Mean Estimation of Truncated Mixtures of Two Gaussians: A Gradient Based Approach

Sai Ganesh Nagarajan[1], **Gerasimos Palaiopanos**[2], **Ioannis Panageas**[3], **Tushar Vaidya**[4], **Samson Yu** [5]

[1] EPFL
[2] University of Pittsburgh
[3] University of California, Irvine
[4] NTU
[5] NUS

Abstract

Even though data is abundant, it is often subjected to some form of censoring or truncation which inherently creates biases. Removing such biases and performing parameter estimation is a classical challenge in Statistics. In this paper, we focus on the problem of estimating the means of a mixture of two balanced d-dimensional Gaussians when the samples are prone to truncation. A recent theoretical study on the performance of the Expectation-Maximization (EM) algorithm for the aforementioned problem showed EM almost surely converges for d=1 and exhibits local convergence for $d > 1$ to the true means. Nevertheless, the EM algorithm for the case of truncated mixture of two Gaussians is not easy to implement as it requires solving a set of *nonlinear* equations at every iteration which makes the algorithm impractical. In this work, we propose a gradient based variant of the EM algorithm that has global convergence guarantees when $d = 1$ and local convergence for $d > 1$ to the true means. Moreover, the update rule at every iteration is easy to compute which makes the proposed method practical. We also provide numerous experiments to obtain more insights into the effect of truncation on the convergence to the true parameters in high dimensions.

Introduction

The performance of algorithms in parameter estimation is crucial for machine learning and its numerous applications. Algorithms such as gradient descent (GD), stochastic gradient descent (SGD), expectation maximization (EM) and their variants are an important part of the modern machine learning toolbox. These algorithms have guarantees, when the data is independent and identically distributed according to the true unknown distribution. However, this is not the case in practice. Data is often subjected to intentional/unintentional censoring or truncation and usually, the modeller has no control over this process. Consequently, an inherent bias could be introduced in the model.

Statisticians, dating back to Pearson (Lee and Pearson 1908) and Fisher (Fisher 1931), tried to address this problem in the early 1900s. Techniques such as method of moments and maximum-likelihood were used for estimating a Gaussian distribution from truncated samples. The seminal work of Rubin (Rubin 1976), on missing/censored data, tried to approach this by studying different models of missing data. Aside from missing data occurring at random, sometimes there might be reasons for missing data and this could be incorporated into the statistical model. However, in many cases such flexibility may not be available.

Gaussian mixtures are ubiquitous in machine learning and statistics with a variety of applications ranging from biology (Boedigheimer and Ferbas 2008; Aristophanous et al. 2007; Brigo and Mercurio 2002; Tasche 2002) to finance, e.g., risk management of financial portfolios. The expected shortfall calculation involves truncated loss distributions. However, the data to compute this risk measure is historical. Hence, inevitably the data is already censored. For example to compute future losses for a portfolio of shares, historical data is used and even with Gaussian returns, we are not sure which distribution the data originates from. In this case, a particular Gaussian distribution with mean μ may reflect a market regime that alternates over time between different means. With truncated data, it will be hard to guess which distribution $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is generating the data, if knowledge of μ is uncertain. In this paper, we focus on the problem of estimating the means of a mixture of two Gaussians that are prone to truncation.

Convergence Guarantees for EM in Gaussian Mixture Models: A standard approach for parameter estimation in Gaussian Mixture models is the EM algorithm. Classically, EM was used to compute the maximum likelihood estimation of parameters in statistical models that depend on hidden (latent) variables. It is well known that there are guarantees for convergence of EM to stationary points (Wu 1983). The idea behind this fact is that the log-likelihood is decreasing along the trajectories of the EM dynamics. Theoretical analysis of EM in mixture of un-truncated Gaussians has been studied extensively, yet the performance of EM is not fully understood. The known theoretical analyses focus on simple cases, i.e., **mean estimation** of a *balanced* mixture of **two Gaussians** with **known covariance**. Recent results indicate that EM works well (converges to true mean) for mixture of two Gaussians (see (Xu, Hsu, and Maleki 2016), (Daskalakis, Tzamos, and Zampetakis 2017) for global convergence and (Balakrishnan et al. 2017) for local convergence), a result that is not true if the number of components is at least three (in (Jin et al. 2016) 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). For a detailed account of the progress made in the theoretical analysis of EM the reader is referred to (Daskalakis, Tzamos, and Zampetakis 2017; Xu, Hsu, and Maleki 2016; Ho et al. 2020). Further convergence analysis can be found in (Ho and Nguyen 2016; Dwivedi et al. 2020a) and (Dwivedi et al. 2020b; Kwon, Ho, and Caramanis 2021).

Learning Under Truncation: Recent work on estimation with truncated data has taken an algorithmic approach and focused on tractable parametric models such as learning the parameters of a *single* multivariate Gaussian with SGD and providing computational guarantees for convergence to the true parameters (Daskalakis et al. 2018). A key part of this involved proving that the population log-likelihood is (globally) *strongly convex*, which is rendered useless in the case of mixture of Gaussians. Other recent works involving truncation are studied by (Daskalakis et al. 2019), where the authors address the problem of truncated regression.

Closer to our work, the results of (Nagarajan and Panageas 2020) analyzed EM for a truncated mixture of two Gaussians for the same settings where global convergence of EM to the true means is known such as (Daskalakis, Tzamos, and Zampetakis 2017; Xu, Hsu, and Maleki 2016) and showed that when the Gaussians are single dimensional (i.e, $d = 1$), EM globally converges to the true means (with local rates of convergence provided). However, when $d > 1$ there are no global convergence guarantees to the true parameters (unless the truncation set or function is rotation invariant under an appropriate transformation). The work done by (Nagarajan and Panageas 2020) analyze the population version of the EM update and although they were able to provide these convergence guarantees, the update rule of EM has an implicit form which makes it impractical for a computer to run the algorithm: it requires a system of non-linear equations to be solved in each step.

