let P be the Gaussian distribution with mean vector $\overline{\mu} = (\mu^1, \ldots, \mu^d)$, and covariance matrix $\mathbb{I}$. Let $X_1, \ldots, X_n$ be independent samples from this Gaussian distribution. As in the proof of the lower bound for product distributions, we define the following quantities.

$$\begin{aligned} Z_i^j &= \left(R^2 - (\mu^j)^2\right)(M^j(X) - \mu^j)(X_i^j - \mu^j) \\ Z_i &= \sum_{j=1}^{d} Z_i^j \end{aligned}$$

Again, our strategy would be to give upper and lower bounds on $\sum_{i=1}^n \mathbb{E}[Z_i]$, which would conradict each other unless n is larger than some specific quantity. To obtain the lower bound, we first prove a lemma similar to Lemma 54.

Lemma 59 (Fingerprinting Lemma for Gaussians) *For every $f : \mathbb{R}^n \to [-R, R]$, we have*

$$\mathop{\mathbb{E}}_{\mu\sim[R,R],X_{1\ldots n}\sim\mathcal{N}(\mu,1)}\left[\left(R^2 - \mu^2\right)\cdot(f(X) - \mu)\cdot\sum_{i=1}^{n}(X_i - \mu) + (f(X) - \mu)^2\right] \geq \frac{R^2}{3}$$

Proof [Proof of Lemma 59] Define the function

$$g(\mu) = \mathop{\mathbb{E}}_{X_{1\ldots n}\sim\mathcal{N}(\mu,1)}[f(X)].$$

For brevity, we will write $\mathop{\mathbb{E}}_{\mu}[\cdot]$ to indicate that the expectation is being taken over μ, where μ is chosen uniformly from $[-R, R]$. We use an adaptation of (15) to the Gaussian setting. From an extension of a similar statement in the full version of Dwork et al. (2015), for every fixed μ,

$$h(\mu) := \mathop{\mathbb{E}}_{X_{1\ldots n}\sim\mathcal{N}(\mu,1)}\left[\left(R^2-\mu^2\right)\cdot(f(X)-\mu)\cdot\sum_{i=1}^{n}(X_i-\mu)\right] = \left(R^2-\mu^2\right)\frac{\partial}{\partial\mu}g(\mu).$$

Therefore, we get:

$$\mathop{\mathbb{E}}_{\mu}[h(\mu)] = 2\mathop{\mathbb{E}}_{\mu}[\mu g(\mu)].$$

Now, using the above, we get:

$$\begin{aligned}\mathop{\mathbb{E}}_{\mu,X_{1\ldots n}\sim\mathcal{N}(\mu,1)}\left[(f(X)-\mu)^2\right] &= \mathop{\mathbb{E}}_{\mu,X_{1\ldots n}\sim\mathcal{N}(\mu,1)}\left[f(X)^2\right] + \mathop{\mathbb{E}}_{\mu}\left[\mu^2\right] - 2\cdot\mathop{\mathbb{E}}_{\mu}[\mu g(\mu)]\\ &\geq \mathop{\mathbb{E}}_{\mu}\left[\mu^2\right] - 2\cdot\mathop{\mathbb{E}}_{\mu}[\mu g(\mu)]\\ &= \mathop{\mathbb{E}}_{\mu}\left[\mu^2\right] - \mathop{\mathbb{E}}_{\mu}[h(\mu)].\end{aligned}$$

Rearranging the above inequality gives:

$$\mathop{\mathbb{E}}_{\mu,X_{1\ldots n}\sim\mathcal{N}(\mu,1)}\left[(f(X)-\mu)^2\right] + \mathop{\mathbb{E}}_{\mu}[h(\mu)] \geq \mathop{\mathbb{E}}_{\mu}\left[\mu^2\right] = \frac{R^2}{3}.$$

■

Henceforth, all expectations are taken over $\overline{\mu}$, X, and M. In the same way as in case of product distributions, we obtain the following bound,

