Learning Mixtures of Gaussians with Maximum-a-posteriori Oracle

Satyaki Mahalanabis
Dept. of Computer Science
University of Rochester
smahalan@cs.rochester.edu

Abstract

We consider the problem of estimating the parameters of a mixture of distributions, where each component distribution is from a given parametric family e.g. exponential, Gaussian etc. We define a learning model in which the learner has access to a "maximum-a-posteriori" oracle which given any sample from a mixture of distributions, tells the learner which component distribution was the most likely to have generated it. We describe a learning algorithm in this setting which accurately estimates the parameters of a mixture of k spherical Gaussians in $\mathbb{R}^d$ assuming the component Gaussians satisfy a mild separation condition. Our algorithm uses only polynomially many (in d, k) samples and oracle calls, and our separation condition is much weaker than those required by unsupervised learning algorithms like [Arora 01, Vempala 02].

1 Introduction

Learning mixtures of distributions, such as mixture of Gaussians, is one of the most well-studied problems in machine learning (see e.g. [Melnykov 10]). However, most of the known learning algorithms work in an unsupervised setting and not much is known theoretically about how side information (e.g. in the form of class labels [Basu 02] or pairwise constraints [Shental 03]) can help the learner. In this paper, we define a learning model in which the learner has access to a "maximum-a-posteriori" oracle, which for any sample from a mixture of distributions, tells the learner which component distribution was most likely to have generated it. Our maximum-a-posteriori oracle can be thought of as an expert who can be asked to guess the most likely class label for a sample point. Then we demonstrate the advantage provided by this oracle by giving an algorithm (Algorithm 2) which recovers the parameters of a mixture of k, d-dimensional spherical Gaussians under a weaker separation assumption and more efficiently than known unsupervised algorithms. The number of oracle calls made by our algorithm is polynomial in d, k and is independent of the desired error.

The problem of learning mixture of Gaussians has a long history. Various learning models have been considered in literature. There is the clustering framework [Dasgupta 99, Arora 01] in which the goal is to correctly label each sample generated by the mixture with the Gaussian that generated it. Then there is the distribution learning model [Feldman 06] in which the goal is to output a hypothesis mixture which is close to the unknown mixture as possible. Probably the most popular model however is one in which the parameters of the original mixture need to be recovered e.g. [Dempster 77, Belkin 10] and this is the model we use. EM [Dempster 77] is perhaps the most well known algorithm in this model. Note that for the clustering problem, it is necessary to have a separation assumption [Arora 01], unlike for the problem of recovering parameters. Note also that our model is more difficult than learning the distribution.

Almost all algorithms for clustering or parameter learning which give provable guarantees also require that each Gaussian in the hypothesis mixture be separated from others [Vempala 02, Arora 01]. For the clustering framework, [Arora 01] give a lower bound on the separation for the case of spherical Gaussians. We also need to make an separation assumption (see (7)) which is much weaker than those required by e.g. [Arora 01, Vempala 02]. However we have access to a maximum-a-posteriori oracle whereas the settings in [Arora 01, Vempala 02, Dasgupta 99, Belkin 10, Moitra 10] are unsupervised. Recently [Belkin 10, Moitra 10] have given algorithms for learning parameters which require no separation assumption. However the absence of a separation assumption means that their algorithms, unlike ours or [Arora 01, Vempala 02], may require exponentially many samples in

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15. Copyright 2011 by the authors.

the number of mixture components. Our algorithm is also, not surprisingly, simpler than theirs.

We also note that our results are for spherical Gaussians only. The algorithm in [Dasgupta 99] works only when all Gaussians have the same covariance matrix. This was extended to Gaussians having arbitrary covariance matrices by [Arora 01]. The results in [Vempala 02], like ours, are for spherical Gaussians. They were the first to use spectral methods, and their technique has since been extended to arbitrary log-concave distributions [Kannan 08, Achlioptas 05]. Algorithms in [Belkin 10, Moitra 10] can learn mixtures of arbitrary gaussians.

Even though semi-supervised clustering has been a popular area of research e.g. [Basu 02, Xing 02, Basu 04], not much is known theoretically about semi-supervised or active learning of mixture of Gaussians. A variant of EM using pairwise constraints is given by [Shental 03]. They have two types of constraints: pairs of points which were generated by the same Gaussian, and pairs which were not. These type of pairwise relations seem to be the most popular model of semi-supervised learning of mixture models [Melnykov 10]. In our model we consider a maximum-a-posteriori oracle and as we point out in Section 2, learning with this oracle can be more challenging than with pairwise constraints. Our setting is different from semi-supervised learning because the learner is allowed to choose the points for which it wants labels. However the algorithm we give for mixtures of Gaussians queries the label of every sample it draws initially and later on uses only unlabeled samples.

Finally, while our aim in this paper is to theoretically demonstrate the advantage of the maximum-a-posteriori oracle, the oracle models an expert in a natural way and should find practical applications. The various UCI datasets discussed in the 'Experimental Results' section of [Shental 03] are examples where our oracle learning model could be applicable. Our Algorithm 2 can be used whenever a dataset can be modeled reasonably accurately using a mixture of Gaussians and side information is available in the form of which clusters selected data points belong to.

The paper is organized as follows. In Section 2 we first define our learning model (subsection 2.1), and then we state our result about learning mixture of Gaussians in this model (subsection 2.2). Then in Section 3 we justify the separation assumption that we require, and then present our algorithm and analyze it. Finally, in Section 4 we discuss how our model and our algorithm could be extended.

2 Summary of Contributions

We first define the maximum-a-posteriori oracle and then give our result for mixtures of gaussians.

2.1 Learning with Maximum-a-posteriori Oracle

We first define the learning model. Consider a mixture density in $\mathbb{R}^d$ defined as $\mu(x) = \sum_{i=1}^{k} w_i\mu_i(x)$, where $\sum_i w_i = 1$ and $w_{min} = \min_i w_i > 0$. The densities $\{\mu_i\}_{i=1}^k$ belong to a parametric family (e.g. uniform, Gaussian etc.) and have parameters $\{\theta_i\}_{i=1}^k$ respectively. Given an unknown mixture density, the goal of a learner is to output estimates $\{(\hat{\theta}_i, \hat{w}_i)\}_{i=1}^k$ of $\{(\theta_i, w_i)\}_{i=1}^k$ respectively. The learner can draw independent (unlabeled) samples from μ by invoking an oracle called SAMP. For any sample x returned by SAMP, the learner can also choose to invoke an oracle, MAP, which labels x with the most likely density that could have generated x i.e.

$$\begin{aligned} \mathrm{MAP}(x) &= \operatorname*{argmax}_i p_i(x) \quad \text{where} \\ \forall\, i,\ p_i(x) &= w_i\mu_i(x) \big/ \sum_j w_j\mu_j(x) \end{aligned} \tag{1}$$

is the posterior distribution on $\{1 \dots k\}$. For the case of mixture of Gaussians (with no two Gaussians being concentric), note that the set of points where 2 or more densities are equally likely has measure 0 and hence for a random sample from μ almost surely there is a unique most likely (a posteriori) density as defined by (1). We point out that in our model the number of mixture components k is known to the learner. Typically we have $k \ll d$.

