Private Distribution Learning with Public Data

by

Alex Bie

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2023

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis is a synthesis of two studies, conducted in collaboration with Shai Ben-David, Clément L. Canonne, Gautam Kamath, and Vikrant Singhal:

- [BKS22]: Alex Bie, Gautam Kamath, and Vikrant Singhal. *Private estimation with public data*. NeurIPS 2022.
- [BDBC+23]: Shai Ben-David, Alex Bie, Clément L. Canonne, Gautam Kamath, and Vikrant Singhal. *Private distribution learning with public data: The view from sample compression*. NeurIPS 2023.

Chapter 5 covers positive results on privately learning Gaussians with public data from [BKS22].

Chapter 6 covers the connection between private distribution learning with public data and sample compression schemes, and applications thereof. Chapter 7 covers a negative result for privately learning Gaussians when we do not have enough public data. These results are from [BDBC+23].

Chapter 8 discusses a question left open by this work. The framing of the question uses a result from [BDBC+23].

Abstract

We study the problem of private distribution learning with access to public data. In this setup, a learner is given both public and private samples drawn from an unknown distribution p belonging to a class $\mathcal{Q}$, and has the task of outputting an estimate of p while adhering to privacy constraints (here, pure differential privacy) only with respect to the private samples.

Our setting is motivated by the privacy-utility tradeoff: algorithms satisfying the mathematical definition of differential privacy offer provable privacy guarantees for the data they operate on, however, owing to such a constraint, exhibit degraded accuracy. In particular, there are classes $\mathcal{Q}$ where learning is possible when privacy is not a concern, but for which any algorithm satisfying the constraint of pure differential privacy will fail on.

We show that in several scenarios, we can use a small amount of public data to evade such impossibility results. Additionally, we complement these positive results with an analysis of how much public data is necessary to see such improvements. Our main result is that to learn the class of all Gaussians in $\mathbb{R}^d$ under pure differential privacy, $d+1$ public samples suffice while d public samples are necessary.

Acknowledgements

I have a great deal of gratitude for Gautam and Shai, for their guidance, support, and pivotal ideas throughout this research. They have been extremely generous with their time, (mental) computational resources, and patience. It was the second time I visited Gautam's office when he wrote down the problem that an unsuspecting reader is about to read about. Gautam has made himself available at all times to help with everything and anything, which I can only attribute to his exceptional clarity of thought, sharp intuitions, and care for his students. Shai has a remarkable skill for chewing through complexity and distilling things to their essence, which I gladly took advantage of. I greatly appreciated his insights, as well as his words of encouragement. I am lucky to have had such great advisors; working on problems together has been a lot of fun, and I've learned a lot.

I am grateful to Vikrant, who I worked with closely on the results in this thesis. Vikrant is a great researcher and wonderful collaborator, and I appreciate his patience in putting up with a strictly less-wonderful collaborator.

Of course, my time at Waterloo was made uncountably-infinitely many times better by all the amazing, talented, and thoughtful friends and colleagues I met here: Tosca, Niki, Matt, Mahbod, Argyris, Sara, Sabrina, the list goes on... I'll miss them and look forward to catching up again soon.

I am so thankful to my parents, from whom I have received endless unconditional support. Because of this, no disaster is ever actually a disaster; 从来不感觉压力山大。

Table of Contents

Chapter 1

Introduction

INTERVIEWER: "Where do you go from here Mike?"

TYSON: "I don't know man... I might just fade into Bolivian."

– Mike Tyson, *after Lewis vs. Tyson.*

In the distribution learning problem, a learner is given a dataset of samples drawn from an unknown distribution p, belonging to a known class of distributions $\mathcal{Q}$ (e.g. all Gaussians over $\mathbb{R}$), and is tasked with outputting an estimate of p.

In addition to outputting an accurate estimate, a key requirement for such a learner may be to preserve the privacy of individuals contributing to the input dataset. Would you allow such a learner access to your data, if you knew that its output may reveal sensitive information about you?

Differential privacy (DP) [DMNS06] is a formal framework that can be used to quantify the privacy risk of a learning algorithm. Roughly speaking, an algorithm satisfying the mathematical definition of differential privacy is guaranteed to output a similar result with or without any single individual's data. Therefore, observing the output of such an algorithm cannot reveal much information about any particular individual.

Differential privacy guarantees the privacy of every point in a dataset. This is a strong requirement, and often gives rise to qualitatively new requirements in learning tasks. For instance: for the problem of learning a d-dimensional, identity covariance Gaussian with unknown mean, $\mathcal{N}(\mu, \mathbb{I})$, under pure differential privacy, the analyst must specify a range

for the unknown parameter μ, and needs more data to get an accurate estimate depending on how large this range is. This cost may be prohibitive in cases where the data domain is unfamiliar. In fact, for μ unbounded and arbitrary, the problem can be solved with $O(d^2)$ samples when privacy is not a concern – yet under pure differential privacy, no finite sample algorithm can exist.

Fortunately, in many cases, it is natural to assume that there exists an additional public dataset. This public dataset may vary in both size and quality. For example, it is common to pretrain models on large amounts of public data from the web, which may be orders of magnitude larger than the private data but significantly out of distribution. On the other hand, one can imagine that a fraction of users opt out of privacy considerations, giving a small set of public in-distribution data.

In a variety of such settings, this public data can yield dramatic theoretical and empirical improvements to utility in private data analysis (see discussion in §4). We seek to answer the following question:

How can one take advantage of public data for private distribution learning?

We initiate the study of differentially private distribution learning with a supplementary public dataset. In particular, our goal is to understand when a small amount of public data can significantly reduce the cost of private distribution learning.

Reading guide for this thesis. In Chapter 2, we focus on a simple running example, and use it to illustrate the main ideas of this thesis. We do not assume any background on learning or privacy.

On the other hand, Chapter 3 offers the general view, giving the formal statements of the results presented in this thesis, with all proofs and technical details delegated to Chapters 5, 6, and 7.

Chapter 4 reviews related work on distribution learning, with or without privacy requirements, as well as other learning problems considered under the public-private setting.

Finally, Chapter 8 states an open question.

Chapter 2

Primer: A distilled example

In this chapter, we go through a minimal running example that illustrates the problem at hand, as well as the core ideas of this thesis. No background regarding privacy or learning are assumed.

Running example: unit variance Gaussians. Let $\mathcal{Q} = \{\mathcal{N}(\mu, 1) : \mu \in \mathbb{R}\}$ denote the class of unit variance Gaussians over $\mathbb{R}$.

Our task is to design an algorithm that, given i.i.d. samples $\boldsymbol{X} = (X_1, ..., X_n)$ drawn from any unknown $\mathcal{N}(\mu, 1) \in \mathcal{Q}$, outputs an estimate of $\mathcal{N}(\mu, 1)$ that is close in total variation distance (denoted by $\mathrm{TV}(\cdot, \cdot)$; see Section 3.1). We call this task *learning* $\mathcal{Q}$. In this case, it turns out that: (1) computing the empirical mean of the samples received; and (2) outputting the unit variance Gaussian centered around it – satisfies such a guarantee.

Algorithm 1: Unit variance Gaussian learner; `UnitVarLearn`($\boldsymbol{x}$).

Input: Data $\boldsymbol{x} = (x_1, ..., x_n)$.
Output: q, a distribution over $\mathbb{R}$.

$$\widehat{\mu} \leftarrow \frac{1}{n} \sum_{j=1}^{n} x_j$$

$$q \leftarrow \mathcal{N}(\widehat{\mu}, 1)$$

Return q.

Claim 2.0.1. *For any unknown unit variance Gaussian, $\mathcal{N}(\mu, 1) \in \mathcal{Q}$, if we draw $n = O(\frac{1}{\alpha^2})$ samples $\boldsymbol{X} = (X_1, ..., X_n)$ i.i.d. from $\mathcal{N}(\mu, 1)$, with probability $\geq \frac{9}{10}$ over the sampling of $\boldsymbol{X}$, $\mathrm{TV}(\mathcal{N}(\mu, 1), \texttt{UnitVarLearn}(\boldsymbol{X})) \leq \alpha$.*

Crucially, a uniform number of samples, n, suffices for any choice of unknown $\mathcal{N}(\mu, 1)$. The proof of the claim comes from: (1) with enough samples, $\widehat{\mu}$ gets close to μ; and (2) when $|\widehat{\mu} - \mu|$ is small, so is $\mathrm{TV}(\mathcal{N}(\mu, 1), \mathcal{N}(\widehat{\mu}, 1))$.

Proof. Let $\mu \in \mathbb{R}$ be arbitrary, and suppose $\boldsymbol{X} = (X_1, ..., X_n)$ are drawn i.i.d. from $\mathcal{N}(\mu, 1)$. We compute the variance of $\widehat{\mu}$:

$$\mathrm{Var}(\widehat{\mu}) = \mathrm{Var}\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) = \frac{1}{n^2}\sum_{j=1}^{n} \mathrm{Var}\left(X_j\right) = \frac{1}{n}.$$

We have that

$$\mathbb{P}\left\{|\widehat{\mu} - \mu| > 2\alpha\right\} = \mathbb{P}\left\{|\widehat{\mu} - \mathbb{E}\widehat{\mu}| > 2\alpha\right\} \leq \frac{\mathrm{Var}(\widehat{\mu})}{4\alpha^2} = \frac{1}{4\alpha^2 n}$$

by Chebyshev's inequality. Taking $n \geq \frac{10}{4\alpha^2} = O(\frac{1}{\alpha^2})$ yields that $\mathbb{P}\left\{|\widehat{\mu} - \mu| \leq 2\alpha\right\} \geq \frac{9}{10}$. When this event occurs, we have

$$\begin{aligned}
\mathrm{TV}(\mathcal{N}(\mu, 1), \texttt{UnitVarLearn}(\boldsymbol{X})) &\coloneqq \mathrm{TV}(\mathcal{N}(\mu, 1), \mathcal{N}(\widehat{\mu}, 1)) \\
&\leq \sqrt{\frac{1}{2}\,\mathrm{KL}\left(\mathcal{N}(\mu, 1) \| \mathcal{N}(\widehat{\mu}, 1)\right)} && \text{(Pinsker's; Fact A.4.2)} \\
&= \sqrt{\frac{1}{2} \cdot \frac{(\widehat{\mu} - \mu)^2}{2}} && \text{(Gaussian KL; Fact 2.0.2)} \\
&\leq \sqrt{\frac{1}{2} \cdot \frac{4\alpha^2}{2}} \\
&= \alpha
\end{aligned}$$

as desired. The equality on the third line is from the following KL divergence identity for Gaussians. □

Fact 2.0.2 (KL divergence between 1-dimensional Gaussians [Gup20])**.** *Let $\mathcal{N}(\mu_1, \sigma_1^2)$, $\mathcal{N}(\mu_2, \sigma_2^2)$ be Gaussians over $\mathbb{R}$. Then*

$$\mathrm{KL}\left(\mathcal{N}\left(\mu_1, \sigma_1^2\right) \| \mathcal{N}\left(\mu_2, \sigma_2^2\right)\right) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.$$

What does the situation look like under privacy? In the above, $\mathcal{Q}$ admits a learning algorithm that is capable of achieving error $\leq \alpha$ with high probability on *any* distribution in $\mathcal{Q}$, so long as we receive $O(\frac{1}{\alpha^2})$ samples from it. For this problem, it is known that this is the best possible: any algorithm that succeeds on all distributions in $\mathcal{Q}$ with high probability requires $\Omega(\frac{1}{\alpha^2})$ samples on some distribution [SOAJ14].

Informally (we do not give the formal definition in this chapter; see Definition 3.2.1), differential privacy is the property of an algorithm to output similar answers with or without any single individual's data; the degree of similarity is quantified by the privacy parameter ε, where smaller ε implies more similar. Requiring such a property constrains the space of permissible algorithms for a given task. Hence, under privacy, more samples are necessarily required to accomplish the same task. As we see below, the requirement of *pure* differential privacy introduces new dependencies in the number of samples required for learning.

Fact 2.0.3 (Packing lower bound for pure DP distribution learning (specialized; see Fact A.3.2))**.** *Let $\mathcal{Q}$ be a set of distributions over $\mathbb{R}$, and let $\varepsilon > 0$. Suppose we have an ε-differentially private algorithm that can take n i.i.d. samples from any unknown $p \in \mathcal{Q}$, and with probability $\geq \frac{9}{10}$, output an estimate q with $\mathrm{TV}(p, q) \leq \alpha$.*

For any $\mathcal{P} \subseteq \mathcal{Q}$ with the property that $\mathrm{TV}(u, v) > 2\alpha$ for all $u \neq v \in \mathcal{Q}$, we have

$$n \geq \frac{\log(|\mathcal{P}|)}{2\varepsilon}.$$

Note that such a $\mathcal{P}$ is referred to as a 2α-*packing of* $\mathcal{Q}$ (Definition A.2.1).

What does this mean for our class of unit variance Gaussians $\mathcal{Q}$? The above Fact 2.0.3 says that if we can find a set of distributions in $\mathcal{Q}$ that are pairwise separated in total variation distance, then the number of samples needed learn $\mathcal{Q}$ under pure DP depends on the size of such a set. In $\mathcal{Q}$, there are infinite sets satisfying this criterion.

Fact 2.0.4 (TV distance between 1-dimensional unit variance Gaussians; (specialized; see Fact 7.0.5))**.** *Let $\mathcal{N}(\mu_1, 1)$, $\mathcal{N}(\mu_2, 1)$ be Gaussians over $\mathbb{R}$. Then*

$$\frac{1}{200} \cdot \min\{1, 40|\mu_1 - \mu_2|\} \leq \mathrm{TV}(\mathcal{N}(\mu_1, 1), \mathcal{N}(\mu_2, 1)).$$

Claim 2.0.5. *$\mathcal{P} = \{\mathcal{N}(k, 1) : k \in \mathbb{N}\}$ is a $\frac{1}{400}$-packing of $\mathcal{Q}$. Therefore for any $\varepsilon > 0$, there is no ε-differentially private algorithm that: takes a finite number of samples from any $\mathcal{N}(\mu, 1) \in \mathcal{Q}$, and with probability $\geq \frac{9}{10}$, outputs an estimate q with $\mathrm{TV}(\mathcal{N}(\mu, 1), q) \leq \frac{1}{800}$.*

To reiterate: while $\Theta(\frac{1}{\alpha^2})$ samples are necessary and sufficient to learn $\mathcal{Q}$ non-privately, no finite sample size suffices under pure differential privacy when we target error $\alpha \leq \frac{1}{800}$.

With a little help from public data. In the above, the impossibility lies in the fact that there is an unbounded range of output distributions that our algorithm must succeed on. It turns out a *single* public data point is enough to resolve this issue.

First, we consider the following pure differentially private learning algorithm, as described and analyzed in [Kam20], that handles the class of R-*bounded* univariate Gaussians, $\mathcal{Q}_R \coloneqq \{\mathcal{N}(\mu, 1) : \mu \in \mathbb{R} \text{ and } |\mu| \leq R\}$ for $R \geq 0$. The algorithm is an application of the Laplace mechanism, which: (1) zeroes out data points outside of a specified bounded range; and (2) adds Laplace-distributed noise calibrated to this range. These two steps respectively *limit* and *mask* the contribution of a single input data point.

Algorithm 2: Private unit variance Gaussian learner. $\texttt{PrivUnitVarLearn}_{R,\varepsilon}(\boldsymbol{x})$

Input: Bound on the unknown distribution mean $R \geq 0$. Target privacy parameter ε. Private data $\boldsymbol{x} = (x_1, ..., x_n)$.

Output: q, a distribution over $\mathbb{R}$.

$$c \leftarrow R + 10 + 2\log n$$

$$\widehat{\mu} \leftarrow \frac{1}{n}\sum_{j=1}^{n} x_j \cdot \mathbb{1}\{x \in \mathbb{R} : -c \leq x \leq c\}$$

$$Z \sim \text{Laplace}\left(\frac{2c}{\varepsilon n}\right)$$

$$\tilde{\mu} \leftarrow \widehat{\mu} + Z$$

$$q \leftarrow \mathcal{N}(\tilde{\mu}, 1)$$

Return q.

Where in the above, $Z \sim \text{Laplace}(b)$ corresponds to sampling from the centered Laplace distribution with density function $p(z) = \frac{1}{2b}\exp\left(-\frac{|z|}{b}\right)$.

Claim 2.0.6. *Let $R \geq 0$. For any unknown unit variance Gaussian $\mathcal{N}(\mu, 1)$ coming from the R-bounded set of Gaussians $\mathcal{Q}_R$, if we draw $n = \tilde{O}(\frac{1}{\alpha^2} + \frac{R}{\alpha\varepsilon})$ samples $\boldsymbol{X} = (X_1, ..., X_n)$ i.i.d. from $\mathcal{N}(\mu, 1)$, with probability $\geq \frac{95}{100}$ over the sampling of $\boldsymbol{X}$ and the randomness of the algorithm, $\text{TV}(\mathcal{N}(\mu, 1), \texttt{PrivUnitVarLearn}_{R,\varepsilon}(\boldsymbol{X})) \leq \alpha$. Also, $\texttt{PrivUnitVarLearn}_{R,\varepsilon}(\cdot)$ is ε-differentially private.*

The proof of Algorithm 2's privacy follows from the privacy guarantee of the Laplace mechanism; we defer it to Section B.1.

Proof. We focus on the accuracy guarantee of $\texttt{PrivUnitVarLearn}_{R,\varepsilon}(\cdot)$. First, we show that when the $X_1, ..., X_n$ are sampled from $\mathcal{N}(\mu, 1)$ with $|\mu| \le R$, with high probability, all the X_j fall within the interval $[-c, c]$.

Fix any j. $|X_j| \le |X_j - \mu| + |\mu| \le |X_j - \mu| + R$. By a Gaussian tail bound (Fact 2.0.8), $|X_j - \mu| \le 10 + 2\log n$ with probability at least

$$1 - 2\exp\left(-\frac{(10 + 2\log n)^2}{2}\right) \ge 1 - 2\exp\left(-\frac{10 + 2\log n}{2}\right) = 1 - \frac{2}{ne^5} \ge 1 - \frac{1}{50n}.$$

Hence, with probability $\ge 1 - \frac{1}{50n}$, $|X_j| \le 10 + 2\log n + R = c$. Applying the union bound yields that this holds for all the X_j simultaneously with probability $\ge \frac{98}{100}$.

When this is the case, $\widehat{\mu}$ is precisely the sample mean of Gaussian random variables, so $\mathrm{Var}(\widehat{\mu}) = \frac{1}{n}$, and we can bound its deviation from μ with Chebyshev's inequality, exactly as in Claim 2.0.1. Taking $n \ge \frac{100}{\alpha^2} = O(\frac{1}{\alpha^2})$, $\mathbb{P}\left\{|\widehat{\mu} - \mu| > \alpha\right\} \le \frac{1}{\alpha^2 n} \le \frac{1}{100}$.

Finally, we need to bound the gap between $\tilde{\mu}$ and $\widehat{\mu}$, which, by definition, is distributed according to Laplace$(\frac{2c}{\varepsilon n})$. Applying a Laplace tail bound (Fact 2.0.7), we obtain that, for $n \ge \frac{10c}{\alpha\varepsilon} = \frac{10R + 100 + 20\log n}{\alpha\varepsilon}$

$$\mathbb{P}\left\{|\tilde{\mu} - \widehat{\mu}| > \alpha\right\} = \exp\left(-\frac{\alpha\varepsilon n}{2c}\right) \le \exp\left(-5\right) \le \frac{1}{100}.$$

Note that some $n = O(\frac{R}{\alpha\varepsilon} \cdot \log\frac{R}{\alpha\varepsilon}) = \tilde{O}(\frac{R}{\alpha\varepsilon})$ indeed solves for the above requirement on n.

Therefore taking $n = \tilde{O}(\frac{1}{\alpha^2} + \frac{R}{\alpha\varepsilon})$ and applying the union bound on the above events allows us to conclude that: with probability $\ge \frac{95}{100}$, we have $|\tilde{\mu} - \widehat{\mu}| \le \alpha$ and $|\widehat{\mu} - \mu| \le \alpha$, so $|\mu - \tilde{\mu}| \le 2\alpha$ by the triangle inequality. Following the final steps in the proof of Claim 2.0.1, we can conclude $\mathrm{TV}(\mathcal{N}(\mu, 1), \texttt{PrivUnitVarLearn}_{R,\varepsilon}(\boldsymbol{X})) \coloneqq \mathrm{TV}(\mathcal{N}(\mu, 1), \mathcal{N}(\tilde{\mu}, 1)) \le \alpha$, as desired. □

Fact 2.0.7 (Laplace tail bound)**.** *Let $Z \sim$ Laplace(b), that is, Z is sampled from the centered Laplace distribution with density function $p(z) = \frac{1}{2b}\exp\left(-\frac{|z|}{b}\right)$. Then for $t \ge 0$*

$$\mathbb{P}\left\{|Z| > t\right\} = \exp\left(-\frac{t}{b}\right).$$

Fact 2.0.8 (Gaussian tail bound)**.** *Let $X \sim \mathcal{N}(\mu, 1)$. Then for $t \ge 0$*

$$\mathbb{P}\left\{|X - \mu| > t\right\} \le 2\exp\left(-\frac{t^2}{2}\right).$$

With a single public point $\tilde{X}$ sampled from our unknown and unbounded mean $\mathcal{N}(\mu, 1)$, we can reduce the unbounded-mean learning problem to a bounded-mean learning problem.

We do this by recentering all our private samples $X_1, ..., X_n$ via subtraction of $\tilde{X}$. After this transformation, our private samples can be thought of as coming from the distribution $\mathcal{N}(\mu - \tilde{X}, 1)$, which, by Gaussian concentration, is likely to be a R-bounded univariate Gaussian for small R. The bounded case can be solved completely privately with Algorithm 2. Then adding back $\tilde{X}$ yields an answer for the original problem. The detailed steps are given below in Algorithm 3.

Algorithm 3: Public-private unit variance Gaussian learner.
$\texttt{PubPrivUnitVarLearn}_{\varepsilon}(\tilde{x}, \boldsymbol{x})$

Input: Target privacy parameter ε. Public data point $\tilde{x}$. Private data $\boldsymbol{x} = (x_1, ..., x_n)$.
Output: q, a distribution over $\mathbb{R}$.

For $j = 1$ **to** n
  $y_j \leftarrow x_j - \tilde{x}$
$\boldsymbol{y} = (y_1, ..., y_n)$
$\mathcal{N}(\tilde{\mu}_Y, 1) \leftarrow \texttt{PrivUnitVarLearn}_{R=3,\varepsilon}(\boldsymbol{y})$
$\tilde{\mu} \leftarrow \tilde{\mu}_Y + \tilde{x}$
$q \leftarrow \mathcal{N}(\tilde{\mu}, 1)$

Return q.

The key aspect to realize is that for *arbitrary* μ, the information given by the single public sample $\tilde{X}$ reduces the problem to the 3-bounded case, which is solvable with private data only.

Claim 2.0.9. *For any unknown unit variance Gaussian $\mathcal{N}(\mu, 1) \in \mathcal{Q}$, if we draw 1 public sample $\tilde{X}$ and $n = \tilde{O}(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon})$ private samples $\boldsymbol{X} = (X_1, ..., X_n)$ i.i.d. from $\mathcal{N}(\mu, 1)$, with probability $\geq \frac{9}{10}$ over the sampling of $\tilde{X}, \boldsymbol{X}$, and the randomness of the algorithm, $\mathrm{TV}(\mathcal{N}(\mu, 1), \texttt{PubPrivUnitVarLearn}(\tilde{X}, \boldsymbol{X})) \leq \alpha$. Also, for any choice of public point $\tilde{x}$, $\texttt{PubPrivUnitVarLearn}_{\varepsilon}(\tilde{x}, \cdot)$ is ε-differentially private (with respect to the private data).*

We defer the proof of Algorithm 3's privacy guarantee to Section B.2. In this section, we prove the accuracy guarantee.

Proof. The inputs to the private subroutine $\texttt{PrivUnitVarLearn}_{R=3,\varepsilon}(\cdot)$, $Y_1, ..., Y_n$, are random variables defined as $Y_j \coloneqq X_j - \tilde{X}$. Hence, they are distributed $\mathcal{N}(\mu_Y, 1)$ with $\mu_Y \coloneqq \mu - \tilde{X}$. By Fact 2.0.8

$$\mathbb{P}\left\{|\mu_Y| > 3\right\} = \mathbb{P}\left\{\left|\mu - \tilde{X}\right| > 3\right\} \leq 2\exp\left(-\frac{9}{2}\right) \leq \frac{5}{100}.$$

Next, conditioned on the event that $|\mu_Y| \leq 3$, $\texttt{PrivUnitVarLearn}_{R=3,\varepsilon}(\cdot)$ indeed succeeds with probability $\geq \frac{95}{100}$, outputting $\tilde{\mu}_Y$ with $|\tilde{\mu}_Y - \mu_Y| \leq 2\alpha$ for some $n = \tilde{O}(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon})$. Note that $|\tilde{\mu} - \mu| \coloneqq |(\tilde{\mu}_Y + \tilde{X}) - (\mu_Y + \tilde{X})| = |\tilde{\mu}_Y - \mu_Y| \leq 2\alpha$. By the same steps as in Claim 2.0.1, we can conclude $\mathrm{TV}(\mathcal{N}(\mu, 1), \mathcal{N}(\tilde{\mu}, 1)) \leq \alpha$. The statement of the claim follows by the union bound. □

Takeaway. As we've seen, a single point public data point suffices to get around boundedness constraints imposed by pure differential privacy.

The strategy we take here – using public data to reduce to the bounded case, applying a known private algorithm, and then finally mapping the solution to the solution for the original unbounded problem – is quite general. In the rest of this thesis, we apply this main conceptual idea to take advantage of public data for private distribution learning of more general and interesting distribution classes.

Chapter 3

Overview

We give an overview of the results in this thesis, providing the required notation and definitions along the way.

3.1 Notation

We denote by $\mathcal{X}$ the *domain of examples*. For a domain $\mathcal{U}$, denote by $\Delta(\mathcal{U})$ the set of all probability distributions over $\mathcal{U}$.[1] We refer to a set $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ as a *class of distributions over* $\mathcal{X}$.

We equip $\Delta(\mathcal{X})$ with the *total variation* metric, which is defined as follows: for $p, q \in \Delta(\mathcal{X})$, $\mathrm{TV}(p, q) \coloneqq \sup_{B \in \mathcal{B}} |p(B) - q(B)|$, where $\mathcal{B}$ are the measurable sets of $\mathcal{X}$. For $p \in \Delta(\mathcal{X})$ and a set of distributions $L \subseteq \Delta(\mathcal{X})$, we denote their *point-set distance* by $\mathrm{dist}(p, L) \coloneqq \inf_{q \in L} \mathrm{TV}(p, q)$.

