Toward Learning Gaussian Mixtures with Arbitrary Separation

Mikhail Belkin
Ohio State University
Columbus, Ohio
mbelkin@cse.ohio-state.edu

Kaushik Sinha
Ohio State University
Columbus, Ohio
sinhak@cse.ohio-state.edu

Abstract

In recent years analysis of complexity of learning Gaussian mixture models from sampled data has received significant attention in computational machine learning and theory communities. In this paper we present the first result showing that polynomial time learning of multidimensional Gaussian Mixture distributions is possible when the separation between the component means is arbitrarily small. Specifically, we present an algorithm for learning the parameters of a mixture of k identical spherical Gaussians in n-dimensional space with an arbitrarily small separation between the components, which is polynomial in dimension, inverse component separation and other input parameters for a fixed number of components k. The algorithm uses a projection to k dimensions and then a reduction to the 1-dimensional case. It relies on a theoretical analysis showing that two 1-dimensional mixtures whose densities are close in the L^2 norm must have similar means and mixing coefficients. To produce the necessary lower bound for the L^2 norm in terms of the distances between the corresponding means, we analyze the behavior of the Fourier transform of a mixture of Gaussians in one dimension around the origin, which turns out to be closely related to the properties of the Vandermonde matrix obtained from the component means. Analysis of minors of the Vandermonde matrix together with basic function approximation results allows us to provide a lower bound for the norm of the mixture in the Fourier domain and hence a bound in the original space. Additionally, we present a separate argument for reconstructing variance.

1 Introduction

Mixture models, particularly Gaussian mixture models, are a widely used tool for many problems of statistical inference (Titterington et al., 1985; McLachlan & Peel, 2000; McLachlan & Basford, 1988; Everitt & Hand, 1981; Lindsay, 1995). The basic problem is to estimate the parameters of a mixture distribution, such as the mixing coefficients, means and variances within some pre-specified precision from a number of sampled data points. While the history of Gaussian mixture models goes back to (Pearson, 1894), in recent years the theoretical aspects of mixture learning have attracted considerable attention in the theoretical computer science, starting with the pioneering work of (Dasgupta, 1999), who showed that a mixture of k spherical Gaussians in n dimensions can be learned in time polynomial in n, provided certain separation conditions between the component means (separation of order $\sqrt{n}$) are satisfied. This work has been refined and extended in a number of recent papers. The first result from (Dasgupta, 1999) was later improved to the order of $\Omega(n^{\frac{1}{4}})$ in (Dasgupta & Schulman, 2000) for spherical Gaussians and in (Arora & Kannan, 2001) for general Gaussians. The separation requirement was further reduced and made independent of n to the order of $\Omega(k^{\frac{1}{4}})$ in (Vempala & Wang, 2002) for spherical Gaussians and to the order of $\Omega(\frac{k^{\frac{3}{2}}}{\epsilon^2})$ in (Kannan et al., 2005) for Logconcave distributions. In a related work (Achlioptas & McSherry, 2005) the separation requirement was reduced to $\Omega(k + \sqrt{k \log n})$. An extension of PCA called isotropic PCA was introduced in (Brubaker & Vempala, 2008) to learn mixtures of Gaussians when any pair of Gaussian components is separated by a hyperplane having very small overlap along the hyperplane direction (so-called "pancake layering problem").

In a slightly different direction the recent work (Feldman et al., 2006) made an important contribution to the subject by providing a polynomial time algorithm for PAC-style learning of mixture of Gaussian distributions with arbitrary separation between the means. The authors used a grid search over the space of parameters to a construct a hypothesis mixture of Gaussians that has density close to the actual mixture generating the data. We note that the problem analyzed in (Feldman et al., 2006) can be viewed as density estimation within

a certain family of distributions and is different from most other work on the subject, including our paper, which address parameter learning[1].

We also note several recent papers dealing with the related problems of learning mixture of product distributions and heavy tailed distributions. See for example, (Feldman et al., 2008; Dasgupta et al., 2005; Chaudhuri & Rao, 2008a; Chaudhuri & Rao, 2008b).

In the statistics literature, (Chen, 1995) showed that optimal convergence rate of MLE estimator for finite mixture of normal distributions is $O(n^{-\frac{1}{2}})$, where n is the sample size, if number of mixing components k is known in advance and is $O(n^{-\frac{1}{4}})$ when the number of mixing components is known up to an upper bound. However, this result does not address the computational aspects, especially in high dimension.

In this paper we develop a polynomial time (for a fixed k) algorithm to identify the parameters of the mixture of k identical spherical Gaussians with potentially unknown variance for an arbitrarily small separation between the components[2]. To the best of our knowledge this is the first result of this kind except for the simultaneous and independent work (Kalai et al., 2010), which analyzes the case of a mixture of two Gaussians with arbitrary covariance matrices using the method of moments. We note that the results in (Kalai et al., 2010) and in our paper are somewhat orthogonal. Each paper deals with a special case of the ultimate goal (two arbitrary Gaussians in (Kalai et al., 2010) and k identical spherical Gaussians with unknown variance in our case), which is to show polynomial learnability for a mixture with an arbitrary number of components and arbitrary variance.

All other existing algorithms for parameter estimation require minimum separation between the components to be an increasing function of at least one of n or k. Our result also implies a density estimate bound along the lines of (Feldman et al., 2006). We note, however, that we do have to pay a price as our procedure (similarly to that in (Feldman et al., 2006)) is super-exponential in k. Despite these limitations we believe that our paper makes a step towards understanding the fundamental problem of polynomial learnability of Gaussian mixture distributions. We also think that the technique used in the paper to obtain the lower bound may be of independent interest.

The main algorithm in our paper involves a grid search over a certain space of parameters, specifically means and mixing coefficients of the mixture (a completely separate argument is given to estimate the variance). By giving appropriate lower and upper bounds for the norm of the difference of two mixture distributions in terms of their means, we show that such a grid search is guaranteed to find a mixture with nearly correct values of the parameters.

To prove that, we need to provide a lower and upper bounds on the norm of the mixture. A key point of our paper is the lower bound showing that two mixtures with different means cannot produce similar density functions. This bound is obtained by reducing the problem to a 1-dimensional mixture distribution and analyzing the behavior of the Fourier transform (closely related to the characteristic function, whose coefficients are moments of a random variable up to multiplication by a power of the imaginary unit i) of the difference between densities near zero. We use certain properties of minors of Vandermonde matrices to show that the norm of the mixture in the Fourier domain is bounded from below. Since the L^2 norm is invariant under the Fourier transform this provides a lower bound on the norm of the mixture in the original space.

We also note the work (Lindsay, 1989), where Vandermonde matrices appear in the analysis of mixture distributions in the context of proving consistency of the method of moments (in fact, we rely on a result from (Lindsay, 1989) to provide an estimate for the variance).

Finally, our lower bound, together with an upper bound and some results from the non-parametric density estimation and spectral projections of mixture distributions allows us to set up a grid search algorithm over the space of parameters with the desired guarantees.

Extended version of this paper is available at `http://arxiv.org/abs/0907.1054`.

2 Outline of the argument

In this section we provide an informal outline of the argument that leads to the main result. To simplify the discussion, we will assume that the variance for the components is known or estimated by using the estimation algorithm provided in Section 3.3. It is straightforward (but requires a lot of technical details) to see that all results go through if the actual variance is replaced by a sufficiently (polynomially) accurate estimate.

We will denote the n-dimensional Gaussian density $\frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{\mu}_i\|^2}{2\sigma^2}\right)$ by $K(\boldsymbol{x},\boldsymbol{\mu})$, where $\boldsymbol{x},\boldsymbol{\mu}\in$

[1] Note that density estimation is generally easier than parameter learning since quite different configurations of parameters could conceivably lead to very similar density functions, while similar configurations of parameters always result in similar density functions.

[2] We point out that some non-zero separation is necessary since the problem of learning parameters without any separation assumptions at all is ill-defined.

$\mathbb{R}^n$ or, when appropriate, in $\mathbb{R}^k$. The notation $\|\cdot\|$ will always be used to represent L^2 norm while $d_H(\cdot,\cdot)$ will be used to denote the Hausdorff distance between sets of points. Let $p(\boldsymbol{x}) = \sum_{i=1}^k \alpha_i K(\boldsymbol{x}, \boldsymbol{\mu}_i)$ be a mixture of k Gaussian components with the covariance matrix $\sigma^2 I$ in $\mathbb{R}^n$. The goal will be to identify the means $\boldsymbol{\mu}_i$ and the mixing coefficients α_i under the assumption that the minimum distance $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|, i \neq j$ is bounded from below by some given (arbitrarily small) $d_{\min}$ and the minimum mixing weight is bounded from below by $\alpha_{\min}$. We note that while σ can also be estimated, we will assume that it is known in advance to simplify the arguments. The number of components needs to be known in advance which is in line with other work on the subject. Our main result is an algorithm guaranteed to produce an approximating mixture $\tilde{p}$, whose means and mixing coefficients are all within ϵ of their true values and whose running time is a polynomial in all parameters other than k. Input to our algorithm is $\alpha_{\min}, \sigma, k$, N points in $\mathbb{R}^n$ sampled from p and an arbitrary small positive ϵ satisfying $\epsilon \leq \frac{d_{\min}}{2}$. The algorithm has the following main steps.

Parameters: $\alpha_{\min}, d_{min}, \sigma, k$.
Input: $\epsilon \leq \frac{d_{min}}{2}$, N points in $\mathbb{R}^n$ sampled from p.
Output: $\boldsymbol{\theta}^*$, the vector of approximated means and mixing coefficients.

Step 1. (Reduction to k dimensions). Given a polynomial number of data points sampled from p it is possible to identify the k-dimensional span of the means $\boldsymbol{\mu_i}$ in $\mathbb{R}^n$ by using Singular Value Decomposition (see (Vempala & Wang, 2002)). By an additional argument the problem can be reduced to analyzing a mixture of k Gaussians in $\mathbb{R}^k$.

