# Learning Overspecified Gaussian Mixtures Exponentially Fast with the EM Algorithm

Zhenisbek Assylbekov[1,2], Alan Legg[1], and Artur Pak[2,3]

[1] Department of Mathematical Sciences, Purdue University Fort Wayne, Fort Wayne IN, USA
{zassylbe,leggar01}@pfw.edu
[2] Department of Mathematics, Nazarbayev University, Astana, Kazakhstan
[3] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
artur.pak@nu.edu.kz

**Abstract.** We investigate the convergence properties of the EM algorithm when applied to overspecified Gaussian mixture models—that is, when the number of components in the fitted model exceeds that of the true underlying distribution. Focusing on a structured configuration where the component means are positioned at the vertices of a regular simplex and the mixture weights satisfy a non-degeneracy condition, we demonstrate that the population EM algorithm converges exponentially fast in terms of the Kullback-Leibler (KL) distance. Our analysis leverages the strong convexity of the negative log-likelihood function in a neighborhood around the optimum and utilizes the Polyak-Łojasiewicz inequality to establish that an $\epsilon$-accurate approximation is achievable in $O(\log(1/\epsilon))$ iterations. Furthermore, we extend these results to a finite-sample setting by deriving explicit statistical convergence guarantees. Numerical experiments on synthetic datasets corroborate our theoretical findings, highlighting the dramatic acceleration in convergence compared to conventional sublinear rates. This work not only deepens the understanding of EM's behavior in overspecified settings but also offers practical insights into initialization strategies and model design for high-dimensional clustering and density estimation tasks.

**Keywords:** Overspecification · Gaussian Mixtures · Expectation-Maximization.

## 1 Introduction and Main Results

Let $Z_1, \ldots, Z_n$ be a random sample from the standard $d$-variate normal distribution $\mathcal{N}_d(0, I)$, where $0 \in \mathbb{R}^d$ is the mean vector, and $I \in \mathbb{R}^{d\times d}$ is the identity covariance matrix. We aim to fit a $k$-component Gaussian mixture model of the form

$$\pi_1 \cdot \mathcal{N}_d(\mu_1, I) + \ldots + \pi_k \cdot \mathcal{N}_d(\mu_k, I) \tag{1}$$

to this sample. When $k \geq 2$, this setting is known as *overspecification*, meaning the fitted model contains more mixture components than the true data-generating process. We assume the location parameters $\mu = (\mu_1^\top, \ldots, \mu_k^\top)^\top$ are unknown, while the mixture weights $(\pi_1, \ldots, \pi_k)$ are fixed and satisfy $\pi_j > 0$ and $\sum_{j=1}^k \pi_j = 1$.

Let $f(x;\mu)$ denote the probability density function of the mixture defined in (1). The maximum likelihood estimator (MLE) of $\mu$ is given by

$$\hat{\mu} \in \arg\max_{\mu} \frac{1}{n} \sum_{i=1}^{n} \log f(Z_i;\mu). \tag{2}$$

For $k \neq 1$, a closed-form solution for $\hat{\mu}$ does not exist. Instead, (2) is typically solved using iterative methods such as the Expectation-Maximization (EM) algorithm [7]. However, since the log-likelihood function in (2) is non-concave, iterative methods generally do not guarantee convergence to the global optimum.

Recent studies have analyzed the behavior of EM in overspecified settings. Dwivedi et al. [8, 9] examined the case $k = 2$, differentiating between balanced mixtures ($\pi_1 = \pi_2 = 1/2$) and unbalanced mixtures ($\pi_1 \neq \pi_2$). Assuming symmetric means ($\mu_1 = -\mu_2$), they showed that in the unbalanced case, the population EM[4] algorithm requires $O(\log(1/\epsilon))$ steps to obtain an $\epsilon$-accurate estimate of the parameter $\mu^* = 0$. In contrast, for balanced mixtures, the algorithm needs $\Theta(\log(1/\epsilon)/\epsilon^2)$ steps, making it exponentially slower.

Xu et al. [17] investigated the behavior of *gradient EM*[5] in the population setting for general $k$. Their results show that gradient EM exhibits a slow convergence rate, requiring $O(1/\epsilon^2)$ iterations to approximate the $k$-component Gaussian mixture (1) to $\mathcal{N}_d(0, I)$ within an accuracy $\epsilon$ in the KL metric. Their work imposes no assumptions on the balance of the mixture weights or the arrangement of Gaussian component centers. From this perspective, their result is more general. However, as demonstrated by Dwivedi et al. [9], in certain overspecified cases, the EM algorithm can achieve exponential convergence. This motivates the following question:

*When learning a mixture of $k$ Gaussians from $\mathcal{N}_d(0, I)$ data, does there exist a configuration of component centers and mixture weights such that the EM algorithm converges exponentially fast?*

Our answer to this question is affirmative, and we present it in the form of the following theorem.

**Theorem 1.** *Let $R \in \mathbb{R}^{d\times d}$ be an orthogonal matrix such that for any nonzero $\theta \in \mathbb{R}^d$, the points*

$$\mu_j(\theta) = R^{j-1}\theta, \quad \text{for } j = 1, \ldots, k.$$

*form the vertices of a regular $(k-1)$-simplex in $\mathbb{R}^d$, $d \geq k-1$, centered at the origin. Consider the $k$-component Gaussian mixture*

$$\mathcal{G}(\theta) := \pi_1 \cdot \mathcal{N}_d(\mu_1(\theta), I) + \pi_2 \cdot \mathcal{N}_d(\mu_2(\theta), I) + \cdots + \pi_k \cdot \mathcal{N}_d(\mu_k(\theta), I),$$

*where the mixture weights $\pi_1, \ldots, \pi_k$ are fixed, positive, satisfy $\sum_{j=1}^{k} \pi_j = 1$, and their discrete Fourier transform has no zero entries. This mixture is fitted to the standard*

[4] Population EM assumes access to the true data-generating distribution, allowing updates to be computed as exact expectations, free from sampling variability.

[5] Gradient EM replaces the M-step of the Expectation-Maximization algorithm with a single gradient ascent step on the Q-function.

