Finite Sample Guarantees of Differentially Private Expectation Maximization Algorithm

Di Wang\textsuperscript{a,∗}, Jiahao Ding\textsuperscript{b}, Lijie Hu\textsuperscript{c}, Zejun Xie\textsuperscript{a}, Miao Pan\textsuperscript{b} and Jinhui Xu\textsuperscript{d}

\textsuperscript{a}PRADA Lab, King Abdullah University of Science and Technology
\textsuperscript{b}University of Houston
\textsuperscript{c}Rutgers University
\textsuperscript{d}The State University of New York at Buffalo

ORCiD ID: Di Wang https://orcid.org/0000-0003-4908-0243

Abstract. (Gradient) Expectation Maximization (EM) is a widely used algorithm for estimating the maximum likelihood of mixture models or incomplete data problems. A major challenge facing this popular technique is how to effectively preserve the privacy of sensitive data. Previous research on this problem has already lead to the discovery of some Differentially Private (DP) algorithms for (Gradient) EM. However, unlike in the non-private case, existing techniques are not yet able to provide finite sample statistical guarantees. To address this issue, we propose in this paper the first DP version of Gradient EM algorithm with statistical guarantees. Specifically, we first propose a new mechanism for privately estimating the mean of a heavy-tailed distribution, which significantly improves a previous result in [25], and it could be extended to the local DP model, which has not been studied before. Next, we apply our general framework to three canonical models: Gaussian Mixture Model (GMM), Mixture of Regressions Model (MRM) and Linear Regression with Missing Covariates (RMC). Specifically, for GMM in the DP model, our estimation error is near optimal in some cases. For the other two models, we provide the first result on finite sample statistical guarantees. Our theory is supported by thorough numerical experiments on both real-world data and synthetic data.

1 Introduction

As one of the most popular techniques for estimating the maximum likelihood of mixture models or incomplete data problems, Expectation Maximization (EM) algorithm has been widely applied to many areas such as genomics [14], finance [10], and crowdsourcing [7]. EM algorithm is well-known for its convergence to an empirically good local estimator [28]. Recent studies have further revealed that it can also provide finite sample statistical guarantees [3, 33, 27, 31]. Specifically, [3] showed that classical EM and its gradient ascent variant (Gradient EM) are capable of achieving the first local convergence (theory) and finite sample statistical rate of convergence. They also provided a (near) optimal minimax rate for some canonical statistical models such as Gaussian mixture model (GMM), mixture of regressions model (MRM) and linear regression with missing covariates (RMC).

The wide applications of EM also present some new challenges to this method. Particularly, due to the existence of sensitive data and their distributed nature in many applications like social science, biomedicine, and genomics, it is often challenging to preserve the privacy of such data as they are extremely difficult to aggregate and learn from. Consider a case where health records are scattered across multiple hospitals (or even countries), it is not possible to process the whole dataset in a central server due to privacy and ownership concerns. A better solution is to use some differentially private mechanisms to conduct the aggregation and learning tasks. Differential Privacy (DP) [8] is a commonly-accepted criterion that provides provable protection against identification and is resilient to arbitrary auxiliary information that might be available to attackers.

Thus, to be able to use (Gradient) EM algorithm to learn from these sensitive data, it is urgent to design some DP versions of the (gradient) EM algorithm. [18] proposed the first DP EM algorithm which mainly focuses on the practical behaviors of the method. Their algorithm needs quite a few assumptions on the model and the data, which make it difficult to extend to some canonical models mentioned above. Furthermore, unlike the aforementioned non-private case, their algorithm does not provide any finite sample statistical guarantee on the solution. Thus, it is still unknown whether there exists any DP variant of the (gradient) EM algorithm that has finite sample statistical guarantees.

To answer this question, we propose in this paper the first \((\epsilon, \delta)\)-DP (Gradient) EM algorithm with finite sample statistical guarantees. Specifically,

\begin{itemize}
  \item We first show that, given an appropriate initialization \(\beta^{\text{init}}\) \((\text{i.e., } \|\beta^{\text{init}} - \beta^*\|_2 \leq \kappa\|\beta^*\|_2 \text{ for some constant } \kappa \in (0, 1))\), if the model satisfies some additional assumptions and the number of sample \(n\) is large enough, the output \(\beta^{\text{priv}}\) of our DP (Gradient) EM algorithm is guaranteed to have a bounded estimation error, \(\|\beta^{\text{priv}} - \beta^*\|_2 \leq \tilde{O}(\frac{\kappa}{\sqrt{n}})\), with high probability, where \(d\) is the dimensionality and \(\tau\) is an upper bound of the second-order moment of each coordinate of the gradient function. To get the result, we propose a new mechanism for privately estimating the mean of a heavy-tailed distribution, which is based on a finer analysis of the mechanism given by [25]. Moreover, our mechanism could be easily extended to the local privacy model, which is the first result on the problem. Thus, we believe our mechanism could be used in
\end{itemize}
other machine learning problems.

