On Learning Mixtures of Well-Separated Gaussians

Oded Regev[*] Aravindan Vijayaraghavan[†]

Abstract

We consider the problem of efficiently learning mixtures of a large number of spherical Gaussians, when the components of the mixture are well separated. In the most basic form of this problem, we are given samples from a uniform mixture of k standard spherical Gaussians with means $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$, and the goal is to estimate the means up to accuracy δ using $\mathrm{poly}(k, d, 1/\delta)$ samples.

In this work, we study the following question: what is the minimum separation needed between the means for solving this task? The best known algorithm due to Vempala and Wang [JCSS 2004] requires a separation of roughly $\min\{k, d\}^{1/4}$. On the other hand, Moitra and Valiant [FOCS 2010] showed that with separation $o(1)$, exponentially many samples are required. We address the significant gap between these two bounds, by showing the following results.

- We show that with separation $o(\sqrt{\log k})$, super-polynomially many samples are required. In fact, this holds even when the k means of the Gaussians are picked at random in $d = O(\log k)$ dimensions.
- We show that with separation $\Omega(\sqrt{\log k})$, $\mathrm{poly}(k, d, 1/\delta)$ samples suffice. Notice that the bound on the separation is independent of δ. This result is based on a new and efficient "accuracy boosting" algorithm that takes as input coarse estimates of the true means and in time (and samples) $\mathrm{poly}(k, d, 1/\delta)$ outputs estimates of the means up to arbitrarily good accuracy δ assuming the separation between the means is $\Omega(\min\{\sqrt{\log k}, \sqrt{d}\})$ (independently of δ). The idea of the algorithm is to iteratively solve a "diagonally dominant" system of non-linear equations.

We also (1) present a *computationally efficient* algorithm in $d = O(1)$ dimensions with only $\Omega(\sqrt{d})$ separation, and (2) extend our results to the case that components might have different weights and variances. These results together essentially characterize the optimal order of separation between components that is needed to learn a mixture of k spherical Gaussians with polynomial samples.

1 Introduction

Gaussian mixture models are one of the most widely used statistical models for clustering. In this model, we are given random samples, where each sample point $x \in \mathbb{R}^d$ is drawn independently from

[*] Courant Institute of Mathematical Sciences, New York University. Supported by the Simons Collaboration on Algorithms and Geometry and by the National Science Foundation (NSF) under Grant No. CCF-1320188.

[†] Department of Electrical Engineering and Computer Science, Northwestern University. Supported by the National Science Foundation (NSF) under Grant No. CCF-1652491 and CCF-1637585. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

one of k Gaussian components according to mixing weights $w_1, w_2, \ldots, w_k$, where each Gaussian component $j \in [k]$ has a mean $\mu_j \in \mathbb{R}^d$ and a covariance $\Sigma_j \in \mathbb{R}^{d \times d}$. We focus on an important special case of the problem where each of the components is a *spherical* Gaussian, i.e., the covariance matrix of each component is a multiple of the identity. If f represents the p.d.f. of the Gaussian mixture $\mathcal{G}$, and g_j represents the p.d.f. of the jth Gaussian component,

$$g_j = \frac{1}{\sigma_j^d} \exp\left(-\pi \|x - \mu_j\|_2^2 / \sigma_j^2\right), \; f(x) = \sum_{j=1}^{k} w_j g_j(x).$$

The goal is to estimate the parameters $\{(w_j, \mu_j, \sigma_j) : j \in [k]\}$ up to required accuracy $\delta > 0$ in time and number of samples that is polynomial in $k, d, 1/\delta$.

Learning mixtures of Gaussians has a long and rich history, starting with the work of Pearson [38]. (See Section 1.2 for an overview of prior work.) Most of the work on this problem, especially in the early years but also recently, is under the assumption that there is some minimum *separation* between the means of the components in the mixture. Starting with work by Dasgupta [16], and continuing with a long line of work (including [4, 45, 2, 29, 39, 17, 14, 30, 7, 8, 46, 19]), efficient algorithms were found under mild separation assumptions. Considering for simplicity the case of uniform mixtures (i.e., all weights are $1/k$) of standard Gaussians (i.e., spherical with $\sigma = 1$), the best known result due to Vempala and Wang [45] provides an efficient algorithm (both in terms of samples and running time) under separation of at least $\min\{k, d\}^{1/4} \text{poly}\log(dk/\delta)$ between any two means.

A big open question in the area is whether efficient algorithms exist under weaker separation assumptions. It is known that when the separation is $o(1)$, a super-polynomial number of samples is required (e.g., [35, 3, 26]), but the gap between this lower bound and the above upper bound of $\min\{k, d\}^{1/4} \text{poly}\log(dk/\delta)$ is quite wide. Can it be that efficient algorithms exist under only $\Omega(1)$ separation? In fact, prior to this work, this was open even in the case of $d = 1$.

Question 1.1. What is the minimum order of separation that is needed to learn the parameters of a mixture of k spherical Gaussians up to accuracy δ using $\text{poly}(d, k, 1/\delta)$ samples?

1.1 Our Results

By improving both the lower bounds and the upper bounds mentioned above, we characterize (up to constants) the minimum separation needed to learn the mixture from polynomially many samples. Our first result shows super-polynomial lower bounds when the separation is of the order $o(\sqrt{\log k})$. In what follows, $\Delta_{\text{param}}(\mathcal{G}, \tilde{\mathcal{G}})$ represents the "distance" between the parameters of the two mixtures of Gaussians $\mathcal{G}, \tilde{\mathcal{G}}$ (see Definition 2.2 for the precise definition).

Informal Theorem 1.2 (Lower Bounds). *For any $\gamma(k) = o(\sqrt{\log k})$, there are two uniform mixtures of standard spherical Gaussians $\mathcal{G}, \tilde{\mathcal{G}}$ in $d = O(\log k)$ dimensions with means $\{\mu_1, \ldots, \mu_k\}, \{\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_k\}$ respectively, that are well separated*

$$\forall i \neq j \in [k] : \|\mu_i - \mu_j\|_2 \geq \gamma(k), \quad \text{and } \|\tilde{\mu}_i - \tilde{\mu}_j\|_2 \geq \gamma(k),$$

and whose parameter distance is large $\Delta_{\text{param}}(\{\mu_1, \ldots, \mu_k\}, \{\tilde{\mu}_1, \ldots, \tilde{\mu}_k\}) = \Omega(1)$, but have very small statistical distance $\|\mathcal{G} - \tilde{\mathcal{G}}\|_{TV} \leq k^{-\omega(1)}$.

The above statement implies that we need at least $k^{\omega(1)}$ many samples to distinguish between $\mathcal{G}, \tilde{\mathcal{G}}$, and identify $\mathcal{G}$. See Theorem 3.1 for a formal statement of the result. In fact, these sample complexity lower bounds hold even when the means of the Gaussians are picked randomly in a ball of radius $\sqrt{d}$ in $d = o(\log k)$ dimensions. This rules out obtaining smoothed analysis guarantees for small dimensions (as opposed to [10, 3] which give polytime algorithms for smoothed mixtures of Gaussians in $k^{\Omega(1)}$ dimensions).

Our next result shows that the separation of $\Omega(\sqrt{\log k})$ is tight – this separation suffices to learn the parameters of the mixture with polynomial samples. We state the theorem for the special case of uniform mixtures of spherical Gaussians. (See Theorem 5.1 for the formal statement.)

Informal Theorem 1.3 (Tight Upper Bound in terms of k)**.** *There exists a universal constant $c > 0$, such that given samples from a uniform mixture of standard spherical Gaussians in $\mathbb{R}^d$ with well-separated means, i.e.,*

$$\forall i, j \in [k], i \neq j : \ \|\mu_i - \mu_j\|_2 \geq c\sqrt{\log k} \tag{1}$$

there is an algorithm that for any $\delta > 0$ uses only $\text{poly}(k, d, 1/\delta)$ samples and with high probability finds $\{\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_k\}$ satisfying $\Delta_{\text{param}}(\{\mu_1, \ldots, \mu_k\}, \{\tilde{\mu}_1, \ldots, \tilde{\mu}_k\}) \leq \delta$.

While the above algorithm uses only $\text{poly}(k, d, 1/\delta)$ samples, it is computationally inefficient.

Our next result shows that in constant dimensions, one can obtain a *computationally efficient* algorithm. In fact, in such low dimensions a separation of order $\Omega(1)$ suffices.

Informal Theorem 1.4 (Efficient algorithm in low dimensions)**.** *There exists a universal constant $c > 0$, such that given samples from a uniform mixture of standard spherical Gaussians in $\mathbb{R}^d$ with well-separated means, i.e.,*

$$\forall i, j \in [k], i \neq j : \ \|\mu_i - \mu_j\|_2 \geq c\sqrt{d} \tag{2}$$

there is an algorithm that for any $\delta > 0$ uses only $\text{poly}_d(k, 1/\delta)$ time (and samples) and with high probability finds $\{\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_k\}$ satisfying $\Delta_{\text{param}}(\{\mu_1, \ldots, \mu_k\}, \{\tilde{\mu}_1, \ldots, \tilde{\mu}_k\}) \leq \delta$.

See Theorem 6.1 for a formal statement. An important feature of the above two algorithmic results is that the separation is independent of the accuracy δ that we desire in parameter estimation (δ can be arbitrarily small compared to k and d). These results together almost give a *tight characterization* (up to constants) for the amount of separation needed to learn with $\text{poly}(k, d, 1/\delta)$ samples.

Iterative Algorithm. The core technical portion of Theorem 1.3 and Theorem 1.4 is a new iterative algorithm, which is the main algorithmic contribution of the paper. This algorithm takes coarse estimates of the means, and iteratively refines them to get arbitrarily good accuracy δ. We now present an informal statement of the guarantees of the iterative algorithm.

Informal Theorem 1.5 (Iterative Algorithm Guarantees)**.** *There exists a universal constant $c > 0$, such that given samples from a uniform mixture of standard spherical Gaussians in $\mathbb{R}^d$ with well-separated means, i.e.*

$$\forall i, j \in [k], i \neq j : \ \|\mu_i - \mu_j\|_2 \geq c \min\{\sqrt{\log k}, \sqrt{d}\} \tag{3}$$

and suppose we are given initializers $\tilde{\mu}_1, \dots, \tilde{\mu}_k$ for the means $\mu_1, \dots, \mu_k$ satisfying

$$\forall j \in [k], \ \frac{1}{\sigma_j} \|\mu_j - \tilde{\mu}_j\|_2 \leq 1/\text{poly}\big(\min\{d, k\}\big).$$

There exists an iterative algorithm that for any $\delta > 0$ that runs in $\text{poly}(k, d, 1/\delta)$ time (and samples), and after $T = O(\log\log(k/\delta))$ iterations, finds with high probability $\mu_1^{(T)}, \dots, \mu_k^{(T)}$ such that $\Delta_{\text{param}}(\{\mu_1, \dots, \mu_k\}, \{\mu_1^{(T)}, \dots, \mu_k^{(T)}\}) \leq \delta$.

The above theorem also holds when the weights and variances are unequal. See Theorem 4.1 for a formal statement. Note that in the above result, the desired accuracy δ can be arbitrarily small compared to k, and the separation required does not depend on δ. To prove the polynomial identifiability results (Theorems 1.3 and 1.4), we first find coarse estimates of the means that serve as initializers to this iterative algorithm, which then recovers the means up to arbitrarily fine accuracy independent of the separation.

The algorithm works by solving a system of non-linear equations that is obtained by estimating simple statistics (e.g., means) of the distribution restricted to certain carefully chosen regions. We prove that the system of non-linear equations satisfies a notion of "diagonal dominance" that allows us to leverage iterative algorithms like Newton's method and achieve rapid (quadratic) convergence.

The techniques developed here can find such initializers using only $\text{poly}(k, d)$ many samples, but use time that is exponential in k. This leads to the following natural open question:

Open Question 1.6. Given a mixture of spherical Gaussians with equal weights and variances, and with separation

$$\forall i \neq j \in [k], \|\mu_i - \mu_j\|_2 \geq c\sqrt{\log k}$$

for some sufficiently large absolute constant $c > 0$, is there an algorithm that recovers the parameters up to δ accuracy in time $\text{poly}(k, d, 1/\delta)$?

Our iterative algorithm shows that to resolve this open question affirmatively, it is enough to find initializers that are reasonably close to the true parameters. In fact, a simple amplification argument shows that initializers that are $c\sqrt{\log k}/8$ close to the true means will suffice for this approach.

Our iterative algorithm is reminiscent of some commonly used iterative heuristics, such as Lloyd's Algorithm and especially Expectation Maximization (EM). While these iterative methods are the practitioners' method-of-choice for learning probabilistic models, they have been notoriously hard to analyze. We believe that the techniques developed here may also be useful to prove guarantees for these heuristics.

1.2 Prior Work and Comparison of Results

Gaussian mixture models are among the most widely used probabilistic models in statistical inference [38, 41, 42]. Algorithmic results fall into two broad classes — separation-based results, and moment-based methods that do not assume explicit geometric separation.

Separation-based results. The body of work that is most relevant to this paper assumes that there is some minimum separation between the means of the components in the mixture. The first polynomial time algorithmic guarantees for mixtures of Gaussians were given by Dasgupta [16], who showed how to learn mixtures of spherical Gaussians when the separation is of the order of $d^{1/2}$. This was later improved by a series of works [4, 45, 2, 29, 17, 14, 30, 7] for both spherical Gaussians and general Gaussians. The algorithm of Vempala and Wang [45] gives the best known result, and uses PCA along with distance-based clustering to learn mixtures of spherical Gaussians with separation

$$\|\mu_i - \mu_j\|_2 \geq (\min\{k, d\}^{1/4} \log^{1/4}(dk/\delta) + \log^{1/2}(dk/\delta))(\sigma_i + \sigma_j).$$

We note that all these clustering-based algorithms require a separation that either implicitly or explicitly depend on the estimation accuracy δ.[1] Finally, although not directly related to our work, we note that a similar separation condition was shown to suffice also for *non-spherical* Gaussians [14], where separation is measured based on the variance along the direction of the line joining the respective means (as opposed, e.g., to the sum of maximum variances $\|\Sigma_i\| + \|\Sigma_j\|$ which could be much larger).

Iterative methods like Expectation Maximization (EM) and Lloyd's algorithm (sometimes called the k-means heuristic) are commonly used in practice to learn mixtures of spherical Gaussians but, as mentioned above, are notoriously hard to analyze. Dasgupta and Schulman [17] proved that a variant of the EM algorithm learns mixtures of spherical Gaussians with separation of the order of $d^{1/4}\text{polylog}(dk)$. Kumar and Kannan [30] and subsequent work [7] showed that the spectral clustering heuristic (i.e., PCA followed by Lloyd's algorithm) provably recovers the clusters in a rather wide family of distributions which includes non-spherical Gaussians; in the special case of spherical Gaussians, their analysis requires separation of order $\sqrt{k}$.

Very recently, the EM algorithm was shown to succeed for mixtures of $k = 2$ spherical Gaussians with $\Omega(\sigma)$ separation [8, 46, 19] (we note that in this setting with $k = O(1)$, polynomial time guarantees are also known using other algorithms like the method-of-moments [28], as we will see in the next paragraph). SDP-based algorithms have also been studied in the context of learning mixtures of spherical Gaussians with a similar separation requirement of $\Omega(k \max_i \sigma_i)$ [34]. The question of how much separation between the components is necessary was also studied empirically by Srebro et al. [39], who observed that iterative heuristics successfully learn the parameters under much smaller separation compared to known theoretical bounds.

Moment-based methods. In a series of influential results, algorithms based on the method-of-moments were developed by [28, 35, 9] for efficiently learning mixtures of $k = O(1)$ Gaussians under arbitrarily small separation. To perform parameter estimation up to accuracy δ, the running time of the algorithms is $\text{poly}(d, 1/w_{\min}, 1/\delta)^{O(k^2)}$ (this holds for mixtures of general Gaussians). This exponential dependence on k is necessary in general, due to statistical lower bound results [35]. The running time dependence on δ was improved in the case of $k = 2$ Gaussians in [26].

Recent work [27, 11, 25, 10, 3, 24] use uniqueness of tensor decompositions (of order 3 and above) to implement the method of moments and give polynomial time algorithms assuming the means are sufficiently high dimensional, and do not lie in certain degenerate configurations. Hsu and Kakade [27] gave a polynomial time algorithm based on tensor decompositions to learn a mixture of spherical Gaussians, when the means are linearly independent. This was extended by [25, 10, 3] to give smoothed analysis guarantees to learn "most" mixtures of spherical Gaussians when the means

[1] Such a dependency on δ seems necessary for clustering-based algorithms that cluster every point accurately with high probability.

are in $d = k^{\Omega(1)}$ dimensions. These algorithms do not assume any strict geometric separation conditions and learn the parameters in $\text{poly}(k, d, 1/\delta)$ time (and samples), when these non-degeneracy assumptions hold. However, there are many settings where the Gaussian mixture consists of many clusters in a low dimensional space, or have their means lying in a low-dimensional subspace or manifold, where these tensor decomposition guarantees do not apply. Besides, these algorithms based on tensor decompositions seem less robust to noise than clustering-based approaches and iterative algorithms, giving further impetus to the study of the latter algorithms as we do in this paper.

Lower Bounds. Moitra and Valiant [35] showed that $\exp(k)$ samples are needed to learn the parameters of a mixture of k Gaussians [35]. In fact, the lower bound instance of [35] is one dimensional, with separation of order $1/\sqrt{k}$. Anderson et al. [3] proved a lower bound on sample complexity that is reminiscent of our Theorem 1.2. Specifically, they obtain a super-polynomial lower bound assuming separation $O(\sigma/\text{poly}\log(k))$ for $d = O(\log k/\log\log k)$. This is in contrast to our lower bound which allows separation greater than σ, or $o(\sigma\sqrt{\log k})$ to be precise.

Other Related work. While most of the previous work deals with estimating the parameters (e.g., means) of the Gaussians components in the given mixture $\mathcal{G}$, a recent line of work [23, 40, 18] focuses on the task of learning a mixture of Gaussians $\mathcal{G}'$ (with possibly very different parameters) that is close in statistical distance i.e., $\|\mathcal{G} - \mathcal{G}'\|_{TV} < \delta$ (this is called "proper learning", since the hypothesis that is output is also a mixture of k Gaussians). When identifiability using polynomial samples is known for the family of distributions, proper learning automatically implies parameter estimation. Algorithms for proper learning mixtures of spherical Gaussians [40, 18] give polynomial sample complexity bounds when $d = 1$ (note that the lower bounds of [35] do not apply here) and have run time dependence that is exponential in k; the result of [23] has sample complexity that is polynomial in d but exponential in k. Algorithms that take time and samples $\text{poly}(k, 1/\delta)^d$ are also known for "improper learning" and density estimation for mixtures of k Gaussians (the hypothesis class that is output may not be a mixture of k Gaussians) [15, 12]. We note that known algorithms have sample complexity that is either exponential in d or k, even though proper learning and improper learning are easier tasks than parameter estimation. To the best of our knowledge, better bounds are not known under additional separation assumptions.

A related problem in the context of clustering graphs and detecting communities is the problem of learning a stochastic block model or planted partitioning model [33]. Here, a sharp threshold phenomenon involving an analogous separation condition (between intra-cluster probability of edges and inter-cluster edge probability) is known under various settings [36, 32, 37] (see the recent survey by Abbe [1] for details). In fact, the algorithm of Kumar and Kannan [30] give a general separation condition that specializes to separation between the means for mixtures of Gaussians, and separation between the intra-cluster and inter-cluster edge probabilities for the stochastic block model.

1.3 Overview of Techniques

Lower bound for $O(\sqrt{\log k})$ separation. The sample complexity lower bound proceeds by showing a more general statement: in any large enough collection of uniform mixtures, for all but a small fraction of the mixtures, there is at least one other mixture in the collection that is close in statistical distance (see Theorem 3.2). For our lower bounds, we will just produce a large collection of uniform mixtures of well-separated spherical Gaussians in $d = c\log k$ dimensions, whose pairwise parameter distances are reasonably large. In fact, we can even pick the means of these mixtures randomly in a ball of radius $\sqrt{d}$ in $d = c \log k$ dimensions; w.h.p. most of these mixtures will need at least $k^{\omega(1)}$ samples to identify.

To show the above pigeonhole style statement about large collections of mixtures, we will associate with a uniform mixture having means $\mu_1, \ldots, \mu_k$, the following quantities that we call "mean moments," and we will use them as a proxy for the actual moments of the distribution:

$$(M_1, \ldots, M_R) \text{ where } \forall 1 \leq r \leq R : M_r = \frac{1}{k} \sum_{j=1}^{k} \mu_j^{\otimes r}.$$

The mean moments just correspond to the usual moments of a mixture of delta functions centered at $\mu_1, \ldots, \mu_k$. Closeness in the first $R = O(1/\varepsilon)$ mean moments (measured in injective tensor norm) implies that the two corresponding distributions are ε close in statistical distance (see Lemma 3.7 and Lemma 3.8). The key step in the proof uses a careful packing argument to show that for most mixtures in a large enough collection, there is a different mixture in the collection that approximately matches in the first R mean moments (see Lemma 3.6).

Iterative Algorithm. Our iterative algorithm will function in both settings of interest: the high-dimensional setting when we have $\Omega(\sqrt{\log k})$ separation, and the low-dimensional setting when $d < \log k$ and we have $\Omega(\sqrt{d})$ separation. For the purpose of this description, let us assume δ is arbitrarily small compared to d and k. In our proposed algorithm, we will consider distributions obtained by restricting the support to just certain regions around the initializers $z_1 = \tilde{\mu}_1, \ldots, z_k = \tilde{\mu}_k$ that are somewhat close to the means $\mu_1, \mu_2, \ldots, \mu_k$ respectively. Roughly speaking, we first partition the space into a Voronoi partition given by $\{z_j : j \in [k]\}$, and then for each component $j \in [k]$ in $\mathcal{G}$, let S_j denote the region containing z_j (see Definition 4.3 for details). For each $j \in [k]$ we consider only the samples in the set S_j and let $u_j \in \mathbb{R}^d$ be the (sample) mean of these points in S_j, after subtracting z_j.

The regions are chosen in such a way that S_j has a large fraction of the probability mass from the jth component, and the total probability mass from the other components is relatively small (it will be at most $1/\text{poly}(k)$ with $\Omega(\sqrt{\log k})$ separation, and $O_d(1)$ with $\Omega(1)$ separation in constant dimensions). However, since δ can be arbitrarily small functions of k, d, there can still be a relatively large contribution from the other components. For instance, in the low-dimensional case with $O(1)$ separation, there can be $\Omega(1)$ mass from a single neighboring component! Hence, u_j does not give a δ-close estimate for μ_j (even up to scaling), unless the separation is at least of order $\sqrt{\log(1/\delta)}$ – this is too large when $\delta = k^{-\omega(1)}$ with $\sqrt{\log k}$ separation, or $\delta = o_d(1)$ with $\Omega(1)$ separation in constant dimensions.

Instead we will use these statistics to set up a system of non-linear equations where the unknowns are the true parameters and solve for them using the Newton method. We will use the initializers $z_j = \mu_j^{(0)}$, to define the statistics that give our equations. Hence the unknown parameters $\{\mu_i : i \in [k]\}$ satisfy the following equation for each $j \in [k]$:

$$\sum_{i=1}^{k} w_i \int_{y \in S_j} (y - z_j) \cdot \sigma_j^{-d} \exp\left(-\frac{\pi \|y - \mu_i\|_2^2}{\sigma_i^2}\right) dy = u_j. \tag{4}$$

Note that in the above equation, the only unknowns or variables are the true means $\{\mu_i : i \in [k]\}$. After scaling the equations, and a suitable change of variables $\mathbf{x}_j = \mu_j/\sigma_j$ to make the system "dimensionless" we get a non-linear system of equations denoted by $F(\mathbf{x}) = b$. For the above system, $\mathbf{x}_i^* = \mu_i/\sigma_i$ represents a solution to the system given by the parameters of $\mathcal{G}$. The Newton algorithm uses the iterative update

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + (F'(\mathbf{x}^{(t)}))^{-1}(b - F(\mathbf{x}^{(t)})).$$

For the Newton method we need access to the estimates for b, and the derivative matrix F' (the Jacobian) evaluated at $\mathbf{x}^{(t)}$. The derivative of the j equation w.r.t. $\mathbf{x}_i$ corresponds to

$$\nabla_{\mathbf{x}_i} F_j(\mathbf{x}) = \frac{2\pi w_i}{w_j \sigma_j \sigma_i} \int_{y \in S_j} (y - z_j)(y - \sigma_i \mathbf{x}_i)^T g_{\sigma_i \mathbf{x}_i, \sigma_i}(y)\, dy \ ,$$

where $g_{\sigma_i \mathbf{x}_i, \sigma_i}(y)$ represents the p.d.f. at a point y due to a spherical Gaussian with mean at $\sigma_i \mathbf{x}_i$ and covariance $\sigma_i^2/(2\pi)$ in each direction. Unlike usual applications of the Newton method, we do not have closed form expressions for F' (the Jacobian), due to our definition of the set S_j. However, we will instead be able to estimate the Jacobian at $\mathbf{x}^{(t)}$ by calculating the above expression (RHS) by considering a Gaussian with mean $\sigma_i \mathbf{x}_i^{(t)}$ and variance $\sigma_i^2/(2\pi)$. The Newton method can be shown to be robust to errors in b, F, F' (see Theorem B.2).

We want to learn each of the k means up to good accuracy; hence we will measure the error and convergence in $\|\cdot\|_\infty$ norm. This is important in low dimensions since measuring convergence in ℓ_2 norm will introduce extra $\sqrt{k}$ factors, that are prohibitive for us since the means are separated only by $\Theta_d(1)$. The convergence of the Newton's method depends on upper bounding the operator norm of the inverse of the Jacobian $\|(F')^{-1}\|$ and the second-derivative $\|F''\|$, with the initializer being chosen δ-close to the true parameters so that $\delta \|(F')^{-1}\| \|F''\| < 1/2$.

The main technical effort for proving convergence is in showing that the inverse $(F')^{-1}$ evaluated at any point in the neighborhood around $\mathbf{x}^*$ is well-conditioned. We will show the convergence of the Newton method by showing "diagonal dominance" properties of the $dk \times dk$ matrix F'. This uses the separation between the means of the components, and the properties of the region S_j that we have defined. For $\Omega(\sqrt{\log k})$ separation, this uses standard facts about Gaussian concentration to argue that each of the $(k-1)$ off-diagonal blocks (in the jth row of F') is at most $1/(2k)$ factor of the corresponding diagonal term. With $\Omega(1)$ separation in $d = O(1)$ dimensions, we can not hope to get such a uniform bound on all the off-diagonal blocks (a single off-diagonal block can itself be $\Omega_d(1)$ times the corresponding diagonal entry). We will instead use careful packing arguments to show that the required diagonal dominance condition (see Lemma 4.13 for a statement).

Hence, the initializers are used to both define the regions S_j, and as initialization for the Newton method. Using this diagonal dominance in conjunction with initializers (Theorem 5.2) and the (robust) guarantees of the Newton method (Theorem B.2) gives rapid convergence to the true parameters.

2 Preliminaries

Consider a mixture of k spherical Gaussians $\mathcal{G}$ in $\mathbb{R}^d$ that has parameters $\{(w_j, \mu_j, \sigma_j) : j \in [k]\}$. The jth component has mean μ_j and covariance $\sigma_j^2/2\pi \cdot I_{d\times d}$. For $\mu \in \mathbb{R}^d, \sigma \in \mathbb{R}_+$, let $g_{\mu,\sigma} : \mathbb{R}^d \to \mathbb{R}_+$ represent the p.d.f. of a spherical Gaussian centered at μ and with covariance $\sigma^2/(2\pi) \cdot I_{d\times d}$. We will use f to represent the p.d.f. of the mixture of Gaussians $\mathcal{G}$, and g_j to represent the p.d.f. of the jth Gaussian component.

