Theoretical guarantees for the EM algorithm when applied to mis-specified Gaussian mixture models

Raaz Dwivedi[⋆] **Nhat Ho**[⋆] **Koulik Khamaru**[⋆]
UC Berkeley
{raaz.rsk, minhnhat, koulik}@berkeley.edu

Martin J. Wainwright
UC Berkeley
Voleon Group
wainwrig@berkeley.edu

Michael I. Jordan
UC Berkeley
jordan@berkeley.edu

Abstract

Recent years have witnessed substantial progress in understanding the behavior of EM for mixture models that are correctly specified. Given that model mis-specification is common in practice, it is important to understand EM in this more general setting. We provide non-asymptotic guarantees for the population and sample-based EM algorithms when used to estimate parameters of certain mis-specified Gaussian mixture models. Due to mis-specification, the EM iterates no longer converge to the true model and instead converge to the projection of the true model onto the fitted model class. We provide two classes of theoretical guarantees: (a) a characterization of the bias introduced due to the mis-specification; and (b) guarantees of geometric convergence of the population EM to the model projection given a suitable initialization. This geometric convergence rate for population EM implies that the EM algorithm based on n samples converges to an estimate with $1/\sqrt{n}$ accuracy. We validate our theoretical findings in different cases via several numerical examples.

1 Introduction

Mixture models play a central role in statistical applications, where they are used to capture heterogeneity of data arising from several underlying subpopulations. However, estimating the parameters of mixture models is a challenging task, due to the non-convexity of the log likelihood function. As shown by classical work, the maximum likelihood estimate (MLE) often has good properties for mixture models, but its computation can be non-trivial. One of the most popular algorithms used to compute the MLE (approximately) is the expectation maximization (EM) algorithm. Although EM is widely used in practice, it does not always converge to the MLE, and its convergence rate can vary as a function of the problem. Classical results provide guarantees about the convergence rates of EM to local maxima [4, 16]. In the specific setting of Gaussian mixtures, population EM (idealized EM with infinite samples) was shown to have a range of behavior from super-linear convergence to slow convergence like a first-order method depending on the overlap between the mixtures [9, 18]. More recently, there has been a renewed interest in providing explicit and non-asymptotic guarantees on the convergence of EM. Notably, Balakrishnan et al. [1] developed a rather general framework for characterizing the convergence of EM. For well-specified problems—including the two-component Gaussian location mixture as a particular example—they provided sufficient conditions for the EM

[⋆] Raaz Dwivedi, Nhat Ho, and Koulik Khamaru contributed equally to this work.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

algorithm to converge to a small neighborhood of global maximum; in addition, they provided explicit bounds on the sample complexity of EM, meaning the number of samples n required, as a function of the tolerance ϵ, problem dimension and other parameters, to achieve an ϵ-accurate solution. A line of follow-up work has generalized and extended results of this type (e.g., see the papers [20, 15, 7, 17, 3, 19, 5, 2]).

A shared assumption common to this body of past work is that either the true distribution of each subpopulations is known, or that the number of components is exactly known; in practice, both of these conditions are often violated. In such settings, it is well known that the MLE, instead of approximating the true parameter, approximates a Kullback-Leibler projection of the data-generating distribution onto the fitted model class. Thus, the MLE exhibits a desirable form of robustness to model mis-specification.

On the other hand, it is not obvious *a priori* that this robustness need be shared by the solutions returned by the EM algorithm. Since these solutions are those actually used in practice, it is important understand under what conditions the EM algorithm, when applied with mis-specified models, converges to an (approximate) KL projection. The main contribution of this paper is to provide some precise answers to this question, and moreover to quantify the bias that arises from model mis-specification. Our analysis focuses on two classes of mis-specified mixture models.

- *Under-specified number of components*: Suppose that the true model is given by location-shifted mixture of $k \geq 3$ univariate Gaussians, but we use EM to fit a location-shifted Gaussian mixture with $k-1$ components. This scenario is very common: it arises naturally when either the mixture components are very close or some of the mixture weights are very small, so that the data generating distribution appears to have fewer components. Analysis of the EM algorithm when the fitting distribution has fewer mixture components than the data-generating distribution poses new challenges; in particular, it requires an understanding of the *model bias*, meaning the Kullback-Leibler discrepancy between the true model from its projection(s) onto the class of fitted models. In this paper, we provide a detailed analysis of the $k = 3$ case. First, we characterize the model bias induced by fitting a two-component mixture to a three-component mixture with unknown means but known variance. We then provide sufficient conditions for the population EM updates to converge at a geometric rate to the KL projection of the true model onto the fitted model class. Finally, using Rademacher-complexity based arguments and the geometric convergence of population EM, we conclude that with high probability, the EM updates with n samples converge to a ball of radius $1/\sqrt{n}$ around the aforementioned KL projection.
- *Incorrectly specified weights or variances*: In our second problem class, we assume that the number of components is correctly specified, but either the mixture weights or the variances are mis-specified. Concretely, suppose that the true model is a two-component location-shifted Gaussian mixture with weights/variances that differ from those in the fitted model class. Our analysis reveals a rather surprising phenomenon with respect to EM convergence: despite the potential non-convexity of the problem, the iterates converge at a geometric rate to a unique fixed point from an arbitrary initialization. Our results suggest that the projection from the true model to the fitted model is actually unique. Finally, we prove that the sample-based EM updates achieve standard minimax convergence rate of order $1/\sqrt{n}$.

Table 1 provides a high-level summary of our results, where we use (θ, σ, α) to denote the Gaussian mixture component with mean θ, variance σ^2 and weight α, i.e., $\alpha\mathcal{N}(\theta, \sigma^2)$.

The remainder of our paper is organized as follows. In Section 2, we introduce the problem set-up and provide the background information on the EM algorithm. In Section 3, we present our results for the first framework and provide expressions for the bias and rate of convergence of EM for different 3 component mixture of Gaussians. Section 4 contains results when the mixture weights and variance are mis-specified. Numerical experiments illustrating our theoretical results are presented in Section 5. Finally in Section 6, we conclude the paper with a discussion of our results and a few possible venues for future work.

Notation: We use c, c', c_1, c_2 to denote universal constants whose value may vary in different contexts. For two distributions $\mathbb{P}$ and $\mathbb{Q}$, the Kullback-Leibler divergence between them is denoted by $\text{KL}(\mathbb{P}, \mathbb{Q})$. We use the standard big-O notation to depict the scaling with respect to a particular quantity and hide constants and other problem parameters.

True Model	**Best fit with two components**	**Bias** $\min\{\lvert\overline{\theta}-\theta^*\rvert, \lvert\overline{\theta}+\theta^*\rvert\}$	**Statistical error** $\lvert\widehat{\theta}_n-\overline{\theta}\rvert$ **of sample EM**
3-component mixture: $(-\theta^*(1{+}\rho), \sigma, 1/4)$; $(-\theta^*(1{-}\rho), \sigma, 1/4)$; $(\theta^*, \sigma, 1/2)$	$(-\overline{\theta}, \sigma, 1/2)$; $(\overline{\theta}, \sigma, 1/2)$ σ known	$\rho\,\lvert\theta^*\rvert + c\,(\rho\,\lvert\theta^*\rvert/\sigma)^{1/4}$	$n^{-1/2}$
3-component mixture: $(-\theta^*, \sigma, (1{-}\omega)/2)$; $(\theta^*, \sigma, (1{-}\omega)/2)$; $(0, \sigma^2, \omega)$	$(-\overline{\theta}, \sigma, 1/2)$; $(\overline{\theta}, \sigma, 1/2)$; σ known	$\dfrac{c\omega^{1/8}\,\lvert\theta^*\rvert^{1/4}}{\sqrt{1{-}\omega}\sigma^{1/4}}$	$n^{-1/2}$
2-component mixture: $(-\theta^*, \sqrt{\sigma^2{-}\theta^{*2}}, 1/2)$; $(\theta^*, \sqrt{\sigma^2{-}\theta^{*2}}, 1/2)$;	$(-\overline{\theta}, \sigma, \pi)$; $(\overline{\theta}, \sigma, 1{-}\pi)$ $\sigma, \pi{\neq}1/2$ known	$\dfrac{c\,\lvert\theta^*\rvert\,((2{-}4\pi)+\theta^{*2})^{1/2}}{\sigma}$	$n^{-1/2}$

Table 1. Summary of the main theoretical gurantees of this paper. Here the parameter θ^* denotes the true parameter value (in the data-generating distribution), $\overline{\theta}$ denotes the value of the parameter of the best fit model, and $\hat{\theta}_n$ denotes the estimate returned by running the EM algorithm. Recall that the true model is not in the class of fitted models, and we can only hope to estimate $\overline{\theta}$; consequently, in the above table lists the performance of the EM algorithm in estimating $\overline{\theta}$ for different settings. The first column lists the true model, while the second column shows the fitted model. In the third column, we summarize the bias of the parameter of the best fitted model (2). When using EM with n samples, the final statistical error $|\widehat{\theta}_n - \overline{\theta}|$ has the statistical rate of order $n^{-1/2}$ in all cases, as depicted in the fourth column (here $\widehat{\theta}_n$ denotes the final sample EM estimate).

2 Problem set-up

Throughout this paper, we assume that data is generated according to some true distribution $\mathbb{P}_*$, which admits a continuous density over $\mathbb{R}$. We are interested in the performance of the EM algorithm when we fit the model below using a two-component mixture of location-shifted Gaussians with known variance σ^2 and known mixture weight $\pi \in (0, 1)$:

$$\mathbb{P}_\theta = \pi\mathcal{N}(\theta, \sigma^2) + (1-\pi)\mathcal{N}(-\theta, \sigma^2) \tag{1}$$

We consider two distinct settings of the mixture weights in model (1):

- **Balanced mixtures**: the mixture weights are assumed to be equal, i.e., $\pi = 1 - \pi = 1/2$.
- **Unbalanced mixtures**: the mixture weights are assumed to be unequal $\pi = \frac{1}{2}(1-\epsilon)$ and $1 - \pi = \frac{1}{2}(1+\epsilon)$ where $|\epsilon| \in (0, 1)$.

In order to estimate the location parameters, we apply the EM algorithm, allowing θ to vary over some compact set Θ. Since the true distribution $\mathbb{P}_*$ may not belong to the class of fitted models, the best possible estimator is the projections of $\mathbb{P}_*$ to the fitted model (1). It is given by

$$\overline{\theta} \in \arg\min_{\theta\in\Theta} \text{KL}\left(\mathbb{P}_*, \mathbb{P}_\theta\right). \tag{2}$$

Our main goal in the paper is to establish the convergence rate of EM updates to $\overline{\theta}$ for various choices of the data-generating model $\mathbb{P}_*$ and the fitted model (1).