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}[Z_i] &= \sum_{j=1}^{d}\mathbb{E}\left[\sum_{i=1}^{n}Z_i^j\right]\\ &\geq \frac{dR^2}{3} - \mathbb{E}\left[\|M(X)-\overline{\mu}\|_2^2\right]\\ &\geq \frac{dR^2}{3} - \alpha^2\\ &\geq \frac{dR^2}{6},\end{aligned} \tag{17}$$

where the second inequality follows from the assumption $\mathbb{E}\left[\|M(X)-\overline{P}\|_2^2\right] \leq \alpha^2 \leq \frac{dR^2}{6}$. Now, to give an upper bound, we first define:

$$\begin{aligned}\widetilde{Z}_i^j &= \left(R^2-(\mu^j)^2\right)(M(X_{\sim i})-\mu^j)(X_i^j-\mu^j)\\ \widetilde{Z}_i &= \sum_{j=1}^{d}\widetilde{Z}_i^j\end{aligned}$$

where $X_{\sim i}$ denotes X with the i-th sample replaced with an independent draw from P. Because $X_{\sim i}$ and X_i are independent conditioned on P, $\mathbb{E}\left[\widetilde{Z}_i\right] = 0$. Using similar calculations as in Lemma 53, we get the following.

$$\mathbb{E}\left[|\widetilde{Z}_i|\right]^2 \leq R^4\alpha^2$$

Observe that, in contrast to Lemma 53, we do not have a worst-case bound on the value of the statistic Z_i, as the support of X_i^j is the real line, rather than just $\{\pm 1\}$ as before. Consequently, we split the computation of the expectation of $Z_{i,+}$ into the intervals $[0, T]$ and (T, ∞), and only apply (ε, δ)-DP to the former. Again, we use the ideas of the same lemma about splitting Z_i into $Z_{i,+}$ and $Z_{i,-}$ to get the following, for any $T > 0$.

$$\mathbb{E}[Z_i] \leq \mathbb{E}\left[\widetilde{Z}_i\right] + 2\varepsilon \cdot \mathbb{E}\left[|\widetilde{Z}_i|\right] + 2\delta \cdot T + 2\int_T^\infty \mathbb{P}[Z_{i,+} > t]dt \tag{18}$$

Now,

$$\begin{aligned} Z_{i,+} &\leq \max\left\{\sum_{j=1}^d (R^2 - (\mu^j)^2)(X_i^j - \mu^j)(M^j(X) - \mu^j), 0\right\} \\ &\leq \max\left\{2R^3\sum_{j=1}^d (X_i^j - \mu^j), 0\right\} \\ &= \max\{Y_i, 0\}, \end{aligned}$$

where $Y_i \sim \mathcal{N}(0, 4R^6 d)$. Let $W_i = \frac{Y_i}{2R^3\sqrt{d}}$, and $S = \frac{T}{2R^3\sqrt{d}}$. This transformation results in W_i being a standard normal random variable. We perform a change of variables, and repeatedly use the inequality $\operatorname{erfc}(x) \leq \exp\left(-x^2\right)$ in the following derivation:

$$\begin{aligned} \int_T^\infty \mathbb{P}[Z_{i,+} > t]dt &= 2R^3\sqrt{d}\int_S^\infty \mathbb{P}[W_i > s]ds \\ &\leq R^3\sqrt{d}\int_S^\infty e^{-s^2/2}ds \\ &= R^3\sqrt{d}\sqrt{\frac{\pi}{2}}\operatorname{erfc}\left(\frac{S}{\sqrt{2}}\right) \\ &\leq R^3\sqrt{\frac{d\pi}{2}}e^{-\frac{S^2}{2}}. \\ &= R^3\sqrt{\frac{d\pi}{2}}e^{-\frac{T^2}{8R^6 d}}. \end{aligned}$$

We will upper-bound this by δT:

$$R^3\sqrt{\frac{d\pi}{2}}e^{-\frac{T^2}{8R^6 d}} \leq \delta T$$

Equivalently,

$$\frac{R^3}{\delta}\sqrt{\frac{d\pi}{2}} \leq Te^{\frac{T^2}{8R^6 d}}$$

Consider setting $T = 2\sqrt{2}R^3\sqrt{d}\sqrt{\log(1/\delta)}$. The right-hand side of this inequality becomes

$$2\sqrt{2}R^3\sqrt{d}\sqrt{\log(1/\delta)} \cdot \frac{1}{\delta},$$

which is greater than the left-hand side for any $\delta < 1$.