- We then apply our general framework to the three canonical models: GMM, MRM and RMC. Our private estimator achieves an estimation error that is upper bounded by $O\left(\frac{d}{\sqrt{\epsilon n}}\right)$, $O\left(\frac{d^2}{\sqrt{\epsilon n}}\right)$ and $O\left(\frac{d^3}{\sqrt{\epsilon n^2}}\right)$ for GMM, MRM and RMC, respectively. We note that they are the first statistical guarantees for MRM and RMC in the Differential Privacy model, and the error bound for GMM is near optimal in some cases. We also conduct thorough experiments on these three models. Experimental results on these models are consistent with our theoretical analysis.

## 2 Related Work

As we mentioned previously, designing DP version of EM algorithm is still not well studied. To our best knowledge, the only previous work on DP EM algorithm is given by [18]. However, their result is incomparable with ours for the following reasons. Firstly, their work aims to achieve finite sample statistical guarantees for the DP EM algorithm, while [18] mainly focuses on designing heuristic methods. Secondly, it has the same sample complexity. Also, although their algorithm is heavily dependent on a previous clustering algorithm; it is unclear whether it can be extended to other mixture models. From these two perspectives, our framework is more general and practical.

### 3 Preliminaries

Let $Y$ and $Z$ be two random variables taking values in the sample space $\mathcal{Y}$ and $\mathcal{Z}$, respectively. Suppose that the pair $(Y, Z)$ has a joint density function $f_{\beta^*}$ that belongs to some parameterized family $f_{\beta^*}(\beta^* \in \Omega)$. Rather than considering the whole pair of $(Y, Z)$, we observe only component $Y$. Thus, component $Z$ can be viewed as the missing or latent structure. We assume that the term $h_{\beta}(y)$ is the marginal distribution over the latent variable $Z$, i.e., $h_{\beta}(y) = \int_Z f_{\beta}(y, z)dz$. Let $k_{\beta}(z|y)$ be the density of $Z$ conditional on the observed variable $Y = y$, that is, $k_{\beta}(z|y) = \frac{f_{\beta}(y, z)}{h_{\beta}(y)}$.

Given $n$ observations $y_1, y_2, \cdots, y_n$ of $Y$, the EM algorithm is to maximize the log-likelihood $\max_{\beta \in \Omega} \ell_n(\beta) = \sum_{i=1}^n \log h_{\beta}(y_i)$. Due to the unobserved latent variable $Z$, it is often difficult to directly evaluate $\ell_n(\beta)$. Thus, we consider the lower bound of $\ell_n(\beta)$ by Jensen’s inequality, we have

$$-\frac{1}{n} \Delta n(\ell_n(\beta) - \ell_n(\beta')) \geq \frac{1}{n} \sum_{i=1}^n \int_Z k_{\beta'}(z|y_i) \log f_{\beta}(y_i, z)dz - \frac{1}{n} \sum_{i=1}^n \int_Z k_{\beta}(z|y_i) \log f_{\beta'}(y_i, z)dz. \quad (1)$$

Let $Q_n(\beta; \beta') = \frac{1}{2} \sum_{i=1}^n q_i(\beta; \beta')$, where

$$q_i(\beta; \beta') = \int_Z k_{\beta'}(z|y_i) \log f_{\beta}(y_i, z)dz. \quad (2)$$

Also, it is convenient to let $Q_n(\beta; \beta')$ denote the expectation of $Q_n(\beta; \beta')$ w.r.t. $\{y_i\}_{i=1}^n$, that is,

$$Q(\beta; \beta') = \mathbb{E}_{y \sim h_{\beta'}} \int_Z k_{\beta'}(z|y) \log f_{\beta}(y, z)dz. \quad (3)$$

We can see that the second term on the right hand side of (1) is independent on $\beta$. Thus, given some fixed $\beta'$, we can maximize the lower bound function $Q_n(\beta; \beta')$ over $\beta$ to obtain sufficiently large $\ell_n(\beta) - \ell_n(\beta')$. Thus, in the $t$-th iteration of the standard EM algorithm, we can evaluate $Q_n(\beta; \beta')$ at the E-step and then perform the operation of $\beta_{t+1} = \max_{\beta \in \Omega} Q_n(\beta; \beta')$ at the M-step. See [16] for more details.

In addition to the exact maximization implementation of the M-step, we add a gradient ascent implementation of the M-step, which performs an approximate maximization via a gradient descent step.

