

Mixtures of Gaussians are Privately Learnable with a Polynomial Number of Samples

Mohammad Afzali

McMaster University

AFZALIKM@MCMASTER.CA

Hassan Ashtiani*

McMaster University

ZOKAEIAM@MCMASTER.CA

Christopher Liaw

Google

CVLIAW@GOOGLE.COM

Editors: Claire Vernade and Daniel Hsu

Abstract

We study the problem of estimating mixtures of Gaussians under the constraint of differential privacy (DP). Our main result is that $\text{poly}(k, d, 1/\alpha, 1/\epsilon, \log(1/\delta))$ 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 solve the problem, we devise a new framework which may be useful for other tasks. 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).

Keywords: Differential Privacy, Distribution Learning, Gaussian Mixture Models

1. Introduction

Density estimation—also known as distribution learning—is the fundamental task of estimating a distribution given samples generated from it. In the most commonly studied setting, the data points are assumed to be generated independently from an unknown distribution f and the goal is to find a distribution \hat{f} that is close to f with respect to the total variation (TV) distance. Assuming that f belongs to (or is close to a member of) a class of distributions \mathcal{F} , an important question is to characterize the number of samples that is required to guarantee that with high probability \hat{f} is close to f in TV distance.

There is a large body of work on characterizing the optimal sample complexity (or the related minimax error rate) of learning various classes of distributions (see Diakonikolas (2016); Devroye and Lugosi (2001); Ashtiani and Mehrabian (2018) for an overview). Nevertheless, determining the sample complexity (or even learnability) of a general class of distributions remains an important open problem (e.g., Open Problem 15.1 in Diakonikolas (2016)).

We study the problem of density estimation under the constraint of differential privacy (Dwork et al., 2006b). Intuitively, differential privacy (DP) requires that changing a single individual’s data does not identifiably alter the outcome of the estimation. There are various formulations of DP. The original pure DP (ϵ -DP) formulation can be somewhat restrictive, as learning some simple classes

* Hassan Ashtiani is also affiliated with Vector Institute, and was supported by an NSERC Discovery Grant.

of distributions (such as univariate Gaussians with unbounded mean) is impossible in this model. An alternative formulation is approximate DP (Dwork et al., 2006a), 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 setting but not learnable in the (ϵ, δ) -DP setting. This is in sharp contrast with other areas of learning such as classification². 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.

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).

In the non-private setting, GMMs with k components in d dimensions are known to be learnable to within total variation distance α with a polynomial number of samples (in terms of $d, k, 1/\alpha$). Nearly tight sample complexity bound of $\tilde{O}(d^2 k / \alpha^2)$ 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?

Indeed, Aden-Ali et al. (2021b) have shown that univariate GMMs are learnable in the (ϵ, δ) -DP setting. 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 GMMs, their approach does not work for GMMs with general covariance matrices³.

Another 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 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.

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) and Tsfadia et al. (2022). In particular, Ashtiani and Liaw (2022) and Tsfadia et al. (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).

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

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 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.

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, 2020, 2018b) 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 a non-private GMM learner as a blackbox since its output can be “unstable”.

In this paper, we bypass the above barriers and show that GMMs are privately learnable with a polynomial number of samples (in terms of $d, k, 1/\alpha, 1/\varepsilon, \log(1/\delta)$). We do this via a fairly general reduction that shows if a class (such as Gaussians) admits a “locally small cover” and is “list decodable”, then the class of *its mixtures* is privately learnable. We find it useful to first define some notations in Section 2 before stating our main results in Section 3. Given the subtleties in the proofs, we give an overview of our technical contributions in Section 4 before delving into the core technical sections.

2. Preliminaries

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 $\{1, 2, \dots, k\}$. 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 $d_{\text{TV}}(f_1, f_2) = \frac{1}{2} \int_{\mathbb{R}^d} |f_1(x) - f_2(x)| dx$. For a matrix A , let $\|A\|_F = \sqrt{\text{Tr}(A^T A)}$ 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 (κ -ball) Consider a metric space (\mathcal{F}, κ) . For a fixed $f \in \mathcal{F}$, define $B_\kappa(r, f, \mathcal{F}) := \{f' \in \mathcal{F} : \kappa(f', f) \leq r\}$ to be the κ -ball of radius r around f .

Definition 2.2 (α -cover) A set $C_\alpha \subseteq \mathcal{F}$ is said to be an α -cover for a metric space (\mathcal{F}, κ) , if for every $f \in \mathcal{F}$, there exists an $f' \in C_\alpha$ such that $\kappa(f, f') \leq \alpha$.

Definition 2.3 (Locally small cover (Bun et al., 2021)) Let C_α be an α -cover for a metric space (\mathcal{F}, κ) . 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 an α -cover exists, we say \mathcal{F} admits a (t, γ) -locally small α -cover.

Definition 2.4 (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 paper, 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 $d_{\text{TV}}(f, g)$ to denote the TV distance from the underlying distribution. In other words, if $f = \sum_{i \in [k]} w_i f_i$ and $g = \sum_{i \in [k']} w'_i g_i$ then $d_{\text{TV}}(f, g) = d_{\text{TV}}(\sum_{i \in [k]} w_i f_i, \sum_{i \in [k']} w'_i g_i)$. 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 6.2).

Definition 2.5 (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. Distribution learning and list decodable learning

Here, we define list decodable learning 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. In other words, 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 allows for choosing the metric based on the application.