Using (18), this gives us the following upper bound:

$$\sum_{i=1}^{n} \mathbb{E}[Z_i] \leq 2R^2\alpha\varepsilon n + 4\sqrt{2}R^3\sqrt{d}n\delta\sqrt{\log(1/\delta)}$$

On the other hand, (17) gives us a lower bound on this quantity, and thus we require that the following inequality is satisfied:

$$\frac{dR^2}{12} \leq n\Big(R^2\alpha\varepsilon + 2\sqrt{2}R^3\sqrt{d}\delta\sqrt{\log(1/\delta)}\Big)$$

Our goal is to find conditions on δ such that the product involving this term is at most $\frac{dR^2}{24}$. If this holds, the corresponding term can be moved to the left-hand side, and we are left with the following inequality:

$$\frac{dR^2}{24} \leq nR^2\alpha\varepsilon,$$

which is satisfied when

$$n \geq \frac{d}{24\alpha\varepsilon},$$

as we desired.

Thus, it remains to find conditions on δ which satisfy the following inequality:

$$2\sqrt{2}R^3\sqrt{d}\delta\sqrt{\log(1/\delta)} \leq \frac{dR^2}{24n}.$$

Rearranging, we get:

$$\delta\sqrt{\log(1/\delta)} \leq \frac{\sqrt{d}}{48\sqrt{2}Rn} \triangleq \phi.$$

Consider setting $\delta = \frac{\phi}{2\sqrt{\log(1/\phi)}}$. This results in

$$\begin{aligned}
\delta\sqrt{\log(1/\delta)} &\leq \frac{\phi}{2\sqrt{\log(1/\phi)}} \cdot \sqrt{\log(1/\phi) + \log(2\sqrt{\log(1/\phi)})} \\
&= \frac{\phi}{2} \cdot \sqrt{1 + \frac{\log(2\sqrt{\log(1/\phi)})}{\log(1/\phi)}} \\
&\leq \phi,
\end{aligned}$$

where the last inequality is because $\sqrt{1 + \frac{2\sqrt{x}}{x}} \leq 2$ for all $x \geq 0$. ■

E.3. Privately Learning Gaussian Distributions with Unknown Covariance

In this section, we prove lower bounds for privately learning a Gaussian with unknown covariance.

Theorem 60 *For any $\alpha \leq 1$ smaller than some absolute constant, any ε-DP mechanism which estimates a Gaussian distribution to accuracy $\leq \alpha$ in total variation distance with probability $\geq 2/3$ requires $n = \Omega\left(\frac{d^2}{\alpha\varepsilon}\right)$ samples.*

Proof The proof is again similar to that of Theorem 52, and we sketch the differences. The primary difference is that instead of considering algorithms which estimate the covariance matrix of the distribution in Frobenius norm, we consider algorithms which estimate the *inverse* of the covariance matrix. The reason is the following theorem of Devroye et al. (2018), which states that if one fails to estimate the inverse of the covariance matrix of a Gaussian in Frobenius norm, then one fails to estimate the Gaussian in total variation distance:

