PRIVATELY LEARNING MIXTURE DISTRIBUTIONS

PRIVATE DENSITY ESTIMATION FOR MIXTURE DISTRIBUTIONS AND GAUSSIAN MIXTURE MODELS

BY

MOHAMMAD AFZALI KHARKOUEI, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTING AND SOFTWARE

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2024) McMaster University

(Computing and Software) Hamilton, Ontario, Canada

TITLE:	Private Density Estimation for Mixture Distributions and Gaussian Mixture Models
AUTHOR:	Mohammad Afzali Kharkouei B.Sc. (Computer Science), Sharif University of Technology, Tehran, Iran
SUPERVISOR:	Dr. Hassan Ashtiani
NUMBER OF PAGES:	xiii, 67

Declaration

I, Mohammad Afzali Kharkouei, declare that this thesis titled "PRIVATE DENSITY ESTIMATION FOR MIXTURE DISTRIBUTIONS AND GAUSSIAN MIXTURE MODELS" is my work and is based on collaboration with Christopher Liaw and Hassan Ashtiani, as presented in our paper (Afzali *et al.*, 2023) published in the 35th International Conference on Algorithmic Learning Theory.

Lay Abstract

The problem of distribution learning, also known as density estimation, has been extensively explored in Statistics over several decades. It involves the task of recovering the original distribution with minimal error given a set of samples from a distribution that belongs to a known family of distributions.

More recently, a branch of research has emerged focusing on private distribution learning. This approach aims to learn a class of distributions while safeguarding the privacy of individuals in the dataset through the fulfillment of the gold standard of differential privacy. A fundamental open question in this domain is: Is there a class of distributions that can be learned without privacy considerations but not with privacy preservation?

To address this question, we delve into the private learnability of the class of mixtures of Gaussians, which represents a diverse and complex set of distributions.

Abstract

We develop a general technique for estimating (mixture) distributions under the constraint of differential privacy (DP). On a high level, we show that if a class of distributions (such as Gaussians) is *(1)* list decodable and *(2)* admits a "locally small" cover (Bun *et al.*, 2021) with respect to total variation distance, then the class of *its mixtures* is privately learnable. The proof circumvents a known barrier indicating that, unlike Gaussians, GMMs do not admit a locally small cover (Aden-Ali *et al.*, 2021b).

As the main application, we study the problem of privately estimating mixtures of Gaussians. Our main result is that $\mathrm{poly}(k, d, 1/\alpha, 1/\varepsilon, \log(1/\delta))$ samples are sufficient to estimate a mixture of k Gaussians in $\mathbb{R}^d$ up to total variation distance α while satisfying (ε, δ)-DP. This is the first finite sample complexity upper bound for the problem that does not make any structural assumptions on the GMMs.

To my beloved family

Acknowledgements

I am profoundly thankful for the exceptional guidance and inspiration provided by my supervisor, Dr. Hassan Ashtiani. Dr. Ashtiani's unwavering support, kindness, and mentorship have played a pivotal role in my academic journey. Despite my initial lack of background in the field, he patiently imparted knowledge and offered invaluable assistance whenever I encountered obstacles. His brilliance as a mentor has significantly contributed to my growth as a researcher.

I extend my sincere appreciation to my co-author, Christopher Liaw, whose invaluable contributions and collaborative spirit were instrumental in overcoming challenges and refining our project.

I would also like to express my gratitude to Dr. Shahab Asoodeh for his continuous inspiration and guidance. His insightful teachings during his course and our regular meetings on Differential Privacy have been both enlightening and enriching.

Special thanks are due to my supervisory committee members, Dr. Shahab Asoodeh and Dr. Sivan Sabato, for their invaluable insights and unwavering support throughout my academic journey.

Additionally, I am grateful for the friendships and companionship shared with Alireza, Hrad, Narges, Vincent, Sohrab, Mohammad, Daei, and Behnoosh, as well as the insightful conversations with my lab mates Qing, Jamil, and Ghazal.

Finally, my deepest appreciation goes to my family for their unwavering support and belief in me. Their endless encouragement has been a constant source of strength and motivation throughout my career.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

1.1.1 Density estimation

The problem of density estimation is a foundational issue in statistics that has been extensively investigated over many decades. Density estimation, also known as distribution learning, involves the task of inferring an unknown distribution based on available samples from it. Specifically, when provided with independent samples from a distribution f, the objective is to derive a distribution $\hat{f}$ that closely approximates f in terms of total variation distance. Given that f is a member of, or close to, a class of distributions $\mathcal{F}$, a crucial question arises: what is the minimum number of samples required to ensure that $\hat{f}$ is close to f in total variation distance with high probability?

Extensive research has been dedicated to characterizing the optimal sample complexity, or the related minimax error rate, of learning different classes of distributions

(for an overview, see Diakonikolas (2016); Devroye and Lugosi (2001); Ashtiani and Mehrabian (2018)). However, determining the sample complexity, or even the learnability, of a general class of distributions still poses a significant open problem (e.g., Open Problem 15.1 in Diakonikolas (2016)). Recently, Lechner *et al.* (2023) has shown that there is no notion of dimension that characterizes the sample complexity of learning distribution classes.

1.1.2 Private learnability

When working with samples and data derived from a distribution, a crucial concern is safeguarding the sensitive information of individuals within the dataset. Here, we consider the problem of density estimation under the constraint of differential privacy (Dwork *et al.*, 2006a).

At a high level, differential privacy (DP) ensures that modifying the data of any single individual in the dataset does not significantly affect the outcome of the estimation, thereby preventing an adversary observing the output from uncovering any sensitive information from the original dataset.

Several formulations of DP exist. The original formulation, pure DP (ϵ-DP), can be somewhat restrictive. For instance, learning certain simple classes of distributions, like univariate Gaussians with unbounded mean, is impossible in this model due to information-theoretic limitations. An alternative formulation is approximate DP (Dwork *et al.*, 2006b), which is also known as (ε, δ)-DP. Interestingly, we are not aware of any class of distributions that is learnable in the non-private (agnostic[1])

[1] In the agnostic setting, we do not assume the true distribution belongs to the class that we are considering. The goal is then finding a distribution in the class that is (approximately) the closest to the true distribution.

setting but not learnable in the (ε, δ)-DP setting. This is in sharp contrast with other areas of learning such as classification[2]. In fact, we conjecture that every class of distributions that is learnable in the non-private agnostic setting is also learnable in the (ε, δ)-DP setting. We believe the conjecture to be true for the agnostic setting, given the recent developments on connections between robustness and privacy Asi *et al.* (2023); Hopkins *et al.* (2023). Recently, the negative result of Bun *et al.* (2024) has shown that there is a class of distribution that is learnable (in realizable setting[3]) with a constant accuracy but not privately learnable.

Nevertheless, we are still far from resolving the above conjecture. Yet, we know that some important classes of distributions are learnable in (ε, δ)-DP setting. For example, finite hypothesis classes are learnable even in the pure DP setting (Bun *et al.*, 2021; Aden-Ali *et al.*, 2021a). Earlier work of Karwa and Vadhan (2018) shows that unbounded univariate Gaussians are also learnable in this setting. More generally, high-dimensional Gaussians with unbounded parameters are private learnable too (Aden-Ali *et al.*, 2021a), even with a polynomial time algorithm (Kamath *et al.*, 2022b; Kothari *et al.*, 2022; Ashtiani and Liaw, 2022; Alabi *et al.*, 2023; Hopkins *et al.*, 2023). A next natural step is studying a richer class of distributions such as Gaussian Mixture Models (GMMs).

[2]For example, initial segments over the real line are learnable in the non-private setting but not in the (ε, δ)-DP model; see Bun *et al.* (2015); Alon *et al.* (2019); Kaplan *et al.* (2020); Bun *et al.* (2020); Cohen *et al.* (2023).

[3]In the realizable setting, unlike agnostic setting, we assume the ture distribution belongs to the class that we are considering.

1.1.3 Learning GMMs

In the non-private setting, GMMs with k components in d dimensions are known to be learnable (with error/TV-distance at most α) with a polynomial number of samples (in terms of $d, k, 1/\alpha$). Perhaps surprisingly, nearly tight sample complexity bound of $\tilde{O}(d^2k/\alpha^2)$ was proved relatively recently using distribution sample compression schemes (Ashtiani *et al.*, 2018a, 2020). But are GMMs learnable in the (ε, δ)-DP setting with a polynomial number of samples?

Several relaxed versions of the aforementioned problem have been investigated in recent years. One idea to resolve the learnability of GMMs is to extend the result of Aden-Ali *et al.* (2021a) for high-dimensional Gaussians. In particular, they show that Gaussians admit a "locally small cover" with respect to the total variation (TV) distance and therefore the class of Gaussians can be learned privately using the private hypothesis selection approach of Bun *et al.* (2021). However, as Aden-Ali *et al.* (2021b) demonstrated, GMMs do not admit such a locally small cover with respect to TV distance. At a high level, this is because there are too many parameter representations for a single GMM, all of which are close with respect to TV distance.

In addition to this negative result, Aden-Ali *et al.* (2021b) have shown that univariate GMMs are learnable in the (ϵ, δ)-differential privacy setting with a polynomial number of samples. Namely, they use stability-based histograms (Bun *et al.*, 2016) in the spirit of Karwa and Vadhan (2018) to come up with a set of candidate parameters for the mixture components, and then choose between these candidates using private hypothesis selection (Bun *et al.*, 2021; Aden-Ali *et al.*, 2021a). While they generalize this idea to learning axis-aligned[4] GMMs, their approach does not work for GMMs

[4]Where the covariance matrices of the mixture components are diagonal matrices.

with general covariance matrices. In fact, it is not clear if it is possible to extend the histogram-based approach to handle arbitrary covariance matrices *even for learning a single* (high-dimensional) Gaussian.

An alternative approach for private learning of GMMs would be using a sample-and-aggregate framework such as those proposed by Ashtiani and Liaw (2022); Tsfadia *et al.* (2022). In particular, Ashtiani and Liaw (2022) show how one can privately learn Gaussians by aggregating the outcomes of multiple non-private Gaussian estimators and then outputting a noisy version of those parameters. In fact, this is the basis of the work by Arbas *et al.* (2023) who showed how to reduce the problem of private *parameter estimation* for GMMs into its non-private counterpart. However, while this reduction is (computationally and statistically) efficient, the non-private version of the problem itself requires an (unavoidable) exponential number of samples with respect to the number of components (Moitra and Valiant, 2010). Can we avoid the (above mentioned) exponential dependence on k if we opt for (private) density estimation rather than (private) parameter estimation? We know this is possible in the non-private setting (Ashtiani *et al.*, 2018a,b, 2020) or when we have access to some "public data" (Ben-David *et al.*, 2023). One idea is to use a sample-and-aggregate approach based on a non-private *density estimator* for GMMs. This turns out to be problematic as GMMs are not uniquely parameterized: two GMMs may be close to each other in terms of total variation distance but with a completely different set of parameters. Thus, it is challenging to use non-private algorithms for learning GMMs as a blackbox since one cannot guarantee that the outputs of these algorithms are "stable".

1.2 Informal problem statement

In this section, we present an informal version of the problem statement following the provided motivation. Our work addresses the following two questions:

- Is the class of d-dimensional Gaussian Mixture Models (GMMs) is (ε, δ)-privately learnable with a polynomial number of samples in terms of d, k (the number of components in the mixture), $1/\alpha$ (where α is the accuracy parameter), $1/\epsilon$, and $\log(1/\delta)$?

- Is there a reduction for privately learning a class of (mixture) distributions, given a non-private learner for that class?

Next, we provide the list of our contributions regarding the above questions.

1.3 List of contributions

In this section we provide the list of our contributions. Indeed, we bypass the barriers mentioned in earlier sections, and show that:

- The class of d-dimensional Gaussian mixtures is (ε, δ)-privately learnable with $\tilde{O}(\frac{k^2 d^4 \log(1/\delta)}{\alpha^4 \varepsilon})$ samples, where k is the number of components in the mixture and α is the accuracy parameter (See Chapter 7). This is the first result for this problem that does not make any structural assumptions on the GMMs.

- Generally, we show that if a class (such as Gaussians) admits a "locally small cover" and is "list decodable", then the class of *its mixtures* is privately learnable (See Chapter 6).

- In our main reduction, we define the problem of private common member selection and propose an algorithm to solve this problem for general (locally small) metric spaces. We believe this problem would have other applications in private density estimation. At a high level, given T lists of objects (e.g., distributions), we say an object is a common member if it is close to a member in each of the lists. The goal of a private common member selector (PCMS) is then to privately find a common member assuming at least one exists.(see Chapter 4).

- We show that the class of Gaussians admit a locally small cover (see Chapter 7). Previously, locally small covers were constructed only for location Gaussians (Gaussians with identity covariance matrix) and scale Gaussians (zero mean Gaussians).

- Given a list decoding algorithm for a class of distributions, we construct a list decoding algorithm for the class of its mixtures with respect to parameter distance (see Chapter 5).

- Given a locally small cover for a class of distributions, we construct a locally small cover for the class of its mixtures with respect to parameter distance (see Chapter 5).

1.4 Thesis organization

We define some notation in Chapter 2 before presenting our main results in Section 2.2. Given the subtleties in the proofs, we offer an overview of our technical contributions in Section 2.3 before delving into the core technical chapters. We also

provide some background information on distribution learning, differential privacy, and related work in Chapter 3.

The formal proof of our main reduction is presented in Chapter 6. As an application of our general framework, we present the first sample complexity upper bound for privately learning GMMs in Chapter 7.

Chapter 2

Problem Formulation and Main Results

In this chapter, we first introduce some notation and definitions to formally state the problem and present our main results. Later, we offer an overview of our technical contributions before delving into the core technical chapters.

2.1 Problem formulation

2.1.1 Notation

For a set $\mathcal{F}$, define $\mathcal{F}^k = \mathcal{F} \times \cdots \times \mathcal{F}$ (k times), and $\mathcal{F}^* = \bigcup_{k=1}^{\infty} \mathcal{F}^k$. We use $[k]$ to denote the set $\{1, 2, \ldots, k\}$. We use S^d to denote the positive-definite cone in $\mathbb{R}^{d\times d}$. Moreover, for two absolutely continuous densities $f_1(x), f_2(x)$ on $\mathbb{R}^d$, the total variation (TV) distance is defined as $\mathrm{d_{TV}}(f_1, f_2) = \frac{1}{2}\int_{\mathbb{R}^d} |f_1(x) - f_2(x)| \,\mathrm{d}x$. For a matrix A, let $\|A\|_F = \sqrt{\mathrm{Tr}(A^T A)}$ be the Frobenius norm and $\|A\|_2$ be the induced

ℓ_2 (spectral) norm. In this paper, if κ is a metric on $\mathcal{F}$, $f \in \mathcal{F}$, and $Y \subseteq \mathcal{F}$ then we define $\kappa(x, Y) = \inf_{y \in Y} \kappa(f, y)$.

