Mean estimation of truncated mixtures of two Gaussians: A gradient based approach

Sai Ganesh Nagarajan
SUTD*
sai_nagarajan@mymail.sutd.edu.sg

Ioannis Panageas
SUTD*
ioannis@sutd.edu.sg

Tushar Vaidya
SUTD*
tushar_vaidya@sutd.edu.sg

Samson Yu Bai Jian
SUTD*
samson_yu@mymail.sutd.edu.sg

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 [17] showed EM converges globally in single dimension $d = 1$ and exhibits local convergence for $d > 1$. 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. In this work, we propose a gradient based variant of the EM algorithm that has global convergence guarantees when $d = 1$. Moreover, the update rule at every iteration is easy to compute. We also provide numerous experiments to obtain insights into the effect of truncation on the convergence to the true parameters in high dimensions.

1 Introduction

The performance of algorithms in parameter estimation problems are 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 is introduced in the model.

Statisticians, dating back to Pearson [12] and Fisher [9], 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 [19] in 1976, on missing/censored data, tried to approach this by a framework of ignorable and non-ignorable missingness, where the reason for missingness is incorporated into the statistical model. However, in many cases such flexibility may not be available.

*Singapore University of Technology and Design.

Gaussian mixtures are ubiquitous in machine learning and statistics with a variety of applications ranging from biology [4, 2] to finance [5, 20], 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.

A standard approach for parameter estimation for Gaussian Mixture models is the so-called 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 [21]. 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. Most theoretical analyses focus on simple cases, i.e., mixtures of two Gaussians and mean estimation with known covariance and balanced mixtures. Recent results indicate that EM works well (converges to true mean) for mixture of two Gaussians (see [22], [8] for global convergence and [3] for local convergence), a result that is not true if the number of components is at least three (in [11] 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 [8], [22] and the references therein.

More 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 [6]. 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, as we shall see later. Using similar techniques, the authors in [7], address the problem of truncated regression.

Closer to our work, the results of [17] analyzed EM for a truncated mixture of two Gaussians and showed that when the Gaussians are single dimensional (i.e, $d = 1$), EM globally converges to the true means (and has an exponential rate locally). 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 authors 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 [16], 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. They impute the missing observations by performing rejection sampling with current parameters of the EM algorithm with appropriate correction terms to the M-step. Similarly, in [13] and [15] 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 ([21]) 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 [17], where the authors provide a two-dimensional example where the truncation set is a box and a spurious stationary point of the truncated EM appears.

To this end, we identify the main challenges which make this problem elusive to theoretical analysis.

Technical Challenges for Truncated Gaussian Mixtures The first challenge is the non-convexity of the problem even in single dimensions and the second one being the inability to overcome the implicit update rule of truncated EM without significant computational cost especially when $d \geq 2$.

As stated before, authors in [17] 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. We know from the results of [17], when $d = 1$, the true parameters, μ, -μ and 0 are the only fixed points for the negative log-likelihood and 0 is a saddle point. The plot shows that even the single dimensional case is non-convex which makes it hard to obtain global convergence guarantees.

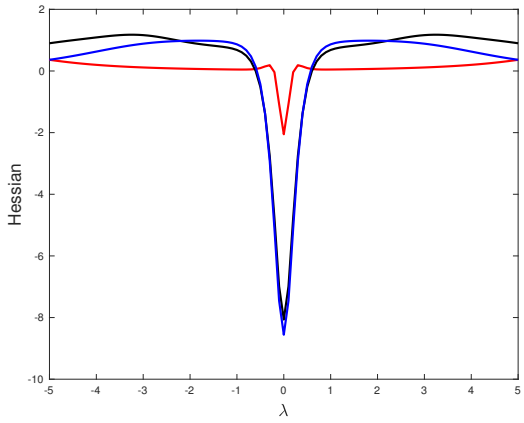


Figure 1: The second-derivative of the negative log-likelihood, when $d = 1$, is computed and shown here for different truncation sets, when the true means are $\mu, -\mu = 3, -3$. As seen here when λ is close to 3 or -3, the second-derivative is positive around the region and when λ is close to 0, it 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.

Secondly, the original truncated population EM which is derived in [17] is an implicit update equation, such that finding the parameters for the next time step involves solving a set of non-linear equations. 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 [1] or similar problems where one can easily solve the implicit equations in finite sample settings.

When $d = 1$, we may still be able to perform binary search to solve for λ_{t+1}, but this becomes impractical in higher dimensions.

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.