Theorem 61 (Theorem 3.8 of Devroye et al. (2018)) *Suppose there are two mean-zero Gaussian distributions $\mathcal{N}_1$ and $\mathcal{N}_2$, with covariance matrices Σ_1 and Σ_2, respectively. Furthermore, suppose that $\Sigma_1^{-1} - I$ and $\Sigma_2^{-1} - I$ are both zero-diagonal and have Frobenius norm smaller than some absolute constant. Then the total variation distance between $\mathcal{N}_1$ and $\mathcal{N}_2$ is at least $c_1\|\Sigma_1^{-1} - \Sigma_2^{-1}\|_F - c_2(\|\Sigma_1^{-1} - I\|_F^2 + \|\Sigma_2^{-1} - I\|_F^2)$, for constants $c_1, c_2 > 0$.*

Therefore, it suffices to show that there does not exist an algorithm which estimates the inverse of the covariance in Frobenius norm with probability $\geq 2/3$, where the inverse of the covariance matrix obeys the conditions of Theorem 61. As before, an algorithm which is accurate in expectation would imply the existence of such an algorithm, so we show that such an algorithm does not exist. We do this by applying a modification of Lemma 62. While this lemma is stated in terms of estimating the covariance matrix, we can obtain an identical statement for estimating the inverse of a covariance matrix by repeating the argument, with Σ replaced by Σ^{-1} at all points. Note that the construction in Lemma 62 obeys the conditions of Theorem 61. Furthermore, the Frobenius norm diameter of the construction is $\Theta(\alpha)$ (rather than $\mathrm{poly}(d)$ as in Theorem 52), we do not lose an $O(\log d)$ factor when converting to an algorithm which is accurate in expectation. Therefore, the application of this modification completes the proof. ■

Lemma 62 *If $M : \mathbb{R}^{n\times d} \to S$ is ε-DP (where S is the space of all $d \times d$ symmetric positive semi-definite matrices), and for every $\mathcal{N}(0, \Sigma)$ over $\mathbb{R}^d$ such that $\frac{1}{2}\mathbb{I} \preceq \Sigma \preceq \frac{3}{2}\mathbb{I}$,*

$$\mathop{\mathbb{E}}_{X\sim\mathcal{N}(0,\Sigma)^{\otimes n}}\left[\|M(X) - \Sigma\|_F^2\right] \leq \alpha^2/64,$$

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

Proof Let P be the uniform distribution over the set of $d \times d$ symmetric matrices with 0 on the diagonal, where the $d^2 - d$ non-zero entries are $\left\{\pm\frac{\alpha}{2d}\right\}$. Note that there are $(d^2 - d)/2$ free parameters, and thus $2^{(d^2-d)/2}$ matrices. For each $v \in \mathrm{supp}(P)$, we will define $\Sigma(v) = I + v$. It is easy to check that for all $v \in \mathrm{supp}(P)$, that $\frac{1}{2}\mathbb{I} \preceq \Sigma \preceq \frac{3}{2}\mathbb{I}$, and furthermore, that for any $v, v' \in \mathrm{supp}(P)$, $d_{\mathrm{TV}}(\mathcal{N}(0, \Sigma(v)), \mathcal{N}(0, \Sigma(v')) \leq O(\alpha)$ (Lemma 21). We assume that our algorithm is aware of

this construction, and thus will always output a symmetric matrix with 1's on the diagonal and off-diagonal entries bounded in magnitude by $\alpha/2d$.

Define the random variables Z and Z', which are sampled according to the following process. Let V and V' be independently sampled accordingly to P. X is a set of n samples from $\mathcal{N}(0, \Sigma(V))$, and similarly, X' is a set of n independent samples from $\mathcal{N}(0, \Sigma(V'))$. Then, $M(X)$ and $M(X')$ are computed with their own (independent) randomness. We then define:

$$Z = \langle M(X), V \rangle = 2 \sum_{i<j} M_{ij}(X) \cdot V_{ij},$$

$$Z' = \langle M(X'), V \rangle = 2 \sum_{i<j} M_{ij}(X') \cdot V_{ij}.$$

We start with the following claim which lower bounds the expectation of Z.

