Empirical Analysis of Gradient EM for Truncated Mixture of Gaussians

Abstract

Abstract here...

1 Introduction

Expectation-Maximization (EM) is an iterative algorithm, widely used to compute the maximum likelihood estimation of parameters in statistical models that depend on hidden (latent) variables $\boldsymbol{z}$ ([12]). Given a probability distribution $p_{\boldsymbol{\lambda}}$ defined on $(\boldsymbol{x}; \boldsymbol{z})$, where $\boldsymbol{x}$ are the observables and $\boldsymbol{\lambda}$ is a parameter vector, and samples $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, one wants to find $\boldsymbol{\lambda}$ in order to maximize the log-likelihood $\sum_{i=1}^{n} \log p_{\boldsymbol{\lambda}}(\boldsymbol{x})$ (finite sample case) or $\mathbb{E}_{\boldsymbol{x}}[\log p_{\boldsymbol{\lambda}}(\boldsymbol{x})]$ (population version). Such a task is not always easy, because computing the log-likelihood involves summations over all the possible values of the latent variables and moreover the log-likelihood might be non-concave. EM algorithm is one way to tackle the described problem and works as follows:

- Guess an initialization of the parameters $\boldsymbol{\lambda}_0$.
- For each iteration:
  (Expectation-step) Compute the posterior $Q_i(\boldsymbol{z})$, which is $p_{\boldsymbol{\lambda}_t}(\boldsymbol{z}|\boldsymbol{x}_i)$ for each sample i.
  (Maximization-step) Compute $\boldsymbol{\lambda}_{t+1}$ as the argmax of $\sum_i \sum_{\boldsymbol{z}} Q_i(\boldsymbol{z}) \log \frac{p_{\boldsymbol{\lambda}}(\boldsymbol{z},\boldsymbol{x})}{Q_i(\boldsymbol{z})}$.

It is well known that there are guarantees for convergence of EM to stationary points [26]. The idea behind this fact is that the log-likelihood is decreasing along the trajectories of the EM dynamics. We note that this is true for the case of the truncated mixture of Gaussians too. One of its main applications is on learning mixture of Gaussians. Recovering the parameters of a mixture of Gaussians with strong guarantees was initiated by Dasgupta [8] and has been extensively studied in theoretical computer science and machine learning communities, e.g., [2], [16], [7], [6], where most of the works assume that the means are well-separated. In addition, the authors in [23], [15] offer stronger guarantees, polynomial time (in dimension d) learnability of Gaussian mixtures.

Recent results indicate that EM works well (converges to true mean) for mixture of two Gaussians (see [27], [11] for global convergence and [3] for local convergence), a result that is not true if the number of components is at least three (in [14] an example is constructed where the log-likelihood landscape has local maxima that are not global and EM converges to these points with positive probability).

Parameter estimation problems involving data that has been censored/truncated is crucial in many statistical problems occurring in practice. Statisticians, dating back to Pearson [17] and Fisher [13], tried to address this problem in the early 1900's. Techniques such as method of moments and maximum-likelihood were used for estimating a Gaussian distribution from truncated samples. The seminal work of Rubin [25] in 1976, on missing/censored data, tried to approach this by a framework of ignorable and non-ignorable missingness, where the reason for missingness is incorporated into the statistical model through. However, in

many cases such flexibilities may not be available. More recent work on estimation with truncated data has focused on tractable parametric models such as Gaussians, thereby providing strong computational guarantees for convergence to the true parameters [9]. Mixture models are ubiquitous in machine learning and statistics with a variety of applications ranging from biology [4, 1] to finance [5]. Many of these practical applications are not devoid of some form of truncation or censoring. To this end, there has been previous work that uses EM algorithm for Gaussian mixtures in this setting [18], [21]. However, they assume truncation sets that are generally boxes and in addition do not provide any convergence guarantees.