2.1 EM algorithm for two-component location-Gaussian mixtures

Let us now introduce some notation as well as a brief description of the EM algorithm for two-component Gaussian location mixtures (1). The population version of EM is based on the function

$$Q(\theta'; \theta) := -\frac{1}{2}\mathbb{E}\left[w_\theta(X)\left(X-\theta'\right)^2 + \left(1-w_\theta(X)\right)\left(X+\theta'\right)^2\right], \tag{3}$$

where the expectation is taken over the true distribution $\mathbb{P}_*$. For any fixed θ, the M-step in the EM updates for the model (1) is obtained by maximizing the minorization function (3); for a detailed derivation, see the paper [1]. More precisely, we denote the population EM operator $M : \mathbb{R} \to \mathbb{R}$ as

$$M(\theta) := \arg\max_{\theta'} Q(\theta', \theta) = \mathbb{E}\left[(2w_\theta(X) - 1)X\right], \tag{4a}$$

where the weighting function w_θ in the above formulation is given by

$$w_\theta(x) := \frac{\pi \exp\left(-\frac{(\theta - x)^2}{2\sigma^2}\right)}{\pi \exp\left(-\frac{(\theta - x)^2}{2\sigma^2}\right) + (1 - \pi) \exp\left(-\frac{(\theta + x)^2}{2\sigma^2}\right)}. \tag{4b}$$

Note that the parameter $\overline{\theta}$, defined in equation (2), minimizes the KL-distance between the fitted model and the true model, thereby ensuring that the log-likelihood is maximized at the model indexed by the parameter $\overline{\theta}$. Consequently, the parameter $\overline{\theta}$ is a fixed point of the population EM update—that is, $M(\overline{\theta}) = \overline{\theta}$. The sample version of the EM algorithm—the method actually used in practice—is obtained by simply replacing the expectations in equations (3) and equation (4a) by the sample-based counterpart. In particular, given a set of n i.i.d. samples $\{X_i\}_{i=1}^n$ from the true model, the sample EM operator $M_n : \mathbb{R} \mapsto \mathbb{R}$ takes the form

$$M_n(\theta) := \frac{1}{n} \sum_{i=1}^n (2w_\theta(X_i) - 1)X_i. \tag{5}$$

With this notation in place, we are now ready to state our main results.

3 Guarantees for EM algorithm for mis-specified number of components

In this section, we study the convergence of the EM algorithm in the setting of under-fitted mixtures, where the number of components in the true model is larger than that in the fitted model. In sharp contrast to the traditional setting of correctly specified mixture models, where the number of components of the true model is known to the EM algorithm, we analyze the performance of the EM algorithm in the setting where the true number of the components is not known. Such a scenario naturally occur in many practical cases, examples include: (1) Some components in the mixture are very close, and it is hard to distinguish them; (2) Some components have very small mixture weights and thereby are difficult to detect. Consequently, in the aforementioned situations, the number of components observed from the data may be much smaller compared to the number of components present in the true model. In this section, we characterize the bias of the two-component fit and analyze the convergence properties of EM for such a fit.

3.1 Three-component mixtures with two close components

First, we consider the case, where the true model has distribution $\mathbb{P}_*$ is a mixture of three-component Gaussian location mixture given by

$$\mathbb{P}_* = \frac{1}{4}\mathcal{N}(-\theta^*(1+\rho), \sigma^2) + \frac{1}{4}\mathcal{N}(-\theta^*(1-\rho), \sigma^2) + \frac{1}{2}\mathcal{N}(\theta^*, \sigma^2) \tag{6}$$

for some θ^* in a compact subset Θ of the real line, and a small positive scalar ρ that characterizes the separation between two cluster means $-\theta^*(1+\rho)$ and $-\theta^*(1-\rho)$. For fitting the model, we assume that the variance σ^2 is known, and we suspect that the true model is a two-component mixture (since ρ is small). Consequently, we fit the data with the model

$$\mathbb{P}_\theta = \frac{1}{2}\mathcal{N}(-\theta, \sigma^2) + \frac{1}{2}\mathcal{N}(\theta, \sigma^2), \tag{7}$$

and we use the EM algorithm to estimate the location parameter θ. Clearly, the performance of model (7) and consequently the EM algorithm depends on the relationship between the separation factor ρ and the SNR $\eta := |\theta^*| / \sigma$ of the true model (6). Since the true model does not belong in the family of two components location-Gaussian mixtures in model class (7), the role of the projection parameter $\overline{\theta} \in \arg\min_{\theta\in\Theta} \mathrm{KL}(\mathbb{P}_*, \mathbb{P}_\theta)$ becomes crucial. In the next proposition, we provide an explicit bound for the bias between $\overline{\theta}$ and θ^* as a function of the problem parameters.

Proposition 1. *Given the true model* (6) *and any $\rho > 0$, we have*

$$\min\left\{\left|\theta^* - \overline{\theta}\right|, \left|\theta^* + \overline{\theta}\right|\right\} \leq \rho\left|\theta^*\right| + c\left(\frac{\rho\left|\theta^*\right|}{\sigma}\right)^{1/4}, \tag{8}$$

where c is a universal positive constant that depends only on the set Θ.

In order to simplify our results in the sequel, we assume that $\eta = \left|\theta^*\right|/\sigma \geq 1$ and use a simpler bound on the bias—viz.:

$$\min\left\{\left|\theta^* - \overline{\theta}\right|, \left|\theta^* + \overline{\theta}\right|\right\} \leq \left(\rho + \frac{\rho^{1/4}}{\eta^{3/4}\sigma}\right)\left|\theta^*\right| \leq \left(\rho + \frac{\rho^{1/4}}{\sigma}\right)\left|\theta^*\right|. \tag{9}$$

The bound above directly implies that $\left|\overline{\theta}\right|$ belongs to the interval $[(1 - C_\rho)\left|\theta^*\right|, (1 + C_\rho)\left|\theta^*\right|]$, assuming that $C_\rho := \rho + \rho^{1/4}/\sigma \leq 1$. As $\rho \to 0$, we have $C_\rho \to 0$ implying that $\left|\theta^*\right|$ and $\left|\overline{\theta}\right|$ are almost identical. In the sequel, we utilize this precise control of $\left|\overline{\theta}\right|$ in terms of $\left|\theta^*\right|$, provided by Proposition 1, to analyze the behavior of the EM algorithm in a neighborhood of $\overline{\theta}$. Defining $\rho_\star := \sup\left\{\rho > 0 | C_\rho \leq 1/9\right\}$, the following result characterizes the behavior population EM operator for the three-component Gaussian location mixture described by equation (6).

Theorem 1. *There exist universal constants c', c'' such that the population EM operator for model* (6) *with $\rho \leq \rho_\star$ and $\eta \geq c'$ satisfies*

$$\left|M(\theta) - \overline{\theta}\right| = \mathbb{E}\left|2(w_\theta(X) - w_{\overline{\theta}}(X))X\right| \leq \gamma\left|\theta - \overline{\theta}\right|, \quad \textit{for any } \theta \in \mathbb{B}(\overline{\theta}, \left|\overline{\theta}\right|/4).$$

In words, Theorem 1 establishes that the population EM iterates (in the ideal, infinite data limit) are γ-contractive with respect to $\overline{\theta}$ over the ball $\mathbb{B}(\overline{\theta}, \left|\overline{\theta}\right|/4)$, where $\gamma \leq e^{-c''\eta^2}$. Combining that result with the condition $C_\rho \leq 1/9$, we can demonstrate that $\left|\overline{\theta}\right|$ is unique (See Section A.1.4 in the Appendix). These results have a direct implication for the *sample-based version* of EM that is implemented in practice. In particular, the next result shows that EM updates with n samples converge in a constant number of steps to a neighborhood of $\overline{\theta}$.

Corollary 1. *Consider any scalar $\delta \in (0, 1)$, sample size $n \geq c_1 \log(1/\delta)$ and starting point $\theta^0 \in \mathbb{B}(\overline{\theta}, \left|\overline{\theta}\right|/4)$. Then under the assumptions of Theorem 1, the sample-based EM sequence $\theta^{t+1} = M_n\left(\theta^t\right)$ for the model* (6) *satisfies*

$$\left|\theta^t - \overline{\theta}\right| \leq \gamma^t\left|\theta^0 - \overline{\theta}\right| + \frac{c_2}{1 - \gamma}\left|\theta^*\right|\left(\theta^{*2} + \sigma^2\right)\sqrt{\frac{\log(1/\delta)}{n}} \tag{10}$$

with probability at least $1 - \delta$, where $\gamma \leq e^{-c'\eta^2}$.

Note that the bound (10) consists of two main terms: the first term captures the geometric convergence of the population EM operator from Theorem 1, while the second term characterizes the radius of convergence in terms of sample complexity, which is $\mathcal{O}(\sqrt{1/n})$. Therefore, with probability at least $1 - \delta$, we have

$$\left|\theta^T - \overline{\theta}\right| \leq \frac{c\left|\theta^*\right|(\theta^{*2} + \sigma^2)}{1 - \gamma}\sqrt{\frac{\log(1/\delta)}{n}} \quad \text{for} \quad T \geq c'\frac{\log(n/(\log(1/\delta)\left|\theta^*\right|(\theta^{*2} + \sigma^2)))}{\log(1/\gamma)},$$

where c, c' are universal constants.

3.2 Three-component mixtures with small weight for one component

Next, we consider the case where the true model $\mathbb{P}_*$ is a three-component Gaussian location mixture model of the form

$$\mathbb{P}_* = \frac{1 - \omega}{2}\mathcal{N}(-\theta^*, \sigma^2) + \omega\mathcal{N}(0, \sigma^2) + \frac{1 - \omega}{2}\mathcal{N}(\theta^*, \sigma^2). \tag{11}$$

In other words, two components are dominant with means $-\theta^*$ and θ^* respectively, and we have a small component at the origin. For such a model, it is again conceivable to fit a 2-component mixture given by equation (7). The primary interest in such a setting is driven by the fact that, when $\omega > 0$ is sufficiently small, recovering the third small component with center at origin is usually hard; consequently clustering that component with one of the other two may be a good idea. Once again, the convergence of EM is governed by the properties of $\overline{\theta}$ that we characterize in the next proposition.

Proposition 2. *For the three components location-Gaussian mixtures in model* (11)*, we have*

$$\min\left\{\left|\theta^* - \overline{\theta}\right|, \left|\theta^* + \overline{\theta}\right|\right\} \leq \frac{c\omega^{1/8}\left|\theta^*\right|^{1/4}}{\sigma^{1/4}\sqrt{1-\omega}}, \tag{12}$$

where c is a universal positive constant that depends only on the set Θ.

In order to simplify further results, we assume under the condition $\eta := \frac{\theta^*}{\sigma} \geq 1$. Then we have $\min\left\{\left|\theta^* - \overline{\theta}\right|, \left|\theta^* + \overline{\theta}\right|\right\} \leq C_\omega \left|\theta^*\right|$, where $C_\omega := c\omega^{1/8}/(\sigma\sqrt{1-\omega})$. Such a bound on bias leads to slightly different conditions for convergence of the EM algorithm for model (11) compared to the EM convergence for model (6). Note that for any fixed variance σ^2, the function C_ω increases with ω and $C_0 = 0$. Let $\omega_\star = \sup\{\omega > 0 | C(\omega) \leq 1/9\}$. Similar to the model (6), we analyze the convergence rate of EM under a strong SNR condition of true model (11). We define $\tilde{\gamma} := \gamma(\eta, \omega) = (1-\omega)e^{-\eta^2/64} + \omega < 1$. With the above notations in place, we now establish the contraction of the population EM operator $M(\theta)$ for the three components location-Gaussian mixture (11).

Theorem 2. *For SNR $\eta \geq 1$ sufficiently large and $\omega \leq \omega_\star$, and for any $\theta^0 \in \mathbb{B}(\overline{\theta}, \left|\overline{\theta}\right|/4)$, the population EM operator for the Gaussian mixture* (11) *satisfies*

$$|M(\theta^0) - \overline{\theta}| = \mathbb{E}\left|2(w_{\theta^0}(X) - w_{\overline{\theta}}(X))X\right| \leq \tilde{\gamma}\left|\theta^0 - \overline{\theta}\right|. \tag{13}$$

Consequently, the population EM sequence $\theta^{t+1} = M(\theta^t)$ converges to $\overline{\theta}$ at a linear rate.

