Abstract

Recent years have witnessed substantial progress in understanding the behavior of EM for mixture models that are correctly specified. Given that model misspecification is common in practice, it is important to understand EM in this more general setting. We provide non-asymptotic guarantees for population and sample-based EM for parameter estimation under a few specific univariate settings of misspecified Gaussian mixture models. Due to misspecification, the EM iterates no longer converge to the true model and instead converge to the projection of the true model over the set of models being searched over. We provide two classes of theoretical guarantees: first, we characterize the bias introduced due to the misspecification; and second, we prove that population EM converges at a geometric rate to the model projection under a suitable initialization condition. This geometric convergence rate for population EM imply a statistical complexity of order $1/pn$ when running EM with $n$ samples. 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 computing the MLE is often non-trivial. One of the most popular algorithm used to compute the MLE is the expectation maximization (EM) algorithm. Although EM is widely used in practice, the convergence properties of EM are less understood. Classical results provide guarantees about the convergence rates of EM to local maxima [5, 18]. 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 [10, 20]. 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 setting which includes two components location-Gaussian mixtures as an example, they provided sufficient conditions under which well-initialized EM algorithm converges to a close
neighborhood of global maximum, and also provided explicit sample-based complexity of EM in this setting. A number of subsequent papers have generalized and extended results of this type (see the papers [22, 17, 8, 19, 4, 21, 6, 3]). However, a shared assumption across these works is that either the true distribution of each subpopulation is known or the number of components is exactly known; in practice, both of these conditions are often violated. In a recent line of work [2, 11], the authors suggested ways to estimate the number of components for the fitted model, nonetheless, the practicality of such techniques are rather limited. The MLE, instead of approximating the true parameter, approximates a Kullback-Leibler projection of the data-generating distribution onto the fitted model class, and hence is known to be robust to various forms of misspecification. Consequently, it is interesting to understand conditions under which the EM algorithm, in mis-specified models, converge to an appropriate KL projection. The contribution of this paper is to provide precise characterization of the EM convergence to the aforementioned KL projection, and quantify the bias that comes from model mis-specification. We provide a number of theoretical guarantees for the EM algorithm applied to two classes of misspecified 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, arises naturally when either the mixture components are very close or the mixture weights are very small, and in both cases the true model appears to have fewer components. Analysis of the EM algorithm in this setting poses new challenges; in particular, it requires an understanding of the **model bias**, meaning that, the Kullback-Leibler discrepancy between the true model from its projection(s) to the fitted model. In this paper, we provide a detailed analysis of the \( k = 3 \) case. First, we characterize the bias of two-component mixture when the true model is a three-component mixture with unknown means but known variance. We then provide sufficient conditions for the bias to be small enough, and prove that under such cases the population EM updates have a geometric rate of convergence to the KL projection of the true model onto the fitted model class. Finally, using Rademacher-complexity based arguments and the linear 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 misspecified. 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 regarding EM convergence: despite the potential non-convexity of the problem, the iterates, with arbitrary initialization, converge at a geometric rate to a unique fixed point. 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 \((\theta, \sigma, \alpha)\) to denote the Gaussian mixture component with mean \( \theta \), variance \( \sigma^2 \) and weight \( \alpha \), i.e., \( \alpha N(\theta, \sigma^2) \). 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 misspecified. Numerical experiments illustrating our theoretical results are presented in Section 5. We conclude the paper with discussion regarding our results as well as 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 \( P \) and \( Q \), the Kullback-Leibler divergence between them is denoted by \( KL(P, Q) \). We use the standard big-O notation to depict the scaling with respect to a particular quantity and hide constants and other problem parameters.

## 2 Problem set-up

Throughout this paper, we assume that data is generated according to some true distribution \( P_* \), which admits a continuous density over \( \mathbb{R} \). We are interested in the performance of the EM algorithm
More precisely, we denote the population EM operator under a few specific settings of true model. Our main goal in the paper is to establish the convergence rate of well initialized EM updates to the fitted model (1).

In order to achieve the estimation of location parameter \( \theta \) 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 \( |\hat{\theta}_n - \hat{\theta}| \) has the statistical rate of order \( n^{-1/2} \) in all cases, as depicted in the fourth column (here \( \hat{\theta}_n \) denotes the final sample EM estimate).

