Sharp Analysis of Expectation-Maximization for Weakly Identifiable Models

Raaz Dwivedi[†,⋆] **Koulik Khamaru**[◊,⋆] **Nhat Ho**[†,⋆]
Martin J. Wainwright[†,◊] **Michael I. Jordan**[†,◊] **Bin Yu**[†,◊]
Department of [†]EECS and [◊]Statistics, UC Berkeley

Abstract

We study a class of weakly identifiable location-scale mixture models for which the maximum likelihood estimates based on n i.i.d. samples are known to have lower accuracy than the classical $n^{-\frac{1}{2}}$ error. We investigate whether the Expectation-Maximization (EM) algorithm also converges slowly for these models. We provide a rigorous characterization of EM for fitting a weakly identifiable Gaussian mixture in a univariate setting where we prove that the EM algorithm converges in order $n^{\frac{3}{4}}$ steps and returns estimates that are at a Euclidean distance of order $n^{-\frac{1}{8}}$ and $n^{-\frac{1}{4}}$ from the true location and scale parameter respectively. Establishing the slow rates in the univariate setting requires a novel localization argument with two stages, with each stage involving an epoch-based argument applied to a different surrogate EM operator at the population level. We demonstrate several multivariate ($d \geq 2$) examples that exhibit the same slow rates as the univariate case. We also prove slow statistical rates in higher dimensions in a special case, when the fitted covariance is constrained to be a multiple of the identity.

1 Introduction

Gaussian mixture models [Pearson, 1894] have been used widely to model heterogeneous data in many applications arising from physical and the biological sciences. In several scenarios, the data has a large number of subpopulations and the mixture components in the data may not be well-separated. In such settings, estimating the true number of components may be difficult, so that one may end up fitting a mixture model with a number of components larger than that present in the data. Such mixture fits, referred to as *over-specified mixture distributions*, are commonly used by practitioners in order to deal with uncertainty in the number of components in the data [Rousseau and Mengersen, 2011, Havre et al., 2015]. However, a deficiency of such models is that they are *singular*, meaning that their Fisher information matrices are degenerate. Given the popularity of over-specified models in practice, it is important to understand how methods for parameter estimation, including maximum likelihood and the EM algorithm, behave when applied to such models.

1.1 Background and past work

In the context of singular mixture models, an important distinction is between those that are *strongly* versus *weakly* identifiable. Chen [Chen, 1995] studied the class of strongly identifiable models in which, while the Fisher information matrix may be degenerate at a point, and it is not degenerate over a larger set. Studying over-specified Gaussian mixtures with known scale parameters, he showed that the accuracy of the MLE for the unknown location parameter is of the order $n^{-\frac{1}{4}}$, which should be contrasted with the classical $n^{-\frac{1}{2}}$ rate achieved in regular settings. A line of follow-up work has extended this type of analysis to other types of strongly identifiable mixture models; see the papers [Ishwaran et al., 2001, Rousseau and Mengersen, 2011, Nguyen, 2013, Heinrich and Kahn, 2018] as well as the references therein for more details.

A more challenging class of mixture models are those that are only *weakly identifiable*, meaning that the Fisher information is degenerate over some larger set. This stronger form of singularity arises, for instance, when the scale parameter in an over-specified Gaussian mixture is also unknown [Chen et al., 2001, Chen and Li, 2009]. Ho et al. [Ho and Nguyen, 2016a] characterized the behavior of MLE for a class of weakly identifiable models. They showed that the conver-

[⋆] The first three authors contributed equally.

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

gence rates of MLE in these models could be very slow, with the precise rates determined by algebraic relations among the partial derivatives. However, this past work has not addressed the computational complexity of computing the MLE in a weakly identifiable model.

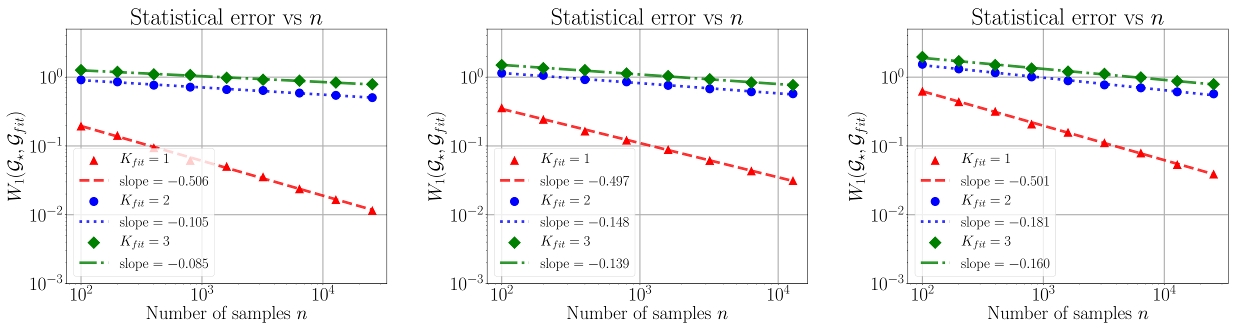
The focus of this paper is the intersection of statistical and computational issues associated with fitting the parameters of weakly identifiable mixture models. In particular, we study the expectation-maximization (EM) algorithm [Dempster et al., 1997, Wu, 1983, Redner and Walker, 1984], which is the most popular algorithm for computing (approximate) MLEs in the mixture models. It is an instance of a minorization-maximization algorithm, in which at each step, a suitably chosen lower bound of the log-likelihood is maximized. There is now a lengthy line of work on the behavior of EM when applied to regular models. The classical papers [Wu, 1983, Tseng, 2004, Chrétien and Hero, 2008] establish the asymptotic convergence of EM to a local maximum of the log-likelihood function for a general class of incomplete data models. Other papers [Jordan and Xu, 1995, Xu and Jordan, 1996, Ma et al., 2000] characterized the rate of convergence of EM for regular Gaussian mixtures. More recent years have witnessed a flurry of work on the behavior of EM for various kinds of regular mixture models [Balakrishnan et al., 2017, Wang et al., 2015, Yi and Caramanis, 2015, Xu et al., 2016, Daskalakis et al., 2017, Yan et al., 2017, Hao et al., 2018, Cai et al., 2019]; as a consequence, our understanding of EM in such cases is now relatively mature. More precisely, it is known that for Gaussian mixtures, EM converges in $\mathcal{O}(\log(n/d))$-steps to parameter estimates that lie within Euclidean distance $\mathcal{O}((d/n)^{1/2})$ of the true location parameters, assuming minimal separation between the mixture components.

