Journal of Machine Learning Research 18 (2018) 1-11 Submitted 12/16; Revised 12/16; Published 4/18

Maximum Likelihood Estimation for Mixtures of Spherical Gaussians is NP-hard

Christopher Tosh CTOSH@CS.UCSD.EDU
Sanjoy Dasgupta DASGUPTA@CS.UCSD.EDU
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093-0404, USA

Editor: Mikhail Belkin

Abstract

This paper presents NP-hardness and hardness of approximation results for maximum likelihood estimation of mixtures of spherical Gaussians.

Keywords: Mixtures of Gaussians, maximum likelihood, NP-completeness

1. Introduction

A *spherical Gaussian* in $\mathbb{R}^d$ is a distribution specified by its mean $\mu \in \mathbb{R}^d$ and variance $\sigma^2 > 0$, with density

$$N(x; \mu, \sigma^2) \;=\; \left(\frac{1}{2\pi\sigma^2}\right)^{d/2} \exp\left(-\frac{\|x-\mu\|^2}{2\sigma^2}\right).$$

(The standard notation for this Gaussian is $N(\mu, \sigma^2 I_d)$, but we will drop the identity matrix as a shorthand.)

When data arise from several sources, or form several clusters, it is common to model each source or cluster by a spherical Gaussian. If there are k sources, the resulting overall distribution is a mixture of k Gaussians,

$$\pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2) + \cdots + \pi_k N(\mu_k, \sigma_k^2),$$

where $\mu_i \in \mathbb{R}^d$ and σ_i^2 are the mean and variance of the ith component, and π_i is the fraction of the distribution that arises from this component. In what follows, we will often package the parameters together as $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k)$.

A standard statistical task is to fit a mixture of k Gaussians to a given data set. This is typically formulated as an optimization problem (Dempster et al., 1977), where given data points $x_1, \ldots, x_n \in \mathbb{R}^d$, the goal is to find the parameters $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ that maximize the *log-likelihood*

$$LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{i=1}^{n} \ln\left(\sum_{j=1}^{k} \pi_j N(x_i; \mu_j, \sigma_j^2)\right). \tag{1}$$

In this brief note, we establish the computational hardness of this estimation problem. This is in contrast with various positive results showing that, when data is in fact generated

from a Gaussian mixture, it is possible to efficiently recover the mixture from a sample of polynomial size, under certain conditions; relevant work includes, for instance, Belkin and Sinha (2010), Moitra and Valiant (2010), Hsu and Kakade (2013), and Hardt and Price (2015), among others.

1.1 Gaussians with the same variance

We start with the simplest subcase, where the variances of the components are constrained to be the same.

> mixtures of spherical Gaussians with same variance: mog-sv
> *Input*: Points $x_1, \ldots, x_n \in \mathbb{R}^d$; positive integer k; unary parameter b.
> *Output*: A mixture of k spherical Gaussians with the same variance, $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$, whose log-likelihood
>
> $$LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{k} \pi_j N(x_i; \mu_j, \sigma^2) \right)$$
>
> is within an additive factor $1/b$ of optimal.

Since the parameters of the optimal mixture are real-valued, they can only be provided to within some precision. The role of the input parameter b is to specify the desired level of accuracy. It is worth pointing out, however, that in our reductions, the coordinates of data points take values in $\{-1, 0, 1\}$ and the hardness does not stem from precision issues but rather from underlying combinatorial structure.

mog-sv is similar to the k-means clustering problem, which is NP-hard (Aloise et al., 2009).

> k-means
> *Input*: Points $x_1, \ldots, x_n \in \mathbb{R}^d$; positive integer k.
> *Output*: A collection of k "centers" $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ in $\mathbb{R}^d$ that minimize the cost function
>
> $$\Phi(\boldsymbol{\mu}) = \sum_{i=1}^{n} \min_{1 \le j \le k} \|x_i - \mu_j\|^2.$$

The biggest difference between the two problems is that k-means assigns each data point x_i to a single center μ_j (a "hard" clustering), while the mixture of Gaussians effectively spreads it out over all the centers (a "soft" clustering). Earlier work (Arora and Kannan, 2001) has established that a "hard clustering" version of the mixture of Gaussians problem is NP-hard. Here we consider the more standard formulation, and show that it is hard even when $k = 2$.