Step 2. (Construction of kernel density estimator). Using Step 1, we can assume that $n = k$. Given a sample of N points in $\mathbb{R}^k$, we construct a density function p_{kde} using an appropriately chosen kernel density estimator. Given sufficiently many points, $\|p - p_{kde}\|$ can be made arbitrarily small. Note that while p_{kde} is a mixture of Gaussians, it is *not* a mixture of k Gaussians.

Step 3. (Grid search). Let $\boldsymbol{\Theta} = (\mathbb{R}^k)^k \times \mathbb{R}^k$ be the $k^2 + k$-dimensional space of parameters (component means and mixing coefficients) to be estimated. Because of Step 1, we can assume (see Lemma 1) $\boldsymbol{\mu}_i$s are in $\mathbb{R}^k$.

For any $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\mu}}_2, \cdots, \tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\alpha}}) = (\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{\alpha}}) \in \boldsymbol{\Theta}$, let $p(\boldsymbol{x}, \tilde{\boldsymbol{\theta}})$ be the corresponding mixture distribution. Note that $\boldsymbol{\theta} = (\boldsymbol{m}, \boldsymbol{\alpha}) \in \boldsymbol{\Theta}$ are the true parameters. We obtain a value G (polynomial in all arguments for a fixed k) from Theorem 4 and take a grid M_G of size G in $\boldsymbol{\Theta}$. The value $\boldsymbol{\theta}^*$ is found from a grid search according to the following equation

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\tilde{\boldsymbol{\theta}} \in M_G} \left\{ \|p(\boldsymbol{x}, \tilde{\boldsymbol{\theta}}) - p_{kde}\| \right\} \tag{1}$$

We show that the means and mixing coefficients obtained by taking $\boldsymbol{\theta}^*$ are close to the true underlying means and mixing coefficients of p with high probability. We note that our algorithm is deterministic and the uncertainty comes only from the sample (through the SVD projection and density estimation).

While a somewhat different grid search algorithm was used in (Feldman et al., 2006), the main novelty of our result is showing that the parameters estimated from the grid search are close to the true underlying parameters of the mixture. In principle, it is conceivable that two different configurations of Gaussians could give rise to very similar mixture distributions. However, we show that this is not the case. Specifically, and this is the theoretical core of this paper, we show that mixtures with different means/mixing coefficients cannot be close in L^2 norm[3] (Theorem 2) and thus the grid search yields parameter values $\boldsymbol{\theta}^*$ that are close to the true values of the means and mixing coefficients.

To provide a better high-level overview of the whole proof we give a high level summary of the argument (Steps 2 and 3).

1. Since we do not know the underlying probability distribution p directly, we construct p_{kde}, which is a proxy for $p = p(\boldsymbol{x}, \boldsymbol{\theta})$. p_{kde} is obtained by taking an appropriate non-parametric density estimate and, given a sufficiently large polynomial sample, can be made to be arbitrarily close to p in L^2 norm (see Lemma 9). Thus the problem of approximating p in L^2 norm can be replaced by approximating p_{kde}.

2. The main technical part of the paper are the lower and upper bounds on the norm $\|p(\boldsymbol{x}, \boldsymbol{\theta}) - p(\boldsymbol{x}, \tilde{\boldsymbol{\theta}})\|$ in terms of the Hausdorff distance between the component means (considered as sets of k points) $\boldsymbol{m}$ and $\tilde{\boldsymbol{m}}$. Specifically, in Theorem 2 and Lemma 3 we prove that for $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{\alpha}})$

$$d_H(\boldsymbol{m}, \tilde{\boldsymbol{m}}) \leq f(\|p(\boldsymbol{x}, \boldsymbol{\theta}) - p(\boldsymbol{x}, \tilde{\boldsymbol{\theta}})\|) \leq h(d_H(\boldsymbol{m}, \tilde{\boldsymbol{m}}) + \|\boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}}\|_1)$$

where f, h are some explicitly given increasing functions. The lower bound shows that $d_H(\boldsymbol{m}, \tilde{\boldsymbol{m}})$ can be controlled by making $\|p(\boldsymbol{x}, \boldsymbol{\theta}) - p(\boldsymbol{x}, \tilde{\boldsymbol{\theta}})\|$ sufficiently small, which (assuming minimum separation

[3]Note that our notion of distance between two density functions is slightly different from the standard ones used in literature, e.g., Hellinger distance or KL divergence. However, our goal is to estimate the parameters and here we use L^2 norm merely as a tool to describe that two distributions are different.

d_{min} between the components of p) immediately implies that each component mean of $\boldsymbol{m}$ is close to exactly one component mean of $\tilde{\boldsymbol{m}}$.

On the other hand, the upper bound guarantees that a search over a sufficiently fine grid in the space $\boldsymbol{\Theta}$ will produce a value $\boldsymbol{\theta}^*$, s.t. $\|p(\boldsymbol{x}, \boldsymbol{\theta}) - p(\boldsymbol{x}, \boldsymbol{\theta}^*)\|$ is small.

3. Once the component means $\boldsymbol{m}$ and $\tilde{\boldsymbol{m}}$ are shown to be close an argument using the Lipschitz property of the mixture with respect to the mean locations can be used to establish that the corresponding mixing coefficient are also close (Corollary 5).

We will now briefly outline the argument for the main theoretical contribution of this paper which is a lower bound on the L^2 norm in terms of the Hausdorff distance (Theorem 2).

1. (Minimum distance, reduction from $\mathbb{R}^k$ to $\mathbb{R}^1$) Suppose a component mean $\boldsymbol{\mu}_i$, is separated from every estimated mean $\tilde{\boldsymbol{\mu}}_j$ by a distance of at least d, then there exists a unit vector $\boldsymbol{v}$ in $\mathbb{R}^k$ such than $\forall_{i,j}$ $|\langle \boldsymbol{v}, (\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_j)\rangle| \geq \frac{d}{4k^2}$. In other words a certain amount of separation is preserved after an appropriate projection to one dimension. See Lemma 10 for a proof.

2. (Norm estimation, reduction from $\mathbb{R}^k$ to $\mathbb{R}^1$). Let p and $\tilde{p}$ be the true and estimated density respectively and let $\boldsymbol{v}$ be a unit vector in $\mathbb{R}^k$. $p_{\boldsymbol{v}}$ and $\tilde{p}_{\boldsymbol{v}}$ will denote the one-dimensional marginal densities obtained by integrating p and $\tilde{p}$ in the directions orthogonal to $\boldsymbol{v}$. It is easy to see that $p_{\boldsymbol{v}}$ and $\tilde{p}_{\boldsymbol{v}}$ are mixtures of 1-dimensional Gaussians, whose means are projections of the original means onto $\boldsymbol{v}$. It is shown in Lemma 11 that
$$\|p - \tilde{p}\|^2 \geq \left(\frac{1}{c\sigma}\right)^k \|p_{\boldsymbol{v}} - \tilde{p}_{\boldsymbol{v}}\|^2$$
and thus to provide a lower bound for $\|p - \tilde{p}\|$ it is sufficient to provide an analogous bound (with a different separation between the means) in one dimension.

3. (1-d lower bound) Finally, we consider a mixture q of $2k$ Gaussians in one dimension, with the assumption that one of the component means is separated from the rest of the component means by at least t and that the (not necessarily positive) mixing weights exceed α_{min} in absolute value. Assuming that the means lie in an interval $[-a, a]$ we show (Theorem 6)
$$\|q\|^2 \geq \alpha_{\min}^{4k} \left(\frac{t}{a^2}\right)^{Ck^2}$$
for some positive constant C independent of k.

   The proof of this result relies on analyzing the Taylor series for the Fourier transform of q near zeros, which turns out to be closely related to a certain Vandermonde matrix.

Combining 1 and 2 above and applying the result in 3, $q = p_{\boldsymbol{v}} - \tilde{p}_{\boldsymbol{v}}$ yields the desired lower bound for $\|p - \tilde{p}\|$.

3 Main Results

In this section we present our main results. First we show that we can reduce the problem in $\mathbb{R}^n$ to a corresponding problem in $\mathbb{R}^k$, where n represents the dimension and k is the number of components, at the cost of an arbitrarily small error. Then we solve the reduced problem in $\mathbb{R}^k$, again allowing for only an arbitrarily small error, by establishing appropriate lower and upper bounds of a mixture norm in $\mathbb{R}^k$.

Lemma 1 (Reduction from $\mathbb{R}^n$ to $\mathbb{R}^k$) *Consider a mixture of k n-dimensional spherical Gaussians $p(\boldsymbol{x}) = \sum_{i=1}^k \alpha_i K(\boldsymbol{x}, \boldsymbol{\mu}_i)$ where the means lie within a cube $[-1, 1]^n$, $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| \geq d_{\min} > 0, \forall_{i \neq j}$ and for all i, $\alpha_i > \alpha_{\min}$. For any positive $\epsilon \leq \frac{d_{\min}}{2}$ and $\delta \in (0, 1)$, given a sample of size $\text{poly}\left(\frac{n}{\epsilon \alpha_{\min}}\right) \cdot \log\left(\frac{1}{\delta}\right)$, with probability greater than $1 - \delta$, the problem of learning the parameters (means and mixing weights) of p within ϵ error can be reduced to learning the parameters of a k-dimensional mixture of spherical Gaussians $p_o(\boldsymbol{x}) = \sum_{i=1}^k \alpha_i K(\boldsymbol{x}, \boldsymbol{\nu}_i)$ where the means lie within a cube $[-\sqrt{\frac{n}{k}}, \sqrt{\frac{n}{k}}]^k$, $\|\boldsymbol{\nu}_i - \boldsymbol{\nu}_j\| > \frac{d_{\min}}{2} > 0, \forall_{i \neq j}$. However, in $\mathbb{R}^k$ we need to learn the means within $\frac{\epsilon}{2}$ error.*

Proof: For $i = 1, \ldots, k$, let $\boldsymbol{v}_i \in \mathbb{R}^n$ be the top k right singular vectors of a data matrix of size $\text{poly}\left(\frac{n}{\epsilon \alpha_{\min}}\right) \cdot \log\left(\frac{1}{\delta}\right)$ sampled from $p(\boldsymbol{x})$. It is well known (see (Vempala & Wang, 2002)) that the space spanned by the

means $\{\boldsymbol{\mu}_i\}_{i=1}^k$ remains arbitrarily close to the space spanned by $\{\boldsymbol{v}_i\}_{i=1}^k$. In particular, with probability greater than $1-\delta$, the projected means $\{\tilde{\boldsymbol{\mu}}_i\}_{i=1}^k$ satisfy $\|\boldsymbol{\mu}_i - \tilde{\boldsymbol{\mu}}_i\| \leq \frac{\epsilon}{2}$ for all i (see Lemma 12).