In our recent work [Dwivedi et al., 2020], we studied the behavior of EM for fitting a class of *non-regular* mixture models, namely those in which the Fisher information is degenerate at a point, but the model remains strongly identifiable. One such class of models are Gaussian location mixtures with known scale parameters that are *over-specified*, meaning that the number of components in the mixture-fit exceeds the number of components in the data generating distribution. For such non-regular but strongly identifiable mixture models, they [Dwivedi et al., 2020] showed that the EM algorithm takes $\mathcal{O}((n/d)^{\frac{1}{2}})$ steps to converge to a Euclidean ball of radius $\mathcal{O}((d/n)^{\frac{1}{4}})$ around the true location parameter. Recall that for such models, the MLE is known to lie at a distance $\mathcal{O}(n^{-\frac{1}{4}})$ from the true parameter [Chen, 1995], so that even though its convergence rate as an optimization algorithm is slow; the EM algorithm nonetheless produces a solution with a statistical error of the same order as the MLE. This past work does not consider the more realistic setting in which both the location and scale parameters are unknown, and the EM algorithm is used to fit both simultaneously. Indeed, as mentioned earlier, such models may become weakly identifiable due to algebraic relations among the partial derivatives [Chen and Li, 2009]. Thus, analyzing EM in the case of weakly identifiable mixtures is challenging for two reasons: the weak separation between the mixture components, and the algebraic interdependence of the partial derivatives of the log-likelihood. The main contributions of this work are (a) to highlight the dramatic differences in the convergence behavior of the EM algorithm, depending on the structure of the fitted model relative to the data-generating distribution; and (b) to analyze the EM algorithm under a few specific yet representative settings of weakly identifiable models, giving a precise analytical characterization of its convergence behavior.

1.2 Some illustrative examples

Before proceeding further, we summarize a few common aspects of the numerical experiments and the associated figures presented in the paper. Computations at the population-level were done via numerical integration on a sufficiently fine grid. For EM with finite sample size n, we track its performance for several values of $n \in \{100, 200, 400, \ldots, \}$ and report the quantity $\widehat{m}_e + 2\widehat{s}_e$ on the y-axis, where $\widehat{m}_e$ and $\widehat{s}_e$, respectively, denote the mean and standard deviation across the experiments for the metric under consideration (as a function of n on the x-axis, e.g., Wasserstein error for parameter estimation in Figure 1. The stopping criteria for sample EM were: (a) the change in the iterates was small enough ($< .001/n$), or (b) the number of iterations was too large (greater than $100,000$); criteria (a) led to convergence in most experiments. Furthermore, whenever we provide a slope, it is the slope for the least-squares fit on the log-log scale for the quantity on y-axis when fitted with the quantity reported on the x-axis. For instance, in Figure 1(a), we plot the Wasserstein error between the estimated mixture and the true mixture on the y-axis value versus the sample size n on the x-axis and also provide the slopes for the least-squares fit. In particular, in panel (a) the green dot-dashed line with the legend 'slope= -0.09' denotes the least-squares fit and the respective slope for the logarithmic error $\log W_1(\mathcal{G}_*, G_{\text{fit}})$ (green diamonds) with respect to the logarithmic sample size $\log n$ when the number of components in the fitted mixture is 3. Such a result implies that the error $W_1(\mathcal{G}_*, G_{\text{fit}})$ scales as $n^{-0.09}$ with the sample size n in our experiments.

To begin with, we consider the simplest case of over-



(a) $d = 1$ (b) $d = 2$ (c) $d = 4$

Figure 1. Scaling of the Wasserstein error between the true parameters and the EM estimates, when EM is used to fit a Gaussian mixture model with $K_{\text{fit}} \in \{1, 2, 3\}$ components, i.e., $\mathcal{G}_{\text{fit}} = \sum_{i=1}^{K_{\text{fit}}} w_i \mathcal{N}(\mu_i, \Sigma_i)$, on an n sample-dataset generated from standard Gaussian distribution $\mathcal{G}_* = \mathcal{N}(0, I_d)$. In all three examples, when the fitted model is over-specified, meaning that the fitted model has more components than the true model ($K_{\text{fit}} \in \{2, 3\}$ in these examples), we observe a significant increase in the Wasserstein error. Stated differently, the simulations suggest that the estimation accuracy of the EM algorithm degrades dramatically when the fitted model is over-specified.

specification with Gaussian mixture models—when the true data is generated from a zero-mean standard Gaussian distribution in d dimensions and EM is used to fit a general multi-component mixture model with different number of mixtures. (We note that fitting by one mixture model is simply a Gaussian fit.) Given the estimates for the mixture weights, location and scale parameters returned by EM, we compute the first order Wasserstein distance[1] between the true and estimated parameters. Results for $d \in \{1, 2, 4\}$ and for various amount of over-specification are plotted in Figure 1. From these results, we notice that the decay in statistical error is $n^{-1/2}$ when the fitted number of components is well-specified and equal to the true number of components but has a much slower rate whenever the number of fitted components is two or more. Moreover, in Section 4 (see Figure 3) we show that such a phenomenon occurs more generally in mixture models.

While a rigorous theoretical analysis of EM under over-specification in general mixture models is desirable, it remains beyond the scope of this paper. Instead, here we provide a full characterization of EM when it is used to fit the following class of models to the data drawn from standard Gaussian $\mathcal{N}(0, I_d)$:

$$\mathcal{G}_{\text{symm}}((\theta, \sigma^2)) = \frac{1}{2}\mathcal{N}(\theta, \sigma^2 I_d) + \frac{1}{2}\mathcal{N}(-\theta, \sigma^2 I_d). \quad (1)$$

In particular, in this symmetric fit, we fix the mixture weights to be equal to $\frac{1}{2}$ and require that the two components have same scale parameter. Given the estimates $\widehat{\theta}, \widehat{\sigma}$, the Wasserstein error (see equation (58) in Appendix D) in this case can be simplified as $\|\widehat{\theta}\|_2 + \sqrt{d}\sqrt{|\widehat{\sigma}^2 - 1|}$. In our results to be stated later, we show that the two terms are of the same order (equations (6), (21)) and hence we primarily focus on the error $\|\widehat{\theta} - \theta_\star\|_2$ going forward to simplify the exposition. We consider our set-up as a simple yet first step towards understanding the behavior of EM in over-specified mixtures when *both* location and scale parameter are unknown. In our prior work [Dwivedi et al., 2020], we studied the slow down of EM with over-specified mixtures for estimating only the location parameter, but they assumed that the scale parameter was known and fixed. Here a more general setting is considered.

