Electronic Journal of Statistics
Vol. 14 (2020) 632–660
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1660

Statistical convergence of the EM algorithm on Gaussian mixture models

Ruofei Zhao and Yuanzhi Li and Yuekai Sun

Department of Statistics
University of Michigan
Ann Arbor, MI
e-mail: rfzhao@umich.edu; yzli@umich.edu; yuekai@umich.edu

Abstract: We study the convergence behavior of the Expectation Maximization (EM) algorithm on Gaussian mixture models with an arbitrary number of mixture components and mixing weights. We show that as long as the means of the components are separated by at least $\Omega(\sqrt{\min\{M, d\}})$, where M is the number of components and d is the dimension, the EM algorithm converges locally to the global optimum of the log-likelihood. Further, we show that the convergence rate is linear and characterize the size of the basin of attraction to the global optimum.

MSC 2010 subject classifications: Primary 62F10; secondary 65K05.
Keywords and phrases: EM algorithm, Gaussian mixture models.

Received October 2018.

1. Introduction

The EM algorithm [12] is an instrumental tool for evaluating the maximum likelihood estimator of latent variable models. It is a Majorization-Minimization (MM) algorithm that minimizes a surrogate objective function to avoid evaluating the intractable marginal (negative) log-likelihood of the latent variable model. However, as is shown by Wu [29] and Tseng [25], the EM algorithm may not converge to a global minimizer of the log-likelihood. Instead, it may converge to a local minimizer or a stationary point. For a method that aims to evaluate the MLE, this is somewhat disappointing.

There is an recent line of research that shows the EM algorithm initialized in a neighborhood of the data generating parameters converges to the global minimizer. Unfortunately, this line of work does not encompass the EM algorithm for fitting mixtures of more than two Gaussians. Assuming the mixture weights are known, this paper fills this gap in the literature by providing conditions under which the EM algorithm for fitting Gaussian mixture models with an arbitrary number of well-separated components and arbitrary mixing weights converges to the global minimizer. We show that the EM algorithm converges linearly as long as it is initialized in a neighborhood of the true centers. Our results are of the same flavor as those by Yan, Yin and Sarkar [31] and Balakrishnan, Wainwright and Yu [3], and our proofs follow the same general route. Importantly, we also consider the ideal case where the mixing weights and the covariance structure are known.

This paper is organized as follows. The rest of this section briefly reviews related work on the EM algorithm. Section 2 describes the EM algorithm for fitting mixtures of Gaussians and introduces a population version of the algorithm that appears in our study. Section 3 states our main results on the convergence of EM. In Section 4, we present simulation results that validate some of our theoretical results. In Section 5, we prove the main results. Finally, in Section 6, we discuss our results and compare them to similar results in the literature.

1.1. Related work

Most closely related to our work is the line of work on the convergence of the EM algorithm for fitting Gaussian mixture models (GMM). On a mixture of two equally-weighted Gaussians, Balakrishnan, Wainwright and Yu [3] first derived statistical convergence results by specializing the general framework they proposed to this model. Their framework also applies to a variant of the EM algorithm known as gradient EM, and they used their framework to obtain similar results for gradient EM. Klusowski and Brinda [20] later obtained results of a similar flavor, but showed that there is a larger neighborhood of the true centers within which the EM algorithm converges. Finally, Xu, Hsu and Maleki [30] and Daskalakis, Tzamos and Zampetakis [11] completely characterized the global convergence behavior of the EM algoritm for fitting two equally weighted Gaussians. When there are more than two components, a result by Jin et al. [17] showed that bad local minima exists even in the idealized case of equally weighted mixtures of well-separated spherical Gaussians. Yan, Yin and Sarkar [31] proved local convergence results for the gradient EM algorithm for fitting mixtures of an arbitrary number of well-separated spherical Gaussians. Despite the recent progress, we are not aware of any results that characterize the local convergence behavior of the EM algorithm on mixtures of two or more Gaussians. In high dimension regime, Dasgupta and Schulman [10] outlines a variant of EM that outputs an estimate that is sufficiently close to the center after merely two iterations. But their result does not fully characterize the convergence behaviour of general EM iterations. For other variants of the EM algorithm for fitting high-dimensional mixture models, we refer readers to Cai, Ma and Zhang [6], Wang et al. [28], Yi and Caramanis [32].

There is also a large body of work on other methods for learning mixtures of Gaussians and, more generally, finite mixtures. One major line of work [9, 26, 1, 2, 19, 5] is based on dimension reduction techniques (such as spectral embeddings). Like the EM algorithm, these methods require the centers of the mixture components to be well-separated. Another more recent line of work [4, 18, 23, 16, 14] employs the method-of-moments, and allows the centers of the mixture components to be arbitrarily close (as long as the sample size is large enough). Some other important algorithms and theoretical work are Brubaker and Vempala [5], Chaudhuri and Rao [7], Chaudhuri et al. [8], Lu and Zhou [21]. For statistical properties such as the rate of convergence of the MLE or the rate

of convergence of the estimated mixing distribution, we refer readers to Ghosal and van der Vaart [13], Nguyen [24], Heinrich and Kahn [15] and the references therein.

2. The EM algorithm on Gaussian mixture models

We consider Gaussian mixture models with known mixture weights and known common covariance structure. Formally, suppose there are M isotropic Gaussian distributions, $\mathcal{N}(\boldsymbol{\mu}_1^*, I_d), \ldots, \mathcal{N}(\boldsymbol{\mu}_M^*, I_d)$, and mixture weights, $\pi_1, \ldots, \pi_M \geq 0$. The Gaussian mixture model we consider is the set of densities

$$\{p(x; \boldsymbol{\mu}^*) = \textstyle\sum_{i=1}^M \pi_i \phi(x - \boldsymbol{\mu}_i^*) : \boldsymbol{\mu}^* = \left[(\boldsymbol{\mu}_1^*)^T, \ldots, (\boldsymbol{\mu}_M^*)^T\right]^T \in \mathbb{R}^{Md}\}, \tag{1}$$

where $\phi(z) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|z\|_2^2}$ is the standard Gaussian density in $\mathbb{R}^d$. The assumption that the components are isotropic leads to no loss of generality as long as (i) the mixture components share a common covariance structure and (ii) this structure is known. Without loss of generality, we also assume mixture is centered; *i.e.* it has mean zero. The task of fitting (1) boils down to estimating the cluster centers $\boldsymbol{\mu}^*$ from observations

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} p(\cdot; \boldsymbol{\mu}^*).$$

The *EM algorithm* for fitting a Gaussian mixture model alternates between evaluating the posterior probabilities (E-step) of the labels and updating the estimates of the parameters (M-step). In the simple situation that we consider, both the mixture weights and the covariance matrix are known. The EM algorithm then only updates the vector $\boldsymbol{\mu}$ iteratively. Combining the E-step and M-step gives the update rule

$$\boldsymbol{\mu}_i^+ \leftarrow \frac{\sum_{j=1}^n w_i(X_j; \boldsymbol{\mu}) X_j}{\sum_{j=1}^n w_i(X_j; \boldsymbol{\mu})}, \quad i = 1, \ldots, M, \tag{2}$$

where the weights $w_i(X_j; \boldsymbol{\mu})$ are defined as

$$w_i(x; \boldsymbol{\mu}) := \frac{\pi_i \phi(x; \boldsymbol{\mu}_i)}{\sum_{k=1}^M \pi_k \phi(x; \boldsymbol{\mu}_k)}. \tag{3}$$

We see that $w_i(X; \boldsymbol{\mu})$ is the probability that X comes from component i and $\boldsymbol{\mu}_i^+$ is a weighted average of samples, where the weights are the $w_i(X_j; \boldsymbol{\mu})$'s. For this reason, the EM algorithm is known as a soft version of the K-means algorithm.

In our analysis, we work with a population version of the EM algorithm. Its update rule is:

$$\boldsymbol{\mu}_i^+ \leftarrow \frac{\mathbb{E}[w_i(X; \boldsymbol{\mu}) X]}{\mathbb{E} w_i(X; \boldsymbol{\mu})} = \frac{\int_{\mathbb{R}^d} w_i(x; \boldsymbol{\mu}) x p(x; \boldsymbol{\mu}^*) dx}{\int_{\mathbb{R}^d} w_i(x; \boldsymbol{\mu}) p(x; \boldsymbol{\mu}^*) dx}, \quad i = 1, \ldots, M. \tag{4}$$

We emphasize that the expectations in (4) are with respect to the true data generating process. We will first derive convergence results for this population version of the EM algorithm and then extend this result to (the sample-version of) the EM algorithm using concentration results.

Notation: We define some the notations in this paper. Let $R_{\max}$, $R_{\min}$ be the largest and smallest distances between the centers of any pair of mixture components:

$$\begin{aligned} R_{\max} &= \max_{i \neq j} \|\boldsymbol{\mu}_i^* - \boldsymbol{\mu}_j^*\|_2, \\ R_{\min} &= \min_{i \neq j} \|\boldsymbol{\mu}_i^* - \boldsymbol{\mu}_j^*\|_2. \end{aligned} \tag{5}$$

Define κ as the smallest mixture weight: $\kappa = \min\{\pi_1, \ldots, \pi_M\}$. Given two positive sequences $\{a_n\}, \{b_n\}$, $a_n = \mathcal{O}(b_n)$ means there exists an absolute constant C such that $a_n \leq C b_n$ for all n; $a_n = \Omega(b_n)$ there exists an absolute constant C such that $a_n \geq C b_n$ for all n. We write $a_n = \Theta(b_n)$ if $a_n = \mathcal{O}(b_n)$ and $a_n = \Omega(b_n)$; we write $a_n = o(b_n)$ if $\frac{a_n}{b_n} \to 0$ as $n \to \infty$. Table 1 below summarizes some useful notations that appears in the main text of this paper.

TABLE 1
Summary of all notations.

Notation	Explanation
d	Dimension of the observations.
n	Sample size.
M	Number of mixture components.
$[M]$	The index set $\{1, 2, \ldots, M\}$.
$(a)_+$	$\max\{0, a\}$.
$X_1, \ldots, X_n$	Observations in $\mathbb{R}^d$.
$\boldsymbol{\mu}^*$	The collection of all M true centers.
$\boldsymbol{\mu}^+$	Population level EM update in (4).
$\boldsymbol{\mu}^0$	The initial iterate in the sample level EM update.
$\{\boldsymbol{\mu}^t\}_{t=1}^{\infty}$	A sequence of sample level EM updates in (2).
$\hat{\boldsymbol{\mu}}$	MLE based on the observations $X_1, \ldots, X_n$.
$\pi_1, \ldots, \pi_M$	Known mixture weights.
κ	The smallest mixing weights as $\min\{\pi_1, \ldots, \pi_M\}$.
$w_1(X, \boldsymbol{\mu}), \ldots, w_M(X, \boldsymbol{\mu})$	Stochastic weights in the EM update rule, defined in (3).
ζ	The contraction coefficient in Theorem 3.1
$R_{\max}$, $R_{\min}$	The largest/smallest distance between centers as in (5).
$\|\cdot\|$	The ℓ-2 norm of a vector.
$\|\cdot\|_{op}$	The operator norm of a square matrix.
$\mathcal{B}(x, r)$	The euclidean ball centers at x with radius r.
$\otimes$	The cartesian product of sets.
C_0, C_1, C_2, C_3	Universal constants.

3. Statement of the main results

In this section, we state our main results for the convergence of the EM algorithm on Gaussian mixture models. First, we show that the population version of the EM algorithm converges linearly to $\boldsymbol{\mu}^*$ as long as (i) certain signal strength

conditions are met and (ii) the algorithm is initialized in a neighborhood of $\boldsymbol{\mu}^*$. We also characterize the size of this neighborhood in terms of the properties of the data generating process.

Theorem 3.1. *Suppose* $R_{\min} \geq 30 \min\{2M, d\}^{\frac{1}{2}}$ *and radius* a *satisfies*

$$a \leq \frac{1}{2} R_{\min} - \min\{d, 2M\}^{\frac{1}{2}} \max\{4\sqrt{2}[\log(\tfrac{R_{\min}}{4})]_+^{\frac{1}{2}}, 8\sqrt{3}, 8\log(\tfrac{4}{\kappa})\}. \quad (6)$$

Then for any iterate $\boldsymbol{\mu}$ *satisfying* $\max_{i\in[M]} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^*\|_2 \leq a$, *the next iterate* $\boldsymbol{\mu}^+$ *given by* (4) *satisfies*

$$\max_{i\in[M]} \|\boldsymbol{\mu}_i^+ - \boldsymbol{\mu}_i^*\|_2 \leq \zeta \max_{i\in[M]} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^*\|_2,$$

where

$$\zeta = \frac{3M}{\kappa^2}(2R_{\max} + \min\{2M, d\})^2 \exp(-\tfrac{1}{8}(\tfrac{1}{2}R_{\min} - a)\min\{d, 2M\}^{\frac{1}{2}}). \quad (7)$$

We remark that Theorem 3.1 does not imply cluster-wise convergence; *i.e.* Theorem 3.1 does not imply

$$\|\boldsymbol{\mu}_i^+ - \boldsymbol{\mu}_i^*\|_2 \leq \zeta \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^*\|_2 \text{ for all } i \in [M].$$

This is hardly surprising. If we initialize the EM algorithm at

$$\left[(\boldsymbol{\mu}_1^*)^T, \boldsymbol{\mu}_2^T, \ldots, \boldsymbol{\mu}_M^T\right]^T,$$

where $\boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_M$ are arbitrary points in $\mathbb{R}^d$, there is no guarantee that $\boldsymbol{\mu}_1^+$ remains at $\boldsymbol{\mu}_1^*$.