We next compare our model with other similar settings. We point out that learning using the MAP oracle is more difficult than in a semi-supervised setting where each point in a random subset of samples is labeled according to which component *actually* generated it i.e. the label is drawn randomly from posterior $p_i()$. To see this consider a mixture of 2 or more identical almost concentric Gaussians in which one has a much greater weight than the others. If points are labeled according to the posterior, enough samples from each Gaussian can obtained to estimate the parameters with arbitrary accuracy whereas the MAP oracle would only return the label of the heaviest Gaussian and hence MAP does not help recover the parameters of the other Gaussians. Learning with the MAP oracle is more challenging than with pairwise constraints [Shental 03] as well because with sufficiently many constraints one can discover the actual label of every sample and recover all the parameters with arbitrary accuracy. Our model is different from that of [Castelli 96] where it is assumed that the parameters of each component are known (the mixing weights are unknown), and the learner just needs to be able to approximate the posterior.

The example of a mixture of 2 or more identical almost concentric Gaussians given above shows that in order for the MAP oracle to help the learner, the component densities need to be sufficiently separated. Such separation assumptions are often required for unsupervised learning (e.g. see

Table 1 for the case of mixture of Gaussians), though the separation we require should be smaller because we have help from MAP.

Finally note that any learning algorithm requires $\Omega(1/w_{min})$ unlabeled samples so that there are sufficiently many samples from each component from which to estimate their parameter.

2.2 Mixture of Gaussians

Let $N_{u,\Sigma}$ to denote a Gaussian[1] with mean u and covariance matrix Σ. The spherical Gaussian in $\mathbb{R}^d$ has density

$$N_{u,\sigma^2 I_d}(x) \;=\; \frac{1}{\left(\sqrt{2\pi}\sigma\right)^d}\mathrm{e}^{-\|x-u\|^2/2\sigma^2}.$$

where I_d is the $d \times d$ identity matrix, and $\| \;\|$ denotes the euclidean norm. We will demonstrate the advantage provided by the MAP oracle by giving an algorithm (Algorithm 2) which learns a mixture of Gaussians $\mu(x) = \sum_{i=1}^k w_i N_{u_i,\sigma_i^2 I_d}(x)$ at a smaller separation than previous unsupervised algorithms like [Dasgupta 99, Arora 01] (see Table 1). Our main result and separation condition are as follows.

Theorem 1. *(Restated from Theorem 11, Section 3.2) There is an algorithm which, for any d, $0 < \varepsilon < \frac{1}{10}$, $0 < \delta < \frac{1}{2}$ and for any mixture of k Gaussians $\mu = \sum_i w_i N_{u_i \sigma_i^2 I_d}$ satisfying the following separation condition*

$$\forall\, i \neq j,\; \frac{\|u_i - u_j\|}{\sqrt{\sigma_i^2+\sigma_j^2}} \geq 30\sqrt{\ln\left(144\sqrt{d}/\varepsilon\right) + \frac{3}{2}\ln\left(\frac{1}{w_{min}}\right)}, \tag{2}$$

draws at most $\frac{2\times 10^6 d}{w_{min}\varepsilon^2}\ln\left(\frac{4k(d+2)}{\delta}\right)$ (unlabeled) samples, makes $\frac{10^6 d}{w_{min}}\ln\left(\frac{4k(d+2)}{\delta}\right)$ calls to MAP, and returns $\{(\hat{u}_i, \hat{\sigma}_i, \hat{w}_i)\}_{i=1}^k$ such that with probability at least $1-\delta$,

$$\forall\, i,\; \|u_i - \hat{u}_i\| \;\leq\; \varepsilon\sigma_i,\; |\hat{\sigma}_i - \sigma_i| \leq \frac{\varepsilon}{\sqrt{d}}\sigma_i \text{ and } |\hat{w}_i - w_i| \leq \varepsilon w_i \tag{3}$$

See Theorem 11 and separation (25) for a more precise statement. Theorem 1 essentially says that the number of calls made by our algorithm to the MAP oracle in order to achieve an error of ε is independent of ε i.e. only $O(d\ln(dk)/w_{min})$, provided the separation between the Gaussians is only polylogarithmic in d, $1/w_{min}$.

The accuracy achieved (3) by our algorithm implies that each estimated component is statistically close to the true component : for each i, $\int \left|N_{u_i,\sigma_i^2 I_d} - N_{\hat{u}_i,\hat{\sigma}_i^2 I_d}\right| \leq 2\sqrt{2}\varepsilon$. This means the hypothesis mxture our algorithm outputs is within $O(\varepsilon)$ (w.r.t. L_1 or total variation distance) of the true mixture distribution.

Table 1: Sample complexity and separation for learning mixture of Gaussians (all algorithms except ours are unsupervised)

	Separation	**Complexity**
[Dasgupta 99]	$\Omega(d^{1/2})\max\{\sigma_i,\sigma_j\}$	$\tilde{O}(d\,\mathrm{poly}(k))$
[Dasgupta 07]	$\Omega(d^{1/4})\max\{\sigma_i,\sigma_j\}$	$\tilde{O}(dk^2)$
[Arora 01]	$\Omega(d^{1/4}\ln\; d)\max\{\sigma_i,\sigma_j\}$	$\tilde{O}(d^2\;\mathrm{poly}(k))$
[Vempala 02]	$\Omega(k^{1/4}\ln(dk))\max\{\sigma_i,\sigma_j\}$	$\tilde{O}(d^3k^2)$
[Belkin 10]	none	$poly(d)\;\mathrm{e}^{\mathrm{poly}(k)}$
[Moitra 10]	none	$poly(d)\;\mathrm{e}^{\mathrm{poly}(k)}$
Ours	$\Omega(\sqrt{\ln(dk)})\max\{\sigma_i,\sigma_j\}$	$\tilde{O}(dk)$

Table 1 compares[2] our separation assumption (2) with that required by previous unsupervised algorithms. We want to be able to learn the mixture parameters for as small a separation as possible, ideally at most polylogarithmic in d, k as in (2). Our required separation is smaller than that required by e.g. [Vempala 02] ($O(\max\{\sigma_i,\sigma_j\}k^{1/4}\mathrm{polylog}(d,k))$). Note that under separation (2), the overlap between the Gaussians is such that unsupervised distance based clustering algorithms (like [Arora 01, Vempala 02]) may not work. We point out that for spherical Gaussians [Arora 01] show that a separation of $\Omega(d^{1/4})\max\{\sigma_i,\sigma_j\}$ is necessary such that one can tell with high probability which Gaussian actually generated each sample from μ.