We now elaborate the choice of our class of models (1) that may appear a bit restrictive at first glance. This model turns out to be the simplest example of a weakly identifiable model in $d = 1$. Let ϕ denote the density of a Gaussian distribution with mean θ and variance σ^2, then we have

$$\frac{\partial^2 \phi}{\partial \theta^2}(x; \theta, \sigma^2) = 2\frac{\partial \phi}{\partial \sigma^2}(x; \theta, \sigma^2), \quad (2)$$

valid for all $x \in \mathbb{R}$, $\theta \in \mathbb{R}$ and $\sigma > 0$. As alluded to earlier, models with algebraic dependence between partial derivatives lead to weak identifiability and slow statistical estimation with MLE. However, in the multivariate setting when the same parameter σ is shared across multiple dimensions, this algebraic relation does not hold and the model is strongly identifiable (since the Fisher information matrix is singular at $(\theta^*, \sigma^*) := (0, 1)$). For this reason, we believe that analysis of EM for the special fit (1) may provide important insight for more general over-specified weakly identifiable models.

Population EM: Given n samples from a d-dimensional standard Gaussian distribution, the sample EM algorithm for location and scale parameters generates a sequence of the form $\theta^{t+1} = M_{n,d}(\theta^t)$ and σ^{t+1}, which is some function of $\|\theta^{t+1}\|_2^2$; see equation (3c) for a precise definition. An abstract counterpart of the sample EM algorithm—not useful in practice but rather for theoretical understanding—is the population EM algorithm $\overline{M}_d$, obtained in the limit of an infinite

[1] First-order Wasserstein distance has been used in prior works to characterize the error between the estimated and true parameters. See section 1.1 [Ho and Nguyen, 2016b].

sample size (cf. equation (11b)).

In practice, running the sample EM algorithm yields an estimate $\widehat{\theta}_{n,d}$ of the unknown location parameter θ^*. Panel (a) in Figure 2 shows the scaling of the statistical estimation error $\|\widehat{\theta}_{n,d} - \theta^*\|_2$ of this sample EM estimate versus the sample size n on a log-log scale. The three curves correspond to dimensions $d \in \{1, 2, 16\}$, along with least-squares fits (on the log-log scale) to the data. In panel (b), we plot the Euclidean norm $\|\theta^t\|_2$ of the population EM iterate[2] versus the iteration number t, with solid red line corresponding to $d = 1$ and the dash-dotted green line corresponding to $d = 2$. Observe that the algorithm converges far more slowly in the univariate case than the multivariate case. The theory to follow in this paper (see Theorems 1, 2 and Lemmas 1 and 3) provides explicit predictions for the rate at which different quantities plotted in Figure 2 should decay. We now summarize our theoretical results that are also consistent with the trends observed in Figure 2.

1.3 Our contributions

The main contribution of this paper is to provide a precise analytical characterization of the behavior of the EM algorithm for certain special cases of over-specified mixture models (1).

Univariate over-specified Gaussian mixtures: In the univariate setting ($d = 1$) of $\mathcal{G}_{\text{symm}}$ in (1), we prove that the EM estimate has statistical estimation error of the order $n^{-\frac{1}{8}}$ and $n^{-\frac{1}{4}}$ after order $n^{\frac{3}{4}}$ steps for the location and scale parameters respectively. In particular, Theorem 1 provides a theoretical justification for the slow rate observed in Figure 2 (a) for $d = 1$ (red dotted line with star marks). Proving these rates requires a novel analysis, and herein lies the main technical contribution of our paper. Indeed, we show that all the analysis techniques introduced in past work on EM, including work on both the regular [Balakrishnan et al., 2017] and strongly identifiable cases [Dwivedi et al., 2020], lead to sub-optimal rates. Our novel method is a *two-stage approach* that makes use of two different population level EM operators. Moreover, we also prove a matching lower bound (see Appendix B) which ensures that the upper bound of order $n^{-\frac{1}{8}}$ for the statistical error of sample EM from Theorem 1 is tight up to constant factors.

Multivariate setting with shared covariance: Given the technical challenges even in the simple univariate case, the symmetric spherical fit $\mathcal{G}_{\text{symm}}$ in (1) serves as a special case for the multivariate setting $d \geq 2$. In this case, we establish that the sharing of scale parameter proves beneficial in the convergence of EM. Theorem 2 shows that sample EM algorithm takes $\mathcal{O}((n/d)^{1/2})$ steps in order to converge to estimates, of the location and scale parameters respectively, that lie within distances $\mathcal{O}(d/n)^{1/4}$ and $\mathcal{O}(nd)^{-\frac{1}{2}}$ of the true location and scale parameters, respectively.

General multivariate setting: We want to remind the readers that we expect the Wasserstein error to scale much slowly than $n^{-\frac{1}{4}}$ (the rate mentioned in the previous paragraph) while estimating over-specified mixtures with no shared covariance. When the fitted variance parameters are not shared across dimensions our simulations under general multi-component fits in Figure 1 demonstrate a much slower convergence of EM (for which a rigorous justification is beyond the scope of this paper).

Notation: In the paper, the expressions $a_n \precsim b_n$ or $a_n \leq O(b_n)$ will be used to denote $a_n \leq cb_n$ for some positive universal constant c that does not change with n. Additionally, we write $a_n \asymp b_n$ if both $a_n \precsim b_n$ and $b_n \precsim a_n$ hold. Furthermore, we denote $[n]$ as the set $\{1, \ldots, n\}$ for any $n \geq 1$. We define $\lceil x \rceil$ as the smallest integer greater than or equal to x for any $x \in \mathbb{R}$. The notation $\|x\|_2$ stands for the ℓ_2 norm of vector $x \in \mathbb{R}^d$. We use c, c', c_1 etc. to denote some universal constants independent of problem parameters (which might change in value each time they appear).