Definition 2.6 (List decodable learning under Huber’s contamination) Let \mathcal{F} be a class of distributions and $\kappa : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R}_{\geq 0}$ 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 $\min_{\hat{f} \in \hat{\mathcal{F}}} \kappa(\hat{f}, f) \leq \alpha$.

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 2.7 (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 $m \geq m(\alpha, \beta)$ 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 $1 - \beta$. If such an algorithm exists, we say \mathcal{F} is (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 2.8 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 .

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 2.9 (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 , it returns an $\hat{f} \in \mathcal{F}$ such that $d_{\text{TV}}(\hat{f}, g) \leq 3 \cdot \min_{f \in \mathcal{F}} d_{\text{TV}}(f, g) + \alpha$ with probability at least $1 - \beta$.

2.2. 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 2.10 ((ε, δ) -Indistinguishable) *Two distributions 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 2.11 ((ε, δ) -Differential Privacy (Dwork et al., 2006a,b)) *A randomized algorithm $\mathcal{M} : \mathcal{X}^n \rightarrow \mathcal{Y}$ 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.*

3. 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. d_{TV}), and (2) a list decoding algorithm (w.r.t. d_{TV}) for a class of distributions, then the class of its mixtures is privately learnable.

Theorem 3.1 (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. d_{TV}), and it is $(L, m, \alpha/15, \beta', 0)$ -list-decodable (w.r.t. d_{TV}), where $\log(1/\beta') = \tilde{\Theta}(\log(mk \log(tL/\alpha\delta)/\varepsilon\beta))$, then k -mix(\mathcal{F}) is (ε, δ) -DP (α, β) -PAC-learnable (w.r.t. 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 3.1, we can use a naive (α, β) -PAC-learner that outputs a single distribution as the list decoding algorithm (see Remark 2.8). Therefore, if we have (1) a locally small cover (w.r.t. d_{TV}), and (2) a (non-private) PAC learner (w.r.t. d_{TV}) for a class of distributions, then the class of its mixtures is privately learnable. The next corollary states this result formally.

Corollary 3.2 *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. d_{TV}), and it is $(\alpha/15, \beta')$ -PAC-learnable (w.r.t. d_{TV}) using $m(\alpha/15, \beta')$ samples, where $\log(1/\beta') = \tilde{\Theta}(\log(mk \log(t/\alpha\delta)/\varepsilon\beta))$, then k -mix(\mathcal{F}) is (ε, δ) -DP (α, β) -PAC-learnable (w.r.t. d_{TV}) with sample complexity*

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

As an application of Theorem 3.1, 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 3.3 (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. d_{TV} with sample complexity*

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

Remark 3.4 Note that if we use Corollary 3.2 and a PAC learner as a naive list decoding algorithm for Gaussians, the resulting sample complexity is worse. To see this, note that \mathcal{G} is (α, β) -PAC-learnable using $m(\alpha, \beta) = O\left(\frac{d^2 \log(1/\beta)}{\alpha^2}\right)$ samples. Using Corollary 3.2 and the existence of a locally small cover for Gaussians, we obtain a sample complexity upper bound of

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

This is a weaker result compared to Theorem 3.3 in terms of α , which is based on a more sophisticated (compression-based) list decoding algorithm for Gaussians.

It is worth mentioning that our approach 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).

4. Technical Challenges and Contributions

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. d_{TV} ? One idea to privately learn (dense) GMMs is to create a locally small cover w.r.t. d_{TV} (see Definition 2.3) 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. d_{TV}) does *not* exist, 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 6.2). 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 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. 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. 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 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. d_{TV}), then class of its dense

mixtures is list decodable (w.r.t. κ_{mix}). Then for the class of Gaussians, 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 Section 5). 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 3.1.

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. 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.

5. Common Member Selection

In this section, 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.

Definition 5.1 (Common Member) *Let (\mathcal{F}, κ) be a metric space and $\alpha, \zeta \in (0, 1]$. We say $f \in \mathcal{F}$ is an (α, ζ) -common-member for $\mathcal{Y} = \{Y_1, Y_2, \dots, Y_T\} \in (\mathcal{F}^*)^T$, 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 5.2 (Common Member Selector (CMS)) *Let (\mathcal{F}, κ) 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 $\mathcal{Y} = \{Y_1, Y_2, \dots, Y_T\} \in (\mathcal{F}^*)^T$ 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 5.3 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 5.4 (Private Common Member Selector (PCMS)) Let (\mathcal{F}, κ) 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 $C_1 = \{Y_1, Y_2, \dots, Y_T\} \in (\mathcal{F}^*)^T$ and $C_2 = \{Y'_1, Y_2, \dots, Y_T\} \in (\mathcal{F}^*)^T$ that differ in only one list, the output distributions of $\mathcal{A}(C_1)$ and $\mathcal{A}(C_2)$ are (ε, δ) -indistinguishable.

In Algorithm 1, we describe an algorithm for privately finding 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 $\{Y_1, \dots, Y_T\}$, 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 an 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 A.1) to make this selection differentially private⁴. In Theorem 5.5, we prove the correctness of this algorithm.

