Comparing EM with GD in Mixture Models of Two Components

Guojun Zhang and Pascal Poupart
Computer Science, Waterloo AI Institute, University of Waterloo
Vector Institute
{guojun.zhang, ppoupart}@uwaterloo.ca

George Trimponias
Noah's Ark Lab
Huawei
g.trimponias@huawei.com

Abstract

The expectation-maximization (EM) algorithm has been widely used in minimizing the negative log likelihood (also known as cross entropy) of mixture models. However, little is understood about the goodness of the fixed points it converges to. In this paper, we study the regions where one component is missing in two-component mixture models, which we call *one-cluster* regions. We analyze the propensity of such regions to trap EM and gradient descent (GD) for mixtures of two Gaussians and mixtures of two Bernoullis. In the case of Gaussian mixtures, EM escapes one-cluster regions exponentially fast, while GD escapes them linearly fast. In the case of mixtures of Bernoullis, we find that there exist one-cluster regions that are stable for GD and therefore trap GD, but those regions are unstable for EM, allowing EM to escape. Those regions are local minima that appear universally in experiments and can be arbitrarily bad. This work implies that EM is less likely than GD to converge to certain bad local optima in mixture models.

1 INTRODUCTION

The EM algorithm has a long history dating back to 1886 [1]. Its modern presentation was given in [2] and it has been applied widely on the maximum likelihood problem in mixture models, or equivalently, minimizing the negative log likelihood (also known as cross entropy). Some basic properties of EM have been derived. For example, EM monotonically decreases the loss [2] and it converges to critical points [3].

Mixture models [4] are important in clustering. The most popular one is the Gaussian mixture model (GMM). The optimization problem is non-convex, making the convergence analysis hard. This is true even in some simple settings such as one-dimensional or spherical covariance matrices. With separability assumptions, there has been a series of works proposing new algorithms for learning GMMs efficiently [5–13]. However, for the EM algorithm, not so much work has been done. One well-studied case is the two equally balanced mixtures [14–16] where local and global convergence guarantees are shown. For more than two components, [17] shows that there exist arbitrarily bad local minima. There has also been some study of EM on the convergence of Gaussian mixtures with more than two components [18, 19].

GMMs are well suited in practice for continuous variables. For discrete data clustering, we need discrete mixture models. One common assumption people make is *naive Bayes*, saying that in each component, the features are independent. This corresponds to the spherical covariance matrix setting in GMMs. Since discrete data can be converted to binary strings, in this paper we will focus on Bernoulli mixture models (BMMs), which have been applied in text classification [20] and digit recognition [21]. However, the theory is less studied. Although the Bernoulli and Gaussian distributions are both in the exponential family, we will show that BMMs have different properties from GMMs in the sense that GD is stable (i.e., gets trapped) at so-called one-cluster regions of BMMs, but unstable at (i.e., escapes) all one-cluster regions of GMMs. On the other hand, the two mixture models share similarities that we will explain.

In the practice of clustering with mixture models, it is common for algorithms to converge to regions with only k clusters out of m desired clusters (where $k < m$). We call them *k-cluster regions*. We observed in experiments that EM escapes such regions exponentially faster than GD for GMMs. For BMMs, GD may get stuck at some k-cluster regions and the probability can be very high, while EM always escapes such regions. These experi-

mental results can be found in Section 6.

Theoretically, k-cluster regions are difficult to study. In this paper, we focus on one-cluster regions in mixtures of two components, which can be considered as a first step towards the more difficult problem. We succeed in explaining the different escape rates for GMMs with two components. We find that GD escapes one-cluster regions linearly while EM escapes exponentially fast. For BMMs, our theorem shows that there exist one-cluster regions where GD will converge to, given a small enough step size and a close enough neighborhood, but EM always escapes such regions exponentially. Such one-cluster regions can be arbitrarily worse than the global minimum. Experiments show that this contrast does not only hold for mixtures of two Bernoullis, but also BMMs with any number of components.

The rest of the paper is structured as follows. In Section 2, we give necessary background and notations. Following that, the main contributions are summarized in Section 3. In Section 4, we analyze the stability of EM and GD around one-cluster regions, for GMMs and BMMs with two components. The properties of such one-cluster regions are shown in Section 5. We provide supporting experiments in Section 6 and conclude in Section 7.

2 BACKGROUND AND NOTATIONS

A mixture model of m components has the distribution:

$$p(\mathbf{x}) = \sum_{c=1}^{m} \pi_c f(\mathbf{x}|\boldsymbol{\mu}_c). \tag{2.1}$$

with the *mixing vector* $\boldsymbol{\pi} := (\pi_1, \ldots, \pi_m)$ on the $m-1$ dimensional probability simplex [22]: $\Delta_m := \{\boldsymbol{\pi} : \boldsymbol{\pi} \succeq 0, \mathbf{1}^T\boldsymbol{\pi} = 1\}$. Here $f(\cdot|\boldsymbol{\mu}_c)$ is the conditional distribution given cluster c, and parameters $\boldsymbol{\mu}_c$. The sample space (i.e., space of $\mathbf{x}$) is D-dimensional.

We study the population likelihood, where there are infinitely many samples and the sample distribution is the true distribution. This assumption is common (e.g., in [14]), and it allows us to separate estimation error from optimization error. Since we consider the problem of minimizing the negative log likelihood, the loss function is:

$$\ell = -\mathbb{E}[\log p(\mathbf{x})], \tag{2.2}$$

where the expectation is over $p^*(\mathbf{x})$, the true distribution. We assume by default this notation of expectation in this section. The loss function above is also known as cross entropy loss.

The true distribution $p^*(\mathbf{x})$ is assumed to have the same form as the model distribution: $p^*(\mathbf{x}) = \sum_{c=1}^{m} \pi_c^* f(\mathbf{x}|\boldsymbol{\mu}_c^*)$, where $\pi_c^* \in (0,1)$. Notice that the population assumption is not necessary in this section. We could replace $p^*(\mathbf{x})$ with a finite sample distribution, with i.i.d. samples drawn from $p^*(\mathbf{x})$.

Gaussian mixtures For Gaussian mixtures, we consider the conditional distributions:

$$f(\mathbf{x}|\boldsymbol{\mu}_c) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, I), \tag{2.3}$$

where $\mathbf{x} \in \mathbb{R}^D$, $\boldsymbol{\mu}_c \in \mathbb{R}^D$ and the covariance matrix of each cluster $\boldsymbol{\Sigma}_c = I$.

Bernoulli mixtures For Bernoulli mixtures, we have $\mathbf{x} \in \{0,1\}^D$, $\boldsymbol{\mu}_c \in [0,1]^D$ and:

$$f(\mathbf{x}|\boldsymbol{\mu}_c) = \prod_{i=1}^{D} B(x_i|\mu_{c,i}) = B(\mathbf{x}|\boldsymbol{\mu}_c), \tag{2.4}$$

where the Bernoulli distribution is denoted as $B(x|\mu) = \mu^x(1-\mu)^{1-x}$. Slightly abusing the notation, we also use $B(\mathbf{x}|\boldsymbol{\mu}_c)$ to represent the joint distribution of $x_1, \ldots, x_D$ as a product of the marginal distributions. The expectation of the conditional distribution $B(\mathbf{x}|\boldsymbol{\mu}_c)$ is:

$$\mathbb{E}_{\mathbf{x}\sim B(\mathbf{x}|\boldsymbol{\mu}_c)}[\mathbf{x}] = \boldsymbol{\mu}_c, \tag{2.5}$$

so $\boldsymbol{\mu}_c$ can be interpreted as the mean of cluster c. The covariance matrix of cluster c is $\boldsymbol{\Sigma}_c = \mathrm{diag}(\boldsymbol{\mu}_c * (\mathbf{1} - \boldsymbol{\mu}_c))$, where we use $*$ to denote element-wise multiplication. Also, we always assume $\boldsymbol{\mu}_1^* \in (0,1)^D$, $\boldsymbol{\mu}_2^* \in (0,1)^D$ and thus $\overline{\mathbf{x}} := \mathbb{E}[\mathbf{x}] \in (0,1)^D$.