When we fit the model below using a two-component mixture of location-shifted Gaussians with known variance \( \sigma^2 \) and known mixture weight \( \pi \in (0, 1) \):

\[
P_\theta = \pi N(\theta, \sigma^2) + (1 - \pi)N(-\theta, \sigma^2)
\]

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 achieve the estimation of location parameter \( \theta \) that fits the true model efficiently, we utilize the vanilla EM algorithm over a bounded search space \( \Theta \). Since the true distribution \( P_\theta \) may not belong to the fit model-class, the best possible estimator is the projections of \( P_\pi \) to the fitted model (1), which can be defined rigorously as follows:

\[
\hat{\theta} \in \arg \min_{\theta \in \Theta} KL(P_\pi, P_\theta).
\]

Our main goal in the paper is to establish the convergence rate of well initialized EM updates to \( \hat{\theta} \) under a few specific settings of true model \( P_\pi \) and fitted model (1).

### 2.1 EM algorithm for two-component location-Gaussian mixtures

Let us now introduce few notations and a brief description of the EM algorithm for two components location-Gaussian mixtures (1). The population version of EM is based on the function

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

where the outer expectation is taken over the true distribution \( P_\pi \). For any fixed \( \theta \), the M-step in the EM updates for the model (1) is obtained by maximizing the minorization function (3) (for detailed derivation see the paper [1]). More precisely, we denote the population EM operator \( M : \mathbb{R} \rightarrow \mathbb{R} \) as

\[
M(\theta) := \arg \max_{\theta'} Q(\theta', \theta) = \mathbb{E} \left[ (2w_\theta(X) - 1)X \right],
\]
where the weighting function \( w_\theta \) in the above formulation is given by

\[
    w_\theta(x) := \frac{\pi \exp\left(-\frac{(\theta-x)^2}{2\sigma^2}\right)}{\pi \exp\left(-\frac{(\theta-x)^2}{2\sigma^2}\right) + (1-\pi) \exp\left(-\frac{(\theta+x)^2}{2\sigma^2}\right)}.
\]

If there exists some \( \theta \in \Theta \) such that the M-step updates satisfy the self-consistency property—that is, \( M(\theta) = \theta \)—then well-initialized population EM updates converge to \( \theta \). It is interesting to point out that the parameter \( \hat{\theta} \), 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 \( \hat{\theta} \). Consequently, the parameter \( \theta \) satisfies self-consistency property, i.e. \( M(\theta) = \theta \).

The sample version for the M-step updates of EM is obtained by simply replacing the expectation in equation (3) and consequently in equation (4) by its sample counterpart. In particular, given a set of \( n \) i.i.d. samples \( \{X_i\}_{i=1}^n \) from the true model, we find that the sample EM operator \( M_n : \mathbb{R} \rightarrow \mathbb{R} \) takes the form

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

With these notations in place, we are now ready to state our main results.

### 3 Guarantees for EM algorithm for misspecified 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 \( P_* \) is a mixture of three components

\[
    P_* = \frac{1}{4} \mathcal{N}(-\theta^*(1+\rho), \sigma^2) + \frac{1}{4} \mathcal{N}(-\theta^*(1-\rho), \sigma^2) + \frac{1}{2} \mathcal{N}(\theta^*, \sigma^2)
\]

for some \( \theta^* \in \Theta \) (compact subset of \( \mathbb{R} \)) and a small positive scalar \( \rho \) that characterizes the separation between two cluster means \(-\theta^*(1+\rho)\) and \(-\theta^*(1-\rho)\). For fitting the model, we assume that the variance \( \sigma^2 \) is known, and we suspect that the true model is a two-component mixture (since \( \rho \) is small). Consequently, we fit the data with the following model:

\[
    P_\theta = \frac{1}{2} \mathcal{N}(-\theta, \sigma^2) + \frac{1}{2} \mathcal{N}(\theta, \sigma^2),
\]

and we estimate the parameter \( \theta \) using the EM algorithm. Clearly, the performance of model (8) and consequently the EM algorithm would depend on the relationship between the separation factor \( \rho \) and the SNR \( \eta := |\theta^*|/\sigma \) of the true model (7). Since the true model does not belong in the family of two components location-Gaussian mixtures in model class (8), the role of the projection parameter \( \hat{\theta} \in \arg \min_{\theta \in \Theta} KL(P_*, P_\theta) \) becomes crucial. In the next proposition, we provide an explicit bound for the bias between \( \theta \) and \( \theta^* \) as a function of the problem parameters.

