NAZARBAYEV UNIVERSITY

MASTER THESIS

Population level analysis of convergence of the EM algorithm for overspecified mixtures

Author:
Artur PAK

Supervisor:
Dr. Rustem TAKHANOV
Co-Supervisor:
Dr. Zhenisbek ASSYLBEKOV

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

School of Sciences and Humanities

April 30, 2025

NAZARBAYEV UNIVERSITY

Abstract

School of Sciences and Humanities

Master of Science

Population level analysis of convergence of the EM algorithm for overspecified mixtures

by Artur PAK

This thesis analyzes the convergence properties of the Expectation-Maximization (EM) algorithm when applied to an overspecified Gaussian Mixture Model (GMM). Specifically, it examines the case where a two-component balanced GMM is fitted to data generated from a single Gaussian distribution. A population-level analysis establishes an upper bound of $\tilde{O}(1/t^2)$ on the Kullback-Leibler (KL) divergence between the learned and true distributions, where t is number of steps of EM algorithm. These theoretical findings are further validated through empirical experiments. This thesis contributes to a broader collaborative study (see Acknowledgments) titled *Convergence of the EM Algorithm in KL Distance for Overspecified Gaussian Mixtures*.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Rustem Takhanov, for his invaluable guidance and support throughout this research. His insights and encouragement have been instrumental in shaping this thesis.

I am also sincerely grateful to my co-supervisor and co-author, Dr. Zhenisbek Assylbekov, for his mentorship and significant contributions to this work. His expertise and feedback have been essential to the development of both the theoretical and experimental aspects of this study.

A special thanks to Dr. Alan Legg for his major theoretical contributions to our joint research, particularly in Parts 1 and 3 of the Lemma 2. His expertise in mathematical analysis has been crucial in refining our theoretical results.

I would also like to acknowledge Dr. Igor Melnykov for his detailed and insightful analysis of the related work, which greatly enriched this research.

Finally, I extend my heartfelt appreciation to Arman Bolatov, my co-author and friend, for his extensive help with experiments and visualizations. His efforts in designing and interpreting the graphs have been invaluable in illustrating our findings.

To all my collaborators, mentors, and friends, thank you for your support and contributions to this research.

Contents

Chapter 1

Introduction

Mixture models are among the most widely used approaches for modeling real-world data distributions, as natural data (e.g., text and images) often consist of multiple subpopulations. As an illustrative example, consider the AG News text classification dataset,[1] which contains news articles categorized by topic (world, sports, business, sci/tech, etc.). Figure 1.1 depicts BERT representations Devlin et al., 2019 of articles from the four largest classes, projected onto two dimensions using UMAP McInnes et al., 2018.

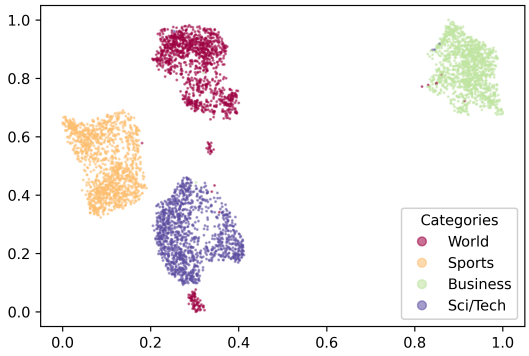

FIGURE 1.1: BERT representations of news articles from the AG News dataset, projected to two dimensions using UMAP. Colors denote different classes.

Within the 'World' category, the data distribution exhibits distinct subpopulations, suggesting that a mixture of multiple components could naturally model this distribution.

However, in practice, the number of components is often unknown. A common strategy is to *overspecify* the number of components, relying on algorithms such as expectation-maximization (EM) to yield a distribution close to the true one, which may have fewer components. This approach was examined by Dwivedi et al. (2020b), who analyzed the fitting of a two-component Gaussian mixture to data drawn from a single Gaussian in $\mathbb{R}^d$. They demonstrated that in the balanced case, the EM algorithm Dempster, Laird, and Rubin, 1977 requires $\widetilde{O}(\sqrt{n/d})$ iterations to estimate the location parameters within $O(\sqrt[4]{d/n})$ of the true parameters (in Euclidean distance), assuming a known variance. Later, Dwivedi et al. (2020a) relaxed this assumption for a balanced mixture and showed that both the algorithmic and statistical convergence rates remain unchanged when the variance is also learned.

In many applications involving Gaussian mixtures, the primary concern is the proximity of the learned distribution to the true underlying distribution in terms of

[1] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

the Kullback–Leibler (KL) divergence, rather than the Euclidean distance between parameters. For example, when modeling class-conditional distributions with Gaussian mixtures, minimizing the KL divergence to the true distributions leads to lower classification error rates (Devroye, Györfi, and Lugosi, 2013, Chapter 2).

Xu, Fazel, and Du (2024) offer population-level guarantees for fitting a k-component mixture with fixed covariance matrices to a single Gaussian, showing that the EM algorithm requires $O(1/\epsilon^2)$ steps to produce a mixture whose KL distance to the true distribution is $O(\epsilon)$.

This thesis work establishes population-level guarantee for a two-component, balanced mixture where both location and scale parameters are learned, specifically, that the population-level EM algorithm requires $O(1/\sqrt{\epsilon})$ iterations to produce a mixture whose KL distance to the true distribution (a single Gaussian) is $O(\epsilon)$.

This thesis is based on a research project conducted in collaboration with Dr. Assylbekov, Dr. Legg, Dr. Melnykov and Arman Bolatov. While all chapters have been adapted to fit the thesis format, certain sections contain contributions from my co-authors. Specifically, Chapter 2 was developed with significant input from Dr. Melnykov, and proof in A.4 incorporates the work of Dr. Legg (specifically items 1 and 3).

Chapter 2

Related work

The study of the Expectation-Maximization (EM) algorithm and its convergence properties in Gaussian mixture models has been a rapidly evolving field of research. Balakrishnan, Wainwright, and Yu (2017) established a framework for determining the convergence region of the algorithm concerning distribution parameters. Their analysis distinguished between a population-level approach and the sample-based implementation typically used in practice. Their characterization of the convergence region focused on the correctly specified case of $k = 2$ components, considering both balanced and unbalanced scenarios.

Within this framework, considerable attention has been given to the initialization of the EM algorithm to ensure convergence to the global optimum. Klusowski and Brinda (2016) demonstrated that local convergence occurs in a broader region than identified by Balakrishnan, Wainwright, and Yu (2017) for the two-component case, while Zhao, Li, and Sun (2020) explored the effect of initialization for an arbitrary number of well-separated components. Daskalakis, Tzamos, and Zampetakis (2017) established global convergence results for a two-component mixture where the mean vectors are symmetrically located around the origin. For k well-separated components, Segol and Nadler (2021) proved that convergence is guaranteed even when the algorithm is initialized near the midpoint between two clusters. They also improved the bound on the resulting estimation error and demonstrated similar results for Gradient EM, a variant of the classical EM algorithm. The convergence rate and local contraction radius of the Gradient EM algorithm for an arbitrary number of mixture components were further analyzed by Yan, Yin, and Sarkar (2017).

Model misspecification has also been a prominent topic in the literature. Dwivedi et al. (2018) examined an underspecified model where a two-component Gaussian mixture is fitted to data generated from a three-component mixture, providing a characterization of the bias induced by such misspecification. They also investigated the impact of initialization on convergence. The practical utility of overspecified mixture models has been recognized by Dwivedi et al. (2020b), Dwivedi et al. (2020a), Chen et al. (2024), and others. Dwivedi et al. (2020b) and Dwivedi et al. (2020a) analyzed the case where two Gaussian components are fitted to data originating from a single Gaussian distribution. They compared balanced and unbalanced cases and showed that in the sample-based EM, the unbalanced scenario exhibits significantly faster $O(1/\sqrt{n})$ statistical convergence compared to the balanced case, which follows $O(\sqrt[4]{1/n})$ when learning mean vectors under both known and estimated isotropic covariance structures. Furthermore, they demonstrated that the algorithmic convergence rate is exponentially faster in the unbalanced scenario.