Our results and techniques Our results can be summarized in the following two theorems (one for single-dimensional and one for multi-dimensional case):

Theorem 1.1 (Single-dimensional case)**.** *Let $S \subset \mathbb{R}$ be an arbitrary measurable set of positive Lebesgue measure, i.e, $\int_S \mathcal{N}(x;\mu,\sigma^2)+\mathcal{N}(x;-\mu,\sigma^2)dx = \alpha > 0$. It holds that under random initialization (under a measure on $\mathbb{R}$ that is absolutely continuous w.r.t Lebesgue), EM algorithm converges with probability one to either μ or $-\mu$. Moreover, if initialization $\lambda_0 > 0$ then EM converges to μ with an exponential rate*

$$|\boldsymbol{\lambda}_{t+1} - \mu| \le \rho_t |\lambda_t - \mu|,$$

with $\rho_t = 1 - \Omega(\alpha^4)\min(\alpha^2 \min(\lambda_t, \mu), 1)$ which is decreasing in t. Analogously if $\lambda_0 < 0$, it converges to $-\mu$ with same rate (substitute $\max(\lambda_t, -\mu)$ in the expression).

Theorem 1.2 (Multi-dimensional case)**.** *Let $S \subset \mathbb{R}^d$ with $d > 1$ be an arbitrary measurable set of positive Lebesgue measure so that $\int_S \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) + \mathcal{N}(\boldsymbol{x};-\boldsymbol{\mu},\boldsymbol{\Sigma})d\boldsymbol{x} = \alpha > 0$. It holds that under random initialization (according to a measure on $\mathbb{R}^d$ that is absolutely continuous with Lebesgue measure), EM algorithm converges with probability one to either $\boldsymbol{\mu}$ or $-\boldsymbol{\mu}$ as long as EM update rule has only $-\boldsymbol{\mu}, \mathbf{0}, \boldsymbol{\mu}$ as fixed points. Moreover, if λ_0 is in a neighborhood of $\boldsymbol{\mu}$ or $-\boldsymbol{\mu}$, it converges with a rate $1 - \Omega(\alpha^6)$*[1].

Remark 1.3. *We would like first to note that we prove the two theorems above in a more general setting where we have truncation functions instead of truncation sets (see Section 2.1 for definitions). Furthermore, in the proof of Theorem 1.2, we show that $\mathbf{0}$ is a repelling fixed point and moreover $-\boldsymbol{\mu}, \boldsymbol{\mu}$ are attracting so if the initialization is close enough to $-\boldsymbol{\mu}$ or $\boldsymbol{\mu}$, then EM actually converges to the true mean. Finally, in Section* **??***, Lemma* **??** *we provide sufficient conditions of the truncated set S (or truncation function) so that the EM update rule has exactly three fixed points. The sufficient condition is that S is rotation invariant under some appropriate transformation.*

To prove the qualitative part of our two main theorems, we perform stability analysis on the fixed points $-\boldsymbol{\mu}, \mathbf{0}, \boldsymbol{\mu}$ of the dynamical system that is induced by EM algorithm and moreover show that the update rule is a diffeomorphism. This is a general approach that has appeared in other papers that talk about first-order methods avoiding saddle points ([22], [20], [24], [19], [10] to name a few).

Nevertheless, computing the update rule of EM for a truncated mixture of two Gaussians is not always possible, because the set/function S is not necessarily symmetric around $\mathbf{0}$ (even for functions). As a result, the techniques of [11] (for the population version) do not carry over to our case. In particular we can find an *implicit* description of the update rule of the EM.