Recently [Belkin 10, Moitra 10] have given algorithms for mixture of Gaussians which do not require any separation assumption. However [Moitra 10] prove that any (unsupervised) learning algorithm for mixture of Gaussians with no separation assumption will require exponentially many samples in k. In contrast, our algorithm (provably) requires only polynomially many (in k) samples under a (weak) separation assumption and is much simpler than theirs.

3 Detailed Results

3.1 Notations and Auxiliary Lemmas

We will use $e_1, e_2, \ldots, e_d$ to denote the usual orthonormal basis of $\mathbb{R}^d$. Given vectors $u, v \in \mathbb{R}^d$, we will use $u \cdot v$ for their inner product. Φ will denote the cdf of the standard normal distribution $N_{0,1}$.

Lemma 2. *(Chernoff Bound, see e.g. [Dubhashi 09]) Let $X = \sum_{i=1}^n X_i$ where $\{X_i\}_{i=1}^n$ are independently dis-*

[1] We will use $N_{u,\Sigma}$ to denote both the density and the probability measure

[2] Table 1 does not explicitly show dependence on parameter w_{min} and instead assumes, for the sake of easy comparison, that $\frac{1}{k} \geq w_{min} \geq \frac{c}{k}$ for some constant $c > 0$. The separation in [Arora 01] is more complicated and can allow for concentric gaussians if σ_i, σ_j are different

tributed in $[0,1]$. *Then for any* $1 > \varepsilon > 0$,

$$\Pr\Big[(1-\varepsilon)\mathrm{E}\big[X\big] \;\le\; X \;\le\; (1+\varepsilon)\mathrm{E}\big[X\big]\Big] \;\le\; 2\mathrm{e}^{-\varepsilon^2\mathrm{E}[X]/3}. \tag{4}$$

Fact 3. *The KL-divergence between two Gaussians* $N_{u,\sigma^2 I_d}, N_{u',\sigma'^2 I_d}$ *is given by*
$KL\left(N_{u,\sigma^2 I_d}\|N_{u',\sigma'^2 I_d}\right) = \frac{\|u-u'\|^2+d\sigma^2}{2\sigma'^2} - \frac{d}{2} + d\ln\left(\frac{\sigma'}{\sigma}\right)$.

We will use the following property of Gaussians in proving Lemma 10.

Fact 4. *Let* $\rho = \frac{\Phi^{-1}(3/4)-\Phi^{-1}(1/4)}{2}$. *For any* $r \le \frac{1}{10}$, $\Delta > 0$ *and* x,

$$\Big|\frac{1}{2} - N_{u,\sigma^2}\left((-\infty,x]\right)\Big| \le r \;\Rightarrow\; |x-u| \le 5\,r\sigma, \text{ and} \tag{5}$$

$$\begin{aligned}&\big|N_{u,\sigma^2}([u-\rho\sigma,u+\rho\sigma]) - N_{u,\sigma^2}([u-\Delta,u+\Delta])\big| \le r\\ &\;\Rightarrow\; \big|\Delta - \rho\sigma\big| \le 2r\sigma.\end{aligned} \tag{6}$$

Definition 5. We will say that a mixture of k Gaussians $\sum_{i=1}^{k} w_i N_{u_i,\sigma_i^2 I_d}$ in $\mathbb{R}^d$ is β-separated, where $\beta > 0$, if the following pairwise separation condition holds -

$$\forall\, i \ne j \quad \frac{\|u_i - u_j\|}{\sqrt{\sigma_i^2+\sigma_j^2}} \;\ge\; 30\sqrt{\ln\left(\frac{2}{\beta}\right) + \frac{1}{2}\left|\ln\left(\frac{w_j}{w_i}\right)\right|}. \tag{7}$$

Our required separation condition (2) implies that the mixture to be learned is $\beta = \frac{\varepsilon w_{min}}{72\sqrt{d}}$-separated. We also need to define, for all $i,j,\ i \ne j$

$$A_{ij} = \left\{w_i N_{u_i,\sigma_i^2 I_d}(x) \;\ge\; w_j N_{u_j,\sigma_j^2 I_d}(x)\right\} \text{ and} A_i = \bigcap_{j\ne i} A_{ij}. \tag{8}$$

i.e A_{ij} is the set where Gaussian i is more likely than j and A_i is the set where i is the most likely Gaussian. Note that A_i is precisely the set of points for which MAP will return label i.

Consider two Gaussians which are far apart, and a third Gaussian which is much closer to the first Gaussian than the second. The following Lemma shows that with high probability, the first Gaussian is more likely than the second one at points drawn from the third Gaussian.

Lemma 6. *Consider two Gaussians* $N_{u,\sigma^2 I_d}, N_{u',\sigma'^2 I_d}$. *Let* $N_{\tilde{u},\tilde{\sigma}^2 I_d}$ *be another Gaussian such that*

$$\begin{aligned}\|u-\tilde{u}\| &\le \frac{1}{10}\min\{\tilde{\sigma},\ \|u'-u\|\} \textit{ and}\\ |\sigma - \tilde{\sigma}| &\le \min\left\{\frac{\tilde{\sigma}}{\sqrt{d}},\ \frac{\tilde{\sigma}}{10}\right\}.\end{aligned} \tag{9}$$

Then for any $t > 0$,

$$\begin{aligned}&N_{\tilde{u},\tilde{\sigma}^2 I_d}\left(\left\{N_{u,\sigma^2 I_d}(x) \;\ge\; t\, N_{u',\sigma'^2 I_d}(x)\right\}\right)\\ &\qquad\qquad \ge 1 - \sqrt{t}\,\mathrm{e}^{-\frac{\|u'-u\|^2}{80(\sigma^2+\sigma'^2)}+3}.\end{aligned} \tag{10}$$

Proof :
Assume w.l.g that $\tilde{u} = 0$. We have from Markov's inequality that for $X \sim N_{\tilde{u},\tilde{\sigma}^2 I_d}$,

$$\begin{aligned}&N_{\tilde{u},\tilde{\sigma}^2 I_d}\left(\left\{N_{u,\sigma^2 I_d}(x) \;\le\; t\, N_{u',\sigma'^2 I_d}(x)\right\}\right)\\ &= \Pr\left[\sqrt{\frac{N_{u',\sigma'^2 I_d}(X)}{N_{u,\sigma^2 I_d}(X)}} \ge \sqrt{\frac{1}{t}}\right]\\ &\le \sqrt{t}\,\mathrm{E}\left[\sqrt{\frac{N_{u',\sigma'^2 I_d}(X)}{N_{u,\sigma^2 I_d}(X)}}\right]. \qquad (11)\\ &= \sqrt{t}\int\left(\frac{\mathrm{e}^{-\frac{\|x-u'\|^2}{4\sigma'^2}}}{\mathrm{e}^{-\frac{\|x-u\|^2}{4\sigma'^2}}}\frac{\sigma^{d/2}}{\sigma'^{d/2}}\right)\frac{\mathrm{e}^{-\frac{\|x\|^2}{2\tilde{\sigma}^2}}}{\left(\sqrt{2\pi}\tilde{\sigma}\right)^d}\mathrm{d}x\\ &= \sqrt{t}\left(\frac{\sigma r^2}{\sigma'\tilde{\sigma}^2}\right)^{d/2}\mathrm{e}^{\frac{v^2r^2}{8}-a},\end{aligned}$$