In practice, there are some works that try to overcome the problem of truncation in Gaussian mixtures by appropriately modifying the EM algorithm. For instance, in astronomy (Melchior and Goulding 2018), where the data is noisy and incomplete/missing, they treat the data according to Rubin's missing at random (MAR) framework where each sample has a "selection bias" (or a truncation function) that is independent of the density, associated with it. Similarly, in (Lee and Scott 2012) and (McLachlan and Jones 1988) the truncation sets are generally boxes and a correction step is proposed by approximating the moments (as the truncation sets are known to be boxes). Firstly, the above methods do not provide any convergence guarantees and the results are mainly empirical. Secondly, the main justification of their algorithm involves the result of (Wu 1983), that guarantees convergence to the stationary points of the log-likelihood. It is not clear how the landscape is modified due to the presence of truncation and the corrections that are applied. This poses a risk for the algorithm to end up at a saddle point or worse at a spurious local maxima. This is evidenced by (Nagarajan and Panageas 2020), where the authors provide a two-dimensional example where the truncation set is a box and a spurious fixed point of the truncated EM appears. To this end, we identify the main challenges which make this problem elusive to theoretical analysis.

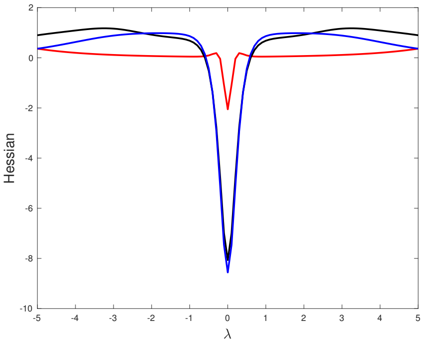


Figure 1: The second-derivative of the negative log-likelihood, when $d = 1$, is computed and shown here for different truncation sets, when the true means are $\mu, -\mu = 3, -3$. As seen here when λ is close to 3 or -3, the second-derivative is positive around the region and when λ is close to 0, it is negative and in both cases the actual bounds depend on the truncation set. This immediately establishes the non-convexity present in the problem and informs us that obtaining uniform rates for arbitrary truncation functions might be challenging.

Technical Challenges for Truncated Gaussian Mixtures: The first challenge is the non-convexity of the problem and the second one being the inability to overcome the implicit update rule of truncated EM without significant computational cost especially when $d > 1$. Work by (Nagarajan and Panageas 2020) showed the existence of truncation sets which are rectangles ($d = 2$) that create spurious fixed points for EM. Additionally, although the single dimension case has no spurious fixed points, the negative log-likelihood function under truncation is still highly non-convex, making it difficult to provide quantitative global convergence guarantees. The Hessian (second derivative since $d = 1$) for such an example is shown in Figure 1.

Secondly, the original truncated population EM which is derived in (Nagarajan and Panageas 2020) is an implicit update equation, such that finding the parameters for the next time step involves solving a set of non-linear equations.

Recall that EM is a heuristic to estimate parameters of statistical model with latent variables. It has an (E-Step) and an (M-Step). The function that we maximize in the M-step is called the Evidence Lower Bound (ELBO), which acts as the lower bound on the likelihood (by Jensen's inequality).

$$ELBO(\boldsymbol{\lambda}) = \sum_i \sum_{\boldsymbol{z}} p_{\boldsymbol{\lambda}_t}(\boldsymbol{z}|\boldsymbol{x}_i) \log \frac{p_{\boldsymbol{\lambda}}(\boldsymbol{z}, \boldsymbol{x})}{p_{\boldsymbol{\lambda}_t}(\boldsymbol{z}|\boldsymbol{x}_i)}. \quad (1)$$

Here $\boldsymbol{x}_i$ are the samples from the observable data, $\boldsymbol{z}$ are the latent variables and $p_{\boldsymbol{\lambda}}$ is the probability model defined on $(\boldsymbol{x}; \boldsymbol{z})$. The implicit updates in (Nagarajan and Panageas 2020) appears during this maximization step due to the presence of truncation.

Moreover, this problem persists even in the finite sample setting due to the non-linearity of the update rule. This scenario is unlike certain cases such as the implicit PCA (Amid and Warmuth 2019) or similar problems where one can solve the implicit equations in finite sample settings.

Thus it becomes clear that we require a method that is easy to implement in *high dimensions* and also can provide some guarantees in the presence of global non-convexity and some local regularities.

We now describe, the truncated mixture model that is analyzed in this paper and follow it by the results obtained.

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 $\boldsymbol{x}$ and any generic parameter estimate of the model is represented by $\boldsymbol{\lambda}$.

Here, we consider the same setting as described in (Nagarajan and Panageas 2020), i.e, the 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 are $(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, such that the mixture model is given by:

$$0.5\mathcal{N}(\boldsymbol{x}; -\boldsymbol{\mu}, \boldsymbol{\Sigma}) + 0.5\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2}$$

We define truncation in a similar fashion, i.e, we call $S \subset \mathbb{R}^d$ the truncation set, which means that we have access only to the samples that fall in the set S. Additionally, we assume that this is of positive measure under the true distribution, i.e.,

Assumption 1.