Moreover, getting inspiration from the *Implicit Function Theorem*, we are able to compute explicitly the Jacobian of the update rule of EM and perform spectral analysis on it (Jacobian is computed at the three fixed points $-\boldsymbol{\mu}, \mathbf{0}, \boldsymbol{\mu}$). We show that the spectral radius of the Jacobian computed at $-\boldsymbol{\mu}, \boldsymbol{\mu}$ is less than one (the fixed points are attracting locally) and moreover the spectral radius of the Jacobian computed at $\mathbf{0}$ is greater

[1] If $\alpha\mu \ll 1$ then the global convergence rate we provide in the single-dimensional case coincides with the local convergence rate of multidimensional.

than one (repelling). Along with the fact that the Jacobian is invertible (hence the update rule of EM is a diffeomorphism[2]), we can use the center-stable manifold theorem to show that the region of attraction of fixed point $\mathbf{0}$ is of measure zero. Due to the fact that EM converges always to stationary points (folklore), our result follows. We note that in the case $d = 1$, the fixed points are exactly three $(-\mu, 0, \mu)$ and we prove this fact using FKG (see Theorem **??**) inequality. As far as the case $d > 1$ is concerned, if S is rotation invariant (under proper transformation so that covariance matrix becomes identity), we can show that there are exactly three fixed points by reducing it to the single dimensional case. Last but not least, for the rates of convergence (quantitative part of our theorems), we prove a quantitative version of the FKG inequality (see Lemma **??**) which also might be of independent interest. Due to space constraints, please see supplementary material for the proofs.

2 Background

2.1 Truncated Mixture Model

Before describing the model, we establish the notations used in this paper. We use bold font to represent vectors, any generic element in $\mathbb{R}^d$ is represented by $\boldsymbol{x}$.

The density of a balanced mixture of two different Gaussians with parameters $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ respectively, is given by $f(\boldsymbol{x}) := \frac{1}{2}\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \frac{1}{2}\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, where $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) := \dfrac{\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))}{(2\pi)^{\frac{d}{2}}\det(\Sigma)^{1/2}}$. For this work we consider the case when true covariances are known and they are equal to $\boldsymbol{\Sigma}$. The means are assumed to be symmetric around the origin and we represent the true parameters of the distribution to be $(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Thus, we can write the density as follows:

$$f_{\boldsymbol{\mu}}(\boldsymbol{x}) := \frac{1}{2}\mathcal{N}(\boldsymbol{x}; -\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2}\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2.1}$$

Under this setting we consider a truncation set $S \subset \mathbb{R}^d$, which means that we have access only to the samples that fall in the set S which is of positive measure under the true distribution, i.e.,

$$\int_{\mathbb{R}^d} (0.5\mathcal{N}(\boldsymbol{x}; -\boldsymbol{\mu}, \boldsymbol{\Sigma}) + 0.5\mathcal{N}(\boldsymbol{x}; -\boldsymbol{\mu}, \boldsymbol{\Sigma}))\mathbf{1}_S d\boldsymbol{x} = \alpha > 0,$$

where $\mathbf{1}_S$ is the indicator function for S, i.e., if $\boldsymbol{x} \in S$ then $\mathbf{1}_S(\boldsymbol{x}) = 1$ and is zero otherwise.

Hence we can write the truncated mixture density as follows:

$$f_{\boldsymbol{\mu},S}(\boldsymbol{x}) = \begin{cases} \dfrac{0.5\mathcal{N}(\boldsymbol{x}; -\boldsymbol{\mu}, \boldsymbol{\Sigma}) + 0.5\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\int_S 0.5\mathcal{N}(\boldsymbol{x}; -\boldsymbol{\mu}, \boldsymbol{\Sigma}) + 0.5\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})d\boldsymbol{x}} & , \boldsymbol{x} \in S \\ 0 & , \boldsymbol{x} \notin S \end{cases} \tag{2.2}$$

The above definition can be generalized for "truncation" *functions* too. Let $S : \mathbb{R}^d \to \mathbb{R}$ be a non-negative, bounded by one, measurable function so that $0 < \alpha = \int_{\mathbb{R}^d} S(\boldsymbol{x}) f_{\boldsymbol{\mu}}(\boldsymbol{x}) d\boldsymbol{x}$ (we say nonnegative function S is of "positive measure" if $S(\boldsymbol{x})$ is *not* almost everywhere zero). The truncated mixture then is defined as follows:

$$f_{\boldsymbol{\mu},S}(\boldsymbol{x}) = \frac{(0.5\mathcal{N}(\boldsymbol{x}; -\boldsymbol{\mu}, \boldsymbol{\Sigma}) + 0.5\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}))S(\boldsymbol{x})}{\int_{\mathbb{R}^d}(0.5\mathcal{N}(\boldsymbol{x}; -\boldsymbol{\mu}, \boldsymbol{\Sigma}) + 0.5\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}))S(\boldsymbol{x})d\boldsymbol{x}}$$