Algorithm 1 Private Common Member Selector (PCMS)

Input: $D = \{Y_1, Y_2, \dots, Y_T\} \in (\mathcal{F}^*)^T$, 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 $\text{score}(h, D) := |\{i \in [T] : \kappa(y, h) \leq 2\alpha \text{ for some } y \in Y_i\}|$.
 - 2: **return** GAP-MAX($C_\alpha, D, \text{score}, 0.1, \beta$)
-

Theorem 5.5 Let (\mathcal{F}, κ) 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 $T = O\left(\frac{\log(1/\delta) + \log(tQ/\beta)}{\varepsilon}\right)$.

6. Mixtures and Their Properties

In this section, 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.

4. This task can also be done using the Choosing Mechanism of Bun et al. (2015).

6.1. 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 2.4, where the number of components is fixed.

Definition 6.1 (General mixtures) *Let \mathcal{F} be an arbitrary class of distributions. We denote the class of mixtures of \mathcal{F} by $\text{mix}(\mathcal{F}) = \bigcup_{k=1}^{\infty} k\text{-mix}(\mathcal{F})$.*

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 6.2 (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 $\kappa_{\text{mix}}: \text{mix}(\mathcal{F}) \times \text{mix}(\mathcal{F}) \rightarrow \mathbb{R}_{\geq 0}$ as*

$$\kappa_{\text{mix}}(g_1, g_2) = \begin{cases} \min_{\pi} \max_{i \in [k_1]} \max\{k_1 \cdot |w_i - w'_{\pi(i)}|, d_{\text{TV}}(f_i, f'_{\pi(i)})\} & k_1 = k_2 \\ \infty & k_1 \neq k_2 \end{cases} \quad (1)$$

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. d_{TV} .

Lemma 6.3 *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 $\kappa_{\text{mix}}(f, f') \leq \alpha$, then $d_{\text{TV}}(f, f') \leq 3\alpha/2$.*

The next lemma states that if a class of distributions has a locally small cover w.r.t. d_{TV} , then the mixtures of that class admit a locally small cover w.r.t. κ_{mix} . The choice of the metric is important here as the next theorem is false if we consider the d_{TV} for mixtures. In other words, there is a class of distributions (e.g. Gaussians) that admits a locally small cover w.r.t. d_{TV} but there is no locally small cover for the mixtures of that class w.r.t. d_{TV} (Proposition 1.3 of [Aden-Ali et al. \(2021b\)](#)).

Theorem 6.4 *For any $0 < \alpha < \gamma < 1$, if a class of distributions \mathcal{F} has a (t, γ) -locally-small α -cover w.r.t. d_{TV} , then the class $k\text{-mix}(\mathcal{F})$ has a $(k!(tk/\alpha)^k, \gamma)$ -locally-small α -cover w.r.t. κ_{mix} .*

6.2. 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 has samples from *all* of the components, with high probability. This allows us to show that if a class of distribution is list decodable w.r.t. 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 6.5 (Dense mixtures) *Let \mathcal{F} be a class of distributions, $k \in \mathbb{N}$, and $\eta \in [0, 1/k]$. The class of k -mixtures of \mathcal{F} without negligible components is given 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 6.6 *For every $k \in \mathbb{N}$, $g \in k\text{-mix}(\mathcal{F})$ 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 $g = \gamma h + (1 - \gamma)g'$.*

Theorem 6.4 shows that if a class of distributions admits a locally small cover (w.r.t. 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. d_{TV}) then the class of its dense mixtures admits a locally small cover (w.r.t. κ_{mix}).

Lemma 6.7 *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. 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} .*

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. d_{TV}), then the class of its dense mixtures is list decodable (w.r.t. κ_{mix}).

Theorem 6.8 *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. 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 $m' = \frac{2m+8 \log(1/k\beta)}{1-\gamma}$.*

7. Proof of the Main Reduction

In this section, we prove Theorem 3.1 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.

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 $f^* \in k\text{-mix}(\mathcal{F})$, 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. d_{TV} . Therefore, we filter out all distributions that have a bad 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. 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. d_{TV} then using Lemma B.2, we get that \mathcal{F} is $(L_1, m_1, \alpha', 2\beta', 1 - \alpha'/k)$ -list-decodable w.r.t. 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 6.6, we can write $f^* = \gamma h + (1 - \gamma)f$, where $f \in (k, \frac{\alpha'}{k})$ -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 6.8, we know that $(k, \frac{\alpha'}{k})$ -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 $T = O\left(\frac{\log(1/\delta) + \log(t_1 L_2/\beta')}{\epsilon}\right)$. 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 $(k, \frac{\alpha'}{k})$ -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 6.3 to get that $d_{TV}(f, \mathcal{L}_i) \leq 3\alpha'/2$.

Note that the size of each \mathcal{L}_i is at most L_2 . By making use of the Minimum Distance Estimator (Theorem 2.9), 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 $d_{TV}(\hat{f}_i, f) \leq 3 \cdot d_{TV}(f, \mathcal{L}_i) + \alpha' \leq 11\alpha'/2$.

We then proceed with a filtering step. For each $i \in [T]$, we define $\mathcal{L}'_i = \{f' \in \mathcal{L}_i : d_{TV}(f', \hat{f}_i) < 11\alpha'/2\}$ 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 $\max_{f' \in \mathcal{L}'_i} d_{TV}(f, f') \leq 11\alpha'$.

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 $D = \{\mathcal{L}'_1, \dots, \mathcal{L}'_T\}$ w.r.t. κ_{mix} , and $\max_{f' \in \mathcal{L}'_i} d_{TV}(f, f') \leq 11\alpha'$ for all $i \in [T]$.