The precise expression for the contraction parameter $\tilde{\gamma}$ provides sufficient conditions for a fast convergence of EM, which involves an interesting trade off between the SNR η and weight ω. More concretely, if the SNR is large enough, the population EM converges fast towards the projection $\overline{\theta}$, which is unique in its absolute value (See Section A.1.4 in the Appendix). This fast convergence of the population EM again enables us to derive the following convergence rate of sample-based EM:

Corollary 2. *Consider the model* (11) *such that the assumptions of Theorem 2 hold. For any fixed $\delta \in (0,1)$, $\theta^0 \in \mathbb{B}(\overline{\theta}, \left|\overline{\theta}\right|/4)$, if $n \geq c_1 \log(1/\delta)$ then the sample EM iterates $\theta^{t+1} = M_n(\theta^t)$ satisfy*

$$\left|\theta^t - \overline{\theta}\right| \leq \tilde{\gamma}^t\left|\theta^0 - \overline{\theta}\right| + \frac{c_2}{1-\tilde{\gamma}}\left|\theta^*\right|\left(\theta^{*2} + \sigma^2\right)\sqrt{\frac{\log(1/\delta)}{n}}$$

with probability at least $1-\delta$.

Similar to the structure of the convergence result of sample EM updates in Corollary 1, the result in Corollary 2 also consists of two key terms: the first term is the linear rate of convergence from the population EM operator in Theorem 2 while the second term characterizes the radius of convergence in terms of sample complexity, which is of $\mathcal{O}(\sqrt{1/n})$ after $T = \mathcal{O}(\log n / \log(1/\tilde{\gamma}))$ iterations.

4 Robustness of EM for mis-specified variances and weights

In this section, we focus on establishing the convergence rate of EM under different mis-specified regime of the fitted model (1). In particular, we assume that the true data distribution $\mathbb{P}_*$ is given by:

$$\mathbb{P}_* = \frac{1}{2}\mathcal{N}(\theta^*, \sigma^2 - \theta^{*2}) + \frac{1}{2}\mathcal{N}(-\theta^*, \sigma^2 - \theta^{*2}), \tag{14}$$

where $\sigma > 0$ is a given positive number, and $|\theta^*| \in (0, \sigma/2)$ is a true but unknown parameter. Note that the assumption that $|\theta^*| \in (0, \sigma/2)$ ensures that the variance $\sigma^2 - \theta^{*2}$ is bounded away from zero. We fit the above model by unbalanced two-component Gaussian location mixture model $\mathbb{P}_\theta$ given by

$$\mathbb{P}_\theta = \pi\mathcal{N}(-\theta, \sigma^2) + (1-\pi)\mathcal{N}(\theta, \sigma^2), \tag{15}$$

where $\pi := \frac{1}{2}(1-\epsilon)$ and $|\epsilon| \in (0,1)$ are known apriori and only the parameter θ is to be estimated. In the fitted model $\mathbb{P}_\theta$, we have mis-specified the variance σ^2 and the weight π, and we wish to understand the rate of convergence of EM to $\overline{\theta}$, where $\overline{\theta}$ is the parameter of the model $\mathbb{P}_{\bar{\theta}}$, and $\mathbb{P}_{\bar{\theta}}$ is the projection of the true model $\mathbb{P}_*$ onto the model class $\mathcal{P}_\theta := \{\mathbb{P}_\theta \ : \ \theta \in \mathbb{R}\}$. We emphasize that the main goal here is to see how the mis-specification with variance and weight affects the statistical inference of EM. We choose variance of the form $\sigma^2 - \theta^{*2}$ because under this setting, we obtain interesting behavior of EM without rendering the proof too technical. We begin with the first result establishing the global linear convergence rate of population EM to $\overline{\theta}$.

Theorem 3. *For a two-component Gaussian location mixture model* (14) *and fitted model* (15)*, the population EM operator $\theta \mapsto M(\theta)$ satisfies*

$$\left|M(\theta) - \overline{\theta}\right| \leq \left(1 - \frac{\epsilon^2}{2}\right) \left|\theta - \bar{\theta}\right|.$$

Hence, the population EM sequence $\{\theta^t\}$ converges geometrically to $\overline{\theta}$ from any initialization θ^0.

There are two interesting features regarding the geometric convergence of population EM updates to $\overline{\theta}$: (1) it does not require an evaluation of bias which was needed for our previous results; (2) it holds under any initialization θ^0. Overall, we have that $\overline{\theta}$ is unique, thereby we conclude that the projection of $\mathbb{P}_*$ to the model class (15) is unique. Before proceeding to the sample-based convergence of EM, we establish the following upper bound on the bias of the parameter $\overline{\theta}$:

Proposition 3. *For the two-component Gaussian location mixture model* (14)*, we have*

$$\min\left\{\left|\overline{\theta} - \theta^*\right|, \left|\overline{\theta} + \theta^*\right|\right\} \leq c(\theta^*, \sigma) \cdot \sqrt{\left[2\left(1 - 2\pi\right)\right]{\theta^*}^2 + {\theta^*}^4},$$

where $c(\theta^*, \sigma)$ is a positive constant depending only on θ^*, σ, and the set Θ.

Given the above bound, we obtain the range of $\left|\overline{\theta}\right|$ as $\left|\overline{\theta}\right| \in \left[(1 - C_{\theta^*})\left|\left|\theta^*\right|, (1 + C_{\theta^*})\left|\theta^*\right|\right]\right.$ where $C_{\theta^*} := c(\theta^*, \sigma)\sqrt{\left[2(1 - 2\pi)\right] + {\theta^*}^2}$. Equipped with this bound on $\left|\overline{\theta}\right|$, we have the following result regarding the convergence of sample-based EM:

Corollary 3. *Consider the model* (14)*. Let radius $r > 0$ and $n \geq c_1 \log(1/\delta)$ and $\theta^0 \in \mathbb{B}\left(\overline{\theta}, r\right)$, then the sample-based EM sequence $\theta^{t+1} = M_n(\theta^t)$, satisfies*

$$\left|\theta^t - \overline{\theta}\right| \leq \left(1 - \frac{\epsilon^2}{2}\right)^t \left|\theta^0 - \overline{\theta}\right| + \frac{c_2\left(\left(1 + C_{\theta^*}\right)\left|\theta^*\right| + r\right)\sigma^2}{\epsilon^2}\sqrt{\frac{\log(1/\delta)}{n}},$$

with probability at least $1 - \delta$ where $\epsilon := 1 - 2\pi$.

The proof of Corollary 3 is similar to those of Corollary 1 or Corollary 2; therefore, it is omitted. The last corollary demonstrates that the sample-based EM iterates converge to ball of radius $\mathcal{O}(\sqrt{1/n})$ around $\overline{\theta}$ after $T = \mathcal{O}(\log n/\log(1/(1 - \epsilon^2/2)))$ iterations.

5 Simulation studies

In this section, we illustrate our theoretical results using a few numerical experiments. In particular, we use the EM algorithm to fit 2-component Gaussian mixtures for the three mis-specified settings considered above. For convenience in discussion, we refer to three settings as follows:

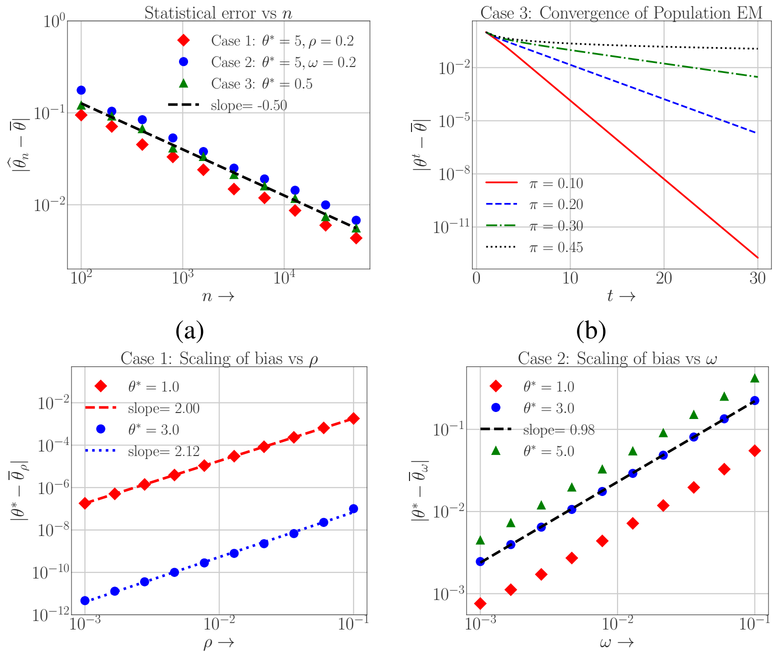
- Case 1 refers to the true model (6) from Section 3.1, namely a three component Gaussian mixture where two of the components very close to each other and the quantity $\rho \in (0, 1)$ denotes the extent of weak separation.
- Case 2 refers to the true model (11) from Section 3.2, namely, a three components Gaussian mixture where one of the components has very small weight at origin and the quantity $\omega \in (0, 1)$ denotes the small mixture-weight.
- Finally, Case 3 refers to the true model (14) from Section 4, namely where the true model is a two-Gaussian mixture.

For cases 1 and 2, we fit a symmetric balanced two-Gaussian mixture given by equation (7); while for the third case we fit the unbalanced two-Gaussian mixture given by equation (15) for different values of π. Let $\widehat{\theta}_n$ denote the final sample EM estimate. Since our results establish that population EM converges to $\overline{\theta}$ (2), we use the final iterate from the population EM sequence to estimate the error $|\widehat{\theta}_n - \overline{\theta}|$. We now summarize our key findings:

(i) In Figure 1(a), we observe that for all cases the final statistical error $|\widehat{\theta}_n - \overline{\theta}|$ has a parametric rate $n^{-1/2}$ which verifies the claims of Corollaries 1, 2 and 3.

(ii) For all cases, the population EM sequence has a geometric convergence (we omit illustrations for Cases 1 and 2). From Figure 1 (b), we note that for Case 3, the linear convergence of the population EM sequence $\theta^{t+1} = M(\theta^t)$ is affected by the extent of unbalancedness: as $\pi \to 0.5$, the rate of decay of the error of population EM sequence decreases which is consistent with the contraction result stated in Theorem 3.

(iii) In panels (c) and (d) of Figure 1, we plot the biases for Case 1 and 2, with respect to ρ and ω respectively. Least squares fit on the log-log scale suggest that the biases stated in Proposition 1 and Proposition 2 are potentially sub-optimal: the numerical scaling of the biases $|\theta^* - \overline{\theta}|$ is of the order ρ^2 and ω for Case 1 and 2 respectively, which is significantly smaller than the corresponding scaling of the order $\rho^{1/4}$ and $\omega^{1/8}$ stated in Propositions 1 and 2. In Appendix B, we illustrate the scaling of the bias with θ^* in these cases via further simulations.