2.1 EM algorithm

We briefly review the EM algorithm [4] for mixture models. Define the responsibility function[1] $\gamma_c(\mathbf{x})$:

$$\gamma_c(\mathbf{x}) := \frac{f(\mathbf{x}|\boldsymbol{\mu}_c)}{p(\mathbf{x})}. \tag{2.6}$$

The EM update map M is:

$$M(\boldsymbol{\mu}_c) := \frac{\mathbb{E}[\mathbf{x}\gamma_c(\mathbf{x})]}{\mathbb{E}[\gamma_c(\mathbf{x})]},\ M(\pi_c) := \pi_c \mathbb{E}[\gamma_c(\mathbf{x})]. \tag{2.7}$$

Sometimes, it is better to interpret (2.7) in a different way. Define an unnormalized distribution $\tilde{q}_c(\mathbf{x}) = p^*(\mathbf{x})\gamma_c(\mathbf{x})$. The corresponding partition function Z_c and the normalized distribution $q_c(\mathbf{x})$ can be written as:

$$Z_c = \int \tilde{q}_c(\mathbf{x}) d\mathbf{x},\ q_c(\mathbf{x}) = \frac{\tilde{q}_c(\mathbf{x})}{Z_c}, \tag{2.8}$$

[1] This is slightly different from the usual definition by a factor of π_c (see, e.g., [4]).

where the integration is replaced with summation, given discrete mixture models. (2.7) can be rewritten as:

$$M(\boldsymbol{\mu}_c) := \mathbb{E}_{\mathbf{x}\sim q_c(\mathbf{x})}[\mathbf{x}],\ M(\pi_c) := \pi_c Z_c. \tag{2.9}$$

This interpretation will be useful for our analysis near one-cluster regions in Section 4.

2.2 Gradient Descent

GD and its variants are the default algorithms in optimization of deep learning [23]. In mixture models, we have to deal with constrained optimization, and thus we consider projected gradient descent (PGD) [24]. From (2.1), (2.2) and (2.6), the derivative over $\boldsymbol{\pi}$ is:

$$\frac{\partial \ell}{\partial \pi_c} = -\mathbb{E}[\gamma_c(\mathbf{x})] = -Z_c. \tag{2.10}$$

With (2.3) and (2.4), the derivatives over $\boldsymbol{\mu}_c$ are:

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}_c} = -\pi_c \mathbb{E}[\gamma_c(\mathbf{x})\boldsymbol{\Sigma}_c^{-1}(\mathbf{x}-\boldsymbol{\mu}_c)] \tag{2.11}$$

$$= -\pi_c \boldsymbol{\Sigma}_c^{-1} Z_c(\mathbb{E}_{\mathbf{x}\sim q_c(\mathbf{x})}[\mathbf{x}] - \boldsymbol{\mu}_c). \tag{2.12}$$

The equation above is valid for both GMMs and BMMs. However, for BMMs, $\boldsymbol{\Sigma}_c^{-1}$ raises numerical instability near the boundary of $[0,1]^D$. Therefore, we use the following equivalent formula instead in the implementation:

$$\frac{\partial \ell}{\partial \mu_{cj}} = -\pi_c \mathbb{E}\left[\gamma_c(\mathbf{x})\frac{(-1)^{1-x_j}}{B(x_j|\mu_{cj})}\right]. \tag{2.13}$$

After a GD step, we have to project $(\boldsymbol{\pi}, \boldsymbol{\mu}_c)$ back into the feasible space $\Delta_m \times [0,1]^D$. For BMMs, we use Euclidean norm projection. The projection of the $\boldsymbol{\mu}_c$ part is simply:

$$P_{\boldsymbol{\mu}_c}(\boldsymbol{\mu}_c) = \min\{\max\{\boldsymbol{\mu}_c, 0\}, \mathbf{1}\}, \tag{2.14}$$

with min and max applied element-wise. Projection of $\boldsymbol{\mu}_c$ is not needed for GMMs. For the $\boldsymbol{\pi}$ part, we borrow the algorithm from [25], which essentially solves the following optimization problem:

$$P_{\boldsymbol{\pi}}(\boldsymbol{\pi}) = \operatorname{argmin}_{\boldsymbol{\pi}'}||\boldsymbol{\pi} - \boldsymbol{\pi}'||_2 \tag{2.15}$$

$$\text{such that } \boldsymbol{\pi}' \succeq 0, \mathbf{1}^T\boldsymbol{\pi}' = 1. \tag{2.16}$$

The projected gradient descent is therefore:

$$\boldsymbol{\pi} \leftarrow P_{\boldsymbol{\pi}}\left(\boldsymbol{\pi} - \alpha\frac{\partial \ell}{\partial \boldsymbol{\pi}}\right),\ \boldsymbol{\mu}_c \leftarrow P_{\boldsymbol{\mu}_c}\left(\boldsymbol{\mu}_c - \alpha\frac{\partial \ell}{\partial \boldsymbol{\mu}_c}\right), \tag{2.17}$$

with α the step size.

2.3 k-cluster region

We define a k-cluster region and a k-cluster point as:

Definition 2.1. *A k-cluster region is a subset of the parameter space where $||\boldsymbol{\pi}||_0 = k$, with $||\cdot||_0$ denoting the number of non-zero elements. An element in a k-cluster region is called a k-cluster point.*

In this work, we focus on one-cluster regions and mixtures of two components. In Section 6, we will show experiments on k-cluster regions for an arbitrary number of components. A key observation throughout our theoretical analysis and experiments, is that with random initialization, GD often converges to a k-cluster region where $k < m$, whereas EM almost always escapes such k-cluster regions, with random initialization and even with initialization in the neighborhood of k-cluster regions.

Now, let us study one-cluster regions for mixture models of two components. WLOG, assume $\pi_1 = 0$. To study the stability near such regions, we consider $\pi_1 = \epsilon$, with ϵ sufficiently small. From (2.6), the responsibility functions can be approximated as:

$$\gamma_1(\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\mu}_1)}{f(\mathbf{x}|\boldsymbol{\mu}_2)},\ \gamma_2(\mathbf{x}) = 1. \tag{2.18}$$

Under this approximation, $q_2(\mathbf{x}) = p^*(\mathbf{x})$. For EM , $\boldsymbol{\mu}_2$ converges to $\overline{\mathbf{x}}$ within one step based on (2.7). For GD, $\boldsymbol{\mu}_2$ converges to $\overline{\mathbf{x}}$ at a linear rate[2] based on (2.12)[3].

For convenience, we define $\mathbf{b} := \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ the difference between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

2.4 Stable fixed point and stable fixed region

A fixed point p is stable under map M if there exists a small enough feasible neighborhood B of p such that for any $p' \in B$, $\lim_{n\to\infty} M^n(p') = p$, where we use M^n to denote a composition for n times. Otherwise, the fixed point is called unstable. Similarly, a fixed region R is stable under map M if there exists a small enough feasible neighborhood B of R such that for any $p' \in B$, $\lim_{n\to\infty} M^n(p')$ exists and $\lim_{n\to\infty} M^n(p') \in R$. We will use these definitions in Section 4.