1.4 EM updates for symmetric fit $\mathcal{G}_{\text{symm}}$

The EM updates for Gaussian mixture models are standard, so we simply state them here. In terms of the shorthand notation $\eta := (\theta, \sigma)$, the E-step in the EM algorithm involves computing the function

$$Q_n(\eta'; \eta) := \frac{1}{n} \sum_{i=1}^{n} \big[w_{\theta,\sigma}(X_i) \log \left(\phi(X_i; \theta', (\sigma')^2 I_d) \right) + (1 - w_{\theta,\sigma}(X_i)) \log \left(\phi(X_i; -\theta', (\sigma')^2 I_d) \right) \big],$$

where the weight function is given by $w_{\theta,\sigma}(x) = (1 + e^{\frac{-2\theta^\top x}{\sigma^2}})^{-1}$. The M-step involves maximizing the Q_n-function over the pair (θ', σ') with η fixed, which yields

$$\theta' = \frac{1}{n} \sum_{i=1}^{n} (2 w_{\theta,\sigma}(X_i) - 1) X_i, \quad \text{and} \tag{3a}$$

$$(\sigma')^2 = \frac{1}{d} \left(\frac{\sum_{i=1}^{n} \|X_i\|_2^2}{n} - \|\theta'\|_2^2 \right), \tag{3b}$$

[2] In fact, our analysis makes use of two slightly different population-level operators $\widetilde{M}_{n,d}$ and $\overline{M}_d$ defined in equations (22) and (11b) respectively. Figure 2(b) shows plots for the operator $\overline{M}_d$, but the results are qualitatively similar for the operator $\widetilde{M}_{n,d}$.

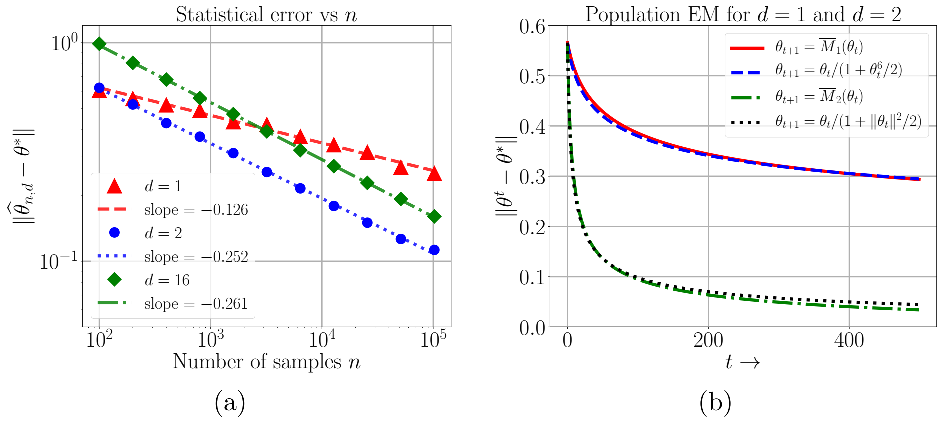


Figure 2. Behavior of the EM algorithm for the fitted model (1), where the data is being generated from $\mathcal{N}(0, I_d)$. (a) Scaling of the Euclidean error $\|\widehat{\theta}_{n,d} - \theta^*\|_2$ with respect to the sample size n for dimension $d \in \{1, 2, 16\}$. Here, $\widehat{\theta}_{n,d}$ denotes the EM algorithm estimate of the mean parameter θ based on n samples. Note that the simulations indicate two distinct error scaling for $d = 1$ and $d > 1$. (b) Convergence behavior of the population-like EM sequence $\theta^{t+1} = \overline{M}_d(\theta^t)$ (11b) in dimensions $d = 1$ and 2. The rate of convergence in dimension $d = 1$ is significantly slower compared to the rate in dimension $d = 2$. Overall, both the plots provide strong empirical evidence towards two distinct behaviors of the EM algorithm for dimension $d = 1$ and dimensions $d > 1$. See the Theorems 1-2, and Lemmas 1 and 3 for a theoretical justification of trends in panels (a) and (b) respectively.

Doing some straightforward algebra, the EM updates (θ_n^t, σ_n^t) can be succinctly defined as

$$\theta_n^{t+1} = \frac{1}{n} \sum_{i=1}^{n} \tanh\left(\frac{X_i^\top \theta_n^t}{\sum_{i=1}^n \|X_i\|_2^2/(nd) - \|\theta_n^t\|_2^2/d} \right) \\ =: M_{n,d}(\theta_n^t), \tag{3c}$$

and $\sigma_n^{t+1} = \sum_{i=1}^n \|X_i\|_2^2/(nd) - \|\theta_n^{t+1}\|_2^2/d$. For simplicity in presentation, we refer to the operator $M_{n,d}$ as the *sample EM operator.*

Organization: We present our main results in Section 2, with Section 2.1 devoted to the univariate case, Section 2.2 to the multivariate case and Section 2.3 to the simulations with more general mixtures. Our proof ideas are summarized in Section 3 and we conclude with a discussion in Section 4. The detailed proofs of all our results are deferred to the Appendices.

2 Main results

In this section, we provide our main results for the behavior of EM with the singular (symmetric) mixtures fit $\mathcal{G}_{\text{symm}}$ (1). Theorem 1 discusses the result for the univariate case, Theorem 2 discusses the result for multivariate case. In Section 2.3 we discuss some simulated experiments for general multivariate location-scale Gaussian mixtures.

2.1 Results for the univariate case

As discussed before, due to the relationship between the location and scale parameter, namely the updates (3c), it suffices to analyze the sample EM operator for the location parameter. For the univariate Gaussian mixtures, given n samples $\{X_i, i \in [n]\}$, the sample EM operator is given by

$$M_{n,1}(\theta) := \frac{1}{n} \sum_{i=1}^{n} X_i \tanh\left[\frac{X_i \theta}{\sum_{j=1}^n X_j^2/n - \theta^2} \right]. \tag{4}$$

We now state our first main result that characterizes the guarantees for EM under the univariate setting. Let I'_β denote the interval $[cn^{-\frac{1}{12}+\beta}, 1/10]$ where c is a positive universal constant.

