A Review of the Expectation-Maximization Algorithm and its Applications to Mixture Models

Quentin Fruytier, School of Mathematics and Statistics

McGill University, Montreal

August, 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Master of Mathematics and Statistics

Abstract

The Expectation-Maximization (EM) algorithm has long been recognized as a powerful tool for approximating the maximum likelihood estimator in parametric models with latent variables. This thesis provides a selective survey of the existing EM literature, spanning from its original formulation in the 1970s to its present-day developments, with the objective of creating a valuable resource for future research. By exploring the evolution of the EM algorithm, we present both earlier and recent results as well as practical applications in mixture models. Chapter 1 serves as a thorough introduction to the EM, contextualizing it within the broader framework of parameter estimation in parametric models with latent variables. In Chapter 2, we study the general convergence properties of the EM; in particular, we present conditions under which the algorithm's fitted iterates converge inside a ball centered around the true parameter of the model. Meanwhile, in Chapter 3, we survey the existing literature on the EM algorithm as it relates to Gaussian mixture models and mixed linear regression models. Finally, in Chapter 4, we conclude with a discussion on important aspects such as initialization, SNR, parameterization, and new research directions for the EM algorithm. By collating the wealth of knowledge available on the EM algorithm, this thesis offers researchers a valuable reference for understanding, applying, and advancing the EM algorithm.

Abrégé

L'algorithme Expectation-Maximization (EM) est depuis longtemps reconnu comme un outil puissant pour approximer les parametres de modèles paramétriques avec des variables latentes. Cette thèse propose une revue sélective de la littérature existante sur l'EM, couvrant sa formulation originale dans les années 1970 jusqu'à ses développements actuels, dans le but de créer une ressource précieuse pour la recherche future. En explorant l'évolution de l'EM, nous présentons à la fois des résultats antérieurs et récents, des applications pratiques et des exemples numériques ou appropriés. Le chapitre 1 sert d'introduction approfondie à l'algorithme dans le cadre de l'estimation des paramètres de modèles paramétriques. Dans le chapitre 2, nous étudions les propriétés générales de convergence de l'EM ; en particulier, nous présentons les conditions dans lesquelles les itérations de l'algorithme convergent autours des vrais paramètres du modèle. Parallèlement, dans le chapitre 3, nous passons en revue la littérature existante sur l'EM appliqué aux modèles de mélange de gaussiennes et aux modèles de régression linéaire mixtes. Enfin, dans le chapitre 4, nous concluons par une discussion sur des aspects importants tels que l'initialisation, le SNR, la paramétrisation, et les nouvelles orientations de recherche pour l'algorithme EM. En rassemblant et en analysant les connaissances disponibles sur l'algorithme EM, cette thèse offre aux chercheurs une référence précieuse pour comprendre, appliquer et faire progresser notre comprehension de l'EM.

Acknowledgements

I would like to express my sincere gratitude to my two supervisors, Professor Tim Hoheisel and Professor Abbas Khalili, for their invaluable support and guidance throughout my research. Their expertise and mentorship have been instrumental in the successful completion of this thesis. I am deeply grateful for their encouragement, patience, and unwavering support. Without their guidance, this journey would not have been possible. Thank you.

Table of Contents

List of Figures

Chapter 1

Introduction

In this Chapter, we commence by providing a concise historical overview of the research on the EM algorithm, highlighting its evolution and significance in the field of parameter estimation. Building upon this foundation, we delve into the core problem of parameter estimation in parametric latent variable models. We define the key components of the estimation process, including the probability densities, the log-likelihood function, and the method of maximum likelihood estimation. With a solid understanding of the underlying principles, we motivate the EM, then unveil its iterative steps. Moreover, we extend the discussion to encompass other relevant algorithms commonly employed for parameter estimation in parametric models with latent variables. Throughout this chapter, we reinforce our explanations and insights by providing detailed examples within the realm of Gaussian mixture models (GMMs) and mixed linear regression (MLR) models, further solidifying the conceptual understanding and practical applications of the EM algorithm.

1.1 History of the EM Algorithm

The problem of accurately estimating parameters in the presence of missing information has puzzled statisticians for almost a century. The EM algorithm, formally presented in 1977, has become widely-used for fitting parametric models with latent variables due to its intuitive nature and appealing algorithmic properties (see Chapter 2). However, the mathematical details surrounding its convergence have been limited – until recently.

Early appearances of the EM can be traced back to the 1950s. Hartley [16, 1958] simplified and unified older known techniques for parameter estimation in latent variable models. He further provided practical examples such as modelling the pollution of Phleum pratense seeds by the presence of weed seeds where the complete data was assumed to follow a Poisson distribution. Orchard and Woodbury [27, 1972] studied parameter estimation with latent variables for mixed linear regression, and mixtures of k-multivariate Gaussians. In these early works, the contributions were iterative algorithms whose iterations are no-more difficult to perform than parameter estimation of parametric models with no missing data – which is well understood. Moreover, they introduced the first appearances of the Expectation (E) and Maximization (M) steps that gave the EM algorithm its name.

It wasn't until 1977 that Dempster et al. [9] generalized previous works (such as Hartley [16, 1958], Baum et al. [2, 1970], Hartley and Hocking [17, 1971], Orchard and Woodbury [27, 1972], and Sundberg [35, 1974]) into a more broadly applicable algorithm for computing maximum likelihood estimates based on incomplete data. They introduced the general form of the EM and proved several important results (see Section 2.2). Ever since, the EM algorithm has been used for a wide variety of problems including but not limited to density estimation, clustering, and regression.

In the years following its introduction, numerous papers were published on the EM providing details on its convergence properties and diverse applications. Early work focused on the convergence of the EM's fitted parameter to local optima or stationary points of the log-likelihood function. Notably, Wu [37, 1983] rectified an error in the original 1977 paper [9, Theorem 2] where the proof made incorrect use of the triangle inequality. In addition, he proved several convergence properties of the EM which we break apart for the reader in Section 2.3. Some years later, Tseng [36, 2004], developed an "entropy-like" proximal point iteration, of which the EM algorithm is a special case, which allowed for more intuitive analysis of the convergence properties. In 2014, Balakrishnan et al. [1] introduced a novel framework for analyzing the local convergence of the EM algorithm, establishing conditions under which the algorithm's iterates converge to the true parameter of the model (see Section 2.4). Their work laid the foundation for subsequent advancements in the field, employing a combination of concentration of measure inequalities and other techniques to bridge the results from population EM to finite-sample EM (see Algorithm 3 and Algorithm 1).

GMMs have been a particularly prominent domain for the application of the EM algorithm. Dempster et al. [1] and Tseng [36] initially included GMMs as practical examples, but it was Balakrishnan et al. [1] who first demonstrated the local convergence of the EM's fitted parameters to the true parameter in a 2-component symmetric Gaussian mixture with unknown mean parameter. Subsequently, Kwon et al. [21, 2020] extended this result to k-component spherical Gaussian mixtures with unknown mean parameters. Dwivedi et al. [11, 2018] analyzed the local convergence of the EM algorithm to the true parameter for over-parametrized 2-component symmetric Gaussian mixtures. For the same class of models, Dwivedi et al. [12, 2018] demonstrated that the EM algorithm retains its fast convergence up to certain constants in the under-specified setting.

While Gaussian mixture models have garnered significant attention, they are not the sole focus of research on mixture models. The class of MLR models has been extensively studied in the past two decades. Balakrishnan et al. [1] made notable contributions to the analysis of symmetric mixtures of two linear regressions, where only the solution parameter is unknown. They characterized the local convergence properties, initially in the high signal-to-noise (SNR) ratio regime. Subsequently, Kwon et al. [23, 2020] obtained tighter bounds in the high SNR regime and extended the analysis to the middle and low SNR regimes. What's more, they provided local convergence results in the same settings when the mixing weights are additionally unknown. one year prior, Kwon et al. [22] explored the local convergence properties of the EM for k-component MLR with only the variance parameter unknown, utilizing the sample-splitting variant of the EM algorithm. Historically, results for MLR appear after similar findings for Gaussian mixtures emerge. Our literature survey suggests that the EM works in similar ways on the two mixture models as similarities in the rates of convergence for both mixture models are consistent throughout. Unlike Gaussian mixtures, the over-parameterized and under-parameterized MLR setting is – to our knowledge – yet to receive a proper look.

Very recent advancements have further enriched the research landscape surrounding the EM algorithm. Ho et al. [18, 2022] extended the framework introduced by Balakrishnan et al. [1] and addressed cases where the EM algorithm exhibits instability. Meanwhile Kunstner et al. [20, 2022] presented an alternative perspective on the EM algorithm in the context of the exponential family of distributions, conceptualizing it as a mirror descent algorithm. Doing so, they obtained non-asymptotic convergence properties that are invariant of the choice of parameterization.

With the growing popularity of the EM algorithm in the field of machine learning, an increasing number of papers have been published on the topic. Consequently, it has become challenging to discern the contributions made under various conditions. In this thesis, we aim to selectively review the extensive literature spanning several decades of research, providing a comprehensive resource on the EM algorithm for current and future researchers, with a particular emphasis on mixture models.

1.2 Notation

We introduce several notations that will be used throughout this thesis. The set of positive semi-definite matrices on $\mathbb{R}^{d\times d}$ is denoted as $\mathbb{S}^d_+$, while the set of positive definite matrices is denoted as $\mathbb{S}^d_{++}$. We use $\mathbb{R}_+$ to represent the interval $[0,\infty)$, $\mathbb{R}_{++}$ to represent $(0,\infty)$, and $\bar{\mathbb{R}}$ to represent $[-\infty,\infty]$. The notation $[k]$ refers to the set $\{0,1,2,...,k-1\}$. For a vector $x\in\mathbb{R}^d$, we denote $\|x\|_2$ to be the euclidean norm of x. For a matrix $M\in\mathbb{R}^{m\times n}$, we denote $\|M\|_F$ to be the Frobenius norm of M, if M is invertible we denote M^{-1} as its inverse and $|M|$ as its determinant. For a set $A\in\mathbb{R}^s$, we denote $\operatorname{int} A$ to be the interior of A and

$\mathcal{P}(A)$ to be the power set of A. Further, for a continuous function $f(x,y) : \mathbb{R}^d \times \mathbb{R}^s \to \mathbb{R}$, we denote $\nabla_1 f(x,y) = \frac{d}{dx} f(x,y)$ and $\nabla_2 f(x,y) = \frac{d}{dy} f(x,y)$. The notation $x \lesssim y$ signifies that x is smaller than or equal to y up to logarithmic factors. Similarly, $x \gtrsim y$ is used to indicate that x is greater than or equal to y up to logarithmic factors. Lastly, we employ the notation y is $\tilde{\mathcal{O}}(x)$ to express that y is $\mathcal{O}(x)$ up to logarithmic factors. The same logic applies to $\tilde{\Omega}$.

1.3 Parameter Estimation with Latent Variables

Parameter estimation is a fundamental problem in statistical inference, with numerous applications in machine learning, data analysis, and other fields. The aim is to recover the true parameter θ^* of a parametric model from the feasible set

$$\Omega := \{\theta \in \mathbb{R}^s : \theta \text{ is a possible vector of parameters}\}.$$

In particular, we assume that Ω is convex for the remainder of this thesis, unless specified otherwise. In this thesis, we deal – specifically – with latent variable parametric models, meaning that some of the data is missing or unobserved. This presents a challenge for traditional parameter estimation methods, such as the method of maximum likelihood estimation which we introduce in Section 1.3.3.

1.3.1 Parametric Models With Latent Variables

We assume the reader has a prior basic understanding of the notion of a random variables (RV) and its probability density. To distinguish between an RV and its realization, we use capital letters and lower-case letters, respectively. For example, Y_i represents an RV while y_i is its realization. Further, We may use the capital letter without its index to mean the same distribution: Y_i and Y have the same distribution. We formally introduce latent variable parametric models below.

Suppose the data we observe is $(y_1, ..., y_n)$ and denote $y_i \in \mathbb{R}^d$ to be the i^{th} of n observations. We denote $(z_1, ..., z_n)$ as the latent unobserved portion of the data where $z_i \in \mathbb{R}^p$

is the i^{th} of n unseen observations. The complete data is written as $((y_1, z_1), ..., (y_n, z_n))$ where $(y_i, z_i) \in \mathbb{R}^d \times \mathbb{R}^p$ is the i^{th} of n complete observations. We define

$$\mathbb{H} : y \mapsto \mathbb{H}(y) := \{z \in \mathbb{R}^p : z \text{ can be sampled from } Z|Y = y\} \tag{1.1}$$

to be the map which takes an observation and outputs the set of all its possible corresponding latent variables.

We make the following standard assumptions for the data. First, $Y_1, ..., Y_n$ are independent and identically distributed (i.i.d) and so are $Z_1, ..., Z_n$. Also, the joint RV of the complete data (Y, Z) has a pdf $f_{\theta^*} : \mathbb{R}^d \times \mathbb{R}^p \mapsto \mathbb{R}_+$ where $\theta^* \in \Omega$ is the true parameter of the model; in particular, for all $\theta \in \Omega$, we assume $f_\theta(y, z)$ exists for all (y,z) in the sample space and $f_\theta(y, z) > 0$ almost everywhere on the sample space. Similarly, Y also has the pdf

$$g_{\theta^*}(y) := \int_{\mathbb{H}(y)} f_{\theta^*}(y, s) ds \tag{1.2}$$

which maps $\mathbb{R}^d \mapsto \mathbb{R}_+$. We make the remark that in cases where the missing latent variable is discrete, the integral term becomes a summation over the support. Lastly, the conditional RV $Z|Y$ has the conditional pdf

$$k_{\theta^*}(z|y) := \frac{f_{\theta^*}(y, z)}{g_{\theta^*}(y)} \tag{1.3}$$

which maps $\mathbb{R}^d \times \mathbb{R}^p \mapsto \mathbb{R}_+$. If the latent variable is discrete, k_{θ^*} is a probability density mass function (pmf) instead. We provide two detailed and practical examples of latent variable models below.

Example 1 (k-component d-dimensional Gaussian mixture models (GMMs))**.** *Gaussian mixture models of k-components in d-dimensions are parametric models with latent variables wherein observations $(y_1, ..., y_n) \in \mathbb{R}^{d \times n}$ are sampled from a linear combination of k independent Gaussian components and the latent variable $z \in [k] := \{0, 1, ..., k-1\}$ is the discrete label expressing which Gaussian component an observation was sampled from. Formally, the observa-*

tions are sampled from

$$Y \sim \sum_{j=0}^{k-1} \pi_j^* \mathcal{N}(\mu_j^*, \Sigma_j^*) \tag{1.4}$$

where $\pi_j^* \in [0,1]$ for all $j \in [k]$ are the mixing weights satisfying $\sum_{j=0}^{k-1} \pi_j^* = 1$ and $\mathcal{N}(\mu_j^*, \Sigma_j^*)$ is a multivariate Gaussian distribution with mean $\mu_j^* \in \mathbb{R}^d$ and covariance matrix $\Sigma_j^* \in \mathbb{S}_{++}^d$. For this class of parametric models, $\mathbb{H}(y)$ is given as

$$\mathbb{H}(y) := \{0, 1, ..., k-1\}$$

and the relevant pdfs and pmfs are provided below as

$$\begin{aligned}
&\textit{The pdf of } \mathcal{N}(\mu_j^*, \Sigma_j^*) \textit{ is} && \mathcal{G}(y; \mu_j^*, \Sigma_j^*) := && \frac{\exp\{-\frac{1}{2}(y-\mu_j^*)^T \Sigma_j^{*-1}(y-\mu_j^*)\}}{(2\pi)^{\frac{d}{2}} |\Sigma_j^*|^{\frac{1}{2}}}; \quad (1.5)\\
&\textit{the pdf of } Y \textit{ is} && g_{\theta^*}(y) := && \sum_{j=0}^{k-1} \pi_j^* \mathcal{G}(y; \mu_j^*, \Sigma_j^*); \quad (1.6)\\
&\textit{the pmf of } Z \textit{ is} && p_{\theta^*}(z) := && \pi_z^*; \quad (1.7)\\
&\textit{the pmf of } Z|Y \textit{ is} && k_{\theta^*}(z|y) := && \frac{\pi_z^* \mathcal{G}(y; \mu_z^*, \Sigma_z^*)}{\sum_{j=0}^{k-1} \pi_j^* \mathcal{G}(y; \mu_j^*, \Sigma_j^*)}; \quad (1.8)\\
&\textit{the pdf of } Y|Z \textit{ is} && v_{\theta^*}(y|z) := && \mathcal{G}(y; \mu_z^*, \Sigma_z^*); \quad (1.9)\\
&\textit{the pdf of } (Y, Z) \textit{ is} && f_{\theta^*}(y, z) := && \pi_z^* \mathcal{G}(y; \mu_z^*, \Sigma_z^*). \quad (1.10)
\end{aligned}$$

In this context, the true parameter vector $\theta^* = (\pi_j^*, \mu_j^*, \Sigma_j^*)_{j \in [k]}$ fully describes the mixture.

One may ask why we restrict Σ_j^* to the open set $\mathbb{S}_{++}^d$ when it is well known that the covariance matrix of a multivariate Gaussian distribution can be singular, therefore belonging to the bigger and closed set $\mathbb{S}_{+}^d$. This is because in the case where the covariance matrix is singular, the multivariate Gaussian distribution is degenerate; it does not have a density with respect to the k-dimensional Lebesgue measure. It is easiest to see why in (1.5) *where Σ_j^{*-1} does not exist when Σ_j^* is singular. One could restrict such a d-dimensional Gaussian to $rank(\Sigma_j^*)$-dimensions in favor of a new distribution whose covariance matrix is now positive definite. However, this is not advisable for d-dimensional Gaussian mixtures where $k > 1$ since there are more than one Gaussian component. This is because information can be lost from the other components that have a covariance* *matrix that is already full rank. Still, the Disintegration Theorem makes it possible to define the density in the case where Σ_j^* is singular (see [13] and [28]). However, we will not use it in this thesis.*

Example 2 (k-component d-dimensional mixed linear regression models (MLR))**.** *Mixtures of k-linear regressions in d-dimensions are parametric models with latent variables wherein observations $((y_1, x_1), ..., (y_n, x_n))$ are sampled from a linear combination of k independent linear regression components and the latent variable $z \in [k] = \{0, 1, ..., k-1\}$ is the discrete label expressing which regression component an observation was sampled from. In particular $y \in \mathbb{R}$ is the response variable while $x \in \mathbb{R}^d$ are the covariates. Formally, we assume the observations (y,x) are sampled from*

$$Y \sim \sum_{j=0}^{k-1} \pi_j^* N(\langle X, \mu_j^* \rangle, \sigma_j^{*2}) \tag{1.11}$$

$$X \sim \mathcal{N}(0, I_d) \tag{1.12}$$

where $\pi_j^* \in [0,1]$ for all $j \in [k]$ are the mixing weights satisfying $\sum_{j=0}^{k-1} \pi_j^* = 1$ and $N(\langle X, \mu_j^* \rangle, \sigma^*)$ is a univariate Gaussian distribution with mean $\langle X, \mu_j^* \rangle$ and variance $\sigma_j^* \in \mathbb{R}_{++}$. For this class of parametric models, $\mathbb{H}(y, x)$ is given as

$$\mathbb{H}(y) := \{0, 1, ..., k-1\}$$

and the relevant pdfs and pmfs are provided below as

$$\textit{pdf of } N(\langle x, \mu_j^* \rangle, \sigma^*)\text{:} \qquad G(y; \mu_j^*, \sigma_j^*) := \frac{\exp\{-\frac{(y-\langle x,\mu_j^*\rangle)^2}{2\sigma_j^{*2}}\}}{\sigma_j^*\sqrt{2\pi}}; \tag{1.13}$$

$$\textit{pdf of } (Y, X)\text{:} \qquad g_{\theta^*}(y, x) := \left[\sum_{j=0}^{k-1} \pi_j^* G(y; \langle x, \mu_j^* \rangle, \sigma_j^{*2})\right] \mathcal{G}(x; 0, I_d); \tag{1.14}$$

$$\textit{pmf of } Z\text{:} \qquad p_{\theta^*}(z) := \pi_z^*; \tag{1.15}$$

$$\textit{pmf of } Z|(Y, X)\text{:} \qquad k_{\theta^*}(z|y, x) := \frac{\pi_z^* G(y; \langle x, \mu_z^* \rangle, \sigma_z^{*2})}{\sum_{j=0}^{k-1} \pi_j^* G(y; \langle x, \mu_j^* \rangle, \sigma_j^{*2})}; \tag{1.16}$$

$$\textit{pdf of } (Y, X)|Z\text{:} \qquad v_{\theta^*}(y, x|z) := G(y; \langle x, \mu_z^* \rangle, \sigma_z^{*2}) \mathcal{G}(x; 0, I_d); \tag{1.17}$$

$$\textit{pdf of } (Y, X, Z)\text{:} \qquad f_{\theta^*}(y, x, z) := \pi_z^* G(y; \langle x, \mu_z^* \rangle, \sigma_z^{*2}) \mathcal{G}(x; 0, I_d); \tag{1.18}$$

where $\mathcal{G}$ is the pdf of a multivariate Gaussian given as (1.5). *In this context, the true parameter vector $\theta^* := (\pi_j^*, \mu_j^*, \sigma_j^*)_{j \in [k]}$ fully describes the mixture.*

1.3.2 Log-likelihood Function

The likelihood function is the function whose output is the probability of observing the sample data viewed as a function of the model parameters $\theta \in \Omega$. For latent variable models, the log-likelihood function,

$$\begin{aligned} L_n(\theta) &:= \frac{1}{n} \sum_{i=1}^{n} [\log(g_\theta(y_i))] \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[\log\left(\int_{\mathbb{H}(y_i)} f_\theta(y_i, s) ds\right)\right], \end{aligned} \tag{1.19}$$

is a convenient alternative to the likelihood function as it allows us to work with summations instead of products. Meanwhile, the population log-likelihood function,

$$\begin{aligned} L(\theta) &:= \mathbb{E}_{\theta^*}[\log(g_\theta(Y))] \\ &= \int_{\mathbb{R}^d} \log(g_\theta(s)) g_{\theta^*}(s) ds, \end{aligned} \tag{1.20}$$

which can be thought of as taking $n \to \infty$ in (1.19), is considered solely for analysis purposes.

1.3.3 Maximum Likelihood Estimation

Defined as the parameter value that maximizes the likelihood function, the maximum likelihood estimator (MLE) is widely-used for estimating the true parameter of a statistical model. In the case of a parametric model with latent variables, the MLE is given as

$$\hat{\theta}_n \in \arg\max_{\theta \in \Omega} L_n(\theta) \overset{(1.19)}{=} \arg\max_{\theta \in \Omega} \frac{1}{n} \sum_{i=1}^{n} \left[\log \left(\int_{\mathbb{H}(y_i)} f_\theta(y_i, s) ds \right) \right]. \tag{1.21}$$

It is well known that, in latent variable models, the MLE is often intractable due to the non-concave nature of the log-likelihood function [23]. This is in large due to the appearance of the integral term in (1.19) making direct optimization with respect to $\theta \in \Omega$ more difficult. To clarify, this integral term is not present for MLE in parametric models where the observed data is complete. To remedy this, optimization algorithms are used to approximate the MLE. Among them is the EM algorithm – whose iterations are no-more difficult than computing the MLE with no missing data. We provide two practical examples below.