We will let $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \ldots, \tilde{x}_m) \in \mathcal{X}^m$ denote a *public dataset* and $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ denote a *private dataset*. Their respective capital versions $\tilde{\boldsymbol{X}}, \boldsymbol{X}$ denote random variables for datasets realized by sampling from some underlying distribution. For $p \in \Delta(\mathcal{X})$, we denote by p^m the distribution over $\mathcal{X}^m$ obtained by concatenating m i.i.d. samples from p.

For symmetric matrices $A, B \in \mathbb{R}^{d \times d}$, we write $A \preceq B$ if $B - A$ is positive semi-definite, and $A \prec B$ if $B - A$ is positive definite.

[1] We will assume the domain $\mathcal{U}$ is a metric space with some metric, which determines $\mathcal{B}$, the set of Borel subsets of $\mathcal{U}$, which determines the set of all probability distributions over $(\mathcal{U}, \mathcal{B})$.

3.2 Problem setup

We describe formally the notion of "public-private algorithms" – algorithms that take as input a public dataset and a private dataset, and always provides differential privacy guarantees with respect to the private dataset – as studied previously in the setting of binary classification [BNS16a, ABM19].

Definition 3.2.1 (Differential privacy [DMNS06])**.** Fix an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. Let $\varepsilon, \delta > 0$. A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \Delta(\mathcal{Y})$ is (ε, δ)-*differentially private* ((ε, δ)-*DP*), if for any private datasets $\boldsymbol{x}, \boldsymbol{x'} \in \mathcal{X}^n$ differing in one entry

$$\mathop{\mathbb{P}}_{Y \sim \mathcal{A}(\boldsymbol{x})} \{Y \in B\} \leq \exp(\varepsilon) \cdot \mathop{\mathbb{P}}_{Y' \sim \mathcal{A}(\boldsymbol{x'})} \{Y' \in B\} + \delta \qquad \text{for all measurable } B \subseteq \mathcal{Y}.$$

In this work, we focus on pure differential privacy (where $\delta = 0$), also referred to as ε-DP.

Definition 3.2.2 (Public-private ε-DP)**.** Fix an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. Let $\varepsilon > 0$. A randomized algorithm $\mathcal{A} : \mathcal{X}^m \times \mathcal{X}^n \to \Delta(\mathcal{Y})$ is *public-private* ε-*DP* if for any public dataset $\tilde{\boldsymbol{x}} \in \mathcal{X}^m$, the randomized algorithm $\mathcal{A}(\tilde{\boldsymbol{x}}, \cdot) : \mathcal{X}^n \to \Delta(\mathcal{Y})$ is ε-DP.

Definition 3.2.3 (Public-private learner)**.** Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$. For $\alpha, \beta \in (0, 1]$ and $\varepsilon > 0$, an $(\alpha, \beta, \varepsilon)$-*public-private learner for* $\mathcal{Q}$ is a public-private ε-DP algorithm $\mathcal{A} : \mathcal{X}^m \times \mathcal{X}^n \to \Delta(\Delta(\mathcal{X}))$, such that for any $p \in \mathcal{Q}$, if we draw datasets $\tilde{\boldsymbol{X}} = (\tilde{X}_1, ..., \tilde{X}_m)$ and $\boldsymbol{X} = (X_1, ..., X_n)$ i.i.d. from p and then $Q \sim \mathcal{A}(\tilde{\boldsymbol{X}}, \boldsymbol{X})$,

$$\mathop{\mathbb{P}}_{\substack{\tilde{\boldsymbol{X}} \sim p^m \\ \boldsymbol{X} \sim p^n \\ Q \sim \mathcal{A}(\tilde{\boldsymbol{X}}, \boldsymbol{X})}} \{\mathrm{TV}(Q, p) \leq \alpha\} \geq 1 - \beta.$$

Crucially, the learner must (1) satisfy DP with respect to the private data, regardless of the public data it receives as input; and (2) given a fixed amount of samples from any $p \in \mathcal{Q}$, output an accurate estimate with high probability.

Definition 3.2.4 (Public-privately learnable class)**.** We say that a class of distributions $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ is *public-privately learnable with* $m(\alpha, \beta, \varepsilon)$ *public and* $n(\alpha, \beta, \varepsilon)$ *private samples* if for any $\alpha, \beta \in (0, 1]$ and $\varepsilon > 0$, there exists an $(\alpha, \beta, \varepsilon)$-public-private learner for $\mathcal{Q}$ that takes $m = m(\alpha, \beta, \varepsilon)$ public samples and $n = n(\alpha, \beta, \varepsilon)$ private samples.

When $\mathcal{Q}$ satisfies the above, we may omit the private sample requirement, and say that $\mathcal{Q}$ is *public-privately learnable with* $m(\alpha, \beta, \varepsilon)$ *public samples.*

If a class $\mathcal{Q}$ is known to be pure privately learnable with $\mathrm{PSC}_{\mathcal{Q}}(\alpha, \beta, \varepsilon)$ (**P**rivate **S**ample **C**omplexity) samples, then $\mathcal{Q}$ is public-privately learnable with $m(\alpha, \beta, \varepsilon) = 0$ public and $n(\alpha, \beta, \varepsilon) = \mathrm{PSC}_{\mathcal{Q}}(\alpha, \beta, \varepsilon)$ private samples. In this case, we also say that $\mathcal{Q}$ is public-privately learnable with no public samples. If a class $\mathcal{Q}$ is known to be non-privately learnable with $\mathrm{SC}_{\mathcal{Q}}(\alpha, \beta)$ (**S**ample **C**omplexity) samples, then $\mathcal{Q}$ is public-privately learnable with $m(\alpha, \beta, \varepsilon) = \mathrm{SC}_{\mathcal{Q}}(\alpha, \beta)$ public and $n(\alpha, \beta, \varepsilon) = 0$ private samples. In this case, we also say that $\mathcal{Q}$ is public-privately learnable with $\mathrm{SC}_{\mathcal{Q}}(\alpha, \beta)$ public samples.

Our primary interest lies in determining when non-privately learnable $\mathcal{Q}$ can be public-privately learned with $m(\alpha, \beta, \varepsilon) = o(\mathrm{SC}_{\mathcal{Q}}(\alpha, \beta))$ public samples, at a target ε.

3.3 Privately learning d-dimensional Gaussians with $d + 1$ public samples

Learning Gaussians. In Chapter 5, we give public-private algorithms for learning Gaussians over $\mathbb{R}^d$. Note that under pure DP, we only have learners for the (R, K)-bounded case

$$\mathcal{Q}_{R,K} = \left\{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d \text{ with } \|\mu\| \leq R, \Sigma \in \mathbb{R}^{d \times d} \text{ with } \mathbb{I} \preceq \Sigma \preceq K\mathbb{I}\right\},$$

where the sample complexity $n(\alpha, \beta, \varepsilon)$ depends on R and K. The boundedness requirement turns out to be necessary: the class of all Gaussians over $\mathbb{R}^d$ cannot be pure privately learned with any finite sample complexity $n(\alpha, \beta, \varepsilon)$. However, a small amount of public data can get around this impossibility result.

Theorem 3.3.1 (Public-privately learning Gaussians (Informal; see Theorem 5.1.6))**.** *The class of d-dimensional Gaussians is public-privately learnable with $d + 1$ public and*

$$n(\alpha, \beta, \varepsilon) = \tilde{O}\left(\frac{d^2 + \log(\frac{1}{\beta})}{\alpha^2} + \frac{d^2 \log(\frac{1}{\beta})}{\alpha\varepsilon}\right)$$

private samples.

Note that $d + 1$ public samples is fewer than the known $\Theta(\frac{d^2}{\alpha^2})$ sample complexity necessary and sufficient to solve the problem non-privately. Furthermore the number of private samples used is only a mild increase over the known $\tilde{O}(\frac{d^2+\log(1/\beta)}{\alpha^2} + \frac{d^2+\log(1/\beta)}{\alpha\varepsilon})$ upper bound on learning the class of constant-bounded Gaussians (e.g. $(2, 2)$-bounded Gaussians) under pure DP (see Fact 5.1.1).

Handling public-private distribution shift. In practical settings, it is quite possibly the case that public and private data do not come from exactly the same distribution. We can relax this assumption, and show that a slight modification of the same algorithm as above works under the assumption that the public and private data come from Gaussians with bounded total variation distance.

Theorem 3.3.2 (Public-privately learning Gaussians under distribution-shift (Informal; see Theorem 5.2.3))**.** *The class of d-dimensional Gaussians is public-privately learnable with $d+1$ public and*

$$n(\alpha, \beta, \varepsilon) = \tilde{O}\left(\frac{d^2 + \log(\frac{1}{\beta})}{\alpha^2} + \frac{d^2 \log(\frac{1}{\beta(1-\gamma)})}{\alpha\varepsilon}\right)$$

private samples, in the case where the public and private data come from different Gaussians with total variation distance $\gamma < 1$.

3.4 Connections to sample compression schemes

Sample compression, public-private learning, and list learning. Ashtiani, Ben-David, Harvey, Liaw, Mehrabian, and Plan (JACM'20) introduced sample compression schemes for distribution classes, which yield nearly tight sample complexity bounds for learning mixtures of Gaussians [ABDH+20]. In Chapter 6, we adapt sample compression schemes to the public-private setting.

Specifically, we establish a connection between learning (in the sense of *distribution learning* or *density estimation*) with public and private data (Definition 3.2.4) and *sample compression schemes for distributions* (Definition 3.4.1; [ABDH+20, Definition 4.2]), as well as an intermediate notion we refer to as *list learning* (Definition 3.4.3). We find that a key parameter of a class $\mathcal{Q}$'s sample compression scheme, the *compression sample size*, corresponds to the number of public samples required to render $\mathcal{Q}$ privately learnable, as well as to the number of samples required for list learning.

We go over the definitions of sample compression schemes and list learning.

Definition 3.4.1 (Robust sample compression schemes [ABDH+20, Definition 4.2])**.** Let $r \geq 0$. We say $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ admits $(\tau(\alpha,\beta), t(\alpha,\beta), m(\alpha,\beta))$ *r-robust sample compression* if for any $\alpha, \beta \in (0,1]$, letting $\tau = \tau(\alpha,\beta)$, $t = t(\alpha,\beta)$, $m = m(\alpha,\beta)$, there exists a decoder $g : \mathcal{X}^\tau \times \{0,1\}^t \to \Delta(\mathcal{X})$, such that the following holds:

For any $q \in \mathcal{Q}$ there exists an encoder $f_q : \mathcal{X}^m \to \mathcal{X}^\tau \times \{0,1\}^t$ satisfying "for all $\boldsymbol{x} \in \mathcal{X}^m$, for all $i \in [\tau]$, $f(\boldsymbol{x})_i = \boldsymbol{x}_j$ for some $j \in [m]$", such that for every $p \in \Delta(\mathcal{X})$ with $\mathrm{TV}(p,q) \le r$, if we draw a dataset $\tilde{\boldsymbol{X}} = (\tilde{X}_1, ..., \tilde{X}_m)$ i.i.d. from p then

$$\mathop{\mathbb{P}}_{\tilde{\boldsymbol{X}} \sim p^m} \Big\{ \mathrm{TV}(g(f_q(\tilde{\boldsymbol{X}})), q) \le \alpha \Big\} \ge 1 - \beta.$$

When $\mathcal{Q}$ satisfies the above, we may omit the compression size complexity $\tau(\alpha,\beta)$ and bit complexity $t(\alpha,\beta)$, and say that $\mathcal{Q}$ is *r-robustly compressible* with $m(\alpha,\beta)$ samples. When $r = 0$, we say that $\mathcal{Q}$ admits $(\tau(\alpha,\beta), t(\alpha,\beta), m(\alpha,\beta))$ *realizable sample compression* and is *realizably compressible* with $m(\alpha,\beta)$ samples.

Definition 3.4.2 (List learner)**.** Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$. For $\alpha, \beta \in (0,1]$ and $\ell \in \mathbb{N}$, an (α,β,ℓ)*-list learner for* $\mathcal{Q}$ is an algorithm $\mathcal{L} : \mathcal{X}^m \to \{L \subseteq \Delta(\mathcal{X}) : |L| \le \ell\}$, such that for any $p \in \mathcal{Q}$, if we draw a dataset $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \ldots, \tilde{X}_m)$ i.i.d. from p then

$$\mathop{\mathbb{P}}_{\tilde{\boldsymbol{X}} \sim p^m} \Big\{ \mathrm{dist}(p, \mathcal{L}(\tilde{\boldsymbol{X}})) \le \alpha \Big\} \ge 1 - \beta.$$

Definition 3.4.3 (List learnable class)**.** A class of distributions $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ is *list learnable to list size $\ell(\alpha,\beta)$ with $m(\alpha,\beta)$ samples* if for every $\alpha, \beta \in (0,1]$, letting $\ell = \ell(\alpha,\beta)$ and $m = m(\alpha,\beta)$, there is an (α,β,ℓ)-list-learner for $\mathcal{Q}$ that takes m samples.

If $\mathcal{Q}$ satisfies the above, irrespective of the list size complexity $\ell(\alpha,\beta)$, we also say $\mathcal{Q}$ is *list learnable with $m(\alpha,\beta)$ samples.*

The following sample complexity equivalence result summarizes the sample-efficient reductions between sample compression, public-private learning, and list learning.

Theorem 3.4.4 (Sample complexity equivalence between sample compression, public-private learning, and list learning (Informal; see Theorem 6.1.1))**.** *Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$. Let $m : (0,1]^2 \to \mathbb{N}$ be a sample complexity function in terms of target error α and failure probability β, with $m(\alpha,\beta) = \mathrm{poly}(\frac{1}{\alpha}, \frac{1}{\beta})$. The following are equivalent.*

1. *$\mathcal{Q}$ is realizably compressible with $O(m(\alpha,\beta))$ samples.*
2. *$\mathcal{Q}$ is public-privately learnable with $O(m(\alpha,\beta))$ public samples.*
3. *$\mathcal{Q}$ is list learnable with $O(m(\alpha,\beta))$ samples.*

The full reductions Propositions 6.1.2, 6.1.3, and 6.1.4 also give the quantitative translations between compression size complexity and bit complexity, public sample complexity, and list size complexity.

Despite its technical simplicity, this sample complexity equivalence turns out to be quite useful, and allows us to derive new public-private learners for an array of key distribution classes by leveraging known results on sample compression schemes. From the connection to sample compression schemes we are able to obtain new public-private learners for: (1) high-dimensional Gaussian distributions (Theorem 3.4.5); (2) arbitrary mixtures of high-dimensional Gaussians (Theorem 3.4.7); (3) mixtures of public-privately learnable distribution classes (Theorem 3.4.8); and (4) products of public-privately learnable distribution classes (Theorem 3.4.10).

Learning Gaussians and mixtures of Gaussians, via compression. Via the above connection, known sample compression schemes for a class $\mathcal{Q}$ translate to public-private learners for $\mathcal{Q}$. The sample compression scheme for Gaussians given in [ABDH+20] approximately recovers the result of Theorem 3.3.1.

Theorem 3.4.5 (Public-privately learning Gaussians, via compression (Informal; see Corollary 6.2.3))**.** *The class of d-dimensional Gaussians is public-privately learnable with $m(\alpha, \beta, \varepsilon) = O(d \log(\frac{1}{\beta}))$ public and*

$$n(\alpha, \beta, \varepsilon) = \tilde{O}\left(\frac{d^2 + \log(\frac{1}{\beta})}{\alpha^2} + \frac{d^2 + \log(\frac{1}{\beta})}{\alpha\varepsilon}\right).$$

private samples.

Compared to Theorem 3.3.1, we pay an extra $O(\log(\frac{1}{\beta})$ factor in the public sample complexity, and save a $\log(\frac{1}{\beta})$ factor in the private sample complexity.

Since there are known sample compression schemes for mixtures of Gaussians, we also get public-private learners for them.

Definition 3.4.6 (Class of k-mixtures)**.** Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ and $k \geq 1$. The *class of k-mixtures of $\mathcal{Q}$*, denoted by $\mathcal{Q}^{\oplus k}$, is given by

$$\mathcal{Q}^{\oplus k} := \left\{\sum_{i=1}^{k} w_i q_i \in \Delta(\mathcal{X}) : \text{each } q_i \in \mathcal{Q}, \text{each } w_i \geq 0 \text{ with } \sum_{i=1}^{k} w_i = 1\right\}.$$

Theorem 3.4.7 (Public-privately learning mixtures of Gaussians (Informal; see Corollary 6.2.4))**.** *The class of k-mixtures of d-dimensional Gaussians is public-privately learnable with*

$$m(\alpha, \beta, \varepsilon) = \tilde{O}\left(\frac{kd\log^2(\frac{1}{\beta})}{\alpha}\right)$$

public and

$$n(\alpha, \beta, \varepsilon) = \tilde{O}\left(\frac{kd^2 + \log(\frac{1}{\beta})}{\alpha^2} + \frac{kd^2 + \log(\frac{1}{\beta})}{\alpha\varepsilon}\right)$$

private samples.

To learn mixtures of Gaussians non-privately, $\tilde{\Theta}(\frac{kd^2}{\alpha^2})$ samples are necessary and sufficient. Our public-private learner uses fewer public samples, and in the regime where $\varepsilon \geq \alpha$, uses about as many total samples.

Closure of public-private learnability under mixtures and products. If a class $\mathcal{Q}$ admits a realizable compression scheme, the class of mixtures of $\mathcal{Q}$ and the class of products of $\mathcal{Q}$ also admit realizable compression schemes. Owing to Theorem 3.4.4, the same holds for public-private learnability.

Theorem 3.4.8 (Public-privately learning mixtures (Informal; see Theorem 6.2.7))**.** *Suppose $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ is public-privately learnable with $m(\alpha, \beta, \varepsilon)$ public samples. Then for any $\varepsilon_0 > 0$, $\mathcal{Q}^{\oplus k}$, the class of k-mixtures of $\mathcal{Q}$, is learnable with*

$$m_k(\alpha, \beta, \varepsilon) = O\left(\frac{k\log(\frac{k}{\beta})}{\alpha} \cdot m\left(\frac{\alpha}{36}, \frac{\beta}{20}, \varepsilon_0\right)\right)$$

public samples.

Definition 3.4.9 (Class of k-products)**.** Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ and $k \geq 1$. The *class of k-products of $\mathcal{Q}$*, denoted by $\mathcal{Q}^{\otimes k}$, is given by

$$\mathcal{Q}^{\otimes k} \coloneqq \left\{(q_1, ..., q_k) \in \Delta(\mathcal{X}^k) : \text{each } q_i \in \mathcal{Q} \text{ and independent}\right\}.$$

Theorem 3.4.10 (Public-privately learning products (Informal; see Theorem 6.2.10))**.** *Suppose $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ is public-privately learnable with $m(\alpha, \beta, \varepsilon)$ public samples. Then for any $\varepsilon_0 > 0$, $\mathcal{Q}^{\otimes k}$, the class of k-products of $\mathcal{Q}$, is learnable with*

$$m(\alpha, \beta, \varepsilon) = O\left(\log\left(\frac{k}{\beta}\right) \cdot m\left(\frac{\alpha}{12k}, \frac{\beta}{20}, \varepsilon_0\right)\right)$$

public samples.

Agnostic and distribution-shifted public-private learning. The setting we have examined up to now has (mostly) relied on the following assumptions on the data generation process.

1. *(Same distribution).* The public and private data are sampled from the same underlying distribution.
2. *(Realizability).* The public and private data are sampled from members of the class $\mathcal{Q}$.

The exception is Theorem 5.2.3, which shows that for Gaussians over $\mathbb{R}^d$, the first condition can be relaxed (we can learn in the case where the public and the private data are generated from different Gaussians with bounded TV distance).

We show that we can relax both of the above assumptions. For distribution classes that admit robust compression schemes, we have public-private learners which: (1) can handle public-private *distribution shifts*; and (2) are *agnostic*, that is, they do not require samples to come from a member of the reference class of distributions $\mathcal{Q}$, and so instead promise error close to the best approximation of the private data distribution by a member of $\mathcal{Q}$.

We formally define the notion of *agnostic and distribution-shifted* public-private learning.

Definition 3.4.11 (Agnostic and distribution-shifted public-private learner)**.** Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$. For $\alpha, \beta \in (0,1]$, $\varepsilon > 0$, $\gamma \in [0,1]$, and $c \geq 1$ a *γ-shifted c-agnostic $(\alpha, \beta, \varepsilon)$-public-private learner for $\mathcal{Q}$* is an ε-DP public-private algorithm $\mathcal{A} : \mathcal{X}^m \times \mathcal{X}^n \to \Delta(\Delta(\mathcal{X}))$, such that for any $\tilde{p}, p \in \Delta(\mathcal{X})$ with $\mathrm{TV}(\tilde{p}, p) \leq \gamma$, if we draw a public dataset $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \ldots, \tilde{X}_m)$ i.i.d. from $\tilde{p}$, a private dataset $\boldsymbol{X} = (X_1, \ldots, X_n)$ i.i.d. from p, and then $Q \sim \mathcal{A}(\tilde{\boldsymbol{X}}, \boldsymbol{X})$

$$\mathop{\mathbb{P}}_{\substack{\tilde{\boldsymbol{X}} \sim \tilde{p}^m \\ \boldsymbol{X} \sim p^n \\ Q \sim \mathcal{A}(\tilde{\boldsymbol{X}}, \boldsymbol{X})}} \left\{ \mathrm{TV}(Q, p) \leq c \cdot \mathrm{dist}(p, \mathcal{Q}) + \alpha \right\} \geq 1 - \beta.$$

We have the following result for learning distributions that can be approximated by Gaussians.

Theorem 3.4.12 (Agnostic and distribution-shifted public-private learner for Gaussians (Informal; see Theorem 6.2.13))**.** *For any $\alpha, \beta \in (0,1]$ and $\varepsilon > 0$, there exists $\frac{1}{3}$-shifted 3-agnostic public-private learner for the class of Gaussians in $\mathbb{R}^d$ that uses m public samples and n private samples, where*

$$m(\alpha, \beta, \varepsilon) = O\left(d \log\left(\frac{1}{\beta}\right)\right) \quad \text{and} \quad n(\alpha, \beta, \varepsilon) = \tilde{O}\left(\frac{d^2 + \log(\frac{1}{\beta})}{\alpha^2} + \frac{d^2 + \log(\frac{1}{\beta})}{\alpha\varepsilon}\right).$$

3.5 A lower bound on how many public samples are needed for Gaussians

Using the aforementioned connection to list learning, we are able to establish a fine-grained lower bound on the number of public data points required to pure privately learn the class of d-dimensional Gaussians.

Theorem 3.5.1 (Almost tight lower bound on the amount of public samples required for public-privately learning Gaussians (Informal; see Theorem 7.0.1))**.** *The class of d-dimensional Gaussians is not public-privately learnable with fewer than d public samples, regardless of the number of private samples.*

Recall that Theorem 3.3.1 shows that d-dimensional Gaussians are public-privately learnable with $\tilde{O}(\frac{d^2}{\alpha^2} + \frac{d^2}{\alpha\varepsilon})$ private samples, as soon as $d+1$ public samples are available. Thus, our result shows a very sharp threshold for the number of public data points necessary and sufficient to make private learning possible.

3.6 Limitations

This work investigates the sample complexity of public-private learning, and does not give computationally efficient learners, or in some cases, *algorithmic* learners that run in finite time. In particular, all public-private learners obtained through the sample compression framework in Chapter 6, either directly by utilizing a sample compression scheme from [ABDH+20] or via the non-constructive[2] reduction of list learning to public-private learning, have exponential or infinite running times respectively. The same is the case for the VC dimension bound of Theorem 8.0.3. In there, we enumerate all labellings of the input sample realizable by the relevant Yatracos class $\mathcal{H}$, which is not a computable task for general $\mathcal{H}$ [AABD+20].

Additionally, in Theorem 8.0.3, the dependence on $\mathrm{VC}^*(\mathcal{H})$ for a general class $\mathcal{H}$ in the private sample complexity is not ideal, as $\mathrm{VC}^*(\mathcal{H}) \leq 2^{\mathrm{VC}(\mathcal{H})+1} - 1$ is the best possible upper bound in terms of $\mathrm{VC}(\mathcal{H})$ [Ass83].

[2] A comment of reviewer rJcv [rJc23, this https URL] points out an approach to address the non-constructive nature of the reduction of list learning to public-private learning (Proposition 6.1.3). To summarize: fixing the m public samples and running the public-private learner on n "null samples" repeatedly (with different random coins) produces a cover containing the true distribution with high probability. By giving up determinism in the list learner, we can get a finite time algorithm.

Finally, our lower bound for public-privately learning Gaussians in Chapter 7 establishes that at least d public samples are necessary for public-private learning to vanishingly small error as d increases – phrased alternatively, the error threshold under which we show learning is impossible decreases exponentially in d. One would hope for an error threshold for impossibility that is independent of d.

Chapter 4

Related work

The results of this thesis can be viewed in light of two main contexts: (1) the literature on distribution learning and statistical estimation under privacy constraints, in which we examine familiar problems under a modified privacy threat model; and (2) the literature on augmenting private (learning) algorithms with access to public data, in which we examine the power of the public-private model for distribution learning.

Private distribution learning. There is a long line of work on distribution learning under privacy constraints [DHS15], especially with regards to Gaussians and Gaussian mixtures. This thesis studies the task of learning arbitrary, unbounded Gaussians while offering differential privacy guarantees. Basic private algorithms for the task (variants of "clip-and-noise") impose boundedness assumptions on the underlying parameters of the unknown Gaussian, since their sample complexities grow to infinity as the bounds widen to include more allowed distributions.

Understanding these dependencies in the fully private setting has been a topic of significant study. [KV18] examined univariate Gaussians, showing that logarithmic dependencies on parameter bounds are necessary and sufficient in the case of pure DP, but can be removed under approximate DP using stability-based histograms [KKMN09, BNS16b]. The same is true in the multivariate setting: [KLSU19, BKSW21] demonstrate the necessity and sufficiency of parameter bounds under pure DP, while later works remove them under approximate DP [AAAK21, KMS+22, TCK+22, AL22, KMV22, LKO22]. Our results demonstrate that instead of relaxing the privacy definition for all the data, we can achieve similar results by employing a small amount of public data. We supplement this result with a lower bound that telling us almost exactly how much public data is needed.

For mixtures of Gaussians, most studies focus on *parameter estimation* of mixture components [NRS07, KSSU19, TCK+22, AAL23b], employing component separation and mixing weight assumptions. For *density estimation*, the setting studied in this work, [BKSW21, AAL21] give learnability results under various structural assumptions (e.g., one-dimensional, boundedness, axis-aligned). Concurrent to the publication of the results in this thesis, the study of Azfali, Ashtiani, and Liaw [AAL23a] gives the first learnability result for general, high-dimensional mixtures of Gaussians under approximate differential privacy. Our results also give learnability results for this class, in the public-private setting.