Claim 63 $\mathbb{E}[Z] \geq \frac{\alpha^2}{16} - \frac{1}{2}\|M(X) - \Sigma(V)\|_F^2 \geq \frac{7\alpha^2}{128}$.

Proof

$$\begin{aligned}
\mathbb{E}\big[2Z + \|M(X) - \Sigma(V)\|_F^2\big] &= \sum_{i<j} \mathbb{E}\big[4M_{ij}(X)V_{ij} + 2(M_{ij}(X) - \Sigma_{ij}(V))^2\big] \\
&= \sum_{i<j} \mathbb{E}\big[2M_{ij}^2(X) + 2\Sigma_{ij}^2(V)\big] \\
&\geq \sum_{i<j} \mathbb{E}\big[2\Sigma_{ij}^2(V)\big] \\
&= \sum_{i<j} \frac{\alpha^2}{2d^2} \\
&= \frac{\alpha^2}{2} \cdot \frac{d^2 - d}{2d^2} \\
&\geq \frac{\alpha^2}{8},
\end{aligned}$$

where the last inequality holds for any $d \geq 2$. The claimed statement follows by rearrangement, and the second inequality by the assumption on $\|M(X) - \Sigma(V)\|_F^2$. ■

Next, we show that Z' will not be too large, with high probability.

Claim 64 $\mathbb{P}\big[Z' > \alpha^2/32\big] \leq \exp(-\Omega(d^2))$.

We begin by observing $M(X')$ and V are independent. Condition on any realization of $M(X')$. Then $Z' \mid M(X')$ is the sum of $(d^2 - d)/2$ independent summands, each with contained in the range $[-\alpha^2/4d^2, \alpha^2/4d^2]$ and with expectation 0 (since $\mathbb{E}[V_{ij}] = 0$). By Hoeffding's inequality, we have that

$$\mathbb{P}\big[Z' > \alpha^2/32 \mid M(X')\big] \leq \exp\left(-\frac{2\frac{\alpha^4}{1024}}{\frac{d^2-d}{2} \cdot \frac{\alpha^4}{4d^4}}\right) \leq \exp\big(-\Omega(d^2)\big).$$

The claim follows by noting that value of $M(X')$ which we conditioned on was arbitrary. ■

Claim 65 $\mathbb{P}\big[Z > \alpha^2/32\big] \le \exp(O(\alpha\varepsilon n)) \cdot \mathbb{P}\big[Z' > \alpha^2/32\big]$.

Proof We start by proving the following lemma:

Lemma 66 *Let $M : \mathbb{R}^{n\times d} \to \mathcal{A}$ be an ε-DP mechanism. Suppose that D, D' are probability distributions over $\mathbb{R}^d$ such that $d_{\mathrm{TV}}(D, D') = \alpha$. Then for $S \subseteq \mathcal{A}$,*

$$\mathop{\mathbb{P}}_{M, X\sim D^n}[M(X) \in S] \le \exp(O(\varepsilon\alpha n)) \mathop{\mathbb{P}}_{M, X'\sim D'^n}\big[M(X') \in S\big].$$

Proof By the definition of D and D', this implies that there exists distributions A, B, C such that

$$D = \alpha B + (1-\alpha)A$$
$$D' = \alpha C + (1-\alpha)A$$

We will actually prove the following for any subset $S \subseteq \mathcal{A}$:

$$\mathop{\mathbb{P}}_{M, X\sim D^n}[M(X) \in S] \le \exp(O(\varepsilon\alpha n)) \mathop{\mathbb{P}}_{M, X'\sim A^n}\big[M(X') \in S\big].$$

A symmetric argument, with D' in place of D, and using the other direction of the definition of differential privacy will give the lemma statement (with an extra factor of 2 in the exponent).