The fact that \mathcal{F} admits a $(t, 2\alpha')$ -locally small α' -cover, along with Lemma 6.7, 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 5.5 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 6.3 implies that \hat{f} is a $(3\alpha', 0.9)$ -common-member w.r.t. d_{TV} .

Using the fact that for every $i \in [T]$, $\max_{f' \in \mathcal{L}'_i} d_{TV}(f', f) \leq 11\alpha'$, we get that $d_{TV}(\hat{f}, f) \leq 14\alpha'$. Finally, triangle inequality implies that $d_{TV}(\hat{f}, f^*) \leq d_{TV}(\hat{f}, f) + d_{TV}(f, f^*) \leq 15\alpha'$ 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 $\beta' = \frac{\beta\varepsilon}{12ek \log(6ekt_1 L_2/\varepsilon\beta\delta)} < 1$. Applying Claim F.3 with $c_1 = 6k/\varepsilon, c_2 = t_1 L_2/\delta$, we get that $d_{TV}(\hat{f}, f^*) \leq \alpha$ 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 5.5. \blacksquare

8. Privately Learning GMMs

Prior to this work, there was no 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 section 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. See Theorem 3.3, and the formal proof in E.

8.1. List-decoding Gaussians using compression

As we stated in Remark 3.4, 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.

Lemma 8.1 *For any $\alpha, \beta \in (0, 1)$, the class \mathcal{G} is $(L, m, \alpha, \beta, 0)$ -list-decodable w.r.t. 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.

8.2. 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 2.5). However, the product of these two covers is not a valid cover for unbounded Gaussians as the 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 d_{TV} -ball of location gaussians around $N(0, I_d)$ and scale it using a small cover for the d_{TV} -ball of scale gaussians around $N(0, I_d)$. Showing that this is a valid and small cover for a small 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 d_{TV} distance between high-dimensional Gaussians. Showing that this is a small cover requires the recently proved lower bound on the d_{TV} distance of high-dimensional Gaussians (Arbas et al., 2023).

Once we created a small cover for a d_{TV} -ball of unbounded Gaussians around $N(0, I_d)$, we can transform it to a small cover for a d_{TV} -ball of unbounded Gaussians around an arbitrary $N(\mu, \Sigma)$. Finally, we use Lemma D.7 to show that there exists a global locally small cover for the whole space of Gaussians. An immediate implication of Lemma D.6 and Lemma D.7 is the existence of a locally small cover for unbounded Gaussians.

Lemma 8.2 *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. d_{TV} .*

9. More on Related Work

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

9.1. Privately learning Gaussians

[Karwa and Vadhan \(2018\)](#) proposed a sample and time efficient algorithm for estimating the mean and variance of bounded univariate Gaussians under pure DP. They also provide a method 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. \(2019a\)](#); [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 Minimum Distance Estimator ([Yatracos, 1985](#)) and inspired by the private hypothesis selection of [Bun et al. \(2021\)](#). One downside of [Aden-Ali et al. \(2021a\)](#) is that their method is not computationally efficient.

There has 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 a near-optimal sample complexity using a sample-and-aggregate-based technique. 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](#)).

9.2. 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. 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., 2019b](#); [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.

Note that in parameter estimation of GMMs (even in the non-private setting), the exponential dependence of the sample complexity on the number of components is inevitable (Moitra and Valiant, 2010).

9.3. 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. 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 axis-aligned GMMs under the constraint of approximate DP. They extended the idea of stable histograms used in Karwa and Vadhan (2018) to learn univariate GMMs. However, this idea can't be generalized to general GMMs, as it is not clear how to learn even a single high-dimensional Gaussian using stability-based histogram.

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.

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.

References

- Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. 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, 2017.
- Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR, 2021.
- Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.
- Ishaq Aden-Ali, Hassan Ashtiani, and Gautam Kamath. On the sample complexity of privately learning unbounded high-dimensional gaussians. In *Algorithmic Learning Theory*, pages 185–216. PMLR, 2021a.
- Ishaq Aden-Ali, Hassan Ashtiani, and Christopher Liaw. Privately learning mixtures of axis-aligned gaussians. *Advances in Neural Information Processing Systems*, 34:3925–3938, 2021b.
- Daniel Alabi, Pravesh K Kothari, Pranay Tankala, Prayaag Venkat, and Fred Zhang. Privately estimating a gaussian: Efficient, robust, and optimal. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 483–496, 2023.
- Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private pac learning implies finite littlestone dimension. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 852–860, 2019.
- Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James Voss. The more, the merrier: the blessing of dimensionality for learning large gaussian mixtures. In *Conference on Learning Theory*, pages 1135–1164. PMLR, 2014.
- Jamil Arbas, Hassan Ashtiani, and Christopher Liaw. Polynomial time and private learning of unbounded gaussian mixture models. In *International Conference on Machine Learning*. PMLR, 2023.
- Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning gaussians and beyond. In *Conference on Learning Theory*, pages 1075–1076. PMLR, 2022.
- Hassan Ashtiani and Abbas Mehrabian. Some techniques in density estimation. *arXiv preprint arXiv:1801.04003*, 2018.
- Hassan Ashtiani, Shai Ben-David, Nicholas Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. *Advances in Neural Information Processing Systems*, 31, 2018a.
- Hassan Ashtiani, Shai Ben-David, and Abbas Mehrabian. Sample-efficient learning of mixtures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.
- Hassan Ashtiani, Shai Ben-David, Nicholas JA Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *Journal of the ACM (JACM)*, 67(6):1–42, 2020.