(a)

(b)

(c)

(d)

Figure 1. Plots depicting behavior of EM when fitting two Gaussian mixture (7) for the three mis-specified mixtures cases (6), (11) and (14), referred to as Case 1, 2 and 3 respectively. (a) For all cases, the statistical error $|\widehat{\theta}_n - \overline{\theta}|$ has the parametric rate $n^{-1/2}$. (b) For Case 3, convergence of population EM sequence $\theta^{t+1} = M(\theta^t)$ is affected by the mixture weight π. The convergence rate slows down as $\pi \to 0.5$. (c) For Case 1, the bias scales quadratically with the extent of weak-separation ρ for different values of θ^*. (d) For Case 2, the bias scales linearly with the weight ω of the third component, for different values of θ^*. Refer to the text for more details.

6 Discussion

In this paper, we analyzed the behavior of the EM algorithm for certain classes of mis-specified mixture models. Analyzing the behavior of the EM algoirithm under general mis-specification is challenging in general, and we view the results in this paper as a first step towards developing a more general framework for the problem. In this paper, we studied the EM algorithm when it is used to fit Gaussian location mixture models to data generated by mixture models with larger numbers of components, and/or differing mixture weights. We considered only univariate mixtures in this paper, but we believe that several of our results can be extended to multivariate mixtures. It is also interesting to investigate the behavior of the EM algorithm when it is used to fit models with scale parameters that vary (in addition to the location parameters). Besides deriving sharper results for the settings considered in this paper, analyzing the behavior of EM for non-Gaussian and more general mixture models is an appealing avenue for future work.

References

[1] S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 45:77–120, 2017.

[2] T. T. Cai, J. Ma, and L. Zhang. CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality. *Annals of Statistics*, To Appear.

[3] C. Daskalakis, C. Tzamos, and M. Zampetakis. Ten steps of EM suffice for mixtures of two Gaussians. In *Proceedings of the 2017 Conference on Learning Theory*, 2017.

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

[5] B. Hao, W. Sun, Y. Liu, and G. Cheng. Simultaneous clustering and estimation of heterogeneous graphical models. *Journal of Machine Learning Research*, To Appear.

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

[7] C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan. Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems 29*, 2016.

[8] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

[9] J. Ma, L. Xu, and M. I. Jordan. Asymptotic convergence rate of the EM algorithm for Gaussian mixtures. *Neural Computation*, 12:2881–2907, 2000.

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

[11] H. Teicher. Identifiability of finite mixtures. *Ann. Math. Statist.*, 32:1265–1269, 1963.

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

[13] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027v7*.

[14] C. Villani. *Optimal transport: Old and New*. Springer, 2008.

[15] Z. Wang, Q. Gu, Y. Ning, and H. Liu. High-dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. In *Advances in Neural Information Processing Systems 28*, 2015.

[16] C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.

[17] J. Xu, D. Hsu, and A. Maleki. Global analysis of expectation maximization for mixtures of two Gaussians. In *Advances in Neural Information Processing Systems 29*, 2016.

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

[19] B. Yan, M. Yin, and P. Sarkar. Convergence of gradient EM on multi-component mixture of Gaussians. In *Advances in Neural Information Processing Systems 30*, 2017.

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

A Supplementary material

In this appendix, we provide self-contained proofs for our results in the paper. In particular, Section A.1 contains the proofs of our theorems, Section A.2 contains the proof of our propositions and in Section A.3 we prove the corollaries stated in the paper.

A.1 Proofs for population EM

In this section, we prove our main results on the contraction properties of the population EM algorithm toward the projection onto the model class—namely, Theorems 1, 2 and 3. We treat each of these theorems one-by-one.

A.1.1 Proof of Theorem 1

The proof of the theorem makes use of Proposition 1 that relates $\left|\overline{\theta}\right|$ in terms of ρ, $|\theta^*|$, and $\eta = |\theta^*|/\sigma$. Without loss of generality, we assume that $\min\left\{\left|\theta^* - \overline{\theta}\right|, \left|\theta^* + \overline{\theta}\right|\right\} = \left|\theta^* - \overline{\theta}\right|$. For each $u \in [0,1]$, we define $\theta_u = \overline{\theta} + u(\theta - \overline{\theta})$. Applying Taylor's theorem along the direction θ_u, we obtain that

$$\left|\mathbb{E}\bigg(2(w_\theta(X) - w_{\overline{\theta}}(X))X\bigg)\right| = 4\Big|\int_0^1 \mathbb{E}\left[\frac{X^2}{\sigma^2(e^{-\theta_u X/\sigma^2} + e^{\theta_u X/\sigma^2})^2}\right](\theta - \overline{\theta})du\Big|$$
$$\leq 4 \sup_{u\in[0,1]} \left\{\mathbb{E}\left[\Gamma_u(X)\right]\right\}\left|\theta - \overline{\theta}\right|, \tag{16}$$

where we have defined

$$\Gamma_u(X) := X^2/(\sigma^2(e^{-\theta_u X/\sigma^2} + e^{\theta_u X/\sigma^2})^2), \tag{17}$$

and the expectation is taken over X which is drawn from the Gaussian mixture given by equation (6). Clearly, we have

$$\mathbb{E}\left[\Gamma_u(X)\right] = \frac{1}{4}\mathbb{E}_{X\sim\mathcal{N}(-\theta^*(1+\rho),\sigma)}\Gamma_u(X) + \frac{1}{4}\mathbb{E}_{X\sim\mathcal{N}(-\theta^*(1-\rho),\sigma)}\Gamma_u(X) + \frac{1}{2}\mathbb{E}_{X\sim\mathcal{N}(\theta^*,\sigma)}\Gamma_u(X).$$

We now bound the three expectations on the right, which we denote by T_1, T_2 and T_3 respectively. We claim that

$$T_1, T_2 \leq e^{-\eta^2/64}/16 \quad \text{and} \quad T_3 \leq e^{-\eta^2/64}/8,$$

which in turn implies that $\mathbb{E}\left[\Gamma_u(X)\right] \leq \gamma = e^{-\eta^2/64}/4$ and our theorem follows.

We now provide a full derivation for an upper bound on T_1. The upper bounds on T_2 and T_3 can be derived in a similar way and their explicit derivation is omitted here. Letting $R = \text{sign}(\theta_u)$ and $V = -RX/\sigma$, we have

$$D_u := 4T_1 = \mathbb{E}_{X\sim\mathcal{N}(-\theta^*(1+\rho),\sigma)}\Gamma_u(X) = \mathbb{E}[V^2/(e^{-|\theta_u|V/\sigma} + e^{|\theta_u|V/\sigma})^2],$$

where the expectation in the last expression is taken with respect to $V \sim \mathcal{N}(R\theta^*(1+\rho)/\sigma, 1)$. We have

$$D_u \leq \mathbb{E}[V^2 e^{-2|\theta_u|V/\sigma}] \leq \mathbb{E}\left[V^2 e^{-2|\theta_u|V/\sigma}\middle|\mathcal{E}\right]\cdot\mathbb{P}[\mathcal{E}] + \mathbb{E}\left[V^2 e^{-2|\theta_u|V/\sigma}\middle|\mathcal{E}^c\right]\cdot\mathbb{P}[\mathcal{E}^c],$$

where we define the event $\mathcal{E} = \{V | V \leq |\theta^*|(1+\rho)/(4\sigma)\}$. Given a scalar μ, consider the real-valued function f such that $f(t) = t^2 e^{-\mu t}$. Observe that $f(t) \leq \frac{4}{e^2\mu^2}$ for all $t \in \mathbb{R}$ and that f is decreasing on the interval $[2/\mu, \infty)$. Invoking these observations with $\mu = 2|\theta_u|/\sigma$, as long as $|\theta^*|(1+\rho)/(4\sigma) \geq 2/\mu$ or equivalently $|\theta^*|(1+\rho)|\theta_u| \geq 4\sigma^2$, we find that

$$D_u \leq \frac{\sigma^2}{e^2\theta_u^2}\cdot\mathbb{P}[\mathcal{E}] + \frac{{\theta^*}^2(1+\rho)^2}{16\sigma^2}e^{-|\theta^*|(1+\rho)|\theta_u|/(2\sigma^2)}. \tag{18}$$

Note that $\theta \in \mathbb{B}(\overline{\theta}, \left|\overline{\theta}\right|/4)$ implies that $|\theta_u| \geq 3\left|\overline{\theta}\right|/4$ and $\text{sign}(\theta_u) = \text{sign}(\overline{\theta})$ for all $u \in [0,1]$. Proposition 1 implies that $\overline{\theta} \in [(1-C_\rho)\theta^*, (1+C_\rho)\theta^*]$. Since ρ is small enough such that $C_\rho < 1$, we also have that $\text{sign}(\overline{\theta}) = \text{sign}(\theta^*)$. As a result $\mathbb{E}[V] = \text{sign}(\theta_u)\theta^*(1+\rho)/\sigma = |\theta^*|\,(1+\rho)/\sigma$. Invoking standard Gaussian tail bounds, we have

$$\mathbb{P}[\mathcal{E}] = \mathbb{P}\left[V - \mathbb{E}[V] \le -\frac{3}{4}\frac{|\theta^*|\,(1+\rho)}{\sigma}\right] \le \exp\left(-\frac{9{\theta^*}^2(1+\rho)^2}{32\sigma^2}\right).$$

Plugging this bound along with the fact that $|\theta_u| \ge (1-C_\rho)\,|\theta^*|$ in the inequality (18), we find that

$$\begin{aligned}
D_u &\le \frac{16\sigma^2 \exp\left(-\frac{9{\theta^*}^2(1+\rho)^2}{32\sigma^2}\right)}{9e^2{\theta^*}^2(1-C_\rho)^2} + \frac{{\theta^*}^2(1+\rho)^2 \exp\left(-\frac{{\theta^*}^2(1+\rho)(1-C_\rho)}{8\sigma^2}\right)}{16\sigma^2} \\
&\le \frac{16}{9e^2}\left(\frac{\sigma^2}{{\theta^*}^2(1-C_\rho)^2} + \frac{{\theta^*}^2(1+\rho)^2}{\sigma^2}\right)\exp\left(-\frac{{\theta^*}^2(1+\rho)(1-C_\rho)}{8\sigma^2}\right) \\
&\le (\eta^2 + \eta^{-2})\exp(-\eta/16) \quad (\text{since } 64 \le 9e^2) \\
&\le 2\eta^2\exp(-\eta^2/16) \qquad (\text{for } \eta \ge 1) \\
&\le \exp(-\eta^2/64)/16 \qquad (\text{for } \eta \ge 14),
\end{aligned}$$

where we have used the fact that $\rho \in (0,1)$ is small enough and that $C_\rho \le 1/9 < 1/2$. The claim follows.

