On Gaussian Mixture Models

THIS IS A LONGER VERSION OF THE PAPER. IT IS INCLUDED TO HELP WITH REVIEW PROCESS.

N/A

Abstract. We investigate the sample complexity of Gaussian mixture models (GMMs). Our results provide the optimal upper bound, in the context of uniform spherical Gaussian mixtures. Furthermore, we highlight the relationship between the sample complexity of GMMs and the distribution of spacings among their means.

Gaussian Mixture Models have been extensively studied in the field of machine learning. Let $\mathcal{N}$ represent the Probability Density Function (PDF) of a multidimensional Gaussian distribution characterized by a mean μ and a covariance matrix Σ. A Gaussian Mixture Model is defined as follows:

$$\Gamma = \sum_{i=1}^{k} \omega_i \mathcal{N}(\mu_i, \Sigma_i), \tag{0.1}$$

where $\sum \omega_i = 1$.

Assume that we are given a sequence of i.i.d. samples, $x_1, \cdots, x_n$, generated from Γ. An important area of research on GMM's is determining the optimal number of samples required to learn the underlying mixture.

Learning the mixture can be divided into two distinct approaches. The first approach aims to yield a distribution that closely approximates Γ, in terms of total variation [2]. However, this approach doesn't inherently offer insights into the specific parameters themselves. The second approach involves the estimation of each individual parameter of Γ, namely ω_i, μ_i, and Σ_i, with a "good precision". To elaborate, let's clarify the term "good precision".

Definition 1. *For a given Γ as in* (0.1)*, an algorithm is said to learn the parameters of Γ with ϵ-precision if, for every μ_i, Σ_i, and ω_i in* (0.1)*, it produces $\hat{\mu}_i, \hat{\Sigma}_i$, and $\hat{\omega}_i$ such that $|\mu_i - \hat{\mu}_i| < \epsilon$, $|\Sigma_i - \hat{\Sigma}_i| < \epsilon$, and $|\omega_i - \hat{\omega}_i| < \epsilon$.*

Parameter estimation of a GMM dates back to the 18th century and was initially introduced by Pearson[12]. Subsequently it has been studied in various aspects [9, 4, 3, 6, 7, 8]. If the parameters of the mixture all well separated, it can be shown that around ϵ^{-2} samples are sufficient to estimate parameters with ϵ-precision [9, 3]. Despite the rich body of work on this problem there are not many results that link the parameter distribution to sample complexity. This constitutes a central theme of the present paper.

Our work is inspired by the paper of Hardt and Price [9]. They explored the case of $k = 2$ and demonstrated that ϵ^{-12} samples are both necessary and sufficient for determining the parameters of a 2-mixture with ϵ-precision. If the means are well-separated, merely ϵ^{-2} samples are enough.

For $k = 2$, there are 5 parameters to learn. Exponent 12 suggests that we need a factor of ϵ^{-2} samples to distinguish between each of these parameters. When this is done, with an extra ϵ^{-2} samples we can reach ϵ-accuracy.

For arbitrary k, there are $3k-1$ parameters to learn. Hardt and Price [9], considered a k-mixture with $3k-2$ parameters [1], and showed that at least ϵ^{-6k+2} samples are necessary for this case. This leads us to anticipate the optimal lower bound, for the general case, to be ϵ^{-6k}, which raises the following question: are ϵ^{-6k} i.i.d. samples adequate for determining the parameters of a GMM with ϵ-precision?

Our Contribution. We answer the above question affirmatively, in the uniform spherical case, where $\omega_i = 1/k$ and $\sigma_i = 1$. Further, we establish a connection between pair correlation of the means of the mixture (μ_i's) and the sample complexity of learning its parameters. In essence, we prove that the distribution of spacing between the μ_i's greatly influences the sample complexity. This connection is a novel part of this paper.

Note that in the uniform spherical case with mean zero, we have $k-1$ parameters to recover.Therefore, we expect ϵ^{-2k} samples to suffice.

Corollary 0.1. *Let Γ be a uniform spherical mixture of k Gaussians distribution with mean equal to zero and variance equal to σ^2. Assume that $\epsilon < \min_{i \neq j}(|\mu_i - \mu_j|)$, and let $c(\sigma, k) = \left|k^2(1+k\sigma^2)^{k/2}(2\sigma)^k e^{0.5(k/\sigma)^2}\right|^2$. Then*

$$10^4 c(\sigma, k)\epsilon^{-2k}$$

samples are sufficient to learn parameters of Γ with $\epsilon/100$-accuracy.