#### Gradient EM Algorithm [3]

When $Q_n(\cdot; \beta')$ is differentiable, the update of $\beta_t$ to $\beta_{t+1}$ consists of the following two steps.

- **E-step**: Evaluate the functions in (2) to compute $Q_n(\cdot; \beta')$.
- **M-step**: Update $\beta_{t+1} = \beta_t + \eta \nabla Q_n(\beta_t; \beta')$, where $\nabla$ is the derivative of $Q_n$ w.r.t. the first component and $\eta$ is the step size.

Next, we give some examples that use the gradient EM algorithm. Note that they are the typical examples for studying the statistical property of EM algorithm [27, 3, 31, 33].

### Gaussian Mixture Model (GMM)

Let $y_1, \cdots, y_n$ be $n$ i.i.d samples from $Y \in \mathbb{R}^d$ with

$$Y = Z \cdot \beta^* + V, \quad (4)$$

We denote the term $q(\beta; \beta')$ for a general sample $y$.
where $Z$ is a Rademacher random variable (i.e., $\mathbb{P}(Z = +1) = \mathbb{P}(Z = -1) = \frac{1}{2}$), and $V \sim \mathcal{N}(0, \sigma^2 I_d)$ is independent of $Z$ for some known standard deviation $\sigma$.

**Mixture of (Linear) Regressions Model (MMR).** Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be $n$ samples i.i.d sampled from $Y \in \mathbb{R}$ and $X \in \mathbb{R}^d$ with

$$ Y = Z(\beta^*, X) + V, \quad (5) $$

where $X \sim \mathcal{N}(0, I_d)$, $V \sim \mathcal{N}(0, \sigma^2)$, $Z$ is a Rademacher random variable, and $X$, $V$, $Z$ are independent.

**Linear Regression with Missing Covariates (RMC).** We assume that $Y \in \mathbb{R}$ and $X \in \mathbb{R}^d$ satisfy

$$ Y = \langle X, \beta^* \rangle + V, \quad (6) $$

where $X \sim \mathcal{N}(0, I_d)$ and $V \sim \mathcal{N}(0, \sigma^2)$ are independent. Let $x_1, x_2, \ldots, x_n$ be $n$ observations of $X$ with each coordinate of $x_i$ missing (unobserved) independently with probability $p_m \in [0, 1]$.

Next, we provide several definitions on the required properties of functions $Q_n(\cdot; \cdot)$ and $Q(\cdot; \cdot)$.

**Definition 1.** Function $Q(\cdot; \beta^*)$ is self-consistent if $\beta^* = \arg \max_{x \in B} Q(\beta; \beta^*)$. That is, $\beta^*$ maximizes the lower bound of the log likelihood function.

**Definition 2 (Lipschitz-Gradient-2(\gamma, B)).** $Q(\cdot; \cdot)$ is called Lipschitz-Gradient-2(\gamma, B), if for the underlying parameter $\beta^*$ and any $\beta \in B$ for some set $B$, the following holds

$$ \|\nabla Q(\beta; \beta^*) - \nabla Q(\beta^*; \beta^*)\|_2 \leq \gamma \|\beta - \beta^*\|_2. \quad (7) $$

**Definition 3 (\mu-smooth).** $Q(\cdot; \beta^*)$ is $\mu$-smooth, that is if for any $\beta, \beta^* \in B$, $Q(\beta; \beta^*) \geq Q(\beta^*; \beta^*) + (\beta - \beta^*)^T \nabla Q(\beta^*; \beta^*) - \frac{\mu}{2} \|\beta - \beta^*\|_2^2$.

**Definition 4 (v-strongly concave).** $Q(\cdot; \beta^*)$ is $v$-strongly concave, that is if for any $\beta, \beta^* \in B$, $Q(\beta; \beta^*) \leq Q(\beta^*; \beta^*) + (\beta - \beta^*)^T \nabla Q(\beta^*; \beta^*) - \frac{v}{2} \|\beta - \beta^*\|_2^2$.

In the following we will propose the assumptions that will be used throughout the whole paper. Note that these assumptions are commonly used in other works on statistical analysis of EM algorithm such as [2, 33, 27, 24].

**Assumption 1.** We assume that function $Q(\cdot; \cdot)$ in (3) is self-consistent, Lipschitz-Gradient-2(\gamma, B), $\mu$-smooth, $v$-strongly concave over some set $B$. Moreover, we assume that $\forall j \in [d]$ and $\beta \in B$, there is some known upper bound $\tau$ on the second-order moment of the $j$-coordinate of $\nabla q(\beta, \beta)$, i.e., $E_q(\nabla^2 q(\beta, \beta))^2 \leq \tau$ and for each $\beta \in [n]$, $\nabla q(\beta, \beta)$ is independent with others.