Note that each projected mean $\tilde{\boldsymbol{\mu}}_i \in \mathbb{R}^n$ can be represented by a k dimensional vector $\boldsymbol{\nu}_i$ which are the coefficients along the singular vectors $\boldsymbol{v}_j$s, that is for all i, $\tilde{\boldsymbol{\mu}}_i = \sum_{j=1}^k \nu_{ij}\boldsymbol{v}_j$. Thus, for any $i \neq j, \|\tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}}_j\| = \|\boldsymbol{\nu}_i - \boldsymbol{\nu}_j\|$. Since $\|\tilde{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}}_j\| \geq d_{\min} - \frac{\epsilon}{2} - \frac{\epsilon}{2} = d_{\min} - \epsilon \geq d_{\min} - \frac{d_{\min}}{2} = \frac{d_{\min}}{2}$, we have $\|\boldsymbol{\nu}_i - \boldsymbol{\nu}_j\| \geq \frac{d_{\min}}{2}$. Also note that each $\boldsymbol{\nu}_i$ lie within a cube of $[-\sqrt{\frac{n}{k}}, \sqrt{\frac{n}{k}}]^k$ where the axes of the cube are along the top k singular vectors $\boldsymbol{v}_j$s.

Now suppose we can estimate each $\boldsymbol{\nu}_i$ by $\tilde{\boldsymbol{\nu}}_i \in \mathbb{R}^k$ such that $\|\boldsymbol{\nu}_i - \tilde{\boldsymbol{\nu}}_i\| \leq \frac{\epsilon}{2}$. Again each $\tilde{\boldsymbol{\nu}}_i$ has a corresponding representation $\hat{\boldsymbol{\mu}}_i \in \mathbb{R}^n$ such that $\hat{\boldsymbol{\mu}}_i = \sum_{j=1}^k \tilde{\nu}_{ij}\boldsymbol{v}_j$ and $\|\tilde{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_i\| = \|\boldsymbol{\nu}_i - \tilde{\boldsymbol{\nu}}_i\|$. This implies for each i, $\|\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i\| \leq \|\boldsymbol{\mu}_i - \tilde{\boldsymbol{\mu}}_i\| + \|\tilde{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_i\| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$. ■

From here onwards we will deal with mixture of Gaussians in $\mathbb{R}^k$. Thus we will assume that p_o denotes the true mixture with means $\{\boldsymbol{\nu}_i\}_{i=1}^k$ while $\tilde{p}_o$ represents any other mixture in $\mathbb{R}^k$ with different means and mixing weights.

We first prove a lower bound for $\|p_o - \tilde{p}_o\|$.

Theorem 2 (Lower bound in $\mathbb{R}^k$) *Consider a mixture of k k-dimensional spherical Gaussians $p_o(\boldsymbol{x}) = \sum_{i=1}^k \alpha_i K(\boldsymbol{x}, \boldsymbol{\nu}_i)$ where the means lie within a cube $[-\sqrt{\frac{n}{k}}, \sqrt{\frac{n}{k}}]^k$, $\|\boldsymbol{\nu}_i - \boldsymbol{\nu}_j\| \geq \frac{d_{\min}}{2} > 0, \forall_{i\neq j}$ and for all i,$\alpha_i > \alpha_{\min}$. Let $\tilde{p}_o(\boldsymbol{x}) = \sum_{i=1}^k \tilde{\alpha}_i K(\boldsymbol{x}, \tilde{\boldsymbol{\nu}}_i)$ be some arbitrary mixture such that the Hausdorff distance between the set of true means $\boldsymbol{m}$ and the estimated means $\tilde{\boldsymbol{m}}$ satisfies $d_H(\boldsymbol{m}, \tilde{\boldsymbol{m}}) \leq \frac{d_{\min}}{4}$. Then $\|p_o - \tilde{p}_o\|^2 \geq \left(\frac{\alpha_{\min}^4}{c\sigma}\right)^k \left(\frac{d_H(\boldsymbol{m},\tilde{\boldsymbol{m}})}{4nk^2}\right)^{Ck^2}$ where C, c are some positive constants independent of n, k.*

Proof: Consider any arbitrary $\boldsymbol{\nu}_i$ such that its closest estimate $\tilde{\boldsymbol{\nu}}_i$ from $\tilde{\boldsymbol{m}}$ is $t = \|\boldsymbol{\nu}_i - \tilde{\boldsymbol{\nu}}_i\|$. Note that $t \leq \frac{d_{\min}}{4}$ and all other $\boldsymbol{\nu}_j, \tilde{\boldsymbol{\nu}}_j, j \neq i$ are at a distance at least t from $\boldsymbol{\nu}_i$. Lemma 10 ensures the existence of a direction $\boldsymbol{v} \in \mathbb{R}^k$ such that upon projecting on which $|\langle \boldsymbol{v}, (\boldsymbol{\nu}_i - \tilde{\boldsymbol{\nu}}_i)\rangle| \geq \frac{t}{4k^2}$ and all other projected means $\langle \boldsymbol{v}, \boldsymbol{\nu}_j\rangle, \langle \boldsymbol{v}, \tilde{\boldsymbol{\nu}}_j\rangle, j \neq i$ are at a distance at least $\frac{t}{4k^2}$ from $\langle \boldsymbol{v}, \boldsymbol{\nu}_i\rangle$. Note that after projecting on $\boldsymbol{v}$, the mixture becomes a mixture of 1-dimensional Gaussians with variance σ^2 and whose projected means lie within $[-\sqrt{n}, \sqrt{n}]$. Let us denote these 1-dimensional mixtures by $p_{\boldsymbol{v}}$ and $\tilde{p}_{\boldsymbol{v}}$ respectively. Then using Theorem 6 $\|p_{\boldsymbol{v}} - \tilde{p}_{\boldsymbol{v}}\|^2 \geq \alpha_{\min}^{4k} \left(\frac{(t/4k^2)}{n}\right)^{Ck^2}$. Note that we obtain $p_{\boldsymbol{v}}$ (respectively $\tilde{p}_{\boldsymbol{v}}$) by integrating p_o (respectively $\tilde{p}_o$) in all $(k-1)$ orthogonal directions to $\boldsymbol{v}$. Now we need to relate $\|p_o - \tilde{p}_o\|$ and $\|p_{\boldsymbol{v}} - \tilde{p}_{\boldsymbol{v}}\|$. This is done in Lemma 11 to ensure that $\|p_o - \tilde{p}_o\|^2 \geq \left(\frac{1}{c\sigma}\right)^k \|p_{\boldsymbol{v}} - \tilde{p}_{\boldsymbol{v}}\|^2$ where $c >$ is in chosen such a way that in any arbitrary direction probability mass of each projected Gaussian on that direction becomes negligible outside the interval of $[-c\sigma/2, c\sigma/2]$. Thus, $\|p_o - \tilde{p}_o\|^2 \geq \left(\frac{\alpha_{\min^4}}{c\sigma}\right)^k \left(\frac{t}{4nk^2}\right)^{Ck^2}$. Since this holds for any arbitrary $\boldsymbol{\nu}_i$, we can replace t by $d_H(\boldsymbol{m}, \tilde{\boldsymbol{m}})$. ■

Next, we prove a straightforward upper bound for $\|p_o - \tilde{p}_o\|$.

Lemma 3 (Upper bound in $\mathbb{R}^k$) *Consider a mixture of k, k-dimensional spherical Gaussians $p_o(\boldsymbol{x}) = \sum_{i=1}^k \alpha_i K(\boldsymbol{x}, \boldsymbol{\nu}_i)$ where the means lie within a cube $[-\sqrt{\frac{n}{k}}, \sqrt{\frac{n}{k}}]^k$, $\|\boldsymbol{\nu}_i - \boldsymbol{\nu}_j\| \geq \frac{d_{\min}}{2} > 0, \forall_{i\neq j}$ and for all i,$\alpha_i > \alpha_{\min}$. Let $\tilde{p}_o(\boldsymbol{x}) = \sum_{i=1}^k \tilde{\alpha}_i K(\boldsymbol{x}, \tilde{\boldsymbol{\nu}}_i)$ be some arbitrary mixture such that the Hausdorff distance between the set of true means $\boldsymbol{m}$ and the estimated means $\tilde{\boldsymbol{m}}$ satisfies $d_H(\boldsymbol{m}, \tilde{\boldsymbol{m}}) \leq \frac{d_{\min}}{4}$. Then there exists a permutation $\pi : \{1, 2, \ldots, k\} \to \{1, 2, , \ldots, k\}$ such that*

$$\|p_o - \tilde{p}_o\| \leq \frac{1}{(2\pi\sigma^2)^{k/2}} \sum_{i=1}^k \left(\sqrt{|\alpha_i - \tilde{\alpha}_{\pi(i)}|^2 + \frac{d_H^2(\boldsymbol{m}, \tilde{\boldsymbol{m}})}{\sigma^2}}\right)$$