Corollary 0.1 shows that in the worst case scenario, we requires ϵ^{-2k} samples to learn parameters with ϵ-accuracy. However, this is not always the case. There are many instances where we can learn parameters of a GMM with fewer samples. Next, we demonstrate how the sample complexity is contingent upon the distribution of spacing between the mixture's means.

0.1. **Pair correlation of means and sample complexity.** Hardt and Price [9] have shown that if $|\mu_1 - \mu_2| = \Omega(\sigma)$, roughly ϵ^{-2} samples are sufficient for determining the parameters of a 2-mixture with ϵ-precision. In general, without this separation, ϵ^{-12} samples would be required. This implies that pair correlations of means can impact the sample complexity. We expand on this phenomenon for general k.

[1] They assume that at least two of the parameters cannot be very close to each other, which saves a factor of ϵ^{-2} samples. See [9] the paragraph before Theorem 2.10.

To explain the pair correlation's role, assume that we have gaps of length ϵ between consecutive means in our mixture. However, these gaps are isolated, meaning that if $\mu_{n+1}-\mu_n = \epsilon$, then the adjacent gaps are significantly larger: $\mu_{n+2} - \mu_{n+1} \gg 1$, and $\mu_n - \mu_{n-1} \gg 1$. Further assume that we are aiming to learn μ_i's with $\epsilon/100$-accuracy. In this situation our result shows having roughly $c(k, \sigma)\epsilon^{-4}$ samples will suffice.

Now let us alter the situation by allowing two consecutive gaps of length ϵ in our mixture .i.e. $\mu_{n+1} - \mu_n = \epsilon$, and $\mu_n - \mu_{n-1} = \epsilon$. As before, we assume that other surrounding gaps remain significantly larger: $\mu_{n+2} - \mu_{n+1} \gg 1$, and $\mu_{n-1} - \mu_{n-2} \gg 1$. Under this modified scenario, our required sample complexity swells to $c(k, \sigma)\epsilon^{-6}$.

Let us define a function that captures the essence of the pair correlation for our purpose. Let

$$\mathcal{P}(\mu_m) = \prod_{\substack{n=1 \\ n\neq m}}^{k} |\mu_m - \mu_n|. \tag{0.2}$$

When an ϵ-vicinity of μ_m contains l means, then $\mathcal{P}(\mu_m) \approx \epsilon^l$. Our next corollary shows the maximum number of means that are located in an interval of length about ϵ determines the sample complexity.

Corollary 0.2. *Let Γ be a uniform spherical mixture of k Gaussians distribution with mean equal to zero and variance equal to σ^2. Then, given*

$$\log\big(\frac{1}{\delta}\big)\frac{c(\sigma, k)}{\big|\min_m \mathcal{P}(\mu_m)\big|^2}\,\epsilon^{-2}, \tag{0.3}$$

samples from Γ, where $c(\sigma, k) = \big|k^2(1+k\sigma^2)^{k/2}(2\sigma)^k e^{0.5(k/\sigma)^2}\big|^2$, with probability $1-\delta$, we can approximate Γ's parameters with ϵ-precision.

Next we state our main theorem.

Theorem 0.3. *Let $\Gamma = \frac{1}{k}\sum_{m=1}^{k}\mathcal{N}(\mu_m, 1)$, where its mean is zero and variance is σ^2. Assume that we are given $\log\big(\frac{1}{\delta}\big)\epsilon^{-2}$ samples from Γ with*

$$\epsilon < \frac{\mathcal{P}(\mu_m)}{k^2(1+|\mu_m|^2)^{(k-1)/2}(2\sigma)^k e^{0.5(k/\sigma)^2}}.$$

Then our algorithm returns $\hat{\mu}_m$ such that

$$|\hat{\mu}_m - \mu_m| < \frac{k^2(1+|\mu_m|^2)^{k/2}(2\sigma)^k e^{0.5(k/\sigma)^2}}{\mathcal{P}(\mu_m)}\epsilon. \tag{0.4}$$

0.2. **Related work.** We have already mentioned the work of Hardt and Price [9]. Parallels to our approach, they too employ the method of moments. Their scope, however, is constrained to mixtures of two Gaussians.

Similarly to our work, a main aspect of their work revolves around the correlation between the separation of parameters and sample complexity.

A substantial separation between parameters intuitively simplifies the task of clustering samples. In this genre, for the Spherical GMM, Liu and Li [3] proposed an algorithm that operates within a $poly(d, k)$ time frame. Their result is contingent on the condition:

$$|\mu_i - \mu_j| > (\log k)^{1/2+\varepsilon}$$

where d signifies the dimension. It's noteworthy that when the separation between parameters exceeds 1, our algorithm attains ϵ-accuracy with approximately ϵ^{-2} samples.