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

where $S(\boldsymbol{x})$ is the truncation function and special cases include, $\mathbf{1}_S$, which is a truncation set, i.e., if $\boldsymbol{x} \in S$ then $S(\boldsymbol{x}) = 1$ and is zero otherwise.

Assumption 1 is a standard assumption that can be found in prior works on truncation, such as (Daskalakis et al. 2018, 2019) and (Nagarajan and Panageas 2020). The truncation function $S(\boldsymbol{x})$ can be seen as a term that controls selection bias, similar to the one analyzed by (Nagarajan and Panageas 2020).

Our results and techniques: We propose the (Gradient-Truncated EM) algorithm (see Algorithm 1), which performs gradient ascent on the ELBO (Equation (1))(or gradient descent on the negative ELBO, which is an upper bound to the population negative log likelihood) to circumvent the issue of practicality of (Truncated EM) (Nagarajan and Panageas 2020). This is the *truncated* variant of Gradient-EM algorithm (Lange 1995; Yan, Yin, and Sarkar 2017).

We show that when $d = 1$, (Gradient-Truncated EM) algorithm converges to the true means globally and when $d > 1$, we guarantee local convergence to the true means.

When $d = 1$, we show that the dynamics induced by (Gradient-Truncated EM) is "well-behaved" in the sense that it always approaches the true mean and is bounded in some interval $[-B, B]$ that contains μ and when (Gradient-Truncated EM) is initialized in this interval. We then compute the Lipschitz constant of the gradient and use the guarantees of convergence to first order stationary points for gradient descent (Nesterov 1998). Finally, to obtain the convergence to the true means, we show that there is a one-one mapping of the fixed points of (Gradient-Truncated EM) and the Truncated-EM updated rule proposed in (Nagarajan and Panageas 2020).

When $d > 1$, it was shown in (Nagarajan and Panageas 2020) that truncation creates *new* fixed points other than the true means and 0, even when the truncation set is a box in 2-dimensions. This prevents us from obtaining global convergence guarantees in general when $d > 1$ and hence we can only provide guarantees of local convergence around the true means by exploiting local regularities such as local strong convexity and local smoothness. We set function $f(.) := -ELBO(.)$, on which we run gradient descent and the true means $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$ are minimizers of this function f, such that $f(\boldsymbol{\mu}) = f(-\boldsymbol{\mu})$. Specifically, we show the following theorems.

Theorem 2 (Single dimensional (global convergence))**.** *Given, any $\epsilon > 0$, when* (Gradient-Truncated EM) *(with η set to $\alpha^{2(B+1)}$) is initialized to λ_0, it finds a point $\tilde{\mu}$ such that $|\tilde{\mu} - \mu| \leq \epsilon$ when $B \geq \lambda_0 > \epsilon$ in at most*

$$\mathcal{O}\left(\frac{|\lambda_0 - \mu|^2}{\alpha^{4(B+2)}|\lambda_0|^3\epsilon^2}\right), \tag{3}$$

steps, with the assumption that the measure under truncation is $\alpha > 0$ and also $|\mu| \leq B$. Analogously, when $-B \leq \lambda_0 < -\epsilon$, then $|\tilde{\mu} + \mu| \leq \epsilon$ in at most $\mathcal{O}\left(\frac{|\lambda_0+\mu|^2}{\alpha^{4(B+2)}|\lambda_0|^3\epsilon^2}\right)$.

Theorem 3 (Single dimensional (local convergence))**.** *Given, any $\epsilon > 0$, there exists a neighborhood of μ and (equivalently $-\mu$) such that when* (Gradient-Truncated EM) *(with η set to α^2) is initialized to λ_0 in this neighborhood outputs a point $\tilde{\mu}$ such that $|\tilde{\mu} - \mu| \leq \epsilon$ in at most*

$$\mathcal{O}\left(\frac{1}{\alpha^6}\log\left(\frac{|\lambda_0 - \mu|}{\epsilon}\right)\right), \tag{4}$$

steps, with the assumption that the measure under truncation is $\alpha > 0$ and also $|\mu| \leq B$. Analogously, when λ_0 in the neighborhood of of $-\mu$, then $|\tilde{\mu} + \mu| \leq \epsilon$ in at most $\mathcal{O}\left(\frac{1}{\alpha^6}\log\left(\frac{|\lambda_0+\mu|}{\epsilon}\right)\right)$.

Theorem 4 (Multi dimensional (local convergence))**.** *Given, any $\epsilon > 0$, there exists a neighborhood of $\boldsymbol{\mu}$ (equivalently $-\boldsymbol{\mu}$) such that when* (Gradient-Truncated EM) *(with η set to α^2) is initialized to $\boldsymbol{\lambda}_0$ in this neighborhood outputs a point $\tilde{\boldsymbol{\mu}}$ such that $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \leq \epsilon$ in at most*

$$\mathcal{O}\left(\frac{1}{\alpha^6}\log\left(\frac{\|\boldsymbol{\lambda}_0 - \boldsymbol{\mu}\|_2}{\epsilon}\right)\right), \tag{5}$$

steps, with the assumption that the measure under truncation is $\alpha > 0$ and also $\|\boldsymbol{\mu}\|_2 \leq B$. Analogously, when $\boldsymbol{\lambda}_0$ in the neighborhood of of $-\boldsymbol{\mu}$, then $\|\tilde{\boldsymbol{\mu}} + \boldsymbol{\mu}\|_2 \leq \epsilon$ in at most $\mathcal{O}\left(\frac{1}{\alpha^6}\log\left(\frac{\|\boldsymbol{\lambda}_0+\boldsymbol{\mu}\|_2}{\epsilon}\right)\right)$.