**Definition 5 (Differential Privacy [8]).** Given a data universe $\mathcal{X}$, we say that two datasets $D, D' \subseteq \mathcal{X}$ are neighbors if they differ by only one entry, which is denoted as $D \sim D'$. A randomized algorithm $A$ is $(\epsilon, \delta)$-differentially private (DP) if for all neighboring datasets $D, D'$ and for all events $S$ in the output space of $A$, we have $\mathbb{P}(A(D) \in S) \leq e^\epsilon \mathbb{P}(A(D') \in S) + \delta$.

**Definition 6 (Gaussian Mechanism).** Given a function $q : \mathcal{X}^n \rightarrow \mathbb{R}^p$, the Gaussian Mechanism is defined as: $\mathcal{M}_G(D, q, \epsilon) = q(D) + Y$, where $Y$ is drawn from a Gaussian Distribution $\mathcal{N}(0, \sigma^2 I_p)$ with $\sigma \geq \sqrt{2 \ln(2(\epsilon^{-1})/\delta)}/2\epsilon$. $\Delta_2(q)$ is the $L_2$-sensitivity of the function $q$, i.e., $\Delta_2(q) = \sup_{D \sim D'} \|q(D) - q(D')\|_2$. Gaussian Mechanism is $(\epsilon, \delta)$-DP.

Due to the similarity with the Gradient Descent algorithm and the simplicity of illustrating our idea compared with the original EM algorithm, in this paper, we will mainly focus on DP Gradient EM algorithm. See the full version for the statistical guarantees of the DP EM algorithm.

**4 Main Method**

4.1 Main Difficulty

In the previous section, we introduced the Gradient EM algorithm, which updates the estimator via the gradient $\nabla Q_n(\beta; \beta)$. It is notable that this idea is quite similar to the Gradient Descent algorithm. Moreover, we know that there are several DP versions of the (Stochastic) Gradient Descent algorithm such as [4, 26, 15, 19, 21, 29]. The key idea of DP Gradient Descent is adding some randomized noise such as Gaussian noise to preserve DP property in each iteration, and by the composition theorem of DP (\cite{9}), the whole algorithm will still be DP. Thus, motivated by this, to design a DP variant of Gradient EM algorithm, the most direct way is adding some Gaussian noise to the gradient $\nabla Q_n(\beta; \beta)$ in each iteration and updating the parameter.

However, it is notable that we cannot add Gaussian noise directly to the gradient in the Gradient EM algorithm. The main reason is that all previous DP Gradient Descent algorithms need to assume that each component of the gradient (which correspond to the function $\nabla q_i$ in (2)) is bounded, or the loss function is $O(1)$-Lipschitz, such as Logistic Regression, so that its $L_2$-norm sensitivity is bounded and thus the Gaussian mechanism can be used. However, in the Gradient EM algorithm, each component of $\nabla q_i(\beta; \beta)$ in (2) is unbounded in most of the cases. For example, we can easily show the following fact.

**Theorem 1.** Consider the GMM in (4), there is a case with fixed $\beta$, such that for each constant $c$, with positive probability w.r.t. $y$ we have $\|\nabla q(y; \beta)\|_2 \geq c$.

Thus, to design a DP (Gradient) EM algorithm, the major difficulty lies in how to process the gradient to make its sensitivity bounded. Two main approaches are used in previous work: (1) \cite{18} assumed that datasets are pre-processed such that the $L_2$ norm of each sample is bounded by $1$. However, as mentioned previously, our goal is to achieve the statistical guarantees for the DP (Gradient) EM algorithm. If a similar approach is adopted in our algorithm, the (manual) normalization can easily destroy many statistical properties of the data and force the private estimator to introduce additional bias, making it inconsistent.\footnote{An estimator $\beta_n$ is consistent if $\lim_{n \rightarrow \infty} \|\beta_n - \beta^*\|_2 = 0$.} (2) Instead of normalizing the datasets, \cite{1, 30} first clipped the gradient to ensure that the $L_2$-norm of each component of the gradient is bounded by the threshold $C$, and then added Gaussian noise (see Algorithm 1 for more details). However, such an approach may cause two issues. First, in general clipping gradient could introduce additional bias even in statistical estimation, which has also been pointed out in \cite{20}. Second, the threshold $C$ heavily affects the convergence speed and selecting the best $C$ is quite difficult (see Experimental section for more details). Due to these two reasons, it is hard to study the statistical guarantees of Algorithm 1. Thus, we need a new approach to pre-process the gradient to ensure that it has not only bounded $L_2$-norm but also consistent statistical guarantee.
Algorithm 1 Clipped DP Gradient EM

Input: $D = \{y_i\}_{i=1}^n \subset \mathbb{R}^d$, privacy parameters $\epsilon, \delta$, $Q_n(\cdot; \cdot)$ and its $q(\cdot, \cdot)$, initial parameter $\beta_0$, gradient norm $C$, step size $\eta$ and the number of iterations $T$.