For scenarios where $d < \log k$, Qiao et al. [4] further refined the results of Liu and Li. Their sample complexity though, with respect to ϵ-precision, is quasi-polynomial i.e., $\epsilon^{-c\log(1/\epsilon)}$.

What sets our method apart is the absence of assumptions regarding parameter separation, and the optimal bound on sample complexity subject to accuracy. Additionally, we clearly describe the circumstances under which more than ϵ^{-2} samples—a theoretical minimum—are required.

0.3. **Proof Overview.** Let P be a polynomial of degree k with roots corresponding to the parameters of our mixture; we'll refer to P as our "parameter polynomial". We aim to approximate coefficients of P with a good precision and relate this approximation to the number of samples we use.

We employ the method of moments combined with Newton's identity to derive a polynomial whose coefficients are close to the coefficients of P. In practice, since we are using samples to calculate the moments of our mixture, we need to consider the empirical error. Therefore, we need to have a sense of how much of this error would spill into our approximation of the parameter polynomial. Following this process we end up with a polynomial whose coefficients are within a certain distance to P. This distance is proportional to the number of samples we use to calculate empirical moments.

Lastly, we must establish an argument that connects the similarity between the coefficients of two polynomials to the difference between their roots. Here we use a theorem from real analysis which measures this relation.

1. The method of Moments

One way of approaching the problem of parameters estimation of GMM's, is through the method of moments, see [6, 7, 9]. For a one dimensional Gaussian we have the r-th moment equals

$$\mathcal{M}_r(\mu, \sigma) := \frac{1}{\sigma\sqrt{2\pi}} \int x^r e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{1.1}$$

which is a polynomial in terms of μ, σ and easily calculable.

For a k-mixture, to simplify the problem, assume that $\omega_i = 1/k$, and consider the one dimensional case that covariance matrices $\Sigma_i = \sigma_i$ are real numbers. Therefore, moments of the k-mixture are

$$\begin{aligned}
M_1 &= \frac{1}{k}\sum_{i=1}^{k} \mathcal{M}_1(\mu_i, \sigma_i) = \frac{1}{k}\sum_{i=1}^{k} \mu_i \\
M_2 &= \frac{1}{k}\sum_{i=1}^{k} \mathcal{M}_2(\mu_i, \sigma_i) = \frac{1}{k}\sum_{i=1}^{k} \mu_i^2 + \frac{1}{k}\sum_{i=1}^{k} \sigma_i^2 \\
&\cdots \\
M_m &= \frac{1}{k}\sum_{i=1}^{k} \mathcal{M}_d(\mu_i, \sigma_i) = \frac{1}{k}\sum_{i=1}^{k} \sum_{j=0}^{\lfloor m/2 \rfloor} c_{m,j}\mu_i^{m-2j}\sigma_i^{2j}.
\end{aligned} \tag{1.2}$$

Here we have $2k$ variables $\mu_1 \cdots, \mu_k$ and $\sigma_1 \cdots, \sigma_k$, and as many equations as we desire. Intuitively, to find our variables it suffices to look at first $2k$ moments. In other words, using moments, we can form a system of $2k$ equations and we try to solve it.

There are various ways to solve a system of equations, however the first barrier here is that we can only approximate M_i, $1 \leq i \leq 2k$, using the empirical data.

Given n i.i.d. samples $x_1 \cdots x_n$ from our mixture, the m-th empirical moment is

$$\hat{M}_m = \tfrac{1}{n}\sum_{j=1}^{n} x_j^m. \tag{1.3}$$

Note that $\hat{M}_m$ is random variable with mean M_i and variance σ^{2m}/n, where σ^2 is the variance of the k-mixture. By Chebyshev's inequality we have

$$\mathbb{P}\big[|\hat{M}_m - M_m| > r\frac{\sigma^m}{\sqrt{n}}\big] < \frac{1}{r^2}. \tag{1.4}$$

Take $n \geq \epsilon^{-2}$, then the portion of samples for which we have

$$|\hat{M}_m - M_m| > r\epsilon\sigma^m,$$

is bounded by r^{-2}. This would suffice for our purposes, take for example $r = 10$, then we know that for 99% of samples we have

$$|\hat{M}_m - M_m| < 10\epsilon\sigma^m.$$

This can be improved by taking samples in groups, calculate $\hat{M}_m$ for each group and look at the median of these estimates. In [5][Lemma 3.2] Hardt and Price proved that given $n \gg \log(1/\delta)\epsilon^{-2}$, samples from the mixture, with probability $1-\delta$ we have:

$$|\hat{M}_i - M_i| < \epsilon\sigma^i. \tag{1.5}$$

Going forward we will use (1.5).