Definition 2.1 (Standard mixtures of Gaussians)**.** A *standard* mixture of k Gaussians with means $\mu_1, \dots, \mu_k \in \mathbb{R}^d$ is a mixture of k spherical Gaussians $\{(\frac{1}{k}, \mu_j, 1) : j \in [k]\}$.

A standard mixture is just a uniform mixture of spherical Gaussians with all covariances $\sigma^2 = 1/(2\pi)$. Before we proceed, we define the following notion of parameter "distance" between mixtures of Gaussians:

Definition 2.2 (Parameter distance)**.** Given two mixtures of Gaussians in $\mathbb{R}^d$, $\mathcal{G} = \{(w_j, \mu_j, \sigma_j) : j \in [k]\}$ and $\mathcal{G}' = \{(w'_j, \mu'_j, \sigma'_j) : j \in [k]\}$, define

$$\Delta_{\text{param}}\left(\mathcal{G}, \mathcal{G}'\right) = \min_{\pi \in \text{Perm}_k} \sum_{j=1}^{k} \frac{|w_j - w_{\pi(j)}|}{\min\{w_j, w_{\pi(j)}\}} + \sum_{j=1}^{k} \frac{\|\mu_j - \mu'_{\pi(j)}\|_2}{\min\{\sigma_j, \sigma'_{\pi(j)}\}} + \sum_{j=1}^{k} \frac{|\sigma_j - \sigma'_{\pi(j)}|}{\min\{\sigma_j, \sigma'_{\pi(j)}\}} \ .$$

For *standard* mixtures, the definition simplifies to

$$\Delta_{\text{param}}\big((\mu_1, \dots, \mu_k), (\mu'_1, \dots, \mu'_k)\big) = \min_{\pi \in \text{Perm}_k} \sum_{j=1}^{k} \|\mu_j - \mu'_{\pi(j)}\|_2 \ .$$

Note that this definition is invariant to scaling the variances (for convenience). We note that parameter distance is not a metric, but it is just a convenient way of measure closeness of parameters between two distributions.

The distance between two individual Gaussian components can also be measured in terms of the total variation distance between the components [35]. For instance, in the case of standard spherical Gaussians, a parameter distance of $c\sqrt{\log k}$ corresponds to a total variation distance of $k^{-O(c^2)}$.

Definition 2.3 (ρ-bounded mixtures)**.** For $\rho \geq 1$, a mixture of spherical Gaussians $\mathcal{G} = \{(w_j, \mu_j, \sigma_j)\}_{j=1}^{k}$ in $\mathbb{R}^d$ is called *ρ-bounded* if for each $j \in [k]$, $\|\mu_j\|_2 \leq \rho$ and $\frac{1}{\rho} \leq \sigma_j \leq \rho$. In particular, a *standard* mixture is ρ-bounded if for each $j \in [k]$, $\|\mu_j\|_2 \leq \rho$.

Also, for a given mixture of k spherical gaussians $\mathcal{G} = \{(w_j, \mu_j, \sigma_j) : j \in [k]\}$, we will denote $w_{\min} = \min_{j \in [k]} w_j$, $\sigma_{\max} = \max_{j \in [k]} \sigma_j$ and $\sigma_{\min} = \min_{j \in [k]} \sigma_j$.

In the above notation the bound ρ can be thought of as a sufficiently large polynomial in k, since we are aiming for bounds that are polynomial in k. Since we can always scale the points by an arbitrary factor without affecting the performance of the algorithm, we can think of ρ as the (multiplicative) range of values taken by the parameters $\{\mu_i, \sigma_i : i \in [k]\}$. Since we want separation bounds independent of k, we will denote individual aspect ratios for variances and weights given by $\rho_\sigma = \max_{i \in [k]} \sigma_i / \min_{i \in [k]} \sigma_i$, and $\rho_w = \max_{i \in [k]} w_i / \min_{i \in [k]} w_i$.

Finally, we list some of the conventions used in this paper. We will denote by $N(0, \sigma^2)$ a normal random variable with mean 0 and variance σ^2. For $x \in \mathbb{R}$ generated according to $N(0, \sigma^2)$, let $\tilde{\Phi}_{0,\sigma}(t)$ denote the probability that $x > t$, and let $\tilde{\Phi}^{-1}_{0,\sigma}(y)$ denote the quantile t at which $\tilde{\Phi}_{0,\sigma}(t) \leq y$. For any function $f : \mathbb{R}^d \to \mathbb{R}$, f' will denote the first derivative (or gradient) of the function, and f'' will denote the second derivative (or Hessian). We define $\|f\|_{1,S} = \int_S |f(x)| dx$ to be the L_1 norm of f restricted to the set S. Typically, we will use indices i, j to represent one of the k components of the mixture, and we will use r (and s) for coordinates. For a vector $x \in \mathbb{R}^d$, we will use $x(r)$ denote the rth coordinate. Finally, we will use *w.h.p.* in statements about the success of algorithms to represent probability at least $1 - \gamma$ where $\gamma = (d + k)^{-\Omega(1)}$.

Norms. For any $p \geq 1$, given a matrix $M \in \mathbb{R}^{d \times d}$, we define the matrix norm

$$\|M\|_{p \to p} = \max_{x \in \mathbb{R}^d : \|x\|_p = 1} \|Mx\|_p.$$

2.1 Notation and Preliminaries about Newton's method

Consider a system of m non-linear equations in variables $u_1, u_2, \dots, u_m$:

$$\forall j \in [m], f_j(u_1, \dots, u_m) = b_j.$$

Let $F' = J(u) \in \mathbb{R}^{m \times m}$ be the Jacobian of the system given by the non-linear functional $f : \mathbb{R}^m \to \mathbb{R}^m$, where the $(j,i)^{th}$ entry of J is the partial derivative $\frac{\partial f_j(u)}{\partial u_i}$ is evaluated at u. Newton's method starts with an initial point $u^{(0)}$, and updates the solution using the iteration:

$$u^{(t+1)} = u^{(t)} + \left(J(u^{(t)})\right)^{-1} \left(b_j - f(u^{(t)})\right).$$

Standard results shows quadratic convergence of the Newton method for general normed spaces [5]. We restrict our attention in the restricted setting where both the range and domain of f is $\mathbb{R}^m$, equipped with an appropriate norm $\|\cdot\|$ to measure convergence.

Theorem 2.4 (Theorem 5.4.1 in [5]). *Assume $u^* \in \mathbb{R}^m$ is a solution to the equation $f(y) = b$ where $f : \mathbb{R}^m \to \mathbb{R}^m$ and the inverse Jacobian J^{-1} exists in a neighborhood $N = \{u : \|u - u^*\| \leq \|u^{(0)} - u^*\|\}$, and $F' : \mathbb{R}^m \to \mathbb{R}^{m \times m}$ is locally L-Lipschitz continuous in the neighborhood N i.e., $\forall u, v \in N, \ \|F'(u) - F'(v)\| \leq L\|u - v\|$. Then we have $\|u^{(t+1)} - u^*\| \leq L \cdot \|J(u^{(t)})^{-1}\| \cdot \|u^{(t)} - u^*\|^2$.*

In particular, for Newton's method to work, $\|u^0 - u^*\| \leq (L \max_{u \in \mathcal{N}} \|J(u)^{-1}\|)^{-1}$ will guarantee convergence. A statement of the robust convergence of Newton's method in the presence of estimates is given in Theorem B.2 and Corollary B.4.

We want to learn each of the k sets of parameters up to good accuracy; hence we will measure the error in ℓ_∞ norm. To upper bound $\|J^{-1}\|_{\infty \to \infty}$, we will use *diagonal dominance* properties of the matrix J. Note that $\|A\|_{\infty \to \infty}$ is just the maximum ℓ_1 norm of the rows of A. The following lemma bound $\|A^{-1}\|_{\infty \to \infty}$ for a diagonally dominant matrix A.

Lemma 2.5 ([44]). *Consider any square matrix A of size $n \times n$ satisfying*

$$\forall i \in [n] \quad a_{ii} - \sum_{j \neq i} |a_{ij}| \geq \alpha.$$

Then, $\|A^{-1}\|_{\infty \to \infty} \leq 1/\alpha$.

3 Lower Bounds with $O(\sqrt{\log k})$ Separation

Here we show a sample complexity lower bound for learning standard mixtures of k spherical Gaussians even when the separation is of the order of $\sqrt{\log k}$. In fact, this lower bound will also hold for a *random* mixture of Gaussians in $d \leq c \cdot \log k$ dimensions (for sufficiently small constant c) with high probability.[2]

[2] In particular, this rules out polynomial-time smoothed analysis guarantees of the kind shown for $d = k^{\Omega(1)}$ in [10, 3].

Theorem 3.1. *For any large enough C there exist $c, c_2 > 0$, such that the following holds for all $k \geq C^8$. Let $\mathcal{D}$ be the distribution over standard mixtures of k spherical Gaussians obtained by picking each of the k means independently and uniformly from a ball of radius $\sqrt{d}$ around the origin in $d = c \log k$ dimensions. Let $\{\mu_1, \mu_2, \ldots, \mu_k\}$ be a mixture chosen according to $\mathcal{D}$. Then with probability at least $1 - 2/k$ there exists another standard mixture of k spherical Gaussians with means $\{\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_k\}$ such that both mixtures are $\sqrt{d}$ bounded and well separated, i.e.,*

$$\forall i, j \in [k], i \neq j : \|\mu_i - \mu_j\| \geq c_2 \sqrt{\log k} \quad \text{and} \quad \|\tilde{\mu}_i - \tilde{\mu}_j\| \geq c_2 \sqrt{\log k},$$

and their p.d.f.s satisfy

$$\|f - \tilde{f}\|_1 \leq k^{-C} \tag{5}$$

even though their parameter distance is at least $c_2\sqrt{\log k}$. Moreover, we can take $c = 1/(4 \log C)$ and $c_2 = C^{-24}$.

Remark. In Theorem 3.1, there is a trade-off between getting a smaller statistical distance $\varepsilon = k^{-C}$, and a larger separation between the means in the Gaussian mixture. When $C = \omega(1)$, with $c_1, c = o(1)$ we see that $\|f - \tilde{f}\|_1 \leq k^{-\omega(1)}$ when the separation is $o(\sqrt{\log k})\sigma$. On the other hand, we can also set $C = k^{\varepsilon'}$ (for some small constant $\varepsilon' > 0$) to get lower bounds for mixtures of spherical Gaussians in $d = 1$ dimension with $\|f - \tilde{f}\|_1 = \exp(-k^{\Omega(1)})$ and separation $1/k^{O(1)}$ between the means. We note that in this setting of parameters, our lower bound is nearly identical to that in [35]. Namely, they achieve statistical distance $\|f - \tilde{f}\|_1 = \exp(-\Omega(k))$ (which is better than ours) with a similar separation of $k^{-O(1)}$ in one dimension. One possible advantage of our bound is that it holds with a random choice of means, unlike their careful choice of means.

3.1 Proof of Theorem 3.1

The key to the proof of Theorem 3.1 is the following pigeonhole statement, which can be viewed as a bound on the covering number (or equivalently, the metric entropy) of the set of Gaussian mixtures.

Theorem 3.2. *Suppose we are given a collection $\mathcal{F}$ of standard mixtures of spherical Gaussians in d dimensions that are $\rho = \sqrt{d}$ bounded, i.e., $\|\mu_j\| \leq \sqrt{d}$ for all $j \in [k]$. There are universal constants $c_0, c_1 \geq 1$, such that for any $\eta > 0, \varepsilon \leq \exp(-c_1 d)$, if*

$$|\mathcal{F}| > \frac{1}{\eta} \exp\left(c_0 \Big(\frac{\log(1/\varepsilon)}{d}\Big)^d \cdot \log(1/\varepsilon) \log(5d) \right), \tag{6}$$

then for at least $(1 - \eta)$ fraction of the mixtures $\{\mu_1, \mu_2, \ldots, \mu_k\}$ from $\mathcal{F}$, there is another mixture $\{\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_k\}$ from $\mathcal{F}$ with p.d.f. $\tilde{f}$ such that $\|f - \tilde{f}\|_1 \leq \varepsilon$. Moreover, we can take $c_0 = 8\pi e$ and $c_1 = 36$.

Remark 3.3. Notice that k plays no role in the statement above. In fact, the proof also holds for mixtures with arbitrary number of components and arbitrary weights.

Claim 3.4. *Let $x_1, \ldots, x_N$ be chosen independently and uniformly from the ball of radius r in $\mathbb{R}^d$. Then for any $0 < \gamma < 1$, with probability at least $1 - N^2\gamma^d$, we have that for all $i \neq j$, $\|x_i - x_j\| \geq \gamma r$.*

Proof. For any fixed $i \neq j$, the probability that $\|x_i - x_j\| \geq \gamma r$ is at most γ^d, because the volume of a ball of radius γr is γ^d times that of a ball of radius r. The claim now follows by a union bound. □

Proof of Theorem 3.1. Set $\gamma := 2^{-6/c}$, and consider the following probabilistic procedure. We first let $\mathcal{X}$ be a set of $(1/\gamma)^{d/3}$ points chosen independently and uniformly from the ball of radius $\sqrt{d}$. We then output a mixture chosen uniformly from the collection $\mathcal{F}$, defined as the collection of all standard mixtures of spherical Gaussians obtained by selecting k distinct means from $\mathcal{X}$. Observe that the output of this procedure is distributed according to $\mathcal{D}$. Our goal is therefore to prove that with probability at least $1-2/k$, the output of the procedure satisfies the property in the theorem.

First, by Claim 3.4, with probability at least $1-\gamma^{d/3} \geq 1-1/k$, any two points in $\mathcal{X}$ are at distance at least $\gamma\sqrt{d}$. It follows that in this case, the means in any mixture in $\mathcal{F}$ are at least $\gamma\sqrt{d}$ apart, and also that any two distinct mixtures in $\mathcal{F}$ have a parameter distance of at least $\gamma\sqrt{d}$ since they must differ in at least one of the means. Note that $\gamma = C^{-24}$ for our choice of c, γ.

To complete the proof, we notice that by our choice of parameters, and denoting $\varepsilon = k^{-C}$,

$$|\mathcal{F}| = \binom{|\mathcal{X}|}{k} \geq \Big(\frac{1}{\gamma}\Big)^{dk/3} \cdot k^{-k} = k^k \geq k \cdot \exp\left(c_0 \Big(\frac{\log(1/\varepsilon)}{d}\Big)^d \cdot \log(1/\varepsilon)\log(5d)\right) .$$

The last inequality follows since for our choice of $\varepsilon = k^{-C}$, $c = \frac{1}{4\log C}$ and C is large enough with $C \geq c_0$, so that

$$\left(\frac{\log(1/\varepsilon)}{d}\right)^d = k^{c\log(C/c)} < \sqrt{k}, \quad \text{and} \;\; c_0 \log(1/\varepsilon)\log(5d) \leq c_0 C \log k \log(5c\log k) < \sqrt{k}.$$

Hence applying Theorem 3.2 to $\mathcal{F}$, for at least $1-1/k$ fraction of the mixtures in $\mathcal{F}$, there is another mixture in $\mathcal{F}$ that is ε close in total variation distance. We conclude that with probability at least $1-2/k$, a random mixture in $\mathcal{F}$ satisfies all the required properties, as desired. □

3.2 Proof of Theorem 3.2

It will be convenient to represent the p.d.f. $f(x)$ of the standard mixture of spherical Gaussians with means $\mu_1, \mu_2, \dots, \mu_k$ as a convolution of a standard mean zero Gaussian with a sum of delta functions centered at $\mu_1, \mu_2, \dots, \mu_k$,

$$f(x) = \left(\frac{1}{k}\sum_{j=1}^{k} \delta(x-\mu_j)\right) * e^{-\pi\|x\|_2^2}.$$

Instead of considering the moments of the mixture of Gaussians, we will consider moments of just the corresponding mixture of delta functions at the means. We will call them "mean moments," and we will use them as a proxy for the actual moments of the distribution.

$$(M_1, \dots, M_R) \text{ where } \; \forall 1 \leq r \leq R : M_r = \frac{1}{k}\sum_{j=1}^{k} \mu_j^{\otimes r}.$$

To prove Theorem 3.2 we will use three main steps. Lemma 3.6 will show using the pigeonhole principle that for any large enough collection of Gaussian mixtures $\mathcal{F}$, most Gaussians mixtures in the family have other mixtures which approximately match in their first $R = O(\log(1/\varepsilon))$ mean moments. This closeness in moments will be measured using the *symmetric injective tensor norm* defined for an order-ℓ tensor $T \in \mathbb{R}^{d^\ell}$ as

$$\|T\|_* = \max_{\substack{y\in\mathbb{R}^d \\ \|y\|=1}} |\langle T, y^{\otimes \ell}\rangle|.$$

Next, Lemma 3.7 shows that the two distributions that are close in the first R mean moments are also close in the L_2 distance. This translates to small statistical distance between the two distributions using Lemma 3.8.

We will use the following standard packing claim.

Claim 3.5. *Let $\|\cdot\|$ be an arbitrary norm on $\mathbb{R}^D$. If $x_1, \ldots, x_N \in \mathbb{R}^D$ are such that $\|x_i\| \le \Delta$ for all i, and for all $i \neq j$, $\|x_i - x_j\| > \delta$, then $N \le (1 + 2\Delta/\delta)^D$. In particular, if $x_1, \ldots, x_N \in \mathbb{R}^D$ are such that $\|x_i\| \le \Delta$ for all i, then for all but $(1 + 2\Delta/\delta)^D$ of the indices $i \in [N]$, there exists a $j \neq i$ such that $\|x_i - x_j\| \le \delta$.*

Proof. Let K be the unit ball of the norm $\|\cdot\|$. Then by assumption, the sets $x_i + \delta K/2$ for $i \in [N]$ are disjoint. But since they are all contained in $(\Delta + \delta/2)K$,

$$N \le \frac{\mathrm{Vol}((\Delta + \delta/2)K)}{\mathrm{Vol}(\delta K/2)} = (1 + 2\Delta/\delta)^D \ .$$

The "in particular" part follows by taking a maximal set of δ-separated points. □

Lemma 3.6. *Suppose we are given a set $\mathcal{F}$ of standard mixtures of spherical Gaussians in d dimensions with means of length at most $\sqrt{d}$. Then for any integer $R \ge d$, if $|\mathcal{F}| > \frac{1}{\eta} \cdot \exp\big((2eR/d)^d R \log(5d)\big)$, it holds that for at least $(1-\eta)$ fraction of the mixtures $\{\mu_1, \mu_2, \ldots, \mu_k\}$ in $\mathcal{F}$, there is another mixture $\{\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_k\}$ in $\mathcal{F}$ satisfying that for $r = 1, \ldots, R$,*

$$\Big\| \frac{1}{k} \sum_{j=1}^{k} \mu_j^{\otimes r} - \frac{1}{k} \sum_{j=1}^{k} (\tilde{\mu}_j)^{\otimes r} \Big\|_* \le (d+1)^{-R/4}. \tag{7}$$

Proof. With any choice of means $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$ we can associate a vector of moments $\psi(\mu_1, \mu_2, \ldots, \mu_k) = (M_1, \ldots, M_R)$ where for $r = 1, \ldots, R$,

$$M_r = \frac{1}{k} \sum_{j=1}^{k} \mu_j^{\otimes r} \ .$$

Notice that the image of ψ lies in a direct sum of symmetric subspaces whose dimension is $D = \binom{d+R}{R}$ (i.e., the number of ways of distributing R identical balls into $(d+1)$ different bins). Since $R \ge d$,

$$D = \binom{d+R}{R} \le \left(\frac{e(d+R)}{d}\right)^d \le \left(\frac{2eR}{d}\right)^d . \tag{8}$$

We define a norm on these vectors $\psi(\mu_1, \ldots, \mu_k)$ in terms of the maximum injective norm of the mean moments,

$$\|\psi(\mu_1, \ldots, \mu_k)\|_{*,\infty} = \max_{r \in [R]} \Big\| \frac{1}{k} \sum_{j=1}^{k} \mu_j^{\otimes r} \Big\|_* \ . \tag{9}$$

Since each of our means has length at most $\sqrt{d}$, we have that

$$\|\psi(\mu_1, \ldots, \mu_k)\|_{*,\infty} \le \max_{r \in [R]} \Big\| \frac{1}{k} \sum_{j=1}^{k} \mu_j^{\otimes r} \Big\|_* \le \max_{r \in [R]} \max_{\|\mu\| \le \sqrt{d}} \|\mu^{\otimes r}\|_* \le d^{R/2}. \tag{10}$$

Using Claim 3.5, if $|\mathcal{F}| > N/\eta$ where

$$N = \Big(1 + 2d^{R/2}/(d+1)^{-R/4}\Big)^D \le \Big(1 + 2(d+1)^{3R/4}\Big)^D \le \exp\left(\frac{3}{4}(2eR/d)^d \cdot R\log(5d)\right),$$

we have that for at least $1-\eta$ fraction of the Gaussian mixtures in $\mathcal{F}$, there is another mixture from $\mathcal{F}$ which is $(d+1)^{-R/4}$ close, as required. □

Next we show that the closeness in moments implies closeness in the L_2 distance. This follows from fairly standard Fourier analytic techniques. We will first show that if the mean moments are close, then the low-order Fourier coefficients are close. This will then imply that the Fourier spectrum of the corresponding Gaussian mixtures f and $\tilde{f}$ are close.

Lemma 3.7. *Suppose $f(x), \tilde{f}(x)$ are the p.d.f. of $\mathcal{G}, \tilde{\mathcal{G}}$ which are both standard mixtures of k Gaussians in d dimensions with means $\{\mu_j : j \in [k]\}$ and $\{\tilde{\mu}_j : j \in [k]\}$ respectively that are both $\rho = \sqrt{d}$ bounded. There exist universal constants $c_1, c_0 \ge 1$, such that for every $\varepsilon \le \exp(-c_1 d)$ if the following holds for $R = c_0 \log(1/\varepsilon)$:*

$$\forall 1 \le r \le R,\ \frac{1}{k}\Big\|\sum_{j=1}^k \mu_j^{\otimes r} - \sum_{j=1}^k (\tilde{\mu}_j)^{\otimes r}\Big\|_* \le \varepsilon_r := \varepsilon\Big(\frac{r}{8\pi e\sqrt{\log(1/\varepsilon)}}\Big)^r, \tag{11}$$

then $\|f - \tilde{f}\|_2 \le \varepsilon$.

Proof. We first show that if the moments are very close, the Fourier transform of the two distributions is very close. This translates to a bound on the L_2 distance by Parseval's identity.

Let $g, h : \mathbb{R}^d \to \mathbb{R}$ be defined by

$$h(x) = \frac{1}{k}\Big(\sum_{j=1}^k \delta_{\mu_j}(x) - \sum_{j=1}^k \delta_{\tilde{\mu}_j}(x)\Big), \qquad g(x) = f(x) - \tilde{f}(x) = h(x) * e^{-\pi\|x\|^2}. \tag{12}$$

Since the Fourier transform of a convolution is the product of the Fourier transforms, we have

$$\forall \zeta \in \mathbb{R}^d,\ \widehat{f}(\zeta) = \frac{1}{k}\Big(\sum_{j=1}^k e^{-2\pi \mathrm{i}\langle \zeta, \mu_j\rangle}\Big) \cdot e^{-\pi\|\zeta\|^2} \tag{13}$$

$$\widehat{h}(\zeta) = \frac{1}{k}\Big(\sum_{j=1}^k e^{-2\pi \mathrm{i}\langle \mu_j, \zeta\rangle} - \sum_{j=1}^k e^{-2\pi \mathrm{i}\langle \tilde{\mu}_j, \zeta\rangle}\Big), \qquad \widehat{g}(\zeta) = \widehat{h}(\zeta) \cdot e^{-\pi\|\zeta\|^2}. \tag{14}$$

We will now show that $\int_{\mathbb{R}^d} |\widehat{g}(\zeta)|^2 d\zeta \le \varepsilon^2$. We first note that the higher order Fourier coefficients of g do not contribute much to the Fourier mass of g because of the Gaussian tails. Let $\tau^2 = 4\log(1/\varepsilon)$. Since $\tau^2 \ge (d + 2\sqrt{d\log(16/\varepsilon^2)}) + 2\log(16/\varepsilon^2))/(2\pi)$, using Lemma A.4

$$\int_{\|\zeta\|>\tau} |\widehat{g}(\zeta)|^2 d\zeta = \int_{\|\zeta\|>\tau} |\widehat{h}(\zeta)|^2 e^{-2\pi\|\zeta\|^2} d\zeta \le \int_{\|\zeta\|>\tau} 4e^{-2\pi\|\zeta\|^2} d\zeta \le \frac{\varepsilon^2}{4}. \tag{15}$$

Now we upper bound $|\widehat{h}(\zeta)|$ for $\|\zeta\| < \tau$.

$$|\widehat{h}(\zeta)| = \frac{1}{k}\left|\sum_{j=1}^{k}(e^{-2\pi \mathrm{i}\langle \mu_j,\zeta\rangle} - e^{-2\pi \mathrm{i}\langle \tilde{\mu}_j,\zeta\rangle})\right| = \frac{1}{k}\left|\sum_{j=1}^{k}\sum_{r=1}^{\infty}\frac{(-2\pi \mathrm{i})^r}{r!}\left(\langle \mu_j,\zeta\rangle^r - \langle \tilde{\mu}_j,\zeta\rangle^r\right)\right|$$
$$\leq \sum_{r=1}^{\infty}\frac{(2\pi)^r\|\zeta\|^r}{r!}\cdot\frac{1}{k}\cdot\Big\|\sum_{j=1}^{k}\mu_j^{\otimes r} - \sum_{j=1}^{k}(\tilde{\mu}_j)^{\otimes r}\Big\|_*.$$

We claim that the injective norm above is at most $k\varepsilon_r$ for all $r \geq 1$. For $r \leq R$, this follows immediately from the assumption in (11). For $r \geq R$, we use the fact that the means are $\rho = \sqrt{d}$ bounded,

$$\frac{1}{k}\Big\|\sum_{j=1}^{k}\mu_j^{\otimes r} - \sum_{j=1}^{k}(\tilde{\mu}_j)^{\otimes r}\Big\|_* \leq 2\max_{\mu\in\{\mu_j,\tilde{\mu}_j\}_{j=1}^k}\|\mu\|_2^r \leq 2d^{r/2} \leq 2^{-r/2+1}(2d)^{r/2}$$
$$\leq \varepsilon\cdot\Big(\frac{R^2}{(8\pi e)^2\log(1/\varepsilon)}\Big)^{r/2} \leq \varepsilon\Big(\frac{r}{8\pi e\sqrt{\log(1/\varepsilon)}}\Big)^r = \varepsilon_r,$$

where the last line follows since $R \geq 16\pi e\log(1/\varepsilon)$ and $2d \leq \log(1/\varepsilon)$. Hence, for $\|\zeta\| \leq \tau$, we have

$$|\widehat{h}(\zeta)| \leq \sum_{r=1}^{\infty}\frac{(2\pi)^r\|\zeta\|^r}{r!}\cdot\varepsilon_r \leq \sum_r \frac{(2\pi\|\zeta\|)^r}{\sqrt{2\pi r}(r/e)^r}\cdot\varepsilon\Big(\frac{r}{8\pi e\sqrt{\log(1/\varepsilon)}}\Big)^r$$
$$\leq \varepsilon\sum_{r\geq 1}\frac{1}{\sqrt{2\pi r}}\Big(\frac{2\pi e\|\zeta\|}{r}\cdot\frac{r}{8\pi e\sqrt{\log(1/\varepsilon)}}\Big)^r \leq \varepsilon\sum_{r\geq 1}\frac{1}{\sqrt{2\pi r}}\left(\frac{\|\zeta\|}{2\tau}\right)^r \leq \frac{\varepsilon}{2},$$

since $\|\zeta\| \leq \tau$. Finally, using this bound along with (15) we have

$$\int_{\mathbb{R}^d}|\widehat{g}(\zeta)|^2\,d\zeta \leq \frac{\varepsilon^2}{4}\int_{\|\zeta\|\leq\tau}e^{-2\pi\|\zeta\|^2}d\zeta + \frac{\varepsilon^2}{4} \leq \frac{\varepsilon^2}{4} + \frac{\varepsilon^2}{4} \leq \varepsilon^2.$$

Hence, by Parseval's identity, the lemma follows.