The study of model overspecification in a Bayesian setting by Rousseau and Mengersen (2011) revealed that learned mixture weights tend to vary significantly in magnitude. When the number of components considerably exceeds the true number, some components tend to become hollowed out, allowing model improvement by

removing those with exceptionally small weights. Regarding spurious components, Chen et al. (2024) showed that all local minima of the negative log-likelihood, including spurious ones, encode structural information useful for identifying component means. They also highlighted the advantages of overspecification over underspecification, describing the comparison as "many-fit-one" versus "one-fit-many," respectively. Dasgupta and Schulman (2013) proposed an approach to finite mixture overspecification, recommending that the model be deliberately overspecified with $\frac{\log(k)}{w_{\min}}$ initial clusters, where $w_{\min}$ is the smallest weight, leading to significantly accelerated algorithmic convergence.

Compared to the extensive literature on convergence in terms of distribution parameters, research measuring the quality of fit using Kullback-Leibler (KL) divergence remains relatively scarce. Ghosal and Vaart (2001) established a statistical convergence rate of $(\log n)^{\kappa}/\sqrt{n}$ in Hellinger distance, translating to a lower bound of $(\log n)^{2\kappa}/n$ in KL distance. However, they did not address algorithmic aspects and worked in a well-specified setting. Dwivedi et al. (2018) employed KL divergence to analyze underspecified mixtures, but, to our knowledge, no study investigated KL divergence in the context of overspecified mixtures until the recent work of Xu, Fazel, and Du (2024). They derived KL distance bounds for the population version of Gradient EM applied to a mixture of k components with known variances. This work extends these results by analyzing population-based EM with learned component variances in a two-component mixture.

Chapter 3

Main Result

Notation We write $\mathbb{R}$ for the real numbers. Boldface lowercase letters (e.g. $\mathbf{x}$) denote vectors in $\mathbb{R}^d$, while boldface uppercase letters (e.g. $\mathbf{A}$, $\mathbf{X}$) denote matrices or random vectors, and regular lowercase letters (e.g. x) denote scalars. The Euclidean norm of $\mathbf{x} \in \mathbb{R}^d$ is $\|\mathbf{x}\| := \sqrt{\mathbf{x}^\top \mathbf{x}}$.

For functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}_+$, we write $f \precsim g$ if there exist $x_0, c \in \mathbb{R}_+$ such that $|f(x)| \le c\, g(x)$ for all $x > x_0$. 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 positive constants that may change in value at each occurrence.

Convergence of the EM algorithm Let $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ be i.i.d. random variables from the d-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{0} \in \mathbb{R}^d$ is the mean vector and $\mathbf{I} \in \mathbb{R}^{d \times d}$ is the identity covariance matrix. We wish to fit a balanced two-component location-scale Gaussian mixture

$$\mathcal{G}(\boldsymbol{\theta}, \sigma^2) \;:=\; \frac{1}{2}\mathcal{N}(-\boldsymbol{\theta},\, \sigma^2\, \mathbf{I}) \;+\; \frac{1}{2}\mathcal{N}(\boldsymbol{\theta},\, \sigma^2\, \mathbf{I}) \tag{3.1}$$

to this sample. Denoting its probability density function by $f(\mathbf{x}; \boldsymbol{\theta}, \sigma^2)$, we define the maximum likelihood estimator (MLE) of $(\boldsymbol{\theta}, \sigma^2)$ as

$$(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) \;\in\; \arg\max_{(\boldsymbol{\theta}, \sigma^2)} \frac{1}{n} \sum_{i=1}^{n} \log\, f(\mathbf{Z}_i; \boldsymbol{\theta}, \sigma^2). \tag{3.2}$$

Since (3.2) admits no closed-form solution, one typically resorts to iterative optimization methods such as the EM algorithm (Dempster, Laird, and Rubin, 1977). Note that the log-likelihood in (3.2) is not concave, hence EM can converge to different local optima depending on the initialization.

To separate the algorithmic complexity from the statistical aspects, this analysis considers the so-called *population* EM, which replaces the empirical mean by the expectation under $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. In that setting, we can show that after running the population EM algorithm for t iterations, it outputs a parameter estimate $(\boldsymbol{\theta}_t, \sigma_t)$ satisfying

$$D_{\mathrm{KL}}\big[\mathcal{N}(\mathbf{0}, \mathbf{I}) \,\big\|\, \mathcal{G}(\boldsymbol{\theta}_t, \sigma_t^2)\big] \precsim \frac{1}{t^2}. \tag{3.3}$$

The next theorem provides a more precise statement.

Theorem 1. *Fix $\epsilon > 0$ and let $d \ge 2$. For any $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ with $\|\boldsymbol{\theta}_0\| \le 1/5$, the population EM algorithm (which has access to an infinite sample) generates a sequence $\{(\boldsymbol{\theta}_t, \sigma_t^2)\}$ such that*

$$D_{\mathrm{KL}}\Big[\mathcal{N}(\mathbf{0}, \mathbf{I}) \,\Big\|\, \mathcal{G}(\boldsymbol{\theta}_T, \sigma_T^2)\Big] \;\precsim\; \epsilon \quad \textit{for} \quad T \;\succsim\; \frac{\log(1/\epsilon)}{\sqrt{\epsilon}}.$$

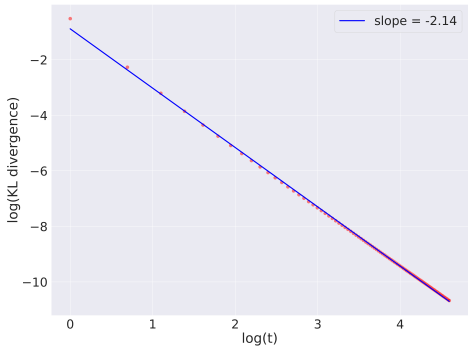

FIGURE 3.1: Log-log plot of KL divergence over iteration ($\log t$) for a 5-dimensional parameter space. The red points represent the computed values of log KL divergence, while the blue line corresponds to the best-fit linear regression with a slope of -2.14 and an intercept of -0.89. The initial parameter values were set as $\boldsymbol{\theta}_0 = (0.7, \ldots, 0.7)$. The expectation in the loss function was approximated using the trapezoidal rule, while the expectation in the update step for $\boldsymbol{\theta}$ was computed via Gauss-Hermite quadrature with 15 points per dimension.

Figure 3.1 exhibits numerical verification of the bound (3.3). As we can see, the KL distance indeed drops at a rate close to $O(1/t^2)$.

Chapter 4

Population Level Analysis

In the population setting, we assume access to infinitely many samples and replace the sample-based log-likelihood (3.2) by the population log-likelihood

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2) \;:=\; \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\Big[\log f(\mathbf{Z};\, \boldsymbol{\theta},\, \sigma^2)\Big]. \tag{4.1}$$

The Population EM algorithm then proceeds iteratively as follows:

- *Expectation step.* Given $(\boldsymbol{\theta}_t, \sigma_t^2)$, compute

$$Q\big(\boldsymbol{\theta}, \sigma^2;\, \boldsymbol{\theta}_t, \sigma_t^2\big) \;:=\; \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\Big[w\big(\mathbf{Z};\, \boldsymbol{\theta}_t, \sigma_t^2\big)\, \log\phi\Big(\tfrac{\mathbf{Z}-\boldsymbol{\theta}}{\sigma}\Big) + \big(1 - w\big(\mathbf{Z};\, \boldsymbol{\theta}_t, \sigma_t^2\big)\big)\, \log\phi\Big(\tfrac{\mathbf{Z}+\boldsymbol{\theta}}{\sigma}\Big)\Big],$$

  where

$$w\big(\mathbf{z};\, \boldsymbol{\theta}_t, \sigma_t^2\big) \;:=\; \Big(1 + \exp\big(-\tfrac{2\,\boldsymbol{\theta}_t^\top \mathbf{z}}{\sigma_t^2}\big)\Big)^{-1}.$$