1: for $t = 1, 2, \ldots, T$ do
2: For each $i \in [n]$, evaluate the function in (2) to compute $q_i(\beta; \beta^{t-1})$.
3: Clip gradient:
   \[ \nabla \hat{q}_i(\beta^{t-1}; \beta^{t-1}) = \frac{\nabla q_i(\beta^{t-1}; \beta^{t-1})}{\max\{1, |\nabla q_i(\beta^{t-1}; \beta^{t-1})|\}} \]
4: Update $\beta^t = \beta^{t-1} + \eta \nabla \hat{q}_i(\beta^{t-1}; \beta^{t-1}) + N(0, C^2 \sigma^2 I_d)$, where $\nabla \hat{q}_i(\beta^{t-1}; \beta^{t-1}) = \frac{1}{n} \sum_{i=1}^n \nabla \hat{q}_i(\beta^{t-1}; \beta^{t-1})$ and $\sigma^2 = \frac{T \log \frac{2}{\eta}}{n \epsilon^2}$ for some constant $c$.
5: end for
6: Return $\beta^T$

4.2 Our Method

In this section, we will propose our method to overcome the aforementioned difficulties. Since our method is motivated by a robust and private mean estimator for heavy-tailed distributions, which was given in [25, 12, 22], and it is derived from the robust mean estimator in [11]. To be self-contained, we first review their estimator. We now consider a 1-dimensional random variable $x$ and assume that $x_1, x_2, \ldots, x_n$ are i.i.d. sampled from $x$. The estimator consists of three steps:

Scaling and Truncation. For each sample $x_i$, we first re-scale it by dividing $s$ (which will be specified later). Then, the re-scaled one was passed through a soft truncation function $\phi$. Finally, we put the truncated mean back to the original scale. That is,
\[ \hat{x} = \frac{1}{n} \sum_{i=1}^n \phi\left(\frac{x_i}{s}\right) \approx \mathbb{E}(x). \]

Here, we use the function given in [6],
\[ \phi(x) = \begin{cases} x - \frac{x^2}{\sqrt{2}}, & x \leq \sqrt{2} \\ 2x - \frac{x^2}{\sqrt{2}}, & \sqrt{2} < x \leq \sqrt{2} \\ 2x - \frac{x^2}{\sqrt{2}}, & x < -\sqrt{2}. \end{cases} \]

A key property for $\phi$ is that $\phi$ is bounded, that is, $|\phi(x)| \leq 2\sqrt{2}$.

Noise Multiplication. Let $\eta_1, \eta_2, \ldots, \eta_n$ be random noise generated from a common distribution $\eta$ with $\mathbb{E}(\eta) = 0$. We multiply each data $x_i$ by a factor of $1 + \eta_i$, and then perform the scaling and truncation step on the term $x_i(1 + \eta_i)$. That is,
\[ \hat{x}(\eta) = \frac{1}{n} \sum_{i=1}^n \phi\left(\frac{x_i + \eta_i x_i}{s}\right). \]

Noise Smoothing. In this final step, we smooth the multiplicative noise by taking the expectation w.r.t. the distributions. That is,
\[ \hat{x} = \mathbb{E}\hat{x}(\eta) = \frac{1}{n} \sum_{i=1}^n \int \phi\left(\frac{x_i + \eta_i x_i}{s}\right) d\mathbb{P}(\eta_i). \]

Computing the explicit form of each integral in (11) depends on the function $\phi(\cdot)$ and the distribution $\mathbb{P}$. Fortunately, [6] showed that when $\phi$ is in (9) and $\chi \sim N(0, \frac{\beta}{2})$ (where $\beta$ will be specified later), we have for any $a$ and $b > 0$
\[ \mathbb{E}\phi(a + b \sqrt{2} \eta) = a(1 - \frac{\beta^2}{2}) - \frac{a^3}{6} + \mathcal{C}(a, b), \]
where $\mathcal{C}(a, b)$ is a correction form which is easy to implement.

To obtain an $(\epsilon, \delta)$-DP estimator, the key observation is that the bounded function $\phi$ in (9) also makes the integral form of (11) bounded by $\frac{2\sqrt{2}}{\epsilon}$.

Finally, we put the result into the following steps:

1. **Scaling and Truncation.**
   - For each sample $x_i$, we first re-scale it by dividing $s$ (which will be specified later). Then, the re-scaled one was passed through a soft truncation function $\phi$.
   - Finally, we put the truncated mean back to the original scale.