Definition 2.1.1 (κ-ball)**.** *Consider a metric space $(\mathcal{F}, \kappa)$. For a fixed $f \in \mathcal{F}$, define $B_\kappa(r, f, \mathcal{F}) \coloneqq \{f' \in \mathcal{F} : \kappa(f', f) \leq r\}$ to be the κ-ball of radius r around f.*

Definition 2.1.2 (α-cover)**.** *A set $C_\alpha \subseteq \mathcal{F}$ is said to be an α-cover for a metric space $(\mathcal{F}, \kappa)$, if for every $f \in \mathcal{F}$, there exists an $f' \in C_\alpha$ such that $\kappa(f, f') \leq \alpha$.*

Definition 2.1.3 (Locally small cover (Bun *et al.*, 2021))**.** *Consider an α-cover C_α for a metric space $(\mathcal{F}, \kappa)$. For $\gamma \geq \alpha$, C_α is said to be (t, γ)-locally small if:*

$$sup_{f \in \mathcal{F}} |B_\kappa(\gamma, f, C_\alpha)| \leq t$$

Moreover, if such a α-cover exists, we say $\mathcal{F}$ admits a (t, γ)-locally small α-cover.

Definition 2.1.4 (k-mixtures)**.** *Let $\mathcal{F}$ be an arbitrary class of distributions. We denote the class of k-mixtures of $\mathcal{F}$ by $k\text{-}mix(\mathcal{F}) = \Delta_k \times \mathcal{F}^k$ where $\Delta_k = \{w \in \mathbb{R}^k : w_i \geq 0, \sum_{i=1}^k w_i = 1\}$ is the $(k-1)$-dimensional probability simplex.*

In this thesis, we abuse notation and simply write $f \in k\text{-mix}(\mathcal{F})$ to denote a k-mixture from the class $\mathcal{F}$ where the representation of the weights is implicit. Further, if $f \in k\text{-mix}(\mathcal{F})$ and $g \in k'\text{-mix}(\mathcal{F})$ then we write $\mathrm{d_{TV}}(f, g)$ to denote the TV distance from the underlying densities. In other words, if $f = \sum_{i \in [k]} w_i f_i$ and $g = \sum_{i \in [k']} w'_i g_i$ then $\mathrm{d_{TV}}(f, g) = \mathrm{d_{TV}}(\sum_{i \in [k]} w_i f_i, \sum_{i \in [k']} w'_i g_i)$. The purpose of defining mixture distributions in this way is that it turns out to be convenient to define a distance between different representations of a distribution (see Definition 5.1.2).

Definition 2.1.5 (Unbounded Gaussians)**.** *Let $\mathcal{G} = \{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in S^d\}$ be the class of d-dimensional Gaussians.*

2.1.2 List decoding distributions

Here, we define the task of list decoding distributions under Huber's contamination model (Huber, 1992), where the samples are drawn from a corrupted version of the original distribution f that we are interested in. The contamination is additive, that is, with probability $1-\gamma$ we receive samples from f, and with probability γ we receive samples from an arbitrary distribution h. Upon receiving corrupted samples from f, the goal of a list decoding algorithm is to output a short list of distributions one of which is close to the original distribution f. In the next definition, we use a general metric to measure the closeness; this allow for choosing the metric based on the application.

Definition 2.1.6 (List decodable learning under Huber's contamination)**.** *Let $\mathcal{F}$ be a class of distributions and $\kappa : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_{\geq 0}$ be a metric defined on it. For $\alpha, \beta, \gamma \in (0,1)$, an algorithm $A_{\mathcal{F}}$ is said to be an $(L, m, \alpha, \beta, \gamma)$-list-decoding algorithm for $\mathcal{F}$ w.r.t. κ if the following holds:*

> *For any $f \in \mathcal{F}$ and arbitrary distribution h, given an i.i.d. sample S of size m from $g = (1-\gamma)f + \gamma h$, $A_{\mathcal{F}}$ outputs a list of distributions $\hat{\mathcal{F}} \subseteq \mathcal{F}$ of size no more than L such that with probability at least $1-\beta$ (over the randomness of S and $A_{\mathcal{F}}$) we have $\min_{\hat{f} \in \hat{\mathcal{F}}} \kappa(\hat{f}, f) \leq \alpha$.*

2.1.3 Distribution learning

A distribution learner is a, possibly randomized, algorithm that receives i.i.d. samples from a distribution f, and outputs a distribution $\hat{f}$ which is close to f.

Definition 2.1.7 (PAC learning)**.** *An algorithm $A_{\mathcal{F}}$ is said to (α, β)-PAC-learn a*

class of distributions $\mathcal{F}$ w.r.t. metric κ with $m(\alpha, \beta)$ samples, if for any $f \in \mathcal{F}$, and any $\alpha, \beta \in (0, 1)$, after receiving $m(\alpha, \beta)$ i.i.d. samples from f, outputs a distribution $\hat{f} \in \mathcal{F}$ such that $\kappa(f, \hat{f}) \leq \alpha$ with probability at least $1 - \beta$. Moreover, if such an algorithm exists, we call $\mathcal{F}$ to be (α, β)-PAC-learnable w.r.t. κ. The sample complexity of learning $\mathcal{F}$ is the minimum $m(\alpha, \beta)$ among all such (α, β)-PAC-learners.

Remark 2.1.8. *Equivalently, an algorithm $A_{\mathcal{F}}$ is said to (α, β)-PAC-learn a class of distributions $\mathcal{F}$ w.r.t. metric κ with $m(\alpha, \beta)$ samples, if for any $\alpha, \beta \in (0, 1)$, $A_{\mathcal{F}}$ is a $(1, m(\alpha, \beta), \alpha, \beta, 0)$-list-decoding algorithm for $\mathcal{F}$ w.r.t. κ.*

2.1.4 Differential privacy

Two datasets $D, D' \in \mathcal{X}^n$ are called neighbouring datasets if they differ by one element. Informally, a differentially private algorithm is required to have close output distributions on neighbouring datasets.

Definition 2.1.9 ((ε, δ)-Indistinguishable)**.** *Two distribution Y, Y' with support $\mathcal{Y}$ are said to be (ε, δ)-indistinguishable if for all measurable subsets $E \in \mathcal{Y}$, $\mathbb{P}_{X \sim Y}[X \in E] \leq e^{\varepsilon}\mathbb{P}_{X \sim Y'}[X \in E] + \delta$ and $\mathbb{P}_{X \sim Y'}[X \in E] \leq e^{\varepsilon}\mathbb{P}_{X \sim Y}[X \in E] + \delta$.*

Definition 2.1.10 ((ε, δ)-Differential Privacy (Dwork *et al.*, 2006a,b))**.** *A randomized algorithm $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ is said to be (ε, δ)-differentially private if for every two neighbouring datasets $D, D' \in \mathcal{X}^n$, the output distributions $\mathcal{M}(D), \mathcal{M}(D')$ are (ε, δ)-indistinguishable.*

2.2 Main results

In this section, we describe our main results. We introduce a general framework for privately learning mixture distributions, and as an application, we propose the first finite upper bound on the sample complexity of privately learning general GMMs. More specifically, we show that if we have *(1)* a locally small cover (w.r.t. $\mathrm{d_{TV}}$), and *(2)* a list decoding algorithm (w.r.t. $\mathrm{d_{TV}}$) for a class of distributions, then the class of its mixtures is privately learnable.

2.2.1 Privately learning mixtures

Theorem 2.2.1 (Reduction)**.** *For $\alpha, \beta \in (0,1)$, if a class of distributions $\mathcal{F}$ admits a $(t, 2\alpha/15)$-locally small $\frac{\alpha}{15}$-cover (w.r.t. $\mathrm{d_{TV}}$), and it is $(L, m, \alpha/15, \beta', 0)$-list-decodable (w.r.t. $\mathrm{d_{TV}}$), where $\log(1/\beta') = \widetilde{\Theta}(\log(mk\log(tL/\alpha\delta)/\varepsilon\beta))$, then k-mix$(\mathcal{F})$ is (ε, δ)-DP (α, β)-PAC-learnable (w.r.t. $\mathrm{d_{TV}}$) with sample complexity*

$$\tilde{O}\left(\left(\frac{\log(1/\delta) + k\log(tL)}{\varepsilon} + \frac{mk + k\log(1/\beta)}{\alpha\varepsilon}\right) \cdot \left(\frac{k\log(L)}{\alpha^2} + \frac{mk + k\log(1/\beta)}{\alpha^3}\right)\right).$$

Note that in Theorem 2.2.1, we can use a naive (α, β)-PAC-learner that outputs *a single* distribution as the list decoding algorithm (see Remark 2.1.8). Therefore, if we have *(1)* a locally small cover (w.r.t. $\mathrm{d_{TV}}$), and *(2)* a (non-private) PAC learner (w.r.t. $\mathrm{d_{TV}}$) for a class of distributions, then the class of its mixtures is privately learnable[1]. The next corollary states this result formally.

[1]Later in Remark 2.2.4, we explain how Theorem 2.2.1 can sometimes give us a better bound compared to Corollary 2.2.2.

Corollary 2.2.2. *For $\alpha, \beta \in (0,1)$, if a class of distributions $\mathcal{F}$ admits a $(t, 2\alpha/15)$-locally small $\frac{\alpha}{15}$-cover (w.r.t. $\mathrm{d_{TV}}$), and it is $(\alpha/15, \beta')$-PAC-learnable (w.r.t. $\mathrm{d_{TV}}$) using $m(\alpha/15, \beta')$ samples, where $\log(1/\beta') = \widetilde{\Theta}(\log(mk\log(t/\alpha\delta)/\varepsilon\beta))$, then k-mix$(\mathcal{F})$ is (ε, δ)-DP (α, β)-PAC-learnable (w.r.t. $\mathrm{d_{TV}}$) with sample complexity*

$$\tilde{O}\left(\left(\frac{\log(1/\delta)}{\varepsilon} + \frac{m(\alpha/15, \beta')k + k\log(1/\beta)}{\alpha\varepsilon}\right) \cdot \left(\frac{m(\alpha/15, \beta')k + k\log(1/\beta)}{\alpha^3}\right)\right).$$

2.2.2 Privately learning GMMs

As an application of the Theorem 2.2.1, we show that the class of GMMs is privately learnable. We need two ingredients to do so. We show that the class of unbounded Gaussians *(1)* has a locally small cover, and *(2)* is list decodable (using compression).

As a result, we prove the first sample complexity upper bound for privately learning general GMMs. Notably, the above upper bound is polynomial in all the parameters of interest.

Theorem 2.2.3 (Private Learning of GMMs)**.** *Let $\alpha, \beta \in (0,1)$. The class k-mix$(\mathcal{G})$ is (ε, δ)-DP (α, β)-PAC-learnable w.r.t. $\mathrm{d_{TV}}$ with sample complexity*

$$\tilde{O}\left(\frac{kd^2\log(1/\delta) + k^2d^4}{\alpha^2\varepsilon} + \frac{kd\log(1/\delta)\log(1/\beta) + k^2d^3\log(1/\beta)}{\alpha^3\varepsilon} + \frac{k^2d^2\log^2(1/\beta)}{\alpha^4\varepsilon}\right).$$

Remark 2.2.4. *Note that if we had used Corollary 2.2.2 and a PAC learner as a naive list decoding algorithm for Gaussians, the resulting sample complexity would have become worse. To see this, note that $\mathcal{G}$ is (α, β)-PAC-learnable using $m(\alpha, \beta) = O(\frac{d^2\log(1/\beta)}{\alpha^2})$ samples. Using Corollary 2.2.2 and the existence of a locally small cover*

for Guassians, we obtain a sample complexity upper bound of

$$\tilde{O}\left(\frac{kd^2\log(1/\delta)\log(1/\beta)}{\alpha^5\varepsilon}+\frac{k^2d^4\log^2(1/\beta)}{\alpha^8\varepsilon}\right).$$

This is a weaker result compared to Theorem 2.2.3 in terms of α, which was based on a more sophisticated (compression-based) list decoding algorithm for Gaussians.

It is worth mentioning that our approach is information-theoretic and does not yield a finite time algorithm for privately learning GMMs, due to the non-constructive cover that we use for Gaussians. Moreover, designing a computationally efficient algorithm (i.e. with a running time that is polynomial in k and d) for learning GMMs even in the non-private setting remains an open problem (Diakonikolas *et al.*, 2017).

2.3 Technical challenges and overview of techniques

Given the subtleties in the proofs, we give an overview of our technical contributions in this chapter before delving into the core technical sections.

Dense mixtures. As a simple first step, we reduce the problem of learning mixture distributions to the problem of learning "dense mixtures" (i.e., those mixtures whose component weights are not too small). Working with dense mixtures is more convenient since a large enough sample from a dense mixture will include samples from *every* component.

Locally small cover for GMMs w.r.t. $\mathrm{d_{TV}}$? One idea to privately learn (dense) GMMs is to create a locally small cover w.r.t. $\mathrm{d_{TV}}$ (see Definition 2.1.3) for this class and then apply "advanced" private hypothesis selection (Bun *et al.*, 2021). However, as Aden-Ali *et al.* (2021b) showed, such a locally small cover (w.r.t.$\mathrm{d_{TV}}$)

does *not* exists, even for a mixtures of two (dense) Gaussians.

A locally small cover for the component-wise distance. An alternative measure for the distance between two mixtures is their component-wise distance, which we denote by κ_{mix} (see Definition 5.1.2). Intuitively, given two mixtures, κ_{mix} measures the distance between their farthest components. Therefore, if two GMMs are close in κ_{mix} then they are close in $\mathrm{d_{TV}}$ distance too. Interestingly, we prove that GMMs *do* admit a locally small cover w.r.t. κ_{mix}. To prove this, we first show that if a class of distributions admits a locally small cover w.r.t. $\mathrm{d_{TV}}$ then the class of its mixtures admits a locally small cover w.r.t. κ_{mix}. Next, we argue that the class of Gaussians admits a locally small cover w.r.t. $\mathrm{d_{TV}}$. Building a locally small cover for the class of Gaussians is challenging due to the complex geometry of this class. We show the existence of such cover using the techniques of Aden-Ali *et al.* (2021a) and the recently proved lower bound for the $\mathrm{d_{TV}}$ distance between two (high dimensional) Gaussians (Arbas *et al.*, 2023).

Hardness of learning GMMs w.r.t. κ_{mix}. Given that we have a locally small cover for GMMs w.r.t. κ_{mix}, one may hope to apply some ideas similar to private hypothesis selection for privately learning GMMs w.r.t. κ_{mix}. Unfortunately, learning GMMs w.r.t. κ_{mix}, even in the non-private setting, requires exponentially many samples in terms of the number of components (Moitra and Valiant, 2010).

