Finite Sample Guarantees of Differentially Private Expectation Maximization Algorithm

Di Wang[a;*], **Jiahao Ding**[b], **Lijie Hu**[a], **Zejun Xie**[c], **Miao Pan**[b] **and Jinhui Xu**[d]

[a]PRADA Lab, King Abdullah University of Science and Technology
[b]University of Houston
[c]Rutgers University
[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).

* Corresponding Author. Email: di.wang@kaust.edu.sa. The first two authors contributed to this paper equally. The full version of the paper could be found at [23].

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 (ϵ, δ)-DP (Gradient) EM algorithm with finite sample statistical guarantees. Specifically,

- We first show that, given an appropriate initialization β^{init} (*i.e.,* $\|\beta^{\text{init}} - \beta^*\|_2 \leq \kappa\|\beta^*\|_2$ 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 β^{priv} of our DP (Gradient) EM algorithm is guaranteed to have a bounded estimation error, $\|\beta^{priv} - \beta^*\|_2 \leq \tilde{O}(\frac{d\sqrt{\tau}}{\sqrt{n}\epsilon})$, with high probability, where d is the dimensionality and τ 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 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 $\tilde{O}(\frac{d}{\sqrt{n\epsilon}})$, $\tilde{O}(\frac{d^{\frac{3}{2}}}{\sqrt{n\epsilon}})$ and $\tilde{O}(\frac{d^{\frac{3}{2}}}{\sqrt{n\epsilon}})$ 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, our work aims to achieve finite sample statistical guarantees for the DP EM algorithm, while [18] mainly focuses on designing heuristic DP EM algorithms that do not provide any statistical guarantees. Particularly, [18] assumed that datasets are pre-processed such that the ℓ_2-norm of each data record is less than 1. This means that their algorithm will likely introduce additional bias on the statistical guarantees. Secondly, the assumptions made in [18] regarding the model's sufficient statistics and their sensitivity are not applicable to all three fundamental models studied in our experiments. Specifically, [18] studied only the exponential family so that noise can be directly added to the sufficient statistics. However, most of the latent variable models do not satisfy such an assumption. This includes the MRM and RMC models to be considered in this paper.

Concurrently, [32] also studied theoretical guarantees of the DP-EM algorithm in the high dimensional sparse setting. However, there are significant differences between their work and ours. First, their method requires two strong assumptions, namely Condition 3.6 and 3.7. Condition 3.6 assumes that $\|\nabla q_i(\beta;\beta) - \mathbb{E}\nabla q(\beta;\beta)\|_\infty$ is bouned with high probability, while in our paper we only need to assume $\mathbb{E}(\nabla_j q(\beta;\beta))^2$ is bounded, i.e., their assumption implies ours, making our assumption more general. Additionally, their Condition 2.7 imposes a peculiar assumption that we do not require in our paper. Secondly, Due to their strong assumption, their algorithm merely truncates the ∇q_i functions and adds noise, whereas our paper utilizes a more complex method to handle our relaxed Assumption 1. Thirdly, their algorithm only applies to the central DP model, and it is unknown if it can be extended to the local DP model. On the other hand, in our paper, we can extend all the methods to the local DP model. Due to these reasons their results are incomparable with ours.

In this paper, we implement our general framework on three specific models, and DP GMM is the only one that has been studied previously. Specifically, [17] provided the first result for the general k-GMM based on the sample-and-aggregate framework. However, their algorithms are impractical as it has been shown that the sample-and-aggregate framework has poor practical performance previously. Later on, [13] improved the result by a factor of $O(\sqrt{d}/\epsilon)$, and also claimed that their sample complexity is near optimal. Compared with their result, our proposed algorithm ensures that when ϵ is some constant, it has the same sample complexity. Also, although their algorithm has polynomial time complexity, it is actually not very practical and thus no practical study has been conducted. Moreover, 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 spaces $\mathcal{Y}$ and $\mathcal{Z}$, respectively. Suppose that the pair (Y,Z) has a joint density function f_{β^*} 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.,* $h_\beta(y)=\int_{\mathcal{Z}} f_\beta(y,z)dz$. 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

$$\frac{1}{n}[\ell_n(\beta)-\ell_n(\beta')] \geq \frac{1}{n}\sum_{i=1}^n\int_{\mathcal{Z}} k_{\beta'}(z|y_i)\log f_\beta(y_i,z)dz \\ -\frac{1}{n}\sum_{i=1}^n\int_{\mathcal{Z}} k_{\beta'}(z|y_i)\log f_{\beta'}(y_i,z)dz. \tag{1}$$