- *Maximization step.* Update $\boldsymbol{\theta}_{t+1}$ and σ_{t+1}^2 by solving

$$(\boldsymbol{\theta}_{t+1}, \sigma_{t+1}^2) \;\in\; \arg\max_{(\boldsymbol{\theta},\, \sigma^2)}\; Q\big(\boldsymbol{\theta}, \sigma^2;\, \boldsymbol{\theta}_t, \sigma_t^2\big).$$

For the mixture model (3.1) applied to data from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, it is possible to derive explicit update formulas for $(\boldsymbol{\theta}_t, \sigma_t^2)$ in the population setting (see Appendix A.1):

$$\boldsymbol{\theta}_{t+1} \;=\; \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\Big[\tanh\Big(\tfrac{\boldsymbol{\theta}_t^\top \mathbf{Z}}{1 - \|\boldsymbol{\theta}_t\|^2/d}\Big)\, \mathbf{Z}\Big], \tag{4.2}$$

$$\sigma_{t+1}^2 \;=\; 1 \;-\; \tfrac{\|\boldsymbol{\theta}_{t+1}\|^2}{d}. \tag{4.3}$$

Observe that the iterates $\big(\boldsymbol{\theta}_t, \sigma_t^2\big)$ lie on the hypersurface

$$\mathcal{S} \;:=\; \Big\{(\boldsymbol{\theta}, \sigma^2) \in \mathbb{R}^{d+1} \;\big|\; \sigma^2 \;=\; 1 \;-\; \tfrac{\|\boldsymbol{\theta}\|^2}{d}\Big\}, \tag{4.4}$$

so that σ_t^2 is entirely determined by $\|\boldsymbol{\theta}_t\|$.

Radial form of the risk function. To study the population log-likelihood (4.1) restricted to $\mathcal{S}$, we note that on $\mathcal{S}$ the log-likelihood depends on $\boldsymbol{\theta}$ only through its norm. Introduce the function

$$L(\boldsymbol{\theta}) \;:=\; -\,\mathcal{L}\Big(\boldsymbol{\theta},\, 1 - \tfrac{\|\boldsymbol{\theta}\|^2}{d}\Big), \tag{4.5}$$

which can be interpreted as a *risk function* by virtue of the leading minus sign. In Appendix A.2, we show that $L(\boldsymbol{\theta})$ is *radial*. Specifically, let $\theta := \|\boldsymbol{\theta}\|$ and define

$$\ell(\theta) \;:=\; \tfrac{d}{2}\,\log\!\Big(2\pi\big(1-\tfrac{\theta^2}{d}\big)\Big) \;+\; \frac{d+\theta^2}{2\big(1-\theta^2/d\big)} \;-\; \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\Big[\log\Big(\cosh\Big(\tfrac{\theta\, Z}{1-\theta^2/d}\Big)\Big)\Big]. \tag{4.6}$$

Then

$$L(\boldsymbol{\theta}) \;=\; \ell\big(\|\boldsymbol{\theta}\|\big), \quad \text{i.e. the risk depends on } \boldsymbol{\theta} \text{ only via } \theta.$$

Population EM operator in one dimension. In view of the updates (4.2)–(4.3), the sequence $\{\boldsymbol{\theta}_t\}$ evolves under the *Population EM operator*

$$M(\boldsymbol{\theta}) \;:=\; \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathcal{N}(\mathbf{0},\,\mathbf{I})}\Big[\tanh\Big(\tfrac{\boldsymbol{\theta}^\top\mathbf{Z}}{1-\|\boldsymbol{\theta}\|^2/d}\Big)\,\mathbf{Z}\Big]. \tag{4.7}$$

Thus $\boldsymbol{\theta}_{t+1} = M(\boldsymbol{\theta}_t)$. Because $L(\boldsymbol{\theta})$ is radial, we only need to track the evolution of $\|\boldsymbol{\theta}_t\|$. Indeed, it is easily shown (Appendix A.3) that

$$\|M(\boldsymbol{\theta})\| \;=\; \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\Big[\tanh\Big(\tfrac{\|\boldsymbol{\theta}\|\, Z}{1-\|\boldsymbol{\theta}\|^2/d}\Big)\,Z\Big]. \tag{4.8}$$

Defining $\theta_t := \|\boldsymbol{\theta}_t\|$, we see that $\theta_{t+1} = m(\theta_t)$ where

$$m(\theta) \;:=\; \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\Big[\tanh\Big(\tfrac{\theta\, Z}{1-\theta^2/d}\Big)\,Z\Big]. \tag{4.9}$$

Hence understanding $\{\theta_t\}$ reduces to analyzing the univariate functions $\ell(\theta)$ and $m(\theta)$. We summarize their properties in the following lemma.

Lemma 2. *Let $\ell(\theta)$ be defined by (4.6) and $m(\theta)$ be defined by (4.9). Then*

1. $m'(\theta) \;<\; 1 \quad$ *for all* $\;|\theta| \;\le\; \tfrac{1}{5}$.
2. *$\ell(\theta)$ is convex for $\theta \in [0, \tfrac{1}{5}]$.*
3. *For $\theta \in [0, \tfrac{1}{5}]$ and $d \ge 2$, we have*

$$\theta\,\big(1 - a_d\,\theta^2\big) \;\le\; m(\theta) \;\le\; \theta\,\big(1 - b_d\,\theta^2\big),$$

   where $a_d = 1 - \frac{22}{25\,d}$ and $b_d = \frac{1}{3}\left(2 - \frac{1}{d} - \frac{1}{d^2}\right)$.

4. *Let $\theta_0 \in [0, \tfrac{1}{5}]$. Then $\theta_t \;<\; \epsilon$ for $t \;\ge\; \frac{\log(\theta_0/\epsilon)}{b_d\,\epsilon^2} \;+\; 1$.*

The convexity of $L(\boldsymbol{\theta})$ follows immediately from Lemma 2 part 2. For $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ with $\|\boldsymbol{\theta}_i\| \le 1/5$, $i = 1, 2$, and $\alpha \in [0, 1]$,

$$\begin{aligned} L\big(\alpha\boldsymbol{\theta}_1 + (1-\alpha)\boldsymbol{\theta}_2\big) \;&=\; \ell\big(\|\alpha\boldsymbol{\theta}_1 + (1-\alpha)\boldsymbol{\theta}_2\|\big) \\ &\le\; \ell\Big(\alpha\|\boldsymbol{\theta}_1\| + (1-\alpha)\|\boldsymbol{\theta}_2\|\Big) \;\le\; \alpha\, L(\boldsymbol{\theta}_1) + (1-\alpha)\, L(\boldsymbol{\theta}_2), \end{aligned}$$

where the second inequality uses the convexity of $\ell(\theta)$.

Figure 4.1 illustrates $L(\boldsymbol{\theta})$ for $d = 2$, and Figure 4.2 shows $\ell(\theta)$. Note that while both functions are convex near the origin, they also flatten significantly as $\theta \to 0$, causing slower convergence for the over-specified mixture model compared to the faster (exponential) convergence known under well-specified conditions Balakrishnan, Wainwright, and Yu, 2017.

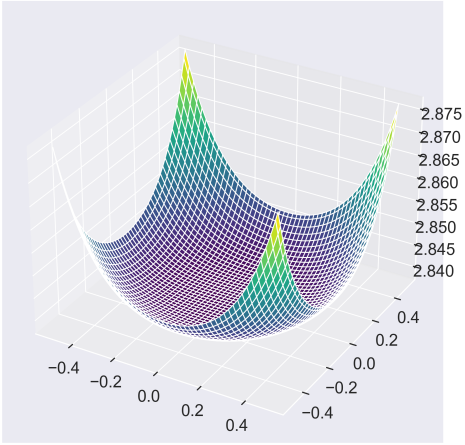

FIGURE 4.1: Plot of $L(\boldsymbol{\theta})$ in (4.5), for $d = 2$.

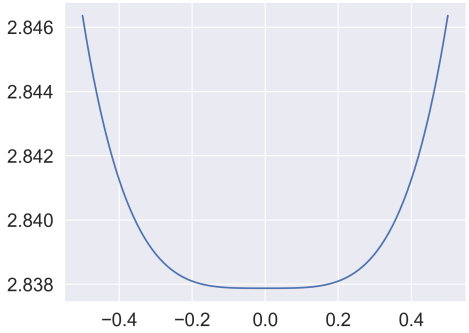

FIGURE 4.2: Plot of $\ell(\theta)$ in (4.6), for $d = 2$.