□

The following lemma shows how to go from L_2 distance to L_1 distance using the Cauchy-Schwartz lemma. Here we use the fact that all the means have length at most $\sqrt{d}$. Hence, we can focus on a ball of radius at most $O(\sqrt{\log(1/\varepsilon)})$, since both $f, \tilde{f}$ have negligible mass outside this ball.

Lemma 3.8. *In the notation above, suppose the p.d.f.s $f, \tilde{f}$ of two standard mixtures of Gaussians in d dimensions that are $\sqrt{d}$-bounded (means having length $\leq \sqrt{d}$) satisfy $\|f - \tilde{f}\|_2 \leq \varepsilon$, for some $\varepsilon \leq \exp(-6d)$. Then, $\|f - \tilde{f}\|_1 \leq 2\sqrt{\varepsilon}$.*

Proof. The means are all length at most $\sqrt{d}$, and $\varepsilon < 2^{-d}$. Let us define as before

$$\tau^2 = \frac{1}{2\pi}\big(d+2\sqrt{d\log(2/\varepsilon)}+2\log(2/\varepsilon)\big),\ \text{ and }\ \gamma = (\max\|\mu\|)+\tau \leq 2(\sqrt{d}+\sqrt{\log(2/\varepsilon)}) \leq 4\sqrt{\log(2/\varepsilon)}.$$

Let $g = f - \tilde{f}$ and let $S = \{x : \|x\| \leq \gamma\}$. Using Gaussian concentration in Lemma A.4, we see that both $f, \tilde{f}$ have negligible mass of at most $\varepsilon/2$ each outside S. Hence, using Cauchy-Schwartz inequality and $\|f - \tilde{f}\|_2 \leq \varepsilon$,

$$\int_{\mathbb{R}^d} |g(x)| dx \leq \int_S |g(x)| dx + \varepsilon \leq \sqrt{\int_S g(x)^2 dx} \cdot \sqrt{\int_S dx} + \varepsilon \leq \varepsilon\sqrt{\mathrm{Vol}(S)} + \varepsilon.$$

Since S is a Euclidean ball of radius γ, by Stirling's approximation (Fact A.6), the volume is

$$\mathrm{Vol}(S) = \frac{\pi^{d/2}\gamma^d}{(d/2)!} \leq \left(\frac{2\pi e \gamma^2}{d}\right)^{d/2} \leq \left(\frac{32\pi e \log(2/\varepsilon)}{d}\right)^{d/2} \leq 2^{\log(1/\varepsilon)} = \frac{1}{\varepsilon},$$

where the third inequality follows by raising both sides to the power $2/d$ and using the fact that for $\alpha \geq 6$, $32\pi e(\alpha + 1) \leq 2^{2\alpha}$. This concludes the proof.

□

Proof of Theorem 3.2. The proof follows by a straightforward combination of Lemmas 3.6, 3.7, and 3.8. As stated earlier, we will choose $R = c_0 \log(1/\varepsilon)$, and constants $c_0 = 8\pi e$, $c_1 = 36$. First, by Lemma 3.6 we have that for at least $(1 - \eta)$ fraction of the Gaussian mixtures in $\mathcal{F}$, there is another mixture in the collection whose first $R = c_0 \log(1/\varepsilon)$ mean moments differ by at most $d^{-R/4}$ in symmetric injective tensor norm. To use Lemma 3.7 with $\varepsilon' = \varepsilon^2/4$, we see that

$$\begin{aligned}\min_{r\in[R]} \varepsilon_r &= \frac{\varepsilon^2}{4} \min_{r\in[R]} \left(\frac{r}{8\pi e\sqrt{\log(4/\varepsilon^2)}}\right)^r \\ &\geq \frac{\varepsilon^2}{4} \cdot 2^{-8\pi e\sqrt{2\log(2/\varepsilon)}} \geq 2^{-2\log(2/\varepsilon) - 8\pi e\sqrt{2\log(2/\varepsilon)}} \geq 2^{-2\pi e \log(1/\varepsilon)} \geq 2^{-R/4} \geq (d+1)^{-R/4},\end{aligned}$$

as required, where for the first inequality we used that $2^{-\alpha} < 1/\alpha$ for $\alpha > 0$, and the second inequality uses $\log(1/\varepsilon) \geq c_1 = 36$. We complete the proof by applying Lemma 3.7 (with ε in Lemma 3.7 taking the value $\varepsilon^2/4$) and Lemma 3.8. □

4 Iterative Algorithms for $\min\{\Omega(\sqrt{\log k}), \sqrt{d}\}$ Separation

We now give a new iterative algorithm that estimates the means of a mixture of k spherical Gaussians up to arbitrary accuracy $\delta > 0$ in $\mathrm{poly}(d, k, \log(1/\delta))$ time when the means have separation of order $\Omega(\sqrt{\log k})$ or $\Omega(\sqrt{d})$, when given coarse initializers. In all the bounds that follow, the most interesting setting of parameters is when $1/\delta$ is arbitrarily small compared to k, d (e.g., $1/w_{\min} \leq \mathrm{poly}(k)$ and $\delta = k^{-\omega(1)}$, or when $d = O(1)$ and $\delta = o(1)$). For sake of exposition, we will restrict our attention to the case when the standard deviations σ_i, and weights w_i are known for all $i \in [k]$. We believe that similar techniques can be used to handle unknown σ_i, w_i as well (see Remark 4.6). We also note that the results of this section are interesting even in the case of uniform mixtures, i.e., $w_i = 1/k$ and $\sigma_i = 1$ for all $i \in [k]$.

We will assume that we are given initializers that are inverse polynomially close in parameter distance. These initializers $\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_k^{(0)}$ will be used to set up an "approximate" system of non-linear equations that has sufficient "diagonal dominance" properties, and then use the Newton method with the same initializers to solve it. In what follows, ρ_w and ρ_σ denote the aspect ratio for the weights and variances respectively as defined in Section 2.

Theorem 4.1. *There exist universal constants $c, c_0 > 0$ such that the following holds. Suppose we are given samples from a mixture of k spherical Gaussians $\mathcal{G}$ having parameters $\{(w_j, \mu_j, \sigma_j) : j \in [k]\}$, where the weights and variances are known, satisfying*

$$\forall i \neq j \in [k],\ \|\mu_i - \mu_j\|_2 \geq c(\sigma_i + \sigma_j) \cdot \min\{\sqrt{d} + \sqrt{\log(\rho_w \rho_\sigma)}, \sqrt{\log(\rho_\sigma / w_{min})}\} \tag{16}$$

and suppose we are given initializers $\mu_1^{(0)}, \mu_2^{(0)}, \dots, \mu_k^{(0)}$ satisfying

$$\forall j \in [k],\ \frac{1}{\sigma_j}\|\mu_j^{(0)} - \mu_j\|_2 \leq \frac{c_0}{\min\{d,k\}^{5/2}}. \tag{17}$$

Then for any $\delta > 0$, there is an iterative algorithm that runs in $\text{poly}(\rho, d, 1/w_{min}, 1/\delta)$ time (and samples), and after $T = O(\log\log(d/\delta))$ iterations recovers $\{\mu_j : j \in [k]\}$ up to δ relative error w.h.p. i.e., finds $\{\mu_j^{(T)} : j \in [k]\}$ such that $\forall j \in [k]$, we have $\|\mu_j^{(T)} - \mu_j\|_2/\sigma_j \leq \delta$.

For standard mixtures, (16) corresponds to a separation of order $\min\{\sqrt{\log k}, \sqrt{d}\}$.

Firstly, we will assume without loss of generality that $d \leq k$, since otherwise we can use a PCA-based dimension-reduction result due to Vempala and Wang [45].

Theorem 4.2. *Let $\{(w_i, \mu_i, \sigma_i) : i \in [k]\}$ be a mixture of k spherical Gaussians that is ρ-bounded, and let w_{min} be the smallest mixing weight. Let $\mu'_1, \mu'_2, \dots, \mu'_k$ be the projections onto the subspace spanned by the top k singular vectors of sample matrix $X \in \mathbb{R}^{d \times N}$. For any $\varepsilon > 0$, with $N \leq \text{poly}(d, \rho, w_{min}^{-1}, \varepsilon^{-1})$ samples we have with high probability*

$$\forall i \in [k],\ \|\mu_i - \mu'_i\|_2 \leq \varepsilon.$$

The above theorem is essentially Corollary 3 in [45]. In [45], however, they take the subspace spanned by the top $\max\{k, \log d\}$ singular vectors, most likely due to an artifact of their analysis. We give a different self-contained proof in Appendix C. We will abuse notation, and use $\{\mu_i : i \in [k]\}$ to refer to the means in the dimension reduced space. Note that after dimension reduction, the means are still well-separated for $c' > (c-1)$ i.e.,

$$\forall i, j \in [k],\ \|\mu_i - \mu_j\| \geq c'(\sigma_i + \sigma_j) \min\{\sqrt{d} + \sqrt{\log(\rho_w \rho_\sigma)}, \sqrt{\log(\rho_\sigma / w_{\min})}\}. \tag{18}$$

4.1 Description of the Non-linear Equations and Iterative Algorithm

For each component $j \in [k]$ in $\mathcal{G}$, we first define a region S_j around z_j as follows. We will show in Lemmas 4.8 and 4.9 that the total probability mass in S_j from other components is smaller than the probability mass from the component j.

Definition 4.3 (Region S_j)**.** Given initializers $z_1, z_2, \dots, z_k \in \mathbb{R}^d$, define $\widehat{e}_{j\ell}$ as the unit vector along $z_\ell - z_j$, and let

$$S_j = \{x \in \mathbb{R}^d : \forall \ell \in [k]\ |\langle x - z_j, \widehat{e}_{j\ell}\rangle| \leq 4\sqrt{\log(\rho_\sigma/w_{\min})}\sigma_j, \text{ and } \|x - z_j\|_2 \leq 4(\sqrt{d} + \sqrt{\log(\rho_\sigma \rho_w)})\sigma_j\}. \tag{19}$$

Based on those regions, we define the function $F : \mathbb{R}^{kd} \to \mathbb{R}^{kd}$ by

$$F_j(\mathbf{x}) := \frac{1}{w_j \sigma_j} \sum_{i=1}^{k} w_i \int_{y \in S_j} (y - z_j) g_{\sigma_i \mathbf{x}_i, \sigma_i}(y)\, dy, \tag{20}$$

for $j = 1, \dots, k$, where $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_k) \in \mathbb{R}^{kd}$. We also define $\mathbf{x}^* = (\mu_i/\sigma_i)_{i=1}^k$ as the intended solution. (Notice that for convenience, our variables correspond to "normalized means" μ/σ instead of with the means μ directly.) The system of non-linear equations is then simply

$$F(\mathbf{x}) = b \tag{21}$$

where $b = F(\mathbf{x}^*)$. This is a system of kd equations in kd unknowns. Obviously $\mathbf{x}^*$ is a solution of the system, and if we could find it we would be able to recover all the means, as desired.

Our algorithm basically just applies Newton's method to solve (21) with initializers given by $\mathbf{x}_i^{(0)} = \mu_i^{(0)}/\sigma_i$ for $i \in [k]$. To recall, Newton's method uses the iterative update

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - (F'(\mathbf{x}^{(t)}))^{-1}(b - F(\mathbf{x}^{(t)})),$$

where $F'(\mathbf{x}^{(t)})$ is the first derivative matrix (Jacobian) of F evaluated at $\mathbf{x}^{(t)}$. One issue, however, is that we are not given $b = F(\mathbf{x}^*)$. Nevertheless, as we will show in Lemma 4.4, we can easily estimate it to within any desired accuracy using a Monte Carlo approach based on the given samples (from the Gaussian mixture corresponding to $\mathbf{x}^*$). A related issue is that we do not have a closed-form expression for F and F', but again, we can easily approximate their evaluation at any point $\mathbf{x}$ and to within any desired accuracy using a Monte Carlo approach (by generating samples from the Gaussian mixture corresponding to $\mathbf{x}$). The algorithm is given below in detail.

Iterative Algorithm for Amplifying Accuracy of Parameter Estimation

Input: Estimation accuracy $\delta > 0$, N samples $y^{(1)}, \dots, y^{(N)} \in \mathbb{R}^d$ from a mixture of well-separated Gaussians $\mathcal{G}$ with parameters $\{(w_j, \mu_j, \sigma_j) : j \in [k]\}$, the weights w_j and variances σ_j^2, as well as initializers $\mu_i^{(0)}$ for each $i \in [k]$ such that $\|\mu_i^{(0)} - \mu_i\|_\infty \le \varepsilon_0 \sigma_i$.

Parameters: Set $T = C \log\log(d/\delta)$, for some sufficiently large constant $C > 0$, $\varepsilon_0 = c_0 d^{-5/2}$ where $c_0 > 0$ is a sufficiently small constant, and $\eta_1, \eta_2, \eta_3 = \delta w_{\min}/(c'\sqrt{d}\rho_\sigma)$ where $c' > 0$ is a sufficiently small constant.

Output: Estimates $(\mu_i^{(T)} : i \in [k])$ for each component $i \in [k]$ such that $\|\mu_i^{(T)} - \mu_i\|_\infty \le \delta\sigma_i$.

1. If $\delta \ge \varepsilon_0$, then we just output $\mu_i^{(T)} = \mu_i^{(0)}$ for each $i \in [k]$.
2. Set $\mathbf{x}_i^{(0)} = \frac{1}{\sigma_i}\mu_i^{(0)}$ for each $i \in [k]$.
3. Obtain using Lemma 4.4 an estimate $\tilde{b}$ of b up to accuracy η_1 (in ℓ_∞ norm). In more detail, define the empirical average

$$\forall j \in [k], \tilde{b}_j = \frac{1}{w_j \sigma_j N} \sum_{\ell \in [N]} \mathbb{I}[y^{(\ell)} \in S_j] \left(y^{(\ell)} - z_j\right).$$

   (This is the only place the given samples are used)

4. For $t = 1$ to $T = O(\log\log(dk/\delta))$ steps do the following:
   (a) Obtain using Lemma 4.4 an estimate $\tilde{F}(\mathbf{x}^{(t)})$ of $F(\mathbf{x}^{(t)})$ at $\mathbf{x}^{(t)}$ up to accuracy η_2 (in ℓ_∞ norm).

(b) Obtain using Lemma 4.5 an estimate $\widetilde{F'}(\mathbf{x}^{(t)})$ of $F'(\mathbf{x}^{(t)}) = \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)})$ up to accuracy η_3 (in $\infty \to \infty$ operator norm).

(c) Update $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \left(\widetilde{F'}(\mathbf{x}^{(t)})\right)^{-1} \left(\tilde{b} - \tilde{F}(\mathbf{x}^{(t)})\right)$.

5. Output $\mu_i^{(T)} = \sigma_i \mathbf{x}_i^{(T)}$ for each $i \in [k]$.

The proof of the two approximation lemmas below is based on a rather standard Chernoff argument; see Section 4.5.

Lemma 4.4 (Estimating $F_j(\mathbf{x})$)**.** *Suppose samples $y^{(1)}, y^{(2)}, \dots, y^{(N)}$ are generated from a mixture of k spherical Gaussians with parameters $\{(w_i, \sigma_i \mathbf{x}_i, \sigma_i)\}_{i \in [k]}$ in d dimensions, and $\max\{\|\mathbf{x}_i\|_\infty, 1\} \le \rho/\sigma_i$ $\forall i \in [k]$. There exists a constant $C > 0$ such that for any $\eta > 0$ the following holds for $N \ge C\rho^3 \log(\frac{dk}{\gamma})/(\eta^2 w_{min})$ samples: with probability at least $1-\gamma$, for each $j \in [k]$ the empirical estimate for $F_j(\mathbf{x})$ has error at most η i.e.,*

$$\left\|F_j(\mathbf{x}) - \frac{1}{w_j \sigma_j N} \sum_{\ell \in [N]} \mathbb{I}[y^{(\ell)} \in S_j] \left(y^{(\ell)} - z_j\right) \right\|_\infty < \eta, \tag{22}$$

where S_j is defined in Definition 4.3.

We now state a similar lemma for the Jacobian $F' : \mathbb{R}^{k \cdot d} \to \mathbb{R}^{k \cdot d}$, which by an easy calculation is given for all $i, j \in [k]$ by

$$\nabla_{\mathbf{x}_i} F_j(\mathbf{x}) = \frac{2\pi w_i}{w_j \sigma_j \sigma_i} \int_{y \in S_j} (y - z_j)(y - \sigma_i \mathbf{x}_i)^T g_{\sigma_i \mathbf{x}_i, \sigma_i}(y)\, dy$$

i.e., for all $r, r_0 \in [d]$,

$$\frac{\partial F_{j,r_0}(\mathbf{x})}{\partial \mathbf{x}_i(r)} = \frac{2\pi w_i}{w_j \sigma_j \sigma_i} \int_{y \in S_j} (y(r_0) - z_j(r_0))(y(r_1) - \sigma_i \mathbf{x}_i(r_1)) g_{\sigma_i \mathbf{x}_i, \sigma_i}(y)\, dy \tag{23}$$

Lemma 4.5 (Estimating $\nabla_{\mathbf{x}_i} F_j(\mathbf{x})$)**.** *Suppose we are given the parameters of a mixture of k spherical Gaussians $\{(w_i, \sigma_i \mathbf{x}_i, \sigma_i)\}_{i \in [k]}$ in d dimensions, with $\max\{\|\mathbf{x}_i\|_\infty, 1\} \le \rho/\sigma_i$ for each $i \in [k]$, and region S_j is defined as in Definition 4.3. There exists a constant $C > 0$ such that for any $\eta > 0$ the following holds for $N \ge C\rho^4 d^2 k^2 \log(\frac{dk}{\gamma})/(\eta^2 w_{min})$. Given N samples $y^{(1)}, y^{(2)}, \dots, y^{(N)}$ generated from spherical Gaussian with mean $\sigma_i \mathbf{x}_i$ and variance $\sigma_i^2/(2\pi)$, we have with probability at least $1-\gamma$, for each $j \in [k]$ the following empirical estimate for $\nabla_{\mathbf{x}_i} F_j$ has error at most η i.e.,*

$$\left\|\nabla_{\mathbf{x}_i} F_j(\mathbf{x}) - \widetilde{\nabla_{\mathbf{x}_i} F_j}(\mathbf{x})\right\|_{\infty \to \infty} < \frac{\eta}{k}, \quad \textit{where } \widetilde{\nabla_{\mathbf{x}_i} F_j}(\mathbf{x}) := \frac{2\pi w_i}{w_j \sigma_i \sigma_j N} \sum_{\ell \in [N]} \mathbb{I}[y^{(\ell)} \in S_j] \left(y^{(\ell)} - z_j\right) \left(y^{(\ell)} - \sigma_i \mathbf{x}_i\right)^T. \tag{24}$$

Furthermore, we have that the estimated second derivative $\widetilde{F'}(\mathbf{x}) = (\widetilde{\nabla_{\mathbf{x}_i} F_j}(\mathbf{x}) : i, j \in [k])$ satisfies $\left\|F'(\mathbf{x}) - \widetilde{F'}(\mathbf{x})\right\|_{\infty \to \infty} \le \eta$.

Remark 4.6. Although for simplicity we focus here on the case that only the means are unknown, we believe that our approach can be extended to the case that the weights and variances are also unknown, and only coarse estimates for them are given. In order to handle that case, we need to collect more statistics about the given samples, in addition to just the mean in each region. Namely, using the samples restricted to S_j, we estimate the total probability mass in each S_j, i.e., $b_j^{(w)} = \mathbb{E}_{y \leftarrow \mathcal{G}} \mathbb{I}[y \in S_j]$, and the average squared Euclidean norm $b_j^{(\sigma)} = \frac{1}{d} \mathbb{E}_{y \leftarrow \mathcal{G}} \mathbb{I}[y \in S_j] \|y\|_2^2$. We now modify (21) by adding new unknowns (for the weights and variances), as well as new equations for $b_j^{(\sigma)}$ and $b_j^{(w)}$. This corresponds to $k(d+2)$ non-linear equations with $k(d+2)$ unknowns.

4.2 Convergence Analysis using the Newton method

We will now analyze the convergence of the Newton algorithm. We want each parameter $\mathbf{x}_i^{(T)} \in \mathbb{R}^d$ to be close to $\mathbf{x}_i^*$ in an appropriate norm (e.g., ℓ_2 or ℓ_∞). Hence, we will measure the convergence and error of $\mathbf{x} = (\mathbf{x}_i : i \in [k])$ to be measured in ℓ_∞ norm.

Definition 4.7 (Neighborhood)**.** Consider a mixture of Gaussians with parameters $((\mu_i, \sigma_i, w_i) : i \in [k])$, and let $(\mathbf{x}_i : i \in [k]) \in \mathbb{R}^{kd}$ be the corresponding parameters of the non-linear system $F(\mathbf{x}) = b$. The neighborhood set

$$\mathcal{N} = \{(\mathbf{x}_i : i \in [k]) \in \mathbb{R}^{kd} \ \mid \ \forall i \in [k], \|\mathbf{x}_i - \mathbf{x}_i^*\|_\infty < \varepsilon_0 = c_0 d^{-5/2}\},$$

is the set of values of the variables that are close to the true values of the variables given by $\mathbf{x}_i^* = \frac{\mu_i}{\sigma_i} \ \forall i \in [k]$, and $c_0 > 0$ is an appropriately large universal constant given in Theorem 4.1.

We will now show the convergence of the Newton method by showing diagonal dominance properties of the non-linear system given in Lemma 2.5. This diagonal dominance arises from the separation between the means of the components. Lemmas 4.8 and 4.9 show that most of the probability mass from jth component around μ_j is confined to S_j, while the other components of $\mathcal{G}$ are far enough from z_j, that they do not contribute much ℓ_1 mass in total to S_j.

The following lemma lower bounds the contribution to region S_j from the jth component.

Lemma 4.8. *In the notation of Theorem 4.1, for all $j \in [k]$, we have*

$$\int_{S_j} g_{\mu_j,\sigma_j}(y)\, dy \geq 1 - \frac{1}{8\pi d}. \tag{25}$$

$$\forall r_1 \in [d], \ \frac{1}{\sigma_j} \left| \int_{S_j} (y(r_1) - \mu_j(r_1))\, g_{\mu_j,\sigma_j}(y)\, dy \right| \leq \frac{1}{8\pi d}. \tag{26}$$

$$\forall r_1, r_2 \in [d], \ \frac{1}{\sigma_j^2} \left| \int_{S_j} (y(r_1) - \mu_j(r_1))\, (y(r_2) - \mu_j(r_2))\, g_{\mu_j,\sigma_j}(y)\, dy - \frac{\sigma_j^2}{2\pi} I[r_1 = r_2] \right| \leq \frac{1}{8\pi d}. \tag{27}$$

The proof of the above lemma follows from concentration bounds for multivariate Gaussians. The following lemma upper bounds the contribution in S_j from the other components.

Lemma 4.9 (Leakage from other components)**.** *In the notation of Theorem 4.1, and for a component $j \in [k]$ in $\mathcal{G}$, the contribution in S_j from the other components is small, namely, for all*

$j \in [k]$

$$\sum_{i \in [k], i \neq j} w_i \int_{S_j} g_{\mu_i, \sigma_i}(y)\, dy < \frac{w_j}{16\pi d}. \tag{28}$$

$$\forall r_1 \in [d], \quad \sum_{i \in [k], i \neq j} \frac{w_i}{\sigma_i} \cdot \frac{\|\mu_i - \mu_j\|_2}{\sigma_i} \int_{S_j} |y(r_1) - \mu_i(r_1)|\, g_{\mu_i, \sigma_i}(y)\, dy < \frac{w_j}{16\pi d \rho_\sigma}. \tag{29}$$

$$\forall r_1, r_2 \in [d], \quad \sum_{i \in [k], i \neq j} \frac{w_i}{\sigma_i \sigma_j} \int_{S_j} |y(r_1) - \mu_i(r_1)|\, |y(r_2) - \mu_i(r_2)|\, g_{\mu_i, \sigma_i}(y)\, dy < \frac{w_j}{16\pi d \rho_\sigma}. \tag{30}$$

The above lemma is the more technical of the two, and it is crucial in showing diagonal dominance of the Jacobian. The proof of the above lemma is very different for separation of order $\sqrt{\log k}$ and $\sqrt{d}$ – hence we will handle this separately in Sections 4.4.1 and 4.4.2 respectively.

We now show the convergence of Newton algorithm assuming the above two lemmas. Theorem 4.1 follows in a straightforward manner from the guarantees of the Newton algorithm. We mainly need to show that $\|F'\|\|F''\|\varepsilon_0 < 1/2$.

We now prove that the function $(F'(\mathbf{x}))^{-1}$ has bounded operator norm using the diagonal dominance properties of F.

Lemma 4.10. *For any point $\mathbf{x} \in \mathcal{N}$, the operator $F' : \mathbb{R}^{d \cdot k} \to \mathbb{R}^{d \cdot k}$ satisfies $\|(F'(\mathbf{x}))^{-1}\|_{\infty \to \infty} \leq 4$.*

Proof. We will divide the matrix F' into $k \times k = k^2$ blocks of size $d \times d$ each, and show that the matrix satisfies the required diagonal dominance property. Let us first consider the diagonal blocks i.e., we have from (23) for each $j \in [k]$

$$\begin{aligned}
\nabla_{\mathbf{x}_j} F_j(\mathbf{x}) &= \frac{2\pi}{\sigma_j^2} \int_{y \in S_j} (y - z_j)(y - \sigma_j \mathbf{x}_j)^T g_{\sigma_j \mathbf{x}_j, \sigma_j}(y)\, dy \\
&= \frac{2\pi}{\sigma_j^2} \int_{y \in S_j} (y - \sigma_j \mathbf{x}_j)(y - \sigma_j \mathbf{x}_j)^T g_{\sigma_j \mathbf{x}_j, \sigma_j}(y)\, dy + \frac{2\pi(\sigma_j \mathbf{x}_j - z_j)}{\sigma_j^2} \int_{y \in S_j} (y - \sigma_j \mathbf{x}_j)^T g_{\sigma_j \mathbf{x}_j, \sigma_j}(y)\, dy
\end{aligned}$$

Consider a mixture of Gaussians where the jth component has mean $\sigma_j \mathbf{x}_j$ and standard deviation $\sigma_j/\sqrt{2\pi}$. It satisfies the required separation conditions since $\|\sigma_i \mathbf{x}_i - \mu_i\|_2 \leq \varepsilon_0 \sigma_i$. Applying Lemma 4.8,

$$\begin{aligned}
\nabla_{\mathbf{x}_j} F_j(\mathbf{x}) = I_{d \times d} &- \frac{2\pi}{\sigma_j^2} \int_{y \notin S_j} (y - \sigma_j \mathbf{x}_j)(y - \sigma_j \mathbf{x}_j)^T g_{\sigma_j \mathbf{x}_j, \sigma_j}(y)\, dy \\
&- \frac{2\pi(\sigma_j \mathbf{x}_j - z_j)}{\sigma_j} \int_{y \notin S_j} \frac{(y - \sigma_j \mathbf{x}_j)^T}{\sigma_j} g_{\sigma_j \mathbf{x}_j, \sigma_j}(y)\, dy \\
= I_{d \times d} - E,
\end{aligned}$$

where $E \in \mathbb{R}^{d \times d}$ satisfies from (27) and (26) that

$$\begin{aligned}
\forall r_1, r_2 \in [d], \quad |E(r_1, r_2)| &\leq \frac{1}{4d} + \varepsilon_0 \cdot \frac{1}{4d} \leq \frac{1}{2d} \\
\forall r_1 \in [d], \quad \sum_{r_2 \in [d]} |E(r_1, r_2)| &\leq \frac{1}{2}.
\end{aligned} \tag{31}$$

Using a similar calculation, we see that for the off-diagonal blocks $M_{ji} = \nabla_{\mathbf{x}_i} F_j(\mathbf{x})$,

$$\begin{aligned} M_{ji} = \nabla_{\mathbf{x}_i} F_j(\mathbf{x}) &= \frac{2\pi w_i}{w_j \sigma_i \sigma_j} \int_{y \in S_j} (y - z_j)(y - \sigma_i \mathbf{x}_i)^T g_{\sigma_i \mathbf{x}_i, \sigma_i}(y)\, dy \\ &= \frac{2\pi w_i}{w_j} \int_{y \in S_j} \frac{(y - \sigma_i \mathbf{x}_i)(y - \sigma_i \mathbf{x}_i)^T}{\sigma_i \sigma_j} g_{\sigma_i \mathbf{x}_i, \sigma_i}(y)\, dy \\ &\quad + \frac{2\pi w_i (\sigma_i \mathbf{x}_i - z_j)}{w_j \sigma_j} \int_{y \in S_j} \frac{(y - \sigma_i \mathbf{x}_i)^T}{\sigma_i} g_{\sigma_i \mathbf{x}_i, \sigma_i}(y)\, dy \end{aligned}$$

Also, $\|\sigma_i \mathbf{x}_i - z_j\|_\infty \le \|\mu_i - \mu_j\|_\infty + \sigma_i + \sigma_j \le 2\|\mu_i - \mu_j\|_2$. For each $r_1, r_2 \in [d]$

$$\begin{aligned} |M_{ji}(r_1, r_2)| &\le \frac{2\pi w_i \rho_\sigma}{w_j \sigma_i^2} \int_{y \in S_j} |y(r_1) - \sigma_i \mathbf{x}_i(r_1)|\, |y(r_2) - \sigma_i \mathbf{x}_i(r_2)|\, g_{\sigma_i \mathbf{x}_i, \sigma_i}(y)\, dy \\ &\quad + \frac{4\pi w_i \rho_\sigma}{w_j} \frac{\|\mu_i - \mu_j\|_2}{\sigma_i} \int_{y \in S_j} \frac{|y(r_2) - \sigma_i \mathbf{x}_i(r_2)|}{\sigma_i} g_{\sigma_i \mathbf{x}_i, \sigma_i}(y)\, dy. \end{aligned}$$

Summing over all $i \neq j$, and using the bounds in Lemma 4.9,

$$\sum_{i \in [k], i \neq j} |M_{ji}(r_1, r_2)| \le \frac{1}{4d}$$

$$\sum_{r_2=1}^{d} \sum_{i \neq j} |M_{ji}(r_1, r_2)| \le \frac{1}{4} \tag{32}$$

Hence using Lemma 2.5 for diagonally dominant matrices, we get from (31) and (32)

$$\|(F'(\mathbf{x}))^{-1}\|_{\infty \to \infty} \le \frac{1}{1 - \frac{1}{2} - \frac{1}{4}} \le 4.$$

□

We now prove that the function $F'(\mathbf{x})$ is locally Lipschitz by upper bounding the second derivative operator.