List decoding (dense mixtures) w.r.t. κ_{mix}. Interestingly, we show that unlike *PAC learning*, *list decoding* GMMs w.r.t. κ_{mix} can be done with a polynomial number of samples. To show this, first, we prove that if a class of distributions is list decodable (w.r.t. $\mathrm{d_{TV}}$), then class of its dense mixtures is list decodable (w.r.t. κ_{mix}). Then for the class of Guassians, we use a compression-based (Ashtiani *et al.*, 2018a)

list decoding method.

Privacy challenges of using the list decoder. Unfortunately, the list decoding method we described is not private. Otherwise, we could have used Private Hypothesis Selection (Bun *et al.*, 2021) to privately choose from the list of candidate GMMs. To alleviate this problem, we define and solve the "private common member selection" problem below.

Private common member selection. Given T lists of objects (e.g., distributions), we say an object is a common member if it is close (w.r.t. some metric κ) to a member in each of the lists (we give a rigorous definition Chapter 4). The goal of a private common member selector (PCMS) is then to privately find a common member assuming at least one exists. We then show *(1)* how to use a PCMS to learn GMMs privately and *(2)* how to solve the PCMS itself. This will conclude the proof of Theorem 2.2.1.

Private learning of GMMs using PCMS. Given a PCMS, we first run the (non-private) list decoding algorithm on T disjoint datasets to generate T lists of dense mixture distributions. At this point, we are guaranteed that with high probability, there exists a common member for these lists w.r.t. κ_{mix}. Therefore, we can simply run the PCMS method to find such a common member. However, note that not all the common members are necessarily "good": there might be some other common members that are far from the true distribution w.r.t. $\mathrm{d_{TV}}$. To resolve this issue, in each list we filter out (non-privately) the distributions that are far from the true distribution. After filtering the lists, we are still guaranteed to have a "good" common member and therefore we can run PCMS to choose it privately.

Designing a PCMS for locally small spaces. Finally, we give a recipe for designing a private common member selector for T lists w.r.t. a generic metric κ. To do so, assume we have access to a locally small cover for the space w.r.t. κ (indeed, we had showed this is the case for the space of GMMs w.r.t. κ_{mix}). We need to privately choose a member from this cover that represents a common member. We then design a score function such that: *(1)* a common member gets a high score and *(2)* the sensitivity of the score function is low (i.e., changing one of the input *lists* does not change the score of any member drastically). Using this score function, we apply the GAP-MAX algorithm of Bun *et al.* (2021, 2018) to privately select a member with a high score from the infinite (but locally small) cover.

In the next chapter we give a background on differential privacy, distribution learning. We also go through some related works.

Chapter 3

Background

In this chapter, we present some standard facts on distribution learning and differential privacy. Later, we discuss related work and results in the field.

3.1 Standard facts

We begin by a well-known result on learning finite class of distributions.

3.1.1 Learning finite classes

The following result on learning a finite class of distributions is based on the Minimum Distance Estimator (Yatracos, 1985); see the excellent book by Devroye and Lugosi (2001) for details.

Theorem 3.1.1 (Learning finite classes, Theorem 6.3 of Devroye and Lugosi (2001))**.** *Let $\alpha, \beta \in (0,1)$. Given a finite class of distributions $\mathcal{F}$, there is an algorithm that upon receiving $O(\frac{\log|\mathcal{F}|+\log(1/\beta)}{\alpha^2})$ i.i.d. samples from a distribution g, returns an $\hat{f} \in \mathcal{F}$ such that $\mathrm{d_{TV}}(\hat{f}, g) \leq 3 \cdot \min_{f\in\mathcal{F}} \mathrm{d_{TV}}(f, g) + \alpha$ with probability at least $1-\beta$.*

Next, we discuss some known results in differential privacy that are helpful in understanding the ideas used in this thesis.

3.1.2 Differential privacy toolkit

A known tool for privately choosing a "good" item from a set of candidates is Exponential Mechanism (McSherry and Talwar, 2007), where the "goodness" of candidates is measured using a score function.

Theorem 3.1.2 (Exponential Mechanism (McSherry and Talwar, 2007)). *Let* $(\mathcal{F}, \kappa)$ *be a metric space and* $\mathcal{X}$ *be an arbitrary set. Let* $\text{score}\colon \mathcal{F} \times \mathcal{X}^T \to \mathbb{R}_{\geq 0}$ *be a function such that for any* $f \in \mathcal{F}$ *and any two neighbouring sets* $D \sim D' \in \mathcal{X}^T$*, we have* $|\text{score}(f, D) - \text{score}(f, D')| \leq \Delta$*. Then there is an algorithm, called the Exponential Mechanism, that is* $(\varepsilon, 0)$*-DP with the following property. For every* $D \in \mathcal{X}^T$*, and* $\beta \in (0, 1)$*, it outputs an element* $\hat{f} \in \mathcal{F}$ *satisfying*

$$\text{score}(\hat{f}, D) \geq \max_{f \in \mathcal{F}} \text{score}(f, D) - \frac{2\Delta \log(|\mathcal{F}|/\beta)}{\varepsilon}$$

with probability at least $1 - \beta$*.*

However, Exponential Mechanism fails in the regimes where the candidate set is not finite. This leads us to use the GAP-MAX algorithm of (Bun *et al.*, 2021, 2018) that has the advantage of compatibility with infinite candidate sets. GAP-MAX guarantees returning a "good" candidate as long as the number of "near good" candidates is small.

Theorem 3.1.3 (GAP-MAX, Theorem IV.6. of Bun *et al.* (2021)). *Let* $(\mathcal{F}, \kappa)$ *be a metric space and* $\mathcal{X}$ *be an arbitrary set. Let* $\text{score}\colon \mathcal{F} \times \mathcal{X}^T \to \mathbb{R}_{\geq 0}$ *be a function*

such that for any $f \in \mathcal{F}$ *and any two neighbouring sets* $D \sim D' \in \mathcal{X}^T$*, we have* $|\operatorname{score}(f, D) - \operatorname{score}(f, D')| \leq 1$*. Then there is an algorithm, called the GAP-MAX algorithm, that is* (ε, δ)*-DP with the following property. For every* $D \in \mathcal{X}^T$ *and* $\alpha' \in (0, 1)$*, if*

$$\left| \left\{ f \in \mathcal{F} : \operatorname{score}(f, D) \geq \sup_{f' \in \mathcal{F}} \operatorname{score}(f', D) - 5\alpha' T \right\} \right| \leq t$$

then

$$\mathbb{P}\left[\operatorname{score}(\textit{GAP-MAX}(\mathcal{F}, D, \operatorname{score}, \alpha', \beta), D) \geq \sup_{f' \in \mathcal{F}} \operatorname{score}(f', D) - \alpha' T \right] \geq 1 - \beta$$

provided $T = \Omega\left(\frac{\min\{\log|\mathcal{F}|, \log(1/\delta)\} + \log(t/\beta)}{\alpha' \varepsilon} \right)$.

In the context of distribution learning, another useful tool for privately selecting a hypothesis from a set of candidate distributions is the PHS method of Aden-Ali *et al.* (2021a).

Theorem 3.1.4. *Let* $\alpha, \beta \in (0, 1)$*, and* $\varepsilon > 0$*. Given a finite class of distributions* $\mathcal{F}$*, there is an* $(\varepsilon, 0)$*-DP algorithm that upon receiving* $O(\frac{\log(|\mathcal{F}|/\beta)}{\alpha^2} + \frac{\log(|\mathcal{F}|/\beta)}{\alpha\varepsilon})$ *i.i.d. samples from a distribution* g*, returns an* $\hat{f} \in \mathcal{F}$ *such that* $\mathrm{d_{TV}}(\hat{f}, g) \leq 3 \cdot \min_{f \in \mathcal{F}} \mathrm{d_{TV}}(f, g) + \alpha$ *with probability at least* $1 - \beta$*.*

3.2 Related Work

In this section, we go through some related works on private learning of Gaussians and their mixtures.

3.2.1 Privately learning Gaussians

The very first and simplest case of this problem is the bounded univariate Gaussian. Indeed, Karwa and Vadhan (2018) proposed a sample- and time-efficient algorithm for estimating the mean and variance of bounded univariate Gaussians under pure DP. We cannot hope to privately estimate the mean and variance without boundedness assumptions. In fact, this is impossible due to information-theoretical lower bounds for this problem. Furthermore, Karwa and Vadhan (2018) also provide a method for estimating the mean and variance of univariate Gaussians for the unbounded setting under approximate DP.

Later, other methods for learning high-dimensional Gaussians with respect to total variation distance were introduced by Kamath *et al.* (2019b); Biswas *et al.* (2020). They assume the parameters of the Gaussians have bounded ranges. As a result, the sample complexity of their method depends on the condition number of the covariance matrix, and the range of the mean.

Afterwards, Aden-Ali *et al.* (2021a) proposed the first finite sample complexity upper bound for privately learning unbounded high-dimensional Gaussians, which was nearly tight and matching the lower bound of Kamath *et al.* (2022a). Their method was based on a privatized version of the Minimum Distance Estimator (Yatracos, 1985) and was inspired by the private hypothesis selection approach of Bun *et al.* (2021). They indeed proved the existence of a "locally small" cover for the space of Gaussian covariance matrices and mean vectors and utilized a stable scoring function (based on the Minimum Distance Estimator) to privately choose one good element from that cover.

One downside of Aden-Ali *et al.* (2021a) is that their method is not computationally efficient. This is because they did not explicitly construct the cover and only showed the existence of such a cover using Zorn's lemma.

There have been several recent results on computationally efficient learning of unbounded Gaussians (Kamath *et al.*, 2022b; Kothari *et al.*, 2022; Ashtiani and Liaw, 2022), with the method of Ashtiani and Liaw (2022) achieving near-optimal sample complexity using a sample-and-aggregate-based technique. The framework works as follows: they run a non-private algorithm for learning Gaussians on many different datasets, and use the propose-test-release approach to privately check whether the output distributions are close together or not. If not, their algorithm halts as the inputs were not good; otherwise, they compute a weighted average of the output distributions to reduce the sensitivity. Finally, they privately mask the averaged distribution and output the noisy distribution.

Another sample-and-aggregate framework that can be used for this task is FriendlyCore (Tsfadia *et al.*, 2022). The methods of Ashtiani and Liaw (2022); Kothari *et al.* (2022) also work in the robust setting achieving sub-optimal sample complexities. Recently, Alabi *et al.* (2023) improved this result in terms of dependence on the dimension. Finally, Hopkins *et al.* (2023) achieved a robust and efficient learner with near-optimal sample complexity for unbounded Gaussians.

In the pure DP setting, Hopkins *et al.* (2022) proposed a method for efficiently learning Gaussians with bounded parameters. There are some other related works on private mean estimation w.r.t. Mahalanobis distance in Brown *et al.* (2021); Duchi *et al.* (2023); Brown *et al.* (2023).

Several other related studies have explored the relationship between robust and

private estimation, as seen in Dwork and Lei (2009); Georgiev and Hopkins (2022); Liu *et al.* (2022b); Hopkins *et al.* (2023); Asi *et al.* (2023). Also, there have been investigations into designing estimators that achieve both privacy and robustness at the same time (Liu *et al.*, 2021).

3.2.2 Parameter estimation for GMMs

In this setting, upon receiving i.i.d. samples from a GMM, the goal is to estimate the parameters of the mixture. Note that parameter estimation of GMMs (even in the non-private setting) is statistically harder than density estimation for GMMs. In other words, while performing parameter estimation, the exponential dependence of the sample complexity on the number of components is inevitable (Moitra and Valiant, 2010). This is because there are so many different parameter representations for a single GMM, all of which are very close with respect to TV distance. Since they are very close with respect to TV distance, one cannot hope to distinguish the underlying parameters unless they have access to an exponentially large number of samples in terms of the number of components.

In the non-private setting there is an extensive line of research for parameter learning of GMMs (Dasgupta, 1999; Sanjeev and Kannan, 2001; Vempala and Wang, 2004; Achlioptas and McSherry, 2005; Brubaker and Vempala, 2008; Kalai *et al.*, 2010; Belkin and Sinha, 2009; Hardt and Price, 2014; Hsu and Kakade, 2013; Anderson *et al.*, 2014; Regev and Vijayaraghavan, 2017; Kothari *et al.*, 2018; Hopkins and Li, 2018; Liu and Li, 2022; Feldman *et al.*, 2006; Moitra and Valiant, 2010; Belkin and Sinha, 2010; Bakshi *et al.*, 2022; Liu and Moitra, 2021, 2022).

Under the boundedness assumption there has been a line of work in privately

learning parameters of GMMs (Nissim *et al.*, 2007; Vempala and Wang, 2004; Chen *et al.*, 2023; Kamath *et al.*, 2019a; Achlioptas and McSherry, 2005; Cohen *et al.*, 2021). The work of Bie *et al.* (2022) approaches the same problem by taking the advantage of public data. Recently, Arbas *et al.* (2023) proposed an efficient method for reducing the private parameter estimation of unbounded GMMs to its non-private counterpart. They extend the method proposed by Ashtiani and Liaw (2022) for privately learning Gaussians. Arbas *et al.* (2023) employ a non-private parameter estimator on various datasets. Initially, they ensure privately that the outputs are stable and close to each other. Subsequently, they devise a masking (noising) mechanism for GMMs based on parameter distance, and finally, they randomly select one output from the non-private algorithm and apply the masking mechanism to it.

3.2.3 Density estimation for GMMs

In density estimation, which is the main focus of this work, the goal is to find a distribution which is close to the underlying distribution w.r.t. $\mathrm{d_{TV}}$. Unlike parameter estimation, the sample complexity of density estimation can be polynomial in both the dimension and the number of components. In the non-private setting, there has been several results about the sample complexity of learning GMMs (Devroye and Lugosi, 2001; Ashtiani *et al.*, 2018b), culminating in the work of (Ashtiani *et al.*, 2018a, 2020) which gives the near-optimal bound of $\tilde{\Theta}(kd^2/\alpha^2)$.

There are some researches on designing computationally efficient learners for one-dimensional GMMs (Chan *et al.*, 2014; Acharya *et al.*, 2017; Liu *et al.*, 2022a; Wu and Xie, 2018; Li and Schmidt, 2017). However, for general GMMs it is hard to come up with a computationally efficient learner due to the known statistical query lower

bounds (Diakonikolas *et al.*, 2017).

In the private setting, and under the assumption of bounded components, one can use the Private Hypothesis Selection of Bun *et al.* (2021) or private Minimum Distance Estimator of Aden-Ali *et al.* (2021a) to learn classes that admit a finite cover under the constraint of pure DP. Bun *et al.* (2021) also proposes a way to construct such finite cover for mixtures of finite classes.