3 MAIN CONTRIBUTIONS

We summarize the main contributions in this paper, starting from mixtures of two Gaussians. This is a well-

[2] In the context of optimization, "a sequence $\{x_k\}_{k=1}^{\infty}$ converges to L at a linear rate" means that there exists a number $\mu \in (0,1)$ such that $\lim_{k\to\infty} ||x_{k+1} - L||/||x_k - L|| = \mu$.

[3] For BMMs, assume $\boldsymbol{\mu}_2$ is not initialized on the boundary. The rate is upper bounded by the initialization as $\boldsymbol{\mu}_2$ converges to $\overline{\mathbf{x}} \in (0,1)^D$.

known model that has been widely studied. However, we are not aware of any result regarding k-cluster points. As a first result in this line of research, we show that EM is better than GD in escaping one-cluster regions:

Result 3.1. *Consider a mixture of two Gaussians with $\pi_1^* \in (0,1)$, unit covariance matrices and true distribution*

$$p^*(x) = \pi_1^* \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}^*, I) + \pi_2^* \mathcal{N}(\mathbf{x}| - \boldsymbol{\mu}^*, I) \quad (3.1)$$

When $\pi_1 = \epsilon$ is initialized to be small enough, $\boldsymbol{\mu}_2 = \overline{\mathbf{x}}$ and $\mathbf{b}^T\boldsymbol{\mu}^* \neq 0$, EM increases π_1 exponentially fast, while GD increases π_1 linearly[4].

This result indicates that EM escapes the neighborhood of one cluster-regions faster than GD by increasing the probability π_1 of cluster 1 at a rate that is exponentially faster than GD. The escape rates of EM and GD are formally proven in Theorems 4.1 and 4.2.

Our second result concerns mixtures of two Bernoullis:

Result 3.2. *For mixtures of two Bernoullis, there exist one-cluster regions that can trap GD, but do not trap EM. If any one-cluster region traps EM, it will also trap GD.*

This result is formally stated in Theorems 4.3 and 4.7. It shows that EM is also better than GD in escaping one-cluster regions in BMMs. Our third result concerns the value of one-cluster regions, which is an informal summary of Theorems 5.1 and 5.2:

Result 3.3. *The one-cluster regions stated in Result 3.2 are local minima, and they can be $O(D)$ worse than the optimal value.*

So far, we have seen that EM is better than GD in mixtures of two components, in the sense that EM escapes one-cluster regions exponentially faster than GD in GMMs, and that EM escapes local minima that trap GD in BMMs. Empirically, we show that this is true in general. In Section 6, we find that for BMMs with an arbitrary number of components and features, when we initialize the parameters randomly, EM always converges to an m-cluster point, i.e., all clusters are used in the model. Comparably, GD converges to a k-cluster point with a high probability, where only some of the clusters are employed in fitting the data (i.e., $k < m$). This result, combined with our analysis for mixtures of two components, implies that EM can be more robust than GD in terms of avoiding certain types of bad local optima when learning mixture models.

[4]Here, we use the usual notion of function growth. Do not confuse it with the rate of convergence.

4 ANALYSIS NEAR ONE-CLUSTER REGIONS

In this section, we analyze the stability of EM and GD near one-cluster regions. Our theoretical results for EM and GD are verified empirically in Section 6.

4.1 Mixture of two Gaussians

We first consider mixtures of two Gaussians with identity covariance matrices, under the infinite-sample assumption. Due to translation invariance, we can choose the origin to be the midpoint of the two cluster means, such that: $\boldsymbol{\mu}_1^* = \boldsymbol{\mu}^*$ and $\boldsymbol{\mu}_2^* = -\boldsymbol{\mu}^*$.

4.1.1 EM algortihm

As long as π_1 is sufficiently small, the following theorem shows that EM will increase π_1 and therefore will not converge to a one-cluster solution. Recall from (2.9) that π_1 is multiplied by Z_1 at every step of EM and therefore we show that $Z_1 > 1$ almost everywhere, ensuring that π_1 will increase. Since $\boldsymbol{\mu}_2$ converges to $\overline{\mathbf{x}}$ within one step, we only consider those points where $\boldsymbol{\mu}_2 = \overline{\mathbf{x}}$.

We can show that under mild assumptions, EM escapes one-cluster regions exponentially fast:

Theorem 4.1. *Consider a mixture of two Gaussians with $\pi_1^* \in (0,1)$, unit covariance matrices and true distribution*

$$p^*(x) = \pi_1^* \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}^*, I) + \pi_2^* \mathcal{N}(\mathbf{x}| - \boldsymbol{\mu}^*, I) \quad (4.1)$$

When $\pi_1 = \epsilon$ is initialized to be small enough, $\boldsymbol{\mu}_2 = \overline{\mathbf{x}}$ and $\mathbf{b}^T\boldsymbol{\mu}^* \neq 0$, EM increases π_1 exponentially fast.

Proof. It suffices to show that $Z_1 > 1$ and Z_1 grows if initially $\mathbf{b}^T\boldsymbol{\mu}^* \neq 0$. To calculate Z_1, we first compute $\tilde{q}_1(\mathbf{x}) = p^*(\mathbf{x})\gamma_1(\mathbf{x})$:

$$\begin{aligned} \tilde{q}_1(\mathbf{x}) &= \pi_1^* e^{\mathbf{b}^T(\boldsymbol{\mu}^* - \boldsymbol{\mu}_2)} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}^* + \mathbf{b}, I) \\ &+ \pi_2^* e^{-\mathbf{b}^T(\boldsymbol{\mu}^* + \boldsymbol{\mu}_2)} \mathcal{N}(\mathbf{x}| - \boldsymbol{\mu}^* + \mathbf{b}, I). \end{aligned} \quad (4.2)$$

This equation shows that $\tilde{q}_1(\mathbf{x})$ corresponds to an unnormalized mixture of Gaussians with their means shifted by b and their mixing coefficients rescaled in comparison to $p^*(\mathbf{x})$. If $\mathbf{b} = \mathbf{0}$, then $\tilde{q}_1(\mathbf{x}) = p^*(\mathbf{x})$. So, $\mathbf{b}$ describes how different $\tilde{q}_1(\mathbf{x})$ deviates from $p^*(\mathbf{x})$. The partition function can be computed by integrating out $\tilde{q}_1(\mathbf{x})$:

$$Z_1 = \pi_1^* e^{\mathbf{b}^T(\boldsymbol{\mu}^* - \boldsymbol{\mu}_2)} + \pi_2^* e^{-\mathbf{b}^T(\boldsymbol{\mu}^* + \boldsymbol{\mu}_2)}. \quad (4.3)$$

In fact, $\boldsymbol{\mu}_2 = \overline{\mathbf{x}} = (\pi_1^* - \pi_2^*)\boldsymbol{\mu}^*$ which can be derived from $p^*(\mathbf{x})$. So, Z_1 becomes:

$$Z_1 = \pi_1^* e^{2\pi_2^* \mathbf{b}^T\boldsymbol{\mu}^*} + \pi_2^* e^{-2\pi_1^* \mathbf{b}^T\boldsymbol{\mu}^*}. \quad (4.4)$$

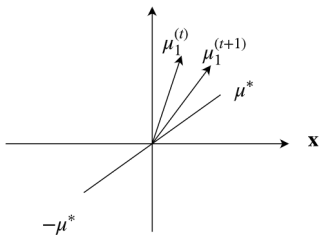

Figure 1: Illustration of an EM update as a rotation of $\boldsymbol{\mu}_1^{(t)}$ to $\boldsymbol{\mu}_1^{(t+1)}$ towards $\boldsymbol{\mu}^*$.