A.1.2 Proof of Theorem 2

Equipped with the bounds for the bias term $\left|\overline{\theta} - \theta^*\right|$ from Proposition 2, the steps in this proof are similar to the ones used in the proof of Theorem 1. Using Taylor expansion along the direction $\theta_u = \overline{\theta} + u(\theta - \overline{\theta})$ for $u \in [0,1]$, we find that

$$\mathbb{E}[2(w_\theta(X) - w_{\overline{\theta}}(X))X] \le 4 \sup_{u\in[0,1]} \mathbb{E}\left[\Gamma_u(X)\right]\left|\theta - \overline{\theta}\right|, \tag{19}$$

where $\Gamma_u(X)$ is the same term defined above in equation (17). The difference compared to the proof of Theorem 1 is in the distribution of X. In particular, now we have

$$\begin{aligned}
\mathbb{E}\left[\Gamma_u(X)\right] &= (1/2 - \omega/2)\mathbb{E}_{X\sim\mathcal{N}(-\theta^*,\sigma)}\Gamma_u(X) + (1/2-\omega/2)\mathbb{E}_{X\sim\mathcal{N}(\theta^*,\sigma)}\Gamma_u(X) \\
&\quad + \omega\mathbb{E}_{X\sim\mathcal{N}(0,\sigma)}\Gamma_u(X) \\
&= (1/2-\omega/2)(S_1+S_2) + \omega S_3.
\end{aligned}$$

Imitating the steps for bounding T_1 in the proof of Theorem 1, we can derive the following bounds for S_1 and S_2:

$$S_1, S_2 \le e^{-\eta^2/64}/4,$$

provided that $C(\eta,\omega) := \dfrac{c(\eta^2\sigma^2\omega)^{1/4}}{\sqrt{(1-\omega)}} \le 1/9 < 1/2$ and η is sufficiently large. Thus it is left to provide a bound for the term S_3. Using the change of variables $V = \text{sign}(\theta_u)X/\sigma$ and the consequent fact that $V \sim \mathcal{N}(0,1)$ we obtain that

$$S_3 = \mathbb{E}_{X\sim\mathcal{N}(0,\sigma)}[\Gamma(X)] = \mathbb{E}\left[\frac{V^2}{(e^{-|\theta_u|V/\sigma} + e^{|\theta_u|V/\sigma})^2}\right] \overset{(i)}{\le} \mathbb{E}\left[\frac{V^2}{4}\right] = \frac{1}{4},$$

where step (i) follows from the inequality that $e^{-y} + e^{y} \ge 2$ for all $y \in \mathbb{R}$. Putting the pieces together yields

$$\mathbb{E}[2(w_\theta(X) - w_{\overline{\theta}}(X))X] \le (1-\omega)e^{-\eta^2/64} + \omega$$

and we are done.

A.1.3 Proof of Theorem 3

Using the definition (4a) of the M-update and the self consistency $M(\bar{\theta}) = \bar{\theta}$, we obtain that

$$\left|M(\theta) - M(\bar{\theta})\right| = \underbrace{\left|\mathbb{E}\left[2(w_\theta(X) - w_{\bar{\theta}}(X))X\right]\right|}_{=:A}.$$

Note that under the unbalanced mixtures, we have

$$w_\theta(X) = \frac{\pi}{\pi + (1-\pi)e^{-2\theta X/\sigma^2}} \quad \text{and} \quad \frac{\partial}{\partial\theta}(w_\theta(X)) = \frac{2\pi(1-\pi)X/\sigma^2}{(\pi e^{-\theta X/\sigma^2} + (1-\pi)e^{\theta X/\sigma^2})^2}.$$

By means of Taylor expansion along the direction $\theta_u = \overline{\theta} + u(\theta - \overline{\theta})$, the following holds

$$\begin{aligned} A &= 4\pi(1-\pi)\left|\int\limits_0^1 \mathbb{E}\left[\frac{X^2}{\sigma^2\left((1-\pi)\exp\left(-\frac{\theta_u X}{\sigma^2}\right) + \pi\exp\left(\frac{\theta_u X}{\sigma^2}\right)\right)^2}\right] du\right| \left|\theta - \overline{\theta}\right| \\ &\leq 4\pi(1-\pi)\left|\theta - \overline{\theta}\right| \max_{u\in[0,1]} \mathbb{E}\left[\Gamma_{\theta_u}(X)\right], \end{aligned} \tag{20}$$

where $\Gamma_{\theta_u}(X) := \dfrac{X^2}{\sigma^2\left(\pi\exp\left(-\frac{\theta_u X}{\sigma^2}\right) + (1-\pi)\exp\left(\frac{\theta_u X}{\sigma^2}\right)\right)}$. Let $\pi = \frac{1}{2}(1-\epsilon)$. We claim that

$$\max_{u\in[0,1]} \mathbb{E}\left[\Gamma_{\theta_u}(X)\right] \leq \frac{1-\epsilon^2/2}{1-\epsilon^2}, \tag{21}$$

which when plugged in the bound (20) implies that the population EM operator is globally contractive towards $\overline{\theta}$, i.e., $\left|M(\theta) - M(\overline{\theta})\right| \leq (1-\epsilon^2/2)\left|\theta - \overline{\theta}\right|$. Therefore, it yields the linear rate of convergence claimed in the theorem.

We now prove the claim (21). Like in proof of Theorem 1, we use $R = \text{sign}(\theta_u)$ and $V = RX/\sigma$. Since $X \sim \frac{1}{2}\mathcal{N}(\theta^*, \sigma^2 - \theta^{*2}) + \frac{1}{2}\mathcal{N}(-\theta^*, \sigma^2 - \theta^{*2})$, it is clear that $\mathbb{E}[V] = 0$ and $\mathbb{E}\left[V^2\right] = 1$. By substituting $X = \sigma V/R$, we have

$$\mathbb{E}\left[\Gamma_{\theta_u}(X)\right] = \mathbb{E}_V\left[\frac{V^2}{(\pi\exp(-\left|\theta_u\right|V/\sigma) + (1-\pi)\exp(\left|\theta_u\right|V/\sigma))^2}\right].$$

Now, observe that

$$\begin{aligned} &(\pi e^{-y} + (1-\pi)e^y) \in [\sqrt{(1-\epsilon^2)},\, 1], \quad && \text{if } e^y \in \left[1, \frac{1+\epsilon}{1-\epsilon}\right], \quad \text{and} \\ &(\pi e^{-y} + (1-\pi)e^y) > 1, && \text{otherwise.} \end{aligned}$$

Let $\mathcal{E}_{\theta_u}$ denote the event such that $\mathcal{E}_{\theta_u} = \left\{e^{\left|\theta_u\right|V/\sigma} \in [1, (1+\epsilon)/(1-\epsilon)]\right\}$. Let $\mathcal{E}^c$ and $\mathbb{I}(\mathcal{E})$ respectively denote the complement and the indicator of any event $\mathcal{E}$. Using the observation above and the fact that $\mathbb{E}\left[V^2\right] = 1$, we obtain that

$$\begin{aligned} \mathbb{E}\left[\Gamma_{\theta_u}(X)\right] &\leq \frac{1}{(1-\epsilon^2)}\mathbb{E}\left[V^2\,\mathbb{I}(\mathcal{E}_{\theta_u})\right] + \mathbb{E}\left[V^2\,\mathbb{I}(\mathcal{E}^c_{\theta_u})\right] \\ &= \frac{1-\epsilon^2 + \epsilon^2\mathbb{E}\left[V^2\,\mathbb{I}(\mathcal{E}_{\theta_u})\right]}{(1-\epsilon^2)}. \end{aligned} \tag{22}$$

Note that whenever $\theta_u \neq 0$, we have that

$$\mathbb{E}\left[V^2\,\mathbb{I}(\mathcal{E}_{\theta_u})\right] \leq \mathbb{E}\left[V^2\,\mathbb{I}(V \geq 0)\right] = \frac{1}{2}. \tag{23}$$

Putting the inequalities (22) and (23) together yields the claim (21).

A.1.4 Proof of uniqueness of $\left|\overline{\theta}\right|$

We provide a proof for the uniqueness of projection in its absolute value from $\mathbb{P}_*$ in (6) or in (11) to the fitted model (7). Due to the similar proof argument between these two cases, we only focus on the case when $\mathbb{P}_*$ is given by (6) and the fitted model is in (7). First, we note that if $\overline{\theta}$ is the projection of $\mathbb{P}_*$ to the fitted model (7), then $-\overline{\theta}$ is also the projection. Therefore, the projection is identifiable in its absolute value. Now, the result of Theorem 1 demonstrates that $\overline{\theta}$ is a unique projection of $\mathbb{P}_*$ to the fitted model (7) within the ball $\mathbb{B}(\overline{\theta}, \left|\overline{\theta}\right|/4)$. Based on the inequality (9) and the condition $C_\rho \leq 1/9$, for any two projections $\overline{\theta}_1$ and $\overline{\theta}_2$ of $\mathbb{P}_*$ to the fitted model (7), we find that

$$\left|\left|\overline{\theta}_1\right| - \left|\overline{\theta}_2\right|\right| \leq \min\left\{\left|\theta^* - \overline{\theta}_1\right|, \left|\theta^* + \overline{\theta}_1\right|\right\} + \min\left\{\left|\theta^* - \overline{\theta}_2\right|, \left|\theta^* + \overline{\theta}_2\right|\right\} \leq 2C_\rho\left|\theta^*\right| \leq 2\left|\theta^*\right|/9.$$

Additionally, for any projection $\overline{\theta}$ of $\mathbb{P}_*$ to the fitted model (7), we obtain that $\left|\overline{\theta}\right| \in [(1 - C_\rho)\left|\theta^*\right|, (1 + C_\rho)\left|\theta^*\right|] \in [8\left|\theta^*\right|/9, 10\left|\theta^*\right|/9]$. Therefore, for any two projections $\overline{\theta}_1$ and $\overline{\theta}_2$ such that $\overline{\theta}_1\overline{\theta}_2 > 0$, we have $\overline{\theta}_2$ in $\mathbb{B}(\overline{\theta}_1, \left|\overline{\theta}_1\right|/4)$. On the other hand, for any projections $\overline{\theta}_1$ and $\overline{\theta}_2$ such that $\overline{\theta}_1\overline{\theta}_2 < 0$, we find that $\left|\overline{\theta}_1 - \overline{\theta}_2\right| = \left|\overline{\theta}_1\right| + \left|\overline{\theta}_2\right| > \left|\overline{\theta}_1\right|/4$, which proves that $\overline{\theta}_2 \notin \mathbb{B}(\overline{\theta}_1, \left|\overline{\theta}_1\right|/4)$. The previous results imply that the projection of $\mathbb{P}_*$ to fitted model (7) is unique in its absolute value.

A.2 Proofs for computing model biases

In this section, we prove our results on the model bias in different cases, namely Propositions 1, 2 and 3. To facilitate further discussion, we begin with introducing some notations and variational formulation of the Wasserstein distance.

A.2.1 Notations

Given two distributions $\mathbb{P}$ and $\mathbb{Q}$, we use $h^2(\mathbb{P}, \mathbb{Q})$ and $\mathrm{KL}(\mathbb{P}, \mathbb{Q})$ to denote the hellinger distance and Kullback-Leibler divergence, respectively, between the two distributions. Let p and q denote the corresponding density of these distributions with respect to the Lebesgue measure. Then we have

$$h^2(\mathbb{P}, \mathbb{Q}) = \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx \quad \text{and} \quad \mathrm{KL}(\mathbb{P}, \mathbb{Q}) = \int p(x) \log \frac{p(x)}{q(x)} dx. \tag{24a}$$

We now introduce some notation to define the Wasserstein distance between two discrete measures. Given any two discrete measures $G = \sum_{i=1}^k \pi_i \delta_{\theta_i}$ and $G' = \sum_{i=1}^{k'} \pi'_i \delta_{\theta'_i}$, where $\theta_i, \theta'_i \in \Theta \subset \mathbb{R}$, and δ_θ denotes the dirac measure at θ. define the set of couplings $\Pi(G, G')$ between the two measures as follows:

$$\Pi(G, G') = \left\{ T \in \mathbb{R}_+^{k \times k'} \,:\, T 1_{k'} = \pi,\ T^\top 1_k = \pi' \right\}, \tag{24b}$$

where $\pi = (\pi_1, \ldots, \pi_k)^T$, $\pi' = (\pi'_1, \ldots, \pi'_{k'})^T$, and 1_k denotes a k-dimensional vector with all entries equal to 1. Put simply, $\Pi(G, G')$ is the set of all joint distributions T on the space $[k] \times [k']$ such that the marginals of the distribution T are equal to π and π'. Furthermore, for any given r, define the matrix $D \in \mathbb{R}^{k \times k'}$ of distances between the parameters of G and G' as

$$D_{ij} = \left|\theta_i - \theta'_j\right|^r, \quad (i, j) \in [k] \times [k']. \tag{24c}$$

With these notations in place, the Wasserstein distance [14] of order $r \geq 1$ between the two measures G and G' is given by

$$W_r^r(G, G') := \inf_{T \in \Pi(G, G')} \sum_{i=1}^k \sum_{j=1}^{k'} T_{ij} D_{ij}. \tag{24d}$$

With this notation in place, we now turn to the proofs of our propositions.