Lemma 4.11. *There exists a universal constant $c' > 0$ such that the derivative F' is locally L-Lipschitz in the neighborhood $\mathcal{N}$ for $L \le c' d^{5/2}$.*

Proof. We proceed by showing that the operator norm of $F'' : \mathbb{R}^{dk \times dk} \to \mathbb{R}^{dk}$ is upper bounded by L at any point $\mathbf{x} \in \mathcal{N}$. We first write down expressions for F'', and then prove that the operator norm of F'' is bounded. We first observe that for each $j, i_1, i_2 \in [k]$,

$$\forall r_0, r_1, r_2 \in [d],\ \frac{\partial^2 F_{j, r_0}(\mathbf{x})}{\partial \mathbf{x}_{i_1}(r_1) \partial \mathbf{x}_{i_2}(r_2)} = 0 \text{ if } i_1 \neq i_2.$$

Hence the second derivatives are non-zero only for diagonal blocks $i_1 = i_2$. For each $j, i \in [k]$ the second derivatives for $\forall r_0, r_1, r_2 \in [d]$ are given by

$$\frac{\partial^2 F_{j,r_0}(\mathbf{x})}{\partial \mathbf{x}_i(r_1)\partial \mathbf{x}_i(r_2)} = \frac{4\pi w_i}{w_j \sigma_j}\int_{y\in S_j}(y(r_0)-z_j(r_0))\Big(\frac{(y(r_1)-\sigma_i\mathbf{x}_i(r_1))(y(r_2)-\sigma_i\mathbf{x}_i(r_2))}{\sigma_i^2} - \mathbb{I}[r_1=r_2]\Big) g_{\sigma_i\mathbf{x}_i,\sigma_i}(y)\,dy$$

$$\left|\frac{\partial^2 F_{j,r_0}(\mathbf{x})}{\partial \mathbf{x}_i(r_1)\partial \mathbf{x}_i(r_2)}\right| \le \frac{4\pi w_i}{w_j}\max_{y\in S_j}\frac{\|y-z_j\|_2}{\sigma_j}\cdot\int_{y\in S_j}\Big(\Big|\frac{(y(r_1)-\sigma_i\mathbf{x}_i(r_1))(y(r_2)-\sigma_i\mathbf{x}_i(r_2))}{\sigma_i^2}\Big|+1\Big) g_{\sigma_i\mathbf{x}_i,\sigma_i}(y)\,dy.$$

$$\le \frac{8\pi w_i\sqrt{d}}{w_j}\int_{y\in S_j}\Big(\Big|\frac{(y(r_1)-\sigma_i\mathbf{x}_i(r_1))(y(r_2)-\sigma_i\mathbf{x}_i(r_2))}{\sigma_i^2}\Big|+1\Big) g_{\sigma_i\mathbf{x}_i,\sigma_i}(y)\,dy.$$

A simple bound on the operator norm of $F'' : \mathbb{R}^{kd}\times\mathbb{R}^{kd}\to\mathbb{R}^{kd}$ is given by summing over all $i\in[k]$ and $r_1, r_2 \in [d]$. Consider a mixture of Gaussians where the jth component has mean $\sigma_j\mathbf{x}_j$ and standard deviation $\sigma_j/\sqrt{2\pi}$. It satisfies the required separation conditions since $\|\sigma_i\mathbf{x}_i - \mu_i\|_2 \le \varepsilon_0\sigma_i$. Applying Lemma 4.9 we get

$$\|F''(\mathbf{x})\|_{\infty,\infty\to\infty} \le \sum_{i\in[k]}\sum_{r_1,r_2\in[d]}\left|\frac{\partial^2 F_{j,r_0}(\mathbf{x})}{\partial \mathbf{x}_i(r_1)\partial \mathbf{x}_i(r_2)}\right|$$

$$\sum_{i\neq j}\sum_{r_1,r_2\in[d]}\left|\frac{\partial^2 F_{j,r_0}(\mathbf{x})}{\partial \mathbf{x}_i(r_1)\partial \mathbf{x}_i(r_2)}\right| \le \frac{16\pi d^{5/2}}{w_j}\cdot\frac{w_j}{16\pi d} \le d^{3/2}, \quad \text{and}$$

$$\sum_{r_1,r_2\in[d]}\left|\frac{\partial^2 F_{j,r_0}(\mathbf{x})}{\partial \mathbf{x}_j(r_1)\partial \mathbf{x}_j(r_2)}\right| \le 8\pi d^{5/2}\int_{y\in\mathbb{R}^d}\left|\frac{(y(r_1)-\sigma_j\mathbf{x}_j(r_1))(y(r_2)-\sigma_j\mathbf{x}_j(r_2))}{\sigma_j^2}\right| g_{\sigma_j\mathbf{x}_j,\sigma_j}(y)$$

$$+ 8\pi d^{5/2}\int_{y\in\mathbb{R}^d} g_{\sigma_j\mathbf{x}_j,\sigma_j}(y)\,dy$$

$$\le 8\pi d^{5/2}\Big(\frac{1}{\pi^2}+1\Big) \le 15\pi d^{5/2}$$

$$\forall \mathbf{x}\in\mathcal{N},\ \|F''(\mathbf{x})\|_{\infty,\infty\to\infty} \le 16\pi d^{5/2}.$$

Hence, applying Lemma B.1 with F' in the open set $K = \mathcal{N}$, the lemma follows. □

Proof of Theorem 4.1. We first note that we have from (17), for each $i\in[k]$ that $\|\mathbf{x}_i^* - \mathbf{x}_i^{(0)}\|_\infty \le \|\mathbf{x}_i^* - \mathbf{x}_i^{(0)}\|_2 \le \|\mu_i - \mu_i^{(0)}\|_2/\sigma_i \le \varepsilon_0$. We will use the algorithm to find $\|\mathbf{x}^{(T)} - \mathbf{x}^*\|_\infty < \delta/\sqrt{d}$ (use the algorithm with $\delta' := \delta/\sqrt{d}$). The proof follows in a straightforward manner from Corollary B.4. Lemma 4.11 shows that F' is locally L-Lipschitz for $L = c'd^{5/2}$. Together with Lemma 4.10 we have $\varepsilon_0 L\|(F')^{-1}\|_{\infty\to\infty} \le 1/2$ for our choice of ε_0. Also from Lemma 4.4, by using $N = O(\delta^{-2}\rho^6 k^6 w_{\min}^{-3})$ samples (and $d\le k$), $\eta_1 + \eta_2 \le \frac{\delta}{4\sqrt{d}\|(F')^{-1}\|_{\infty\to\infty}}$. It is easy to check that $B = \max_y\|F'(y)\| \le \frac{\max_i \sigma_i}{w_{\min}\min_i\sigma_j} \le 4\rho_\sigma/w_{\min}$. Hence, similarly from Lemma 4.5 $\eta_3 \le \frac{\delta}{4\sqrt{d}B\|(F')^{-1}\|^2_{\infty\to\infty}}$ as required. Hence, from Corollary B.4, $T = O(\log\log(d/\delta))$ iterations of the Newton's method suffices. Further, each iteration mainly involves inverting a $dk\times dk$ matrix which is polynomial time in d, k. □

4.3 Bounding Leakage from Own Component: Proof of Lemma 4.8

Lemma 4.12. *Let $\widehat{v}_0 \in \mathbb{R}^d$ be a unit vector. Then for any coordinates $r_1, r_2 \in [d]$ we have for any $t \geq 3$*

$$\frac{1}{\sigma^d}\int_{y:\langle y,\widehat{v}_0\rangle \geq t\sigma} \exp(-\pi\|y\|_2^2/\sigma^2)\,dy < \tfrac{1}{16\pi}\exp(-t^2) \tag{33}$$

$$\frac{1}{\sigma^d}\int_{y:\langle y,\widehat{v}_0\rangle \geq t\sigma} |y(r_1)|\exp(-\pi\|y\|_2^2/\sigma^2)\,dy < \tfrac{1}{16\pi}\exp(-t^2)\sigma \tag{34}$$

$$\frac{1}{\sigma^d}\int_{y:\langle y,\widehat{v}_0\rangle \geq t\sigma} |y(r_1)y(r_2)|\exp(-\pi\|y\|_2^2/\sigma^2)\,dy < \tfrac{1}{16\pi}\exp(-t^2)\sigma^2 \tag{35}$$

Proof. The first equation (33) follows easily from the rotational invariance of a spherical Gaussian. Suppose $g_0 = \langle y, \widehat{v}_0\rangle$, then $g_0 \sim N(0, \sigma^2/(2\pi))$. Hence

$$\frac{1}{\sigma^d}\int_{y:\langle y,\widehat{v}_0\rangle \geq t\sigma} \exp(-\pi\|y\|_2^2/\sigma^2)\,dy = \frac{1}{\sigma}\int_{g_0 \geq t\sigma} \exp(-\pi g_0^2/\sigma^2)\;dg_0 \;\leq \Phi_{0,1}\left(\sqrt{2\pi}t\right) < \exp(-t^2).$$

We will now prove (35); the proof of (34) follows the same argument. We first observe that using the rotational invariance, it suffices to consider the span of $\widehat{v}_0, \widehat{e}_{r_1}, \widehat{e}_{r_2}$.

Consider the case $r_1 \neq r_2$. Let $\widehat{v}_1, \widehat{v}_2$ be the unit vectors along $\widehat{e}_{r_1} - \langle \widehat{e}_{r_1}, \widehat{v}_0\rangle\widehat{v}_0$ and $\widehat{e}_{r_2} - \langle \widehat{e}_{r_2}, \widehat{v}_0\rangle\widehat{v}_0 - \langle \widehat{e}_{r_2}, \widehat{v}_1\rangle\widehat{v}_1$ respectively. If $g_\ell = \langle y, \widehat{v}_\ell\rangle$ for $\ell = 0, 1, 2$, then $\{g_\ell\}$ are mutually independent and distributed as $N(0, \sigma^2/(2\pi))$. Let $\alpha_1 = \langle \widehat{v}_0, \widehat{e}_{r_1}\rangle$, $\alpha_2 = \langle \widehat{v}_0, \widehat{e}_{r_2}\rangle$ and $\beta_2 = \langle \widehat{v}_1, \widehat{e}_{r_2}\rangle$. Then,

$$y(r_1) = \alpha_1 g_0 + \sqrt{1-\alpha_1^2}\cdot g_1, \text{ and } y(r_2) = \alpha_2 g_0 + \beta_2 g_1 + \sqrt{1-\alpha_1^2-\beta_2^2}\cdot g_2,$$
$$|y(r_1)| \leq |g_0| + |g_1|, \text{ and } |y(r_1)y(r_2)| \leq (|g_0|+|g_1|)(|g_0|+|g_1|+|g_2|) \leq 3(|g_0|^2+|g_1|^2+|g_2|^2).$$

Hence,

$$\begin{aligned}
&\frac{1}{\sigma^d}\int_{y:\langle y,\widehat{v}_0\rangle \geq t\sigma} |y(r_1)y(r_2)|\exp(-\pi\|y\|_2^2)\,dy \leq \frac{3}{\sigma^d}\int_{y:\langle y,\widehat{v}_0\rangle \geq t\sigma}\left(|g_0|^2+|g_1|^2+|g_2|^2\right)\exp(-\pi\|y\|_2^2)\,dy \\
&\leq \frac{3}{\sigma^3}\int_{g_0>t\sigma}\int_{g_1}\int_{g_2}(|g_0|^2+|g_1|^2+|g_2|^2)\cdot\exp(-\pi g_2^2)\;dg_2\cdot\exp(-\pi g_1^2)\;dg_1\cdot\exp(-\pi g_0^2)\;dg_0 \\
&\leq \frac{3}{\sigma}\int_{g_0>t\sigma}|g_0|^2\exp(-\pi g_0^2)\;dg_0 + \frac{6}{\sigma}\int_{g_1}|g_1|^2\exp(-\pi g_1^2)\;dg_1\cdot\frac{1}{\sigma}\int_{g_0>t\sigma}\exp(-\pi g_0^2)\;dg_0 \;\;\text{(by symmetry)} \\
&\leq 3\exp(-2t^2)\sigma^2 + 6\sigma^2\exp(-2t^2) \leq 9\sigma^2\exp(-2t^2),
\end{aligned}$$

where the last line follows from truncated moments of a normal variable with variance $\sigma^2/(2\pi)$ (Lemma A.2). Using the fact that $\exp(t^2) > 9\cdot 16\pi$ for $t \geq 3$, the lemma follows. The proof for $r_1 = r_2$, and for (34) follows from an identical argument. □

Proof of Lemma 4.8. We will consider the loss from the region outside of S_j. The region S_j is defined by k linear inequalities, one for each of the components $\ell \in [k]$, and a constraint that points lie in an ℓ_2 ball around z_j. Consider y drawn from g_{μ_j,σ_j}. Since $\|z_j - \mu_j\|_2 \leq \sigma_j$ and Lemma A.4 (applied with $t = 6d$),

$$\mathbb{P}[\|y - z_j\|_2 \geq 4\sqrt{d}\sigma_j] \leq \mathbb{P}[\|y-\mu_j\|_2 \geq 3\sqrt{d}\sigma_j] \leq \exp(-6d) \leq \frac{1}{16\pi d}.$$

When y is drawn from g_{μ_j,σ_j}, from Lemma 4.12, each of the k linear constraints are satisfied with high probability i.e., for a fixed $\ell \in [k]$,

$$\mathbb{P}\left[|\langle y - z_j, \widehat{e}_{j\ell}\rangle| > 4\sigma_j\sqrt{\log(1/w_{\min})}\right] \leq \mathbb{P}\left[|\langle y - \mu_j, \widehat{e}_{j\ell}\rangle| > 3\sigma_j\sqrt{\log(1/w_{\min})}\right] \leq \frac{w_{\min}^2}{16\pi} \leq \frac{1}{k}\cdot\frac{w_{\min}}{16\pi}, \tag{36}$$

where the last step follows since $w_{\min} \leq 1/k$. Hence, performing a union bound over the k components and the ℓ_2 constraint, and using $d \leq k$, we get

$$\int_{S_j} g_{\mu_j,\sigma_j}(y)\,dy \geq 1 - \frac{1}{16\pi d} - \frac{1}{16\pi d} \geq 1 - \frac{1}{8\pi d}.$$

To prove (26), $\forall r_1 \in [d]$ we have

$$\begin{aligned}\int_{S_j} (y(r_1) - \mu_j(r_1))\,g_{\mu_j,\sigma_j}(y)\,dy &= \int_{\mathbb{R}^d} (y(r_1) - \mu_j(r_1))\,g_{\mu_j,\sigma_j}(y)\,dy - \int_{y\notin S_j} (y(r_1) - \mu_j(r_1))\,g_{\mu_j,\sigma_j}(y)\,dy \\ &= 0 - \int_{y\notin S_j} (y(r_1) - \mu_j(r_1))\,g_{\mu_j,\sigma_j}(y)\,dy\end{aligned}$$

Any point $y \notin S_j$ implies that one of the k linear constraints $\langle y - \mu_j, \widehat{e}_{ji}\rangle > 3\sqrt{\log(1/w_{\min})}\sigma_j$, or $\|y - \mu_j\|_2 > 3\sqrt{d}\sigma_j$ is true. Again, for the ℓ_2 ball constraint, from Lemma A.5 with $q \leq 2$,

$$\left|\int_{\|y-\mu_j\|_2 > 3\sqrt{d}\sigma_j} \frac{(y(r_1) - \mu_j(r_1))}{\sigma_j}\cdot g_{\mu_j,\sigma_j}(y)\,dy\right| \leq \left|\int_{\|y-\mu_j\|_2 > 3\sqrt{d}\sigma_j} \frac{\|y - \mu_j\|_2}{\sigma_j}\cdot g_{\mu_j,\sigma_j}(y)\,dy\right| \leq \frac{1}{16\pi d}.$$

Let $Z_{ji} = \{y : \langle y - \mu_j, \widehat{e}_{ji}\rangle > 3\sigma_j\sqrt{\log(\rho_\sigma/w_{\min})}\}$ be the set y that do not satisfy the constraint along $\widehat{e}_{ji}$. Since $\|z_j - \mu_j\|_2 \leq \sigma_j$, we have that the total loss from those $y \in Z_{ji}$ is

$$\begin{aligned}\left|\int_{y\in Z_{ji}} (y(r_1) - \mu_j(r_1))\,g_{\mu_j,\sigma_j}(y)\,dy\right| &\leq \int_{y\in Z_{ji}} |y(r_1) - \mu_j(r_1)|\,g_{\mu_j,\sigma_j}(y)\,dy \\ &\leq \sigma_j\left(\frac{w_{\min}^2}{16\pi\rho_\sigma^2}\right),\end{aligned}$$

by applying Lemma 4.12 with the Gaussian $(y - \mu_j)$, and $t = 3\sqrt{\log(\rho_\sigma/w_{\min})}$. Hence, we have

$$\begin{aligned}\left|\int_{S_j} (y(r_1) - \mu_j(r_1))\,g_{\mu_j,\sigma_j}(y)\,dy\right| &\leq \left|\int_{y\notin S_j} (y(r_1) - \mu_j(r_1))\,g_{\mu_j,\sigma_j}(y)\,dy\right| \\ &\leq \sigma_j\left(\frac{w_{\min}^2}{16\pi\rho_\sigma^2}\right)k + \frac{\sigma_j}{16\pi d} \leq \sigma_j\Big(\frac{1}{8\pi d}\Big),\end{aligned}$$

since $w_{\min} \leq 1/k$. The proof of the last equation follows in an identical fashion. From (35) of Lemma 4.12, we have

$$\forall r_1, r_2 \in [d],\ \frac{1}{\sigma_j^2}\left|\int_{y\notin S_j} \big(y(r_1) - \mu_j(r_1)\big)\big(y(r_2) - \mu_j(r_2)\big)g_{\mu_j,\sigma_j}(y)\,dy\right| \leq \frac{1}{8\pi d}.$$

Further,

$$\forall r_1, r_2 \in [d], \quad \int_{\mathbb{R}^d} (y(r_1) - \mu_j(r_1))\,(y(r_2) - \mu_j(r_2))\, g_{\mu_j,\sigma_j}(y)\, dy = \frac{\sigma_j^2}{2\pi} I[r_1 = r_2]$$

$$\text{Hence,} \quad \left| \frac{1}{\sigma_j^2} \int_{S_j} (y(r_1) - \mu_j(r_1))\,(y(r_2) - \mu_j(r_2))\, g_{\mu_j,\sigma_j}(y)\, dy - \frac{1}{2\pi} I[r_1 = r_2] \right| \leq \frac{1}{8\pi d}.$$

□

4.4 Bounding Leakage from Other Components: Proof of Lemma 4.9

Since our algorithm works when we either have separation of order $\sqrt{\log k}$ or separation of order $\sqrt{d}$, we have two different proofs for Lemma 4.9 depending on whether $\sqrt{\log(1/w_{\min})} \leq \sqrt{d}$ or not[3].

4.4.1 Separation of Order $\sqrt{\log k}$

In this case, we have a separation of

$$\forall i \neq j \in [k], \ \|\mu_i - \mu_j\|_2 \geq c(\sigma_i + \sigma_j)\sqrt{\log(\rho_\sigma/w_{\min})}.$$

Let $\widehat{e}_{ji}$ be the unit vector along $z_i - z_j$. We will now show that every point $y \in S_j$, is far from z_i along the direction $\widehat{e}_{ji}$.

$$\langle z_i - z_j, \widehat{e}_{ji}\rangle = \|z_j - z_i\|_2 \geq \|\mu_i - \mu_j\|_2 - \sigma_i - \sigma_j \geq 5\sqrt{\log(\rho_\sigma/w_{\min})}(\sigma_i + \sigma_j) + \tfrac{1}{2}\|\mu_i - \mu_j\|_2$$

Since $y \in S_j$, $|\langle y - z_j, \widehat{e}_{ji}\rangle| \leq 4\sigma_j\sqrt{\log(\rho_\sigma/w_{\min})}$, and

$$\langle z_i - y, \widehat{e}_{ji}\rangle \geq \langle z_i - z_j, \widehat{e}_{ji}\rangle - |\langle y - z_j, \widehat{e}_{ji}\rangle| \geq 5\sigma_i\sqrt{\log(\rho_\sigma/w_{\min})} + \tfrac{1}{2}\|\mu_i - \mu_j\|_2$$
$$\langle \mu_i - y, \widehat{e}_{ji}\rangle \geq \langle z_i - y, \widehat{e}_{ji}\rangle - \|\mu_i - z_i\|_2 \geq 4\sigma_i\sqrt{\log(\rho_\sigma/w_{\min})} + \tfrac{1}{2}\|\mu_i - \mu_j\|_2$$

We will now use Lemma 4.12 for each component $i \in [k]$ with mean zero Gaussian $y - \mu_i$ and $t^2 = 16\log(\rho_\sigma/w_{\min}) + \frac{1}{8\pi\sigma_i^2}\|\mu_j - \mu_i\|_2^2$. We first prove (28); using equation (33) with Gaussian $(y - \mu_i)$ of variance $\sigma_i^2/(2\pi)$,

$$w_i \int_{S_j} g_{\mu_i,\sigma_i}(y)\, dy \leq w_i \exp(-t^2) < w_i \cdot \frac{w_{\min}^2}{16\pi} \exp\Big(-\frac{\|\mu_i - \mu_j\|_2^2}{16\sigma_i^2}\Big) < \frac{w_{\min}^2}{16\pi}.$$

$$\text{Hence,} \quad \sum_{i \in [k], i \neq j} w_i \int_{S_j} g_{\mu_i,\sigma_i}(y)\, dy < \frac{w_j}{16\pi d}.$$

For (29), we use (34) similarly with $(y - \mu_i)$ and variance $\sigma_i^2/(2\pi)$ and t as above:

$$w_i \int_{S_j} \frac{|y(r_1) - \mu_i(r_1)|}{\sigma_i} g_{\mu_i,\sigma_i}(y)\, dy \leq w_i \cdot \frac{w_{\min}^2}{16\pi\rho_\sigma} \exp\Big(-\frac{\|\mu_i - \mu_j\|_2^2}{16\sigma_i^2}\Big).$$

$$\sum_{i \in [k], i \neq j} \frac{w_i}{\sigma_i} \cdot \frac{\|\mu_i - \mu_j\|_2}{\sigma_i} \int_{S_j} |y(r_1) - \mu_i(r_1)|\, g_{\mu_i,\sigma_i}(y)\, dy < \frac{w_j}{16\pi d\rho_\sigma}.$$

[3] More accurately whether $\sqrt{\log(\rho_\sigma/w_{\min})} \leq \sqrt{d} + \sqrt{\log(\rho_w\rho_\sigma)}$, where $\rho_w \geq 1/w_{\min}$.

The proof of (30) follows in an identical fashion from (35) for $r_1, r_2 \in [d]$

$$\frac{w_i}{\sigma_i \sigma_j} \int_{S_j} |y(r_1) - \mu_i(r_1)|\, |y(r_2) - \mu_i(r_2)|\, g_{\mu_i,\sigma_i}(y)\, dy \le \frac{w_i \sigma_i}{\sigma_j} \cdot \frac{w_{\min}^2}{16\pi \rho_\sigma^2} \le \frac{w_j w_i}{16\pi d \rho_\sigma}$$

$$\sum_{i \in [k], i \neq j} \frac{w_i}{\sigma_i \sigma_j} \int_{S_j} |y(r_1) - \mu_i(r_1)|\, |y(r_2) - \mu_i(r_2)|\, g_{\mu_i,\sigma_i}(y)\, dy < \frac{w_j}{16\pi d \rho_\sigma}.$$

4.4.2 Separation of Order $\sqrt{d}$

The proof of Lemma 4.9 uses the following useful lemma that shows that for any point within $\sigma_j \sqrt{d}$ distance of μ_j, the total probability mass from the other Gaussian components is negligible. Note that in the following lemma, k can be arbitrary large in terms of d.

Lemma 4.13. *There exists a universal constant $C_0 > 1$ such that for any $C \ge C_0$, if $\mathcal{G}$ is a mixture of k spherical Gaussians $\{(w_j, \mu_j, \sigma_j) : j \in [k]\}$ in $\mathbb{R}^d$ satisfying*

$$\forall i \neq j \in [k],\ \|\mu_i - \mu_j\|_2 \ge C(\sigma_i + \sigma_j)\left(\sqrt{d} + \sqrt{\log(\rho_w \rho_\sigma)}\right), \tag{37}$$

then for every component $j \in [k]$ and $\forall x^* : \|x^* - \mu_j\| \le 4\sigma_j \sqrt{d}$ and $\forall 0 \le m \le 2$,

$$\sum_{i \in [k], i \neq j} w_i g_{\mu_i,\sigma_i}(x^*) < \exp(-C^2 d/8) \cdot w_j g_j(x). \tag{38}$$

$$\sum_{i \in [k], i \neq j} w_i \left(\frac{\|x - \mu_i\|^m}{\sigma_i^m}\right) \cdot g_{\mu_i,\sigma_i}(x^*) \le \frac{\exp(-C^2 d/8)}{\rho_\sigma^m} \cdot w_j g_j(x^*), \tag{39}$$

where $\rho_\sigma = \max_i \sigma_i / (\min_i \sigma_i)$ and $\rho_w = \max_i w_i / (\min_i w_i)$.