Proof: Due to the constraint on the Hausdorff distance and constraint on the pair wise distance between the means of $\boldsymbol{m}$, there exists a permutation $\pi : \{1, 2, \ldots, k\} \to \{1, 2, , \ldots, k\}$ such that $\|\boldsymbol{\nu}_i - \tilde{\boldsymbol{\nu}}_{\pi(i)}\| \leq d_H(\boldsymbol{m}, \tilde{\boldsymbol{m}})$. Due to one-to-one correspondence, without loss of generality we can write,
$\|p_o - \tilde{p}_o\| \leq \sum_{i=1}^k ||g_i\|$ where $g_i(\boldsymbol{x}) = \alpha_i K(\boldsymbol{x}, \boldsymbol{\nu}_i) - \tilde{\alpha}_{\pi(i)} K(\boldsymbol{x}, \tilde{\boldsymbol{\nu}}_{\pi(i)})$. Now using Lemma 13,
$\|g_i\|^2 \leq \frac{1}{(2\pi\sigma^2)^k}\left(\alpha_i^2 + \tilde{\alpha}_{\pi(i)}^2 - 2\alpha_i\tilde{\alpha}_{\pi(i)} \exp\left(-\frac{\|\boldsymbol{\nu}_i - \tilde{\boldsymbol{\nu}}_{\pi(i)}\|^2}{2\sigma^2}\right)\right)$
$= \frac{1}{(2\pi\sigma^2)^k}\left((\alpha_i - \tilde{\alpha}_{\pi(i)})^2 + 2\alpha_i\tilde{\alpha}_{\pi(i)}\left(1 - \exp\left(-\frac{\|\boldsymbol{\nu}_i - \tilde{\boldsymbol{\nu}}_{\pi(i)}\|^2}{2\sigma^2}\right)\right)\right)$
$\leq \frac{1}{(2\pi\sigma^2)^k}\left((\alpha_i - \tilde{\alpha}_{\pi(i)})^2 + \frac{\alpha_i\tilde{\alpha}_{\pi(i)}\|\boldsymbol{\nu}_i - \tilde{\boldsymbol{\nu}}_{\pi(i)}\|^2}{\sigma^2}\right)$ ■

We now present our main result for learning mixture of Gaussians with arbitrary small separation.

Theorem 4 *Consider a mixture of k n-dimensional spherical Gaussians $p(\boldsymbol{x}) = \sum_{i=1}^{k} \alpha_i K(\boldsymbol{x}, \boldsymbol{\mu}_i)$ where the means lie within a cube $[-1,1]^n$, $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| > d_{\min} > 0, \forall_{i \neq j}$ and for all i, $\alpha_i > \alpha_{\min}$. Then given any positive $\epsilon \leq \frac{d_{\min}}{2}$ and $\delta \in (0,1)$, there exists a positive C_1 independent of n and k such that using a sample of size $N = \text{poly}\left(\left(\frac{nk^2}{\epsilon}\right)^{k^3} \cdot \log^k\left(\frac{2}{\delta}\right)\right)$ and a grid M_G of size $G = \frac{(\alpha_{\min}^4)^k}{k^{3/2}} \left(\frac{\epsilon}{8nk^2}\right)^{C_1 k^2}$, our algorithm given by Equation 1 runs in time $\frac{k^{3/2}}{(\alpha_{\min}^4 \sigma)^k} \left(\frac{n^{3/2}k^{1/2}}{\epsilon}\right)^{C_1 k^2}$ and provides mean estimates which, with probability greater than $1-\delta$, are within ϵ of their corresponding true values.*

Proof: The proof has several parts.
SVD projection: We have shown in Lemma 1 that after projecting to SVD space (using a sample of size $\text{poly}\left(\frac{n}{\alpha_{\min}\epsilon}\right) \cdot \log\left(\frac{2}{\delta}\right)$), we need to estimate the parameters of the mixture in $\mathbb{R}^k$, $p_o(\boldsymbol{x}) = \sum_{i=1}^{k} \alpha_i K(\boldsymbol{x}, \boldsymbol{\nu}_i)$ where we must estimate the means within $\frac{\epsilon}{2}$ error.
Grid Search: Let us denote the parameters[4] of the underlying mixture $p_o(\boldsymbol{x}, \boldsymbol{\theta})$ by $\boldsymbol{\theta} = (\boldsymbol{m}, \boldsymbol{\alpha}) = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_k, \boldsymbol{\alpha}) \in \mathbb{R}^{k^2+k}$ and any approximating mixture $p_o(\boldsymbol{x}, \tilde{\boldsymbol{\theta}})$ has parameters $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{\alpha}})$. We have proved the bounds $f_1\left(d_H(\boldsymbol{m}, \tilde{\boldsymbol{m}})\right) \leq \|p(\boldsymbol{x}, \boldsymbol{\theta}) - p(\boldsymbol{x}, \tilde{\boldsymbol{\theta}})\| \leq f_2(d_H(\boldsymbol{m}, \tilde{\boldsymbol{m}}) + \|\boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}}\|_1)$ (see Theorem 2, Lemma 3), where f_1 and f_2 are increasing functions. Let G be the step/grid size (whose value we need to set) that we use for gridding along each of the $k^2 + k$ parameters over the grid M_G. We note that the L^2 norm of the difference can be computed efficiently by multidimensional trapezoidal rule or any other standard numerical analysis technique (see e.g., (Burden & Faires, 1993)). Since this integration needs to be preformed on a (k^2+k)-dimensional space, for any pre-specified precision parameter ϵ, this can be done in time $\left(\frac{1}{\epsilon}\right)^{O(k^2)}$. Now note that there exists a point $\boldsymbol{\theta}^* = (\boldsymbol{m}^*, \boldsymbol{\alpha}^*)$ on the grid M_G , such that if somehow we can identify this point as our parameter estimate then we make an error at most $G/2$ in estimating each mixing weight and make an error at most $G\sqrt{k}/2$ in estimating each mean. Since there are k mixing weights and k means to be estimated, $\|p_o(\boldsymbol{x}, \boldsymbol{\theta}) - p_o(\boldsymbol{x}, \boldsymbol{\theta}^*)\| \leq f_2(d_H(\boldsymbol{m}, \boldsymbol{m}^*) + \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_1) \leq f_2(G) = \frac{k\sqrt{1+k/\sigma^2}}{2(2\pi\sigma^2)^{k/2}} G$. Thus,

$$f_1\left(d_H(\boldsymbol{m}, \boldsymbol{m}^*)\right) \leq \|p_o(\boldsymbol{x}, \boldsymbol{\theta}) - p_o(\boldsymbol{x}, \boldsymbol{\theta}^*)\| \leq f_2(G)$$

Now, according to Lemma 9, using a sample of size $\Omega\left(\left[\frac{\log(2/\delta)}{\epsilon_*^2}\right]^k\right)$ we can obtain a kernel density estimate such that with probability greater than $1 - \frac{\delta}{2}$,

$$\|p_{kde} - p_o(\boldsymbol{x}, \boldsymbol{\theta})\| \leq \epsilon_* \tag{2}$$

By triangular inequality this implies,

$$f_1\left(d_H(\boldsymbol{m}, \boldsymbol{m}^*)\right) - \epsilon_* \leq \|p_{kde} - p_o(\boldsymbol{x}, \boldsymbol{\theta}^*)\| \leq f_2(G) + \epsilon_* \tag{3}$$

Since there is a one-to-one correspondence between the set of means of $\boldsymbol{m}$ and $\boldsymbol{m}^*$, $d_H(\boldsymbol{m}, \boldsymbol{m}^*)$ essentially provides the maximum estimation error for any pair of true mean and its corresponding estimate. Suppose we choose G such that it satisfies

$$2\epsilon_* + f_2(G) \leq f_1\left(\frac{\epsilon}{2}\right) \tag{4}$$

For this choice of grid size, Equation 3 and Equation 4 ensures that $f_1\left(d_H(\boldsymbol{m}, \boldsymbol{m}^*)\right) \leq f_2(G) + 2\epsilon_* \leq f_1\left(\frac{\epsilon}{2}\right)$. Hence $d_H(\boldsymbol{m}, \boldsymbol{m}^*) \leq \frac{\epsilon}{2}$. Now consider a point $\boldsymbol{\theta}^N = (\boldsymbol{m}^N, \boldsymbol{\alpha}^N)$ on the grid M_G such that $d_H(\boldsymbol{m}, \boldsymbol{m}^N) > \frac{\epsilon}{2}$. This implies,

$$f_1\left(d_H(\boldsymbol{m}, \boldsymbol{m}^N)\right) > f_1\left(\frac{\epsilon}{2}\right) \tag{5}$$

Now,

$$\begin{aligned}
\|p_o(\boldsymbol{x}, \boldsymbol{\theta}^N) - p_{kde}\| &\overset{a}{\geq} \|p_o(\boldsymbol{x}, \boldsymbol{\theta}^N) - p_o(\boldsymbol{x}, \boldsymbol{\theta})\| - \|p_o(\boldsymbol{x}, \boldsymbol{\theta}) - p_{kde}\| \\
&\overset{b}{\geq} f_1\left(d_H(\boldsymbol{m}, \boldsymbol{m}^N)\right) - \epsilon_* \\
&\overset{c}{>} f_1\left(\tfrac{\epsilon}{2}\right) - \epsilon_* \\
&\overset{d}{\geq} f_2(G) + \epsilon_* \\
&\overset{e}{\geq} \|p_o(\boldsymbol{x}, \boldsymbol{\theta}^*) - p_{kde}\|
\end{aligned}$$

[4]To make our presentation simple we assume that the single parameter variance is fixed and known. Note that it can also be estimated.

where, inequality a follows from triangular inequality, inequality b follows from Equation 2, strict inequality c follows from Equation 5, inequality d follows from Equation 4 and finally inequality e follows from Equation 3. Setting $\epsilon_* = \frac{1}{3} f_1\left(\frac{\epsilon}{2}\right)$, Equation 4 and the above strict inequality guarantees that for a choice of Grid size $G = f_2^{-1}\left(\frac{1}{3} f_1\left(\frac{\epsilon}{2}\right)\right) = \left(\frac{\alpha_{\min}^{4k}}{k^{3/2}}\right)\left(\frac{\epsilon}{8nk^2}\right)^{C_1 k^2}$ the solution obtained by equation 1 can have mean estimation error at most $\frac{\epsilon}{2}$. Once projected onto SVD space each projected mean lies within a cube $[-\sqrt{\frac{n}{k}}, \sqrt{\frac{n}{k}}]^k$. With the above chosen grid size, grid search for the means runs in time $\left(\frac{k^{3/2}}{\alpha_{\min}^{4k}}\right) \cdot \left(\frac{n^{3/2}k^{1/2}}{\epsilon}\right)^{C_1 k^2}$. Note that grid search for the mixing weights runs in time $\left(\frac{k^{3/2}}{\alpha_{\min}^{4k}}\right) \cdot \left(\frac{nk^2}{\epsilon}\right)^{C_1 k^2}$. ■

We now show that not only the mean estimates but also the mixing weights obtained by solving Equation 1 satisfy $|\alpha_i - \tilde{\alpha}_i| \leq \epsilon$ for all i. In particular we show that if two mixtures have almost same means and the L^2 norm of difference of their densities is small then the difference of the corresponding mixing weights must also be small.