Theorem 1. *Fix $\delta \in (0, 1)$, $\beta \in (0, 1/8]$, and let $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ for $i = 1, \ldots, n$ such that $n \gtrsim \log \frac{\log(1/\beta)}{\delta}$. Then for any initialization θ_n^0 that satisfies $|\theta_n^0| \in I'_\beta$, the sample EM sequence $\theta_n^{t+1} = M_{n,1}(\theta_n^t)$, satisfies*

$$|\theta_n^t - \theta^*| \leq c_1 \frac{1}{n^{1/8-\beta}} \log^{5/4}\left(\frac{10n \log(8/\beta)}{\delta} \right), \tag{5}$$

for all $t \geq c_2 n^{\frac{3}{4}-6\beta} \cdot \log n \log \frac{1}{\beta}$ with probability at least $1 - \delta$.

See Appendix A.1 for the proof.

The bound (5) shows that with high probability after $\mathcal{O}(n^{3/4})$ steps the sample EM iterates converge to a ball around θ^* whose radius is arbitrarily close to $n^{-1/8}$. Moreover, as a direct consequence of the relation (3b), we conclude that the EM estimate for the scale parameter is of order $n^{-\frac{1}{4}}$ with high probability:

$$\left|(\sigma_n^t)^2 - (\sigma^*)^2\right| = \left| \frac{\sum_{i=1}^n X_i^{\,2}}{n} - \left(\theta_n^t - \theta^*\right)^2 - (\sigma^*)^2 \right| \\ \precsim n^{-\frac{1}{2}} + n^{-\frac{1}{4}} = O(n^{-\frac{1}{4}}) \tag{6}$$

where we have used the standard chi-squared concentration for the sum $\sum_{i=1}^n X_i^{\,2}/n$.

Matching lower bound: In Appendix B, we prove a matching lower bound and thereby conclude that the upper bound of order $n^{-\frac{1}{8}}$ for the statistical error of sample EM from Theorem 1 is tight up to constant factors. In Section 2.3, we provide further evidence (cf. Figure 3) that the slow statistical rates of EM with location parameter that we derived in Theorem 1 might appear in more general settings of location-scale Gaussian mixtures as well.

2.2 Results for the multivariate case

Analyzing the EM updates for higher dimensions turns out to be challenging. However, for the symmetric fit in higher dimensions given by

$$\mathcal{G}_{\text{symm}}((\theta,\sigma^2)) = \frac{1}{2}\mathcal{N}(\theta,\sigma^2 I_d) + \frac{1}{2}\mathcal{N}(-\theta,\sigma^2 I_d), \quad (7)$$

the sample EM operator $M_{n,d}(\theta)$ has a closed form as already noted in the updates (3b) and (3c). Note that for the fit (7), we have assumed the same scale parameter for all dimensions. Such a fit is over-specified for data drawn from Gaussian distribution $\mathcal{N}(0, I_d)$. We now show that the sharing of scale parameter in the model fit across dimensions (7), leads to a faster convergence of EM in $d \geq 2$—both in terms of number of steps and the final statistical accuracy. In the following result, we denote $I_\beta := [5\left(\frac{d}{n}\right)^{\frac{1}{4}+\beta}, \frac{1}{8}]$.

Theorem 2. *Fix $\delta \in (0,1)$, $\beta \in (0, 1/4]$, and let $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$ for $i = 1, \ldots, n$ such that $d \geq 2$ and $n \gtrsim d \log^{\frac{1}{4\beta}}(\log \frac{1/\beta}{\delta})$. Then with any starting point θ_n^0 such that $\|\theta_n^0\|_2 \in I_\beta$, the sample EM sequence $\theta_n^{t+1} = M_{n,d}(\theta_n^t)$ satisfies*

$$\|\theta_n^t - \theta^*\|_2 \leq c_1 \left(\frac{d}{n}\log\frac{\log(1/\beta)}{\delta}\right)^{\frac{1}{4}-\beta}, \quad (8)$$

for all $t \geq c_2\left(\frac{n}{d}\right)^{\frac{1}{2}-2\beta}\log\frac{n}{d}\log\frac{1}{\beta}$ with probability at least $1-\delta$.

See Appendix A.2 for the proof.

The results in Theorem 2 show that the that the sample EM updates converge to a ball around $\theta^* = 0$ with radius arbitrarily close to $(d/n)^{\frac{1}{4}}$ when $d \geq 2$. At first sight, the initialization condition $\|\theta_n^0\|_2 \leq 1/8$, assumed in Theorem 2, might seem pretty restrictive but Lemma 6 (in Appendix C.6) shows that for any θ_n^0 satisfying $\|\theta_n^0\|_2 \leq \sqrt{d}$, we have $\widetilde{M}_{n,d}(\theta_n^0) \leq \sqrt{2/\pi}$, with high probability. In light of this result, we may conclude that the initialization condition is Theorem 2 is not overly restrictive.

Comparison with Theorem 1: The scaling of order $n^{-\frac{1}{4}}$ with n is significantly better than the univariate case $(n^{-\frac{1}{8}})$ stated in Theorem 1. We note that this faster statistical rate is a consequence of the sharing of the scale parameter across dimensions, and does not hold when the fit (7) has different variance parameters. Indeed, as we demonstrated in Figure 1, when the fitted components have freely varying scale parameter, the statistical rate slows down (and can be of the order $n^{-\frac{1}{8}}$ in higher dimensions).

2.3 Simulations with general cases

We now present preliminary evidence that the slow statistical rates of EM with location parameter that we derived in Theorem 1 might appear in more general settings. In Figure 3, we plot the statistical error of estimates returned by sample EM when estimating *all* the parameters (namely weights, location and scale) simultaneously, as a function of sample size n, for the following two cases:

$$\mathcal{G}_\star^{d=1} = \frac{1}{6}\mathcal{N}(-5,1) + \frac{1}{2}\mathcal{N}(1,3) + \frac{1}{3}\mathcal{N}(7,2); \quad (9)$$

$$\mathcal{G}_\star^{d=2} = \frac{1}{2}\mathcal{N}(\begin{bmatrix}0\\0\end{bmatrix}, I) + \frac{1}{6}\mathcal{N}(\begin{bmatrix}7\\5\end{bmatrix}, 2I)\frac{1}{3}\mathcal{N}(\begin{bmatrix}-4\\-7\end{bmatrix}, 3I). \quad (10)$$

We plot the results for a $K_{\text{fit}} \in \{3, 4, 5\}$-mixture Gaussian model fit. When K_{fit} is equal to the number of components $(= 3)$ in the true mixture the statistical rate is $n^{-1/2}$. When it is larger, i.e., $K_{\text{fit}} \in \{4, 5\}$, the statistical rate of EM is much larger, $n^{-0.12}$ in panel (a) (for $K_{\text{fit}} = 5$) and $n^{-0.20}$ in panel (b) (for $K_{\text{fit}} = 5$) of Figure 3. These simulations suggest that the statistical rates slower than $n^{-\frac{1}{4}}$ and of order $n^{-\frac{1}{8}}$ may arise in more general settings, and moreover that the rates get slower as the over-specification of the number of mixtures increases. See Section 4 for possible future work in this direction.