Theorem 1 mog-sv *is NP-hard on instances with* $k = 2$.

The proof follows from the observation that an additive approximation to the best mog-sv solution yields a multiplicative approximation to the best k-means solution:

Lemma 2 *Fix any data set $x_1, \ldots, x_n \in \mathbb{R}^d$ and any positive integer k. Let LL_{OPT} denote the log-likelihood of the optimal solution to* MOG-SV*, and Φ_{OPT} the lowest achievable k-means cost. For any parameters $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$, we have*

$$\ln \frac{\Phi(\boldsymbol{\mu})}{\Phi_{OPT}} \;\leq\; \frac{4 \ln k}{d} + \frac{2}{nd} \left(LL_{OPT} - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \right).$$

The first term on the right-hand side comes from the discrepancy between hard and soft clustering. It can be made negligible by increasing the dimension, for instance by padding each point with extra zero-valued coordinates.

Lemma 2 can also be combined with a recent hardness of approximation result for k-means (Awasthi et al., 2015) to show that, if k is allowed to be large, MOG-SV cannot be approximated within an additive factor of $o(nd)$.

Theorem 3 *There is a family of* MOG-SV *instances with the following properties:*

- *An instance with n points has dimension $O(n)$.*
- *Each point is $\{0, 1\}$-valued and has $O(1)$ nonzero coordinates.*
- *$k = \Theta(n)$.*

For some absolute constant c_o, it is NP-hard to approximate MOG-SV *on such instances within an additive factor of $c_o dn$.*

The specific form of this result (additive versus multiplicative approximation, interpoint distances that are small constants) is motivated by the unusual properties of the log-likelihood objective. To begin with, consider the problem of fitting a single Gaussian to a data set $\mathcal{X} \subset \mathbb{R}^d$ of size n. A quick calculation shows that the log-likelihood (of the maximum likelihood estimate) is

$$\frac{dn}{2} \ln \frac{d}{2\pi e} - \frac{dn}{2} \ln \text{radius}(\mathcal{X}), \quad \textit{where} \quad \text{radius}(\mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \|x - \text{mean}(\mathcal{X})\|^2.$$

Depending on the scale of the data, this log-likelihood could be positive, negative, or zero. When fitting a mixture of k Gaussians, the log-likelihood has a term of this sort for each cluster, plus an additional term of size $\pm n \ln k$ due to the mixing weights. For the kind of instance described in the theorem, any cluster with at least two points has radius $\Theta(1)$ and thus the log-likelihoods of all reasonable mixture models lie in an interval of size $O(dn)$.

The proofs of these results appear in Section 2.

1.2 Gaussians with differing variances

When the different Gaussian components are allowed to have different variances, and $k > 1$, the maximum-likelihood solution is always degenerate. This is because it is possible to make the log-likelihood go to infinity by centering one of the Gaussians at a single data point and letting its variance go to zero. Thus, in order for the problem to be well-defined, an additional constraint must be introduced. One option is to force all variances to be non-negligible.

mixtures of spherical Gaussians with constrained variances: mog
Input: Points $x_1, \ldots, x_n \in \mathbb{R}^d$; positive integer k; value $\sigma_o > 0$; unary integer b.
Output: A mixture of k spherical Gaussians $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ whose log-likelihood is within an additive factor $1/b$ of optimal, subject to the constraint $\sigma_1, \ldots, \sigma_k \geq \sigma_o$.

This problem is slightly further from k-means, but remains intractable.

Theorem 4 mog *is NP-hard on instances with* $k = 2$.

The proof appears in Section 3.