Recall the mixture components are isotropic, so $R_{\min}$ is the signal strength. Theorem 3.1 requires the signal strength to grow as $\Omega(\min\{M, d\}^{\frac{1}{2}})$. Theorem 3.1 also shows that the contraction coefficient ζ is decreasing in $R_{\min}$ and κ. It also shows that the contraction radius a approaches $\frac{1}{2}R_{\min}$ as $R_{\min}$ goes to infinity. This contraction radius is essentially optimal because there are examples of the EM algorithm converging to non-global local minima if $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^*\| = \frac{1}{2}R_{\min}$. For example, consider the task of fitting a mixture of two Gaussians. If we initialize the EM algorithm at

$$\left[\tfrac{1}{2}(\boldsymbol{\mu}_1^* + \boldsymbol{\mu}_2^*)^T, \tfrac{1}{2}(\boldsymbol{\mu}_1^* + \boldsymbol{\mu}_2^*)^T\right]^T,$$

the algorithm will never separate the centers and converges to a stationary point with identical estimates of the two centers. It is not clear whether the $\min\{d, 2M\}^{\frac{1}{2}}$ term in (6) is optimal. We suspect not, because in the experiments we run, we have never seen EM fail to converge to the truth when the initializer satisfies $\max \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^*\|_2 < R_{\min}/2$. The improved dependence of the contraction radius on $\min\{d, 2M\}$ instead of d is intuitive. Informally, if $d > 2M$, the components of the noise orthogonal to the subspace spanned by the centers and estimated centers cancel out when expectation is taken, which reduces the

effective dimension of the problem from d to $\min\{d, 2M\}$. The details are in the proof of Theorem 3.1. We remark that this improvement is not present in studies of the convergence of the EM algorithm that does not go through a population level analysis (*cf.* the result of Dasgupta and Schulman [10]).

Here we lay out a technical comparision of our result with Dasgupta and Schulman [10], where they characterize the behavior of the first two rounds of EM. When the dimension d is high, the first iteration brings the randomly initiated centers into our convergent radius (6). The result regarding the next iteration is in a similar flavor of ours. When $d \gg M$, they show the estimation error is already exponentially small in d after the second iteration [10, Theorem 17], which requires the minimal separation to grow with the dimension, i.e. $R_{\min} \approx d^{\frac{1}{2}} \to \infty$. Under the same condition, our characterization of ζ in (7) only reveals an error exponentially small in $d^{\frac{1}{2}}$. However, our focus is to show that EM can work well even when $R_{\min} \approx M^{\frac{1}{2}}$. Replacing $\min\{d, 2M\}^{\frac{1}{2}}$ by d in the proof of Theorem 3.1 characterizes a convergence behaviour comparable to Dasgupta and Schulman [10].

In the statement of Theorem 3.1, there is no provision that $\zeta < 1$. Rearranging $\zeta < 1$ leads to additional constraints on the convergent radius a, which leads to a tighter bound. Removing the redundant constraints gives Corollary 3.2. It is easier to parse but obscures the dependency of the contraction coefficient on the data generating process. Specifically, the additional term $R_{\max}$ dominates the logarithm term of $R_{\min}$ in (6). We relegate the details to Section 5.

Corollary 3.2. *As long as* $R_{\min} \geq C_0 \min\{d, M\}^{\frac{1}{2}}$ *and*

$$\max_{i\in[M]} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^*\|_2 \leq \frac{1}{2} R_{\min} - C_1 \min\{d, M\}^{\frac{1}{2}} \log(\max\{\tfrac{M}{\kappa^2}, R_{\max}, \min\{d, M\}\})^{\frac{1}{2}},$$

where $C_0, C_1 > 0$ *are universal constants, we have*

$$\max_{i\in[M]} \|\boldsymbol{\mu}_i^+ - \boldsymbol{\mu}_i^*\|_2 \leq \frac{1}{2} \max_{i\in[M]} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^*\|_2. \tag{8}$$

Second, we carry out a perturbation analysis to extend Corollary 3.2 to (the sample version of) the EM algorithm. At a high level, our result states that the iterates of the EM algorithm converge linearly to $\boldsymbol{\mu}^*$ up to the statistical precision of the estimation task. Our proof hinges on the idea in Mei, Bai and Montanari [22] for proving concentration results. They the uniform convergence for general non-convex loss function and its gradient. We take the same path to show the sample update rule (4) converges to its population counterpart (2) uniformly. Compared to the proof of an analogous result for the gradient EM algorithm by Yan, Yin and Sarkar [31], our proof is considerably simpler.

Theorem 3.3. *Suppose* $R_{\min} \geq C_0 \min\{d, M\}^{\frac{1}{2}}$ *and the initial iterate* $\boldsymbol{\mu}^0$ *satisfies*

$$\max_{i\in[M]} \|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i^*\|_2 \leq \frac{1}{2} R_{\min} - C_1 \min\{d, M\}^{\frac{1}{2}} \log(\max\{\tfrac{M}{\kappa^2}, R_{\max}, \min\{d, M\}\})^{\frac{1}{2}}, \tag{9}$$

where $C_0, C_1 > 0$ are universal constants. As long as the sample size n is large enough so that

$$\frac{\log n}{n} \leq \min\left\{\frac{\kappa^2}{144\tilde{C}_2 Md}, \frac{\kappa^2 \max_{i\in[M]} \|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i^*\|_2^2}{9\tilde{C}_3 R_{\max}^2 Md}\right\}, \tag{10}$$

where $C_2, C_3 > 0$ are universal constants and $\tilde{C}_2 = C_2 \log(M(2R_{\max} + \sqrt{d}))$, $\tilde{C}_3 = C_3 \log(M(3R_{\max}^2 + d))$, the subsequent EM iterates $\{\boldsymbol{\mu}^t\}_{t=1}^{\infty}$ given by (2) satisfy

$$\max_{i\in[M]} \|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\|_2 \leq \frac{1}{2^t} \max_{i\in[M]} \|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i^*\|_2 + \frac{3R_{\max}}{\kappa}\left(\frac{\tilde{C}_3 Md \log n}{n}\right)^{\frac{1}{2}} \tag{11}$$

with probability at least $1 - \frac{2M}{n}$.

We see that the first term on the right side of (11) converges to zero linearly while the second term does not depend on t. Initially, the first term on the right side dominates the second term and the right side decreases linearly. However, after sufficiently many iterations, the second term on the right side dominates the first term, and the right side settles down to a limit that is $\mathcal{O}((\frac{Md}{\kappa^2 n})^{\frac{1}{2}})$ (modulo constant and logarithmic factors). We recognize this limit as the statistical precision.

Convergence to the maximum likelihood estimator. We remark that Theorem 3.3 implies EM converges to a stationary point within a ball of radius $\mathcal{O}((\frac{Md}{\kappa^2 n})^{\frac{1}{2}})$ of the true centers. It is known that the maximum likelihood estimator (MLE) falls inside this ball with high probability. As long as there are no other stationary points in this ball, Theorem 3.3 implies EM converges to the MLE. This is a consequence of the gradient stability result by Yan, Yin and Sarkar [31], which implies $\boldsymbol{\mu}^*$ is the only stationary point of the expected log-likelihood function in $\otimes_i(\boldsymbol{\mu}_i^*, a)$, and concentration results by Mei, Bai and Montanari [22], which imply the log-likelihood function only has one stationary point in $\otimes_i(\boldsymbol{\mu}_i^*, a)$.

4. Simulation results

This section presents numerical results to demonstrate the local-convergence behaviour of the EM. First we operate under our analytical framework, where the mixture weights and variances are both known. Then we investigate the case when the mixture weights and variances are unknown and updated via EM. Though lacking theoretical understanding, empirical evidences shows similar convergence result.

4.1. Known mixture weights and variances

In this section, we assume the mixture weights π_k's are known, and the each mixing component is isotropic with unit variance. In the first set of experiments,

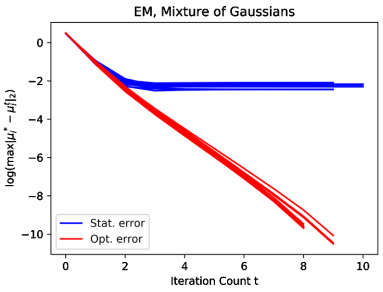

Figure 1.a

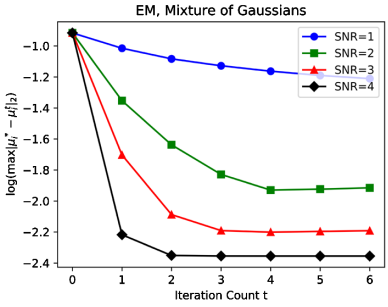

Figure 1.b

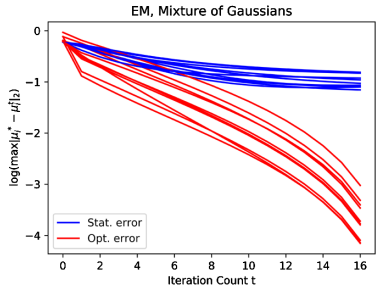

Figure 1.c

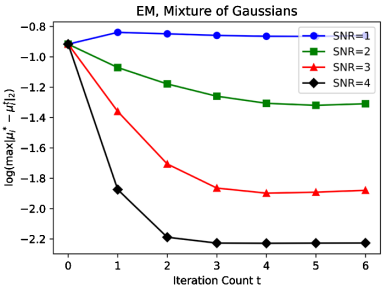

Figure 1.d

FIG 1. *Convergence of EM on the mixture of $M = 5$ Gaussians in $\mathbb{R}^{10}$. Left column shows the statistical error and optimization error for ten trials. Right column shows average statistical error over ten trials under different SNR (SNR is R_{min}). While top row has balanced cluster weights, bottom row does not.*

we empirically verify the prediction of Theorem 3.1 that the statistical error decreases linearly initially and eventually reaches a plateau. The data generating process is a mixture of $M = 5$ isotropic Gaussians in $\mathbb{R}^{10}$ with one mixture component centered at the origin and the remaining four centers at the vertices of $R_{\min}\Delta^9$, where Δ^9 is the probability simplex in $\mathbb{R}^{10}$, so $\frac{R_{\max}}{R_{\min}} = \sqrt{2}$. The components are equally weighted ($\pi_i = \frac{1}{5}$).

The top left panel of Figure 1 shows the decrease of the optimization error $\max_i \|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i^t\|_2$ and statistical error $\max_i \|\boldsymbol{\mu}_i^* - \boldsymbol{\mu}_i^t\|_2$ over 10 runs of the EM algorithm. Here $\hat{\boldsymbol{\mu}}$ is the MLE at which EM converges. We set $R_{\min} = 2$ and generate $n = 8000$ samples from the Gaussian mixture model. We initialize the EM algorithm at

$$\left[(\boldsymbol{\mu}_1^* + \delta_1)^T, \ldots, (\boldsymbol{\mu}_M^* + \delta_M)^T\right]^T,$$

where δ_i is uniformly distributed on the sphere of radius $0.4 \cdot R_{\min}$. We see that the statistical error decreases linearly initially and reaches a plateau after merely 4–5 iterations. This agrees with the implications of Theorem 3.3. The top right panel of Figure 1 shows the average log statistical errors over 10 trails as $R_{\min}$ varies. We see that larger $R_{\min}$ values lead to faster convergence, which agrees with the implications of Theorem 3.1.