We now prove Lemma 4.9 under separation $\|\mu_i - \mu_j\|_2 \ge c(\sigma_i + \sigma_j)(\sqrt{d} + \sqrt{\log(\rho_\sigma \rho_w)})$ assuming the above lemma.

Proof for Lemma 4.9. Our proof will follow by applying Lemma 4.13, and integrating over all $y \in S_j$. For any point $y \in S_j$, $\|y - \mu_j\|_2 \le 3\sigma_j(\sqrt{d} + \sqrt{\log(\rho_w \rho_\sigma)})$, and the separation of the means satisfies (37). To prove (25) we get from (38),

$$\sum_{i \in [k], i \neq j} w_i g_{\mu_i,\sigma_i}(y) \le w_j \exp(-C^2 d/8) \cdot g_{\mu_j,\sigma_j}(y).$$

Integrating over S_j, $\displaystyle\sum_{i \in [k], i \neq j} w_i \int_{S_j} g_{\mu_i,\sigma_i}(y)\, dy \le w_j \exp(-C^2 d/8) \int_{\mathbb{R}^d} g_{\mu_j,\sigma_j}(y)\, dy \le \frac{w_j}{16\pi d}.$

To prove (27), we get from (39) that for each $y \in S_j$ and $r_1, r_2 \in [d]$,

$$\begin{aligned}\sum_{i \in [k] \setminus \{j\}} \frac{w_i}{\sigma_i \sigma_j} |y(r_1) - \mu_i(r_1)|\, |y(r_2) - \mu_i(r_2)|\, g_{\mu_i,\sigma_i}(y) &\le \sum_{i \in [k] \setminus \{j\}} \frac{w_i \rho_\sigma}{w_j} \frac{\|y - \mu_i\|_2^2}{\sigma_i^2} g_{\mu_i,\sigma_i}(y) \\ &\le \frac{w_j \exp(-C^2 d/8)}{\rho_\sigma} \cdot g_{\mu_j,\sigma_j}(y).\end{aligned}$$

Integrating over all $y \in S_j$,

$$\sum_{i\in[k],i\neq j} \frac{w_i}{\sigma_i\sigma_j}\int_{S_j} |y(r_1)-\mu_i(r_1)|\,|y(r_2)-\mu_i(r_2)|\, g_{\mu_i,\sigma_i}(y)\,dy \le \frac{w_j}{\rho_\sigma}\exp(-C^2d/8)\cdot\int_{S_j} g_{\mu_j,\sigma_j}(y) \le \frac{w_j}{16\pi d\rho_\sigma}$$

To prove (26), we first note that for each $y \in S_j$, $\|\mu_i-\mu_j\|_2 \le \|\mu_i - y\| + (\sqrt{d}+\sqrt{\log\rho_w})\sigma_j \le 2\|\mu_i - y\|$. Hence, by applying (39) in Lemma 4.13, the following inequality holds:

$$\sum_{i\in[k],i\neq j}\frac{w_i\|\mu_i-\mu_j\|_2\|y-\mu_i\|_2}{\sigma_i\sigma_j}g_{\mu_i,\sigma_i}(y) \le 2\sum_{i\in[k],i\neq j} w_i\Big(\frac{\|y-\mu_i\|}{\sigma_i}\Big)^2\cdot g_{\mu_i,\sigma_i}(y) \le 2\exp(-C^2d/8)\rho_\sigma^{-2}\cdot w_j g_j(y). \tag{40}$$

Hence (26) follows from a similar argument as before. □

The proof of Lemma 4.13 proceeds by using a packing-style argument. Roughly speaking, in the uniform case when all the variances are roughly equal, the separation condition can be used to establish an upper bound on the number of other Gaussian means that are at a distance of r from a certain mean. We now present the proof for the case when the variances are roughly equal, and this will also be important for the general case.

Lemma 4.14. *In the notation of Lemma 4.13, let us denote for convenience, the jth component by (w_0,μ_0,σ_0). Then, the total probability mass at any point x^* s.t. $\|x^*-\mu_0\| \le 4\sqrt{d}\sigma_0$, from components $(w_1,\mu_1,\sigma_1),\dots,(w_i,\mu_i,\sigma_i),\dots,(w_{k'},\mu_{k'},\sigma_{k'})$ satisfying $\forall i \in [k'],\ \sigma_0 \le \sigma_i \le 2\sigma_0$, and $\|\mu_i-\mu_0\|_2 \ge C\left(\sqrt{d}+\sqrt{\log\rho_w}\right)$ is given by*

$$\sum_{i\in[k']} w_i g_i(x^*) < w_0\sigma_0^{-d}\exp(-C^2d/4). \tag{41}$$

Furthermore, we have $\sum_{i\in[k']} w_i g_i^{1/2}(x^*) < w_0\sigma_0^{-d}\exp(-C^2d/4)$.

Proof. We can assume without loss of generality that $\mu_0 = 0, \sigma_0 = 1$. Suppose $r_i = \|\mu_i - x\|_2$. Considering $1 \le \sigma_i \le 2$ and $g_i(x) = \sigma_i^{-d}\exp(-\pi\|x-\mu_i\|_2^2/\sigma_i^2)$, the lemma will follow, if we prove the following inequality for any $C \ge C_0 \ge 16$:

$$\sum_{i\in[k']} w_i \exp\left(-\tfrac{\pi}{2}\cdot r_i^2\right) < w_0\exp(-C^2d/4), \tag{42}$$

Let $B_i = \{y : \|y-\mu_i\|_2 \le \sqrt{d}\}$. From the separation conditions, we have that

- the balls $B_1,\dots,B_{k'}$ are pairwise disjoint, and
- the balls are far from the origin i.e, $\forall y \in B_i$, $\|y\| \ge C(\sqrt{d}+\sqrt{\log\rho_w})$.

We first relate the p.d.f. value $g_i(x)$ to the Gaussian measure of a ball of radius $\sqrt{d}$ around μ_i. The volume of the ball $\mathrm{Vol}(B_i) \ge 1$ and for every $y \in B_i$, the separation conditions and triangle inequality imply that $\|y\| \ge \|\mu_i - x^*\| - \|y-\mu_i\| - \|x^*\| \ge r_i/\sqrt{2}$. Hence,

$$\exp\left(-\frac{\pi r_i^2}{2}\right) \le \int_{y\in B_i}\exp\left(-\pi\|y\|^2\right)\,dy. \tag{43}$$

By using the disjointness of balls B_i, we get that

$$\begin{aligned}\sum_{i\in[k']} w_i \exp\left(-\frac{\pi r_i^2}{2}\right) &\leq \sum_{i\in[k']} w_i \int_{y\in B_i} \exp\left(-\pi\|y\|^2\right)\,dy \leq w_{\max}\sum_{i\in[k']}\int_{y\in B_i}\exp\left(-\pi\|y\|^2\right)\,dy\\ &\leq w_{\max}\int_{y:\|y\|\geq C(\sqrt{d}+\sqrt{\log\rho_w})}\exp\left(-\pi\|y\|^2\right)\,dy\\ &\leq \rho_w w_0\cdot\frac{1}{\rho_w}\exp(-C^2d/4)\leq w_0\exp(-C^2d/4),\end{aligned}$$

where the last line follows since a Gaussian random variable in d dimensions with mean 0 and unit variance in each direction has measure at most $\exp(-s^2/2)$ outside a ball of radius $(\sqrt{d}+s)$ (see Lemma A.4). □

We now proceed to the proof of Lemma 4.13.

Proof of Lemma 4.13. We may again assume without loss of generality that $\mu_j = 0$ and $\sigma_j = 1$ (by shifting the origin and scaling). Hence $\|x^*\|_2 \leq \sqrt{d}/\sqrt{\pi}$. We will divide the components $i \in [k]\setminus\{j\}$ depending on their standard deviation σ_i into buckets $I_0, I_1, \ldots, I_s$ where $s \leq \lceil\log(\max_i \sigma_i/\sigma_j)\rceil$ as follows:

$$I_0 = \{i\in[k]\setminus\{j\} : \sigma_i \leq \sigma_j\}, \quad \forall q\in[s]\;\; I_q = \{i\in[k] : 2^{q-1}\sigma_j < \sigma_i \leq 2^q\sigma_j\}.$$

Let us first consider the components in the bucket I_0, and suppose we scale so that $\sigma_j = 1$. As before let $r_i = \|x^*-\mu_i\|_2$. We first note that if $\sigma_i \leq \sigma_j$, since $\|x^*-\mu_i\|_2 \geq C\sqrt{d}(\sigma_i+\sigma_j)$, a simple calculation shows that $g_{\mu_i,\sigma_i}(x^*) \leq g_{\mu_i,\sigma_j}(x^*)$. Hence, by applying Lemma 4.14 to I_0 with a uniform variance of $\sigma_j^2 = 1$ for all Gaussians, we see that

$$\sum_{i\in I_0} w_i g_{\mu_i,\sigma_i}(x^*) \leq \sum_{i\in I_0} w_i g_{\mu_i,\sigma_j}(x^*) \leq w_j\exp(-C^2d/4). \tag{44}$$

Consider any bucket I_q. We will scale the points down by $2^{q-1}\sigma_j$ so that $y' = y/(2^{q-1}\sigma_j)$, and let $\forall i\in I_q\; r_i' = r_i/(2^{q-1}\sigma_j)$. Again from Lemma 4.14 we have that

$$\sum_{i\in I_q} w_i\exp(-\pi r_i^2/\sigma_i^2) \leq \sum_{i\in I_q} w_i\exp(-\pi r_i'^2/4) \leq w_j\exp(-C^2d/4). \tag{45}$$

$$\text{Hence,}\quad \sum_{i\in I_q} w_i g_i(x) \leq \frac{1}{2^{(q-1)d}\sigma_j^d}\sum_{i\in I_q} w_i\exp(-\pi r_i^2/\sigma_i^2) \leq w_j 2^{-(q-1)d}\exp(-C^2d/4). \tag{46}$$

Hence, by summing up over all buckets (using (44) and (46)), the total contribution $\sum_{i\in[k]\setminus\{j\}} w_i g_i(x) \leq 4w_j\exp(-C^d/4) \leq \exp(-C^2d/8)\cdot w_j g_j(x)$.

The final equation (39) follows from separation conditions, since $\|x-\mu_i\| > C\sqrt{d}\sigma_i$, hence $\exp(\pi\|x-\mu_i\|^2/2\sigma_i^2) > \|x-\mu_i\|^m/\sigma_i^m$ for some sufficiently large constant $C = C(m) \geq 1$. Hence, by using an identical argument with the furthermore part of Lemma 4.14, it follows.

□

4.5 Sampling Errors

Lemma 4.15 (Error estimates for Gaussians). *Let $S \subset \{x \in \mathbb{R}^d : \|x\|_2 \leq \rho\}$ be any region, and suppose samples $x^{(1)}, x^{(2)}, \dots, x^{(N)}$ are generated from a mixture of k spherical Gaussians with parameters $\{(w_i, \mu_i, \sigma_i)\}_{i \in [k]}$ in d dimensions, with $\forall i$, $\|\mu_i\|, \sigma_i \leq \rho$. There exists a constant $C > 0$ such that for any $\varepsilon > 0$, with $N \geq C\left(\varepsilon^{-2}\rho^2 \log d \log(1/\gamma)\right)$ samples, we have for all $r, r' \in [d]$ with probability at least $1 - \gamma$*

$$\left| \int_{x \in S} \sum_{i \in [k]} (x(r) - \mu_i(r)) g_{\mu_i,\sigma_i}(x) dx - \frac{1}{N} \sum_{\ell \in [N]} \left(x^{(\ell)}(r) - \mu_i(r) \right) \mathbb{I}[x^{(\ell)} \in S] \right| < \varepsilon. \tag{47}$$

$$\begin{aligned} \Big| \int_{x \in S} \sum_{i \in [k]} & (x(r) - \mu_i(r))(x(r') - \mu_i(r'))^T g_{\mu_i,\sigma_i}(x) dx \\ & - \frac{1}{N} \sum_{\ell \in [N]} \left(x^{(\ell)}(r) - \mu_i(r) \right) \left(x^{(\ell)}(r') - \mu_i(r') \right)^T \mathbb{I}[x^{(\ell)} \in S] \Big| < \varepsilon. \end{aligned} \tag{48}$$

Proof. Fix an element $r \in [d]$. Each term $\ell \in [N]$ in the sum corresponds to an i.i.d random variable $Z_\ell = \left(x^{(\ell)}(r) - \mu_i(r) \right) \mathbb{I}[x^{(\ell)} \in S]$. We are interested in the deviation of the sum $Z = \frac{1}{N} \sum_{\ell \in [N]} Z_\ell$.

Firstly, $\mathbb{E}[Z] = \int_{x \in S} \sum_{i \in [k]} (x(r) - \mu_i(r)) g_{\mu_i,\sigma_i}(x) dx$. Further, each of the i.i.d r.v.s has value $|Z_\ell - \mathbb{E} Z_\ell| \leq |x^{(\ell)}(r) - \mu_i(r)| + \rho$. Hence, $|Z_\ell| > (2\rho + t \max_i \sigma_i)$ with probability $O\left(\exp(-t^2/2)\right)$. Hence, by using standard sub-gaussian tail inequalities, we get

$$\mathbb{P}[|Z - \mathbb{E} Z| > \varepsilon] < \exp\left(-\frac{\varepsilon^2 N}{(2\rho + \max_i \sigma_i)^2} \right)$$

Hence, to union bound over all d events $N = O\left(\varepsilon^{-2}\rho^2 \log d \log(1/\gamma)\right)$ suffices.

A similar proof also works for the second equation (48). □

Proof of Lemma 4.4. Let $Z = \frac{1}{w_j \sigma_j N} \sum_{\ell \in [N]} \mathbb{I}[y^{(\ell)} \in S_j] \left(y^{(\ell)} - z_j \right)$. We see that

$$\mathbb{E}[Z] = \frac{1}{w_j \sigma_j} \sum_{i=1}^{k} w_i \int_{y \in \mathbb{R}^d} \mathcal{I}_j(y)(y - z_j) g_{\sigma_i \mathbf{x}_i, \sigma_i}(y) \, dy = F_j(\mathbf{x}).$$

Further, we can write $Z = Z_1 + Z_2$, where

$$\begin{aligned} Z_1 &= \frac{1}{w_j \sigma_j N} \sum_{\ell \in [N]} \mathbb{I}[y^{(\ell)} \in S_j] \left(y^{(\ell)} - \sigma_i \mathbf{x}_i \right), \\ Z_2 &= \frac{1}{w_j \sigma_j N} \left(\sigma_i \mathbf{x}_i - z_j \right) \sum_{\ell \in [N]} \mathbb{I}[y^{(\ell)} \in S_j]. \end{aligned}$$

By applying Lemma 4.15, we get

$$\|Z - \mathbb{E}[Z]\|_\infty \leq \|Z_1 - \mathbb{E} Z_1\|_\infty + \|Z_2 - \mathbb{E} Z_2\|_\infty \leq \frac{\eta}{2} + \frac{\eta}{2} = \eta.$$

□

Proof of Lemma 4.5. Let $Z = \frac{w_i}{w_j \sigma_i \sigma_j N} \sum_{\ell \in [N]} \mathbb{I}[y^{(\ell)} \in S_j] \left(y^{(\ell)} - z_j\right) \left(y^{(\ell)} - \sigma_i \mathbf{x}_i\right)^T$, where the samples are drawn just from the spherical Gaussian with mean $\sigma_i \mathbf{x}_i$ and variance σ_i^2. Hence,

$$\mathbb{E}[Z] = \frac{w_i}{w_j \sigma_j \sigma_i} \int_{y \in \mathbb{R}^d} \mathcal{I}_j(y)(y - z_j)(y - \sigma_i \mathbf{x}_i)^T g_{\sigma_i \mathbf{x}_i, \sigma_i}(y)\, dy = F_j(\mathbf{x}).$$

Also $Z = Z_1 + Z_2$ where

$$Z_1 = \frac{w_i}{w_j \sigma_i \sigma_j N} \sum_{\ell \in [N]} \mathbb{I}[y^{(\ell)} \in S_j] \left(\sigma_i \mathbf{x}_i - z_j\right) \left(y^{(\ell)} - \sigma_i \mathbf{x}_i\right)^T \text{ and}$$

$$Z_2 = \frac{w_i}{w_j \sigma_i \sigma_j N} \sum_{\ell \in [N]} \mathbb{I}[y^{(\ell)} \in S_j] \left(y^{(\ell)} - \sigma_i \mathbf{x}_i\right) \left(y^{(\ell)} - \sigma_i \mathbf{x}_i\right)^T.$$

The $\|Z - \mathbb{E}\, Z\|_{\infty \to \infty}$ is the sum of the absolute values of the d entries in a row. From Lemma 4.15,

$$\|Z - \mathbb{E}\, Z\|_{\infty \to \infty} = \|Z_1 - \mathbb{E}\, Z_1\|_{\infty \to \infty} + \|Z_2 - \mathbb{E}\, Z_2\|_{\infty \to \infty} \le \frac{\eta}{2dk} \cdot d + \frac{\eta}{2dk} \cdot d = \eta/k.$$

Finally, we have from upper bound of the error in the individual blocks that

$$\left\|F'(\mathbf{x}) - \widetilde{F'}(\mathbf{x})\right\|_{\infty \to \infty} \le \max_{j \in [k]} \sum_{i \in [k]} \left\|\nabla_{\mathbf{x}_i} F_j(\mathbf{x}) - \widetilde{\nabla_{\mathbf{x}_i} F_j}(\mathbf{x})\right\|_{\infty \to \infty} \le \eta.$$

□

5 Sample Complexity Upper Bounds with $\Omega(\sqrt{\log k})$ Separation

We now show that a mean separation of order $\Omega(\sqrt{\log k})$ suffices to learn the model parameters up to arbitrary accuracy $\delta > 0$, with $\text{poly}(d, k, \log(1/\delta))$ samples. In all the bounds that follow, the interesting settings of parameters are when $\rho, 1/w_{\min} \le \text{poly}(k)$. We note that these upper bounds are interesting even in the case of uniform mixtures: w_i, σ_i being equal across the components.

For sake of exposition, we will restrict our attention to the case when the standard deviations σ_i, and weights w_i are known for all $i \in [k]$. We believe that similar techniques can also see used to handle unknown σ_i, w_i as well (see Remark 4.6). In what follows ρ_σ corresponds to the aspect ratio of the covariances i.e., $\rho_\sigma = \max_{i \in [k]} \sigma_i / \min_{i \in [k]} \sigma_i$.

Theorem 5.1 (Same as Theorem 1.3). *There exists a universal constant $c > 0$ such that suppose we are given samples from a mixture of spherical Gaussians $\mathcal{G} = \{(w_i, \mu_i, \sigma_i) : i \in [k]\}$ (with known weights and variances) that are ρ-bounded and the means are well-separated i.e.*

$$\forall i, j \in [k], i \ne j : \ \|\mu_i - \mu_j\|_2 \ge c\sqrt{\log(\rho_\sigma / w_{min})}(\sigma_i + \sigma_j), \tag{49}$$

there is an algorithm that for any $\delta > 0$, uses $\text{poly}(k, d, \rho, \log(1/w_{min}), 1/\delta)$ samples and recovers with high probability the means up to δ relative error i.e., finds $\{\mu_i' : i \in [k]\}$ such that $\Delta_{\text{param}}\Big(\mathcal{G}, \{(w_i, \mu_i', \sigma_i) : i \in [k]\}\Big) \le \delta$.

Such results are commonly referred to as *polynomial identifiability* or *robust identifiability* results. We can again assume as in Section 4 that without loss of generality that $d \le k$ due to the following dimension-reduction technique using PCA [45]. Theorem 5.1 follows in a straightforward manner by combining the iterative algorithm, with initializers given by the following theorem.

Theorem 5.2 (Initializers Using Polynomial Samples)**.** *For any constant $c \geq 10$, suppose we are given samples from a mixture of spherical Gaussians $\mathcal{G} = \{(w_i, \mu_i, \sigma_i) : i \in [k]\}$ that are ρ-bounded and the means are well-separated i.e.*

$$\forall i, j \in [k], i \neq j : \; \|\mu_i - \mu_j\|_2 \geq 4c\sqrt{\log(\rho_\sigma / w_{min})}(\sigma_i + \sigma_j). \tag{50}$$

There is an algorithm that uses $\mathrm{poly}(k^c, d, \rho)$ *samples and with high probability learns the parameters of $\mathcal{G}$ up to k^{-c} accuracy, i.e., finds another mixture of spherical Gaussians $\tilde{\mathcal{G}}$ that has parameter distance $\Delta_{\mathrm{param}}(\mathcal{G}, \tilde{\mathcal{G}}) \leq k^{-c}$.*

The key difference between Theorem 5.1 and Theorem 5.2 is that in the former, the parameter estimation accuracy is *independent* of the separation (the constant $c > 0$ does not depend on δ). If $\rho_s = k^{O(1)}$ and $w_{\min} \geq k^{-O(1)}$, then we need means μ_j, μ_ℓ to be separated by $\Omega(\sqrt{\log k})(\sigma_j + \sigma_\ell)$ to get reasonable estimates of the parameters with poly(k) samples. While the algorithm is sample efficient, it takes time that is $(\rho / w_{\min})^{O(ck^2)}$ time since it runs over all possible settings of the $O(k)$ parameters in $d \leq k$ dimensions. Note that the above theorem holds even with unknown variances and weights that are unequal.

We can use Theorem 5.2 as a black-box to obtain a mixture $\mathcal{G}^{(0)}$ such that $\Delta_{\mathrm{param}}(\mathcal{G}, \mathcal{G}^{(0)}) \leq k^{-c}$ whose parameters will serve as initializers for the iterative algorithm in Section 4. Theorem 5.2 follows by exhibiting a lower bound on the statistical distance between any two sufficiently separated spherical Gaussian mixtures which have non-negligible parameter distance.

Proposition 5.3. *Consider any two spherical Gaussian mixtures $\mathcal{G}, \mathcal{G}^*$ in $\mathbb{R}^d$ as in Theorem 5.2, with their corresponding p.d.fs being f and f^* respectively. There is a universal constant $c' > 0$, such that for any $c \geq 5$, if the parameter distance $\Delta_{\mathrm{param}}(\mathcal{G}, \mathcal{G}^*) \geq \frac{1}{k^{c-1}}$, and $\mathcal{G}$ is well-separated:*

$$\forall i, j \in [k], i \neq j : \; \|\mu_i - \mu_j\|_2 \geq 4c\sqrt{\log(\rho_s / w_{min})}(\sigma_i + \sigma_j),$$

and both mixtures have minimum weight $w_{min} \geq 1/k^c$, then

$$\|f - f^*\|_1 = \int_{\mathbb{R}^d} |f(x) - f^*(x)| \, dx \geq \frac{c'}{dk^{2c}\rho_s^2}, \tag{51}$$

where $\rho_s = \max\left\{ \frac{\max_{j \in [k]} \sigma_j^*}{\min_{j \in [k]} \sigma_j}, \frac{\max_{j \in [k]} \sigma_j}{\min_{j \in [k]} \sigma_j^*} \right\}$.

In the above proposition, $\rho_s \leq \rho^2$ is a simple upper bound since we will only search for all parameters of magnitude at most ρ. But it can be much smaller if we have a better knowledge of the range of $\{\sigma_j : j \in [k]\}$. We note that in very recent independent work, Diakonikolas et al. established a similar statement about mixtures of Gaussians where the components have small overlap (see Appendix B in [20]). We first see how the above proposition implies Theorem 5.2.

5.1 Proposition 5.3 to Theorem 5.2

The following simple lemma gives a sample-efficient algorithm to find a distribution from a net of distributions $\mathcal{T}$ that is close to the given distribution. This tournament-based argument is a commonly used technique in learning distributions.

Lemma 5.4. *Suppose $\mathcal{T}$ is a set of probability distributions over $\mathcal{X}$, and we are given m samples from a distribution D, which is δ close to some distribution $D'' \in \mathcal{T}$ in statistical distance, i.e., $\|D - D''\|_{TV} \leq \delta$. Then there is an algorithm that uses $m = O(\delta^{-2} \log|\mathcal{T}|)$ samples from D and with probability at least $1 - 1/|\mathcal{T}|$ finds a distribution D^* such that $\|D - D^*\|_{TV} \leq 4\delta$.*

Proof. Let $\mathcal{T} = \{D_1, D_2, \dots, D_{|\mathcal{T}|}\}$ be the set of distributions over $\mathcal{X}$. For any $i \neq j$ let $A_{ij} \subset \mathcal{X}$ be such that

$$\underset{x \leftarrow D_i}{\mathbb{P}}[x \in A_{ij}] - \underset{x \leftarrow D_j}{\mathbb{P}}[x \in A_{ij}] = \|D_i - D_j\|_{TV} \ . \tag{52}$$

The algorithm is as follows.

- Use $m = O(\delta^{-2} \log|\mathcal{T}|)$ samples to obtain estimates p_{ij} satisfying with probability at least $1 - 1/|\mathcal{T}|$ that
$$\forall i, j, \ |p_{ij} - \underset{x \leftarrow D}{\mathbb{P}}[x \in A_{ij}]| \leq \delta/2 \ . \tag{53}$$
- Output the first distribution $D_i \in \mathcal{T}$ that satisfies
$$\forall j \in \{1, \dots, |\mathcal{T}|\}, \ \left|p_{ij} - \underset{x \leftarrow D_i}{\mathbb{P}}[x \in A_{ij}]\right| \leq 3\delta/2 \ . \tag{54}$$

First, notice that by (53) and the assumption that $\|D - D''\|_{TV} \leq \delta$, D'' satisfies the test in (54). We next observe by the definition of A_{ij} in (52) that for any $i \neq j$ such that both D_i and D_j pass the test, we must have $\|D_i - D_j\|_{TV} \leq 3\delta$. This implies that the output of the algorithm must be within statistical distance 4δ of D, as desired. □

We now prove Proposition 5.2 assuming the above Proposition 5.3. This follows in a straightforward manner by using the algorithm from Lemma 5.4 where $\mathcal{T}$ is chosen to be a net over all possible configuration of means, variances and weights. We give the proof below for completeness.

Proof of Proposition 5.2. Set $\delta := c'/(8\rho^8 dk^{c+1})$, where c' is given in Proposition 5.3, and $\varepsilon = \delta/(6kd\rho)$. We first pick $\mathcal{T}$ to be distributions given by an ε-net over the parameter space, and the algorithm will only output one of the mixtures in the net. Each Gaussian mixture has a standard deviation $\sigma_j \in \mathbb{R}$ and k components each of which has a mean in $\mathbb{R}^d$, and weight in $[0, 1]$. Further, all the means, variances are ρ bounded. Hence, considering a net $\mathcal{T}$ of size $M \leq (\rho/(w_{\min}\varepsilon))^{(d+2)k}$ of ρ-bounded well-separated mixtures with weights at least $w_{\min}$, we can ensure that every Gaussian mixture is ε close to some ρ bounded mixture in the net in parameter distance Δ_{param}. This corresponds to a total variation distance distance of at most $\delta = 6k\sqrt{d}\rho\varepsilon$ as well, using Lemma A.3. Hence, there is a mixture of well-separated ρ-bounded spherical Gaussian mixture $\mathcal{G}'' \in \mathcal{T}$ such that $\|\mathcal{G}'' - \mathcal{G}\|_{TV} \leq 6k\sqrt{d}\rho\varepsilon = \delta$.

By using Lemma 5.4, we can use $m = O(\delta^{-2} \log|\mathcal{T}|) = O(k^{2c+3}(d+2)^3\rho^{16} \log(\rho kd/w_{\min}))$ and with high probability, find some well-separated ρ-bounded mixture of spherical Gaussians $\mathcal{G}^*$ such that $\|\mathcal{G} - \mathcal{G}^*\|_{TV} \leq 4\delta$. Proposition 5.3 implies that the parameters of $\mathcal{G}^*$ satisfy $\Delta_{\text{param}}(\mathcal{G}, \mathcal{G}^*) \leq k^{-c}$.