Later, Aden-Ali *et al.* (2021b) introduced the first polynomial sample complexity upper bound for learning unbounded univariate GMMs under the constraint of approximate differential privacy (DP). They also extended this result to the case of axis-aligned GMMs, where the covariance matrices of the mixture components are assumed to be diagonal matrices. They expanded on the idea of stable histograms used in Karwa and Vadhan (2018) to learn univariate GMMs. At a high level, they privately output a list of possible parameters (mean and variance) of the GMM at each coordinate. This is feasible because the utility of stable histograms in the approximate setting does not depend on the number of bins, which is infinite in this case. Finally, given a finite set of possible parameters, one can then privately select a good set of parameters at each coordinate using private hypothesis selection (Bun *et al.*, 2021; Aden-Ali *et al.*, 2021a).

However, this idea cannot be generalized to general GMMs, as it is not clear how to learn even a single high-dimensional Gaussian using a stability-based histogram. This is simply because there is too much room in high dimensions, and samples are scattered among exponentially many candidate bins, making it difficult to determine which ones are truly effective.

Another related work is the lower bound on the sample complexity of privately

learning GMMs with known covariance matrices (Acharya *et al.*, 2021).

In an independent and concurrent work, Ben-David *et al.* (2023) proposed a pure DP method for learning general GMMs, assuming they have access to additional public samples. In fact, they use sample compression (with public samples) to find a list of candidate GMMs. Since this list is created using public data, they can simply choose a good member of it using private hypothesis selection. We also create such lists in the process. However, most of the technical part of this paper is dedicated to guarantee privacy (with only access to private data). In fact, it is challenging to privatize compression-based approaches since by definition, they heavily rely on a few data points in the data set. Ben-David *et al.* (2023) also study an equivalence between public-private learning, list decodable learning and sample compression schemes.

We also use sample compression in our list decoding algorithm for Gaussians (w.r.t component-wise distance).

Our work is the first polynomial sample complexity upper bound for privately learning mixtures of general GMMs without any restrictive assumptions. Designing a private and computationally efficient density estimator for GMMs remains an open problem even in the one-dimensional setting.

In the next few chapters we dive into technical details of our proofs. We start by introducing the problem of private common member selection, which is the main component in our reduction.

Chapter 4

Private Common Member Selection

In this chapter, we introduce the problem of Common Member Selection, which is used in our main reduction for privately learning mixture distributions. At a high-level, given a set of lists, a common member is an element that is close to some element in most of the lists. The problem of common member selection is to find such an item.

4.1 Problem statement

Definition 4.1.1 (Common Member)**.** *Let $(\mathcal{F}, \kappa)$ be a metric space and $\alpha, \zeta \in (0, 1]$. We say $f \in \mathcal{F}$ is an (α, ζ)-common-member for $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_T\} \in (\mathcal{F}^*)^T$, if there exists a subset $\mathcal{Y}' \subseteq \mathcal{Y}$ of size at least ζT, such that $\max_{Y \in \mathcal{Y}'} \kappa(f, Y) \leq \alpha$.*

Definition 4.1.2 (Common Member Selector (CMS))**.** *Let $(\mathcal{F}, \kappa)$ be a metric space, and $\alpha, \zeta, \beta \in (0, 1]$. An algorithm $\mathcal{A}$ is said to be a $(T_0, Q, \alpha, \zeta, \beta)$-common-member*

selector w.r.t. κ if the following holds for all $T \geq T_0$:

> *Given any $\mathcal{Y} = \{Y_1, Y_2, ..., Y_T\} \in (\mathcal{F}^*)^T$ that satisfies $|Y_i| \leq Q$ for all $i \in [T]$, if there exists at least one $(\alpha, 1)$-common-member for $\mathcal{Y}$, then $\mathcal{A}$ outputs a $(2\alpha, \zeta)$-common-member with probability at least $1 - \beta$.*

Remark 4.1.3. *Note that the CMS problem on its own is a trivial task and can be done using a simple brute force algorithm. However, we are interested in the non-trivial privatized version of this problem. The formal definition of the private CMS is give below.*

Definition 4.1.4 (Private Common Member Selector (PCMS))**.** *Let $(\mathcal{F}, \kappa)$ be a metric space. Further, let $\alpha, \beta, \zeta, \delta \in (0, 1]$ and $\varepsilon \geq 0$ be parameters. An algorithm $\mathcal{A}$ is an (ε, δ)-DP $(T_0, Q, \alpha, \zeta, \beta)$-common-member selector w.r.t. κ if (1) it is a $(T_0, Q, \alpha, \zeta, \beta)$-CMS and (2) for any $T \geq T_0$ and any two collections of lists $C_1 = \{Y_1, Y_2, ..., Y_T\} \in (\mathcal{F}^*)^T$ and $C_2 = \{Y_1', Y_2, ..., Y_T\} \in (\mathcal{F}^*)^T$ that differ in only one list, the output distributions of $\mathcal{A}(C_1)$ and $\mathcal{A}(C_2)$ are (ε, δ)-indistinguishable.*

4.2 The proposed algorithm

In Algorithm 1, we describe an algorithm for privately fining a common member provided that one exists. We note that one requirement is that we have access to a locally small cover $\mathcal{C}$ for $\mathcal{F}$. At a high-level, given a family of sets $\{Y_1, \ldots, Y_T\}$, where each Y_t is a set of elements, we can assign a score to each point $c \in \mathcal{C}$ to be the number of Y_t's that contain a element close to c. We observe that the sensitivity of this score function is 1, meaning that by changing one set, the score of each point will change by at most 1. We remark that this is because the score of a point c is defined

to be the number of lists that contains an element close to it, not the total number of elements. Note that any point c with a sufficiently high score is a common member. Further, since $\mathcal{C}$ is locally small, this allows us to apply the GAP-MAX algorithm (Theorem 3.1.3) to make this selection differentially private [1]. In Theorem 4.2.1, we prove the correctness of this algorithm.

Algorithm 1 Private Common Member Selector (PCMS)

Input: $D = \{Y_1, Y_2, ..., Y_T\} \in (\mathcal{F}^*)^T$, metric κ over $\mathcal{F}$, $(t, 2\alpha)$-locally-small α-cover C_α for $\mathcal{F}$ w.r.t. κ.

Output: $(2\alpha, 0.9)$-common-member of D (assuming D has an $(\alpha, 1)$-common-member)

1: For all $h \in C_\alpha$, set $\text{score}(h, D) \coloneqq |\{i \in [T] : \kappa(y, h) \leq 2\alpha \text{ for some } y \in Y_i\}|$.
2: **return** GAP-MAX$(C_\alpha, D, \text{score}, 0.1, \beta)$

Theorem 4.2.1. *Let $(\mathcal{F}, \kappa)$ be a metric space, $\alpha, \beta, \delta \in (0, 1]$, $Q \in \mathbb{N}$, $\varepsilon > 0$, and C_α be a $(t, 2\alpha)$-locally-small α-cover for $\mathcal{F}$ (w.r.t. κ). Algorithm 1 is an (ε, δ)-DP $(T, Q, \alpha, 0.9, \beta)$-common-member selector w.r.t. κ for some $T = O\left(\frac{\log(1/\delta)+\log(tQ/\beta)}{\varepsilon}\right)$.*

Proof. We first prove the utility of the algorithm.

Utility. We show that the output of the Algorithm 1 is an $(2\alpha, 0.9)$-common-member provided that there exists an $(\alpha, 1)$-common-member (recall Definition 4.1.2). Let the score function be defined as in Algorithm 1. Since C_α is $(t, 2\alpha)$-locally-small, we have

$$|\{h \in C_\alpha : \text{score}(h, D) \geq 1\}| \leq tQT$$

since for any $i \in [T]$ and any $y \in Y_i$, y contributes to at most t candidates' scores. As a result, there are at most tQT candidates with non-zero scores. Assuming that there exists an $(\alpha, 1)$-common-member for $\{Y_1, Y_2, ..., Y_T\}$, we have $\sup_{h \in C_\alpha} \text{score}(h, D) =$

[1] This task can also be done using the Choosing Mechanism of Bun *et al.* (2015).

T. Thus,

$$\begin{aligned} & \left|\left\{h \in C_\alpha : \text{score}(h, D) \geq \sup_{h \in C_\alpha} \text{score}(h, D) - T/2\right\}\right| \\ = \; & |\{h \in C_\alpha : \text{score}(h, D) \geq T - T/2\}| \\ = \; & |\{h \in C_\alpha : \text{score}(h, D) \geq T/2\}| \\ \leq \; & |\{h \in C_\alpha : \text{score}(h, D) \geq 1\}| \leq tQT. \end{aligned}$$

Using this bound, we can apply GAP-MAX algorithm in Theorem 3.1.3 with $\alpha' = 0.1$, and $T = O\left(\frac{\log(1/\delta)+\log(tQ/\beta)}{\varepsilon}\right)$. In particular, if $\hat{h} = \text{GAP-MAX}(C_\alpha, D, \text{score}, 0.1, \beta)$ then

$$\mathbb{P}\left[\text{score}(\hat{h}, D) \geq 0.9T\right] = \mathbb{P}\left[\left|\left\{i \in [T] \,:\, \exists y \in Y_i \text{ such that } \kappa(y, \hat{h}) \leq 2\alpha\right\}\right| \geq 0.9T\right] \geq 1 - \beta.$$

Privacy. Note that for any $h \in C_\alpha$ and any two neighbouring sets $D \sim D'$, we have $|\,\text{score}(h, D) - \text{score}(h, D')| \leq 1$ since each list $y \in D$ contributes to any h's score by at most 1. Thus, Theorem 3.1.3 implies that GAP-MAX is (ε, δ)-DP.

□

In the next chapter, we develop some results for mixture distributions. These properties will be useful later in our reduction.

Chapter 5

Mixtures Distributions and Their Properties

In this chapter, we study some general properties of mixture distributions. First, we introduce a component-wise distance between two mixture distributions which will be useful for constructing locally small covers. Generally, if we have a locally small cover for a class of distributions w.r.t. d_{TV}, then there exists a locally small cover w.r.t. component-wise distance for mixtures of that class. Later, we define dense mixtures and will show that if a class of distributions is list decodable w.r.t. d_{TV}, then the dense mixtures of that class are list decodable w.r.t. component-wise distance.

5.1 Component-wise distance between mixtures

Here, we define the class of general mixtures which are the mixtures with arbitrary number of components, as opposed to Definition 2.1.4, where the number of components is fixed.

Definition 5.1.1 (general mixtures)**.** *Let $\mathcal{F}$ be an arbitrary class of distributions. We denote the class of mixtures of $\mathcal{F}$ by $mix(\mathcal{F}) = \bigcup_{k=1}^{\infty} k\text{-}mix(\mathcal{F})$.*

Below, we define the component-wise distance between two mixture distributions with arbitrary number of components. The definition is inspired by Moitra and Valiant (2010). We set the distance between two mixtures with different number of components to be ∞. Otherwise, the distance between two mixtures is the distance between their farthest components.

Definition 5.1.2 (Component-wise distance between two mixtures)**.** *For a class $\mathcal{F}$ and every $g_1 = \sum_{i\in[k_1]} w_i f_i \in k_1\text{-}mix(\mathcal{F})$, $g_2 = \sum_{i\in[k_2]} w'_i f'_i \in k_2\text{-}mix(\mathcal{F})$, we define the distance $\kappa_{mix}\colon mix(\mathcal{F}) \times mix(\mathcal{F}) \to \mathbb{R}_{\geq 0}$ as*

$$\kappa_{mix}(g_1, g_2) = \begin{cases} \min_\pi \max_{i\in[k_1]} \max\{k_1.|w_i - w'_{\pi(i)}|, \mathrm{d_{TV}}(f_i, f'_{\pi(i)})\} & k_1 = k_2 \\ \infty & k_1 \neq k_2 \end{cases} \tag{5.1.1}$$

where π is chosen from all permutations over $[k_1]$.

The next lemma states that if two mixture distributions are close w.r.t. κ_{mix}, then they are also close w.r.t. $\mathrm{d_{TV}}$.

Lemma 5.1.3. *Let $\alpha \in [0,1]$ and $f = \sum_{i\in[k]} w_i f_i, f' = \sum_{i\in[k]} w'_i f'_i \in k\text{-}mix(\mathcal{F})$. If $\kappa_{mix}(f, f') \leq \alpha$, then $\mathrm{d_{TV}}(f, f') \leq 3\alpha/2$.*

Proof. Using the definition of κ_{mix}, we get that for every $i \in [k]$, $|w_i - w'_i| \leq \alpha/k$ and

$\mathrm{d_{TV}}(f_i, f'_i) \leq \alpha$. Therefore,

$$\begin{aligned}
\mathrm{d_{TV}}(f, f') = \frac{1}{2}||f - f'||_1 &= \frac{1}{2}||\sum_{i\in[k]} w_i f_i - \sum_{i\in[k]} w'_i f'_i||_1 \\
&\leq \frac{1}{2}\sum_{i\in[k]} ||w_i f_i - w'_i f'_i||_1 \\
&\leq \frac{1}{2}\sum_{i\in[k]} ||w_i f_i - w'_i f_i||_1 + ||w'_i f_i - w'_i f'_i||_1 \\
&\leq \frac{1}{2}\sum_{i\in[k]} ||w_i f_i - w'_i f_i||_1 + ||w'_i f_i - w'_i f'_i||_1 \\
&\leq \frac{1}{2}\sum_{i\in[k]} \frac{\alpha}{k} + \frac{1}{2}\sum_{i\in[k]} 2\alpha w'_i = 3\alpha/2. \qquad \square
\end{aligned}$$

5.2 Locally small cover for mixtures w.r.t. component-wise distance

The following simple proposition gives a locally small cover for weight vectors used to construct a mixture.

Proposition 5.2.1. *Let $\alpha \in (0, 1]$. There is an α-cover for $\Delta_k = \{(w_1, w_2, ..., w_k) \in \mathbb{R}^k_{\geq 0} : \sum_{i\in[k]} w_i = 1\}$ w.r.t. ℓ_∞ of size at most $(1/\alpha)^k$.*

Proof. Partition the cube $[0, 1]^k$ into small cubes of side-length $1/\alpha$. If for a cube c, we have $c \cap \Delta_k \neq \emptyset$, put one arbitrary point from $c \cap \Delta_k$ into the cover. The size of the constructed cover is no more than $(1/\alpha)^k$ which is the total number of small cubes. $\square$

The next lemma states that if a class of distributions has a locally small cover w.r.t. $\mathrm{d_{TV}}$, then the mixtures of that class admit a locally small cover w.r.t. κ_{mix}.

Note that the choice of the metric is important as the next theorem is false if we consider the $\mathrm{d_{TV}}$ metric for mixtures. In other words, there is a class of distributions (e.g. Gaussians) that admits a locally small cover w.r.t. $\mathrm{d_{TV}}$ but there is no locally small cover for the mixtures of that class w.r.t. $\mathrm{d_{TV}}$ (Proposition 1.3 of Aden-Ali *et al.* (2021b)).