4.1 Proof of Theorem 1.

Proof. We can now bound the KL divergence $\mathrm{KL}\big[\mathcal{N}(\mathbf{0}, \mathbf{I}) \,\big\|\, \mathcal{G}(\boldsymbol{\theta}_t, \sigma_t^2)\big]$. Since

$$\mathrm{KL}\big[\mathcal{N}(\mathbf{0}, \mathbf{I}) \,\big\|\, \mathcal{G}(\boldsymbol{\theta}_t, \sigma_t^2)\big] \;=\; L(\boldsymbol{\theta}_t) \;-\; L(\mathbf{0}),$$

it suffices to show $L(\boldsymbol{\theta}_t) - L(\mathbf{0}) \precsim \theta_t^4$. From Lemma 2 part 2 and the fact that $L(\boldsymbol{\theta})$ is radial, we have

$$L(\boldsymbol{\theta}_t) - L(\mathbf{0}) \;=\; \ell(\theta_t) - \ell(0) \;\le\; \ell'(\theta_t)\,\theta_t \;=\; \frac{1+\frac{\theta_t^2}{d}}{\left(1-\frac{\theta_t^2}{d}\right)^2}\,\big(\theta_t - m(\theta_t)\big)\,\theta_t. \tag{4.10}$$

By Lemma 2 part 3, $\theta_t - m(\theta_t) \le a_d\,\theta_t^3$ for some constant a_d. Hence

$$L(\boldsymbol{\theta}_t) - L(\mathbf{0}) \;\le\; \underbrace{\frac{1+\frac{\theta_0^2}{d}}{\left(1-\frac{\theta_0^2}{d}\right)^2}\,a_d}_{\precsim 1}\,\theta_t^4 \;\precsim\; \theta_t^4.$$

Lemma 4 implies that $\theta_T \le \sqrt[4]{\epsilon}$ once $T \succsim (\log(1/\epsilon))/\sqrt{\epsilon}$. Consequently,

$$L(\boldsymbol{\theta}_T) - L(\mathbf{0}) \;\precsim\; \theta_T^4 \;\precsim\; \epsilon,$$

which completes the proof of Theorem 1. □

Chapter 5

Conclusion

This study establishes algorithmic complexity of the EM algorithm in the overspecified setup. In the extended work, this result aids in a finite sample analysis to establish statistical complexity. Additionally, this bound is applied in the analysis of binary classification error using over-specified MDA. In a broader context, it allows us to compare the efficiency of different algorithms in suitable settings.

Finally, study of the behavior of the EM algorithm for fitting two components of GMM to a single Gaussian distribution can be extended in the future in multiple ways: examining unbalanced mixture components; examining a single Gaussian GMM fitting k component for $k > 2$; examining convergence in KL divergence of GMM to more complex distributions.

Appendix A

Appendix A

A.1 Explicit formulae for the Population EM iterates

We first give more details on the EM algorithm. It will be convenient to represent the mixture distribution (3.1) using a hidden Bernoulli random variable K, which serves as an identifier of the mixture components. Since both components have equal weight, we assume that $\Pr[K = 0] = Pr[K = 1] = 1/2$. Next, we define the conditional distribution

$$(\mathbf{X} \mid K = 0) \sim \mathcal{N}(-\boldsymbol{\theta}, \sigma^2 \mathbf{I}), \qquad (\mathbf{X} \mid K = 1) \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 \mathbf{I}).$$

This determines the joint distribution of the tuple $(\mathbf{X}, K)$, and by construction, the marginal distribution of $\mathbf{X}$ is the Gaussian mixture $\mathcal{G}(\boldsymbol{\theta}, \sigma^2)$ given by (3.1). The Population EM algorithm attempts to maximize the expected log-likelihood (4.1) by iteratively applying these two steps:

- *E step:* Given the current estimate $(\boldsymbol{\theta}_t, \sigma_t^2)$, do a soft assignment of any $\mathbf{x} \in \mathbb{R}^d$ to the component $K = 1$, i.e. compute the conditional probability of $K = 1$ given $\mathbf{X} = \mathbf{x}$:

$$w(\mathbf{x}; \boldsymbol{\theta}_t, \sigma_t^2) = \frac{\phi\left(\frac{\mathbf{x}-\boldsymbol{\theta}_t}{\sigma_t}\right)}{\phi\left(\frac{\mathbf{x}-\boldsymbol{\theta}_t}{\sigma_t}\right) + \phi\left(\frac{\mathbf{x}+\boldsymbol{\theta}_t}{\sigma_t}\right)}. \tag{A.1}$$

- *M step:* Use the assignment (A.1) to update the parameters by computing the weighted mean and variance:

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \frac{\mathbb{E}_{\mathbf{Z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[w(\mathbf{Z}; \boldsymbol{\theta}_t, \sigma_t^2)\mathbf{Z}]}{\mathbb{E}_{\mathbf{Z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[w(\mathbf{Z}; \boldsymbol{\theta}_t, \sigma_t^2)]}, \\ \sigma_{t+1}^2 &= \frac{1}{d} \cdot \frac{\mathbb{E}_{\mathbf{Z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[w(\mathbf{Z}; \boldsymbol{\theta}_t, \sigma_t^2)\|\mathbf{Z} - \boldsymbol{\theta}_{t+1}\|^2]}{\mathbb{E}_{\mathbf{Z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[w(\mathbf{Z}; \boldsymbol{\theta}_t, \sigma_t^2)]}. \end{aligned} \tag{A.2}$$

The derivation of formulas (A.1) and (A.2) for a more general case is given in the work of Cai, Ma, and Zhang, 2019, see their formulas (3.6) and (3.7).

Lemma 3. *Let the expected log-likelihood (4.1) be maximized by the Population EM algorithm. Then the parameter updates are given by*

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\tanh\left(\frac{\boldsymbol{\theta}_t^\top \mathbf{Z}}{1 - \frac{\|\boldsymbol{\theta}_t\|^2}{d}}\right)\mathbf{Z}\right], \\ \sigma_{t+1}^2 &= 1 - \frac{\|\boldsymbol{\theta}_{t+1}\|^2}{d}. \end{aligned}$$

Proof. First, we rewrite the weight function (A.1) as

$$w(\mathbf{z};\boldsymbol{\theta},\sigma^2) = \frac{1}{1+\exp\left(-\frac{2\boldsymbol{\theta}^\top \mathbf{z}}{\sigma^2}\right)} = s\left(\frac{2\boldsymbol{\theta}^\top \mathbf{z}}{\sigma^2}\right),$$

where $s(x) = (1+\exp(-x))^{-1}$ is the logistic sigmoid function. Then the population EM update for $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\theta}_{t+1} = \frac{\mathbb{E}_{\mathbf{Z}}\left[s\left(\frac{2\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\mathbf{Z}\right]}{\mathbb{E}_{\mathbf{Z}}\left[s\left(\frac{2\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\right]}. \tag{A.3}$$

Taking into account that $s(x) = 1 - s(-x)$, the denominator in (A.3) can be written as

$$\mathop{\mathbb{E}}_{\mathbf{Z}}\left[s\left(\frac{2\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\right] = \mathop{\mathbb{E}}_{\mathbf{Z}}\left[1 - s\left(-\frac{2\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\right] = 1 - \mathop{\mathbb{E}}_{\mathbf{Z}}\left[s\left(\frac{2\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\right], \tag{A.4}$$

where we used the fact that $\boldsymbol{\theta}^\top \mathbf{Z}$ and $-\boldsymbol{\theta}^\top \mathbf{Z}$ are identically distributed. From (A.4) we have

$$\mathop{\mathbb{E}}_{\mathbf{Z}}\left[s\left(\frac{2\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\right] = \frac{1}{2}. \tag{A.5}$$