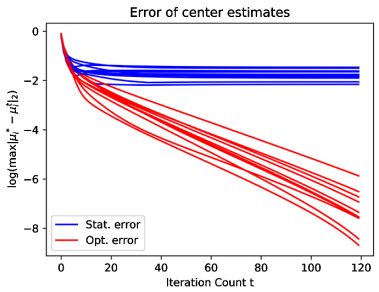

Figure 2.a

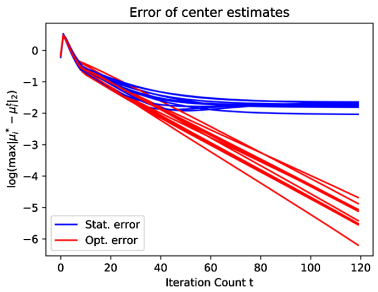

Figure 2.b

FIG 2. *Convergence of EM with unknown mixture weights among 10 independent trails. Left panel initializes the mixture weights near the truth. Right panel uses extreme initializations that are far from the truth.*

The plots in the bottom panels are analogous to the plots above. We keep all simulation parameters, except we change the mixture weights from uniform to $\pi_i = \frac{i}{15}$, $i \in [M]$. We see that non-uniform mixture weights hurt the performance of the EM algorithm. Comparing the two plots on the left, we see that non-uniform weights causes the algorithm to converge slower and reduces the statistical precision of the output. The slower convergence agrees with the implications of Theorem 3.1, which shows that the contraction coefficient is inversely proportional to the minimum mixture weight κ. The reduced statistical precision is due to the fact that the centers of mixture components with smaller weights are estimated less accurately. This is because the mixture components with smaller weights have smaller effective sample sizes. We also see greater variation across the ten runs of the algorithm.

4.2. Unknown mixture weights

In this section, we consider the scenario when the mixture weights are unknown and updated via EM, while each component is still isotropic with known variance. We focus on the case with equal mixing weights. The underlying data generating process is identical to the one in Section 4.1 with equal mixing weights. The mean vectors are initialized as in Section 4.1. For initializing the weights $\pi_i^{(0)}$, we consider the following two methods:

(a) $\{\pi_i^{(0)} : k = 1, \ldots, 5\} \sim \text{Dir}(5, 5, 5, 5, 5)$,
(b) $\{\pi_i^{(0)} : k = 1, \ldots, 5\} \sim \text{Dir}(100, 1, 1, 1, 1)$,

where Dir() is the Dirichlet distribution. Method (a) initializes the weights that are close to the truth: $\pi_i^{(0)}$ will be centered at $\pi_i = 1/5$ with a small variace. Method (b) is an extreme initialization, in the sense that it places most weight only to the first mixing component, which is wildly different from the truth.

Figure 2 shows the errors among 10 runs of EM algorithms, under initializations (a) and (b) respectively. We only present the error for estimating the mean

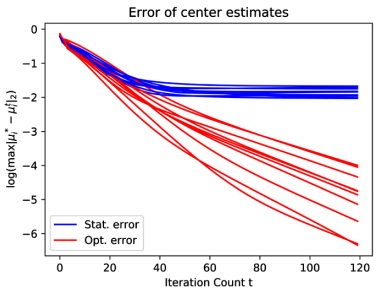

Figure 3.a

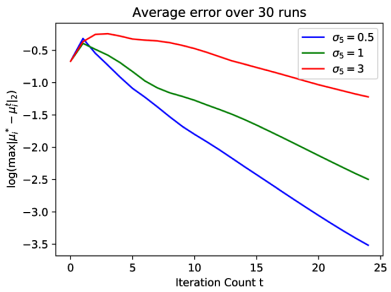

Figure 3.b

FIG 3. *Convergence of EM with unknown mixture weights and variances. Left panel shows 10 independent runs of EM. Right panel shows average statistical error over 30 trails under different variances for the fifth component.*

vectors. Two comments are in place. First, comparing with Figure 1.a, we find the local-convergence phenomena persists with a similar statistical precision. Though the unknown mixture weights does slows down the convergence. Second, while the initialization Method (b) does jeopardize the rate of convergence, even this 'worst-case' scenario does not fully destroy the convergence, as shown on the right panel of Figure 2. The result seems to suggest the initialization of weights can be quite arbitrary, provided that the centers are sufficiently close to the truth.

4.3. Unknown mixture weights and variances

Now we consider the situation where both mixture weights and variances are unknown and updated via EM. The ith mixing component is an isotropic Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_i^*, \sigma_i^2 I_d)$, where σ_i could be different acorss i. The first set of simulation shares the same data generating process with Section 4.1 with equal mixing weights and variances. The left panel of Figure 3 shows the error of center estimates among 10 runs of the EM. We initialize the EM as in Section 4.2, using Method (a) for the weights. The variances are initiated from

$$\sigma_i^2(0) = \sigma_i^2 \cdot Z_i,$$

for each $i = 1, \ldots, 5$. Here each Z_i follows a chi-square distribution with $df = 2$. We observe a similar local-convergence behaviour in line with Theorem 3.3: The statistical error decreases linearly at first, eventually stabilizes at the statistical precision. Comparing with Figure 1.a and 2.a, we can see that the unknown variances further slow down the convergence of EM. Through more extensive simulation studies, we find that the convergence of EM seem to hold if the initial variances are reasonably close to the true variances.

In the next set of simulation, we vary σ_5, the standard deviation of the fifth mixing component, while keeping the other variances at 1. The right panel of Figure 3 shows the average log-statistical errors over 30 independent trails for

each value of σ_5. We see a larger σ_5 leads to slower convergence rate. When $\sigma_5 \geq 1$, the phenomena agrees with Dasgupta and Schulman [10]. They show the key factor dominating the convergence rate is the ratio of $R_{\min}$ with $\sigma_{\max}$, the maximum standard deviation among all mixing components. On the other hand, the convergence becomes faster when σ_5 goes below 1, although $\sigma_{\max}$ remains to be 1. Intuitively, a smaller σ_5 allows for a more accurate estimate for the fifth component. This component-specific benefit turns out to improve the convergence rate for all other components in the end.

5. Proofs of the main results

We prove Theorem 3.1, 3.2, and 3.3 in this section, deferring the proofs of the technical lemmas to the appendices.

5.1. *Proof of Theorem 3.1*

We make a few observations before proceeding to the proof. Without loss of generality, we focus on the update rule for the first center $\boldsymbol{\mu}_1$:

$$\boldsymbol{\mu}_1^+ - \boldsymbol{\mu}_1^* = \frac{\mathbb{E}\big[w_1(X;\boldsymbol{\mu})(X-\boldsymbol{\mu}_1^*)\big]}{\mathbb{E}\big[w_1(X;\boldsymbol{\mu})\big]}$$

The vector of true centers $\boldsymbol{\mu}^*$ is a fixed point of (4), which implies

$$\mathbb{E}\big[w_1(X;\boldsymbol{\mu}^*)(X-\boldsymbol{\mu}_1^*)\big] = \mathbf{0}.$$

Thus

$$\boldsymbol{\mu}_1^+ - \boldsymbol{\mu}_1^* = \frac{\mathbb{E}\big[\big(w_1(X;\boldsymbol{\mu}) - w_1(X;\boldsymbol{\mu}^*)\big)(X-\boldsymbol{\mu}_1^*)\big]}{\mathbb{E}\big[w_1(X;\boldsymbol{\mu})\big]}. \tag{12}$$

In the first step of the proof, we establish an upper bound on the norm of the numerator. In the second step, we establish a lower bound on the denominator in (12). Finally, in the third step, we combine the upper and lower bounds to show the EM update rule is a contraction.

Step 1 (upper bounding the numerator). Define $\boldsymbol{\mu}^t := \boldsymbol{\mu}^* + t(\boldsymbol{\mu} - \boldsymbol{\mu}^*)$ and define $g_X(t) := w_1(X;\boldsymbol{\mu}^t)$. We have

$$w_1(X;\boldsymbol{\mu}) - w_1(X;\boldsymbol{\mu}^*) = \int_0^1 g_X'(t)dt = \int_0^1 \nabla_{\boldsymbol{\mu}} w_1\big(x;\boldsymbol{\mu}^t\big)^T(\boldsymbol{\mu}^t - \boldsymbol{\mu}^*)dt,$$

where

$$\nabla_{\boldsymbol{\mu}} w_1(X;\boldsymbol{\mu}) = \begin{bmatrix} -w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))(\boldsymbol{\mu}_1 - X) \\ w_1(X;\boldsymbol{\mu})w_2(X;\boldsymbol{\mu})(\boldsymbol{\mu}_2 - X) \\ \vdots \\ w_1(X;\boldsymbol{\mu})w_M(X;\boldsymbol{\mu})(\boldsymbol{\mu}_M - X) \end{bmatrix}. \tag{13}$$

Let $|\cdot|_{op}$ be the operator norm of a matrix. We thus have

$$\begin{aligned}
&\big\|\mathbb{E}\big[(w_1(X;\boldsymbol{\mu}) - w_1(X;\boldsymbol{\mu}^*))(X-\boldsymbol{\mu}_1^*)\big]\big\|_2 \\
&\quad= \Big\| \int_0^1 \mathbb{E}\big[w_1(X;\boldsymbol{\mu}^t)(1-w_1(X;\boldsymbol{\mu}^t))(X-\boldsymbol{\mu}_1^t)^T(\boldsymbol{\mu}_1^t-\boldsymbol{\mu}_1^*)(X-\boldsymbol{\mu}_1^*)\big]dt \\
&\qquad - \sum_{i\neq 1}\int_0^1 \mathbb{E}\big[w_1(X;\boldsymbol{\mu}^t)w_i(X;\boldsymbol{\mu}^t)(X-\boldsymbol{\mu}_i^t)^T(\boldsymbol{\mu}_i^t-\boldsymbol{\mu}_i^*)(X-\boldsymbol{\mu}_1^*)\big]dt\Big\|_2 \\
&\quad\le \int_0^1 \big\|\mathbb{E}\big[w_1(X;\boldsymbol{\mu}^t)(1-w_1(X;\boldsymbol{\mu}^t))(X-\boldsymbol{\mu}_1^*)(X-\boldsymbol{\mu}_1^t)^T\big]\big\|_{op}\|\boldsymbol{\mu}_1^t-\boldsymbol{\mu}_1^*\|_2 dt \\
&\qquad + \sum_{i\neq 1}\int_0^1 \big\|\mathbb{E}\big[w_1(X;\boldsymbol{\mu}^t)w_i(X;\boldsymbol{\mu}^t)(X-\boldsymbol{\mu}_1^*)(X-\boldsymbol{\mu}_i^t)^T\big]\big\|_{op}\|\boldsymbol{\mu}_i^t-\boldsymbol{\mu}_i^*\|_2 dt \\
&\quad\le V_1\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_1^*\|_2 + \sum_{i\neq 1} V_i\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_i^*\|_2 \\
&\quad\le M(\max_i V_i)\cdot\max_i\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_i^*\|_2, \qquad (14)
\end{aligned}$$

where

$$V_1 = \sup_{t\in[0,1]}\big\|\mathbb{E}\big[w_1(X;\boldsymbol{\mu}^t)(1-w_1(X;\boldsymbol{\mu}^t))(X-\boldsymbol{\mu}_1^*)(X-\boldsymbol{\mu}_1^t)^T\big]\big\|_{op}, \qquad (15)$$

$$V_i = \sup_{t\in[0,1]}\big\|\mathbb{E}\big[w_1(X;\boldsymbol{\mu}^t)w_i(X;\boldsymbol{\mu}^t)(X-\boldsymbol{\mu}_1^*)(X-\boldsymbol{\mu}_i^t)^T\big]\big\|_{op}. \qquad (16)$$

The rest of the first step consists of establishing bounds on the V_i's. We state the result here and defer the details to Appendix A.

Lemma 5.1. *As long as* $R_{\min} \ge 30\min\{d, 2M\}^{\frac{1}{2}}$ *and*

$$a \le \frac{1}{2}R_{\min} - \min\{d,2M\}^{\frac{1}{2}}\cdot\max\{4\sqrt{2}[\log(R_{\min}/4)]_+^{\frac{1}{2}}, 8\sqrt{3}\}, \qquad (17)$$

then for any $\boldsymbol{\mu}$ *such that* $\boldsymbol{\mu}_i \in \mathcal{B}(\boldsymbol{\mu}_i^*, a), i\in[M]$, *we have*

$$\max_{i\in[M]} V_i \le \frac{2}{\kappa}(2R_{\max}+\min\{2M,d\})^2\exp(-\tfrac{1}{8}(\tfrac{1}{2}R_{\min}-a)\min\{d,2M\}^{\frac{1}{2}}). \qquad (18)$$

Step 2 (lower bounding the denominator). We state the lower bound and describe the underlying intuition, deferring the proof to the appendix. Let Z be the label of X. We observe that

$$\mathbb{E}\big[w_1(X;\boldsymbol{\mu}^*)\big] = \mathbb{E}\big[\mathbb{P}_{\boldsymbol{\mu}^*}(Z=1|X)\big] = \pi_1 > \kappa.$$

As long as $\boldsymbol{\mu}\approx\boldsymbol{\mu}^*$, $\mathbb{E}\big[w_1(X;\boldsymbol{\mu})\big]\approx\mathbb{E}\big[w_1(X;\boldsymbol{\mu}^*)\big]$, so $\mathbb{E}\big[w_1(X;\boldsymbol{\mu})\big]$ cannot be much smaller than κ.