Theorem 5.2.2. *For any $0 < \alpha < \gamma < 1$, if a class of distributions $\mathcal{F}$ has a (t, γ)-locally-small α-cover w.r.t. $\mathrm{d_{TV}}$, then the class k-mix$(\mathcal{F})$ has a $(k!(tk/\alpha)^k, \gamma)$-locally-small α-cover w.r.t. κ_{mix}.*

Proof. Let C_α be the (t, γ)-locally small α-cover for $\mathcal{F}$, and $\hat{\Delta}_k$ be an $\frac{\alpha}{k}$-cover for the probability simplex Δ_k from Proposition 5.2.1. Construct the set $\mathcal{J} = \{\sum_{i\in[k]} \hat{w}_i \hat{f}_i : \hat{w} \in \hat{\Delta}_k, \hat{f}_i \in C_\alpha\}$. Note that $\mathcal{J}$ is an α-cover for k-mix$(\mathcal{F})$ w.r.t. κ_{mix} since for any $g = \sum_{i\in[k]} w_i f_i \in k$-mix$(\mathcal{F})$, by construction, there exists an $g' \in \mathcal{J}$ such that $\kappa_{mix}(g, g') \le \alpha$. Moreover, we have $|B_{\kappa_{mix}}(\gamma, g, \mathcal{J})| \le |B_{\mathrm{d_{TV}}}(\gamma, f_i, C_\alpha)|^k \cdot |\hat{\Delta}_k| \cdot k! = t^k \cdot (k/\alpha)^k \cdot k!$, where the first term is because of composing the cover for a single component k times. The term $|\hat{\Delta}_k|$ comes from the size of cover for mixing weights of k components, and the $k!$ term is the result of permutation over k unordered components in the mixture. □

5.3 Dense mixtures

Dense mixtures are mixture distributions where each component has a non-negligible weight. Intuitively, a dense mixture is technically easier to deal with since given a large enough sample from the dense mixture, one would get samples from *all* of the components. This will allow us to show that if a class of distribution is list decodable

w.r.t. $\mathrm{d_{TV}}$, then the class of its dense mixtures is list decodable w.r.t. κ_{mix}. Later, we reduce the problem of learning mixture distributions to the problem of learning dense mixtures.

Definition 5.3.1 (Dense mixtures)**.** *Let $\mathcal{F}$ be an arbitrary class of distributions, $k \in \mathbb{N}$, and $\eta \in [0, 1/k]$. We denote the class of k-mixtures of $\mathcal{F}$ without negligible components by (k, η)-dense-mix$(\mathcal{F}) = \{\sum_{i=1}^{s} w_i f_i : s \leq k, w_i \geq \eta, \sum_{i=1}^{s} w_i = 1, f_i \in \mathcal{F}\}$.*

The next lemma states that every mixture distribution can be approximated using a dense mixture.

Lemma 5.3.2. *For every $k \in \mathbb{N}$, $g \in k$-mix$(\mathcal{F})$ and $\alpha \in [0, 1)$, there exists $\gamma \in [0, \alpha)$, $g' \in (k, \alpha/k)$-dense-mix$(\mathcal{F})$, and a distribution h such that $g = \gamma h + (1 - \gamma)g'$.*

Proof. For any $g = \sum_{i\in[k]} w_i f_i \in k$-mix$(\mathcal{F})$, let $N = \{i \in [k] : w_i < \alpha/k\}$ be the set of negligible weights, and $\gamma = \sum_{i\in N} w_i < \alpha$. Then g can be written as $g = (1 - \gamma) \sum_{i\in[k]\setminus N} \frac{w_i}{1-\gamma} f_i + \gamma \sum_{i\in N} \frac{w_i}{\gamma} f_i$. Note that $\sum_{i\in[k]\setminus N} \frac{w_i}{1-\gamma} f_i \in (k, \frac{\alpha}{k})$-dense-mix$(\mathcal{F})$. □

Theorem 5.2.2 shows that if a class of distributions admits a locally small cover (w.r.t. $\mathrm{d_{TV}}$) then the class of its mixtures admits a locally small cover (w.r.t. κ_{mix}). In the next lemma, we see that this is also the case for dense mixtures, i.e. if a class of distributions admits a locally small cover (w.r.t. $\mathrm{d_{TV}}$) then the class of its dense mixtures admits a locally small cover (w.r.t. κ_{mix}).

Lemma 5.3.3. *For any $0 < \alpha < \gamma < 1$, and $\alpha' \in (0, 1]$, if a class of distributions $\mathcal{F}$ has a (t, γ)-locally-small α-cover w.r.t. $\mathrm{d_{TV}}$, then the class $(k, \frac{\alpha'}{k})$-dense-mix$(\mathcal{F})$ has a $(k!(tk/\alpha)^k, \gamma)$-locally-small α-cover w.r.t. κ_{mix}.*

Proof. Using Theorem 5.2.2 we know that if $\mathcal{F}$ has a (t, γ)-locally-small α-cover w.r.t. $\mathrm{d_{TV}}$, then for every $i \in [k]$, there exists an $(i!(ti/\alpha)^i, \gamma)$-locally-small α-cover C_i for i-mix$(\mathcal{F})$ w.r.t. κ_{mix}. Since $(k, \frac{\alpha'}{k})$-dense-mix$(\mathcal{F}) \subseteq \bigcup_{i\in[k]} i$-mix$(\mathcal{F})$, we get that $\mathcal{J} = \bigcup_{i\in[k]} C_i$ is an α-cover for $(k, \frac{\alpha'}{k})$-dense-mix$(\mathcal{F})$. Moreover, $\mathcal{J}$ is $(k!(tk/\alpha)^k, \gamma)$-locally-small since the κ_{mix} distance is ∞ for two mixtures with different number of components. □

5.4 List decoding algorithm for dense mixtures

The following theorem is one of the main ingredients used for reducing the problem of privately learning mixtures to common member selection. It states that if a class of distributions is list decodable (w.r.t. $\mathrm{d_{TV}}$), then the class of its dense mixtures is list decodable (w.r.t. κ_{mix}).

Theorem 5.4.1. *For any $\alpha, \beta, \gamma \in (0, 1)$, if a class of distributions $\mathcal{F}$ is $(L, m, \alpha, \beta, 1-\alpha/k)$-list-decodable w.r.t. $\mathrm{d_{TV}}$, then the class $(k, \frac{\alpha}{k})$-dense-mix$(\mathcal{F})$ is $(L', m', \alpha, 2k\beta, \gamma)$-list-decodable w.r.t. κ_{mix}, where $L' = (\frac{kL}{\alpha})^{k+1} \cdot (\frac{10e\log(1/k\beta)}{1-\gamma})^m$, and $m' = \frac{2m+8\log(1/k\beta)}{1-\gamma}$.*

In order to prove Theorem 5.4.1, we need the following lemma, which states that if a class of distributions is list decodable with contamination level $\gamma = 0$, it is also list decodable with $\gamma > 0$, at the cost of additional number of samples *and an increased list size.*

Lemma 5.4.2. *For any $\alpha, \beta, \gamma \in (0, 1)$, if a class of distributions $\mathcal{F}$ is $(L, m, \alpha, \beta, 0)$-list-decodable w.r.t. κ, then it is $\left(L(\frac{10e\log(1/\beta)}{1-\gamma})^m, \frac{2m+8\log(1/\beta)}{1-\gamma}, \alpha, 2\beta, \gamma\right)$-list-decodable w.r.t. κ.*

Proof. Let $f \in \mathcal{F}$ and h be an arbitrary distribution. Consider $g = (1-\gamma)f + \gamma h$, where $\gamma \in (0,1)$. Upon drawing N samples from g, let X_N be the random variable indicating the number of samples coming from f. Note that X_N has binomial distribution. Setting $N \geq \frac{2m+8\log(1/\beta)}{1-\gamma}$, results in $\mathbb{E}[X_N]/2 \geq m$, $\mathbb{E}[X_N] \geq 8\log(1/\beta)$. Using the Chernoff bound (Theorem 4.5(2) of (Mitzenmacher and Upfal, 2005)), we have $\mathbb{P}[X_N \leq m] \leq \mathbb{P}[X_N \leq \mathbb{E}[X_N]/2] \leq \exp(-\mathbb{E}[X_N]/8) \leq \beta$. Meaning that after drawing $N \geq \frac{2m+8\log(1/\beta)}{1-\gamma}$ samples from g, with probability at least $1-\beta$, we will have m samples coming from f, which is enough for list decoding $\mathcal{F}$. Let $S_1, \cdots, S_K$ be all subsets of these N samples with size m, where $K = \binom{N}{m}$. Now, run the list decoding algorithm on these subsets and let $\mathcal{L}_i$ be the outputted list. Let $\mathcal{L} = \cup_{i\in[K]}\mathcal{L}_i$. Using the fact that among N samples there are m samples from f, we get that there exists $\hat{f} \in \mathcal{L}$ such that with probability at least $1-\beta$, we have $\kappa(f,\hat{f}) \leq \alpha$. Note that using Stirling's approximation we have $|\mathcal{L}| = L \cdot \binom{N}{m} \leq L \cdot (\frac{2em+8e\log(1/\beta)}{(1-\gamma)m})^m \leq L \cdot (\frac{10e\log(1/\beta)}{1-\gamma})^m$. Finally, using a union bound we will get that $\mathcal{F}$ is $\left(L(\frac{10e\log(1/\beta)}{1-\gamma})^m, \frac{2m+8\log(1/\beta)}{1-\gamma}, \alpha, 2\beta, \gamma\right)$-list-decodable w.r.t. κ. □

Proof of Theorem 5.4.1. Consider the algorithm $\mathcal{A}$ to be an $(L, m, \alpha, \beta, 1-\alpha/k)$-list-decodable learner for $\mathcal{F}$. Fix any distribution $g = \sum_{i\in[s]} w_i f_i \in (k, \frac{\alpha}{k})$-dense-mix($\mathcal{F}$), where $s \leq k$. Note that for any $i \in [s]$, g can be written as $g = w_i f_i + (1-w_i)\sum_{j\neq i} \frac{w_j f_j}{1-w_j} = w_i f_i + (1-w_i)h$. Knowing that $w_i \geq \frac{\alpha}{k}$, allows us to apply the algorithm $\mathcal{A}$ on the m samples generated from g and get a list of distributions $\mathcal{L}_s$ such that with probability at least $1-\beta$ we have $\min_{f'\in\mathcal{L}_s} \mathrm{d_{TV}}(f', f_i) \leq \alpha$. Let $\hat{\Delta}_s$ be an $\frac{\alpha}{s}$-cover for Δ_s from Proposition 5.2.1. Now construct a set $\mathcal{J} = \bigcup_{s\in[k]}\{\sum_{i\in[s]} \hat{w}_i \hat{f}_i : \hat{w} \in \hat{\Delta}_s, \hat{f}_i \in \mathcal{L}_s\}$. Note that with probability at least $1-k\beta$ we have $\min_{g'\in\mathcal{J}} \kappa_{mix}(g,g') \leq \max\{\alpha,\alpha\} = \alpha$. Moreover, we have $|\mathcal{J}| =$

$\sum_{s\in[k]} |\mathcal{L}_s|^s |\hat{\Delta}_s| = \sum_{s\in[k]} L^s (\frac{s}{\alpha})^s \leq (\frac{kL}{\alpha})^{k+1}$. Thus, the class of $(k, \frac{\alpha}{k})$-dense-mix$(\mathcal{F})$ is $\big((\frac{kL}{\alpha})^{k+1}, m, \alpha, k\beta, 0\big)$-list-decodable w.r.t. κ_{mix}. Using Lemma 5.4.2, we get that $(k, \frac{\alpha}{k})$-dense-mix$(\mathcal{F})$ is $\Big((\frac{kL}{\alpha})^{k+1} \cdot (\frac{10e\log(1/k\beta)}{1-\gamma})^m, \frac{2m+8\log(1/k\beta)}{1-\gamma}, \alpha, 2k\beta, \gamma\Big)$-list-decodable w.r.t. κ_{mix}.

□

In the next chapter, we put all the pieces together and prove our main theorem on privately learning mixture distributions.

Chapter 6

Proof of the Main Reduction

In this chapter, we prove our main reduction which states that if a class of distributions admits a locally small cover and is list decodable, then its *mixture class* can be learned privately. Let us first state the theorem again.

Theorem 2.2.1 (Reduction)**.** *For $\alpha, \beta \in (0,1)$, if a class of distributions $\mathcal{F}$ admits a $(t, 2\alpha/15)$-locally small $\frac{\alpha}{15}$-cover (w.r.t. $\mathrm{d_{TV}}$), and it is $(L, m, \alpha/15, \beta', 0)$-list-decodable (w.r.t. $\mathrm{d_{TV}}$), where $\log(1/\beta') = \widetilde{\Theta}(\log(mk\log(tL/\alpha\delta)/\varepsilon\beta))$, then k-mix$(\mathcal{F})$ is (ε, δ)-DP (α, β)-PAC-learnable (w.r.t. $\mathrm{d_{TV}}$) with sample complexity*

$$\tilde{O}\left(\left(\frac{\log(1/\delta) + k\log(tL)}{\varepsilon} + \frac{mk + k\log(1/\beta)}{\alpha\varepsilon}\right) \cdot \left(\frac{k\log(L)}{\alpha^2} + \frac{mk + k\log(1/\beta)}{\alpha^3}\right)\right).$$

The high level idea of the proof is based on a connection to the private common member selection problem. To see this, assume class $\mathcal{F}$ is list decodable and admits a locally small cover. Then we show that given some samples from any $f^* \in k$-mix$(\mathcal{F})$, one can generate T lists of dense mixtures (i.e., member of $(k, \frac{\alpha}{k})$-dense-mix$(\mathcal{F})$) such that they have a common member (w.r.t. κ_{mix}). However, there could be multiple

common members w.r.t. κ_{mix} that are far from f^* w.r.t. $\mathrm{d_{TV}}$. Therefore, we filter out all distributions that have a bad $\mathrm{d_{TV}}$ distance with the original distribution f^* (This can be done using Minimum Distance Estimator). Afterwards, we are guaranteed that all common members are also good w.r.t. $\mathrm{d_{TV}}$. Also, note that changing a data point can change at most one of the lists. Therefore, by using the private common member selector, we can choose a common member from these filtered lists while maintaining privacy. The formal proof is given below.

Proof. Let $\alpha', \beta' \in (0,1)$. If $\mathcal{F}$ is $(L, m, \alpha', \beta', 0)$-list-decodable w.r.t. $\mathrm{d_{TV}}$ then using Lemma 5.4.2, we get that $\mathcal{F}$ is $(L_1, m_1, \alpha', 2\beta', 1-\alpha/k)$-list-decodable w.r.t. $\mathrm{d_{TV}}$, where $L_1 = L \cdot (\frac{10ek\log(1/\beta')}{\alpha'})^m$, and $m_1 = \frac{2mk+8k\log(1/\beta')}{\alpha'}$.