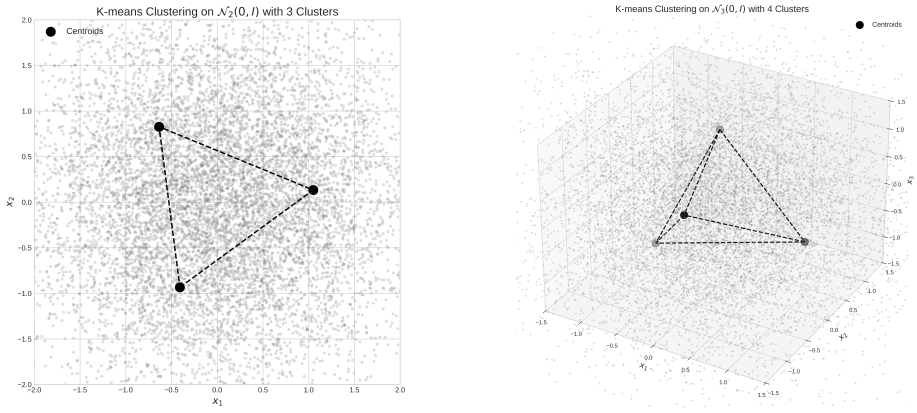
*Gaussian distribution $\mathcal{N}(0, I)$ using the Population EM algorithm. Let $\theta_t$ denote the parameter value at iteration $t$. Then there exists $\gamma > 0$ such that the following holds:*

$$D_{KL}[\mathcal{N}(0, I) \parallel \mathcal{G}(\theta_t)] \leq \kappa^t D_{KL}[\mathcal{N}(0, I) \parallel \mathcal{G}(\theta_0)],$$

*for $\theta_0$ satisfying $\|\theta_0\| \leq \gamma$ and for some constant $\kappa \in (0, 1)$.*

At first glance, the choice of placing Gaussian component centers at the vertices of a regular $(k-1)$-simplex may seem arbitrary. However, this configuration naturally arises in the context of Gaussian mixture learning.

A common approach to initializing Gaussian mixture components in the EM algorithm is via Lloyd's variant of the $k$-means algorithm [13]. We can show that the vertices of a regular $(k-1)$-simplex (with a particular radius) form a fixed point of Lloyd's algorithm when applied to $\mathcal{N}(0, I)$ at the population level (Section 2). Figure 1 illustrates this for finite samples and $k = 2, 3$. This suggests that the regular $(k-1)$-simplex




(a) $k = 3$ on $\mathcal{N}_2(0, I)$ (b) $k = 4$ on $\mathcal{N}_3(0, I)$

Fig. 1: K-means clustering on standard Gaussian data. *Left:* 10,000 samples in $\mathbb{R}^2$ are clustered into 3 groups; centroids (black markers) are connected by dashed lines to form a near-equilateral triangle. *Right:* 10,000 samples in $\mathbb{R}^3$ are clustered into 4 groups; centroids (black markers) connected by dashed lines approximate a regular tetrahedron.

is a natural initialization choice for the EM algorithm when learning an overspecified Gaussian mixture from data generated by a single Gaussian.

Following Xu et al. [17], we focus on the convergence of the fitted distribution to the true distribution in the KL metric rather than the convergence of the parameters to zero in the Euclidean metric, as studied by Dwivedi et al. [9]. However, our analysis fundamentally differs from both works. We find that the expected negative log-likelihood function is strongly convex in the neighborhood of the optimum and satisfies the so-called *Polyak-Łojasiewicz* inequality [1, 14]. This significantly simplifies the analysis of the convergence of the KL distance between the fitted model and the true distribution.

An immediate consequence of Theorem 1 is that the Population EM algorithm requires $O(\log(1/\epsilon))$ steps to approximate the mixture $\mathcal{G}(\theta)$ to $\mathcal{N}(0, I)$ within $\epsilon$ in the KL metric. This is exponentially faster than the general result of Xu et al. [17]. Moreover, by leveraging the now-standard approach of Balakrishnan et al. [3], we can translate the fast convergence of the population EM into the following finite-sample guarantee for the sample-based EM algorithm.

**Theorem 2.** *Under the assumptions of Theorem 1 on the structure of the Gaussian mixture, there exists $\gamma > 0$ such that for any initialization $\theta_0$ with $\|\theta_0\| \leq \gamma$, the EM algorithm produces a sequence of parameter estimates $\hat{\theta}_t$ satisfying*

$$D_{KL}\left[\mathcal{N}(0, I) \parallel \mathcal{G}(\hat{\theta}_T)\right] \leq c_1 \|\theta_0\|^2 \frac{\log(1/\delta)}{n}, \tag{3}$$

*for $T \geq c_2 \log \frac{n}{\log(1/\delta)}$ with probability at least $1 - \delta$.*

The proof of Theorem 2 is based on a perturbation bound that relates the sample-based EM operator to its population-level counterpart (Lemma 8 in Section C). In turn, the proof of the latter utilizes standard arguments to derive Rademacher complexity bounds.

The theoretical insights presented above are supported by our numerical experiments. In particular, Figure 2 demonstrates the exponential decay of the KL divergence over EM iterations under various mixture weight configurations, while Figure 3 reveals how the divergence decreases as the sample size increases. Together, these figures provide an intuitive visualization of the convergence dynamics and statistical guarantees established by our analysis.

To summarize, our work makes the following key contributions:

- We demonstrate that the EM algorithm can achieve *exponential convergence* in the KL metric when learning an overspecified mixture of $k$ Gaussian components under a specific structured configuration of mixture centers and weights. This contrasts with prior work [17], which establishes only sublinear convergence rates in general settings.
- We develop a novel analytical framework based on the *Polyak-Łojasiewicz inequality*, leveraging the strong convexity of the expected negative log-likelihood function near the optimum. This significantly simplifies the convergence analysis compared to previous approaches.
- We establish an explicit *finite-sample guarantee* for learning an overspecified mixture of $k$ Gaussians with the EM algorithm.

These contributions provide new insights into the role of mixture structure in the efficiency of EM and identify settings where the algorithm achieves fast convergence rates.

*Notation.* Lowercase letters ($x$) denote vectors in $\mathbb{R}^d$, uppercase letters ($A$, $X$) denote matrices and random vectors. The Euclidean norm is denoted by $\|x\| := \sqrt{x^\top x}$. We denote $\{1, 2, \ldots, k\}$ with $[k]$.

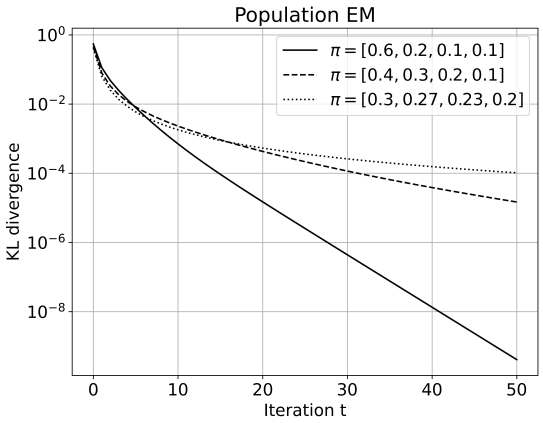



Fig. 2: Convergence of Population EM: The plot shows the evolution of the KL divergence versus the number of EM iterations for three different sets of mixture weights. The curves correspond to varying levels of imbalance illustrating how the choice of weights influences convergence speed.