Our results and techniques We propose the (Gradient-Truncated EM) algorithm which performs gradient descent on the population negative log-likelihood to circumvent the issue of the practicality of (Truncated EM) [17]. By taking this algorithmic approach, even when the problem is non-convex, we are able to obtain convergence guarantees when $d = 1$, by using recent techniques from [10] that hinges on global properties of the function such as smoothness of the gradient and the Hessian.

Theorem 1.1 (Informal Statement)**.** *Given any $\delta > 0$, $0 < \epsilon \leq \frac{1}{\alpha^{B+1}}$ and $\Delta_f \geq f(\lambda_0) - f^*$, the* (Gradient-Truncated EM) *rule on the population negative log-likelihood $f(.)$ in the single dimensional case converges to an estimate $\tilde{\mu}$ (or equivalently $-\tilde{\mu}$) with probability $1 - \delta$ such that, $|\frac{\mu - \tilde{\mu}}{\sigma}| \leq \epsilon$,in the following number of iterations:*

$$\mathcal{O}\left(\frac{f(\boldsymbol{\lambda}_0) - f^*}{poly(\alpha)\epsilon^2} \log^4\left(\frac{\Delta_f}{poly(\alpha)\epsilon^2\delta}\right)\right), \tag{1.1}$$

under the assumption that the measure under truncation is $\alpha > 0$ and also $|\mu| \leq B$. Where, λ_0 is the initial condition, $f^* = f(\mu) = f(-\mu)$ is the optimal value.

Finally, we provide experimental results and some insights into the convergence guarantees in high dimensional settings, relating the convergence rates of the gradient based EM algorithm to the measure of the truncation sets. Specifically, we focus on the example provided by the authors in [17], where for an appropriate choice of true means and a truncation set which is a particular rectangle, truncated EM has a spurious fixed point.

2 Background

2.1 Truncated Mixture Model

Before describing the model, we establish the notations used in this paper. The notations and the settings are going to be same as the notations introduced for the truncated mixture setting in[17] . 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 a similar setting as described in [17], 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 to be $(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

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 2.1.

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

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.

The truncation function $S(\boldsymbol{x})$ can be seen as a term that controls selection bias, similar to the one analyzed by the authors in [17].

We will denote the expected value with respect to the truncated mixture distribution with parameters $-\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}$ by $\mathbb{E}_{\boldsymbol{\lambda},S}[.]$. The population EM update rule for the truncated setting which was described in [17] is given below.

$$\text{Define function } h(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}) := \mathbb{E}_{\boldsymbol{\mu},S}\left[\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda}_t)\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\right] - \mathbb{E}_{\boldsymbol{\lambda},S}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right]. \quad (2.1)$$

The next iterate is

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

2.2 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 [18]. 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 [14]. 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 [10], propose a "perturbed" gradient method that recovers the original rates by Nesterov [18] up to *polylog* factors in the dimension d.

To obtain some convergence guarantees on smooth non-convex functions, [10] require some weaker notions such as an approximate first order stationary points and approximate second order stationary points.

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 2.2 (Lipschitzness)**.** *A twice differentiable function f is L-smooth if it satisfies the following condition:*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \ \ \forall x, y \quad (2.2)$$

A twice differentiable function f is said to have ρ-Hessian Lipschitzness, if the following holds:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq \rho\|x - y\|_2 \ \ \forall x, y \quad (2.3)$$

In order to get a complete picture we require the following definitions, that appear in [10].

Definition 2.3 (Approximate first/second order stationary points)**.**

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

A point x^* is an ϵ-strict saddle point of f if it is an ϵ- first order stationary point and $\lambda_{\min}(\nabla^2 f(x^*)) \leq -\sqrt{\rho\epsilon}$.

The ϵ-first order points that are not ϵ-strict saddles are ϵ- second order stationary points.

Given the above definitions, we state the following theorem:

Theorem 2.4 (Theorem 3 in [10])**.** *When a function f satisfies the conditions stated in Definitions 2.2, for any $\delta > 0$, $0 < \epsilon \leq \frac{L^2}{\rho}$, $\Delta_f \geq f(\boldsymbol{x}_0) - f^*$, Perturbed Gradient Descent outputs an ϵ-second order stationary point with probability $1 - \delta$ in the following number of iterations:*

$$\mathcal{O}\left(\frac{L(f(\boldsymbol{x}_0) - f^*)}{\epsilon^2}\log^4\left(\frac{dL\Delta_f}{\epsilon^2\delta}\right)\right) \quad (2.4)$$