Let $f^* \in k\text{-mix}(\mathcal{F})$ be the true distribution. Using Lemma 5.3.2, we can write $f^* = \gamma h + (1-\ \gamma)f$, where $f \in (k, \frac{\alpha'}{k})\text{-dense-mix}(\mathcal{F})$, and $\gamma \in [0, \alpha')$. Let $L_2 = (\frac{kL_1}{\alpha'})^{k+1} \cdot (\frac{10e\log(1/2k\beta')}{1-\alpha'})^{m_1}$, and $m_2 = \frac{2m_1+8\log(1/2k\beta')}{1-\alpha'}$. By Theorem 5.4.1, we know that $(k, \frac{\alpha'}{k})\text{-dense-mix}(\mathcal{F})$ is $(L_2, m_2, \alpha', 4k\beta', \alpha')$-list-decodable w.r.t. κ_{mix}.

Let $t_1 = k!(tk/\alpha')^k$, and $T = O\left(\frac{\log(1/\delta)+\log(t_1 L_2/\beta')}{\varepsilon}\right)$. For $i \in [T]$, draw T disjoint datasets each of size $m_3 = O\left(\frac{\log(L_2)+\log(1/\beta')}{\alpha'^2}\right) + m_2$. For each dataset, run the list decoding algorithm using m_2 samples from that dataset, and let $\mathcal{L}_i$ denote the outputted list.

As mentioned above, we know that $(k, \frac{\alpha'}{k})\text{-dense-mix}(\mathcal{F})$ is $(L_2, m_2, \alpha', 4k\beta', \alpha')$-list-decodable w.r.t. κ_{mix}. Thus, for each $i \in [T]$, with probability at least $1 - 4k\beta'$, we have $\kappa_{mix}(f, \mathcal{L}_i) \leq \alpha'$. To convert this bound back to total variation distance, we use Lemma 5.1.3 to get that $\mathrm{d_{TV}}(f, \mathcal{L}_i) \leq 3\alpha'/2$.

Note that the size of each $\mathcal{L}_i$ is at most L_2. By making use of the Minimum Distance Estimator (Theorem 3.1.1), we can use the other $O(\frac{\log(L_2)+\log(1/\beta')}{\alpha'^2})$ samples

from each datasets to find $\hat{f}_i \in \mathcal{L}_i$ such that with probability at least $1-\beta'$ we have $\mathrm{d_{TV}}(\hat{f}_i, f) \leq 3 \cdot \mathrm{d_{TV}}(f, \mathcal{L}_i) + \alpha' \leq 11\alpha'/2$.

We then proceed with a filtering step. For each $i \in [T]$, we define $\mathcal{L}'_i = \{f' \in \mathcal{L}_i : \mathrm{d_{TV}}(f', \hat{f}_i) < 11\alpha'/2\}$ to be the elements in $\mathcal{L}_i$ that are close to $\hat{f}_i$.

Using a union bound and a triangle inequality, we get that with probability at least $1-(4k+1)\beta'$ we have $\max_{f' \in \mathcal{L}'_i} \mathrm{d_{TV}}(f, f') \leq 11\alpha'$.

Applying a union bound over all T datasets, we conclude that with probability at least $1-(4k+1)\beta' T$, f is a $(\alpha', 1)$-common-member for $D = \{\mathcal{L}'_1, \cdots, \mathcal{L}'_T\}$ w.r.t. κ_{mix}, and $\max_{f' \in \mathcal{L}'_i} \mathrm{d_{TV}}(f, f') \leq 11\alpha'$ for all $i \in [T]$.

The fact that $\mathcal{F}$ admits a $(t, 2\alpha')$-locally small α'-cover, along with Lemma 5.3.3, implies that there exists an $(t_1, 2\alpha')$-locally small α'-cover $\mathcal{C}$ for $(k, \frac{\alpha'}{k})$-dense-mix$(\mathcal{F})$.

Note that $|\mathcal{L}'_i| \leq |\mathcal{L}_i| \leq L_2$. Now, we run the private common member selector (Algorithm 1) on $(D, \kappa_{mix}, \mathcal{C})$ to obtain $\hat{f}$. Using Theorem 4.2.1 and a union bound, we get that with probability at least $1-((4k+1)T+1)\beta'$, $\hat{f}$ is a $(2\alpha', 0.9)$-common-member w.r.t. κ_{mix}. Therefore, Lemma 5.1.3 implies that $\hat{f}$ is a $(3\alpha', 0.9)$-common-member w.r.t. $\mathrm{d_{TV}}$.

Using the fact that for every $i \in [T]$, $\max_{f' \in \mathcal{L}'_i} \mathrm{d_{TV}}(f', f) \leq 11\alpha'$, we get that $\mathrm{d_{TV}}(\hat{f}, f) \leq 14\alpha'$. Finally, triangle inequality implies that $\mathrm{d_{TV}}(\hat{f}, f^*) \leq \mathrm{d_{TV}}(\hat{f}, f) + \mathrm{d_{TV}}(f, f^*) \leq 15\alpha'$ with probability at least $1-((4k+1)T+1)\beta' \geq 1-6kT\beta'$. The total sample complexity is:

$$
\begin{aligned}
T \cdot m_3 &= O\left(\frac{\log(1/\delta) + \log(t_1 L_2/\beta')}{\varepsilon} \cdot \left(\frac{\log(L_2) + \log(1/\beta')}{\alpha'^2} + m_2\right)\right) \\
&= \tilde{O}\left(\left(\frac{\log(1/\delta) + k\log(tL)}{\varepsilon} + \frac{mk + k\log(1/\beta')}{\alpha'\varepsilon}\right) \cdot \left(\frac{k\log(L)}{\alpha'^2} + \frac{mk + k\log(1/\beta')}{\alpha'^3}\right)\right).
\end{aligned}
$$

Finally, we substitute $\alpha' = \alpha/15$ and $\beta' = \frac{\beta\varepsilon}{12ek\log(6ekt_1L_2/\varepsilon\beta\delta)} < 1$. Applying Claim A.0.3 with $c_1 = 6k/\varepsilon, c_2 = t_1L_2/\delta$, we get that $\mathrm{d_{TV}}(\hat{f}, f^*) \leq \alpha$ with probability at least $1-\beta$. Moreover the order of the sample complexity remains unchanged. The given approach is private since changing one data point alters only one of the T datasets, and therefore affects at most one $\mathcal{L}'_i$. The privacy guarantee then follows from Theorem 4.2.1.

□

As an application of our main theorem, we prove the first sample complexity upper bound for privately learning Gaussian mixtures (GMMs) in the next chapter.

Chapter 7

Privately Learning GMMs

Prior to this work, there was no a finite sample complexity upper bound for privately learning mixtures of unbounded Gaussians. As an application of our general framework, we study the problem of privately learning GMMs. This chapter provides the necessary ingredients for using our proposed framework. First, we show that the class of Gaussians is list decodable using sample compression schemes (Ashtiani *et al.*, 2018a). Second, we show that this class admits a locally small cover. Putting together, we prove the first sample complexity upper bound for privately learning GMMs.

7.1 List-decoding Gaussians using compression

As we stated in Remark 2.2.4, it is possible to use a PAC learner as a naive list decoding algorithm that outputs *a single* Gaussian. However, doing so, results in a poor sample complexity for privately learning GMMs. In this section, we provide a carefully designed list decoding algorithm for the class of Gaussians which results

in a much better sample complexity for privately learning GMMs due to its mild dependence on the accuracy parameter $(1/\alpha)$.

We reduce the problem of list decoding a class of Gaussians to the problem of compressing this class. Next, we use the result of Ashtiani *et al.* (2018a) that the class of Gaussians is compressible. Finally, we conclude that this class is list decodable.

Remark 7.1.1. *Our reduction is fairly general and works for any class of distributions. Informally, if a class of distributions is compressible, then it is list decodable. However, for the sake of simplicity we state this result only for the class of Gaussians in Lemma 7.1.4.*

The method of sample compression schemes introduced by Ashtiani *et al.* (2018a) is used for distribution learning. At a high level, given m samples from a distribution and t additional bits, if there exists an algorithm (i.e., decoder) that can approximately recover the original distribution given a small subset of (i.e., τ many) samples and t bits, then one can create a list of all possible combinations of choosing τ samples and t bits. Then one can pick a "good" distribution from the generated list of candidates using Minimum Distance Estimator (Yatracos, 1985). Below, we provide the formal definition of sample compression schemes for learning distributions.

Definition 7.1.2 (Compression schemes for distributions (Ashtiani *et al.*, 2018a))**.** *For a set of functions $\tau, m, t : (0,1) \to \mathbb{Z}_{\geq 0}$, the class of distributions $\mathcal{F}$ is said to be (τ, t, m)-compressible if there exists a decoder $\mathcal{D}$ such that for any $f \in \mathcal{F}$ and any $\alpha \in (0,1)$, after receiving an i.i.d sample $\mathcal{S}$ of size $m(\alpha)\log(1/\beta)$ from f, with probability at least $1-\beta$, there exists a sequence L of at most $\tau(\alpha)$ members of $\mathcal{S}$ and a sequence B of at most $t(\alpha)$ bits, such that $\mathrm{d_{TV}}(D(L,B), f) \leq \alpha$.*

The following result states that the class of Gaussians is compressible.

Lemma 7.1.3 (Lemma 4.1 of Ashtiani *et al.* (2018a))**.** *Let $\alpha \in (0,1)$. The class $\mathcal{G}$ is $(O(d), \tilde{O}(d^2 \log(1/\alpha)), O(d))$-compressible.*

Finally, the next lemma uses the above result to show that Gaussians are list decodable.

Lemma 7.1.4. *For any $\alpha, \beta \in (0,1)$, the class $\mathcal{G}$ is $(L, m, \alpha, \beta, 0)$-list-decodable w.r.t. $\mathrm{d_{TV}}$ for $L = (d \log(1/\beta))^{\tilde{O}(d^2 \log(1/\alpha))}$, and $m = O(d \log(1/\beta))$.*

An important feature of the above lemma (that is inherited from the compression result) is the mild dependence of m and L on $1/\alpha$: m does not depend on $1/\alpha$ and L has a mild polynomial dependence on it.

Proof. We will prove that if a class of distributions $\mathcal{F}$ is (τ, t, s)-compressible, then it is
$(O((s \log(1/\beta))^{t+\tau}), s \log(1/\beta), \alpha, \beta, 0)$-list-decodable w.r.t. $\mathrm{d_{TV}}$. Let $f \in \mathcal{F}$, and $\mathcal{S}$ be a set of $s \log(1/\beta)$ i.i.d. samples drawn from f. Now using the decoder D, construct the list $\mathcal{L} = \{D(L, B) : L \subseteq \mathcal{S}, |L| \leq \tau, B \in \{0,1\}^t\}$. Using the Definition 7.1.2, with probability at least $1 - \beta$ there exists $\hat{f} \in \mathcal{L}$ such that $\mathrm{d_{TV}}(\hat{f}, f) \leq \alpha$. Note that $|\mathcal{L}| = O((s \log(1/\beta))^{(t+\tau)})$. Putting together with Lemma 7.1.3 concludes the result. □

7.2 A locally small cover for Gaussians

In this section, we construct a locally small cover for the class of Gaussians using the techniques of Aden-Ali *et al.* (2021a). Explicitly constructing a locally small cover for Gaussians is a challenging task due to the complex geometry of this class. Previously, Aden-Ali *et al.* (2021a) constructed locally small covers for location Gaussians

(Gaussians with I_d covariance matrix) and scale Gaussians (zero mean Gaussians). One might think by taking the product of these two covers we can simply construct a locally small cover for unbounded Gaussians (see Definition 2.1.5). However, the product of these two covers is not a valid cover for unbounded Gaussians as the $\mathrm{d_{TV}}$ between two Gaussians can be very large even if their means are close to each other in euclidean distance.

To resolve this issue, we take a small cover for a small $\mathrm{d_{TV}}$-ball of location gaussians around $N(0, I_d)$ and scale it using a small cover for the $\mathrm{d_{TV}}$-ball of scale gaussians around $N(0, I_d)$. Showing that this is a valid and small cover for a small $\mathrm{d_{TV}}$-ball of unbounded Gaussians around $N(0, I_d)$ is a delicate matter. To argue that this is a valid cover, we use the upper bound of Devroye *et al.* (2018) for the $\mathrm{d_{TV}}$ distance between high-dimensional Gaussians. Showing that this is a small cover requires the recently proved lower bound on the $\mathrm{d_{TV}}$ distance of high-dimensional Gaussians (Arbas *et al.*, 2023).

Once we created a small cover for a $\mathrm{d_{TV}}$-ball of unbounded Gaussians around $N(0, I_d)$, we can transform it to a small cover for a $\mathrm{d_{TV}}$-ball of unbounded Gaussians around an arbitrary $N(\mu, \Sigma)$. Finally, we use Lemma 7.2.7 to show that there exists a global locally small cover for the whole space of Guassians.

Definition 7.2.1 (Location Gaussians)**.** *Let* $\mathcal{G}^L = \{\mathcal{N}(\mu, I_d) : \mu \in \mathbb{R}^d\}$ *be the class of d-dimensional location Gaussians.*

Definition 7.2.2 (Scale Gaussians)**.** *Let* $\mathcal{G}^S = \{\mathcal{N}(0, \Sigma) : \Sigma \in S^d\}$ *be the class of d-dimensional scale Gaussians.*

The next two lemmas propose small covers for location and scale Gaussians near $N(0, I)$. Recall that $B_{\mathrm{d_{TV}}}(r, f, \mathcal{F})$ stands for the $\mathrm{d_{TV}}$-ball of radius r around f (See

Definition 2.1.1).

Lemma 7.2.3 (Lemma 30 of Aden-Ali *et al.* (2021a)). *For any $0 < \alpha < \gamma < c$, where c is a universal constant, there exists an α-cover C^L for the set of distributions $B_{\mathrm{d_{TV}}}(\gamma, N(0, I), \mathcal{G}^L)$ of size $(\frac{\gamma}{\alpha})^{O(d)}$ w.r.t. $\mathrm{d_{TV}}$.*

Lemma 7.2.4 (Corollary 33 of Aden-Ali *et al.* (2021a)). *For any $0 < \alpha < \gamma < c$, and $\Sigma \in S^d$, where c is a universal constant, there exists an α-cover C^S for the set of distributions $B_{\mathrm{d_{TV}}}(\gamma, N(0, \Sigma), \mathcal{G}^S)$ of size $(\frac{\gamma}{\alpha})^{O(d^2)}$ w.r.t. $\mathrm{d_{TV}}$.*

The next theorem provides upper and lower bounds for $\mathrm{d_{TV}}$ between two Gaussians, which will be used for constructing a small cover for unbounded Guassians near $N(0, I)$.