Using the fact $e^x \geqslant 1 + x$ and that equality holds iff $x = 0$, we can show that when $\mathbf{b}^T\boldsymbol{\mu}^* \neq 0$, $Z_1 > 1$.

Now, let us show that Z_1 increases. It suffices to prove that $|\mathbf{b}^T\boldsymbol{\mu}^*|$ increases. From (4.2) and (2.9), we have the update equation for $\boldsymbol{\mu}_1$: $\boldsymbol{\mu}_1 \leftarrow (\pi_1' - \pi_2')\boldsymbol{\mu}^* + \mathbf{b}$, with

$$\pi_1' = \pi_1^* Z_1^{-1} e^{2\pi_2^* \mathbf{b}^T\boldsymbol{\mu}^*}, \pi_2' = \pi_2^* Z_1^{-1} e^{-2\pi_1^* \mathbf{b}^T\boldsymbol{\mu}^*}, \quad (4.5)$$

If $\mathbf{b}^T\boldsymbol{\mu}^* > 0$, then $\pi_1' > \pi_1^*$ and $\pi_2' < \pi_2^*$. So, $\boldsymbol{\mu}_1^{(t+1)} = \boldsymbol{\mu}_1^{(t)} + \delta\boldsymbol{\mu}^*$ with $\delta = \pi_1' - \pi_1^* + \pi_2^* - \pi_2' > 0$, and

$$(\mathbf{b}^{(t+1)})^T\boldsymbol{\mu}^* = (\mathbf{b}^{(t)})^T\boldsymbol{\mu}^* + \delta||\boldsymbol{\mu}^*||_2^2 > (\mathbf{b}^{(t)})^T\boldsymbol{\mu}^*, \quad (4.6)$$

where we use the superscript $(t+1)$ to denote the updated values and (t) to denote the old values.

Similarly, we can prove that $\mathbf{b}^T\boldsymbol{\mu}^*$ will decrease if $\mathbf{b}^T\boldsymbol{\mu}^* < 0$. Hence, from (4.4), Z_1 increases under EM if initially $\mathbf{b}^T\boldsymbol{\mu}^* \neq 0$. □

One may wonder what happens if originally $\mathbf{b}^T\boldsymbol{\mu}^* = 0$. If we choose $\boldsymbol{\mu}_1$ randomly, $\mathbf{b}^T\boldsymbol{\mu}^* = 0$ happens with probability zero since the corresponding Lebesgue measure is zero. Moreover, in numerical calculation such points are extremely unstable and thus unlikely. A similar condition appears in balanced mixtures of Gaussians as well, e.g., Theorem 1 in [15].

Rotation The proof of Theorem 4.1 also shows an interesting phenomenon: $\boldsymbol{\mu}_1$ rotates towards $\boldsymbol{\mu}^*$ or $-\boldsymbol{\mu}^*$ (see Fig. 1). In fact, one can show that if $(\mathbf{b}^{(t)})^T\boldsymbol{\mu}^* > 0$, then:

$$\frac{\langle \boldsymbol{\mu}_1^{(t+1)}, \boldsymbol{\mu}^* \rangle}{||\boldsymbol{\mu}_1^{(t+1)}|| \cdot ||\boldsymbol{\mu}^*||} \geqslant \frac{\langle \boldsymbol{\mu}_1^{(t)}, \boldsymbol{\mu}^* \rangle}{||\boldsymbol{\mu}_1^{t}|| \cdot ||\boldsymbol{\mu}^*||}, \quad (4.7)$$

where the equality holds iff $\boldsymbol{\mu}_1$ is a multiple of $\boldsymbol{\mu}^*$. A similar result can be shown for $(\mathbf{b}^{(t)})^T\boldsymbol{\mu}^* < 0$.

Effect of Separation Another interesting observation is that one can understand how the degree of separation affects the escape rate. If $||\boldsymbol{\mu}^*||$ is small, from (4.4), Z_1 will be close to 1. Also, from (4.5) and (4.6), $\mathbf{b}^T\boldsymbol{\mu}^*$ will change more slowly, adding to the effect of slowing down the escape rate.

4.1.2 GD algorithm

Now, let us analyze the behavior of GD near one-cluster regions. After one GD iteration, the mixing coefficients become $(\pi_1 + \alpha Z_1, \pi_2 + \alpha)$ before projection (where $\pi_2 + \alpha Z_2 = \pi_2 + \alpha$ since $Z_2 = 1$). Assume a small step size α such that $\pi_2 + \alpha > \pi_1 + \alpha Z_1$. After projection based on (2.17), the updated mixing coefficients are:

$$(\pi_1 + (\alpha/2)(Z_1 - 1), \pi_2 - (\alpha/2)(Z_1 - 1)). \quad (4.8)$$

Combining with (2.9), whenever $Z_1 > 1$, both EM and GD increase π_1. In this sense, EM and GD achieve an agreement. However, in GD $\boldsymbol{\mu}_1$ changes little since the change is proportional to π_1. This argument is true for any mixture models with two components.

For GMMs, the update of $\boldsymbol{\mu}_2$ is:

$$\boldsymbol{\mu}_2 \leftarrow \boldsymbol{\mu}_2 + \alpha(\overline{\mathbf{x}} - \boldsymbol{\mu}_2). \quad (4.9)$$

Hence, $\boldsymbol{\mu}_2^{(t)}$ is on the line segment between $\overline{\mathbf{x}}$ and $\boldsymbol{\mu}_2^{(0)}$, and $\boldsymbol{\mu}_2$ converges to $\overline{\mathbf{x}}$ at a linear rate $1 - \alpha$.

If $\boldsymbol{\mu}_2 \neq \overline{\mathbf{x}}$, it is possible to have $Z_1 < 1$. For example, take $D = 1$, $\mu_2 = 3$, $\mu_1 = 5$, $\pi_1^* = \pi_2^* = 1/2$ and $\mu^* = 2$, we have $Z_1 = (1/2)(e^{-10} + e^{-2}) < 1$. In such cases, π_1 could either increase or decrease.

In the worst case, π_1 stays small until $\boldsymbol{\mu}_2$ converges to $\overline{\mathbf{x}}$. Then, from (4.4), $Z_1 > 1$ given $\mathbf{b}^T\boldsymbol{\mu}^* \neq 0$. Assuming $\boldsymbol{\mu}_1$ changes little in this process, we can conclude that with probability one GD escapes one-cluster regions. However, the growth of π_1 is linear compared to EM. Therefore, we have the following theorem:

Theorem 4.2. *For mixtures of two Gaussians and arbitrary initialization $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$, there exists a π_1 small enough such that when the mixture is initialized at $(\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$, GD increases π_1 as a linear function of the number of time steps.*

Proof. For π_1 small enough, $\boldsymbol{\mu}_1$ stays in a small neighborhood, and $\boldsymbol{\mu}_2$ converges to $\overline{\mathbf{x}}$ according to (4.9). Around $\boldsymbol{\mu}_2 = \overline{\mathbf{x}}$, Z_1 stays close to a constant and $Z_1 \geqslant 1$. The linear growth of π_1 follows from (4.8). □

Notice that (4.8) is true for both GMMs and BMMs. The only difference is the computation of Z_1. From this equation, we know that if any one-cluster region traps EM, then $Z_1 < 1$ and $\boldsymbol{\mu}_2 = \overline{\mathbf{x}}$. On this configuration, for the GD algorithm, we know that π_1 decreases if it is not zero, and $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ stays the same. Hence, this region traps GD as well. So, the following theorem holds:

