# On Learning Mixtures of Well-Separated Gaussians

Oded Regev
*Courant Institute of Mathematical Sciences*
*New York University*
*New York, USA*
*regev@cims.nyu.edu*

Aravindan Vijayaraghavan
*Department of EECS*
*Northwestern University*
*Evanston, USA*
*aravindv@northwestern.edu*

***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 $\mu_1, \dots, \mu_k \in \mathbb{R}^d$, and the goal is to estimate the means up to accuracy $\delta$ 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 $\delta$. 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 $\delta$ assuming the separation between the means is $\Omega(\min\{\sqrt{\log k}, \sqrt{d}\})$ (independently of $\delta$). 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.**

***Keywords*-mixtures of Gaussians; unsupervised learning; clustering; parameter estimation; sample complexity; iterative algorithms**

## I. 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 one of $k$ Gaussian components according to mixing weights $w_1, w_2, \dots, w_k$, 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 $j$th 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 [22]. (See Section I-B 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 [11], and continuing with a long line of work (including [3, 27, 1, 19, 23, 12, 10, 20, 5, 6, 28, 13]), 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 [27] provides an efficient algorithm (both in terms of samples and running time) under separation of at least $\min\{k, d\}^{1/4}$ (up to polylog factors) 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., [21, 2, 16]), but the gap between this lower bound and the above upper bound of roughly $\min\{k, d\}^{1/4}$ is quite

0272-5428/17 $31.00 © 2017 IEEE
DOI 10.1109/FOCS.2017.17

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 I.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 $\delta$ using $\text{poly}(d, k, 1/\delta)$ samples?

### *A. 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 II.2 for the precise definition).

**Informal Theorem I.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*

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

*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 III.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 [8, 2] 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 V.1 for the formal statement.)

**Informal Theorem I.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.,*

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

*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 dimension a separation of order $\Omega(1)$ suffices.

**Informal Theorem I.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.,*

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

*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 V.3 for a formal statement. An important feature of the above two algorithmic results is that the separation is independent of the accuracy $\delta$ that we desire in parameter estimation ($\delta$ can be arbitrarily small compared to $k$ and $d$). These results together essentially 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 I.3 and Theorem I.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 $\delta$. We now present an informal statement of the guarantees of the iterative algorithm.

**Informal Theorem I.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.*

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

*and suppose we are given initializers $\tilde{\mu}_1, \ldots, \tilde{\mu}_k$ for the means $\mu_1, \ldots, \mu_k$ 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 $\mu_1^{(T)}, \ldots, \mu_k^{(T)}$ such that $\Delta_{\text{param}}(\{\mu_1, \ldots, \mu_k\}, \{\mu_1^{(T)}, \ldots, \mu_k^{(T)}\}) \leq \delta$.*

The above theorem also holds when the weights and variances are unequal. See Theorem IV.1 for a formal statement. Note that in the above result, the desired accuracy $\delta$ can be arbitrarily small compared to $k$, and the separation required does not depend on $\delta$. To prove the polynomial identifiability results (Theorems I.3 and I.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 I.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 $\delta$ 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.

### *B. Prior Work and Comparison of Results*

Gaussian mixture models are among the most widely used probabilistic models in statistical inference [22, 24, 25]. 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 [11], 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 [3, 27, 1, 19, 12, 10, 20] for both spherical Gaussians and general Gaussians. The work of Vempala and Wang [27] uses PCA along with distance-based clustering to learn mixtures of spherical Gaussians when the separation $\|\mu_i - \mu_j\|_2$ is at least

$$(\min\{k, d\}^{1/4} \log^{1/4}(dk/\delta) + \log^{1/2}(dk/\delta))(\sigma_i + \sigma_j).$$

For non-spherical Gaussians, the result of [10] assumes a similar separation condition, but involving just the variance along the direction of the line joining the respective means, as opposed to $(\|\Sigma_i\| + \|\Sigma_j\|)$. We also note that all these clustering-based algorithms required a separation that either implicitly or explicitly depend on the estimation accuracy $\delta$.[1]

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. Dasgupta and Schulman [12] proved that a variant of the EM algorithm learns mixtures of Gaussians with separation of the order of $d^{1/4}\text{polylog}(dk)$. Kumar and Kannan [20] showed that spectral clustering (PCA followed by $k$-means) recovers the clusters under deterministic conditions about the data, that specializes to a separation of order $\sqrt{k}$ for mixtures of spherical Gaussians [5]. Very recently, the EM algorithm was shown to succeed for mixtures of $k = 2$ spherical Gaussians with $\Omega(\sigma)$ separation [6, 28, 13] (we note that in this setting with $k = O(1)$, polynomial time guarantees are also known using other algorithms like the method-of-moments [18], as we will see in the next paragraph).