Example 3 (Parameter estimation of GMMs)**.** *We consider the problem of estimating the true parameter of the parametric model described in Example 1* ***when the covariance matrices $(\Sigma_j^*)_{j \in [k]}$ are known****. For this task, we turn to the MLE defined in* (1.21)*. First, assuming the covariance matrices $(\Sigma_j^*)_{j=0}^{k-1}$ are known, we specify the set of feasible parameters as*

$$\Omega = \{(\pi_j, \mu_j)_{j=0}^{k-1} : \sum_{j=0}^{k-1} \pi_j = 1;\ \pi_j \in [0, 1], \text{ and } \mu_j \in \mathbb{R}^d, \text{for all } j \in [k]\} \tag{1.22}$$

and make the remark that Ω is closed, convex, but not bounded. Next, we derive the log-likelihood function according to its definition in (1.19)*; the function evaluates to*

$$L_n((\pi_j, \mu_j)_{j=0}^{k-1}) := \frac{1}{n} \sum_{i=1}^{n} \left[\log \left(\sum_{j=0}^{k-1} \pi_j \mathcal{G}(y_i; \mu_j, \Sigma_j^*) \right) \right]. \tag{1.23}$$

At this stage, we would ideally like to maximize L_n over Ω directly, yielding the MLE of the model. However, this cannot be done directly, in large part, due to the 'log of a sum' term in (1.23)*. Were the missing labels known, this summation term would disappear and the maximization would be tractable. To remedy this, we will need to consider optimization algorithms to approximate the MLE of the model (see Example 5).*

Example 4 (Parameter estimation of MLR models)**.** *We consider the problem of estimating the true parameter of the parametric model described in Example 2. For this task, we turn to the MLE defined in* (1.21)*. First, we specify the set of feasible parameters as*

$$\Omega := \{(\pi_j, \mu_j, \sigma_j)_{j=0}^{k-1} : \sum_{j=0}^{k-1} \pi_j = 1; \pi_j \in [0,1],\, \mu_j \in \mathbb{R}^d, \textit{ and } \sigma_j^* \in \mathbb{R}_{++}, \textit{for all } j \in [k]\} \tag{1.24}$$

and make the remark that Ω is convex, but open and unbounded. Next, we derive the log-likelihood function according to its definition in (1.19)*; the function evaluates to*

$$L_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \left[\log \left(\sum_{j=0}^{k-1} \pi_j G(y_i; \langle x_i, \mu_j \rangle, {\sigma_j}^2) \mathcal{G}(x_i; 0, I_d) \right) \right]. \tag{1.25}$$

At this stage, we would ideally like to maximize L_n over Ω directly, yielding the MLE of the model. However, this cannot be done directly, in large part, due to the 'log of a sum' term in (1.25)*. Were the missing labels known, this summation term would disappear and the maximization would be tractable. To remedy this, we will need to consider optimization algorithms to approximate the MLE of the model (see Example 6).*

We may wonder why the MLE is so appealing for recovering the true parameter of a statistical model. It was established in [7] that the true parameter vector of a statistical model belongs to the set of global maximizers of the population log-likelihood given as (1.20). Finally, under certain regularity conditions outlined in [19, Theorem 6.1.3], the MLE converges in probability to the true parameters; that is, $\hat{\theta}_n \underset{n\to\infty}{\xrightarrow{p}} \theta^*$. This property, known in statistics as consistency, guarantees the MLE $\hat{\theta}_n$ to approach the true parameter of the model for large n. Because of this, the method of maximum likelihood estimation is, by and large, the most popular approach for estimating the true parameter of a statistical model.

1.4 Expectation-Maximization (EM) Algorithm

The EM algorithm is an iterative algorithm used to approximate the MLE in parametric models with incomplete data. In this section, we formally introduce the EM algorithm; we make the distinction between the finite sample EM (Algorithm 1) which is used in practice, the sample-splitting EM (Algorithm 2) which splits the data set so as to have iterations based on independent sub-samples, the fully deterministic population EM algorithm (Algorithm 3) indispensable for the analysis of convergence of the previous two algorithms, and finally, the General EM (Algorithm 4) where the requirement for global maximization in the M-step is relaxed.

1.4.1 Finite-Sample EM (EM)

Introduced in 1977 by Dempster et al. [9], the finite-sample EM (see Algorithm 1 below) is the sate-of-the-art for approximating the MLE of latent variable parametric models. The EM is an iterative algorithm that alternates between the E-step and M-step. The E-step, is analogous to forming a complete data log-likelihood by giving conditional probabilities or weights to the unknown latent variables $(z_1, ..., z_n)$. On the other hand, the M-step consists in finding the global maximizers of the expression obtained from the E-step. For ease of writing, we may write the EM's iterations with respect to the EM operator function

$$M_n(\theta^{(t)}) := \arg\max_{\theta\in\Omega} Q_n(\theta|\theta^{(t)}) \tag{1.26}$$

which maps $\Omega \mapsto \mathcal{P}(\Omega)$ and where Q_n is given below as (1.27). We make the remark that solutions to the above optimization problem are not always guaranteed to exist in Ω.

Algorithm 1 Finite-Sample EM

```
**Input**: Initial $\theta^{(0)} \in \Omega$, $\{y_1, ..., y_n\}$
**for** $t = 0, ..., T-1$ **do**
    **E-Step**: Let $Q_n(\theta|\theta^{(t)}) \overset{(1.27)}{=} \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\theta^{(t)}}[\log f_\theta(Y_i, Z_i)|Y_i = y_i]$
    **M-Step:** Let $\theta^{(t+1)} \in M_n(\theta^{(t)}) = \arg\max_{\theta\in\Omega} Q_n(\theta|\theta^{(t)})$
**end for**
**Output**: $\theta^{(T)}$
```

Provided a sample of size n, each iteration of the EM revolves around the function

$$\begin{aligned} Q_n(\phi|\theta) &:= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta[\log f_\phi(Y_i, Z_i)|Y_i = y_i] \\ &= \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{H}(y_i)} \log(f_\phi(y_i, s))k_\theta(s|y_i)ds \end{aligned} \tag{1.27}$$

which maps $\Omega \times \Omega \mapsto \mathbb{R}$. We see that unlike (1.19), $Q_n(\theta'|\theta^{(t)})$ has the integral outside the log. The advantage of this form was pointed out by Hartley [16, 1958] and is most well seen in the M-step as each iteration is akin to calculating the MLE of a parametric model with no missing data, a simpler problem. We provide two practical examples below.

Example 5 (Iterations of the EM for GMMs when the variance is known)**.** *We continue the estimation in Example 3 and obtain an approximation using the EM algorithm (see Algorithm 1). Since we assume the covariance matrices $(\Sigma_j^*)_{j\in[k]}$ are known, we denote $\theta := (\pi_j, \mu_j)_{j\in[k]} \in \Omega$ and $\theta^{(t)} := (\pi_j^{(t)}, \mu_j^{(t)})_{j\in[k]} \in \Omega$ where Ω is given as* (1.22)*. We begin with the evaluation of Q_n,* *deriving it as follows:*

$$\begin{aligned} Q_n(\theta|\theta^{(t)}) &= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=0}^{k-1}\log(f_\theta(y_i,j))k_{\theta^{(t)}}(j|y_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=0}^{k-1}\left(\log\left(\frac{\pi_j}{(2\pi)^{\frac{d}{2}}|\Sigma_j^*|^{\frac{1}{2}}}\right)-\frac{1}{2}(y_i-\mu_j)^T{\Sigma_j^*}^{-1}(y_i-\mu_j)\right)k_{\theta^{(t)}}(j|y_i)\right]. \end{aligned} \tag{1.28}$$

Now, we are ready to evaluate the EM operator $M_n(\theta)$ given as (1.26)*. Fortunately, the EM operator exists and has a closed form solution for the model that is given as*

$$M_n(\theta^{(t)}) = (\pi_j^{(t+1)}, \mu_j^{(t+1)})_{j\in[k]} \tag{1.29}$$

where for all $j \in [k]$,

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} k_{\theta^{(t)}}(j|y_i)}{n}, \tag{1.30}$$

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} y_i k_{\theta^{(t)}}(j|y_i)}{\sum_{i=1}^{n} k_{\theta^{(t)}}(j|y_i)}. \tag{1.31}$$

The complete derivation is given in Section A.1.1 of the Appendix.

Example 6 (Iterations of the EM for MLR models)**.** *We continue the estimation in Example 4 and obtain an approximation using the EM algorithm (see Algorithm 1). We denote $\theta := (\pi_j, \mu_j, \sigma_j)_{j\in[k]} \in \Omega$ and $\theta^{(t)} := (\pi_j^{(t)}, \mu_j^{(t)}, \sigma_j^{(t)})_{j\in[k]} \in \Omega$ where Ω is given as* (1.24)*. We begin with the evaluation of Q_n, deriving it as follows:*

$$\begin{aligned} Q_n(\theta|\theta^{(t)}) &= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=0}^{k-1}\log(f_\theta(y_i,x_i,j))k_{\theta^{(t)}}(j|y_i,x_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=0}^{k-1}\left(\log\left(\frac{\pi_j\mathcal{G}(x_i;0,I_d)}{\sqrt{2\pi\sigma_j^2}}\right)-\frac{(y_i-\langle x_i,\mu_j\rangle)^2}{2\sigma_j^2}\right)k_{\theta^{(t)}}(j|y_i,x_i)\right]. \end{aligned} \tag{1.32}$$

Now we are ready to evaluate the EM operator $M_n(\theta)$ given as (1.26)*. Fortunately, the EM operator exists and has a closed form solution for the model. It is given in [22] as*

$$M_n(\theta^{(t)}) = (\pi_j^{(t+1)}, \mu_j^{(t+1)}, \sigma_j^{(t+1)})_{j=0}^{k-1} \tag{1.33}$$

where for all $j \in [k]$,

$$\begin{aligned}
\pi_j^{(t+1)} &= \frac{\sum_{i=1}^n k_{\theta^{(t)}}(j|y_i, x_i)}{n}, && (1.34)\\
\mu_j^{(t+1)} &= \left(\sum_{i=1}^n k_{\theta^{(t)}}(j|y_i, x_i) x_i x_i^T\right)^{-1} \left(\sum_{i=1}^n k_{\theta^{(t)}}(j|y_i, x_i) y_i x_i\right), && (1.35)\\
\sigma_j^{(t+1)} &= \sqrt{\frac{\sum_{i=1}^n (y_i - \langle x_i, \mu_j^{(t+1)}\rangle)^2 k_{\theta^{(t)}}(j|y_i, x_i)}{\sum_{i=1}^n k_{\theta^{(t)}}(j|y_i, x_i)}}. && (1.36)
\end{aligned}$$

One drawback of the EM is that the initial estimate $\theta^{(0)}$ is not specified by the algorithm. It is often not clear how to find these initial estimates. As a result, the initialization of the EM can be critical to whether a global or only local maximum is attained (see Section 2.4.2). In Section 4.2 we provide a brief discussion on the topic of initialization for the EM algorithm in the context of mixture models.

Because the EM is an algorithm used for approximating the MLE of parametric models with latent variables, we aim to obtain the maximizers of the log-likelihood function L_n given as (1.19). So why don't we see L_n in the iterations of Algorithm 1? It happens that Q_n is intimately connected to L_n as follows:

$$Q_n(\phi|\theta) := L_n(\phi) + H_n(\phi|\theta)\ \forall \theta, \phi \in \Omega. \tag{1.37}$$

Where if Q_n is the expected complete data log-likelihood at the current parameter estimate,

$$\begin{aligned}
H_n(\phi|\theta) &:= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta[\log k_\phi(Z_i|Y_i)|Y_i = y_i]\\
&= \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{H}(y_i)} \log(k_\phi(s|y_i)) k_\theta(s|y_i) ds
\end{aligned} \tag{1.38}$$

is the expected log of the conditional probabilities of the latent variables at the current parameter estimate. Relationship (1.37), proven below, is essential for understanding the EM and will be used throughout this thesis.

Proof of (1.37). Let $\theta, \phi \in \Omega$, we first expand Q_n according to (1.27).

$$Q_n(\phi|\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta[\log(f_\phi(Y_i, Z_i))|Y_i = y_i]$$

Next, we expand $f_\phi(\cdot)$ according to (1.3) and use properties of logarithms.

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta[\log(f_\phi(Y_i, Z_i))|Y_i = y_i] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta[\log(g_\phi(Y_i)) + \log(k_\phi(Z_i|Y_i))|Y_i = y_i]$$

We now use linearity of expectation, then (1.19) and (1.38), to write the above in terms of L_n and H_n respectively.

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta[\log(g_\phi(Y_i)) + \log(k_\phi(Z_i|Y_i))|Y_i = y_i] = L_n(\phi) + H_n(\phi|\theta)$$

Combining the above steps, the desired result is obtained. Here, one subtlety was to notice that $\mathbb{E}_\theta[\log(g_\phi(Y_i))|Y_i = y_i] = \log(g_\phi(y_i))$. □

1.4.2 Sample-Splitting EM

It is not uncommon to see a slightly different formulation of Algorithm 1 where the data set is separated into T sub-datasets. This is referred to as the sample-splitting finite-sample EM algorithm written below. It is seldom used in practice and exists for the purpose of simplifying the analysis as the iterations are now based on independent sub-samples.

Algorithm 2 Sample-Splitting Finite-Sample EM

```
Input: Initial $\theta^{(0)} \in \Omega$, $\{\{y_i : i = (t)\frac{n}{T} + 1, (t)\frac{n}{T} + 2, .., (t+1)\frac{n}{T}\}$ for $t \in [T]\}$
for $t = 0, ..., T-1$ do
    Choose $t^{th}$ sub-dataset
    E-Step: Let $Q_{\frac{n}{T}}(\theta|\theta^{(t)}) = \frac{1}{n} \sum_{i=1}^{\frac{n}{T}} \mathbb{E}_{\theta^{(t)}}[\log f_\theta(Y_i, Z_i)|Y_i = y_i]$
    M-Step: Let $\theta^{(t+1)} \in M_{\frac{n}{T}}(\theta^{(t)}) = \arg\max_{\theta \in \Omega} Q_{\frac{n}{T}}(\theta|\theta^{(t)})$
end for
Output: $\theta^{(T)}$
```

1.4.3 Population EM

The population EM algorithm written below as Algorithm 3, is a fully deterministic form of Algorithm 1 and Algorithm 2. In each iteration, we require $\theta^{(t+1)} \in M(\theta^{(t)})$ where

$$M(\theta^{(t)}) := \arg\max_{\theta \in \Omega} Q(\theta|\theta^{(t)}). \tag{1.39}$$

and Q is given below as (1.40). Popularized by Balakrishnan et al. [1], this non-tractable algorithm is used solely for analysis of convergence of the EM as its fully deterministic nature allows for the use of previously derived results from optimization.

Algorithm 3 Population EM

```
Input: Initial $\theta^{(0)} \in \Omega$
for $t = 0, ..., T-1$ do
    E-Step: Let $Q(\theta|\theta^{(t)}) = \mathbb{E}_{\theta^*}[\mathbb{E}_{\theta^{(t)}}[\log f_\theta(Y_i, Z_i)|Y_i]]$
    M-Step: Let $\theta^{(t+1)} \in M(\theta^{(t)}) = \arg\max_{\theta \in \Omega} Q(\theta|\theta^{(t)})$
end for
Output: $\theta^{(T)}$
```

The population versions of the Q_n (1.27) and H_n (1.38)are given as

$$\begin{aligned} Q(\phi|\theta) &:= \mathbb{E}_{\theta^*}[\mathbb{E}_\theta[\log(f_\phi(Y, Z))|Y]] \\ &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{H}(y)} \log(f_\phi(y, s)) k_\theta(s|y) ds \right) g_{\theta^*}(y) dy, \end{aligned} \tag{1.40}$$

$$\begin{aligned} H(\phi|\theta) &:= \mathbb{E}_{\theta^*}[\mathbb{E}_\theta[\log(k_\phi(Y, Z))|Y]] \\ &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{H}(y)} \log(k_\phi(s|y)) k_\theta(s|y) ds \right) g_{\theta^*}(y) dy, \end{aligned} \tag{1.41}$$

where the outer-most expectation is taken with respect to $Y \sim g_{\theta^*}$. It is impossible to use Algorithm 3, since we do not know the true parameter θ^*. Therefore, this algorithm is considered for analysis purposes only.

1.4.4 General EM (GEM)

Finally, the General EM algorithms (GEM) (see Algorithm 4 below) are a class of algorithms where for any parameter estimate $\theta^{(t)}$, we are satisfied with any $\theta^{(t+1)} \in \Omega$ that satisfy $Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)})$. For ease of writing, we will write the GEM's iterations with respective to the GEM operator function

$$M_n^{GEM}(\theta^{(t)}) := \{\theta \in \Omega : Q(\theta|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)})\}. \tag{1.42}$$

The GEM algorithms are considered, specifically, in cases where it is too difficult to perform the global maximization of Q_n in the M-step of Algorithm 1; this is often the case in practice.

Algorithm 4 General EM

```
Input: Initial $\theta^{(0)} \in \Omega$, $\{y_1, ..., y_n\}$
for $t = 0, ..., T-1$ do
    E-Step: Let $Q_n(\theta|\theta^{(t)}) \overset{(1.27)}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta^{(t)}}[\log(f_\theta(Y_i, Z_i))|Y_i = y_i]$
    M-Step: Let $\theta^{(t+1)} \in M_n^{GEM}(\theta^{(t)}) \overset{(1.42)}{=} \{\theta \in \Omega : Q(\theta|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)})\}$
end for
Output: $\theta^{(T)}$
```

We make the remark that if $M_n(\theta^{(t)}) \subseteq \Omega$ exists for all $\theta^{(t)} \in \Omega$, then Algorithm 1 belongs to the class of GEM algorithms by definition; that is to say, $M_n(\theta^{(t)}) \subseteq M_n^{GEM}(\theta^{(t)})$ for all $\theta^{(t)} \in \Omega$. In particular this implies that any positive results for GEM algorithms automatically applies to the EM algorithm. We explore general convergence problems for this class of algorithms in Section 2.3.1.

1.5 Other Algorithms

In addition to the EM algorithm discussed in Section 1.4, several other algorithms are commonly employed to approximate the Maximum Likelihood Estimation (MLE) in parametric models with incomplete data. In this section, we provide a brief overview of three such algorithms: **Gradient Descent** [5], the **Newton-Raphson** algorithm [5], and **Gradient EM**.

Gradient Descent (see Algorithm 5) is a popular optimization algorithm for obtaining the minimizer(s) of a function. In the case of maximum likelihood estimation for parametric models with incomplete data, it involves computing the gradient of L_n with respect to the parameters. The parameter estimates are updated iteratively by taking steps in the direction of the computed gradient.

On the other hand, the Newton-Raphson algorithm is a popular optimization algorithm for obtaining the zeroes of a function. In the case of maximum likelihood estimation in parametric models with incomplete data, it involves computing the gradient and hessian of L_n with respect to the parameters. The parameter estimates are updated iteratively by solving the first-order Taylor polynomial of ∇L_n at the previous estimates.

Finally, the Gradient EM algorithm, initially proposed by [1], bears a close relation to the EM algorithm. In each iteration, the Gradient EM algorithm described in Algorithm 6 takes a step in the direction of $\nabla_1 Q_n(\theta^{(t)}|\theta^{(t)})$. However, it is important to note that Gradient EM is not guaranteed to fall under the category of GEM algorithms.

Due to the variety in parametric models with latent variables, , it is challenging to determine a universally superior algorithm. Yet, work has been completed in an attempt to shed light on this issue. Salakhutdinov et al. [34] demonstrated a relationship between Gradient Descent and the EM algorithm for specific latent variable models, showing that they are connected through a positive definite matrix. Meanwhile, Salakhutdinov and Ghahramani [33] compared the EM algorithm to Newton-like methods and proposed a faster converging algorithm for maximum likelihood estimation in parametric models with incomplete data. The framework proposed by [1] for analyzing the local convergence properties of the EM algorithm can also be applied to Gradient EM. Notably, Yan et al. [38] investigated the local convergence properties of mixtures of k spherical Gaussians under the assumption of identity covariance ($\sigma_j^* = 1 \forall j \in [k]$). Their results required a separation of the order $\tilde{\Omega}(\sqrt{k})$ for each of the k components. They demonstrated that, with appropriate initialization, Algorithm 6 converges, with high probability, to a point inside a ball of radius $\tilde{\mathcal{O}}(\frac{d}{\sqrt{n}})$ centered around the true parameter θ^* after $\mathcal{O}(\log(\frac{\sqrt{n}}{d}))$ iterations. However, as we will discuss in Chapter 3, the local convergence rate of the EM algorithm in this setting is better.

Algorithm 5 Gradient Descent [5]

```
**Input**: $f : \mathbb{R}^d \to \mathbb{R}$, $\alpha_t \in [0, 1]$, and $\theta^{(0)} \in \mathbb{R}^d$
**for** $t = 0, ..., T - 1$ **do**
    $\theta^{(t+1)} := \theta^{(t)} - \alpha_t \nabla f(\theta^{(t)})$
**end for**
**Output**: $\theta^{(T)}$
```

Algorithm 6 Gradient EM [1]

```
**Input**: Initialize $\theta^{(0)} \in \Omega$ and $\{y_1, ..., y_n\}$. Let $\alpha \in \mathbb{R}^+$:
**for** $t = 0, ..., T - 1$ **do**
    **E-Step**: Let $Q_n(\theta^{'}|\theta^{(t)}) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\theta^{(t)}}[\log(f_{\theta'}(Z_i, Y_i))|Y_i = y_i]$
    **M-Step:** Let $\theta^{(t+1)} = G_n(\theta^{(t)}) := \theta^{(t)} + \alpha \nabla_1 Q_n(\theta^{(t)}|\theta^{(t)})$
**end for**
**Output**: $\theta^{(T)}$
```

Chapter 2

Selective Review of General Properties of the EM

In the previous chapter, we discussed the EM algorithm as a widely used approach for parameter estimation in parametric models with latent variables. We presented the iterations of the EM algorithm used to approximate the MLE of a model with latent variables, focusing on its application to GMMs and MLR models. However, an essential question remains: Can we rely on the fitted parameter estimate obtained from the EM algorithm as a good approximation for the MLE?

To address this question, this chapter examines the general convergence properties of the EM algorithm (see Algorithm 1). We explore seminal works by Dempster et al. [9], Wu [37], Tseng [36], and Balakrishnan et al. [1], which shed light on the convergence behavior of the EM algorithm. By delving into the findings of these papers, we aim to gain an understanding of the convergence properties of the EM algorithm.

2.1 Preliminary Notation

Throughout this chapter, we adopt the following notation conventions unless explicitly stated otherwise. The sequences of parameter estimates obtained from executing Algorithm 1, Algorithm 4, Algorithm 3, and sample-splitting Algorithm 2 are denoted as $\{\theta^{(t)}\}_{t\geq 0}$, $\{\psi^{(t)}\}_{t\geq 0}$, $\{\phi^{(t)}\}_{t\geq 0}$, and $\{\theta^{(t)}\}_{t=0}^{T}$, respectively. Similarly, the corresponding sequence of log-likelihood values are denoted as $\{L_n(\theta^{(t)})\}_{t\geq 0}$, $\{L_n(\psi^{(t)})\}_{t\geq 0}$, $\{L_n(\phi^{(t)})\}_{t\geq 0}$, and $\{L_n(\theta^{(t)})\}_{t=0}^{T}$.

We refer to a point $\theta \in \Omega$ as a stationary point of the log-likelihood function L_n if its gradient vanishes, i.e., $\nabla L_n(\theta) = 0$. Also, a point $\theta \in \Omega$ is considered a limit point of a sequence if there exists a subsequence that converges to θ.

Furthermore, we utilize the term 'superlevel set' to denote the set $\Omega_{\theta^{(0)}}(L_n)$, which corresponds to the collection of points in Ω where the log-likelihood function L_n is greater than or equal to its value at the initial parameter $\theta^{(0)} \in \Omega$ (see Definition A.2.3).

2.2 Convergence Properties of $\{L_n(\theta^{(t)})\}_{t\geq 0}$

In this section, we present convergence properties of the sequence $\{L_n(\theta^{(t)})\}_{t\geq 0}$ where we recall that $\{\theta^{(t)}\}_{t\geq 0}$ is a sequence obtained from executing iterations of Algorithm 1. We begin with Lemma 2.2.1, instrumental to the proof of the main result of this section.

Lemma 2.2.1. *Let $\phi \in \Omega$, then it follows that*

$$\phi \in \underset{\theta\in\Omega}{\arg\max} H_n(\theta|\phi).$$

Proof. (Originally stated by Dempster et al. [9])
Let $\phi \in \Omega$. First, observe that

$$\phi \in \arg\max_{\theta\in\Omega} H_n(\theta|\phi) \iff H_n(\theta|\phi) \leq H_n(\phi|\phi)\ \forall\theta \in \Omega.$$

We dedicate the rest of the proof to showing $H_n(\theta|\phi) - H_n(\phi|\phi) \leq 0\ \forall\theta \in \Omega$. Let $\theta \in \Omega$, we expand H_n according to its definition in (1.38), then use linearity of expectation, yielding

$$H_n(\theta|\phi) - H_n(\phi|\phi) = \frac{1}{n} \textstyle\sum_{i=1}^{n} \mathbb{E}_\phi[\log(k_\theta(Z_i|y_i)) - \log(k_\phi(Z_i|y_i))|Y_i = y_i]).$$

Next, because $\log$ is concave, we use properties of logarithms and the Jensen's inequality given in Theorem A.2.1 to obtain

$$\begin{aligned} H_n(\theta|\phi) - H_n(\phi|\phi) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\phi \left[\log\left(\frac{k_\theta(Z_i|y_i)}{k_\phi(Z_i|y_i)} \right) \middle| Y_i = y_i \right] \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \log\left(\mathbb{E}_\phi \left[\frac{k_\theta(Z_i|y_i)}{k_\phi(Z_i|y_i)} \middle| Y_i = y_i \right] \right). \end{aligned}$$

Finally, we evaluate the above expectation as

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \log\left(\mathbb{E}_\phi \left[\frac{k_\theta(Z_i|y_i)}{k_\phi(Z_i|y_i)} \middle| Y_i = y_i \right] \right) &= \frac{1}{n}\sum_{i=1}^{n} \log\left(\int_{\mathbb{H}(y_i)} \frac{k_\theta(s|y_i)}{\cancel{k_\phi(s|y_i)}} \cancel{k_\phi(s|y_i)} ds \right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \log \underbrace{\left(\int_{\mathbb{H}(y_i)} k_\theta(s|y_i) ds \right)}_{=1} \\ &= 0. \end{aligned}$$

The proof is complete. □

We make the remark that it directly follows from the above lemma that for all $\phi \in \Omega$, $H_n(\theta|\phi) \leq H_n(\phi|\phi)$. The importance of the above result is made obvious in the main result of this section, presented below.