**Proposition 1.** Given the true model (7) and any \( \rho > 0 \), we have

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

where \( c \) is a universal positive constant that depends only on the set \( \Theta \).
In order to simplify our results in the sequel, we assume that \( \eta = |\theta^*|/\sigma \geq 1 \) and use the following simpler bound on the bias from now onwards:

\[
\min \{ |\theta^* - \bar{\theta}|, |\theta^* + \bar{\theta}| \} \leq \left( \rho + \frac{\rho^{1/4}}{\eta^{3/4} \sigma} \right) |\theta^*| \leq \left( \rho + \frac{\rho^{1/4}}{\sigma} \right) |\theta^*|. \tag{10}
\]

The bound above directly implies that \( |\bar{\theta}| \in \left[ (1 - C_\rho) |\theta^*|, (1 + C_\rho) |\theta^*| \right] \), assuming \( C_\rho := \rho + \rho^{1/4}/\sigma \leq 1 \). As \( \rho \to 0 \), we have \( C_\rho \to 0 \) implying that \( |\bar{\theta}| \) and \( |\theta^*| \) are almost identical. In the sequel, we utilize this precise control of \( |\bar{\theta}| \) in terms of \( |\theta^*| \), provided by Proposition 1, to analyze the behavior of the EM algorithm in a neighborhood of \( \bar{\theta} \). Let \( \rho_* = \sup \{ \rho > 0 | C_\rho \leq 1/2 \} \). In the next theorem, we characterize the behavior of the population EM operator for the three components location-Gaussian mixtures described by equation (7).

**Theorem 1.** There exist universal constants \( c', c'' \) such that the population EM operator for model (7) with \( \rho \leq \rho_* \) and \( \eta \geq c' \) satisfies

\[
|M(\theta) - \bar{\theta}| = \mathbb{E}[2(w_\theta(X) - w_{\bar{\theta}}(X))X] \leq \gamma |\theta - \bar{\theta}|, \quad \text{for any } \theta \in \mathbb{B}(\bar{\theta}, |\bar{\theta}|/4).
\]

In words, Theorem 1 establishes that the population EM iterates (in the ideal, infinite data limit) are \( \gamma \)-contractive with respect to \( \bar{\theta} \) over the ball \( \mathbb{B}(\bar{\theta}, |\bar{\theta}|/4) \), where \( \gamma \leq e^{-c'' \eta^2} \). Such a result has 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 converges in a constant number of steps to a neighborhood of \( \bar{\theta} \).

**Corollary 1.** Consider any scalar \( \delta \in (0, 1) \), sample size \( n \geq c_1 \log(1/\delta) \) and starting point \( \theta^0 \in \mathbb{B}(\bar{\theta}, |\bar{\theta}|/4) \). Then under the assumptions of Theorem 1, the sample-based EM sequence \( \theta^{t+1} = M_n(\theta^t) \) for the model (7) satisfies

\[
|\theta^t - \bar{\theta}| \leq \gamma^t |\theta^0 - \bar{\theta}| + \frac{c_2}{1 - \gamma} |\theta^*| \left( \theta^{t+1} + \sigma^2 \right) \sqrt{\log(1/\delta) \over n} \tag{11}
\]

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

Note that the bound (11) 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 \( O(\sqrt{1/n}) \). Therefore, with probability at least \( 1 - \delta \), we have

\[
|\theta^T - \bar{\theta}| \leq \frac{c |\theta^*| \left( \theta^{t+1} + \sigma^2 \right)}{1 - \gamma} \sqrt{\log(1/\delta) \over n} \quad \text{for } T \geq c \log(n/\log(1/\delta) |\theta^*| \left( \theta^{t+1} + \sigma^2 \right)),
\]

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 \( P_* \) is a mixture of three components location-Gaussian mixtures as follows:

\[
P_* = \frac{1 - \omega}{2} N(-\theta^*, \sigma^2) + \omega N(0, \sigma^2) + \frac{1 - \omega}{2} N(\theta^*, \sigma^2). \tag{12}
\]

In other words, two components are dominant with means \(-\theta^* \) and \( \theta^* \) 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 (8). 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 \( \bar{\theta} \) that we characterize in the next proposition.