3 Gradient-Truncated EM

As mentioned in the previous section equation (2.1) describes the truncated EM update rule in the population setting derived by [17]. 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 that is more amenable to analysis and that allows us to easily compute the parameters at every step. We describe the rule below:

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

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

For convenience, we are going to look at the negative log likelihood and use (Gradient-Truncated EM) as a gradient descent algorithm for a non-convex minimization problem. Thus we can re-write it as follows:

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

The gradient in the above equation, is the gradient of the population negative log-likelihood function $f(\boldsymbol{\lambda})$ and is written as (i.e., $g(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}} f(.)$):

$$g(\boldsymbol{\lambda}) = \mathbb{E}_{\boldsymbol{\lambda},S}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right] - \mathbb{E}_{\boldsymbol{\mu},S}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\tanh(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\lambda})\right]. \tag{3.2}$$

Finally, we state the following assumption

Assumption 3.1. *The true mean has bounded norm, i.e,* $\|\boldsymbol{\mu}\|_2 \leq B$.

With the above assumption in place, the negative log-likelihood is now defined on the d-dimensional ball of radius B, i.e, $\mathcal{X} = \{\boldsymbol{\lambda} : \|\boldsymbol{\lambda}\|_2 \leq B\}$. Let $\mathcal{X}^* \subseteq \mathcal{X}$ such that $\mathcal{X}^* = \mathcal{X}^*_- \cup \mathcal{X}^*_+$, where $\mathcal{X}^*_-$ is the neighbourhood of $-\boldsymbol{\mu}$ and $\mathcal{X}^*_+$ is the neighbourhood of $\boldsymbol{\mu}$. Let f^* minimum value of the negative log-likelihood which is at $\boldsymbol{\mu}$ or $-\boldsymbol{\mu}$.

Now we state the following useful lemma that is an adaption of Lemma 7 that appears in [6] to the setting where the covariances are same and known.

Lemma 3.2. *Suppose* $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ *and* $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ *are two multivariate Gaussians such that:*

$$\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\|_2 \leq B. \tag{3.3}$$

If $\Pr(S; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \geq \alpha$, *then* $\Pr(S; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \geq 7\exp(-B^2)\left(\frac{\alpha}{8}\right)^{B+1}$.
Where, $\Pr(S; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = \int_{\mathbb{R}^d} S(\boldsymbol{x})\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})d\boldsymbol{x}$

Proof. Following the proof of Lemma 7 in [6], we can write the following, we first assume without loss of generality that $\boldsymbol{\mu}_1 = 0$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$

$$\Pr(S; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \mathbb{E}_{\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[S(\boldsymbol{x})\frac{\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{I})}{\mathcal{N}(\boldsymbol{x}; \boldsymbol{0}, \boldsymbol{I})}\right] \tag{3.4}$$

Call the above ratio of the two pdfs to be $R(\boldsymbol{x})$. Then, we have that

$$R(\boldsymbol{x}) = \exp\left(-\left(\sum_i \mu_{2i}^2 - 2x_i\mu_{2i}\right)\right) \tag{3.5}$$

If suppose we show that $R(\boldsymbol{x}) \geq l$ with probability $1 - \frac{\alpha}{8}$, then it follows from Equation 3.4 that $\Pr(S; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \geq \frac{7}{8}\alpha l$.

Define the set $S^{'} := \{\boldsymbol{x} : R(\boldsymbol{x}) \geq l\}$. By the above claim $\Pr(S^{'}; \mathbf{0}, \boldsymbol{I}) \geq 1 - \frac{\alpha}{8}$. Also from the given constraints, we know that $\Pr(S; \mathbf{0}, \boldsymbol{I}) \geq \alpha$. Then we have the following:

$$\begin{aligned}
\int_{\mathbb{R}^d} S(\boldsymbol{x})\mathcal{N}(\boldsymbol{x}; \mathbf{0}, \boldsymbol{I})d\boldsymbol{x} &= \int_{S'} S(\boldsymbol{x})\mathcal{N}(\boldsymbol{x}; \mathbf{0}, \boldsymbol{I})d\boldsymbol{x} + \int_{\mathbb{R}^d \setminus S'} S(\boldsymbol{x})\mathcal{N}(\boldsymbol{x}; \mathbf{0}, \boldsymbol{I})d\boldsymbol{x} \\
\int_{S'} S(\boldsymbol{x})\mathcal{N}(\boldsymbol{x}; \mathbf{0}, \boldsymbol{I})d\boldsymbol{x} &\geq \alpha - \int_{\mathbb{R}^d \setminus S'} S(\boldsymbol{x})\mathcal{N}(\boldsymbol{x}; \mathbf{0}, \boldsymbol{I})d\boldsymbol{x} \\
&\geq \alpha - \int_{\mathbb{R}^d \setminus S'} \mathcal{N}(\boldsymbol{x}; \mathbf{0}, \boldsymbol{I})d\boldsymbol{x} \\
&\geq \alpha - \frac{\alpha}{8} \\
&= 7\frac{\alpha}{8}
\end{aligned} \tag{3.6}$$