Theorem 2.2.1. *Let $\theta^{(t)} \in \Omega$ and $\theta^{(t+1)} \in M_n(\theta^{(t)})$ where M_n is given as* (1.26). *The following holds:*

a) $Q_n(\theta^{(t+1)}|\theta^{(t)}) \geq Q_n(\theta^{(t)}|\theta^{(t)})$;

b) $H_n(\theta^{(t+1)}|\theta^{(t)}) \leq H_n(\theta^{(t)}|\theta^{(t)})$;

c) $L_n(\theta^{(t+1)}) \geq L_n(\theta^{(t)})$.

Proof. (Originally proved by Dempster et al. [9, Theorem 1].)
Proof of a): The proof follows by definition: $\theta^{(t+1)} \in M_n(\theta^{(t)}) \overset{(1.26)}{\iff} \theta^{(t+1)} \in \underset{\theta\in\Omega}{\arg\max} Q_n(\theta|\theta^{(t)})$.
Proof of b): The proof follows from Lemma 2.2.1: $\theta^{(t)} \in \underset{\theta\in\Omega}{\arg\max} H_n(\theta|\theta^{(t)})$.
Proof of c): First, we remember that

$$L_n(\theta^{(t)}) \overset{(1.37)}{=} Q_n(\theta^{(t)}|\theta^{(t)}) - H_n(\theta^{(t)}|\theta^{(t)}).$$

Next, we bound the right-hand side using $a)$ and $b)$, yielding

$$Q_n(\theta^{(t)}|\theta^{(t)}) - H_n(\theta^{(t)}|\theta^{(t)}) \overset{a),\, b)}{\leq} Q_n(\theta^{(t+1)}|\theta^{(t)}) - H_n(\theta^{(t+1)}|\theta^{(t)}).$$

Lastly, we apply (1.37) again to obtain

$$Q_n(\theta^{(t+1)}|\theta^{(t)}) - H_n(\theta^{(t+1)}|\theta^{(t)}) \overset{(1.37)}{=} L_n(\theta^{(t+1)}|\theta^{(t)}).$$

This completes the proof. □

We make the following remarks. First, for any $\theta^{(0)} \in \Omega$, it follows from the above result that $\{L_n(\theta^{(t)})\}_{t\geq}$ is a non-decreasing sequence. Also, it is clear from the above proof that $\theta^{(t+1)} \in M_n(\theta^{(t)})$ can be relaxed and still guarantee $L_n(\theta^{(t+1)}) \geq L_n(\theta^{(t)})$. In fact, it is enough for $\theta^{(t+1)}$ to only satisfy $Q_n(\theta^{(t+1)}|\theta^{(t)}) \geq Q_n(\theta^{(t)}|\theta^{(t)})$. We formalize the latter in Corollary 2.3.1. Below, we present the last result of this section pertaining to the convergence of the sequence $\{L_n(\theta^{(t)})\}_{t\geq 0}$.

Corollary 2.2.1. *Let $\{\theta^{(t)}\}_{t\geq 0}$ be a sequence obtained from executing Algorithm 1. Then, if $L_n(\theta^{(t)})$ is bounded for all $t \geq 0$, it follows that*

$$\bar{L_n} := \lim_{t\to\infty} L_n(\theta^{(t)}) \textit{ exists and is finite.}$$

Proof. (Originally stated by Wu [37].)
We recall that Theorem 2.2.1 guarantees the sequence $\{L_n(\theta^t)\}_{t\geq 0}$ to be non-decreasing. If the sequence is additionally bounded, it is a standard result from analysis that it converges monotonically to its finite limit (see [32, Theorem 3.14]). □

So far, we have learned that for any $\theta^{(0)} \in \Omega$, the sequence $\{L_n(\theta^{(t)})\}_{t\geq 0}$ obtained from Algorithm 1 is non-decreasing and will converge if bounded; a desirable property for an algorithm used to approximate the MLE. But many things are still unclear. Under which conditions is the log-likelihood bounded? More importantly, what of the convergence properties of $\{\theta^{(t)}\}_{t\geq 0}$? Does this sequence converge? And if so, what does it converge to? We explore these questions in the remaining sections of this chapter.

2.3 Convergence of $\{\theta^{(t)}\}_{t\geq 0}$ to Stationary Point(s) of L_n

In this section, we explore a selection of results that provide conditions under which $\{\theta^{(t)}\}_{t\geq 0}$ converges to the stationary point(s) of L_n. We begin by introducing core assumptions that are required for the results that follow:

A_1: $\Omega_{\theta^{(0)}}(L_n)$ is compact for all $\theta^{(0)} \in \Omega$ such that $L_n(\theta^{(0)}) > -\infty$;

A_2: $L_n(\cdot)$ and $Q_n(\cdot|\theta)$ are continuous on Ω for all $\theta \in \Omega$ and
$L_n(\cdot)$ and $Q_n(\cdot|\theta)$ are differentiable in $\operatorname{int}\Omega$ for all $\theta \in \Omega$;

A_3: $\Omega_{\theta^{(0)}}(L_n) \subseteq \operatorname{int}\Omega$ for all $\theta^{(0)} \in \Omega$ such that $L_n(\theta^{(0)}) > -\infty$.

We make the following remarks to help the interpretability of the assumptions A_1, A_2, and A_3:

I) for all $\theta^{(0)} \in \Omega$, it holds that $\{\theta^{(t)}\}_{t\geq 0} \subseteq \Omega_{\theta^{(0)}}(L_n)$. If in addition, A_1 holds and $L_n(\theta^{(0)}) > -\infty$, then $\{\theta^{(t)}\}_{t\geq 0}$ is closed, bounded, has at least one limit point, and its limit point(s) are in $\Omega_{\theta^{(0)}}(L_n)$;

II) if A_2 holds, then $H_n(\cdot|\theta)$ is also continuous on Ω for all $\theta \in \Omega$ and differentiable in the interior of Ω for all $\theta \in \Omega$;

III) if A_1, A_2 hold, and $\theta^{(0)} \in \Omega$ such that $L_n(\theta^{(0)}) > \infty$, then L_n is bounded on $\Omega_{\theta^{(0)}}(L_n)$ and the sequence $\{L_n(\theta^{(t)})\}_{t\geq 0}$ is bounded;

IV) if A_1, A_2 hold, and $\theta^{(0)} \in \Omega$ such that $L_n(\theta^{(0)}) > -\infty$, then L_n takes a maximum over $\Omega_{\theta^{(0)}}(L_n)$. If in addition A_3 holds, then that maximizer is in $\operatorname{int}\Omega$;

V) Suppose A_2, A_3 hold and $\theta^{(0)} \in \Omega$ such that $L_n(\theta^{(0)}) > -\infty$. If in addition, there is some $t \geq 0$ such that $\theta^{(t)} = \theta^{(t+1)}$, then $\theta^{(t)}$ is a stationary point of L_n.

Proof of I). For all $\theta^{(0)} \in \Omega$, it follows from Theorem 2.2.1 that $\{\theta^{(t)}\}_{t\geq 0} \subseteq \Omega_{\theta^{(0)}}(L_n)$. When A_1 holds and $L_n(\theta^{(0)}) > -\infty$, then $\{\theta^{(t)}\}_{t\geq 0} \subseteq \Omega_{\theta^{(0)}}(L_n)$ is a bounded sequence and it is a consequence of the Bolzano-Weierstrass theorem that it has at least one limit point. Furthermore, since $\Omega_{\theta^{(0)}}(L_n)$ is closed, the limit point is in $\Omega_{\theta^{(0)}}(L_n)$. □

Proof of II). This follows from (1.37) together with the fact that the sum of two continuous function is continuous and the sum of two differentiable function is also differentiable. □

Proof of III). Let $\theta^{(0)} \in \Omega$ such that $L_n(\theta^{(0)}) > \infty$; in particular, $\Omega_{\theta^{(0)}}(L_n)$ is compact by A_1. Further, because A_2 guarantees the continuity of L_n over $\Omega_{\theta^{(0)}}(L_n) \subseteq \Omega$ and since continuous functions map compact sets to compact sets, it follows that $L_n(\Omega_{\theta^{(0)}}(L_n))$ is compact; thus $L_n(\Omega_{\theta^{(0)}}(L_n))$ is also bounded. Since $\{\theta^{(t)}\}_{t\geq 0} \subseteq \Omega_{\theta^{(0)}}(L_n)$, it follows that $\{L_n(\theta^{(t)})\}_{\{t\geq 0\}} \subseteq L_n(\Omega_{\theta^{(0)}}(L_n))$; $\{L_n(\theta^{(t)})\}_{\{t\geq 0\}}$ is bounded. □

Proof of IV). Let $\theta^{(0)} \in \Omega$ such that $L_n(\theta^{(0)}) > \infty$; in particular, $\Omega_{\theta^{(0)}}(L_n)$ is compact by A_1. Because L_n is continuous over Ω, it follows that it takes a maximum over the compact set $\Omega_{\theta^{(0)}}(L_n)$. □

Proof of V). We want to show $\nabla L_n(\theta^{(t+1)}) = 0$. We observe that differentiating (1.37) in the first variable yields

$$\nabla L_n(\theta^{(t+1)}) = \nabla_1 Q_n(\theta^{(t+1)}|\theta^{(t)}) - \nabla_1 H_n(\theta^{(t+1)}|\theta^{(t)}).$$

Recall $\theta^{(t+1)} \in \arg\max_{\theta\in\Omega} Q_n(\theta|\theta^{(t)})$ by definition and $\theta^{(t+1)} \in \arg\max_{\theta\in\Omega} H_n(\theta|\theta^{(t+1)})$ by Lemma 2.2.1. Because A_3 guarantees $\theta^{(t+1)} \in \operatorname{int}\Omega$, the following first order optimality conditions hold: $\nabla_1 Q_n(\theta^{(t+1)}|\theta^{(t)}) = 0$ and $\nabla_1 H_n(\theta^{(t+1)}|\theta^{(t+1)}) = 0$. We conclude that $\nabla_1 L_n(\theta^{(t)}) = 0$; the fitted parameter $\theta^{(t)}$ is a stationary point of L_n. □

We acknowledge that assumptions A_1 and A_2 are commonly made in literature [1,9, 36,37]. However, assumption A_3 is less prevalent as it is challenging to verify and does not hold for a large number of parametric models with latent variables. Nonetheless, as we will see, it plays a crucial role in the results of this section. Below, we explore the properties of $L_n(\theta^{(t+1)})$ in the case where $\theta^{(t)} \in \Omega$ is not a stationary point of L_n.

Lemma 2.3.1. *Assume A_2, A_3 hold, and $\theta^{(0)} \in \Omega$ such that $L_n(\theta^{(0)}) > \infty$. If $\theta^{(t)}$ is not a stationary point of L_n, it follows that*

$$L_n(\theta^{(t+1)}) > L_n(\theta^{(t)})$$

where $\theta^{(t+1)} \in M_n(\theta^{(t)})$.

Proof. (Originally stated by Wu [37] and Tseng [36])
We recall from Lemma 2.2.1 that $\theta^{(t)} \in \arg\max_{\theta\in\Omega} H_n(\theta|\theta^{(t)})$, which directly implies

$$H_n(\theta^{(t+1)}|\theta^{(t)}) \leq H_n(\theta^{(t)}|\theta^{(t)}). \tag{2.1}$$

Further, because A_3 guarantees $\theta^{(t)} \in \operatorname{int}\Omega$, the first order optimality condition that $\nabla_1 H_n(\theta^{(t)}|\theta^{(t)}) = 0$ follows. Next, we differentiate (1.37) in the first variable and obtain

$$\nabla L_n(\theta^{(t)}) \overset{(1.37)}{=} \nabla_1 Q_n(\theta^{(t)}|\theta^{(t)}) - \nabla_1 H_n(\theta^{(t)}|\theta^{(t)}).$$

Since $\nabla_1 H_n(\theta^{(t)}|\theta^{(t)}) = 0$, the above expression simplifies to $\nabla L_n(\theta^{(t)}) = \nabla_1 Q_n(\theta^{(t)}|\theta^{(t)})$. Therefore, we observe that $\nabla_1 Q_n(\theta^{(t)}|\theta^{(t)}) \neq 0$ because we assume $\theta^{(t)}$ is not a stationary point of L_n. This means that – unlike $\theta^{(t+1)}$ – the parameter estimate $\theta^{(t)} \in \operatorname{int}\Omega$ is not a maximizer of $\max_{\theta\in\Omega} Q_n(\theta|\theta^{(t)})$ and thus

$$Q_n(\theta^{(t+1)}|\theta^{(t)}) > Q_n(\theta^{(t)}|\theta^{(t)}). \tag{2.2}$$

Combining (1.37), (2.1), and (2.2), we obtain

$$\begin{aligned} L_n(\theta^{(t+1)}) &\overset{(1.37)}{=} \overbrace{Q_n(\theta^{(t+1)}|\theta^{(t)})}^{\overset{(2.2)}{>} Q_n(\theta^{(t)}|\theta^{(t)})} - \overbrace{H_n(\theta^{(t+1)}|\theta^{(t)})}^{\overset{(2.1)}{\leq} H_n(\theta^{(t)}|\theta^{(t)})} \\ &> Q_n(\theta^{(t)}|\theta^{(t)}) - H_n(\theta^{(t)}|\theta^{(t)}) \\ &\overset{(1.37)}{=} L_n(\theta^{(t)}) \end{aligned}$$

and this completes the proof. □

The above lemma extends the result from Theorem 2.2.1 and provides conditions for which an iteration of the EM will strictly improve the log-likelihood L_n. We make the remark that the result crucially depends on A_3 and the fact that $\theta^{(t+1)} \in \underset{\theta \in \Omega}{\arg\max} Q_n(\theta|\theta^{(t)})$. Below, we obtain conditions under which the limit points $\bar{\theta}$ of $\{\theta^{(t)}\}_{t\geq 0}$ are stationary points of L_n satisfying $L_n(\bar{\theta}) := \bar{L_n} = \lim_{t\to\infty} L_n(\theta^{(t)})$.

Theorem 2.3.1. *Let $\{\theta^{(t)}\}_{t\geq 0}$ be a sequence obtained from executing Algorithm 1 where $\theta^{(0)} \in \Omega$ such that $L_n(\theta^{(0)}) > -\infty$. If A_1, A_2, A_3 hold, and Q_n is continuous on $\Omega \times \Omega$, then, all limit points $\bar{\theta}$ of $\{\theta^{(t)}\}_{t\geq 0}$ are stationary points of L_n and $L_n(\bar{\theta}) = \bar{L_n}$.*

Proof. (Originally stated by Wu [37, Theorem 2] and Tseng [36].)
To prove this result, we use Zangwill's Global Convergence Theorem given as Theorem A.2.4 of the Appendix. Recall the point-to-set map M_n defined as (1.26), let $\theta^{(0)} \in \Omega$, and denote the solution set as

$$\mathcal{T} := \{\theta \in \operatorname{int} \Omega : \nabla L_n(\theta) = 0\}. \tag{2.3}$$

Below, we prove conditions a), b), and c) of the theorem are satisfied.
Condition a): The condition is satisfied since, choosing $K = \Omega_{\theta^{(0)}}(L_n)$, it follows that

- $K \subseteq \Omega$;
- $\{\theta^{(t)}\}_{t\geq 0} \subseteq K$ by remark I);
- K compact by A_1.

Condition b): The condition is satisfied since

- L_n is continuous by A_2;
- Theorem 2.2.1 guarantees that $L_n(\theta^{(t+1)}) \geq L_n(\theta^{(t)})$ for any $\theta^{(t)} \in \Omega$ and $\theta^{(t+1)} \in M_n(\theta^{(t)})$;
- Lemma 2.3.1 guarantees that $L_n(\theta^{'}) > L_n(\theta)$ for any $\theta \in \Omega/\mathcal{T}$ and $\theta^{'} \in M_n(\theta)$.

Condition c): Referring to Definition A.2.4, we show M_n is a closed point-to-set map on $\Omega/\mathcal{T}$. Let $\bar{\theta} \in \Omega/\mathcal{T}$ and suppose there exists $\{\theta^{(l)}\}_{l\geq 0} \subseteq \Omega$ and $\{\phi^{(l)}\}_{l\geq 0} \subseteq \Omega$ where $\phi^{(l)} \in M_n(\theta^{(l)})$ such that $\theta^{(l)} \underset{l\to\infty}{\to} \bar{\theta}$ and $\phi^{(l)} \underset{l\to\infty}{\to} \bar{\phi}$. First, since $\phi^{(l)} \in M_n(\theta^{(l)})$, it holds that $Q_n(\phi^{(l)}|\theta^{(l)}) \geq Q_n(\phi|\theta^{(l)})$ for all $\phi \in \Omega$. Next, we take the $\lim\limits_{l\to\infty}$ on both sides of the inequality to obtain $\lim\limits_{l\to\infty} Q_n(\phi^{(l)}|\theta^{(l)}) \geq \lim\limits_{l\to\infty} Q_n(\phi|\theta^{(l)})$ for all $\phi \in \Omega$. Lastly, because Q_n is continuous on $\Omega \times \Omega$, it follows that $Q_n(\bar{\phi}|\bar{\theta}) \geq Q_n(\phi|\bar{\theta})$ for all $\phi \in \Omega$. Therefore, $\bar{\phi} \in M_n(\bar{\theta})$ and M_n is a closed point-to-set map.

This completes the proof. □

The aforementioned theorem establishes conditions under which the limit points of the sequence $\{\theta^{(t)}\}_{t\geq 0}$ correspond to stationary points of the log-likelihood function L_n. Tseng [36] uses the result to prove that, with appropriate initialization, the limit points of $\{\theta^{(t)}\}_{t\geq 0}$ obtained for a 2-component 1-dimensional GMM with known unit variance will also be stationary points of L_n. It is worth noting that the initialization criterion is easily verifiable in this case.

In this section, we have identified conditions that ensure the limit points of $\{\theta^{(t)}\}_{t\geq 0}$ are stationary points of L_n. However, these conditions can be challenging to verify and may not hold in general. Furthermore, convergence to a stationary point of L_n alone does not provide insights into the quality of the obtained approximation for the MLE. To address this concern, Section 2.4 delves into the topic of assessing the proximity of EM's parameter estimates to the true parameter θ^*. Before doing so, we extend some of the results from the previous two sections to the GEM algorithms.

2.3.1 Convergence of $\{\psi^{(t)}\}_{t\geq 0}$ to stationary point(s) of L_n

In this section, we present results that provide conditions under which $\{\psi^{(t)}\}_{t\geq 0}$ converges to stationary point(s) of L_n. We begin with the first result for GEM.

Corollary 2.3.1. *Let $\psi^{(t)} \in \Omega$ and $\psi^{(t+1)} \in M_n^{GEM}(\psi^{(t)})$ where M_n^{GEM} is given as* (1.42)*. The following holds:*

a) $Q_n(\psi^{(t+1)}|\psi^{(t)}) \geq Q_n(\psi^{(t)}|\psi^{(t)})$*;*

b) $H_n(\psi^{(t+1)}|\psi^{(t)}) \leq H_n(\psi^{(t)}|\psi^{(t)})$*;*

c) $L_n(\psi^{(t+1)}) \geq L_n(\psi^{(t)})$*.*

Proof. The proof follows from an identical argument to that of Theorem 2.2.1. □

It follows from the above theorem that the sequence $\{L_n(\psi^{(t)})\}_{t\geq 0}$ is non-decreasing and thus, each iteration of Algorithm 1 can only increase the log-likelihood L_n. Next, we present a corollary pertaining to the convergence of the sequence $\{L_n(\psi^{(t)})\}_{t\geq 0}$.

Corollary 2.3.2. *Let* $\{\psi^{(t)}\}_{t\geq 0}$ *be a sequence obtained from executing Algorithm 4. Then, if* $L_n(\psi^{(t)})$ *is bounded for all* $t \geq 0$*, it follows that*

$$\bar{L_n} := \lim_{t\to\infty} L_n(\psi^{(t)}) \text{ exists and is finite.}$$

Proof. The proof follows from an identical argument to that of Corollary 2.2.1. □

The next result provides conditions under which the limit points of $\{\psi^{(t)}\}_{t\geq 0}$ will be stationary point(s) of L_n.

Corollary 2.3.3. *Let* $\{\psi^{(t)}\}_{t\geq 0}$ *be a sequence obtained from executing Algorithm 4 where* $\psi^{(0)} \in \Omega$ *such that* $L_n(\psi^{(0)}) > -\infty$*. If the following holds:*

- A_1*,* A_2*,* A_3*;*
- Q_n *is continuous on* $\Omega \times \Omega$*;*
- $L_n(\psi^{(t+1)}) > L_n(\psi^{(t)})$ *for all* $\psi^{(t)}$ *such that* $\nabla L_n(\psi^{(t)}) \neq 0$*.*

Then, all limit points $\bar{\psi}$ *of* $\{\psi^{(t)}\}_{t\geq 0}$ *are stationary points of* L_n *and* $L_n(\bar{\psi}) = \lim_{t\to\infty} L_n(\psi^{(t)})$*.*

Proof. The proof follows from an identical argument to that of Theorem 2.3.1. □

The above result summarizes the general properties of GEM algorithms. Similar to Algorithm 1, the sequence $\{L_n(\psi^{(t)})\}_{t\geq 0}$ remains non-decreasing and converges if it is bounded. Hence, GEM algorithms also possess desirable properties for approximating the MLE. Furthermore, the limit points of $\{\phi^{(t)}\}_{t\geq 0}$ can be ensured to be stationary points of L_n, although this requires stricter conditions compared to $\{\theta^{(t)}\}_{t\geq 0}$. In light of these findings, it is preferable, in general, to compute $\theta^{(t+1)} \in M_n(\theta^{(t)})$ when possible.

2.4 Convergence of $\{\theta^{(t)}\}_{t\geq 0}$ to the True Parameter θ^*

In this section, we present the framework developed by Balakrishnan et al. [1] to analyze the local convergence properties of the EM algorithm towards the true parameter θ^*. The framework involves two main steps, outlined below:

$$\|\theta^{(t+1)} - \theta^*\|_2 \leq \underbrace{\|\phi^{(t+1)} - \theta^*\|_2}_{\text{Step 1}} + \underbrace{\|\theta^{(t+1)} - \phi^{(t+1)}\|_2}_{\text{Step 2}} \tag{2.4}$$

Here, $\theta^{(t+1)} \in M_n(\theta^{(t)})$ and $\phi^{(t+1)} \in M(\theta^{(t)})$. The first step aims to establish general conditions under which the sequence $\{\phi^{(t)}\}_{t\geq 0}$ converges to θ^*. The second step utilizes a combination of probability theory and other techniques to bound the disparity between $\{\phi^{(t)}\}_{t\geq 0}$ and $\{\theta^{(t)}\}_{t\geq 0}$. In Section 2.4.1, we address the first step, while Section 2.4.2 focuses on the second step. Before proceeding, let us introduce another assumption.

We refer to A_4 when we require $Q(\cdot|\theta^*)$ to be λ-strongly-concave in a neighborhood of θ^*, where Q is given as (1.40):

A_4: $\exists\, r > 0, \lambda > 0$ such that

$$Q(\theta_1|\theta^*) - Q(\theta_2|\theta^*) \leq \langle \nabla_1 Q(\theta_2|\theta^*), \theta_1 - \theta_2\rangle - \frac{\lambda}{2}\|\theta_1 - \theta_2\|_2^2$$

$$\forall\, \theta_1, \theta_2 \in \mathbb{B}_2(\theta^*; r). \tag{2.5}$$

We make the remark that A_4 only makes sense if $\nabla_1 Q(\cdot|\theta^*)$ exists everywhere on Ω. Further, we clarify that A_4 does not extend to Q_n; this assumption is specific to Q and is therefore solely related to the population EM described in Algorithm 3.