**Proposition 2.** For the three components location-Gaussian mixtures in model (12), we have

\[
\min \{ |\theta^* - \bar{\theta}|, |\theta^* + \bar{\theta}| \} \leq \frac{c \omega^{1/8} |\theta^*|^{1/4}}{\sigma^{1/4} \sqrt{1 - \omega}}. \tag{13}
\]

where \( c \) is a universal positive constant that depends only on the set \( \Theta \).
In order to simplify further results, we assume under the condition \( \eta := \frac{\alpha}{\sigma} \geq 1 \). Then we have \( \min \{ |\theta^* - \bar{\theta}|, |\theta^* + \bar{\theta}| \} \leq C_\omega |\theta^*| \), where \( C_\omega := \omega^{1/8}/(\sigma \sqrt{1 - \omega}) \). Such a bound on bias leads to slightly different conditions for convergence of the EM algorithm for model (12) compared to the EM convergence for model (7). Note that for any fixed variance \( \sigma^2 \), the function \( C_\omega \) increases with \( \omega \) and \( C_0 = 0 \). Let \( \omega_* = \sup \{ \omega > 0 | C(\omega) \leq 1/2 \} \). Similar to the model (7), we analyze the convergence rate of EM under a strong SNR condition of true model (12). 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 (12).

**Theorem 2.** For SNR \( \eta \geq 1 \) sufficiently large and \( \omega \leq \omega_* \), and for any \( \theta^0 \in \mathbb{B}(\bar{\theta}, |\bar{\theta}|/4) \), the population EM operator for the Gaussian mixture (12) satisfies

\[
|M(\theta^0) - \bar{\theta}| = \mathbb{E} |2(w_{\theta^0}(X) - w_{\bar{\theta}}(X))X| \leq \tilde{\gamma} |\theta^0 - \bar{\theta}|.
\]

Consequently, the population EM sequence \( \theta^{t+1} = M(\theta^t) \) converges to \( \bar{\theta} \) 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 \( \eta \) and weight \( \omega \). More concretely, if the SNR is large enough, the population EM converges fast towards the projection \( \bar{\theta} \). 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 (12) such that the assumptions of Theorem 2 hold. For any fixed \( \delta \in (0, 1) \), \( \theta^0 \in \mathbb{B}(\bar{\theta}, |\bar{\theta}|/4) \), if \( n \geq c_1 \log(1/\delta) \) then the sample EM iterates \( \theta^{t+1} = M_n(\theta^t) \) satisfy

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

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 \( O(\sqrt{1/n}) \) after \( T = O(\log n / \log(1/\delta)) \) iterations.

### 4 Robustness of EM for misspecified variances and weights

In this section, we focus on establishing the convergence rate of EM under different misspecified regime of the fitted model (1). In particular, we assume that the true data distribution \( P \), is given by:

\[
P_\pi = \frac{1}{2} \mathcal{N}(\theta^*, \sigma^2 - \theta^*^2) + \frac{1}{2} \mathcal{N}(\theta^*, \sigma^2 - \theta^*^2),
\]

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 components location-Gaussian mixture \( P_\theta \) that is given by

\[
P_\theta = \pi \mathcal{N}(\theta, \sigma^2) + (1 - \pi) \mathcal{N}(\theta^*, \sigma^2),
\]

where \( \pi := \frac{1}{2}(1 - \epsilon) \) and \( |\epsilon| \in (0, 1) \) are known apriori and only the parameter \( \theta \) is to be estimated.

In the fitted model \( P_\theta \), we have misspecified the variance \( \sigma^2 \) and the weight \( \pi \), and we wish to understand the rate of convergence of EM to \( \theta \), where \( \bar{\theta} \) is the parameter of the model \( P_\theta \), and \( P_\theta \) is the projection of the true model \( P_\pi \) onto the model class \( \mathcal{P}_{\theta} := \{ P_\theta : \theta \in \mathbb{R} \} \). We emphasize that the main goal here is to see how the misspecification 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 \( \theta \).