*Moment-based methods:* In a series of influential results, algorithms based on the method-of-moments were developed by [18, 21, 7] for efficiently learning mixtures of $k = O(1)$ Gaussians under arbitrarily small separation. To perform parameter estimation up to accuracy $\delta$, the running time of the algorithms is $\text{poly}(d, 1/w_{\text{min}}, 1/\delta)^{O(k^2)}$ (this holds for mixtures of

[1] Such a dependency on $\delta$ seems necessary for such clustering-based algorithm that clusters every point accurately with high probability.

general Gaussians). This exponential dependence on $k$ is necessary in general, due to statistical lower bound results [21].

Recent work [17, 9, 15, 8, 2, 14] 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 [17] 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 [15, 8, 2] to give smoothed analysis guarantees to learn "most" mixtures of spherical Gaussians when the means are in $d = k^{\Omega(1)}$ dimensions. These algorithms do not assume any strict geometric separation conditions and learn the parameters in $\mathrm{poly}(k, d, 1/\delta)$ 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.

*Lower Bounds:* Moitra and Valiant [21] showed that $\exp(k)$ samples are needed to learn the parameters of a mixture of $k$ Gaussians [21]. In fact, the lower bound instance of [21] is one dimensional, with separation of order $1/\sqrt{k}$. Anderson et al. [2] proved a lower bound on sample complexity that is reminiscent of our Theorem I.2. Specifically, they obtain a super-polynomial lower bound assuming separation $O(\sigma/\mathrm{poly}\log(k))$ for $d = O(\log k/\log\log k)$. This is in contrast to our lower bound which allows separation greater than $\sigma$, or $o(\sigma\sqrt{\log k})$ to be precise.

### C. Overview of Techniques

*Iterative Algorithm:* Our iterative algorithm will function in both the 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 $\delta$ is arbitrarily small compared to $(kd)^{-\omega(1)}$ (for instance, think of $k, d$ as small). 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 IV.2 for details). For each $j \in [k]$ we consider only the samples in the set $S_j$ and be $u_j \in \mathbb{R}^d$ 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 $j$th component, and the total probability mass from the other components is relatively small (it will be at most $1/\mathrm{poly}(k)$ with $\Omega(\sqrt{\log k})$ separation, and $O_d(1)$ with $\Omega(1)$ separation in constant dimensions). However, since $\delta$ 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 $\delta$-close estimate for $\mu_j$ (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. $\mathbf{x}_i$ is $\nabla_{\mathbf{x}_i} F_j(\mathbf{x})$ which equals

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

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 $b, F, F'$.

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 $\ell_2$ 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 $\delta$-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 $j$th 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. 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 gives rapid convergence to the true parameters.

*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 III.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 $\mu_1, \dots, \mu_k$, the following quantities that we call "mean moments," and we will use them as a proxy for the actual moments of the distribution:

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

The mean moments just correspond to the usual moments of a mixture of delta functions centered at $\mu_1, \dots, \mu_k$. Closeness in the first $R = O(1/\varepsilon)$ mean moments (measured in injective tensor norm) implies that the two corresponding distributions are $\varepsilon$ close in statistical distance (see Lemma III.7 and Lemma III.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 III.6).

## II. 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 $j$th component has mean $\mu_j$ 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 $\mu$ 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 $j$th Gaussian component.

**Definition II.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 II.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

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

For *standard* mixtures, the definition simplifies to $\Delta_{\text{param}}((\mu_1,\ldots,\mu_k),(\mu'_1,\ldots,\mu'_k)) = \min_{\pi\in\text{Perm}_k}\sum_{j=1}^{k}\|\mu_j-\mu'_{\pi(j)}\|_2$.

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 [21].