We will draw X, X' from a coupling of (D, A). In particular, let $W \in \{0,1\}^n$ be a random vector, where each entry is independently set to be 1 with probability α and 0 otherwise. Note that $Y(w) \triangleq \sum_i w_i$ is distributed as $\mathrm{Bin}(n, \alpha)$. Then there exists a coupling C of (D, A) such that $X' \sim A^n$, and X_i is equal to X'_i when $W_i = 1$, and is an independent draw from B otherwise.

$$\begin{aligned}
\mathop{\mathbb{P}}_{\substack{X\sim D^n\\ M}}[M(X) \in S] &= \sum_w \mathbb{P}[W = w] \mathop{\mathbb{P}}_{\substack{(X,X')\sim C\\ M}}[M(X) \in S \mid w] \\
&\le \sum_w \mathbb{P}[W = w] \exp(\varepsilon Y(w)) \mathop{\mathbb{P}}_{\substack{(X,X')\sim C\\ M}}\big[M(X') \in S \mid w\big] \\
&= \mathop{\mathbb{P}}_{\substack{X'\sim A^n\\ M}}\big[M(X') \in S\big] \sum_w \mathbb{P}[W = w] \exp(\varepsilon Y(w)) \\
&= \mathop{\mathbb{P}}_{\substack{X'\sim A^n\\ M}}\big[M(X') \in S\big] \mathbb{E}[\exp(\varepsilon Y(w))] \\
&= (1 - \alpha + \alpha e^{\varepsilon})^n \mathop{\mathbb{P}}_{\substack{X'\sim A^n\\ M}}\big[M(X') \in S\big] \\
&\le \exp(\alpha(e^{\varepsilon} - 1)n) \mathop{\mathbb{P}}_{\substack{X'\sim A^n\\ M}}\big[M(X') \in S\big] \\
&\le \exp(O(\alpha\varepsilon n)) \mathop{\mathbb{P}}_{\substack{X'\sim A^n\\ M}}\big[M(X') \in S\big],
\end{aligned}$$

as desired. The first inequality uses the definition of differential privacy. We also used the moment generating function of the binomial distribution, and the fact that $e^{\varepsilon} - 1 = O(\varepsilon)$ for $\varepsilon < 1$. ■

With this in hand, the proof is as follows:

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\substack{V,V'\sim P\\ X\sim\mathcal{N}(0,\Sigma(V))\\ M}}\left[Z > \alpha^2/32\right] \\
&= \sum_{v,v'} \mathop{\mathbb{P}}_{V,V'\sim P}\left[V = v, V' = v'\right] \mathop{\mathbb{P}}_{\substack{X\sim\mathcal{N}(0,\Sigma(V))\\ M}}\left[Z > \alpha^2/32 \mid V, V'\right] \\
&\leq \exp(O(\alpha\varepsilon n)) \sum_{v,v'} \mathop{\mathbb{P}}_{V,V'\sim P}\left[V = v, V' = v'\right] \mathop{\mathbb{P}}_{\substack{X\sim\mathcal{N}(0,\Sigma(V'))\\ M}}\left[Z' > \alpha^2/32 \mid V, V'\right] \\
&= \exp(O(\alpha\varepsilon n)) \mathop{\mathbb{P}}_{\substack{V,V'\sim P\\ X\sim\mathcal{N}(0,\Sigma(V'))\\ M}}\left[Z' > \alpha^2/32\right]
\end{aligned}$$

The inequality follows using the lemma above, and noting that for any $v, v' \in \mathrm{supp}(P)$, that $d_{\mathrm{TV}}(\mathcal{N}(0,\Sigma(V)),\mathcal{N}(0,\Sigma(V'))) \leq O(\alpha)$. ■

With this in hand, we make the following observations. Claim 63 implies that $\Omega(1) \leq \mathbb{P}\left[Z > \alpha^2/32\right]$. Claim 64 states that $\mathbb{P}\left[Z' > \alpha^2/32\right] \leq \exp(-\Omega(d^2))$. Using these together with Claim 65 gives us that $\Omega(1) \leq \exp(O(\alpha\varepsilon n) - \Omega(d^2))$, which implies that we require $n \geq \Omega(d^2/\alpha\varepsilon)$ to avoid a contradiction.