3 Analysis of EM

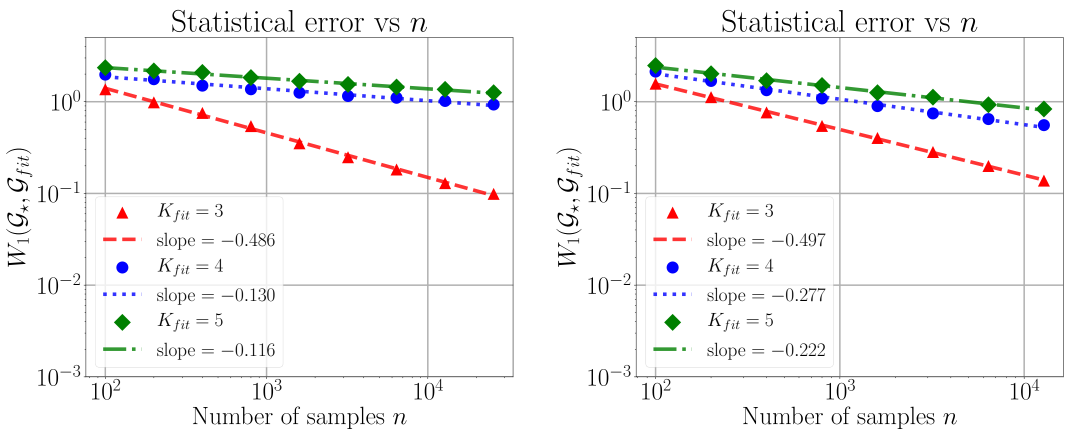
We now provide a road-map for the technical analysis of EM. Deriving a sharp rate for univariate case (Theorem 1) turns out to

Our proof makes use of the population-to-sample analysis framework of Balakrishnan et al. [Balakrishnan et al., 2017] albeit with several new ideas. Let $Y \sim \mathcal{N}(0, 1)$, then the population-level analog of the operator (3c) can be defined in two ways:

$$\widetilde{M}_{n,1}(\theta) := \mathbb{E}_Y\left[Y \tanh\left(\frac{Y\theta}{\sum_{j=1}^n X_j^2/n - \theta^2}\right)\right], \quad (11a)$$

$$\overline{M}_1(\theta) := \mathbb{E}_Y\left[Y \tanh\left(\frac{Y\theta}{1-\theta}\right)\right]. \quad (11b)$$

The particular choice of the population-like operator $\widetilde{M}_{n,1}$ in equation (11a) was motivated by the previous



(a) $d = 1$ with $\mathcal{G}_*$ given by equation (9) (a) $d = 2$ with $\mathcal{G}_*$ given by equation (10)

Figure 3. Scaling of the first-order Wasserstein error for EM estimates when fitting a Gaussian mixture with $K_{\text{fit}} \in \{3, 4, 5\}$, i.e., $\mathcal{G}_{\text{fit}} = \sum_{i=1}^{K_{\text{fit}}} w_i \mathcal{N}(\mu_i, \Sigma_i)$, on n i.i.d. samples from a 3-Gaussian mixture model (equations (9) and (10)). In the case of no over-specification, i.e., $K_{\text{fit}} = K_{\text{true}} = 3$, the error scales as $n^{-1/2}$, but when the fitted model is over-specified ($K_{\text{fit}} \in \{4, 5\}$), the scaling is much worse (and degrades further for any given n as K_{fit} gets large). See Section 2.3 for further details.

works [Cai et al., 2019] with the location-scale Gaussian mixtures. We refer to this operator as the *pseudo-population operator* since it depends on the samples $\{X_i, i = 1, \ldots n\}$ and involves an expectation. Nonetheless, as we show in the sequel, analyzing $\widetilde{M}_{n,1}$ is not enough to derive sharp rates for sample EM in the over-specified setting considered in Theorem 1. A careful inspection reveals that a "better" choice of the population operator is required, which leads us to define the operator $\overline{M}_1$ in equation (11b). Unlike the pseudo-population operator $\widetilde{M}_{n,1}$, the operator $\overline{M}_1$ is indeed a population operator as it does not depend on samples $X_1, \ldots, X_n$. Note that, this operator is obtained when we replace the sum $\sum_{j=1}^{n} {X_j}^2/n$ in the definition (11a) of the operator $\widetilde{M}_{n,1}$ by its corresponding expectation $\mathbb{E}[\|X\|_2^2] = 1$. For this reason, we also refer to this operator $\overline{M}_1$ as the *corrected population operator*. In the next lemma, we state the properties of the operators defined above (here I'_β denotes the interval $[cn^{-\frac{1}{12}+\beta}, 1/10]$).