- Hilal Asi, Jonathan Ullman, and Lydia Zakyntinou. From robustness to privacy and back. *arXiv preprint arXiv:2302.01855*, 2023.
- Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M Kane, Pravesh K Kothari, and Santosh S Vempala. Robustly learning mixtures of k arbitrary gaussians. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1234–1247, 2022.
- Mikhail Belkin and Kaushik Sinha. Learning gaussian mixtures with arbitrary separation. *arXiv preprint arXiv:0907.1054*, 2009.
- Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 103–112. IEEE, 2010.
- Shai Ben-David, Alex Bie, Clément L Canonne, Gautam Kamath, and Vikrant Singhal. Private distribution learning with public data: The view from sample compression. *arXiv preprint arXiv:2308.06239*, 2023.
- Alex Bie, Gautam Kamath, and Vikrant Singhal. Private estimation with public data. *Advances in Neural Information Processing Systems*, 35:18653–18666, 2022.
- Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. *Advances in Neural Information Processing Systems*, 33:14475–14485, 2020.
- Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, and Lydia Zakyntinou. Covariance-aware private mean estimation without private covariance estimation. *Advances in Neural Information Processing Systems*, 34:7950–7964, 2021.
- Gavin Brown, Samuel Hopkins, and Adam Smith. 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, 2023.
- S Charles Brubaker and Santosh S Vempala. Isotropic pca and affine-invariant clustering. *Building Bridges: Between Mathematics and Computer Science*, pages 241–281, 2008.
- Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 634–649. IEEE, 2015.
- Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 369–380, 2016.
- Mark Bun, Cynthia Dwork, Guy N Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–86, 2018.
- Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 389–402. IEEE, 2020.

- Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. *IEEE Transactions on Information Theory*, 67(3), 2021.
- Siu-On Chan, Ilias Diakonikolas, Rocco A Servedio, and Xiaorui Sun. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 604–613, 2014.
- Hongjie Chen, Vincent Cohen-Addad, Tommaso d’Orsi, Alessandro Epasto, Jacob Imola, David Steurer, and Stefan Tiegel. Private estimation algorithms for stochastic block models and mixture models. *arXiv preprint arXiv:2301.04822*, 2023.
- Edith Cohen, Haim Kaplan, Yishay Mansour, Uri Stemmer, and Eliad Tsfadia. Differentially-private clustering of easy instances. In *International Conference on Machine Learning*, pages 2049–2059. PMLR, 2021.
- Edith Cohen, Xin Lyu, Jelani Nelson, Tamás Sarlós, and Uri Stemmer. 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, 2023.
- Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE, 1999.
- Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001.
- Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*, 2018.
- Ilias Diakonikolas. Learning structured distributions. *Handbook of Big Data*, 267:10–1201, 2016.
- Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. 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, 2017.
- John Duchi, Saminul Haque, and Rohith Kuditipudi. A fast algorithm for adaptive private mean estimation. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 2023.
- Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.
- Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 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, 2006a.
- Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 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, 2006b.

- Jon Feldman, Rocco A Servedio, and Ryan O’Donnell. Pac learning axis-aligned mixtures of gaussians with no separation assumption. In *Proceedings of the 19th annual conference on Learning Theory*, pages 20–34, 2006.
- Kristian Georgiev and Samuel Hopkins. Privacy induces robustness: Information-computation gaps and sparse mean estimation. *Advances in Neural Information Processing Systems*, 35:6829–6842, 2022.
- Moritz Hardt and Eric Price. Sharp bounds for learning a mixture of two gaussians. 2014.
- Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018.
- Samuel B Hopkins, Gautam Kamath, and Mahbod Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1406–1417, 2022.
- Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 497–506, 2023.
- Daniel Hsu and Sham M Kakade. 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, 2013.
- Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.
- Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562, 2010.
- Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902. PMLR, 2019a.
- Gautam Kamath, Or Sheffet, Vikrant Singhal, and Jonathan Ullman. Differentially private algorithms for learning mixtures of separated gaussians. *Advances in Neural Information Processing Systems*, 32, 2019b.
- Gautam Kamath, Argyris Mouzakis, and Vikrant Singhal. New lower bounds for private estimation and a generalized fingerprinting lemma. *Advances in Neural Information Processing Systems*, 35:24405–24418, 2022a.
- Gautam Kamath, Argyris Mouzakis, Vikrant Singhal, Thomas Steinke, and Jonathan Ullman. A private and computationally-efficient estimator for unbounded gaussians. In *Conference on Learning Theory*, pages 544–572. PMLR, 2022b.
- Haim Kaplan, Katrina Ligett, Yishay Mansour, Moni Naor, and Uri Stemmer. Privately learning thresholds: Closing the exponential gap. In *Conference on Learning Theory*, pages 2263–2285. PMLR, 2020.

- Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.
- Pravesh Kothari, Pasin Manurangsi, and Ameya Velingker. Private robust estimation by stabilizing convex relaxations. In *Conference on Learning Theory*, pages 723–777. PMLR, 2022.
- Pravesh K Kothari, Jacob Steinhardt, and David Steurer. 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, 2018.
- Jerry Li and Ludwig Schmidt. Robust and proper learning for mixtures of gaussians via systems of polynomial inequalities. In *Conference on Learning Theory*, pages 1302–1382. PMLR, 2017.
- Allen Liu and Jerry Li. 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, 2022.
- Allen Liu and Ankur Moitra. Settling the robust learnability of mixtures of gaussians. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 518–531, 2021.
- Allen Liu and Ankur Moitra. Learning gmms with nearly optimal robustness guarantees. In *Conference on Learning Theory*, pages 2815–2895. PMLR, 2022.
- Allen Liu, Jerry Li, and Ankur Moitra. Robust model selection and nearly-proper learning for gmms. *Advances in Neural Information Processing Systems*, 35:22830–22843, 2022a.
- Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. *Advances in neural information processing systems*, 34:3887–3901, 2021.
- Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, pages 1167–1246. PMLR, 2022b.
- Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS’07)*, pages 94–103. IEEE, 2007.
- M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005. ISBN 9780521835404. URL <https://books.google.ca/books?id=0bAY16d7hvkC>.
- Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010.
- Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84, 2007.

- Oded Regev and Aravindan Vijayaraghavan. On learning mixtures of well-separated gaussians. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 85–96. IEEE, 2017.
- Arora Sanjeev and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257, 2001.
- Eliad Tsfadia, Edith Cohen, Haim Kaplan, Yishay Mansour, and Uri Stemmer. Friendlycore: Practical differentially private aggregation. In *International Conference on Machine Learning*, pages 21828–21863. PMLR, 2022.
- Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.
- Xuan Wu and Changzhi Xie. Improved algorithms for properly learning mixture of gaussians. In *National Conference of Theoretical Computer Science*, pages 8–26. Springer, 2018.
- Yannis G Yatracos. Rates of convergence of minimum distance estimators and kolmogorov’s entropy. *The Annals of Statistics*, 13(2):768–774, 1985.

Appendix A. Missing proofs of section 5

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. 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 A.1 (GAP-MAX, Theorem IV.6. of Bun et al. (2021)) *Let (\mathcal{F}, κ) be a metric space and \mathcal{X} be an arbitrary set. Let $\text{score}: \mathcal{F} \times \mathcal{X}^T \rightarrow \mathbb{R}_{\geq 0}$ 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 $|\text{score}(f, D) - \text{score}(f, D')| \leq 1$. 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*

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

then

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

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

Proof of Theorem 5.5 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 5.2). 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 $\{Y_1, Y_2, \dots, Y_T\}$, we have $\sup_{h \in C_\alpha} \text{score}(h, D) = T$. 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 A.1 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 A.1 implies that GAP-MAX is (ε, δ) -DP. \blacksquare

Appendix B. Missing proofs of section 6

Proof of Lemma 6.6 For any $g = \sum_{i \in [k]} w_i f_i \in k\text{-mix}(\mathcal{F})$, 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 $\sum_{i \in [k] \setminus N} \frac{w_i}{1 - \gamma} f_i \in (k, \frac{\alpha}{k})\text{-dense-mix}(\mathcal{F})$. \blacksquare

Proof of Lemma 6.3 Using the definition of κ_{mix} , we get that for every $i \in [k]$, $|w_i - w'_i| \leq \alpha/k$ and $d_{TV}(f_i, f'_i) \leq \alpha$. Therefore,

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

as required. \blacksquare

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

Proposition B.1 *Let $\alpha \in (0, 1]$. There is an α -cover for $\Delta_k = \{(w_1, w_2, \dots, w_k) \in \mathbb{R}_{\geq 0}^k : \sum_{i \in [k]} w_i = 1\}$ 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. \blacksquare

Proof of Theorem 6.4 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 B.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 $\kappa_{mix}(g, g') \leq \alpha$. Moreover, we have $|B_{\kappa_{mix}}(\gamma, g, \mathcal{J})| \leq |B_{d_{TV}}(\gamma, f_i, C_\alpha)|^k \cdot |\hat{\Delta}_k| \cdot k! = t^k \cdot (k/\alpha)^k \cdot k!$, 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. \blacksquare

Proof of Theorem 6.7 Using Theorem 6.4 we know that if \mathcal{F} has a (t, γ) -locally-small α -cover w.r.t. 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. \blacksquare

In order to prove Theorem 6.8, 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 B.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 $\left(L\left(\frac{10e \log(1/\beta)}{1-\gamma}\right)^m, \frac{2m+8 \log(1/\beta)}{1-\gamma}, \alpha, 2\beta, \gamma\right)$ -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, \dots, 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 $|\mathcal{L}| = L \cdot \binom{N}{m} \leq L \cdot \left(\frac{2em+8e \log(1/\beta)}{(1-\gamma)m}\right)^m \leq L \cdot \left(\frac{10e \log(1/\beta)}{1-\gamma}\right)^m$. Finally, using a union bound we will get that \mathcal{F} is $\left(L\left(\frac{10e \log(1/\beta)}{1-\gamma}\right)^m, \frac{2m+8 \log(1/\beta)}{1-\gamma}, \alpha, 2\beta, \gamma\right)$ -list-decodable w.r.t. κ . \blacksquare