1.1. **System of Equations and Perturbation of Coefficients.** In practice, we often have to approximate these moments. Consequently, a relevant question arises: How robust is the system (1.2) when subjected to perturbations in its constant terms?

Problem 1. *Let $\mathcal{M}_r$ be defined as in* (1.1)*, and consider the following system of equations:*

$$\begin{aligned} M_1 + \varepsilon_1 &= \frac{1}{k}\sum_{i=1}^{k} \mathcal{M}_1(\mu_i, \sigma_i) \\ M_2 + \varepsilon_2 &= \frac{1}{k}\sum_{i=1}^{k} \mathcal{M}_2(\mu_i, \sigma_i) \\ &\cdots \\ M_{2k} + \varepsilon_{2k} &= \frac{1}{k}\sum_{i=1}^{k} \mathcal{M}_{2k}(\mu_i, \sigma_i). \end{aligned} \tag{1.6}$$

We know that for $\varepsilon_1 = \varepsilon_2 = \cdots = \varepsilon_{2k} = 0$, the system has a real solution:

$$X = (\mu_1, \cdots, \mu_k, \sigma_1, \cdots, \sigma_k).$$

Now consider the system with perturbation $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_{2k})$ subject to $\varepsilon_i \ll \epsilon\sigma^i$, where $\sigma > 0$ is fixed. Let $X_\varepsilon = (\hat{\mu}_1, \cdots, \hat{\mu}_k, \hat{\sigma}_1, \cdots, \hat{\sigma}_k)$ be a solution to the system with perturbation ε.

Is it possible to establish a bound such as

$$\|X - X_\varepsilon\|_2 \ll C_k \epsilon^{\alpha_k}, \tag{1.7}$$

where C_k and α_k are constants that depend on k?

We solve this problem in the context of the uniform spherical case. Our approach involves constructing a polynomial whose coefficients closely resemble those of our parameter polynomial. Subsequently, our objective is to demonstrate that the roots of this polynomial provide a robust approximation of the parameters we seek to learn.

In general, roots of polynomials are sensitive to the perturbation of the coefficients. A famous example is Wilkinson's polynomial:

$$\omega(x) = (x-1)(x-2)\cdots(x-20).$$

If the coefficient of x^{19} is decreased from -210 by 2^{-23} to -210.0000001192, then the the root at $x = 20$ grows to $x \simeq 20.8$

2. The uniform spherical case

In this section we prove some necessary lemmas that will help us recover parameters of the mixture. We assume that all variances are equal to 1. This assumption simplifies equation (1.2).

Our final goal is to estimate coefficients of the following polynomial:

$$P(x) := \prod_{i=1}^{k}(x-\mu_i),$$

using information we get from (1.2).

Define

$$P_m(\mu_1,\cdots,\mu_k) = \mu_1^m + \cdots + \mu_k^m.$$

We can write each moments in (1.2), in terms of P_m and vice versa. For example

$$\begin{aligned} M_1 &= \frac{1}{k}P_1(\mu_1,\cdots,\mu_k) \\ M_2 &= 1+\frac{1}{k}P_2(\mu_1,\cdots,\mu_k) \\ &\cdots \\ M_m &= \frac{1}{k}\sum_{j=0}^{\lfloor m/2\rfloor} c_{m,j}P_{m-2j}. \end{aligned} \tag{2.1}$$

Using (2.1), and given M_i for $i \le m$, recursively we can find P_i for $i \le m$, in terms of M_i. For example

$$\begin{aligned} P_1 &= kM_1 \\ P_2 &= kM_2 - k \\ P_3 &= kM_3 - 3kM_1 \cdots . \end{aligned} \tag{2.2}$$

Let us precisely calculate $c_{m,j}$ first, then we move to approximate P_m using imperial moments we obtain in (1.3).

Lemma 2.1. *For $m \ge 1$ we have that*

$$\mathcal{M}_m(\mu,1) = \mu^m + \sum_{i=1}^{m/2}\binom{m}{2i}\mu^{m-2i}(2i-1)!!, \tag{2.3}$$

where $(2i-1)!!$ is the product of odd numbers less than $2i-1$.