where $v = \frac{u'}{\sigma'^2} - \frac{u}{\sigma^2}$, $\frac{1}{r^2} = \frac{1}{2\sigma'^2} + \frac{1}{\tilde{\sigma}^2} - \frac{1}{2\sigma^2}$ and $a = \frac{u'^2}{4\sigma'^2} - \frac{u^2}{4\sigma^2}$. For $|\sigma - \tilde{\sigma}| \le \frac{\tilde{\sigma}}{\sqrt{d}}$, $\frac{\sigma r^2}{\sigma'\tilde{\sigma}^2} \le \frac{1}{1-5/d}$. Also, under assumption (9), $\frac{v^2r^2}{8} - a \le -\frac{\|u'-u\|^2}{80(\sigma^2+\sigma'^2)} + \frac{1}{81}$. On substituting these in (11), we get

$$\begin{aligned}&N_{\tilde{u},\tilde{\sigma}^2 I_d}\left(\left\{N_{u,\sigma^2 I_d}(x) \;\le\; t\, N_{u',\sigma'^2 I_d}(x)\right\}\right)\\ &\qquad \le\; \sqrt{t}\mathrm{e}^{-\frac{\|u'-u\|^2}{80(\sigma^2+\sigma'^2)}+3}\ ,\end{aligned}$$

which gives us (10). ■

Next using Lemma 6 we show that in a β-separated mixture, with high probability, component i is more likely *a posteriori* than any other component $j \ne i$ at a random point generated by component i itself.

Lemma 7. *For any* $\beta > 0$ *and any* β*-separated mixture of* k *Gaussians* $\sum_i\ w_i N_{u_i,\sigma_i^2 I_d}$, *for all* $i,j,\ i \ne j$,

$$N_{u_i,\sigma_i^2 I_d}\left(A_{ij}\right) > 1 - \beta. \tag{12}$$

Proof :
(Proof of Lemma 7) Applying Lemma 6 to Gaussians i,j, we have

$$N_{u_i,\sigma_i^2 I_d}\left(A_{ij}\right) \ge 1 - \sqrt{\frac{w_j}{w_i}}\mathrm{e}^{-\frac{\|u_j-u_i\|^2}{80(\sigma_i^2+\sigma_j^2)}+3}.$$

Note that condition (9) required for Lemma 6 is satisfied trivially since in this case $N_{u,\sigma^2 I_d}, N_{\tilde{u},\tilde{\sigma}^2 I_d}$ in Lemma 6 are the same. Lemma 7 now follows by substituting separation (7) in the bound above. ■

Remark 8. The separation in (7) is the minimum required (up to a constant factor) such that (12) holds i.e. such that the i^{th} Gaussian is more likely than others at most points generated by itself. One can show this by using the following inequality due to Bretagnolle and Huber (see e.g. [Devroye 01] Chapter 5, Exercise 5.6) for densities f_1, f_2-

$$\left(\frac{1}{2}\int |f_1 - f_2|\right)^2 \leq 1 - \mathrm{e}^{KL(f_1 \| f_2)}.$$

Applying this to $N_{u_i,\sigma_i^2 I_d}, N_{u_j,\sigma_j^2 I_d}$ for the case $w_i = w_j$ yields, using Fact 3,

$$\begin{aligned} N_{u_i,\sigma_i^2 I_d}(A_{ij}) &= \frac{1}{2}\int \left|N_{u_i,\sigma_i^2 I_d} - N_{u_j,\sigma_j^2 I_d}\right| \\ &\leq 1 - \mathrm{e}^{-\frac{1}{2}\left(\frac{\|u_i - u_j\|}{\max\{\sigma_i,\sigma_j\}}\right)^2} \end{aligned}$$

which shows that separation (7) is tight.

The following Lemma will be used to show that in Algorithm 2 step 2, when the estimated parameters $\{(\bar{u}_i, \bar{\sigma}_i, \bar{w}_i\}_{i=1}^k$ are sufficiently close to the true parameters $\{(u_i, \sigma_i, w_i)\}_{i=1}^k$, component $N_{\bar{u}_i,\bar{\sigma}_i^2 I_d}$ is more likely than any other $N_{\bar{u}_j,\bar{\sigma}_j^2 I_d}$ at most points generated by the i^{th} component $N_{u_i,\sigma_i^2 I_d}$. This means that once our estimates have converged close to the true parameters, we need not make any more calls to the MAP oracle and instead approximate MAP using $\{(\bar{u}_i, \bar{\sigma}_i, \bar{w}_i\}_{i=1}^k$.

Lemma 9. *Let $\beta > 0$. Let $\sum_i w_i \; N_{u_i,\sigma_i^2 I_d}$ be a β-separated mixture of k Gaussians, and $\sum_i \; \bar{w}_i \; N_{\bar{u}_i,\bar{\sigma}_i^2 I_d}$ be another mixture of k Gaussians such that*

$$\forall\, i \quad \|\bar{u}_i - u_i\| \leq \frac{\sigma_i}{10},\; |\bar{\sigma}_i - \sigma_i| \leq \frac{\sigma_i}{\sqrt{d}} \text{ and } |\bar{w}_i - w_i| \leq \frac{w_i}{10}. \tag{13}$$

Then for all $i, j,\; i \neq j$,

$$N_{u_i,\sigma_i^2 I_d}\left(\left\{\bar{w}_i N_{\bar{u}_i,\bar{\sigma}_i^2 I_d} \geq \bar{w}_j N_{\bar{u}_j,\bar{\sigma}_j^2 I_d}\right\}\right) \geq 1 - \beta$$