2. Mixtures of spherical Gaussians with the same variance

2.1 Induced partitions

We start with a basic relation between hard and soft clustering that applies to arbitrary mixture models, not just those with Gaussian components of the same variance.

Although a mixture model represents a soft clustering, it also induces a natural hard partition. For data set $\mathcal{X}$ and mixture of Gaussians $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$, this hard partition has clusters

$$\mathcal{X}_j \;=\; \left\{ x \in \mathcal{X} \;:\; j = \operatorname*{argmax}_{\ell} \; \pi_\ell N(x; \mu_\ell, \sigma_\ell^2) \right\} \tag{2}$$

(breaking ties arbitrarily). The log-likelihood of a mixture is easily bounded in terms of the likelihood of the corresponding hard partition.

Lemma 5 *Pick any mixture* $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ *and data set* $\mathcal{X} = \{x_1, \ldots, x_n\}$.

(a) For any partition $(\mathcal{X}'_1, \ldots, \mathcal{X}'_k)$ *of* $\mathcal{X}$*, we have*

$$LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \;\geq\; \sum_{j=1}^{k} \sum_{x \in \mathcal{X}'_j} \ln(\pi_j N(x; \mu_j, \sigma_j^2)).$$

(b) For the partition $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ *induced by* $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$*, as in Eq (2), we have*

$$LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \;\leq\; n \ln k + \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \ln(\pi_j N(x; \mu_j, \sigma_j^2)).$$

Proof Recall from (1) that the contribution of each data point x_i to $LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ is

$$\ln \left(\sum_{j=1}^{k} \pi_j N(x_i; \mu_j, \sigma_j^2) \right).$$

For $x_i \in \mathcal{X}'_j$, we can lower-bound this contribution by $\ln(\pi_j N(x_i; \mu_j, \sigma_j^2))$. Similarly, if $x_i \in \mathcal{X}_j$, then we can upper-bound the contribution by $\ln(k\pi_j N(x_i; \mu_j, \sigma_j^2))$, by the manner in which the hard partition $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ is defined. ■

2.2 Proof of Lemma 2

As in the statement of Lemma 2, fix data $x_1, \ldots, x_n \in \mathbb{R}^d$, and define LL_{OPT} to be the log-likelihood of the optimal solution of MOG-SV. Let Φ_{OPT} be the optimal k-means cost.

Pick any parameters $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$, and let $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ be the induced hard partition of the data set, as per Eq (2). From Lemma 5,

$$\begin{aligned}
LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) &\leq n \ln k + \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \left(\ln \pi_j + \frac{d}{2} \ln \left(\frac{1}{2\pi\sigma^2} \right) - \frac{\|x - \mu_j\|^2}{2\sigma^2} \right) \\
&\leq n \ln k + \frac{nd}{2} \ln \left(\frac{1}{2\pi\sigma^2} \right) - \frac{1}{2\sigma^2} \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \|x - \mu_j\|^2 \\
&\leq n \ln k + \frac{nd}{2} \ln \left(\frac{1}{2\pi\sigma^2} \right) - \frac{\Phi(\boldsymbol{\mu})}{2\sigma^2} \\
&\leq n \ln k + \frac{nd}{2} \ln \left(\frac{nd}{2\pi\Phi(\boldsymbol{\mu})} \right) - \frac{nd}{2},
\end{aligned}$$

where the last inequality comes from solving for the optimal value of σ^2 (namely, $\Phi(\boldsymbol{\mu})/nd$) in the preceding line.