Now, the upper bound on $\int_{\mathbb{R}^d \setminus S'} S(\boldsymbol{x})\mathcal{N}(\boldsymbol{x}; \mathbf{0}, \boldsymbol{I})d\boldsymbol{x}$ is because $S(\boldsymbol{x})$ is non-negative and bounded above by 1.

Thus, we can see that the probability can be bounded below as follows:

$$\begin{aligned}
\Pr(S; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) &= \int_{\mathbb{R}^d} S(\boldsymbol{x})R(\boldsymbol{x})\mathcal{N}(\boldsymbol{x}; \mathbf{0}, \boldsymbol{I})d\boldsymbol{x} \\
&\geq l \int_{\mathbf{1}(\boldsymbol{x} \in S')} S(\boldsymbol{x})\mathcal{N}(\boldsymbol{x}; \mathbf{0}, \boldsymbol{I})d\boldsymbol{x} \\
&\geq \frac{7}{8}\alpha l
\end{aligned} \tag{3.7}$$

To find l, we simply use the concentration lemma used in [6] lemma 7 which is as follows:

$$\Pr_{\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})} \left[-\sum_i x_i \mu_{2i} \geq \|\boldsymbol{\mu}_2\| \log\left(\frac{8}{\alpha}\right) \right] \leq \frac{\alpha}{8} \tag{3.8}$$

Then, the $R(\boldsymbol{x}) \geq \exp\left(-\left(B^2 + B\log\left(\frac{8}{\alpha}\right)\right)\right)$ with probability $1 - \frac{\alpha}{8}$ This implies, the lower bound $l = \exp(-B^2)\left(\frac{\alpha}{8}\right)^B$. Combining with the previous claim, finally we obtain that $\Pr(S; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \geq 7\exp(-B^2)\left(\frac{\alpha}{8}\right)^{B+1}$, which gives us the required result. □

Finally, we end this section by characterizing an equivalence between the Gradient Truncated EM and the Truncated EM framework of [17]. We show that there is a one-one mapping of the fixed points.

Lemma 3.3. *The fixed points of* (Gradient-Truncated EM) *and* (Truncated EM) *have a one-one mapping.*

Proof. If $\boldsymbol{\gamma}$ is a fixed point then in general we have, $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t = \boldsymbol{\gamma}$. Let $\boldsymbol{\gamma}$ be a fixed point of Gradient-Truncated EM. Then we have that:

$$\mathbb{E}_{\boldsymbol{\gamma}, S}\left[\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \tanh(\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma})\right] = \mathbb{E}_{\boldsymbol{\mu}, S}\left[\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \tanh(\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma})\right] \tag{3.9}$$

Given that $\boldsymbol{\lambda}_t = \boldsymbol{\gamma}$, we know that: $h(\boldsymbol{\gamma}, \boldsymbol{\lambda}_{t+1}) = 0$. From Lemma 9 of [17], we know that there exists a unique $\boldsymbol{\lambda}_{t+1}$, that satisfies the Truncated EM update rule. Also from Equation 3.9,we have that

$h(\boldsymbol{\gamma}, \boldsymbol{\gamma}) = 0$, combining the above statements, this means that $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\gamma}$, which is a fixed point of Truncated EM.

In the other direction, if $\boldsymbol{\gamma}$ is a fixed point of Truncated EM, then we have that $h(\boldsymbol{\gamma}, \boldsymbol{\gamma}) = 0$ and $\boldsymbol{\lambda}_t = \boldsymbol{\gamma}$. Thus, writing out the Gradient-Truncated EM when $\boldsymbol{\lambda}_t = \boldsymbol{\gamma}$, it holds that:

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\gamma} - \eta * h(\boldsymbol{\gamma}, \boldsymbol{\gamma}) \tag{3.10}$$

The above equation directly implies that $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\gamma}$ and hence is a fixed point of Gradient-Truncated EM. □

From lemma 3.3, we can conclude that when $d = 1$, (Gradient-Truncated EM) has three fixed points, which are μ, $-\mu$ and 0. Additionally, local stability properties carry over from [17] (Lemmas 12 and 13), which means that μ and $-\mu$ are locally stable and 0 is unstable. This means that μ and $-\mu$ are local minima (or a second order stationary point) and 0 is a saddle point.