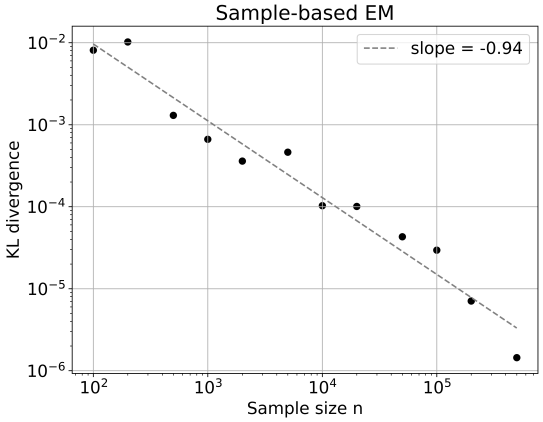



Fig. 3: Sample-Based EM Performance: The figure plots the final KL divergence against the sample size $n$ on a log-log scale. It demonstrates how increasing the number of samples improves the accuracy of the EM estimate by reducing the divergence between the fitted mixture and the true $\mathcal{N}(0, I)$ distribution

The probability density function of $Z \sim \mathcal{N}(0, I)$, where $I$ is a $d \times d$ identity matrix, is denoted by $\phi(z)$. The cumulative distribution function of $Z \sim \mathcal{N}(0, 1)$ is denoted by $\Phi(z)$.

Given $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}_+$, we write $f \precsim g$ if there exist $x_0 \in \mathbb{R}$, $c \in \mathbb{R}_+$ such that for all $x > x_0$ we have $|f(x)| \leq cg(x)$. When $f : \mathbb{R} \to \mathbb{R}_+$, we write $f \asymp g$ if $f \precsim g$ and $g \precsim f$. We use $c$, $c_1$, $c_2$, etc. to denote some universal constants (which might change in value each time they appear).

**Related Work**

Research on the Expectation-Maximization (EM) algorithm and its convergence behavior in Gaussian mixture models has advanced rapidly. Balakrishnan et al. [3] introduced a framework to delineate the region of convergence in terms of distribution parameters. Their work contrasted a population-level analysis with the sample-based implementation commonly used in practice, focusing specifically on the well-specified case of $k = 2$ components and addressing both balanced and unbalanced scenarios.

Within this context, significant effort has been devoted to developing initialization strategies that guarantee convergence to the global optimum. Klusowski and Brinda [12] demonstrated that local convergence can occur over a broader region than previously identified for the two-component case, while Zhao et al. [19] investigated how initialization affects mixtures with an arbitrary number of well-separated components. In addition, Daskalakis et al. [6] provided global convergence guarantees for a two-component model with symmetrically positioned mean vectors. For mixtures with $k$ well-separated components, Segol and Nadler [16] proved that convergence is assured even when the algorithm is initialized near the midpoint between clusters, refining the estimation error bounds and extending the analysis to Gradient EM—a variant of the classical EM algorithm. Moreover, Yan et al. [18] further analyzed the convergence rate and local contraction radius of Gradient EM for an arbitrary number of mixture components.

Another major line of inquiry has focused on model misspecification. Dwivedi et al. [10] examined an underspecified scenario, where a two-component Gaussian mixture is fitted to data generated by a three-component mixture, and characterized the resulting bias while also exploring the influence of initialization on convergence. The benefits of overspecified mixture models have been recognized by Dwivedi et al. [9], Dwivedi et al. [8], Chen et al. [4], and others. In particular, Dwivedi et al. [9] and Dwivedi et al. [8] studied the case of fitting two Gaussian components to data from a single Gaussian distribution. They compared balanced and unbalanced scenarios, demonstrating that in sample-based EM the unbalanced case converges at a statistical rate of $O(1/\sqrt{n})$, in contrast to $O(\sqrt[4]{1/n})$ for the balanced case when estimating mean vectors under both known and estimated isotropic covariance structures. They further showed that the algorithmic convergence rate is exponentially faster in the unbalanced setting.

Bayesian approaches to model overspecification, as discussed in [15], have revealed that the estimated mixture weights can vary greatly, often causing some components to become redundant and allowing for model refinement by discarding those with very small weights. In addition, Chen et al. [4] found that even spurious local minima of the negative log-likelihood retain structural information that is valuable for identifying

component means, highlighting the advantages of overspecification over underspecification—a contrast often described as "many-fit-one" versus "one-fit-many." Furthermore, Dasgupta and Schulman [5] proposed a method for finite mixture overspecification by recommending that models be deliberately initialized with $\frac{\log(k)}{w_{\min}}$ clusters, where $w_{\min}$ denotes the smallest weight, to substantially accelerate convergence.

While much of the literature has focused on convergence in terms of distribution parameters, investigations measuring the quality of fit using the Kullback-Leibler (KL) divergence are relatively few. Ghosal and van der Vaart [11] derived a statistical convergence rate of $(\log n)^{\kappa}/\sqrt{n}$ in Hellinger distance, which translates to a lower bound of $(\log n)^{2\kappa}/n$ in KL divergence; however, their analysis was confined to well-specified models and did not consider algorithmic factors. Dwivedi et al. [10] also employed KL divergence in the context of underspecified mixtures, but, to our knowledge, the use of KL divergence in overspecified mixtures was first explored by Xu et al. [17]. They obtained KL divergence bounds for the population version of Gradient EM applied to a $k$-component mixture with known variances. In contrast, our work extends these results by analyzing both population and sample-based EM under a structured configuration of mixture centers and weights, and importantly, we establish an *exponentially faster* algorithmic convergence rate in KL divergence than that reported by Xu et al. [17].

## 2 Initialization with $k$-means

Initialization is a critical step in the Expectation–Maximization (EM) algorithm, particularly in overspecified settings where the number of mixture components exceeds the true number. A common strategy is to first run the $k$–means algorithm (i.e., Lloyd's algorithm) on the data and then use the resulting cluster centers to initialize the EM algorithm.

When the data are generated from a single Gaussian distribution $\mathcal{N}(0, I)$, one observes that Lloyd's algorithm exhibits a natural fixed–point property under a symmetric configuration. In particular, consider initializing the $k$ centers at the vertices of a regular $(k-1)$–simplex centered at the origin. That is, let

$$\mu_i = r\, v_i, \quad i = 1, \ldots, k,$$

where the vectors $v_1, \ldots, v_k \in \mathbb{R}^d$ (with $d \geq k-1$) are unit vectors forming the vertices of a regular simplex, and $r > 0$ is a scaling factor.