One can think of $S(\boldsymbol{x})$ as the probability to actually see sample $\boldsymbol{x}$.

[2] A function is called a diffeomorphism if it is differentiable and a bijection and its inverse is differentiable.

Remark 2.1 (Results proven for truncation functions)**.** *Our main Theorems 1.1 and 1.2 provided in the introduction, hold in the general setting where we have non-negative truncation functions $S(\boldsymbol{x})$ of "positive measure". Our proofs are written in the general setting (not only the case of indicator functions).*

We will use the following short hand for the truncated EM density with means $\boldsymbol{\mu}$ and truncation set or function S such that $f_{\boldsymbol{\mu},S}(\boldsymbol{x}) = \frac{f_{\boldsymbol{\mu}}(\boldsymbol{x})\mathbf{1}_S}{\int_{\mathbb{R}^d} f_{\boldsymbol{\mu}}(\boldsymbol{x})\mathbf{1}_S d\boldsymbol{x}}$ or $f_{\boldsymbol{\mu},S}(\boldsymbol{x}) = \frac{f_{\boldsymbol{\mu}}(\boldsymbol{x})S(\boldsymbol{x})}{\int_{\mathbb{R}^d} f_{\boldsymbol{\mu}}(\boldsymbol{x})S(\boldsymbol{x})d\boldsymbol{x}}$. Also, we will denote the expected value with respect to the truncated mixture distribution with parameters $-\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}$ by $\mathbb{E}_{\boldsymbol{\lambda},S}[.]$. We conclude the subsection with an important definition that will be needed for the multi-dimensional case.

Definition 2.2 (Rotation invariant/Symmetric)**.** *We call a "truncation" function $S(\boldsymbol{x})$ rotation invariant if $S(Q\boldsymbol{x}) = S(\boldsymbol{x})$ for all orthogonal matrices Q. It is clear that every rotation invariant "truncation" function is also even (choose $Q = -\boldsymbol{I}$, where $\boldsymbol{I}$ denotes the identity matrix). A set S is called rotation invariant if $\mathbf{1}_S$ is rotation invariant function and moreover it is called symmetric if $\mathbf{1}_S$ is an even function.*

Next, we derive the EM-update rule to estimate the mean under the "truncated" setting.

2.2 EM Algorithm

The EM algorithm tries to maximize a lower bound of the likelihood at every time step. The population version of the update rule to estimate the mean of a truncated balanced Gaussian mixture with symmetric means $(-\boldsymbol{\mu}, \boldsymbol{\mu})$ and covariance $\boldsymbol{\Sigma}$ with truncation set S boils down to:

$$h(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}) := \mathbb{E}_{\boldsymbol{\mu},S}\left[\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_t)\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\right] - \mathbb{E}_{\boldsymbol{\lambda},S}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right] \tag{2.3}$$

such that

$$\boldsymbol{\lambda}_{t+1} = \{\boldsymbol{\lambda} : h(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}) = \mathbf{0}\}. \tag{2.4}$$

The above population EM update rule for the truncated setting was derived in [**?**]. Although the authors were able to analyze the stability of fixed points of the update rule, it is computationally challenging to compute the update rule at every step (especially in higher dimensions) as it accommodates only an implicit form. Thus we use a "gradient" version of the above rule that is more amenable to analysis and that allows us to easily compute the parameters at every step. We describe the rule below:

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \eta\left(\mathbb{E}_{\boldsymbol{\mu},S}\left[\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_t)\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\right] - \mathbb{E}_{\boldsymbol{\lambda}_t,S}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_t)\right]\right) \tag{2.5}$$

, where $\eta > 0$ is the step size.