Corollary 5 *With sample size and grid size as in Theorem 4, the solution of Equation 1 provides mixing weight estimates which are, with high probability, within ϵ of their true values.*

Due to space limitation we defer the proof is omitted.

3.1 Lower Bound in 1-Dimensional Setting

In this section we provide the proof of our main theoretical result in 1-dimensional setting. Before we present the actual proof, we provide high level arguments that lead us to this result. First note that Fourier transform of a mixture of k univariate Gaussians $q(x) = \sum_{i=1}^k \alpha_i K(x, \mu_i)$ is given by

$\mathcal{F}(q)(u) = \frac{1}{\sqrt{2\pi}} \int q(x) \exp(-iux) dx = \frac{1}{\sqrt{2\pi}} \sum_{j=1}^k \alpha_j \exp\left(-\frac{1}{2}(\sigma^2 u^2 + i2u\mu_j)\right)$

$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\sigma^2 u^2}{2}\right) \sum_{j=1}^k \alpha_j \exp(-iu\mu_j)$

Thus, $\|\mathcal{F}(q)\|^2 = \frac{1}{2\pi} \int |\sum_{j=1}^k \alpha_j \exp(-iu\mu_j)|^2 \exp(-\sigma^2 u^2) du$. Since L^2 norm of a function and its Fourier transform are the same, we can write,

$\|q\|^2 = \frac{1}{2\pi} \int |\sum_{j=1}^k \alpha_j \exp(-iu\mu_j)|^2 \exp(-\sigma^2 u^2) du$.

Further, $\frac{1}{2\pi} \int |\sum_{j=1}^k \alpha_j \exp(-iu\mu_j)|^2 \exp(-\sigma^2 u^2) du = \frac{1}{2\pi} \int |\sum_{j=1}^k \alpha_j \exp(iu\mu_j)|^2 \exp(-\sigma^2 u^2) du$ and we can write,

$$\|q\|^2 = \frac{1}{2\pi} \int |g(u)|^2 \exp(-\sigma^2 u^2) du$$

where $g(u) = \sum_{j=1}^k \alpha_j \exp(i\mu_j u)$. This a complex valued function of a real variable which is infinitely differentiable everywhere. In order to bound the above square norm from below, now our goal is to find an interval where $|g(u)|^2$ is bounded away from zero. In order to achieve this, we write Taylor series expansion of $g(u)$ at the origin using $(k-1)$ terms. This can be written in matrix vector multiplication format $g(u) = \boldsymbol{u^t A \alpha} + O(u^k)$, where $\boldsymbol{u^t} = [1 \; u \; \frac{u^2}{2!} \cdots \; \frac{u^{k-1}}{(k-1)!}]$, such that $\boldsymbol{A\alpha}$ captures the function value and $(k-1)$ derivative values at origin. In particular, $\|\boldsymbol{A\alpha}\|^2$ is the sum of the squares of the function g and $k-1$ derivatives at origin. Noting that $\boldsymbol{A}$ is a Vandermonde matrix we establish (see Lemma 16) $\|\boldsymbol{A\alpha}\| \geq \alpha_{\min}\left(\frac{t}{2\sqrt{n}}\right)^{k-1}$. This implies that at least one of the $(k-1)$ derivatives, say the j^{th} one, of g is bounded away from zero at origin. Once this fact is established, and noting that $(j+1)^{th}$ derivative of g is bounded from above everywhere, it is easy to show (see Lemma 14) that it is possible to find an interval $(0, a)$ where j^{th} derivative of g is bounded away from zero in this whole interval. Then using Lemma 15, it can be shown that, it is possible to find a subinterval of $(0, a)$ where the $(j-1)^{th}$ derivative of g is bounded away from zero. And thus, successively repeating this Lemma j times, it is easy to show that there exists a subinterval of $(0, a)$ where $|g|$ is bounded away from zero. Once this subinterval is found, it is easy to show that $\|q\|^2$ is lower bounded as well.

Now we present the formal statement of our result.

Theorem 6 (Lower bound in $\mathbb{R}$) *Consider a mixture of k univariate Gaussians $q(x) = \sum_{i=1}^k \alpha_i K(x, \mu_i)$ where, for all i, the mixing coefficients $\alpha_i \in (-1, 1)$ and the means $\mu_i \in [-\sqrt{n}, \sqrt{n}]$. Suppose there exists a μ_l such that $\min_j |\mu_l - \mu_j| \geq t$, and for all $i, |\alpha_i| \geq \alpha_{\min}$. Then the L^2 norm of q satisfies $||q||^2 \geq \alpha_{\min}^{2k}\left(\frac{t}{n}\right)^{Ck^2}$ where C is some positive constant independent of k.*

Proof: Note that,

$$\|q\|^2 = \frac{1}{2\pi}\int |g(u)|^2 \exp(-\sigma^2 u^2)du$$

where, $g(u) = \sum_{j=1}^{k} \alpha_j \exp(i\mu_j u)$. Thus, in order to bound the above square norm from below, we need to find an interval where $g(u)$ is bounded away from zero. Note that $g(u)$ is an infinitely differentiable function with n^{th} order derivative [5] $g^{(n)}(u) = \sum_{j=1}^{k} \alpha_j (i\mu_j)^n \exp(i\mu_j u)$. Now we can write the Taylor series expansion of $g(u)$ about origin as,

$$g(u) = g(0) + g^{(1)}(0)\frac{u}{1!} + g^{(2)}(0)\frac{u^2}{2!} + ... + g^{(k-1)}(0)\frac{u^{(k-1)}}{(k-1)!} + O(u^k)$$

which can be written as

$$g(u) = \begin{bmatrix} 1 & u & \frac{u^2}{2!} & \cdots & \frac{u^{k-1}}{(k-1)!} \end{bmatrix} \underbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ i\mu_1 & i\mu_2 & i\mu_3 & \cdots & i\mu_k \\ (i\mu_1)^2 & (i\mu_2)^2 & (i\mu_3)^2 & \cdots & (i\mu_k)^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ (i\mu_1)^{k-1} & (i\mu_2)^{k-1} & (i\mu_3)^{k-1} & \cdots & (i\mu_k)^{k-1} \end{bmatrix}}_{\boldsymbol{A}} \underbrace{\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \alpha_k \end{bmatrix}}_{\boldsymbol{\alpha}} + O(u^k)$$

Note that matrix $\boldsymbol{A}$ is Vandermonde matrix thus, using Lemma 16 this implies $|g(0)|^2 + |g^{(1)}(0)|^2 + \cdots + |g^{(k-1)}(0)|^2 = \|\boldsymbol{A}\boldsymbol{\alpha}\|^2 \geq \alpha_{\min}^2 \left(\frac{t}{1+\sqrt{n}}\right)^{2(k-1)} \geq \alpha_{\min}^2 \left(\frac{t}{2\sqrt{n}}\right)^{2(k-1)}$. This further implies that either $|g(0)|^2 \geq \frac{\alpha_{\min}^2}{k}\left(\frac{t}{2\sqrt{n}}\right)^{2(k-1)}$ or there exists a $j \in \{1, 2, \cdots, k-1\}$ such that $|g^{(j)}(0)|^2 \geq \frac{\alpha_{\min}^2}{k}\left(\frac{t}{2\sqrt{n}}\right)^{2(k-1)}$. In the worst case we can have $j = k-1$, i.e. the $(k-1)$-th derivative of g is lower bounded at origin and we need to find an interval where g itself is lower bounded.