Lemma 1. *The operators $\widetilde{M}_{n,1}$ and $\overline{M}_1$ satisfy*

$$\left(1 - \frac{3\theta^6}{2}\right)|\theta| \leq \left|\widetilde{M}_{n,1}(\theta)\right| \leq \left(1 - \frac{\theta^6}{5}\right)|\theta|, \tag{12a}$$

$$\left(1 - \frac{\theta^6}{2}\right)|\theta| \leq \left|\overline{M}_1(\theta)\right| \leq \left(1 - \frac{\theta^6}{5}\right)|\theta|, \tag{12b}$$

where bound (12a) holds for all $|\theta| \in I'_\beta$ with high probability[3] *and the bound (12b) is deterministic and holds for all $|\theta| \in \left[0, \frac{3}{20}\right]$. Furthermore, for any fixed $\delta \in (0, 1)$ and any fixed $r \geq O(n^{-\frac{1}{12}})$, we have that*

$$\mathbb{P}\left[\sup_{\theta \in \mathbb{B}(0,r)} \left|M_{n,1}(\theta) - \widetilde{M}_{n,1}\right| \leq cr\sqrt{\frac{\log(1/\delta)}{n}}\right] \geq 1 - \delta. \tag{12c}$$

On the other hand, for any fixed $r \leq O(n^{-\frac{1}{16}})$, we have

$$\mathbb{P}\left[\sup_{\theta \in \mathbb{B}(0,r)} \left|M_{n,1}(\theta) - \overline{M}_1(\theta)\right| \leq c_2 r^3 \sqrt{\frac{\log^{10}(5n/\delta)}{n}}\right] \geq 1 - \delta. \tag{12d}$$

See Appendix A.3 for its proof where we also numerically verify the sharpness of the results above (see Figure 4). Lemma 1 establishes that, as $\theta \to 0$, both the operators have similar contraction coefficient $\gamma(\theta) \asymp 1 - c\theta^6$; thereby justifying the rates observed for $d = 1$ in Figure 2(b). However, their perturbation bounds are significantly different: while the error $\sup_{\theta \in \mathbb{B}(0,r)} \left|M_{n,1}(\theta) - \widetilde{M}_{n,1}(\theta)\right|$ scales linearly with the radius r, the deviation error $\sup_{\theta \in \mathbb{B}(0,r)} \left|M_{n,1}(\theta) - \overline{M}_1(\theta)\right|$ has a cubic scaling r^3.

Remark: A notable difference between the two bounds (12c) and (12d) is the range of radius r over which we *prove* the validity of the bounds (12c) and (12d). With our tools, we establish that the perturbation bound (12c) for the operator $\widetilde{M}_{n,1}$ is valid for any $r \gtrsim n^{-\frac{1}{12}}$. On the other hand, the corresponding bound (12d) for the operator $\overline{M}_1$ is valid for any $r \precsim n^{-\frac{1}{16}}$. We now elaborate why these different ranges of radii are helpful and make both the operators crucial to in the analysis to follow.

[3] Since the operator $\widetilde{M}_{n,1}$ depends on the samples $\{X_j, j \in [n]\}$, only a high probability bound (and not a deterministic one) is possible.

3.1 A sub-optimal analysis

Using the properties of the operator $\widetilde{M}_{n,1}$ from Lemma 1, we now sketch the statistical rates for the sample EM sequence, $\theta_n^{t+1} = M_{n,1}(\theta_n^t)$, that can be obtained using (a) the generic procedure outlined by Balakrishnan et al. [Balakrishnan et al., 2017] and (b) the localization argument introduced in our previous work [Dwivedi et al., 2020]. As we show, both these arguments end up being *sub-optimal* as they do not provide us the rate of order $n^{-\frac{1}{8}}$ stated in Theorem 1. We use the notation:

$$\sup_{|\theta|\geq\epsilon} \left|\widetilde{M}_{n,1}(\theta)\right| / |\theta| \precsim 1 - \epsilon^6 =: \gamma(\epsilon).$$

Sub-optimal rate I: The eventual radius of convergence obtained using Theorem 5(a) from the paper [Balakrishnan et al., 2017] can be determined by

$$\frac{r/\sqrt{n}}{1-\gamma(\epsilon)} = \epsilon \quad \Longrightarrow \quad \epsilon \sim n^{-1/14}, \tag{13a}$$

where r denotes the bound on the initialization radius $|\theta^0|$ but we have tracked dependency only on n. This informal computation suggests that the the sample EM iterates for location parameter are bounded by a term of order $n^{-1/14}$. This rate is clearly sub-optimal when compared to the EM rate of order $n^{-\frac{1}{8}}$ from Theorem 1.

Sub-optimal rate II: Next we apply the more sophisticated localization argument from the paper [Dwivedi et al., 2020] in order to obtain a sharper rate. In contrast to the computation (13a), this argument leads to solving the equation

$$\frac{\epsilon \cdot r/\sqrt{n}}{1-\gamma(\epsilon)} = \epsilon \implies \frac{\epsilon r/\sqrt{n}}{\epsilon^6} = \epsilon \implies \epsilon \sim n^{-\frac{1}{12}}, \tag{13b}$$

where, as before, we have only tracked dependency on n. This calculation allows us to conclude that the EM algorithm converges to an estimate which is at a distance of order $n^{-\frac{1}{12}}$ from the true parameter, which is again sub-optimal compared to the $n^{-\frac{1}{8}}$ rate of EM from Theorem 1.

Indeed both the conclusions above can be made rigorous (See Corollary 1 for a formal statement) to conclude that, with high probability for any $\beta \in (0, \frac{1}{12}]$

$$\left|\theta_n^t - \theta^*\right| \leq O(n^{-\frac{1}{12}+\beta}) \text{ for } t \geq O(n^{\frac{1}{2}-6\beta}). \tag{14}$$

3.2 A two-staged analysis for sharp rates

In lieu of the above observations, the proof of the sharp upper bound (5) in Theorem 1 proceeds in two stages. In the first stage, invoking Corollary 1 with $\beta = \frac{1}{48}$, we conclude that with high probability the sample EM iterates converge to a ball of radius at most r after $\sqrt{n}$ steps, where $r \ll n^{-1/16}$. Consequently, the sample EM iterates after $\sqrt{n}$ steps satisfy the assumptions required to invoke the perturbation bounds for the operator $\overline{M}_1$ from Lemma 1. Thereby, in the second stage of the proof, we apply the $1 - c\theta^6$ contraction bound (12b) of the operator $\overline{M}_1$ in conjunction with the cubic perturbation bound (12d). Using localization argument for this stage, we establish that the EM iterates obtain a statistical error of order $n^{-1/8}$ in $\mathcal{O}\left(n^{3/4}\right)$ steps as stated in Theorem 1. See Appendix A.1 for a detailed proof.