Remark 1. Using the above lemma we immediately have the expansion of M_m in terms of P_i :

$$kM_m = P_m + \sum_{i=1}^{m/2}\binom{m}{2i}(2i-1)!!P_{m-2i}. \tag{2.4}$$

In the next lemma we show that

$$P_m = kM_m + k\sum_{i=1}^{m/2}(-1)^i\binom{m}{2i}(2i-1)!!M_{m-2i}. \tag{2.5}$$

Proof of Lemma 2.1. We have

$$\begin{aligned}\mathcal{M}_m(\mu,1) &= \frac{1}{\sqrt{2\pi}}\int x^r e^{-\frac{1}{2}\left(x-\mu\right)^2} \\ &= \frac{1}{\sqrt{2\pi}}\int (x+\mu)^m e^{-\frac{1}{2}x^2} = \sum_i \binom{m}{i}\mu^{m-i}E(x^i),\end{aligned}$$

where $x \sim \mathcal{N}(0,1)$.

We use $E(zf(z)) = E(f'(z))$, therefore we have

$$E(x^m) = E(xx^{m-1}) = (m-1)E(x^{m-2}).$$

Also note that $E(x^m)$ is zero if m is odd. This will give the lemma. □

We now define a new object that is a empirical approximation of P_m.

Definition 2. *Let $\hat{M}_m$ be as* (1.3), *following* (2.5), *we define*

$$\hat{P}_m = k\hat{M}_m + k\sum_{i=1}^{m/2}(-1)^i\binom{m}{2i}(2i-1)!!\hat{M}_{m-2i}. \tag{2.6}$$

Lemma 2.2. *We have that P_m satisfy:*

$$P_m = kM_m + k\sum_{i=1}^{m/2}(-1)^i\binom{m}{2i}(2i-1)!!M_{m-2i}. \tag{2.7}$$

Moreover, Let $\varepsilon_m = \hat{P}_m - P_m$ and $\Delta_m = \hat{M}_m - M_m$. We have that

$$\varepsilon_m = k\Delta_m + k\sum_{i=1}^{m/2}(-1)^i\binom{m}{2i}(2i-1)!!\Delta_{m-2i}. \tag{2.8}$$

Proof. We proceed with using induction, we have that $kM_1 = P_1(\mu_1,\cdots,\mu_k)$ and $kM_2 = 1 + P_2(\mu_1,\cdots,\mu_k)$. This gives us (2.7) and (2.8) for the base cases. We assume (2.8) for $m-2, m-4, \cdots$ and we prove it for m.

Using (2.4), we have the following recursive identity relating ε_m to Δ_m and $\varepsilon_{m-2}, \varepsilon_{m-4}, \cdots$.

$$\varepsilon_m = k\Delta_m - \sum_{i=1}^{m/2}\binom{m}{2i}(2i-1)!!\varepsilon_{m-2i}. \tag{2.9}$$

We use our induction hypothesis in (2.9) and we get everything in terms of $\Delta_m, \Delta_{m-2}, \cdots$. Only thing remaining is to calculate the coefficients of Δ_{m-2i} for $0 < i \le m/2$.

The term Δ_{m-2i} appears in (2.9) expansion of $\varepsilon_m, \varepsilon_{m-2}, \cdots, \varepsilon_{m-2i}$. Therefore its coefficients equal to

$$-\binom{m}{2i}(2i-1)!!+\sum_{n=1}^{i-1}(-1)^{n-1}\binom{m}{2n}\binom{m-2n}{2i-2n}(2n-1)!!(2i-2n-1)!! \tag{2.10}$$

If i is odd, then all terms cancel each other except $n=i$, which gives

$$(-1)^i\binom{m}{2i}(2i-1)!!.$$

When i is even we have (2.10) equals to $m(m-1)\cdots(m-2i+1)$ times

$$\begin{aligned} &-\frac{(2i-1)!!}{2i!}-\sum_{n=1}^{i-1}(-1)^n\frac{(2n-1)!!(2i-2n-1)!!}{2n!(2i-2n)!} \\ =\ &-\sum_{n=1}^{i}\frac{(-1)^n}{2^n n! 2^{i-n}(i-n)!} = -\frac{1}{2^i i!}\sum_{n=1}^{i}\frac{(-1)^n i!}{n!(i-n)!} \\ =\ &-\frac{1}{2^i i!}\big((1-1)^i-1\big)=\frac{1}{2^i i!}. \end{aligned}$$

Therefore the coefficient of Δ_{m-2i} equals to

$$\frac{1}{2^i i!}m(m-1)\cdots(m-2i+1)=\binom{m}{2i}(2i-1)!!.$$

□

Now if we have $\gg \log(1/\delta)\epsilon^{-2}$ samples, using (1.5) we have that

$$\Delta_m \ll \epsilon\sigma^m.$$

Therefore, Using the above lemma we get

$$\varepsilon_m=\hat{P}_m-P_m \ll k\sigma^m e^{0.5(\frac{m}{\sigma})^2}. \tag{2.11}$$

We obtain (2.11) by considering the Taylor expansion of $e^{0.5(\frac{m}{\sigma})^2}$ and comparing it to (2.8).