A.2.2 Proof of Proposition 1

In order to prove this proposition, we utilize several bounds between KL divergence, Hellinger distance, and Wasserstein distance. The road-map of the proof is as follows: First, we relate the KL divergences between the mixture distributions to the Wasserstein distances between the corresponding discrete mixing measures. Then, using carefully constructed couplings, we derive lower and upper bounds on the Wasserstein distances in terms of the bias term $\left|\overline{\theta} - \theta^*\right|$ and other problem parameters to obtain the claimed result.

For any mixing-measure (discrete mixture measure) G on Θ, let $\mathbb{P}_G$ denote the Gaussian mixture distribution induced by G on $\mathbb{R}$ whose density is given by $p_G(x) = \int_\Theta \phi(x; \theta, \sigma) dG$, where $\phi(\cdot; \theta, \sigma)$ denotes the density of the Gaussian distribution $\mathcal{N}(\theta, \sigma^2)$. We introduce the following notation for the mixing-measures:

$$G_* = \frac{1}{4}\delta_{-\theta^*(1-\rho)} + \frac{1}{4}\delta_{-\theta^*(1+\rho)} + \frac{1}{2}\delta_{\theta^*}, \quad \text{and} \quad G(\theta) = \frac{1}{2}\delta_{-\theta} + \frac{1}{2}\delta_{\theta}. \tag{25a}$$

Note that in our notation, $\mathbb{P}_* = \mathbb{P}_{G_*}$, $G_* \neq G(\theta^*)$ and consequently $\mathbb{P}_{G_*} \neq \mathbb{P}_{G(\theta^*)}$. Define

$$\bar{G} := G(\bar{\theta}) \quad \text{where} \quad \bar{\theta} \in \arg\min_{\theta\in\Theta} \mathrm{KL}(\mathbb{P}_{G_*}, \mathbb{P}_{G(\theta)}). \tag{25b}$$

Applying Lemma 1 from the paper [10], we obtain the following relationship between the KL divergence between the Gaussian-mixture measures $\mathbb{P}_{G^*}$ and $\mathbb{P}_{G(\theta)}$ and the Wasserstein distance between the corresponding mixing measures G and $G(\theta)$:

$$\mathrm{KL}(\mathbb{P}_{G_*}, p_{G(\theta)}) \leq W_2^2(G_*, G(\theta))/(2\sigma^2) \quad \text{for any } \theta \in \Theta.$$

Consequently, we find that

$$\mathrm{KL}(\mathbb{P}_{G_*}, \mathbb{P}_{\bar{G}}) = \min_{\theta\in\Theta} \mathrm{KL}(\mathbb{P}_{G_*}, \mathbb{P}_{G(\theta)}) \leq \min_{\theta\in\Theta} W_2^2(G_*, G(\theta))/(2\sigma^2). \tag{26}$$

On the other hand, from the classical bound between KL divergence and Hellinger distance, we have

$$\mathrm{KL}(\mathbb{P}_{G_*}, \mathbb{P}_{\bar{G}}) \geq 2h^2(\mathbb{P}_{G_*}, \mathbb{P}_{\bar{G}}). \tag{27}$$

Noting that, the univariate location Gaussian distribution is 4-strongly identifiable (cf. Definition 2.2 in [6] for the definition of 4-strongly identifiable condition and Theorem 2.4 in [6] for the result with univariate location Gaussian), with an application of the result of Theorem 6.3 in [6], we obtain that

$$h^2(\mathbb{P}_{G_*}, \mathbb{P}_{\bar{G}}) \geq C W_2^8(G_*, \bar{G}), \tag{28}$$

where C is a universal constant depending only on Θ. The results from (26), (27), and (28) lead to

$$2CW_2^8(G_*, \bar{G}) \leq \mathrm{KL}(\mathbb{P}_{G_*}, \mathbb{P}_{\bar{G}}) = \min_{\theta\in\Theta} \mathrm{KL}(\mathbb{P}_{G_*}, \mathbb{P}_{G(\theta)}) \leq \frac{1}{2\sigma^2} \min_{\theta\in\Theta} W_2^2(G_*, G(\theta)),$$

which implies that

$$2\sigma\sqrt{C}W_2^4(G_*, \bar{G}) \leq \min_{\theta\in\Theta} W_2(G_*, G(\theta)) \leq W_2(G_*, G(\theta^*)). \tag{29}$$

(Recall in our notation $G(\theta^*) \neq G_*$.) Now we derive obtain an upper bound for the distance $W_2(G_*, \bar{G})$, by deriving an upper bound for the distance $W_2(G_*, G(\theta^*))$ using the variational formulation (24d) of the Wasserstein distance. In particular, we use a particular coupling to derive an upper bound for $W_2^2(G_*, G(\theta^*))$. Recalling the definitions (24b) and (24c) of the coupling $\Pi(G_*, G(\theta^*))$ and the corresponding distance matrix D for $G = G_*, G' = G(\theta^*), r = 2$, we find that

$$T = \begin{bmatrix} 1/4 & 0 \\ 1/4 & 0 \\ 0 & 1/2 \end{bmatrix} \in \Pi(G_*, G(\theta^*)) \quad \text{and} \quad D = \begin{bmatrix} \rho^2\theta^{*2} & (2-\rho)^2\theta^{*2} \\ \rho^2\theta^{*2} & (2+\rho)^2\theta^{*2} \\ 4\theta^{*2} & 0 \end{bmatrix}.$$

Now applying the definition (24d), we obtain that

$$W_2^2(G_*, G(\theta^*)) \leq \sum_{i=1}^{3}\sum_{j=1}^{2} T_{ij}D_{ij} = \frac{1}{4}\rho^2\theta^{*2} + \frac{1}{4}\rho^2\theta^{*2} + \frac{1}{2}0 = \rho^2\theta^{*2}. \tag{30}$$

Putting the previous inequalities (29) and (30) together, we conclude that

$$W_2(G_*, \bar{G}) \leq c\left(\frac{\rho\,|\theta^*|}{\sigma}\right)^{1/4}, \tag{31}$$

where $c = 1/(4C)^{1/8}$ is a universal positive constant that depends only on the set Θ.

Now we directly obtain a lower bound for $W_2(G_*, \bar{G})$ by invoking the definition (24d) for the pair $(G_*, \bar{G})$. The corresponding distance matrix is given by

$$\bar{D} = \begin{bmatrix} \left(-\bar{\theta} + \theta^*(1-\rho)\right)^2 & \left(\bar{\theta} + \theta^*(1-\rho)\right)^2 \\ \left(-\bar{\theta} + \theta^*(1+\rho)\right)^2 & \left(\bar{\theta} + \theta^*(1+\rho)\right)^2 \\ \left(-\bar{\theta} + \theta^*\right)^2 & \left(\bar{\theta} + \theta^*\right)^2. \end{bmatrix}.$$

Noting that for any coupling $\bar{T} \in \Pi(G_*, \bar{G})$, we have $\bar{T}_{ij} \geq 0, \sum_{i,j} \bar{T}_{ij} = 1$, we have that

$$\sum_{i,j} \bar{T}_{ij} \bar{D}_{ij} \geq \min_{i,,j} \bar{D}_{ij},$$

and hence

$$\begin{aligned} W_2(G_*, \bar{G}) \geq \min_{i,,j} \sqrt{\bar{D}_{ij}} &= \min \Big\{ \left|-\bar{\theta} + \theta^*(1+\rho)\right|, \left|-\bar{\theta} + \theta^*(1-\rho)\right|, \left|-\bar{\theta} + \theta^*\right| \\ &\qquad \left|\bar{\theta} + \theta^*(1+\rho)\right|, \left|\bar{\theta} + \theta^*(1-\rho)\right|, \left|\bar{\theta} + \theta^*\right| \Big\} \\ &\geq \min \left\{ \left|\theta^* - \bar{\theta}\right|, \left|\theta^* + \bar{\theta}\right| \right\} - \rho \left|\theta^*\right|, \end{aligned} \tag{32}$$

where the last step follows from the triangle inequality. Putting the inequalities (31) and (32) together yields that

$$\min \left\{ \left|\theta^* - \bar{\theta}\right|, \left|\theta^* + \bar{\theta}\right| \right\} \leq \rho \left|\theta^*\right| + c \left(\frac{\rho \left|\theta^*\right|}{\sigma} \right)^{1/4},$$

and the proposition follows.

A.2.3 Proof of Proposition 2

The proof of the proposition is similar to the proof of Proposition 1 except for a few differences which we point to now. For any mixing-measure (discrete mixture measure) G on Θ, we denote by $\mathbb{P}_G$ the corresponding Gaussian mixture distribution with density $p_G(x) = \int_\Theta \phi(x; \theta, \sigma) dG$, where $\phi(\cdot; \theta, \sigma)$ denotes the density of the Gaussian distribution $\mathcal{N}(\theta, \sigma^2)$. For this proof, we define

$$G_* = \frac{(1-\omega)}{2} \delta_{-\theta^*} + \omega \delta_0 + \frac{(1-\omega)}{2} \delta_{\theta^*} \quad \text{and} \quad G(\theta) = \frac{1}{2} \delta_{(-\theta, \sigma)} + \frac{1}{2} \delta_{(\theta, \sigma)}.$$

Once again in our notation, we have $\mathbb{P}_* = \mathbb{P}_{G_*}$, $G_* \neq G(\theta^*)$ and consequently $\mathbb{P}_{G_*} \neq \mathbb{P}_{G(\theta^*)}$. Rewriting equation (29), we have

$$2\sigma\sqrt{C} W_2^4(G_*, \bar{G}) \leq \min_{\theta \in \Theta} W_2(G_*, G(\theta)) \leq W_2(G_*, G(\theta^*)),$$

where C is some universal positive constant only depending on Θ. Once again, we derive an upper bound for $W_2(G_*, \bar{G})$ by deriving an upper bound on $W_2(G_*, G(\theta^*))$. We now provide a coupling T and the matrix D (refer to equations (24b),(24c)) for the pair $G_*, G(\theta^*)$:

$$T = \begin{bmatrix} (1-\omega)/2 & 0 \\ \omega/2 & \omega/2 \\ 0 & (1-\omega)/2 \end{bmatrix} \in \Pi(G_*, G(\theta^*)) \quad \text{and} \quad D = \begin{bmatrix} 0 & 4{\theta^*}^2 \\ {\theta^*}^2 & {\theta^*}^2 \\ 4{\theta^*}^2 & 0 \end{bmatrix}.$$

Using the definition (24d), we have that $\sum_{ij} T_{ij} D_{ij}$ is an upper bound for $W_2^2(G_*, G(\theta^*))$. Doing some algebra yields that

$$W_2(G_*, \bar{G}) \leq c W_2^{1/4}(G_*, G(\theta^*)) \leq c \frac{\omega^{1/8} \left|\theta^*\right|^{1/4}}{\sigma^{1/4}} \tag{33}$$

where $c = 1/(4C)^{1/8}$ is a universal positive constant that only depends on Θ.