Theorem 4.3. *For two-component mixture models, if any one-cluster region is stable for EM, it is stable for GD.*

Proof. Near any one-cluster region that is stable for EM, $\boldsymbol{\mu}_2 = \overline{\mathbf{x}}$ and from (2.9), $Z_1 < 1$. Also, $\boldsymbol{\mu}_1$ changes very little since the update is proportional to π_1. So, Z_1 is a constant and π_1 decreases linearly according to (4.8). The parameters stop changing at $\pi_1 = 0$. □

4.2 Mixture of two Bernoullis

Now, let us see if Bernoulli mixtures have similar results. With the EM algorithm, $\boldsymbol{\mu}_2 = \overline{\mathbf{x}}$ can be easily obtained. From the definition of $\tilde{q}_c(\mathbf{x})$, we have:

$$\begin{aligned}\tilde{q}_1(\mathbf{x}) &= p^*(\mathbf{x})\frac{B(\mathbf{x}|\boldsymbol{\mu}_1)}{B(\mathbf{x}|\overline{\mathbf{x}})} \\ &= \pi_1^*\prod_{i=1}^{D}\left(\mu_{1i}^*\frac{\mu_{1i}}{\overline{x}_i}\right)^{x_i}\left((1-\mu_{1i}^*)\frac{1-\mu_{1i}}{1-\overline{x}_i}\right)^{1-x_i} \\ &+ \pi_2^*\prod_{i=1}^{D}\left(\mu_{2i}^*\frac{\mu_{1i}}{\overline{x}_i}\right)^{x_i}\left((1-\mu_{2i}^*)\frac{1-\mu_{1i}}{1-\overline{x}_i}\right)^{1-x_i}\end{aligned} \tag{4.10}$$

Consider the unormalized binary distribution $B(x|\mu_1', \mu_2') = (\mu_1')^x(\mu_2')^{1-x}$ with $\mu_1' + \mu_2' \neq 1$. Then the normalization factor (partition function) is $\mu_1' + \mu_2'$. With this fact, one can derive that

$$\begin{aligned}Z_1 &= \pi_1^*\prod_{i=1}^{D}\left(\mu_{1i}^*\frac{\mu_{1i}}{\overline{x}_i} + (1-\mu_{1i}^*)\frac{1-\mu_{1i}}{1-\overline{x}_i}\right) \\ &+ \pi_2^*\prod_{i=1}^{D}\left(\mu_{2i}^*\frac{\mu_{1i}}{\overline{x}_i} + (1-\mu_{2i}^*)\frac{1-\mu_{1i}}{1-\overline{x}_i}\right)\end{aligned} \tag{4.11}$$

From the intuition of Gaussian mixtures, we similarly define $2\boldsymbol{\mu}^* = \boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*$ as the separation between the two true cluster means. As usual, $\mathbf{b}$ describes the separation between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. With the notations above, we obtain:

$$Z_1 = \pi_1^*\prod_{i=1}^{D}(1+\pi_2^*\lambda_i) + \pi_2^*\prod_{i=1}^{D}(1-\pi_1^*\lambda_i), \tag{4.12}$$

with

$$\lambda_i = 2S_i^{-1}\mu_i^* b_i \text{ and } S_i = \mathrm{var}[x_i] = \overline{x}_i(1-\overline{x}_i). \tag{4.13}$$

Approximating $1 + x \sim e^x$ and taking the covariance matrix to be the identity, we retrieve GMM (see (4.4)).

The update of $\boldsymbol{\lambda}$ can be computed from (2.9):

$$M(\boldsymbol{\lambda})_i = \lambda_i + (2S_i^{-1}\mu_i^*)^2\pi_1^*\pi_2^*\frac{\Lambda_i}{Z_1}(B_{1i} - B_{2i}), \tag{4.14}$$

where we denote $\Lambda_i = \mu_{1i}(1-\mu_{1i})$ and

$$B_{1i} := \prod_{j\neq i}^{D}(1+\pi_2^*\lambda_j), B_{2i} := \prod_{j\neq i}^{D}(1-\pi_1^*\lambda_j). \tag{4.15}$$

The derivation of (4.14) is presented in the supplementary material. Notice that each λ_i is an affine transformation of μ_{1i}, as defined by (4.13). Hence, the domain of $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_D)$ is a Cartesian product of closed intervals. We assume by default that $\boldsymbol{\lambda}$ is in the domain. Studying the update of $\boldsymbol{\lambda}$ instead of $\boldsymbol{\mu}_1$ is more convenient, as we will see in the following subsections that $\boldsymbol{\lambda}$ plays an important role in expressing the convergence results.

Also, we denote the covariance between features i and j:

$$\sigma_{ij} = \mathbb{E}[x_i x_j] - \mathbb{E}[x_i]\mathbb{E}[x_j] = 4\pi_1^*\pi_2^*\mu_i^*\mu_j^*. \tag{4.16}$$

We consider the case when there are no independent pairs, i.e., $\sigma_{ij} \neq 0$ for any $i \neq j$.

4.2.1 Attractive one-cluster regions for GD

When $Z_1 < 1$ and $\boldsymbol{\mu}_2 = \overline{\mathbf{x}}$, GD will get stuck in one-cluster regions, as shown in the proof of Theorem 4.3. If $\boldsymbol{\mu}_2 \neq \overline{\mathbf{x}}$, $\boldsymbol{\mu}_2$ converges to $\overline{\mathbf{x}}$ at a linear rate:

$$\boldsymbol{\mu}_2 \leftarrow \boldsymbol{\mu}_2 + \alpha\boldsymbol{\Sigma}_2^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}_2), \tag{4.17}$$

as can be seen from (2.12). Hence, we assume that $\boldsymbol{\mu}_2$ has already converged to $\overline{\mathbf{x}}$. With (4.12), we can define attractive one-cluster regions for GD:

Proposition 4.1 (attractive one-cluster regions for GD). *Denote $T := \{\boldsymbol{\lambda}|Z_1(\boldsymbol{\lambda}) < 1\}$. The one-cluster regions $(\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ with $\boldsymbol{\lambda}(\boldsymbol{\mu}_1) \in T$, $\boldsymbol{\mu}_2 = \overline{\mathbf{x}}$ and $\pi_1 = 0$ are attractive for GD, given small enough step size. Here, we denote $\boldsymbol{\lambda}(\boldsymbol{\mu}_1)$ as an affine function of $\boldsymbol{\mu}_1$, in the form of (4.13).*

For example, when $D = 2$, from (4.12), we have $Z_1 = 1 + \pi_1^*\pi_2^*\lambda_1\lambda_2$. If $\lambda_1\lambda_2 < 0$, then $Z_1 < 1$, and GD will get stuck. If $\lambda_1\lambda_2 > 0$, $Z_1 > 1$, and then GD will escape one-cluster regions. In the latter case, $\boldsymbol{\lambda} \in \mathbb{R}_{++}^2$ or $\boldsymbol{\lambda} \in -\mathbb{R}_{++}^2$.

4.2.2 Positive regions

Now, let us study the behavior of EM. We will show that EM escapes one-cluster regions when $\boldsymbol{\lambda} \in \mathbb{R}_{++}^2 \cup (-\mathbb{R}_{++}^2)$, for the two-feature case. More generally with D features, we should consider $\boldsymbol{\lambda} \in \mathbb{R}_{++}^D \cup (-\mathbb{R}_{++}^D)$.