Lemma 5.2. *As long as* $R_{\min}\ge 30\min\{M,d\}^{\frac{1}{2}}$ *and*

$$a \le \frac{1}{2}R_{\min} - \min\{M,d\}^{\frac{1}{2}}\cdot\max\{4\sqrt{2}[\log(\tfrac{R_{\min}}{4})]_+^{\frac{1}{2}}, 8\sqrt{3}, 8\log(\tfrac{4}{\kappa})\}, \qquad (19)$$

then for any $\boldsymbol{\mu}$ *such that* $\boldsymbol{\mu}_i \in \mathcal{B}(\boldsymbol{\mu}_i^*, a), i \in [M]$, *we have*

$$\mathbb{E}\big[w_i(X;\boldsymbol{\mu})\big] \geq \frac{3}{4}\kappa, \quad i \in [M]. \tag{20}$$

Step 3: We combine the bounds for $\max_i V_i$ and $\mathbb{E}\big[w_1(X;\boldsymbol{\mu})\big]$ to show that the EM update rule is a contraction. Without loss of generality, we focus on the update rule for the first cluster. By (12), (14), and Lemma 5.2, we have

$$\begin{aligned} \|\boldsymbol{\mu}_1^+ - \boldsymbol{\mu}_1^*\|_2 &= \frac{\|\mathbb{E}[\big(w_1(X;\boldsymbol{\mu}) - w_1(X;\boldsymbol{\mu}^*)\big)(X - \boldsymbol{\mu}_1^*)]\|_2}{\mathbb{E}\big[w_1(X;\boldsymbol{\mu})\big]} \\ &\leq \frac{4M}{3\kappa}(\max_i V_i)(\max_i \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^*\|_2). \end{aligned} \tag{21}$$

Plugging in (18), we have

$$\begin{aligned} \|\boldsymbol{\mu}_1^+ - \boldsymbol{\mu}_1^*\|_2 \leq & \frac{8M}{3\kappa^2}(2R_{\max} + \min\{2M, d\})^2 \\ & \cdot \exp(-\tfrac{1}{8}(\tfrac{1}{2}R_{\min} - a)\min\{d, 2M\}^{\frac{1}{2}})(\max_i \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^*\|_2). \end{aligned} \tag{22}$$

We recognize the factor in front of $\max_i \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^*\|_2$ is smaller than the contraction factor ζ in Theorem 3.1. We can also check that (6) implies (17) and (19) so the conditions of Lemmas 5.1 and 5.2 are satisfied. Putting the two parts together, we see that Theorem 3.1 is correct.

5.2. *Proof of Corollary 3.2*

To prove Corollary 3.2, we start from (21) and solve for the contraction radius a that implies $\zeta < \frac{1}{2}$. It is enough to find α such that

$$\frac{M \max_i V_i}{\frac{2}{3}\kappa} \leq \frac{1}{2}.$$

By Lemma 5.1, the above condition is implied by

$$\frac{2}{\kappa}(2R_{\max} + \min\{2M, d\})^2 \exp(-\tfrac{1}{8}(\tfrac{1}{2}R_{\min} - a)\min\{d, 2M\}^{\frac{1}{2}}) \leq \frac{\kappa}{3M}.$$

Rearranging, we have

$$a \leq \frac{1}{2}R_{\min} - \frac{8}{\min\{d, 2M\}^{\frac{1}{2}}}\log(\tfrac{6M(2R_{\max} + \min\{2M,d\})^2}{\kappa^2}). \tag{23}$$

Finally, we check that there is a universal constant C_1 such that

$$a \leq \frac{1}{2}R_{\min} - C_1 \min\{d, 2M\}^{\frac{1}{2}} \log(\max\{\tfrac{M}{\kappa^2}, R_{\max}, \min\{2M, d\}\})^{\frac{1}{2}},$$

implies (17), (19), and (23).

5.3. Proof of Theorem 3.3

The intuition underlying the proof of Theorem 3.3 is the population and sample update rules ((2) and (4) respectively) are similar. In the proof, we appeal to the following technical lemmas on the uniform convergence of $\frac{1}{n}\sum_{j=1}^{n} w_i(X_j;\boldsymbol{\mu}) \times (X_j - \boldsymbol{\mu}_1^*)$ and $\frac{1}{n}\sum_{j=1}^{n} w_i(X_j;\boldsymbol{\mu})$ to their population counterparts.

Lemma 5.3. *Define the event* $\mathcal{E}_{1,i}$ *as*

$$\begin{aligned}&\sup_{\boldsymbol{\mu}\in\mathcal{U}} \|\tfrac{1}{n}\textstyle\sum_{j=1}^{n} w_i(X_j;\boldsymbol{\mu})(X_j - \boldsymbol{\mu}_i^*) - \mathbb{E}\big[w_i(X;\boldsymbol{\mu})(X - \boldsymbol{\mu}_i^*)\big]\|_2 \\ &\quad \geq 1.5R_{\max}(\tfrac{\tilde{C}_3 M d\log n}{n})^{\frac{1}{2}},\end{aligned}$$

where $\mathcal{U} = \otimes_{i=1}^{M}\mathcal{B}(\boldsymbol{\mu}_i^*, 1.5R_{\max})$, C_3 *is a universal constant, and*

$$\tilde{C}_3 = C_3\log(M(3R_{\max}^2 + d)).$$

We have $\mathbb{P}(\mathcal{E}_{1,i}) \leq \frac{1}{n}$.

Lemma 5.4. *Define the event*

$$\mathcal{E}_{2,i} = \sup_{\boldsymbol{\mu}\in\mathcal{U}} |\tfrac{1}{n}\textstyle\sum_{j=1}^{n} w_i(X_j,\boldsymbol{\mu}) - \mathbb{E}\big[w_i(X;\boldsymbol{\mu})\big]| \geq (\tfrac{\tilde{C}_2 M d\log n}{n})^{\frac{1}{2}},$$

where $\boldsymbol{\mu}\in\mathcal{U} = \otimes_{i=1}^{M}\mathcal{B}(\boldsymbol{\mu}_i^*, R_{\max})$, C_2 *is a universal constant, and*

$$\tilde{C}_2 = C_2\log(M(2R_{\max} + \sqrt{d})).$$

We have $\mathbb{P}(\mathcal{E}_{2,i}) \leq \frac{1}{n}$.

Proof of Theorem 3.3. On the event $\{\cap_{i\in[M]}\mathcal{E}_{1,i}^c\} \cap \{\cap_{i\in[M]}\mathcal{E}_{2,i}^c\}$, we have

$$\begin{aligned}&\sup_{\boldsymbol{\mu}\in\mathcal{U}} |\tfrac{1}{n}\textstyle\sum_{j=1}^{n} w_i(X_j,\boldsymbol{\mu}) - \mathbb{E}\big[w_i(X;\boldsymbol{\mu})\big]| \leq (\tfrac{\tilde{C}_2 M d\log n}{n})^{\frac{1}{2}}, \\ &\sup_{\boldsymbol{\mu}\in\mathcal{U}} \|\tfrac{1}{n}\textstyle\sum_{j=1}^{n} w_i(X_j;\boldsymbol{\mu})(X_j - \boldsymbol{\mu}_i^*) - \mathbb{E}\big[w_i(X;\boldsymbol{\mu})(X - \boldsymbol{\mu}_i^*)\big]\|_2 \\ &\quad \leq 1.5R_{\max}(\frac{\tilde{C}_3 M d\log n}{n})^{\frac{1}{2}}\end{aligned}$$

for all $i \in [M]$. By Lemmas 5.3 and 5.4, this event occurs with probability at least $1 - \frac{2M}{n}$. The minimum sample size condition (10) implies

$$R_{\max}(\tfrac{\tilde{C}_3 M d\log n}{n})^{\frac{1}{2}} \leq \frac{\kappa}{3}\max_{i\in[M]}\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i^*\|_2, \tag{24}$$

$$(\tfrac{\tilde{C}_2 M d\log n}{n})^{\frac{1}{2}} \leq \frac{\kappa}{12}. \tag{25}$$

The second inequality (25) in turn implies

$$\sup_{\boldsymbol{\mu}\in\mathcal{U}} |\tfrac{1}{n}\textstyle\sum_{j=1}^{n} w_i(X_j,\boldsymbol{\mu}) - \mathbb{E}\big[w_i(X;\boldsymbol{\mu})\big]| \leq \tfrac{1}{12}\kappa$$

for all $i \in [M]$. Let $\boldsymbol{\mu}^0$ be the initial iterate. We have

$$\begin{aligned}
\|\boldsymbol{\mu}_i^1 - \boldsymbol{\mu}_i^*\|_2 &= \frac{\|\frac{1}{n}\sum_{j=1}^n w_i(X_j;\boldsymbol{\mu}^0)(X_j - \boldsymbol{\mu}_i^*)\|_2}{\frac{1}{n}\sum_{j=1}^n w_i(X_j;\boldsymbol{\mu}^0)} \\
&\le \frac{\|\mathbb{E}w_i(X;\boldsymbol{\mu}^0)(X-\boldsymbol{\mu}_i^*)\|_2 + R_{\max}(\frac{\tilde{C}_3 M d \log n}{n})^{\frac{1}{2}}}{\mathbb{E}[w_i(X;\boldsymbol{\mu}^0)] - \frac{\kappa}{12}} \\
&\overset{(i)}{\le} \frac{\big\|\mathbb{E}w_i(X;\boldsymbol{\mu}^0)(X-\boldsymbol{\mu}_i^*)\big\|_2 + R_{\max}(\frac{\tilde{C}_3 M d \log n}{n})^{\frac{1}{2}}}{\frac{2\kappa}{3}} \\
&\overset{(ii)}{\le} \frac{1}{2}\max_{i\in[M]}\big\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i^*\big\|_2 + \frac{3R_{\max}(\tilde{C}_3 M d\frac{\log n}{n})^{\frac{1}{2}}}{2\kappa},
\end{aligned} \tag{26}$$

where we appealed to

$$\begin{aligned}
\mathbb{E}\big[w_i(X;\boldsymbol{\mu}^0)\big] &\ge \frac{3}{4}\kappa, \\
\|\mathbb{E}\big[w_i(X;\boldsymbol{\mu}^0)(X-\boldsymbol{\mu}_i^*)\big]\|_2 &\le \frac{\kappa}{3}\max_{i\in[M]}\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i^*\|_2
\end{aligned}$$

in steps (i) and (ii). Both are intermediate results established in the proof of Corollary 3.2. Finally, we have

$$\begin{aligned}
\big\|\boldsymbol{\mu}_i^1 - \boldsymbol{\mu}_i^*\big\|_2 &\le \frac{1}{2}\max_{i\in[M]}\big\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i^*\big\|_2 + \frac{3R_{\max}(\tilde{C}_3 M d\frac{\log n}{n})^{\frac{1}{2}}}{2\kappa} \\
&\le \max_{i\in[M]}\big\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i^*\big\|_2,
\end{aligned}$$

where the second step is a consequence of (24). This implies $\boldsymbol{\mu}^1$ is also in the contraction region. By applying (26) iteratively, we have

$$\begin{aligned}
\max_{i\in[M]}\|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^*\|_2 &\le \frac{1}{2^t}\max_{i\in[M]}\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i^*\|_2 \\
&\quad + \Big(1 + \frac{1}{2} + \ldots + \frac{1}{2^{t-1}}\Big)\frac{3R_{\max}(\tilde{C}_3 M d\frac{\log n}{n})^{\frac{1}{2}}}{2\kappa} \qquad (27) \\
&\le \frac{1}{2^t}\max_{i\in[M]}\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i^*\|_2 + \frac{3R_{\max}(\tilde{C}_3 M d\frac{\log n}{n})^{\frac{1}{2}}}{\kappa}. \qquad (28)
\end{aligned}$$

6. Summary and discussion

Initialization. We emphasize that our convergence results are local: they assume the EM algorithm is initialized in a neighborhood of the true centers. To obtain a such an initial iterate, we appeal to approaches based on the method of moments, such as the method proposed by Hsu and Kakade [16]. These methods are consistent, but its sample complexity is worse than that of the EM algorithm. Under certain conditions on the true centers $\boldsymbol{\mu}^*$ (see [16, Theorem 3] the

detailed conditions), the algorithm in Hsu and Kakade [16] gives estimates of the center $\hat{\boldsymbol{\mu}}$ that satisfy

$$\|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i^*\|_2 \leq (\|\boldsymbol{\mu}_i^*\|_2 + \sqrt{\lambda_{max}(M_2)})\epsilon \text{ with probability at least } 1-\delta,$$

where $\lambda_{max}(M_2)$ is the largest eigenvalue of the matrix $M_2 \triangleq \sum_{i=1}^{M} w_i \boldsymbol{\mu}_i^* (\boldsymbol{\mu}_i^*)^T$. By combining a spectral method with the EM algorithm, we have the best of both: the combined estimator is both consistent and (asymptotically) efficient.