□

5.2 Proof of Proposition 5.3

To show Proposition 5.3, we will consider any two mixtures of well-separated Gaussians, and show that the statistical distance is at least inverse polynomial in k. This argument becomes particularly tricky when the different components can have different values of σ_i e.g., instances where one component of $\mathcal{G}$ with large σ_i is covered by multiple components from $\mathcal{G}^*$ with small σ_j^* values.

For each component $j \in [k]$ in $\mathcal{G}$, we define the region S_j around μ_j, where we hope to show a statistical distance.

Definition 5.5 (Region S_j)**.** In the notation above, for the component of $\mathcal{G}$ centered around μ_j, define $\widehat{e}_{j\ell}$ as the unit vector along $\mu^*_\ell - \mu_j$.

$$S_j = \{x \in \mathbb{R}^d : \forall \ell \in [k] \ \ |\langle x - \mu_j, \widehat{e}_{j\ell}\rangle| \le 2c\sqrt{\log(\rho_\sigma/w_{\min})}\sigma_j\}. \tag{55}$$

The following lemma shows that most of the probability mass from jth component around μ_j is confined to S_j.

Lemma 5.6.

$$\forall j \in [k], \int_{S_j} g_{\mu_j,\sigma}(x)dx \ge 1 - \left(\frac{w_{min}}{\rho_\sigma}\right)^{4c}. \tag{56}$$

Proof. The region S_j is defined by k equations, one for each of the components ℓ in $\mathcal{G}^*$. When x is drawn from the spherical Gaussian g_{μ_j,σ_j}

$$\underset{x \leftarrow G_{\mu_j,\sigma_j}}{\mathbb{P}}\left[|\langle x - \mu_j, \widehat{e}_{j\ell}\rangle| > 2c\sigma_j\sqrt{\log(\rho_\sigma/w_{\min})}\right] \le \Phi_{0,1}\left(2c\log(\rho_\sigma/w_{\min})\right) \le \left(\frac{w_{\min}}{\rho_\sigma}\right)^{2c^2} \le \frac{1}{k}\left(\frac{w_{\min}}{\rho_\sigma}\right)^{4c},$$

where the last step follows since $c \ge 5$ and $w_{\min} \le 1/k$. Hence, performing a union bound over the k components in $\mathcal{G}^*$ completes the proof. □

The following lemma shows that components of $\mathcal{G}^*$ that are far from μ_j do not contribute much ℓ_1 mass to S_j.

Lemma 5.7. *For a component $j \in [k]$ in $\mathcal{G}$, and let $\ell \in [k]$ be a component of $\mathcal{G}^*$ such that $\|\mu_j - \mu^*_\ell\|_2 \ge 2c\sqrt{\log(\rho_\sigma/w_{min})}(\sigma_j + \sigma^*_\ell)$. Then the total probability mass from the ℓth component of $\mathcal{G}^*$ is negligible, i.e.,*

$$\int_{S_j} w^*_\ell g_{\mu^*_\ell,\sigma^*_\ell}(x)dx < w^*_\ell\left(\frac{w_{min}}{\rho_\sigma}\right)^{2c^2}. \tag{57}$$

Proof. Let $\widehat{e}_{j\ell}$ be the unit vector along $\mu^*_\ell - \mu_j$. We will now show that every point $x \in S_j$, is far from μ^*_ℓ along the direction e_{jl}:

$$\begin{aligned} x \in S_j &\implies \ |\langle x - \mu_j, \widehat{e}_{j\ell}\rangle| \le 2c\sqrt{\log(\rho_\sigma/w_{\min})} \cdot \sigma_j \\ \|\mu_j - \mu^*_\ell\|_2 \ge 2c\sqrt{\log(\rho_\sigma/w_{\min})}(\sigma_j + \sigma^*_\ell) &\implies \ \langle \mu^*_\ell - \mu_j, \widehat{e}_{j\ell}\rangle \ge 2c\sqrt{\log(\rho_\sigma/w_{\min})}(\sigma_j + \sigma^*_\ell) \\ \text{Hence } \forall x \in S_j & \ \ \ \langle \mu^*_\ell - x, \widehat{e}_{j\ell}\rangle \ge 2c\sqrt{\log(\rho_\sigma/w_{\min})} \cdot \sigma^*_\ell \end{aligned}$$

Hence the ℓ_1 contribution from $g_{\mu_\ell^*,\sigma^*}$ restricted to S_j is bounded as

$$\int_{S_j} w_\ell^* g_{\mu_\ell^*,\sigma_\ell^*}(x)dx \leq \int_{x:\langle \mu_\ell^*-x,\widehat{e}_{j\ell}\rangle \geq 2c\sqrt{\log(\rho_\sigma/w_{\min})}\sigma_\ell^*} w_\ell^* g_{\mu_\ell^*,\sigma_\ell^*}(x)\ dx$$

$$\leq w_\ell^* \Phi_{0,1}\left(2c\log(\rho_\sigma/w_{\min})\right) \leq w_\ell^* \left(\frac{w_{\min}}{\rho_\sigma}\right)^{2c^2},$$

as required. □

The following lemma shows that there is at most one component of $\mathcal{G}^*$ that is close to the component $(w_j, \mu_j, \sigma_j^2 I)$ in $\mathcal{G}$.

Lemma 5.8 (Mapping between centers)**.** *Suppose we are given two spherical mixtures of Gaussians $\mathcal{G}$, $\mathcal{G}^*$ as in Proposition 5.3. Then for every $j \in [k]$ there is at most one $\ell \in [k]$ such that*

$$\|\mu_j - \mu_\ell^*\|_2 \leq 4c\sqrt{\log(\rho_\sigma/w_{min})}\sigma_\ell^*. \tag{58}$$

Proof. This follows from triangle inequality. Suppose for contradiction that there are two centers corresponding to indices ℓ_r with $r = 1, 2$, that satisfy (58). By using triangle inequality this shows that $\|\mu_{\ell_1}^* - \mu_{\ell_2}^*\| \leq 4c\sqrt{\log(\rho_\sigma/w_{\min})}(\sigma_{\ell_1}^* + \sigma_{\ell_2}^*)$ which contradicts the assumption. □

We now proceed to the proof of the main proposition (Proposition 5.3) of this section. We will try to match up components in $\mathcal{G}, \mathcal{G}^*$ that are very close to each other in parameter distance and remove them from their respective mixtures. Then we will consider among unmatched components the one with the *smallest variance*. Suppose (w_j, μ_j, σ_j) were this component, we will show a significant statistical distance in the region S_j around μ_j.

The following lemma considers two components, $G_j = (w_j, \mu_j, \sigma_j)$ from $\mathcal{G}$, and $G_j^* = (w_j^*, \mu_j^*, \sigma_j^*)$ from $\mathcal{G}^*$ that have a non-negligible difference in parameters (we use the same index j for convenience, since this is without loss of generality). This lemma shows that if $\sigma_j \leq \sigma_j^*$, then there is some region $S \subset S_j$ where the component G_j has significantly larger probability mass than G_j^*. We note that it is crucial for our purposes that we obtain non-negligible statistical distance in a region around S_j. Suppose f_j and f_j^* are the p.d.f.s of the two components (with weights), it is easier to lower bound $\|f_j - f_j^*\|_1$ (e.g. Lemma 38 in [35]). However, this does not translate to a corresponding lower bound restricted to region S i.e., $\|f_j - f_j^*\|_{1,S}$ since f_j and f_j^* do not represent distributions (e.g. $\|f_j\|_1 = w_j < 1$).

Lemma 5.9. *For some universal constant $c_1 > 0$, suppose we are given two spherical Gaussian components with parameters (w_j, μ_j, σ_j) and $(w_j^*, \mu_j^*, \sigma_j^*)$ that are ρ_σ-bounded satisfying*

$$\sigma_j \leq \sigma_j^* \text{ and } \frac{\|\mu_j - \mu_j^*\|_2}{\sigma_j} + \frac{|\sigma_j - \sigma_j^*|}{\sigma_j} + |w_j - w_j^*| \geq \gamma. \tag{59}$$

Then, there exists a set $S \subset S_j$ such that

$$\int_S \left| w_j g_{\mu_j,\sigma_j}(x) - w_j^* g_{\mu_j^*,\sigma_j^*}(x) \right| dx > \frac{c_1\gamma^2}{d\rho_s^2}. \tag{60}$$

Before we proceed, we present two lemmas which lower bound the statistical distance when the means of the components differ, or if the means are identical but the variances differ.

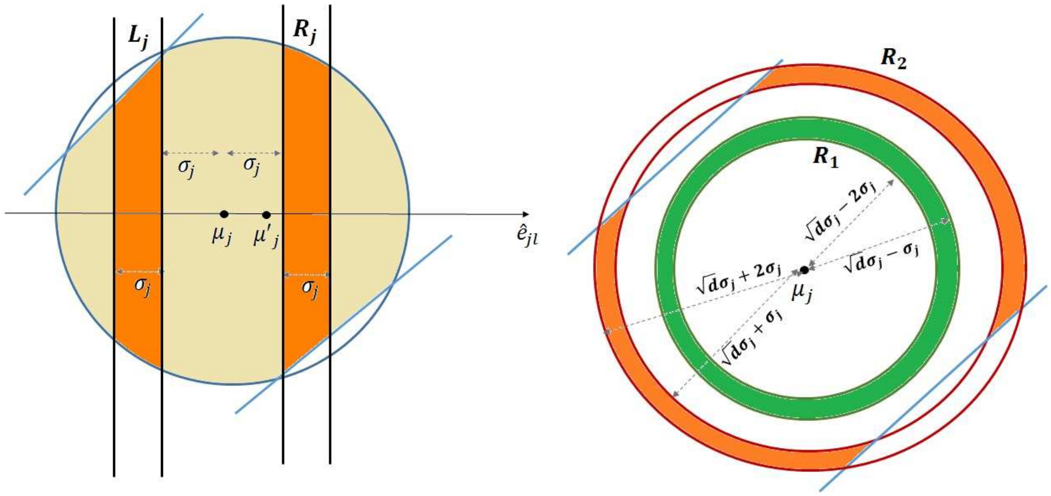
L_j
R_j
σ_j
σ_j
μ_j
μ'_j
$\hat{e}_{jl}$
σ_j
σ_j
R_2
R_1
$\sqrt{d}\sigma_j - 2\sigma_j$
$\sqrt{d}\sigma_j + 2\sigma_j$
$\sqrt{d}\sigma_j - \sigma_j$
μ_j
$\sqrt{d}\sigma_j + \sigma_j$

Figure 1: *Regions S_j used in Lemma 5.10 and Lemma 5.11:* Lemma 5.10 shows a statistical difference in one of the strips L_j or R_j. Lemma 5.11 shows a statistical difference in one of the annular strips R_1 or R_2

Lemma 5.10 (Statistical Distance from Separation of Means)**.** *In the notation of Lemma 5.9, suppose $\|\mu_j - \mu_j^*\|_2 \geq \gamma\sigma_j$. Then, there exists a set $S \subset S_j$ such that*

$$\int_S \left| w_j g_{\mu_j,\sigma_j}(x) - w_j^* g_{\mu_j^*,\sigma_j^*}(x) \right| dx > \frac{\gamma\sigma_j^2}{8(\sigma_j^*)^2} > \frac{\gamma}{8\rho_s^2}. \tag{61}$$

Proof. Without loss of generality we can assume that $\mu_j = 0$ (by shifting the origin to μ_j). Let us consider two regions

$$R_j = S_j \cap \{x : \langle x, \widehat{e}_{jj}\rangle \in [\tfrac{1}{\sqrt{2\pi}}\sigma_j, \tfrac{2}{\sqrt{2\pi}}\sigma_j]\} \text{ and } L_j = S_j \cap \{x : \langle x, \widehat{e}_{jj}\rangle \in [-\tfrac{2}{\sqrt{2\pi}}\sigma_j - \tfrac{1}{\sqrt{2\pi}}\sigma_j]\}.$$

Due to the symmetry of S_j (Definition 5.5), it is easy to see that

$$\int_{L_j} f(x)dx = \int_{R_j} f(x)dx \geq \frac{1}{8}. \tag{62}$$

On the other hand, we will now show that

$$\int_{R_j} f^*(x)dx > \left(1 + \frac{2\gamma\sigma_j^2}{\sigma_j^{*2}}\right) \int_{L_j} f^*(x)dx \tag{63}$$

For any point x, let $x_{jl} = \langle x, \widehat{e}_{jj}\rangle$. Further, $\mu_j^* = \|\mu_j^*\|\widehat{e}_{jj}$. We first note that $x \in R_j \iff (-x) \in L_j$ due to the symmetric definition of S_j. Hence,

$$\begin{aligned}\int_{L_j} f^*(x)dx &= \int_{R_j} \frac{f^*(-x)}{f^*(x)} f^*(x)dx = \int_{R_j} \exp\left(\frac{-\pi}{\sigma_j^{*2}}(\|-x-\mu_j^*\|_2^2 - \|x-\mu_j^*\|_2^2)\right) f^*(x)dx \\ &= \int_{R_j} \exp\left(-2\pi\langle x, \mu_j^*\rangle/\sigma_j^{*2}\right) f^*(x)dx = \int_{R_j} \exp\left(-2\pi\|\mu_j\|_2\langle x, \widehat{e}_{jj}\rangle/(\sigma_j^*)^2\right) f^*(x)dx \\ &\leq \exp(-\sqrt{2\pi}\sigma_j^2\gamma/\sigma_j^{*2}) \int_{R_j} f^*(x)dx, \ \text{ since } \langle x, \widehat{e}_{jj}\rangle \geq \frac{\sigma_j}{\sqrt{2\pi}} \text{ for } x \in R_j.\end{aligned}$$

Hence from (62) and (63) either

$$\frac{\int_{L_j} f(x)dx}{\int_{L_j} f^*(x)dx} > 1 + \frac{\sqrt{2\pi}\gamma\sigma_j^2}{\sigma_j^{*2}} \quad \text{or} \quad \frac{\int_{R_j} f^*(x)dx}{\int_{R_j} f(x)dx} > 1 + \frac{\sqrt{2\pi}\gamma\sigma_j^2}{\sigma_j^{*2}},$$

which completes the proof, since $\int_{L_j} f(x)dx = \int_{R_j} f(x)dx \geq 1/8$. □

Lemma 5.11 (Statistical Distance from Different Variances)**.** *In the notation of Lemma 5.9, suppose $\mu_j = \mu_j^*$, but $\sigma_j = (1-\eta)\sigma_j^*$. Then, there exists a set $S \subset S_j$ such that*

$$\int_S \left| w_j g_{\mu_j,\sigma_j}(x) - w_j^* g_{\mu_j^*,\sigma_j^*}(x) \right| dx > \min\{c_2\eta, 1\}, \tag{64}$$

where $c_2 > 0$ is a universal constant.

Proof. Without loss of generality, let us assume that $\mu_j = \mu_j^* = 0$ (by shifting the origin), and $\sigma_j = 1$ (by scaling). Let $R_1 = [\frac{1}{\sqrt{2\pi}}(\sqrt{d}-2), \frac{1}{\sqrt{2\pi}}(\sqrt{d}-1)]$ and $R_2 = [\frac{1}{\sqrt{2\pi}}(\sqrt{d}+1), \frac{1}{\sqrt{2\pi}}(\sqrt{d}+2)]$ and consider two annular strips $T_1 = \{x : \|x\| \in R_1\}$ and $T_2 = \{x : \|x\| \in R_2\}$.

First, we note that using standard facts about the $\chi^2(d)$ distribution,for some appropriately chosen universal constant $c_2 > 0$ we have

$$\int_{T_1} f(x)dx \geq c_2 \ , \quad \text{and} \quad \int_{T_2} f(x)dx \geq c_2.$$

We will show that there is a significant statistical difference between f, f^* in either T_1 or T_2. Let us assume for contradiction that

$$(1-\eta) \leq \frac{\int_{T_1} f(x)dx}{\int_{T_1} f^*(x)dx} \leq (1+\eta) \quad \text{and} \quad (1-\eta) \leq \frac{\int_{T_2} f(x)dx}{\int_{T_2} f^*(x)dx} \leq (1+\eta). \tag{65}$$

Let us consider the range of values that f takes in T_1 and T_2. We

$$\forall x \in T_1 \ \frac{w_j}{\sigma_j^d} \exp\left(-(\sqrt{d}-1)^2/2\right) \leq f(x) \leq \frac{w_j}{\sigma_j^d} \exp\left(-(\sqrt{d}-2)^2/2\right) \tag{66}$$

$$\forall x \in T_2 \ \frac{w_j}{\sigma_j^d} \exp\left(-(\sqrt{d}+2)^2/2\right) \leq f(x) \leq \frac{w_j}{\sigma_j^d} \exp\left(-(\sqrt{d}+1)^2/2\right) \tag{67}$$

$$\forall x_1 \in T_1, x_2 \in T_2, \ \frac{f(x_1)}{f(x_2)} \geq \exp\left(-\tfrac{1}{2}\left((\sqrt{d}-1)^2) - (\sqrt{d}+1)^2\right)\right) \geq \exp(2\sqrt{d}) \geq e^2. \tag{68}$$

Let $A(r)$ be the $d-1$ dimensional volume of $\{x \in S_j : \|x\| = r\}$. From (65), we have that

$$\begin{aligned}
\frac{\int_{T_1} f(x)dx}{\int_{T_1} f^*(x)dx} &= \frac{w_j \int_{r\in R_1} \exp(-r^2/2)A(r)\,dr}{w_j^*(1-\eta)^d \int_{r\in R_1} \exp(-(1-\eta)^2 r^2/2)A(r)\,dr} \\
&= \frac{w_j \int_{r\in R_1} \exp(-r^2/2)A(r)\,dr}{w_j^*(1-\eta)^d \int_{r\in R_1} \exp\left((2\eta-\eta^2)r^2/2\right) \cdot \exp(-r^2/2)A(r)\,dr} \\
&\geq \frac{w_j}{w_j^*(1-\eta)^d} \cdot \exp\left(-\eta(\sqrt{d}-1)^2/2\right).
\end{aligned}$$

Hence, $\dfrac{w_j}{w_j^*(1-\eta)^d} \leq (1+\eta)\exp\left(\eta(\sqrt{d}-1)^2/2\right).$

Using a similar argument for T_2 we get

$$\frac{w_j}{w_j^*(1-\eta)^d} \geq (1-\eta)\exp\left(\eta(\sqrt{d}+1)^2/2\right).$$

Combining the two we get

$$\begin{aligned}(1+\eta)\exp\left(\eta(\sqrt{d}-1)^2/2\right) &\geq (1-\eta)\exp\left(\eta(\sqrt{d}+1)^2/2\right)\\ \exp\left(2\eta\sqrt{d}\right) &\leq (1+\eta)^2 \leq e^{2\eta},\end{aligned}$$

which is a contradiction. □

Proof of Lemma 5.9. Let $f_j(x) = w_j g_{\mu_j,\sigma_j}(x), f_j^*(x) = w_j^* g_{\mu_j^*,\sigma_j^*}(x)$ be the p.d.f. of the two components. From (59) we know that there is some non-negligible separation in the parameters. Hence either the weights, or means, or variances are separated by at least $\Omega(\gamma)$. We will now consider three cases depending on whether there is non-negligible separation in the means, variances or weights (in that order).

Set $\gamma_1 := \gamma^2/(256d)$, $\gamma_2 := c_2\gamma_1/2$ where c_2 is the constant in Lemma 5.11. Suppose there is some non-negligible separation in the means i.e., $\|\mu_j - \mu_j^*\| \geq \gamma_2\sigma_j/\sqrt{2\pi}$. Lemma 5.10 shows that there is a set $S \subset S_j$ that has

$$\|f - f^*\|_{1,S} \geq \frac{\gamma_2\sigma_j^2}{8\sigma_j^{*2}} \geq \frac{c_1\gamma^2}{d\rho_s^2},$$

where $c_1 = c_2/(512\sqrt{2\pi})$.

Otherwise we have that $\|\mu_j - \mu_j^*\| < \gamma_2\sigma_j/\sqrt{2\pi}$. Let g_j^* be the p.d.f. of the Gaussian component $(w_j^*, \mu_j, \sigma_j^*)$. From Lemma A.3 we know that $\|f_j^* - g_j^*\|_1 \leq \gamma_2$. Now, f_j and g_j^* are p.d.f. of components that have the same mean.

If $\sigma_j^* - \sigma_j > \gamma_1\sigma_j$. From Lemma 5.11 we see that

$$\|f - g^*\|_{1,S} \geq c'\gamma_1 \implies \|f - f^*\|_{1,S} \geq c'\gamma_1 - \gamma_2 \geq c'\gamma_1/2.$$

Finally if $\sigma_j^* - \sigma_j \leq \gamma_1\sigma_j$, then one can bound the statistical distance over $S = S_j$ between two components with equal means and variances, but different weights:

$$\begin{aligned}\|f_j - f_j^*\|_{1,S} &\geq \|f_j - g_j^*\|_{1,S} - \|g_j^* - f_j^*\| = \int_S \left|w_j g_{\mu_j,\sigma_j}(x) - w_j^* g_{\mu_j,\sigma_\ell}(x)\right| dx - \gamma_2\\ &\geq \int_S \left|w_j g_{\mu_j,\sigma_j}(x) - w_j^* g_{\mu_j,\sigma_j}(x)\right| dx - \int_S \left|w_j^* g_{\mu_j,\sigma_j}(x) - w_j^* g_{\mu_j,\sigma_j^*}(x)\right| dx - \gamma_2\\ &\geq |w_j - w_j^*| \int_S g_{\mu_j,\sigma_j}(x) - 2\sqrt{d\gamma_1} - \gamma_2 \qquad (\text{ from Lemma A.3})\\ &\geq \frac{2\gamma}{3}\left(1 - (w_{\min}/\rho_\sigma)^{4c}\right) - \frac{\gamma}{8} - \frac{\gamma}{8} \geq \gamma/4,\end{aligned}$$

where the last line follows from Lemma 5.6. □

We now complete the proof of the main proposition of this section, that lower bounds the statistical difference between two mixtures of well-separated Gaussians which differ in their parameters.

Proof of Proposition 5.3. We follow the approach outlined earlier in the section. We first match up components from $(w_j, \mu_j, \sigma_j) \in \mathcal{G}$ and $(w_j^*, \mu_j^*, \sigma_j^*) \in \mathcal{G}^*$ that are at most $\gamma = k^{-c}$ close in parameter distance. From triangle inequality and well-separatedness of $\mathcal{G}, \mathcal{G}^*$, it is easy to see that one component from $\mathcal{G}$ (or $\mathcal{G}^*$) cannot be matched to more than one component from $\mathcal{G}^*$ (or $\mathcal{G}$ respectively). Let $\{1, 2, \dots, k'\}$ be the indices of the components in $\mathcal{G}$ and $\mathcal{G}^*$ that are unmatched, for $\ell \in \{k'+1, \dots, k\}$, let the component $(w_\ell, \mu_\ell, \sigma_\ell)$ of $\mathcal{G}$ be matched to the component $(w_\ell^*, \mu_\ell^*, \sigma_\ell^*)$ of $\mathcal{G}^*$. Since the parameter distance $\Delta_{\text{param}}(\mathcal{G}, \mathcal{G}^*) \geq k^{-c+1}$, we have that $k' \geq 1$.

Consider among the unmatched components in both $\mathcal{G}$ and $\mathcal{G}^*$, the one with the smallest variance: let this component be (w_j, μ_j, σ_j) from $\mathcal{G}$ without loss of generality. From Lemma 5.8, we know that at most one component of $\mathcal{G}^*$ satisfies (58). Again, without loss of generality, let $(w_j^*, \mu_j^*, \sigma_j^*)$ be this component of $\mathcal{G}^*$ (if it exists). Hence

$$\forall \ell \in [k'], \ell \neq j, \ \|\mu_j - \mu_\ell^*\|_2 > 4c\sqrt{\log(\rho_\sigma/w_{\min})}\sigma_\ell^* \geq 2c\sqrt{\log(\rho_\sigma/w_{\min})}(\sigma_j + \sigma_\ell^*),$$

where the last inequality follows because σ_j^2 is the smallest variance.[4] Further, all the matched components $\ell \in \{k'+1, \dots k\}$ in $\mathcal{G}^*$ are all far from μ_j since

$$\begin{aligned}
\forall \ell \in \{k'+1, \dots, k\}, \ \|\mu_j - \mu_\ell^*\|_2 &\geq \|\mu_j - \mu_\ell\|_2 - \|\mu_\ell - \mu_\ell^*\| \\
&\geq 4c\sqrt{\log(\rho_\sigma/w_{\min})}(\sigma_j + \sigma_\ell) - \gamma \\
&\geq 4c\sqrt{\log(\rho_\sigma/w_{\min})}(\sigma_j + \sigma_\ell^*) - \gamma(1 + 4c\sqrt{\log(\rho_\sigma/w_{\min})}) \\
&\geq 2c\sqrt{\log(\rho_\sigma/w_{\min})}(\sigma_j + \sigma_\ell),
\end{aligned}$$

since $\gamma < 2/(5\rho_\sigma)$. From Lemma 5.7, there is negligible contribution from the rest of the components (using $c \geq 3$):

$$\sum_{\ell \in [k]:\ell \neq j} w_\ell^* \int_{S_j} g_{\mu_\ell^*, \sigma_\ell^*}(x) dx < \left(\frac{w_{\min}}{\rho_\sigma}\right)^{4c}. \tag{69}$$

From Lemma 5.9 there is a subset $S \subset S_j$ where there is significant statistical distance

$$\int_S \left| w_j g_{\mu_j, \sigma_j}(x) - w_j^* g_{\mu_j^*, \sigma_j^*}(x) \right| dx > \frac{c_1 \gamma^2}{d\rho_s^2}.$$

Combining the last two equations, we have

$$\begin{aligned}
\|f - f^*\|_1 &\geq \int_S \left| \sum_\ell w_\ell g_{\mu_\ell, \sigma_\ell}(x) - \sum_\ell w_\ell^* g_{\mu_\ell^*, \sigma_\ell^*}(x) \right| \ dx \\
&\geq \int_S \left| w_j g_{\mu_j, \sigma_j}(x) - w_j^* g_{\mu_j^*, \sigma_j^*}(x) \right| dx - \sum_{\ell \neq j} \int_S w_\ell^* g_{\mu_\ell^*, \sigma_\ell^*}(x) dx \\
&\geq \frac{c_1 \gamma^2}{d\rho_s^4} - \left(\frac{1}{\rho_\sigma k}\right)^{4c} \geq \frac{c' \gamma^2}{d\rho_s^2}
\end{aligned}$$

for some universal constant $c' > 0$, since $\gamma \geq k^{-c}$. □

[4] This is in fact the only point where we use the fact that we picked the small variance Gaussian.

6 Efficient Algorithms in Low Dimensions

In this section, we give a computationally efficient algorithm that works in $d = O(1)$ dimensions, and learns the mixture of k spherical Gaussians even when the separation between centers is $O(\sigma)$. In comparison, previous algorithms need separation of the order of $\Omega(\sigma\sqrt{\log(k/\delta)})$. We prove the following theorem.

Theorem 6.1. *There exists universal constants $c > 0$ such that the following holds. Suppose we are given samples from a mixture of spherical Gaussians $\mathcal{G} = \{(w_j, \mu_j, \sigma_j) : j \in [k]\}$, where the weights and covariances are known, such that $\|\mu_j\| \le \rho \ \forall j \in [k]$ and*

$$\forall i, j \in [k], i \neq j : \ \|\mu_i - \mu_j\|_2 \ge c\left(\sqrt{d} + \sqrt{\log(\rho_\sigma \rho_w)}\right) \cdot (\sigma_i + \sigma_j). \tag{70}$$

For any $\delta > 0$, there is an algorithm using time (and samples) $\operatorname{poly}\left(w_{min}^{-1}, \delta^{-1}, \rho, \rho_\sigma\right)^{O(d)}$ that with high probability recovers the means up to δ accuracy i.e. finds for each $j \in [k]$, $\tilde{\mu}_j$ such that $\|\tilde{\mu}_j - \mu_j\|_2 \le \delta\sigma_j$.

In the above theorem, when both $\rho_w, \rho_\sigma = O(1)$ as in the case of uniform mixtures, this corresponds to a separation of order $\Omega(\sqrt{d})$.