Proof :
(Proof of Lemma 9) Condition (13) and β-separation (7) imply that condition (9) required for Lemma 6 is satisified for Gaussians $N_{\bar{u}_i,\bar{\sigma}_i^2 I_d}, N_{\bar{u}_j,\bar{\sigma}_j^2 I_d}$ and $N_{u_i,\sigma_i^2 I_d}$. Hence by Lemma 6,

$$\begin{aligned} &N_{u_i,\sigma_i^2 I_d}\left(\left\{\bar{w}_i N_{\bar{u}_i,\bar{\sigma}_i^2 I_d} \geq \bar{w}_j N_{\bar{u}_j,\bar{\sigma}_j^2 I_d}\right\}\right) \\ &\geq 1 - \sqrt{\frac{\bar{w}_j}{\bar{w}_i}}\, \mathrm{e}^{-\frac{\|\bar{u}_i - \bar{u}_j\|^2}{80(\bar{\sigma}_i^2 + \bar{\sigma}_j^2)} + 3} \geq 1 - \sqrt{\frac{w_j}{w_i}}\, \mathrm{e}^{-\frac{\|u_i - u_j\|^2}{160(\sigma_i^2 + \sigma_j^2)} + 4} \end{aligned} \tag{14}$$

where we have used the fact that under separation (7) and (13), $\|\bar{u}_i - \bar{u}_j\| \geq \frac{4}{5}\|u_i - u_j\|$ and $\frac{9}{10} \leq \frac{w_i}{\bar{w}_i}, \frac{w_j}{\bar{w}_j} \leq \frac{11}{10}$ and $\frac{9}{10} \leq \frac{\sigma_i}{\bar{\sigma}_i}, \frac{\sigma_j}{\bar{\sigma}_j} \leq \frac{11}{10}$. Substituting (7) in (14) gives the Lemma. ∎

3.2 Learning Algorithm and Analysis

We first state a subroutine EstimateParameters which takes a set of labeled samples as input and uses a sample median-based estimator for computing the parameters of each Gaussian component. This subroutine is used by the main Algorithm 2.

```
Subroutine 1: EstimateParameters
Input: Set $S$ of labeled samples.
Output: $\{(\tilde{u}_i, \tilde{\sigma}_i, \tilde{w}_i)\}_{i=1}^k$
begin
    for $i = 1 \ldots k$ do
        $S_i \leftarrow \{x \mid (x,i) \in S\}$ ;
1       $\tilde{u}_i \leftarrow \left(\operatorname*{median}_{x \in S_i} x \cdot e_1, \operatorname*{median}_{x \in S_i} x \cdot e_2, \ldots, \operatorname*{median}_{x \in S_i} x \cdot e_d\right)$;
2       $\tilde{\sigma}_i \leftarrow \frac{2}{\Phi^{-1}(3/4) - \Phi^{-1}(1/4)} \operatorname*{median}_{x \in S_i} |x \cdot e_1 - \tilde{u}_i \cdot e_1|$ ;
        $\tilde{w}_i \leftarrow |S_i| / (|S_1| + |S_2| + \ldots + |S_k|)$ ;
    end
end
```

For estimating $\tilde{\sigma}_i$ in step 1 of EstimateParameters, we could have projected the points in S_i along any e_l instead of e_1. Intuitively, if the points in S_i are drawn from the exact distribution $N_{u_i,\sigma^2 I_d}$, then $\tilde{u}_i = u_i,\; \tilde{\sigma}_i = \sigma_i$ in expectation. However in our case, because of the overlap between Gaussians there will be some samples in S_i which were generated by other components $j \neq i$. We use the sample median to filter out these outliers.

A naïve way learning algorithm would be to draw $O\left(\frac{d \ln(dk)}{w_{min}\; \varepsilon^2}\right)$ unlabeled samples and then invoke MAP on each of them. The points labeled i can then be used to compute the parameters of the i^{th} Gaussian. However we show below that Algorithm 2 actually makes fewer calls to MAP by requesting labels only initially. The number of labels required thus decreases to $O\left(\frac{d \ln(dk)}{w_{min}}\right)$ (i.e. independent of ε).

Algorithm 2 is based on Lemma 7 and Lemma 9. The parameters m, m' are set as stated in Theorem 11. Algorithm 2 works in two phases. In the first phase (steps 1, 2) it uses the MAP oracle to label each random sample.The number of samples m is sufficiently large for the estimated centres $\{\bar{u}_i\}_{i=1}^k$ to be within constant distance $O(1)$ of the respective true centres $\{u_i\}_{i=1}^k$. In the second phase (steps 3, 4, 5) Algorithm 2 does not need to use MAP at all. Instead it uses estimates $\{(\bar{u}_i, \bar{\sigma}_i, \bar{w}_i)\}_{i=1}^k$ from the first phase to approximate the MAP oracle (see Lemma 9), and the final estimated centres $\{\bar{u}_i\}_{i=1}^k$ are guaranteed to be within distance ε of the respective true ones.

We first analyze subroutine EstimateParameters before an-

Algorithm 2: Learning Algorithm for Mixture of Gaussians

```
Input: Sample sizes $m, m'$
Output: $\{(\hat u_i, \hat\sigma_i, \hat w_i)\}_{i=1}^k$
begin
    $S \leftarrow \emptyset$; $S' \leftarrow \emptyset$ ;
    for $i = 1 \dots m$ do
        $x \leftarrow$ SAMP() ; $l \leftarrow$ MAP($x$) ;
        $S \leftarrow S \bigcup \{(x, l)\}$ ;
    end
    $\{(\bar u_i, \bar\sigma_i, \bar w_i)\}_{i=1}^k \leftarrow$ EstimateParameters($S$) ;
    for $i = 1 \dots m'$ do
        $x \leftarrow$ SAMP() ; $l^* \leftarrow \operatorname{argmax}_l \bar w_l N_{\bar u_l, \bar\sigma_l^2 I_d}(x)$ ;
        $S' \leftarrow S' \bigcup \{(x, l^*)\}$ ;
    end
    $\{(\hat u_i, \hat\sigma_i, \hat w_i)\}_{i=1}^k \leftarrow$ EstimateParameters($S'$) ;
end
```

alyzing Algorithm 2.

Lemma 10. *Let $\frac{1}{2} > \delta' > 0$ and $\frac{1}{100} > \varepsilon' > 0$, $\frac{1}{100} > \beta' > 0$. Let the input S to EstimateParameters be such that the points $\{x \mid \exists\ i\ (x, i) \in S\}$ are independent samples from the mixture of k Gaussians $\sum\ w_i N_{u_i,\sigma_i^2 I_d}$. Assume $|S| \geq \frac{6}{w_{min}\varepsilon'^2}\ln(2k(d+2)/\delta')$. Further, in EstimateParameters assume that for each $i = 1 \dots k$, the set S_i (i.e. points labeled i) consists of independent samples from a density μ_i', where*

$$\int \Big|\mu_i' - N_{u_i,\sigma_i^2 I_d}\Big| \;\leq\; \beta' , \qquad \textit{and that} \tag{15}$$

$$w_i(1-\beta') \leq \mathrm{E}\big[|S_i|/|S|\big] \leq w_i(1+\beta'). \tag{16}$$

Then with probability at least $1 - \delta'$,

$$\forall\ i\ \|\tilde u_i - u_i\| \leq 5(\varepsilon' + \beta')\sqrt{d}\sigma_i,\ |\tilde\sigma_i - \sigma_i| \leq 9(\varepsilon' + \beta')\sigma_i \textit{ and } |\tilde w_i - w_i| \leq 2(\varepsilon' + \beta')w_i.$$

Proof :
(Proof of Lemma 10) From Chernoff bound (4) and assumption (16), it follows that given $|S| \geq \frac{6}{w_{min}\varepsilon'^2}\ln(2k(d+2)/\delta')$, with probability at least $1 - \frac{\delta'}{d+2}$, for each $i \leq k$, $|S_i| \geq \frac{6}{\varepsilon'^2}\ln(2(d+2)k/\delta')$ and $|\tilde w_i - w_i| \leq 2(\varepsilon' + \beta')w_i$.