Theory for private algorithms with public data. Beyond distribution learning, there is a lot of work investigating how public data can be used improve private algorithms. Some specific areas include private query release, synthetic data generation, and prediction [JE13, BNS16a, ABM19, NB20, BCM+20, BMN20, LVS+21]. The definition of public-private algorithms that we adopt is from [BNS16a], which studied classification in the PAC model. The VC dimension bound in Chapter 8 for public-private distribution learning relies on results from public-private classification [ABM19] and uniform convergence [BCM+20].

A concurrent and independent work [LLHR23] also studies learning with public and private data, focusing on the problems of mean estimation, empirical risk minimization, and stochastic convex optimization.

Private machine learning with public data. Within the context of private machine learning, there has been significant interest in how to best employ public data. There are a variety of ways of using this data, including pretraining [ACG+16, PCS+19, TB21, LWAF21, YZCL21, LTLH22, YNB+22] (though some caution about this practice [TKC22]), computing statistics about the private gradients [ZWB21, YZCL21, KDRT21, AGM+22, GKW23], or using unlabelled public data to train a student model [PAE+17, PSM+18, BTT18]. For more discussion of public data for private learning, see Section 3.1 of [CDE+23].

Sample compression schemes. The focus of Chapter 6 is on establishing connections between public-private distribution learning and distribution sample compression schemes as introduced by [ABDH+20], as well as directly applying their results to establish new results for public-private learning. Related compression schemes for PAC learning for binary classification have been shown to be necessary and sufficient for learnability [LW86, MY16].

We also use the notion of list learning in this thesis, which is a non-robust version of the well known list-decodable learning [AAL21, RY20, BBV08, CSV17, DKS18, KS17], where

the goal is still to output a list of distributions that contains one that is accurate with respect to the true distribution, but the sampling may happen from a corrupted version of the underlying distribution. List learning and list-decodable learning are respectively strongly related to realizable and robust compression schemes [ABDM18, AAL21, AAL23a], and are useful tools for designing learners for mixture distributions.

Hybrid model of differential privacy. A related setting is the hybrid model, in which samples require either local or central differential privacy [AKZ+17]. Some learning tasks studied in this model include mean estimation [ADK20] and transfer learning [KS22].

Chapter 5

Privately learning Gaussians, with a little help from public data

To learn d-dimensional Gaussians in total variation distance under the constraint of pure differential privacy, $d+1$ public samples suffice. Moreover, the public samples can come from a different Gaussian of bounded total variation distance away. In contrast, with only private data, no finite sample complexity suffices.

5.1 $d+1$ public samples suffice to privately learn d-dimensional Gaussians

Assume we are given $d+1$ public samples $\tilde{\boldsymbol{X}} = (\tilde{X}_1, ..., \tilde{X}_{d+1})$ and n private samples $\boldsymbol{X} = (X_1, ..., X_n)$, where $\tilde{\boldsymbol{X}}$ and $\boldsymbol{X}$ are drawn i.i.d. from an unknown, d-dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$. We use the public samples to do coarse estimation of the unknown mean and covariance, and then use the coarse estimate to transform the private data, reducing to the bounded case that can be solved using existing private algorithms (such as the one given in the following Fact 5.1.1).

Fact 5.1.1 (Pure DP Gaussian estimator [BKSW21, Corollary 6.11])**.** *Let $R, K > 0$, and consider the class of (R, K)-bounded d-dimensional Gaussians*

$$\mathcal{Q} = \left\{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d \text{ with } \|\mu\| \leq R, \Sigma \in \mathbb{R}^{d \times d} \text{ with } \mathbb{I} \preceq \Sigma \preceq K\mathbb{I}\right\}.$$

Let $\alpha, \beta \in (0,1]$, and $\varepsilon > 0$. There exists an ε-DP algorithm $\texttt{DPGE}_{\alpha,\beta,\varepsilon,R,K}(\boldsymbol{x})$ that upon receiving

$$n = O\left(\frac{d^2 + \log(\frac{1}{\beta})}{\alpha^2} + \frac{d\log(\frac{dR}{\alpha}) + d^2\log(\frac{dK}{\alpha}) + \log(\frac{1}{\beta})}{\alpha\varepsilon}\right)$$

i.i.d. samples $\boldsymbol{X} = (X_1, ..., X_n)$ from any $\mathcal{N}(\mu, \Sigma) \in \mathcal{Q}$, outputs $\mu^* \in \mathbb{R}^d$ and $\Sigma^* \in \mathbb{R}^{d\times d}$ such that with probability $\geq 1-\beta$, $\mathrm{TV}(\mathcal{N}(\mu^*, \Sigma^*), \mathcal{N}(\mu, \Sigma)) \leq \alpha$.

Specifically, we use the $d+1$ public samples for a "public data preconditioning" step (Algorithm 4).

Algorithm 4: Public data preconditioning; $\texttt{PubPrecond}_{\beta}(\tilde{\boldsymbol{x}})$

Input: Public data $\tilde{\boldsymbol{x}} = (\tilde{x}_1, ..., \tilde{x}_{d+1})$. Failure probability $\beta \in (0,1]$.
Output: $\widehat{\mu} \in \mathbb{R}^d$, $\widehat{\Sigma} \in \mathbb{R}^{d\times d}$. $L, U > 0$.

`// Compute the empirical mean and covariance of` $\tilde{\boldsymbol{x}}$.

$$\widehat{\mu} \leftarrow \frac{1}{d+1}\sum_{i=1}^{d+1}\tilde{x}_i \quad \text{and} \quad \widehat{\Sigma} \leftarrow \frac{1}{d}\sum_{i=1}^{d+1}(\tilde{x}_i - \widehat{\mu})(\tilde{x}_i - \widehat{\mu})^\top$$

`// Compute` L `and` U.

$$L \leftarrow \frac{d}{4d + 4\sqrt{2d\log(\frac{3}{\beta})} + 2\log(\frac{3}{\beta})} \quad \text{and} \quad U \leftarrow \frac{9d^2}{\beta^2}$$

Return $(\widehat{\mu}, \widehat{\Sigma}, L, U)$.

The preconditioning parameters output by Algorithm 4 are used to recenter, then rescale our private samples $\boldsymbol{X} = (X_1, ..., X_n)$. The transformed private samples, which we denote by $\boldsymbol{Y} = (Y_1, ..., Y_n)$, are then fed as input to an existing DP Gaussian estimator (Fact 5.1.1), which outputs estimates μ_Y^* and Σ_Y^*. We apply the inverse of the preconditioning transform to μ_Y^* and Σ_Y^* to obtain our final estimates μ^* and Σ^*. This process is summarised in Algorithm 5.

The key step needed to establish the correctness of Algorithm 5 is to ensure that, with high probability over the sampling of public data, the parameters of the true distribution

Algorithm 5: Public-private Gaussian estimator; $\texttt{PubPrivGE}_{\alpha,\beta,\varepsilon}(\tilde{\boldsymbol{x}}, \boldsymbol{x})$

Input: Public data $\tilde{\boldsymbol{x}} = (\tilde{x}_1, ..., \tilde{x}_{d+1})$. Private data $\boldsymbol{x} = (x_1, ..., x_n)$. Desired error and failure probability $\alpha, \beta \in (0, 1]$. Privacy budget $\varepsilon > 0$.

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

`// Precondition the private data using the public data.`

$(\widehat{\mu}, \widehat{\Sigma}, L, U) \leftarrow \texttt{PubPrecond}_{\beta/2}(\tilde{\boldsymbol{x}})$

For $j \in [n]$

$$y_j \leftarrow \frac{1}{\sqrt{L}} \widehat{\Sigma}^{-1/2}(x_j - \widehat{\mu})$$

$\boldsymbol{y} = (y_1, ..., y_n)$

`// Set` R, K `parameters for the bounded DP Gaussian estimator.`

$$R \leftarrow \sqrt{\frac{U}{L}} \cdot \sqrt{5 \log\left(\frac{6}{\beta}\right)} \quad \text{and} \quad K \leftarrow \frac{U}{L}$$

`// Run the DP Gaussian estimator and invert the preconditioning.`

$(\mu_Y^*, \Sigma_Y^*) \leftarrow \texttt{DPGE}_{\alpha,\beta/2,\varepsilon,R,K}(\boldsymbol{y})$
$\mu^* \leftarrow \sqrt{L}\widehat{\Sigma}^{1/2}\mu_Y^* + \widehat{\mu} \quad \text{and} \quad \Sigma^* \leftarrow L\widehat{\Sigma}^{1/2}\Sigma_Y^*\widehat{\Sigma}^{1/2}$

Return (μ^*, Σ^*).

underlying $\boldsymbol{Y} = (Y_1, ..., Y_n)$ indeed satisfy tight range bounds that enable an existing DP Gaussian estimator to provide the desired success guarantee with private sample complexity free of dependence on R and K. That is, we want the mean of the transformed Gaussian to lie in a known $\text{poly}(d, \frac{1}{\beta})$ ball, and the condition number of its covariance to be $\text{poly}(d, \frac{1}{\beta})$.

Lemma 5.1.2 (Public data preconditioning)**.** *Let $\beta \in (0, 1]$. There exists an algorithm that takes $d+1$ samples $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \ldots, \tilde{X}_{d+1})$ drawn i.i.d. from any Gaussian $\mathcal{N}(\mu, \Sigma)$ over $\mathbb{R}^d$ and outputs $\widehat{\mu} \in \mathbb{R}^d$, $\widehat{\Sigma} \in \mathbb{R}^{d \times d}$, and $L, U > 0$, such that letting $\Sigma_Y \coloneqq \frac{1}{L}\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2}$ and $\mu_Y \coloneqq \frac{1}{\sqrt{L}}\widehat{\Sigma}^{-1/2}(\mu - \widehat{\mu})$, with probability $\geq 1 - \beta$ over the sampling of $\tilde{\boldsymbol{X}}$,*

1. $\mathbb{I} \preceq \Sigma_Y \preceq \frac{U}{L}\mathbb{I}$; *and*

2. $\|\mu_Y\| \leq \sqrt{\frac{U}{L}}\sqrt{5\log(\frac{3}{\beta})}$,

where $\frac{U}{L} = O(d^2 \log(\frac{1}{\beta})/\beta^2)$.

Lemma 5.1.2 follows from bounds on the singular values of a random Gaussian matrix. We employ the following bounds stated in [SST06].[1]

Fact 5.1.3 (Singular values of Gaussian matrices [SST06])**.** *Let $Z \in \mathbb{R}^{d\times d}$ be a matrix with each $Z_{ij} \sim N(0,1)$ independently. Denote by $\sigma_d(Z)$ the smallest singular value of Z, and by $\sigma_1(Z)$ its largest singular value. We have that*

1. $\mathbb{P}\left\{|\sigma_d(Z)| \leq \frac{\beta}{\sqrt{d}}\right\} \leq \beta$; *and*

2. $\mathbb{P}\left\{|\sigma_1(Z)| \geq 2\sqrt{d} + \sqrt{2\log(\frac{1}{\beta})}\right\} \leq \beta$.

We also use a fact about the distribution of the empirical covariance of d-dimensional standard Gaussian random variables.

Fact 5.1.4 (Properties of the Wishart Distribution)**.**[2] *Let $Z_1, ..., Z_{m+1}$ be sampled i.i.d. from $\mathcal{N}(0, \mathbb{I})$ over $\mathbb{R}^d$. Let $\widehat{\mu} = \frac{1}{m+1}\sum_{i=1}^{m+1} Z_i$. Then*

$$\frac{1}{m}\sum_{i=1}^{m+1}(Z_i - \widehat{\mu})(Z_i - \widehat{\mu})^\top \sim \frac{1}{m}\sum_{i=1}^{m} Z_i Z_i^\top.$$

That is, the two random variables are identically distributed. The random variable on the right can be recognized as a scaled d-dimensional Wishart distribution with m degrees of freedom, $\frac{1}{m}\mathcal{W}_d(m, \mathbb{I})$.

We also use the Hanson-Wright inequality for quadratic forms of d-dimensional Gaussian random variables.

Fact 5.1.5 (Hanson-Wright inequality [HW71])**.** *Let $Z \sim \mathcal{N}(0, \mathbb{I})$ and let $A \in \mathbb{R}^{d\times d}$. Then for all $t > 0$, we have*

[1] [SST06] attributes the bound on the smallest eigenvalue to Edelman [Ede88], and the bound on the largest to Davidson and Szarek [DS01].

[2] See Theorem 6 from `https://www.stat.pitt.edu/sungkyu/course/2221Fall13/lec2.pdf`.

1. $\mathbb{P}\left\{Z^\top AZ - \mathrm{tr}(A) \geq 2\|A\|_F\sqrt{t} + 2\|A\|_2 t\right\} \leq \exp(-t)$; *and*
2. $\mathbb{P}\left\{Z^\top AZ - \mathrm{tr}(A) \leq -2\|A\|_F\sqrt{t}\right\} \leq \exp(-t)$.

Proof of Lemma 5.1.2. We prove the lemma by proving the utility of Algorithm 4. Note that the quantities L, U as defined in Algorithm 4 satisfy $\frac{U}{L} = O(d^2 \log(\frac{1}{\beta})/\beta^2)$.

We start with part (1), where we wish to bound the eigenvalues of Σ_Y. Using the properties of Loewner ordering ($\preceq$) and the symmetric nature of $\widehat{\Sigma}, \Sigma$, we note that

$$\begin{aligned}\mathbb{I} \preceq \frac{1}{L}\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2} \preceq \frac{U}{L}\mathbb{I} &\iff L\widehat{\Sigma} \preceq \Sigma \preceq U\widehat{\Sigma} \\ &\iff L\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} \preceq \mathbb{I} \preceq U\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}.\end{aligned}$$

Therefore, it is sufficient to prove the final inequality. Denote $\widehat{\Sigma}_Z \coloneqq \Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}$. Then

$$\widehat{\Sigma}_Z = \frac{1}{d}\sum_{i=1}^{d+1}(\Sigma^{-1/2}(\tilde{X}_i - \widehat{\mu}))(\Sigma^{-1/2}(\tilde{X}_i - \widehat{\mu}))^\top = \frac{1}{d}\sum_{i=1}^{d+1}(Z_i - \widehat{\mu}_Z)(Z_i - \widehat{\mu}_Z)^\top$$

where for each $i \in [d+1]$, $Z_i \coloneqq \Sigma^{-1/2}(\tilde{X}_i - \mu) \sim N(0, \mathbb{I})$ independently and $\widehat{\mu}_Z \coloneqq \Sigma^{-1/2}(\widehat{\mu} - \mu) = \frac{1}{d+1}\sum_{i=1}^{d+1} Z_i$.

We show $L\widehat{\Sigma}_Z \preceq \mathbb{I}$ and $\mathbb{I} \preceq U\widehat{\Sigma}_Z$. For $i \in [d]$, let $Z_i \sim \mathcal{N}(0, \mathbb{I})$ independently. Then Fact 5.1.4 says that $\widehat{\Sigma}_Z$ is *identically distributed* to $\frac{1}{d}\sum_{i=1}^{d} Z_i Z_i^\top$. From here, we can apply the bounds from Fact 5.1.3 by noting that $\lambda_d(\widehat{\Sigma}_Z) \sim \frac{1}{d}\sigma_d\left(\left[Z_1, ..., Z_d\right]\right)^2$ and $\lambda_1(\widehat{\Sigma}_Z) \sim \frac{1}{d}\sigma_1\left(\left[Z_1, ..., Z_d\right]\right)^2$.

1. With probability $\geq 1 - \frac{\beta}{3}$, $\frac{1}{d}\sigma_d\left(\left[Z_1, ..., Z_d\right]\right)^2 > \frac{(\beta/3)^2}{d^2} \implies \lambda_d(\widehat{\Sigma}_Z) > \frac{(\beta/3)^2}{d^2} \implies U\widehat{\Sigma}_Z \succeq \mathbb{I}$.
2. With probability $\geq 1 - \frac{\beta}{3}$, $\frac{1}{d}\sigma_1\left(\left[Z_1, \ldots, Z_d\right]\right)^2 < \frac{\left(2\sqrt{d} + \sqrt{2\log(3/\beta)}\right)^2}{d} \implies \lambda_1(\widehat{\Sigma}_Z) < \frac{\left(2\sqrt{d} + \sqrt{2\log(3/\beta)}\right)^2}{d} \implies L\widehat{\Sigma}_Z \preceq \mathbb{I}$.

Taking the union bound, part (1) of Lemma 5.1.2 holds with probability $\geq 1 - \frac{2\beta}{3}$.

Next, we prove the bound on μ_Y stated in part (2).

$$\mu_Y = \frac{1}{\sqrt{L}}\widehat{\Sigma}^{-1/2}(\mu - \widehat{\mu}) = \frac{1}{\sqrt{L}}\widehat{\Sigma}^{-1/2}\left(\mu - \frac{1}{d+1}\sum_{i=1}^{d+1}(\Sigma^{1/2}Z_i + \mu)\right) = -\frac{1}{\sqrt{L}}\widehat{\Sigma}^{-1/2}\Sigma^{1/2}\widehat{\mu}_Z.$$

Since $\widehat{\mu}_Z$ is identically distributed to $\frac{1}{\sqrt{d+1}}Z_1$, applying Lemma 5.1.5 with $t = \log(\frac{3}{\beta})$ and $A = \mathbb{I}$ implies that with probability $\geq 1 - \frac{\beta}{3}$,

$$\|\widehat{\mu}_Z\| \leq \sqrt{\frac{d + 2\sqrt{d\log(\frac{3}{\beta})} + 2\log(\frac{3}{\beta})}{d+1}}$$

From (1), we know that $\left\|\frac{1}{L}\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2}\right\|_2 \leq \frac{U}{L}$. Hence, $\left\|-\frac{1}{\sqrt{L}}\widehat{\Sigma}^{-1/2}\Sigma^{1/2}\right\|_2 \leq \sqrt{\frac{U}{L}}$. This implies that

$$\|\mu_Y\| \leq \left\|-\frac{1}{\sqrt{L}}\widehat{\Sigma}^{-1/2}\Sigma^{1/2}\right\|_2 \cdot \|\widehat{\mu}_Z\| \leq \sqrt{\frac{U}{L}} \cdot \sqrt{\frac{d + 2\sqrt{d\log(\frac{3}{\beta})} + 2\log(\frac{3}{\beta})}{d+1}}$$

which implies part (2) of Lemma 5.1.2. Applying the union bound again completes the proof. □

Algorithm 5 indeed is a $(\alpha, \beta, \varepsilon)$-public-private learner using $d+1$ public samples that (a) satisfies privacy with respect to private data $\boldsymbol{x} = (x_1, ..., x_n)$; and (b) achieves the desired accuracy and success probability. This follows from the guarantees of public data preconditioning (Lemma 5.1.2) combined with the guarantees of existing DP Gaussian estimators (Fact 5.1.1). We note that our sample complexity no longer depends on the *a priori* bounds on the mean and the covariance of the unknown private data distribution.

Theorem 5.1.6 (Public-private Gaussian estimator using $d+1$ public samples)**.** *Let $\alpha, \beta \in (0, 1]$, and $\varepsilon > 0$. There exists a public-private ε-DP algorithm that takes $d+1$ public samples $\tilde{\boldsymbol{X}} = (\tilde{X}_1, ..., \tilde{X}_{d+1})$ and*

$$n = O\left(\frac{d^2 + \log(\frac{1}{\beta})}{\alpha^2} + \frac{d^2\log(\frac{d}{\alpha\beta})}{\alpha\varepsilon}\right)$$

private samples $\boldsymbol{X} = (X_1, ..., X_n)$ drawn i.i.d. from any unknown d-dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$, and outputs $\mu^* \in \mathbb{R}^d$ and $\Sigma^* \in \mathbb{R}^{d\times d}$ such that with probability $\geq 1 - \beta$, $\mathrm{TV}(\mathcal{N}(\mu^*, \Sigma^*), \mathcal{N}(\mu, \Sigma)) \leq \alpha$.

Proof. We prove the privacy and the utility guarantees for Algorithm 5.

We start with the utility guarantee. Using the public data preconditioning parameters obtained from running Algorithm 4 on public samples $\tilde{\boldsymbol{X}} = (\tilde{X}_1, ...\tilde{X}_{d+1})$ and targeting

failure probability $\frac{\beta}{2}$, we apply a shift and scale on the private samples $\boldsymbol{X} = (X_1, ..., X_n)$, yielding $\boldsymbol{Y} = (Y_1, ..., Y_n)$. For each $j \in [n]$, we have that $Y_j \sim \mathcal{N}(\mu_Y, \Sigma_Y)$, where μ_Y and Σ_Y are quantities as defined in Lemma 5.1.2. Then with probability $\geq 1 - \frac{\beta}{2}$ over the sampling of public data, we have $\mathbb{I} \preceq \Sigma_Y \preceq \frac{U}{L}\mathbb{I}$ and $\|\mu_Y\| \leq \sqrt{\frac{U}{L}} \cdot \sqrt{5\log(\frac{6}{\beta})}$ (Lemma 5.1.2).

Hence, we can set $K = \frac{U}{L} = O(d^2 \log(\frac{1}{\beta})/\beta^2)$, and $R = \sqrt{\frac{U}{L}} \cdot \sqrt{5\log(\frac{6}{\beta})} = O(d\log(\frac{1}{\beta})/\beta)$, and run the ε-DP Gaussian estimator (Fact 5.1.1) on $\boldsymbol{Y}$ with target accuracy α and failure probability $\frac{\beta}{2}$. We obtain our private sample complexity, which is now independent of mean and covariance of the underlying distribution, by plugging in these values into the DP Gaussian estimator's sample complexity. Under these parameter settings and number of private samples used, by Fact 5.1.1 and the union bound, we have that with probability $\geq 1-\beta$, the algorithm succeeds in outputting μ_Y^* and Σ_Y^*, such that $\mathrm{TV}(\mathcal{N}(\mu_Y^*, \Sigma_Y^*), \mathcal{N}(\mu_Y, \Sigma_Y)) \leq \alpha$. We output the estimates $\Sigma^* := L\widehat{\Sigma}^{1/2}\Sigma_Y^*\widehat{\Sigma}^{1/2}$ and $\mu^* := \sqrt{L}\widehat{\Sigma}^{1/2}\mu_Y^* + \widehat{\mu}$. Denoting $A := \frac{1}{L}\widehat{\Sigma}^{-1}$, by the properties of the Mahalanobis norm $\|\cdot\|_\Sigma$ (see Section 3.1)

$$\begin{aligned}
\|\Sigma^* - \Sigma\|_\Sigma &= \|A^{1/2}\Sigma^* A^{1/2} - A^{1/2}\Sigma A^{1/2}\|_{A^{1/2}\Sigma A^{1/2}} \\
&= \|\Sigma_Y^* - \Sigma_Y\|_{\Sigma_Y}, \\
\|\mu^* - \mu\|_\Sigma &= \|A^{1/2}\mu^* - A^{1/2}\mu\|_{A^{1/2}\Sigma A^{1/2}} \\
&= \|(\mu_Y^* + A^{1/2}\widehat{\mu}) - (\mu_Y + A^{1/2}\widehat{\mu})\|_{\Sigma_Y} \\
&= \|\mu_Y^* - \mu_Y\|_{\Sigma_Y}.
\end{aligned}$$

which implies that $\mathrm{TV}(\mathcal{N}(\mu^*, \Sigma^*), \mathcal{N}(\mu, \Sigma)) = \mathrm{TV}(\mathcal{N}(\mu_Y^*, \Sigma_Y^*), \mathcal{N}(\mu_Y, \Sigma_Y)) \leq \alpha$.

To argue about privacy, note that releasing (μ^*, Σ^*) is ε-DP with respect to $\boldsymbol{y} = (y_1, ..., y_n)$, since it is a post-processing (Fact A.3.1) of the output (μ_Y^*, Σ_Y^*) of an ε-DP algorithm. To argue about ε-DP with respect to the private dataset $\boldsymbol{x} = (x_1, ..., x_n)$, note that for any fixed public dataset $\tilde{\boldsymbol{x}}$, the ε-DP Gaussian estimator $\texttt{DPGE}_{\alpha,\beta/2,\varepsilon,R,K}(\cdot)$'s privacy guarantee holds for the arbitrary replacement of any single example y_j. Since each x_j maps to exactly one x_j, $\texttt{DPGE}_{\alpha,\beta/2,\varepsilon,R,K}(\cdot)$'s privacy guarantee holds for the arbitrary replacement of any single x_j as well. This gives us the final privacy guarantee with respect to $\boldsymbol{x}$. □

5.2 Handling public-private distribution shift

A natural question to ask is: what if our public data does not come from the same distribution as our private data? In practical settings, there may be distribution shift between data that is publicly available and our private data.

We relax the assumption that our public and private data are sampled from the same underlying distribution, and show that it suffices that our $d+1$ public samples comes from another Gaussian of bounded total variation distance from the private data distribution.

It turns out that the same Algorithm 5 works for this case, after slight modifications: R and K must be set based on a known upper bound γ on the total variation distance between the public and private data distributions. We employ the following Lemma 5.2.1 that translates a total variation bound into a bound on the difference between Gaussian parameters.

Lemma 5.2.1 (Total variation to Gaussian parameters bound)**.** *Let $\gamma > 0$. Let $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\tilde{\mu}, \tilde{\Sigma})$ be d-dimensional Gaussians over $\mathbb{R}^d$ such that $\mathrm{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\tilde{\mu}, \tilde{\Sigma})) \leq \gamma$. We have*

1. $\frac{(1-\gamma)^4}{4}\Sigma \preceq \tilde{\Sigma} \preceq \frac{4}{(1-\gamma)^4}\Sigma$; *and*
2. $(\mu - \tilde{\mu})(\mu - \tilde{\mu})^\top \preceq \frac{8\gamma}{1-\gamma}(\Sigma + \tilde{\Sigma})$.

The proof of Lemma 5.2.1 can be found in Section B.3.

Now we discuss the modifications. Let L, U be quantities as defined in Algorithm 4. Given a known upper bound $1 > \gamma \geq \mathrm{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\tilde{\mu}, \tilde{\Sigma}))$ between the private and public data distributions, we let $L_\gamma = \frac{(1-\gamma)^4}{4}L$ and $U_\gamma = \frac{4}{(1-\gamma)^4}U$. The following is an analogue of Lemma 5.1.2 when we run the modified Algorithm 4 on public data that comes from a Gaussian of at most total variation distance $\gamma < 1$ from the private data distribution.