4 Convergence of Population Gradient -Truncated EM

As seen previously, the problem is highly non-convex, especially for arbitrary truncation sets. Thus we are only able to characterize some global aspects of the population negative log-likelihood such as Lipschitzness. Then, to argue about the global convergence we turn to the guarantees of perturbed gradient descent for non-convex minimization as shown by the authors in [10] to finally provide the convergence results.

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 3.1.

We show some global properties of the population negative log-likelihood in the single dimensional case. The proof requires an upper bound on the magnitude of the second derivative and third derivative respectively.

Lemma 4.1. *The population negative log-likelihood function, when $d = 1$ is $\mathcal{O}\left(\frac{1}{\alpha^{2(B+1)}}\right)$-smooth and $\mathcal{O}\left(\frac{1}{\alpha^{3(B+1)}}\right)$-Hessian Lipschitz.*

Proof. The proof requires an upper bound on the magnitude of the second derivative and third derivative respectively.

For the smoothness, It suffices to find an upper bound on the second derivative of the function. The second derivative is obtained as follows:

$$g^{'}(\lambda) = \frac{1}{\sigma^2}\left(\mathbb{E}_{\lambda,S}[x^2] - \mathbb{E}^2_{\lambda,S}\left[x\tanh\left(\frac{x\lambda}{\sigma^2}\right)\right] - \mathbb{E}_{\mu,S}[x^2] + \mathbb{E}_{\mu,S}\left[x^2\tanh^2\left(\frac{x\lambda}{\sigma^2}\right)\right]\right) \tag{4.1}$$

To bound the magnitude of the second derivative:

$$\sigma^2|g^{'}(\lambda)| = \mathbb{E}_{\lambda,S}[x^2] - \mathbb{E}^2_{\lambda,S}\left[x\tanh\left(\frac{x\lambda}{\sigma^2}\right)\right] - \mathbb{E}_{\mu,S}[x^2] + \mathbb{E}_{\mu,S}\left[x^2\tanh^2\left(\frac{x\lambda}{\sigma^2}\right)\right] \tag{4.2}$$

$$\leq |\mathbb{E}_{\lambda,S}[x^2]| + \left|\mathbb{E}^2_{\lambda,S}\left[x\tanh\left(\frac{x\lambda}{\sigma^2}\right)\right]\right| + |\mathbb{E}_{\mu,S}[x^2]| + \left|\mathbb{E}_{\mu,S}\left[x^2\tanh^2\left(\frac{x\lambda}{\sigma^2}\right)\right]\right| \tag{4.3}$$

$$\leq |\mathbb{E}_{\lambda,S}[x^2]| + 2|\mathbb{E}_{\mu,S}[x^2]| + |\mathbb{E}^2_{\lambda,S}[x]| \tag{4.4}$$

$$\leq |\mathbb{E}_{\mathcal{N}(\lambda,\sigma^2,\frac{S+S'}{2})}[x^2]| + 2|\mathbb{E}_{\mathcal{N}(\mu,\sigma^2,\frac{S+S'}{2})}[x^2]| + |\mathbb{E}^2_{\mathcal{N}(\lambda,\sigma^2,\frac{S+S'}{2})}[x]| \tag{4.5}$$

$$\leq \left(B^2+\sigma^2\right)\mathcal{O}\left(\frac{1}{\alpha^{B+1}}\right) + \frac{2}{\alpha}\left(\mu^2+\sigma^2\right) + B^2\mathcal{O}\left(\frac{1}{\alpha^{2(B+1)}}\right) \tag{4.6}$$

$$\leq \mathcal{O}\left(\frac{1}{\alpha^{2(B+1)}}\right) \tag{4.7}$$

We use the fact that the expectation of an even function over any truncated Gaussian mixture distribution reduces to the expectation over a truncated normal distribution with an appropriate truncation. This is described in Lemma 19 of [17]. We obtain the upper bound to be the second moment and then the lower bound on the measure assigned is due to Lemma 3.2 and finally using Assumption 3.1 we get the desired bound.

Now for Hessian Lipschitzness, we require an upper bound on the third derivative.