Next, note that for any $u, g^{(k)}(u) = \sum_{j=1}^{k} \alpha_j (i\mu_j)^k \exp(iu\mu_j)$. Thus, $|g^{(k)}| \leq \sum_{j=1}^{k} |\alpha_j||(i\mu_j)^k| \leq \alpha_{\max}(\sqrt{n})^k$. Assuming $t \leq 2\sqrt{n}$, if we let $M = \frac{\alpha_{\min}}{\sqrt{k}}\left(\frac{t}{2\sqrt{n}}\right)^k$, then using Lemma 14, if we choose $a = \frac{M}{2\sqrt{2}\alpha_{\max}(\sqrt{n})^k} = \frac{\alpha_{\min}}{\alpha_{\max}2\sqrt{2k}}\left(\frac{t}{2n}\right)^k$, and thus, in the interval $[0, a], |g^{(k-1)}| > \frac{M}{2} = \frac{\alpha_{\min}}{2\sqrt{k}}\left(\frac{t}{2\sqrt{n}}\right)^k$. This implies $|Re[g^{(k-1)}]|^2 + |Im[g^{(k-1)}]|^2 > \frac{\alpha_{\min}^2}{4k}\left(\frac{t}{2\sqrt{n}}\right)^{2k}$. For simplicity denote by $h = Re[g]$, thus, $h^{(k-1)} = Re[g^{(k-1)}]$ and without loss of generality assume $|h^{(k-1)}| > \frac{\alpha_{\min}}{2\sqrt{2k}}\left(\frac{t}{2\sqrt{n}}\right)^k = \frac{M}{2\sqrt{2}}$ in the interval $(0, a)$. Now repeatedly applying Lemma 15 $(k-1)$ times yields that in the interval $\left(\frac{(3^{k-1}-1)}{3^{k-1}}a, a\right)$, (or in any other subinterval of length $\frac{a}{3^{k-1}}$ within $[0, a]$)

$|h| > \frac{M}{2\sqrt{2}}(\frac{a}{6})(\frac{a}{6.3})(\frac{a}{6.3^2})\cdots(\frac{a}{6.3^{k-1}}) = \left(\frac{M}{2\sqrt{2}}\right)\left(\frac{a}{6}\right)^k\left(\frac{1}{3^{\frac{k(k-1)}{2}}}\right) = \frac{\alpha_{\max}(\sqrt{n})^k a^{k+1}}{2^k 3^{\frac{k^2+k}{2}}}$

In particular, this implies, $|g|^2 \geq |h|^2 > \frac{\alpha_{\max}^2 n^k a^{2(k+1)}}{2^{2k}3^{k^2+k}}$ in an interval $\left(\frac{(3^{k-1}-1)}{3^{k-1}}a, a\right)$.

Next, note that $0 < a \leq 1 \Rightarrow \exp(-\sigma^2) \leq \exp(-\sigma^2 a^2)$. Now, denoting $\beta_1 = \frac{(3^{k-1}-1)}{3^{k-1}}a, \beta_2 = a$, we have,

$\|q\|^2 \geq \frac{1}{2\pi}\int_{\beta_1}^{\beta_2} |g(u)|^2 \exp(-\sigma^2 u^2)du \geq \frac{\beta_2-\beta_1}{2\pi}|g(\beta_2)|^2 \exp(-\sigma^2)$

$= \left(\frac{\exp(-\sigma^2)}{2\pi}\right)\frac{\alpha_{\max}^2 n^k a^{2k+3}}{2^{2k}3^{k^2+2k-1}} = \left(\frac{\exp(-\sigma^2)\alpha_{\min}^{2k+3}}{2\pi}\right)\left(\frac{t^{2k^2+3k}}{2^{2k^2+5k+9/2}3^{k^2+2k-1}(\alpha_{\max})^{2k+1}k^{k+3/2}n^{2k^2+2k}}\right)$

$\geq \left(\frac{\exp(-\sigma^2)\alpha_{\min}^{2k+3}}{2\pi}\right)\left(\frac{t^{2k^2+3k}}{2^{2k^2+5k+9/2}3^{k^2+2k-1}k^{k+3/2}n^{2k^2+2k}}\right)$

$\geq \alpha_{\min}^{2k}\left(\frac{t^{2k^2+3k}}{2^{O(k^2\log n)}}\right) = \alpha_{\min}^{2k}\left(\frac{t}{n}\right)^{O(k^2)}$

where the last inequality follows from the fact that if we let,

$F(k) = 2^{2k^2+5k+9/2}3^{k^2+2k-1}k^{k+3/2}n^{k^2+2k}$ then taking log with base 2 on both sides yields,

$\log(F(k)) = (2k^2+5k+9/2) + (k^2+2k-1)\log 3 + (k+3/2)\log k + (2k^2+2k)\log n = O(k^2 \log n)$.

Thus, $F(k) = 2^{O(k^2\log n)} = n^{O(k^2)}$. ■

[5] Note that Fourier transform is closely related to the characteristics function and the n^{th} derivative of g at origin is related to the n^{th} order moment of the mixture in the Fourier domain.

3.2 Determinant of Vandermonde Like Matrices

In this section we derive a result for the determinant of a Vandermonde-like matrix. This result will be useful in finding the angle made by any column of a Vandermonde matrix to the space spanned by the rest of the columns and will be useful in deriving the lower bound in Theorem 6.

Consider any $(n+1) \times n$ matrix B of the form

$$B = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_n^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_1^n & x_2^n & x_3^n & \cdots & x_n^n \end{bmatrix}$$

If the last row is removed then it exactly becomes an $n \times n$ Vandermonde matrix having determinant $\Pi_{i>j}(x_i - x_j)$. The interesting fact is that if any other row except the last one is removed then the corresponding $n \times n$ matrix has a structure very similar to that of a Vandermonde matrix. The following result shows how the determinants of such matrices are related to $\Pi_{i>j}(x_i - x_j)$.

Lemma 7 *For $1 \leq i \leq (n-1)$, let B_i represents the $n \times n$ matrix obtained by removing the i^{th} row from B. Then $\det(B_i) = c_i \Pi_{s>t}(x_s - x_t)$ where c_i is a polynomial having $\binom{n}{i-1}$ terms with each term having degree $(n-i+1)$. Terms of the polynomial c_i represent the possible ways in which $(n-i+1)$ x_js can be chosen from $\{x_i\}_{i=1}^n$.*

Proof: First note that if a matrix has elements that are monomials in some set of variables, then its determinant will in general be polynomial in those variables. Next, by the basic property of a determinant, that it is zero if two of its columns are same, we can deduce that for $1 \leq i < n, \det(B_i) = 0$ if $x_s = x_t$ for some $s \neq t, 1 \leq s, t < n$, and hence $q_i(x_1, x_2, ..., x_n) = \det(B_i)$ contains a factor $p(x_1, x_2, ..., x_n) = \Pi_{s>t}(x_s - x_t)$. Let $q_i(x_1, x_2, ..., x_n) = p(x_1, x_2, ..., x_n) r_i(x_1, x_2, ..., x_n)$.

Now, note that each term of $p(x_1, x_2, ..., x_n)$ has degree $0+1+2+...+(n-1) = \frac{n(n-1)}{2}$. Similarly, each term of the polynomial $q_i(x_1, x_2, ..., x_n)$ has degree $(0+1+2+...+n)-(i-1) = \frac{n(n+1)}{2} - (i-1)$. Hence each term of the polynomial $r_i(x_1, x_2, ..., x_n)$ must be of degree $\frac{n(n+1)}{2} - (i-1) - \frac{n(n-1)}{2} = (n-i+1)$. However in each term of $r_i(x_1, x_2, ..., x_n)$, the maximum power of any x_j can not be greater than 1. This follows from the fact that maximum power of x_j in any term of $q_i(x_1, x_2, ..., x_n)$ is n and in any term of $p(x_1, x_2, ..., x_n)$ is $(n-1)$. Hence each term of $r_i(x_1, x_2, ..., x_n)$ consists of $(n-i+1)$ different x_js and represents the different ways in which $(n-i+1)$ x_js can be chosen from $\{x_i\}_{i=1}^n$. And since it can be done in $\binom{n}{n-i+1} = \binom{n}{i-1}$ ways there will be $\binom{n}{i-1}$ terms in $r_i(x_1, x_2, ..., x_n)$. ■

3.3 Estimation of Unknown Variance

In this section we discuss a procedure for consistent estimation of the unknown variance due to (Lindsay, 1989) (for the one-dimensional case) and prove that the estimate is polynomial. This estimated variance can then be used in place of true variance in our main algorithm discussed earlier and the remaining mixture parameters can be estimated subsequently.

We start by noting a mixture of k identical spherical Gaussians $\sum_{i=1}^k \alpha_i \mathcal{N}(\boldsymbol{\mu}_i, \sigma^2 I)$ in $\mathbb{R}^n$ projected on an arbitrary line becomes a mixture of identical 1-dimensional Gaussians $p(x) = \sum_{i=1}^k \alpha_i \mathcal{N}(\mu_i, \sigma^2)$. While the means of components may no longer be different, the variance does not change. Thus, the problem is easily reduced to the 1-dimensional case.

We will now show that the variance of a mixture of k Gaussians in 1 dimension can be estimated from a sample of size $\text{poly}\left(\frac{1}{\epsilon}, \frac{1}{\delta}\right)$, where $\epsilon > 0$ is the precision ,with probability $1-\delta$ in time $\text{poly}\left(\frac{1}{\epsilon}, \frac{1}{\delta}\right)$. This will lead to an estimate for the n-dimensional mixture using $\text{poly}\left(n, \frac{1}{\epsilon}, \frac{1}{\delta}\right)$ sample points/operations.

Consider now the set of Hermite polynomials $\gamma_i(x, \tau)$ given by the recurrence relation $\gamma_i(x, \tau) = x\gamma_{i-1}(x, \tau) - (i-1)\tau^2 \gamma_{i-2}(x, \tau)$, where $\gamma_0(x, \tau) = 1$ and $\gamma_1(x, \tau) = x$. Take M to be the $(k+1) \times (k+1)$ matrix defined by

$$M_{ij} = \mathbb{E}_p[\gamma_{i+j}(X, \tau)], \;\; 0 \leq i+j \leq 2k.$$

It is shown in Lemma 5A of (Lindsay, 1989) that the determinant $\det(M)$ is a polynomial in τ and, moreover, that the smallest positive root of $\det(M)$, viewed is a function of τ, is equal to the variance σ of the original mixture p. We will use $d(\tau)$ to represent $\det(M)$.