Consider any e_l. For each i, it follows from condition (15) that

$$\left|\Pr_{X\sim\mu_i'}\big[X \cdot e_l \leq \tilde u_i \cdot e_l\big] - N_{u_i\cdot e_l,\sigma_i^2}\big((-\infty, \tilde u_i \cdot e_l]\big)\right| \;\leq\; \beta'. \tag{17}$$

Let $F_{i,l}$ denote the empirical distribution of points in S_i projected onto e_l i.e. $F_{i,l}(A) = \big|\{x \in S_i \mid x \cdot e_l \in A\}\big|/\big|S_i\big|$. Now $\tilde u_i \cdot e_l$ is the median of $F_{i,l}$. If $|S_i| \geq \frac{6}{\varepsilon'^2}\ln(2(d+2)k/\delta')$, it follows from the Chernoff bound (4) that with probability at least $1 - \frac{\delta'}{k(d+2)}$,

$$\begin{aligned} &\left|F_{i,l}((-\infty, \tilde u_i \cdot e_l]) - \Pr_{X\sim\mu_i'}\big[X \cdot e_l \leq \tilde u_i \cdot e_l\big]\right| \\ &= \left|\frac{1}{2} - \Pr_{X\sim\mu_i'}\big[X \cdot e_l \leq \tilde u_i \cdot e_l\big]\right| \;\leq\; \varepsilon'. \end{aligned} \tag{18}$$

Combining (17), (18) we have

$$\left|\frac{1}{2} \;-\; N_{u_i\cdot e_l,\sigma_i^2}\big((-\infty, \tilde u_i \cdot e_l]\big)\right| \;\leq\; \beta' + \varepsilon'. \tag{19}$$

Now (19) and bound (5) in Fact 4 together imply that $|\tilde u_i \cdot e_l - u_i \cdot e_l| \leq 5(\beta' + \varepsilon')\sigma_i$. Hence by applying the union bound over all directions $l = 1 \dots d$, we have that with probability at least $1 - \frac{d\delta'}{k(d+2)}$, $\|\tilde u_i - u_i\| \leq 5(\beta' + \varepsilon')\sqrt{d}\sigma_i$.

For $\tilde\sigma_i$ a similar argument applies. Define $\rho = \frac{\Phi^{-1}(3/4) - \Phi^{-1}(1/4)}{2}$. From condition (15), it follows that

$$\begin{aligned} &\bigg|\Pr_{X\sim\mu_i'}\big[|X \cdot e_1 - \tilde u_i \cdot e_1| \leq \rho\tilde\sigma_i\big] \\ &\quad - \; N_{u_i\cdot e_1,\sigma_i^2}\big([\tilde u_i \cdot e_1 - \rho\tilde\sigma_i, \tilde u_i \cdot e_1 + \rho\tilde\sigma_i]\big)\bigg| \;\leq\; \beta'. \end{aligned} \tag{20}$$

Note that $F_{i,l}\big([\tilde u_i \cdot e_1 - \rho\tilde\sigma_i, \tilde u_i \cdot e_1 + \rho\tilde\sigma_i]\big) = \frac{1}{2}$ by definition. If $|S_i| \geq \frac{6}{\varepsilon'^2}\ln(2k(d+2)/\delta')$ then by the Chernoff bound (4), with probability at least $1 - \frac{\delta'}{k(d+2)}$,

$$\begin{aligned} &\bigg|F_{i,l}\big([\tilde u_i \cdot e_1 - \rho\tilde\sigma_i, \tilde u_i \cdot e_1 + \rho\tilde\sigma_i]\big) \\ &\quad - \Pr_{X\sim\mu_i'}\big[|X \cdot e_1 - \tilde u_i \cdot e_1| \leq \rho\tilde\sigma_i\big]\bigg| \\ &= \left|\frac{1}{2} - \Pr_{X\sim\mu_i'}\big[|X \cdot e_1 - \tilde u_i \cdot e_1| \leq \rho\tilde\sigma_i\big]\right| \;\leq\; \varepsilon'. \end{aligned} \tag{21}$$

Combining (20), (21) we get

$$\begin{aligned} &\bigg|N_{u_i\cdot e_1,\sigma_i^2}\big([u_i \cdot e_1 - \rho\sigma_i, u_i \cdot e_1 + \rho\sigma_i]\big) \\ &\quad - N_{u_i\cdot e_1,\sigma_i^2}\big([\tilde u_i \cdot e_1 - \rho\tilde\sigma_i, \tilde u_i \cdot e_1 + \rho\tilde\sigma_i]\big)\bigg| \\ &= \left|\frac{1}{2} - N_{u_i\cdot e_1,\sigma_i^2}\big([\tilde u_i \cdot e_1 - \rho\tilde\sigma_i, \tilde u_i \cdot e_1 + \rho\tilde\sigma_i]\big)\right| \\ &\leq \beta' + \varepsilon'. \end{aligned} \tag{22}$$

Now it follows from (19) that

$$\begin{aligned} &\bigg|N_{u_i\cdot e_1,\sigma_i^2}\big([\tilde u_i \cdot e_1 - \rho\tilde\sigma_i, \tilde u_i \cdot e_1 + \rho\tilde\sigma_i]\big) \\ &\quad - N_{u_i\cdot e_1,\sigma_i^2}\big([u_i \cdot e_1 - \rho\tilde\sigma_i, u_i \cdot e_1 + \rho\tilde\sigma_i]\big)\bigg| \\ &\leq\; 2(\beta' + \varepsilon'). \end{aligned} \tag{23}$$

Combining (22),(23) we have

$$\begin{aligned}
&\Big| N_{u_i\cdot e_1,\sigma_i^2}\big([u_i\cdot e_1-\rho\sigma_i, u_i\cdot e_1+\rho\sigma_i]\big)\\
&\quad - N_{u_i\cdot e_1,\sigma_i^2}\big([u_i\cdot e_1-\rho\tilde{\sigma}_i, u_i\cdot e_1+\rho\tilde{\sigma}_i]\big)\Big|\\
&\le \Big| N_{u_i\cdot e_1,\sigma_i^2}\big([u_i\cdot e_1-\rho\sigma_i, u_i\cdot e_1+\rho\sigma_i]\big)\\
&\quad - N_{u_i\cdot e_1,\sigma_i^2}\big([\tilde{u}_i\cdot e_1-\rho\tilde{\sigma}_i, \tilde{u}_i\cdot e_1+\rho\tilde{\sigma}_i]\big)\Big|\\
&\quad + \Big| N_{u_i\cdot e_1,\sigma_i^2}\big([\tilde{u}_i\cdot e_1-\rho\tilde{\sigma}_i, \tilde{u}_i\cdot e_1+\rho\tilde{\sigma}_i]\big)\\
&\quad - N_{u_i\cdot e_1,\sigma_i^2}\big([u_i\cdot e_1-\rho\tilde{\sigma}_i, u_i\cdot e_1+\rho\tilde{\sigma}_i]\big)\Big|\\
&\le (\beta'+\varepsilon')+2(\beta'+\varepsilon') \;=\; 3(\beta'+\varepsilon').
\end{aligned} \tag{24}$$