Minimum separation between centers. Theorems 3.1 and 3.3 require the minimum separation between centers to grow as $\Omega(\min\{M, d\}^{\frac{1}{2}})$. Compared to other methods for fitting mixtures of isotropic Gaussians, this dependence is sub-optimal. For example, Vempala and Wang [26] showed that spectral clustering can accurately recover the labels in a mixture of spherical Gaussians provided that the minimum separation is at least $\Omega((M \log d)^{\frac{1}{4}})$. Some approaches based on the method of moments are able to learn mixtures of Gaussians in which the centers are arbitrarily close together (as long as the sample size is large enough). However, the sample complexity of such methods are usually worse than that of the EM algorithm.

If we restrict to studies of the EM algorithm and its variants (including gradient EM and the K-means algorithm), our requirement on the minimum separation between centers is optimal. Yan, Yin and Sarkar [31] imposes the same condition in their study of the convergence of the gradient EM algorithm. Lu and Zhou [21] requires the minimum separation to grow proportionally to M in their study of the convergence of the K-means algorithm.

Contraction radius and convergence rate. This contraction radius in (6) is optimal in the sense that it is approximately $\frac{1}{2}R_{min}$ when R_{min} is large and we can find examples of the EM algorithm converging to non-global local minima if $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^*\|_2 = \frac{1}{2}R_{\min}$ (see the remarks after Theorem 3.1). By comparison, the contraction radius for the gradient EM algorithm is very similar to (6) [31], and the contraction radius for K-means is roughly $\frac{1}{2}R_{\min} - CM^{\frac{3}{4}}$ [21, Theorem 6.2].

We show that the EM algorithm converges linearly up to statistical precision. This agrees with our simulation results and previous studies on the convergence of EM [3, 25]. We see that the convergence rate decreases as κ increases, which agrees with the folklore that the EM algorithm converges slowly on imbalanced mixture models.

Minimum sample size. In terms of minimum sample size, our result is valid as long as $\frac{n}{\log n} \geq C\frac{MdR_{\max}^2}{\kappa^2 R_{\min}^2}$. Yan, Yin and Sarkar [31] established linear convergence of the gradient EM algorithm as long as $n \geq C\frac{M^6 R_{\max}^6 d}{R_{\min}^2}$ and Lu and Zhou [21] established linear convergence of the K-mean algorithm as long as $\frac{n}{\log n} \geq C\frac{M}{\kappa^2}$. The variations in the minimum sample size are due to differences in the concentration results that appear in the proofs. We believe it is possible to avoid the $\log n$ factor and improve the dependence on $\frac{R_{\max}}{R_{\min}}$ in the minimum sample size by refining the concentration results in our proofs.

Appendix A

In appendix A, we prove Lemma 5.1 and 5.2.

A.1. Preliminaries

Before jumping into the proof, we need the following preliminary result, which is Lemma 12 and 13 in Yan, Yin and Sarkar [31].

Lemma A.1. *Suppose the minimum separation $R_{\min}$ and radius a satisfy $R_{\min} \geq 30\sqrt{d}$ and*

$$a \leq \frac{1}{2}R_{\min} - \sqrt{d}\max\left(4\sqrt{2[\log(R_{\min}/4)]_+}, 8\sqrt{3}\right),$$

then for any $\boldsymbol{\mu}$ such that $\boldsymbol{\mu}_i \in \mathcal{B}(\boldsymbol{\mu}_i^*, a), \forall i \in [M]$, we have the following inequalities for any $p = 0, 1, 2$ and any $i \neq j \in [M]$

$$\begin{aligned}
&\mathbb{E}w_i(X;\boldsymbol{\mu})(1 - w_i(X;\boldsymbol{\mu}))\|X - \boldsymbol{\mu}_i\|_2^p \leq \\
&\qquad 2M\Big(\frac{3}{2}R_{\max} + d\Big)^p \exp\Big(-\Big(\frac{1}{2}R_{\min} - a\Big)\frac{\sqrt{d}}{8}\Big), \\
&\mathbb{E}w_i(X;\boldsymbol{\mu})w_j(X;\boldsymbol{\mu})\|X - \boldsymbol{\mu}_i\|\|X - \boldsymbol{\mu}_j\| \leq \\
&\qquad \frac{2-\kappa}{\kappa}\Big(\frac{3}{2}R_{\max} + d\Big)^2 \exp\Big(-\Big(\frac{1}{2}R_{\min} - a\Big)\frac{\sqrt{d}}{8}\Big).
\end{aligned}$$

A.2. Proof for Lemma 5.1

We start from bounding V_1:

$$\begin{aligned}
V_1 &\leq \sup_{t\in[0,1]} \big\|\mathbb{E}w_1(X;\boldsymbol{\mu}^t)(1 - w_1(X;\boldsymbol{\mu}^t))(X - \boldsymbol{\mu}_1^t)(X - \boldsymbol{\mu}_1^t)^T\big\|_{op} \\
&\quad + \sup_{t\in[0,1]} \big\|\mathbb{E}w_1(X;\boldsymbol{\mu}^t)(1 - w_1(X;\boldsymbol{\mu}^t))(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_1^t)(X - \boldsymbol{\mu}_1^t)^T\big\|_{op} \\
&\leq \sup_{t\in[0,1]} \big\|\mathbb{E}w_1(X;\boldsymbol{\mu}^t)(1 - w_1(X;\boldsymbol{\mu}^t))(X - \boldsymbol{\mu}_1^t)(X - \boldsymbol{\mu}_1^t)^T\big\|_{op} \\
&\quad + a \sup_{t\in[0,1]} \big\|\mathbb{E}w_1(X;\boldsymbol{\mu}^t)(1 - w_1(X;\boldsymbol{\mu}^t))(X - \boldsymbol{\mu}_1^t)^T\big\|_2, \qquad (29)
\end{aligned}$$

where a is the radius of the contraction region $\otimes_i \mathcal{B}(\boldsymbol{\mu}_i^*, a)$. For any $t \in [0,1]$, there exists a rotation matrix Γ, such that all $\Gamma\boldsymbol{\mu}_i^t, \Gamma\boldsymbol{\mu}_i^*$, $i \in [M]$ have zero entries in the last $[d - 2M]_+$ coordinates. Assume $d > 2M$ for now, because this is the case where the rotation can yield a tighter bound. If $d \leq 2M$, this rotation is unhelpful but innocuous, and we can derive the same results without much modification.

Let $\tilde{X} = \Gamma X$, then $\tilde{X}|Z = i \sim \mathcal{N}(\Gamma\boldsymbol{\mu}_i^*, I)$ and $\mathbb{E}\tilde{X} = \mathbb{E}X = 0$. Write

$$\Gamma\boldsymbol{\mu}_i^t = [\tilde{\boldsymbol{\mu}}, \mathbf{0}_{d-2M}], \quad \Gamma\boldsymbol{\mu}_i^* = [\tilde{\boldsymbol{\mu}}_i^*, \mathbf{0}_{d-2M}], \quad \tilde{\boldsymbol{\mu}}_i^*, \tilde{\boldsymbol{\mu}} \in \mathbb{R}^{d-2M}.$$

It thus follows

$$(X-\boldsymbol{\mu}_1^t)(X-\boldsymbol{\mu}_1^t)^T=\Gamma^T\begin{bmatrix}(\tilde{X}^{2M}-\tilde{\boldsymbol{\mu}}_1^t)(\tilde{X}^{2M}-\tilde{\boldsymbol{\mu}}_1^t)^T & (\tilde{X}^{2M}-\tilde{\boldsymbol{\mu}}_1^t)(\tilde{X}^{d-2M})^T\\ (\tilde{X}^{d-2M})(\tilde{X}^{2M}-\tilde{\boldsymbol{\mu}}_1^t)^T & (\tilde{X}^{d-2M})(\tilde{X}^{d-2M})^T\end{bmatrix}\Gamma.$$

Since $\tilde{X}^{d-2M}\sim\mathcal{N}(0,I_{d-2M})$, it is independent of $\tilde{X}^{2M}$. Also note that $w_1(X;\boldsymbol{\mu}^t)$ only depends on $\tilde{X}^{2M}$ (the part involving X^{d-2M} cancels out), we have

$$\begin{aligned}&\left\|\mathbb{E}w_1(X;\boldsymbol{\mu}^t)(1-w_1(X;\boldsymbol{\mu}^t))(X-\boldsymbol{\mu}_1^t)(X-\boldsymbol{\mu}_1^t)^T\right\|_{op}=\left\|\begin{bmatrix}D_1 & 0\\ 0 & D_2\end{bmatrix}\right\|_{op}\\ &\le\max(\|D_1\|_{op},\|D_2\|_{op}).\end{aligned}$$

Applying Lemma A.1 with dimension $\min\{2M,d\}$ and $p=2$, we have

$$\|D_1\|_{op}\le 2M\Big(\frac{3}{2}R_{\max}+\min\{2M,d\}\Big)^2\exp\Big(-\Big(\frac{1}{2}R_{\min}-a\Big)\min\{d,2M\}^{\frac{1}{2}}/8\Big).\tag{30}$$

Applying Lemma A.1 with dimension $\min\{2M,d\}$ and $p=0$, we have

$$\begin{aligned}\|D_2\|_{op}&=\mathbb{E}\Big[w_1(\tilde{X}_{2M};\tilde{\boldsymbol{\mu}}^t)(1-w_1(\tilde{X}_{2M};\tilde{\boldsymbol{\mu}}^t))\Big]\\ &\le 2M\exp\Big(-\Big(\frac{1}{2}R_{\min}-a\Big)\min\{d,2M\}^{\frac{1}{2}}/8\Big).\end{aligned}\tag{31}$$

Combining (30), (31) with (29), we see

$$\begin{aligned}V_1&\le 2M\bigg(\Big(\frac{3}{2}R_{\max}+\min\{2M,d\}\Big)^2+a\Big(\frac{3}{2}R_{\max}+\min\{2M,d\}\Big)\bigg)\\ &\quad\cdot\exp\Big(-\Big(\frac{1}{2}R_{\min}-a\Big)\min\{d,2M\}^{\frac{1}{2}}/8\Big)\\ &\le 2M\Big(2R_{\max}+\min\{2M,d\}\Big)^2\exp\Big(-\Big(\frac{1}{2}R_{\min}-a\Big)\min\{d,2M\}^{\frac{1}{2}}/8\Big).\end{aligned}\tag{32}$$

Next we move to $V_i, i\neq 1$. Using the same decomposition

$$\begin{aligned}V_i&\le\sup_{t\in[0,1]}\left\|\mathbb{E}w_1(X;\boldsymbol{\mu}^t)w_2(X;\boldsymbol{\mu}^t)(X-\boldsymbol{\mu}_1^t)(X-\boldsymbol{\mu}_i^t)^T\right\|_{op}\\ &\quad+a\sup_{t\in[0,1]}\left\|\mathbb{E}w_1(X;\boldsymbol{\mu}^t)w_i(X;\boldsymbol{\mu}^t)(X-\boldsymbol{\mu}_i^t)\right\|_2.\end{aligned}\tag{33}$$

Apply the same rotation trick, we are able to show

$$V_i\le\frac{2}{\kappa}\Big(2R_{\max}+\min\{2M,d\}\Big)^2\exp\Big(-\Big(\frac{1}{2}R_{\min}-a\Big)\min\{d,2M\}^{\frac{1}{2}}/8\Big).\tag{34}$$

Since $\kappa\le\frac{1}{M}$, we see

$$\max_{i\in[M]}V_i\le\frac{2}{\kappa}\Big(2R_{\max}+\min\{2M,d\}\Big)^2\exp\Big(-\Big(\frac{1}{2}R_{\min}-a\Big)\min\{d,2M\}^{\frac{1}{2}}/8\Big).$$

The proof is now complete.