3. Newton's identities and roots of polynomials

So far we established that in the case of spherical Gaussian using the method of moments we can easily calculate P_m. Now consider the following polynomial:

$$P(x):=\prod_{i=1}^{k}(x-\mu_i)=x^k-e_1x^{k-1}+\cdots+(-1)^k e_k \tag{3.1}$$

Recall that P is our parameter polynomial. By using Newton's identities we can write coefficients e_n, in terms of P_m, with $m\le n$. For example we have

$$\begin{aligned} e_1 &= P_1 \\ 2e_2 &= P_1^2 - P_2 \\ 3e_3 &= \tfrac{1}{2}P_1^3 - \tfrac{3}{2}P_1P_2 + P_3, \cdots \end{aligned}$$

In general we have

$$e_n = \sum_{j=1}^{n} (-1)^j e_{n-j} P_j. \tag{3.2}$$

Similar to the definition 2.6 we define

Definition 3. *Let* $\hat{e}_0 = 1$ *and* $\hat{e}_1 = \hat{P}_1$*. Recursively we define*

$$\hat{e}_n = \sum_{j=1}^{n} (-1)^j \hat{e}_{n-j} \hat{P}_j. \tag{3.3}$$

By this definition and (1.5) and by some calculation we obtain

$$\begin{aligned} \hat{e}_1 &= O(\epsilon k \sigma) \\ 2\hat{e}_2 &= -k(\sigma^2 - 1) + O(\epsilon k \sigma^2) \\ 3\hat{e}_3 &= KM_3 + O(\epsilon k^2(\sigma^3 + 1)). \end{aligned}$$

Next we estimate how close $\hat{e}_n$ is to e_n.

Lemma 3.1. *Assuming we have* $\asymp \log(1/\delta)\epsilon^{-2}$ *samples from our mixture, and* $P_1 = 0$*, and we have that*

$$|\hat{e}_m - e_m| \ll \epsilon K (2\sigma)^m e^{0.5\left(\frac{m}{\sigma}\right)^2} \tag{3.4}$$

Proof. Let $E_n = \hat{e}_n - e_n$. Using (3.3) and (3.2), we have that

$$E_n \ll \sum_{i=1}^{n} \Big(e_{n-i}\varepsilon_i + E_{n-i}P_i + E_{n-i}\varepsilon_i \Big) \tag{3.5}$$

We use following bounds on each of the terms:

(1) For $i \geq$ we have

$$P_m < \sigma^m.$$

(2) Using (2.11) we have

$$\varepsilon_m \ll k\sigma^m e^{0.5(m/\sigma)^2}.$$

(3) Using induction hypothesis for $i < n$

$$E_{n-i} \ll k(2\sigma)^{n-i} e^{0.5\left(\frac{n-i}{\sigma}\right)^2}.$$

(4) From [11] we have

$$e_n \ll \big(\frac{6e}{n}\big)^{n/2} \sigma^n.$$

Applying these bound to the first term inside the summation in (3.5) we get

$$\sum_{i=1}^{n} e_{n-i}\varepsilon_i \ll \epsilon \sum_{i=1}^{n} \big(\frac{6e}{n-i}\big)^{(n-i)/2} \sigma^{n-i}\sigma^{i} e^{0.5(i/\sigma)^2}$$

$$\epsilon k \sigma^{n} \sum_{i=1}^{n} \big(\frac{6e}{n-i}\big)^{(n-i)/2} e^{0.5(i/\sigma)^2} \ll \epsilon k \sigma^{n} e^{0.5(n/\sigma)^2}.$$

As for the term

$$\sum_{i=1}^{n} E_{n-i} P_i$$

using mentioned bounds gives (3.4). The third error term is obviously smaller than the other two. □

We prove that in the spherical case having first k-moment is enough to uniquely determine parameters of the mixture.