The above theorem follows by applying the guarantees of the iterative algorithm (Theorem 4.1) along with a computationally efficient procedure that finds appropriate initializers. The following theorem shows how to find reasonable initializers for μ_j, σ_j, w_j for each of the k components.

Theorem 6.2. *Let $c_0 \ge 2$ be any constant, and suppose $\varepsilon_0 = \exp(-c_0 d)$. There is an algorithm running in time $\left(\frac{\rho d}{\varepsilon_0^3}\right)^{O(d)} \operatorname{poly}(1/w_{min})$ that given samples from a ρ-bounded mixture of k spherical Gaussians $\mathcal{G} = \{(w_j, \mu_j, \sigma_j) : j \in [k]\}$ in d dimensions satisfying*

$$\forall i \neq j \in [k], \ \|\mu_i - \mu_j\|_2 \ge 4c_0\left(\sqrt{d} + \sqrt{\log(\rho_w \rho_\sigma)}\right)(\sigma_i + \sigma_j), \tag{71}$$

can find with high probability $\{(\tilde{\mu}_j, \tilde{\sigma}_j, \tilde{w}_j) : j \in [k]\}$ s.t.

$$\forall j \in [k], \|\tilde{\mu}_j - \mu_j\|_2 \le \varepsilon_0 \sigma_j \sqrt{d}, \ |\tilde{\sigma}_j - \sigma_j| \le \varepsilon_0 \sigma_j, \ \textit{and } |\tilde{w}_j - w_j| \le \varepsilon_0 w_j.$$

Note that the above theorem also finds initializers for the weights and variances when they are unknown. Hence, Theorem 6.1 will also apply to the setting with unknown weights and variances if we get similar guarantees for Theorem 4.1 (see Remark 4.6).

To prove Theorem 6.2, we will first find estimates $\tilde{\mu}_j$ for the means μ_j (Proposition 6.3), and then obtain good estimates $\tilde{\sigma}_j, \tilde{w}_j$ for each $j \in [k]$. Proposition 6.3 already suffices in the case of known weights and variances.

Proposition 6.3. *In the notation of Theorem 6.2, there is an algorithm running in $\left(\frac{\rho}{\varepsilon_0^3 w_{min}}\right)^{O(d)}$ time that finds w.h.p. $\tilde{\mu}_1, \ldots, \tilde{\mu}_k$ s.t. $\forall j \in [k], \|\tilde{\mu}_j - \mu_j\|_2 \le \varepsilon_0 \sigma_j \sqrt{d}$.*

We first start with expressions for the first derivative (gradient) and second derivative (Hessian) at a point x in terms of the model parameters.

$$f'(x) = \nabla f(x) = -2\pi \sum_{j=1}^{k} \frac{w_j}{\sigma_j} \cdot g_{\mu_j,\sigma_j}(x) \cdot \frac{(x-\mu_j)}{\sigma_j} \tag{72}$$

$$f''(x) = \nabla^2 f(x) = 4\pi^2 \sum_{j=1}^{k} \frac{w_j}{\sigma_j^2} \cdot g_{\mu_j,\sigma_j}(x) \cdot \left(\frac{1}{\sigma_j^2}(x-\mu_j)^{\otimes 2} - \frac{1}{2\pi} I_{d\times d} \right). \tag{73}$$

Note that using $\text{poly}_d(k,\rho)$ samples, we have access to the p.d.f. $f(x)$ and its derivatives $f' = \nabla f$, $f'' = \nabla^2 f$ up to $1/\text{poly}_d(k)$ accuracy at any point x.

The algorithm will consider a δ-net of points, and find "approximate local-maxima" of the p.d.f., which are defined as follows.

Definition 6.4 (Approximate Local Maximum)**.** Consider a mixture of k spherical Gaussians in $\mathbb{R}^d$ with parameters $\{(w_j,\mu_j,\sigma_j) : j \in [k]\}$ and p.d.f. f. Then $x \in \mathbb{R}^d$ is an approximate local-maxima iff

$$\text{(i) } f(x) \geq \frac{w_{\min}}{2\sigma_{\max}^d}, \qquad \text{(ii) } \|f'(x)\|_2 \leq \frac{\pi\varepsilon_0\sqrt{d}\sigma_{\min}}{4\sigma_{\max}^2} f(x), \tag{74}$$
$$\text{(iii) } f''(x) = \nabla^2 f(x) \preceq -\frac{\pi}{2\sigma_{\max}^2} f(x) I_{d\times d} \qquad .$$

To show the above proposition, we will show that all approximate local-maxima are close to one of the means μ_j (Lemma 6.5 and Lemma 6.6), and there is at least one such approximate local maxima near each mean μ_j (Lemma 6.7). Further, since the parameters are separated, this will allow us to pick all approximate local-maxima in a net, cluster them geometrically and pick one such point from each cluster to get good initializers for each mean μ_j. These statements will also allow for some slack to tolerate estimation errors.

We start with some inequalities that use the separation between means to show the p.d.f. and the first few "moments" near one of the means μ_j is dominated by the the jth component. By applying Lemma 4.13 we get that at any point $x \in \mathbb{R}^d$ such that $\|x-\mu_j\|_2 \leq \sigma_j\sqrt{d/\pi}$

$$\sum_{i\neq j\in[k]} w_i g_{\mu_i,\sigma_i}(x) < w_j g_{\mu_j,\sigma_j}(x) \exp(-c_0 d)\Big(\frac{\sigma_{\min}}{\sigma_{\max}}\Big)^2. \tag{75}$$

$$\sum_{i\neq j\in[k]} w_i \cdot \frac{\|x-\mu_i\|_2}{\sigma_i} \cdot g_{\mu_i,\sigma_i}(x) < w_j g_{\mu_j,\sigma_j}(x) \exp(-c_0 d) \cdot \Big(\frac{\sigma_{\min}}{\sigma_{\max}}\Big)^4. \tag{76}$$

$$\sum_{i\neq j\in[k]} w_i \cdot \frac{\|x-\mu_i\|_2^2}{\sigma_i^2} \cdot g_{\mu_i,\sigma_i}(x) < w_j g_{\mu_j,\sigma_j}(x) \exp(-c_0 d) \cdot \Big(\frac{\sigma_{\min}}{\sigma_{\max}}\Big)^4. \tag{77}$$

The following lemma shows that any point that is far from all of the means is *not* an approximate local maximum (does not satisfy condition (iii) of Def. 6.4).

Lemma 6.5. *In the notation of Proposition 6.3, for any $x \in \mathbb{R}^d$ that satisfies*

$$\forall j \in [k] \quad \|x-\mu_j\|_2 > \sigma_j\sqrt{\frac{d}{\pi}}, \exists u \in \mathbb{R}^d \text{ s.t. } u^T H_x u > 0,$$

where where $H_x = f''(x)$ represents the Hessian evaluated at x. Hence, such a point x is not an approximate local maxima.

Proof. Let $H_{x,j} = (x-\mu_j)^{\otimes 2} - \sigma_j^2 I/(2\pi)$. Then

$$\forall j \in [k],\ \operatorname{tr}(H_{x,j}) = \|x-\mu_j\|^2 - \frac{d\sigma_j^2}{2\pi} > \frac{d\sigma_j^2}{2\pi} \quad \text{since } \frac{\|x-\mu_j\|_2}{\sigma_j} > \sqrt{\frac{d}{\pi}}.$$

Let v a random unit standard Gaussian vector drawn from $\mathcal{N}(0,1)^d$.

$$\begin{aligned}\frac{\mathbb{E}[v^T H_x v]}{\mathbb{E}[\|v\|_2^2]} = \frac{\operatorname{tr}(H_x)}{d} &= \frac{4\pi^2}{d}\sum_{j=1}^k \frac{w_j}{\sigma_j^{d+4}} \cdot \exp\left(-\frac{\pi\|x-\mu_j\|^2}{\sigma_j^2}\right) \cdot \operatorname{tr}(H_j)\\ &> \frac{2\pi}{d}\sum_{j=1}^k \frac{w_j}{\sigma_j^d}\exp\left(-\frac{\pi\|x-\mu_j\|^2}{\sigma_j^2}\right)\cdot \frac{d}{\sigma_j^2} > 2\pi \cdot f(x) \cdot \frac{1}{\max_j \sigma_j^2}.\end{aligned}$$

Hence, there is a direction $v^T H_x v \geq 2\pi f(x)\|v\|_2^2/\sigma_{\max}^2$. □

The following lemma shows that we cannot have approximate local maxima (or more generally, critical points) whose distance from μ_j is between $[\varepsilon_0\sqrt{d}\sigma_{\min}, \sqrt{d/\pi}\cdot\sigma_j]$. Hence, together with Lemma 6.5, this shows that every approximate local maximum is within $\varepsilon_0\sqrt{d}\sigma_{\min}$ from one of the true means.

Lemma 6.6. *In the notation of Proposition 6.3, for any $\varepsilon_0 = 2\exp(-c_0 d)$ with $c_0 \geq 2$, suppose $x \in \mathbb{R}^d$ satisfies for some component $j \in [k]$, $\varepsilon_0\sqrt{d}\sigma_{\min} < \|x-\mu_j\|_2 \leq \sqrt{d/\pi}\sigma_j$, then*

$$\|f'(x)\|_2 = \|\nabla f(x)\|_2 > \frac{\varepsilon_0\sqrt{d}\sigma_{\min}}{4\sigma_j^2}\cdot f(x) \geq \frac{\varepsilon_0\sqrt{d}\sigma_{\min}}{4\sigma_{\max}^2}\cdot f(x).$$

Proof. From (72), the first derivative satisfies

$$\|f'(x)\|_2 \geq \frac{2\pi w_j\|x-\mu_j\|_2}{\sigma_j^2}\cdot g_{\mu_j,\sigma_j}(x) - \sum_{i\neq j\in[k]} \frac{2\pi w_i}{\sigma_i}\cdot\frac{\|x-\mu_i\|_2}{\sigma_i}\cdot g_{\mu_i,\sigma_i}(x). \tag{78}$$

Applying (76) with the given separation,

$$\sum_{i\neq j\in[k]}\frac{w_i}{\sigma_i}\cdot\frac{\|x-\mu_i\|_2}{\sigma_i}\cdot g_{\mu_i,\sigma_i}(x) < \frac{1}{\sigma_{\min}}\cdot w_j g_{\mu_j,\sigma_j}(x)\exp(-c_0 d)\cdot\left(\frac{\sigma_{\min}}{\sigma_{\max}}\right)^2 < \frac{\varepsilon_0 w_j\sigma_{\min}}{2\sigma_{\max}^2} g_{\mu_j,\sigma_j}(x). \tag{79}$$

Further, $\|x-\mu_j\|_2 \geq \varepsilon_0\sqrt{d}\sigma_j$. Using (78) and (79),

$$\begin{aligned}\|f'(x)\|_2 &\geq 2\pi w_j g_{\mu_j,\sigma_j}(x)\cdot\frac{\|x-\mu_j\|_2^2}{\sigma_j^2} - \sum_{i\neq j}^k 2\pi w_i g_{\mu_i,\sigma_i}(x)\cdot\frac{\|x-\mu_i\|_2^2}{\sigma_i^2}\\ &\geq \frac{2\pi\|x-\mu_j\|_2}{\sigma_j^2}\cdot w_j g_{\mu_j,\sigma_j}(x) - \frac{\pi\varepsilon_0 w_j\sigma_{\min}}{\sigma_{\max}^2} g_{\mu_j,\sigma_j}(x) \quad \text{(from (79))},\\ &\geq \frac{\pi\varepsilon_0\sqrt{d}\sigma_{\min}}{4\sigma_{\max}^2}\cdot f(x),\end{aligned}$$

where the last inequality follows from (75) and using $\|x-\mu_j\|_2 > \varepsilon_0\sqrt{d}\sigma_{\min}$. □

We now proceed to the proof of Lemma 6.7, which shows that any point that is sufficiently close to one of the component means is an approximate local maxima. This shows that in any δ-net (for sufficiently small $\delta < \varepsilon_0\sqrt{d}\sigma_{\min}^2/\sigma_{\max}$) , there will be an approximate local maxima.

Lemma 6.7. *In the notation of Proposition 6.3, for any $\varepsilon' \le \exp(-c_0 d)$ with $c_0 \ge 2$ and $\forall j \in [k]$, any point $x \in \mathbb{R}^d$ with $\|x-\mu_j\|_2 \le \frac{\varepsilon'\sqrt{d}}{32}\cdot\frac{\sigma_j^2}{\sigma_{\max}}$ is an approximate local maxima. In particular,*

$$\textit{(i) } f(x) \ge \frac{3w_{min}}{4\sigma_{\max}^d}, \qquad \textit{(ii) } \|f'(x)\|_2 \le \frac{4\pi\varepsilon'\sqrt{d}}{\sigma_{\max}} f(x), \qquad (80)$$
$$\textit{(iii) } f''(x) = \nabla^2 f(x) \preceq -\frac{\pi}{\sigma_j^2} f(x) I \preceq -\frac{3\pi}{4\sigma_{\max}^2} f(x) I \qquad .$$

Proof of Lemma 6.7. The lemma follows in a straightforward way from (75), (76), (77), since f, f', f'' at x are dominated by the jth component. Firstly by considering just the contribution to the p.d.f. from the jth component, the lower bound on $f(x)$ follows. Now we bound $\|f'(x)\|_2$.

$$\begin{aligned} \|f'(x)\|_2 &\le 2\pi \sum_{i=1}^{k} w_i g_{\mu_i,\sigma_i}(x)\cdot\frac{\|x-\mu_i\|_2}{\sigma_i^2} \le 4\pi w_j g_{\mu_j,\sigma_j}(x)\cdot\frac{\|x-\mu_j\|_2}{\sigma_j^2} \quad \text{(from (76)).} \\ &\le \frac{4\pi\varepsilon'\sqrt{d}}{\sigma_{\max}}\cdot w_j g_{\mu_j,\sigma_j}(x) \le \frac{4\pi\varepsilon'\sqrt{d}}{\sigma_{\max}}\cdot f(x). \end{aligned}$$

We argue about $f''(x)$ similarly. Suppose we use M to denote the following matrix, and $\|M\|$ to represent its maximum singular value,

$$\begin{aligned} M &= 4\pi^2 \sum_{i\ne j} w_i g_{\mu_i,\sigma_i}(x)\cdot\frac{1}{\sigma_i^2}\left(\frac{1}{\sigma_i^2}(x-\mu_i)^{\otimes 2} - \frac{1}{2\pi} I_{d\times d}\right). \\ \|M\| &\le 4\pi^2 \sum_{i\ne j} w_i g_{\mu_i,\sigma_i}(x)\cdot\frac{1}{\sigma_i^2}\left(\frac{\|x-\mu_i\|_2^2}{\sigma_i^2} + \frac{1}{2\pi}\right) \\ &\le \frac{4\pi^2}{\sigma_{\max}^2} w_j g_{\mu_j,\sigma_j}(x)\exp(-c_0 d) < \frac{\pi}{2\sigma_j^2} w_j g_{\mu_j,\sigma_j}(x), \end{aligned}$$

where the last line follows from (75), (77) and since $\exp(-c_0 d) < 1/(8\pi)$. Further $2\pi\|x-\mu_j\|_2^2/\sigma_j^2 < 1/4$. Substituting in (73), we get

$$\begin{aligned} f''(x) &\preceq \frac{2\pi}{\sigma_j^2} w_j g_{\mu_j,\sigma_j}(x)\left(-I + \frac{2\pi\|x-\mu_j\|_2^2}{\sigma_j^2} I\right) + \|M\| I \\ &\preceq -\frac{2\pi}{\sigma_j^2} w_j g_{\mu_j,\sigma_j}(x)\left(I - \tfrac{1}{4}I - \tfrac{1}{4}I\right) \preceq \frac{-\pi}{\sigma_j^2} w_j g_{\mu_j,\sigma_j}(x) \preceq \frac{-3\pi}{4\sigma_j^2} f(x), \end{aligned}$$

where the last inequality follows from (75). □

We now proceed to the algorithm and proof of Proposition 6.3.

Proof of Proposition 6.3. The algorithm first considers a δ-net $\mathcal{X}_\delta$ in $\mathbb{R}^d$ over a ball of radius 2ρ, and estimates $f(y)$ up to additive accuracy γ where $\gamma = w_{\min}\sigma_{\max}^{-(d+4)}\varepsilon_0^3\delta^2/4$ and $\delta = \varepsilon_0\sqrt{d}\sigma_{\min}^3/(64\sigma_{\max}^2)$

will suffice[5]. Similarly, we can also estimate $f'(y)$ in ℓ_2 norm, and $f''(y)$ in operator norm within additive accuracy γ. The size of the net is $|\mathcal{X}_\delta| = (\rho/\delta)^{O(d)}$, and the sample complexity is $O(1/\gamma^2) \cdot (\rho/\delta)^{O(d)}$ samples.

From Lemma 6.6 and Lemma 6.5 we have that if

$$f(x) \geq w_{\min}\sigma_{\max}^{-d}/2, \quad f''(x) \preceq -\frac{\pi}{2\sigma_{\max}^2} f(x)I, \text{ and } \frac{\|f'(x)\|_2}{f(x)} \leq \frac{\pi\varepsilon_0\sqrt{d}\sigma_{\min}}{4\sigma_{\max}^2}, \tag{81}$$

then there exists $j \in [k]$ s.t. $\|x - \mu_j\|_2 \leq \varepsilon_0\sqrt{d}\sigma_{\min}$. On the other hand, applying Lemma 6.7 with $\varepsilon' = \varepsilon_0\sigma_{\min}/(32\sigma_{\max})$, any point that is within $O(\varepsilon_0\sqrt{d}\sigma_{\min}^3/\sigma_{\max}^2)$ close to μ_j satisfy

$$f(x) \geq \frac{3w_{\min}}{4\sigma_{\max}^d}, \quad f''(x) = \nabla^2 f(x) \preceq -\frac{3\pi}{4\sigma_{\max}^2} f(x)I, \text{ and } \frac{\|f'(x)\|_2}{f(x)} \leq \frac{\pi\varepsilon_0\sqrt{d}\sigma_{\min}}{8\sigma_{\max}^2}. \tag{82}$$

Our accuracy γ of estimating f, f', f'' is chosen so that we can distinguish between the bounds in (81) and (82). For convenience, since we have sufficiently accurate estimates, we will abuse notation and also use $f(x), f'(x), f''(x)$ to represent the estimate of the f, f', f'' at x.

First using our estimates, we consider all points

$$T = \Big\{ y \in \mathcal{X}_\delta \mid f(y) \geq w_{\min}\sigma_{\max}^{-d}/2, \ f''(y) \preceq -\frac{\pi}{2\sigma_{\max}^2} f(y)I, \ \ \frac{\|f'(y)\|_2}{f(y)} \leq \frac{\pi\varepsilon_0\sqrt{d}\sigma_{\min}}{4\sigma_{\max}^2} \Big\}.$$

We can find T from our estimates since $\gamma < w_{\min}\sigma_{\max}^{-(d+2)}/8$. From (81), we have that for every $y \in T$, there is some $j \in [k]$, such that $\|y - \mu_j\|_2 \leq \varepsilon_0\sqrt{d}\sigma_{\min}$.

Further the means are well separated i.e., $\|\mu_i - \mu_j\|_2 > 4(\sigma_i + \sigma_j)$. Hence, suppose we define

$$T_j^* = \{ y \in T \mid \|\mu_j - y\|_2 \leq \varepsilon_0\sqrt{d}\sigma_j \},$$

the sets $\{T_j^* : j \in [k]\}$ are disjoint and form a partition of T that is consistent with the k components of the Gaussian mixture $\mathcal{G}$. From the separation conditions we see that the distances between any two points in the same cluster T_j^* are smaller than the distance between any two points in different clusters $T_{j_1}^*$ and $T_{j_2}^*$ ($j_1 \neq j_2$). Hence, we can use single-linkage clustering algorithm (see Awasthi et al. [6] for a proof that single-linkage algorithm suffices)to find the clustering $\{T_j^* : j \in [k]\}$ in time $\text{poly}(|T|) \leq \text{poly}(|\mathcal{X}_\delta|)$.

Finally, for each $j \in [k]$, let $\tilde{\mu}_j$ be any point in T_j^*. Since, the coarseness of the net δ satisfies $\delta < \varepsilon_0\sqrt{d}\sigma_{\min}^3/(64\sigma_{\max}^2)$, we have from (82) that there is at least one point $y \in T$ close to each μ_j. Hence, $\|\tilde{\mu}_j - \mu_j\|_2 \leq \varepsilon_0\sqrt{d}\sigma_{\min}$, as required.

□

Note. In fact, the above proposition can also be used to show that f has exactly k local maxima $r_1, r_2, \ldots, r_k$, such that there is a unique r_j satisfying $\|r_j - \mu_j\|_2 \leq \varepsilon\sigma_j\sqrt{d}$. This is by using a quantitative version of the inverse function mapping theorem with the function $h(x) = \nabla f(x) = 0$ (one can use the Newton method as in Section 4).

Again, in what follows, when it is clear that we have sufficiently accurate estimates, we will abuse notation and also use $f(x), f'(x), f''(x)$ to represent the estimate of the f, f', f'' at x.

[5] For our purposes, it suffices to have estimates of f, f', f'' at each of the points of the net $\mathcal{X}_\delta$. Hence we can just take histogram counts in a small ball around each of the net points, and estimate f up to accuracy γ with $|\mathcal{X}_\delta|/\delta^2$ samples. Similarly the derivatives can also estimated using $\text{poly}(|\mathcal{X}_\delta|, \frac{1}{\delta}, \frac{1}{\gamma})$ samples.

Lemma 6.8. *Assume the conditions in Theorem 6.2, and let* $\varepsilon_0 = \exp(-c_0 d)$*. Suppose we have* $\tilde{\mu}_j \in \mathbb{R}^d$ *such that* $\|\tilde{\mu}_j - \mu_j\|_2 \le \exp(-2c_0 d)\sigma_j$*. Then there is an algorithm running in* $\left(\frac{\rho}{\varepsilon_0^3 w_{min}}\right)^{O(d)}$ *time, that w.h.p. finds* $\tilde{\sigma}_j \in R_+$ *such that* $|\tilde{\sigma}_j - \sigma_j| \le \varepsilon_0 \sigma_j$*.*

Proof. Let κ be a fixed number chosen so that $\kappa \le \sigma_j$, and pick any point $y \in \mathbb{R}^d$ such that $\|y - \tilde{\mu}_j\|_2 = \frac{1}{\sqrt{\pi}}\kappa\sqrt{d}$.[6] Based on the estimates of the p.d.f. at $\tilde{\mu}_j, y$, we will set

$$\tilde{\sigma}_j = \frac{\kappa\sqrt{d}}{\sqrt{\log\left(\frac{f(\tilde{\mu}_j)}{f(y)}\right)}}.$$

Both $\tilde{\mu}_j, y$ are in a ball of radius at most $\sigma_j\sqrt{d}/\pi$ around μ_j. Further from Lemma 4.13, and since we have good estimate for the p.d.f. $f(\tilde{\mu}_j)$ and $f(y)$ at these points, we have $f(\tilde{\mu}_j) = w_j\sigma_j^{-d}(1 \pm \exp(-2c_0 d))$, and $f(y) = w_j\sigma_j^{-d}\exp\left(-\frac{\kappa^2 d}{\sigma_j^2}\right)(1 \pm \exp(-2c_0 d))$. Dividing,

$$\left|\log\left(\frac{f(\tilde{\mu}_j)}{f(y)}\right) - \frac{\kappa^2 d}{\sigma_j^2}\right| \le 2\exp(-2c_0 d).$$

Substituting for $\tilde{\sigma}_j$ we have

$$\begin{aligned}\sigma_j^2 &= \frac{\tilde{\sigma}_j^2 \log\left(\frac{f(\tilde{\mu}_j)}{f(y)}\right)}{\log\left(\frac{f(\tilde{\mu}_j)}{f(y)}\right) + \eta} \quad \text{where } |\eta| \le 2\exp(-2c_0 d)\\ &= \tilde{\sigma}_j^2\left(1 + \frac{\eta}{\log\left(\frac{f(\tilde{\mu}_j)}{f(y)}\right)}\right).\end{aligned}$$

Hence, $\left|\frac{\sigma_j^2}{\tilde{\sigma}_j^2} - 1\right| \le \varepsilon_0$, where the last inequality follows from our choice of y, since $\log(f(\tilde{\mu}_j)/f(y)) \le d < \exp(c_0 d)$. □

Lemma 6.9. *Assume the conditions in Theorem 6.2, and let* $\varepsilon_0 = \exp(-c_0 d)$*. Suppose we have* $\tilde{\mu}_j \in \mathbb{R}^d, \tilde{\sigma}_j \in \mathbb{R}^d$ *such that* $\|\tilde{\mu}_j - \mu_j\|_2 + |\tilde{\sigma}_j - \sigma_j| \le \exp(-2c_0 d)\sigma$*. Then there is an algorithm running in* $\left(\frac{\rho}{\varepsilon_0^3 w_{min}}\right)^{O(d)}$ *time, that w.h.p. finds* $\tilde{w}_j \in [0,1]$ *such that* $|\tilde{w}_j - w_j| \le \varepsilon_0 w_j$*.*

Proof. Let $\eta = w_j\exp(-2c_0 d)$. Let c_d be the constant that is only dependent on d given by

$$c_d = \int_{x\in\mathbb{R}^d : \|x\|_2 \le \frac{1}{\sqrt{2\pi}}\sqrt{d}} \exp\left(-\pi\|x\|_2^2\right) dx.$$

In fact, $c_d = \gamma(\frac{d}{2} - 1, \frac{1}{2})$ is the incomplete Gamma function evaluated at $(\frac{d}{2} - 1, \frac{1}{2})$ which has an asymptotic approximation given in [21, 8.11(ii)]. Also $c_d \ge 2^{-d/2}/d$. To get an estimate of $\tilde{w}_j$, we will consider the set

$$T_j = \{x \in \mathbb{R}^d \mid \|x - \tilde{\mu}_j\|_2 \le \tfrac{1}{\sqrt{2\pi}}\sqrt{d}\tilde{\sigma}_j\}.$$

[6] We can guess such a κ by either doing a binary search, or set it to be one-eighth of the diameter of cluster T_j^* defined in proof of Proposition 6.3

We will now generate $N = O(\rho \log(dk)/\eta^2)$ samples $x^{(1)}, \dots, x^{(N)}$ from the mixture of k Gaussians and estimate the fraction of samples that are in T_j:[7]

$$\tilde{w}_j = \frac{1}{c_d N} \sum_{\ell=1}^{N} \mathbb{I}[x^{(\ell)} \in T_j]. \tag{83}$$

From Lemma 4.15, we have small contribution from the other components

$$|\tilde{w}_j - \mathbb{E}[\tilde{w}_j]| = \left| \tilde{w}_j - \frac{1}{c_d} \int_{y \in T_j} f(y)\, dy \right| \le \eta \le \exp(-2cd) w_j.$$

From Lemma 4.13, we have

$$\begin{aligned} &\forall y \in T_j, \ \left| f(y) - w_j g_{\mu_j,\sigma_j}(y) \right| \le \exp(-2c_0 d) w_j g_{\mu_j,\sigma_j}(y) \\ \text{Hence, } &\left| \mathbb{E}[\tilde{w}_j] - \frac{w_j}{c_d} \int_{y \in T_j} g_{\mu_j,\sigma_j}(y)\, dy \right| \le \exp(-2c_0 d) \cdot \frac{w_j}{c_d} \int_{y \in T_j} g_{\mu_j,\sigma_j}(y)\, dy < \frac{w_j}{c_d} \exp(-2c_0 d). \end{aligned}$$

Let $B = \{ y \mid \|y - \mu_j\|_2 \le \frac{1}{2\pi} \sqrt{d} \sigma_j \}$, and let S_d be the surface area of the unit ball in d-dimensions (volume of the d-sphere). Since $\|\tilde{\mu}_j - \mu_j\| + |\sigma_j - \tilde{\sigma}_j| \le \exp(-2c_0 d)\sigma_j$, the probability mass on $T_j \setminus B$ is small

$$\begin{aligned} \left| \int_{y \in T_j} g_{\mu_j,\sigma_j}(y)\, dy - \int_{y \in B} g_{\mu_j,\sigma_j}(y)\, dy \right| &\le \int_{\left| \frac{\sqrt{2\pi}\|y-\mu_j\|_2}{\sigma_j} - 1 \right| \le \exp(-2c_0 d)} g_{\mu_j,\sigma_j}(y)\, dy. \\ &\le S_d \left(\frac{d}{2\pi} \right)^{d/2} \times 2\exp(-2c_0 d) \le \exp(-2(c_0 - 1)d). \end{aligned}$$

Further, the probability mass inside B is given by $\frac{w_j}{c_d} \int_{y \in B} g_{\mu_j,\sigma_j}(y)\, dy = w_j$. Hence,

$$\begin{aligned} |\mathbb{E}[\tilde{w}_j] - w_j| &\le \left| \mathbb{E}[\tilde{w}_j] - \frac{w_j}{c_d} \int_{y \in T_j} g_{\mu_j,\sigma_j}(y)\, dy \right| + \left| \frac{w_j}{c_d} \int_{y \in B} g_{\mu_j,\sigma_j}(y)\, dy - \frac{w_j}{c_d} \int_{y \in T_j} g_{\mu_j,\sigma_j}(y)\, dy \right| \\ &\le \frac{w_j}{c_d} \exp(-2(c_0 - 1)d) + \frac{w_j}{c_d} \exp(-2c_0 d) \le w_j \exp(-c_0 d). \end{aligned}$$

□

Proof of Theorem 6.2. The proof follows by using Proposition 6.3, followed by Lemma 6.8 and Lemma 6.9 in that order. Set $\varepsilon_0 = \exp(-c_0 d)$. First we use Proposition 6.3 to find w.h.p. initializers for the means $(\tilde{\mu}_j : j \in [k])$ such that $\|\tilde{\mu}_j - \mu_j\|_2 \le \exp(-4c_0 d)\sigma_{\min}$. Then using Lemma 6.8, we find w.h.p. initializers $(\tilde{\sigma}_j : j \in [k])$ such that $|\tilde{\sigma}_j - \sigma_j| \le \exp(-2c_0 d)\sigma_j$. Finally, these initializers $\tilde{\mu}_j, \tilde{\sigma}_j$ can be used in Lemma 6.9 to find w.h.p. $\tilde{w}_j \ \forall j \in [k]$ such that $|\tilde{w}_j - w_j| \le \exp(-c_0 d) w_j$. By choosing the failure probability of at most $1/3k^2$ in each step, we see that the algorithm succeeds w.h.p. and runs in time $\left(\frac{\rho}{\varepsilon_0^3 w_{\min}} \right)^{O(d)}$. □

Appendix

[7] We could also integrate the estimated p.d.f. over the set T_j to get this estimate.