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

$$q_i(\beta;\beta')=\int_{\mathcal{Z}} k_{\beta'}(z|y_i)\log f_\beta(y_i,z)dz.^1 \tag{2}$$

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

$$Q(\beta;\beta')=\mathbb{E}_{y\sim h_{\beta^*}}\int_{\mathcal{Z}} k_{\beta'}(z|y)\log f_\beta(y,z)dz. \tag{3}$$

We can see that the second term on the right hand side of (1) is independent on β. Thus, given some fixed β', we can maximize the lower bound function $Q_n(\beta;\beta')$ over β to obtain sufficiently large $\ell_n(\beta)-\ell_n(\beta')$. Thus, in the t-th iteration of the standard EM algorithm, we can evaluate $Q_n(\cdot;\beta^t)$ at the E-step and then perform the operation of $\beta^{t+1}=\max_{\beta\in\Omega}Q_n(\beta;\beta^t)$ 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 $Q_n(\cdot;\beta^t)$ is differentiable, the update of β^t to β^{t+1} consists of the following two steps.

- E-step: Evaluate the functions in (2) to compute $Q_n(\cdot;\beta^t)$.
- M-step: Update $\beta^{t+1}=\beta^t+\eta\nabla Q_n(\beta^t;\beta^t)$, where ∇ is the derivative of Q_n w.r.t. the first component and η 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

$$Y=Z\cdot\beta^*+V, \tag{4}$$

[1] 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 σ.

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

$$Y=Z\langle\beta^*,X\rangle+V, \tag{5}$$

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

$$Y=\langle X,\beta^*\rangle+V, \tag{6}$$

where $X\sim\mathcal{N}(0,I_d)$ and $V\sim\mathcal{N}(0,\sigma^2)$ are independent. Let $x_1,x_2,\cdots,x_n$ 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 $\beta^*=\arg\max_{\beta\in\Omega}Q(\beta;\beta^*)$. That is, β^* maximizes the lower bound of the log likelihood function.

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

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

Definition 3 (μ-smooth)**.** $Q(\cdot;\beta^*)$ is μ-smooth, that is if for any $\beta,\beta'\in\mathcal{B}$, $Q(\beta;\beta^*)\geq Q(\beta';\beta^*)+(\beta-\beta')^T\nabla Q(\beta';\beta^*)-\frac{\mu}{2}\|\beta'-\beta\|_2^2$.

Definition 4 (υ-strongly concave)**.** $Q(\cdot;\beta^*)$ is υ-strongly concave, that is if for any $\beta,\beta'\in\mathcal{B}$, $Q(\beta;\beta^*)\leq Q(\beta';\beta^*)+(\beta-\beta')^T\nabla Q(\beta';\beta^*)-\frac{\upsilon}{2}\|\beta'-\beta\|_2^2$.

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($\gamma,\mathcal{B}$), μ-smooth, υ-strongly concave over some set $\mathcal{B}$. Moreover, we assume that $\forall j\in[d]$ and $\beta\in\mathcal{B}$, there is some known upper bound τ on the second-order moment of the j-coordinate of $\nabla q(\beta,\beta)$, *i.e.,* $\mathbb{E}_y(\nabla_j q(\beta,\beta))^2\leq\tau$ and for each $i\in[n]$, $\nabla_j q_i(\beta,\beta)$ 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 $\mathcal{A}$ is (ϵ,δ)-differentially private (DP) if for all neighboring datasets D,D' and for all events S in the output space of $\mathcal{A}$, we have $\mathbb{P}(\mathcal{A}(D)\in S)\leq e^\epsilon\mathbb{P}(\mathcal{A}(D')\in S)+\delta$.

Definition 6 (Gaussian Mechanism)**.** Given a function $q:\mathcal{X}^n\to\mathbb{R}^p$, 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 $\sigma\geq\frac{\sqrt{2\ln(1.25/\delta)}\Delta_2(q)}{\epsilon}$. $\Delta_2(q)$ 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 (ϵ,δ)-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 $\nabla Q_n(\beta^t;\beta^t)$. 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 ([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 $\nabla Q_n(\beta^t;\beta^t)$ 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 ∇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 ($\nabla q_i(\beta^t;\beta^t)$ 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 β, such that for each constant c, with* **positive probability** *w.r.t. y we have $\|\nabla q(\beta;\beta)\|_2\geq c$.*

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) [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.[2] (2) Instead of normalizing the datasets, [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 [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.