Since $\mathbb{E}_{\mathbf{Z}}[\mathbf{Z}] = \mathbf{0}$, and using (A.5), the update (A.3) can be rewritten as

$$\boldsymbol{\theta}_{t+1} = \mathop{\mathbb{E}}_{\mathbf{Z}}\left[\left(2s\left(\frac{2\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right) - 1\right)\mathbf{Z}\right] = \mathop{\mathbb{E}}_{\mathbf{Z}}\left[\tanh\left(\frac{\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\mathbf{Z}\right]. \tag{A.6}$$

Now, let us focus on the population EM update for σ^2:

$$\begin{aligned} d\sigma_{t+1}^2 =& \frac{\mathbb{E}_{\mathbf{Z}}[s(2\boldsymbol{\theta}_t^\top \mathbf{Z}/\sigma_t^2)\|\mathbf{Z}-\boldsymbol{\theta}_{t+1}\|^2]}{\mathbb{E}_{\mathbf{Z}}[s(2\boldsymbol{\theta}_t^\top \mathbf{Z}/\sigma_t^2)]} \\ =& \mathop{\mathbb{E}}_{\mathbf{Z}}\left[\left(2s\left(\frac{2\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma^2}\right) - 1\right)\|\mathbf{Z}-\boldsymbol{\theta}_{t+1}\|^2\right] + \mathop{\mathbb{E}}_{\mathbf{Z}}\left[\|\mathbf{Z}-\boldsymbol{\theta}_{t+1}\|^2\right] \\ =& \mathop{\mathbb{E}}_{\mathbf{Z}}\left[\tanh\left(\frac{\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\|\mathbf{Z}\|^2\right] + \mathop{\mathbb{E}}_{\mathbf{Z}}\left[\tanh\left(\frac{\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\|\boldsymbol{\theta}_{t+1}\|^2\right] \\ & - 2\mathop{\mathbb{E}}_{\mathbf{Z}}\left[\tanh\left(\frac{\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\boldsymbol{\theta}_{t+1}^\top \mathbf{Z}\right] + \mathop{\mathbb{E}}_{\mathbf{Z}}[\|\mathbf{Z}\|^2] + \|\boldsymbol{\theta}_{t+1}\|^2 - 2\cdot\boldsymbol{\theta}_{t+1}\mathop{\mathbb{E}}_{\mathbf{Z}}[\mathbf{Z}] \end{aligned} \tag{A.7}$$

Since $\tanh(x) = -\tanh(-x)$, we have

$$\mathop{\mathbb{E}}_{\mathbf{Z}}\left[\tanh\left(\frac{\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\|\mathbf{Z}\|^2\right] = \mathop{\mathbb{E}}_{\mathbf{Z}}\left[-\tanh\left(-\frac{\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\|-\mathbf{Z}\|^2\right] = -\mathop{\mathbb{E}}_{\mathbf{Z}}\left[\tanh\left(\frac{\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\|\mathbf{Z}\|^2\right],$$

which implies $\mathbb{E}_{\mathbf{Z}}\left[\tanh\left(\frac{\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\|\mathbf{Z}\|^2\right] = \mathbf{0}$. Similarly, $\mathbb{E}_{\mathbf{Z}}\left[\tanh\left(\frac{\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\|\boldsymbol{\theta}_{t+1}\|^2\right] = \mathbf{0}$. Also, notice that $\|\mathbf{Z}\|^2 \sim \chi_d^2$, and thus $\mathbb{E}_{\mathbf{Z}}[\|\mathbf{Z}\|^2] = d$. Plugging these into (A.7) we

get

$$d\sigma_{t+1}^2 = -2\boldsymbol{\theta}_{t+1}^\top \underbrace{\mathop{\mathbb{E}}_{\mathbf{Z}}\left[\tanh\left(\frac{\boldsymbol{\theta}_t^\top \mathbf{Z}}{\sigma_t^2}\right)\mathbf{Z}\right]}_{\boldsymbol{\theta}_{t+1}} + d + \|\boldsymbol{\theta}_{t+1}\|^2 = d - \|\boldsymbol{\theta}_{t+1}\|^2,$$

which implies (4.3). Now, plugging $\sigma_t^2 = 1 - \frac{\|\boldsymbol{\theta}_t\|^2}{d}$ into (A.6) we obtain (4.2). □

A.2 Radiality of the function $L(\boldsymbol{\theta})$

Lemma 4. *Consider the function $L : \mathbb{R}^d \to \mathbb{R}$ defined as*

$$L(\boldsymbol{\theta}) := -\mathcal{L}(\boldsymbol{\theta}, 1 - \|\boldsymbol{\theta}\|^2/d).$$

Then $L(\boldsymbol{\theta})$ is a radial function of $\boldsymbol{\theta} \in \mathbb{R}^d$. It can be explicitly written as

$$L(\boldsymbol{\theta}) = \frac{d}{2}\log(2\pi(1 - \|\boldsymbol{\theta}\|^2/d)) + \frac{d + \|\boldsymbol{\theta}\|^2}{2(1 - \|\boldsymbol{\theta}\|^2/d)} - \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\log\left(\cosh\left(\frac{\|\boldsymbol{\theta}\|Z}{1 - \|\boldsymbol{\theta}\|^2/d}\right)\right)\right].$$

Proof. Since $f(\mathbf{x}; \boldsymbol{\theta}, \sigma^2)$ is the p.d.f. of $\frac{1}{2}\mathcal{N}(-\boldsymbol{\theta}, \sigma^2\mathbf{I}) + \frac{1}{2}\mathcal{N}(+\boldsymbol{\theta}, \sigma^2\mathbf{I})$, it can be written as

$$\begin{aligned} f(\mathbf{x}; \boldsymbol{\theta}, \sigma^2) &= \frac{1}{2\sigma^d}\phi\left(\frac{\mathbf{x} + \boldsymbol{\theta}}{\sigma}\right) + \frac{1}{2\sigma^d}\phi\left(\frac{\mathbf{x} - \boldsymbol{\theta}}{\sigma}\right) \\ &= (2\pi\sigma^2)^{-d/2} \cdot \exp\left(-\frac{\|\mathbf{x}\|^2 + \|\boldsymbol{\theta}\|^2}{2\sigma^2}\right) \cdot \cosh\left(\frac{\boldsymbol{\theta}^\top \mathbf{x}}{\sigma^2}\right). \end{aligned}$$

Hence

$$\log f(\mathbf{x}; \boldsymbol{\theta}, \sigma^2) = -\frac{d}{2}\log(2\pi\sigma^2) - \frac{\|\mathbf{x}\|^2 + \|\boldsymbol{\theta}\|^2}{2\sigma^2} + \log\left(\cosh\left(\frac{\boldsymbol{\theta}^\top \mathbf{x}}{\sigma^2}\right)\right),$$

and thus the negative population log-likelihood is

$$\begin{aligned} -\mathcal{L}(\boldsymbol{\theta}, \sigma^2) &= -\mathop{\mathbb{E}}_{\mathbf{Z}}\left[\log f(\mathbf{Z}; \boldsymbol{\theta}, \sigma^2)\right] \\ &= \frac{d}{2}\log(2\pi\sigma^2) + \frac{d + \|\boldsymbol{\theta}\|^2}{2\sigma^2} - \mathop{\mathbb{E}}_{\mathbf{Z}}\left[\log\left(\cosh\left(\frac{\boldsymbol{\theta}^\top \mathbf{Z}}{\sigma^2}\right)\right)\right] \\ &= \{\text{since } \boldsymbol{\theta}^\top \mathbf{Z} \sim \mathcal{N}(0, \|\boldsymbol{\theta}\|^2)\} \\ &= \underbrace{\frac{d}{2}\log(2\pi\sigma^2) + \frac{d + \|\boldsymbol{\theta}\|^2}{2\sigma^2} - \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\log\left(\cosh\left(\frac{\|\boldsymbol{\theta}\|Z}{\sigma^2}\right)\right)\right]}_{q(\|\boldsymbol{\theta}\|, \sigma^2)}. \end{aligned} \tag{A.8}$$