It follows from (24) and bound (6), Fact 4 that $|\tilde{\sigma}_i-\sigma_i| \le \frac{6}{\rho}(\beta'+\varepsilon')\sigma_i < 9(\beta'+\varepsilon')\sigma_i$ with probability at least $1-\frac{\delta'}{k(d+2)}$.
Hence if $|S_i| \ge \frac{6}{\varepsilon'^2}\ln(2k(d+2)/\delta')$ then with probability at least $1-\frac{\delta'}{k}$, $\|\tilde{u}_i-u_i\| \le 5(\varepsilon'+\beta')\sqrt{d}\sigma_i$, $|\tilde{\sigma}_i-\sigma_i| \le 9(\varepsilon'+\beta')\sigma_i$ and $|\tilde{w}_i-w_i| \le 2(\varepsilon'+\beta')w_i$. The Lemma now follows by applying the union bound to all $i=1\ldots k$. ■

We finally state our main theorem.

Theorem 11. *Let $0<\varepsilon<\frac{1}{10}$, $0<\delta<\frac{1}{2}$. Then for any mixture of k Gaussians $\mu=\sum_i w_i N_{u_i\sigma_i^2 I_d}$ satisfying the following separation condition*

$$\forall\, i\neq j,\ \frac{\|u_i-u_j\|}{\sqrt{\sigma_i^2+\sigma_j^2}} \ge 30\sqrt{\ln\left(144\sqrt{d}/\varepsilon\right)+\frac{3}{2}\ln\left(\frac{1}{w_{min}}\right)}. \tag{25}$$

Algorithm 2 with $m=\frac{10^6 d}{w_{min}}\ln\left(\frac{4k(d+2)}{\delta}\right)$ and $m'=\frac{2000d}{w_{min}\varepsilon^2}\ln\left(\frac{4k(d+2)}{\delta}\right)$ returns $\{(\hat{u}_i,\hat{\sigma}_i,\hat{w}_i)\}_{i=1}^k$ such that with probability at least $1-\delta$,

$$\forall\, i,\ \|u_i-\hat{u}_i\| \le \varepsilon\sigma_i,\ |\hat{\sigma}_i-\sigma_i| \le \frac{\varepsilon}{\sqrt{d}}\sigma_i \text{ and } |\hat{w}_i-w_i| \le \varepsilon w_i \tag{26}$$

Proof :
(Proof of Theorem 11) Consider running Algorithm 2 with parameters $m=\frac{10^6 d}{w_{min}}\ln(4k(d+2)/\delta)$, $m'=\frac{2000d}{w_{min}\varepsilon^2}\ln(4k(d+2)/\delta)$. Note that Algorithm 2 makes only m calls to MAP.

We first analyze steps 1, 2. Under separation (25), our mixture μ is (βw_{min})-separated where $\beta=\frac{\varepsilon}{72\sqrt{d}}$. Therefore by Lemma 7 we have

$$\forall\, i\neq j \quad N_{u_i,\sigma_i^2 I_d}(A_{ij}) \;\ge\; 1-\beta w_{min} \;\ge\; 1-\frac{\beta}{k} \tag{27}$$

which implies that

$$\forall\, i\neq j,\ N_{u_j,\sigma_j^2 I_d}(A_i) \le \beta w_{min} \text{ and } N_{u_i,\sigma_i^2 I_d}(A_i) \ge 1-\beta, \tag{28}$$

where we have used the union bound. From (28) it follows that

$$w_i(1-\beta) \le \sum_j w_j N_{u_j,\sigma_j^2 I_d}(A_i) \le w_i(1+\beta). \tag{29}$$

Now note that for each i, set of points S_i which are labeled i in step 1 are drawn independently from μ_i' where

$$\mu_i'(x)=\begin{cases}\frac{\sum_j w_j N_{u_j,\sigma_j^2 I_d}(x)}{\sum_j w_j N_{u_j,\sigma_j^2 I_d}(A_i)} & \text{if } x\in A_i\\ 0 & \text{otherwise.}\end{cases} \tag{30}$$

Using (27), (28) and (29) we have that for each i,

$$\begin{aligned}
&\int \left|\mu_i' - N_{u_i,\sigma_i^2 I_d}\right|\\
&= N_{u_i,\sigma_i^2 I_d}\left(\mathbb{R}^d\setminus A_i\right) + \int_{A_i}\left|\mu_i' - N_{u_i,\sigma_i^2 I_d}\right|\\
&\le N_{u_i,\sigma_i^2 I_d}\left(\mathbb{R}^d\setminus A_i\right) + \sum_{j\neq i}\int_{A_i}\frac{w_j N_{u_j,\sigma_j^2 I_d}}{\sum_j w_j N_{u_j,\sigma_j^2 I_d}(A_i)}\\
&+\int_{A_i}\left|1-\frac{w_i}{\sum_j w_j N_{u_j,\sigma_j^2 I_d}(A_i)}\right| N_{u_i,\sigma_i^2 I_d}\\
&\le \beta + \frac{\beta w_{min}}{w_i(1-\beta)} + \frac{\beta}{1-\beta} \;<\; 4\beta.
\end{aligned} \tag{31}$$

Now $|S|=m\ge\frac{10^6 d}{w_{min}}\ln(4k(d+2)/\delta)$ and by (29), $w_i(1-\beta) \le \mathrm{E}\big[|S_i|/|S|\big] = \sum_j w_j N_{u_j,\sigma_j^2 I_d}(A_i) \le w_i(1+\beta)$. Hence we can apply Lemma 10 (with $\varepsilon'=\frac{1}{180\sqrt{d}}, \beta'=4\beta=\frac{\varepsilon}{18\sqrt{d}}, \delta'=\frac{\delta}{2}$) to the call to EstimateParameters in step 2. Thus Lemma 10 implies that with probability at least $1-\frac{\delta}{2}$

$$\begin{aligned}
&\forall\, i \quad \|\bar{u}_i-u_i\| \le \frac{1}{18}\sigma_i,\ |\bar{\sigma}_i-\sigma_i| \le \frac{1}{10\sqrt{d}}\sigma_i,\\
&\text{and } |\bar{w}_i-w_i| \le \frac{1}{45\sqrt{d}}w_i.
\end{aligned} \tag{32}$$

Next we analyze steps 3, 4 and 5. For each i, the set of samples S_i' which are labeled i in step 3 are drawn independently from distribution μ_i'', defined as

$$\mu_i''(x) \;=\begin{cases}\frac{\sum_j w_j N_{u_j,\sigma_j^2 I_d}(x)}{\sum_j w_j N_{u_j,\sigma_j^2 I_d}(\bar{A}_i)} & \text{if } x\in \bar{A}_i\\ 0 & \text{otherwise,}\end{cases} \tag{33}$$

where $\bar{A}_i=\bigcap_{j\neq i}\bar{A}_{ij}$ and $\bar{A}_{ij}=\big\{\bar{w}_i N_{\bar{u}_j,\bar{\sigma}_j^2 I_d}(x) \ge \bar{w}_j N_{\bar{u}_j,\bar{\sigma}_j^2 I_d}(x)\big\}$.