Theorem 7.2.5 (Theorem 1.8 of Arbas *et al.* (2023)). *Let $N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2) \in \mathcal{G}$, and $\Delta = \max\{||\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - I_d||_F, ||\Sigma_1^{-1/2}(\mu_1 - \mu_2)||_2\}$. Then*

$$\mathrm{d_{TV}}(N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)) \leq \frac{1}{\sqrt{2}}\Delta.$$

Also, if $\mathrm{d_{TV}}(N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)) \leq \frac{1}{600}$, we have:

$$\frac{1}{200}\Delta \leq \mathrm{d_{TV}}(N(\mu_1, \Sigma_1), N(\mu_2, \Sigma_2)).$$

The next lemma proposes a small cover for the unbounded Guassians near $N(\mu, \Sigma)$ for any given μ and Σ. To do so, we combine the small covers from Lemma 7.2.3 and Lemma 7.2.4 in a way that it approximates any Gaussian near $N(0, I)$.

Lemma 7.2.6. *Let $0 < \alpha < \gamma \leq \frac{1}{600}$, $\mu \in \mathbb{R}^d$, and $\Sigma \in S^d$. There exists an α-cover for $B_{\mathrm{d_{TV}}}(\gamma, N(\mu, \Sigma), \mathcal{G})$ of size at most $(\frac{\gamma}{\alpha})^{O(d^2)}$.*

Proof. First, we construct a cover for $B_{\mathrm{d_{TV}}}(\gamma, N(0,I), \mathcal{G})$, then we extend it to a cover for $B_{\mathrm{d_{TV}}}(\gamma, N(\mu,\Sigma), \mathcal{G})$ using a linear transformation.

Let $\gamma \in (\alpha, \frac{1}{600})$ and consider the ball $\mathcal{B} := B_{\mathrm{d_{TV}}}(\gamma, N(0,I), \mathcal{G})$. Let $\gamma_1 = 200\gamma$ and C^L be an $\frac{\sqrt{2}}{200}\alpha$-cover for $B_{\mathrm{d_{TV}}}(\gamma_1, N(0,I), \mathcal{G}^L)$ from Lemma 7.2.3. Also, let $\gamma_2 = 200\gamma$ and C^S be an $\frac{\sqrt{2}}{200}\alpha$-cover for $B_{\mathrm{d_{TV}}}(\gamma_2, N(0,I), \mathcal{G}^S)$ from Lemma 7.2.4.

We claim that $C^B = \{N(\Sigma^{1/2}\mu, \Sigma) : N(\mu, I) \in C^L, N(0,\Sigma) \in C^S\}$ is an α-cover for $\mathcal{B}$. Let $N(\hat{\mu}, \hat{\Sigma}) \in \mathcal{B}$ so that $\mathrm{d_{TV}}(N(0,I), N(\hat{\mu}, \hat{\Sigma})) \leq \gamma \leq \frac{1}{600}$. Applying the lower bound of Theorem 7.2.5 with $\Sigma_1 = I, \Sigma_2 = \hat{\Sigma}, \mu_1 = 0, \mu_2 = \hat{\mu}$ gives that

$$||\hat{\Sigma} - I||_F \leq 200\, \mathrm{d_{TV}}(N(0,I), N(\hat{\mu}, \hat{\Sigma})) \leq 200\gamma \quad \text{and} \tag{7.2.1}$$

$$||\hat{\mu}||_2 \leq 200\, \mathrm{d_{TV}}(N(0,I), N(\hat{\mu}, \hat{\Sigma})) \leq 200\gamma. \tag{7.2.2}$$

Moreover, applying Theorem 7.2.5 with $\Sigma_1 = \hat{\Sigma}, \Sigma_2 = I, \mu_1 = \hat{\mu}, \mu_2 = 0$ gives that

$$||\hat{\Sigma}^{-1/2}\hat{\mu}||_2 \leq 200\, \mathrm{d_{TV}}(N(0,I), N(\hat{\mu}, \hat{\Sigma})) \leq 200\gamma.$$

Next, applying the upper bound of Theorem 7.2.5 with $\mu_1 = \mu_2 = 0, \Sigma_1 = I, \Sigma_2 = \hat{\Sigma}$ gives $\mathrm{d_{TV}}(N(0,I), N(0,\hat{\Sigma})) \leq \frac{1}{\sqrt{2}}||\hat{\Sigma} - I||_F \leq \frac{200}{\sqrt{2}}\gamma < 200\gamma = \gamma_2$. Therefore $N(0,\hat{\Sigma}) \in B_{\mathrm{d_{TV}}}(\gamma_2, N(0,1), \mathcal{G}^S)$. Recall that C^S is an $\frac{\sqrt{2}}{200}\alpha$-cover for $B_{\mathrm{d_{TV}}}(\gamma_2, N(0,I), \mathcal{G}^S)$. Thus, there exists $N(0,\tilde{\Sigma}) \in C^S$ such that $\mathrm{d_{TV}}(N(0,\hat{\Sigma}), N(0,\tilde{\Sigma})) \leq \frac{\sqrt{2}}{200}\alpha$. Using the lower bound of Theorem 7.2.5 with $\Sigma_1 = \tilde{\Sigma}, \Sigma_2 = \hat{\Sigma}, \mu_1 = 0, \mu_2 = 0$ results in

$$||\tilde{\Sigma}^{-1/2}\hat{\Sigma}\tilde{\Sigma}^{-1/2} - I||_2 \leq ||\tilde{\Sigma}^{-1/2}\hat{\Sigma}\tilde{\Sigma}^{-1/2} - I||_F \leq 200\, \mathrm{d_{TV}}(N(0,\hat{\Sigma}), N(0,\tilde{\Sigma})) \leq \sqrt{2}\alpha. \tag{7.2.3}$$

Therefore $||\tilde{\Sigma}^{-1/2}\hat{\Sigma}\tilde{\Sigma}^{-1/2}||_2 = ||(\tilde{\Sigma}^{-1/2}\hat{\Sigma}^{1/2})(\tilde{\Sigma}^{-1/2}\hat{\Sigma}^{1/2})^T||_2 \leq 1 + \sqrt{2}\alpha$. Finally, we

get $||\tilde{\Sigma}^{-1/2}\hat{\Sigma}^{1/2}||_2 \leq \sqrt{1+\sqrt{2}\alpha} \leq \sqrt{2}$.

Now let $\hat{v} = \hat{\Sigma}^{-1/2}\hat{\mu}$. From 7.2.2 we know that $||\hat{v}||_2 \leq 200\gamma$. Therefore we have $||\tilde{\Sigma}^{-1/2}\hat{\mu}||_2 = ||\tilde{\Sigma}^{-1/2}\hat{\Sigma}^{1/2}\hat{v}||_2 \leq ||\tilde{\Sigma}^{-1/2}\hat{\Sigma}^{1/2}||_2||\hat{v}||_2 \leq 200\gamma\sqrt{2}$. Using the upper bound of Theorem 7.2.5 with $\Sigma_1 = I$, $\Sigma_2 = I$, $\mu_1 = \tilde{\Sigma}^{-1/2}\hat{\mu}$, and $\mu_2 = 0$ gives $\mathrm{d_{TV}}(N(\tilde{\Sigma}^{-1/2}\hat{\mu}, I), N(0, I)) \leq \frac{1}{\sqrt{2}}||\tilde{\Sigma}^{-1/2}\hat{\mu}||_2 \leq \frac{200\gamma\sqrt{2}}{\sqrt{2}} = \gamma_1$. Thus $N(\tilde{\Sigma}^{-1/2}\hat{\mu}, I) \in B_{\mathrm{d_{TV}}}(\gamma_1, N(0,1), \mathcal{G}^L)$. Recall that C^L is an $\frac{\sqrt{2}}{200}\alpha$-cover for $B_{\mathrm{d_{TV}}}(\gamma_1, N(0, I), \mathcal{G}^L)$. Therefore, there exists $N(\tilde{\mu}, I) \in C^L$ such that $\mathrm{d_{TV}}(N(\tilde{\mu}, I), N(\tilde{\Sigma}^{-1/2}\hat{\mu}, I)) \leq \frac{\sqrt{2}}{200}\alpha$. Using the lower bound of Theorem 7.2.5 with $\Sigma_1 = I, \Sigma_2 = I, \mu_1 = \tilde{\Sigma}^{-1/2}\hat{\mu}, \mu_2 = \tilde{\mu}$, we can write $||\tilde{\Sigma}^{-1/2}\hat{\mu} - \tilde{\mu}||_2 = ||\tilde{\Sigma}^{-1/2}(\tilde{\Sigma}^{1/2}\tilde{\mu} - \hat{\mu})||_2 \leq \sqrt{2}\alpha$. Putting together with 7.2.3, we can use the upper bound in Theorem 7.2.5 with $\Sigma_1 = \tilde{\Sigma}$, $\Sigma_2 = \hat{\Sigma}$, $\mu_1 = \tilde{\Sigma}^{1/2}\tilde{\mu}$, and $\mu_2 = \hat{\mu}$ to get that $\mathrm{d_{TV}}(N(\tilde{\Sigma}^{1/2}\tilde{\mu}, \tilde{\Sigma}), N(\hat{\mu}, \hat{\Sigma})) \leq \alpha$. Note that $N(\tilde{\Sigma}^{1/2}\tilde{\mu}, \tilde{\Sigma}) \in C^B$. Hence, C^B is an α-cover for $\mathcal{B}$. Moreover, we have $|C^B| = |C^L||C^S| \leq (\frac{\gamma_1}{\alpha})^{O(d)}(\frac{\gamma_2}{\alpha})^{O(d^2)} = (\frac{\gamma}{\alpha})^{O(d^2)}$.

Now, we propose a cover for $B_{\mathrm{d_{TV}}}(\gamma, N(\mu, \Sigma), \mathcal{G})$. Note that using Lemma A.0.1, for any $\Sigma, \Sigma_1, \Sigma_2 \in S^d$, we have:

$$\mathrm{d_{TV}}(N(0, \Sigma^{1/2}\Sigma_1\Sigma^{1/2}), N(0, \Sigma^{1/2}\Sigma_2\Sigma^{1/2})) = \mathrm{d_{TV}}(N(0, \Sigma_1), N(0, \Sigma_2)).$$

Note that equality holds since the mapping is bijection. Next, create the set $C^{B_\Sigma} = \{N(\mu, \Sigma^{1/2}\Sigma'\Sigma^{1/2}) : N(\mu, \Sigma') \in C^B\}$. Note that C^{B_Σ} is an α-cover for $B_{\mathrm{d_{TV}}}(\gamma, N(0, \Sigma), \mathcal{G})$ since the $\mathrm{d_{TV}}$ distance between every two distributions in C^B remains same (i.e. does not increase) after this transformation. Finally, the set $C^{B_{\mu,\Sigma}} = \{N(\mu + \mu', \Sigma') : N(\mu', \Sigma') \in C^{B_\Sigma}\}$ is the desired α-cover for $B_{\mathrm{d_{TV}}}(\gamma, N(\mu, \Sigma), \mathcal{G})$ since it is the shifted version of C^{B_Σ}. Also, we have $|C^{B_{\mu,\Sigma}}| = |C^{B_\Sigma}| = |C^B| \leq (\frac{\gamma}{\alpha})^{O(d^2)}$. □

The next lemma provides a useful tool for creating (global) locally small covers. Informally, given a class of distributions, if there exists a small cover for a small ball around each distribution in the class, then there exists a (global) locally small cover for the whole class.

Lemma 7.2.7 (Lemma 29 of Aden-Ali *et al.* (2021a))**.** *Consider a class of distributions $\mathcal{F}$ and let $0 < \alpha < \gamma < 1$. If for every $f \in \mathcal{F}$ the $B_{\mathrm{d_{TV}}}(\gamma, f, \mathcal{F})$ has an α-cover of size no more than t, then there exists a (t, γ)-locally small 2α-cover for $\mathcal{F}$ w.r.t. $\mathrm{d_{TV}}$.*

The proof of the above lemma is non-constructive and uses Zorn's lemma. An immediate implication of Lemma 7.2.6 and Lemma 7.2.7 is the existence of a locally small cover for unbounded Gaussians.

Lemma 7.2.8. *For any $0 < \alpha < \gamma \leq \frac{1}{600}$, there exists a $((2\gamma/\alpha)^{O(d^2)}, \gamma)$-locally small α-cover for the class $\mathcal{G}$ w.r.t. $\mathrm{d_{TV}}$.*

7.3 Learning GMMs

In this section, we prove the first sample complexity upper bound for privately learning mixtures of unbounded Gaussians. We use the fact that Gaussians are *(1)* list decodable, and *(2)* admit a locally small cover. Putting this together with our main theorem for learning mixture distributions, we conclude that GMMs are privately learnable.

Proof. According to Lemma 7.1.4, the class $\mathcal{G}$ is $(L, m, \alpha, \beta, 0)$-list-decodable w.r.t. $\mathrm{d_{TV}}$ for $L = (d\log(1/\beta))^{\tilde{O}(d^2\log(1/\alpha))}$ and $m = O(d\log(1/\beta))$. Moreover, Lemma 7.2.8

implies that $\mathcal{G}$ admits a $(t, 2\alpha)$-locally small α-cover, where $t = 4^{O(d^2)}$. Using Theorem 2.2.1, we get that k-mix$(\mathcal{G})$ is (ε, δ)-DP (α, β)-PAC-learnable using

$$\begin{aligned}
&\tilde{O}\left(\left(\frac{\log(1/\delta) + k\log(tL)}{\varepsilon} + \frac{mk + k\log(1/\beta)}{\alpha\varepsilon}\right) \cdot \left(\frac{k\log(L)}{\alpha^2} + \frac{mk + k\log(1/\beta)}{\alpha^3}\right)\right) \\
&= \tilde{O}\left(\left(\frac{\log(1/\delta) + kd^2}{\varepsilon} + \frac{kd\log(1/\beta)}{\alpha\varepsilon}\right) \cdot \left(\frac{kd^2}{\alpha^2} + \frac{kd\log(1/\beta)}{\alpha^3}\right)\right) \\
&= \tilde{O}\left(\frac{kd^2\log(1/\delta) + k^2d^4}{\alpha^2\varepsilon} + \frac{kd\log(1/\delta)\log(1/\beta) + k^2d^3\log(1/\beta)}{\alpha^3\varepsilon} + \frac{k^2d^2\log^2(1/\beta)}{\alpha^4\varepsilon}\right)
\end{aligned}$$

samples. □

Chapter 8

Conclusion

In this work, we proposed the first sample complexity upper bound for privately learning unbounded GMMs. We demonstrated that $\tilde{O}(k^2d^4)$ samples are sufficient to estimate a mixture of k Gaussians in $\mathbb{R}^d$ up to a small total variation distance while satisfying approximate differential privacy. We achieve this through a fairly general reduction, which posits that if a class of distributions, such as Gaussians, is (1) list-decodable (or simply learnable) and (2) admits a locally small cover, then the class of its mixtures is privately learnable.