[2] An estimator β_n is consistent if $\lim_{n\to\infty}\|\beta_n-\beta^*\|_2=0$.

Algorithm 1 Clipped DP Gradient EM

Input: $D = \{y_i\}_{i=1}^n \subset \mathbb{R}^d$, privacy parameters ϵ, δ; $Q_n(\cdot;\cdot)$ and its $q(\cdot;\cdot)$, initial parameter β^0, gradient norm C, step size η and the number of iterations T.

1: **for** $t = 1, 2, \cdots, T$ **do**

2: For each $i \in [n]$, evaluate the function in (2) to compute $q_i(\beta; \beta^{t-1})$.

3: Clip gradient:

$$\nabla \bar{q}_i(\beta^{t-1}; \beta^{t-1}) = \frac{\nabla q_i(\beta^{t-1}; \beta^{t-1})}{\max\{1, \frac{\|\nabla q_i(\beta^{t-1}; \beta^{t-1})\|_2}{C}\}}.$$

4: Update $\beta^t = \beta^t + \eta(\nabla \bar{Q}_n(\beta^{t-1}; \beta^{t-1}) + \mathcal{N}(0, C^2\sigma^2 I_d)$, where $\nabla \bar{Q}_n(\beta^{t-1}; \beta^{t-1})) = \frac{1}{n}\sum_{i=1}^n \nabla \bar{q}_i(\beta^{t-1}; \beta^{t-1})$ and $\sigma^2 = c\frac{T \log \frac{1}{\delta}}{n^2\epsilon^2}$ for some constant c.

5: **end for**

6: Return β^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 $x_1, x_2, \cdots, x_n$ 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 ϕ. Finally, we put the truncated mean back to the original scale. That is,

$$\frac{s}{n}\sum_{i=1}^n \phi(\frac{x_i}{s}) \approx \mathbb{E}(x). \tag{8}$$

Here, we use the function given in [6],

$$\phi(x) = \begin{cases} x - \frac{x^3}{6}, & -\sqrt{2} \le x \le \sqrt{2} \\ \frac{2\sqrt{2}}{3}, & x > \sqrt{2} \\ -\frac{2\sqrt{2}}{3}, & x < -\sqrt{2}. \end{cases} \tag{9}$$

A key property for ϕ is that ϕ is bounded, that is, $|\phi(x)| \le \frac{2\sqrt{2}}{3}$.

Noise Multiplication. Let $\eta_1, \eta_2, \cdots, \eta_n$ be random noise generated from a common distribution $\eta \sim \chi$ 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,

$$\tilde{x}(\eta) = \frac{s}{n}\sum_{i=1}^n \phi(\frac{x_i + \eta_i x_i}{s}). \tag{10}$$

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

$$\hat{x} = \mathbb{E}\tilde{x}(\eta) = \frac{s}{n}\sum_{i=1}^n \int \phi(\frac{x_i + \eta_i x_i}{s}) d\chi(\eta_i). \tag{11}$$

Computing the explicit form of each integral in (11) depends on the function $\phi(\cdot)$ and the distribution χ. Fortunately, [6] showed that when ϕ is in (9) and $\chi \sim \mathcal{N}(0, \frac{1}{\beta})$ (where β will be specified later), we have for any a and $b > 0$

$$\mathbb{E}_\eta \phi(a + b\sqrt{\beta}\eta) = a(1 - \frac{b^2}{2}) - \frac{a^3}{6} + \hat{C}(a, b), \tag{12}$$

where $\hat{C}(a, b)$ is a correction form which is easy to implement.

To obtain an (ϵ, δ)-DP estimator, the key observation is that the bounded function ϕ in (9) also makes the integral form of (11) bounded by $\frac{2\sqrt{2}}{3}$. Thus, we know that the ℓ_2-norm sensitivity is $\frac{s}{n}\frac{4\sqrt{2}}{3}$. Hence, the query

$$\mathcal{A}(D) = \hat{x} + Z, Z \sim \mathcal{N}(0, \sigma^2), \sigma^2 = O(\frac{s^2 \log \frac{1}{\delta}}{\epsilon^2 n^2}) \tag{13}$$

will be (ϵ, δ)-DP, which leads to the following result.