Finally, we provide experimental results and some insights into the convergence of (Gradient-Truncated EM) in high dimensional settings. We first focus on the example provided by the authors in (Nagarajan and Panageas 2020), where for an appropriate choice of true means and a truncation set which is a particular rectangle, truncated EM has a spurious fixed point and empirically show that (Gradient-Truncated EM) converges to the true means. In addition, we provide examples with more complicated sets in three dimensions and show that our update rule can converge to the true means.

Background

Truncated EM

The population EM update rule for the truncated setting which was described in (Nagarajan and Panageas 2020) is given below.

$$\text{Let } h(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}) := \mathbb{E}_{\boldsymbol{\mu},S}\left[\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_t)\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\right] - \mathbb{E}_{\boldsymbol{\lambda},S}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right]. \tag{6}$$

The next iterate is

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda} \text{ where } \boldsymbol{\lambda} \text{ is the solution of } h(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}) = \mathbf{0}. \tag{Truncated EM}$$

In the above equation, the expected value with respect to the truncated mixture distribution with parameters $-\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}$ is denoted by $\mathbb{E}_{\boldsymbol{\lambda},S}\left[.\right]$.

Convergence Theorems for Gradient Based Methods

Numerous works on non-convex optimization have analyzed how gradient descent converges to a first order stationary point (FOSP), starting from the works of Nesterov (Nesterov 1998). However, no guarantees were known for second order stationary points (SOSP) as an FOSP could be a local minima or a saddle point. Only recently, it was shown that gradient descent avoids saddle points (Lee et al. 2019). However, they do not quantify the rates of convergence, as the dynamics may get stuck at saddles for an arbitrarily long time. Recent work by (Jin et al. 2017), propose a "perturbed" gradient method that recovers the original rates by Nesterov (Nesterov 1998) up to *polylog* factors in the dimension d. Although, convergence to SOSP is desirable in general, we do not require this additional machinery and we are able to leverage Nesterov's result (Nesterov 1998) on convergence to ϵ-FOSP to guarantee global convergence in our case ($d = 1$).

We state the following definitions for the function $f : \mathbb{R}^d \mapsto \mathbb{R}$, which is assumed to be twice differentiable. In addition, let the optimum value of f be f^*.

Definition 5 (Gradient Lipschitzness or Smoothness)**.** *A twice differentiable function f is L-smooth or L-gradient Lipschitz if it satisfies the following condition:*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x-y\|_2 \;\; \forall x, y. \tag{7}$$

Definition 6 (Approximate first order stationary points)**.** *A point x^* is an ϵ-first order stationary point (or critical point) of f if $\|\nabla f(x^*)\|_2 \leq \epsilon$.*

Theorem 7 ((Nesterov 1998))**.** *Assume that the function $f(.)$ is L-smooth. Then for any $\epsilon > 0$, if we run gradient descent with step size $\eta = \frac{1}{L}$ and termination condition $\|\nabla f(x)\|_2 \leq \epsilon$, the output will be a ϵ- first order stationary point and the algorithm will terminate in the following number of iterations:*

$$\frac{L\left(f(x_0) - f^*\right)}{\epsilon^2}. \tag{8}$$

Definition 8 (Local smoothness and local strong convexity)**.**

If a function f is locally ν-strongly convex and β-smooth in $\mathcal{X} \subset \mathbb{R}^d$, then for all $x, y \in \mathcal{X}$, we have :

$$f(y) \geq f(x) + \langle y - x, \nabla f(x)\rangle + \frac{\nu}{2}\left\|y - x\right\|_2^2, \tag{9}$$

$$\left\|\nabla f(y) - \nabla f(x)\right\|_2 \leq \beta\left\|y - x\right\|_2. \tag{10}$$

Given the local regularities of the function (locally strongly convex and locally smooth), we can state the following theorem which applies to general convex functions.

Theorem 9 ((Bubeck 2014))**.** *Assume that the function $f(.)$ is L-smooth and ν-strongly convex. For any $\epsilon > 0$, if we run gradient descent with step size $\eta = \frac{1}{L}$, then iterate x_t will be ϵ close to x^* (the global minimizer of f) in iterations:*

$$\frac{2L}{\nu}\log\left(\frac{\|x_0 - x^*\|_2}{\epsilon}\right). \tag{11}$$

Gradient-Truncated EM

As mentioned in the previous section equation (6) describes the truncated EM update rule in the population setting derived by (Nagarajan and Panageas 2020). Although they were able to analyze the stability of fixed points of the aforementioned update rule, it is impractical to compute the update rule at every step (especially in higher dimensions) as it accommodates only an implicit form and one has to solve a set of *nonlinear equations*.

Thus we propose a "gradient" version of the above rule which performs gradient ascent on $ELBO(\boldsymbol{\lambda})$, that is more amenable to analysis and that allows us to easily compute the parameters at every step.

For convenience, we are going to look at $-ELBO(\boldsymbol{\lambda})$ and use (Gradient-Truncated EM) as a gradient descent algorithm for a non-convex minimization problem. Thus we can rewrite it as follows:

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

The gradient of the $-ELBO(\boldsymbol{\lambda})$ is given by:

$$g(\boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{\lambda},S}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right] - \mathbb{E}_{\boldsymbol{\mu},S}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right]. \tag{12}$$

As a short hand, we will set $f(\boldsymbol{\lambda}) := -ELBO(\boldsymbol{\lambda})$. In other words $\nabla f(\boldsymbol{\lambda}) = g(\boldsymbol{\lambda})$.