We also propose and address the problem of common member selection for general locally small metric spaces. We believe this has future applications and can be utilized in various private estimation tasks in a black-box manner.

There are still some important open directions related to this work, which we discuss here.

Firstly, we are not aware of any class of distributions that is learnable in the non-private (agnostic) setting but not learnable in the (ϵ, δ)-DP setting. In fact, we conjecture that every class of distributions that is learnable in the non-private

agnostic setting is also learnable in the (ϵ, δ)-DP setting. We believe this conjecture to be true for the agnostic setting, given the recent developments on connections between robustness and privacy Asi *et al.* (2023); Hopkins *et al.* (2023). Recently, the negative result of Bun *et al.* (2024) has shown that there is a class of distribution that is learnable (in realizable setting) with a constant accuracy but not privately learnable.

Another important open question is whether it is possible to privately and robustly learn GMMs simultaneously.

Furthermore, there is a gap between the sample complexity of private versus non-private learning of GMMs. For the non-private case, the nearly tight sample complexity is known to be $\tilde{\Theta}(kd^2)$. This remains an open problem, particularly in the one-dimensional setting (Aden-Ali *et al.*, 2021b).

Lastly, as stated before, our result is information-theoretic. It remains an open question to design a finite-time algorithm for the problem of privately learning GMMs. Designing a computationally efficient algorithm, even in the non-private setting, remains an open problem (Diakonikolas *et al.*, 2017).

Appendix A

Additional Facts

Lemma A.0.1. *Let f be an arbitrary function, and X, Y be two random variables with the same support. Then $\mathrm{d_{TV}}(f(X), f(Y)) \leq \mathrm{d_{TV}}(X, Y)$.*

Proof.

$$\begin{aligned}\mathrm{d_{TV}}(f(X), f(Y)) &= \sup_{A \in \mathcal{X}} \mathbb{P}\left[f(X) \in A\right] - \mathbb{P}\left[f(Y) \in A\right] \\ &= \sup_{A \in \mathcal{X}} \mathbb{P}\left[X \in f^{-1}(A)\right] - \mathbb{P}\left[Y \in f^{-1}(A)\right] \leq \mathrm{d_{TV}}(X, Y)\end{aligned}$$

□

Claim A.0.2. *Let $x \geq 1$. Then $1 + \frac{\log 2}{x} + \frac{\log x}{x} < 2$.*

Proof. Let $f(x) = 1 + \frac{\log 2}{x} + \frac{\log x}{x}$. Then $f'(x) = -\frac{\log 2}{x^2} + \frac{1 - \log x}{x^2} = \frac{1 - \log(2x)}{x^2}$. Note that $f'(x)$ is decreasing so f is concave. In addition, $x = e/2$ is the only root of f' so f is maximized at $e/2$. Thus, $f(x) \leq f(e/2) = 1 + \frac{2}{e} < 2$. □

Claim A.0.3. *For $c_1, c_2, 1/\beta \geq 1$, let $\beta' = \frac{\beta}{2ec_1 \log(ec_1c_2/\beta)}$, then $\beta' \leq \frac{\beta}{2ec_1} \leq \frac{1}{ec_1}$, and $c_1\beta' \log(c_2/\beta') \leq \beta$.*

Proof.

$$\begin{aligned} c_1\beta' \log(c_2/\beta') &= c_1 \frac{\beta}{2ec_1 \log(ec_1c_2/\beta)} \cdot [\log(ec_1c_2/\beta) + \log(2) + \log\log(ec_1c_2/\beta)] \\ &= \frac{\beta}{2e} \cdot \left[1 + \frac{\log 2}{\log(ec_1c_2/\beta)} + \frac{\log\log(ec_1c_2/\beta)}{\log(ec_1c_2/\beta)}\right] \\ &\leq \beta/e. \end{aligned}$$

where in the last inequality, we used Claim A.0.2 with $x = \log(ec_1c_2/\beta) \geq 1$. □

Bibliography

Acharya, J., Diakonikolas, I., Li, J., and Schmidt, L. (2017). Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1278–1289. SIAM.

Acharya, J., Sun, Z., and Zhang, H. (2021). Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR.

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

Aden-Ali, I., Ashtiani, H., and Kamath, G. (2021a). On the sample complexity of privately learning unbounded high-dimensional gaussians. In *Algorithmic Learning Theory*, pages 185–216. PMLR.

Aden-Ali, I., Ashtiani, H., and Liaw, C. (2021b). Privately learning mixtures of axis-aligned gaussians. *Advances in Neural Information Processing Systems*, **34**, 3925–3938.

Afzali, M., Ashtiani, H., and Liaw, C. (2023). Mixtures of gaussians are privately learnable with a polynomial number of samples. *arXiv preprint arXiv:2309.03847*.

Alabi, D., Kothari, P. K., Tankala, P., Venkat, P., and Zhang, F. (2023). Privately estimating a gaussian: Efficient, robust, and optimal. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 483–496.

Alon, N., Livni, R., Malliaris, M., and Moran, S. (2019). Private pac learning implies finite littlestone dimension. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 852–860.

Anderson, J., Belkin, M., Goyal, N., Rademacher, L., and Voss, J. (2014). The more, the merrier: the blessing of dimensionality for learning large gaussian mixtures. In *Conference on Learning Theory*, pages 1135–1164. PMLR.

Arbas, J., Ashtiani, H., and Liaw, C. (2023). Polynomial time and private learning of unbounded gaussian mixture models. In *International Conference on Machine Learning*. PMLR.

Ashtiani, H. and Liaw, C. (2022). Private and polynomial time algorithms for learning gaussians and beyond. In *Conference on Learning Theory*, pages 1075–1076. PMLR.

Ashtiani, H. and Mehrabian, A. (2018). Some techniques in density estimation. *arXiv preprint arXiv:1801.04003*.

Ashtiani, H., Ben-David, S., Harvey, N., Liaw, C., Mehrabian, A., and Plan, Y. (2018a). Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. *Advances in Neural Information Processing Systems*, **31**.

Ashtiani, H., Ben-David, S., and Mehrabian, A. (2018b). Sample-efficient learning

of mixtures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ashtiani, H., Ben-David, S., Harvey, N. J., Liaw, C., Mehrabian, A., and Plan, Y. (2020). Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *Journal of the ACM (JACM)*, **67**(6), 1–42.

Asi, H., Ullman, J., and Zakynthinou, L. (2023). From robustness to privacy and back. *arXiv preprint arXiv:2302.01855*.

Bakshi, A., Diakonikolas, I., Jia, H., Kane, D. M., Kothari, P. K., and Vempala, S. S. (2022). Robustly learning mixtures of k arbitrary gaussians. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1234–1247.

Belkin, M. and Sinha, K. (2009). Learning gaussian mixtures with arbitrary separation. *arXiv preprint arXiv:0907.1054*.

Belkin, M. and Sinha, K. (2010). Polynomial learning of distribution families. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 103–112. IEEE.

Ben-David, S., Bie, A., Canonne, C. L., Kamath, G., and Singhal, V. (2023). Private distribution learning with public data: The view from sample compression. *arXiv preprint arXiv:2308.06239*.

Bie, A., Kamath, G., and Singhal, V. (2022). Private estimation with public data. *Advances in Neural Information Processing Systems*, **35**, 18653–18666.

Biswas, S., Dong, Y., Kamath, G., and Ullman, J. (2020). Coinpress: Practical private mean and covariance estimation. *Advances in Neural Information Processing Systems*, **33**, 14475–14485.

Brown, G., Gaboardi, M., Smith, A., Ullman, J., and Zakynthinou, L. (2021). Covariance-aware private mean estimation without private covariance estimation. *Advances in Neural Information Processing Systems*, **34**, 7950–7964.

Brown, G., Hopkins, S., and Smith, A. (2023). Fast, sample-efficient, affine-invariant private mean and covariance estimation for subgaussian distributions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5578–5579. PMLR.

Brubaker, S. C. and Vempala, S. S. (2008). Isotropic pca and affine-invariant clustering. *Building Bridges: Between Mathematics and Computer Science*, pages 241–281.

Bun, M., Nissim, K., Stemmer, U., and Vadhan, S. (2015). Differentially private release and learning of threshold functions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 634–649. IEEE.

Bun, M., Nissim, K., and Stemmer, U. (2016). Simultaneous private learning of multiple concepts. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 369–380.

Bun, M., Dwork, C., Rothblum, G. N., and Steinke, T. (2018). Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–86.

Bun, M., Livni, R., and Moran, S. (2020). An equivalence between private classification and online prediction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 389–402. IEEE.

Bun, M., Kamath, G., Steinke, T., and Wu, Z. S. (2021). Private hypothesis selection. *IEEE Transactions on Information Theory*, **67**(3).

Bun, M., Kamath, G., Mouzakis, A., and Singhal, V. (2024). Not all learnable distribution classes are privately learnable. *arXiv preprint arXiv:2402.00267*.

Chan, S.-O., Diakonikolas, I., Servedio, R. A., and Sun, X. (2014). Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 604–613.

Chen, H., Cohen-Addad, V., d'Orsi, T., Epasto, A., Imola, J., Steurer, D., and Tiegel, S. (2023). Private estimation algorithms for stochastic block models and mixture models. *arXiv preprint arXiv:2301.04822*.

Cohen, E., Kaplan, H., Mansour, Y., Stemmer, U., and Tsfadia, E. (2021). Differentially-private clustering of easy instances. In *International Conference on Machine Learning*, pages 2049–2059. PMLR.

Cohen, E., Lyu, X., Nelson, J., Sarlós, T., and Stemmer, U. (2023). Optimal differentially private learning of thresholds and quasi-concave optimization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 472–482.

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

Devroye, L. and Lugosi, G. (2001). *Combinatorial methods in density estimation*. Springer Science & Business Media.

Devroye, L., Mehrabian, A., and Reddad, T. (2018). The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*.

Diakonikolas, I. (2016). Learning structured distributions. *Handbook of Big Data*, **267**, 10–1201.

Diakonikolas, I., Kane, D. M., and Stewart, A. (2017). Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE.

Duchi, J., Haque, S., and Kuditipudi, R. (2023). A fast algorithm for adaptive private mean estimation. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR.

Dwork, C. and Lei, J. (2009). Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006a). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. (2006b). Our data, ourselves: Privacy via distributed noise generation. In *Advances in*

Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25, pages 486–503. Springer.

Feldman, J., Servedio, R. A., and O'Donnell, R. (2006). Pac learning axis-aligned mixtures of gaussians with no separation assumption. In *Proceedings of the 19th annual conference on Learning Theory*, pages 20–34.

Georgiev, K. and Hopkins, S. (2022). Privacy induces robustness: Information-computation gaps and sparse mean estimation. *Advances in Neural Information Processing Systems*, **35**, 6829–6842.

Hardt, M. and Price, E. (2014). Sharp bounds for learning a mixture of two gaussians.

Hopkins, S. B. and Li, J. (2018). Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034.

Hopkins, S. B., Kamath, G., and Majid, M. (2022). Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1406–1417.

Hopkins, S. B., Kamath, G., Majid, M., and Narayanan, S. (2023). Robustness implies privacy in statistical estimation. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 497–506.

Hsu, D. and Kakade, S. M. (2013). Learning mixtures of spherical gaussians: moment

methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer.

Kalai, A. T., Moitra, A., and Valiant, G. (2010). Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562.

Kamath, G., Sheffet, O., Singhal, V., and Ullman, J. (2019a). Differentially private algorithms for learning mixtures of separated gaussians. *Advances in Neural Information Processing Systems*, **32**.

Kamath, G., Li, J., Singhal, V., and Ullman, J. (2019b). Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902. PMLR.

Kamath, G., Mouzakis, A., and Singhal, V. (2022a). New lower bounds for private estimation and a generalized fingerprinting lemma. *Advances in Neural Information Processing Systems*, **35**, 24405–24418.

Kamath, G., Mouzakis, A., Singhal, V., Steinke, T., and Ullman, J. (2022b). A private and computationally-efficient estimator for unbounded gaussians. In *Conference on Learning Theory*, pages 544–572. PMLR.

Kaplan, H., Ligett, K., Mansour, Y., Naor, M., and Stemmer, U. (2020). Privately learning thresholds: Closing the exponential gap. In *Conference on Learning Theory*, pages 2263–2285. PMLR.

Karwa, V. and Vadhan, S. (2018). Finite sample differentially private confidence intervals. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Kothari, P., Manurangsi, P., and Velingker, A. (2022). Private robust estimation by stabilizing convex relaxations. In *Conference on Learning Theory*, pages 723–777. PMLR.

Kothari, P. K., Steinhardt, J., and Steurer, D. (2018). Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046.

Lechner, T. *et al.* (2023). Impossibility of characterizing distribution learning–a simple solution to a long-standing problem. *arXiv preprint arXiv:2304.08712*.

Li, J. and Schmidt, L. (2017). Robust and proper learning for mixtures of gaussians via systems of polynomial inequalities. In *Conference on Learning Theory*, pages 1302–1382. PMLR.

Liu, A. and Li, J. (2022). Clustering mixtures with almost optimal separation in polynomial time. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1248–1261.

Liu, A. and Moitra, A. (2021). Settling the robust learnability of mixtures of gaussians. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 518–531.

Liu, A. and Moitra, A. (2022). Learning gmms with nearly optimal robustness guarantees. In *Conference on Learning Theory*, pages 2815–2895. PMLR.

Liu, A., Li, J., and Moitra, A. (2022a). Robust model selection and nearly-proper learning for gmms. *Advances in Neural Information Processing Systems*, **35**, 22830–22843.

Liu, X., Kong, W., Kakade, S., and Oh, S. (2021). Robust and differentially private mean estimation. *Advances in neural information processing systems*, **34**, 3887–3901.

Liu, X., Kong, W., and Oh, S. (2022b). Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, pages 1167–1246. PMLR.

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE.

Mitzenmacher, M. and Upfal, E. (2005). *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press.

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

Nissim, K., Raskhodnikova, S., and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84.

Regev, O. and Vijayaraghavan, A. (2017). On learning mixtures of well-separated gaussians. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 85–96. IEEE.

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

Tsfadia, E., Cohen, E., Kaplan, H., Mansour, Y., and Stemmer, U. (2022). Friendlycore: Practical differentially private aggregation. In *International Conference on Machine Learning*, pages 21828–21863. PMLR.

Vempala, S. and Wang, G. (2004). A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, **68**(4), 841–860.

Wu, X. and Xie, C. (2018). Improved algorithms for properly learning mixture of gaussians. In *National Conference of Theoretical Computer Science*, pages 8–26. Springer.

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