2. **Noise Multiplication.**
   - Let $\eta_1, \eta_2, \ldots, \eta_n$ be random noise generated from a common distribution $\eta$ with $\mathbb{E}(\eta) = 0$. We multiply each data $x_i$ by a factor of $1 + \eta_i$.
   - Then, perform the scaling and truncation step on the term $x_i(1 + \eta_i)$.

3. **Noise Smoothing.**
   - In this final step, we smooth the multiplicative noise by taking the expectation w.r.t. the distributions.

   \[ \hat{x} = \mathbb{E}\hat{x}(\eta) = \frac{1}{n} \sum_{i=1}^n \int \phi\left(\frac{x_i + \eta_i x_i}{s}\right) d\mathbb{P}(\eta_i). \]
Algorithm 2 DP Gradient EM Algorithm

**Input:** $D = \{y_i\}_{i=1}^n \subset \mathbb{R}^d$, privacy parameters $\epsilon, \delta, Q(\cdot)$ and its $q_k(\cdot)$, initial parameter $\beta^0 \in B$, $\tau$ which satisfies Assumption 1, the number of iterations $T$ (to be specified later), step size $\eta$ and failure probability $\zeta > 0$. 

1. Let $\tilde{\epsilon} = \sqrt{\frac{\log \frac{1}{\tau}}{\epsilon}}$, $s = \sqrt{\frac{m\epsilon}{2\log \frac{1}{\tau}}}$, $\beta = \sqrt{\log \frac{1}{\tau}}$. Partite the data $D$ into $T$ subsets $\{D_t\}_{t=1}^T$ with $|D_t| = m = \frac{n}{T}$.

2. For $t = 1, 2, \ldots, T$ do

3. For each $j \in [d]$, calculate the robust gradient by using (11) and add Gaussian noise over the dataset $D_t$, that is

$$g_{t-1}(\beta^{t-1}) = \frac{1}{m} \sum_{i \in D_t} \left( \nabla_j q_i(\beta^{t-1}, \beta^{t-1})(1 - \nabla_j q_i(\beta^{t-1}, \beta^{t-1}) - \nabla_j q_i(\beta^{t-1}, \beta^{t-1})) \right)$$

$$+ s \sum_{i \in D_t} \frac{C}{\sqrt{\beta}} \left( \nabla_j q_i(\beta^{t-1}, \beta^{t-1}), \nabla_j q_i(\beta^{t-1}, \beta^{t-1}) \right) + Z_{t-1}^{j-1}, \quad (14)$$

where $y_i \in D_t$ for $i \in [m]$, $Z_{t-1}^{j-1} \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = \frac{\kappa \epsilon m}{2\log \frac{1}{\tau}} = \frac{4\tau T^2}{m \log \frac{1}{\tau}}$.

4. Let vector $\tilde{\nabla}Q_{\alpha}(\beta^{t-1}) \in \mathbb{R}^d$ denote $\tilde{\nabla}Q_{\alpha}(\beta^{t-1}) = (g_{t-1}^{(1)}(\beta^{t-1}), g_{t-1}^{(2)}(\beta^{t-1}), \ldots, g_{t-1}^{(d)}(\beta^{t-1}))$.

5. Update $\beta^t = \beta^{t-1} + \frac{\eta}{T} \tilde{\nabla}Q_{\alpha}(\beta^{t-1})$.

6. end for

similar analysis, we can have a local DP version of (13) with an error bound of $O\left(\sqrt{\frac{\tau \log 1/2}{\epsilon}} \frac{1}{\sqrt{m}} \sqrt{\log \frac{1}{\tau}} \right)$. To our best knowledge, this is the first result on private mean estimation of heavy-tailed distribution in the local DP model.

Inspired by the previous private 1-dimensional mean estimation, we propose our method (Algorithm 2). In Algorithm 2, the key idea is that, in the $t$-th iteration of Gradient EM algorithm, we first apply the previous private estimator to each coordinate of the gradient $\nabla Q_{\alpha}(\beta^{t-1}; \beta^{t-1})$, and then perform the $M$-step.