Lemma 1 (Theorem 6 in [25]). Let $x_1, x_2, \cdots, x_n$ be i.i.d. samples from distribution $x \sim \mu$. Assume that there is some known upper bound on the second-order moment, *i.e.,* $\mathbb{E}_\mu x^2 \le \tau$. For a given failure probability ζ, if set $\beta = 2\log\frac{1}{\zeta}$ and $s = \sqrt{\frac{n\tau}{2\log\frac{1}{\zeta}}}$, with probability at least $1 - \zeta$ the following holds

$$|\mathcal{A}(D) - \mathbb{E}(x)| \le O(\sqrt{\frac{\tau \log\frac{1}{\delta}\log\frac{1}{\zeta}}{n\epsilon^2}}). \tag{15}$$

Although in Lemma 1 we just need to assume that x has bounded second order moment instead of bounded norm, there are still other two problems: First, Lemma 1 is directly followed by a result in [11] with the same parameter s and β. However, due to the noise we add, is it possible that we can further improve the result by some other specific s and β? Second, by using the previous parameters we cannot extend to the local DP model since it will have a huge error (we can easily see that in the local DP setting, the mechanism is equivalent to (13) with $\sigma^2 = O(\frac{s^2 \log\frac{1}{\delta}}{n\epsilon^2}) = O(\frac{\tau}{\epsilon^2})$, which could be considered as a constant error since it is not decayed to zero when n increases. Thus, can we extend the method to the local DP model? In the following we provide affirmative answer of these two questions through finer analysis of the mechanism (13).

Theorem 2. Let $x_1, x_2, \cdots, x_n$ be i.i.d. samples from distribution $x \sim \mu$. Assume that there is some known upper bound on the second-order moment, *i.e.,* $\mathbb{E}_\mu x^2 \le \tau$. For a given failure probability ζ, if set $\beta = \sqrt{\log\frac{1}{\zeta}}$ and $s = \frac{\sqrt{n\epsilon\tau}}{\log\frac{1}{\zeta}\log^{1/4}\frac{1}{\delta}}$, then with probability at least $1 - \zeta$ mechanism (13) satisfies

$$|\mathcal{A}(D) - \mathbb{E}x| \le O(\sqrt{\frac{\tau \log^{1/2}\frac{1}{\delta}\log\frac{1}{\zeta}}{n\epsilon}}). \tag{16}$$

Comparison with [25]. Although our private estimator has a similar form as the one in [25]. From Theorem 2, we can see there are several critical differences. (1) We have provided a more refined analysis of the estimator and showed that with some specific parameters, we can get an improved upper bound. Specifically, compared with (15) given by [25], we can see the parameter s in Theorem 2 depends on n, ϵ and τ, where s in (15) only depends on n and τ. This is due to different trade-offs between the bias, variance in the estimation error and the noises we added. Moreover, we can see the error bound in (16) improves a factor of $O(\frac{1}{\sqrt{\epsilon}})$. The theoretical analysis on the trade-offs is non-trivial, which is started from a Legendre transform of the mapping given by [5]. (2) We will also see that, by using a

Algorithm 2 DP Gradient EM Algorithm

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

Let $\tilde{\epsilon} = \sqrt{\log\frac{1}{\delta} + \epsilon} - \sqrt{\log\frac{1}{\delta}}$, $s = \frac{\sqrt{m\tau\tilde{\epsilon}}}{2\log\frac{d}{\zeta}}$, $\beta = \sqrt{\log\frac{d}{\zeta}}$. Partite the data D into T subsets $\{D_i\}_{i=1}^T$ with $|D_i| = m = \frac{n}{T}$.

for $t = 1, 2, \cdots, T$ **do**

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

$$g_j^{t-1}(\beta^{t-1}) = \frac{1}{m}\sum_{i\in D_t}\left(\nabla_j q_i(\beta^{t-1}, \beta^{t-1})\big(1 - \frac{\nabla_j^2 q_i(\beta^{t-1}, \beta^{t-1})}{2s^2\beta}\big) - \frac{\nabla_j^3 q_i(\beta^{t-1}, \beta^{t-1})}{6s^2}\right) + \frac{s}{m}\sum_{i\in D_t}\hat{C}\left(\frac{\nabla_j q_i(\beta^{t-1}, \beta^{t-1})}{s}, \frac{|\nabla_j q_i(\beta^{t-1}, \beta^{t-1})|}{s\sqrt{\beta}}\right) + Z_j^{t-1}, \quad (14)$$

where $y_i \in D_t$ for $i \in [m]$, $Z_j^{t-1} \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = \frac{16s^2d}{9m^2\tilde{\epsilon}^2} = \frac{4dT\tau}{9n\beta^2\tilde{\epsilon}}$.