**Theorem 3.** Under the setting of two components location-Gaussian mixtures (14) and fitted model (15), the population EM operator \( \theta \mapsto M(\theta) \) satisfies

\[
|M(\theta) - \bar{\theta}| \leq \left( 1 - \frac{\epsilon^2}{2} \right) |\theta - \bar{\theta}|
\]

Hence, the population EM sequence \( \{\theta^t\} \) converges geometrically to \( \bar{\theta} \) from any initialization \( \theta^0 \).
There are two interesting features regarding the geometric convergence of population EM updates to \( \hat{\theta} \): (1) it does not require an evaluation of bias which was needed for our previous results; (2) it holds under any initialization \( \theta^0 \). Overall, we have that \( \hat{\theta} \) is unique, thereby we conclude that the projection of \( P_{\hat{\theta}} \) 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 \( \hat{\theta} \):

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

\[
\min \{|\hat{\theta} - \theta^*|, |\hat{\theta} + \theta^*|\} \leq c(\theta^*) \cdot \sqrt{(2 - 4\pi)\theta^*2 + \theta^*4}/\sigma,
\]

where \( c(\theta^*) \) is a positive constant depending only on \( \theta^* \) and the set \( \Theta \).

Since \( |\theta^*| < \sigma/2 \), the above bound implies that \( |\hat{\theta}| \leq |(1 - C)|\theta^*|, (1 + C)|\theta^*| \) where \( C = c(\theta^*)\sqrt{(2 - 4\pi) + \sigma^2}/\sigma \). Equipped with this bound on \( |\hat{\theta}| \), 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 \( \theta^0 \in B(\bar{\theta}, r) \), then the sample-based EM sequence \( \theta^{t+1} = M_n(\theta^t) \), satisfies

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

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 \( O(\sqrt{1/n}) \) around \( \bar{\theta} \) after \( T = O(\log n / \log(1/(1 - \epsilon^2/2))) \) iterations.

## 5 Simulation studies

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

- Case 1 refers to the true model (7) 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 (12) 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 (8); while for the third case we fit the unbalanced two-Gaussian mixture given by equation (15) for different values of \( \pi \). Let \( \hat{\theta}_n \) denote the final sample EM estimate. Since our results establish that population EM converges to \( \bar{\theta} \) (2), we use the final iterate from the population EM sequence to estimate the error \( |\hat{\theta}_n - \bar{\theta}| \). We now summarize our key findings:

1. In Figure 1(a), we observe that for all cases the final statistical error \( |\hat{\theta}_n - \bar{\theta}| \) has a parametric rate \( n^{-1/2} \) which verifies the claims of Corollaries 1, 2 and 3.
2. 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 \( \pi \rightarrow 0.5 \), the rate of decay of the error of population EM sequence decreases which is consistent with the contraction result stated in Theorem 3.
3. In panels (c) and (d) of Figure 1, we plot the biases for Case 1 and 2, with respect to \( \rho \) and \( \omega \) 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
$|\theta^* - \tilde{\theta}|$ is of the order $\rho^2$ and $\omega$ 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 $\theta^*$ in these cases via further simulations.

In Appendix B, we illustrate the scaling of the bias with $\theta^*$ in these cases via further simulations.

In this paper, we established several new convergence results for population and sample-based EM updates under few specific settings of misspecified mixture models. Analyzing behavior of EM under general misspecification is challenging in general and we consider our work as a first step towards developing a framework for the same. We considered only univariate mixtures in this paper, but we strongly believe that several of our results can be extended to multivariate mixtures. While this paper focused on estimating the location parameter, it is natural to investigate the behavior of EM algorithm, when we fit a model with scale as well. Besides deriving sharper results for the settings considered in this paper, developing a framework to analyze the behavior of EM for non-Gaussian and more general mixture models is an appealing venue for future work.

6 Discussion

In this paper, we established several new convergence results for population and sample-based EM updates under few specific settings of misspecified mixture models. Analyzing behavior of EM under general misspecification is challenging in general and we consider our work as a first step towards developing a framework for the same. We considered only univariate mixtures in this paper, but we strongly believe that several of our results can be extended to multivariate mixtures. While this paper focused on estimating the location parameter, it is natural to investigate the behavior of EM algorithm, when we fit a model with scale as well. Besides deriving sharper results for the settings considered in this paper, developing a framework to analyze the behavior of EM for non-Gaussian and more general mixture models is an appealing venue for future work.
References