**Definition II.3** ($\rho$-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 *$\rho$-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 $\rho$-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 $\sigma_{\min} = \min_{j\in[k]}\sigma_j$. 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 $\rho_w = \max_{i\in[k]} w_i/\min_{i\in[k]} w_i$.

In the above notation the bound $\rho$ 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 $\rho$ as the (multiplicative) range of values taken by the parameters $\{\mu_i,\sigma_i : i\in[k]\}$.

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 $\sigma^2$. 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 $r$th 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 will denote the matrix operator norms by:

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

### A. Notation and Preliminaries about Newton's method

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

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

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 [4]. 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 II.4** (Theorem 5.4.1 in [4])**.** *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 the full version of the paper.

We want to learn each of the $k$ sets of parameters up to good accuracy; hence we will measure the error in $\ell_\infty$ 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 $\ell_1$ norm of the rows of $A$. The following lemma bound $\|A^{-1}\|_{\infty\to\infty}$ for a diagonally dominant matrix $A$.

**Lemma II.5** ([26])**.** *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,* $\|A^{-1}\|_{\infty\to\infty}\leq 1/\alpha$.

Finally, we will use some standard facts about high-dimensonsal Gaussians. Using concentration bounds for the $\chi^2$ random variables, we have the following bounds for the lengths of vectors picked according to a standard Gaussian in $d$ dimensions.

**Lemma II.6.** *For a standard Gaussian in $d$ dimensions (mean $0$ and variance $1/(2\pi)$ in each direction), and any $t > 0$*

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

## III. 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]

**Theorem III.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 III.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.

### A. Proof of Theorem III.1

The key to the proof of Theorem III.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 III.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*

$$|\mathcal{F}| > \frac{1}{\eta}\exp\left(c_0\Big(\frac{\log(1/\varepsilon)}{d}\Big)^d \cdot \log(1/\varepsilon)\log(3d)\right), \tag{6}$$

*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, $c_0 = 8\pi e$ and $c_1 = 36$ suffice for our purposes.*

*Remark* III.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.

We start with a simple claim (proof in full version).

**Claim III.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 of Theorem III.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 III.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 $c, \gamma$.

To complete the proof, we notice that by our choice

[2] In particular, this rules out polynomial-time smoothed analysis guarantees of the kind shown for $d = k^{\Omega(1)}$ in [8, 2].

of parameters, and denoting $\varepsilon = k^{-C}$,

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

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

$$\Big(\frac{\log(1/\varepsilon)}{d}\Big)^d = k^{c\log(C/c)} < \sqrt{k}, \quad \text{and}$$
$$c_0 \log(1/\varepsilon)\log(3d) \leq c_0 C \log k \log(3c\log k) < \sqrt{k}.$$

Hence applying Theorem III.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 $\varepsilon$ 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. ■

### *B. Proof of Theorem III.2*

It will be convenient to represent the p.d.f. $f(x)$ of the standard mixture of spherical Gaussians with means $\mu_1, \mu_2, \ldots, \mu_k$ as a convolution of a standard mean zero Gaussian with a sum of delta functions centered at $\mu_1, \mu_2, \ldots, \mu_k$, $f(x) = \Big(\frac{1}{k}\sum_{j=1}^k \delta(x-\mu_j)\Big) * e^{-\pi\|x\|_2^2}$.

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.

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

To prove Theorem III.2 we will use three main steps. Lemma III.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. Lemma III.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 III.8.

We will use the following standard packing claim, whose proof we defer to the full version.

**Claim III.5.** *Let $\|\cdot\|$ be an arbitrary norm on $\mathbb{R}^D$. If $x_1, \ldots, x_N \in \mathbb{R}^D$ are such that $\|x_i\| \leq \Delta$ for all $i$, and for all $i \neq j$, $\|x_i - x_j\| > \delta$, then $N \leq (1+2\Delta/\delta)^D$. In particular, if $x_1, \ldots, x_N \in \mathbb{R}^D$ are such that $\|x_i\| \leq \Delta$ 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 $\|x_i - x_j\| \leq \delta$.*

**Lemma III.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 $R \geq d$, if $|\mathcal{F}| > \frac{1}{\eta} \cdot \exp\big((2eR/d)^d R\log(3d)\big)$, 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 $r = 1, \ldots, R$,*

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

With any choice of means $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^d$ we can associate a vector of moments $\psi(\mu_1, \mu_2, \ldots, \mu_k) = (M_1, \ldots, M_R)$, whose dimension $D = \binom{d+R}{R} < (2eR/d)^d$ since $R \geq d$. The proof then follows using a packing argument in this $D$ dimensional space. We defer the details to the full version.