Remark 2.3 (Fixed Points)**.** *We first characterize the fixed points of the dynamical system given in equation (2.4). We can identify that there are 3 fixed points, namely, $\boldsymbol{\mu}$,$-\boldsymbol{\mu}$ and $\mathbf{0}$, since*

$$h(\boldsymbol{\mu}, \boldsymbol{\mu}) = \mathbf{0}, \;\; h(-\boldsymbol{\mu}, -\boldsymbol{\mu}) = \mathbf{0} \;\; \textit{and} \;\; h(\mathbf{0}, \mathbf{0}) = \mathbf{0} \tag{2.6}$$

In general there may be more fixed points in the dynamics for any arbitrary truncation function $S(\boldsymbol{x})$ or set S (see Section **??***). However, in the single dimension case we prove that there are only three fixed points (see Lemma* **??***). In multi-dimensional ($d > 1$) case we can also show that if S is rotation invariant, then there are only three fixed points as well (see Lemma* **??***).*

3 Properties of the EM Update Rule

In the section we analyze the dynamical system arising from the EM update rule. To this end, we first describe the derivative $\nabla_{\boldsymbol{\lambda}_t}\boldsymbol{\lambda}_{t+1}$ of the dynamical system, by invoking the *Implicit Function Theorem*. Then we present some derivatives that are essential to characterize the dynamics and argue about the stability of fixed points.

3.1 Properties of the Dynamics

We use the *Implicit Function Theorem* to represent the derivative of $\boldsymbol{\lambda}_{t+1}$ with respect to $\boldsymbol{\lambda}_t$ to analyze the dynamical system around some point say $\boldsymbol{\gamma}$.

$$\nabla_{\boldsymbol{\lambda}_t}\boldsymbol{\lambda}_{t+1}\Big|_{\boldsymbol{\gamma}} = \nabla_{\boldsymbol{\lambda}_{t+1}}\mathbb{E}_{\boldsymbol{\lambda}_{t+1},S}\left[\boldsymbol{x}^T\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_{t+1})\right]^{-1}\Big|_{\boldsymbol{\gamma}} \cdot \nabla_{\boldsymbol{\lambda}_t}\mathbb{E}_{\boldsymbol{\mu},S}\left[\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_t)\boldsymbol{x}^T\right]\Big|_{\boldsymbol{\gamma}} \tag{3.1}$$

The analogue of the above ratio in the single dimension setting is given by:

$$\frac{d\lambda_{t+1}}{d\lambda_t}\Big|_{\gamma} = \frac{\frac{d}{d\lambda_t}\mathbb{E}_{\mu,S}\left[x\tanh\left(\frac{x\lambda_t}{\sigma^2}\right)\right]\Big|_{\lambda_t=\gamma}}{\frac{d}{d\lambda_{t+1}}\mathbb{E}_{\lambda_{t+1},S}\left[x\tanh\left(\frac{x\lambda_{t+1}}{\sigma^2}\right)\right]\Big|_{\lambda_t=\gamma}} \tag{3.2}$$

To this end, we state the following lemma which describes certain derivatives of the terms involved in the above ratio to argue about local stability of the fixed points.