A key observation is that the Voronoi partition induced by these centers depends only on the directions $v_i$ and not on the scalar $r$. Consequently, the conditional expectations computed in the Lloyd update—i.e., the new centers—are also determined solely by the angular configuration. In fact, one may show that the Lloyd update maps the configuration to

$$\mu_i' = R_0\, v_i, \quad i = 1, \ldots, k,$$

where $R_0 > 0$ is determined by the radial integrals of the Gaussian density. Thus, the fixed–point condition $\mu_i' = \mu_i$ for all $i$ is equivalent to choosing $r = R_0$.

This fixed–point property suggests that initializing the EM algorithm with a regular simplex (properly scaled) is natural in the context of overspecified Gaussian mixtures.

In the proof of the following proposition (Section A), we rigorously analyze this phenomenon by first characterizing the Voronoi partition induced by a regular simplex and then proving that there exists a unique scaling $r > 0$ such that the configuration

$$\{\mu_i = r\, v_i : i = 1, \dots, k\}$$

remains invariant under the population–level Lloyd update.

**Proposition 1.** *Let $d \geq k - 1$, and suppose that*

$$v_1, \dots, v_k \in \mathbb{R}^d$$

*are unit vectors that form the vertices of a regular simplex in some $(k-1)$–dimensional subspace of $\mathbb{R}^d$,*

$$\|v_i\| = 1, \quad \textit{for } i = 1, \dots, k, \quad \textit{with} \quad \sum_{i=1}^{k} v_i = 0,$$

*with the pairwise inner products being constant for $i \neq j$. For any $r > 0$, define centers*

$$\mu_i = r\, v_i, \quad i = 1, \dots, k,$$

*and let the Voronoi cells be*

$$V_i = \{\, x \in \mathbb{R}^d : \|x - \mu_i\| \leq \|x - \mu_j\| \textit{ for all } j \neq i \,\}.$$

*Then there exists a unique $r > 0$ such that if one performs the population-level Lloyd update*

$$\mu_i' = \frac{\displaystyle\int_{V_i} x\, \phi(x)\, dx}{\displaystyle\int_{V_i} \phi(x)\, dx},$$

*one obtains $\mu_i' = \mu_i$ for all $i = 1, \dots, k$. That is, the configuration*

$$\{\mu_i = r\, v_i : i = 1, \dots, k\}$$

*is a fixed point of Lloyd's algorithm.*

## 3 Population-Level Analysis

We begin by analyzing the behavior of the so-called *population EM*, a theoretical construct that isolates algorithmic complexity from sample complexity. Population EM assumes direct access to the data-generating distribution $\mathcal{N}(0, I)$ and, instead of maximizing the sample-based log-likelihood in (2), optimizes the population log-likelihood:

$$\mathcal{L}(\theta) := \mathop{\mathbb{E}}_{Z \sim \mathcal{N}(0,I)} [\log f(Z; \theta)]. \tag{4}$$

The algorithm proceeds iteratively by applying the following two steps:

– *Expectation step*: Given the current estimate $\theta_t$, compute the function

$$Q(\theta, \theta_t) := \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,I)} \left[ \sum_{j=1}^{k} w_j(Z;\theta_t) \log\left(\pi_j \cdot \phi\left(Z - R^{j-1}\theta\right)\right) \right],$$

where

$$w_j(Z;\theta_t) = \frac{\pi_j \cdot \phi(Z - R^{j-1}\theta_t)}{\sum_{\ell=1}^{k} \pi_\ell \cdot \phi(Z - R^{\ell-1}\theta_t)}.$$

– *Maximization step*: Update the parameters by solving the optimization problem:

$$\theta_{t+1} \in \arg\max_{\theta} Q(\theta, \theta_t).$$

In this specific case, where the population EM algorithm is used to fit the mixture (1) to $\mathcal{N}(0, I)$, the recurrence relations governing the parameter updates can be explicitly derived (Section B.1). The parameter updates follow the recursion $\theta_{t+1} = M(\theta_t)$, where

$$M(\theta) := \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,I)} \left[ \sum_{j=1}^{k} w_j(Z;\theta)(R^{j-1})^\top Z \right]. \tag{5}$$

The mapping $M(\theta)$ is referred to as the *population EM operator*. Notably, it is closely related to the population negative log-likelihood, as stated in the following equation (Section B.2):

$$\nabla_\theta[-\mathcal{L}(\theta)] = \theta - M(\theta). \tag{6}$$

Denote $L(\theta) := -\mathcal{L}(\theta)$. The equation (6) implies that

$$\theta_{t+1} = \theta_t - \nabla_\theta[L(\theta_t)],$$

which means that in the given setting, the EM algorithm is equivalent to gradient descent (GD) on $L(\theta)$ with a step size 1. This suggests that standard techniques used in the analysis of GD can be applied to study the convergence of the EM algorithm. One such technique is the Polyak-Łojasiewicz inequality, a sufficient condition for the exponential convergence of GD. We establish this property for $L(\theta)$ in the following lemma.

**Lemma 1 (Local PL Inequality).** *Let $\mathcal{L}(\theta)$ be the population log-likelihood function defined by* (4)*. Suppose $\pi_1, \ldots, \pi_k > 0$ are positive real numbers whose discrete Fourier transform has no zero entries. Then there exists $\delta > 0$ such that $L(\theta) := -\mathcal{L}(\theta)$ satisfies the following local Polyak–Łojasiewicz (PL) inequality in $\{\theta : \|\theta\| \le \delta\}$:*

$$\left\|\nabla L(\theta)\right\|^2 \;\geq\; \lambda_{\min}\Big(L(\theta) \;-\; L(0)\Big), \tag{7}$$

*where $\lambda_{\min} \le 1$ is the smallest eigenvalue of $\nabla^2 L(0)$.*

A key step in establishing the local Polyak–Łojasiewicz (PL) inequality around $\theta^* = 0$ is to show that the Hessian $\nabla^2 L(\theta)$ remains positive definite in a sufficiently small neighborhood of $\theta^* = 0$. Concretely, we need the Jacobian of the EM operator at $\theta^* = 0$ to have spectral properties that ensure strong convexity of the population negative log-likelihood $L$.