4 Discussion

In this paper, we established several results characterizing the convergence behavior of EM algorithm for over-specified location-scale Gaussian mixtures. We view our analysis of EM for the symmetric singular Gaussian mixtures as the first step toward a rigorous understanding of EM for a broader class of weakly identifiable mixture models. Such a study would provide a better understanding of the singular models with weak identifiability which do arise in practice since: (a) over-specification is a common phenomenon in fitting mixture models due to weak separation between mixture components, and, (b) the parameters being estimated are often inherently dependent due to the algebraic structures of the class of kernel densities being fitted and the associated partial derivatives. We now discuss a few other directions that can serve as a natural follow-up of our work.

The slow rate of order $n^{-\frac{1}{8}}$ for EM updates with location parameter is in a sense a worst-case guarantee. In the univariate case, for the entire class of two mixture Gaussian fits, MLE exhibits the slowest known statistical rate $n^{-\frac{1}{8}}$ for the settings that we analyzed. More precisely, for certain asymmetric Gaussian mixture fits, the MLE convergence rate for the location parameter is faster than that of the symmetric equal-weighted mixture considered in this paper E.g., for the fit $1/3\mathcal{N}(-2\theta, \sigma^2) + 2/3\mathcal{N}(\theta, \sigma^2)$ on $\mathcal{N}(0, 1)$ data, the MLE converges at the rate $n^{-1/6}$ and $n^{-1/3}$ respectively [Ho and Nguyen, 2016b]. It is interesting to understand the effect of such a geometric structure of the global maxima on the convergence of the EM algorithm.

Our work analyzed over-specified mixtures with a specific structure and only one extra component. As demonstrated above, the statistical rates for EM appear to be slow for general covariance fits and further appear to slow down as the number of over-specified components increases. The convergence rate of the MLE for such over-specified models is known to further deteriorate as a function of the number of extra components. It remains to understand how the EM algorithm responds to these more severe—and practically relevant—instances of over-specification.

Acknowledgments

This work was partially supported by Office of Naval Research grant DOD ONR-N00014-18-1-2640 and National Science Foundation grant NSF-DMS-1612948 to MJW, and National Science Foundation grant NSF-DMS-1613002 to BY, and by Army Research Office grant W911NF-17-1-0304 to MIJ.

References

[Balakrishnan et al., 2017] Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 45:77–120. (Cited on pages 2, 4, 6, 8, and 15.)

[Cai et al., 2019] Cai, T. T., Ma, J., and Zhang, L. (2019). CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality. *Annals of Statistics*. (Cited on pages 2 and 7.)

[Chen et al., 2001] Chen, H., Chen, J., and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63:19–29. (Cited on page 1.)

[Chen, 1995] Chen, J. (1995). Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1):221–233. (Cited on pages 1 and 2.)

[Chen and Li, 2009] Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *Annals of Statistics*, 37:2523–2542. (Cited on pages 1 and 2.)

[Chrétien and Hero, 2008] Chrétien, S. and Hero, A. O. (2008). On EM algorithms and their proximal generalizations. *ESAIM: Probability and Statistics*, 12:308–326. (Cited on page 2.)

[Daskalakis et al., 2017] 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*. (Cited on page 2.)

[Dempster et al., 1997] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1997). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39:1–38. (Cited on page 2.)

[Dwivedi et al., 2020] Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M. J., Jordan, M. I., and Yu, B. (2020). Singularity, misspecification, and the convergence rate of EM. *Annals of Statistics, to appear.* (Cited on pages 2, 3, 4, 8, 15, and 18.)

[Hao et al., 2018] Hao, B., Sun, W., Liu, Y., and Cheng, G. (2018). Simultaneous clustering and estimation of heterogeneous graphical models. *Journal of Machine Learning Research*. (Cited on page 2.)

[Havre et al., 2015] Havre, Z. V., White, N., Rousseau, J., and Mengersen, K. (2015). Overfitting Bayesian mixture models with an unknown number of components. *PLOS One*, 10. (Cited on page 1.)

[Heinrich and Kahn, 2018] Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *Annals of Statistics*, 46:2844–2870. (Cited on page 1.)

[Ho and Nguyen, 2016a] Ho, N. and Nguyen, X. (2016a). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755. (Cited on page 1.)

[Ho and Nguyen, 2016b] Ho, N. and Nguyen, X. (2016b). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*. (Cited on pages 3, 8, and 30.)

[Ishwaran et al., 2001] Ishwaran, H., James, L. F., and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96:1316–1332. (Cited on page 1.)

[Jordan and Xu, 1995] Jordan, M. I. and Xu, L. (1995). Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, 8. (Cited on page 2.)

[Ledoux and Talagrand, 1991] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY. (Cited on page 18.)

[Ma et al., 2000] Ma, J., Xu, L., and Jordan, M. I. (2000). Asymptotic convergence rate of the EM algorithm for Gaussian mixtures. *Neural Computation*, 12:2881–2907. (Cited on page 2.)

[Nguyen, 2013] Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 4(1):370–400. (Cited on page 1.)

[Pearson, 1894] Pearson, K. (1894). Contributions to the theory of mathematical evolution. *Philosophical Transactions of the Royal Society of London A*, 185:71–110. (Cited on page 1.)

[Redner and Walker, 1984] Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239. (Cited on page 2.)

[Rousseau and Mengersen, 2011] Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:689–710. (Cited on page 1.)

[Tseng, 2004] Tseng, P. (2004). An analysis of the EM algorithm and entropy-like proximal point methods. *Mathematics of Operations Research*, 29(1):27–44. (Cited on page 2.)

[van der Vaart and Wellner, 2000] van der Vaart, A. W. and Wellner, J. A. (2000). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York, NY. (Cited on page 18.)

[Villani, 2008] Villani, C. (2008). *Optimal transport: Old and New*. Springer. (Cited on page 30.)

[Wang et al., 2015] Wang, Z., Gu, Q., Ning, Y., and Liu, H. (2015). High-dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. In *Advances in Neural Information Processing Systems 28*. (Cited on page 2.)

[Wu, 1983] Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103. (Cited on page 2.)

[Xu et al., 2016] 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*. (Cited on page 2.)

[Xu and Jordan, 1996] Xu, L. and Jordan, M. I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8:129–151. (Cited on page 2.)

[Yan et al., 2017] 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*. (Cited on page 2.)

[Yi and Caramanis, 2015] Yi, X. and Caramanis, C. (2015). Regularized EM algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems 28*. (Cited on page 2.)

[Yu, 1997] Yu, B. (1997). Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, pages 423–435. (Cited on page 22.)