Lemma 3.1 (Some Useful Derivatives)**.** *The following equations hold:*

1. $\nabla_{\boldsymbol{\lambda}}\mathbb{E}_{\boldsymbol{\lambda},S}\left[\boldsymbol{x}^T\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right] = \boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{\lambda},S}\left[\boldsymbol{x}\boldsymbol{x}^T\right] - \boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{\lambda},S}\left[\boldsymbol{x}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right]\mathbb{E}_{\boldsymbol{\lambda},S}\left[\boldsymbol{x}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right]^T$
2. $\nabla_{\boldsymbol{\mu}}\mathbb{E}_{\boldsymbol{\mu},S}\left[\boldsymbol{x}^T\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right] = \boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{\mu},S}\left[\boldsymbol{x}\boldsymbol{x}^T\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right] - \boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{\mu},S}\left[\boldsymbol{x}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right]\mathbb{E}_{\boldsymbol{\mu},S}\left[\boldsymbol{x}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right]^T$
3. $\nabla_{\boldsymbol{\lambda}}\mathbb{E}_{\boldsymbol{\mu},S}\left[\boldsymbol{x}^T\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right] = \boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{\mu},S}\left[\boldsymbol{x}\boldsymbol{x}^T\frac{1}{\cosh^2(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})}\right] = \boldsymbol{\Sigma}^{-1}\mathbb{E}_{\boldsymbol{\mu},S}\left[\boldsymbol{x}\boldsymbol{x}^T\left(1-\tanh^2(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right)\right]$

3.2 Two Important Lemmas

We end the section about the update rule of EM by proving that is well-defined (in the sense that for every λ_t there exists a *unique* $\boldsymbol{\lambda}_{t+1}$) and moreover, we show that the update rule has Jacobian that is invertible for all $\boldsymbol{x} \in \mathbb{R}^d$. The first Lemma that is needed to argue about global convergence (in case there are three fixed points), with the use of center-stable manifold (as the proof appears in [19]) is the following:

Lemma 3.2 (Local Diffeomorphism)**.** *Let J be the Jacobian of the update rule of the EM dynamics (of size $d \times d$). It holds that J is invertible.*

Proof. It suffices to prove that $\nabla_{\boldsymbol{\lambda}}\mathbb{E}_{\boldsymbol{\lambda},S}\left[\boldsymbol{x}^T\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right], \nabla_{\boldsymbol{\lambda}}\mathbb{E}_{\boldsymbol{\mu},S}\left[\boldsymbol{x}^T\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right]$ have non zero eigenvalues (thus invertible) for all $\boldsymbol{\lambda}\in\mathbb{R}^d$ and hence the result follows by Equation (3.1). Observe that

$$M := \mathbb{E}_{\boldsymbol{\lambda},S}[\boldsymbol{x}\boldsymbol{x}^T(1-\tanh^2(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}))] = Cov\left(\boldsymbol{x}\sqrt{1-\tanh^2(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})}, \boldsymbol{x}\sqrt{1-\tanh^2(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})}\right) + \mathbb{E}_{\boldsymbol{\lambda},S}[\boldsymbol{x}\sqrt{1-\tanh^2(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})}]\mathbb{E}_{\boldsymbol{\lambda},S}[\boldsymbol{x}\sqrt{1-\tanh^2(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})}]^T$$

(where $\boldsymbol{x}$ follows a truncated mixture with parameters $\boldsymbol{\lambda},\boldsymbol{\Sigma}$ and truncated function S of "positive measure") which is positive definite (not positive semidefinite) since the function S is of "positive measure" and $-1<\tanh(y)<1$ for all $y\in\mathbb{R}$ (otherwise the variables $\boldsymbol{x}_1,...,\boldsymbol{x}_d$ would live in a lower dimensional subspace). Moreover, from Lemma 3.1 it is clear that

$$\boldsymbol{\Sigma}\nabla_{\boldsymbol{\lambda}}\mathbb{E}_{\boldsymbol{\lambda},S}\left[\boldsymbol{x}^T\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right] - M = Cov\left(\boldsymbol{x}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}), \boldsymbol{x}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right),$$

which is positive definite as well. Hence we conclude that

$$\boldsymbol{\Sigma}\cdot\nabla_{\boldsymbol{\lambda}}\mathbb{E}_{\boldsymbol{\lambda},S}\left[\boldsymbol{x}^T\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right]$$