A.3. Proof of Lemma 5.2

First, we can apply the same rotation trick to reduce the effective dimension to $\min\{M, d\}$. To do so, let Γ be a rotation matrix such that the last $[d-M]_+$ coordinates of $\Gamma\boldsymbol{\mu}_i$ are zero for all $i \in [M]$. Write $\Gamma\boldsymbol{\mu}_i = (\tilde{\boldsymbol{\mu}}_i, \mathbf{0}_{[d-M]^+})$, $\tilde{X} = \Gamma X$ and we have

$$\|X - \boldsymbol{\mu}_i\|_2^2 = \|\Gamma X - \Gamma\boldsymbol{\mu}_i\|_2^2 = \|\tilde{X}_M - \tilde{\boldsymbol{\mu}}_i\|_2^2 + \|\tilde{X}_{[d-M]_+}\|_2^2.$$

This implies $w_1(X;\boldsymbol{\mu}) = w_1(\tilde{X}^M;\tilde{\boldsymbol{\mu}})$ and $\mathbb{E}w_1(X;\boldsymbol{\mu}) = \mathbb{E}w_1(\tilde{X}^M;\tilde{\boldsymbol{\mu}})$ where $\tilde{X}_M|Z = i \sim \mathcal{N}((\Gamma\boldsymbol{\mu}_i^*)_M, I_M)$. We thus have successfully reduced the effective dimension to $\min\{M, d\}$.

The rotation step is optional and can only reduce dimension when $d > M$. For ease of notation, let us assume $M \geq d$ and we opt not to do it. The next step in bounding $\mathbb{E}w_1(X;\boldsymbol{\mu})$ is to restrict ourselves to the event where a) X is generated by the first cluster, and b) X lies in a ball $\mathcal{B}(\boldsymbol{\mu}_1^*, r)$ for some radius to be selected later. Specifically, we have

$$\mathbb{E}w_1(X;\boldsymbol{\mu}) \geq \pi_1 \mathbb{E}_{X\sim\mathcal{N}(\boldsymbol{\mu}_1^*, I)} w_1(X;\boldsymbol{\mu}) \geq \pi_1 \int_{\mathcal{B}(\boldsymbol{\mu}_1^*, r)} w_1(x;\boldsymbol{\mu})\phi(x;\boldsymbol{\mu}_1^*)dx. \quad (35)$$

Notice on $\mathcal{B}(\boldsymbol{\mu}_i^*, r)$, by triangular inequality we have

$$\begin{aligned} \|x - \mu_1\|_2 &\leq r + a \\ \|x - \mu_i\|_2 &\geq R_{\min} - r - a, \forall i \neq 1. \end{aligned}$$

Also, since $w_1(x;\mu)$ is decreasing in $\|x - \mu_1\|_2$ and increasing in $\|x - \mu_i\|_2$, we have

$$\begin{aligned} w_1(x;\mu) &\geq \frac{\pi_1 e^{-\frac{(r+a)^2}{2}}}{\pi_1 e^{-\frac{(r+a)^2}{2}} + (1-\pi_1)e^{-\frac{(R_{\min}-r-a)^2}{2}}} \\ &\geq 1 - \frac{1-\pi_1}{\pi_1}\exp\Big(-\frac{1}{2}R_{\min}(R_{\min} - 2r - 2a)\Big) \\ &\geq 1 - \frac{1-\kappa}{\kappa}\exp\Big(-\frac{1}{2}R_{\min}(R_{\min} - 2r - 2a)\Big), \end{aligned}$$

where in the second to the last step, we used numerical inequality $\frac{a}{a+b} \geq 1 - \frac{b}{a}$. It thus follows

$$\begin{aligned} &\pi_1 \int_{\mathcal{B}(\boldsymbol{\mu}_1^*, r)} w_1(x;\boldsymbol{\mu})\phi(x;\boldsymbol{\mu}_1^*)dx \\ &\geq \pi_1 \left(1 - \frac{1-\kappa}{\kappa}\exp\Big(-\frac{1}{2}R_{\min}(R_{\min} - 2r - 2a)\Big)\right) \int_{\mathcal{B}(\boldsymbol{\mu}_1^*, r)} \phi(x;\boldsymbol{\mu}_1^*)dx \\ &\geq \kappa \left(1 - \frac{1-\kappa}{\kappa}\exp\Big(-\frac{1}{2}R_{\min}(R_{\min} - 2r - 2a)\Big)\right) \mathbb{P}(\|\varepsilon\|_2 \leq r), \end{aligned} \quad (36)$$

where $\varepsilon \sim \mathcal{N}(\mathbf{0}, I_d)$. Moving forward, we naturally want to lower bound $\mathbb{P}(\|\varepsilon\|_2 \leq r)$, and the following lemma (Lemma 8 in [31]) allows us to do so.

Lemma A.2. *Let $X \sim \mathcal{N}(\mathbf{0}, I_d)$, for $r \geq 2\sqrt{d}$, we have*

$$\mathbb{P}(\|X\|_2 \geq r) \leq \exp\big(-\frac{r\sqrt{d}}{2}\big).$$

Let us pretend for now that $\frac{1}{2}R_{\min}(R_{\min} - 2r - 2a) \geq \frac{r\sqrt{d}}{2}$, then by chaining (35), (36) and applying Lemma A.2, we have

$$\begin{aligned}
&\mathbb{E}w_1(X;\boldsymbol{\mu})\\
\geq&\kappa\bigg(1 - \frac{1-\kappa}{\kappa}\exp\big(-\frac{r\sqrt{d}}{2}\big)\bigg)\bigg(1 - \exp\big(-\frac{r\sqrt{d}}{2}\big)\bigg)\\
\geq&\kappa - \exp\big(-\frac{r\sqrt{d}}{2}\big).
\end{aligned}$$

Therefore to let $\mathbb{E}w_1(X;\boldsymbol{\mu}) \geq \frac{3\kappa}{4}$, it suffices to let

$$\exp\big(-\frac{r\sqrt{d}}{2}\big) \leq \frac{\kappa}{4}.$$

Now we collect all the conditions we need for all the inequalities to go through; they are

(C1) $r \geq 2\sqrt{d}$. (Lemma A.2)
(C2) $\frac{1}{2}R_{\min}(R_{\min} - 2r - 2a) \geq \frac{r\sqrt{d}}{2}$
(C3) $\exp\big(-\frac{r}{2}\sqrt{d}\big) \leq \kappa/4$.

Setting $r = \frac{R_{\min}/2 - a}{4}$, we can check

1. $a \leq \frac{1}{2}R_{\min} - 8\sqrt{d}$ implies (C1).
2. $R_{\min} \geq \sqrt{d}/6$ implies (C2).
3. $a \leq \frac{1}{2}R_{\min} - \frac{8}{\sqrt{d}}\log(\frac{4}{\kappa})$ implies (C3).

If we replace all d by $\min\{M, d\}$, the proof goes through with only notational changes. Now, we can readily check the conditions on $R_{\min}$ and a in Lemma 5.2 imply the three conditions above. The proof is complete.

Appendix B

In appendix B, we prove Lemma 5.4 and Lemma 5.3 which facilitate the proof of Theorem 3.3. We first introduce some preliminary results on sub-gaussian random variables and then prove Lemma 5.4 and Lemma 5.3.

B.1. Preliminaries

Lemma B.1. *Let X be a random variable.*

1. ***(Sub-gaussian random variable).*** *X is called sub-gaussian if there exists a finite t such that $\mathbb{E}\exp\big(X^2/t^2\big) \le 2$. For a sub-gaussian X, its sub-gaussian norm $\|X\|_{\psi_2}$ is defined as*

$$\|X\|_{\psi_2} = \inf\{t > 0 : \ \mathbb{E}\exp\big(X^2/t^2\big) < \infty\}.$$

2. ***(Hoeffding's inequality).*** *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-gaussian random variables. Then, for every $t \ge 0$, we have*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{N} X_i\Big| \ge t\Big) \le 2\exp\Big(-\frac{ct^2}{\sum_{i=1}^{N}\|X_i\|_{\psi_2}^2}\Big),$$

where c is an absolute constant.

3. ***(Centering).*** *If X is a sub-gaussian random variable, then $X - \mathbb{E}X$ is sub-gaussian too, and*

$$\|X - \mathbb{E}X\|_{\psi_2} \le C\|X\|_{\psi_2},$$

where C is an absolute constant.

4. ***(Bounded random variable is sub-gaussian).*** *Any bounded random variable X is sub-gaussian, with*

$$\|X\|_{\psi_2} \le C\|X\|_{\infty},$$

where $C = 1/\sqrt{\log 2}$.

5. *Let X be a bounded random variable on $[0, 1]$, Y be a sub-gaussian random variable. Then,*

$$\|XY\|_{\psi_2} \le \|Y\|_{\psi_2}.$$

Proof of Lemma B.1. Properties 1–4 are standard results from chapter 2 of [27]; 5 is a consequence of taking expectation and infimum on both sides of

$$\exp\big(\frac{X^2Y^2}{t^2}\big) \le \exp\big(\frac{Y^2}{t^2}\big).$$

Lemma B.2. *Let X be the mixture of M unit variance gaussian distributions on $\mathbb{R}$, with centers denoted by $\{\theta_i\}_{i\in[M]}$ and mixing proportions by $\{\pi_i\}_{i\in[M]}$. Suppose $\max_{i\in[M]}|\theta_i| \le R$ for some constant R. Then X is sub-gaussian with sub-gaussian norm*

$$\|X\|_{\psi_2} \le C\max(R, 1)$$

for some absolute constant C.

Proof of Lemma B.2. Let Y be a random draw from centers $\{\theta_i\}_{i\in[M]}$ according to probabilities $\{\pi_i\}_{i\in[M]}$, i.e. $P(Y = \theta_i) = \pi_i$. It follows that Y is a bounded random variable and $\|Y\|_{\psi_2} \le C_1R$. Let $\varepsilon \sim \mathcal{N}(0, 1)$. From standard results we know $\|\varepsilon\|_{\psi_2} \le C_2$. Note that X has the same distribution as $Y + \varepsilon$, we have

$$\|X\|_{\psi_2} = \|Y + \varepsilon\|_{\psi_2} \le \|Y\|_{\psi_2} + \|\varepsilon\|_{\psi_2} \le C_1R + C_2 \le C\max(R, 1).$$

B.2. Proof of Lemma 5.3

Define $L := 1.5R_{\max}$, then for $a \le \frac{1}{2}R_{\min}$, we have $\otimes_i \mathcal{B}(\boldsymbol{\mu}_i^*, a) \in \otimes_i \mathcal{B}(0, L)$. This is a natural consequence of $\|\boldsymbol{\mu}_i^*\|_2 \le R_{\max}$ for all i. To see why $\|\boldsymbol{\mu}_i^*\|_2 \le R_{\max}$, suppose the opposite and, without loss of generality, let $\|\boldsymbol{\mu}_1^*\|_2 > R_{\max}$, then all $\boldsymbol{\mu}_i^* \in \mathcal{B}(\boldsymbol{\mu}_1^*, R_{\max})$. Since $\mathbb{E}X = \sum \pi_i \boldsymbol{\mu}_i^* = 0$ but $\mathcal{B}(\boldsymbol{\mu}_1^*, R_{\max})$ does not contain the origin, we get a contradiction. Also note that since Theorem 3.1 requires $R_{\min} \ge C_0 \min\{d, 2M\}^{\frac{1}{2}}$ for a large C_0, we can work under the premise that $R_{\max} \ge 1$, because otherwise, even the population level convergence result does not apply.

Denote $\mathcal{U} = \otimes_i \mathcal{B}(0, L)$, and we establish all uniform convergence results on $\mathcal{U}$. Let n_ε be the ε-covering number of $\mathcal{B}(0, L)$, then standard results [27] have it $\log(n_\varepsilon) \le d \log(3L/\varepsilon)$. By doing cartesian product on such covers, we can get a cover on $\mathcal{U}$. We denote this cover by $M_\varepsilon = \{\boldsymbol{\mu}^1, \dots, \boldsymbol{\mu}^{N_\varepsilon}\}$ with $M_\varepsilon \subset \mathbb{R}^{Md}$ and $\log(N_\varepsilon) \le Md \log(3L/\varepsilon)$. For any $\boldsymbol{\mu} \in \otimes_i \mathcal{B}(0, L)$, let $j(\boldsymbol{\mu}) = \arg\min_{j \in [N_\varepsilon]} \|\boldsymbol{\mu} - \boldsymbol{\mu}^j\|_2$. Then for all $\boldsymbol{\mu} \in \mathcal{U}$, $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^{j(\boldsymbol{\mu})}\|_2 \le \varepsilon$ for $\forall i \in [M]$.