2.4.1 Convergence of population $\{\phi^{(t)}\}_{t\geq 0}$ to θ^*

In this section, we explore the first step of the framework in (2.4):

$$\|\theta^{(t)} - \theta^*\|_2 \leq \underbrace{\|\phi^{(t)} - \theta^*\|_2}_{\text{Step 1}} + \underbrace{\|\theta^{(t)} - \phi^{(t)}\|_2}_{\text{Step 2}}.$$

where $\theta^{(t+1)} \in M_n(\theta^{(t)}), \phi^{(t+1)} \in M(\theta^{(t)})$. We aim to understand the local convergence properties of $\|\phi^{(t)} - \theta^*\|_2$. We introduce a notion essential for the main result of this section: first order stability. We say the functions $\{Q(\cdot|\theta) : \theta \in \Omega\}$ satisfy FOS(γ) over $\mathbb{B}_2(r;\theta^*)$ if

$$\|\nabla_1 Q(\phi^{(t+1)}|\theta^*) - \nabla_1 Q(\phi^{(t+1)}|\phi^{(t)})\|_2 \leq \gamma\|\theta - \theta^*\|_2$$
$$\text{for all } \phi^{(t)} \in \mathbb{B}_2(r;\theta^*), \phi^{(t+1)} \in M(\phi^{(t)}). \quad (2.6)$$

We make the remark that the above only makes sense if $\nabla_1 Q$ exists everywhere on $\Omega \times \Omega$. Further, given a parametric model, there is no general guarantee that the FOS(γ) conditions can be satisfied for some λ and r. Interestingly, Balakrishnan et al. [1] suggest that since (2.6) is satisfied for $\phi^{(t)} = \theta^*$, it may be that under some regularity conditions on Q, FOS(γ) holds for some choice of γ and r. After careful examination, we see that if A_3 holds, it follows that $\nabla_1 Q(\phi^{(t+1)}|\phi^{(t)}) = 0$ where $\phi^{(t+1)} \in M(\phi^{(t)})$; in particular, the FOS(γ) conditions simplify to $\|\nabla_1 Q(\phi^{(t+1)}|\theta^*)\|_2 \leq \gamma\|\phi^{(t)} - \theta^*\|_2$. Under this condition, first order stability guarantees that the gradient do not explode, in fact they are bounded above by the distance of the current parameter estimate to the true parameter θ^* up to a constant factor. Below, we provide a result which may prove useful to verify the conditions of FOS(γ).

Lemma 2.4.1. *Assume A_2 holds. The functions $\{Q(\cdot|\theta) : \theta \in \Omega\}$ satisfy FOS(γ) over $\mathbb{B}_2(r;\theta^*)$ if and only if the functions $\{H(\cdot|\theta) : \theta \in \Omega\}$ satisfy FOS(γ) over $\mathbb{B}_2(r;\theta^*)$.*

Proof. Letting $\phi^{(t)} \in \mathbb{B}_2(r;\theta^*)$ and $\phi^{(t+1)} \in M(\phi^{(t)})$, it follows from (1.37) and A_2 that

$$\begin{aligned}
&\|\nabla_1 Q(\phi^{(t+1)}|\theta^*) - \nabla_1 Q(\phi^{(t+1)}|\phi^{(t)})\|_2 \\
&\quad \overset{1.37}{=} \|[\nabla L(\phi^{(t+1)}) + \nabla_1 H(\phi^{(t+1)}|\theta^*)] - [\nabla L(\phi^{(t+1)}) - \nabla_1 H(\phi^{(t+1)}|\phi^{(t)})]\|_2 \\
&\quad = \|\nabla_1 H(\phi^{(t+1)}|\theta^*) - \nabla_1 H(\phi^{(t+1)})|\phi^{(t)})\|_2.
\end{aligned}$$

$'' \Longrightarrow ''$: If the functions $\{Q(\cdot|\theta) : \theta \in \Omega\}$ satisfy FOS(γ) over $\mathbb{B}_2(r;\theta^*)$, it follows from the above that

$$\|\nabla_1 H(\phi^{(t+1)}|\theta^*) - \nabla_1 H(\phi^{(t+1)}|\phi^{(t)})\|_2 \leq \gamma\|\phi^{(t)} - \theta^*\|_2.$$

$'' \Longleftarrow ''$: Similarly, if the functions $\{H(\cdot|\theta) : \theta \in \Omega\}$ satisfy FOS(γ) $\mathbb{B}_2(r;\theta^*)$, it follows from the former that

$$\|\nabla_1 Q(\phi^{(t+1)}|\theta^*) - \nabla_1 Q(\phi^{(t+1)}|\phi^{(t)})\|_2 \leq \gamma\|\phi^{(t)} - \theta^*\|_2.$$

This completes the proof. □

Because given some $\phi^{(0)} \in \Omega$, the sequence $\{\phi^{(t)}\}_{t\geq 0}$ obtained from Algorithm 3 is completely deterministic, we use convex optimization theory to analyze its convergence properties. We give the main result of the section below.

Theorem 2.4.1. *Assume A_2, A_4 hold, and $\exists$ some $r > 0$ and $0 \leq \gamma < \lambda$ such that the functions $\{Q(\cdot|\theta) : \theta \in \Omega\}$ satisfy $FOS(\gamma)$ over $\mathbb{B}_2(r;\theta^*)$. Then, it follows that*

$$\|\phi^{(t+1)} - \theta^*\| \leq \frac{\gamma}{\lambda}\|\phi^{(t)} - \theta^*\| \tag{2.7}$$

for all $\phi^{(t)} \in \mathbb{B}_2(r;\theta^*)$ and $\phi^{(t+1)} \in M(\phi^{(t)})$ where M is given as (1.39).

Proof. (Originally states by Balakrishnan et al. [1, Theorem 1].)
Let $\phi^{(t)} \in \mathbb{B}_2(r_1;\theta^*)$ and $\phi^{(t+1)} \in M(\phi^{(t)})$. Because $\theta^* = \arg\max_{\theta\in\Omega} Q(\theta|\theta^*)$ and Ω is convex,

it follows from Theorem A.2.3 that $\langle \nabla_1 Q(\theta^*|\theta^*), \theta - \theta^* \rangle \leq 0 \; \forall \theta \in \Omega$; in particular, we have

$$\langle \nabla_1 Q(\theta^*|\theta^*), \phi^{(t+1)} - \theta^* \rangle \leq 0. \tag{2.8}$$

Similarly, because $\phi^{(t+1)} \in \arg\max_{\theta \in \Omega} Q(\theta|\phi^{(t)})$ and Ω is convex, it follows from Theorem A.2.3 that $\langle \nabla_1 Q(\phi^{(t+1)}|\theta^{(t)}), \theta - \phi^{(t+1)} \rangle \leq 0 \; \forall \theta \in \Omega$; in particular, we have

$$\langle \nabla_1 Q(\phi^{(t+1)}|\phi^{(t)}), \theta^* - \phi^{(t+1)} \rangle \leq 0. \tag{2.9}$$

Combining (2.8) and (2.9) yields

$$\langle -\nabla_1 Q(\theta^*|\theta^*), \theta^* - \phi^{(t+1)} \rangle \leq \langle -\nabla_1 Q(\phi^{(t+1)}|\theta^{(t)}), \theta^* - \phi^{(t+1)} \rangle. \tag{2.10}$$

Adding $\langle \nabla_1 Q(\phi^{(t+1)}|\theta^*), \theta^* - \phi^{(t+1)} \rangle$ to (2.10), we obtain

$$\begin{aligned} &\overbrace{\langle \nabla_1 Q(\phi^{(t+1)}|\theta^*) - \nabla_1 Q(\theta^*|\theta^*), \theta^* - \phi^{(t+1)} \rangle}^{i)} \leq \\ &\qquad\qquad \underbrace{\langle \nabla_1 Q(\phi^{(t+1)}|\theta^*) - \nabla_1 Q(\phi^{(t+1)}|\theta^{(t)}), \theta^* - \phi^{(t+1)} \rangle}_{ii)}. \end{aligned} \tag{2.11}$$

We now look to find a lower bound for $i)$. It follows from A_4 that

$$\langle \nabla_1 Q(\phi^{(t+1)}|\theta^*), \theta^* - \phi^{(t+1)} \rangle + Q(\phi^{(t+1)}|\theta^*) - Q(\theta^*|\theta^*) \geq \frac{\lambda}{2} \|\theta^* - \phi^{(t+1)}\|_2^2 \tag{2.12}$$

$$\langle \nabla_1 Q(\theta^*|\theta^*), \phi^{(t+1)} - \theta^* \rangle + Q(\theta^*|\theta^*) - Q(\phi^{(t+1)}|\theta^*) \geq \frac{\lambda}{2} \|\theta^* - \phi^{(t+1)}\|_2^2. \tag{2.13}$$

Summing (2.12) and (2.13) together, we obtain

$$\overbrace{\langle \nabla_1 Q(\phi^{(t+1)}|\theta^*), \theta^* - \phi^{(t+1)} \rangle + \langle \nabla_1 Q(\theta^*|\theta^*), \phi^{(t+1)} - \theta^* \rangle}^{i)} \geq \lambda \|\theta^* - \phi^{(t+1)}\|_2^2. \tag{2.14}$$

We have obtained the lower bound. Next, we look to find an upper bound for $ii)$. It follows from the Cauchy Schwartz inequality and FOS(γ) that

$$\begin{aligned} ii) &= \langle \nabla_1 Q(\phi^{(t+1)}|\theta^*) - \nabla_1 Q(\phi^{(t+1)}|\phi^{(t)}), \theta^* - \phi^{(t+1)} \rangle \\ &\leq \|\nabla_1 Q(\phi^{(t+1)}|\theta^*) - \nabla_1 Q(\phi^{(t+1)}|\phi^{(t)})\|_2 \|\theta^* - \phi^{(t+1)}\|_2 \\ &\leq \gamma \|\theta^* - \phi^{(t)}\|_2 \|\theta^* - \phi^{(t+1)}\|_2 \end{aligned} \tag{2.15}$$

Substituting (2.14) and (2.15) into (2.11), we obtain

$$\overbrace{\lambda\|\theta^* - \phi^{(t+1)}\|_2^2}^{(2.14)} \leq \overbrace{\gamma\|\theta^* - \phi^{(t)}\|_2\|\theta^* - \phi^{(t+1)}\|_2}^{(2.15)}.$$

Finally, rearranging yields $\|\theta^* - \phi^{(t+1)}\|_2 \leq \frac{\gamma}{\lambda}\|\theta^* - \phi^{(t)}\|_2$ and this completes the proof. □

Unravelling the recurrence in the above theorem, we see that

$$\|\phi^{(t)} - \theta^*\|_2 \leq \left(\frac{\gamma}{\lambda}\right)^t \|\phi^{(0)} - \theta^*\|_2.$$

Therefore, under the conditions of the theorem, the sequence $\{\phi^{(t)}\}_{t\geq 0}$ converges geometrically to the true parameter θ^*. However, since we don't have direct access to $\{\phi^{(t)}\}_{t\geq 0}$, our goal is to establish a similar result for the sequence $\{\theta^{(t)\}}_{t\geq 0}$. We address this objective in the next section.

We make the remark that in many practical applications of the EM algorithm, such as GMMs and MLR models, it is not always necessary to rely on the aforementioned theorem to obtain a result of the form $\|\phi^{(t+1)} - \theta^*\|_2 \leq \kappa\|\phi^{(t)} - \theta^*\|_2$. As explored in Chapter 3, many researchers tailor their analysis to the specific class of models considered to obtain such a result in the first step of the framework. We consider step 2 in the subsequent section.

2.4.2 Local Convergence of $\{\theta^{(t)}\}_{t\geq 0}$ to θ^*

We recall (2.4) and consider the second step of the framework proposed by [1]:

$$\|\theta^{(t)} - \theta^*\|_2 \leq \underbrace{\|\phi^{(t)} - \theta^*\|_2}_{\text{deterministic analysis}} + \underbrace{\|\theta^{(t)} - \phi^{(t)}\|_2}_{\text{stability } (\epsilon_M^{unif})}$$

where $\theta^{(t)} \in M_n(\theta^{(t-1)}), \phi^{(t)} \in M(\theta^{(t-1)})$. We aim to bound the disparity between the population EM operator M and the EM operator M_n; Ho et al. [18] refers to this as the stability of the EM Operator. Also, unlike $\{\phi^{(t)}\}_{t\geq 0}$ which is deterministic, Algorithm 1 deals with data generated from a probability distribution and so, the sequence $\{\theta^{(t)}\}_{t\geq 0}$ is random. Therefore, we must introduce some new notions.

Because each iterate $\theta^{(t)}$ is based on the same sample, the stability bound must hold uniformly over the ball $\mathbb{B}_2(r;\theta^*)$. Consider $\epsilon_M^{unif}(n,\delta)$ defined below.

Definition 2.4.1. *For a given sample size n and tolerance parameter $\delta \in (0,1)$, we define $\epsilon_M^{unif}(n,\delta)$ to be the smallest scalar such that*

$$\mathbb{P}\left(\sup_{\mathcal{S}}\|\theta^{(t+1)} - \phi^{(t+1)}\|_2 \leq \epsilon_M^{unif}(n,\delta)\right) \geq 1-\delta \tag{2.16}$$

where $\mathcal{S} := \left\{\theta^{(t+1)} \in M_n(\theta^{(t)}), \phi^{(t+1)} \in M(\theta^{(t)}) : \theta^{(t)} \in \mathbb{B}_2(r;\theta^*)\right\}$.

For some $\delta \in (0,1)$ and sample size n, the quantity $\epsilon_M^{unif}(n,\delta)$ serves to uniformly bound the disparity between M and M_n over the ball $\mathbb{B}_2(r;\theta^*)$ with probability at least $1-\delta$. We make the remark that there is no guarantee in general that, given some n and δ, $\epsilon_M^{unif}(n,\delta)$ exists. Even if $\epsilon_M^{unif}(n,\delta)$ exists, finding it can be challenging and specific to the family of parametric models. We now present the main result of this section.

Theorem 2.4.2. *If*

a) there exists n large enough that $\epsilon_M^{unif}(n,\delta) \leq (1-\kappa)r$ with probability at least $1-\delta$;

b) there exists $r > 0$ such that for all $t \geq 0$ and $\theta^{(0)} \in \mathbb{B}_2(r;\theta^*)$,

$$\|\phi^{(t+1)} - \theta^*\| \leq \kappa\|\theta^{(t)} - \theta^*\|$$

where $\theta^{(t)} \in \mathbb{B}_2(r;\theta^*)$ and $\phi^{(t+1)} \in M(\theta^{(t)})$;

it follows that the sequence $\{\theta^{(t)}\}_{t\geq 0}$ obtained from executing iterations of Algorithm 1 satisfies

$$\|\theta^{(t)} - \theta^*\|_2 \leq \kappa^t \|\theta^{(0)} - \theta^*\|_2 + \frac{1}{1-\kappa}\epsilon_M^{unif}(n,\delta) \tag{2.17}$$

with probability at least $1-\delta$*.*

Proof. (Originally stated by Balakrishnan et al. [1, Theorem 2].)
For the remainder of this proof, we only consider the event, of probability at least $1-\delta$, guaranteeing $\epsilon_M^{unif}(n,\delta) \leq (1-\kappa)r$.

First, we prove by induction that if r satisfies b) and $\theta^{(0)} \in \mathbb{B}_2(r;\theta^*)$, then $\theta^{(t)} \in \mathbb{B}_2(r;\theta^*)$ for all $t \geq 0$, with probability $1-\delta$.

Base Case: $\theta^{(0)} \in \mathbb{B}_2(r;\theta^*)$ by assumption.

Induction Step: Show $\theta^{(t)} \in \mathbb{B}_2(r;\theta^*) \implies \theta^{(t+1)} \in \mathbb{B}_2(r;\theta^*)$.

Since $\theta^{(t)} \in \mathbb{B}_2(r;\theta^*)$, it follows from b) that for all $\phi^{(t+1)} \in M(\theta^{(t)})$,

$$\|\phi^{(t+1)} - \theta^*\|_2 \leq \kappa \|\theta^{(t)} - \theta^*\|_2. \tag{2.18}$$

Also it follows from the definition of $\epsilon_M^{unif}(n,\delta)$ that for all $\theta^{(t+1)} \in M_n(\theta^{(t)}), \phi^{(t+1)} \in M(\theta^{(t)})$

$$\|\theta^{(t+1)} - \phi^{(t+1)}\|_2 \leq \epsilon_M^{unif}(n,\delta). \tag{2.19}$$

Therefore, if $\theta^{(t+1)} \in M_n(\theta^{(t)}), \phi^{(t+1)} \in M(\theta^{(t)})$, it follows from (2.18), (2.19), and the triangle inequality that

$$\begin{aligned}
\|\theta^{(t+1)} - \theta^*\|_2 &= \|\phi^{(t+1)} + \theta^{(t+1)} - \phi^{(t+1)} - \theta^*\|_2 \\
&\leq \|\phi^{(t+1)} - \theta^*\|_2 + \|\theta^{(t+1)} - \phi^{(t+1)}\|_2 \\
&\leq \kappa \underbrace{\|\theta^{(t)} - \theta^*\|_2}_{<r} + \underbrace{\epsilon_M^{unif}(n,\delta)}_{\leq(1-\kappa)r} \\
&< r
\end{aligned}$$

where last step follows because $\theta^{(t)} \in \mathbb{B}_2(r;\theta^*) \iff \|\theta^{(t)} - \theta^*\|_2 < r$ and $\epsilon_M^{unif}(n,\delta) \leq (1-\kappa)r$ under this event. This completes the induction.

So far, if r satisfies b) and $\theta^{(0)} \in \mathbb{B}_2(r;\theta^*)$, then the following recurrence relation holds for all $t \geq 0$:

$$\begin{aligned}\|\theta^{(t)} - \theta^*\|_2 &\leq \|\phi^{(t)} - \theta^*\|_2 + \|\phi^{(t)} - \theta^{(t)}\|_2, \text{where } \phi^{(t)} \in M(\theta^{(t-1)}) \\ &\leq \kappa\|\theta^{(t-1)} - \theta^*\|_2 + \epsilon_M^{unif}(n,\delta).\end{aligned}$$

where $\phi^{(t+1)} \in M(\theta^{(t)})$. We solve the recurrence below:

$$\begin{aligned}\|\theta^{(t)} - \theta^*\|_2 &\leq \kappa \overbrace{\|\theta^{(t-1)} - \theta^*\|_2}^{\leq \kappa\|\theta^{(t-2)} - \theta^*\|_2 + \epsilon_M^{unif}(n,\delta)} + \epsilon_M^{unif}(n,\delta) \\ &\leq \kappa^t\|\theta^{(0)} - \theta^*\|_2 + \left[\sum_{s=0}^{t} \kappa^s\right] \epsilon_M^{unif}(n,\delta) \\ &\leq \kappa^t\|\theta^{(0)} - \theta^*\|_2 + \frac{1}{1-\kappa}\epsilon_M^{unif}(n,\delta).\end{aligned}$$

In the last step, we used that a geometric series with $\kappa \in (0,1)$ converges to $\frac{1}{1-\kappa}$. This completes the proof. □

It follows from the above theorem that if $T = \log_{1/\kappa}\left(\frac{(1-\kappa)\|\theta^{(0)}-\theta^*\|_2}{\epsilon_M^{unif}(n,\delta)}\right)$, then $\|\theta^{(T)} - \theta^*\|_2 \leq \frac{2}{1-\kappa}\epsilon_M^{unif}(n,\delta)$ with probability at least $1-\delta$. Therefore, for all $t \geq T$, it follows that $\theta^{(t)} \in \mathbb{B}_2\left(\frac{2}{1-\kappa}\epsilon_M^{unif}(n,\delta);\theta^*\right)$ with probability at least $1-\delta$. This concludes the general properties of convergence on Algorithm 1.

We make the remark that the main challenge with using the above theorem lies in calculating a good bound for $\epsilon_n^{unif}(n,\delta)$. Still, in our literature survey in Chapter 3, this theorem is often referenced and used to obtain local convergence properties in a wide variety of settings. Still, in cases where it proves too difficult to obtain $\epsilon_m^{unif}(n\delta)$, it is possible to turn to sample-splitting EM for easier analysis. We explore this scenario in the next section.

2.4.3 Local Convergence of the Sample-Splitting $\{\theta^{(t)}\}_{t=0}^T$ to θ^*

In this section, we re-consider the second step of the framework proposed by [1] when the sequence $\{\theta^{(t)}\}_{t=0}^T$ is obtained from executing Algorithm 2 instead:

$$\|\theta^{(t)} - \theta^*\|_2 \leq \underbrace{\|\phi^{(t)} - \theta^*\|_2}_{\text{deterministic analysis}} + \underbrace{\|\theta^{(t)} - \phi^{(t)}\|_2}_{\text{stability } (\epsilon_M)}.$$

where $\theta^{(t)} \in M_{\frac{n}{T}}(\theta^{(t-1)}), \phi^{(t)} \in M(\theta^{(t-1)})$. Unlike $\{\phi^{(t)}\}_{t\geq 0}$ which is deterministic, Algorithm 2 deals with data generated from a probability distribution and so, the sequence $\{\theta^{(t)}\}_{t=0}^T$ is random.

This time, because $\{\theta^{(t)}\}_{t=0}^T$ is based on independent sub-samples, we no longer need a bound that holds uniformly over the ball $\mathbb{B}_2(r;\theta^*)$. Instead, we consider $\epsilon_M(n,\delta)$ defined below.

Definition 2.4.2. *For a given sample size n and tolerance parameter $\delta \in (0,1)$, we define $\epsilon_M(n,\delta)$ to be the smallest scalar such that fixing any $\theta^{(t)} \in \mathbb{B}_2(r,\theta^*)$,*

$$\mathbb{P}(\|\theta^{(t+1)} - \phi^{(t+1)}\|_2 \leq \epsilon_M(n,\delta)) \geq 1-\delta \tag{2.20}$$

where $\theta^{(t+1)} \in M_n(\theta^{(t)})$ and $\phi^{(t+1)} \in M(\theta^{(t)})$.

For some $\delta \in (0,1)$ and sample size n, the quantity $\epsilon_M(n,\delta)$ serves to bound the disparity between M and M_n with probability $1-\delta$. We make the remark that there is still no guarantee in general that, given some n and δ, $\epsilon_M(n,\delta)$ exists. We begin with the below-lemma comparing the two quantities $\epsilon_M(n,\delta)$ and $\epsilon_M^{unif}(n,\delta)$.

Lemma 2.4.2. *Fix $n \in \mathbb{N}$ and $\delta \in (0,1)$. If $\epsilon_M^{unif}(n,\delta)$ (see Definition 2.4.1) and $\epsilon_M(n,\delta)$ (see Definition 2.4.1) exist then,*

$$\epsilon_M(n,\delta) \leq \epsilon_M^{unif}(n,\delta).$$

Proof. Fix $n \in \mathbb{N}$ and $\delta \in (0,1)$ such that $\epsilon_M^{unif}(n,\delta)$ and $\epsilon_M(n,\delta)$ both exist. In addition, let $\theta_1^{(t)} \in \mathbb{B}_2(r;\theta^*)$, $\theta_1^{(t+1)} \in M_n(\theta^{(t)})$, and $\phi_1^{(t+1)} \in M(\theta^{(t)})$.