Finally, we state the following assumption

Assumption 10. *The true mean has bounded norm, i.e,* $\|\boldsymbol{\mu}\|_2 \leq B$.

```
Algorithm 1: Gradient-Truncated EM
Output mean estimate $\tilde{\boldsymbol{\mu}}$
Initialize $\boldsymbol{\lambda}_0$, choose $\epsilon > 0$ and set $\eta = \alpha^{2(B+1)}$ (see Theorems 2, 3, 4 for specific choices of $\eta$)
  While $\|\nabla f(\boldsymbol{\lambda_t})\|_2 > \epsilon$
      $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \eta \left( \mathbb{E}_{\boldsymbol{\lambda}_t, S} \left[ \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \tanh(\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\lambda}_t) \right] - \mathbb{E}_{\boldsymbol{\mu}, S} \left[ \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \tanh(\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\lambda}_t) \right] \right)$
```

We end this section by characterizing an equivalence between the Gradient Truncated EM and the Truncated EM framework of (Nagarajan and Panageas 2020). We show that there is a one-one mapping of the fixed points.

Lemma 11. *The fixed points of* (Gradient-Truncated EM) *and* (Truncated EM) *have a one-one mapping.*

From lemma 11, we can conclude that when $d = 1$, (Gradient-Truncated EM) has three fixed points, which are μ, $-\mu$ and 0.

Convergence of Gradient-Truncated EM

As seen previously, the problem is highly non-convex, especially for arbitrary truncation sets. In addition, the gradient based rule proposed in this work does not correspond to the EM trajectories as in (Nagarajan and Panageas 2020), which means we cannot take for granted the convergence to stationary points of Truncated EM from (Nagarajan and Panageas 2020).

Thus to characterize the global convergence properties, we first look at the single-dimension case and show that when Gradient-Truncated EM is initialized in some interval $[-B, B]$, which contains the true means, the dynamics stays in the interval at all times. Finally, to argue about the global convergence we turn to the guarantees gradient descent for non-convex minimization of smooth functions. Additionally, we show that the function is locally strongly convex and locally smooth which enables us to provide faster rates of convergence to the true means when Gradient-Truncated EM is initialized close to the true means.

For higher dimensions, we provide guarantees of local convergence by characterizing the strong convexity and smoothness around $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$, similar to (Nagarajan and Panageas 2020) which guarantees in the general setting, local convergence in higher dimensions. Before we can state the main theorem, we require the following lemmas about Gradient-Truncated EM. All the lemmas and theorems in this section are stated with Assumption 10.

Convergence in Single-Dimensions

First, we show that for a small enough (constant learning rate) η, the dynamics stays inside the interval $[-B, B]$.

Lemma 12. *The dynamical system induced by Gradient-Truncated EM when $d = 1$ is such that, when $-B \leq \lambda_0 \leq B$ and when Assumption 10 holds, then $\lambda_t \in [-B, B]$ for all $t \in \mathbb{Z}^+ \cup \{0\}$.*

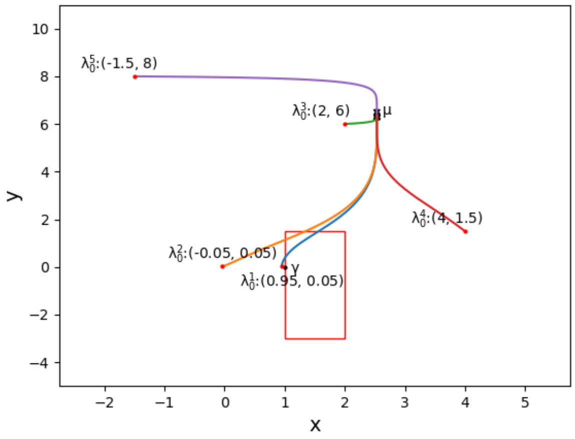


Figure 2: Trajectories of (Gradient-Truncated EM) for five starting points. The red box is the truncation set.

Using, the above lemma, we can see that the dynamics is restricted to the interval $[-B, B]$ and this ensures, that we can obtain smoothness guarantees of the gradient and the Hessian in this interval. This “regularity" helps us avoid a projection step. Now, we are able to derive the smoothness constant.

The computation of gradient Lipschitzness (smoothness) requires an upper bound on the magnitude of the first derivative of the gradient.

Lemma 13. *The function $f(\lambda)$, when $d = 1$ is $\mathcal{O}\left(\frac{1}{\alpha^{2(B+1)}}\right)$-gradient Lipschitz.*

Before going to the global convergence guarantees, we need the following lemma to say that the gradient outside the true means is “large", so when we check the termination condition in Algorithm 1, we can run it long enough so as to ensure that we are close to the true mean.

Lemma 14. *The magnitude of the gradient $|g(\lambda)| > \Omega(c\alpha^2(\tanh(\alpha c))^2)$ when,*

$$\lambda \in [-B, B] \backslash ([\mu - c, \mu + c] \cup [-\mu - c, -\mu + c] \cup [-c, c]) .$$

Now we can state the main theorem that guarantees global convergence to the true means.