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 III.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 $c_1, c_0 \geq 1$, such that for every $\varepsilon \leq \exp(-c_1 d)$ if the following holds for $R = c_0\log(1/\varepsilon)$: $\forall 1 \leq r \leq R$,*

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

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

We defer the proof to the full version of the paper.

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 III.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, $\|f - \tilde{f}\|_1 \leq 2\sqrt{\varepsilon}$.*

We defer the proof to the full version.

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

We now briefly describe the new iterative algorithm that estimates the means of a mixture of $k$ spherical Gaussians up to arbitrary accuracy $\delta > 0$ in $\text{poly}(d, k, \log(1/\delta))$ 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., $1/w_{\min} \le \text{poly}(k)$ and $\delta = k^{-\omega(1)}$, or when $d = O(1)$ and $\delta = o(1)$). Please see the full version of the paper for the details.

We assume that we are given coarse initializers $\mu_1^{(0)}, \mu_2^{(0)}, \dots, \mu_k^{(0)}$; we use them to set up an "approximate" system of non-linear equations with "diagonal dominance" properties, and then use the Newton method with the same initializers to solve it. In what follows, $\rho_w$ and $\rho_\sigma$ denote the aspect ratio for the weights and variances respectively as defined in Section II.

**Theorem IV.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 $\mathcal{G}$ with parameters $\{(w_j, \mu_j, \sigma_j) : j \in [k]\}$, where the weights and covariances are known, satisfying*

$$\forall i \neq j \in [k], \quad \|\mu_i - \mu_j\|_2 \qquad (9)$$
$$\ge c(\sigma_i + \sigma_j)\min\{\sqrt{d} + \sqrt{\log(\rho_w\rho_\sigma)}, \sqrt{\log(\rho_\sigma/w_{min})}\}$$

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

$$\forall j \in [k], \ \frac{1}{\sigma_j}\|\mu_j^{(0)} - \mu_j\|_2 \le \frac{c_0}{\min\{d, k\}^{5/2}}. \qquad (10)$$

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

For standard mixtures, (9) corresponds to a separation of order $\min\{\sqrt{\log k}, \sqrt{d}\}$. Firstly, we will assume without loss of generality that $d \le k$, since otherwise we can use a PCA-based dimension-reduction result due to Vempala and Wang [27] (see the full version for a self-contained proof).

### A. 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 that the total probability mass in $S_j$ from other components is much smaller than the probability mass from the component $j$. Since we assume that variances and weights are known, we will use $\tau_j$ to refer to the known $\sigma_j$ in the algorithm description and analysis that follows.

**Definition IV.2** (Region $S_j$)**.** For the set of (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 $\tau_j = \sigma_j$. Then $S_j = \{x : \ |\langle x - z_j, \widehat{e}_{j\ell}\rangle| \le 4\tau_j\sqrt{\log\left(\frac{\rho_\sigma}{w_{\min}}\right)}\ \forall \ell \in [k]$ and $\|x - z_j\|_2 \le 4\tau_j(\sqrt{d} + \sqrt{\log(\rho_\sigma\rho_w)})\}$.

We will use a simple change of variables (so that they are "dimension-free"), that will make our system easier to analyze.

**Definition IV.3.** For $i \in [k]$, let $\mathbf{x}_i \in \mathbb{R}^d$ represent variables of the non-linear system, where $\tau_i\mathbf{x}_i$ represents the (unknown) mean of the $i$th component. Also, let $\mathbf{x}_i^* = \mu_i/\tau_i$, represents the desired solution to the system, and $\tau_i = \sigma_i\ \forall i \in [k]$.

*System of non-linear equations.:* We now describe the system of non-linear equations that we use for the algorithm. In what follows for each $j$, $\tau_j = \sigma_j$, and $z_j = \mu_j^{(0)}$ corresponds to the initializer close to $\mu_j$.