First, we observe that

$$\|\theta_1^{(t+1)} - \phi_1^{(t+1)}\|_2 \leq \sup_{\mathcal{S}} \|\theta^{(t+1)} - \phi^{(t+1)}\|_2.$$

where the set $\mathcal{S}$ is described in Definition 2.4.1. Next, it follows that

$$\{(y_1, ..., y_n) : \sup_{\mathcal{S}} \|\theta^{(t+1)} - \phi^{(t+1)}\|_2 \leq \epsilon_m^{unif}\} \subseteq \{(y_1, ..., y_n) : \|\theta_1^{(t+1)} - \phi_1^{(t+1)}\|_2 \leq \epsilon_m^{unif}\}$$

Therefore, it follows from the definition of ϵ_M^{unif} that

$$\begin{aligned} 1 - \delta &\leq \mathbb{P}\left[\sup_{\mathcal{S}} \|\theta^{(t+1)} - \phi^{(t+1)}\|_2 \leq \epsilon_M^{unif}(n, \delta)\right] \\ &\leq \mathbb{P}\left[\|\theta_1^{(t+1)} - \phi_1^{(t+1)}\|_2 \leq \epsilon_M^{unif}(n, \delta)\right] \end{aligned}$$

The above holds for all $\theta_1^{(t)} \in \mathbb{B}_2(r; \theta^*)$, $\theta_1^{(t+1)} \in M_n(\theta^{(t)})$, and $\phi_1^{(t+1)} \in M(\theta^{(t)})$. Thus, by definition of $\epsilon_M(n, \delta)$, it follows that $\epsilon_M(n, \delta) \leq \epsilon_M^{unif}(n, \delta)$. This completes the proof. □

We make the remark that the above implies that while $\epsilon_M^{unif}(n, \delta)$ can be used as an upper bound for $\epsilon_M(n, \delta)$, the converse is not true. This means that whenever a convergence result is obtained for Algorithm 2, encouraging though it is, the behavior cannot be expected to hold for Algorithm 1. We now present the main result of this section.

Theorem 2.4.3. *Let $T \in \mathbb{N}$, if*

a) there exists n large enough such that $\epsilon_M(\frac{n}{T}, \frac{\delta}{T}) \leq (1 - \kappa)r$ with probability at least $1 - \frac{\delta}{T}$;

b) there exists $r > 0$ such that for all $t \geq 0$ and $\theta^{(0)} \in \mathbb{B}_2(r; \theta^*)$,

$$\|\phi^{(t+1)} - \theta^*\| \leq \kappa \|\theta^{(t)} - \theta^*\|$$

where $\theta^{(t)} \in \mathbb{B}_2(r; \theta^*)$ and $\phi^{(t+1)} \in M(\theta^{(t)})$;

it follows that the sequence $\{\theta^{(t)}\}_{t=0}^{T}$ obtained from executing iterations of Algorithm 2 satisfies

$$\|\theta^{(t)} - \theta^*\|_2 \leq \kappa^t \|\theta^{(0)} - \theta^*\|_2 + \frac{1}{1-\kappa}\epsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right) \tag{2.21}$$

for all $t \in \{0, 1, , ..., T\}$ with probability at least $1-\delta$.

Proof. (Originally states by Balakrishnan et al. [1][Theorem 2].)
Unlike the proof of Theorem (2.4.2), iterates of the sequence $\{\theta^{(t)}\}_{t=0}^{T}$ are based on independent sub-samples. Therefore, for each $t \in \{0, 1, ..., T\}$ and with probability at least $1 - \frac{\delta}{T}$, we must see if the bound $\|\theta^{(t+1)} - \phi^{(t+1)}\|_2 \leq \epsilon_M(\frac{n}{T}, \frac{\delta}{T})$ is satisfied where $\theta^{(t+1)} \in M_{\frac{n}{T}}(\theta^{(t)})$ and $\phi^{(t+1)} \in M(\theta^{(t)})$.

We resolve this by performing a union bound over all T iterations. Let $\theta^{(t)} \in \mathbb{B}_2(r; \theta^*)$, $\theta^{(t+1)} \in M_{\frac{n}{T}}(\theta^{(t)})$ and $\phi^{(t+1)} \in M(\theta^{(t)})$,

$$\begin{aligned}
\mathbb{P}\left(\max_{t\in[T]}\|\theta^{(t+1)} - \phi^{(t+1)}\|_2 \leq \epsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right)\right) &= 1 - \mathbb{P}\left(\cup_{t=0}^{T-1}\left[\|\theta^{(t+1)} - \phi^{(t+1)}\|_2 > \epsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right)\right]\right) \\
&\overset{(i)}{\geq} 1 - \sum_{t=0}^{T-1} \underbrace{\mathbb{P}\left(\left[\|\theta^{(t+1)} - \phi^{(t+1)}\|_2 > \epsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right)\right]\right)}_{\leq \delta/T} \\
&\geq 1 - \sum_{t=0}^{T-1} \frac{\delta}{T} \\
&= 1 - \delta
\end{aligned}$$

Where in (i), we used Bonferroni's inequality. Under the above event of probability at least $1-\delta$ where $\max_{t\in[T]}\|\theta^{(t+1)} - \phi^{(t+1)}\|_2 \leq \epsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right)$, we can continue with an identical argument to that of Theorem 2.4.2. This completes the proof. □

It follows from the above theorem that if $T = \log_{1/\kappa}\left(\frac{(1-\kappa)\|\theta^{(0)} - \theta^*\|_2}{\epsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right)}\right)$, then $\|\theta^{(T)} - \theta^*\|_2 \leq \frac{2}{1-\kappa}\epsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right)$ with probability at least $1-\delta$. Therefore, for all $t \geq T$, it follows that $\theta^{(t)} \in \mathbb{B}_2\left(\frac{2}{1-\kappa}\epsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right); \theta^*\right)$ with probability at least $1-\delta$. This concludes the general properties of convergence on Algorithm 2.

2.5 Conclusion

In Chapter 2, we explored the existing EM literature to understand the general properties of the EM algorithm and investigated its convergence behavior. We have drawn the following conclusions regarding the convergence properties of the EM algorithm.

First and foremost, we established the connection between the surrogate function Q_n and maximum likelihood estimation, as outlined in equation (1.37). Building upon this foundation, we demonstrated that for any parameter value ϕ within the parameter space Ω, the $\max\limits_{\theta \in \Omega} H_n(\theta|\phi)$ is attained at $\theta = \phi$. This result serves as a crucial underpinning for the subsequent finding. By leveraging the above result, we were able to establish that the sequence $\{L_n(\theta^{(t)})\}_{t \geq 0}$ is non-decreasing. This observation holds significance as it provides assurance of progress in likelihood improvement during the iterations of the EM algorithm. Furthermore, we delved into the conditions under which each iteration of the algorithm yields parameter estimates that strictly enhance the likelihood function L_n. This finding crucial in allowing us to derive conditions under which the limit points of the parameter sequence $\{\theta^{(t)}\}_{t \geq 0}$ coincide with stationary points of L_n.

To enhance our understanding of the convergence properties of the EM algorithm, we explored the general framework proposed by [1] for analyzing the local convergence properties of the EM algorithm. By considering the iterates of the population EM, denoted as $\{\phi^{(t)}\}_{t \geq 0}$ and under appropriate regularity conditions, we established geometric convergence of these iterates towards the true parameter θ^*. Additionally, we investigated the concept of stability of the EM operator, as introduced by Ho et al. [18]. Notably, we found that when the EM operator satisfies the stability criterion, the fitted parameters are guaranteed to converge geometrically to a point inside a ball centered around the true parameter, with high probability.

In conclusion, our analysis of the general properties of the EM algorithm has shed light on its convergence behavior. We have uncovered and re-assembled general properties of the EM into a series of conditions and results.

Chapter 3

Selective Review of the EM on Mixture Models

In this chapter, we delve deeper into the application of the EM algorithm within the context of Gaussian mixture models (GMMs) and mixed linear regression (MLR) models that were introduced in Examples 1 and 2 of this thesis. Building upon this foundation, we provide a selective survey of the existing literature on the EM for these mixture models. We aim to shed light on the key developments, particularly those relevant to the results presented in Chapter 2. By examining the known rates of convergence of the EM algorithm for various subclasses of GMMs and MLR, we hope to gain insights on the underlying similarities in the inner workings of the algorithm for these distinct classes of models.

3.1 Gaussian Mixture Models (GMMs)

From Example 1, we recall that for GMMs, the observed sample data $(y_1, ..., y_n)$ where $y \in \mathbb{R}^d$ is distributed according to (1.4) and the latent variable $z \in [k]$ is the discrete label expressing which Gaussian component an observation was sampled from. In particular, the vector parameter $(\pi_j^*, \mu_j^*, \Sigma_j^*)_{j\in[k]}$ fully describes the mixture. In Gaussian mixture models, the signal-to-noise ratio (SNR) is used to measure the separation of the components:

$$\text{SNR} := \frac{\min\limits_{i,j\in[k]} \|\mu_i^* - \mu_j^*\|_2}{\max\limits_{j\in[k]} \|\Sigma_j^*\|_F}. \tag{3.1}$$

High SNR implies the different Gaussian components are well separated from each other and therefore identifiable. On the other hand, low SNR means the Gaussian components are poorly separated and therefore weakly identifiable.

In this section, we survey the relevant literature and unveil the convergence properties of the EM algorithm when applied to GMMs. More specifically, we consider two sub-classes of Gaussian mixtures: 2-component symmetric Gaussian mixtures and k-component spherical Gaussian mixtures.

3.1.1 2-Component Symmetric Gaussian Mixtures

Recall that for general Gaussian mixtures, the sampled data $(y_1, ..., y_n)$ is distributed according to (1.29) and the latent variable $z \in [k]$ is the discrete label expressing which Gaussian component an observation was sampled from. If in addition, $k = 2$, $\mu_0^* = -\mu_1^*$, and $\Sigma_0^* = \Sigma_1^*$, it follows that the data is sampled from a 2-component symmetric Gaussian mixture. In addition, we say the mixture is balanced if $\pi_0^* = \pi_1^* = \frac{1}{2}$. In particular, the parameter vector $(\pi_0^*, \mu_0^*, \Sigma_0^*)$ fully describes the mixture.

The 2-component symmetric Gaussian mixture is widely considered the simplest subclass among Gaussian mixture models. As a result, researchers often use it as a benchmark when testing properties of the EM algorithm for Gaussian mixtures. This particular subclass serves as a foundation for understanding the algorithm's behavior and is subsequently extended to more complex subclasses. Hence, the 2-component symmetric Gaussian mixture stands as the most comprehensively studied subclass within the realm of Gaussian mixture models. In the following section, we present a collection of well-established properties of the EM algorithm as applied to this specific class of mixture models.

3.1.1.1 Case where $\boldsymbol{\pi_0^*}$ and $\boldsymbol{\Sigma_0^*}$ are known

In this setting, $\theta^{(t)} = \mu_0^{(t)}$ and the EM operator M_n is given in [1] as

$$M_n(\theta^{(t)}) := \frac{1}{n}\sum_{i=0}^{n}(2k_{\theta^{(t)}}(0|y_i) - 1)y_i.$$

In 2014, Balakrishnan et al. [1] considered the balanced case where $\pi_0^* = \pi_1^* = \frac{1}{2}$. They used the framework presented in Section 4.1 to separate the convergence analysis into two easier steps:

$$\|\theta^{(t)} - \theta^*\|_2 \le \underbrace{\|M(\theta^{(t)}) - \theta^*\|_2}_{\text{Step 1}} + \underbrace{\|M_n(\theta^{(t)}) - M(\theta^{(t)})\|_2}_{\text{Step 2}}.$$

They began with Theorem 2.4.1. They showed that if the SNR is bounded below by $\eta > 0$ and $\theta^{(0)} \in \mathbb{B}_2(\frac{\|\theta^*\|_2}{4}; \theta^*)$, then it follows that there is some $c > 0$ such that

$$\|M(\theta^{(t)}) - \theta^*\|_2 \le \kappa\|\theta^{(t)} - \theta^*\|_2$$

where $\kappa \le \exp\{-c\eta^2\}$. Building from this result, they used Theorem 2.4.2 and showed that, with probability at least $1 - \delta$,

$$\|\theta^{(t)} - \theta^*\|_2 \le \kappa^t\|\theta^{(0)} - \theta^*\|_2 + \frac{c_2}{1-\kappa}\phi(\sigma; \|\theta^*\|_2)\sqrt{\frac{d}{n}\log(1/\delta)}$$

where $c_2 > 0$ and $\phi(\sigma; \|\theta^*\|_2) = \|\theta^*\|_2\sqrt{\|\theta^*\|_2^2 + \sigma^{2*}}$. Therefore, for any r, δ, and $t \ge T$ where

$$T \ge \log_{1/\kappa}\left(\frac{\|\theta^{(t)} - \theta^*\|_2(1-\kappa)}{\phi(\sigma; \|\theta^*\|_2)}\sqrt{\frac{n}{d}\frac{1}{\log(1/\delta)}}\right) \sim \mathcal{O}(\log_{1/\kappa}(n/d)),$$

the statistical error after t iterations is guaranteed to be bounded by

$$\|\theta^{(t)} - \theta^*\|_2 \le \frac{(1+c_2)}{1-\kappa}\phi(\sigma; \|\theta^*\|_2)\sqrt{\frac{d}{n}\log(1/\delta)} \sim \mathcal{O}(\sqrt{d/n}) \tag{3.2}$$

with probability at least $1 - \delta$. In other words, for SNR sufficiently large and suitable initialization, they showed Algorithm 1 converges – with high probability – to a point inside a ball of radius $\mathcal{O}(\sqrt{d/n})$ centered around the true parameters θ^* after $\mathcal{O}(\log_{1/\kappa}(n/d))$ iterations.

Two years later, Daskalakis et al. [8] extended the above result to unbalanced cases (i.e. $\pi_0^* \neq \frac{1}{2}$) and any non-zero value of the SNR. They showed that if $\|\theta^{(0)} - \theta^*\|_2 \leq \text{SNR}$, the EM converges – with high probability – within $\mathcal{O}(\sqrt{d/n})$ Mahalanobis distance (see Definition A.2.1) of the true parameter θ^* after $\mathcal{O}\left(\frac{d}{\text{SNR}^2}\log(\sqrt{n/d})\right)$ iterations.

The above results are guaranteed to hold if the problem is well-specified meaning that the model we are fitting has the same number of components as the true model. This poses a natural concern since, in practice, we seldom have the guarantee that the problem is well-specified. In 2018, Dwivedi et al. [11] looked at over-parametrized balanced and unbalanced 2-component symmetric Gaussian mixtures; the setting where the EM fits a model with 2 components while the true model has only 1. To think about this, they thought of the true model as the limiting case of 2-component symmetric Gaussian mixture where the SNR gradually goes to 0.

In the unbalanced case, [11] found that the convergence properties were similar to that of the well-specified setting. First, they showed that the population EM iterates $\{\phi^{(t)}\}_{t\geq 0}$ obtained from Algorithm 3 satisfy

$$\|\phi^{(t)} - \theta^*\|_2 \leq \kappa^t \|\phi^{(0)} - \theta^*\|_2$$

where $\kappa = 1 - \frac{|1-2\pi_0^*|^2}{2}$. Then, they used Theorem 2.4.2 to obtain that Algorithm 1 iterates converge – with high probability – to a point inside a ball of radius $\mathcal{O}(\sqrt{d/n})$ centered around the true parameter θ^* after $\mathcal{O}(\log(n/d))$ iterations. This is pretty remarkable as they showed there isn't any significant change in the rate of convergence when compared to the well-specified setting.

In the balanced case, however, [11] found that the rate of convergence was slower when compared to the well-specified setting. First, it is clear that in this setting, $\kappa > 0$ no longer holds; this complicated the analysis. Instead, they showed that the population iterates $\{\phi^{(t)}\}_{t\geq 0}$ converge to the true parameter from an arbitrary initialization, but the progress made at $\phi^{(t+1)}$ slows down exponentially as a function of $\|\phi^{(t)} - \theta^*\|_2$. Finally, they showed that Algorithm 1 iterates converge – with high probability – to a point inside a ball of radius $\mathcal{O}((d/n)^{1/4})$ centered around θ^* after $\mathcal{O}(\sqrt{n/d})$ iterations. These rates are tight and comparatively much slower than that of the unbalanced case or that of the well-specified setting. Last year, Ren et al. [30] explained that this is because, in the unbalanced case, $L(\theta)$ given as (1.20) is no longer locally strongly concave with respect to $\theta = \mu_0$.

3.1.1.2 Case where π_0^* is known

In this setting and assuming $\Sigma_0^* = \sigma_0^* I_d$ for some $\sigma_0^* \in \mathbb{R}_{++}$, we denote $\theta^{(t)} = (\mu_0^{(t)}, \sigma_0^{(t)} I_d)$ and the EM operator M_n is given in [11] as

$$M_n(\theta^{(t)}) = (\mu_0^{(t+1)}, \sigma_0^{(t+1)})$$

where

$$\begin{aligned} \mu_0^{(t+1)} &:= \frac{1}{n}\sum_{i=0}^{n}(2k_{\theta^{(t)}}(0|y_i) - 1)y_i \\ \sigma_0^{(t+1)^2} &:= \frac{1}{d}\left(\frac{\sum_{i=1}^{n}\|y_i\|_2^2}{n} - \|\mu_0^{(t+1)}\|_2^2\right). \end{aligned}$$

There have been some efforts to understand the convergence properties of the EM when applied to 2-component symmetric Gaussian mixtures where only π_0^* is known. In 2019, Dwiveldi et al. [10] characterized the local convergence properties in all SNR regimes. They showed that under suitable initialization, there is n large enough to guarantee Algorithm 1 converges – with high probability – to a point inside a ball of radius $\tilde{\mathcal{O}}((d/n)^{1/4})$ centered around the true parameters θ^* after $\tilde{\mathcal{O}}(\sqrt{n/d})$ iterations. Their obtained rate of convergence is significantly slower than that of Case 3.1.1.1 in the well-specified setting. But interestingly, it matches very closely the rate obtained for the balanced over-parameterized setting.

3.1.2 k-Component Spherical Gaussian Mixtures

Recall that for general Gaussian mixtures, the sampled data $(y_1, ..., y_n)$ is distributed according to (1.29) and the latent variable $z \in [k]$ is the discrete label expressing which Gaussian component an observation was sampled from. If in addition, $\Sigma_j^* = \sigma_j^* I_d$ where $\sigma_j^* \in \mathbb{R}_{++}$ for all $j \in [k]$, it follows that the data is sampled from a k-component spherical Gaussian mixture; we say the mixture is balanced if $\pi_j^* = \frac{1}{k}$ for all $j \in [k]$. In particular, the parameter vector $(\pi_j^*, \mu_j^*, \sigma_j^* I_d)_{j\in[k]}$ fully describes the mixture. Already, it is clear that the previously considered 2-component symmetric Gaussian mixtures form a subclass of the k-component spherical Gaussian mixtures. Therefore, any result covered in this section also applies to the former. Further, the k-component spherical Gaussian mixture is, to our knowledge, the most complex subclass of Gaussian mixture models to have received significant attention in the EM literature. Below, we unveil a selection of known properties of the EM algorithm on this class of mixture models.

3.1.2.1 Case where $(\pi_j^*, \sigma_j^* I_d)_{j\in[k]}$ is known

In this setting, $\theta^{(t)} = (\mu_j^{(t)})_{j\in[k]}$ and the EM operator M_n is derived in Section A.1.1 as

$$M_n(\theta^{(t)}) = \frac{\sum_{i=1}^n y_i k_{\theta^{(t)}}(j|y_i)}{\sum_{i=1}^n k_{\theta^{(t)}}(y_i, j|y_i)}$$

In 2018, Zhao et al. [41] described the local convergence properties of the EM algorithm for suitably initialized and separated mixtures of k spherical Gaussians with identity covariance (i.e. $\sigma_j^* = 1$ for all $j \in [k]$). They followed the framework for convergence analysis introduced in [1]:

$$\|\theta^{(t+1)} - \theta^*\|_2 \leq \underbrace{\|M(\theta^{(t)}) - \theta^*\|_2}_{\text{Step 1}} + \underbrace{\|M_n(\theta^{(t)}) - M(\theta^{(t)})\|_2}_{\text{Step 2}}.$$

First, for SNR lower bounded by $\tilde{\Omega}(\sqrt{k})$ and suitable initialization, they established that for any $\theta^{(t)}$, one iteration of Algorithm 3 will satisfy

$$\|M(\theta^{(t)}) - \theta^*\|_2 \leq \frac{1}{2}\|\theta^{(t)} - \theta^*\|_2.$$

Next, they connected the latter result to Algorithm 1. Letting $\delta = \frac{2k}{n}, C_3 > 0$ and with suitable initialization $\theta^{(0)}$, they showed

$$\max_{j\in[k]} \|\mu_j^{(t)} - \mu_j^*\|_2 \leq \frac{1}{2^t} \max_{j\in[k]} \|\mu_j^{(0)} - \mu_j^*\|_2 + \frac{3R_{\max}}{\pi_{\min}}\sqrt{\frac{C_3 kd \log n}{n}}$$

with probability at least $1 - \delta$. Therefore, for any $t \geq T$ where

$$T \geq \log_2\left(\frac{\max_{j\in[k]} \|\mu_i^{(t)} - \mu_j^*|_2 \pi_{\min}\sqrt{n}}{3R_{\max}\sqrt{C_3 kd \log n}}\right) \sim \mathcal{O}\left(\log_2\sqrt{\frac{n}{kd}}\right),$$

the statistical error after t iterations is guaranteed to be bounded by

$$\|\theta^{(t)} - \theta^*\|_2 \lesssim \mathcal{O}(\sqrt{kd/n}) \tag{3.3}$$

with probability at least $1-\delta$. In other words, they showed Algorithm 1 iterates converges – with high probability – to a point inside a ball of radius $\mathcal{O}(\sqrt{\frac{kd}{\pi_{\min}^2 n}})$ centered around the true parameter θ^* after $\mathcal{O}\left(\log_2\sqrt{\frac{n}{kd}}\right)$ iterations.

3.1.2.2 Case where no parameter is known

In this setting, $\theta^{(t)} = (\pi_j^{(t)}, \mu_j^{(t)}, \sigma_j^{(t)} I_d)_{j\in[k]}$ and the EM operator M_n and population EM operator M are given in [21] as

$$M_n(\theta^{(t)}) = (\pi_j^{(t+1)}, \mu_j^{(t+1)}, \sigma_j^{(t+1)} I_d)_{j\in[k]})$$

where

$$\begin{aligned}
\pi_j^{(t+1)} &= \frac{\sum_{i=1}^n k_{\theta^{(t)}}(j|y_i)}{n} && \text{, for } j \in [k];\\
\mu_j^{(t+1)} &= \frac{\sum_{i=1}^n y_i k_{\theta^{(t)}}(j|y_i)}{\sum_{i=1}^n k_{\theta^{(t)}}(j|y_i)} && \text{, for } j \in [k];\\
\sigma_j^{(t+1)^2} &= \frac{\sum_{i=1}^n \|y_i - \mu_j^{(t+1)}\|_2^2 k_{\theta^{(t)}}(j|y_i)}{\sum_{i=1}^n k_{\theta^{(t)}}(j|y_i)} && \text{, for } j \in [k].
\end{aligned}$$

Around a decade ago, Moitra et al. [26] and Hardt et al. [15]) showed that with no restrictions on the SNR, the worst case instances can require as much as $\Omega(e^k)$ samples to recover the true parameter with high probability. Then in 2017, Regev and Vijayaraghavan [29] established that $\Omega(\sqrt{\log k})$ separation is necessary and sufficient for recovering – with high probability – the true parameter within ϵ distance; a polynomial (i.e. poly$(k, d, 1/\epsilon)$) number of samples are necessary. Also, they extended this result and showed that restricted to $o(\sqrt{\log k})$ separation, a super-polynomial number of samples is required to recover the true parameter with high probability. Three years later, Kwon et al. [21] tightened this bound in the $\Omega(\sqrt{\log k})$ separation regime. They showed that with suitable initialization (i.e. $\epsilon \lesssim 1/k$), only $\tilde{\mathcal{O}}(kd/\epsilon^2)$ samples are necessary for EM to converge within a ball of radius $\mathcal{O}(\epsilon)$ of the true parameter θ^*. Kwon et al. [21] achieved this by extending results of Section 3.1.2.1 that were obtained two years prior by Zhao et al. [41]; we explain below.