Suppose the optimal k-means solution is realized by centers $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_k^*)$. Let $\pi_1^* = \cdots = \pi_k^* = 1/k$ and $\sigma^{*2} = \Phi(\boldsymbol{\mu}^*)/nd$. To bound the log-likelihood of the mixture model $(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \sigma^*)$, we look at the hard partition that it induces, $(\mathcal{X}_1^*, \ldots, \mathcal{X}_k^*)$, and notice that $\mathcal{X}_j^*$ consists of points whose closest center is μ_j^*. We then apply Lemma 5 to get

$$\begin{aligned}
LL(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \sigma^*) &\geq \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j^*} \left(\ln \pi_j^* + \frac{d}{2} \ln \left(\frac{1}{2\pi\sigma^{*2}} \right) - \frac{\|x - \mu_j^*\|^2}{2\sigma^{*2}} \right) \\
&= -n \ln k + \frac{nd}{2} \ln \left(\frac{1}{2\pi\sigma^{*2}} \right) - \frac{1}{2\sigma^{*2}} \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j^*} \|x - \mu_j^*\|^2 \\
&= -n \ln k + \frac{nd}{2} \ln \left(\frac{1}{2\pi\sigma^{*2}} \right) - \frac{1}{2\sigma^{*2}} \Phi(\boldsymbol{\mu}^*) \\
&= -n \ln k + \frac{nd}{2} \ln \left(\frac{nd}{2\pi\Phi(\boldsymbol{\mu}^*)} \right) - \frac{nd}{2},
\end{aligned}$$

where the last equality comes from substituting in the value of σ^{*2}. Combining our bounds for the two mixtures, we get

$$\begin{aligned}
LL_{OPT} - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) &\geq LL(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \sigma^*) - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \\
&\geq \frac{nd}{2} \ln \left(\frac{\Phi(\boldsymbol{\mu})}{\Phi(\boldsymbol{\mu}^*)} \right) - 2n \ln k.
\end{aligned}$$

Rearranging terms yields the lemma statement.

2.3 Proof of Theorem 1

With Lemma 2 in place, a reduction from k-means to MOG-SV is almost immediate. There are various hardness results available for k-means (Aloise et al., 2009; Dasgupta and Freund, 2009; Mahajan et al., 2009; Awasthi et al., 2015); of these, we use Aloise et al. (2009) as a starting point.

Theorem 6 (Aloise et al. (2009)) *There exists a family of k-means instances with the following properties, for some low-order polynomials $\alpha(\cdot)$ and $\beta(\cdot)$:*

- *For an instance containing n points, each point has dimension at most $\alpha(n)$, with individual coordinates taking values in $\{-1, 0, 1\}$.*
- *It is NP-hard to approximate the best k-means solution, with $k = 2$, within a factor of $1 + 1/\beta(n)$.*

To prove Theorem 1, we reduce the problem of finding a $(1 + 1/\beta(n))$-approximate k-means solution to MOG-SV. Given an instance $x_1, \ldots, x_n$ of k-means:

- Pad each point with additional zero-valued coordinates until the dimension d exceeds $16\beta(n) \ln k$. This has no effect on interpoint distances or on the optimal k-means cost.
- Solve MOG-SV for these modified points, with precision parameter $b = 1$. This yields $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ such that $LL_{OPT} - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \leq 1$, where LL_{OPT} is the optimal log-likelihood. It follows from Lemma 2 that

$$\ln \frac{\Phi(\boldsymbol{\mu})}{\Phi_{OPT}} \leq \frac{4 \ln k}{d} + \frac{2}{nd} \leq \frac{1}{2\beta(n)},$$

  whereupon $\Phi(\boldsymbol{\mu}) \leq \Phi_{OPT}(1 + 1/\beta(n))$.

2.4 Proof of Theorem 3

A recent hardness of approximation result for k-means shows the following.

Theorem 7 (Awasthi et al. (2015)) *There is a family of k-means instances with the following properties:*

- *An instance with n points has dimension at most n, points that are $\{0, 1\}$-valued (and have at most two non-zero coordinates), and a target number of clusters $k = \Omega(n)$.*
- *It is NP-hard to approximate the optimal k-means solution within a factor c, for some absolute constant $c > 1$.*

Pick any $c_o < (1/2) \ln c$. To see that it is hard to approximate MOG-SV within an additive factor $c_o nd$, we reduce from k-means as follows. Start with an instance $x_1, \ldots, x_n \in \mathbb{R}^d$ of the type described in Theorem 7. Then:

- If necessary, pad points with zero-valued coordinates to bring the dimension up to

$$d \geq \frac{4 \ln k}{(\ln c) - 2c_o}.$$

- Obtain an approximate solution $(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma)$ to MOG-SV on these points such that $LL_{OPT} - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) \leq c_o nd$.