Proof of Theorem 6.8 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 $\min_{f' \in \mathcal{L}_s} d_{\text{TV}}(f', f_i) \leq \alpha$. Let $\hat{\Delta}_s$ be an $\frac{\alpha}{s}$ -cover for Δ_s from Proposition B.1. Now construct a set $\mathcal{J} = \cup_{s \in [k]} \{\sum_{i \in [s]} \hat{w}_i \hat{f}_i : \hat{w} \in \hat{\Delta}_s, \hat{f}_i \in \mathcal{L}_s\}$. Note that with probability at least $1-k\beta$ we have $\min_{g' \in \mathcal{J}} \kappa_{\text{mix}}(g, g') \leq \max\{\alpha, \alpha\} = \alpha$. Moreover, we have $|\mathcal{J}| = \sum_{s \in [k]} |\mathcal{L}_s|^s |\hat{\Delta}_s| = \sum_{s \in [k]} L^s \left(\frac{s}{\alpha}\right)^s \leq \left(\frac{kL}{\alpha}\right)^{k+1}$. Thus, the class of $(k, \frac{\alpha}{k})$ -dense-mix(\mathcal{F}) is $\left(\left(\frac{kL}{\alpha}\right)^{k+1}, m, \alpha, k\beta, 0\right)$ -list-decodable w.r.t. κ_{mix} . Using Lemma B.2, we get that $(k, \frac{\alpha}{k})$ -dense-mix(\mathcal{F}) is $\left(\left(\frac{kL}{\alpha}\right)^{k+1} \cdot \left(\frac{10e \log(1/k\beta)}{1-\gamma}\right)^m, \frac{2m+8 \log(1/k\beta)}{1-\gamma}, \alpha, 2k\beta, \gamma\right)$ -list-decodable w.r.t. κ_{mix} . \blacksquare

Appendix C. List decoding Gaussians using compression

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 C.1 (Compression schemes for distributions (Ashtiani et al., 2018a)) For a set of functions $\tau, m, t : (0, 1) \rightarrow \mathbb{Z}_{\geq 0}$, 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 $d_{\text{TV}}(D(L, B), f) \leq \alpha$.

The following result states that the class of Gaussians is compressible (see Lemma 4.1 of [Ash-tiani et al. \(2018a\)](#)).

Lemma C.2 *Let $\alpha \in (0, 1)$. The class \mathcal{G} is $(O(d), \tilde{O}(d^2 \log(1/\alpha)), O(d))$ -compressible.*

Remark C.3 *Our reduction from list decoding the class of Gaussians to the problem of compressing this class 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 8.1.*

Proof of Lemma 8.1 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. 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 C.1, with probability at least $1 - \beta$ there exists $\hat{f} \in \mathcal{L}$ such that $d_{\text{TV}}(\hat{f}, f) \leq \alpha$. Note that $|\mathcal{L}| = O((s \log(1/\beta))^{(t+\tau)})$. Putting together with Lemma C.2 concludes the result. \blacksquare

Appendix D. Locally small cover for Gaussians

Definition D.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 D.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 $B_{d_{\text{TV}}}(r, f, \mathcal{F})$ stands for the d_{TV} -ball of radius r around f (See Definition 2.1).

Lemma D.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 $B_{d_{\text{TV}}}(\gamma, N(0, I), \mathcal{G}^L)$ of size $(\frac{\gamma}{\alpha})^{O(d)}$ w.r.t. d_{TV} .*

Lemma D.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 $B_{d_{\text{TV}}}(\gamma, N(0, \Sigma), \mathcal{G}^S)$ of size $(\frac{\gamma}{\alpha})^{O(d^2)}$ w.r.t. d_{TV} .*

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

Theorem D.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*

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

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

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

The next lemma proposes a small cover for the unbounded Gaussians near $N(\mu, \Sigma)$ for any given μ and Σ . To do so, we combine the small covers from Lemma D.3 and Lemma D.4 in a way that it approximates any Gaussian near $N(0, I)$.

Lemma D.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 $B_{\text{dTV}}(\gamma, N(\mu, \Sigma), \mathcal{G})$ of size at most $(\frac{\gamma}{\alpha})^{O(d^2)}$.*

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

Let $\gamma \in (\alpha, \frac{1}{600})$ and consider the ball $\mathcal{B} := B_{\text{dTV}}(\gamma, N(0, I), \mathcal{G})$. Let $\gamma_1 = 200\gamma$ and C^L be an $\frac{\sqrt{2}}{200}\alpha$ -cover for $B_{\text{dTV}}(\gamma_1, N(0, I), \mathcal{G}^L)$ from Lemma D.3. Also, let $\gamma_2 = 200\gamma$ and C^S be an $\frac{\sqrt{2}}{200}\alpha$ -cover for $B_{\text{dTV}}(\gamma_2, N(0, I), \mathcal{G}^S)$ from Lemma D.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 $d_{\text{TV}}(N(0, I), N(\hat{\mu}, \hat{\Sigma})) \leq \gamma \leq \frac{1}{600}$. Applying the lower bound of Theorem D.5 with $\Sigma_1 = I, \Sigma_2 = \hat{\Sigma}, \mu_1 = 0, \mu_2 = \hat{\mu}$ gives that

$$\|\hat{\Sigma} - I\|_F \leq 200 d_{\text{TV}}(N(0, I), N(\hat{\mu}, \hat{\Sigma})) \leq 200\gamma \quad \text{and} \quad (2)$$

$$\|\hat{\mu}\|_2 \leq 200 d_{\text{TV}}(N(0, I), N(\hat{\mu}, \hat{\Sigma})) \leq 200\gamma. \quad (3)$$

Moreover, applying Theorem D.5 with $\Sigma_1 = \hat{\Sigma}, \Sigma_2 = I, \mu_1 = \hat{\mu}, \mu_2 = 0$ gives that