As we can notice, the $\mathcal{L}(\boldsymbol{\theta}, \sigma^2)$ depends on $\boldsymbol{\theta}$ only through its norm. Plugging in $\sigma^2 = 1 - \|\boldsymbol{\theta}\|^2/d$, we get

$$\begin{aligned} L(\boldsymbol{\theta}) &= -\mathcal{L}(\boldsymbol{\theta}, 1 - \|\boldsymbol{\theta}\|^2/d) \\ &= \frac{d}{2}\log(2\pi(1 - \|\boldsymbol{\theta}\|^2/d)) + \frac{d + \|\boldsymbol{\theta}\|^2}{2(1 - \|\boldsymbol{\theta}\|^2/d)} - \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\log\left(\cosh\left(\frac{\|\boldsymbol{\theta}\|Z}{1 - \|\boldsymbol{\theta}\|^2/d}\right)\right)\right], \end{aligned}$$

which immediately implies the statement of the lemma. □

A.3 Radiality of $\|M(\boldsymbol{\theta})\|$

Lemma 5. *For the population EM operator $M(\boldsymbol{\theta})$ defined by (4.7), we have*

$$\|M(\boldsymbol{\theta})\| = \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)}\left[\tanh\left(\frac{\|\boldsymbol{\theta}\|Z}{1 - \frac{\|\boldsymbol{\theta}\|^2}{d}}\right) Z\right].$$

Proof. Let $\mathbf{R}$ be an orthonormal matrix such that $\mathbf{R}\boldsymbol{\theta} = \|\boldsymbol{\theta}\|\mathbf{e}_1$, where $\mathbf{e}_1$ is the first canonical basis vector in $\mathbb{R}^d$. Let $\mathbf{Y} = \mathbf{R}\mathbf{Z}$, then $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{Z} = \mathbf{R}^\top\mathbf{Y}$. Thus, we have

$$\begin{aligned} \|M(\boldsymbol{\theta})\| &= \left\| \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\tanh\left(\frac{\boldsymbol{\theta}^\top\mathbf{Z}}{1 - \|\boldsymbol{\theta}\|^2/d}\right)\mathbf{Z}\right]\right\| = \left\| \mathop{\mathbb{E}}_{\mathbf{Y}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\tanh\left(\frac{\|\boldsymbol{\theta}\|Y_1}{1 - \|\boldsymbol{\theta}\|^2/d}\right)\mathbf{R}^\top\mathbf{Y}\right]\right\| \\ &= \left\| \mathbf{R}^\top \mathop{\mathbb{E}}_{\mathbf{Y}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\tanh\left(\frac{\|\boldsymbol{\theta}\|Y_1}{1 - \|\boldsymbol{\theta}\|^2/d}\right)\cdot\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}\right]\right\| \\ &= \left\| \mathbf{R}^\top \begin{bmatrix} \mathbb{E}_{Y_1\sim\mathcal{N}(0,1)}\left[\tanh\left(\frac{\|\boldsymbol{\theta}\|Y_1}{1 - \|\boldsymbol{\theta}\|^2/d}\right)Y_1\right] \\ 0 \\ \vdots \\ 0 \end{bmatrix}\right\| \\ &= \mathop{\mathbb{E}}_{Y_1\sim\mathcal{N}(0,1)}\left[\tanh\left(\frac{\|\boldsymbol{\theta}\|Y_1}{1 - \|\boldsymbol{\theta}\|^2/d}\right)Y_1\right]. \end{aligned}$$

□

A.4 Proof of Lemma 2

1. We first consider the case $d = 1$.

   Lemma 6. *Define*

   $$m_1(\theta) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\tanh\left(\frac{\theta x}{1 - \theta^2}\right) x\, e^{-x^2/2}dx.$$

   Then $m_1'(\theta) \le 1$ for $0 \le \theta \le 0.2$.

   Proof. By differentiating under the integral sign

   $$m_1'(\theta) = \frac{1}{\sqrt{2\pi}}\frac{1 + \theta^2}{(1 - \theta^2)^2}\int_{-\infty}^{\infty}\operatorname{sech}^2\left(\frac{\theta x}{1 - \theta^2}\right) x^2\, e^{-x^2/2}dx.$$

We use the Maclaurin expansion of $\text{sech}^2(x)$ of order 8, which is

$$T_8(x) := 1 - x^2 + (2/3)\,x^4 - (17/45)\,x^6 + (62/315)\,x^8.$$

Define the approximation F by

$$F(\theta) := \frac{1}{\sqrt{2\pi}}\,\frac{1+\theta^2}{(1-\theta^2)^2}\int_{-\infty}^{\infty} T_8\left(\frac{\theta x}{1-\theta^2}\right)\,x^2\,e^{-x^2/2}dx.$$

By Taylor's Theorem (and evenness) $R(x) := |\text{sech}^2(x) - T_8(x)|$ is bounded above by $K\,x^{10}/10!$, where K is any upper bound on $|d^{10}/dx^{10}(\text{sech}^2(x))|$. Now this tenth derivative is the composition

$$d^{10}/dx^{10}(\text{sech}^2 x) = q(\text{sech(x)}),$$

where q is the polynomial

$$q(x) := 1024x^2 - 523776x^4 + 10813440x^6 - 50561280x^8 + 79833600x^{10} - 39916800x^{12}.$$

Since $0 \le \text{sech(x)} \le 1$ for all x, we need only bound $|q(x)|$ on the domain $0 \le x \le 1$, and it is easily checked by calculus that the maximum on this interval occurs at $x = 1$. Thus

$$R(x) \le |q(1)|\,x^{10}/10! \le 0.0975x^{10}.$$

It is also simple to verify that T_8 is positive-valued everywhere, so that the triangle inequality gives

$$m_1'(\theta) \le F(\theta) + \frac{0.0975}{\sqrt{2\pi}}\,\frac{1+\theta^2}{(1-\theta^2)^2}\int_{-\infty}^{\infty}\left(\frac{\theta x}{1-\theta^2}\right)^{10}\,x^2\,e^{-x^2/2}dx.$$

The right hand side can be explicitly computed and gives

$$m_1'(\theta) \le \frac{1+\theta^2}{(1-\theta^2)^{12}}\sum_{j=0}^{10}(-1)^j\,c_j\,\theta^{2j},$$

with coefficients

$$c_0 = c_{10} = 1, \quad c_1 = c_9 = 13, \quad c_2 = c_8 = 79,$$
$$c_3 = c_7 = \frac{911}{3}, \quad c_4 = c_6 = \frac{2618}{3}, \quad c_5 = \frac{20679}{80}.$$

Notice that $m_1'(0) = 1$. Calculus and a root-finding algorithm will show that $(1 - m_1'(\theta)) > 0$ for $0 < \theta < 0.226$. □

We now prove the statement for any $d \ge 1$. By differentiation under the integral sign

$$m'(\theta) = \frac{1}{\sqrt{2\pi}}\,\frac{1+\theta^2/d}{(1-\theta^2/d)^2}\int_{-\infty}^{\infty}\text{sech}^2\left(\frac{\theta x}{1-\theta^2/d}\right)\,x^2\,e^{-x^2/2}dx.$$

It is now advantageous to consider m' as a function of both θ and d, so define

$$\mu(d,\theta) := \frac{1}{\sqrt{2\pi}} \frac{1+\theta^2/d}{(1-\theta^2/d)^2} \int_{-\infty}^{\infty} \operatorname{sech}^2\left(\frac{\theta x}{1-\theta^2/d}\right) x^2 e^{-x^2/2} dx.$$

In the previous lemma we showed that $\mu(1,\theta) < 1$ for $0 < \theta < 1/5$. To show that $m'(\theta) < 1$ for such θ and for all $d \geq 1$, it will suffice to show that $\mu(d,\theta)$ is always decreasing in $d \geq 1$ for each fixed θ. To that end compute the partial derivative and simplify, getting

$$\frac{\partial \mu}{\partial d} = \frac{2\theta^2}{(d-\theta^2)^4 \sqrt{2\pi}} \int_{-\infty}^{\infty} g(d,\theta,x) \operatorname{sech}^2\left(\frac{\theta x}{1-\theta^2/d}\right) x^2 e^{-x^2/2} dx,$$

where the function g is defined as

$$g(d,\theta,x) := \left(-\frac{3}{2} - \theta\, x\, \tanh\left(\frac{\theta x}{1-\theta^2/d}\right)\right) d^2 + \left(\theta^2 - \theta^3 x \tanh\left(\frac{\theta x}{1-\theta^2/d}\right)\right) d + \frac{\theta^4}{2}.$$

Since we consider $d > 1$ and $\theta < 1$ it is clear that

$$g(d,\theta,x) \leq -\frac{3}{2}d^2 + \theta^2 d + \frac{\theta^4}{2};$$

even more, since we consider $0 \leq \theta \leq 1/5$, we get

$$g(d,\theta,x) \leq -\frac{3}{2}d^2 + \frac{1}{25}d + \frac{1}{1250}.$$

This in turn guarantees, uniformly in x and for all $d > 1/25$, $0 \leq \theta \leq 1/5$, that

$$g(d,\theta,x) < 0.$$

Hence we obviously have by the integral formula above

$$\frac{\partial \mu}{\partial d} < 0$$

for all such parameter values.