**Theorem 3** (Privacy guarantee). For any $0 < \epsilon, \delta < 1$, Algorithm 2 is $(\epsilon, \delta)$-DP.

**Theorem 4.** Let the parameter set $B = \{\beta : \|\beta - \beta^*\|_2 \leq R\}$ for $R = \kappa\|\beta^*\|_2$ for some constant $\kappa \in (0, 1)$. Assume that Assumption 1 holds for parameters $\gamma, B, \beta, v, \tau$ satisfying the condition of $1 - \frac{2 \gamma \epsilon}{m \log \frac{1}{\tau}} \in (0, 1)$. Also, assume that $\|\beta^0 - \beta^*\|_2 \leq \frac{R}{\eta}$, $n$ is large enough so that

$$\tilde{\Omega}(\frac{1}{v - \gamma}) \frac{d^2 \tau T \sqrt{\log \frac{1}{\tau} \log \frac{1}{\tau}}}{\epsilon R^2} \leq n. \quad (17)$$

Then, with probability at least $1 - \zeta$, we have for all $t \in [T], \beta^t \in B$. If it holds and if taking $T = O(\frac{\log \frac{1}{\tau} \log \frac{1}{\tau}}{\epsilon})$ and $\eta = \frac{2}{\tau \sqrt{n}}$, we have

$$\|\beta^T - \beta^*\|_2 \leq \tilde{\Omega}(R \frac{d^2 \sqrt{\log \frac{1}{\tau} \log \frac{1}{\tau}}}{\sqrt{\epsilon} R^2}), \quad (18)$$

where the $\tilde{O}$-term and $\tilde{\Omega}$-term omit $\log d, \log n$ and other factors (see Appendix for the explicit form of the result).

**Remark 1.** There are several points that need to note. Firstly, the assumptions of the parameter set $\beta$ and the initial parameter $\beta^0$ are commonly used in other papers on statistical guarantees of (Gradient) EM algorithm such as [2, 33, 27]. Even though Theorem 4 requires that the initial estimator be close enough to the optimal one, our experiments show that the algorithm actually performs quite well for any random initialization. Secondly, in (17) we need to assume that $n \propto \frac{1}{\tau^2}$, where $R$ is the radius of $B$. This is due to that in Algorithm 2, we need to keep each $\beta^t \in B$ under perturbation. When $R$ is small, we have to let the noise be small enough, which means that $n$ should be large enough. Finally, for specific models, $R, v, \mu, \gamma$ are constants, this means that the error in (18) is $\tilde{O}(\frac{1}{\sqrt{n} \tau})$. However, here $\tau$ depends on the model, which may also depend on $d$ and $\|\beta^*\|_2$.

### 5 Implications for Some Specific Models

In this section, we apply our framework (i.e., Algorithm 2) to the models mentioned in the Preliminaries section. To obtain results for these models, we only need to find the corresponding $B, \gamma, k, R, v, \mu, \tau$ to ensure that Assumption 1 and the assumptions in Theorem 4 hold. Due to the space limit, the results of RMC are included in the full version.

#### 5.1 Gaussian Mixture Model

**Lemma 2** ([3, 31]). If $\|\beta^0\|_2 \geq r$, where $r$ is a sufficiently large constant denoting the minimum signal-to-noise ratio (SNR), then there exists an absolute constant $C > 0$ such that the properties of self-consistent, Lipschitz-Gradient-2$(\gamma, B)$, $\nu$-smoothness and $\nu$-strongly concave hold for function $Q(\cdot|\cdot)$ with $\gamma = \text{exp}(-Cr^2), \nu = \frac{1}{v}, R = k\|\beta^*\|_2, k = \frac{1}{2}$, and $B = \{\beta : \|\beta - \beta^*\|_2 \leq R\}$.

**Lemma 3.** With the same notations as in Lemma 2, for each $\beta \in B$, the $j$-th coordinate of $\nabla q(\beta; \beta)$ (i.e., $\nabla_j q(\beta; \beta)$) satisfies the following inequality

$$E_v(\nabla_j q(\beta; \beta))^2 \leq O((\|\beta^*\|_2^2 + \sigma^2)).$$

Also, for fixed $j \in [d], \nabla_j q(\beta; \beta)$, where $i \in [n]$, is independent with others.

**Theorem 5.** With the same notations as in Lemma 2, in Algorithm 2 assume that $\|\beta^0 - \beta^*\|_2 \leq \frac{1}{\sqrt{\kappa}}\|\beta^*\|_2$ and $n$ is large enough so that

$$\tilde{\Omega}(\frac{d^2 \sqrt{\|\beta^*\|_2^2 + \sigma^2 \log \frac{1}{\tau} \log \frac{1}{\tau}}}{\sqrt{\epsilon} \|\beta^*\|_2}) \leq \frac{n}{\tau} \tilde{\Omega}(\frac{d^2 \sqrt{\|\beta^*\|_2^2 + \sigma^2 \log \frac{1}{\tau} \log \frac{1}{\tau}}}{\sqrt{\epsilon} \|\beta^*\|_2}). \quad (19)$$

Moreover, if take $T = O(\log n)$ and $\eta = O(1)$, then we have with probability at least $1 - \zeta$

$$\|\beta^T - \beta^*\|_2^2 \leq \tilde{O}(\|\beta^*\|_2^2 \frac{d^2 \sqrt{\log \frac{1}{\tau} \log \frac{1}{\tau}} \sqrt{\|\beta^*\|_2^2 + \sigma^2}}{\sqrt{\epsilon} \|\beta^*\|_2}). \quad (20)$$
5.2 Mixture of Regressions Model