In our setup, this boils down to proving that the matrix $A = \sum_{j=1}^{k} \pi_j R^{j-1}$ is invertible (cf. Lemmas 2 and 3 in Section 5), since one can then show $I - \frac{\partial M}{\partial \theta}(0) = A^\top A$ is positive definite. The invertibility of $A$ follows from the assumption that the $\pi_j$'s have a discrete Fourier transform with no zero entries. Intuitively, if the discrete Fourier transform of $\{\pi_j\}$ vanished at one of the $k$-th roots of unity, then certain "rotational symmetries" in the update equations would cause degeneracies, preventing $A$ from being invertible. By ruling out such degeneracies, the condition $\widehat{\pi}(\ell) \neq 0$ for all $\ell$ guarantees the necessary full rank of $A$.

Under these conditions, the Hessian $\nabla^2 L(\theta)$ remains uniformly positive definite in a neighborhood of $\theta^* = 0$, which yields the strong convexity of $L$ around $\theta^* = 0$. From strong convexity, the usual argument then gives the local PL inequality (7) demonstrating the sharpness of the landscape near the stationary point $\theta^* = 0$.

Since the local PL inequality plays a central role in our convergence analysis, we present the proof of Lemma 1 in the main text (Section 5) to ensure the core argument remains transparent. The proofs of the remaining supporting lemmas are deferred to the Appendix.

We are ready to prove the exponential decay of the KL divergence between the true distribution and the sequence of fitted mixtures.

*Proof (Proof of Theorem 1).* We start by noting that

$$\mathrm{D}_{\mathrm{KL}}\big[\mathcal{N}(0, I) \,\big\|\, \mathcal{G}(\theta_t)\big] \;=\; L(\theta_t) \;-\; L(0).$$

Equation (6) implies that the Hessian of $L$ is given by

$$\nabla^2 L(\theta) \;=\; I \;-\; \frac{\partial M}{\partial \theta}.$$

Furthermore, we can show (see Lemma 2 in Section 5) that at $\theta^* = 0$, we have

$$\nabla^2 L(0) \;=\; AA^\top, \quad \text{where} \quad A \;:=\; \sum_{j=1}^{k} \pi_j \, R^{j-1}.$$

Since $R$ is an orthogonal matrix, its eigenvalues are among the $k$-th roots of unity $\{e^{2\pi i \ell / k}\}_{\ell=0}^{k-1}$, implying $\|R\|_{\mathrm{op}} = 1$. Hence, by the triangle inequality,

$$\|\nabla^2 L(0)\|_{\mathrm{op}} \;\le\; \left(\sum_{j=1}^{k} \pi_j \|R^{j-1}\|_{\mathrm{op}}\right)^2 \;=\; \left(\sum_{j=1}^{k} \pi_j\right)^2 \;=\; 1.$$

By smoothness of $L(\theta)$, there is therefore a neighborhood of $\theta^* = 0$ in which $\|\nabla^2 L(\theta)\|_{\mathrm{op}} \le 3/2$. Consequently, for $\theta$ and $\theta'$ in that neighborhood,

$$L(\theta') \;\le\; L(\theta) \;+\; \nabla L(\theta)^\top (\theta' - \theta) \;+\; \tfrac{3}{4} \, \|\theta' - \theta\|^2.$$

In particular, applying this to $\theta_{t+1}$ and $\theta_t$ yields

$$\begin{aligned} L(\theta_{t+1}) \;&\le\; L(\theta_t) \;+\; \nabla L(\theta_t)^\top \big(\theta_{t+1} - \theta_t\big) \;+\; \tfrac{3}{4}\,\|\theta_{t+1} - \theta_t\|^2 \\ &=\; L(\theta_t) \;+\; \nabla L(\theta_t)^\top \big(M(\theta_t) - \theta_t\big) \;+\; \tfrac{3}{4}\,\|M(\theta_t) - \theta_t\|^2 \\ &=\; L(\theta_t) \;-\; \|\nabla L(\theta_t)\|^2 \;+\; \tfrac{3}{4}\,\|\nabla L(\theta_t)\|^2 \\ &=\; L(\theta_t) \;-\; \tfrac{1}{4}\,\|\nabla L(\theta_t)\|^2. \end{aligned}$$

Next, using the Polyak–Lojasiewicz inequality (7), we obtain

$$L(\theta_{t+1}) \;\le\; L(\theta_t) - \;\tfrac{\lambda_{\min}}{4}\,\big(L(\theta_t) - L(0)\big).$$

Subtracting $L(0)$ from both sides gives

$$L(\theta_{t+1}) - L(0) \;\le\; \Big(1 - \tfrac{\lambda_{\min}}{4}\Big)\,\big(L(\theta_t) - L(0)\big).$$

Applying this inequality recursively completes the proof.

## 4 Finite-Sample Analysis

When the sample-based averaged log-likelihood in (2) is maximized via the EM algorithm, the parameter updates can be expressed explicitly by replacing the expectation $\mathbb{E}$ in (5) with the empirical average over the sample:

$$\begin{aligned} &\hat{\theta}_{t+1} = M_n(\hat{\theta}_t), \\ \text{where}\quad &M_n(\theta) := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} w_j(Z_i;\theta)(R^{j-1})^\top Z_i. \end{aligned} \tag{8}$$

The following perturbation bound (Section C) relates the sample-based EM operator to its population-level counterpart:

$$\Pr\left[\sup_{\|\theta\|\le r} \|M_n(\theta) - M(\theta)\| \le cr\sqrt{\frac{d + \log(1/\delta)}{n}}\right] \ge 1 - \delta, \tag{9}$$

for any radius $r > 0$, threshold $\delta \in (0, 1)$, and sufficiently large $n$.

Due to the strict contractivity of the population EM operator in a neighborhood of $\theta^* = 0$ (Lemma 4 in Section 5) and the perturbation bound above, we can establish that the sequence of EM iterates $\hat{\theta}_t$ satisfies, with probability at least $1 - \delta$,

$$\|\hat{\theta}_T\| \precsim \|\theta_0\|\sqrt{\frac{\log(1/\delta)}{n}}, \tag{10}$$

for $T \succsim \log\left(\frac{n}{\log(1/\delta)}\right)$, provided that $\theta_0$ lies within the contraction neighborhood (see the proof of Theorem 2 in [3]).

With this, we are ready to establish our key result on convergence in KL distance for the finite-sample case.

*Proof (Proof of Theorem 2).* By the convexity of $L$ in a neighborhood of $\theta^* = 0$ (Lemma 5 in Section 5), we have

$$L(\hat{\theta}_t) - L(0) \leq \nabla L(\hat{\theta}_t)^\top \hat{\theta}_t. \tag{11}$$

From (6), it follows that

$$\nabla L(\hat{\theta}_t)^\top \hat{\theta}_t = \|\hat{\theta}_t\|^2 - [M(\hat{\theta}_t)]^\top \hat{\theta}_t \precsim \|\hat{\theta}_t\|^2, \tag{12}$$

where we used the contraction property of $M$ near $\theta = 0$ (Corollary 1 in Section 5).