Lemma 5.2.2 (Distribution-shifted public data preconditioning)**.** *Let $\beta \in (0, 1]$ and $\gamma \in [0, 1)$. There exists an algorithm that takes $d+1$ samples $\tilde{\boldsymbol{X}} = (\tilde{X}_1, ..., \tilde{X}_{d+1})$ drawn i.i.d. from any Gaussian $\mathcal{N}(\tilde{\mu}, \tilde{\Sigma})$ over $\mathbb{R}^d$, and outputs $\widehat{\mu} \in \mathbb{R}^d$, $\widehat{\Sigma} \in \mathbb{R}^{d\times d}$, and $L_\gamma, U_\gamma > 0$, such that for any Gaussian $\mathcal{N}(\mu, \Sigma)$ over $\mathbb{R}^d$ with $\mathrm{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\tilde{\mu}, \tilde{\Sigma})) \leq \gamma$, letting $\mu_Y = \frac{1}{\sqrt{L_\gamma}}\widehat{\Sigma}^{-1/2}(\mu - \widehat{\mu})$ and $\Sigma_Y = \frac{1}{L_\gamma}\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2}$, with probability $\geq 1 - \beta$ over the sampling of $\tilde{\boldsymbol{X}}$,*

1. $\mathbb{I} \preceq \Sigma_Y \preceq \frac{U_\gamma}{L_\gamma}\mathbb{I}$; *and*
2. $\|\mu_Y\| \leq \sqrt{\frac{U_\gamma}{L_\gamma}}\left(\sqrt{\frac{10\gamma}{1-\gamma}} + \sqrt{5\log(\frac{3}{\beta})}\right)$,

where $\frac{U_\gamma}{L_\gamma} = O(d^2\log(\frac{1}{\beta})/\beta^2(1-\gamma)^8)$.

Proof. We prove the lemma by proving the utility of a modified version of Algorithm 4, which returns L_γ and U_γ instead of L and U. The result follows from tracing through the proof of Lemma 5.1.2, and applying Lemma 5.2.1 as necessary. We highlight the differences.

We start with part (1). By the same chain of equivalences in the proof of Lemma 5.1.2, it suffices to show that $L_\gamma \Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} \preceq \mathbb{I} \preceq U_\gamma \Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}$. We have the following

$$\begin{aligned}
\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} &= \Sigma^{-1/2}\left(\frac{1}{d}\sum_{i=1}^{d+1}((\tilde{\Sigma}^{1/2}Z_i + \tilde{\mu}) - \widehat{\mu})((\tilde{\Sigma}^{1/2}Z_i + \tilde{\mu}) - \widehat{\mu})^\top\right)\Sigma^{-1/2} \\
&= \Sigma^{-1/2}\left(\frac{1}{d}\sum_{i=1}^{d+1}(\tilde{\Sigma}^{1/2}(Z_i - \widehat{\mu}_Z))(\tilde{\Sigma}^{1/2}(Z_i - \widehat{\mu}_Z))^\top\right)\Sigma^{-1/2} \\
&= \Sigma^{-1/2}\tilde{\Sigma}^{1/2}\widehat{\Sigma}_Z\tilde{\Sigma}^{1/2}\Sigma^{-1/2}.
\end{aligned}$$

In the above, $Z_i \coloneqq \tilde{\Sigma}^{-1/2}(\tilde{X}_i - \tilde{\mu}) \sim \mathcal{N}(0, \mathbb{I})$ independently, $\widehat{\mu}_Z \coloneqq \tilde{\Sigma}^{-1/2}(\widehat{\mu} - \tilde{\mu}) = \frac{1}{d+1}\sum_{i=1}^{d+1} Z_i$, and $\widehat{\Sigma}_Z \coloneqq \frac{1}{d}\sum_{i=1}^{d+1}(Z_i - \widehat{\mu}_Z)(Z_i - \widehat{\mu}_Z)^\top$ as in the proof of Lemma 5.1.2. From the same proof, we know that with probability $\geq 1 - \frac{\beta}{3}$, we have $U\widehat{\Sigma}_Z \succeq \mathbb{I}$, which implies that

$$U\Sigma^{-1/2}\tilde{\Sigma}^{1/2}\widehat{\Sigma}_Z\tilde{\Sigma}^{1/2}\Sigma^{-1/2} \succeq \Sigma^{-1/2}\tilde{\Sigma}\Sigma^{-1/2} \succeq \frac{(1-\gamma)^4}{4}\mathbb{I},$$

where the last inequality follows from (1) in Lemma 5.2.1. Recalling that we set $U_\gamma = \frac{4}{(1-\gamma)^4}U$, rearranging gives us that $U_\gamma\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} \succeq \mathbb{I}$, as desired.

Similarly, with probability $\geq 1 - \frac{\beta}{3}$, $L\widehat{\Sigma}_Z \preceq \mathbb{I}$, which implies that

$$L\Sigma^{-1/2}\tilde{\Sigma}^{1/2}\widehat{\Sigma}_Z\tilde{\Sigma}^{1/2}\Sigma^{-1/2} \preceq \Sigma^{-1/2}\tilde{\Sigma}\Sigma^{-1/2} \preceq \frac{4}{(1-\gamma)^4}\mathbb{I},$$

where the last inequality follows from (1) in Lemma 5.2.1. Recalling that we set $L_\gamma = \frac{(1-\gamma)^4}{4}L$, rearranging allows us to conclude $L_\gamma\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} \preceq \mathbb{I}$.

It remains to verify that part (2) holds. Write

$$\begin{aligned}
\mu_Y \coloneqq \frac{1}{\sqrt{L_\gamma}}\widehat{\Sigma}^{-1/2}(\mu - \widehat{\mu}) &= \frac{1}{\sqrt{L_\gamma}}\widehat{\Sigma}^{-1/2}(\mu - \tilde{\mu} - \tilde{\Sigma}^{1/2}\widehat{\mu}_Z) \\
&= \frac{1}{\sqrt{L_\gamma}}\widehat{\Sigma}^{-1/2}(\mu - \tilde{\mu}) - \frac{1}{\sqrt{L_\gamma}}\widehat{\Sigma}^{-1/2}\tilde{\Sigma}^{1/2}\widehat{\mu}_Z
\end{aligned}$$

We bound the two terms separately. Note that the second term appears in the proof of Lemma 5.1.2. By the same argument as in that proof, we claim that with probability $\geq 1 - \frac{\beta}{3}$,

$$\|\widehat{\mu}_Z\| \leq \sqrt{\frac{d + 2\sqrt{d\log(\frac{3}{\beta})} + 2\log(\frac{3}{\beta})}{d+1}}.$$

When we indeed have $U\widehat{\Sigma}_Z \succeq \mathbb{I}$ (that is, our events from (1) occur), we have $\left\|\widehat{\Sigma}^{-1/2}\tilde{\Sigma}^{1/2}\right\|_2 \leq \sqrt{U}$. Taking the union bound, with probability $\geq 1 - \beta$,

$$\left\| -\frac{1}{\sqrt{L_\gamma}}\widehat{\Sigma}^{-1/2}\tilde{\Sigma}^{1/2}\widehat{\mu}_Z \right\| \leq \sqrt{\frac{U}{L_\gamma}} \cdot \sqrt{\frac{d + 2\sqrt{d\log(\frac{3}{\beta})} + 2\log(\frac{3}{\beta})}{d+1}}.$$

Now, we argue that the first term is also bounded. First, we apply Lemma 5.2.1 to get

$$(\mu - \tilde{\mu})(\mu - \tilde{\mu})^\top \preceq \frac{8\gamma}{1-\gamma}(\Sigma + \tilde{\Sigma}) \preceq \frac{8\gamma}{1-\gamma}\left(\frac{4}{(1-\gamma)^4}\tilde{\Sigma} + \tilde{\Sigma}\right) \preceq \frac{40\gamma}{(1-\gamma)^5}\tilde{\Sigma}.$$

Note that $U\widehat{\Sigma}_Z \succeq \mathbb{I} \implies \tilde{\Sigma} \preceq U\widehat{\Sigma}$. Plugging this in above and rearranging gives

$$\widehat{\Sigma}^{-1/2}(\mu - \tilde{\mu})(\mu - \tilde{\mu})^\top\widehat{\Sigma}^{-1/2} \preceq U\frac{40\gamma}{(1-\gamma)^5}\mathbb{I} = U_\gamma\frac{10\gamma}{1-\gamma}\mathbb{I}.$$

Thus, we have

$$\begin{aligned}\left\|\frac{1}{\sqrt{L_\gamma}}\widehat{\Sigma}^{-1/2}(\mu - \tilde{\mu})\right\| &\leq \frac{1}{\sqrt{L_\gamma}} \cdot \sqrt{\left\|\widehat{\Sigma}^{-1/2}(\mu - \tilde{\mu})(\mu - \tilde{\mu})^\top\widehat{\Sigma}^{-1/2}\right\|_2} \\ &\leq \sqrt{\frac{U_\gamma}{L_\gamma}} \cdot \sqrt{\frac{10\gamma}{1-\gamma}}.\end{aligned}$$

Combining the two terms gives us

$$\|\mu_Y\| \leq \sqrt{\frac{U_\gamma}{L_\gamma}} \cdot \left(\sqrt{\frac{10\gamma}{1-\gamma}} + \sqrt{\frac{d + 2\sqrt{d\log(\frac{3}{\beta})} + 2\log(\frac{3}{\beta})}{d+1}}\right),$$

which completes the proof. □

Lemma 5.2.2, combined with guarantees of $\texttt{DPGE}_{\alpha,\beta,\varepsilon,R,K}(\cdot)$ from Fact 5.1.1, gives us the following analogue to Theorem 5.1.6 for the public-private distribution shift case.

Theorem 5.2.3 (Distribution-shifted public-private Gaussian estimator using $d+1$ public samples)**.** *Let $\alpha, \beta \in (0,1]$, $\gamma \in [0,1)$, and $\varepsilon > 0$. There exists a public-private ε-DP algorithm that takes $d+1$ public samples $\tilde{\boldsymbol{X}} = (\tilde{X}_1, ..., \tilde{X}_{d+1})$ drawn i.i.d. from an unknown d-dimensional Gaussian $\mathcal{N}(\tilde{\mu}, \tilde{\Sigma})$, along with*

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

private samples $\boldsymbol{X} = (X_1, ..., X_n)$ drawn i.i.d. from an unknown d-dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$, such that $\mathrm{TV}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\tilde{\mu}, \tilde{\Sigma})) \leq \gamma$, and outputs $\mu^* \in \mathbb{R}^d$ and $\Sigma^* \in \mathbb{R}^{d \times d}$ such that with probability $\geq 1 - \beta$, $\mathrm{TV}(\mathcal{N}(\mu^*, \Sigma^*), \mathcal{N}(\mu, \Sigma)) \leq \alpha$.

Theorem 5.2.3 follows from the privacy and the utility guarantees of a modified version of Algorithm 5, which uses the modified version of Algorithm 4 as outlined in Lemma 5.2.2 (outputting L_γ and U_γ instead of L and U). The proof is the same as that of Theorem 5.1.6.

Chapter 6

The connection to sample compression schemes

When privacy is not a concern, sample compression schemes for distribution classes [ABDH+20] yield nearly tight sample complexity bounds for learning mixtures of Gaussians. We adapt them to the public-private setting, finding that a key parameter of sample compression schemes, the *compression sample size*, is within a constant factor of the number of public samples necessary and sufficient to render a class privately learnable.

Leveraging the connection (approximately) recovers the sample-efficient public-private learner for Gaussians discussed in the previous chapter, and also yields sample-efficient learners for arbitrary k-mixtures of Gaussians. It also gives us results on the closure properties of public-private learnability, as well as the agnostic and distribution-shifted case.

6.1 Reductions between sample compression, public-private learning, and list learning

In this section, we give sample-efficient reductions between sample compression, public private learning, and an intermediate notion we refer to as *list learning*.

Sample compression schemes. If each member q of a class of distributions $\mathcal{Q}$ admits a way to encode enough information about itself in a small number of samples from q and

extra bits (with high probability over sampling from q), such that it can be approximately reconstructed by a fixed and deterministic decoder, then $\mathcal{Q}$ can be learned.

The formal definition of sample compression schemes for distribution classes is given in Definition 3.4.1.

Both robust and realizable compression schemes satisfy certain useful properties that we use to develop public-private distribution learners in different settings. For example, the existence of a realizable compression scheme is closed under taking mixtures or products of a distribution class.

List learning. A list learner for a class of distributions $\mathcal{Q}$ takes a fixed number of samples from any $p \in \mathcal{Q}$, and outputs a finite list of distributions L such that with high probability, L contains at least one distribution q satisfying $\mathrm{TV}(p, q) \leq \alpha$.

The formal definition of list learning is given in Definition 3.4.2 and Definition 3.4.3.

List learning can be viewed a a relaxation of the normal learning setting, where we only require the algorithm output a finite list (who's length can depend on the desired error α and failure probability β), rather than a single distribution.

Sample complexity equivalence. The following result relates sample compression, public-private learning, and list learning.

Theorem 6.1.1 (Sample complexity equivalence between sample compression, public-private learning, and list learning)**.** *Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$. Let $m : (0,1]^2 \to \mathbb{N}$ be a sample complexity function such that $m(\alpha, \beta) = \mathrm{poly}(\frac{1}{\alpha}, \frac{1}{\beta})$.*[1] *Then the following are equivalent:*

1. *$\mathcal{Q}$ is realizably compressible with $m_C(\alpha, \beta) = O(m(\alpha, \beta))$ samples.*
2. *$\mathcal{Q}$ is public-privately learnable with $m_P(\alpha, \beta, \epsilon) = O(m(\alpha, \beta))$ public samples.*
3. *$\mathcal{Q}$ is list learnable with $m_L(\alpha, \beta) = O(m(\alpha, \beta))$ samples.*

The functions m_C, m_P, and m_L are related to one another as: $m_P(\alpha, \beta, \varepsilon) = m_C(\frac{\alpha}{6}, \frac{\beta}{2})$; $m_L(\alpha, \beta) = m_P(\frac{\alpha}{2}, \frac{\beta}{10}, \varepsilon_0)$ for any $\varepsilon_0 > 0$; and $m_C(\alpha, \beta) = m_L(\alpha, \beta)$.

Hence, if there exists a polynomial $m : (0,1]^2 \to \mathbb{N}$, such that $m_C(\alpha, \beta) = O(m(\alpha, \beta))$, then $m_C(\alpha, \beta), m_P(\alpha, \beta), m_L(\alpha, \beta)$ are all within constant factors of each other.

[1] The reductions between the learners do not need this assumption, it is only used to state the sample complexity equivalence.

The proof of Theorem 6.1.1 follows from the reductions in Propositions 6.1.2, 6.1.3, and 6.1.4. The propositions also state the quantitative translations between: the compression size $\tau(\alpha, \beta)$ and bit size $t(\alpha, \beta)$, the number of private samples $n(\alpha, \beta, \varepsilon)$, and the list size $\ell(\alpha, \beta)$.

Sample compression implies public-private learning. We start by establishing that the existence of a sample compression scheme for $\mathcal{Q}$ implies the existence of a public-private learner for $\mathcal{Q}$.

Proposition 6.1.2 (Sample compression $\implies$ public-private learning)**.** *Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$. Suppose $\mathcal{Q}$ admits $(\tau(\alpha, \beta), t(\alpha, \beta), m_C(\alpha, \beta))$ realizable sample compression. Then $\mathcal{Q}$ is public-privately learnable with*

$$m(\alpha, \beta, \varepsilon) = m_C\left(\frac{\alpha}{6}, \frac{\beta}{2}\right)$$

public and

$$n(\alpha, \beta, \varepsilon) = O\left(\left(\frac{1}{\alpha^2} + \frac{1}{\alpha\epsilon}\right) \cdot \left(t\left(\frac{\alpha}{6}, \frac{\beta}{2}\right) + \tau\left(\frac{\alpha}{6}, \frac{\beta}{2}\right) \cdot \log\left(m_C\left(\frac{\alpha}{6}, \frac{\beta}{2}\right)\right) + \log\left(\frac{1}{\beta}\right)\right)\right)$$

private samples.

Proof. The proof of Proposition 6.1.2 closely mirrors that of Theorem 4.5 from [ABDH+20]. We adapt their result to the public-private setting.

Fix $\alpha, \beta \in (0, 1]$ and $\varepsilon > 0$. Let $\tau = \tau(\frac{\alpha}{6}, \frac{\beta}{2})$, $t = t(\frac{\alpha}{6}, \frac{\beta}{2})$, and $m = m_C(\frac{\alpha}{6}, \frac{\beta}{2})$. We draw a public dataset $\tilde{\boldsymbol{X}}$ of size m i.i.d. from p. Consider

$$\mathcal{S} \coloneqq \left\{(\boldsymbol{S'}, b) : \boldsymbol{S'} \subseteq \tilde{\boldsymbol{X}} \text{ where } |\boldsymbol{S'}| = \tau, \text{ and } b \in \{0, 1\}^t\right\}.$$

Note that the encoding $f_p(\tilde{\boldsymbol{X}}) \in \mathcal{S}$, so forming $\widehat{\mathcal{Q}} = \{g(S', b) : (S', b) \in \mathcal{S}\}$ means that with probability $\geq 1 - \frac{\beta}{2}$ over the sampling of $\tilde{\boldsymbol{X}}$, $q = g(f_p(\tilde{\boldsymbol{X}})) \in \widehat{\mathcal{Q}}$ has $\mathrm{TV}(q, p) \leq \frac{\alpha}{6}$.

Now, we run the ε-DP 3-agnostic learner from Fact A.3.3 on $\widehat{\mathcal{Q}}$, targeting error $\frac{\alpha}{2}$ and failure probability $\frac{\beta}{2}$, which is achieved as long as we have n private samples (given in the statement of Proposition 6.1.2), which is logarithmic in $|\mathcal{S}|$. With probability $\geq 1 - \beta$, we approximately recover p with the compression scheme and the DP learner succeeds, and so the output Q satisfies

$$\begin{aligned}\mathrm{TV}(Q, p) &\leq 3 \cdot \min_{q \in \widehat{\mathcal{Q}}} \mathrm{TV}(p, q) + \frac{\alpha}{2} \\ &\leq 3 \cdot \frac{\alpha}{6} + \frac{\alpha}{2} = \alpha.\end{aligned}$$

□

Public-private learning implies list learning. The key step of the reduction of list learning to public-private learning is showing that, upon receiving samples $\tilde{\boldsymbol{x}}$, outputting a finite cover of the set of distributions that a public-private learner would succeed on, *given public data* $\tilde{\boldsymbol{x}}$, is a successful strategy for list learning.

Proposition 6.1.3 (Public-private learning $\Longrightarrow$ list learning)**.** *Suppose $\mathcal{Q}$ is public-privately learnable with $m_P(\alpha, \beta, \varepsilon)$ public and $n(\alpha, \beta, \varepsilon)$ private samples. Then for all $\varepsilon_0 > 0$, $\mathcal{Q}$ is list learnable to list size*

$$\ell(\alpha, \beta) = \frac{10}{9} \exp\left(\left(\varepsilon_0 \cdot n\left(\frac{\alpha}{2}, \frac{\beta}{10}, \varepsilon_0\right)\right)\right)$$

with $m(\alpha, \beta) = m_P(\frac{\alpha}{2}, \frac{\beta}{10}, \varepsilon_0)$ samples.

Proof. Let $\varepsilon_0 > 0$ be arbitrary. Fix any $\alpha, \beta \in (0, 1]$. By assumption, $\mathcal{Q}$ admits a $(\frac{\alpha}{2}, \frac{\beta}{10}, \varepsilon_0)$-public-private learner $\mathcal{A}$, which uses $m := m_P(\frac{\alpha}{2}, \frac{\beta}{10}, \varepsilon_0)$ public and $n := n(\frac{\alpha}{2}, \frac{\beta}{10}, \varepsilon_0)$ private samples. We use $\mathcal{A}$ to construct a $(\alpha, \beta, \frac{10}{9}\exp(\varepsilon_0 n))$-list learner that uses m samples.

Consider any $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \ldots, \tilde{x}_m) \in \mathcal{X}^m$ and the class

$$\mathcal{Q}_{\tilde{\boldsymbol{x}}} = \left\{ q \in \mathcal{Q} : \mathop{\mathbb{P}}_{\substack{\boldsymbol{X} \sim q^n \\ Q \sim \mathcal{A}(\tilde{\boldsymbol{x}}, \boldsymbol{X})}} \left\{ \mathrm{TV}(Q, q) \leq \tfrac{\alpha}{2} \right\} \geq \tfrac{9}{10} \right\}.$$

Note that by definition $\mathcal{Q}_{\tilde{\boldsymbol{x}}}$ has a $(\frac{\alpha}{2}, \frac{1}{10})$-learner under ε_0-DP that takes n samples. Hence, by Fact A.3.2 it follows that any α-packing of $\mathcal{Q}_{\tilde{\boldsymbol{x}}}$ must have size $\leq \frac{10}{9}\exp(\varepsilon_0 n) := \ell$. Let $\widehat{Q}_{\tilde{\boldsymbol{x}}}$ be such a maximal α-packing, hence it is also an α-cover of $\mathcal{Q}_{\tilde{\boldsymbol{x}}}$ with $|\widehat{Q}_{\tilde{\boldsymbol{x}}}| \leq \ell$. We define our list learner $\mathcal{L}(\tilde{\boldsymbol{x}}) = \widehat{Q}_{\tilde{\boldsymbol{x}}}$.

It remains to show that for any $p \in \mathcal{Q}$, with probability $\geq 1 - \beta$ over the sampling of $\tilde{\boldsymbol{X}} \sim p^m$, $\mathrm{dist}(p, \mathcal{L}(\tilde{\boldsymbol{X}})) \leq \alpha$. Suppose otherwise, that is, there exists $p_0 \in \mathcal{Q}$, such that,

$$\mathop{\mathbb{P}}_{\tilde{\boldsymbol{X}} \sim p_0^m} \left\{ \mathrm{dist}(p_0, \mathcal{L}(\tilde{\boldsymbol{X}})) > \alpha \right\} > \beta.$$

Since $\mathcal{L}(\tilde{\boldsymbol{X}})$ is a α-cover of $\mathcal{Q}_{\tilde{\boldsymbol{X}}}$, we have that with probability $> \beta$ over the sampling of $\tilde{\boldsymbol{X}} \sim p_0^m$, $p_0 \notin \mathcal{Q}_{\tilde{\boldsymbol{X}}}$. This contradicts the success guarantee of $\mathcal{A}$:

$$\begin{aligned} \mathop{\mathbb{P}}_{\substack{\tilde{\boldsymbol{X}} \sim p_0^m, \\ \boldsymbol{X} \sim p_0^n, \\ Q \sim \mathcal{A}(\tilde{\boldsymbol{X}}, \boldsymbol{X})}} \left\{ \mathrm{TV}(Q, p_0) > \tfrac{\alpha}{2} \right\} &\geq \mathbb{P}\left\{ \mathrm{TV}(Q, p_0) > \tfrac{\alpha}{2} | p_0 \notin \mathcal{Q}_{\tilde{\boldsymbol{X}}} \right\} \cdot \mathbb{P}\left\{ p_0 \notin \mathcal{Q}_{\tilde{X}} \right\} \\ &> \frac{1}{10} \cdot \beta = \frac{\beta}{10}. \end{aligned}$$

The second inequality follows by the definition of $Q_{\tilde{\boldsymbol{X}}}$: conditioned on the event $p_0 \notin \mathcal{Q}_{\tilde{\boldsymbol{X}}}$, the probability, over the private samples $\tilde{\boldsymbol{X}} \sim p_0^n$ and the randomness of the algorithm $\mathcal{A}$, that the output Q of our algorithm satisfies $\mathrm{TV}(Q, p_0) \le \frac{\alpha}{2}$ is $< \frac{9}{10}$. □

List learning implies sample compression. We state the final component of Theorem 6.1.1: the existence of a list learner for a class $\mathcal{Q}$ implies the existence of a sample compression scheme for $\mathcal{Q}$. This follows from the definitions. Given samples $\tilde{\boldsymbol{x}}$, the encoder passes along the entire $\tilde{\boldsymbol{x}}$, and, with knowledge of the target distribution q, the index i of the distribution in $\mathcal{L}(\tilde{\boldsymbol{x}})$ close to q. The decoder receives this information and outputs $\mathcal{L}(\tilde{\boldsymbol{x}})_i$.

Proposition 6.1.4 (List learning $\implies$ sample compression)**.** *Suppose $\mathcal{Q}$ is list learnable to list size $\ell(\alpha, \beta)$ with $m_L(\alpha, \beta)$ samples. Then $\mathcal{Q}$ admits*

$$(m_L(\alpha, \beta), \log_2(\ell(\alpha, \beta)), m_L(\alpha, \beta))$$

realizable sample compression.

Proof. Fix any $\alpha, \beta \in (0, 1]$. Let $m = m_L(\alpha, \beta)$ and $\ell = \ell(\alpha, \beta)$. By assumption, $\mathcal{Q}$ admits an (α, β, ℓ)-list learner $\mathcal{L} : \mathcal{X}^m \to \{L \subseteq \Delta(\mathcal{X}) : |L| \le \ell\}$ that takes m samples. Letting $\tau = m$ and $t = \log_2(\ell)$, we define the compression scheme as follows.

- Encoder: for any $q \in \mathcal{Q}$, the encoder $f_q : \mathcal{X}^m \to \mathcal{X}^\tau \times \{0, 1\}^t$ produces the following, given an input $\tilde{\boldsymbol{x}} \in \mathcal{X}^m$. It first runs the list learner on $\tilde{\boldsymbol{x}}$, obtaining $\mathcal{L}(\tilde{\boldsymbol{x}})$. Then, it finds the smallest index i with $\mathrm{TV}(q, \mathcal{L}(\tilde{\boldsymbol{x}})_i) = \mathrm{dist}(q, \mathcal{L}(\tilde{\boldsymbol{x}}))$, where $\mathcal{L}(\tilde{\boldsymbol{x}})_i$ denotes the i-th element of the the list $\mathcal{L}(\tilde{\boldsymbol{x}})$. The output of the list learner is $(\tilde{\boldsymbol{x}}, i)$. Note that $\tilde{\boldsymbol{x}} \in \mathcal{X}^\tau$ and that i can be represented with $\log_2(\ell) = t$ bits.
- Decoder: the fixed decoder $g : \mathcal{X}^\tau \times \{0, 1\}^t \to \Delta(\mathcal{X})$ takes $\tilde{\boldsymbol{x}}$ and i, runs the list learner $\mathcal{L}$ on $\tilde{\boldsymbol{x}}$, and produces $\mathcal{L}(\tilde{\boldsymbol{x}})_i$.

By the guarantee of the list learner, we indeed have for any $q \in \mathcal{Q}$, with probability $\ge 1 - \beta$ over the sampling of $\tilde{\boldsymbol{X}} \sim q^m$, $\mathrm{TV}(q, g(f_q(S))) \le \alpha$. □

6.2 Applications

Here, we state a few applications of the connections we determined via Theorem 6.1.1. First, we recover and extend results on the public-private learnability of high-dimensional

Gaussians and mixtures of Gaussians, using known results on sample compression schemes. Second, we describe the closure properties of public-private learnability: if a class $\mathcal{Q}$ is public-privately learnable, the class of mixtures of $\mathcal{Q}$ and the class of products of $\mathcal{Q}$ are also public-privately learnable. Finally, we use robust compression schemes to give learners for the agnostic and distribution-shifted case.