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

Next, applying the upper bound of Theorem D.5 with $\mu_1 = \mu_2 = 0, \Sigma_1 = I, \Sigma_2 = \hat{\Sigma}$ gives $d_{\text{TV}}(N(0, I), N(0, \hat{\Sigma})) \leq \frac{1}{\sqrt{2}}\|\hat{\Sigma} - I\|_F \leq \frac{200}{\sqrt{2}}\gamma < 200\gamma = \gamma_2$. This implies that $N(0, \hat{\Sigma}) \in B_{\text{dTV}}(\gamma_2, N(0, I), \mathcal{G}^S)$. Recall that C^S is an $\frac{\sqrt{2}}{200}\alpha$ -cover for $B_{\text{dTV}}(\gamma_2, N(0, I), \mathcal{G}^S)$. Thus, there exists $N(0, \tilde{\Sigma}) \in C^S$ such that $d_{\text{TV}}(N(0, \hat{\Sigma}), N(0, \tilde{\Sigma})) \leq \frac{\sqrt{2}}{200}\alpha$. Using the lower bound of Theorem D.5 with $\Sigma_1 = \tilde{\Sigma}, \Sigma_2 = \hat{\Sigma}, \mu_1 = 0, \mu_2 = 0$ results in

$$\|\tilde{\Sigma}^{-1/2}\hat{\Sigma}\tilde{\Sigma}^{-1/2} - I\|_2 \leq \|\tilde{\Sigma}^{-1/2}\hat{\Sigma}\tilde{\Sigma}^{-1/2} - I\|_F \leq 200 d_{\text{TV}}(N(0, \hat{\Sigma}), N(0, \tilde{\Sigma})) \leq \sqrt{2}\alpha. \quad (4)$$

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 3 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 D.5 with $\Sigma_1 = I, \Sigma_2 = I, \mu_1 = \tilde{\Sigma}^{-1/2}\hat{\mu}, \mu_2 = 0$ gives $d_{\text{TV}}(N(\tilde{\Sigma}^{-1/2}\hat{\mu}, I), N(0, I)) \leq \frac{1}{\sqrt{2}}\|\tilde{\Sigma}^{-1/2}\hat{\mu}\|_2 \leq \frac{200\gamma\sqrt{2}}{\sqrt{2}} = \gamma_1$. Thus $N(\tilde{\Sigma}^{-1/2}\hat{\mu}, I) \in B_{\text{dTV}}(\gamma_1, N(0, I), \mathcal{G}^L)$. Recall that C^L is an $\frac{\sqrt{2}}{200}\alpha$ -cover for $B_{\text{dTV}}(\gamma_1, N(0, I), \mathcal{G}^L)$. Therefore, there exists $N(\tilde{\mu}, I) \in C^L$ such that $d_{\text{TV}}(N(\tilde{\mu}, I), N(\tilde{\Sigma}^{-1/2}\hat{\mu}, I)) \leq \frac{\sqrt{2}}{200}\alpha$. Using the lower bound of Theorem D.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 4, we can use the upper bound in Theorem D.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 $d_{\text{TV}}(N(\tilde{\Sigma}^{1/2}\tilde{\mu}, \tilde{\Sigma}), N(\hat{\mu}, \hat{\Sigma})) \leq \alpha$. 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 $B_{d_{\text{TV}}}(\gamma, N(\mu, \Sigma), \mathcal{G})$. Note that using Lemma F.1, for any $\Sigma, \Sigma_1, \Sigma_2 \in S^d$, we have:

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

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 $C^{B\Sigma}$ is an α -cover for $B_{d_{\text{TV}}}(\gamma, N(0, \Sigma), \mathcal{G})$ since the 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 $B_{d_{\text{TV}}}(\gamma, N(\mu, \Sigma), \mathcal{G})$ since it is the shifted version of $C^{B\Sigma}$. 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 D.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 $B_{d_{\text{TV}}}(\gamma, f, \mathcal{F})$ 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. d_{TV} .*

The proof of the above lemma is non-constructive and uses Zorn's lemma.

Appendix E. Proof of Theorem 3.3

Proof According to Lemma 8.1, the class \mathcal{G} is $(L, m, \alpha, \beta, 0)$ -list-decodable w.r.t. d_{TV} for $L = (d \log(1/\beta))^{O(d^2 \log(1/\alpha))}$ and $m = O(d \log(1/\beta))$. Moreover, Lemma 8.2 implies that \mathcal{G} admits a $(t, 2\alpha)$ -locally small α -cover, where $t = 4^{O(d^2)}$. Using Theorem 3.1, we get that k -mix(\mathcal{G}) is (ε, δ) -DP (α, β) -PAC-learnable using

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

samples. ■

Appendix F. Useful facts

Lemma F.1 *Let f be an arbitrary function, and X, Y be two random variables with the same support. Then $d_{\text{TV}}(f(X), f(Y)) \leq d_{\text{TV}}(X, Y)$.*

Claim F.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 F.3 *Let $c_1, c_2, 1/\beta \geq 1$ and define $\beta' = \frac{\beta}{2ec_1 \log(ec_1 c_2/\beta)}$. We have that $\beta' \leq \frac{\beta}{2ec_1} \leq \frac{1}{ec_1}$ and $c_1 \beta' \log(c_2/\beta') \leq \beta$.*

Proof We have

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

where in the last inequality, we used Claim F.2 with $x = \log(ec_1 c_2/\beta) \geq 1$. ■