The theorem follows directly from (10), (11), and (12).

## 5 Proof of the Local PL Inequality

In this section, we prove the local Polyak–Lojasiewicz (PL) inequality for the negative log-likelihood $L(\theta)$ of our overspecified Gaussian mixture model. By analyzing the Jacobian of the population EM operator $M(\theta)$ at $\theta^* = 0$, we show that $M(\theta)$ is locally contractive, which implies that the Hessian of $L(\theta)$ is uniformly positive definite near $\theta^* = 0$.

The subsequent lemmas establish these properties and lead directly to the local PL inequality, ensuring the exponential convergence of the EM algorithm in terms of the KL divergence.

**Lemma 2.** *Let $M(\theta)$ be the EM operator defined by* (5)*. Then the Jacobian of $M(\theta)$ at $\theta^* = 0$ is given by $\frac{\partial M}{\partial \theta}(0) = I - \left(\sum_{j=1}^k \pi_j R^{j-1}\right)\left(\sum_{j=1}^k \pi_j R^{j-1}\right)^\top$.*

*Proof.* Let $S(\theta, Z) = \sum_{\ell=1}^k \pi_\ell \exp\big((R^{\ell-1}\theta)^\top Z\big)$. Then

$$w_j(Z;\theta) = \frac{\pi_j \exp\big((R^{j-1}\theta)^\top Z\big)}{S(\theta, Z)}.$$

At $\theta^* = 0$, each exponential term is $\exp(0) = 1$, so $S(0, Z) = \sum_{\ell=1}^k \pi_\ell = 1$, $w_j(Z;0) = \pi_j$. Since

$$\nabla_\theta \left((R^{j-1}\theta)^\top Z\right) = (R^{j-1})^\top Z,$$

$$\nabla_\theta S(\theta, Z) = \sum_{\ell=1}^k \pi_\ell \exp\big((R^{\ell-1}\theta)^\top Z\big)\left[(R^{\ell-1})^\top Z\right],$$

using the quotient rule for $\nabla_\theta w_j(Z;\theta)$, and then evaluating it at $\theta^* = 0$, we get

$$\nabla_\theta w_j(Z;\theta)\big|_{\theta^*=0} = \pi_j\Big[(R^{j-1})^\top Z - \sum_{\ell=1}^k \pi_\ell\,(R^{\ell-1})^\top Z\Big].$$

Define $g(\theta, Z) = \sum_{j=1}^k w_j(Z;\theta)\,(R^{j-1})^\top Z$. Then

$$\frac{\partial g}{\partial \theta}(\theta, Z) = \sum_{j=1}^k (R^{j-1})^\top Z\left[\nabla_\theta w_j(Z;\theta)\right]^\top.$$

At $\theta^* = 0$,

$$\frac{\partial g}{\partial \theta}(\theta, Z)\big|_{\theta^*=0} \;=\; \sum_{j=1}^{k} \pi_j (R^{j-1})^\top Z \left[ (R^{j-1})^\top Z \;-\; \sum_{\ell=1}^{k} \pi_\ell \, (R^{\ell-1})^\top Z \right]^\top .$$

Then the sought Jacobian of $M(\theta)$ at $\theta^* = 0$ is

$$\frac{\partial M}{\partial \theta}(0) \;=\; \mathbb{E}_Z\big[\nabla_\theta g(0, Z)\big].$$

Since $Z \sim \mathcal{N}_d(0, I)$, we have $\mathbb{E}[ZZ^\top] = I$. Each $R^{j-1}$ is orthogonal, hence

$$\mathbb{E}\Big[(R^{j-1})^\top Z \, Z^\top \, R^{j-1}\Big] \;=\; (R^{j-1})^\top \, I \, R^{j-1} \;=\; I.$$

Collecting terms, the result is

$$\frac{\partial M}{\partial \theta}(0) \;=\; I \;-\; \Big(\sum_{j=1}^{k} \pi_j \, R^{j-1}\Big)\Big(\sum_{j=1}^{k} \pi_j \, R^{j-1}\Big)^\top,$$

which completes the proof.

**Lemma 3.** *Let $R \in \mathbb{R}^{d\times d}$ be a (real) matrix whose eigenvalues lie among the $k$-th roots of unity (except 1), i.e., $\operatorname{spec}(R) \;\subseteq\; \big\{ e^{\,i\frac{2\pi\ell}{k}} \;:\; \ell = 1, 2, \ldots, k-1 \big\}$. Let $\pi_1, \ldots, \pi_k > 0$ be positive real numbers whose discrete Fourier transform $\widehat{\pi}(\ell) \;=\; \sum_{j=0}^{k-1} \pi_{j+1} \, e^{\,i\frac{2\pi\ell}{k} j}$, $\ell = 0, 1, \ldots, k-1$, has no zero entries (i.e. $\widehat{\pi}(\ell) \neq 0$ for all $\ell$). Define the matrix $A \;:=\; \sum_{j=1}^{k} \pi_j \, R^{j-1}$. Then $A$ is invertible.*

*Proof.* Since all eigenvalues of $R$ lie among the $k$-th roots of unity (except 1), we can work over $\mathbb{C}$ and bring $R$ into a Jordan (or block-diagonal) form. Concretely, there exists an invertible matrix $V \in \mathbb{C}^{d\times d}$ such that $R \;=\; V \, \Lambda \, V^{-1}$, where $\Lambda$ is block-diagonal and each block corresponds to an eigenvalue of the form $e^{\,i\frac{2\pi\ell}{k}}$ (with $1 \le \ell \le k-1$). In particular, $R^{j-1} \;=\; V \, \Lambda^{j-1} \, V^{-1}$ for all $j = 1, \ldots, k$. Thus we can rewrite $A$ as

$$A \;=\; \sum_{j=1}^{k} \pi_j \, R^{j-1} \;=\; \sum_{j=1}^{k} \pi_j \big(V \, \Lambda^{j-1} \, V^{-1}\big) \;=\; V\Big(\sum_{j=1}^{k} \pi_j \, \Lambda^{j-1}\Big) V^{-1}.$$

Since $V$ is invertible, $A$ is invertible if and only if $\sum_{j=1}^{k} \pi_j \, \Lambda^{j-1}$ is invertible.