1) For each $j \in [k]$ we consider only the samples $y^{(1)}, y^{(2)}, \dots, y^{(N)} \in \mathbb{R}^d$ in the set $S_j$ and let $\tilde{u}_j$ be the sample average of $(y^{(\ell)} - z_j)$ for $\ell \in [N]$. Let $\tilde{b}_j^{(\mu)} = \frac{1}{w_j\tau_j}\tilde{u}_j$
2) Consider the system $F(\mathbf{x}) = b$ of non-linear equations: $\forall j \in [k],\ F_j(\mathbf{x}) = \tilde{b}_j^{(\mu)}$ with
$$F_j(\mathbf{x}) := \frac{1}{w_j\tau_j}\sum_{i=1}^{k} w_i \int_{y \in S_j}(y - z_j)g_{\tau_i\mathbf{x}_i,\sigma_i}(y)\,dy, \qquad (11)$$
where $g_{\tau_i\mathbf{x}_i,\sigma_i}(y)$ is the p.d.f. of a Gaussian with mean $\tau_i\mathbf{x}_i \in \mathbb{R}^d$, and variance $\sigma_i^2/2\pi$ in each direction. The above constraints are equations involving the variables $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_k)$ (though not in closed-form).

We observe that the population average (i.e., with infinite samples) $b_j^{(\mu)} \in \mathbb{R}^d$ equals $F_j(\mu_i/\tau_i : i \in [k])$ in (11); hence $\mathbf{x}^*$ is indeed a solution to the system with infinite samples.

*Iterative Algorithm to solve the non-linear system.:* We will use Newton's method to solve the non-linear system of equations with initializers $\mathbf{x}_i^{(0)} = \frac{1}{\tau_i}\mu_i^{(0)}$ for each $i \in [k]$. The Newton method uses the following 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) evaluated at $\mathbf{x}^{(t)}$.

We first derive the expression for $F' : \mathbb{R}^{k\cdot d} \to \mathbb{R}^{k\cdot d}$ assuming $\sigma_j = \tau_j$. For all $i, j \in [k]$, $\nabla_{\mathbf{x}_i} F_j(\mathbf{x}) =$

$$\frac{\pi w_i}{w_j \tau_j \tau_i} \int_{y\in S_j} (y - z_j)(y - \tau_i \mathbf{x}_i)^T g_{\tau_i \mathbf{x}_i, \tau_i}(y)\, dy$$

However, we do not have a closed-form expression for $F', F$. Instead we will estimate the values of $F'(\mathbf{x}^{(t)}), F(\mathbf{x}^{(t)})$ from samples. The full version of the paper shows how $F_j, F_j'$ can be estimated at any point $\mathbf{x}^{(t)}$ from samples drawn from the distribution with parameters $\{(w_j, \tau_j \mathbf{x}_j^{(t)}, \tau_j) : j \in [k]\}$ up to any desired inverse polynomial accuracy $\eta > 0$ with polynomial samples. This also shows that the RHS of the equations $b$ can be estimated up to accuracy $\eta$.

In what follows $\varepsilon_0 < c_0 d^{-5/2}$, where $c_0 > 0$ is an appropriate constant.

**Iterative Algorithm for Amplifying Accuracy of Parameter Estimation**

```
Input: Estimation accuracy $\delta > 0$, $N$ samples from a mixture of well-separated Gaussians $\mathcal{G}$ (known weights and variances) and initializers $\mu_i^{(0)}$ for each $i \in [k]$ such that $\|\mu_i^{(0)} - \mu_i\|_\infty \leq \varepsilon_0$, and set $\tau_i = \sigma_i$. Set $T = C \log\log(dk/\delta)$, for some sufficiently large constant $C > 0$.
Output: Estimates $(\mu_i^{(T)} : i \in [k])$ for each component $i \in [k]$ such that $\|\mu_i^{(T)} - \mu_i\|_\infty \leq \delta\sigma_i$.
  1) If $\delta \geq \varepsilon_0 \sqrt{d}$, then we just output $\mu_i^{(T)} = \mu_i^{(0)}$ for each $i \in [k]$.
  2) Set $\mathbf{x}_i^{(0)} = \frac{1}{\tau_i}\mu_i^{(0)}$ for each $i \in [k]$. Set $\eta_1, \eta_2, \eta_3 = \delta/(8c'\rho_\sigma^2 k^6)$, where $c' > 0$ is an appropriately small absolute constant.
  3) Obtain an estimate $\tilde{b}^{(\mu)}$ of $b^{(\mu)}$ from the $N$ samples of the given mixture of $k$ Gaussians.
  4) For $t = 1$ to $T = O(\log\log(1/\delta)$ steps do the following:
      a) Estimate for each $j \in [k]$ an estimate $\tilde{F}(\mathbf{x}^{(t)})$ of $F(\mathbf{x}^{(t)})$ at $\mathbf{x}^{(t)}$.
      b) Estimate for each $j \in [k]$ an estimate $\tilde{J}(\mathbf{x}^{(t)})$ of $F'(\mathbf{x}^{(t)}) = \nabla_{\mathbf{x}} F(\mathbf{x}^{(t)})$.
      c) Update with the Newton iteration $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \left(\tilde{J}(\mathbf{x}^{(t)})\right)^{-1}\left(\tilde{b} - \tilde{F}(\mathbf{x}^{(t)})\right)$.
  5) Output $\mu_i^{(T)} = \tau_i \mathbf{x}_i^{(T)}$ for each $i \in [k]$.
```