Lemma 4 ([3, 31]). If $\|\beta^*\|_2 = O(1)$, then the error in (20) is upper bounded by $O\left(\frac{d}{\sqrt{m}}\right)$. This means that to achieve the error $\alpha$ is $O(1)$, the sample complexity is $O\left(\frac{d}{\sqrt{\alpha m}}\right)$. It is notable that for GMM, the near optimal rate is $O(d^2(\frac{1}{\sqrt{\epsilon}} + \frac{1}{\epsilon}))$ [13]. Thus when $\epsilon$ is some constant, our result matches their near optimal rate. However, as mentioned in previous section, their algorithm has extremely large hidden constants in their parameters and thus is impractical and it is difficult to extend their method to other mixture models.

6 Experiments

In this section, we evaluate the performance of Algorithm 2 on three canonical models: GMM, MRM, and RMC. We evaluate our algorithm on both the synthetic data and the real world datasets: ADULT, IPUMS-BR, and IPUMS-US.

Baseline Methods. As we mentioned in the related work section, [18] only provides heuristic methods without any finite sample statistical guarantees, and its method cannot be applied to our models (i.e., using their method to our models cannot guarantee DP). Thus, we will not compare with their method. [32] needs strong assumptions on the statistical guarantee and thus it is incomparable with our work. Thus, here we compare our approach against two baseline algorithms. One is the Gradient EM algorithm [3], namely, EM, as our non-private baseline method. The other is clipped DP Gradient EM (Algorithm 1), namely, clipped, as our private baseline method.

Experimental Results. Firstly, we will show that the performance of Algorithm 1 is heavily affected by the clipping threshold $C$. As shown in Figure 1, we conduct the algorithm on three canonical models with fixed data size $n$, dimension data $d$, and privacy budget $\epsilon$. If $C$ is set to be a small value (e.g., 0.1), it significantly reduces the adding noise in each iteration but at the same time it leads much information loss in gradient estimation. Conversely, if $C$ is set too high (e.g., 5 or 10), the noise variance becomes high, resulting in introducing too much noise to the estimation. Thus, selecting the optimal $C$ is quite difficult since too large or too small values of $C$ has a negative effect on the performance of Algorithm 1. Even for $C = 1$ that achieves lowest estimation error among other threshold values, the estimation error does not decay as the number of iterations increases, whereas under the same privacy guarantee, our proposed algorithm achieves the same convergence behavior as EM, and thoroughly outperforms Algorithm 1. For fair comparison, we fixed $C = 1$ for Algorithm 1 in the following experiments.

In Figure 2, 3 and 4, we test how privacy budget $\epsilon$, data dimension $d$ and data size $n$ affect the estimation error $\|\beta - \beta^*\|_2$ of all algorithms on three canonical models over iteration $t$. We can see that the estimation error of our proposed algorithm in each of the three models decreases when $\epsilon$ increases, $d$ decreases or $n$ increases, which are consistent with our theoretical results. In these figures, our algorithm exhibits nearly the same convergence behavior as the non-private baseline method and outperforms Algorithm 1.

We further present the estimation error of different algorithms on GMM model over three real world datasets, as shown in Figure 5. We can observe that our proposed algorithm still outperforms the baseline algorithms under different privacy budgets.

7 Conclusion

We provided the first study on the finite sample statistical guarantees of (Gradient) EM algorithm in the Differential Privacy (DP) model. Previous DP Gradient Descent based methods cannot be directly extended to the Gradient EM algorithm. We proposed a new and improved private algorithm for estimating the mean of heavy-tailed distributions, which could also be extended to the local DP model. We also implemented our algorithms to several canonical latent variable models. Finally, we conducted extensive experiments on both of the synthetic and real-world data, and these results outperform previous heuristic methods and show the effectiveness of our algorithm.

![Figure 1. Estimation error of Algorithm 1 (clipped) v.s. iteration $t$ under different clipping threshold $C$](http://example.com/image1.png)
Figure 2. Estimation error of GMM w.r.t. privacy budget $\epsilon$, data dimension $d$, data size $n$ and iteration $t$

Figure 3. Estimation error of MRM w.r.t. privacy budget $\epsilon$, data dimension $d$, data size $n$ and iteration $t$

Figure 4. Estimation error of RMC w.r.t. privacy budget $\epsilon$, data dimension $d$, data size $n$ and iteration $t$

Figure 5. Estimation error of GMM over three real datasets: ADULT, IPUMS-US and IPUMS-BR