Theorem 15 (Single dimensional (global convergence))**.** *Given, any $\epsilon > 0$, when* (Gradient-Truncated EM) *(with η set to $\alpha^{2(B+1)}$) is initialized to λ_0, it finds a point $\tilde{\mu}$ such that $|\tilde{\mu} - \mu| \leq \epsilon$ when $B \geq \lambda_0 > \epsilon$ in at most*

$$\mathcal{O}\left(\frac{|\lambda_0 - \mu|^2}{\alpha^{4(B+2)}|\lambda_0|^3\epsilon^2}\right), \tag{13}$$

steps, with the assumption that the measure under truncation is $\alpha > 0$ and also $|\mu| \leq B$. Analogously, when $-B \leq \lambda_0 < -\epsilon$, then $|\tilde{\mu} + \mu| \leq \epsilon$ in at most $\mathcal{O}\left(\frac{|\lambda_0 + \mu|^2}{\alpha^{4(B+2)}|\lambda_0|^3\epsilon^2}\right)$.

Proof. We first apply the equivalence between (Gradient-Truncated EM) and (Truncated EM) shown in Lemma 11 and thus (Gradient-Truncated EM) has only three fixed points μ, $-\mu$ and 0. Now given that $B \geq \lambda_0 > \epsilon$ or $-B \leq \lambda_0 < -\epsilon$, the iterates never visit $[-\epsilon, \epsilon]$ due to Lemma 12. Now, Theorem 7, guarantees that gradient descent outputs a point $\tilde{\mu}$ such that $|g(\tilde{\mu})| \leq \epsilon$ in at most:

$$\mathcal{O}\left(\frac{(f(\lambda_0) - f(\mu))}{\alpha^{2(B+1)}\epsilon^2}\right), \tag{14}$$

by Lemma 13 for the appropriate smoothness constant. Using, the standard fact of L-smooth functions on a convex domain, we obtain that $f(\lambda_0) - f(\mu) \leq \frac{L}{2}|\lambda_0 - \mu|^2$. In addition, since this method terminates $|f'(\tilde{\lambda})| \leq \epsilon$, to ensure that we are ϵ close to μ, we must run (Gradient-Truncated EM) for $\frac{1}{\epsilon^2|\lambda_0|^3\alpha^4}$ iterations, to find a point $\tilde{\mu}$ that is ϵ close to μ due to Lemma 14. This gives us the required iteration complexity. □

Finally, we study the local properties of f around the true parameters. We can show that the function is locally strongly convex and locally smooth (with a better smoothness guarantee as compared to global smoothness). These properties guarantee better convergence rates when the initial beliefs are close to the true parameters.

Lemma 16. *The function $f(\lambda)$ when $d = 1$ is $\Omega(\alpha^4)$-locally strongly convex and $\mathcal{O}\left(\frac{1}{\alpha^2}\right)$-locally smooth around μ (or equivalently -μ).*

Now, we can show the local convergence guarantees in the single dimensional case the proof of which follows directly by applying Lemma 16 in conjunction with Theorem 9.

Theorem 17 (Single dimensional (local convergence))**.** *Given, any $\epsilon > 0$, there exists a neighborhood of μ and (equivalently $-\mu$) such that when* (Gradient-Truncated EM) *(with η set to α^2) is initialized to λ_0 in this neighborhood outputs a point $\tilde{\mu}$ such that $|\tilde{\mu} - \mu| \leq \epsilon$ in at most*

$$\mathcal{O}\left(\frac{1}{\alpha^6}\log\left(\frac{|\lambda_0 - \mu|}{\epsilon}\right)\right), \tag{15}$$

steps, with the assumption that the measure under truncation is $\alpha > 0$ and also $|\mu| \leq B$. Analogously, when λ_0 in the neighborhood of of $-\mu$, then $|\tilde{\mu} + \mu| \leq \epsilon$ in at the most $\mathcal{O}\left(\frac{1}{\alpha^6}\log\left(\frac{|\lambda_0+\mu|}{\epsilon}\right)\right)$.

Convergence in Higher Dimensions

As stated earlier, in general we are able to guarantee only local convergence, similar to (Nagarajan and Panageas 2020), as in higher dimensions, the truncation may introduce new fixed points. This problem is ported to the (Gradient-Truncated EM) as Lemma 11 shows that there is a one-one mapping between the fixed points of (Gradient-Truncated EM) and (Truncated EM).

We now state, the following Lemma that guarantees the local properties of the negative log-likelihood function is locally strongly convex and locally smooth in higher dimensions as well.

Lemma 18. *The function $f(\boldsymbol{\lambda})$ when $d > 1$ is $\Omega(\alpha^4)$-locally strongly convex and $\mathcal{O}\left(\frac{1}{\alpha^2}\right)$-locally smooth around $\boldsymbol{\mu}$ (or equivalently -$\boldsymbol{\mu}$).*

Finally, we can state the local convergence results in high dimensions the proof of which follows directly by applying Lemma 18 in conjunction with Theorem 9.

Theorem 19 (Multi dimensional (local convergence))**.** *Given, any $\epsilon > 0$, there exists a neighborhood of $\boldsymbol{\mu}$ (equivalently $-\boldsymbol{\mu}$) such that when* (Gradient-Truncated EM) *(with η set to α^2) is initialized to $\boldsymbol{\lambda}_0$ in this neighborhood outputs a point $\tilde{\boldsymbol{\mu}}$ such that $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \leq \epsilon$ in at most*

$$\mathcal{O}\left(\frac{1}{\alpha^6}\log\left(\frac{\|\boldsymbol{\lambda}_0 - \boldsymbol{\mu}\|_2}{\epsilon}\right)\right), \tag{16}$$

steps, with the assumption that the measure under truncation is $\alpha > 0$ and also $\|\boldsymbol{\mu}\|_2 \leq B$. Analogously, when $\boldsymbol{\lambda}_0$ in the neighborhood of of $-\boldsymbol{\mu}$, then $\|\tilde{\boldsymbol{\mu}} + \boldsymbol{\mu}\|_2 \leq \epsilon$ in at most $\mathcal{O}\left(\frac{1}{\alpha^6}\log\left(\frac{\|\boldsymbol{\lambda}_0+\boldsymbol{\mu}\|_2}{\epsilon}\right)\right)$.