This result leads to an estimation procedure, after observing that $\mathbb{E}_p[\gamma_{i+j}(X, \tau)]$ can be replaced by its empirical value given a sample $X_1, X_2, ..., X_N$ from the mixture distribution p. Indeed, one can construct

the empirical version of the matrix M by putting

$$\hat{M}_{ij} = \frac{1}{N}\sum_{t=1}^{N}[\gamma_{i+j}(X_t,\tau)],\ \ 0 \le i+j \le 2k. \tag{6}$$

It is clear that $\hat{d}(\tau) = \det(\hat{M})(\tau)$ is a polynomial in τ. Thus we can provide an estimate σ^* for the variance σ by taking the smallest positive root of $\hat{d}(\tau)$. This leads to the following estimation procedure :

Parameter: Number of components k.
Input: N points in $\mathbb{R}^n$ sampled from $\sum_{i=1}^{k}\alpha_i\mathcal{N}(\boldsymbol{\mu}_i,\sigma^2 I)$.
Output: σ^*, estimate of the unknown variance.

Step 1. Select an arbitrary direction $\boldsymbol{v} \in \mathbb{R}^n$ and project the data points onto this direction.

Step 2. Construct the $(k+1)\times(k+1)$ matrix $\hat{M}(\tau)$ using Eq. 6

Step 3. Compute the polynomial $\hat{d}(\tau) = \det(\hat{M})(\tau)$. Obtain the estimated variance σ^* by approximating the smallest positive root of $\hat{d}(\tau)$. This can be done efficiently by using any standard numerical method or even a grid search.

We will now state our main result in this section, which establishes that this algorithm for variance estimation is indeed polynomial in both the ambient dimension n and the inverse of the desired accuracy ϵ.

Theorem 8 *For any $\epsilon > 0, 0 < \delta < 1$, if sample size $N > O\left(\frac{n^{\mathrm{poly}(k)}}{\epsilon^2\delta}\right)$, then the above procedure provides an estimate σ^* of the unknown variance σ such that $|\sigma - \sigma^*| \le \epsilon$ with probability greater than $1-\delta$.*

The idea of the proof is to show that the coefficients of the polynomials $d(\tau)$ and $\hat{d}(\tau)$ are polynomially close, given enough samples from p. That (under some additional technical conditions) can be shown to imply that the smallest positive roots of these polynomials are also close. To verify that $d(\tau)$ and $\hat{d}(\tau)$ are close, we use the fact that the coefficients of $d(\tau)$ are polynomial functions of the first $2k$ moments of p, while coefficients of $\hat{d}(\tau)$ are the same functions of the empirical moment estimates. Using standard concentration inequalities for the first $2k$ moments and providing a bound for these functions the result.

Proof: It is shown in Lemma 5A of (Lindsay, 1989) that the smallest positive root of the determinant $d(\tau) = \det(M)(\tau)$, viewed is a function of τ, is equal to the variance σ of the original mixture p and also that $d(\tau)$ undergoes a sign change at its smallest positive root. Let the smallest positive root of $\hat{d}(\tau) = \det(\hat{M})(\tau)$ be $\hat{\sigma}$. We now show for any $\epsilon > 0$ that σ and $\hat{\sigma}$ are within ϵ given $O\left(\frac{n^{\mathrm{poly}(k)}}{\epsilon^2\delta}\right)$ samples.

In Corollary 18 we show that both $d(\tau)$ and $\hat{d}(\tau)$ are polynomials of degree $k(k+1)$ and the highest degree coefficient of $\hat{d}(\tau)$ is independent of the sample. The rest of the coefficients of $d(\tau)$ and $\hat{d}(\tau)$ are sums of products of the coefficients of individual entries of the matrices M and $\hat{M}$ respectively.

Note that $\mathbb{E}(\hat{M}) = M$, i.e., for any $1 \le i,j, \le (k+1), \mathbb{E}(\hat{M}_{i,j}(\tau)) = M_{i,j}(\tau)$. Since $M_{i,j}(\tau)$ is a polynomial in τ, using standard concentration results we can show that coefficients of the polynomial $\hat{M}_{i,j}(\tau)$ are close to the corresponding coefficients of the polynomial $M_{i,j}(\tau)$ given large enough sample size. Specifically, we show in Lemma 21 that given a sample of size $O\left(\frac{n^{\mathrm{poly}(k)}}{\epsilon^2\delta}\right)$ each of the coefficients of each of the polynomials $M_{i,j}(\tau)$ can be estimated within error $O\left(\frac{\epsilon}{n^{\mathrm{poly}(k)}}\right)$ with probability at least $1-\delta$.

Next, in Lemma 22 we show that estimating each of the coefficients of the polynomial $M_{i,j}(\tau)$ for all i,j with accuracy $O\left(\frac{\epsilon}{n^{\mathrm{poly}(k)}}\right)$ ensures that all coefficients of $d(\hat{\tau})$ are $O\left(\frac{\epsilon}{k}\right)$ close to the corresponding coefficients of $d(\tau)$ with high probability.

Consequently, in Lemma 20 we show that when all coefficients of $\hat{d}(\tau)$ are within $O\left(\frac{\epsilon}{k}\right)$ of the corresponding coefficients of $d(\tau)$, the smallest positive root of $\hat{d}(\tau)$, $\hat{\sigma}$, is at most ϵ away from the smallest positive root σ of $d(\tau)$.

Observing that there exist many efficient numerical methods for estimating roots of polynomial of one variable within the desires accuracy completes the proof. ■

References

Achlioptas, D., & McSherry, F. (2005). On Spectral Learning of Mixture of Distributions. *The 18th Annual Conference on Learning Theory*.

Arora, S., & Kannan, R. (2001). Learning Mixtures of Arbitrary Gaussians. *33rd ACM Symposium on Theory of Computing*.

Brubaker, S. C., & Vempala, S. (2008). Isotropic pca and affine-invariant clustering. *49th Annual Symposium on Foundations of Computer Science*.

Burden, R. L., & Faires, J. D. (1993). *Numerical Analysis*. Harcourt Trade Publisher.

Chaudhuri, K., & Rao, S. (2008a). Beyond Gaussians: Spectral Methods for Learning Mixtures of Heavy Tailed Distributions. *The 21st Annual Conference on Learning Theory*.

Chaudhuri, K., & Rao, S. (2008b). Learning Mixtures of Product Distributions Using Correlations and Independence. *The 21st Annual Conference on Learning Theory*.

Chen, H. (1995). Optimal Rate of Convergence for Finite Mixture Models. *The Annals of Statistics*, *23(1)*, 221–233.

Dasgupta, A., Hopcroft, J. E., Kleinberg, J., & Sandler, M. (2005). On Learning Mixture of Heavy Tailed Distributions. *46th Annual Symposium on Foundations of Computer Science*.

Dasgupta, S. (1999). Learning Mixture of Gaussians. *40th Annual Symposium on Foundations of Computer Science*.

Dasgupta, S., & Schulman, L. (2000). A Two Round Variant of EM for Gaussian Mixtures. *16th Conference on Uncertainty in Artificial Intelligence*.

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions*. Chapman & Hall.

Feldman, J., D'Onnell, O., & Servedio, R. (2008). Learning Mixture of Product Distributions over Discrete Domains. *SIAM Journal on Computing*, *37(5)*, 1536–1564.

Feldman, J., Servedio, R. A., & O'Donnell, R. (2006). PAC Learning Mixture of Axis Aligned Gaussians with No Separation Assumption. *The 19th Annual Conference on Learning Theory*.

Kalai, A. T., Moitra, A., & Valiant, G. (2010). Efficiently Learning Mixture of Two Gaussians. *42nd ACM Symposium on Theory of Computing*.

Kannan, R., Salmasian, H., & Vempala, S. (2005). The Spectral Method for General Mixture Models. *The 18th Annual Conference on Learning Theory*.

Lindsay, B. G. (1989). Moment Matrices: Applications in Mixtures. *The Annals of Statistics*, *17(2)*, 722–740.

Lindsay, B. G. (1995). *Mixture Models: Theory Geometry and Applications*. Institute of Mathematical Statistics.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. Marcel Dekker.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. John Wiley & Sons.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Phil. Trans. Royal Soc.*, 71110.

Titterington, D., Smith, A., & Makov, U. (1985). *Statistical analysis of finite mixture distributions*. John Wiley & Sons.

Vempala, S., & Wang, G. (2002). A Spectral Algorithm for Learning Mixture of Distributions. *43rd Annual Symposium on Foundations of Computer Science*.

A Appendix

In this appendix we provide some of the auxiliary lemmas that are required in the main text. Due to space limitation the proofs are omitted. Extended version of this paper containing the proofs is available at `http://arxiv.org/abs/0907.1054`.
For the purpose of estimating the sample size requirement for an appropriate non-parametric density estimator that is arbitrarily close in L^2 norm sense, we define the Sobolev class as follows. In the following, Sobolev space $W^{2,2}$ is defined as the subset of L^2 such that if $f \in W^{2,2}$ then f and its weak derivatives up to order 2 have finite L^2 norm.

Definition 1 *Let $L > 0$. The Sobolev class $\mathcal{S}(2, L)$ is defined as the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $f \in W^{2,2}$, and all the second order partial derivatives $\frac{\partial^2 f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$, where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ is a multi-index with $|\alpha| = 2$, satisfy*

$$\left\| \frac{\partial^2 f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} \right\|_2 \leq L$$

Note that when the parameters are bounded, mixture of k Gaussian distributions belongs to Sobolev class as defined above and the following Lemma shows that we can approximate the density of such a mixture arbitrarily well in L^2 norm sense.