Now for the lower bound on $W_2(G_*, \bar{G})$, suppose that we are a given coupling $\bar{T} \in \Pi(G_*, \bar{G})$, and the distance matrix with elements

$$\bar{D} = \begin{bmatrix} \left(\theta^* - \bar{\theta}\right)^2 & \left(\theta^* + \bar{\theta}\right)^2 \\ \bar{\theta}^2 & \bar{\theta}^2 \\ \left(\theta^* + \bar{\theta}\right)^2 & \left(\theta^* - \bar{\theta}\right)^2 \end{bmatrix}.$$

Direct computation leads to

$$\begin{aligned} \sum_{ij} \bar{T}_{ij} \bar{D}_{ij} &= (\bar{T}_{11} + \bar{T}_{32}) \left(\theta^* - \bar{\theta}\right)^2 + (\bar{T}_{12} + \bar{T}_{31}) \left(\theta^* + \bar{\theta}\right)^2 + (T_{21} + T_{22}) \bar{\theta}^2 \\ &\geq (\bar{T}_{11} + \bar{T}_{32} + \bar{T}_{12} + \bar{T}_{31}) \cdot \min\{ \left(\bar{\theta} + \theta^*\right)^2, \left(\bar{\theta} - \theta^*\right)^2 \} \\ &= (1-\omega) \min \left\{ \left(\bar{\theta} + \theta^*\right)^2, \left(\bar{\theta} - \theta^*\right)^2 \right\}, \end{aligned}$$

where the final inequality is due to the constraint $\bar{T}_{11} + T_{21} = T_{13} + T_{23} = (1-\omega)/2$. As a result, we have

$$W_2^2(G_*, \bar{G}) \geq (1-\omega) \min\left\{ \left(\overline{\theta} + \theta^*\right)^2, \left(\overline{\theta} - \theta^*\right)^2 \right\}. \tag{34}$$

Combining the inequalities (33) and (34), we obtain that

$$\min\left\{ \left|\overline{\theta} - \theta^*\right|, \left|\overline{\theta} + \theta^*\right| \right\} \leq \frac{c\omega^{1/8} \left|\theta^*\right|^{1/4}}{\sigma^{1/4}\sqrt{1-\omega}},$$

thereby yielding the desired result.

A.2.4 Proof of Proposition 3

While the proof of the proposition follows several ideas from the other proofs on controlling the biases, a key difference for this case is that the two measures do not have same variance, and that forces us to use a couple new ideas in the proof. Define

$$G_* = \frac{1}{2}\delta_{(-\theta^*, \sigma^2 - \theta^{*2})} + \frac{1}{2}\delta_{(\theta^*, \sigma^2 - \theta^{*2})} \quad \text{and} \quad G(\theta) = \pi\delta_{(-\theta, \sigma^2)} + (1-\pi)\frac{1}{2}\delta_{(\theta, \sigma^2)}.$$

Note that we have $G_* \neq G(\theta^*)$. Unlike the cases considered in Proposition 1 and 2, the key lower bound $h^2(\mathbb{P}_{G_*}, \mathbb{P}_{\bar{G}}) \geq CW_2^8(G_*, \bar{G})$ does not apply hear for C some universal constant depending only on Θ. Such an issue is caused due to the fact that the variance of the components corresponding to G_* and $\bar{G}$ are different as $\theta^* \neq 0$. To overcome this issue, we claim the following point-wise bound:

$$h(\mathbb{P}_{G_*}, \mathbb{P}_{\bar{G}}) \geq C(G_*) W_2(G_*, \bar{G}), \tag{35}$$

where $C(G_*)$ is a positive constant depending only on G_* and Θ. To simplify notation, we substitute $C = C(G_*)$. Deferring the proof of the claim (35) to the end of this section, we proceed to finishing the proof.

Note that the relationship (27) between the Hellinger distance and the KL divergence is still valid but the bound (26) needs to be modified as follows (again applying Lemma 1 from the paper [10]):

$$\mathrm{KL}(\mathbb{P}_{G_*}, \mathbb{P}_{\bar{G}}) = \min_{\theta \in \Theta} \mathrm{KL}(\mathbb{P}_{G_*}, \mathbb{P}_{G(\theta)}) \leq \min_{\theta \in \Theta} \frac{C}{\sigma^2} W_2^2(G_*, G(\theta)) \tag{36}$$

for some large constant C. Putting the pieces together yields that

$$W_2(G_*, \bar{G}) \leq \frac{C}{\sigma} W_2(G_*, G(\theta^*)).$$

We now consider the following coupling and the distance matrix (refer to equations (24b) and (24c) respectively) for the mixing-measure pair $(G_*, G(\theta^*))$:

$$T = \begin{bmatrix} \pi & 0 \\ (1/2 - \pi) & 1/2 \end{bmatrix} \in \Pi(G_*, G(\theta^*)) \quad \text{and} \quad D = \begin{bmatrix} \theta^{*4} & 4\theta^{*2} + \theta^{*4} \\ 4\theta^{*2} + \theta^{*4} & \theta^{*4}. \end{bmatrix}$$

Invoking the variational formulation (24d), we find that

$$W_2^2(G_*, G(\theta^*)) \leq \sum_{i,j} T_{ij} D_{ij} = (2 - 4\pi)\theta^{*2} + \theta^{*4}.$$

Therefore, the following inequality holds

$$W_2(G_*, \bar{G}) \leq \frac{c(\theta^*)}{\sigma} \sqrt{(2-4\pi)\theta^{*2} + \theta^{*4}} \tag{37}$$

where $c(\theta^*)$ is a positive constant that only depends on G_* and Θ. On the other hand, for any coupling $\bar{T} \in \Pi(G_*, \bar{G})$, arguing as in the previous proofs, we have that

$$\sum_{i,j} \bar{T}_{ij} \bar{D}_{ij} \geq \min_{i,j} \bar{D}_{ij} \quad \text{where} \quad \bar{D} = \begin{bmatrix} \left(\theta^* - \overline{\theta}\right)^2 + \theta^{*4} & \left(\theta^* + \overline{\theta}\right)^2 + \theta^{*4} \\ \left(\theta^* + \overline{\theta}\right)^2 + \theta^{*4} & \left(\theta^* - \overline{\theta}\right)^2 + \theta^{*4} \end{bmatrix}$$

and consequently we have that

$$W_2(G_*, \bar{G}) \geq \min_{i,j} \sqrt{\bar{D}_{ij}} \geq \min\left\{ \left|\overline{\theta} - \theta^*\right|, \left|\overline{\theta} + \theta^*\right| \right\}. \tag{38}$$

Combining the inequalities (37) and (38) yields the result.

Proof of claim (35) In order to prove inequality (35), it suffices to show that

$$\inf_{\theta \in \Theta} h(\mathbb{P}_{G_*}, \mathbb{P}_{G(\theta)})/W_2(G_*, G(\theta)) > 0.$$

We proceed via proof by contraction: assume that the above bound does not hold. It implies that we can find a sequence of $\{\theta_n\}_{n \geq 1}$ such that $h(\mathbb{P}_{G_*}, \mathbb{P}_{G(\theta_n)})/W_2(G_*, G(\theta_n)) \to 0$ as $n \to \infty$. Since Θ is a compact subset of $\mathbb{R}$, there must exist a subsequence of θ_n such that $\theta_n \to \theta'$ for some $\theta' \in \Theta$. Without loss of generality, we can replace this subsequence of θ_n by its whole sequence. Applying Fatou's lemma, we obtain that

$$\begin{aligned} 0 = \lim_{n \to \infty} h^2(\mathbb{P}_{G_*}, \mathbb{P}_{G(\theta_n)}) &= \frac{1}{2} \int \liminf_{n \to \infty} \left(\sqrt{\mathbb{P}_{G_*}(x)} - \sqrt{\mathbb{P}_{G(\theta_n)}(x)} \right)^2 d\mu(x) \\ &= h^2(\mathbb{P}_{G_*}, \mathbb{P}_{G(\theta')}). \end{aligned}$$

The above result implies that $\mathbb{P}_{G_*}(x) = \mathbb{P}_{G(\theta')}(x)$ almost surely. Due to the general identifiability of finite location-scale Gaussian mixtures [11], the previous equations implies that $G_* \equiv G(\theta')$, which is a contradiction as $\pi \in (0, 1/2)$ and $\theta^* \neq 0$. Therefore, we have established the claim (35).

A.3 Proofs for sample-based EM

In this section, we prove the Corollaries 1 and 2. The proof for Corollary 3 is rather similar and is omitted.

A.3.1 Proof of Corollary 1

To prove this corollary, we use Theorem 2 by Balakrishnan et al. [1] and note that it suffices to establish the following lemma:

Lemma 1. *For any threshold $\delta \in (0, 1)$, we have*

$$\mathbb{P}\left[\sup_{\theta \in \Omega} |M_n(\theta) - M(\theta)| - c_2 \left(1 + C_\rho\right) |\theta^*| \left((1 + \rho^2)\theta^{*2} + \sigma^2\right) \sqrt{\frac{\log(1/\delta)}{n}} \geq 0\right] \leq \delta,$$

where $\Omega := \mathbb{B}(\overline{\theta}, \left|\overline{\theta}\right|/4)$ for sample size $n \geq c_1 \log(1/\delta)$ where c_1 and c_2 are universal positive constants.

Proof. The proof of this lemma makes use of standard arguments to derive Rademacher complexity bounds. We denote

$$Z = \sup_{\theta \in \Omega} |M_n(\theta) - M(\theta)|.$$

By means of standard symmetrization argument with empirical processes [12], the following holds

$$\mathbb{E}\left[\exp\left(\lambda Z\right)\right] \leq \mathbb{E}\left[\exp\left(\sup_{\theta \in \Omega} \frac{2\lambda}{n} \left|\sum_{i=1}^{n} \epsilon_i (2w_\theta(X_i) - 1) X_i\right|\right)\right], \text{ for any } \lambda > 0.$$