- Return the centers $\boldsymbol{\mu}$.

By Lemma 2, we have

$$\ln \frac{\Phi(\boldsymbol{\mu})}{\Phi_{OPT}} \;\leq\; \frac{4 \ln k}{d} + \frac{2}{nd} c_o nd \;\leq\; \ln c,$$

so that $\boldsymbol{\mu}$ is a c-approximate solution to the k-means instance.

3. The general case

We now consider the case where the variances are allowed to differ but are uniformly lower bounded. Specifically, a mixture model $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ is *admissible* if all $\sigma_j \geq \sigma_o$, where σ_o is supplied as part of the input.

The basic reduction still applies, with an additional device to force all variances to be close to the lower bound—and therefore approximately equal.

3.1 Controlling the variances

Lemma 8 *Fix any data set $\mathcal{X} = \{x_1, \ldots x_n\}$ in $\mathbb{R}^d$, and let $D = \max_{i \neq i'} \|x_i - x_{i'}\|$ denote its diameter. Pick any $\Delta, \delta > 0$. If the dimension d satisfies*

$$d \geq \frac{4}{\delta} \left(\frac{nD^2}{2\sigma_o^2} + n \ln k + \Delta \right), \tag{3}$$

then any admissible mixture $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ within an additive factor Δ of optimal (that is, $LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \geq LL_{OPT} - \Delta$) has the following property: in the associated hard partition $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$, any nonempty cluster $\mathcal{X}_j$ has $\sigma_j^2 \leq \sigma_o^2(1 + \delta)$.

Proof Pick any admissible mixture $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ that is within Δ of optimal, and let $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ be the associated hard partition. Let $\widetilde{\mu}_j$ denote the cluster means:

$$\widetilde{\mu}_j \;=\; \frac{1}{|\mathcal{X}_j|} \sum_{x \in \mathcal{X}_j} x.$$

Using Lemma 5, we can compare the log-likelihood of $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ to that of the adjusted parameters $(\boldsymbol{\pi}, \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\sigma}})$, where each $\widetilde{\sigma}_j = \sigma_o$.

$$\begin{aligned}
&LL(\boldsymbol{\pi}, \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\sigma}}) - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \\
&\geq \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \left(\ln(\pi_j N(x; \widetilde{\mu}_j, \sigma_o^2)) - \ln(\pi_j N(x; \mu_j, \sigma_j^2)) \right) - n \ln k \\
&= \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \left(\frac{d}{2} \ln \frac{1}{2\pi\sigma_o^2} - \frac{\|x - \widetilde{\mu}_j\|^2}{2\sigma_o^2} - \frac{d}{2} \ln \frac{1}{2\pi\sigma_j^2} + \frac{\|x - \mu_j\|^2}{2\sigma_j^2} \right) - n \ln k \\
&= \sum_{j=1}^{k} \left(\frac{d|\mathcal{X}_j|}{2} \ln \frac{\sigma_j^2}{\sigma_o^2} + \sum_{x \in \mathcal{X}_j} \left(\frac{\|x - \mu_j\|^2}{2\sigma_j^2} - \frac{\|x - \widetilde{\mu}_j\|^2}{2\sigma_o^2} \right) \right) - n \ln k \\
&\geq \sum_{j=1}^{k} \left(\frac{d|\mathcal{X}_j|}{2} \ln \frac{\sigma_j^2}{\sigma_o^2} + \left(\frac{1}{2\sigma_j^2} - \frac{1}{2\sigma_o^2} \right) \sum_{x \in \mathcal{X}_j} \|x - \widetilde{\mu}_j\|^2 \right) - n \ln k \\
&\geq \sum_{j=1}^{k} |\mathcal{X}_j| \left(d \ln \frac{\sigma_j}{\sigma_o} - \frac{D^2}{2\sigma_o^2} \right) - n \ln k \quad \geq \quad d \ln \frac{\max_{j: \mathcal{X}_j \neq \emptyset} \sigma_j}{\sigma_o} - \frac{nD^2}{2\sigma_o^2} - n \ln k.
\end{aligned}$$