### *B. Outline of the 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., $\ell_2$ or $\ell_\infty$). Hence, we will measure the convergence and error of $\mathbf{x} = (\mathbf{x}_i : i \in [k])$ to be measured in $\ell_\infty$ norm.

**Definition IV.4** (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 $\mathbf{x}_i^* = \frac{\mu_i}{\tau_i}\ \forall i \in [k]$, and $c_0 > 0$ is an appropriately large universal constant given in Theorem IV.1.

We will now show the convergence of the Newton method by showing diagonal dominance properties of the non-linear system given in Lemma II.5. To prove Theorem IV.1, we show that $\|F'\|\|F''\|\varepsilon_0 < 1/2$, and use the guarantees of the Newton algorithm. The main technical component of the proof is to show that the function $(F'(\mathbf{x}))^{-1}$ has bounded operator norm using the diagonal dominance properties of $F$.

**Lemma IV.5.** *For any point $\mathbf{x} \in \mathcal{N}$, the operator $F' : \mathbb{R}^{d\cdot k} \to \mathbb{R}^{d\cdot k}$ satisfies $\|(F'(\mathbf{x}))^{-1}\|_{\infty\to\infty} \leq 8$.*

The following lemma shows that the function $F'(\mathbf{x})$ is locally Lipschitz i.e., we bound the second derivative operator.

**Lemma IV.6.** *At any $\mathbf{x} \in \mathcal{N}$, the operator $F'' : \mathbb{R}^{d\cdot k} \times \mathbb{R}^{d\cdot k} \to \mathbb{R}^{d\cdot k}$ satisfies $\|F''\|_{\infty,\infty\to\infty} \leq c' d^{5/2}$, for some absolute constant $c' > 0$.*

This above two lemmas (especially Lemma IV.5) use diagonal dominance that arises from the separation between the means of the components. We show that most of the probability mass from $j$th component around $\mu_j$ is confined to $S_j$, while the other components of $\mathcal{G}$ are far enough from $z_j$ that they do not contribute much $\ell_1$ mass in total to $S_j$. The proof of the latter statement is the more technical of the two, and it is very different for separation of order $\sqrt{\log k}$ and $\sqrt{d}$ – hence they are handled separately in the full version. We defer all the proofs to the full version of the paper.

## V. Initialization and Algorithmic Guarantees

*Initialization for High-dimensions:* We now give the general statement of the theorem showing that a mean separation of order $\Omega(\sqrt{\log k})$ suffices to learn model parameters can be learned 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 $\rho, 1/w_{\min} \leq \text{poly}(k)$. In what follows $\rho_\sigma$

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 V.1** (Same as Theorem I.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 $\rho$-bounded and the means are well-separated i.e. $\forall i, j \in [k], i \neq j$:*

$$\|\mu_i - \mu_j\|_2 \geq c\sqrt{\log(\rho_\sigma / w_{\mathit{min}})}(\sigma_i + \sigma_j), \tag{12}$$

*there is an algorithm that for any $\delta > 0$, uses* $\mathrm{poly}(d, \rho, 1/w_{\mathit{min}}, 1/\delta)$ *samples and recovers with high probability the means up to $\delta$ relative error i.e., finds $\mu'_1, \dots, \mu'_k$ s.t. $\|\mu'_j - \mu_j\|_2 \leq \delta\sigma_j$ for all $j \in [k]$.*

Such results are commonly referred to as *polynomial identifiability* or *robust identifiability* results. Theorem V.1 follows in a straightforward manner by combining the iterative algorithm, with initializers given by the following theorem (note that the separation here does depend on the accuracy $k^{-c}$).