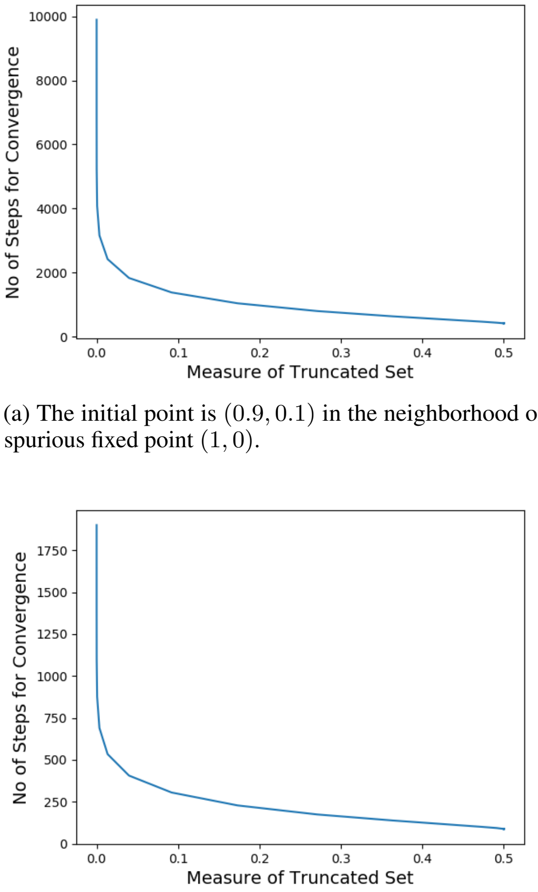
Experiments

Since the higher dimensional settings are prone to spurious fixed points, we try to perform some experiments to understand the convergence rates of (Gradient-Truncated EM) both globally and locally, when the truncation sets are boxes (when $d = 2$). Specifically for box truncation we focus on the case identified by (Nagarajan and Panageas 2020) that has additional fixed points; see Figure 2). In Figure 2), you can see the trajectories of (Gradient-Truncated EM) for five starting points. The red box is the truncation set. All trajectories converge to the true mean $(2.534, 6.395)$. Specifically $\boldsymbol{\lambda}_0^1 = (0.95, 0.05)$ which is in the neighborhood of $\boldsymbol{\gamma} = (1, 0)$ and is a spurious fixed point in this setting. Similarly, $\boldsymbol{\lambda}_0^2 = (-0.05, 0.05)$ is close to $(0, 0)$ which is a saddle point and $\boldsymbol{\lambda}_0^3 = (2, 6)$ is in the neighborhood of the true mean. The remaining starting points are far away. Notice that even if there are spurious fixed points, (Gradient-Truncated EM) converges fast to the true mean.

In addition, we study more complex truncation sets such as a tetrahedron and a set which is an intersection of a paraboloid and a sphere (when $d = 3$). Finally, we study how the convergence rates depend on the measure of the truncation sets.

We define certain terms that will be useful in the experimental context. We perform the experiments with respect to the true mean which is indicated by the vector (μ_1, μ_2) and thus the mean for the other component is $(-\mu_1, -\mu_2)$. Let the threshold to reach a certain error be defined by ϵ. When we mention the average rates, we consider the number of iterations required to reach within ϵ of the true parameters from a particular initial condition and then take the average number of iterations over 50 randomly chosen starting points.

The performance analysis of (Gradient-Truncated EM) on box-truncation sets in 2-dimensions is deferred to the corresponding section of the supplementary material. Our experimental results in higher dimensions, provides us hope that even if spurious fixed points arise due to truncation, some gradient variant of EM may still perform well in these cases. We let the measure of the truncated set vary and then record the number of iterations required to a threshold of error $\epsilon = 0.1$.



(a) The initial point is $(0.9, 0.1)$ in the neighborhood of the spurious fixed point $(1, 0)$.

(b) The initial point is $(2.4, 6.2)$ (neighborhood of $(2.534, 6.395)$).

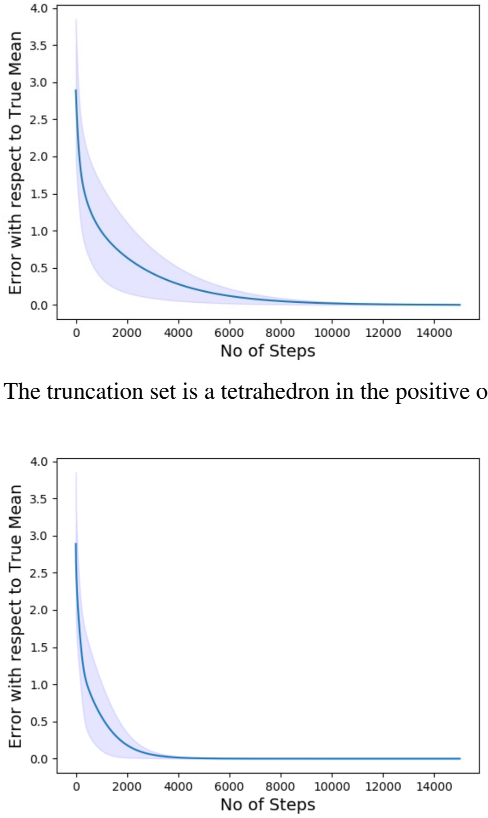
Figure 3: The # of iterations required to reach an error threshold of $\epsilon = 0.1$ vs measure when the true mean is $(2.534, 6.395)$ and the truncation set varies from $(1, 2), (-3, 1.5)$ to $(1, 22), (-3, 21.5)$ with 0.5 increments in the x and y coordinates.