In the second-last line, we have exploited the fact that $\widetilde{\mu}_j$ is the mean of cluster $\mathcal{X}_j$, so that $\sum_{x \in \mathcal{X}_j} \|x - \widetilde{\mu}_j\|^2 \leq \sum_{x \in \mathcal{X}_j} \|x - \mu_j\|^2$, and for the last line we have used $\|x - \widetilde{\mu}_j\| \leq D$.

The difference above is at most Δ, and thus for each nonempty cluster $\mathcal{X}_j$,

$$d \ln \frac{\sigma_j}{\sigma_o} - \frac{nD^2}{2\sigma_o^2} - n \ln k \leq \Delta,$$

whereupon $\sigma_j^2 \leq \sigma_o^2(1 + \delta)$ given the bound (3) on the dimension d. ■

This observation allows us to prove following analog of Lemma 2.

Lemma 9 *Following the terminology of Lemma 8, pick $\delta, \Delta > 0$ and suppose that the dimension satisfies (3). Pick any admissible mixture $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ whose log-likelihood is within an additive factor Δ of the optimal. Then*

$$\Phi(\boldsymbol{\mu}) \ \leq \ (1 + \delta) \left(2\sigma_o^2(\Delta + 2n \ln k) + \Phi_{OPT} \right).$$

Proof Let $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ be the hard partition of the data set induced by $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$. By Lemma 8, we know that for any nonempty cluster $\mathcal{X}_j$, the variance σ_j^2 is at most $(1 + \delta)\sigma_o^2$.

Thus, using Lemma 5, we have

$$\begin{aligned}
LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) &\leq n \ln k + \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \left(\ln \pi_j + \frac{d}{2} \ln \frac{1}{2\pi\sigma_j^2} - \frac{\|x - \mu_j\|^2}{2\sigma_j^2} \right) \\
&\leq n \ln k + \sum_{j=1}^{k} \left(\frac{|\mathcal{X}_j| d}{2} \ln \frac{1}{2\pi\sigma_j^2} - \frac{1}{2\sigma_j^2} \sum_{x \in \mathcal{X}_j} \|x - \mu_j\|^2 \right) \\
&\leq n \ln k + \frac{nd}{2} \ln \frac{1}{2\pi\sigma_o^2} - \frac{1}{2(1+\delta)\sigma_o^2} \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j} \|x - \mu_j\|^2 \\
&\leq n \ln k + \frac{nd}{2} \ln \left(\frac{1}{2\pi\sigma_o^2} \right) - \frac{\Phi(\boldsymbol{\mu})}{2(1+\delta)\sigma_o^2}
\end{aligned}$$

Let $\mu_1^*, \ldots, \mu_k^*$ be an optimal k-means solution and let $(\mathcal{X}_1^*, \ldots, \mathcal{X}_k^*)$ be the hard partition of the data set induced by the mixture model $(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \boldsymbol{\sigma}^*)$ where $\pi_j^* = 1/k$ and $\sigma_j^* = \sigma_o$ for all j. Again using Lemma 5,

$$\begin{aligned}
LL(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \boldsymbol{\sigma}^*) &\geq \sum_{j=1}^{k} \sum_{x \in \mathcal{X}_j^*} \left(\ln \pi_j^* + \frac{d}{2} \ln \frac{1}{2\pi\sigma_j^{*2}} - \frac{\|x - \mu_j^*\|^2}{2\sigma_j^{*2}} \right) \\
&= -n \ln k + \frac{nd}{2} \ln \left(\frac{1}{2\pi\sigma_o^2} \right) - \frac{\Phi(\boldsymbol{\mu}^*)}{2\sigma_o^2}
\end{aligned}$$