Like many others, Kwon et al. used the framework for convergence analysis introduced in [1]:

$$\|\theta^{(t+1)} - \theta^*\|_2 \leq \underbrace{\|M(\theta^{(t)}) - \theta^*\|_2}_{\text{Step 1}} + \underbrace{\|M_n(\theta^{(t)}) - M(\theta^{(t)})\|_2}_{\text{Step 2}}.$$

But, their analysis differs in two major ways. First, they use the sample-splitting Algorithm 2 which assumes each iteration is based on independent sub-samples. Also, in step 1, they condition the expectation on so-called "good" and "bad" samples. When the initialization is close enough to the true parameters, many samples dubbed "good" have approximately the right weights $k_\theta(z|y)$ according to which Gaussian component they come from. Denoting ϵ_{good} as the set where the E-step assigns approximately the right labels the estimation error in the means after one EM step is given by

$$\mathbb{E}[\beta Y|\theta^*] - \mu^* = (\mathbb{E}[\beta Y|\epsilon_{good}, \theta^*] - \mu^*)\mathbb{P}(\epsilon_{good}) + (\mathbb{E}[\beta Y|\epsilon_{good}^C, \theta^*] - \mu^*)\mathbb{P}(\epsilon_{good}^C)$$

where $\beta = (k_{\theta^{(t)}}(1|Y), ..., k_{\theta^{(t)}}(k|Y))$. Under suitable separation and if we are close enough to the true parameters, we will see that $\mathbb{E}[kY|\epsilon_{good}, \theta^*] \approx \mu^*$ and $\mathbb{P}(\epsilon_{good}^C) \approx 0$. For Algorithm 2, they proved that the EM converges locally to the true parameters under $\Omega(\sqrt{\log k})$ separation. In fact, they show that for suitable initialization, Algorithm 2 converges locally – with high probability – to a point inside a ball of radius $\mathcal{O}(\sqrt{kd/n})$ centered around the true parameter θ^* after $\mathcal{O}(\log\sqrt{\frac{n}{kd}})$ iterations. Although the analysis is performed with Algorithm 2, it is interesting to note that the rates are similar to that obtained above by Zhao et al. [41]. What's more [21] obtain the rates in a tighter separation regime.

3.2 Mixed Linear Regression Models

From Example 2, we recall that for general mixtures of k-linear regression, the observed sample data $((y_1, x_1), ..., (y_n, x_n))$ is distributed according to (1.11)(1.12) and the latent variable $z \in [k]$ is the discrete label expressing which regression component an observation was sampled from. In particular, the vector parameter $(\pi_j^*, \mu_j^*, \sigma_j^*)_{j\in[k]}$ fully describes the mixture. In MLR models, the SNR is used to measure the separation of the regression components:

$$\text{SNR} := \frac{\min\limits_{i,j\in[k]} \|\mu_i^* - \mu_j^*\|_2}{\max\limits_{j\in[k]} \sigma_j^*}. \tag{3.4}$$

The SNR measure the identifiability of the parametric models. In particular, a low SNR means the models is weakly identifiable.

In this section, we survey the relevant literature and unveil the convergence properties of the EM algorithm when applied to MLR models. More specifically, we consider two sub-classes of MLR models: 2-component symmetric MLR and k-component MLR.

3.2.1 2-Component Symmetric Mixed Linear Regression

Recall from Example 2 that for general mixtures of k-linear regression, the observed sample data
$((y_1, x_1), ..., (y_n, x_n))$ is distributed according to (1.11)(1.12) and the latent variable $z \in [k]$ is the discrete label expressing which regression component an observation was sampled from. If in addition, $k = 2$, $\mu_0^* = \mu_1^*$ and $\sigma_0^* = \sigma_1^*$, it follows that the data is sampled from a 2-component symmetric MLR model; we say the mixture is balanced if $\pi_0^* = \pi_1^* = \frac{1}{2}$. In particular, the parameter vector $(\pi_0^*, \mu_0^*, \sigma_0^*)$ fully describes the mixture.

Similarly to its Gaussian mixture model counter-part, the 2-component symmetric MLR class of models is regarded as the simplest sub-class of MLR models. As a result, whenever a property of the EM for MLR is to be tested, it will more often than not be checked for this sub-class of models first. Then, it is extended, if possible, to more general sub-classes. For this reason, the 2-component symmetric MLR is easily the most well understood sub-class of MLR models. Below, we unveil a selection of the known properties of the EM algorithm on this class of mixture models.

3.2.1.1 Case where π_0^* and σ_0^* are known

In this setting, $\theta^{(t)} = \mu_0^{(t)}$ and the EM operator M_n is given in [] as

$$M_n(\theta^{(t)}) := \left(\frac{1}{n}\sum_{i=1}^{n} y_{E_i} y_{E_i}^T\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \tanh\left(\frac{y_{R_i} y_{E_i}^T \theta^{(t)}}{\sigma*^2}\right) y_{R_i} y_{E_i}\right) \tag{3.5}$$

In 2020, Kwon et al. [23] completely characterized the EM algorithm's convergence behavior for 2-component symmetric MLR models. They did so for every SNR regime whilst not making the restrictive use of sample-splitting Algorithm 2 in the analysis. They considered the following two categories of SNR:

$$\begin{aligned} &\text{Low SNR:} && \|\theta^*\|_2 \lesssim (d/n)^{1/4} \\ &\text{High SNR:} && (d/n)^{1/4} \lesssim \|\theta^*\|_2. \end{aligned}$$

In the high SNR regime, they followed the framework introduced in Section 4.1 to separate the convergence analysis into two easier steps:

$$\|\theta^{(t)} - \theta^*\|_2 \leq \underbrace{\|M(\theta^{(t)}) - \theta^*\|_2}_{\text{Step 1}} + \underbrace{\|M_n(\theta^{(t)}) - M(\theta^{(t)})\|_2}_{\text{Step 2}}.$$

First, they showed that for any initialization, it follows that

$$\|M(\theta^{(t)}) - \theta^*\|_2 \leq \kappa\|\theta^{(t)} - \theta^*\|_2$$

where $\kappa \leq 1 - \frac{1}{8}\|\theta^*\|_2^2$. Next, they calculated $\epsilon_M^{unif}(n, \delta) = cr\sqrt{d\log^2(n/\delta)/n}$ for some $c, r > 0$. Finally, they used Theorem 2.4.2 and obtained

$$\|\theta^{(t)} - \theta^*\|_2 \leq \kappa\|\theta^{(t-1)} - \theta^*\|_2 + cr\sqrt{d\log^2(n/\delta)/n}.$$

The remainder of their proof is spent unravelling the above recurrence relation to show Algorithm 1 iterates converges – with high probability – to a point inside a ball of radius $\tilde{\mathcal{O}}(\max\{1, \|\theta^*\|_2^{-1}\}\sqrt{d/n})$ centered around the true parameters θ^* after

$\mathcal{O}(\max\{1, \|\theta^*\|_2^{-2}\}\log(n/d))$ iterations. Interestingly, the statistical error's upper bound dependency on $\sqrt{d/n}$ is consistent, in all SNR regimes, with the rates obtained in 2013 by Chen et al. [6] for their proposed information-theoretically optimal algorithm. In any case, they also conjectured that in the high SNR regime, the EM actually has a superlinear rate of convergence; this was already proven, the same year, by Gosh et al. [14] in the noiseless setting (i.e. $\sigma_0^* = 0$).

In the low SNR regime, the rates worsen. Contrarily to the high SNR regime, they did not follow the framework for analysis described above. Instead, their proof hinges on the idea that EM cannot distinguish between $\theta^* = 0$ and $\theta^* \neq 0$. They realized that, given the low SNR regime, showing $\theta^{(t)} \leq \mathcal{O}((d/n)^{1/4})$ is sufficient for $\|\theta^{(t)} - \theta^*\|_2 \leq \mathcal{O}((d/n)^{1/4})$. In the end, they obtained the guarantee that Algorithm 1 iterates converge – with high probability – to a point inside a ball of radius $\tilde{\mathcal{O}}((d/n)^{1/4})$ centered around the true parameters θ^* after $\tilde{\mathcal{O}}(\sqrt{n/d})$ iterations. As expected, the radius of the ball decreases while the number of iterations required to reach it increases.

3.2.1.2 Case where π_0^* is known

In this setting, we denote $\theta^{(t)} = (\mu_0^{(t)}, \sigma_0^{(t)})$ and the EM operator M_n is given in [23] as

$$M_n(\theta^{(t)}) = (\mu_0^{(t+1)}, \sigma_0^{(t+1)})$$

where

$$\begin{aligned}\mu_0^{(t+1)} &:= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\tanh\left(\frac{y_i x_i^T\theta^{(t)}}{\sigma*^2}\right) y_i x_i\right)\\ \sigma_0^{(t+1)} &:= \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \langle\mu_0^{(t+1)}, x_i\rangle^2}.\end{aligned}$$

In 2020, Kwon et al. [23] attempted to characterize the convergence rate of the EM algorithm for balanced 2-component symmetric MLR models where only π_0^* is known. Whilst they remarked that the main challenge with more unknown variables comes from the analysis of the population EM, they still obtained a result in the low SNR regime. Given appropriate initialization satisfying $|\sigma^{2^{(0)}} - 1| \leq 0.04$, they showed the EM iterates for $\mu^{(t)}, \sigma^{(t)}$ converge – with high probability – to a point inside a ball of radius $\tilde{\mathcal{O}}((d/n)^{1/4})$), $\tilde{\mathcal{O}}(\sqrt{d/n})$ respectively centered around the true parameters mu^*, σ^* after $\tilde{\mathcal{O}}(\sqrt{n/d})$ iterations. Formally, for T suitably large, it follows that

$$\|\theta^{(T)} - \theta^*\|_2 \lesssim \mathcal{O}((d/n)^{1/4})$$
$$\|\sigma^{(T)} - \sigma^*\|_2 \lesssim \mathcal{O}((d/n)^{1/2})$$

with high probability.

3.2.2 k-Component Mixed Linear Regression

Recall from Example 2 that for k-component MLR models, the observed sample data $((y_1, x_1), ..., (y_n, x_n))$ is distributed according to (1.11)(1.12) and the latent variable $z \in [k]$ is the discrete label expressing which regression component an observation was sampled from. In particular, the parameter vector $(\pi_j^*, \mu_j^*, \sigma_j^*)_{j\in[k]}$ fully describes the mixture. It is clear the previously considered 2-component symmetric MLR models form a subclass of the k-component MLR models. Therefore, any result covered in this section also applies to the former. Below, we unveil a selection of known properties of the EM algorithm on this class of mixture models.

3.2.2.1 Case where $(\sigma_j^*)_{j\in[k]}$ is known

In this setting, we denote $\theta^{(t)} = (\pi_j^{(t)}, \mu_j^{(t)})_{j\in[k]}$ and the EM operator M_n is given in [22] as

$$M_n(\theta^{(t)}) = (\pi_j^{(t+1)}, \mu_j^{(t+1)})_{j\in[k]}$$

where for all $j \in [k]$,

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^n k_{\theta^{(t)}}(j|y_i, x_i)}{n}$$
$$\mu_j^{(t+1)} = \left(\sum_{i=1}^n k_{\theta^{(t)}}(j|y_i, x_i) x_i x_i^T\right)^{-1} \left(\sum_{i=1}^n k_{\theta^{(t)}}(j|y_i, x_i) y_i x_i\right).$$

In 2019, Kwon et al. [22] characterized the convergence of the sample-splitting Algorithm 2 for SNR $\geq \tilde{\Omega}(k)$. They showed that when $\theta^{(0)}$ satisfies $\|\theta^{(0)} - \theta^*\|_2 \leq \mathcal{O}(1/k)$, Algorithm 2 iterates converges – with high probability – to a point within a ball of radius $\mathcal{O}(\epsilon)$ centered around the true parameters θ^* after $\tilde{\mathcal{O}}(k^2 d/\epsilon^2)$ iterations. They also considered the analysis with Algorithm 1; no sample-splitting. Their results differ in that their is an added polynomial dependence on $\max_{i,j} \|\theta_i^* - \theta_j^*\|_2$ when bounding the statistical error. They conjecture that this dependence is an artifact of the analysis.

3.3 Conclusion

After conducting an extensive literature survey on the EM algorithm applied to mixture models, we have observed that the considered settings can be categorized into either fast or slow convergence regimes. In the fast convergence category, the finite-sample EM algorithm has been found to converge, with high probability, to a point inside a ball of radius $\tilde{\mathcal{O}}(\sqrt{d/n})$ centered around the true parameters θ^* after $\tilde{\mathcal{O}}(\log(n/d))$ iterations. This category primarily encompasses scenarios where $(\mu_j^*)_{j\in[k]}$ are unknown. Specifically, the fast convergence is observed in the 2-component symmetric Gaussian mixture, 2-component symmetric MLR in the high SNR regime, k-component spherical Gaussian mixture, and the over-parametrized 2-component symmetric Gaussian mixture in the unbalanced regime. On the other hand, the slow convergence category entails scenarios where the finite-sample EM algorithm iterates converge, with high probability, to a point inside a ball of radius $\mathcal{O}((d/n)^{1/4})$ centered around θ^* after $\mathcal{O}(\sqrt{n/d})$ iterations. Our literature survey revealed that latent variable models falling under this category include the 2-component symmetric Gaussian mixture where π_0^* is known, 2-component symmetric mixed linear regression in the regime of low SNR and known (π_0^*, σ_0^*), 2-component symmetric mixed linear regression with known π_0^*, and the over-parametrized balanced 2-component symmetric Gaussian mixture with known (π_0^*, Σ_0^*). Interestingly, we noted that the convergence category for each setting is not determined by the number of components in the mixture model. Instead, a combination of multiple unknown parameters and low SNR, leading to weakly identifiable models and a more complicated objective function, appears to be the primary factor contributing to a slower convergence rate. It is especially interesting that convergence properties of the EM hold across different classes of mixture models, hinting at some inner-structure of the EM for mixture models. Finally, we note a strong lack of understand for the convergence of the EM on GMMs or MLR models when the variance parameters is unknown. According to our literature survey, this is mainly due to the objective function L_n losing regularity properties with respect to that parameterization.

Chapter 4

Discussion of Related Topics

Throughout this thesis, we have examined various aspects of the EM algorithm, including its convergence to stationary points of L_n or the true parameter θ^*, as well as local convergence properties within GMMs and MLR models. In this chapter, we delve into the contentious issue of the initialization of the EM, SNR, parameterization, and also present some recent research advancements from the past year. By addressing these subjects, we aim to close some of the gaps that were not properly discussed in the previous chapters as well as explore novel directions in the field.

4.1 Simulation Study: Effect of SNR and Parameterization On Convergence of the EM

In this section we perform a simulation study to study the effects of SNR and parameterization of the parameteric model of interest on the convergence of the EM (Algorithm 1). To allow for comparison with known results presented in Chapter 3, we consider, as benchmark, the 2-component symmetric GMM class of models.

4.1.1 Preliminaries

In our simulation study, the true model considered is the 2-component 10 dimensional symmetric Gaussian mixture class of models described in Section 3.1.1. Each observation is to be sampled from

$$Y \sim \frac{1}{2}\mathcal{N}(\mu_0^*, \sigma_0^* I_d) + \frac{1}{2}\mathcal{N}(-\mu_0^*, \sigma_0^* I_d) \tag{4.1}$$

where $\mathcal{N}(\mu_0^*, \Sigma_0^*)$ is a multivariate Gaussian distribution with mean $\mu_0^* \in \mathbb{R}^{10}$ and $\sigma_0^* \in \mathbb{R}_{++}$.

4.1.1.1 SNR regimes and Data Generation

Our simulation study explores several different regimes of SNR; in particular, SNR$\in$ $0.5, 0.75, 1, 1.8, 2, 2.5$. The SNR given for GMMS as (3.1) measures the ratio of the signal strength to the noise level and plays a crucial role in assessing the performance of the EM algorithm. It quantifies the separability between the underlying components of the GMMs.

In each SNR regime, we set the true means μ_0^* as a k-dimensional vector of ones (i.e. $\mu_0^* = (1)_{j\in[k]}$). To achieve the desired SNR, we adjust the value of σ_0^* while keeping the means fixed. By manipulating the variance, we control the overlap or separability between the Gaussian components, thereby exploring a range of identifiable scenarios. This allows us to examine the impact of SNR on the convergence behavior of the EM algorithm.

Once the true parameters are determined for each SNR regime, we generate artificial data sets using Python. We employ numpy's random.multivariate_normal function to sample a total of 200 observations from the Gaussian mixture model. This function ensures that the generated data reflect the characteristics of the specified GMM in (4.1).

To ensure the robustness of our findings, we repeat the data generation process ten times for each SNR regime. By generating multiple datasets, we account for the inherent randomness and variability in the simulation. This allows us to obtain more robust results.

4.1.1.2 Performance Metrics

We employ two distinct performance metrics to evaluate the accuracy of the estimates obtained from the iterates of Algorithm 1: optimization error and statistical error. These metrics enable us to assess how well the algorithm approximates the true parameters of the underlying model.

The statistical error measures the dissimilarity between the parameter estimate at iteration t, denoted as $\theta^{(t)}$, and the true parameter θ^*. It quantifies the closeness of the estimated parameter to the ground truth and is defined as $\|\theta^{(t)} - \theta^*\|_2$.

In addition to the statistical error, we employ the optimization error as a performance metric. This error quantifies the discrepancy between the parameter estimate at iteration t, and the limit of the iterates as t approaches infinity, denoted as θ^∞. The optimization error is computed as $\|\theta^{(t)} - \theta^\infty\|_2$. This metric allows us to assess the convergence rate and efficiency of the algorithm by measuring the proximity of the current estimate to the ultimate parameter value.

It is important to note that the optimization error provides insights into the rate at which the algorithm converges towards the final parameter estimate, while the statistical error evaluates the accuracy of the estimates with respect to the true model parameters. By considering both metrics, we obtain a comprehensive understanding of the algorithm's performance throughout the iterations and its ability to capture the underlying characteristics of the symmetric Gaussian mixture model.

4.1.1.3 Experimental Set-up

Our simulation study consists of two distinct numerical experiments: Experiment 1 and Experiment 2.

In Experiment 1, we investigate the impact of SNR on the convergence of Algorithm 1 in the context of a 2-component 10-dimensional symmetric Gaussian mixture, with only μ_0^* being unknown. To explore the effect of different SNR regimes, we initialize 10 artificial datasets, each comprising 200 samples, as described previously. For each dataset, we perform 20 iterations of Algorithm 1 and record the optimization error and statistical error after each iteration. This process is repeated for 5 different SNR regimes. By examining the optimization and statistical errors across these regimes, we gain insights into how SNR influences the convergence behavior of the algorithm.

In Experiment 2, we simultaneously examine the impact of unknown variance in different SNR regimes for the 2-component 10-dimensional symmetric Gaussian mixture model. We consider three distinct SNR regimes and initialize 10 artificial datasets, each consisting of 200 samples, as described previously. For each dataset, we execute 20 iterations of Algorithm 1 in two scenarios: when only μ_0 is unknown and when both μ_0 and σ_0^* are unknown. We calculate the optimization error and statistical error after each iteration. By comparing the results across the different SNR regimes and unknown parameter scenarios, we gain valuable insights into the combined effect of SNR and parameter uncertainty on the algorithm's convergence.

Throughout both experiments, we ensure consistency in the initial parameter values by calculating $\theta^{(0)}$ in such a way that $\|\theta^{(0)} - \theta^*\|_2$ remains constant across all regimes and experiments. This approach guarantees that the initial discrepancy between the estimated and true parameters is the same across all experiments, allowing for a fair and meaningful comparison of the optimization and statistical errors.

4.1.1.4 Implementation Details

We do not use any specific python libraries for evaluating iterations of the EM algorithm other than standard libraries such as numpy, matplotlib, etc. Instead, we implement the algorithm ourselves and perform the iterations according to the EM operator M_n described in Section 3.1.1.2 as

$$M_n(\theta^{(t)}) = (\mu_0^{(t+1)}, \sigma_0^{(t+1)})$$

where

$$\mu_0^{(t+1)} := \frac{1}{n}\sum_{i=0}^{n}(2k_{\theta^{(t)}}(0|y_i) - 1)y_i$$
$$\sigma_0^{(t+1)^2} := \frac{1}{d}\left(\frac{\sum_{i=1}^{n}\|y_i\|_2^2}{n} - \|\mu_0^{(t+1)}\|_2^2\right).$$

4.1.2 Results

We present the results obtained from our numerical experiments on the EM algorithm in the context of 2-component symmetric GMMs. The primary objective of these experiments is to highlight the algorithm's limitations and gain insights into its inner workings.

In the first experiment, we aimed to examine the effect of the SNR of the true model on the convergence of the EM algorithm. As depicted in Figure A.10, we observe a consistent trend where both the optimization error and statistical error worsen as the SNR decreases. This finding aligns with the well-established understanding that the EM algorithm tends to perform less effectively on weakly identifiable models.

The second experiment focuses on comparing the local convergence rate of the EM algorithm for a 2-component 10-dimensional symmetric GMM when the variance is known or unknown. Our goal was to investigate whether the slower convergence rate observed when σ_0^* is unknown can be mitigated for models with large SNR and strong identifiability. To explore this, we conducted the experiment across three SNR regimes: $\frac{1}{2}$, 1, and 2. The results are summarized in Figure A.7, providing insights into the impact on the EM algorithm's convergence.

Firstly, we note that the optimization error remains consistent across all tested SNR regimes, regardless of whether the variance is known or unknown. However, this is not the case for the statistical error. We find that decreasing the SNR leads to larger statistical errors, regardless of the knowledge of the variance. In fact, when SNR $= \frac{1}{2}$, we even observe that the EM's fitted parameters exhibit a larger statistical error compared to the initial parameters. Furthermore, the disparity in the convergence rates between the scenarios where σ_0^* is known and unknown aligns with the findings discussed in Chapter 3.

4.2 Brief Discussion on the Topic of Initialization

Initialization poses a challenge in the application of the EM algorithm, as the convergence results presented in Chapter 3 often rely on assumptions about the initial parameter $\theta^{(0)}$. However, such assumptions are unlikely to hold in general. In this section, we explore strategies that researchers have developed to ensure a favorable initialization in the context of MLR models and GMMs, which are widely studied in the literature.

For MLR models, a promising approach involves leveraging both method of moments estimators and the spectral structure of the data. In the case of a 2-component MLR, Yi et al. [39] proposed an initialization method based a specific matrix estimated from the data as $\frac{1}{n}\sum_i^n y_i^2 x_i x_i^T$. More precisely, they estimate the leading two eigen-vectors of said matrix that span the same space as $\theta^* = (\mu_0^*, \mu_1^*)$. This technique guarantees an initial parameter $\theta^{(0)} = (\mu_0^{(0)}, \mu_1^{(0)})$ that satisfies $\|\theta^{(0)} - \theta^*\|_2 \leq \|\mu_1^* - \mu_0^*\|_2$. Three years later, Yi et al. [40] extended the approach to noiseless k-component MLR models. They employed the method of moments to estimate $(\mu_j^*)_{j\in[k]}$ and then used a tensor factorization algorithm to obtain the initial estimates $(\mu_j^{(0)})_{j\in[k]}$. Remarkably, this initialization method guarantees that $\|\mu_j^{(0)} - \mu_j^*\|_2 \leq \epsilon$ for all $j \in [k]$ with high probability, provided the sample size $n \geq \mathcal{O}(\frac{1}{\epsilon^2})$.

In contrast, GMM initialization is relatively simpler. The widely-used k-means algorithm has proven to be effective in providing reasonable estimates for initializing the EM. In 2020, Kwon et al. [21] used the k-means algorithm to ensure initialization satisfying $\|\mu_i^{(0)} - \mu_i^*\|_2 \leq \frac{1}{4}\min_{i\neq j}\|\mu_i^* - \mu_j^*\|_2$. This initialization method is not only straightforward to implement but also well understood, making it a practical choice.

By addressing the crucial issue of initialization, these strategies enhance the effectiveness and reliability of the EM algorithm in both MLR and GMM settings. The presented techniques offer researchers practical and theoretically grounded approaches to initialize the EM algorithm, fostering its widespread applicability in real-world scenarios.

4.3 New Research Directions

Throughout this thesis, we have observed the extensive research conducted on the EM algorithm in the past decade. However, with the abundance of work, it can be challenging to identify the latest and most impactful directions for EM research. In light of this, we dedicate this section to shed light on significant contributions that we believe will shape the future of EM research. By highlighting these influential contributions, we aim to provide valuable insights and inspire further exploration in EM research.

4.3.1 (In)stability of the EM Operator

In Chapter 2, we introduced the framework proposed by Balakrishnan et al. [1] for analyzing the local convergence of the EM algorithm. This framework, encapsulated by the inequality

$$\|\theta^{(t)} - \theta^*\|_2 \leq \underbrace{\|M(\theta^{(t)}) - \theta^*\|_2}_{\text{Step 1}} + \underbrace{\|M_n(\theta^{(t)}) - M(\theta^{(t)})\|_2}_{\text{Step 2}},$$

has significantly influenced subsequent research. However, it falls short in addressing one critical aspect: the analysis of local convergence when the EM operator M_n is unstable, meaning that $\epsilon_m^{unif}(n, \delta)$ does not exist. This limitation has led to the misconception that unstable algorithms are inherently sluggish and undesirable. But is this really the case?