**Theorem V.2.** *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 $\rho$-bounded and the means are well-separated i.e. $\forall i, j \in [k], i \neq j$:*

$$\|\mu_i - \mu_j\|_2 \geq 4c\sqrt{\log(\rho_\sigma / w_{\mathit{min}})}(\sigma_i + \sigma_j). \tag{13}$$

*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 $\Delta_{\mathrm{param}}(\mathcal{G}, \tilde{\mathcal{G}}) \leq k^{-c}$.*

*Initialization for Low-dimensions:* We now state our general result giving a computationally efficient algorithm that works in $d = O(1)$ dimensions, even when the separation is of order $O(1)$. In comparison, previous algorithms need separation of the order of $\Omega(\sqrt{\log k})$. We prove the following theorem.

**Theorem V.3.** *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 $\|\mu_j\| \leq \rho \ \forall j \in [k]$ and $\forall i, j \in [k], i \neq j$:*

$$\|\mu_i - \mu_j\|_2 \geq c\left(\sqrt{d} + \sqrt{\log(\rho_\sigma \rho_w)}\right) \cdot (\sigma_i + \sigma_j). \tag{14}$$

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

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 IV.1) along with a computationally efficient procedure that finds appropriate initializers. The following theorem shows how to find reasonable initializers for $\{\tilde{\mu}_j : j \in [k]\}$ that can be used by the iterative algorithm. We will show that for any $\varepsilon_0 = \exp(-c_0 d)$, we have an algorithm running in time $(\rho/\varepsilon_0^3 w_{\min})^{O(d)}$ that with separation of order $\sqrt{d}$ will find initializers $\{\tilde{\mu}_j : j \in [k]\}$ such that $\|\mu_j - \tilde{\mu}_j\| \leq \varepsilon_0 \sigma_j \ \forall j \in [k]$.

**Theorem V.4.** *Let $c_0 \geq 2$ be any constant, and $\varepsilon_0 = \exp(-c_0 d)$. There is an algorithm running in $(\rho/(\varepsilon_0^3 w_{\mathit{min}}))^{O(d)}$ time that given samples from a $\rho$-bounded mixture of $k$ spherical Gaussians $\mathcal{G} = \{(w_j, \mu_j, \sigma_j) : j \in [k]\}$ in $d$ dimensions satisfying $\forall i \neq j \in [k]$, $\|\mu_i - \mu_j\|_2 \geq 4c_0\left(\sqrt{d} + \sqrt{\log(\rho_w \rho_\sigma)}\right)(\sigma_i + \sigma_j)$, can find with high probability $\tilde{\mu}_1, \dots, \tilde{\mu}_k$ s.t. $\|\tilde{\mu}_j - \mu_j\|_2 \leq \varepsilon_0 \sigma_j \sqrt{d}$ for all $j \in [k]$.*

## Acknowledgements

A.V. is grateful to Santosh Vempala for suggesting the problem of learning one-dimensional Gaussians. O.R. is supported by the Simons Collaboration on Algorithms and Geometry and by the National Science Foundation (NSF) under Grant No. CCF-1320188. A.V. is supported by the National Science Foundation (NSF) under Grant No. CCF-1652491 and CCF-1637585, and a startup grant from Northwestern University. 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.

## References

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

[2] 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 COLT 2014*, pages 1135–1164, 2014.

[3] 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.

[4] Kendall Atkinson and Weimin Han. *Theoretical numerical analysis : a functional analysis framework.* Springer, 2001.

[5] Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. In *APPROX/RANDOM*, pages 37–49. Springer, 2012.

[6] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 2017.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

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

[14] 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.

[15] 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.

[16] Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two Gaussians. In *Proceedings of STOC 2015*, pages 753–760, 2015.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

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

[23] 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.

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

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

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

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

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