6.2.1 Gaussians and mixtures of Gaussians

There are known realizable sample compression schemes for the class of Gaussians in $\mathbb{R}^d$, as well as for the class of all k-mixtures of Gaussians in $\mathbb{R}^d$ [ABDH+20]. Hence, these classes are public-privately learnable.

Fact 6.2.1 (Robust compression scheme for Gaussians [ABDH+20, Lemma 5.3])**.** *The class of Gaussians over $\mathbb{R}^d$ admits*

$$\left(O(d), O\left(d^2 \log\left(\frac{d}{\alpha}\right)\right), O\left(d \log\left(\frac{1}{\beta}\right)\right)\right)$$

$\frac{2}{3}$-robust sample compression.

Fact 6.2.2 (Realizable compression scheme for mixtures of Gaussians [ABDH+20, Lemma 4.8 applied to Lemma 5.3])**.** *The class of k-mixtures of Gaussians over $\mathbb{R}^d$ admits*

$$\left(O(kd), O\left(kd^2 \log\left(\frac{d}{\alpha}\right) + \log_2\left(\frac{k}{\alpha}\right)\right), O\left(\frac{kd \log(\frac{k}{\beta}) \log(\frac{1}{\beta})}{\alpha}\right)\right)$$

realizable sample compression.

We get a public-private learner for Gaussians over $\mathbb{R}^d$ directly as a result of Theorem 6.1.1 and Fact 6.2.1. This recovers the upper-bound on public-private learning of high-dimensional Gaussians from Theorem 5.1.6 up to a factor of $O(\log(\frac{1}{\beta}))$ in m, and improves the private sample complexity by a polylog$(\frac{1}{\beta})$ factor.

Corollary 6.2.3 (Public-privately learning Gaussians, via compression)**.** *Let $d \geq 1$. The class of Gaussians over $\mathbb{R}^d$ is public-privately learnable with $m(\alpha, \beta, \varepsilon)$ public samples and $n(\alpha, \beta, \varepsilon)$ private samples, where*

$$m(\alpha, \beta, \varepsilon) = O\left(d \log\left(\frac{1}{\beta}\right)\right) \quad \text{and}$$

$$n(\alpha, \beta, \varepsilon) = O\left(\frac{d^2 \log(\frac{d}{\alpha}) + \log(\frac{1}{\beta})}{\alpha^2} + \frac{d^2 \log(\frac{d}{\alpha}) + \log(\frac{1}{\beta})}{\alpha\varepsilon}\right).$$

As a result of combining Theorem 6.1.1 and Fact 6.2.2, we obtain public-private learnability for the class of k-mixtures of Gaussians in $\mathbb{R}^d$.

Corollary 6.2.4 (Public-privately learning mixtures of Gaussians)**.** *Let $d, k \geq 1$. The class of all k-mixtures of Gaussians over $\mathbb{R}^d$ is public-privately learnable with $m(\alpha, \beta, \varepsilon)$ public samples and $n(\alpha, \beta, \varepsilon)$ private samples, where*

$$m(\alpha, \beta, \varepsilon) = O\left(\frac{kd \log(\frac{k}{\beta}) \log(\frac{1}{\beta})}{\alpha}\right) \quad \text{and}$$

$$n(\alpha, \beta, \varepsilon) = O\left(\left(\frac{1}{\alpha^2} + \frac{1}{\varepsilon\alpha}\right) \cdot \left(kd^2 \log\left(\frac{d}{\alpha}\right) + kd \log\left(\frac{kd \log(\frac{k}{\beta})}{\alpha}\right) + \log\left(\frac{1}{\beta}\right)\right)\right).$$

6.2.2 Closure properties of public-private learnability

If $\mathcal{Q}$ has a public-private learner, we have public-private learners for the class of k-mixtures $\mathcal{Q}^{\oplus k} \subseteq \Delta(\mathcal{X})$ and the class of k-products $\mathcal{Q}^{\otimes k} \subseteq \Delta(\mathcal{X}^k)$.

Mixture distributions. We first mention a fact from [ABDH+20], which says that if a compression scheme exists for a class of distributions $\mathcal{Q}$, then there exists a compression scheme for the class of k-mixtures of $\mathcal{Q}$.

Fact 6.2.5 (Compression for mixture distributions [ABDH+20, Lemma 4.8])**.** *If a class of distributions $\mathcal{Q}$ admits $(\tau(\alpha, \beta), t(\alpha, \beta), m(\alpha, \beta))$ realizable sample compression, then for any $k \geq 1$, the class of k-mixtures of $\mathcal{Q}$ admits $(\tau_k(\alpha, \beta), t_k(\alpha, \beta), m_k(\alpha, \beta))$ realizable sample compression, where $\tau_k, t_k, m_k : (0, 1]^2 \to \mathbb{N}$ are as follows:*

$$\tau_k(\alpha, \beta) = k\tau\left(\frac{\alpha}{3}, \beta\right),$$

$$t_k(\alpha, \beta) = kt\left(\frac{\alpha}{3}, \beta\right) + \log_2\left(\frac{3k}{\alpha}\right), \text{ and}$$

$$m_k(\alpha, \beta) = \frac{48k \log(\frac{6k}{\beta})}{\alpha} \cdot m\left(\frac{\alpha}{3}, \beta\right).$$

Next, we state a corollary of Propositions 6.1.3 and 6.1.4, which describes the existence of a compression scheme, given the existence of a public-private learner.

Corollary 6.2.6 (Public-private learning $\implies$ sample compression)**.** *Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ be a class of distributions. Suppose $\mathcal{Q}$ is public-privately learnable with $m_P(\alpha, \beta, \varepsilon)$ public samples and $n(\alpha, \beta, \varepsilon)$ private samples. Then for any $\varepsilon_0 > 0$, $\mathcal{Q}$ admits*

$$(\tau(\alpha,\beta), t(\alpha,\beta), m(\alpha,\beta)) = \left(m_P\left(\frac{\alpha}{2}, \frac{\beta}{10}, \varepsilon_0\right), \frac{\log(\frac{10}{9}) + \varepsilon_0 \cdot n(\frac{\alpha}{2}, \frac{\beta}{10}, \varepsilon_0)}{\log(2)}, m_P\left(\frac{\alpha}{2}, \frac{\beta}{10}, \varepsilon_0\right)\right)$$

realizable sample compression.

Proof. Fix $\varepsilon_0 > 0$. From Proposition 6.1.3, if $\mathcal{Q}$ is public-privately learnable, then it is list learnable to list size $\ell(\alpha, \beta)$ with $m_L(\alpha, \beta)$ samples, where

$$\ell(\alpha,\beta) = \frac{10}{9}\exp\left(\varepsilon_0 \cdot n\left(\frac{\alpha}{2}, \frac{\beta}{10}, \varepsilon_0\right)\right) \quad \text{and} \quad m_L(\alpha,\beta) = m_P\left(\frac{\alpha}{2}, \frac{\beta}{10}, \varepsilon_0\right).$$

Proposition 6.1.4 implies $\mathcal{Q}$ admits $(\tau(\alpha, \beta), t(\alpha, \beta), m_C(\alpha, \beta))$ sample compression, where

$$\begin{aligned}
\tau(\alpha,\beta) &= m_L(\alpha,\beta) = m_P\left(\frac{\alpha}{2}, \frac{\beta}{10}, \varepsilon_0\right), \\
t(\alpha,\beta) &= \log_2(\ell(\alpha,\beta)) = \frac{\log(\frac{10}{9}) + \varepsilon_0 \cdot n(\frac{\alpha}{2}, \frac{\beta}{10}, \varepsilon_0)}{\log(2)}, \text{ and} \\
m_C(\alpha,\beta) &= m_L(\alpha,\beta) = m_P\left(\frac{\alpha}{2}, \frac{\beta}{10}, \varepsilon_0\right).
\end{aligned}$$

□

Applying in sequence Corollary 6.2.6 (public-private learning $\implies$ compression), Fact 6.2.5 (compression $\implies$ compression of mixtures), and Proposition 6.1.2 (compression $\implies$ public-private learning), we have the following result about the public-private learnability of mixture distributions.

Theorem 6.2.7 (Public-private learning mixtures)**.** *Suppose $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ is public-privately learnable with $m(\alpha, \beta, \varepsilon)$ public samples and $n(\alpha, \beta, \varepsilon)$ private samples. Then for any $k \geq 1$, $\mathcal{Q}^{\oplus k}$, the class of k-mixtures of $\mathcal{Q}$, is public-privately learnable with $m_k(\alpha, \beta, \varepsilon)$ public samples and $n_k(\alpha, \beta, \varepsilon)$ private samples, where*

$$\begin{aligned}
m_k(\alpha,\beta,\varepsilon) &= O\left(\frac{k\log(\frac{k}{\beta})}{\alpha} \cdot m\left(\frac{\alpha}{36}, \frac{\beta}{20}, \varepsilon_0\right)\right) \quad \textit{and} \\
n_k(\alpha,\beta,\varepsilon) &= O\left(\left(\frac{1}{\alpha^2} + \frac{1}{\varepsilon\alpha}\right) \cdot \left(\varepsilon_0 k \cdot n\left(\frac{\alpha}{36}, \frac{\beta}{20}, \varepsilon_0\right) + \right.\right. \\
&\qquad \left.\left. k\log\left(\frac{k\log(\frac{k}{\beta})}{\alpha} \cdot m\left(\frac{\alpha}{36}, \frac{\beta}{20}, \varepsilon_0\right)\right) \cdot m\left(\frac{\alpha}{36}, \frac{\beta}{20}, \varepsilon_0\right) + \log\left(\frac{1}{\beta}\right)\right)\right)
\end{aligned}$$

for any choice of $\varepsilon_0 > 0$.

Example 6.2.8. We give an example of an application of Theorem 6.2.7. Consider the class of Gaussians over $\mathbb{R}^d$, for which there exists a public-private learner that uses $m = O(d)$ public samples and $n = O(\frac{d^2}{\alpha^2} + \frac{d^2}{\varepsilon\alpha}) \cdot \text{polylog}(d, \frac{1}{\alpha}, \frac{1}{\beta})$ private samples (Theorem 5.1.6). Then Theorem 6.2.7 implies that there exists a public-private learner for the class of k-mixtures of Gaussians that uses

$$m_k(\alpha, \beta, \varepsilon) = O\left(\frac{kd \log(\frac{k}{\beta})}{\alpha}\right)$$

public samples and

$$n_k(\alpha, \beta, \varepsilon) = O\left(\left(\frac{1}{\alpha^2} + \frac{1}{\varepsilon\alpha}\right) \cdot \left(\varepsilon_0 k \left(\frac{d^2}{\alpha^2} + \frac{d^2}{\varepsilon_0\alpha}\right) + kd\right)\right) \cdot \text{polylog}\left(d, k, \frac{1}{\alpha}, \frac{1}{\beta}\right)$$

private samples for any $\varepsilon_0 > 0$.

With the choice of $\varepsilon_0 = \alpha$, we get a private sample complexity of $n_k(\alpha, \beta, \varepsilon) = O(\frac{kd^2}{\alpha^3} + \frac{kd^2}{\alpha^2\varepsilon}) \cdot \text{polylog}(d, k, \frac{1}{\alpha}, \frac{1}{\beta})$. Notably, this private sample complexity, obtained by specializing the general result of Theorem 6.2.7, suffers an $\tilde{O}(\frac{1}{\alpha})$ factor loss compared to our learner for mixtures of Gaussians from Corollary 6.2.4.

Product distributions. We start by mentioning a fact from [ABDH+20], which says that if a compression scheme exists for a class of distributions $\mathcal{Q}$, then there exists a compression scheme for the class of k-products of $\mathcal{Q}$.

Fact 6.2.9 (Compression for product distributions [ABDH+20, Lemma 4.6])**.** *If a class of distributions* $\mathcal{Q}$ *admits* $(\tau(\alpha, \beta), t(\alpha, \beta), m(\alpha, \beta))$ *r-robust sample compression, then for any* $k \geq 1$*, the class of* k*-products of* $\mathcal{Q}$ *admits* $(\tau_k(\alpha, \beta), t_k(\alpha, \beta), m_k(\alpha, \beta))$ *r-robust sample compression, where* $\tau_k, t_k, m_k : (0, 1]^2 \to \mathbb{N}$ *are as follows:*

$$\begin{aligned}
\tau_k(\alpha, \beta) &= k\tau\left(\frac{\alpha}{k}, \beta\right), \\
t_k(\alpha, \beta) &= kt\left(\frac{\alpha}{k}, \beta\right), \text{ and} \\
m_k(\alpha, \beta) &= \log_3\left(\frac{3k}{\beta}\right) \cdot m\left(\frac{\alpha}{k}, \beta\right).
\end{aligned}$$

Applying in sequence Corollary 6.2.6, Fact 6.2.9, and Proposition 6.1.2, we have the following result about the public-private learnability of product distributions.

Theorem 6.2.10 (Public-privately learning products)**.** *Suppose $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ is public-privately learnable with $m(\alpha, \beta, \varepsilon)$ public samples and $n(\alpha, \beta, \varepsilon)$ private samples. Then for any $k \geq 1$, $\mathcal{Q}^{\otimes k}$, the class of k-products of $\mathcal{Q}$ over $\mathcal{X}^k$, is public-privately learnable with $m_k(\alpha, \beta, \varepsilon)$ public samples and $n_k(\alpha, \beta, \varepsilon)$ private samples, where*

$$m_k(\alpha, \beta, \varepsilon) = O\left(\log\left(\frac{k}{\beta}\right) \cdot m\left(\frac{\alpha}{12k}, \frac{\beta}{20}, \varepsilon_0\right)\right),$$

$$n_k(\alpha, \beta, \varepsilon) = O\left(\left(\frac{1}{\alpha^2} + \frac{1}{\varepsilon\alpha}\right) \cdot \left(\varepsilon_0 k \cdot n\left(\frac{\alpha}{12k}, \frac{\beta}{20}, \varepsilon_0\right) + k \log\left(\log\left(\frac{k}{\beta}\right) \cdot m\left(\frac{\alpha}{12k}, \frac{\beta}{20}, \varepsilon_0\right)\right) \cdot m\left(\frac{\alpha}{12k}, \frac{\beta}{20}, \varepsilon_0\right) + \log\left(\frac{1}{\beta}\right)\right)\right)$$

for any choice of $\varepsilon_0 > 0$.

Example 6.2.11. For the class of Gaussians over $\mathbb{R}$, there exists a public-private learner that requires $m = O(1)$ public samples and $n = O(\frac{1}{\alpha^2} + \frac{1}{\varepsilon\alpha}) \cdot \text{polylog}(\frac{1}{\alpha}, \frac{1}{\beta})$ private samples (Theorem 5.1.6). Then Theorem 6.2.10 implies that there exists a public-private learner for the class of k-products of Gaussians that requires

$$m_k(\alpha, \beta, \varepsilon) = O\left(\log\left(\frac{k}{\beta}\right)\right)$$

public samples and

$$n_k(\alpha, \beta, \varepsilon) = O\left(\left(\frac{1}{\alpha^2} + \frac{1}{\varepsilon\alpha}\right) \cdot \left(\varepsilon_0 k \left(\frac{1}{\alpha^2} + \frac{1}{\varepsilon_0 \alpha}\right) + k\right)\right) \cdot \text{polylog}\left(k, \frac{1}{\alpha}, \frac{1}{\beta}\right)$$

private samples, for any choice of $\varepsilon_0 > 0$. With the choice of $\varepsilon_0 = \alpha$, we get a private sample complexity of $O(\frac{k}{\alpha^3} + \frac{k}{\varepsilon\alpha^2}) \cdot \text{polylog}(k, \frac{1}{\alpha}, \frac{1}{\beta})$.

Again note that we can do better polynomially in the private sample complexity than this generic approach. Starting directly from the compression scheme for Gaussians (Fact 6.2.1) and setting $d = 1$, then applying Fact 6.2.9 to get a compression scheme for k-products of Gaussians, and finally Proposition 6.1.2, we obtain a public-private learner with a public sample complexity of $O(\log(\frac{k}{\beta}) \log(\frac{1}{\beta}))$ and a private sample complexity of $\tilde{O}(\frac{k \log(1/\beta)}{\alpha^2} + \frac{k \log(1/\beta)}{\alpha\varepsilon})$.

6.2.3 The agnostic and distribution-shifted case

Robust sample compression schemes yield public-private learners that make fewer assumptions about their input.

Namely, the learners we have considered up until now have been for the realizable case (all data is generated from a member of $\mathcal{Q}$) and, outside of Section 5.2.3, have assumed that the public and private data distributions are the same.

We relax these assumptions, and give an *agnostic and distribution shifted* public-private learner (Definition 3.4.11) for Gaussians.

Theorem 6.2.12 (Robust compression scheme $\implies$ agnostic and distribution-shifted public-private learning)**.** *Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ and $r > 0$. If $\mathcal{Q}$ admits $(\tau(\alpha,\beta), t(\alpha,\beta), m_C(\alpha,\beta))$ r-robust compression, then for every $\alpha, \beta \in (0,1]$ and $\varepsilon > 0$, there exists a $\frac{r}{2}$-shifted $\frac{2}{r}$-agnostic $(\alpha,\beta,\varepsilon)$-public-private learner for $\mathcal{Q}$ that uses*

$$m(\alpha,\beta,\varepsilon) = m_C\left(\frac{\alpha}{12}, \frac{\beta}{2}\right)$$

public samples and

$$n(\alpha,\beta,\varepsilon) = O\left(\left(\tfrac{1}{\alpha^2} + \tfrac{1}{\alpha\epsilon}\right) \cdot \left(t\left(\tfrac{\alpha}{12}, \tfrac{\beta}{2}\right) + \tau\left(\tfrac{\alpha}{12}, \tfrac{\beta}{2}\right) \cdot \log\left(m_C\left(\tfrac{\alpha}{12}, \tfrac{\beta}{2}\right)\right) + \log\left(\tfrac{1}{\beta}\right)\right)\right)$$

private samples.

Proof. The proof again mirrors the proof of Theorem 4.5 in [ABDH+20]. The key observation (and difference from the proof of Proposition 6.1.2) is the following: for the unknown distribution $p \in \Delta(\mathcal{X})$, consider $\mathrm{dist}(p, \mathcal{Q})$. If $\mathrm{dist}(p, \mathcal{Q}) \geq \frac{r}{2}$, the output Q of any algorithm satisfies $\mathrm{TV}(p, Q) \leq 1 \leq \frac{2}{r} \cdot \mathrm{dist}(p, \mathcal{Q})$. Hence, we can assume $\mathrm{dist}(p, \mathcal{Q}) < \frac{r}{2}$, and let $q_* \in \mathcal{Q}$ with $\mathrm{TV}(p, q_*) < \min\left\{\frac{r}{2}, \mathrm{dist}(p, \mathcal{Q}) + \frac{\alpha}{12}\right\}$ as guaranteed by such.

By triangle inequality, $\mathrm{TV}(\tilde{p}, q_*) < r$. This implies that when we generate hypotheses $\widehat{\mathcal{Q}}$ to choose from using the r-robust sample compression with samples from $\tilde{p}$, with high probability there will be some $q \in \widehat{\mathcal{Q}}$ with $\mathrm{TV}(q, q_*) \leq \frac{\alpha}{12}$. We have

$$\mathrm{TV}(p, q) \leq \mathrm{TV}(p, q_*) + \mathrm{TV}(q_*, q) \leq \mathrm{dist}(p, \mathcal{Q}) + \frac{\alpha}{12} + \frac{\alpha}{12} = \mathrm{dist}(p, \mathcal{Q}) + \frac{\alpha}{6}.$$

This gives us that with probability $\geq 1 - \frac{\beta}{2}$, $\mathrm{dist}(p, \widehat{\mathcal{Q}}) \leq \mathrm{dist}(p, \mathcal{Q}) + \frac{\alpha}{6}$. Applying the 3-agnostic ε-DP learner for finite classes from [AAAK21] (Fact A.3.3) on $\widehat{\mathcal{Q}}$ targeting error $\frac{\alpha}{2}$ and failure probability $\frac{\beta}{2}$, and then applying the union bound gives us the result. It can be verified that the above setting of n suffices. □

Theorem 6.2.12 gives us an agnostic and a distribution-shifted learner for Gaussians over $\mathbb{R}^d$, as stated in the following corollary.

Theorem 6.2.13 (Agnostic and distribution-shifted public-private learner for Gaussians)**.** *Let $d \geq 1$. For any $\alpha, \beta \in (0, 1]$ and $\varepsilon > 0$, there exists $\frac{1}{3}$-shifted 3-agnostic public-private learner for the class of Gaussians in $\mathbb{R}^d$ that uses m public samples and n private samples, where*

$$m = O\left(d \log\left(\frac{1}{\beta}\right)\right) \quad \text{and}$$
$$n = O\left(\frac{d^2 \log\left(\frac{d}{\alpha}\right) + \log\left(\frac{1}{\beta}\right)}{\alpha^2} + \frac{d^2 \log\left(\frac{d}{\alpha}\right) + \log\left(\frac{1}{\beta}\right)}{\alpha\varepsilon}\right).$$

Chapter 7

d public samples are necessary to privately learn d-dimensional Gaussians

In this chapter we prove lower bounds on the number of public samples required for public-privately learning Gaussians in $\mathbb{R}^d$.

We know that Gaussians in $\mathbb{R}^d$ are privately learnable with $d+1$ public samples. We show that this within 1 of the optimal: the class of Gaussians in $\mathbb{R}^d$ is not public-privately learnable with $d-1$ public samples.

Theorem 7.0.1. *Let $\mathcal{Q}$ be the class of all Gaussians in $\mathbb{R}^d$. $\mathcal{Q}$ is not public-private learnable with $m(\alpha, \beta, \varepsilon) = d-1$ public samples. That is, there exists $\alpha_d, \beta_d > 0$ such that for any $n \in \mathbb{N}$, $\mathcal{Q}$ does not admit a $(\alpha_d, \beta_d, 1)$-public-private learner using $d-1$ public and n private samples.*

Our result leverages the connection between public-private learning and list learning. The existence of such a public-private learner described above would imply the existence of a list learner for d-dimensional Gaussians taking $d-1$ samples as input. We show, using a "no-free-lunch"-style argument (in the sense of Theorem 5.1 from [SSBD14]) that such a list learner cannot exist. The proof of Theorem 7.0.1 goes through the following steps.

1. We show that a public-private learner for the problem would imply a list learner for the problem, via Proposition 6.1.3;

2. We establish a technical lemma that relates the PAC guarantee of a list learner with its average performance over a set of problem instances, via a "no-free-lunch"-style argument (Lemma 7.0.2);

3. For every $d \geq 2$, we find a sequence of hard subclasses of Gaussians over $\mathbb{R}^d$, which satisfy the conditions of Lemma 7.0.2. This forms the set of hard problem instances that imply a lower bound on the error of any list learner for the class which does not receive enough samples;

4. Since list learning to arbitrary error with few samples is impossible, public-private learning to arbitrary error with few public samples must also be impossible.

Lemma 7.0.2. *Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ and $m \in \mathbb{N}$. For a subclass $\mathcal{C} \subseteq \mathcal{Q}$, denote by $\mathcal{U}(\mathcal{C})$ the uniform distribution over $\mathcal{C}$. Suppose there exists a sequence of distribution classes $(\mathcal{Q}_k)_{k=1}^{\infty}$, with each $\mathcal{Q}_k \subseteq \mathcal{Q}$, and a set $B \subseteq \mathcal{X}^m$ such that following holds:*

1. *There exists $\eta \in (0, 1]$ and $k_\eta \in \mathbb{N}$ with*

$$\mathop{\mathbb{P}}_{\substack{Q \sim \mathcal{U}(\mathcal{Q}_k) \\ \boldsymbol{X} \sim Q^m}} \{\boldsymbol{X} \in B\} \geq \eta$$

   for all $k \geq k_\eta$.

2. *There exist $c > 0$ and $\alpha \in (0, 1]$ such that, defining $(u_k)_{k=1}^{\infty}$, $(r_k)_{k=1}^{\infty}$, and $(s_k)_{k=1}^{\infty}$ as*

$$\begin{aligned} u_k &\coloneqq \sup_{\substack{\boldsymbol{x} \in B \\ q \in \mathcal{Q}_k}} q^m(\boldsymbol{x}), \\ r_k &\coloneqq \sup_{p \in \mathcal{Q}_k} \mathop{\mathbb{P}}_{Q \sim \mathcal{U}(\mathcal{Q}_k)} \{\mathrm{TV}(p, Q) \leq 2\alpha\}, \\ s_k &\coloneqq \inf_{\boldsymbol{x} \in B} \mathop{\mathbb{P}}_{Q \sim \mathcal{U}(\mathcal{Q}_k)} \{Q^m(\boldsymbol{x}) \geq c \cdot u_k\}, \end{aligned}$$

   we have that

$$\lim_{k \to \infty} \frac{r_k}{s_k} = 0.$$

Then for any $\ell \in \mathbb{N}$, there does not exist any $(\frac{\alpha\eta}{4}, \frac{\alpha\eta}{4}, \ell)$-list learner for $\mathcal{Q}$ that uses m samples.

The above result is a technical lemma that gives a set of sufficient conditions which rule out the list learnability of a class, to a particular accuracy and failure probability requirement, given a specified number of samples.

Proof. We provide a proof by contradiction. Suppose for some $\ell \in \mathbb{N}$, we have an $(\frac{\alpha\eta}{4}, \frac{\alpha\eta}{4}, \ell)$-list learner for $\mathcal{Q}$ using m samples, denoted by $\mathcal{L} : \mathcal{X}^m \to \{L \subseteq \Delta(\mathcal{X}) : |L| \leq \ell\}$. Then for all $k \in \mathbb{N}$, we have that

$$\mathop{\mathbb{E}}_{\substack{Q\sim\mathcal{U}(\mathcal{Q}_k)\\ \boldsymbol{X}\sim Q^m}} \operatorname{dist}(Q, \mathcal{L}(\boldsymbol{X})) \leq (1 - \tfrac{\alpha\eta}{4}) \cdot \tfrac{\alpha\eta}{4} + \tfrac{\alpha\eta}{4} \cdot 1 \leq \tfrac{\alpha\eta}{2}. \tag{7.1}$$

Now, since $\lim_{k\to\infty} \frac{r_k}{s_k} = 0$, there exists $k_0 \geq k_\eta \in \mathbb{N}$, such that

$$\frac{r_{k_0} \cdot u_{k_0} \cdot \ell}{s_{k_0} \cdot cu_{k_0}} \leq \frac{1}{11}, \tag{7.2}$$

and

$$\mathop{\mathbb{P}}_{\substack{Q\sim\mathcal{U}(\mathcal{Q}_{k_0})\\ \boldsymbol{X}\sim Q^m}} \{\boldsymbol{X} \in B\} \geq \eta. \tag{7.3}$$

Fix any $\boldsymbol{x} \in B$, and let $R = \{q \in \mathcal{Q}_{k_0} : \operatorname{dist}(q, \mathcal{L}(\boldsymbol{x})) \leq \alpha\}$ and $S = \{q \in \mathcal{Q}_{k_0} : q^m(\boldsymbol{x}) \geq cu_{k_0}\}$ (note that both R and S depend on $\boldsymbol{x}$).