In the field of optimization, both $\mathbb{R}_{++}^D$ and $-\mathbb{R}_{++}^D$ are known as proper cones and a relevant order can be defined. They are critical in our analysis, so, we define these two cones as positive regions:

Definition 4.1 (positive regions). *A positive region of $\boldsymbol{\lambda}$ is defined to be one of $\{\mathbb{R}^D_{++}, -\mathbb{R}^D_{++}\}$. We denote it as $P := \mathbb{R}^D_{++} \cup (-\mathbb{R}^D_{++})$.*

The positive regions are interesting because $\mathbb{R}^D_{++}$ ensures that each element $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ has the same sign as $\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*$ and $-\mathbb{R}^D_{++}$ ensures that each $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ has the opposite sign of $\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*$.

Theorem 4.4. *For any number of features, $Z_1(\boldsymbol{\lambda}) > 1$ in the positive regions.*

Proof. From (4.12),

$$\nabla Z_1(\boldsymbol{\lambda}) = \pi_1^*\pi_2^* \left(\prod_{i=1}^{D} (1 + \pi_2^* \lambda_i) - \prod_{i=1}^{D} (1 - \pi_1^* \lambda_i) \right).$$

So, $\nabla Z_1(\boldsymbol{\lambda}) \succ 0$ if $\boldsymbol{\lambda} \succ 0$. Denote $g(t) = Z_1(t\boldsymbol{\lambda})$, then $Z_1(\boldsymbol{\lambda}) - 1 = g(1) - g(0) = \int_0^1 \boldsymbol{\lambda}^T \nabla Z_1(t\boldsymbol{\lambda}) dt > 0$, given $\boldsymbol{\lambda} \succ 0$. we can similarly prove $Z_1(\boldsymbol{\lambda}) > 1$ given $\boldsymbol{\lambda} \prec 0$. □

We also notice that positive regions are stable. Define $M(\boldsymbol{\lambda})$ to be the EM update of $\boldsymbol{\lambda}$ based on the EM update of $\boldsymbol{\mu}_1$ and (4.13). For any number of features, the following lemma holds:

Lemma 4.1 (stability of positive regions). *For any number of features, given $\pi_1 = \epsilon$ small, the two positive regions $\mathbb{R}^D_{++}$ and $-\mathbb{R}^D_{++}$ are stable for EM, i.e., $M(\boldsymbol{\lambda}) \succ \boldsymbol{\lambda}$ for all $\boldsymbol{\lambda} \succ 0$ and $M(\boldsymbol{\lambda}) \prec \boldsymbol{\lambda}$ for all $\boldsymbol{\lambda} \prec 0$.*

Proof. If $\boldsymbol{\lambda} \succ 0$, then $B_{1i} > B_{2i}$ for any i (based on (4.14) and (4.15)), and thus $M(\boldsymbol{\lambda}) \succ \boldsymbol{\lambda}$. A similar argument demonstrates that $M(\boldsymbol{\lambda}) \prec \boldsymbol{\lambda}$ when $\boldsymbol{\lambda} \prec 0$. □

From the lemma above, we conclude that in the positive regions, $|\mathbf{b}^T \boldsymbol{\mu}^*|$ will increase, similar to Theorem 4.1. Lemma 4.1 and Theorem 4.4 lead to the following:

Corollary 4.1. *For any number of features, if $\boldsymbol{\lambda}$ is initialized in a positive region, then EM will escape one-cluster regions exponentially fast.*

Proof. WLOG, we assume $\boldsymbol{\lambda} \succ 0$. From Theorem 4.4, at positive regions, $Z_1 > 1$. Define $g(t) = Z_1(\boldsymbol{\lambda} + t(M(\boldsymbol{\lambda}) - \boldsymbol{\lambda}))$. Z_1 increases after an EM update because:

$$\begin{aligned} & Z_1(M(\boldsymbol{\lambda})) - Z(\boldsymbol{\lambda}) = g(1) - g(0) \\ = & \int_0^1 (M(\boldsymbol{\lambda}) - \boldsymbol{\lambda})^T \nabla Z_1(\boldsymbol{\lambda} + t(M(\boldsymbol{\lambda}) - \boldsymbol{\lambda})) dt > 0, \end{aligned}$$

where we use $M(\boldsymbol{\lambda}) \succ \boldsymbol{\lambda}$ from Lemma 4.1 and $\nabla Z_1(\boldsymbol{\lambda}) \succ 0$ if $\boldsymbol{\lambda} \succ 0$ from the proof of Theorem 4.4. From (2.9), π_1 increases exponentially. □

With GD, we have similar results as Corollary 4.1, but GD escapes one-cluster regions much more slowly, which can be derived in a similar way as in Section 4.1.

What if $\boldsymbol{\lambda}$ is not initialized in positive regions? The following theorem tells us that if $D = 2$, no matter where $\boldsymbol{\lambda}$ is initialized, EM will almost always escape the one-cluster regions.

Theorem 4.5. *For $m = D = 2$, given $\sigma_{12} \neq 0$ and $\overline{\mathbf{x}} \in (0,1)^D$, with the EM algorithm, $\pi_1 = \epsilon$, $\boldsymbol{\mu}_2 = \overline{\mathbf{x}}$ and uniform random initialization for $\boldsymbol{\mu}_1$, $\boldsymbol{\lambda}$ will converge to the positive regions at a linear rate with probability 1. Therefore, EM will almost surely escape one-cluster regions.*

Proof. Here, we give a proof sketch, and the more detailed proof of Theorem 4.5 can be found in the supplementary material. First, in the worst case, $||\boldsymbol{\lambda}||$ shrinks to a neighborhood of the origin. Then, near the origin, EM rotates $\boldsymbol{\lambda}$ towards positive regions. □

For general m and D, we conjecture that EM almost always escapes k-cluster regions where $k < m$, as described in the supplementary material.

4.2.3 EM as an ascent method

So far, we have studied positive regions P. We showed some nice properties of P and proved that $\boldsymbol{\lambda}$ converges to P almost everywhere when $D = 2$. In this section, we show that EM can be treated as an ascent method for Z_1:

Theorem 4.6. *For $\boldsymbol{\lambda}$ in the feasible region, we have $\nabla Z_1(\boldsymbol{\lambda})^T (M(\boldsymbol{\lambda}) - \boldsymbol{\lambda}) \geqslant 0$, with equality holds iff $\boldsymbol{\lambda} = 0$.*

Proof. From (4.12), $\nabla_i Z_1(\boldsymbol{\lambda}) = \pi_1^*\pi_2^*(B_{1i} - B_{2i})$. So, from (4.14), for any i, $\nabla_i Z_1(\boldsymbol{\lambda})(M(\boldsymbol{\lambda})_i - \lambda_i)$ is nonnegative. Equality holds iff for all $i \in [D]$, $B_{1i} = B_{2i}$. It follows that for any i, $1 + \pi_1^*\lambda_i = 1 - \pi_2^*\lambda_i$. Hence, equality holds iff $\boldsymbol{\lambda} = 0$. □

With this theorem, we can take a small step in the EM update direction: $\boldsymbol{\lambda} + \alpha(M(\boldsymbol{\lambda}) - \boldsymbol{\lambda})$, with α small. In this way, we can always increase Z_1 until $Z_1 > 1$ and EM escapes the one-cluster regions.

4.3 EM vs GD

Now, we are ready to prove one of the main results in our paper: for mixtures of two Bernoullis, there exist one-cluster regions that can trap GD but not EM. We first prove a lemma similar to Lemma 4.1.