The third derivative is obtained as follows:

$$\sigma^2 g^{''}(\lambda) = \mathbb{E}_{\lambda,S}[x^3] - \mathbb{E}_{\lambda,S}[x^2]\mathbb{E}_{\lambda,S}[x] + \frac{2}{\sigma^2}\left(\mathbb{E}_{\lambda,S}\left[x\tanh\left(\frac{x\lambda}{\sigma^2}\right)\right]\mathbb{E}_{\lambda,S}[x^2] - \mathbb{E}^3_{\lambda,S}\left[x\tanh\left(\frac{x\lambda}{\sigma^2}\right)\right]\right) \tag{4.8}$$

$$+ \frac{2}{\sigma^2}\left(\mathbb{E}_{\mu,S}\left[x^3\tanh\left(\frac{x\lambda}{\sigma^2}\right)\right] - \mathbb{E}_{\mu,S}\left[x^3\tanh^3\left(\frac{x\lambda}{\sigma^2}\right)\right]\right) \tag{4.9}$$

To bound the magnitude of the third derivative:

$$\sigma^2|g^{'''}(\lambda)| \leq |\mathbb{E}_{\lambda,S}[x^3]| + |\mathbb{E}_{\lambda,S}[x^2]\mathbb{E}_{\lambda,S}[x]| \tag{4.10}$$

$$+ \frac{2}{\sigma^2}\left(\left|\mathbb{E}_{\lambda,S}\left[x\tanh\left(\frac{x\lambda}{\sigma^2}\right)\right]\mathbb{E}_{\lambda,S}[x^2]\right| + \left|\mathbb{E}^3_{\lambda,S}\left[x\tanh\left(\frac{x\lambda}{\sigma^2}\right)\right]\right|\right) \tag{4.11}$$

$$+ \frac{2}{\sigma^2}\left(\left|\mathbb{E}_{\mu,S}\left[x^3\tanh\left(\frac{x\lambda}{\sigma^2}\right)\right]\right| + \left|\mathbb{E}_{\mu,S}\left[x^3\tanh^3\left(\frac{x\lambda}{\sigma^2}\right)\right]\right|\right) \tag{4.12}$$

$$\leq \mathbb{E}_{\lambda,S}[|x^3|] + |\mathbb{E}_{\lambda,S}[x^2]\mathbb{E}_{\lambda,S}[|x|]| + \frac{2}{\sigma^2}\left(\mathbb{E}_{\lambda,S}\left[|x|\right]\mathbb{E}_{\lambda,S}[x^2] + \mathbb{E}^3_{\lambda,S}\left[|x|\right]\right) \tag{4.13}$$

$$+ \frac{2}{\sigma^2}\left(\mathbb{E}_{\mu,S}\left[|x^3|\right] + \mathbb{E}_{\mu,S}\left[|x^3|\right]\right) \tag{4.14}$$

$$= \mathbb{E}_{\mathcal{N}(\lambda,\sigma^2,\frac{S+S'}{2}}\left[|x^3|\right] + |\mathbb{E}_{\mathcal{N}(\lambda,\sigma^2,\frac{S+S'}{2}}[x^2]\mathbb{E}_{\mathcal{N}(\lambda,\sigma^2,\frac{S+S'}{2}}\left[|x|\right]| \tag{4.15}$$

$$+ \frac{2}{\sigma^2}\left(\mathbb{E}_{\mathcal{N}(\lambda,\sigma^2,\frac{S+S'}{2})}\left[|x|\right]\mathbb{E}_{\mathcal{N}(\lambda,\sigma^2,\frac{S+S'}{2})}[x^2] + \mathbb{E}^3_{\mathcal{N}(\lambda,\sigma^2,\frac{S+S'}{2})}\left[|x|\right]\right) \tag{4.16}$$

$$+ \frac{2}{\sigma^2}\left(\mathbb{E}_{\mathcal{N}(\mu,\sigma^2,\frac{S+S'}{2})}\left[|x^3|\right] + \mathbb{E}_{\mathcal{N}(\mu,\sigma^2,\frac{S+S'}{2})}\left[|x^3|\right]\right) \tag{4.17}$$

$$\leq \mathcal{O}\left(\frac{1}{\alpha^{3(B+1)}}\right) \tag{4.18}$$

Similar to the previous case, we use the fact that the expectation of an even function over any truncated Gaussian mixture distribution reduces to the expectation over a truncated normal distribution with an appropriate truncation. We obtain the upper bounds by using the lower bound on the measure assigned, which is due to Lemma 3.2 and finally using Assumption 3.1 we get the desired bound. □