Then by the near-optimality of $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$, we have

$$\begin{aligned}
\Delta &\geq LL_{OPT} - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \\
&\geq LL(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \boldsymbol{\sigma}^*) - LL(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \\
&\geq \left(-n \ln k + \frac{nd}{2} \ln \frac{1}{2\pi\sigma_o^2} - \frac{\Phi(\boldsymbol{\mu}^*)}{2\sigma_o^2} \right) - \left(n \ln k + \frac{nd}{2} \ln \left(\frac{1}{2\pi\sigma_o^2} \right) - \frac{\Phi(\boldsymbol{\mu})}{2(1+\delta)\sigma_o^2} \right) \\
&= \frac{\Phi(\boldsymbol{\mu})}{2(1+\delta)\sigma_o^2} - \frac{\Phi_{OPT}}{2\sigma_o^2} - 2n \ln k
\end{aligned}$$

Rearranging gives the theorem statement. ■

3.2 Proof of Theorem 4

Once again we reduce from k-means, using the hardness result of Aloise et al. (2009), summarized in Theorem 6. Recall that the family of instances for which k-means was shown to be hard has $k = 2$, $d = \text{poly}(n)$, and points with $\{-1, 0, 1\}$-valued coordinates.

Starting with such an instance $x_1, \ldots, x_n \in \mathbb{R}^d$, we show how MOG can be used to find a $(1 + 1/\beta(n))$-approximate solution to k-means.

- Let D denote the diameter of the points; it is polynomial in n.
- Set $\delta = 1/(5\beta(n))$ and
$$\sigma_o^2 = \frac{\delta}{2(1 + 2n \ln k)}.$$

- Pad the points with zero-valued coordinates to bring the dimension up to at least

$$d = \frac{4}{\delta}\left(\frac{nD^2}{2\sigma_o^2} + n\ln k + 1\right).$$

- Invoke MOG on these modified points, with target precision $b = 1$ and variance lower bound σ_o^2. This returns a mixture $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ whose log-likelihood is at least $LL_{OPT} - 1$, subject to the variance constraint.

- Return centers $\boldsymbol{\mu}$.

Lemma 9, with $\Delta = 1$, asserts that

$$\Phi(\boldsymbol{\mu}) \;\le\; (1+\delta)(2\sigma_o^2(1+2n\ln k) + \Phi_{OPT}) \;\le\; (1+\delta)(\delta + \Phi_{OPT}) \;\le\; (1+5\delta)\Phi_{OPT},$$

which is at most $(1 + 1/\beta(n))\Phi_{OPT}$. For the last inequality, we have used the fact that $\Phi_{OPT} \geq 1/2$ since all interpoint distances are ≥ 1.

Acknowledgments

The authors are grateful to the anonymous reviewers for their feedback and to the NSF for support under grants IIS-1162581 and DGE-1144086.

References

D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of the 33rd ACM Symposium on Theory of Computing*, pages 247–257, 2001.

P. Awasthi, M. Charikar, R. Krishnaswamy, and A. Sinop. The hardness of approximation of Euclidean k-means. In *Proceedings of the 31st International Symposium on Computational Geometry*, 2015.

M. Belkin and K. Sinha. Polynomial learning of distribution families. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, pages 103–112, 2010.

S. Dasgupta and Y. Freund. Random projection trees for vector quantization. *IEEE Transactions on Information Theory*, 55(7):3229–3242, 2009.

A.P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38, 1977.

M. Hardt and E. Price. Tight bounds for learning a mixture of two Gaussians. In *Proceedings of the 47th Annual ACM on Symposium on Theory of Computing*, pages 753–760, 2015.

D.J. Hsu and S.M. Kakade. Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, pages 11–20, 2013.

M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k-means problem is NP-hard. In *Proceedings of the 3rd International Workshop on Algorithms and Computation*, pages 274–285, 2009.

A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, pages 93–102, 2010.