Lemma 4.2. *For all $\boldsymbol{\lambda} \neq 0$ and $\boldsymbol{\lambda} \succeq 0$, $M(\boldsymbol{\lambda}) \succ 0$ and $M(\boldsymbol{\lambda}) \succeq \boldsymbol{\lambda}$. Similarly, for all $\boldsymbol{\lambda} \neq 0$ and $\boldsymbol{\lambda} \preceq 0$, $M(\boldsymbol{\lambda}) \prec 0$ and $M(\boldsymbol{\lambda}) \preceq \boldsymbol{\lambda}$.*

Proof. WLOG, we prove the lemma for $\boldsymbol{\lambda} \succeq 0$. Notice that for any $\boldsymbol{\lambda} \succeq 0$, $M(\boldsymbol{\lambda}) \succeq \boldsymbol{\lambda}$ can be read directly from (4.14). For $\lambda_i = 0$, $B_{1i} > B_{2i}$ since $\boldsymbol{\lambda} \neq 0$. Hence, $M(\boldsymbol{\lambda}) \succ 0$. □

This lemma leads to a main result in our work: for mixtures of two Bernoullis, there exist one cluster regions with nonzero measure that trap GD, but not EM.

Theorem 4.7. *For mixtures of two Bernoullis, given $\pi_1^* \in (0,1)$, $\overline{\mathbf{x}} \in (0,1)^D$ and $||\boldsymbol{\mu}^*||_0 = D$, there exist one-cluster regions B that trap GD, but EM escapes such one-cluster regions exponentially fast.*

Proof. Denote $\mathbf{e}_i$ as the i^{th} unit vector in the standard basis. For any $i \in [D]$, at $\boldsymbol{\lambda} = \lambda_i \mathbf{e}_i$ and $\lambda_i > 0$, from Lemma 4.2 and Theorem 4.4, we have $M(\boldsymbol{\lambda}) \succ 0$ and $Z_1(M(\boldsymbol{\lambda})) > 1$. Notice also that $Z_1(\boldsymbol{\lambda}) = 1$ and $\nabla_j Z_1(\boldsymbol{\lambda}) > 0$ for all $j \neq i$. Therefore, moving in the opposite direction of the gradient, it is possible to find a neighborhood B of $\boldsymbol{\lambda}$, such that for all $\boldsymbol{\lambda}' \in B$, $Z_1(\boldsymbol{\lambda}') < 1$ and $Z_1(M(\boldsymbol{\lambda}')) > 1$. This region traps GD due to Proposition 4.1, and EM escapes such one-cluster regions exponentially fast due to Corollary 4.1. □

5 ONE-CLUSTER LOCAL MINIMA

It is possible to show that some one-cluster regions are local minima, as shown in the following theorem:

Theorem 5.1. *The attractive one-cluster regions for GD defined in Proposition 4.1 are local minima of the cross entropy loss $\ell = -\mathbb{E}[\log(\pi_1 B(\mathbf{x}|\boldsymbol{\mu}_1) + \pi_2 B(\mathbf{x}|\boldsymbol{\mu}_2))]$.*

Proof. Impose $\pi_2 = 1 - \pi_1$ and treat ℓ as a function of $(\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$. Consider any small perturbation $(\delta\pi_1, \delta\boldsymbol{\mu}_1, \delta\boldsymbol{\mu}_2)$. If $\delta\pi_1 = 0$, then the loss will not decrease as $\boldsymbol{\mu}_2 = \overline{\mathbf{x}}$ is a local minimum of $-\mathbb{E}[\log B(\mathbf{x}|\boldsymbol{\mu}_2)]$. If $\delta\pi_1 \neq 0$, the change of ℓ is determined by the first order. Only $\partial\ell/\partial\pi_1$ is nonzero, so the change is dominated by the first order:

$$\frac{\partial \ell}{\partial \pi_1}\delta\pi_1 = (1 - Z_1|_{\boldsymbol{\mu}_1 = \boldsymbol{\mu}_1 + \delta\boldsymbol{\mu}_1})\delta\pi_1 > 0, \tag{5.1}$$

when $(\delta\pi_1, \delta\boldsymbol{\mu}_1, \delta\boldsymbol{\mu}_2)$ is small enough. Therefore, we conclude that $(\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ is a local minimum. □

We call such minima one-cluster local minima. The following theorem shows that they are not global minima.

Theorem 5.2. *Assume $\boldsymbol{\mu}^* \neq 0$ and $\pi_1^* \in (0,1)$. For mixtures of two Bernoullis, one-cluster local minima cannot be global. The gap between the one-cluster local minima and the global minimum could be as large as $\Theta(D)$.*

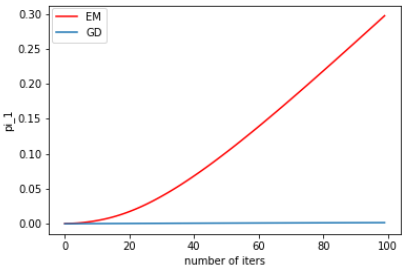

Figure 2: Initialized from $\pi_1 = 10^{-4}$, EM escapes the 1-cluster point exponentially while GD escapes linearly with step size 0.01, achieving 0.0015 in 100 steps.

Proof. The global minimum is obtained when $p(\mathbf{x}) = p^*(\mathbf{x})$. Denote the optimal value as ℓ^*. We have:

$$\ell^* - \ell_1 = -\mathrm{KL}(p^*(\mathbf{x})|\prod_{i=1}^{D} p^*(x_i)) \leqslant 0, \tag{5.2}$$

with $\mathrm{KL}(p||q)$ the Kullback-Leibler divergence. Equality holds iff $p^*(\mathbf{x}) = \prod_{i=1}^D p^*(x_i)$, which means that the features are mutually independent. A direct consequence is that the features are pairwise independent. Nevertheless, we assumed at the end of Section 4.2 that $\sigma_{ij} \neq 0$ for any pair (i,j). This is a contradiction.

Moreover, the difference could be very large. For example, taking $\pi_1^* = 1/2$, $\boldsymbol{\mu}_1^* = \mathbf{1}$ and $\boldsymbol{\mu}_2^* = 0$, the difference is $\ell^* - \ell_1 = -(D-1)\log 2$. □

6 EXPERIMENTS

We analyzed the differences in behavior between EM and GD in Section 4 for mixtures of two components. For Gaussian mixtures with two components, we have done some experiments for the $D = 2$ case with $10,000$ samples to approximate the population case, which can be shown in Figure 2. From this figure, we see that EM increases π_1 exponentially and GD increases π_1 linearly, when we initialize near a one-cluster region.

For Bernoulli mixtures, we aim to understand, at least empirically, how often EM and GD converge to a k-cluster point with $k < m$ with an arbitrary number of components. Since k-cluster points only utilize some of the components, this typically implies that the fixed points are bad. We will show that EM never converges to such bad critical points, whereas GD does.

We experiment with the infinite sample case for mixtures of Bernoullis of m components with D binary variables, where both m and D take values in $\{2,3,4,5,6\}$ for a total of 25 combinations. For each combination

$(m, D) \in \{2, 3, 4, 5, 6\}^2$, We run (i) EM until convergence with at most 20000 iterations; (ii) GD until convergence with a step size of 0.02 and at most 10000 iterations, for 60 times each from random initialization. We find experimentally that EM always converges to an m-cluster point that is reasonably good compared to the optimal value (greater than 0.999 in terms of the likelihood ratio). On the other hand, GD is much more likely to converge to a k-cluster point with $k < m$, especially as m or D increase. Table 1 summarizes our empirical findings for GD.