Now, $\Lambda$ is block-diagonal with Jordan blocks corresponding to eigenvalues $\lambda \in \big\{ e^{\,i\frac{2\pi\ell}{k}} \;:\; \ell = 1, \ldots, k-1 \big\}$. Consider a single eigenvalue $\lambda$. The diagonal entry of the diagonal block of $\sum_{j=1}^{k} \pi_j \, \Lambda^{j-1}$ is $\sum_{j=1}^{k} \pi_j \, \lambda^{j-1} \;=\; \sum_{j=0}^{k-1} \pi_{j+1} \, \lambda^{j}$. Since $\lambda^j = e^{\,i\frac{2\pi\ell}{k} j}$ for some $\ell \in \{1, \ldots, k-1\}$, this sum is precisely the discrete Fourier transform of $(\pi_1, \ldots, \pi_k)$: $\sum_{j=0}^{k-1} \pi_{j+1} \, e^{\,i\frac{2\pi\ell}{k} j} \;=\; \widehat{\pi}(\ell)$. By hypothesis, $\widehat{\pi}(\ell) \neq 0$ for all $\ell = 0, \ldots, k-1$, hence each diagonal entry is nonzero. Therefore, every diagonal block of $\displaystyle\sum_{j=1}^{k} \pi_j \, \Lambda^{j-1}$ is invertible, so the entire block-diagonal matrix is invertible.

**Lemma 4.** *Under the conditions of Lemmas 2 and 3, the matrix $I - \frac{\partial M}{\partial \theta}(0)$ is positive definite.*

*Proof.* From Lemma 2, the Jacobian of the EM operator at $\theta^* = 0$ is given by $\frac{\partial M}{\partial \theta}(0) = I - AA^\top$, where $A = \sum_{j=1}^{k} \pi_j R^{j-1}$. Rearranging this equation, we obtain $I - \frac{\partial M}{\partial \theta}(0) = AA^\top$. By Lemma 3, the matrix $A$ is invertible. Since $AA^\top$ is the product of $A$ and $A^\top$, it follows that $AA^\top$ is symmetric and positive definite. To see this, note that for any non-zero vector $x \in \mathbb{R}^d$, $x^\top AA^\top x = \left(A^\top x\right)^\top \left(A^\top x\right) = \|A^\top x\|^2 > 0$, since $A$ is invertible and thus $A^\top x \neq 0$ for $x \neq 0$. Therefore, the matrix $I - \frac{\partial M}{\partial \theta}(0) = AA^\top$ is positive definite.

**Corollary 1.** *All eigenvalues of $\frac{\partial M}{\partial \theta}(0)$ lie strictly below 1, which in turn implies that $M$ is a contraction near $\theta^* = 0$.*

**Lemma 5.** *Let $\mathcal{L}(\theta)$ be the population log-likelihood function defined by* (4)*. Suppose $\pi_1, \ldots, \pi_k > 0$ are positive real numbers whose discrete Fourier transform has no zero entries. Then there exists $\delta > 0$ such that $L$ is strongly convex in $\{\theta : \|\theta\| \leq \delta\}$:*

*Proof.* Since $\nabla L(\theta) = \theta - M(\theta)$, its Hessian is given by $\nabla^2 L(\theta) = I - \frac{\partial M}{\partial \theta}(\theta)$. By Lemma 4, $S := I - \frac{\partial M}{\partial \theta}(0)$ is positive definite. Let $\lambda_{\min} = \lambda_{\min}(S) > 0$ be the smallest eigenvalue of $S$.

Next, by continuity of $\frac{\partial M}{\partial \theta}(\theta)$ at $\theta = 0$, for any $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\|\theta\| < \delta \quad \Longrightarrow \quad \left\| \frac{\partial M}{\partial \theta}(\theta) - \frac{\partial M}{\partial \theta}(0) \right\|_{\text{op}} < \varepsilon,$$

where $\|\cdot\|_{\text{op}}$ denotes the operator norm.

Choose $\varepsilon = \frac{1}{2}\lambda_{\min}$. Then for $\|\theta\| < \delta$, $\left\| \frac{\partial M}{\partial \theta}(\theta) - \frac{\partial M}{\partial \theta}(0) \right\|_{\text{op}} < \frac{1}{2}\lambda_{\min}$. Thus, for any vector $v \in \mathbb{R}^d$ with $\|v\| = 1$,

$$\begin{aligned} v^\top \Big(I - \frac{\partial M}{\partial \theta}(\theta)\Big) v &= v^\top \big(I - \tfrac{\partial M}{\partial \theta}(0)\big) v - v^\top \Big(\tfrac{\partial M}{\partial \theta}(\theta) - \tfrac{\partial M}{\partial \theta}(0)\Big) v \\ &\geq v^\top \Big(I - \tfrac{\partial M}{\partial \theta}(0)\Big) v - \left\| \tfrac{\partial M}{\partial \theta}(\theta) - \tfrac{\partial M}{\partial \theta}(0) \right\|_{\text{op}} \\ &\geq \lambda_{\min} - \tfrac{1}{2}\lambda_{\min} = \tfrac{1}{2}\lambda_{\min}. \end{aligned}$$

Therefore, $I - \frac{\partial M}{\partial \theta}(\theta) \succeq \frac{1}{2}\lambda_{\min} I$ for all $\|\theta\| < \delta$. Since $\nabla^2 L(\theta) = I - \frac{\partial M}{\partial \theta}(\theta)$, we deduce $\nabla^2 L(\theta) \succeq \frac{1}{2}\lambda_{\min} I$ whenever $\|\theta\| < \delta$. Hence $L(\theta)$ is $\left(\frac{1}{2}\lambda_{\min}\right)$-strongly convex in the ball $\{\theta : \|\theta\| < \delta\}$.

*Proof (Proof of Lemma 1).* By Lemma 5, $L$ is $\frac{\lambda_{\min}}{2}$-strongly convex in a neighborhood of $\theta^* = 0$, i.e. there exists $\delta > 0$ such that for $\theta, \theta' \in \{\theta : \|\theta\| \leq \delta\}$

$$L(\theta') \geq L(\theta) + \nabla L(\theta)^T(\theta' - \theta) + \frac{\lambda_{\min}}{4}\|\theta' - \theta\|^2.$$

Minimizing both sides with respect to $\theta'$, we get

$$L(0) \geq L(\theta) - \frac{1}{\lambda_{\min}} \|\nabla L(\theta)\|^2.$$

Re-arranging the terms we have the PL inequality.

**Acknowledgments.** This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP27510283). Artur Pak's work was supported by Nazarbayev University under Faculty-development competitive research grants program for 2023-2025 Grant #20122022FD4131, PI R. Takhanov. Zhenisbek Assylbekov's work was supported by Purdue University Fort Wayne under Summer Research Grant Program 2024, and he would like to thank Igor Melnykov and Francesco Sica for useful discussions.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

# A Proof of Proposition 1

*Proof.* See the full version [2].