For $i \in [\ell]$, further let $R_i = \{q \in \mathcal{Q}_{k_0} : \mathrm{TV}(q, \mathcal{L}(\boldsymbol{x})_i) \leq \alpha\}$, so that $R = \cup_{i=1}^{\ell} R_i$.

Now, fix $i \in [\ell]$. Assuming that $R_i \neq \emptyset$, consider any $p \in R_i$. For any $q \in R_i$, we have $\mathrm{TV}(p, q) \leq 2\alpha$. Hence, $R_i \subseteq \{q \in \mathcal{Q}_{k_0} : \mathrm{TV}(p, q) \leq 2\alpha\}$. Regardless of whether R_i is empty,

$$\mathop{\mathbb{P}}_{Q\sim\mathcal{U}(\mathcal{Q}_{k_0})} \{Q \in R_i\} \leq \sup_{p\in\mathcal{Q}_{k_0}} \mathop{\mathbb{P}}_{Q\sim\mathcal{U}(\mathcal{Q}_{k_0})} \{\mathrm{TV}(p, Q) \leq 2\alpha\} = r_{k_0}.$$

Moreover, we can conclude that

$$\mathop{\mathbb{P}}_{Q\sim\mathcal{U}(\mathcal{Q}_{k_0})} \{Q \in R\} \leq \sum_{i=1}^{\ell} \mathop{\mathbb{P}}_{Q\sim\mathcal{U}(\mathcal{Q}_{k_0})} \{Q \in R_i\} \leq r_{k_0} \cdot \ell. \tag{7.4}$$

Observe that this implies, since $u_{k_0} \geq q^m(\boldsymbol{x})$,

$$\int_R q^m(\boldsymbol{x}) f_Q(q) dq \leq u_{k_0} \int_R f_Q(q) dq = u_{k_0} \mathop{\mathbb{P}}_{Q\sim\mathcal{U}(\mathcal{Q}_{k_0})} \{Q \in R\} \leq u_{k_0} \cdot r_{k_0} \cdot \ell \tag{7.5}$$

an inequality we will use momentarily. We can now write

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{Q\sim\mathcal{U}(\mathcal{Q}_{k_0})\\ \boldsymbol{X}\sim Q^m}} \operatorname{dist}(Q,\mathcal{L}(\boldsymbol{X}))\mid \boldsymbol{X}=\boldsymbol{x} &= \int_{\mathcal{Q}_{k_0}} f_{Q|\boldsymbol{X}}(q\mid \boldsymbol{x})\cdot \operatorname{dist}(q,\mathcal{L}(\boldsymbol{x}))dq \\
&\geq \int_{S\setminus R} f_{Q|\boldsymbol{X}}(q\mid \boldsymbol{x})\cdot \operatorname{dist}(q,\mathcal{L}(\boldsymbol{x}))dq \\
&\geq \alpha\int_{S\setminus R} f_{Q|\boldsymbol{X}}(q\mid \boldsymbol{x})dq \\
&\geq \alpha\left(\int_S f_{Q|\boldsymbol{X}}(q\mid \boldsymbol{x})dq - \int_R f_{Q|\boldsymbol{X}}(q\mid \boldsymbol{x})dq\right) \\
&= \alpha\left(\int_S \frac{q^m(\boldsymbol{x})f_Q(q)}{f_X(\boldsymbol{x})}dq - \int_R \frac{q^m(\boldsymbol{x})f_Q(q)}{f_X(\boldsymbol{x})}dq\right) \\
&= \alpha\frac{1}{f_X(\boldsymbol{x})}\left(\int_S q^m(\boldsymbol{x})f_Q(q)dq - \int_R q^m(\boldsymbol{x})f_Q(q)dq\right) \\
&\geq \alpha\frac{1}{f_X(\boldsymbol{x})}\left(cu_{k_0}\int_S f_Q(q)dq - u_{k_0}\cdot\ell\cdot r_{k_0}\right) \\
&\qquad\qquad \text{(By definition of } S \text{ and Equation 7.5)} \\
&= \alpha\frac{1}{f_X(\boldsymbol{x})}\left(cu_{k_0}\mathop{\mathbb{P}}_{Q\sim\mathcal{U}(\mathcal{Q}_{k_0})}\{Q\in S\} - u_{k_0}\cdot\ell\cdot r_{k_0}\right).
\end{aligned}$$

Plugging Equation 7.4 in, along with the definition of s_{k_0}, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{Q\sim\mathcal{U}(\mathcal{Q}_{k_0})\\ \boldsymbol{X}\sim Q^m}} \operatorname{dist}(Q,\mathcal{L}(\boldsymbol{X}))\mid \boldsymbol{X}=\boldsymbol{x} &\geq \alpha\frac{1}{f_X(\boldsymbol{x})}\left(cu_{k_0}\cdot s_{k_0} - u_{k_0}\cdot\ell\cdot r_{k_0}\right) \\
&\geq \alpha\frac{1}{f_X(\boldsymbol{x})}(10\cdot u_{k_0}\cdot\ell\cdot r_{k_0}) \qquad (k_0 \text{ from Equation 7.2}) \\
&\geq 10\alpha\int_R \frac{q^m(\boldsymbol{x})f_Q(q)}{f_X(\boldsymbol{x})}dq \qquad \text{(by Equation 7.5)} \\
&= 10\alpha\cdot\mathop{\mathbb{P}}_{\substack{Q\sim\mathcal{U}(\mathcal{Q}_{k_0})\\ X\sim Q^m}}\{Q\in R\mid \boldsymbol{X}=\boldsymbol{x}\}.
\end{aligned}$$

Integrating over all $\boldsymbol{x} \in B$ and using Inequality 7.3,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{Q\sim\mathcal{U}(\mathcal{Q}_{k_0})\\ \boldsymbol{X}\sim Q^m}} \operatorname{dist}(Q, \mathcal{L}(\boldsymbol{X})) &\geq \mathop{\mathbb{P}}_{\substack{Q\sim\mathcal{U}(\mathcal{Q}_{k_0})\\ \boldsymbol{X}\sim Q^m}} \{\boldsymbol{X} \in B\} \cdot \mathop{\mathbb{E}}_{\substack{Q\sim\mathcal{U}(\mathcal{Q}_{k_0})\\ \boldsymbol{X}\sim Q^m}} \operatorname{dist}(Q, \mathcal{L}(\boldsymbol{X}))|\boldsymbol{X} \in B \\
&\geq \eta \cdot \mathop{\mathbb{E}}_{\substack{Q\sim\mathcal{U}(\mathcal{Q}_{k_0})\\ \boldsymbol{X}\sim Q^m}} \operatorname{dist}(Q, \mathcal{L}(\boldsymbol{X}))|\boldsymbol{X} \in B \\
&\geq \eta \cdot 10\alpha \cdot \mathop{\mathbb{P}}_{\substack{Q\sim\mathcal{U}(\mathcal{Q}_{k_0})\\ \boldsymbol{X}\sim Q^m}} \{\operatorname{dist}(Q, \mathcal{L}(\boldsymbol{X})) \leq \alpha \mid \boldsymbol{X} \in B\}.
\end{aligned}$$

If $\mathbb{P}\{\operatorname{dist}(Q, \mathcal{L}(\boldsymbol{X})) \leq \alpha \mid \boldsymbol{X} \in B\} \geq \frac{1}{10}$, then $\mathbb{E}\operatorname{dist}(Q, \mathcal{L}(\boldsymbol{X})) \geq \alpha\eta$, contradicting Equation 7.1. Otherwise,

$$\begin{aligned}
\mathbb{E}\operatorname{dist}(Q, \mathcal{L}(\boldsymbol{X})) &\geq \eta \cdot \mathbb{E}\operatorname{dist}(Q, \mathcal{L}(\boldsymbol{X})) \mid \boldsymbol{X} \in B \\
&\geq \eta \cdot \alpha \cdot \mathbb{P}\{\operatorname{dist}(Q, \mathcal{L}(\boldsymbol{X})) > \alpha \mid \boldsymbol{X} \in B\} \\
&\geq \eta \cdot (\alpha \cdot (1 - \tfrac{1}{10}))
\end{aligned}$$

also contradicting Equation 7.1. □

Proof of Theorem 7.0.1. To prove Theorem 7.0.1, it suffices to find, for every $d \geq 2$, a sequence of subclasses $(\mathcal{Q}_k)_{k=1}^{\infty}$ and a set $B \in (\mathbb{R}^d)^{d-1}$ that indeed satisfy the conditions of Lemma 7.0.2. In what follows, we fix an arbitrary $d \geq 2$.

The construction of the sequence of hard subclasses. Let $e_d = [0, 0, \ldots, 1]^\top \in \mathbb{R}^d$. We define the following sets:

$$\begin{aligned}
T &= \left\{ \begin{bmatrix} t \\ 0 \end{bmatrix} \in \mathbb{R}^d : t \in \mathbb{R}^{d-1} \text{ with } \|t\|_2 \leq \frac{1}{2} \right\}, \\
C &= \left\{ \begin{bmatrix} t \\ \lambda \end{bmatrix} \in \mathbb{R}^d : t \in \mathbb{R}^{d-1} \text{ with } \|t\|_2 \leq \frac{1}{2} \text{ and } \lambda \in [1, 2] \subseteq \mathbb{R} \right\}.
\end{aligned}$$

That is, T is a $\frac{1}{2}$-disk (a disk with radius $\frac{1}{2}$) in $\mathbb{R}^{d-1}$ embedded onto the $(d-1)$-dimensional hyperplane in $\mathbb{R}^d$ spanning the first $(d-1)$ dimensions (axes), centered at the origin. C is a cylinder of unit length and radius $\frac{1}{2}$ placed unit distance away from T in the positive e_d-direction.

Let $S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ be the unit-sphere, centered at the origin, in $\mathbb{R}^d$, and let

$$N = \left\{ u \in S^{d-1} : |u \cdot e_d| \leq \frac{\sqrt{3}}{2} \right\}.$$

That is, N is the set of vectors u on the unit hypersphere with angle $\geq \frac{\pi}{6}$ from e_d. For $u \in N$, define the "rotation" matrix

$$R_u = \begin{bmatrix} | & | & & | \\ u & v_2 & \dots & v_d \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{d \times d}$$

where $\{v_2, \dots, v_d\}$ is any orthonormal basis for $\{u\}^\perp$ (where $\{u\}^\perp$ denotes the subspace orthogonal to the subspace spanned by the set of vectors $\{u\}$).[1]

Now, for $\sigma > 0$, $t \in T$, and $u \in N$, define the Gaussian

$$G(\sigma, t, u) = \mathcal{N}\left(t, R_u \begin{bmatrix} \sigma^2 & & & \\ & 1 & O & \\ & O & \ddots & \\ & & & 1 \end{bmatrix} R_u^\top\right) \in \Delta(\mathbb{R}^d).$$

For all $k \geq 1$, let

$$\mathcal{Q}_k = \left\{ G\left(\frac{1}{k}, t, u\right) : t \in T, u \in N \right\}.$$

That is, each $\mathcal{Q}_k$ is a class of "flat" (i.e., near $(d-1)$-dimensional) Gaussians in $\mathbb{R}^d$, with $\sigma^2 = \frac{1}{k^2}$ variance on a single thin direction u and unit variance in all other directions. Their mean vectors come from a point on the hyperplanar disk T (which we recall is a $(d-1)$-dimensional disk orthogonal to e_d), and the thin direction u comes from N (which is S^{d-1} excluding points that form angle $< \frac{\pi}{6}$ with e_d). As $k \to \infty$, the Gaussians get flatter.

Lower bounding the weight of B. We start with the following claim, which shows the probability that $d-1$ samples drawn the uniform mixture of $\mathcal{Q}_k^{d-1}$ all fall into the cylinder C can be uniformly lower bounded by an absolute constant, independent of k.

Claim 7.0.3. *Let B be the set of all possible vectors of $d-1$ points in the cylinder C, i..e, $B = C^{d-1} \in (\mathbb{R}^d)^{d-1}$. There exists $\eta > 0$ such that for $k \geq 10$,*

$$\mathop{\mathbb{P}}_{\substack{Q \sim \mathcal{U}(\mathcal{Q}_k) \\ \boldsymbol{X} \sim Q^{d-1}}} \{\boldsymbol{X} \in B\} \geq \eta.$$

[1]Technically, R_u is an equivalence class of matrices since we do not specify which orthonormal basis of $\{u\}^\perp$. However, as it turns out, the choice of the orthonormal basis of $\{u\}^\perp$ does not matter since they all result in the same Gaussian densities in the proceeding definition of $G(\sigma, t, u)$.

Proof of Claim 7.0.3. Consider the inscribed cylinder $C' \subseteq C$

$$C' = \left\{ \begin{bmatrix} t \\ \lambda \end{bmatrix} \in \mathbb{R}^d : t \in \mathbb{R}^{d-1} \text{ with } \|t\|_2 \leq \frac{1}{3} \text{ and } \lambda \in \left[\frac{4}{3}, \frac{5}{3}\right] \subseteq \mathbb{R} \right\}.$$

Also, consider $T' \subseteq T$ and $N' \subseteq N$:

$$T' = \left\{ \begin{bmatrix} t \\ 0 \end{bmatrix} \in \mathbb{R}^d : t \in \mathbb{R}^{d-1} \text{ with } \|t\|_2 \leq \frac{1}{4} \right\},$$
$$N' = \left\{ u \in S^{d-1} : |u \cdot e_d| \leq \frac{1}{36} \right\}.$$

Now, fix $u \in N'$ and $t \in T'$. Define the plane going through t with normal vector u as,

$$P(u,t) = \left\{ t + x : x \in \mathbb{R}^d \text{ with } x \cdot u = 0 \right\}.$$

First, we show $P(u,t) \cap C'$ contains a $(d-1)$-dimensional region. Consider,

$$y = \begin{bmatrix} t \\ \frac{3}{2} \end{bmatrix}.$$

The projection onto $P(u,t)$ of y is given by,

$$y' = (y - t) - ((y - t) \cdot u)u + t = \begin{bmatrix} t \\ \frac{3}{2} \end{bmatrix} - cu,$$

where $|c| = |(y-t) \cdot u| \leq \frac{3}{2} \cdot \frac{1}{36} = \frac{1}{24}$. Since $\|t\|_2 \leq \frac{1}{4}$, the norm of the first $(d-1)$ dimensions of y' is $\leq \frac{1}{4} + \frac{1}{24} \leq \frac{1}{3}$ and $y'_d \in [\frac{3}{2} - \frac{1}{24}, \frac{3}{2} + \frac{1}{24}]$, and so $y' \in C'$. Moreover, adding any z with $z \cdot u = 0$ and $\|z\|_2 \leq \frac{1}{24}$ results in $y' + z$ with the norm of the first $d-1$ dimensions being at most $\frac{1}{4} + \frac{1}{24} + \frac{1}{24} \leq \frac{1}{3}$ and $(y' + z)_d \in [\frac{3}{2} - \frac{1}{12}, \frac{3}{2} + \frac{1}{12}]$. Hence, $y' + z \in C'$. This shows that $P(u,t) \cap C'$ contains a $(d-1)$-dimensional subspace, since it contains a $(d-1)$-dimensional disk of radius $\frac{1}{24}$.

Next, let

$$M = \left\{ p + su : p \in C' \cap P(u,t), s \in \left[-\frac{1}{6}, \frac{1}{6}\right] \subseteq \mathbb{R} \right\}.$$

That is, M is a rectangular "extrusion" of $C' \cap P(u,t)$ along both its normal vectors. Indeed, we have $M \subseteq C$, since adding a vector of length $\leq \frac{1}{6}$ cannot take a point in C' outside of C. We also have that M is a d-dimensional region, so

$$\mathop{\mathbb{P}}_{X \sim G(1/10,t,u)} \{X \in C\} \geq \mathop{\mathbb{P}}_{X \sim G(1/10,t,u)} \{X \in M\} > 0.$$

Note that for $\sigma \le \frac{1}{10}$, we have

$$\mathop{\mathbb{P}}_{X\sim G(\sigma,t,u)}\{X \in M\} \ge \mathop{\mathbb{P}}_{X\sim G(1/10,t,u)}\{X \in M\}.$$

This is because any $x \in M$ can be written as $t + x + cu$, where x is such that $x \cdot u = 0$, and $|c| \le \frac{1}{6}$. Plugging in this decomposition of x into the densities of $G(1/10, u, t)$ and $G(\sigma, u, t)$, and simplifying yields the above.

To conclude, for $k \ge 10$, we have

$$\begin{aligned}
\mathop{\mathbb{P}}_{\substack{Q\sim\mathcal{U}(\mathcal{Q}_k)\\ \boldsymbol{X}\sim Q^{d-1}}}\left\{\boldsymbol{X} \in C^{d-1}\right\} &= \mathop{\mathbb{P}}_{\substack{t\sim\mathcal{U}(T)\\ u\sim\mathcal{U}(N)\\ \boldsymbol{X}\sim G(1/k,t,u)^{d-1}}}\left\{\boldsymbol{X} \in C^{d-1}\right\} \\
&= c\int_T\int_N \mathop{\mathbb{P}}_{\boldsymbol{X}\sim G(1/k,t,u)^{d-1}}\left\{\boldsymbol{X} \in C^{d-1}\right\} du\, dt \\
&\ge c\int_{T'}\int_{N'} \mathop{\mathbb{P}}_{\boldsymbol{X}\sim G(1/k,t,u)^{d-1}}\left\{\boldsymbol{X} \in C^{d-1}\right\} du\, dt \\
&= c\int_{T'}\int_{N'} \left(\mathop{\mathbb{P}}_{X\sim G(1/k,t,u)}\{X \in C\}\right)^{d-1} du\, dt \\
&\ge c\int_{T'}\int_{N'} \left(\mathop{\mathbb{P}}_{X\sim G(1/k,t,u)}\{X \in M\}\right)^{d-1} du\, dt \\
&\ge c\int_{T'}\int_{N'} \left(\mathop{\mathbb{P}}_{X\sim G(1/10,t,u)}\{X \in M\}\right)^{d-1} du\, dt \\
&=: \eta > 0, \qquad (7.6)
\end{aligned}$$

where $c = f_T(t) \cdot f_N(u) > 0$ is the uniform density over $T \times N$. Note that the final integral is non-zero since $T' \times N'$ has non-zero measure in $T \times N$ and that $\mathop{\mathbb{P}}_{X\sim G(1/10,t,u)}\{X \in M\}$ is indeed non-zero for all $t \in T', u \in N'$. □

Upper bounding r_k, the weight of α-TV balls. We prove the following.

Claim 7.0.4. *For $k \ge 1$, let*

$$r_k \coloneqq \sup_{p\in\mathcal{Q}_k} \mathop{\mathbb{P}}_{Q\sim\mathcal{U}(\mathcal{Q}_k)}\left\{\mathrm{TV}(p, Q) \le \tfrac{1}{400}\right\}.$$

Then we have,

$$r_k = O\left(\frac{1}{k^d}\right) \to 0 \quad \text{as } k \to \infty.$$

We use the following facts regarding total variation distance between 1-dimensional Gaussians, and the surface area of hyperspherical caps.

Fact 7.0.5 (TV distance between 1-dimensional Gaussians [DMR18, Theorem 1.3])**.** *Let $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ be Gaussians over $\mathbb{R}$. Then*

$$\frac{1}{200} \cdot \min\left\{1, \max\left\{\frac{|\sigma_1^2 - \sigma_2^2|}{\sigma_1^2}, \frac{40|\mu_1 - \mu_2|}{\sigma_1}\right\}\right\} \leq \mathrm{TV}\left(\mathcal{N}\left(\mu_1, \sigma_1^2\right), \mathcal{N}\left(\mu_2, \sigma_2^2\right)\right).$$

Fact 7.0.6 (Surface area of hyperspherical caps [Li10])**.** *For $u \in S^{d-1}$ and $\theta \in [0, \frac{\pi}{2}]$, define*

$$C(u, \theta) = \left\{x \in S^{d-1} : \angle(x, u) \leq \theta\right\}$$

where for $u, v \in S^{d-1}$, $\angle(u, v) \coloneqq \cos^{-1}(u \cdot v)$. We have

$$\mathrm{Area}(C(u, \theta)) = \frac{2\pi^{(d-1)/2}}{\Gamma(\frac{d-1}{2})} \cdot \int_0^\theta \sin^{d-2}(x) dx.$$

Note that

$$\mathrm{Area}(S^{d-1}) = \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})}.$$

Proof of Claim 7.0.4. Let $\sigma > 0$. Let $t_1, t_2 \in T$ and $u_1, u_2, \in N$. We will compare the total variation distance of the Gaussians defined by these parameters. Let

$$D_\sigma = \begin{bmatrix} \sigma^2 & & & \\ & 1 & O & \\ & O & \ddots & \\ & & & 1 \end{bmatrix}.$$

By Fact A.4.1, taking $f \colon \mathbb{R}^d \to \mathbb{R}$ to be $f(x) = u_1^\top (x - t_1)$,

$$\begin{aligned} \mathrm{TV}(G(\sigma, t_1, u_2), G(\sigma, t_2, u_2)) &\geq \mathrm{TV}(\mathcal{N}(u_1^\top (t_1 - t_1), u_1^\top R_{u_1} D_\sigma R_{u_1}^\top u_1), \mathcal{N}(u_1^\top (t_2 - t_1), u_1^\top R_{u_2} D_\sigma R_{u_2}^\top u_1)) \\ &= \mathrm{TV}(\mathcal{N}(0, \sigma^2), \mathcal{N}(u_1 \cdot \Delta t, \sigma^2 \cos^2(\angle(u_1, u_2)) + \sin^2(\angle(u_1, u_2)))), \end{aligned}$$

where $\Delta t = t_2 - t_1$. For the last line above, we take $R_{u_2} = [u_2, v_2, \ldots, v_d]$, where $\{v_2, \ldots, v_d\}$ is an orthonormal basis for $\{u_2\}^\perp$. Then the equality in the last line for the variance of the second Gaussian uses,

$$\begin{aligned} u_1^\top R_{u_2} D_\sigma R_{u_2}^\top u_1 &= \sigma^2 (u_1 \cdot u_2)^2 + (v_2 \cdot u_2)^2 + \cdots + (v_d \cdot u_2)^2 \\ &= \sigma^2 (u_1 \cdot u_2)^2 + (1 - (u_1 \cdot u_2)^2) \\ &= \sigma^2 \cos^2(\angle(u_1, u_2)) + (1 - \cos^2(\angle(u_1, u_2))) \\ &= \sigma^2 \cos^2(\angle(u_1, u_2)) + \sin^2(\angle(u_1, u_2)), \end{aligned}$$

where R_{u_2} being unitary implies that $(u_1 \cdot u_2)^2 + (u_1 \cdot v_2)^2 + \cdots + (u_2 \cdot v_d)^2 = 1$, yielding the second equality in the above.

We show that if $\angle(u_1, u_2) \in [\frac{\sqrt{2}\pi}{2}\sigma, \pi - \frac{\sqrt{2}\pi}{2}\sigma]$, $\mathrm{TV}(G(\sigma, t_1, u_1), G(\sigma, t_2, u_2)) \geq \frac{1}{200}$. First, we consider the case where $\angle(u_1, u_2) \in [\frac{\sqrt{2}\pi}{2}\sigma, \frac{\pi}{2}]$. Using that on $[0, \frac{\pi}{2}]$, we have $\sin(x) \geq \frac{2}{\pi}x$ and $\cos(x) \geq 0$, we get

$$\sigma^2 \cos^2(\angle(u_1, u_2)) + \sin^2(\angle(u_1, u_2)) \geq \frac{4}{\pi^2}\angle(u_1, u_2)^2 \geq 2\sigma^2. \tag{7.7}$$

Therefore,

$$\frac{\sigma_1^2 - \sigma_2^2}{\sigma_1^2} \leq \frac{\sigma^2 - 2\sigma^2}{\sigma^2} \leq -1,$$

and by Fact 7.0.5, we can conclude that $\mathrm{TV}(G(\sigma, t_1, u_1), G(\sigma, t_2, u_2)) \geq \frac{1}{200}$. Now, consider the case where $\angle(u_1, u_2) \in [\frac{\pi}{2}, \pi - \frac{\sqrt{2}\pi}{2}]$. Note that in this case, there exists $u_2' = -u_2 \in [\frac{\sqrt{2}\pi}{2}, \frac{\pi}{2}]$ with $G(\sigma, t_2, u_2) = G(\sigma, t_2, u_2')$, bringing us back to the previous case.

Next, note that since $\|u_1\|_2 = 1$ and $|u_1^{(d)}| = |u_1 \cdot e_d| \leq \frac{\sqrt{3}}{2}$ (by the definition of N), letting $r = [u_1^{(1)}, \ldots, u_1^{(d-1)}]^\top \in \mathbb{R}^{d-1}$, we have $\|r\|_2 \geq \frac{1}{2}$. Let $\hat{r} = \frac{r}{\|r\|_2}$. We have that if $[\Delta t_1, \ldots, \Delta t_{d-1}]^\top \cdot \hat{r} \geq \frac{1}{20}\sigma$, then,

$$\begin{aligned} u_1 \cdot \Delta t &= r \cdot [\Delta t_1, \ldots, \Delta t_{d-1}]^\top \\ &\geq \frac{r}{2\|r\|_2} \cdot [\Delta t_1, \ldots, \Delta t_{d-1}]^\top \\ &\geq \tfrac{1}{2}\hat{r} \cdot [\Delta t_1, \ldots, \Delta t_{d-1}]^\top \\ &\geq \frac{1}{40}\sigma. \end{aligned}$$

This implies that

$$\frac{40(\mu_1 - \mu_2)}{\sigma_1} = \frac{40(-u_1 \cdot \Delta t)}{\sigma} \leq -1,$$

and by Fact 7.0.5, we can conclude $\mathrm{TV}(G(\sigma, t_1, u_1), G(\sigma, t_2, u_2)) \geq \frac{1}{200}$. Therefore, for any

$u \in N$, $t \in T$,

$$\begin{aligned}
\mathop{\mathbb{P}}_{Q\sim\mathcal{U}(\mathcal{Q}_k)}\left\{\mathrm{TV}(G(\tfrac{1}{k},t,u),Q) \le \frac{1}{400}\right\} &= \mathop{\mathbb{P}}_{\substack{t'\sim\mathcal{U}(T)\\ u'\sim\mathcal{U}(N)}}\left\{\mathrm{TV}(G(\tfrac{1}{k},t,u),G(\tfrac{1}{k},t',u')) \le \frac{1}{400}\right\} \\
&\le \mathop{\mathbb{P}}_{\substack{t'\sim\mathcal{U}(T)\\ u'\sim\mathcal{U}(N)}}\left\{\mathrm{TV}(G(\tfrac{1}{k},t,u),G(\tfrac{1}{k},t',u')) < \frac{1}{200}\right\} \\
&\le \mathop{\mathbb{P}}_{t'\sim\mathcal{U}(T)}\left\{[\Delta t_1,\dots,\Delta t_{d-1}]^\top \cdot \hat{r} < \tfrac{1}{20k}\right\} \cdot \\
&\qquad \mathop{\mathbb{P}}_{u'\sim\mathcal{U}(N)}\left\{(\angle(u,u') \in [0,\tfrac{\sqrt{2}\pi}{2k}) \cup (\pi - \tfrac{\sqrt{2}\pi}{2k},\pi])\right\}.
\end{aligned}$$

For the first term, note that the event

$$\left\{[\Delta t_1,\dots,\Delta t_{d-1}]\cdot\hat{r} < \tfrac{1}{20k}\right\} \subseteq \left\{t' \in \left\{t + \begin{bmatrix}x\\0\end{bmatrix} + \lambda\begin{bmatrix}\hat{r}\\0\end{bmatrix} : \|x\|_2 \le 1, x\cdot\hat{r}=0, \lambda \le \frac{1}{20k}\right\}\right\},$$

which under $\mathcal{U}(T)$, for some $c_d > 0$ depending only on d, has probability $\le c_d \cdot \frac{1}{20k}$.