	$D = 2$	$D = 3$	$D = 4$	$D = 5$	$D = 6$
$m = 2$	0.217	0.133	0.100	0.217	0.183
$m = 3$	0.467	0.150	0.567	0.267	0.267
$m = 4$	0.183	0.267	0.300	0.367	0.483
$m = 5$	0.483	0.567	0.500	0.583	0.633
$m = 6$	0.350	0.467	0.683	0.583	0.533

Table 1: Average fraction of times when GD converges to a k-cluster point with $k < m$ from random initialization. Infinite sample case.

We also experiment with datasets of 2000 samples generated from mixtures of Bernoullis of m components with D binary variables. We choose the same hyperparameters as in the previous experiment for EM and GD. As before, EM always converges to m-cluster points. The results for GD are summarized in Table 2. From this table, we see the similarity between infinite sample and finite sample cases. Hence, it is possible to extend our analysis for the population likelihood to a likelihood with a finite number of samples.

	$D = 2$	$D = 3$	$D = 4$	$D = 5$	$D = 6$
$m = 2$	0.100	0.050	0.000	0.133	0.100
$m = 3$	0.100	0.133	0.216	0.367	0.400
$m = 4$	0.333	0.350	0.100	0.367	0.567
$m = 5$	0.333	0.467	0.250	0.400	0.700
$m = 6$	0.433	0.500	0.583	0.600	0.800

Table 2: Average fraction of times when GD converges to a k-cluster point with $k < m$ from random initialization. Finite sample case.

A remaining question is about the quality of the k-cluster points that GD converges to. Denote the loss of the k-cluster point that GD converges to as ℓ_{GD}, and the optimal loss as ℓ^*. In Table 3, we compute $\exp(\ell^* - \ell_{\text{GD}})$ as the ratio of the likelihood of the k-cluster points that GD converges to, to the optimal likelihood. We randomly choose 10 true distributions, compute the ratio and take the average, as shown in Table 3. We also take the worst-case ratio of the 10 true distributions, summarized in Table 4. From these two tables, we see that the k-cluster points become worse as D increases. This agrees with our intuition for one-cluster points in Theorem 5.2. Also, with more clusters, the ratio is larger since it is easier to fit the data, especially when D is small, and thus a k-cluster point may still behave well in this case.

	$D = 4$	$D = 5$	$D = 6$	$D = 8$	$D = 10$
$m = 2$	0.959	0.892	0.915	0.752	0.720
$m = 3$	0.987	0.979	0.971	0.919	0.834
$m = 4$	0.995	0.981	0.991	0.951	0.884
$m = 5$	0.999	0.998	0.989	0.949	0.903
$m = 6$	1.000	0.997	0.996	0.959	0.926

Table 3: Average ratio of the likelihood of the k-cluster points ($k < m$) that GD converges to, to the optimal likelihood. Infinite sample case.

	$D = 4$	$D = 5$	$D = 6$	$D = 8$	$D = 10$
$m = 2$	0.833	0.743	0.677	0.591	0.527
$m = 3$	0.913	0.913	0.826	0.732	0.606
$m = 4$	0.908	0.831	0.920	0.833	0.715
$m = 5$	0.962	0.944	0.945	0.801	0.823
$m = 6$	0.986	0.984	0.971	0.945	0.703

Table 4: Worst-case ratio of the likelihood of the k-cluster points ($k < m$) that GD converges to, to the optimal likelihood. Infinite sample case.

7 CONCLUSIONS

In this paper, we identified k-cluster regions and studied one-cluster regions in mixture models of two components carefully. For mixtures of two Gaussians, when initialized around one-cluster regions, EM escapes such regions exponentially faster than GD, as also shown in experiments. For mixtures of two Bernoullis, we proved that there exist one-cluster local minima that always trap GD, but EM can escape such minima. For Bernoulli mixtures with a general number of components and any number of features, we showed experimentally that with random initialization, EM always converges to an m-cluster point where all components are used, but GD often converges to a k-cluster point where $k < m$. This means that only a subset of the components are employed.

Our work opens up a new direction of research. It would be interesting to know theoretically if EM almost always escapes k-cluster regions in mixture models. A conjecture for BMMs is given in the supplementary material.

Acknowledgements

Guojun would like to thank Stephen Vavasis and Yaoliang Yu for useful discussions.

References

[1] Simon Newcomb. A generalized theory of the combination of observations so as to obtain the best result. *American journal of Mathematics*, pages 343–366, 1886.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[3] CF Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[5] Nikos Vlassis and Aristidis Likas. A Greedy EM Algorithm for Gaussian Mixture Learning. *Neural Process. Lett.*, 15(1):77–87, February 2002.

[6] Sanjoy Dasgupta. Learning Mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, FOCS '99, pages 634–, Washington, DC, USA, 1999. IEEE Computer Society.

[7] Sanjoy Dasgupta and Leonard J. Schulman. A Two-round Variant of EM for Gaussian Mixtures. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, UAI'00, pages 152–159, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[8] Sanjeev Arora and Ravi Kannan. Learning Mixtures of Arbitrary Gaussians. In *Proceedings of the Thirty-third Annual ACM Symposium on Theory of Computing*, STOC '01, pages 247–257, New York, NY, USA, 2001. ACM.

[9] Sanjoy Dasgupta and Leonard Schulman. A Probabilistic Analysis of EM for Mixtures of Separated, Spherical Gaussians. *Journal of Machine Learning Research*, 8:203–226, May 2007.

[10] Santosh Vempala and Grant Wang. A Spectral Algorithm for Learning Mixture Models. *J. Comput. Syst. Sci.*, 68(4):841–860, June 2004.

[11] Mikhail Belkin and Kaushik Sinha. Polynomial Learning of Distribution Families. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 103–112, Washington, DC, USA, 2010. IEEE Computer Society.

[12] Ankur Moitra and Gregory Valiant. Settling the Polynomial Learnability of Mixtures of Gaussians. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 93–102, Washington, DC, USA, 2010. IEEE Computer Society.

[13] Daniel Hsu and Sham M. Kakade. Learning Mixtures of Spherical Gaussians: Moment Methods and Spectral Decompositions. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, pages 11–20, New York, NY, USA, 2013. ACM.

[14] 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.

[15] Ji Xu, Daniel Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 2684–2692, USA, 2016. Curran Associates Inc.

[16] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten Steps of EM Suffice for Mixtures of Two Gaussians. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 704–710, 2017.

[17] 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 *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4123–4131, USA, 2016. Curran Associates Inc.

[18] Bowei Yan, Mingzhang Yin, and Purnamrita Sarkar. Convergence of Gradient EM on Multi-component Mixture of Gaussians. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6959–6969, USA, 2017. Curran Associates Inc.

[19] Ruofei Zhao, Yuanzhi Li, and Yuekai Sun. Statistical Convergence of the EM Algorithm on Gaussian Mixture Models. *arXiv preprint arXiv:1810.04090*, 2018.

[20] Alfons Juan and Enrique Vidal. On the use of Bernoulli mixture models for text classification. *Pattern Recognition*, 35(12):2705–2710, 2002.

[21] Alfons Juan and Enrique Vidal. Bernoulli mixture models for binary images. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 367–370. IEEE, 2004.

[22] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[23] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[24] Eli M Gafni and Dimitri P Bertsekas. Two-metric projection methods for constrained optimization. *SIAM Journal on Control and Optimization*, 22(6):936–964, 1984.

[25] Shai Shalev-Shwartz and Yoram Singer. Efficient learning of label ranking by soft projections onto polyhedra. *Journal of Machine Learning Research*, 7(Jul):1567–1599, 2006.