A Standard Properties of Gaussians

Lemma A.1. *Suppose $x \in \mathbb{R}$ be generated according to $N(0, \sigma^2)$, let $\tilde{\Phi}_{0,\sigma}(t)$ represents the probability that $x > t$, and let $\tilde{\Phi}^{-1}_{0,\sigma}(y)$ represent the quantile t at which $\tilde{\Phi}_{0,\sigma}(t) \leq y$. Then*

$$\frac{\frac{t}{\sigma}}{(\frac{t^2}{\sigma^2}+1)} e^{-\frac{t^2}{2\sigma^2}} \leq \tilde{\Phi}_{0,\sigma}(t) \leq \frac{\sigma}{t} e^{-\frac{t^2}{2\sigma^2}}. \tag{84}$$

Further, there exists a universal constant $c \in (1, 4)$ such that

$$\frac{1}{c}\sqrt{\log(1/y)} \leq \frac{t}{\sigma} \leq c\sqrt{\log(1/y)}. \tag{85}$$

Lemma A.2. *For any $\sigma > 0$, $q \in \mathbb{Z}_+$ and any $\tau \geq 2q$,*

$$\int_{|x| \geq \tau\sigma} |x|^q \exp(-\pi x^2/\sigma^2) dx \leq \sigma^q \exp(-2\tau^2). \tag{86}$$

Lemma A.3. *Let p, q correspond to the (weighted) probability density functions of the spherical Gaussian components in d dimensions with parameters (w_1, μ_1, σ_1^2) and (w_2, μ_2, σ_2^2) respectively. Then*

$$\|p - q\|_1 \leq |w_1 - w_2| + \min\{w_1, w_2\} \left(\frac{\sqrt{2\pi} \cdot \|\mu_1 - \mu_2\|_2}{\sigma_2} + \frac{\sqrt{d|\sigma_1^2 - \sigma_2^2|}}{\sigma_2} + \sqrt{2d \ln\Big(\frac{\sigma_2}{\sigma_1}\Big)} \right). \tag{87}$$

Proof. Without loss of generality let $w_2 \leq w_1$. The KL divergence between any two multivariate Gaussian distributions with means μ_1, μ_2 and covariances Σ_1, Σ_2 respectively is given by [22]

$$d_{KL}\left(N(\mu_1, \Sigma_1) \| N(\mu_2, \Sigma_2)\right) = \frac{1}{2}\left(\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_1 - \mu_2)^T \Sigma_2^{-1}(\mu_1 - \mu_2) - d + \ln\Big(\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\Big) \right).$$

Applying this to $p' := p/w_1, q' := q/w_2$ we get

$$\begin{aligned} d_{KL}(p' \| q') &= \frac{1}{2}\left(\frac{2\pi\|\mu_1 - \mu_2\|_2^2}{\sigma_2^2} + \frac{d\sigma_1^2}{\sigma_2^2} - d + 2d\ln\Big(\frac{\sigma_2}{\sigma_1}\Big) \right) \\ &= \frac{\pi\|\mu_1 - \mu_2\|_2^2}{\sigma_2^2} + \frac{d(\sigma_1^2 - \sigma_2^2)}{2\sigma_2^2} + d\ln\Big(\frac{\sigma_2}{\sigma_1}\Big). \end{aligned}$$

Hence, by Pinsker inequality

$$\|p' - q'\|_1 \leq \sqrt{2d_{KL}(p'\|q')} \leq \frac{\sqrt{2\pi}\|\mu_1 - \mu_2\|_2}{\sigma_2} + \frac{\sqrt{d|\sigma_1^2 - \sigma_2^2|}}{\sigma_2} + \sqrt{2d\log\Big(\frac{\sigma_2}{\sigma_1}\Big)}.$$

By triangle inequality,

$$\|p - q\|_1 \leq \|p - w_2 p'\|_1 + \|w_2 p' - w_2 q'\|_1 + \|w_2 q' - q\|_1 \leq |w_1 - w_2| + w_2\|p' - q'\|_1,$$

which gives the required bound. An identical proof works when $w_1 \leq w_2$. □

Higher Dimensional Gaussians and Approximations. Let γ_d be the Gaussian measure associated with a standard Gaussian with mean 0 and variance 1 in each direction.

Using concentration bounds for the χ^2 random variables, we have the following bounds for the lengths of vectors picked according to a standard Gaussian in d dimensions (see (4.3) in [31]).

Lemma A.4. *For a standard Gaussian in d dimensions (mean 0 and variance $1/(2\pi)$ in each direction), and any $t > 0$*

$$\mathop{\mathbb{P}}_{x\sim\gamma_d}\left[\|x\|^2 \geq \frac{1}{2\pi}(d + 2\sqrt{dt} + 2t)\right] \leq e^{-t}.$$

$$\mathop{\mathbb{P}}_{x\sim\gamma_d}\left[\|x\|^2 \leq \frac{1}{2\pi}(d - 2\sqrt{dt})\right] \leq e^{-t}.$$

Similarly, the following lemma shows a simple bound for the truncated moments, when x is generated according to $N(0, \sigma^2/2\pi)^d$.

Lemma A.5. *For any $\tau \geq q$, and any $q \in \mathbb{Z}_+$*

$$\int_{\|x\|_2 \geq 2q\sqrt{d}\sigma} \|x\|_2^q \exp(-\pi\|x\|_2^2/\sigma^2)dx \leq \sigma^q \exp(-4d). \tag{88}$$

Proof. Assume w.l.o.g that $\sigma = 1$. For $\|x\|_2 \geq 2q\sqrt{d}/\sqrt{2\pi}$, $\|x\|_2^q \leq \exp(\pi\|x\|_2^2/2)$. Hence,

$$\int_{\|x\|_2 \geq 2q\sqrt{d}} \|x\|_2^q \exp(-\pi\|x\|_2^2)dx \leq \int_{\|x\|_2 \geq 2q\sqrt{d}} \exp(-\pi\|x\|^2/2)dx$$
$$\leq 2^{d/2} \int_{\|y\|_2 \geq \frac{2q\sqrt{d}}{\sqrt{2}}} \exp\Big(-\pi\|y\|_2^2/2\Big)d\ y \leq \exp(-5d + d/2) \leq \frac{1}{16\pi d^2},$$

where y is distributed as a normal d-dimensional r.v. with mean 0 and variance 1 in each direction. □

Fact A.6 (Stirling Approximation). *For any $n \geq 1$, $\sqrt{2\pi n}(n/e)^n \leq n! \leq e\sqrt{n}(n/e)^n$.*

B Newton's method for solving non-linear equations

We use a standard theorem that shows quadratic convergence of the Newton method in any normed space [5] in the restricted setting where both the range and domain of f is $\mathbb{R}^m$.

Consider a system of m non-linear equations in variables $u_1, u_2, \ldots, u_m$:

$$\forall j \in [m], f_j(u_1, \ldots, u_m) = b_j.$$

Newton's method starts with an initial point $u^{(0)}$ close to a solution u^* of the non-linear system. Formally, $u^{(0)} \in \mathcal{N}$ where $\mathcal{N}$ is an appropriately defined neighborhood set $\mathcal{N} = \{y : \|y - u^*\| \leq \varepsilon_0\}$. Let $F'(u) = J_f(u) \in \mathbb{R}^{m\times m}$ be the Jacobian of the system given by the non-linear functional $f : \mathbb{R}^m \to \mathbb{R}^m$, where the $(j,i)^{th}$ entry is the partial derivate $J_f(j,i) = \frac{\partial f_j(u)}{\partial u_i}|_y$ is evaluated at y. Additionally, for our algorithm, we assume that given any $y \in \mathcal{N}$, we only have access to an estimates $\tilde{b}, \tilde{F}(u), \tilde{F}'(u)$ of vector $b \in \mathbb{R}^m$, $F(u) \in \mathbb{R}^m$ and $F'(u) \in \mathbb{R}^{m\times m}$ respectively [8].

[8] These errors in the estimate may occur due to sampling errors or precision errors.

Newton's method starts with the initializer $u^{(0)}$, and updates the solution using the iteration:

$$u^{(t+1)} = u^{(t)} + \left(\tilde{F}'(u^{(t)})\right)^{-1} \left(\tilde{b} - f(u^{(t)})\right). \tag{89}$$

The convergence error will be measured in the ℓ_p norm for any $p \geq 1$. In what follows, $\|x\| := \|x\|_p$ for $x \in \mathbb{R}^m$, $\|M\| := \|M\|_{p \to p}$ for $M \in \mathbb{R}^{m \times m}$ and $\|T\| := \|T\|_{p \times p \to p}$ for $T \in \mathbb{R}^{m \times m \times m}$. We first state a simple mean-value theorem that will be useful in the analysis of the Newton method, as well as in applying the guarantees in the context of mixtures of Gaussians.

Lemma B.1 (Proposition 5.3.11 in [5]). *Consider a function $H : K \subset \mathbb{R}^{m_1} \to \mathbb{R}^{m_2}$, with K being an open set. Assume H is differentiable on K and that $F'(u)$ is a continuous function of u on K. Assume $u, w \in K$ and assume line segment from joining them is also contained in K. Then*

$$\|F(u) - F(w)\| \leq \sup_{0 \leq \theta \leq 1} \|F'((1-\theta)u + \theta w)\| \cdot \|u - w\|.$$

The following theorem gives robustness guarantees for Newton's method. It is obtained by using matrix perturbation analysis along with a standard theorem regarding the quadratic convergence of the Newton's method (see Theorem 5.4.1 in [5]).

Theorem B.2. *Assume $u^* \in \mathbb{R}^m$ is a solution to the equation $F(u) = b$ where $F : \mathbb{R}^m \to \mathbb{R}^m$ such that $J^{-1} = (F')^{-1}$ exists in a neighborhood $\mathcal{N} = \{y : \|y - u^*\| \leq \|u^{(0)} - u^*\|\}$, and $F' : \mathbb{R}^m \to \mathbb{R}^{m \times m}$ is locally L-Lipschitz continuous in the neighborhood $\mathcal{N}$ i.e.,*

$$\|F'(u) - F'(v)\| \leq L\|u - v\| \quad \forall u, v \in \mathcal{N}.$$

Further, let the estimates $\tilde{b}, \tilde{F}(u), \tilde{F}'(u)$ satisfy for some $\eta_1, \eta_2, \eta_3 > 0$ and all $y \in \mathcal{N}$

$$\|\tilde{b} - b\| \leq \eta_1, \|\tilde{F}(u) - F(u)\| \leq \eta_2, \|\tilde{F}'(u) - F'(u)\| \leq \eta_3.$$

Then if $\eta_3 \|F'(u^{(t)})^{-1}\| < 1$, $\|F'(u)\| \leq B$, then for all $u \in \mathcal{N}$, the error $\varepsilon_t = \|u^{(t)} - u^*\|$ after the t iterations of (89) satisfies

$$\varepsilon_{t+1} \leq \varepsilon_t^2 \cdot L\|F'(u^{(t)})^{-1}\| + \|F'(u^{(t)})^{-1}\| \left(\eta_1 + \eta_2 + 4\eta_3 \varepsilon_t \|F'(u^{(t)})^{-1}\| B\right) \tag{90}$$

Proof. We sketch the proof here. For convenience, let us use $A := F'(u^{(t)})$ and $\tilde{A} := \tilde{F}'(u^{(t)})$ to denote the derivate and its estimate at $u^{(t)}$. Let $z^{(t+1)}$ be the Newton iterate after the $(t+1)$th step if we had the exact values of $F(u^{(t)}$ and $F'(u^{(t)})$ i.e., $z^{(t+1)} = u^{(t)} + A^{-1}\left(b - F(u^{(t)})\right)$. From the standard analysis of the Newton method (see Theorem 5.4.1 in [5]),

$$\|z^{(t+1)} - u^*\| \leq L\|A^{-1}\| \|u^{(t)} - u^*\|^2.$$

Further, the error between the actual Newton update $u^{(t+1)}$ and $z^{(t)}$ due to the estimates $\tilde{A}$ and $\tilde{F}(u^{(t)})$ is

$$\begin{aligned} u^{(t+1)} - z^{(t+1)} &= \tilde{A}^{-1}\left(\tilde{b} - \tilde{F}(u^{(t)})\right) - A^{-1}\left(b - F(u^{(t)})\right) \\ &= (\tilde{A}^{-1} - A^{-1})\left(\tilde{b} - \tilde{F}(u^{(t)})\right) + A^{-1}\left(\tilde{b} - b + \tilde{F}(u^{(t)}) - F(u^{(t)})\right) \\ \|u^{(t+1)} - z^{(t+1)}\| &\leq \|\tilde{A}^{-1} - A^{-1}\| \|\tilde{b} - \tilde{F}(u^{(t)})\| + \|A^{-1}\|(\eta_1 + \eta_2). \end{aligned}$$

From perturbation bounds on matrix inverses [13], if $\|A^{-1}E\| < 1$,

$$\|A^{-1} - (A+E)^{-1}\| \leq \|A^{-1}\| \cdot \frac{\|A^{-1}E\|}{1 - \|A^{-1}E\|}.$$

Also from Lemma B.1, $\|b - F(u^{(t)})\| \leq \|F'(u')\| \|u^{(t)} - u^*\|$ for some $u' \in \mathcal{N}$. Substituting $E = \tilde{A} - A$,

$$\begin{aligned}\|u^{(t+1)} - z^{(t+1)}\| &\leq \|A^{-1}\| \frac{\|A^{-1}(\tilde{A} - A)\|}{1 - \|A^{-1}(\tilde{A} - A)\|} \left(\eta_1 + \eta_2 + \|b - F(u^{(t)})\|\right) + \|A^{-1}\|(\eta_1 + \eta_2) \\ &\leq 2\eta_3 \|A^{-1}\|^2 \left(\eta_1 + \eta_2 + \|F'(u')\| \|u^{(t)} - u^*\|\right) + \|A^{-1}\|(\eta_1 + \eta_2) \\ &\leq 4\eta_3 \varepsilon_t \|A^{-1}\|^2 \|F'(u')\| + \|A^{-1}\|(\eta_1 + \eta_2).\end{aligned}$$

□

Remark B.3. While the above theorem requires that the derivative F' is locally L-Lipschitz, this is a weaker condition than requiring a upper bound on the operator norm of the second derivative F''. Lemma B.1 shows that it also suffices if $\|F''(u)\| \leq L$ for all $u \in \mathcal{N}$.

Corollary B.4. *Under the conditions of Theorem B.2, there exists* $0 < \varepsilon_0 < \frac{1}{2L\|F'(u^{(t)})^{-1}\|}$*, such that for any given* $\delta \in (0,1)$*, there is an* $\eta_1, \eta_2, \eta_3 > 0$ *with*

$$(\eta_1 + \eta_2) < \frac{\delta}{4\|F'(u^{(t)})^{-1}\|} \text{ and } \eta_3 < \frac{\delta}{4\|F'(u^{(t)})^{-1}\|^2} \cdot \min\{1, \frac{1}{B}\}$$

such that after $T = \log\log(1/\delta)$ *iterations of the Newton's method, we have*

$$\|u^{(T)} - u^*\| \leq \delta.$$

Proof. For the given setting of η_1, η_2, η_3, we have $(\eta_1 + \eta_2)\|F'(u^{(t)})^{-1}\| < \frac{\delta}{4}$, and $B\|F'(u^{(t)})^{-1}\|^2 \varepsilon_t \eta_3 \leq \delta/4$. From Theorem B.2, we have that for any t,

$$\varepsilon_{t+1} \leq \varepsilon_t^2 L \|F'(u^{(t)})^{-1}\| + \frac{\delta}{2}.$$

Further $\varepsilon_1 \leq \varepsilon_0^2 L \|F'(u^{(t)})^{-1}\| < \frac{1}{4L\|F'(u^{(t)})^{-1}\|}$. By induction, it follows that

$$\varepsilon_t \leq \frac{2^{-2^t}}{L\|F'(u^{(t)})^{-1}\|}.$$

Hence, this gives the required guarantee. □

C Dimension Reduction using PCA

Here we give a proof of the assertion that for mixtures of spherical Gaussians, we can assume without loss of generality that $d \leq k$.

Theorem C.1 (Same as Theorem 4.2)**.** *Let $\{(w_i, \mu_i, \sigma_i) : i \in [k]\}$ be a mixture of k spherical Gaussians that is ρ bounded, and let w_{min} be the smallest mixing weight. Let $\mu'_1, \mu'_2, \dots, \mu'_k$ be the projections onto the subspace spanned by the top k singular vectors of sample matrix $X \in \mathbb{R}^{d\times N}$. For any $\varepsilon > 0$, with $N = \mathrm{poly}(d, \rho, w_{min}^{-1}, \varepsilon^{-1})$ samples we have with high probability*

$$\forall i \in [k],\ \|\mu_i - \mu'_i\|_2 \le \varepsilon.$$

Proof. Let $\delta = w_{\min}\varepsilon^4/(2\rho^2)$ and $\eta = \delta^2/2$. Let A be the population average, i.e.,

$$A = \mathbb{E}[xx^T] = M + \bar{\sigma}^2 I, \text{ where } M = \sum_i w_i \mu_i \mu_i^T, \text{ and } \bar{\sigma}^2 = \frac{1}{2\pi}\sum_{i\in[k]} w_i\sigma_i^2.$$

Let $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_d \ge 0$ be the eigenvalues of M sorted in non-increasing order. Since M is of rank at most k, we have $\lambda_{k+1} = \cdots = \lambda_d = 0$. Let $r \le k$ be defined as the smallest index such that $\lambda_{r+1} < \delta$. Let U be represent the orthogonal projector onto the top-r eigenspace of M (and hence of A too), and $U^\perp = I - U$. Notice that $\|U^\perp M U^\perp\| = \lambda_{r+1} < \delta$. Then, using the positive semidefinite inequality $w_i U^\perp \mu_i \mu_i^T U^\perp \preceq U^\perp M U^\perp$, we obtain that

$$\|U^\perp \mu_i\|_2^2 \le w_i^{-1}\|U^\perp M U^\perp\| = \lambda_{r+1}/w_i < \delta/w_{\min}\ .$$

Next, let $\tilde{A}$ represent the sample average $\tilde{A} = \frac{1}{N}XX^T$ (so $\tilde{A}$ converges to A as $N \to \infty$). Since $N \ge \mathrm{poly}(d)/\eta^2$, using standard concentration bounds (see Theorem 6.1.1 in [43] and related notes) we have with high probability that $\|A - \tilde{A}\| < \eta$. Let $\tilde{U}$ be the orthogonal projector onto the top-k eigenspace of $\tilde{A}$, and let $\tilde{V} = I - \tilde{U}$ be the orthogonal projector onto the bottom $(d-k)$ eigenspace of $\tilde{A}$. Hence for each $i \in [k]$, $\mu'_i = \tilde{U}\mu_i$. Notice that from Weyl's perturbation bounds for eigenvalues (see [13], Theorem III.2.1) $\lambda_{k+1}(\tilde{A}) \le \bar{\sigma}^2 + \eta$. Therefore, the eigenvalues of $\tilde{A}$ corresponding to $\tilde{V}$ (which are all at most $\bar{\sigma}^2 + \eta$), and the eigenvalues of A corresponding to U (which are all at least $\bar{\sigma}^2 + \delta$) are separated by at least $\delta - \eta$. From standard perturbation bounds for eigenvectors (see [13], Theorem VII.3.1), we have

$$\|U\tilde{V}\| \le \frac{\|A - \tilde{A}\|}{\delta - \eta} \le \frac{\eta}{\delta - \eta} \le \delta.$$

Hence, for each $i \in [k]$,

$$\begin{aligned}\|\mu_i - \mu'_i\|_2^2 &= \|\tilde{V}\mu_i\|_2^2 = \langle \mu_i, \tilde{V}\mu_i\rangle \\ &= \langle U\mu_i, \tilde{V}\mu_i\rangle + \langle U^\perp\mu_i, \tilde{V}\mu_i\rangle = \langle \mu_i, U\tilde{V}\mu_i\rangle + \langle U^\perp\mu_i, \tilde{V}\mu_i\rangle \\ &\le \|U\tilde{V}\|\|\mu_i\|_2^2 + \|U^\perp\mu_i\|_2\|\mu_i\|_2 \\ &\le \delta\|\mu_i\|_2^2 + \sqrt{\frac{\delta}{w_{\min}}}\cdot\|\mu_i\|_2 \le \varepsilon^2\end{aligned}$$

by our choice of $\delta = w_{\min}\varepsilon^4/(2\rho^2)$. □

Acknowledgements

The authors thank Santosh Vempala for suggesting the problem of learning one-dimensional Gaussians, and other helpful discussions. Part of this work was done when the second author was at the Courant Institute and the Simons Collaboration on Algorithms and Geometry.

References

[1] Emmanuel Abbe. Community detection and the stochastic block model. 2016. Available from `http://www.ee.princeton.edu/research/eabbe`.

[2] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *Learning Theory*, pages 458–469. Springer, 2005.

[3] Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James R. Voss. The more, the merrier: the blessing of dimensionality for learning large Gaussian mixtures. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pages 1135–1164, 2014.

[4] Sanjeev Arora and Ravi Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. ACM, 2001.

[5] Kendall Atkinson and Weimin Han. *Theoretical numerical analysis : a functional analysis framework*. Texts in applied mathematics. Springer, New York, Berlin, Paris, etc., 2001.

[6] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Center-based clustering under perturbation stability. *Information Processing Letters*, 112(12):49 – 54, 2012.

[7] Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 37–49. Springer, 2012.

[8] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *CoRR*, abs/1408.2156, 2014.

[9] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 103–112. IEEE, 2010.

[10] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the 46th Symposium on Theory of Computing (STOC)*. ACM, 2014.

[11] Aditya Bhaskara, Moses Charikar, and Aravindan Vijayaraghavan. Uniqueness of tensor decompositions with applications to polynomial identifiability. *Proceedings of the Conference on Learning Theory (COLT).*, 2014.

[12] Aditya Bhaskara, Ananda Suresh, and Morteza Zadimoghaddam. Sparse Solutions to Nonnegative Linear Systems and Applications. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 83–92, San Diego, California, USA, 09–12 May 2015. PMLR.

[13] Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer, 1997.

[14] Spencer Charles Brubaker and Santosh Vempala. Isotropic pca and affine-invariant clustering. In *Proceedings of the 2008 49th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '08, pages 551–560, Washington, DC, USA, 2008. IEEE Computer Society.

[15] 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*, STOC '14, pages 604–613, New York, NY, USA, 2014. ACM.

[16] Sanjoy Dasgupta. Learning mixtures of Gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644. IEEE, 1999.

[17] Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *The Journal of Machine Learning Research*, 8:203–226, 2007.

[18] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians. In Maria-Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 1183–1213. JMLR.org, 2014.

[19] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of EM suffice for mixtures of two Gaussians. *CoRR*, abs/1609.00368, 2016.

[20] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures. *CoRR*, abs/1611.03473, 2016.

[21] *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.0.13 of 2016-09-16. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds.

[22] John Duchi. Derivations for linear algebra and optimization. `web.stanford.edu/~jduchi/projects/general_notes.pdf`. [Online].

[23] 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*, COLT'06, pages 20–34, Berlin, Heidelberg, 2006. Springer-Verlag.

[24] Rong Ge, Qingqing Huang, and Sham M. Kakade. Learning mixtures of Gaussians in high dimensions. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 761–770, 2015.

[25] Navin Goyal, Santosh Vempala, and Ying Xiao. Fourier PCA and robust tensor decomposition. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 584–593, 2014.

[26] Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two Gaussians. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 753–760, 2015.

[27] 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. ACM, 2013.

[28] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two Gaussians. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 553–562. ACM, 2010.

[29] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. *SIAM J. Comput.*, 38(3):1141–1156, 2008.

[30] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 299–308. IEEE, 2010.

[31] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 10 2000.

[32] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 694–703, New York, NY, USA, 2014. ACM.

[33] F. McSherry. Spectral partitioning of random graphs. In *Proceedings of the 42nd IEEE symposium on Foundations of Computer Science*, FOCS '01, pages 529–, Washington, DC, USA, 2001. IEEE Computer Society.

[34] D. G. Mixon, S. Villar, and R. Ward. Clustering subgaussian mixtures with k-means. In *2016 IEEE Information Theory Workshop (ITW)*, pages 211–215, Sept 2016.

[35] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

[36] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *CoRR*, abs/1311.4115, 2013.

[37] Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propagation, robust reconstruction and optimal recovery of block models. In *Proceedings of the Conference on Learning Theory*, pages 356–370, 2014.

[38] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[39] Nathan Srebro, Gregory Shakhnarovich, and Sam Roweis. An investigation of computational and informational limits in Gaussian mixture clustering. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 865–872, New York, NY, USA, 2006. ACM.

[40] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical Gaussian mixtures. In Z. Ghahramani, M. Welling,

C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1395–1403. Curran Associates, Inc., 2014.

[41] Henry Teicher. Identifiability of mixtures. *The annals of Mathematical statistics*, 32(1):244–248, 1961.

[42] Henry Teicher. Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics*, 38(4):1300–1302, 1967.

[43] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, Aug 2012.

[44] J.M. Varah. A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications*, 11(1):3 – 5, 1975.

[45] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

[46] Ji Xu, Daniel J. Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two Gaussians. In *NIPS*, 2016.