For any θ and θ', we have

$$|2w_\theta(x) - 2w'_\theta(x)| \leq |\theta - \theta'| \, |x|,$$

for all $x \in \mathbb{R}$. Invoking the Ledoux-Talagrand contraction result for Rademacher processes [8] yields

$$\begin{aligned} \mathbb{E}\left[\exp\left(\sup_{\theta \in \Omega} \frac{2\lambda}{n} \left|\sum_{i=1}^{n} \epsilon_i (2w_\theta(X_i) - 1) X_i\right|\right)\right] &\leq \mathbb{E}\left[\exp\left(\sup_{\theta \in \Omega} \frac{4\lambda}{n} \left|\sum_{i=1}^{n} \epsilon_i X_i^2 \theta\right|\right)\right] \\ &= \mathbb{E}\left[\exp\left(\frac{5\lambda \left|\overline{\theta}\right|}{n} \sum_{i=1}^{n} \epsilon_i X_i^2\right)\right], \end{aligned} \quad (39)$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables which are independent of X_i's. Recalling the distribution of X_i's, we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda X_i\right)\right] &= \exp\left(\lambda^2\sigma^2/2\right)\left(\frac{1}{4}\exp\left(-\lambda\theta^*(1+\rho)\right) + \frac{1}{4}\exp\left(-\lambda\theta^*(1-\rho)\right) + \frac{1}{2}\exp(\lambda\theta^*)\right)\\ &\overset{(i)}{\leq} \exp\left(\lambda^2\sigma^2/2\right)\left(\frac{1}{4}\exp\left(-\lambda\theta^*\rho\right) + \frac{1}{4}\exp\left(\lambda\theta^*\rho\right)\right)\left(\exp(-\lambda\theta^*) + \exp(\lambda\theta^*)\right)\\ &\overset{(ii)}{\leq} \exp\left(\lambda^2\frac{(1+\rho^2){\theta^*}^2 + \sigma^2}{2}\right),\end{aligned}$$

where step (i) and step (ii), respectively, follow from the basic inequalities $\exp(-y) + \exp(y) \geq 2$ and $\exp(-y) + \exp(y) \leq 2\exp(y^2/2)$ for all $y \in \mathbb{R}$. Thus, the random variable X_i is sub-Gaussian with parameter at most $\gamma = ((1+\rho^2){\theta^*}^2 + \sigma^2)^{1/2}$ for any $i \in [n]$. Since any squared sub-Gaussian random variable is a sub-exponential random variable, the following inequality holds [13]:

$$\mathbb{E}\left[\exp\left(tX_i^2 - t\mathbb{E}\left[X_i^2\right]\right)\right] \leq \exp\left[16t^2\gamma^4\right] \text{ for all } |t| \leq \frac{1}{4\gamma^2}.$$

Furthermore, we can bound the second moment of X_i as follows:

$$\begin{aligned}\mathbb{E}\left[X_i^2\right] &= \frac{1}{4}\left((\theta^*(1+\rho))^2 + \sigma^2\right) + \frac{1}{4}\left((\theta^*(1-\rho))^2 + \sigma^2\right) + \frac{1}{2}\left({\theta^*}^2 + \sigma^2\right)\\ &\leq (1+\rho^2){\theta^*}^2 + \sigma^2 = \gamma^2.\end{aligned}$$

Using these MGF and moment bounds, we find that

$$\begin{aligned}\mathbb{E}\left[\exp\left(t\epsilon_i X_i^2\right)\right] &= \frac{1}{2}\mathbb{E}\left[\exp\left(tX_i^2\right)\right] + \frac{1}{2}\mathbb{E}\left[\exp\left(-tX_i^2\right)\right]\\ &\leq \exp\left(16t^2\gamma^4\right)\frac{1}{2}\left(\exp(t\gamma^2) + \exp(-t\gamma^2)\right)\\ &\leq \exp(17t^2\gamma^4), \end{aligned} \tag{40}$$

for all $|t| \leq \frac{1}{4\gamma^2}$. Plugging in $t = 5\lambda\left|\overline{\theta}\right|/n$ in the bound (40) and combining with the bound (39) yields the following MGF bound

$$\mathbb{E}\left[\exp\left(\lambda Z\right)\right] \leq \exp\left(425\lambda^2\overline{\theta}^2\gamma^4/n\right) \leq \exp\left(425\lambda^2\left(1+C_\rho\right)^2{\theta^*}^2\gamma^4/n\right)$$

for $|\lambda| \leq n/(20\gamma^2\left|\overline{\theta}\right|)$. Here the second inequality in the above display is due to the upper bound $\left|\overline{\theta}\right| \leq \left(1+C_\rho\right)|\theta^*|$ from Proposition 1. By virtue of standard Chernoff's approach, the above MGF bound implies that

$$Z \leq c_2\left(1+C_\rho\right)|\theta^*|\gamma^2\sqrt{\frac{\log(1/\delta)}{n}} = c_2\left(1+C_\rho\right)|\theta^*|\left((1+\rho^2){\theta^*}^2 + \sigma^2\right)\sqrt{\frac{\log(1/\delta)}{n}},$$

with probability at least $1-\delta$ as long as $n \geq c_1\log(1/\delta)$ for sufficiently large positive constants c_1 and c_2. The lemma now follows. □

A.3.2 Proof of Corollary 2

Similar to the argument of Corollary 1, to prove this corollary it is sufficient to establish the following lemma:

Lemma 2. *For any threshold $\delta \in (0,1)$, we have*

$$\mathbb{P}\left[\sup_{\theta\in\Omega}|M_n(\theta) - M(\theta)| - c_2\left(1+C_\omega\right)|\theta^*|\left({\theta^*}^2 + \sigma^2\right)\sqrt{\frac{\log(1/\delta)}{n}} \geq 0\right] \leq \delta, \tag{41}$$

where $\Omega := \mathbb{B}(\overline{\theta}, \left|\overline{\theta}\right|/4)$ for sample size $n \geq c_1\log(1/\delta)$ where c_1 and c_2 are universal positive constants.

Proof. Following the argument used in the proof of Lemma 1, we obtain that

$$\mathbb{E}\left[\exp\left(\lambda Z\right)\right] \leq \mathbb{E}\left[\exp\left(\frac{5\lambda\left|\overline{\theta}\right|}{n}\sum_{i=1}^{n}\epsilon_i X_i^2\right)\right],$$

where $Z = \sup_{\theta\in\Omega}|M_n(\theta) - M(\theta)|$ and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables independent of X_i's. Recalling the distribution of X_i's, we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda X_i\right)\right] &= \exp\left(\lambda\sigma^2/2\right)\left(\left(\frac{1-\omega}{2}\right)\left(\exp\left(-\lambda\theta^*\right) + \exp\left(\lambda\theta^*\right)\right) + \omega\right) \\ &\leq \exp\left(\lambda^2\sigma^2/2\right)\left((1-\omega)\exp\left(\frac{\lambda^2{\theta^*}^2}{2}\right) + \omega\right) \\ &\leq \exp\left(\lambda^2(\sigma^2 + {\theta^*}^2)/2\right).\end{aligned}$$

Thus, the random variables X_i are independent sub-Gaussian with parameter at most $\gamma = \sqrt{{\theta^*}^2 + \sigma^2}$ for all $i \in [n]$. Furthermore, we can bound the second moment of X_i as follows:

$$\mathbb{E}\left[X_i^2\right] = (1-\omega)({\theta^*}^2 + \sigma^2) + \omega\sigma^2 \leq ({\theta^*}^2 + \sigma^2). \tag{42a}$$

Using these MGF and moment bounds, we have

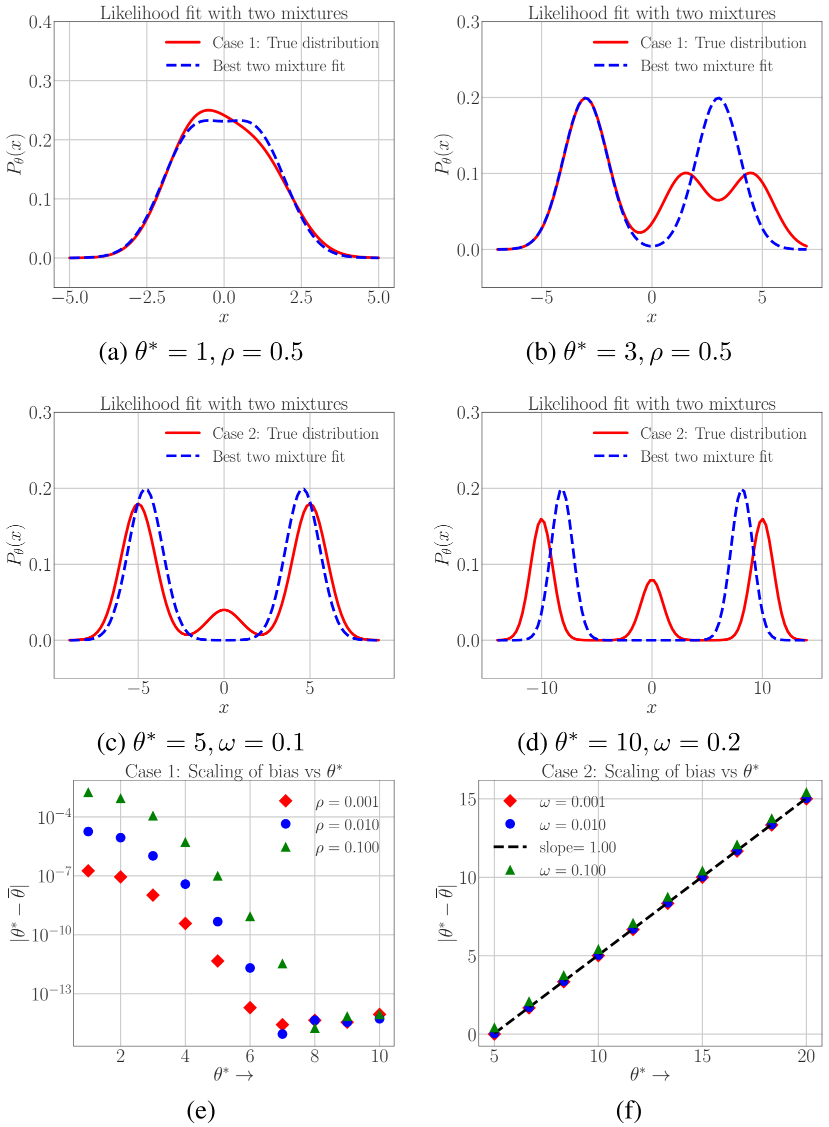
$$\mathbb{E}\left[\exp\left(t\epsilon_i X_i^2\right)\right] \leq \exp(17t^2\gamma^4) \quad \text{for all } |t| \leq 1/4\gamma^2. \tag{42b}$$

Finally, performing computations similar to those in the proof of Lemma 1 yields the claim.

□

B Further numerical experiments

Here we provide supplementary material for the numerical experiments presented in Section 5 of the main text. In particular, we numerically illustrate the scalings of the bias $\left|\theta^* - \overline{\theta}\right|$ as a function of θ^* in Figure 2. We use population EM (the final iterate) to estimate $\overline{\theta}$ (2) for the different settings. We simulated two different settings for Case 1 corresponding to the two mixture fit (7) for the three mixture model (6) and Case 2 corresponding to the two mixture fit (7) for the three mixture model (11) and report the results in Figure 2. The behavior of the bias $|\theta^* - \overline{\theta}|$ is rather different in the two cases. We see that for Case 1 the bias decreases with increase in θ^*, while for Case 2, it increases with increase in θ^*. Such a behavior is not captured in our results stated in Propositions 1 and 2. Thus, the bias analysis presented in this paper should be considered only a first step towards understanding the under-fitted mixtures. Providing a sharper framework that yields optimal bounds for such biases remains an interesting future direction.



(a) $\theta^* = 1, \rho = 0.5$

(b) $\theta^* = 3, \rho = 0.5$

(c) $\theta^* = 5, \omega = 0.1$

(d) $\theta^* = 10, \omega = 0.2$

(e)

(f)

Figure 2. Plots of the best two mixture Gaussian fits for the data generated from three Gaussian mixtures. In (a) & (b), we consider two settings of Case 1 (6) and in panels (c) & (d), two settings of Case 2 (11). We see that for Case 1 as θ^* increases, $\overline{\theta} \rightarrow \theta^*$ and for Case 2 an increase in θ^* leads to an increase in the bias $|\theta^* - \overline{\theta}|$. Indeed as we plot the bias term in panels (e) and (f), we see that for Case 1, larger θ^* has a smaller bias and on the contrary for Case 2, the bias increases with increase in θ^*.