Last year, Ho et al. [18] developed a framework that specifically tackles Step 2 in scenarios where M_n exhibits instability. They provide [18, Theorem 2] that gives conditions under which the sequence $\{\theta^{(t)}\}$ converges locally around the true parameter θ^* with high probability. They perceive their framework as a natural extension of Theorem 2.4.2 providing a comprehensive understanding of the EM's local convergence properties even when the EM operator is unstable. Notably, their framework extends beyond the EM algorithm itself and also encompasses Newton's method and gradient descent for approximating the maximum likelihood estimator (MLE) in parametric models with latent variables. Intriguingly, they. demonstrated that unstable algorithms can achieve faster convergence and superior accuracy compared to their stable counterparts in the context of GMMs and non-linear mixed regression.

The work of Ho et al. [18] challenges the conventional belief that stability is always advantageous and encourages researchers to explore the potential benefits of embracing instability. Furthermore, their framework's versatility extends its applicability to other optimization algorithms beyond the EM, promising fresh insights into a wide range of latent variable models. By expanding our understanding of instability in the EM operator and its implications, this research has the potential to pave the way for more efficient and accurate estimation techniques in various practical settings.

4.3.2 Mirror Descent

Recent work in 2022 by Kunstner et al. [20] offered a new perspective for analyzing the non-asymptotic convergence properties of the EM. Their results stand out from the rest because they do not depend on problem specific constants. This is a big deal since, as we demonstrated through our literature survey in Chapter 3, any slight variations of the specification for a parametric model with latent variable can considerably complexity the analysis. Moreover, the approach with which the results are obtained makes no smoothness or concavity assumption for Q_n, which is commonly assumed for general results on the EM. However, one main draw-back is that their analysis is restricted to latent variable models where the complete data distribution belongs to the exponential family which GMMs are a part of.

First, they establish that the EM algorithm is equivalent to a mirror descent algorithm in the sense of Beck et al. [4] with Bregman divergence (see Definition A.2.2). In particular, Kunstner et al. [20] demonstrate that the EM operator can be seen as minimizing the following objective:

$$\min_{\phi \in \Omega} L_n(\phi) \leq L_n(\theta) + \langle \nabla L_n(\theta), \phi - \theta \rangle + D_A(\phi, \theta),$$

where $D_A(\phi, \theta)$ denotes the Bregman divergence induced by $A(\cdot)$, which is the log-partition of the exponential family distribution.

Building upon this equivalence, they utilize a result from Lu et al. [24, Theorem 3.1] for mirror descent algorithms. This result enables them to establish a non-asymptotic bound of the form:

$$L_n(\theta^{(T)}) - L_n(\theta^*) \leq \tfrac{1}{T} D_A(\theta^{,}\theta^{(0)}),$$

when the EM algorithm is initialized within a locally-convex region with a minimum at θ^*. Importantly, this bound is independent of problem-specific constants, providing a more general framework for convergence analysis. Furthermore, Kunstner et al. demonstrate that under mild and commonly adopted assumptions, it follows that $\min_{t \leq T} D_A(\theta^{(t)}, \theta^{(t+1)}) \leq \frac{L_n(\theta^{(0)}) - L_n(\theta^*)}{T}$ holds.

This contribution offers fresh insights into the non-asymptotic convergence properties of the EM algorithm. Moreover, by avoiding reliance on problem-specific constants and relaxing commonly made smoothness and concavity assumptions, their results provide a more general and flexible framework for analyzing the convergence properties of the EM algorithm.

Chapter 5

Conclusion

This thesis has provided an exploration of the Expectation-Maximization (EM) algorithm and its various applications in parameter estimation with latent variables. In Chapter 1, we began by tracing the historical development of the EM algorithm, introducing the problem of parameter estimation and presenting the formal framework of the EM algorithm for approximating the MLE. Important variants of the EM algorithm, such as sample splitting EM, Population EM, and General EM, were presented and discussed. Throughout, we provided examples in the form of the Gaussian Mixture Models (GMMs) and Mixed Linear Regression (MLR) models and provided derivation where necessary.

Chapter 2 focused on establishing general convergence properties for the EM algorithm as we uncovered some of its most famous results. Through rigorous proofs from [9] [37] [36] [1], we demonstrated the non-decreasing nature of the likelihood function and derived conditions under which the EM algorithm yields parameter estimates that improve the likelihood. Furthermore, we explored convergence properties of the EM algorithm, including the convergence to stationary points of the likelihood and the contractive behavior leading to convergence to the true parameter. Notably, we highlighted the significance of the framework for analyzing the local convergence properties of the EM. Stability conditions, which guarantee geometric convergence of the EM algorithm to a ball centered around the true parameter.

Moving on to Chapter 3, we conducted a thorough survey of the EM literature in the context of GMMs and MLR models. Our review revealed distinct convergence regimes, namely fast convergence and slow convergence. In the fast convergence regime, we observed that the finite-sample EM algorithm converges to a ball of radius $\mathcal{O}(\sqrt{d/n})$ around the true parameters after $\mathcal{O}(\log(\frac{n}{d}))$ iterations. On the other hand, the slow convergence regime entailed scenarios where the finite-sample EM algorithm required $\mathcal{O}(\sqrt{\frac{n}{d}})$ iterations to converge to a ball of radius $\mathcal{O}((d/n)^{1/4})$ centered around the true parameters. Remarkably, we found that the convergence category for each setting was not solely determined by the number of components in the mixture model, but rather by the combination of multiple unknown parameters and low SNR conditions, leading to weaker identifiability and more challenging optimization landscapes. Moreover, we were surprised to notice very consistent rates of convergence across the two distinct classes of mixture models. The consistent convergence rates observed across distinct classes of mixture models highlight its robustness and motivate further investigations into its underlying mechanisms.

Finally, in Chapter 4, we explored additional topics related to the EM algorithm. We began with a simulation study investigating the effects of the SNR and parameterization on the local convergence properties of the EM. We discussed the contentious issue of EM initialization and examined recent research directions from the past year, aiming to identify emerging areas of interest and potential future advancements.

In conclusion, our selective review of the existing EM literature reveals the EM algorithm as a practical and widely adopted approach for estimating parameters in models with latent variables. While its convergence properties are still being uncovered, the EM algorithm remains – arguably – the most popular algorithm for estimating the parameters of parametric models with latent variables. This, on its own, is a testament to its effectiveness and reason enough to dedicate resources to unveiling its secrets.

Appendix A

Appendix

A.1 Derivations

A.1.1 Derivation of the EM Operator in Example 5

For the parameter estimation problem described in Example 3, the EM operator is given in Example 5 as (1.29); we perform the complete derivation hereunder. Referring to the EM operator's formal definition in (1.26) we set-out to evaluate the global maximizer of (1.28) over the set Ω given as (1.22). Treating $\theta^{(t)}$ as a constant, the derivation of (1.29) translates to solving the constrained maximization problem in $2 \times k$ variables that is depicted below:

$$\max_{(\pi_j,\mu_j)_{j=0}^{k-1} \in \Omega} \mathcal{Q}_n^1((\pi_j)_{j\in[k]}) + \mathcal{Q}_n^2((\mu_j)_{j\in[k]}) \tag{A.1}$$

where

$$\mathcal{Q}_n^1((\pi_j)_{j=0}^{k-1}) := \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=0}^{k-1}\log\left(\frac{\pi_j}{(2\pi)^{\frac{d}{2}}|\Sigma_j^*|^{\frac{1}{2}}}\right)k_{\theta^{(t)}}(j|y_i)\right], \tag{A.2}$$

$$\mathcal{Q}_n^2((\mu_j)_{j=0}^{k-1}) := \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=0}^{k-1} -\frac{1}{2}(y_i-\mu_j)^T\Sigma_j^{*-1}(y_i-\mu_j)k_{\theta^{(t)}}(j|y_i)\right]. \tag{A.3}$$

Because the objective function and the feasible set are separable in $(\pi_j)_{j\in[k]}$ and $(\mu_j)_{j\in[k]}$, we split our approach into the following two steps:

A) Global maximization of (A.2) w.r.t. $(\pi_j)_{j=0}^{k-1}$, yielding (1.30);

B) Global maximization of (A.3) w.r.t. $(\mu_j)_{j=0}^{k-1}$, yielding (1.31).

The vector of solutions $\theta^{(t+1)} \in \Omega$ obtained from A) and B) is the global maximizer of the optimization problem in (A.1). The rest of the derivation follows below.

A.1.1.1 A): Derivation of (1.30)

We are dealing with the constrained maximization of

$$\max_{\pi_j \in \mathbb{R}, \text{ for all } j} \mathcal{Q}_n^1((\pi_j)_{j=0}^{k-1})$$

subject to

- $\gamma((\pi_j)_{j=0}^{k-1}) := \sum_{j=0}^{k-1} \pi_j - 1 = 0$,
- $(\pi_j)_{j=0}^{k-1} \in [0,1]^k \iff h_{1j}((\pi_j)_{j=0}^{k-1}) := -\pi_j \leq 0$ and $h_{2j}((\pi_j)_{j=0}^{k-1}) := \pi_j - 1 \leq 0$ for all j.

Because the feasible set $\Omega^1 := \{(\pi_j)_{j\in[k]} \in [0,1]^k : \sum_{j=0}^{k-1} \pi_j = 1\}$ is compact and the objective function given as (A.2) is continuous on Ω^1, we conclude that the objective function in (A.2) takes a maximum in Ω.

With this in mind, we use the Karush-Khun-Tucker Necessary Conditions Theorem written for the reader in Theorem A.2.5 of the appendix (see Bertsekas [5, Proposition 3.3.1 and Proposition 3.1.2] for additional details). In our setting, The Karush-Khun-Tucker Theorem says that if $(\pi_j^{(t+1)})_{j\in[k]}$ is a local maximum of (A.1), there exists $(\pi_j^{(t+1)})_{j=0}^{k-1} \in \Omega$, $\lambda^* \in \mathbb{R}$, $\beta_1^* \in \mathbb{R}^k$, $\beta_2^* \in \mathbb{R}^k$ satisfying:

i) $\nabla_1 \mathcal{L}((\pi_j^{(t+1)})_{j=0}^{k-1}, \lambda^*, \beta_1^*, \beta_2^*) = 0$;

ii) $\gamma((\pi_j^{(t+1)})_{j=0}^{k-1}) = 0$;

iii) $0 \leq (\pi_0^{(t+1)}, ..., \pi_{k-1}^{(t+1)}) \leq 1$;

iv) $\beta_1^*, \beta_2^* \geq 0$;

v) For all j, $(\beta_1^*)_j = 0$ if $\pi_j^{(t+1)} > 0$ and $(\beta_2^*)_j = 0$ if $\pi_j^{(t+1)} < 1$;

vi) $x^T \nabla_{11}^2 \mathcal{L}(\pi_j^{(t+1)}, \lambda^*, \beta_1^*, \beta_2^*) x \leq 0$, for all $x \neq 0$ such that

- $\nabla \gamma((\pi_j^{(t+1)})_{j=0}^{k-1})^T x = 0$;
- $(\frac{d}{d\pi_j^{(t+1)}} - \pi_j^{(t+1)})x = 0$ for all j that are active constraints (i.e. $(\beta_1^*)_j > 0$);
- $(\frac{d}{d\pi_j^{(t+1)}} \pi_j^{(t+1)} - 1)x = 0$for all j that are active constraints (i.e. $(\beta_2^*)_j > 0$);

where the Lagrangian function, $\mathcal{L}((\pi_j)_{j=0}^{k-1}, \lambda, \beta_1, \beta_2)$, is given as

$$\mathcal{Q}_n^1((\pi_j)_{j=0}^{k-1}) + \lambda\gamma((\pi_j)_{j=0}^{k-1}) + \sum_{j=0}^{k-1} (\beta_1)_j h_{1j}((\pi_j)_{j=0}^{k-1}) + \sum_{j=0}^{k-1} (\beta_2)_j h_{2j}((\pi_j)_{j=0}^{k-1}). \tag{A.4}$$

We make the remark that the preliminary conditions of the theorem are satisfied: $\mathcal{Q}_n$, γ, $(h_{1l})_{l\in[k]}$, and $(h_{2l})_{l\in[k]}$ are all twice continuously differentiable w.r.t. π_j for all $j \in [k]$. We proceed, beginning with i):

$$\begin{aligned}
&(\nabla_1 \mathcal{L}((\pi_j)_{j=0}^{k-1}, \lambda, \beta_1, \beta_2))_j = \\
&\quad = \frac{d}{d\pi_j} \left[\mathcal{Q}_n^1((\pi_j)_{j=0}^{k-1}) + \lambda(\sum_{j=0}^{k-1} \pi_j - 1) - (\pi_0, ..., \pi_{k-1})\beta_1 + (\pi_0 - 1, ..., \pi_{k-1} - 1)\beta_2 \right] \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{d}{d\pi_j} \log \left(\frac{\pi_j}{(2\pi)^{\frac{d}{2}} |\Sigma_j^*|^{\frac{1}{2}}} \right) k_{\theta^{(t)}}(j|y_i) \right] + \lambda - (\beta_1)_j + (\beta_2)_j \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{d}{d\pi_j} (\log(\pi_j))\, k_{\theta^{(t)}}(j|y_i) \right] + \lambda - (\beta_1)_j + (\beta_2)_j \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{\pi_j} k_{\theta^{(t)}}(j|y_i) \right] + \lambda - (\beta_1)_j + (\beta_2)_j \\
&\quad = 0.
\end{aligned}$$

Therefore, $\frac{1}{n}\sum_{i=1}^n [k_{\theta^{(t)}}(j|y_i)] = (-\lambda + (\beta_1)_j - (\beta_2)_j)\pi_j$ is satisfied for all $j \in [k]$. Using $\sum_{j=0}^{k-1} k_{\theta^{(t)}}(j|y_i) = \sum_{j=0}^{k-1} \left[\frac{\pi_j^{(t)} \mathcal{G}(y_i; \mu_j^{(t)}, \Sigma_j^*)}{\sum_{l=0}^{k-1} \pi_l^{(t)} \mathcal{G}(y_i; \mu_l^{(t)}, \Sigma_l^*)} \right] = 1$ and summing over all $j \in [k]$ we obtain

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=0}^{k-1} k_{\theta^{(t)}}(j|y_i)\right] &= -\lambda \overbrace{\sum_{j=0}^{k-1}\pi_j}^{\overset{ii)}{=}1} + \sum_{j=0}^{k-1}(\beta_1)_j - (\beta_2)_j)\pi_j \\
&= -\lambda + \sum_{j=0}^{k-1}((\beta_1)_j - (\beta_2)_j)\pi_j \\
&= -\lambda + (\pi_0, ..., \pi_{k-1})(\beta_1 - \beta_2).
\end{aligned}
$$

So far, from i) and ii), we have showed the solution $(\pi_j^{(t+1)})_{j=0}^{k-1}$, λ^*, β_1^*, β_2^* must satisfy

$$
\frac{1}{n}\sum_{i=1}^{n}\left[k_{\theta^{(t)}}(j|y_i)\right] = (-\lambda^* + (\beta_1^*)_j - (\beta_2^*)_j)\pi_j^{(t+1)} \tag{A.5}
$$

where

$$
\lambda^* = (\pi_0^{(t+1)}, ..., \pi_{k-1}^{(t+1)})(\beta_1^* - \beta_2^*) - 1. \tag{A.6}
$$

We now consider individual cases.

<u>Case 1: $\beta_1^* = \beta_2^* = 0$:</u>

Under this case (A.6) simplifies to $\lambda^* = -1$. Plugging everything into (A.5) and solving for $\pi_j^{(t+1)}$, we obtain:

$$
\pi_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\left[k_{\theta^{(t)}}(j|y_i)\right].
$$

We now check the remainder of the optimality conditions:

- Condition iii) is satisfied since $k_{\theta^{(t)}}$ is the pmf of the discrete RV Z and so $0 \leq k_{\theta^{(t)}}(j|y_i) \leq 1$ for all $j, y_i, \theta^{(t)}$;

- Condition iv) is satisfied since $\beta_1^* = 0$ and $\beta_2^* = 0$;

- Condition v) is satisfied since $0 < \frac{1}{n}\sum_{i=1}^{n}\left[k_{\theta^{(t)}}(j|y_i)\right] < 1$ is guaranteed by $\Sigma_j^* \in \mathbb{S}_{++}^d$;

- Lastly, condition vi) is satisfied since the Hessian satisfies $\nabla^2_{11}\mathcal{L}((\pi_j^{(t+1)})_{j=0}^{k-1}, \lambda^*, \beta_1^*, \beta_2^*) \preceq 0$:

$$(\nabla^2_{11}\mathcal{L}((\pi_j^{(t+1)})_{j=0}^{k-1}, \lambda^*, \beta_1^*, \beta_2^*))_{lj} = \begin{cases} \frac{d^2}{d\pi_j^{(t+1)^2}}\mathcal{Q}_n^1((\pi_j^{(t+1)})_{j=0}^{k-1}) = \frac{-1}{n}\sum_{i=1}^n \frac{k_{\theta^{(t)}}(j|y_i)}{\pi_j^{(t+1)^2}} < 0 & ; l = j \\ \frac{d}{d\pi_j^{(t+1)}}\frac{d}{d\pi_l^{(t+1)}}\mathcal{Q}_n^1((\pi_j^{(t+1)})_{j=0}^{k-1}) = 0 & ; l \neq j. \end{cases}$$

Because all the conditions are satisfied, the point $\pi_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^n [k_{\theta^{(t)}}(j|y_i)]$ satisfies the necessary conditions for optimality and could be the local maximizer of (A.1).

Case 2: $\mathcal{H}_1 := \{l \in [k] : (\beta_1^*)_l > 0\} \neq \emptyset$ or $\mathcal{H}_2 := \{l \in [k] : (\beta_2^*)_l > 0\} \neq \emptyset$:
Firstly, it follows from v) that:

x_i: $\mathcal{H}_1 \cap \mathcal{H}_2 = \emptyset$,

x_{ii}: $j \in \mathcal{H}_1 \implies (\beta_1^*)_j > 0 \overset{\text{v)}}{\implies} \pi_j^{(t+1)} = 0 \implies \pi_j^{(t+1)}\mathbb{1}_{\mathcal{H}_1}(j) = 0$,

x_{iii}: $j \in \mathcal{H}_2 \implies (\beta_2^*)_j > 0 \overset{\text{v)}}{\implies} \pi_j^{(t+1)} = 1 \implies \pi_j^{(t+1)}\mathbb{1}_{\mathcal{H}_2}(j) = \mathbb{1}_{\mathcal{H}_2}(j)$.

Using the above, we simplify (A.6) and obtain

$$\lambda^* = \left[\sum_{l=0}^{k-1}\pi_l^{(t+1)}[(\beta_1^*)_l\mathbb{1}_{\mathcal{H}_1}(l) - (\beta_2^*)_l\mathbb{1}_{\mathcal{H}_2}(l)]\right] - 1 = -\left[1 + \sum_{l=0}^{k-1}(\beta_2^*)_l\mathbb{1}_{\mathcal{H}_2}(l)\right].$$

Plugging this result into (A.5) yields

$$\frac{1}{n}\sum_{i=1}^n k_{\theta^{(t)}}(j|y_i) = \pi_j^{(t+1)}\left[1 + \left(\sum_{l=0}^{k-1}(\beta_2^*)_l\mathbb{1}_{\mathcal{H}_2}(l)\right) + (\beta_1^*)_j\mathbb{1}_{\mathcal{H}_1}(j) - (\beta_2^*)_j\mathbb{1}_{\mathcal{H}_2}(j)\right].$$

With a bit more work, we get

$$\frac{1}{n}\sum_{i=1}^n k_{\theta^{(t)}}(j|y_i) = \begin{cases} 0 & ; j \in \mathcal{H}_1 \\ 1 + \left(\sum_{l=0}^{k-1}(\beta_2^*)_l\mathbb{1}_{\mathcal{H}_2}(l)\right) - (\beta_2^*)_j\mathbb{1}_{\mathcal{H}_2}(j) & ; j \in \mathcal{H}_2 \\ \pi_j^{(t+1)} + \pi_j^{(t+1)}\left(\sum_{l=0}^{k-1}(\beta_2^*)_j\mathbb{1}_{\mathcal{H}_2}(j)\right) & ;\ \text{otherwise.} \end{cases} \tag{A.7}$$

In the case where there is some $j \in [k]$ such that $j \in \mathcal{H}_1$, we get from (A.7) that $\frac{1}{n}\sum_{i=1}^{n} k_{\theta^{(t)}}(j|y_i) = 0$. However, this is not possible because $\Sigma_j^* \in \mathbb{S}_{++}^d$. Therefore $\mathcal{H}_1 = \emptyset$. Next, if there is some $j \in [k]$ such that $j \in \mathcal{H}_2$, x_{iii} guarantees that $\pi_j^{(t+1)} = 1$ which, together with ii), implies $\pi_l^{(t+1)} = 0$ for all $l \neq j$. It then follows that for all $l \neq j$, $\frac{1}{n}\sum_{i=1}^{n} k_{\theta^{(t)}}(j|y_i) = 0$; this is not possible because $\Sigma_j^* \in \mathbb{S}_{++}^d$. Therefore, $\mathcal{H}_2 = \emptyset$. We conclude that case 2 cannot occur and the solution $\pi_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}[k_{\theta^{(t)}}(j|y_i)]$ obtained in case 1 is the only possible point satisfying all the necessary conditions for optimality of the Karush-Khun-Tucker Theorem. Because we know (A.2) takes a maximimizer over Ω^1, that maximimizer can only be $\pi_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}[k_{\theta^{(t)}}(j|y_i)]$.

A.1.1.2 B): Derivation of (1.31)

We are dealing with the unconstrained maximization of

$$\max_{(\mu_j)\in\mathbb{R}^d,\text{ for all } j} \mathcal{Q}_n^2((\mu_j)_{j\in[k]}).$$

We make the remark that $\mathcal{Q}_n^2((\mu_j)_{j\in[k]}) \underset{\|\mu_j\|_2\to\infty}{\longrightarrow} -\infty$ for all $j \in [k]$. As a result, the super-level sets of $\mathcal{Q}_n^2$ are bounded. It follows from Theorem A.2.2 of the appendix that since $\mathcal{Q}_n^2$ is additionally continuous, the objective function takes a maximum over $\mathbb{R}$.

To find it, we check the stationary point(s) meaning we solve for all $\mu_j \in \mathbb{R}^d$ such that $(\nabla_1 \mathcal{Q}_n^2((\mu_j)_{j=0}^{k-1}))_j = 0$. We proceed and obtain

$$\begin{aligned}
(\nabla_1 \mathcal{Q}_n^2((\mu_j)_{j=0}^{k-1}))_j &= \frac{d}{d\mu_j}\mathcal{Q}_n^2((\mu_j)_{j=0}^{k-1}) = \\
&= \frac{d}{d\mu_j}\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=0}^{k-1} -\frac{1}{2}(y_i-\mu_j)^T\Sigma_j^{*-1}(y_i-\mu_j)k_{\theta^{(t)}}(j|y_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[-\frac{d}{d\mu_j}\frac{1}{2}(y_i-\mu_j)^T\Sigma_j^{*-1}(y_i-\mu_j)k_{\theta^{(t)}}(j|y_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\Sigma_j^{*-1}(y_i-\mu_j)k_{\theta^{(t)}}(j|y_i)\right] \\
&= 0
\end{aligned}$$

where the last line relies on the fact that Σ_j^* is a symmetric matrix. Solving for μ_j, we get the stationary point is uniquely given as

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^n y_i k_{\theta^{(t)}}(j|y_i)}{\sum_{i=1}^n k_{\theta^{(t)}}(j|y_i)}.$$

Finally, we check the second order optimality conditions. We differentiate a second time and obtain a negative semi-definite Hessian matrix; this indicates the concavity of $\mathcal{Q}_n^2((\mu_j)_{j=0}^{k-1})$ w.r.t. μ_j and guarantees that $\mu_j^{(t+1)}$ is indeed the local maximizer. This follows because

$$(\nabla^2_{11}\mathcal{Q}_n^2((\mu_j)_{j=0}^{k-1}))_{lj} = \begin{cases} \frac{d^2}{d\mu_j^2}\mathcal{Q}_n^2((\mu_j)_{j=0}^{k-1}) = \frac{1}{n}\sum_{i=1}^n -\Sigma_j^{*-1}k_{\theta^{(t)}}(j|y_i) \preceq 0 & ; l = j \\ \frac{d}{d\mu_j}\frac{d}{d\mu_l}\mathcal{Q}_n^2((\mu_j)_{j=0}^{k-1}) = 0 & ; l \neq j \end{cases}$$

and Σ_j^* is positive definite $\iff$ Σ_j^{*-1} is positive definite.

Since $\mu_j^{(t+1)}$ is the unique stationary point of this unconstrained maximization problem and because we know (A.1) takes a maximum over Ω, $\mu_j^{(t+1)}$ is the global maximizer in B).