For the second term, note that $\angle(u,u') \in [0,\frac{\sqrt{2}\pi}{2k}) \cup (\pi - \frac{\sqrt{2}\pi}{2k},\pi]$ means $u' \in C(u,\frac{\sqrt{2}\pi}{2k}) \cup C(-u,\frac{\sqrt{2}\pi}{2k})$. By Fact 7.0.6, we know that under $\mathcal{U}(N)$, for some c_d depending only on d,

$$\begin{aligned}
\mathop{\mathbb{P}}_{u'\sim\mathcal{U}(N)}\left\{u' \in C(u,\tfrac{\sqrt{2}\pi}{2k})\right\} &= c_d \cdot \int_0^{\sqrt{2}\pi/2k} \sin^{d-2}(x)dx \\
&\le c_d \cdot \int_0^{\sqrt{2}\pi/2k} x^{d-2}dx \\
&= \frac{c_d}{d-1}\left(\frac{\sqrt{2}\pi}{2}\right)^{d-1}\frac{1}{k^{d-1}}.
\end{aligned}$$

The bound is the same for $C(-u,\frac{\sqrt{2}\pi}{2k})$. Plugging these into the above, we can conclude that

$$r_k = \sup_{p\in\mathcal{Q}_k}\mathop{\mathbb{P}}_{Q\sim\mathcal{U}(\mathcal{Q}_k)}\left\{\mathrm{TV}(p,Q) \le \frac{1}{400}\right\} \le O\left(\frac{1}{k^d}\right) \to 0 \quad \text{as } k\to\infty.$$

This proves the claim. □

Lower bounding s_k, the weight of alternative hypotheses. First, we note that

$$u_k = \sup_{\substack{\boldsymbol{x}\in B\\ q\in\mathcal{Q}_k}} q^{d-1}(\boldsymbol{x}) = \left(\frac{1}{(2\pi)^{d/2}} k \exp(-\tfrac{1}{2})\right)^{d-1},$$

which is achieved by $G(\frac{1}{k}, \mathbf{0}, e_1)$ (where $\mathbf{0} \in \mathbb{R}^d$ is the origin) and $\boldsymbol{x} = (e_d, \ldots, e_d)$. Let

$$c = \frac{\exp(-5)^{d-1}}{\exp(-\frac{1}{2})^{d-1}} = \exp\left(\frac{9(d-1)}{2}\right).$$

Claim 7.0.7. *For $k \geq 1$, letting $(u_k)_{k=1}^{\infty}$ and c be defined as above, define*

$$s_k \coloneqq \inf_{\boldsymbol{x}\in B} \mathop{\mathbb{P}}_{Q\sim\mathcal{U}(\mathcal{Q}_k)} \left\{Q^{d-1}(\boldsymbol{x}) \geq c u_k\right\}.$$

Then we have,

$$s_k = \Omega\left(\frac{1}{k^{d-1}}\right) \to 0 \quad \text{as } k \to \infty.$$

Proof of Claim 7.0.7. Let $k \geq 1$. Fix any $\boldsymbol{x} = (x_1, \ldots, x_{d-1}) \in B$. For every $t \in T$, there exists $u \in \{x_1 - t, x_2 - t, \ldots, x_{d-1} - t\}^{\perp}$. We show $\angle(u, e_d) \geq \frac{\pi}{4}$. Suppose otherwise, that is, $\angle(u, e_d) < \frac{\pi}{4} \implies |u \cdot e_d| = |u^{(d)}| > \frac{\sqrt{2}}{2}$. Then,

$$u \cdot (x_1 - t) = u^{(1)}(x_1^{(1)} - t^{(1)}) + \cdots + u^{(d-1)}(x_1^{(d-1)} - t^{(d-1)}) + u^{(d)} x_1^{(d)}.$$

By our assumption on $u^{(d)}$, and by the fact that $x_1 \in C$, we have that $|u^{(d)} x_1^{(d)}| > \frac{\sqrt{2}}{2}$. By Cauchy-Schwarz in $\mathbb{R}^{d-1}$, we have that,

$$\begin{aligned}
&|u^{(1)}(x_1^{(1)} - t^{(1)}) + \ldots + u^{(d-1)}(x_1^{(d-1)} - t^{(d-1)})| \\
&\qquad \leq \|[u^{(1)}, \ldots, u^{(d-1)}]^{\top}\|_2 \cdot \|[x_1^{(1)}, \ldots, x_1^{(d-1)}]^{\top} - [t^{(1),\ldots,t^{(d-1)}}]^{\top}\|_2 \\
&\qquad < \frac{\sqrt{2}}{2} \cdot 1.
\end{aligned}$$

The last inequality uses that $\|u\|_2 = 1$ and $(u^{(d)})^2 > \frac{1}{2}$, so the norm of the first $(d-1)$ coordinates is $< \frac{\sqrt{2}}{2}$, and also the fact that the first $(d-1)$ coordinates of x and t are in the $\frac{1}{2}$-disk. This inequality, combined with the fact that $|u^{(d)} x_1^{(d)}| > \frac{\sqrt{2}}{2}$ contradicts that that $(x_1 - t) \cdot u = 0$.

Now, for the t and the u from above, consider an arbitrary u' with $\angle(u,u') \leq \frac{1}{k}$, and the Gaussian with mean t and normal vector u', $G(\frac{1}{k}, t, u')$. We will show that any such Gaussian assigns high mass to the point x, and furthermore that there is a high density of such Gaussians. Note that for $k \geq 5$, $\frac{1}{k} \leq \frac{\pi}{3} - \frac{\pi}{4} \implies \angle(u', e_d) \geq \frac{\pi}{3} \implies u' \in N$. We compute the minimum density this Gaussian assigns to $\boldsymbol{x}$. Consider, for $i \in [d-1]$,

$$\begin{aligned}(x_i - t)^\top (R_{u'} D_{1/k} R_{u'}^\top)^{-1} (x_i - t) &= \|D_{\sqrt{k}} R_{u'}^\top (x_i - t)\|^2 \\ &= k^2 |u' \cdot (x_i - t)|^2 + |v_2 \cdot (x_i - t)|^2 + \cdots + |v_d \cdot (x_i - t)|^2 \\ &\leq 5(k^2 |u' \cdot \hat{r}|^2 + 1),\end{aligned}$$

where $\hat{r} = (x_i - t)/\|x_i - t\|$ and $\{v_2, \ldots, v_d\}$ is an orthonormal basis of $\{u'\}^\perp$. We have that,

$$\begin{aligned}|u' \cdot \hat{r}|^2 &= |(u + (u - u')) \cdot \hat{r}|^2 \\ &\leq \|u - u'\|_2^2 \cdot \|\hat{r}\|_2^2 \\ &= u \cdot u - 2u \cdot u' + u' \cdot u' \\ &= 2 - 2\cos(\angle(u', u)) \\ &\leq 2 - 2(1 - \tfrac{\angle(u',u)^2}{2}) \\ &= \angle(u', u)^2 \leq \frac{1}{k^2}.\end{aligned}$$

Hence, the density of $G(\frac{1}{k}, t, u')$ on $\boldsymbol{x}$ is lower bounded by,

$$\left(\frac{1}{(2\pi)^{d/2}} k \exp(-5)\right)^{d-1} = cu_k.$$

For every $t \in T$, we found a set of $u' \in N$ such that the density $G(\frac{1}{k}, t, u')$ assigns to $\boldsymbol{x}$ is greater than cu_k. Since for some constant $c_d > 0$ depending only on d,

$$\begin{aligned}\mathop{\mathbb{P}}_{u' \sim \mathcal{U}(N)} \left\{u' \in C(u, \tfrac{1}{k})\right\} &= c_d \int_0^{1/k} \sin^{d-2}(x) dx \\ &\geq c_d \int_0^{1/k} \left(\frac{2}{\pi} x\right)^{d-2} dx \\ &= c_d \left(\frac{2}{\pi}\right)^{d-2} \frac{1}{d-1} \cdot \frac{1}{k^{d-1}},\end{aligned}$$

and since $\boldsymbol{x} \in B$ was arbitrary, we indeed have,

$$s_k = \inf_{\boldsymbol{x} \in B} \mathop{\mathbb{P}}_{Q \sim \mathcal{U}(\mathcal{Q}_k)} \left\{Q^{d-1}(\boldsymbol{x}) \geq cu_k\right\} = \Omega\left(\frac{1}{k^{d-1}}\right).$$

This completes the proof of the claim. □

With the three claims, applying Lemma 7.0.2 allows us to conclude that the class of all Gaussians in $\mathbb{R}^d$ is not list learnable with $m(\alpha, \beta) = d - 1$ samples. This implies that the class is also not public-privately learnable with $m(\alpha, \beta, \varepsilon) = d - 1$ public samples. $\square$

Chapter 8

Open question: Characterizing the number of public samples needed for pure private learnability

In our analysis of private distribution learning with public data, a quantity of interest that emerges is the amount of public data *necessary and sufficient* to render a class of distributions pure privately learnable.

As we've seen, this is also the amount of samples that renders a class *list learnable* (Definition 3.4.3). Consider the following concrete example: for the class of unit variance Gaussians over $\mathbb{R}$, $\mathcal{Q} = \{\mathcal{N}(\mu, 1) : \mu \in \mathbb{R}\}$ (as analyzed in Chapter 2), this quantity is 1; 1 public sample suffices, and 0 is not enough.

In this section, we state several questions on this topic. The first of which asks for the resolution of the gap between our upper bound (Theorem 3.3.1) and lower bound (Theorem 3.5.1) for Gaussians.

Question 8.0.1. *Is d or $d+1$ the number of public samples necessary and sufficient for pure privately learning Gaussians in $\mathbb{R}^d$?*

Our next question is on the topic of the problem parameters the number of public samples required for privately learning a class depends on. We first state another result for context.

A VC dimension upper bound for public-private learning. The following gives a public-private learner for classes of distributions whose *Yatracos classes* have finite VC dimension.

Definition 8.0.2 (Yatracos class)**.** For $\mathcal{Q} \subseteq \Delta(\mathcal{X})$, the *Yatracos class* of $\mathcal{Q}$ is given by

$$\mathcal{H} = \{\{x \in \mathcal{X} : p(x) > q(x)\} : p, q \in \mathcal{Q} \text{ with } p \neq q\}.^{1}$$

Theorem 8.0.3. *Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$. Let $\mathcal{H}$ be the Yatracos class of $\mathcal{Q}$. Denote by $\mathrm{VC}(\mathcal{H})$ the VC dimension of $\mathcal{H}$, and $\mathrm{VC}^*(\mathcal{H})$ the dual VC dimension of $\mathcal{H}$. $\mathcal{Q}$ is public-privately learnable with m public and n private samples, where*

$$m = O\left(\frac{\mathrm{VC}(\mathcal{H}) \log\left(\frac{1}{\alpha}\right) + \log\left(\frac{1}{\beta}\right)}{\alpha}\right) \qquad \text{and} \qquad n = O\left(\frac{\mathrm{VC}(\mathcal{H})^2 \, \mathrm{VC}^*(\mathcal{H}) + \log(\frac{1}{\beta})}{\varepsilon\alpha^3}\right).$$

For the proof of this result, see Section B.4.

The above result applies to general classes of distributions, but is not tight. Applying this result to Gaussians in $\mathbb{R}^d$ (noting the VC dimension in this case is $O(d^2)$ [AM18]; [AB99, Theorem 8.14]), Theorem 8.0.3 yields a public sample complexity of $\tilde{O}(\frac{d^2}{\alpha})$, compared to the tight result of $\Theta(d)$ from earlier. In this example, the bound is loose *qualitatively* – the VC dimension analysis fails to capture the α*-independent* nature of the public sample complexity for pure private learning of Gaussians in $\mathbb{R}^d$. One can ask why the VC bound fails to capture this dependency, and more generally

Question 8.0.4. *Is there a necessary and sufficient condition for a class of distributions $\mathcal{Q}$ to be list learnable with an α-independent number of samples?*

The dependence on α in the VC bound comes from Fact B.4.2, the "public data cover" lemma from [ABM19]. In that work, which studies distribution-free binary classification, it is shown that any class learnable with $o(\frac{1}{\alpha})$ public samples is learnable with private data only, which is not the case for density estimation. Note that this establishes that for binary classification, the answer to Question 8.0.4 is affirmative: a necessary and sufficient condition is pure private learnability.

For an example of a distribution class requiring $\Omega(\frac{1}{\alpha})$ public samples for pure private learnability, consider the class of all distributions supported on 2 elements over $\mathbb{N}$. For

[1] This is for when the distributions in $\mathcal{Q}$ are discrete. For classes of continuous distributions, we substitute p and q for their respective density functions.

some $\alpha > 0$, suppose we receive samples from a member of $\{(1-\alpha)\cdot\delta_0 + \alpha\cdot\delta_1, (1-\alpha)\cdot\delta_0 + \alpha\cdot\delta_2, (1-\alpha)\cdot\delta_0 + \alpha\cdot\delta_3, ...\}$. Without $\Omega(\frac{1}{\alpha})$ public samples, we are likely to only receive the sample 0, which does not let us distinguish between an infinite list of plausible and mutually exclusive candidates.

To offer some non-rigorous but potentially helpful intuitions: examples of α-independent list learning complexity occur when $\mathcal{Q}$ is "finite-dimensional". For Gaussians, we can get an estimate within constant error with an α-independent amount of samples, then discretize the volume around the estimate according to whatever target α that is desired; α only affects the length of the list.

However, in the 2 element support example, there is no additional parameterization, and the space is infinite dimensional. The volume around a constant error estimate cannot be discretized into a finite list given a target granularity. Another thing to note is that for Gaussians, the quantity of *numbers* in the necessary and sufficient public sample is approximately equal to the number of *parameters* that define the Gaussian; this is the case for both arbitrary and identity covariance cases.

List learning seems to also be related to the topological notion of *total-boundedness*.[2]

The previous questions are specific, and centre around the core curiosity of understanding what properties of distribution classes affect their list learning complexity. We state the general question, vaguely:

Question 8.0.5. *Can we offer new, satisfying characterizations of the number of samples that renders a class $\mathcal{Q}$ list learnable?*

[2]See this Wikipedia link.

References

[AAAK21] Ishaq Aden-Ali, Hassan Ashtiani, and Gautam Kamath. On the sample complexity of privately learning unbounded high-dimensional Gaussians. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory (ALT'21)*, 2021.

[AABD+20] Sushant Agarwal, Nivasini Ananthakrishnan, Shai Ben-David, Tosca Lechner, and Ruth Urner. On learnability wih computable learners. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory (ALT'20)*, 2020.

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

[AAL23a] Mohammad Afzali, Hassan Ashtiani, and Christopher Liaw. Mixtures of gaussians are privately learnable with a polynomial number of samples. *CoRR*, abs/2309.03847, 2023.

[AAL23b] Jamil Arbas, Hassan Ashtiani, and Christopher Liaw. Polynomial time and private learning of unbounded Gaussian mixture models. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, 2023.

[AB99] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[ABDH+20] Hassan Ashtiani, Shai Ben-David, Nicholas J. A. Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of Gaussian mixtures via compression schemes. *J. ACM*, 67(6), 2020.

[ABDM18] Hassan Ashtiani, Shai Ben-David, and Abbas Mehrabian. Sample-efficient learning of mixtures. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[ABM19] Noga Alon, Raef Bassily, and Shay Moran. Limits of private learning with access to public data. In *Advances in Neural Information Processing Systems 32 (NeurIPS'19)*, 2019.

[ACG+16] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *CCS'16: 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[ADK20] Brendan Avent, Yatharth Dubey, and Aleksandra Korolova. The power of the hybrid model for mean estimation. *Proc. Priv. Enhancing Technol.*, 2020(4):48–68, 2020.

[AGM+22] Ehsan Amid, Arun Ganesh, Rajiv Mathews, Swaroop Ramaswamy, Shuang Song, Thomas Steinke, Vinith M. Suriyakumar, Om Thakkar, and Abhradeep Thakurta. Public data-assisted mirror descent for private model training. In *Proceedings of the 39th International Conference on Machine Learning (ICML'22)*, 2022.

[AKZ+17] Brendan Avent, Aleksandra Korolova, David Zeber, Torgeir Hovden, and Benjamin Livshits. BLENDER: Enabling local search with a hybrid differential privacy model. In *26th USENIX Security Symposium (USENIX Security'17)*, 2017.

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

[AM18] Hassan Ashtiani and Abbas Mehrabian. Some techniques in density estimation. *CoRR*, abs/1801.04003, 2018.

[Ass83] Patrick Assouad. Densité et dimension. *Annales de l'Institut Fourier*, 33(3):233–282, 1983.

[BBV08] Maria-Florina Balcan, Avrim Blum, and Santosh S. Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC'08)*, 2008.

[BCM+20] Raef Bassily, Albert Cheu, Shay Moran, Aleksandar Nikolov, Jonathan R. Ullman, and Steven Wu. Private query release assisted by public data. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, 2020.

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

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

[BKSW21] Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. *IEEE Transactions on Information Theory*, 67(3):1981–2000, 2021.

[BLR08] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC'08)*, 2008.

[BMN20] Raef Bassily, Shay Moran, and Anupama Nandi. Learning from mixtures of private and public populations. In *Advances in Neural Information Processing Systems 33 (NeurIPS'20)*, 2020.

[BNS16a] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. *Theory of Computing*, 12(1):1–61, 2016.

[BNS16b] Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. In *Proceedings of the 7th Conference on Innovations in Theoretical Computer Science (ITCS'16)*, 2016.

[BTT18] Raef Bassily, Abhradeep Guha Thakurta, and Om Dipakbhai Thakkar. Model-agnostic private learning. In *Advances in Neural Information Processing Systems 31 (NeurIPS'18)*, 2018.

[CDE+23] Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Matthew Jagielski, Yangsibo Huang, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, Nicolas Papernot, Ryan Rogers, Milan Shen, Shuang Song, Weijie J. Su, Andreas Terzis, Abhradeep Thakurta, Sergei Vassilvitskii,

Yu-Xiang Wang, Li Xiong, Sergey Yekhanin, Da Yu, Huanyu Zhang, and Wanrong Zhang. Challenges towards the next frontier in privacy. *CoRR*, abs/2304.06929, 2023.

[CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC'17)*, 2017.

[DHS15] Ilias Diakonikolas, Moritz Hardt, and Ludwig Schmidt. Differentially private learning of structured discrete distributions. In *Advances in Neural Information Processing Systems 28 (NIPS'15)*, 2015.

[DKS18] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical Gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC'18)*, 2018.

[DL01] Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation*. Springer Series in Statistics. Springer New York, 2001.

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

[DMR18] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians with the same mean, 2018.

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

[DS01] Kenneth R. Davidson and Stanislaw J. Szarek. Local operator theory, random matrices and Banach spaces. *Handbook of the Geometry of Banach spaces*, 1:317–366, 2001.

[Ede88] Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM J. Matrix Anal. Appl.*, 9(4):543–560, 1988.

[GKW23] Xin Gu, Gautam Kamath, and Zhiwei Steven Wu. Choosing public datasets for private machine learning via gradient subspace distance. *CoRR*, abs/2303.01256, 2023.

[Gup20] Rishabh Gupta. Kl divergence between 2 Gaussian distributions. https://mr-easy.github.io/2020-04-16-kl-divergence-between-2-gaussian-distributions, 2020. Accessed 08/31/2023.

[HW71] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

[JE13] Zhanglong Ji and Charles Elkan. Differential privacy based on importance weighting. *Mach. Learn.*, 93(1):163–183, 2013.

[Kam20] Gautam Kamath. Lecture 15 — Private ML and stats: Mean estimation. http://www.gautamkamath.com/CS860notes/lec15.pdf, 2020. Accessed 09/3/2023.

[KDRT21] Peter Kairouz, Mónica Ribero Diaz, Keith Rush, and Abhradeep Thakurta. (nearly) dimension independent private ERM with adagrad rates via publicly estimated subspaces. In *Proceedings of the 3rth Annual Conference on Learning Theory (COLT'21)*, 2021.

[KKMN09] Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. Releasing search queries and clicks privately. In *Proceedings of the 18th International World Wide Web Conference (WWW'09)*, 2009.

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

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

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

[KS17] Pravesh K. Kothari and Jacob Steinhardt. Better agnostic clustering via relaxed tensor norms. *CoRR*, abs/1711.07465, 2017.

[KS22] Refael Kohen and Or Sheffet. Transfer learning in differential privacy's hybrid-model. In *Proceedings of the 39th International Conference on Machine Learning (ICML'22)*, 2022.

[KSSU19] Gautam Kamath, Or Sheffet, Vikrant Singhal, and Jonathan R. Ullman. Differentially private algorithms for learning mixtures of separated gaussians. In *Advances in Neural Information Processing Systems 28 (NeurIPS'19)*, 2019.

[KV18] Vishesh Karwa and Salil P. Vadhan. Finite sample differentially private confidence intervals. In *9th Innovations in Theoretical Computer Science Conference (ITCS'18)*, 2018.

[Li10] Shengqiao Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics*, 4(1):66–70, 2010.

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

[LLHR23] Andrew Lowy, Zeman Li, Tianjian Huang, and Meisam Razaviyayn. Optimal differentially private learning with public data. *CoRR*, abs/2306.15056, 2023.

[LTLH22] Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *10th International Conference on Learning Representations (ICLR'22)*, 2022.

[LVS+21] Terrance Liu, Giuseppe Vietri, Thomas Steinke, Jonathan R. Ullman, and Zhiwei Steven Wu. Leveraging public data for practical private query release. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21)*, 2021.

[LW86] Nick Littlestone and Manfred Warmuth. Relating data compression and learnability, 1986.

[LWAF21] Zelun Luo, Daniel J. Wu, Ehsan Adeli, and Li Fei-Fei. Scalable differential privacy with sparse network finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, 2021.

[MY16] Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *J. ACM*, 63(3), 2016.

[NB20] Anupama Nandi and Raef Bassily. Privately answering classification queries in the agnostic PAC model. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory (ALT'20)*, 2020.

[NRS07] 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)*, 2007.

[PAE+17] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian J. Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *5th International Conference on Learning Representations (ICLR'17)*, 2017.

[Par05] Leandro Pardo. *Statistical inference based on divergence measures*. CRC Press, 2005.

[PCS+19] Nicolas Papernot, Steve Chien, Shuang Song, Abhradeep Thakurta, and Ulfar Erlingsson. Making the shoe fit: Architectures, initializations, and tuning for learning with privacy, 2019.

[PSM+18] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with PATE. In *6th International Conference on Learning Representations (ICLR'18)*, 2018.

[rJc23] Reviewer rJcv. Official comment by reviewer rjcv. https://openreview.net/forum?id=nDIrJmKPd5&noteId=3cQ2kC9y2p, 2023. Accessed 10/28/2023.

[RY20] Prasad Raghavendra and Morris Yau. List decodable learning via sum of squares. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms (SODA 2020)*, 2020.

[SOAJ14] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical Gaussian mixtures. In *Advances in Neural Information Processing Systems 27 (NIPS'14)*, 2014.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[SST06] Arvind Sankar, Daniel A. Spielman, and Shang-Hua Teng. Smoothed analysis of the condition numbers and growth factors of matrices. *SIAM J. Matrix Anal. Appl.*, 28(2):446–476, 2006.

[TB21] Florian Tramèr and Dan Boneh. Differentially private learning needs better features (or much more data). In *9th International Conference on Learning Representations (ICLR'21)*, 2021.

[TCK+22] Eliad Tsfadia, Edith Cohen, Haim Kaplan, Yishay Mansour, and Uri Stemmer. FriendlyCore: Practical differentially private aggregation. In *Proceedings of the 39th International Conference on Machine Learning (ICML'22)*, 2022.

[TKC22] Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Considerations for differentially private learning with large-scale public pretraining. *CoRR*, abs/2212.06470, 2022.

[Vad17] Salil Vadhan. *The Complexity of Differential Privacy*, pages 347–450. Springer International Publishing, 2017.

[Yat85] Yannis G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13(2):768–774, 1985.

[YNB+22] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *10th International Conference on Learning Representations (ICLR'22)*, 2022.

[YZCL21] Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. In *9th International Conference on Learning Representations (ICLR'21)*, 2021.

[ZWB21] Yingxue Zhou, Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private SGD with gradient subspace identification. In *9th International Conference on Learning Representations (ICLR'21)*, 2021.

APPENDICES

Appendix A

Additional Background

For completeness, we compile some known results on learning, privacy, and probability used throughout the thesis. We also include some notation trimmed from the main body.

A.1 Notation

For positive semi-definite $\Sigma \in \mathbb{R}^{d\times d}$, the *Mahalanobis norm* for $v \in \mathbb{R}^d$ is denoted by $\|v\|_\Sigma \coloneqq \|\Sigma^{-1/2}v\|$. The matrix version for $M \in \mathbb{R}^{d\times d}$ is denoted by $\|M\|_\Sigma \coloneqq \|\Sigma^{-1/2}M\Sigma^{-1/2}\|_F$.

For $p, q \in \Delta(\mathcal{X})$, denote by $\mathrm{KL}\left(p\|q\right)$ the *KL divergence of p with respect to q*. Denote by $\mathrm{H}^2(p,q)$ the *Hellinger distance between p and q*.