is positive definite, thus $\nabla_{\boldsymbol{\lambda}}\mathbb{E}_{\boldsymbol{\lambda},S}\left[\boldsymbol{x}^T\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right]$ is invertible. The proof for $\nabla_{\boldsymbol{\lambda}}\mathbb{E}_{\boldsymbol{\mu},S}\left[\boldsymbol{x}^T\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right]$ is simpler since

$$\boldsymbol{\Sigma}\nabla_{\boldsymbol{\lambda}}\mathbb{E}_{\boldsymbol{\mu},S}\left[\boldsymbol{x}^T\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right] = Cov\left(\boldsymbol{x}\sqrt{1-\tanh^2(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})}, \boldsymbol{x}\sqrt{1-\tanh^2(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})}\right) + \mathbb{E}_{\boldsymbol{\mu},S}[\boldsymbol{x}\sqrt{1-\tanh^2(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})}]\mathbb{E}_{\boldsymbol{\mu},S}[\boldsymbol{x}\sqrt{1-\tanh^2(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})}]^T,$$

(where $\boldsymbol{x}$ follows a truncated mixture with parameters $\boldsymbol{\mu},\boldsymbol{\Sigma}$ and truncated function S of "positive measure").
□

The second lemma is about the fact that the update rule of EM is well defined.

Lemma 3.3 (Well defined)**.** *Let $\lambda_t\in\mathbb{R}^d$. There exists a unique $\boldsymbol{\lambda'}$ such that*

$$\mathbb{E}_{\boldsymbol{\mu},S}\left[\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_t)\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\right] = \mathbb{E}_{\boldsymbol{\lambda'},S}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda'})\right].$$

Proof. Let $H(\boldsymbol{w}) = \boldsymbol{\Sigma}\mathbb{E}_{\boldsymbol{w},S}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{w})\right]$. In the Lemma 3.2 we showed that $\nabla_{\boldsymbol{w}}H(\boldsymbol{w})$ is positive definite since S is of positive measure. Assume there exist $\lambda,\tilde{\lambda}$ so that $H(\boldsymbol{\lambda}) = H(\tilde{\boldsymbol{\lambda}})$. Let $\boldsymbol{y}_t = t\boldsymbol{\lambda}+(1-t)\tilde{\boldsymbol{\lambda}}$ for $t\in[0,1]$. Using standard techniques from calculus and that $\nabla_{\boldsymbol{w}}H(\boldsymbol{w})$ is symmetric we get that

$$(\boldsymbol{\lambda}-\tilde{\boldsymbol{\lambda}})^T(H(\boldsymbol{\lambda})-H(\tilde{\boldsymbol{\lambda}})) \geq \min_{t\in[0,1]}\lambda_{\min}(\nabla_{\boldsymbol{w}}H(\boldsymbol{w})\big|_{\boldsymbol{w}=\boldsymbol{y}_t})\left\|\boldsymbol{\lambda}-\tilde{\boldsymbol{\lambda}}\right\|^2, \tag{3.3}$$

where $\lambda_{\min}(A)$ denotes the minimum eigenvalue of matrix A. It is clear that the left hand side is zero, and also the matrix $\nabla_{\boldsymbol{w}}H(\boldsymbol{w})\big|_{\boldsymbol{w}=\boldsymbol{y}_t}$ has all its eigenvalues positive for every $t\in[0,1]$ (using the fact that $\nabla_{\boldsymbol{w}}H(\boldsymbol{w})$ is positive definite for all w from the proof of Lemma 3.2 above). We conclude that $\boldsymbol{\lambda}=\tilde{\boldsymbol{\lambda}}$.
□

Remark 3.4. *In this remark, we would like to argue why there is always a $\boldsymbol{\lambda}_{t+1}$ such that*

$$\mathbb{E}_{\boldsymbol{\mu},S}\left[\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_t)\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\right] = \mathbb{E}_{\boldsymbol{\lambda}_{t+1},S}\left[\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_{t+1})\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\right].$$

The reason is that $\boldsymbol{\lambda}_{t+1}$ is chosen to maximize a particular quantity. If the gradient of that quantity has no roots, it means that $\|\boldsymbol{\lambda}_{t+1}\|_2$ should be infinity. But the quantity is a concave function (in the proof of Lemma 3.2 we showed that $-\nabla_{\boldsymbol{\lambda}}\mathbb{E}_{\boldsymbol{\lambda},S}\left[\boldsymbol{x}^T\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right]\boldsymbol{\Sigma}^{-1}$ is negative definite which is the Hessian of the quantity to be maximized), so the maximum should be attained in the interior (i.e., λ_{t+1} cannot have ℓ_2 norm infinity).