Theorem 4.2. *Given any $\delta > 0$, $0 < \epsilon \leq \frac{1}{\alpha^{B+1}}$ and $\Delta_f \geq f(\lambda_0) - f^*$, the Gradient-Truncated EM rule on the population negative log-likelihood in the single dimensional case converges to an estimate $\tilde{\mu}$ (or equivalently $-\tilde{\mu}$) with probability $1 - \delta$ such that, $|\frac{\mu - \tilde{\mu}}{\sigma}| \leq \epsilon$,in the following number of iterations:*

$$\mathcal{O}\left(\frac{f(\boldsymbol{\lambda}_0) - f^*}{\alpha^{2(B+1)}\epsilon^2} \log^4\left(\frac{\Delta_f}{\alpha^{2(B+1)}\epsilon^2\delta}\right)\right), \tag{4.19}$$

under Assumptions 2.1 and 3.1.

Proof. We invoke lemma 4.1 and to use Theorem 2.4, we set $\epsilon \leq \frac{\alpha^{3B+3}}{\alpha^{4B+4}} = \frac{1}{\alpha^{B+1}}$. Theorem 2.4 guarantees convergence to an ϵ-SOSP . When $d = 1$, by using Lemma 3.3 and Lemma 13 from [17] we get that the only stationary points of Gradient-Truncated EM are 0, μ and $-\mu$. Additionally using Lemma 12 from [17] we conclude that μ and $-\mu$ are locally stable and hence are not saddle points. This means that when Theorem 2.4 guarantees convergence to an ϵ-SOSP, it converges to $\tilde{\mu}$ (or equivalently $-\tilde{\mu}$) such that $|\frac{\mu - \tilde{\mu}}{\sigma}| \leq \epsilon$ in the number of iterations as given by Theorem 2.4. □

5 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 $d = 2$). Particularly, we study how the convergence rates depend on the measure of the truncation sets.

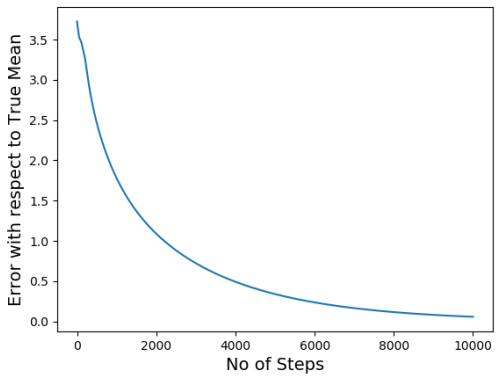
We define certain terms that will be useful in the experimental context. We perform the experiments with respect to the true mean which is indicated by the vector (μ_1, μ_2) and thus the mean for the other component is $(-\mu_1, -\mu_2)$. Let the threshold to reach a certain error be defined by ϵ. In addition we 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 100 randomly chosen starting points. Also we refer to the truncation set by pairs $(x_1, x_2), (y_1, y_2)$ which are the coordinates on the x and y axis respectively. Finally the learning rate is a constant which is fixed to be $\eta = 0.01$.

In the first example, we look at how the error varies with the number of iterations (averaged over 100 random initial conditions) in two examples, where one case is the example given in [17] to show that additional fixed points exist when $d = 2$.

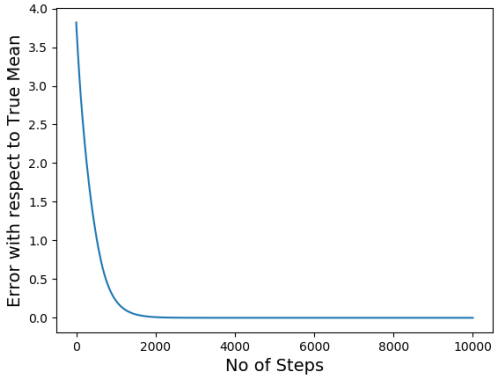
From Figure 2b, we observe convergence to the true parameter; this rules out (at least empirically) that the spurious fixed point is likely to be a spurious maxima for the log-likelihood and that it might just be a saddle or a strict saddle point. Now we let the measure of the truncated set vary and then we record the number of iterations required to a threshold of error given by $\epsilon = 0.1$.

Finally, we visualize how the trajectories of (Gradient-Truncated EM) behave in the 2-D plane for a few starting points.