In summary, since now μ is decreasing in d, we have for each fixed $d > 1$ and for all $0 \leq \theta \leq 1/5$,

$$m'(\theta) = \mu(d,\theta) \leq \mu(1,\theta) \leq 1.$$

2. Recall the notation $\theta := \|\boldsymbol{\theta}\|$ and $\ell(\theta) := L(\boldsymbol{\theta})$. Let $q(\theta,\sigma^2)$ be the function defined in (A.8), then $\ell(\theta) := q(\theta, 1-\theta^2/d)$. To analyze the derivative ℓ', we first find partial derivatives of q:

$$\frac{\partial q}{\partial \theta}(\theta,\sigma^2) = \frac{\theta}{\sigma^2} - \frac{1}{\sigma^2} \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)} \left[\tanh\left(\frac{\theta Z}{\sigma^2}\right) Z\right],$$
$$\frac{\partial q}{\partial \sigma^2}(\theta,\sigma^2) = \frac{d}{2\sigma^2} - \frac{d+\theta^2}{2\sigma^4} + \frac{1}{\sigma^4} \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)} \left[\tanh\left(\frac{\theta Z}{\sigma^2}\right) \theta Z\right]$$

Using the notation $m(\theta) := \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\tanh\left(\frac{\theta Z}{1-\theta^2/d}\right) Z\right]$ introduced in (4.9), we have

$$\begin{aligned}
\ell'(\theta) &= \frac{\partial q}{\partial \theta}(\theta, 1-\theta^2/d) + \frac{\partial q}{\partial \sigma^2}(\theta, 1-\theta^2/d)\cdot\left(-\frac{2\theta}{d}\right) \\
&= \frac{\theta}{1-\theta^2/d} - \frac{M(\theta)}{1-\theta^2/d} - \frac{\theta}{1-\theta^2/d} + \frac{(1+\theta^2/d)\theta}{(1-\theta^2/d)^2} - \frac{2M(\theta)\cdot\theta^2/d}{(1-\theta^2/d)^2} \\
&= \frac{-M(\theta)(1-\theta^2/d) + (1+\theta^2/d)\theta - 2M(\theta)\theta^2/d}{(1-\theta^2/d)^2} \\
&= \frac{1+\theta^2/d}{(1-\theta^2/d)^2}\cdot(\theta - m(\theta))\,.
\end{aligned}$$

From the previous part, we have $m'(\theta) \le 1$ for $\theta \in [0, \frac{1}{5}]$, implying $\theta - m(\theta)$ is nondecreasing on that interval. Consequently, $\ell'(\theta)$ is an increasing function on $[0, \frac{1}{5}]$, which shows that $\ell(\theta)$ is convex there. Also, $\ell'(0) = 0$ and $\ell'(\theta) > 0$ for $\theta > 0$, so $\ell(\theta)$ is increasing on that range.

3. We begin by noting that for all real z, $1 - z^2 \le \operatorname{sech}^2(z) \le 1 - z^2 + \frac{2}{3}z^4$. For z in the interval of convergence of the MacLaurin series, this comes by noticing $\operatorname{sech}^2(z)$ has alternating decreasing MacLaurin coefficients; and then the inequality extends beyond the interval of convergence by simple calculus.

   Integrating by parts, $m(\theta)$ is equivalently

$$m(\theta) = \frac{1}{\sqrt{2\pi}}\cdot\frac{\theta}{1-\theta^2/d}\cdot\int_{-\infty}^{\infty}\operatorname{sech}^2\!\left(\frac{\theta x}{1-\theta^2/d}\right) e^{-x^2/2}\,dx.$$

   From the above inequality together with the integral values

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^2 e^{-x^2/2} = 1,$$

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x^4\, e^{-x^2/2}\,dx = 3,$$

   we see immediately that

$$\frac{\theta}{1-\theta^2/d}\left(1 - \frac{\theta^2}{(1-\theta^2/d)^2}\right) \le m(\theta) \le \frac{\theta}{1-\theta^2/d}\left(1 - \frac{\theta^2}{(1-\theta^2/d)^2} + \frac{2\,\theta^4}{(1-\theta^2/d)^4}\right)$$

   We analyze each side of this inequality in turn.

   From the lower bound on $m(\theta)$, algebraic manipulation shows that

$$-\frac{1}{\theta^2}\left(\frac{m(\theta)}{\theta} - 1\right) \le q(\theta, d),$$

   where we use

$$q(\theta, d) := \frac{d^3 - d^2 + 2\theta^2 d - \theta^4}{(d-\theta^2)^3}.$$

   The partial derivative in θ is

$$\frac{\partial q}{\partial \theta} = \frac{2\theta}{(d-\theta^2)^4}\left(3d^3 - d^2 + 2\theta^2 d - \theta^4\right).$$

By the quadratic formula, for each fixed $d \geq 1$, the polynomial in parentheses in the previous line has roots only at

$$\theta = \pm\sqrt{d + \sqrt{3d^3}}$$

and so it is clear that $\partial q/\partial\theta \geq 0$ for all $0 \leq \theta \leq 1$ and for all $d \geq 1$.

Let now T be a positive constant $T \leq 1/5$, and we restrict θ to the domain $[0, T]$. From what we have determined so far, for such θ we have

$$-\frac{1}{\theta^2}\left(\frac{m(\theta)}{\theta} - 1\right) \leq q(T, d),$$

whenever $d \geq 1$.

Consider now the function

$$r(d) := d\,(\,1 - q(T, d)\,).$$

Some computation reveals that

$$r'(d) = \frac{(6T^4 - T^2)d^2 + (2T^4 - 4T^6)d + T^8 - T^6}{(d - T^2)^3.}$$

Notice the numerator is quadratic in d and by our assumption that $T \leq 1/5$, the leading term is negative. Thus $r'(d) < 0$ whenever d is greater than the greater root of the numerator. That root is

$$\frac{2T^2 - 1 - \sqrt{3T^2 - 2T^4}}{6T^2 - 1},$$

which is less than 2 when $T \leq 1/5$ as can be checked with calculus. Hence $r'(d) < 0$ for all $d \geq 2$.

We conclude that for $d \geq 2$, $r(d)$ decreases asymptotically to $\lim_{d\to\infty} r(d) = 1 - 3T^2$. In particular, $r(d) > 1 - 3T^2$ for all $d \geq 2$. Appealing to the definition of $r(d)$ above, we manipulate algebraically to see that

$$q(T, d) \leq 1 - \frac{1 - 3T^2}{d},$$

from which it follows that

$$-\frac{1}{\theta^2}\left(\frac{m(\theta)}{\theta} - 1\right) \leq q(T, d) \leq 1 - \frac{1 - 3T^2}{d},$$

and after manipulating

$$m(\theta) \geq \theta \cdot (1 - \frac{1 - 3T^2}{d} \cdot \theta^2).$$

Choosing $T = 1/5$, we see that

$$\theta \cdot (1 - a_d \cdot \theta^2) \leq m(\theta),$$

with $a_d := 1 - 22/(25d)$, valid for all $0 \leq \theta \leq 1/5$ and $d \geq 2$.