Define

$$I_{n,1}(\boldsymbol{\mu}) := \Big\| \frac{1}{n} \Big(\sum_{i=1}^n w_1(X_i; \boldsymbol{\mu})(X_i - \boldsymbol{\mu}_1^*) \Big) - \mathbb{E}[w_1(X; \boldsymbol{\mu})(X - \boldsymbol{\mu}_1^*)] \Big\|_2$$

Start by noting

$$\begin{aligned} I_{n,1}(\boldsymbol{\mu}) \le & \Big\| \frac{1}{n} \Big(\sum_{i=1}^n w_1(X_i; \boldsymbol{\mu})(X_i - \boldsymbol{\mu}_1^*) \Big) - \frac{1}{n} \Big(\sum_{i=1}^n w_1(X_i; \boldsymbol{\mu}^{j(\boldsymbol{\mu})})(X_i - \boldsymbol{\mu}_1^*) \Big) \Big\|_2 \\ & + \Big\| \frac{1}{n} \Big(\sum_{i=1}^n w_1(X_i; \boldsymbol{\mu}^{j(\boldsymbol{\mu})})(X_i - \boldsymbol{\mu}_1^*) \Big) - \mathbb{E}[w_1(X; \boldsymbol{\mu}^{j(\boldsymbol{\mu})})(X - \boldsymbol{\mu}_1^*)] \Big\|_2 \\ & + \Big\| \mathbb{E}[w_1(X; \boldsymbol{\mu}^{j(\boldsymbol{\mu})})(X - \boldsymbol{\mu}_1^*)] - \mathbb{E}[w_1(X; \boldsymbol{\mu})(X - \boldsymbol{\mu}_1^*)] \Big\|_2. \end{aligned}$$

Then we have

$$\mathbb{P}\Big(\sup_{\boldsymbol{\mu} \in \mathcal{U}} I_{n,1}(\boldsymbol{\mu}) \ge t \Big) \le \mathbb{P}(A_t) + \mathbb{P}(B_t) + \mathbb{P}(C_t),$$

where the events A_t, B_t, C_t are defined as

$$\begin{aligned} A_t = & \Big\{ \sup_{\boldsymbol{\mu} \in \mathcal{U}} \Big\| \frac{1}{n} \Big(\sum_{i=1}^n w_1(X_i; \boldsymbol{\mu})(X_i - \boldsymbol{\mu}_1^*) \Big) - \\ & \qquad \frac{1}{n} \Big(\sum_{i=1}^n w_1(X_i; \boldsymbol{\mu}^{j(\boldsymbol{\mu})})(X_i - \boldsymbol{\mu}_1^*) \Big) \Big\|_2 \ge \frac{t}{3} \Big\}, \\ B_t = & \Big\{ \sup_{j \in [N_\varepsilon]} \Big\| \frac{1}{n} \Big(\sum_{i=1}^n w_1(X_i; \boldsymbol{\mu}^j)(X_i - \boldsymbol{\mu}_1^*) \Big) - \mathbb{E}[w_1(X; \boldsymbol{\mu}^j)(X - \boldsymbol{\mu}_1^*)] \Big\|_2 \ge \frac{t}{3} \Big\}, \\ C_t = & \Big\{ \sup_{\boldsymbol{\mu} \in \mathcal{U}} \Big\| \mathbb{E}[w_1(X; \boldsymbol{\mu}^{j(\boldsymbol{\mu})})(X - \boldsymbol{\mu}_1^*)] - \mathbb{E}[w_1(X; \boldsymbol{\mu})(X - \boldsymbol{\mu}_1^*)] \Big\|_2 \ge \frac{t}{3} \Big\}. \end{aligned}$$

For some $\delta > 0$, we next derive conditions on t that suffice to let

$$\mathbb{P}(A_t) \le \frac{\delta}{2}, \quad \mathbb{P}(B_t) \le \frac{\delta}{2}, \quad \mathbb{P}(C_t) = 0.$$

Then replacing δ with $\frac{1}{n}$ completes the proof.

Upper bounding $\mathbb{P}(B_t)$:

Let $V_{1/2}$ be a $(1/2)$-cover of $\mathcal{B}_d(0,1)$ with $\log|V_{1/2}| \le d\log 6$. Then we know from standard result that

$$\begin{aligned}
&\Big\|\frac{1}{n}\Big(\sum_{i=1}^n w_1(X_i;\boldsymbol{\mu}^j)(X_i-\boldsymbol{\mu}_1^*)\Big) - \mathbb{E}[w_1(X;\boldsymbol{\mu}^j)(X-\boldsymbol{\mu}_1^*)]\Big\|_2 \\
\le& 2\sup_{\mathbf{v}\in V_{1/2}}\Big\langle \mathbf{v}, \frac{1}{n}\Big(\sum_{i=1}^n w_1(X_i;\boldsymbol{\mu}^j)(X_i-\boldsymbol{\mu}_1^*)\Big) - \mathbb{E}[w_1(X;\boldsymbol{\mu}^j)(X-\boldsymbol{\mu}_1^*)]\Big\rangle.
\end{aligned}$$

Taking union bound, we have

$$\begin{aligned}
&\mathbb{P}(B_t) \\
\le& \mathbb{P}\Big(\sup_{\substack{j\in[N_\varepsilon]\\ \mathbf{v}\in V_{1/2}}}\Big\{\frac{1}{n}\sum_{i=1}^n\Big\langle \mathbf{v}, w_1(X_i;\boldsymbol{\mu}^j)(X_i-\boldsymbol{\mu}_1^*) - \mathbb{E}[w_1(X;\boldsymbol{\mu}^j)(X-\boldsymbol{\mu}_1^*)]\Big\rangle\Big\} \ge \frac{t}{6}\Big) \\
\le& \exp\big(Md\log(3L/\varepsilon) + d\log 6\big) \\
&\cdot \sup_{\substack{j\in[N_\varepsilon]\\ \mathbf{v}\in V_{1/2}}} \mathbb{P}\Big(\Big\{\frac{1}{n}\sum_{i=1}^n w_1(X_i;\boldsymbol{\mu}^j)\langle X_i-\boldsymbol{\mu}_1^*,\mathbf{v}\rangle - \mathbb{E}[w_1(X;\boldsymbol{\mu}^j)\langle X-\boldsymbol{\mu}_1^*,\mathbf{v}\rangle]\Big\} \ge \frac{t}{6}\Big).
\end{aligned}$$

By part 5 of Lemma B.1,

$$\|w_1(X;\boldsymbol{\mu}^j)\langle X-\boldsymbol{\mu}_1^*,\mathbf{v}\rangle\|_{\psi_2} \le \|\langle X-\boldsymbol{\mu}_1^*,\mathbf{v}\rangle\|_{\psi_2}.$$

Note that $\langle X-\boldsymbol{\mu}_1^*,\mathbf{v}\rangle$ follows a one dimensional gaussian mixture model with centers at $\{\langle\boldsymbol{\mu}_i^*-\boldsymbol{\mu}_1^*,\mathbf{v}\rangle\}_{i\in[M]}$. Since $|\langle\boldsymbol{\mu}_i^*-\boldsymbol{\mu}_1^*,\mathbf{v}\rangle| \le R_{\max}$ and $R_{\max}\ge 1$, we conclude from Lemma B.2

$$\|\langle X-\boldsymbol{\mu}_1^*,\mathbf{v}\rangle\|_{\psi_2} \le CR_{\max} \le CL.$$

Consequently, by Hoeffding's inequality, we have

$$\mathbb{P}(B_t) \le \exp\big(Md\log(3L/\varepsilon) + d\log 6 - \frac{cnt^2}{L^2}\big).$$

To ensure $\mathbb{P}(B_t) \le \frac{\delta}{2}$, it suffices to let

$$t \ge C\sqrt{\frac{L^2\big(Md\log(\frac{18L}{\varepsilon}) + \log(\frac{2}{\delta})\big)}{n}}.$$

Upper bound $\mathbb{P}(C_t)$:

Using the same integration expression (14) as in section 4.1, we have

$$\sup_{\boldsymbol{\mu}\in\mathcal{U}} \Big\| \mathbb{E}[w_1(X;\boldsymbol{\mu}^{j(\boldsymbol{\mu})})(X-\boldsymbol{\mu}_1^*)] - \mathbb{E}[w_1(X;\boldsymbol{\mu})(X-\boldsymbol{\mu}_1^*)] \Big\|_2 \le \varepsilon \sum_{i=1}^{M} U_i,$$

where

$$U_1 = \mathbb{E} \sup_{\boldsymbol{\mu}\in\mathcal{U}} \big\| w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))(X-\boldsymbol{\mu}_1^*)(X-\boldsymbol{\mu}_1)^T \big\|_{op}, \tag{37}$$

$$U_i = \mathbb{E} \sup_{\boldsymbol{\mu}\in\mathcal{U}} \big\| w_1(X;\boldsymbol{\mu}) w_i(X;\boldsymbol{\mu})(X-\boldsymbol{\mu}_1^*)(X-\boldsymbol{\mu}_i)^T \big\|_{op} \quad \text{for } i \ne 1. \tag{38}$$

Since C_t is deterministic, so as long as we have $\varepsilon \sum_{i=1}^{M} U_i < \frac{t}{3}$, C_t will never happen.

Upper bound $\mathbb{P}(A_t)$**:**

Using Markov inequality, we have

$$\begin{aligned}
&\mathbb{P}(A_t) \\
&\le \frac{3}{t}\mathbb{E}\Big[\sup_{\boldsymbol{\mu}\in\mathcal{U}} \Big\| \frac{1}{n}\Big(\sum_{i=1}^{n} w_1(X_i;\boldsymbol{\mu})(X_i-\boldsymbol{\mu}_1^*)\Big) - \frac{1}{n}\Big(\sum_{i=1}^{n} w_1(X_i;\boldsymbol{\mu}^{j(\boldsymbol{\mu})})(X_i-\boldsymbol{\mu}_1^*)\Big) \Big\|_2 \Big] \\
&\le \frac{3}{t}\mathbb{E} \sup_{\boldsymbol{\mu}\in\mathcal{U}} \Big\| \big(w_1(X;\boldsymbol{\mu}^{j(\boldsymbol{\mu})}) - w_1(X;\boldsymbol{\mu})\big)(X-\boldsymbol{\mu}_1^*) \Big\|_2 \\
&\le \frac{3\varepsilon}{t} \sum_{i=1}^{M} U_i. \quad \text{(due to mean value theorem)}
\end{aligned}$$

To ensure $P(A_t) \le \frac{\delta}{2}$, it suffices to let $t \ge 6\varepsilon(\sum U_i)/\delta$. Note that whenever this holds, the condition that ensures $\mathbb{P}(C_t) = 0$ is implied.

Bounding U_i:

Knowing that all $w \in [0,1]$, we see

$$U_1 \le \mathbb{E}\Big[\sup_{\boldsymbol{\mu}\in\mathcal{U}} \|X-\boldsymbol{\mu}_1^*\| \|X-\boldsymbol{\mu}_1\| \Big] \le \mathbb{E}(\|X\|+L)^2 \le \mathbb{E}(\|Y\|+\|\varepsilon\|+L)^2,$$

where Y is a random draw from centers $\{\boldsymbol{\mu}_i^*\}_{i\in[M]}$ according to probabilities $\{\pi_i\}_{i\in[M]}$ and $\varepsilon \sim \mathcal{N}(\mathbf{0}, I_d)$. With a bit more calculation, we see $U_1 \le C'(L^2+d)$, and the same bound also hold for other U_i. It thus follows

$$\sum_{i=1}^{M} U_i \le C' M(L^2+d).$$

Conclusion:

We set $\varepsilon = \frac{\delta L}{6C' n M(L^2+d)}, \delta = \frac{1}{n}$, then any t satisfy the following ensures bad events happen with probability less than δ

$$t \ge \max\left\{ \frac{L}{n}, CL\sqrt{\frac{Md\log\left(\frac{108C'M(L^2+d)n}{\delta}\right) + \log\frac{2}{\delta}}{n}} \right\}.$$

The second argument in maximum apparently dominates the first argument. After meticulously checking, we conclude there exists a universal constant C_3, such that

$$\mathbb{P}\left(I_{n,1}(\boldsymbol{\mu}) \geq R_{\max}\sqrt{\frac{\tilde{C}_3 M d \log n}{n}}\right) \leq \frac{1}{n},$$

where $\tilde{C}_3 = C_3 \log(M(3R_{\max}^2 + d))$.