If (32) holds, by Lemma 9 we have for each $i \neq j$, $N_{u_i,\sigma_i^2 I_d}\left(\bar{A}_{ij}\right) \geq 1 - \beta$. This implies (proof is similar to that of (31), (29) above) that for each i,

$$\int \left|\mu_i'' - N_{u_i,\sigma_i^2 I_d}\right| \leq 4\beta = \frac{\varepsilon}{18\sqrt{d}} \text{ and}$$
$$w_i\left(1-\beta\right) \leq \mathrm{E}\left[\left|S_i'\right|/\left|S'\right|\right] \leq w_i\left(1+\beta\right).$$

Also $|S'| = m' \geq \frac{2000d}{w_{min}\varepsilon^2}\ln(4k(d+2)/\delta)$. So if (32) holds, we can apply Lemma 10 to the call to EstimateParameters in step 5 with $\varepsilon' = \frac{\varepsilon}{18\sqrt{d}}, \beta' = 4\beta = \frac{\varepsilon}{18\sqrt{d}}, \delta' = \frac{\delta}{2}$. This gives us that with probability at least $1 - \frac{\delta}{2}$,

$$\begin{aligned} &\forall\, i \quad \|\hat{u}_i - u_i\| < \frac{5\varepsilon}{9}\sigma_i,\ |\hat{\sigma}_i - \sigma_i| < \frac{\varepsilon}{\sqrt{d}}\sigma_i \\ &\text{and } |\hat{w}_i - w_i| < \frac{\varepsilon}{4\sqrt{d}}w_i. \end{aligned} \tag{34}$$

Finally by the union bound (over the two calls to EstimateParameters) we have that (34) holds with probability at least $1 - \delta$. ■

4 Conclusion

We have described an algorithm, Algorithm 2 which, given random samples generated by a unknown mixture of Gaussians and the maximum-a-posteriori oracle (1), is able to recover its parameters assuming only a mild separation condition (2) between the component Gaussians holds. There are a number of ways of extending our algorithm. For instance we can prove the correctness of our algorithm only for spherical Gaussians, and it is likely that one can modify our algorithm and the separation condition to work for arbitrary Gaussians as well. One can also ask how our algorithm could better utilize unlabeled samples - this could probably give a better error bound and a better label complexity. Finally, one can try to give algorithms for more realistic settings by allowing for the unlabeled samples or the oracle to be noisy.

Acknowledgment

I am grateful to Daniel Štefankovič for helpful discussions and for suggesting improvements to an earlier version of this paper. I would also like to thank the anonymous reviewers for helpful comments. This material is based upon work supported by National Science Foundation grants CCF-0910415, IIS-0428020, ISS-0325646 and University of Pennsylvania subcontract NSF # 5419844. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of above named organizations.

References

[Achlioptas 05] Dimitris Achlioptas & Frank McSherry. *On Spectral Learning of Mixtures of Distributions*. In COLT, pages 458–469, 2005.

[Arora 01] Sanjeev Arora & Ravi Kannan. *Learning mixtures of arbitrary gaussians*. In STOC, pages 247–257, 2001.

[Basu 02] Sugato Basu, Arindam Banerjee & Raymond J. Mooney. *Semi-supervised Clustering by Seeding*. In ICML, pages 27–34, 2002.

[Basu 04] Sugato Basu, Arindam Banerjee & Raymond J. Mooney. *Active Semi-Supervision for Pairwise Constrained Clustering*. In SDM, 2004.

[Belkin 10] Mikhail Belkin & Kaushik Sinha. *Polynomial Learning of Distribution Families*. In IEEE FOCS, 2010.

[Castelli 96] Vittorio Castelli & Thomas M. Cover. *The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter*. IEEE Transactions on Information Theory, vol. 42, no. 6, pages 2102–2117, 1996.

[Dasgupta 99] Sanjoy Dasgupta. *Learning Mixtures of Gaussians*. In FOCS, pages 634–644, 1999.

[Dasgupta 07] Sanjoy Dasgupta & Leonard J. Schulman. *A Probabilistic Analysis of EM for Mixtures of Separated, Spherical Gaussians*. Journal of Machine Learning Research, vol. 8, pages 203–226, 2007.

[Dempster 77] A. P. Dempster, N. M. Laird & D. B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm*. J. Roy. Statist. Soc. Ser. B, vol. 1, pages 1–38, 1977.

[Devroye 01] Luc Devroye & Gabor Lugosi. Combinatorial methods in density estimation. Springer Series in Statistics. Springer-Verlag, New York, 2001.

[Dubhashi 09] Devdatt Dubhashi & Alessandro Panconesi. Concentration of measure for the analysis of randomized algorithms. Cambridge University Press, 2009.

[Feldman 06] Jon Feldman, Rocco A. Servedio & Ryan O'Donnell. *PAC Learning Axis-Aligned Mixtures of Gaussians with No Separation Assumption*. In COLT, pages 20–34, 2006.

[Kannan 08] Ravindran Kannan, Hadi Salmasian & Santosh Vempala. *The Spectral Method for General Mixture Models*. SIAM J. Comput., vol. 38, no. 3, pages 1141–1156, 2008.

[Melnykov 10] Volodymyr Melnykov & Ranjan Maitra. *Finite mixture models and model-based clustering*. Statist. Surv., vol. 4, 2010.

[Moitra 10] Ankur Moitra & Gregory Valiant. *Settling the Polynomial Learnability of Mixtures of Gaussians*. In IEEE FOCS, 2010.

[Shental 03] Noam Shental, Aharon Bar-Hillel, Tomer Hertz & Daphna Weinshall. *Computing Gaussian Mixture Models with EM Using Equivalence Constraints*. In NIPS, 2003.

[Vempala 02] Santosh Vempala & Grant Wang. *A Spectral Algorithm for Learning Mixtures of Distributions*. In IEEE FOCS, 2002.

[Xing 02] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan & Stuart J. Russell. *Distance Metric Learning with Application to Clustering with Side-Information*. In NIPS, pages 505–512, 2002.