A.2 Learning

Definition A.2.1 (Covers and packings)**.** Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ and $\alpha > 0$. We say $\mathcal{C} \subseteq \Delta(\mathcal{X})$ is an *α-cover of $\mathcal{Q}$* if for every $q \in \mathcal{Q}$, there exists $p \in \mathcal{C}$ with $\mathrm{TV}(p,q) \le \alpha$.

We say $\mathcal{P} \subseteq \mathcal{Q}$ is an *α-packing of $\mathcal{Q}$* if for every $p, q \in \mathcal{P}$ with $p \neq q$, $\mathrm{TV}(p,q) > \alpha$.

Fact A.2.2. *Let $\mathcal{P}$ an α-packing of $\mathcal{Q}$. Suppose $\mathcal{P}$ is maximal, that is, there is no α-packing $\mathcal{P}'$ with $\mathcal{P}' \supsetneq \mathcal{P}$. Then $\mathcal{P}$ is a α-cover of $\mathcal{Q}$.*

A.3 Privacy

The definition of differential privacy is given in Definition 3.2.1. DP is closed under post-processing.

Fact A.3.1 (Post-processing [DR14, Proposition 2.1]). *Fix an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. Suppose the randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \Delta(\mathcal{Y})$ is (ε, δ)-DP. Consider another output space $\mathcal{O}$, and let $f : \mathcal{Y} \to \mathcal{O}$. Then the randomized algorithm $f \circ \mathcal{A} : \mathcal{X}^n \to \Delta(\mathcal{O})$ is also (ε, δ)-DP.*

We have the following facts about learnability under pure DP.

Fact A.3.2 (Packing lower bound [BKSW21, Lemma 5.1]). *Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$, $\alpha \in (0, 1]$, and $\varepsilon > 0$. Let $\widehat{\mathcal{Q}}$ be a α-packing of $\mathcal{Q}$. Any ε-DP algorithm $\mathcal{A} : \mathcal{X}^n \to \Delta(\Delta(\mathcal{X}))$ that upon receiving n samples $\boldsymbol{X} = (X_1, ..., X_n)$ drawn i.i.d. from any $p \in \mathcal{Q}$ and then outputting $Q \sim \mathcal{A}(\boldsymbol{X})$ satisfying*

$$\mathop{\mathbb{P}}_{\substack{\boldsymbol{X} \sim p^n \\ Q \sim \mathcal{A}(\boldsymbol{X})}} \left\{ \mathrm{TV}(Q, p) \leq \frac{\alpha}{2} \right\} \geq \frac{9}{10}$$

must have

$$n \geq \frac{\log(|\widehat{\mathcal{Q}}|) - \log(\frac{10}{9})}{\varepsilon}.$$

Fact A.3.3 (Pure DP learner for finite classes [BKSW21], [AAAK21, Theorem 2.24]). *Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$ with $|\mathcal{Q}| < \infty$. For every $\alpha, \beta \in (0, 1]$ and $\varepsilon > 0$, there exists an ε-DP algorithm $\mathcal{A} : \mathcal{X}^n \to \Delta(\Delta(\mathcal{X}))$, such that for any $p \in \Delta(\mathcal{X})$, if we draw a dataset $\boldsymbol{X} = (X_1, ..., X_n)$ i.i.d. from p and then $Q \sim \mathcal{A}(\boldsymbol{X})$,*

$$\mathop{\mathbb{P}}_{\substack{\boldsymbol{X} \sim p^n \\ Q \sim \mathcal{A}(\boldsymbol{X})}} \{ \mathrm{TV}(Q, p) \leq 3 \cdot \mathrm{dist}(p, \mathcal{Q}) + \alpha \} \geq 1 - \beta,$$

where

$$n = O\left(\frac{\log(|\mathcal{Q}|) + \log(\frac{1}{\beta})}{\alpha^2} + \frac{\log(|\mathcal{Q}|) + \log(\frac{1}{\beta})}{\alpha\varepsilon} \right).$$

A.4 Statistical distances

The following fact says that we cannot post-process distributions and make them more distinguishable, in terms of total variation.

Fact A.4.1 (Data-processing inequality for total variation)**.** *Let $p, q \in \Delta(\mathcal{X})$. For any measurable $f : \mathcal{X} \to \mathcal{Y}$,*

$$\mathrm{TV}(f(p), f(q)) \leq \mathrm{TV}(p, q),$$

where for $p \in \Delta(\mathcal{X})$, $f(p)$ denotes the push-forward distribution which assigns $f(p)(A) = p(f^{-1}(A))$ for all measurable $A \subseteq \mathcal{Y}$.

The following two facts relate total variation distance to KL divergence and Hellinger distance.

Fact A.4.2 (Pinsker's inequalty)**.** *Let $p, q \in \Delta(\mathcal{X})$. Then*

$$\mathrm{TV}(p, q) \leq \sqrt{\frac{1}{2} \mathrm{KL}\left(p \| q\right)}.$$

Fact A.4.3 (Hellinger distance vs. total variation)**.** *Let $p, q \in \Delta(\mathcal{X})$. Then*

$$\mathrm{H}^2(p, q) \leq \mathrm{TV}(p, q).$$

Appendix B

Deferred proofs

B.1 Proof of the privacy guarantee in Claim 2.0.6 (Pure DP unit variance Gaussian learner)

Proof. The privacy proof of Algorithm 2 follows by the Laplace mechanism [DMNS06], [Vad17, Theorem 1.3]. We give a recap of the proof. Denote by $\mathcal{M}$ our algorithm that outputs the noisy mean $\tilde{\mu}$. Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be datasets of size n varying in at most 1 entry. The density function of $\mathcal{M}(\boldsymbol{x}_1)$ is $f_1(z) = \frac{\varepsilon n}{4c} \exp\left(-\frac{\varepsilon n|z-\widehat{\mu}_1|}{2c}\right)$; $f_2(z) = \frac{\varepsilon n}{4c} \exp\left(-\frac{\varepsilon n|z-\widehat{\mu}_2|}{2c}\right)$ for $\mathcal{M}(\boldsymbol{x}_2)$. We have that for all $z \in \mathbb{R}$

$$\frac{f_1(z)}{f_2(z)} = \exp\left(\frac{\varepsilon n}{2c} \cdot (|z-\widehat{\mu}_2| - |z-\widehat{\mu}_1|)\right) \leq \exp\left(\frac{\varepsilon n}{2c} \cdot |\widehat{\mu}_2 - \widehat{\mu}_1|\right) \leq \exp(\varepsilon)$$

where the second inequality comes from the reverse triangle inequality, and the last comes from $|\widehat{\mu}_1 - \widehat{\mu}_2| \leq \frac{2c}{n}$ by the design of the algorithm. Now for any measurable $B \subseteq \mathbb{R}$

$$\mathop{\mathbb{P}}_{Z_1 \sim \mathcal{M}(\boldsymbol{x}_1)}\{Z_1 \in B\} = \int_B f_1(z)dz \leq \int_B \exp(\varepsilon) \cdot f_2(z)dz = \exp(\varepsilon) \cdot \mathop{\mathbb{P}}_{Z_2 \sim \mathcal{M}(\boldsymbol{x}_2)}\{Z_2 \in B\}$$

which is precisely Definition 3.2.1). □

B.2 Proof of the privacy guarantee in Claim 2.0.9 (Public-private unit variance Gaussian learner)

Proof. Let $\tilde{x} \in \mathbb{R}$ be arbitrary. Let $\boldsymbol{x}, \boldsymbol{x}'$ be private datasets of size n that differ in at most one entry. In Algorithm 3, we first perform element-wise subtraction by $\tilde{x}$ and then pass the results $\boldsymbol{y}$, $\boldsymbol{y}'$ to $\texttt{PrivUnitVarLearn}_{R=3,\varepsilon}(\cdot)$. Note that $\boldsymbol{y}$, $\boldsymbol{y}'$ differs in at most one entry, and therefore the outputs necessarily satisfy DP guarantees, and we can conclude that

$$\mathop{\mathbb{P}}_{\tilde{\mu}_Y \sim \mathcal{M}(\boldsymbol{x}-\tilde{x})} \{\tilde{\mu}_Y \in B\} \leq \exp(\varepsilon) \cdot \mathop{\mathbb{P}}_{\tilde{\mu}'_Y \sim \mathcal{M}(\boldsymbol{x}'-\tilde{x})} \{\tilde{\mu}'_Y \in B\} \qquad \text{for all measurable } B \subseteq \mathbb{R}.$$

The final output of $\texttt{PubPrivUnitVarLearn}_{\varepsilon}(\cdot)$ is a post processing of $\tilde{\mu}_Y$, and satisfies the same DP inequality. □

B.3 Proof of Lemma 5.2.1 (Total variation to Gaussian parameters bound)

Here we prove Lemma 5.2.1. We first establish the 1-dimensional case, which requires some intermediate results. We start with the following fact relating Gaussian parameters and squared Hellinger distance.

Fact B.3.1 (Hellinger distance between 1-dimensional Gaussians [Par05, Chapter 1, Exercises 11 and 14])**.** *Let $\mathcal{N}(\mu, \sigma^2)$, $\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$ be Gaussians over $\mathbb{R}$. Then*

$$\mathrm{H}^2(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)) = 1 - \sqrt{\frac{2\sigma\tilde{\sigma}}{\sigma^2 + \tilde{\sigma}^2}} \exp\left(-\frac{(\mu - \tilde{\mu})^2}{4(\sigma^2 + \tilde{\sigma}^2)}\right).$$

As a consequence, we obtain the following lower bound.

Lemma B.3.2 (Lower bound on Hellinger distance between 1-dimensional Gaussians)**.** *Let $\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$ be Gaussians over $\mathbb{R}$. Denote $\sigma_{\max} \coloneqq \max\{\sigma, \tilde{\sigma}\}$. Then*

1. $\mathrm{H}^2(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)) \geq \mathrm{H}^2(\mathcal{N}(0, \sigma^2), \mathcal{N}(0, \tilde{\sigma}^2))$; *and*
2. $\mathrm{H}^2(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)) \geq \mathrm{H}^2(\mathcal{N}(\mu, \sigma_{\max}^2), \mathcal{N}(\tilde{\mu}, \sigma_{\max}^2))$.

Proof. Using Fact B.3.1, we have

$$\begin{aligned}
\mathrm{H}^2(\mathcal{N}(\mu,\sigma^2),\mathcal{N}(\tilde{\mu},\tilde{\sigma}^2)) &= 1-\sqrt{\frac{2\sigma\tilde{\sigma}}{\sigma^2+\tilde{\sigma}^2}}\exp\left(-\frac{(\mu-\tilde{\mu})^2}{4(\sigma^2+\tilde{\sigma}^2)}\right)\\
&\geq 1-\sqrt{\frac{2\sigma\tilde{\sigma}}{\sigma^2+\tilde{\sigma}^2}}\\
&= \mathrm{H}^2(\mathcal{N}(0,\sigma^2),\mathcal{N}(0,\tilde{\sigma}^2)).
\end{aligned}$$

In the above, the second line follows from the fact that $\exp(-x)\leq 1$ for $x\geq 0$. This gives part (1). For part (2), we have

$$\begin{aligned}
\mathrm{H}^2(\mathcal{N}(\mu,\sigma^2),\mathcal{N}(\tilde{\mu},\tilde{\sigma}^2)) &= 1-\sqrt{\frac{2\sigma\tilde{\sigma}}{\sigma^2+\tilde{\sigma}^2}}\exp\left(-\frac{(\mu-\tilde{\mu})^2}{4(\sigma^2+\tilde{\sigma}^2)}\right)\\
&\geq 1-\exp\left(-\frac{(\mu-\tilde{\mu})^2}{4(\sigma^2+\tilde{\sigma}^2)}\right)\\
&\geq 1-\exp\left(-\frac{(\mu-\tilde{\mu})^2}{8_{\max}\sigma^2}\right)\\
&= \mathrm{H}^2(\mathcal{N}(\mu,\sigma^2_{\max}),\mathcal{N}(\tilde{\mu},\sigma^2_{\max})).
\end{aligned}$$

In the above, the second line follows from AM-GM inequality, that is, $\frac{\sigma^2+\tilde{\sigma}^2}{2}\geq\sigma\tilde{\sigma}$. □

The following Lemma B.3.3 is the 1-dimensional analogue of Lemma B.3.3.

Lemma B.3.3 (Total variation to Gaussian parameters bound, 1-dimensional case)**.** *Let $\mathcal{N}(\mu,\sigma^2),\mathcal{N}(\tilde{\mu},\tilde{\sigma}^2)$ be Gaussians over $\mathbb{R}$. Denote $\sigma_{\max}\coloneqq\max\{\sigma,\tilde{\sigma}\}$ and $\sigma_{\min}\coloneqq\min\{\sigma,\tilde{\sigma}\}$. Suppose $\mathrm{TV}(\mathcal{N}(\mu,\sigma^2),\mathcal{N}(\tilde{\mu},\tilde{\sigma}^2))\leq\gamma$. Then*

1. $\dfrac{\sigma_{\max}}{\sigma_{\min}}\leq\dfrac{2}{(1-\gamma)^2}$; *and*
2. $\dfrac{(\mu-\tilde{\mu})^2}{\sigma^2_{\max}}\leq\dfrac{8\gamma}{1-\gamma}$.

Proof. By Fact A.4.3, Lemma B.3.2, and Fact B.3.1, we have

$$\begin{aligned}
\gamma\geq\mathrm{TV}(\mathcal{N}(\mu,\sigma^2),\mathcal{N}(\tilde{\mu},\tilde{\sigma}^2)) &\geq \mathrm{H}^2(\mathcal{N}(\mu,\sigma^2),\mathcal{N}(\tilde{\mu},\tilde{\sigma}^2))\\
&\geq \mathrm{H}^2(\mathcal{N}(0,\sigma^2),\mathcal{N}(0,\tilde{\sigma}^2))\\
&= 1-\sqrt{\frac{2\sigma_{\max}\sigma_{\min}}{\sigma^2_{\max}+\sigma^2_{\min}}}.
\end{aligned}$$

Rearranging gives

$$\frac{2}{(1-\gamma)^2} \geq \frac{\sigma_{\max}^2 + \sigma_{\min}^2}{\sigma_{\max}\sigma_{\min}} \geq \frac{\sigma_{\max}}{\sigma_{\min}}$$

which is the desired inequality (1). For part (2), applying the same results, except using part (2) of Lemma B.3.2 instead, gives

$$\begin{aligned}\gamma \geq \mathrm{TV}(\mathcal{N}(\mu,\sigma^2),\tilde{\mu},\tilde{\sigma}^2)) &\geq \mathrm{H}^2(\mathcal{N}(\mu,\sigma^2),\mathcal{N}(\tilde{\mu},\tilde{\sigma}^2)) \\ &\geq \mathrm{H}^2(\mathcal{N}(\mu,\sigma_{\max}^2),\mathcal{N}(\tilde{\mu},\sigma_{\max}^2)) \\ &= 1 - \exp\left(-\frac{(\mu-\tilde{\mu})^2}{8\sigma_{\max}^2}\right)\end{aligned}$$

Rearranging gives

$$\frac{1}{1-\gamma} \geq \exp\left(\frac{(\mu-\tilde{\mu})^2}{8\sigma_{\max}^2}\right) \geq 1 + \frac{(\mu-\tilde{\mu})^2}{8\sigma_{\max}^2}$$

which is the desired inequality (2), after some further rearranging. □

With Lemma B.3.3 in place, we can now give the proof of Lemma 5.2.1.

Proof of Lemma 5.2.1. Let $g \coloneqq \mathcal{N}(\mu,\Sigma), \tilde{g} \coloneqq \mathcal{N}(\tilde{\mu},\tilde{\Sigma})$ be Gaussians over $\mathbb{R}^d$. For a unit vector $v \in \mathbb{R}^d$ and $p \in \Delta(\mathbb{R}^d)$, denote by $v^\top p$ the distribution over $\mathbb{R}$ obtained by sampling $x \sim p$ and outputting $v^\top x$. By a data-processing inequality for total variation distance (Fact A.4.1), we have that for any unit vector $v \in \mathbb{R}^d$,

$$\mathrm{TV}(\mathcal{N}(v^\top\mu, v^\top\Sigma v), \mathcal{N}(v^\top\tilde{\mu}, v^\top\tilde{\Sigma}v)) = \mathrm{TV}(v^\top g, v^\top\tilde{g}) \leq \mathrm{TV}(g,\tilde{g}) \leq \gamma.$$

The first equality comes from the fact that the projection of a Gaussian is also Gaussian, with the above parameters. By (1) in the 1-dimensional bound (Lemma B.3.3), we have that for every unit vector $v \in \mathbb{R}^d$,

$$\frac{(1-\gamma)^4}{4} \leq \frac{v^\top\tilde{\Sigma}v}{v^\top\Sigma v} \leq \frac{4}{(1-\gamma)^4}.$$

Rearranging the above gives us (1) in the statement of Lemma 5.2.1.

For (2), we have that for every unit vector $v \in \mathbb{R}^d$,

$$\frac{v^\top(\mu-\tilde{\mu})(\mu-\tilde{\mu})^\top v}{v^\top(\Sigma+\tilde{\Sigma})v} = \frac{(v^\top\mu - v^\top\tilde{\mu})^2}{v^\top\Sigma v + v^\top\tilde{\Sigma}v} \leq \frac{(v^\top\mu - v^\top\tilde{\mu})^2}{\max\{v^\top\Sigma v, v^\top\tilde{\Sigma}v\}} \leq \frac{8\gamma}{1-\gamma}$$

where the last inequality comes from applying the (2) in the 1-dimensional bound. Rearranging the above gives us (2) in the statement of Lemma 5.2.1. □

B.4 Statements and proofs regarding Theorem 8.0.3 (VC dimension upper bound for public-private learning)

Non-private Yatracos learner. When the VC dimension of the Yatracos class of $\mathcal{Q}$ is finite, the following gives an upper bound on the number of samples required to *non-privately* learn $\mathcal{Q}$.

Fact B.4.1 ([Yat85], [DL01, Theorem 6.4])**.** *Let $\mathcal{Q} \subseteq \Delta(\mathcal{X})$. Let $\mathcal{H}$ be the Yatracos class of $\mathcal{Q}$. Denote by $\mathrm{VC}(\mathcal{H})$ the VC dimension of $\mathcal{H}$. $\mathcal{Q}$ is learnable with*

$$m = O\left(\frac{\mathrm{VC}(\mathcal{H}) + \log(\frac{1}{\beta})}{\alpha^2}\right)$$

samples.

For some classes of distributions, the above bound is tight. For example, it recovers the $\Theta(\frac{d^2}{\alpha^2})$ sample complexity for learning Gaussians in $\mathbb{R}^d$ [AM18].

Proof of Theorem 8.0.3. Theorem 8.0.3 is a consequence of a known public-private uniform convergence result [BCM+20, Theorem 10]. To adapt it to our setting, we (1) modify their result for pure DP (rather than approximate DP); and (2) conclude that uniform convergence over the Yatracos sets of $\mathcal{Q}$ suffices to implement the learner from Fact B.4.1.

We employ the following result on generating distribution-dependent covers for binary hypothesis classes with public data.

Fact B.4.2 (Public data cover [ABM19, Lemma 3.3 restated])**.** *Let $\mathcal{H} \subseteq 2^{\mathcal{X}}$. There exists $\mathcal{A}\colon \mathcal{X}^* \to \{H \subseteq 2^{\mathcal{X}} : |H| < \infty\} \times (\mathcal{H} \to \mathcal{H})$, such that for any $\alpha, \beta \in (0, 1]$ there exists*

$$m = O\left(\frac{\mathrm{VC}(\mathcal{H})\log(\frac{1}{\alpha}) + \log(\frac{1}{\beta})}{\alpha}\right)$$

such that for any $p \in \Delta(\mathcal{X})$, if we draw $\boldsymbol{X} = (X_1, \ldots, X_m)$ i.i.d. from p, with probability $\geq 1 - \beta$, $\mathcal{A}(\boldsymbol{X})$ outputs $\widehat{\mathcal{H}} \subseteq 2^{\mathcal{X}}$ and a mapping $f\colon \mathcal{H} \to \widehat{\mathcal{H}}$ with

$$p(h \triangle f(h)) \leq \alpha \qquad \text{for all } h \in \mathcal{H}$$

(where for $A, B \subseteq \mathcal{X}$, $A \triangle B$ denotes the symmetric set difference $(A \setminus B) \cup (B \setminus A)$). Furthermore, we have $|\widehat{\mathcal{H}}| \leq (\frac{em}{d})^{2d}$.

Using samples from an unknown distribution p, the above Fact B.4.2 uses fewer than the $O(\frac{\mathrm{VC}(\mathcal{H})}{\alpha^2})$ samples for uniform convergence to construct a finite approximation of $\mathcal{H}$, $\widehat{\mathcal{H}}$. That is, for every $h \in \mathcal{H}$, we can find $f(h) \in \widehat{\mathcal{H}}$ which approximately agrees with h over the probability mass of p.

We also use the following pure DP algorithm for answering counting queries on finite domains.

Fact B.4.3 (SmallDB [BLR08], [DR14, Theorem 4.5])**.** *Let $\mathcal{X}$ be a finite domain. Let $\mathcal{H} \subseteq 2^{\mathcal{X}}$. Let $\alpha, \beta \in (0, 1]$ and $\varepsilon > 0$, There is an ε-DP randomized algorithm, that on any dataset $\boldsymbol{x} = (x_1, \ldots, x_n)$ with*

$$n = \Omega\left(\frac{\log(|\mathcal{X}|)\log(|\mathcal{H}|) + \log(\frac{1}{\beta})}{\varepsilon\alpha^3}\right)$$

outputs estimates $\hat{g}\colon \mathcal{H} \to \mathbb{R}$ such that with probability $\geq 1 - \beta$,

$$\left|\hat{g}(h) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_h(x_i)\right| \leq \alpha \qquad \text{for all } h \in \mathcal{H}.$$

Proof of Theorem 8.0.3. We use our m public samples from the unknown $p \in \mathcal{Q}$ to generate a public data cover $\widehat{\mathcal{H}}$ and mapping $f\colon \mathcal{H} \to \widehat{\mathcal{H}}$ courtesy of Fact B.4.2, selecting m to target error $\frac{\alpha}{6}$ and failure probability $\frac{\beta}{3}$. Note that this implies that with probability $\geq 1 - \frac{\beta}{3}$, for every $h \in \mathcal{H}$, $|p(h) - p(f(h))| \leq |p(h \triangle f(h))| \leq \frac{\alpha}{6}$.

Next, we consider the representative domain of $\mathcal{X}$ with respect to $\widehat{\mathcal{H}}$, denoted by $\mathcal{X}_{\widehat{\mathcal{H}}}$. In other words, for every unique behaviour $(\mathbb{1}_{\hat{h}}(x))_{\hat{h}\in\widehat{\mathcal{H}}} \in \{0,1\}^{|\widehat{\mathcal{H}}|}$ induced by a point $x \in \mathcal{X}$ on $\widehat{\mathcal{H}}$, we include exactly one representative $[x]$ in $\mathcal{X}_{\widehat{\mathcal{H}}}$. By Sauer's lemma we can conclude that

$$|\mathcal{X}_{\widehat{\mathcal{H}}}| \leq \left(\frac{e|\widehat{\mathcal{H}}|}{d^*}\right)^{\mathrm{VC}^*(\mathcal{H})}.$$

Then, we take our n private samples $\boldsymbol{X} = (X_1, \ldots, X_n)$ and map each point X_i to its representative $[X_i] \in \mathcal{X}_{\widehat{\mathcal{H}}}$, yielding a dataset of n examples $[\boldsymbol{X}]$ on the finite domain $\mathcal{X}_{\widehat{\mathcal{H}}}$. Note that for any $\hat{h} \in \widehat{\mathcal{H}}$, $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\hat{h}}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\hat{h}}([X_i])$. Hence when we run SmallDB (Fact B.4.3) on the input $[\boldsymbol{X}]$ over the finite domain $\mathcal{X}_{\widehat{\mathcal{H}}}$ with finite class $\widehat{\mathcal{H}}$, choosing n large

enough, we obtain $\hat{g} : \widehat{\mathcal{H}} \to \mathbb{R}$ such that with probability $\geq 1 - \frac{\beta}{3}$, $|\hat{g}(\hat{h}) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\hat{h}}(X_i)| \leq \frac{\alpha}{6}$ for all $\hat{h} \in \widehat{\mathcal{H}}$.

We also ensure n is large enough so that we get the uniform convergence property on $\widehat{\mathcal{H}}$, which has VC dimension d, with the private samples. That is, for all $\hat{h} \in \widehat{\mathcal{H}}$, with probability $\geq 1 - \frac{\beta}{3}$, $|p(\hat{h}) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\hat{h}}(X_i)| \leq \frac{\alpha}{6}$.

As a post-processing of $\hat{g}$, our learner outputs

$$\hat{q} := \arg\min_{q \in \mathcal{Q}} \sup_{h \in \mathcal{H}} |q(h) - \hat{g}(f(h))|.$$

By the union bound, with probability $\geq 1 - \beta$, all of our good events occur. In this case, we have for all $h \in \mathcal{H}$,

$$\begin{aligned} |p(h) - p(f(h))| &\leq \tfrac{\alpha}{6} \\ \left| p(f(h)) - \tfrac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{f(h)}(X_i) \right| &\leq \tfrac{\alpha}{6} \\ \left| \tfrac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{f(h)}(X_i) - \hat{g}(f(h)) \right| &\leq \tfrac{\alpha}{6} \end{aligned}$$

which implies $|p(h) - \hat{g}(f(h))| \leq \frac{\alpha}{2}$. So for any $q \in \mathcal{Q}$,

$$\begin{aligned} |q(h) - p(h)| - \frac{\alpha}{2} \leq |q(h) - \hat{g}(f(h))| &\leq |q(h) - p(h)| + \frac{\alpha}{2} \\ \implies \mathrm{TV}(q,p) - \frac{\alpha}{2} \leq \sup_{h \in \mathcal{H}} |q(h) - \hat{g}(f(h))| &\leq \mathrm{TV}(q,p) + \frac{\alpha}{2}. \end{aligned}$$

We have that

$$\sup_{h \in \mathcal{H}} |\hat{q}(h) - \hat{g}(f(h)| \leq \sup_{h \in \mathcal{H}} |p(h) - \hat{g}(f(h)| \leq \mathrm{TV}(p,p) + \frac{\alpha}{2} \leq \frac{\alpha}{2}.$$

Therefore,

$$\mathrm{TV}(\hat{q}, p) \leq \sup_{h \in \mathcal{H}} |\hat{q}(h) - \hat{g}(f(h)| + \frac{\alpha}{2} \leq \alpha.$$

It can be verified that the choices of m and n in the statement of Theorem 8.0.3 suffice. □