B.3. Proof of Lemma 5.4

The proof of Lemma 5.4 is essentially the same as the proof of Lemma 5.3. Start by noticing

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} w_1(X_i;\boldsymbol{\mu}) - \mathbb{E}w_1(X;\boldsymbol{\mu})\right| \leq{}& \left|\frac{1}{n}\sum_{i=1}^{n} w_1(X_i;\boldsymbol{\mu}) - \frac{1}{n}\sum_{i=1}^{n} w_1(X_i;\boldsymbol{\mu}^{j(\boldsymbol{\mu})})\right| \\ &+ \left|\frac{1}{n}\sum_{i=1}^{n} w_1(X_i;\boldsymbol{\mu}^{j(\boldsymbol{\mu})}) - \mathbb{E}w_1(X;\boldsymbol{\mu}^{j(\boldsymbol{\mu})})\right| \\ &+ \left|\mathbb{E}w_1(X;\boldsymbol{\mu}^{j(\boldsymbol{\mu})}) - \mathbb{E}w_1(X;\boldsymbol{\mu})\right|.\end{aligned}$$

Then we have

$$\mathbb{P}\left(\sup_{\boldsymbol{\mu}\in\mathcal{U}}\left|\frac{1}{n}\sum_{i=1}^{n} w_1(X_i;\boldsymbol{\mu}) - \mathbb{E}w_1(X;\boldsymbol{\mu})\right| \geq t\right) \leq \mathbb{P}(A_t) + \mathbb{P}(B_t) + \mathbb{P}(C_t),$$

where the events A_t, B_t, C_t are defined as

$$\begin{aligned}A_t &= \left\{\sup_{\boldsymbol{\mu}\in\mathcal{U}}\left|\frac{1}{n}\sum_{i=1}^{n} w_1(X_i;\boldsymbol{\mu}) - \frac{1}{n}\sum_{i=1}^{n} w_1(X_i;\boldsymbol{\mu}^{j(\boldsymbol{\mu})})\right| \geq \frac{t}{3}\right\}, \\ B_t &= \left\{\sup_{j\in[N_\varepsilon]}\left|\frac{1}{n}\sum_{i=1}^{n} w_1(X_i;\boldsymbol{\mu}^{j(\boldsymbol{\mu})}) - \mathbb{E}w_1(X;\boldsymbol{\mu}^{j(\boldsymbol{\mu})})\right| \geq \frac{t}{3}\right\}, \\ C_t &= \left\{\sup_{\boldsymbol{\mu}\in\mathcal{U}}\left|\mathbb{E}w_1(X;\boldsymbol{\mu}^{j(\boldsymbol{\mu})}) - \mathbb{E}w_1(X;\boldsymbol{\mu})\right| \geq \frac{t}{3}\right\}.\end{aligned}$$

For some $\delta > 0$, we next derive conditions on t that suffice to let

$$\mathbb{P}(A_t) \leq \frac{\delta}{2}, \quad \mathbb{P}(B_t) \leq \frac{\delta}{2}, \quad \mathbb{P}(C_t) = 0.$$

Then replacing δ with $\frac{1}{n}$ completes the proof.

Upper bounding $\mathbb{P}(B_t)$:

Since $w_1(X, \boldsymbol{\mu})$ is bounded between [0, 1], it is sub-gaussian with a bounded norm $\|w_1(X, \boldsymbol{\mu})\|_{\psi_2} \leq C$ for some absolute constant C. We can thus directly apply union bound and Hoeffding's inequality:

$$\mathbb{P}(B_t) \leq \exp\big(Md\log(3L/\varepsilon)\big) \cdot \sup_{j\in[N_\varepsilon]} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} w_1(X_i;\boldsymbol{\mu}^j) - \mathbb{E}w_1(X;\boldsymbol{\mu}^j)\right| \geq \frac{t}{3}\right)$$

$$\leq 2\exp\bigl(Md\log(3L/\varepsilon) - cnt^2\bigr).$$

To ensure $\mathbb{P}(B_t) \leq \frac{\delta}{2}$, it suffices to let

$$t \geq C\sqrt{\frac{Md\log(\frac{3L}{\varepsilon}) + \log(\frac{4}{\delta})}{n}}.$$

Upper bound $\mathbb{P}(C_t)$:

Using the same integration expression (14) as in section 4.1, we have

$$\sup_{\boldsymbol{\mu}\in\mathcal{U}} \Bigl|\mathbb{E}w_1(X;\boldsymbol{\mu}^{j(\boldsymbol{\mu})}) - \mathbb{E}w_1(X;\boldsymbol{\mu})\Bigr| \leq \varepsilon\sum_{i=1}^{M} W_i,$$

where

$$W_1 = \mathbb{E}\sup_{\boldsymbol{\mu}\in\mathcal{U}} \bigl\|w_1(X;\boldsymbol{\mu})(1-w_1(X;\boldsymbol{\mu}))(X-\boldsymbol{\mu}_1)\bigr\|_2, \tag{39}$$

$$W_i = \mathbb{E}\sup_{\boldsymbol{\mu}\in\mathcal{U}} \bigl\|w_1(X;\boldsymbol{\mu})w_i(X;\boldsymbol{\mu})(X-\boldsymbol{\mu}_i)\bigr\|_2 \quad \text{for } i\neq 1. \tag{40}$$

Since C_t is deterministic, so as long as we have $\varepsilon\sum_{i=1}^{M} W_i < \frac{t}{3}$, C_t will never happen.

Upper bound $\mathbb{P}(A_t)$**:**

Using Markov inequality, we have

$$\begin{aligned}
\mathbb{P}(A_t) &\leq \frac{3}{t}\mathbb{E}\Biggl[\sup_{\boldsymbol{\mu}\in\mathcal{U}}\Bigl|\frac{1}{n}\sum_{i=1}^{n} w_1(X_i;\boldsymbol{\mu}) - \frac{1}{n}\sum_{i=1}^{n} w_1(X_i;\boldsymbol{\mu}^{j(\boldsymbol{\mu})})\Bigr|\Biggr] \\
&\leq \frac{3}{t}\mathbb{E}\Biggl[\sup_{\boldsymbol{\mu}\in\mathcal{U}}\Bigl|w_1(X;\boldsymbol{\mu}^{j(\boldsymbol{\mu})}) - w_1(X;\boldsymbol{\mu})\Bigr|\Biggr] \\
&\leq \frac{3\varepsilon}{t}\sum_{i=1}^{M} W_i. \quad \text{(due to mean value theorem)}
\end{aligned}$$

To ensure $P(A_t) \leq \frac{\delta}{2}$, it suffices to let $t \geq 6\varepsilon(\sum U_i)/\delta$. Note that whenever this holds, the condition that ensures $\mathbb{P}(C_t) = 0$ is implied.

Bounding W_i:

Knowing that all $w \in [0,1]$, we see

$$W_1 \leq \mathbb{E}\Bigl[\sup_{\boldsymbol{\mu}\in\mathcal{U}} \|X-\boldsymbol{\mu}_1\|\Bigr] \leq \mathbb{E}\bigl[\|X\| + L\bigr] \leq \mathbb{E}\bigl[\|Y\| + \|\varepsilon\| + L\bigr],$$

where Y is a random draw from centers $\{\boldsymbol{\mu}_i^*\}_{i\in[M]}$ according to probabilities $\{\pi_i\}_{i\in[M]}$ and $\varepsilon \sim \mathcal{N}(\mathbf{0}, I_d)$. With a bit more calculation, we see $W_1 \leq C'(L + \sqrt{d})$, and the same bound also hold for other W_i. It thus follows

$$\sum_{i=1}^{M} W_i \leq C'M(L+\sqrt{d}).$$

Conclusion:

We set $\varepsilon = \frac{\delta}{6C'nM(L+\sqrt{d})}, \delta = \frac{1}{n}$, then any t satisfy the following ensures bad events happen with probability less than δ

$$t \geq \max\left\{\frac{1}{n}, C\sqrt{\frac{Md\log\left(\frac{18C'ML(L+\sqrt{d})n}{\delta}\right)+\log\frac{4}{\delta}}{n}}\right\}.$$

The second argument in maximum apparently dominates the first argument. After meticulously checking, we conclude there exists a universal constant C_2, such that

$$\mathbb{P}\left(\sup_{\boldsymbol{\mu}\in\mathcal{U}}\left|\sum_{i=1}^{n} w_1(X_i,\boldsymbol{\mu}) - \mathbb{E}w_1(X;\boldsymbol{\mu})\right| \geq \sqrt{\frac{\tilde{C}_2 Md\log n}{n}}\right) \leq \frac{1}{n},$$

where $\tilde{C}_2 = C_2\log(2MR_{\max}(2R_{\max}+\sqrt{d}))$.

References

[1] Achlioptas, D. and McSherry, F. (2005). On Spectral Learning of Mixtures of Distributions. In *COLT*. MR2203280

[2] Arora, S. and Kannan, R. (2005). Learning mixtures of separated nonspherical Gaussians. *The Annals of Applied Probability* 69–92. MR2115036

[3] Balakrishnan, S., Wainwright, M. J. and Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics* 77–120. MR3611487

[4] Belkin, M. and Sinha, K. (2010). Polynomial Learning of Distribution Families. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science* 103–112. MR3024780

[5] Brubaker, S. C. and Vempala, S. (2008). Isotropic PCA and Affine-Invariant Clustering. In *Proceedings of the 2008 49th Annual IEEE Symposium on Foundations of Computer Science* 551–560. MR2484643

[6] Cai, T., Ma, J. and Zhang, L. CHIME: Clustering of High-dimensional Gaussian Mixtures with EM Algorithm and its Optimality. *The Annals of Statistics*. To appear. MR3911111

[7] Chaudhuri, K. and Rao, S. (2008). Learning Mixtures of Product Distributions using Correlations and Independence. In *Twenty-First Annual Conference on Learning Theory* 9–20.

[8] Chaudhuri, K., Kakade, S. M., Livescu, K. and Sridharan, K. (2009). Multi-view Clustering via Canonical Correlation Analysis. In *Proceedings of the 26th Annual International Conference on Machine Learning* 129–136.

[9] Dasgupta, S. (1999). Learning mixtures of Gaussians. In *40th Annual Symposium on Foundations of Computer Science* 634–644. MR1917603

[10] Dasgupta, S. and Schulman, L. J. (2007). A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research* **8** 203–226. MR2320668

[11] Daskalakis, C., Tzamos, C. and Zampetakis, M. (2017). Ten Steps of EM Suffice for Mixtures of Two Gaussians. In *Proceedings of the 2017 Conference on Learning Theory* **65** 704–710.
[12] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39** 1–38. MR0501537
[13] Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics* 1233–1263. MR1873329
[14] Hardt, M. and Price, E. (2015). Tight Bounds for Learning a Mixture of Two Gaussians. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing* 753–760. MR3388255
[15] Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics* 2844–2870. MR3851757
[16] Hsu, D. and Kakade, S. M. (2013). Learning Mixtures of Spherical Gaussians: Moment Methods and Spectral Decompositions. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*. MR3385380
[17] Jin, C., Zhang, Y., Balakrishnan, S., J. Wainwright, M. and Jordan, M. (2016). Local Maxima in the Likelihood of Gaussian Mixture Models: Structural Results and Algorithmic Consequences. In *Advances in Neural Information Processing Systems* **29**.
[18] Kalai, A. T., Moitra, A. and Valiant, G. (2010). Efficiently Learning Mixtures of Two Gaussians. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing* 553–562. MR2743304
[19] Kannan, R., Salmasian, H. and Vempala, S. (2008). The spectral method for general mixture models. *SIAM Journal on Computing* **38** 1141–1156. MR2421081
[20] Klusowski, J. M. and Brinda, W. D. (2016). Statistical Guarantees for Estimating the Centers of a Two-component Gaussian Mixture by EM. *arXiv preprint. arXiv:1608.02280*.
[21] Lu, Y. and Zhou, H. H. (2016). Statistical and Computational Guarantees of Lloyd's Algorithm and its Variants. *arXiv preprint.* arXiv:1612.02099.
[22] Mei, S., Bai, Y. and Montanari, A. The Landscape of Empirical Risk for Non-convex Losses. *arXiv preprint.* arXiv:1607.06534. MR3851754
[23] Moitra, A. and Valiant, G. (2010). Settling the Polynomial Learnability of Mixtures of Gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science* 93–102. MR3024779
[24] Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics* 370–400. MR3059422
[25] Tseng, P. (2004). An analysis of the EM algorithm and entropy-like proximal point methods. *Mathematics of Operations Research* **29** 27–44. MR2065712
[26] Vempala, S. and Wang, G. (2004). A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences* **68** 841–860.

MR2059647
[27] Vershynin, R. *High-Dimensional Probability: An introduction with Applications in Data Science*. MR3837109
[28] Wang, Z., Gu, Q., Ning, Y. and Liu, H. (2015). High Dimensional EM Algorithm: Statistical Optimization and Asymptotic Normality. In *Advances in Neural Information Processing Systems* **28** 2521–2529.
[29] Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* 95–103. MR0684867
[30] Xu, J., Hsu, D. and Maleki, A. (2016). Global Analysis of Expectation Maximization for Mixtures of Two Gaussians. In *Advances in Neural Information Processing Systems* **29**.
[31] Yan, B., Yin, M. and Sarkar, P. (2017). Convergence of Gradient EM on Multi-component Mixture of Gaussians. In *Advances in Neural Information Processing Systems* **30**.
[32] Yi, X. and Caramanis, C. (2015). Regularized EM Algorithms: A Unified Framework and Statistical Guarantees. In *Advances in Neural Information Processing Systems* **28** 1567–1575.