Next we turn to the upper bound on $m(\theta)$. This time, after rearranging we have

$$-\frac{1}{\theta^2}\left(\frac{m(\theta)}{\theta}-1\right) \geq Q(\theta, d),$$

where we have used

$$Q(\theta, d) := -\frac{\theta^8 - 4d\theta^6 + (6d^2 - d^3)\theta^4 + (2d^5 + 2d^4 - 4d^3)\theta^2 - d^5 + d^4}{(d-\theta^2)^5}.$$

The partial derivative in θ is

$$\frac{\partial Q}{\partial \theta} = -\frac{2\,\theta}{(d-\theta^2)^6}\, p(\theta, d),$$

where p is the polynomial

$$p(\theta, d) := \theta^8 + 8d^5\theta^2 - 3d^3\theta^4 - 4d\theta^6 + 2d^6 + 6d^4\theta^2 + 2d^2\theta^4 - 3d^5 - 4d^3\theta^2 + d^4.$$

To control $p(\theta, d)$, we note that for $0 \leq \theta \leq 1/5$, we have in particular $0 \leq \theta \leq 1/\sqrt{6}$. We make a rough estimate plugging in 0 and $1/\sqrt{6}$ for θ into the terms with positive and negative coefficients respectively:

$$p(\theta, d) \geq 2d^6 - 3d^5 + d^4 - \frac{3}{4}\,d^3 - \frac{1}{54}\,d.$$

Now the greatest real root of the right side of the inequality is around $d \approx 1.33$, implying that $p(\theta, d) > 0$ for all $d \geq 2$ and $0 \leq \theta \leq 1/\sqrt{6}$. In turn, $Q(\theta, d)$ is decreasing for such values. And so for $0 \leq \theta \leq 1/5$ and all $d \geq 2$ we have

$$Q(\theta, d) \geq Q(1/\sqrt{6}, d).$$

To finish, we define the function R via

$$R(d) := \frac{Q(1/\sqrt{6}, d)}{2 - \frac{1}{d} - \frac{1}{d^2}}.$$

A direct computation will show that

$$R'(d) = \frac{6\,d}{(2d^2 - d - 1)^2(6d-1)^2} \cdot \rho(d),$$

where ρ is the following seventh-degree polynomial with positive leading coefficient:

$$\rho(d) = 6912d^7 - 12960d^6 + 15552d^5 - 9648d^4 + 3000d^3 - 552d^2 + 53d - 2.$$

The greatest real root of ρ is $d \approx 0.53$, which means that $\rho(d) > 0$ for all $d \geq 1$, and so also $R(d)$ is increasing for all $d \geq 1$. But $R(d)$ also has the limiting value $\lim_{d\to\infty} R(d) = 1/3$.

We conclude that for all $d \geq 1$,

$$Q(\theta, d) \geq \left(2 - \frac{1}{d} - \frac{1}{d^2}\right) \cdot \frac{1}{3},$$

and by algebraic manipulation

$$m(\theta) \leq \theta^2 \cdot (1 - b_d \cdot \theta),$$

where

$$b_d = \frac{2}{3} - \frac{1}{3d} - \frac{1}{3d^2},$$

and the inequality in particular is valid for all $0 \leq \theta \leq 1/5$ and for all $d \geq 1$. That finishes the proof.

4. Because $\theta_{t+1} = m(\theta_t) \leq \theta_t(1 - b_d\,\theta_t^2)$ for $\theta_t \in [0, \frac{1}{5}]$, the sequence $\{\theta_t\}$ is a strictly decreasing sequence in $[0, \frac{1}{5}]$ that converges to zero. Let τ be the last index such that $\theta_\tau \geq \epsilon$. Then

$$\theta_\tau \;=\; m(\theta_{\tau-1}) \;\leq\; \theta_{\tau-1}\left(1 - b_d\,\epsilon^2\right) \;\leq\; \cdots \;\leq\; \theta_0\left(1 - b_d\,\epsilon^2\right)^\tau.$$

Taking logarithms and combining with $\theta_\tau \geq \epsilon$ gives

$$\log\epsilon \;\leq\; \log(\theta_\tau) \;\leq\; \log(\theta_0) \;+\; \tau\,\log\!\left(1 - b_d\,\epsilon^2\right) \;\leq\; \log(\theta_0) \;-\; \tau\,b_d\,\epsilon^2.$$

Rearranging completes the proof.

Bibliography

Balakrishnan, Sivaraman, Martin J. Wainwright, and Bin Yu (2017). "Statistical guarantees for the EM algorithm: From population to sample-based analysis". In: *The Annals of Statistics* 45.1, pp. 77 –120. DOI: 10.1214/16-AOS1435. URL: https://doi.org/10.1214/16-AOS1435.

Cai, T. Tony, Jing Ma, and Linjun Zhang (2019). "CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality". In: *The Annals of Statistics* 47.3, pp. 1234 –1267. DOI: 10.1214/18-AOS1711. URL: https://doi.org/10.1214/18-AOS1711.

Chen, Yudong et al. (2024). "Local Minima Structures in Gaussian Mixture Models". In: *IEEE Transactions on Information Theory*.

Dasgupta, Sanjoy and Leonard Schulman (2013). "A two-round variant of em for gaussian mixtures". In: *arXiv preprint arXiv:1301.3850*.

Daskalakis, Constantinos, Christos Tzamos, and Manolis Zampetakis (2017). "Ten steps of EM suffice for mixtures of two Gaussians". In: *Conference on Learning Theory*. PMLR, pp. 704–710.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data Via the EM Algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22. DOI: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. USA: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/V1/N19-1423.

Devroye, Luc, László Györfi, and Gábor Lugosi (2013). *A probabilistic theory of pattern recognition*. Vol. 31. New York: Springer Science & Business Media.

Dwivedi, Raaz et al. (2018). "Theoretical guarantees for EM under misspecified Gaussian mixture models". In: *Advances in Neural Information Processing Systems* 31.

Dwivedi, Raaz et al. (2020a). "Sharp Analysis of Expectation-Maximization for Weakly Identifiable Models". In: *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. USA: PMLR, pp. 1866–1876. URL: http://proceedings.mlr.press/v108/dwivedi20a.html.

Dwivedi, Raaz et al. (2020b). "Singularity, misspecification and the convergence rate of EM". In: *The Annals of Statistics* 48.6, pp. 3161–3182.

Ghosal, Subhashis and Aad W. van der Vaart (2001). "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities". In: *The Annals of Statistics* 29.5, pp. 1233 –1263. DOI: 10.1214/aos/1013203452. URL: https://doi.org/10.1214/aos/1013203452.

Klusowski, Jason M and WD Brinda (2016). "Statistical guarantees for estimating the centers of a two-component Gaussian mixture by EM". In: *arXiv preprint arXiv:1608.02280*.

McInnes, Leland et al. (2018). "UMAP: Uniform Manifold Approximation and Projection". In: *Journal of Open Source Software* 3.29, p. 861. DOI: 10.21105/joss.00861. URL: https://doi.org/10.21105/joss.00861.

Rousseau, Judith and Kerrie Mengersen (Aug. 2011). "Asymptotic Behaviour of the Posterior Distribution in Overfitted Mixture Models". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 73.5, pp. 689–710. ISSN: 1369-7412. DOI: 10.1111/j.1467-9868.2011.00781.x.

Segol, Nimrod and Boaz Nadler (2021). "Improved convergence guarantees for learning Gaussian mixture models by EM and gradient EM". In: *Electronic journal of statistics* 15.2, pp. 4510–4544.

Xu, Weihang, Maryam Fazel, and Simon S Du (2024). "Toward Global Convergence of Gradient EM for Over-Parameterized Gaussian Mixture Models". In: *arXiv preprint arXiv:2407.00490*.

Yan, Bowei, Mingzhang Yin, and Purnamrita Sarkar (2017). "Convergence of gradient EM on multi-component mixture of Gaussians". In: *Advances in Neural Information Processing Systems* 30.

Zhao, Ruofei, Yuanzhi Li, and Yuekai Sun (2020). "Statistical convergence of the EM algorithm on Gaussian mixture models". In: *Electronic Journal of Statistics* 14, pp. 632 –660. URL: https://doi.org/10.1214/19-EJS1660.