4 Experiments

Experiments section goes here...

References

[1] Michalis Aristophanous, Bill C Penney, Mary K Martel, and Charles A Pelizzari. A gaussian mixture model for definition of lung tumor volumes in positron emission tomography. *Medical physics*, 34(11):4223–4235, 2007.

[2] Sanjeev Arora and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings on 33rd Annual ACM Symposium on Theory of Computing, July 6-8, 2001, Heraklion, Crete, Greece*, pages 247–257, 2001.

[3] Sivaraman Balakrishnan, Martin J Wainwright, Bin Yu, et al. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

[4] Michael J Boedigheimer and John Ferbas. Mixture modeling approach to flow cytometry data. *Cytometry Part A: The Journal of the International Society for Analytical Cytology*, 73(5):421–429, 2008.

[5] Damiano Brigo and Fabio Mercurio. Displaced and mixture diffusions for analytically-tractable smile models. In *Mathematical Finance—Bachelier Congress 2000*, pages 151–174. Springer, 2002.

[6] Kamalika Chaudhuri, Sanjoy Dasgupta, and Andrea Vattani. Learning mixtures of gaussians using the k-means algorithm. *arXiv preprint arXiv:0912.0086*, 2009.

[7] Kamalika Chaudhuri and Satish Rao. Learning mixtures of product distributions using correlations and independence. In *COLT*, volume 4, pages 9–20, 2008.

[8] Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science, FOCS '99, 17-18 October, 1999, New York, NY, USA*, pages 634–644, 1999.

[9] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Efficient statistics, in high dimensions, from truncated samples. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 639–649, 2018.

[10] Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 9256–9266, 2018.

[11] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of EM suffice for mixtures of two gaussians. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 704–710, 2017.

[12] Arthur Dempster, Nan Laird, and Donald Rubin. Maximum likelihood from incomplete data via the em algorithm. In *Journal of the royal statistical society*, pages 1–38, 1977.

[13] RA Fisher. Properties and applications of hh functions. *Mathematical tables*, 1:815–852, 1931.

[14] Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J. Wainwright, and Michael I. Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4116–4124, 2016.

[15] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562. ACM, 2010.

[16] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. In *International Conference on Computational Learning Theory*, pages 444–457. Springer, 2005.

[17] Alice Lee and Karl Pearson. On the Generalised Probable Error in Multiple Normal Correlation. *Biometrika*, 6(1):59–68, 1908.

[18] Gyemin Lee and Clayton Scott. Em algorithms for multivariate gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 56(9):2816–2829, 2012.

[19] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht. First-order methods almost always avoid saddle points. In *To appear in Math. Programming*, 2017.

[20] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

[21] GJ McLachlan and PN Jones. Fitting mixture models to grouped and truncated data via the em algorithm. *Biometrics*, pages 571–578, 1988.

[22] Ruta Mehta, Ioannis Panageas, and Georgios Piliouras. Natural selection as an inhibitor of genetic diversity: Multiplicative weights updates algorithm and a conjecture of haploid genetics. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, ITCS 2015, Rehovot, Israel, January 11-13, 2015*, page 73, 2015.

[23] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

[24] Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. In *Innovations of Theoretical Computer Science (ITCS)*, 2017.

[25] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[26] C.F. Jeff Wu. On the convergence properties of the em algorithm. In *The Annals of statistics*, pages 95–103, 1983.

[27] Ji Xu, Daniel J. Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2676–2684, 2016.