A.2 Complementary results and definitions from Optimization

This section contains algorithms, results, and definitions from existing optimization theory which are used in various sections of this thesis.

A.2.1 Results and Definitions

Definition A.2.1 (Mahalabonis Distance [25]). *Given a probability distribution $X \in \mathbb{R}^d$, the mahalabonis distance between two points x_1, x_2 w.r.t. X is*

$$\|x_1 - x_2\|_m := \sqrt{(x_1 - x_2)^T S^{-1}(x_1 - x_2)}$$

where S is the covariance matrix of X.

Definition A.2.2 (Bregman Divergence [20])**.** *For $h : \Omega \to \mathbb{R}$ convex, continuously differentiable, and defined on the closed convex set Ω, the Bregman divergence induced by h*

$$D_h(\phi, \theta) := h(\phi) - h(\theta) - \langle \nabla h(\theta), \phi - \theta \rangle$$

is given as the difference between the function and its first Taylor expansion at θ around ϕ.

Theorem A.2.1. *(Jensen's Inequality [31, 2.2 b)]) Let $g(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ be a convex function, and X be a real-valued random variable. Then*

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)].$$

Definition A.2.3 (superlevel sets)**.** *For $f : \Omega \to \bar{\mathbb{R}}$, we denote the superlevel set of f at $f(\phi)$ as*

$$\Omega_\phi(f) := \{\theta \in \Omega : f(\theta) \geq f(\phi)\}$$

where $\phi \in \Omega$.

Theorem A.2.2. *Let $f : \mathbb{R}^s \to \mathbb{R} \cup \{\pm\infty\}$ (continuous).*
f takes a maximum on $\mathbb{R}^s$ if the superlevel sets ${\mathbb{R}^s}_\phi(f)$ are bounded for all $\phi \in \mathbb{R}^s$.

Theorem A.2.3. *Let x^* be a local maximum of $f(\cdot)$ over the convex set Ω. Then for all $x \in \Omega$, $\nabla f(x^*)^T(x - x^*) \leq 0$.*

Proof. **We prove this by contradiction**

Suppose $\exists x^{'} \in \Omega$ with $\nabla f(x^*)^T(x^{'} - x^*) > 0$. Then,
$\nabla f(x^*)^T(x^{'} - x^*) = f^{'}(x^*, x^{'} - x^*) = \lim\limits_{(1-\lambda)\to 0} \dfrac{f(x^* + (1-\lambda)(x - x^*)) - f(x^*)}{1-\lambda} > 0$
Let $\{\lambda_k\}_{k=1}^{\infty}$ with $\lambda_k \in (0, 1)$ be a sequence converging to 1 ($\{\lambda_k\}_{k\geq 0} \to 1$). This means $\{1 - \lambda_k\}_{k=1}^{\infty} \to 0$
Consider the points $z(\lambda_k) = x^* + (1 - \lambda_k)(x - x^*) = (\lambda_k)x^* + (1 - \lambda_k)x^{'}$. $z(\lambda_k) \in \Omega$ by convexity.
Continuing, we know $\dfrac{f(z(\lambda_k)) - f(x^*)}{1 - \lambda_k} > 0$ for all k sufficiently large.
This then means $f(z(\lambda_k)) > f(x^*)$ for all k sufficiently large. Therefore, it follows that

$\{z(\lambda_k)\} \to x^*$ as $k \to \infty$ with $f(z(\lambda_k)) > f(x^*)$ contradicts that x^* is a local maximum of $f(\cdot)$ over Ω.

This completes the proof. □

Definition A.2.4. *(Closed Point-To-Set Map) We say a point-to-set map $\Psi : \Omega \subseteq \mathbb{R}^s \to \mathcal{P}(\mathbb{R}^s)$ is closed or outer-semi continuous at $\theta \in \Omega$ if:*

There exists $\{\theta^{(l)}\}_{l \geq 0} \subseteq \Omega$ such that $\theta^{(l)} \underset{l \to \infty}{\to} \theta$, $\{\phi^{(l)}\}_{l \geq 0}$ where $\phi^{(l)} \in \Psi(\theta^{(l)})$ and $\phi^{(l)} \underset{l \to \infty}{\to} \phi$ $\implies \phi \in \Psi(\theta)$.

Theorem A.2.4 (Global Convergence Theorem (Zangwill, [3]))**.** *Let the sequence $\{\theta^{(t)}\}_{t \geq 0}$ be generated by $\theta^{(t+1)} \in M_n(\theta^{(t)})$ where M_n is a point-to-set map on Ω. Let $\mathcal{T} \subset \Omega$ be a solution set. If the following assertions hold:*

a) There exists $K \subseteq \Omega$, K compact set such that for any $\theta^{(0)} \in \Omega$ and $\{\theta^{(t)}\}_{t \geq 0}$ generated by M_n we have $\theta^{(t)} \in K$, for any t.

b) There exists a continuous map L_n such that for any $\theta^{(t)} \in \Omega$, $\theta^{(t+1)} \in M_n(\theta^{(t)})$, it follows that $L_n(\theta^{(t+1)}) \geq L_n(\theta^{(t)})$, and for any $\theta \in \Omega/\mathcal{T}$, $\theta' \in M_n(\theta)$, it follows that $L_n(\theta') > L_n(\theta)$.

c) We have $M_n(\theta) \neq 0$ for any $\theta \in \Omega$, and M_n is closed on $\Omega/\mathcal{T}$ (see Definition A.2.4).

Then for any $\theta^{(0)} \in \Omega$, every limit point $\bar{\theta}$ of $\{\theta^{(t)}\}_{t \geq 0}$ belongs to $\mathcal{T}$ and $L_n(\theta^{(t)}) \to L_n(\bar{\theta})$.

Definition A.2.5 (Lagrangian Function)**.** *Let $f : \Omega \to \mathbb{R}$ be a continuous function over a convex set. For a minimization problem*

$$\min f(x)$$

subject to

- $h_1(x) = 0, \ldots, h_m(x) = 0$
- $g_1(x) \leq 0, \ldots, g_r(x) \leq 0,$

the Lagrangian function $\mathcal{L}(x, \lambda, \beta) : \mathbb{R}^s \mapsto \mathbb{R}$ is defined as

$$\mathcal{L}(x, \lambda, \beta) := f(x) + \textstyle\sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \beta_j g_j(x).$$

Definition A.2.6 (Regular point). *Let $f : \Omega \to \mathbb{R}$ be a continuous function over a convex set. For a minimization problem*

$$\min f(x)$$

subject to

- $h_1(x) = 0$,..., $h_m(x) = 0$
- $g_1(x) \leq 0$,..., $g_r(x) \leq 0$,

the feasible point x is regular if the gradients of the active constraints at x are linearly independent.

Theorem A.2.5 (Karush-Khun-Tucker Necessary Conditions: Proposition 3.3.1 of [5]). *Let $f : \Omega \to \mathbb{R}$ be a continuous function over a convex set. Let x^* be a regular local minimum (see Definition A.2.6) of the problem*

$$\min f(x)$$

subject to

- $h_1(x) = 0$,..., $h_m(x) = 0$
- $g_1(x) \leq 0$,..., $g_r(x) \leq 0$,

where f, h_i, g_j are continuously differentiable functions from $\mathbb{R}^s$ to $\mathbb{R}$. Then there exists unique Lagrange multiplier vectors $\lambda^* = (\lambda_1^*, ..., \lambda_m^*)$, $\beta^* = (\beta_1^*, ..., \beta_r^*)$, such that

- $\nabla_1 \mathcal{L}(x^*, \lambda^*, \beta^*) = 0$
- $\beta_j^* \geq 0$ *for all* $j = 1, ..., r$,
- $\beta_j^* = 0$ *for all* $j \notin A(x^*)$

where $A(x^*)$ is the set of active constraints at x^* and $\mathcal{L}(x^*, \lambda^*, \beta^*)$ is the Lagrangian defined in A.2.5. For any feasible point x, the set of active constraints at x is defined as $A(x) := \{j : g_j(x) = 0\}$. Further, if f, h, g are twice continuously differentiable, it holds that

$$x^T \nabla^2_{11} \mathcal{L}(x^*, \lambda^*, \mu^*) x \geq 0,$$

for all $x \in \mathbb{R}^s$ such that

- $\nabla h_i(x^*)^T y = 0$ *for all $i = 1, ..., m$ and*
- $\nabla_j(x^*)^T x = 0$ *for all $j \in A(x^*)$.*

A.3 Figures

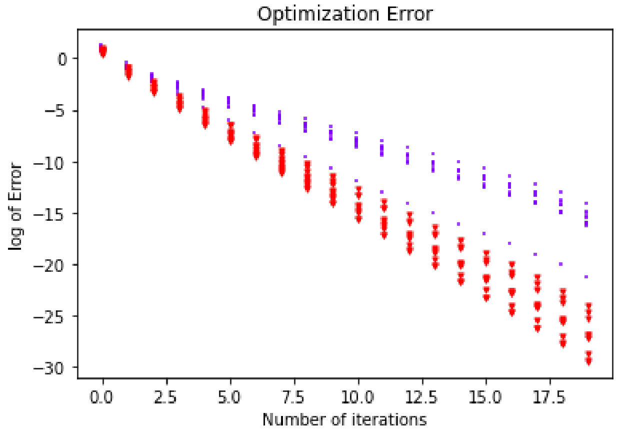

Figure A.1: Optimization Error for SNR of $\frac{1}{2}$

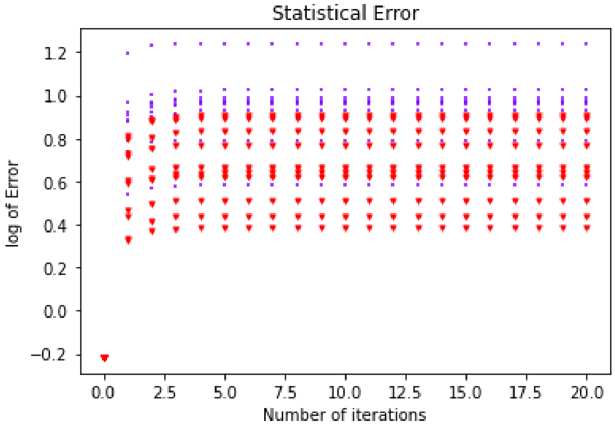

Figure A.2: Statistical Error for SNR of $\frac{1}{2}$

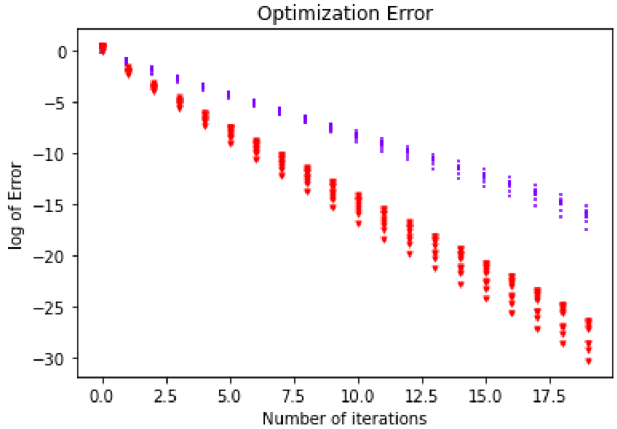

Figure A.3: Optimization Error for SNR of 1

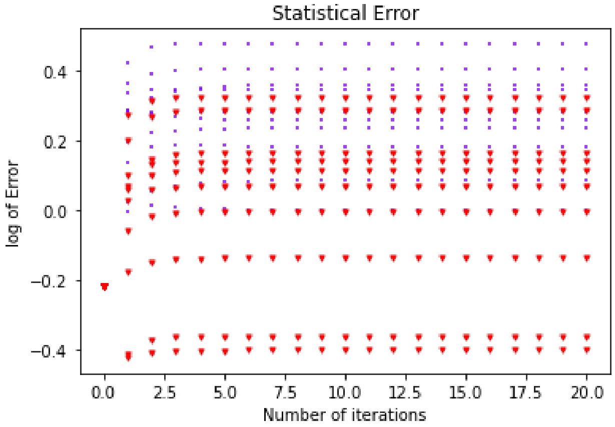

Figure A.4: Statistical Error for SNR of 1

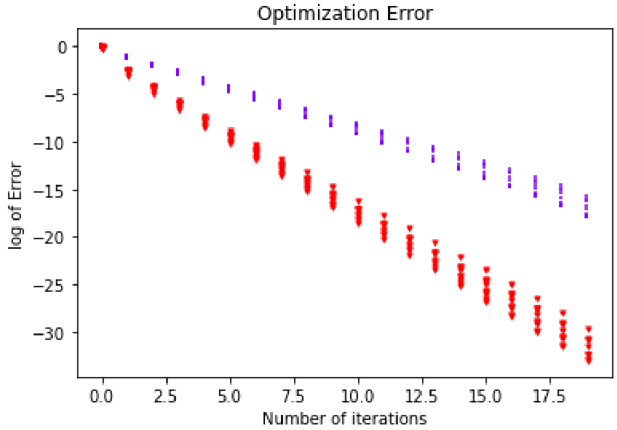

Figure A.5: Optimization Error for SNR of 2

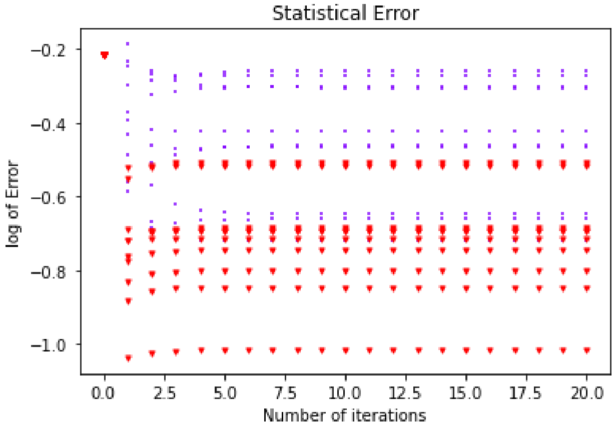

Figure A.6: Statistical Error for SNR of 2

Figure A.7: This figure compiles results for Numerical Experiment 1. The optimization and statistical error is shown for 20 iterations of Algorithm 1. The results for the case where σ_0^* is unknown is shown in purple while the case where σ_0^* is known is shown in red.

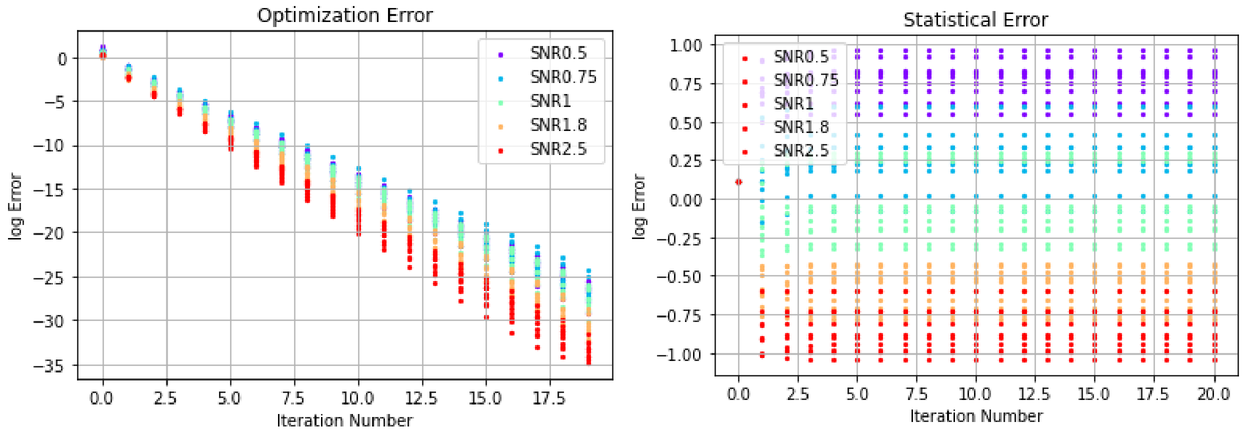

Figure A.8: Optimization Error

Figure A.9: Statistical Error

Figure A.10: This figure compiles results for Numerical Experiment 2. The optimization and statistical error is shown for 20 iterations of Algorithm 1.

Bibliography

[1] BALAKRISHNAN, S., WAINWRIGHT, M. J., AND YU, B. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics 45*, 1 (2017), 77 – 120. 3, 4, 17, 19, 20, 21, 26, 31, 32, 33, 35, 37, 39, 41, 42, 45, 49, 51, 64, 67

[2] BAUM, L. E. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In *Inequalities III: Proceedings of the Third Symposium on Inequalities* (University of California, Los Angeles, 1972), O. Shisha, Ed., Academic Press, pp. 1–8. 2

[3] BEALE, E. M. L. Nonlinear Programming: A Unified Approach. *Journal of the Royal Statistical Society. Series A (General) 133*, 2 (1970), 264–265. 77

[4] BECK, A., AND TEBOULLE, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters 31*, 3 (2003), 167–175. 65

[5] BERTSEKAS, D. *Nonlinear Programming*. Athena Scientific, 1999. 19, 20, 70, 78

[6] CHEN, Y., YI, X., AND CARAMANIS, C. A convex formulation for mixed regression with two components: Minimax optimal rates. *IEEE Transactions on Information Theory 35* (01 2014), 560–604. 54

[7] CONNIFFE, D. Expected maximum log likelihood estimation. *Journal of the Royal Statistical Society. Series D (The Statistician) 36*, 4 (1987), 317–329. 11

[8] DASKALAKIS, C., TZAMOS, C., AND ZAMPETAKIS, M. Ten steps of em suffice for mixtures of two gaussians. In *Proceedings of the 2017 Conference on Learning Theory* (07–10 Jul 2017), S. Kale and O. Shamir, Eds., vol. 65 of *Proceedings of Machine Learning Research*, PMLR, pp. 704–710. 46

[9] DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*, 1 (1977), 1–22. 2, 12, 21, 22, 23, 26, 67

[10] DWIVEDI, R., HO, N., KHAMARU, K., WAINWRIGHT, M., JORDAN, M., AND YU, B. Sharp analysis of expectation-maximization for weakly identifiable models. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (26–28 Aug 2020), S. Chiappa and R. Calandra, Eds., vol. 108 of *Proceedings of Machine Learning Research*, PMLR, pp. 1866–1876. 47

[11] DWIVEDI, R., HO, N., KHAMARU, K., WAINWRIGHT, M., JORDAN, M., AND YU, B. Singularity, misspecification and the convergence rate of em. *Annals of Statistics 48* (12 2020), 3161–3182. 3, 46, 47

[12] DWIVEDI, R., H, N., KHAMARU, K., WAINWRIGHT, M. J., AND JORDAN, M. I. Theoretical guarantees for em under misspecified gaussian mixture models. In *Advances in Neural Information Processing Systems* (2018), S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc. 3

[13] FREMLIN, D. *Measure Theory*. No. v. 1 in Measure theory. Torres Fremlin, 2000. 8

[14] GHOSH, A., AND KANNAN, R. Alternating minimization converges super-linearly for mixed linear regression. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (26–28 Aug 2020), S. Chiappa and R. Calandra, Eds., vol. 108 of *Proceedings of Machine Learning Research*, PMLR, pp. 1093–1103. 54

[15] HARDT, M., AND PRICE, E. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing* (New York, NY, USA, 2015), STOC '15, Association for Computing Machinery, p. 753–760. 50

[16] HARTLEY, H. O. Maximum likelihood estimation from incomplete data. *Biometrics 14*, 2 (1958), 174–194. 2, 13

[17] HARTLEY, H. O., AND HOCKING, R. R. The analysis of incomplete data. *Biometrics 27*, 4 (1971), 783–823. 2

[18] HO, N., KHAMARU, K., DWIVEDI, R., WAINWRIGHT, M. J., JORDAN, M. I., AND YU, B. Instability, computational efficiency and statistical accuracy. *Journal of Machine Learning Research* (2022). 4, 36, 42, 64, 65

[19] HOGG, R. V., MCKEAN, J., AND CRAIG, A. T. Introduction to mathematical statistics (6th edition). 12

[20] KUNSTNER, F., KUMAR, R., AND SCHMIDT, M. Homeomorphic-invariance of em: Non-asymptotic convergence in kl divergence for exponential families via mirror descent. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* (13–15 Apr 2021), A. Banerjee and K. Fukumizu, Eds., vol. 130 of *Proceedings of Machine Learning Research*, PMLR, pp. 3295–3303. 4, 65, 66, 76

[21] KWON, J., AND CARAMANIS, C. The em algorithm gives sample-optimality for learning mixtures of well-separated gaussians. In *Proceedings of Thirty Third Conference on Learning Theory* (09–12 Jul 2020), J. Abernethy and S. Agarwal, Eds., vol. 125 of *Proceedings of Machine Learning Research*, PMLR, pp. 2425–2487. 3, 50, 51, 63

[22] KWON, J., AND CARAMANIS, C. Em converges for a mixture of many linear regressions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (26–28 Aug 2020), S. Chiappa and R. Calandra, Eds., vol. 108 of *Proceedings of Machine Learning Research*, PMLR, pp. 1727–1736. 3, 15, 55, 56

[23] KWON, J., HO, N., AND CARAMANIS, C. On the minimax optimality of the em algorithm for learning two-component mixed linear regression. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* (13–15 Apr 2021), A. Banerjee and K. Fukumizu, Eds., vol. 130 of *Proceedings of Machine Learning Research*, PMLR, pp. 1405–1413. 3, 10, 53, 54

[24] LU, H., FREUND, R. M., AND NESTEROV, Y. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization 28*, 1 (2018), 333–354. 66

[25] MCLACHLAN, G. J. Mahalanobis distance. *Resonance 4* (1999), 20–26. 75

[26] MOITRA, A., AND VALIANT, G. Settling the polynomial learnability of mixtures of gaussians. *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS* (04 2010). 50

[27] ORCHARD, T., AND WOODBURY, M. A. *A MISSING INFORMATION PRINCIPLE: THEORY AND APPLICATIONS*. University of California Press, 1972, pp. 697–716. 2

[28] PACHL, J. K. Disintegration and compact measures. *Mathematica Scandinavica 43*, 1 (1978), 157–168. 8

[29] REGEV, O., AND VIJAYARAGHAVAN, A. On learning mixtures of well-separated gaussians. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)* (2017), pp. 85–96. 50

[30] REN, T., CUI, F., SANGHAVI, S., AND HO, N. Beyond EM algorithm on over-specified two-component location-scale gaussian mixtures. *CoRR abs/2205.11078* (2022). 47

[31] ROCKAFELLAR, R., AND WETS, R. J.-B. *Variational Analysis*. Springer Verlag, Heidelberg, Berlin, New York, 1998. 76

[32] RUDIN, W. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1976. 24

[33] SALAKHUTDINOV, R., ROWEIS, S., AND GHAHRAMANI, Z. Optimization with em and expectation-conjugate-gradient. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning* (2003), ICML'03, AAAI Press, p. 672–679. 19

[34] SALAKHUTDINOV, R., AND ROWEIS, S. T. Relationship between gradient and em steps in latent variable models. Technical Report Unpublished, University of Toronto Department of Computer Science. 19

[35] SUNDBERG, R. Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics 1*, 2 (1974), 49–58. 2

[36] TSENG, P. An analysis of the em algorithm and entropy-like proximal point methods. *Mathematics of Operations Research 29*, 1 (2004), 27–44. 2, 3, 21, 26, 27, 28, 29, 67

[37] WU, C. F. J. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics 11*, 1 (1983), 95 – 103. 2, 21, 24, 26, 27, 28, 67

[38] YAN, B., YIN, M., AND SARKAR, P. Convergence of gradient em on multi-component mixture of gaussians. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2017), NIPS'17, Curran Associates Inc., p. 6959–6969. 19

[39] YI, X., CARAMANIS, C., AND SANGHAVI, S. Alternating minimization for mixed linear regression. In *Proceedings of the 31st International Conference on Machine Learning* (Bejing, China, 22–24 Jun 2014), E. P. Xing and T. Jebara, Eds., vol. 32 of *Proceedings of Machine Learning Research*, PMLR, pp. 613–621. 63

[40] YI, X., CARAMANIS, C., AND SANGHAVI, S. Solving a mixture of many random linear equations by tensor decomposition and alternating minimization, 2016. 63

[41] ZHAO, R., LI, Y., AND SUN, Y. Statistical convergence of the EM algorithm on Gaussian mixture models. *Electronic Journal of Statistics 14*, 1 (2020), 632 – 660. 48, 50, 51