Lemma 3.2. *Let G be a k-mixture of spherical Gaussian (KGMM):*

$$G = \tfrac{1}{k} \sum_{i=1}^{k} \mathcal{N}(\mu_i, 1).$$

We have that first k moments of G uniquely determine parameters $\mu_1, \cdots, \mu_k$.

Proof. By using (2.2) knowing $M_1, \cdots, M_k$, we can uniquely find $p_1, \cdots, p_k$. Assume that $\check{M}_1 \cdots, \check{M}_k$, would also give us $p_1, \cdots, p_k$. Then, using the first equation in (2.2) we have $M_1 = \check{M}_1$, by the second equation, we have $M_2 = \check{M}_2$, and so on.

Since we have $P_1, \cdots, P_k$, using Newton's identities we can determine coefficients of $P(x)$ in (3.1). This polynomial has at most k-distinct roots, and we know that $\mu_1, \cdots, \mu_k$ are roots of P, therefore first k moments uniquely determine parameters of the mixture. □

4. APPROXIMATING COEFFICIENTS USING SAMPLES

Using results in the previous section we can construct a polynomial whose roots are close to $\mu_1, \cdots, \mu_k$. Let

$$\hat{P}(x) = \sum_{n=0}^{k} \hat{e}_n x^n,$$

where we obtain $\hat{e}_n$ as in (3.3). By using Lemma 3.1 we have that $\hat{P}(x)$ is an approximation of P. Our task is to measure how close are roots of $\hat{P}$ to the roots of P. We apply the following theorem of Beauzamy [10]

Let $Q(x) = \sum_{i=0}^{k} a_i x^i$ be a polynomial with complex coefficients and degree k. The Bombieri's norm of Q is defined as

$$B(Q) = \Big(\sum_{i=0}^{k} \frac{|a_i|^2}{\binom{k}{i}} \Big)^{1/2}. \tag{4.1}$$

Theorem (Beauzamy)**.** *Let P and $\hat{P}$ be polynomials of degree k. Assume that $B(\hat{P} - P) < \varepsilon$. If x is any zero of P there exist zero y of $\hat{P}$ with*

$$|x - y| < \frac{k(1 + |x|^2)^{k/2}}{|\hat{P}'(x)|} \varepsilon. \tag{4.2}$$

If ε is small enough, namely

$$\varepsilon < \frac{|P'(x)|}{k(1 + |x|^2)^{(k-1)/2}},$$

then

$$|x - y| < \frac{k(1 + |x|^2)^{k/2}}{|P'(x)|} \varepsilon. \tag{4.3}$$

Proof of Theorem 0.3. We apply the above theorem to examine how close are roots of $\hat{P}$ to roots of P. Using Lemma 3.1 we know that if we have $\asymp \log(1/\delta)\varepsilon^{-2}$ samples from the mixture, then the Bombieri's norm of $\hat{P} - P$ is bounded by

$$B(\hat{P} - P) \ll \varepsilon \Big(\sum_{i=0}^{k} \frac{k^2 e^{(i/\sigma)^2} (2\sigma)^{2i}}{\binom{k}{i}} \Big)^{1/2} \ll \varepsilon k e^{0.5(k/\sigma)^2} (2\sigma)^k.$$

Therefore by Beauzamy's theorem, if

$$\varepsilon < \frac{\prod_{j \neq m} (\mu_m - \mu_j)}{k^2 (1 + |\mu_m|^2)^{(k-1)/2} (2\sigma)^k e^{0.5(k/\sigma)^2}}$$

there exist a root $\hat{\mu}_m$ of $\hat{P}$ such that

$$|\hat{\mu}_m - \mu_m| < \frac{k^2 (1 + |\mu_m|^2)^{k/2} (2\sigma)^k e^{0.5(k/\sigma)^2}}{\prod_{j \neq m} (\mu_m - \mu_j)} \varepsilon. \tag{4.4}$$

This gives us the proof of Theorem 0.3. □

Proof of corollary 0.2. First note that in the statement of Beauzamy's theorem and the proof of Theorem 0.3, we used ε to indicate the distance between coefficients of polynomial. Going forward we use ϵ to indicate the accuracy we expect the statement of the corollary.

To get Corollary 0.2, LHS of (4.4) must be smaller than the accuracy :

$$|\hat{\mu}_m - \mu_m| < \epsilon,$$

for all m.