# B Population EM properties

## B.1 Population EM updates

We begin by providing additional details on the EM algorithm. It is convenient to represent the mixture distribution (1) using a latent categorical random variable $K$, which identifies the mixture components. Given the mixture weights $(\pi_1, \ldots, \pi_k)$, we assume that

$$\Pr[K = j] = \pi_j.$$

The conditional distribution of $X$ given $K = j$ is then defined as

$$(X \mid K = j) \sim \mathcal{N}_d(R^{j-1}\theta, I), \quad \text{for } j \in [k].$$

This specifies the joint distribution of the tuple $(X, K)$, ensuring that the marginal distribution of $X$ corresponds to the Gaussian mixture $\mathcal{G}(\theta)$ in (1). The Population EM algorithm maximizes the expected log-likelihood (4) through the following iterative steps:

– *E-step:* Given the current estimate $\theta_t$, compute the soft assignment of any $x \in \mathbb{R}^d$ to component $K = j$, i.e., evaluate the posterior probability:

$$w_j(x; \theta_t) = \frac{\pi_j \cdot \phi(x - R^{j-1}\theta_t)}{\sum_{\ell=1}^{k} \pi_\ell \cdot \phi(x - R^{\ell-1}\theta_t)}, \tag{13}$$

and use it to compute the $Q$-function:

$$Q(\theta, \theta_t) := \mathbb{E}_{Z \sim \mathcal{N}(0,I)} \left[ \sum_{j=1}^{k} w_j(Z; \theta_t) \log \left( \pi_j \cdot \phi(Z - R^{j-1}\theta) \right) \right].$$

– *M-step:* Update the parameter by solving the following optimization problem:

$$\theta_{t+1} \in \arg\max_{\theta} Q(\theta, \theta_t). \tag{14}$$

**Lemma 6.** *Let the population EM algorithm maximize the expected log-likelihood* (4)*. Then, the parameter updates follow the recursion $\theta_{t+1} = M(\theta_t)$, where*

$$M(\theta) := \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,I)} \left[ \frac{\sum_{j=1}^{k} \pi_j \cdot \exp\left((R^{j-1}\theta)^\top Z\right) (R^{j-1})^\top Z}{\sum_{\ell=1}^{k} \pi_\ell \cdot \exp\left((R^{\ell-1}\theta)^\top Z\right)} \right].$$

*Proof.* See the full version [2].

### B.2 Population EM operator and Log-likelihood

**Lemma 7.** *The population log-likelihood $\mathcal{L}(\theta)$ defined in* (4) *and the population EM operator $M(\theta)$ are related by the equation:*

$$\nabla_\theta[-\mathcal{L}(\theta)] = \theta - M(\theta).$$

*Proof.* See the full version [2].

## C Perturbation bound

**Lemma 8.** *There exist universal constants $c, c' > 0$ such that for any radius $r > 0$, confidence level $\delta \in (0,1)$, and sample size $n \geq c'\left(d + \log(1/\delta)\right)$, the following holds with probability at least $1 - \delta$:*

$$\sup_{\|\theta\|\leq r} \| M_n(\theta) - M(\theta)\| \leq cr\sqrt{\frac{d + \log(1/\delta)}{n}}.$$

*Proof.* See the full version [2].

## Bibliography

[1] Gradient methods for the minimisation of functionals. USSR Computational Mathematics and Mathematical Physics **3**(4), 864–878 (1963), ISSN 0041-5553

[2] Assylbekov, Z., Legg, A., Pak, A.: Learning overspecified gaussian mixtures exponentially fast with the em algorithm (2025), URL https://arxiv.org/abs/2506.11850

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

[4] Chen, Y., Song, D., Xi, X., Zhang, Y.: Local minima structures in gaussian mixture models. IEEE Transactions on Information Theory (2024)

[5] Dasgupta, S., Schulman, L.J.: A two-round variant of em for gaussian mixtures. In: Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, pp. 152–159 (2000)

[6] Daskalakis, C., Tzamos, C., Zampetakis, M.: Ten steps of em suffice for mixtures of two gaussians. In: Conference on Learning Theory, pp. 704–710, PMLR (2017)
[7] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**(1), 1–22 (1977)
[8] Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M., Jordan, M., Yu, B.: Sharp analysis of expectation-maximization for weakly identifiable models. In: International Conference on Artificial Intelligence and Statistics, pp. 1866–1876, PMLR (2020)
[9] Dwivedi, R., Ho, N., Khamaru, K., Wainwright, M.J., Jordan, M.I., Yu, B.: Singularity, misspecification and the convergence rate of EM. The Annals of Statistics **48**(6), 3161–3182 (2020)
[10] Dwivedi, R., Khamaru, K., Wainwright, M.J., Jordan, M.I., et al.: Theoretical guarantees for em under misspecified gaussian mixture models. Advances in Neural Information Processing Systems **31** (2018)
[11] Ghosal, S., van der Vaart, A.W.: Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. The Annals of Statistics **29**(5), 1233 – 1263 (2001)
[12] Klusowski, J.M., Brinda, W.: Statistical guarantees for estimating the centers of a two-component gaussian mixture by em. arXiv preprint arXiv:1608.02280 (2016)
[13] Lloyd, S.P.: Least squares quantization in PCM. IEEE Trans. Inf. Theory **28**(2), 129–136 (1982)
[14] Lojasiewicz, S.: A topological property of real analytic subsets. Coll. du CNRS, Les équations aux dérivées partielles **117**(87-89), 2 (1963)
[15] Rousseau, J., Mengersen, K.: Asymptotic Behaviour of the Posterior Distribution in Overfitted Mixture Models. Journal of the Royal Statistical Society Series B: Statistical Methodology **73**(5), 689–710 (08 2011)
[16] Segol, N., Nadler, B.: Improved convergence guarantees for learning gaussian mixture models by em and gradient em. Electronic journal of statistics **15**(2), 4510–4544 (2021)
[17] Xu, W., Fazel, M., Du, S.S.: Toward global convergence of gradient em for overparamterized gaussian mixture models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
[18] Yan, B., Yin, M., Sarkar, P.: Convergence of gradient em on multi-component mixture of gaussians. Advances in Neural Information Processing Systems **30** (2017)
[19] Zhao, R., Li, Y., Sun, Y.: Statistical convergence of the em algorithm on gaussian mixture models. Electronic Journal of Statistics **14**, 632 – 660 (2020)