Figure 4 shows that even in the neighborhood of the spurious fixed point $\boldsymbol{\gamma}$ or the saddle point $(0, 0)$, we have convergence to the true means. At least empirically this provides us hope that these spurious fixed points arising from some of these truncation sets may be strict saddles.

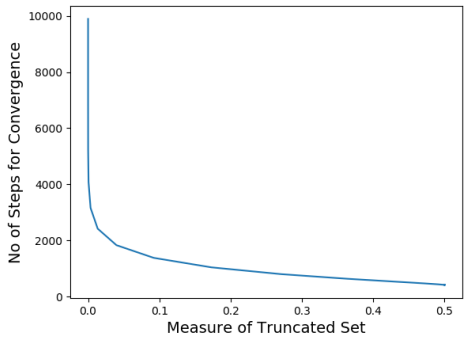


(a) The truncation set is $(1, 2), (-3, 1.5)$ and the true mean is $(2.534, 6.395)$.

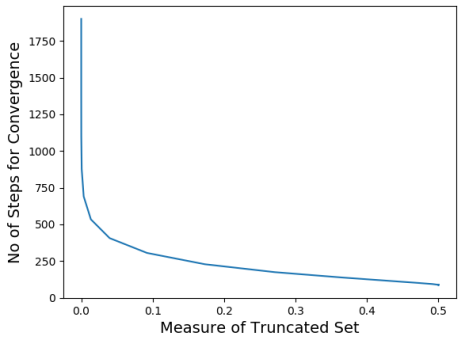


(b) The truncation set is $(-2, 2), (-2, 2)$ and true mean is $(2, 1)$.

Figure 2: The error with respect to the true mean vs the number of iterations, averaged over 100 starting points for two cases. The example on the left has a known spurious fixed point and the example on the right does not.



(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. We can see that the local convergence is effective but its dependence on the measure of the truncated set is still similar to the one on the left where the initial condition is farther away from the true means.

6 Conclusion

We studied the problem of mean estimation for truncated two component Gaussian mixtures and we proposed a gradient based rule, given that the original EM update rule has an implicit form which makes it impractical as an algorithm. We also showed through experiments that although truncation may introduce spurious fixed points, it may be likely that those points are strict saddles which then automatically becomes amenable to analysis. Thus, the characterization of spurious fixed points arising in certain truncation functions or sets of interest and analyzing the sample likelihood are tantalizing future directions that intertwines non-convex optimization and statistics.

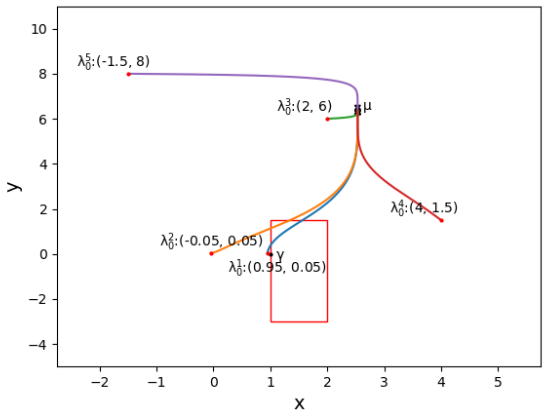


Figure 4: The trajectories of (Gradient-Truncated EM) for five different starting points. The red box is the truncation set. Here trajectories for five different initial conditions are shown, where they all converge to the true mean $(2.534, 6.395)$. Specifically $\boldsymbol{\lambda}_0^1 = (0.95, 0.05)$ which is in the neighborhood of $\boldsymbol{\gamma} = (1, 0)$ which 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 points are starting conditions that are far away. This shows that even though there might be spurious fixed points, (Gradient-Truncated EM) converges to the true mean efficiently and does not get stuck in the neighborhood of that point.

References

[1] Ehsan Amid and Manfred K Warmuth. An implicit form of krasulina's k-pca update without the orthonormality constraint. *arXiv preprint arXiv:1909.04803*, 2019.

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

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

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

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

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

[7] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Computationally and statistically efficient truncated regression. In *Conference on Learning Theory*, pages 955–960, 2019.

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

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

[10] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org, 2017.

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

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

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

[14] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176(1-2):311–337, 2019.

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

[16] Peter Melchior and Andy D Goulding. Filling the gaps: Gaussian mixture models from noisy, truncated or incomplete samples. *Astronomy and computing*, 25:183–194, 2018.

[17] Sai Ganesh Nagarajan and Ioannis Panageas. On the analysis of em for truncated mixtures of two gaussians. In *Algorithmic Learning Theory*, pages 634–659, 2020.

[18] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

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

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

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

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