In order to have this, ε in the RHS of (4.4), must be smaller than ϵ. Therefore we must have

$$\varepsilon = \epsilon \frac{\prod_{j\neq m}(\mu_m - \mu_j)}{k^2(1+|\mu_m|^2)^{k/2}(2\sigma)^k e^{0.5(k/\sigma)^2}}.$$

Number of samples we need to get this accuracy is

$$\gg \log(\delta^{-1})\varepsilon^{-2} = \log(\delta^{-1})\epsilon^{-2}\left|\frac{\prod_{j\neq m}(\mu_m - \mu_j)}{k^2(1+|\mu_m|^2)^{k/2}(2\sigma)^k e^{0.5(k/\sigma)^2}}\right|^{-2}.$$

We use

$$1+\mu_m^2 \ll 1+k\sigma^2,$$

and we set

$$c(\sigma,k) = \left|k^2(1+k\sigma^2)^{k/2}(2\sigma)^k e^{0.5(k/\sigma)^2}\right|^2$$

Therefore, the sample complexity is

$$\log(\delta^{-1})\frac{c(\sigma,k)}{\left|\mathcal{P}_m\right|^2}\epsilon^{-2}. \tag{4.5}$$

We complete the proof of the corollary by taking the minimum over m. □

Proof of corollary 0.1. We assumed $\min_{i\neq j}(\mu_i - \mu_j) > \epsilon$, therefore applying Corollary 0.2, we have that $\min_m \mathcal{P}(\mu_m) > \epsilon^{k-1}$. This indicates that the denominator in (4.5) is bigger than ϵ^{k-1}. We would like to learn our parameters with $\epsilon/100$-accuracy. Therefore the number of samples we need is

$$10^4 \log\left(\frac{1}{\delta}\right)\epsilon^{-2k+2}c(\sigma,k)\,\epsilon^{-2}.$$

This completes the proof. □

Description of the algorithm. Using samples from the mixture we we easily compute $\hat{M}_1,\cdots,\hat{M}_k$ by (1.3).

We use (2.6) and from $\hat{M}_1,\cdots,\hat{M}_k$ we recover $\hat{P}_1,\cdots,\hat{P}_k$.

Next, by using (3.4) from $\hat{P}_1,\cdots,\hat{P}_k$, recursively we obtain $\hat{e}_1,\cdots,\hat{e}_k$. At this point we have coefficients of our empirical parameters polynomial.

Finally we use one of algorithms for finding roots of univariant polynomials. For example one can use the Aberth method [1]. Theorem 0.3 gives us a guarantee in terms of how close are roots of the empirical parameters polynomial to parameters of the mixture.

References

[1] Aberth, Oliver, Iteration methods for finding all zeros of a polynomial simultaneously. Math. Comp. Mathematics of Computation, Vol. 27, No. 122. 27 (122): 339–344. doi:10.2307/2005621. (1973).

[2] Hassan Ashtiani, Shai Ben-David, Nicholas J. A. Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. J. ACM, 67(6), October 2020.

[3] Allen Liu and Jerry Li. 2022. Clustering mixtures with almost optimal separation in polynomial time. In Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC 2022). Association for Computing Machinery, New York, NY, USA, 1248–1261.

[4] Mingda Qiao, Guru Guruganesh, Ankit Singh Rawat, Avinava Dubey, Manzil Zaheer. A Fourier Approach to Mixture Learning. To appear at NeurIPS 2022.

[5] Moritz Hardt and Eric Price. 2015. Tight Bounds for Learning a Mixture of Two Gaussians. https://doi.org/10.48550/arXiv.1404.4997.

[6] A. Moitra and G. Valiant. 2010. Settling the polynomial learnability of mixtures of Gaussians. In FOCS.

[7] A. Moitra and G. Valiant. 2010. Settling the polynomial learnability of mixtures of Gaussians. In FOCS.

[8] M. Belkin and K. Sinha. 2010. Polynomial Learning of Distribution Families. In FOCS.

[9] Moritz Hardt and Eric Price. 2015. Tight Bounds for Learning a Mixture of Two Gaussians. In Proceedings of the forty-seventh annual ACM symposium on Theory of Computing (STOC '15). Association for Computing Machinery, New York, NY, USA, 753–760.

[10] Bernard Beauzamy, How the Roots of a Polynomial Vary with its Coefficients: A Local Quantitative Result. Canad. Math. Bull. Vol. 42 (1), 1999 pp. 3–12.

[11] Parikshit Gopalan, Amir Yehudayoff, Inequalities and tail bounds for elementary symmetric polynomials with applications. arXiv:1402.3543. 2015.

[12] K. Pearson. 1894. Contribution to the mathematical theory of evolution. Phil. Trans. Roy. Soc. A 185 (1894), 71ś110.