Lemma 9 *Let $p \in \mathcal{S}(2, L)$ be a d-dimensional probability density function and $K : \mathbb{R}^d \to \mathbb{R}$ be any kernel function with diagonal bandwidth matrix $h^2\boldsymbol{I}, h > 0$, satisfying $\int K(\boldsymbol{x})d\boldsymbol{x} = 1, \int \boldsymbol{x}K(\boldsymbol{x})d\boldsymbol{x} = \boldsymbol{0}, \int \boldsymbol{x}^T\boldsymbol{x}K(\boldsymbol{x})d\boldsymbol{x} < C_1$ and $\int K^2(\boldsymbol{x})d\boldsymbol{x} < C_2$ for positive C_1, C_2. Then for any $\epsilon_0 > 0$ and any $\delta \in (0, 1)$, with probability grater than $1 - \delta$, the kernel density estimate $\hat{p}_S$ obtained using a sample S of size $\Omega\left(\left[\frac{\log(1/\delta)}{\epsilon_0^2}\right]^d\right)$ satisfies, $\int \left(p(\boldsymbol{x}) - \hat{p}_S(\boldsymbol{x})\right)^2 d\boldsymbol{x} \leq \epsilon_0$.*

Lemma 10 *Consider any set of k points $\{\boldsymbol{x}_i\}_{i=1}^k$ in $\mathbb{R}^d$. There exists a direction $\boldsymbol{v} \in \mathbb{R}^d, \|\boldsymbol{v}\| = 1$ such for any i, j $|\langle \boldsymbol{x}_i, \boldsymbol{v}\rangle - \langle \boldsymbol{x}_j, \boldsymbol{v}\rangle| > \frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}{k^2}$.*

Note that in the above Lemma dimensionality of the space $\mathbb{R}^d$ is irrelevant but the number of samples k is important. This Lemma can also be considered as a special kind of one sided version of Johnson-Lindenstraus Lemma, and by choosing $\boldsymbol{v}$ at random from $\mathbb{R}^d$ the same result can be shown to hold with high probability. However, the above result is deterministic.

Lemma 11 *Let $g : \mathbb{R}^k \to \mathbb{R}$ be a continuous bounded function. Let $\boldsymbol{v}, \boldsymbol{u}_1, ..., \boldsymbol{u}_{k-1} \in \mathbb{R}^k$ be an orthonormal basis of $\mathbb{R}^k$ and let $g_1 : \mathbb{R} \to \mathbb{R}$ be defined as $g_1(v) = \int \cdots \int g(v, u_1, ...u_{k-1})du_1 \cdots du_{k-1}$. Then for some $c > 0, \|g\|^2 \geq \left(\frac{1}{c\sigma}\right)^k \|g_1\|^2$.*

A version of the following Lemma was proved in (Vempala & Wang, 2002). We tailor it for our purpose.

Lemma 12 *Let the rows of $A \in \mathbb{R}^{N \times n}$ be picked according to a mixture of Gaussians with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^n$, common variance σ^2 and mixing weights $\alpha_1, \alpha_2, \ldots, \alpha_k$ with minimum mixing weight being $\alpha_{\min}$. Let $\tilde{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\mu}}_2, \ldots, \tilde{\boldsymbol{\mu}}_k$ be the projections of these means on to the subspace spanned by the top k right singular vectors of the sample matrix A. Then for any $0 < \epsilon < 1, 0 < \delta < 1$, with probability at least $1 - \delta$, $\|\boldsymbol{\mu}_i - \tilde{\boldsymbol{\mu}}_i\| \leq \frac{\epsilon}{2}$, provided $N = \Omega\left(\frac{n^3\sigma^4}{\alpha_{min}^3\epsilon^4}\left(\log\left(\frac{n\sigma}{\epsilon\alpha_{min}}\right) + \frac{1}{n(n-k)}\log(\frac{1}{\delta})\right)\right)$,*

In the following Lemma we consider a mixture of Gaussians where the mixing weights are allowed to take negative values. This might sound counter intuitive since mixture of Gaussians are never allowed to take negative mixing weights. However, if we have two separate mixtures, for example, one true mixture density $p(\boldsymbol{x})$ and one its estimate $\hat{p}(\boldsymbol{x})$, the function $(p - \hat{p})(\boldsymbol{x})$ that describes the difference between the two densities can be thought of as a mixture of Gaussians with negative coefficients. Our goal is to find a bound of the L^2 norm of such a function.

Lemma 13 *Consider a mixture of m k-dimensional Gaussians $f(\boldsymbol{x}) = \sum_{i=1}^m \alpha_i K(\boldsymbol{x}, \boldsymbol{\nu}_i)$ where the mixing coefficients $\alpha_i \in (-1, 1),\ i = 1, 2, \ldots, m$. Then the L^2 norm of f satisfies $\|f\|^2 \leq \left(\frac{1}{(2\pi\sigma^2)^k}\right)\boldsymbol{\alpha}^T\hat{\boldsymbol{K}}\boldsymbol{\alpha}$, where $\hat{\boldsymbol{K}}$ is a $m \times m$ matrix with $\hat{\boldsymbol{K}}_{ij} = \exp\left(-\frac{\|\boldsymbol{\nu}_i - \boldsymbol{\nu}_j\|^2}{2\sigma^2}\right)$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m)^T$.*

Lemma 14 *Let $h : \mathbb{R} \to \mathbb{C}$ be an infinitely differentiable function such that for some positive integer n and real $M, T > 0, |h^{(n)}(0)| > M$ and $|h^{(n+1)}| < T$. Then for any $0 < a < \frac{M}{T\sqrt{2}}$, $|h^{(n)}| > M - \sqrt{2}Ta$ in the interval $[0, a]$.*

Lemma 15 *Let $h : \mathbb{R} \to \mathbb{R}$ be an infinitely differentiable function such that for some positive integer n and real $M > 0, |h^{(n)}| > M$ in an interval (a, b). Then $|h^{(n-1)}| > M(b-a)/6$ in a smaller interval either in $(a, \frac{2a+b}{3})$ or in $(\frac{a+2b}{3}, b)$.*

Let A be a $k \times k$ Vandermonde matrix defined as follows.

$$A = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_k \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_k^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_1^{k-1} & x_2^{k-1} & x_3^{k-1} & \cdots & x_k^{k-1} \end{bmatrix}$$

Then we can prove the following result.

Lemma 16 *For any integer $k > 1$, and positive $a, t \in \mathbb{R}$, let $x_1, x_2, ..., x_k \in [-a, a]$ and there exists an x_i such that $t = \min_{j, j \neq i} |x_i - x_j|$. Let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k) \in \mathbb{R}^k$ with $\min_i |\alpha_i| \geq \alpha_{\min}$. Then for A as defined above, $\|A\boldsymbol{\alpha}\| \geq \alpha_{\min} \left(\frac{t}{1+a}\right)^{k-1}$.*

Lemma 17 *Consider the $(k+1) \times (k+1)$ Hankel matrix Γ, $\Gamma_{ij} = (\gamma_{i+j}(x, \tau))$ for $i, j = 0, 1, ..., k$, where $\gamma_n(x, \tau)$ is the n^{th} Hermite polynomial as described above. Then $\det(\Gamma)(x, \tau)$ is a homogeneous polynomial of degree $k(k+1)$ of two variables x and τ.*

Using the above Lemma, we have the following simple corollary.

Corollary 18 *$d(\tau)$ is a polynomial of of degree $k(k+1)$, with the coefficient of the leading term independent of the probability distribution p. Similarly, $\hat{d}(\tau)$ is a polynomial of of degree $k(k+1)$, with the leading term having coefficient independent of the coefficients of the sampled data.*

Lemma 19 *Let $f(x) = x^m + a_{m-1}x^{m-1} + a_{m-2}x^{m-2} + \cdots + a_1 x + a_0$ be a polynomial having a smallest positive real root x_0 with multiplicity one and $f'(x_0) \neq 0$. Let $\hat{f}(x) = x^m + \hat{a}_{m-1}x^{m-1} + \hat{a}_{m-2}x^{m-2} + \cdots + \hat{a}_1 x + \hat{a}_0$ be another polynomial such that $\|a - \hat{a}\| \leq \epsilon$ for some sufficiently small $\epsilon > 0$, where $a = (a_0, a_1, \ldots, a_{m-1})$ and $\hat{a} = (\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_{m-1})$. Then there exists a $C > 0$ such that the smallest positive root $\hat{x}_0$ of $\hat{f}(x)$ satisfies $\|x_0 - \hat{x}_0\| \leq C\epsilon$.*

Lemma 20 *Let σ be the smallest positive root of $d(\tau)$. Suppose $\hat{d}(\tau)$ be the polynomial where each of the coefficients of $d(\tau)$ are estimated within ϵ error for some sufficiently small $\epsilon > 0$. Let $\hat{\sigma}$ be the smallest positive root of $\hat{d}(\tau)$. Then $|\hat{\sigma} - \sigma| = O(k\epsilon)$.*

The matrix M defined in Section 3.3 has $2k$ different entries which is clear from its construction. Each such entry is a polynomial in τ. Let us denote these distinct entries by $m_i(\tau) = \mathbb{E}[\gamma_i(x, \tau)], i = 1, 2, \ldots, 2k$. Note that the empirical version of the matrix M is $\hat{M}$ where each entry $m_i(\tau)$ is replaced by its empirical counterpart $\hat{m}_i(\tau)$. Using standard concentration inequality we show that for any $m_i(\tau)$, its coefficients are arbitrarily close to the corresponding coefficients of $\hat{m}_i(\tau)$ provided a large enough sample size is used to estimate $\hat{m}_i(\tau)$.

Lemma 21 *For any $m_i(\tau), i = 1, , 2, \ldots, 2k$, let β be any arbitrary coefficient of the polynomial $m_i(\tau)$. Suppose $X_1, X_2, \ldots, X_N$ iid samples from p is used to estimate $\hat{m}_i(\tau)$ and the corresponding coefficient is $\hat{\beta}$. Then there exists a polynomial $\eta_1(k)$ such that for any $\epsilon > 0$ and $0 < \delta, |\beta - \hat{\beta}| \leq \epsilon$ with probability at least $1 - \delta$, provided $N > \frac{n^{\eta_1(k)}}{\epsilon^2 \delta}$.*

Lemma 22 *There exists a polynomial $\eta_2(k)$ such that if coefficients of each of the entries of matrix M (where each such entry is a polynomial of τ) are estimated within error $\frac{\epsilon}{n^{\eta_2(k)}}$ then each of the coefficients of $d(\tau)$ are estimated within ϵ error.*