Conclusion

We studied the problem of mean estimation for truncated two component Gaussian mixtures and we proposed a gradient based rule, (Gradient-Truncated EM), given that the original EM update rule under truncation has an implicit form which makes it impractical as an algorithm. Characterization of spurious fixed points arising in certain truncation functions or sets of interest and analyzing the finite sample settings are tantalizing future directions to investigate.

Acknowledgements

Sai Ganesh Nagarajan acknowledges the SUTD Presidents Graduate PhD Fellowship, as majority of the work was done when he was a PhD student at SUTD. Ioannis Panageas would like to acknowledge a start-up grant.



(a) The truncation set is a tetrahedron in the positive orthant.

(b) The truncation set is the intersection of a 3-D sphere and a paraboloid around the origin.

Figure 4: The error with respect to the true mean vs # of iterations, averaged over multiple starting points. The shaded region indicates the +/- 1 stddev from the mean performance. This indicates that convergence to the true means $\boldsymbol{\mu} = (3, 2, 1)$ is observed, albeit at different rates.

References

Amid, E.; and Warmuth, M. K. 2019. An Implicit Form of Krasulina's k-PCA Update without the Orthonormality Constraint. *arXiv preprint arXiv:1909.04803*.

Aristophanous, M.; Penney, B. C.; Martel, M. K.; and Pelizzari, C. A. 2007. A Gaussian mixture model for definition of lung tumor volumes in positron emission tomography. *Medical physics*, 34(11): 4223–4235.

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

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

Brigo, D.; and Mercurio, F. 2002. Displaced and mixture diffusions for analytically-tractable smile models. In *Mathematical Finance—Bachelier Congress 2000*, 151–174. Springer.

Bubeck, S. 2014. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*.

Daskalakis, C.; Gouleakis, T.; Tzamos, C.; and Zampetakis, M. 2018. 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*, 639–649.

Daskalakis, C.; Gouleakis, T.; Tzamos, C.; and Zampetakis, M. 2019. Computationally and Statistically Efficient Truncated Regression. In *Conference on Learning Theory*, 955–960.

Daskalakis, C.; Tzamos, C.; and Zampetakis, M. 2017. 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*, 704–710.

Dwivedi, R.; Ho, N.; Khamaru, K.; Wainwright, M.; Jordan, M.; and Yu, B. 2020a. Sharp analysis of expectation-maximization for weakly identifiable models. In *International Conference on Artificial Intelligence and Statistics*, 1866–1876. PMLR.

Dwivedi, R.; Ho, N.; Khamaru, K.; Wainwright, M. J.; Jordan, M. I.; and Yu, B. 2020b. Singularity, misspecification and the convergence rate of EM. *The Annals of Statistics*, 48(6): 3161–3182.

Fisher, R. 1931. Properties and applications of Hh functions. *Mathematical tables*, 1: 815–852.

Ho, N.; Khamaru, K.; Dwivedi, R.; Wainwright, M. J.; Jordan, M. I.; and Yu, B. 2020. Instability, computational efficiency and statistical accuracy. *arXiv preprint arXiv:2005.11411*.

Ho, N.; and Nguyen, X. 2016. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6): 2726–2755.

Jin, C.; Ge, R.; Netrapalli, P.; Kakade, S. M.; and Jordan, M. I. 2017. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1724–1732. JMLR. org.

Jin, C.; Zhang, Y.; Balakrishnan, S.; Wainwright, M. J.; and Jordan, M. I. 2016. 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*, 4116–4124.

Kwon, J.; Ho, N.; and Caramanis, C. 2021. On the minimax optimality of the EM algorithm for learning two-component mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, 1405–1413. PMLR.

Lange, K. 1995. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(2): 425–437.

Lee, A.; and Pearson, K. 1908. On the Generalised Probable Error in Multiple Normal Correlation. *Biometrika*, 6(1): 59–68.

Lee, G.; and Scott, C. 2012. EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 56(9): 2816–2829.

Lee, J. D.; Panageas, I.; Piliouras, G.; Simchowitz, M.; Jordan, M. I.; and Recht, B. 2019. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176(1-2): 311–337.

McLachlan, G.; and Jones, P. 1988. Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics*, 571–578.

Melchior, P.; and Goulding, A. D. 2018. Filling the gaps: Gaussian mixture models from noisy, truncated or incomplete samples. *Astronomy and computing*, 25: 183–194.

Nagarajan, S. G.; and Panageas, I. 2020. On the analysis of EM for truncated mixtures of two gaussians. In *Algorithmic Learning Theory*, 634–659.

Nesterov, Y. 1998. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4): 5.

Rubin, D. B. 1976. Inference and missing data. *Biometrika*, 63(3): 581–592.

Tasche, D. 2002. Expected shortfall and beyond. *Journal of Banking & Finance*, 26(7): 1519–1533.

Wu, C. J. 1983. On the convergence properties of the EM algorithm. In *The Annals of statistics*, 95–103.

Xu, J.; Hsu, D. J.; and Maleki, A. 2016. 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*, 2676–2684.

Yan, B.; Yin, M.; and Sarkar, P. 2017. Convergence of gradient EM on multi-component mixture of Gaussians. In *Advances in Neural Information Processing Systems*, 6956–6966.