Let vector $\tilde{\nabla}Q_n(\beta^{t-1}) \in \mathbb{R}^d$ denote $\tilde{\nabla}Q_n(\beta^{t-1}) = (g_1^{t-1}(\beta^{t-1}), g_2^{t-1}(\beta^{t-1}), \cdots, g_d^{t-1}(\beta^{t-1}))$.

Update $\beta^t = \beta^{t-1} + \eta\tilde{\nabla}Q_n(\beta^{t-1})$.

end for

similar analysis, we can have a local DP version of (13) with an error bound of $O(\sqrt{\frac{\tau\log^{1/2}\frac{1}{\delta}\log\frac{1}{\zeta}}{\sqrt{n}\epsilon}})$. 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 $\nabla Q_n(\beta^{t-1}; \beta^{t-1})$, and then perform the M-step.

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

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

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

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

$$\|\beta^T - \beta^*\|_2 \leq \tilde{O}\big(R\sqrt{\frac{v+\mu}{(v-\gamma)^3}}\frac{d\sqrt[4]{\log\frac{1}{\delta}}\log\frac{1}{\zeta}\sqrt{\tau}}{\sqrt{n}\epsilon}\big), \quad (18)$$

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 β and the initial parameter β^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 $n \propto \frac{1}{R^2}$, where R is the radius of $\mathcal{B}$. This is due to that in Algorithm 2, we need to keep each $\beta^t \in \mathcal{B}$ 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, μ, γ are constants, this means that the error in (18) is $\tilde{O}(\frac{d\sqrt{\tau}}{\sqrt{n}\epsilon})$. However, here τ 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 $\mathcal{B}, \gamma, k, R, v, \mu, \tau$ 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 $\frac{\|\beta^*\|_2}{\sigma} \geq r$, 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$(\gamma, \mathcal{B})$, μ-smoothness and v-strongly concave hold for function $Q(\cdot;\cdot)$ with $\gamma = \exp(-Cr^2), \mu = v = 1, R = k\|\beta^*\|_2, k = \frac{1}{4}$, and $\mathcal{B} = \{\beta : \|\beta - \beta^*\|_2 \leq R\}$.

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

$$\mathbb{E}_y(\nabla_j q(\beta;\beta))^2 \leq O((\|\beta^*\|_\infty^2 + \sigma^2)).$$

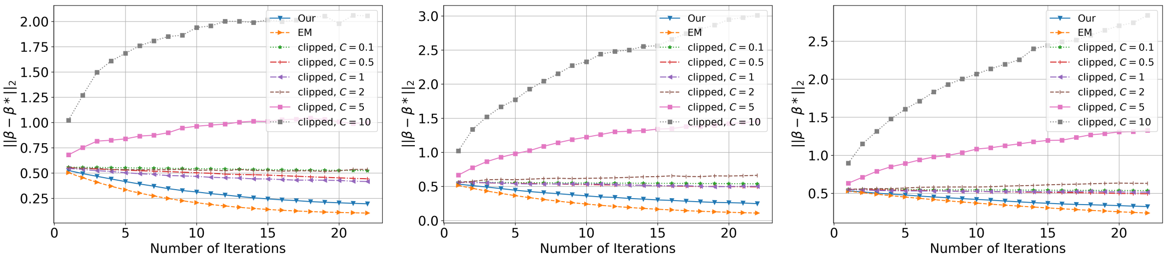
Also, for fixed $j \in [d]$, each $\nabla_j q_i(\beta;\beta)$, where $i \in [n]$, is independent with others.

Theorem 5. With the same notations as in Lemma 2, in Algorithm 2 assume that $\|\beta^0 - \beta^*\|_2 \leq \frac{1}{8}\|\beta^*\|_2$ and n is large enough so that

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

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

$$\|\beta^T - \beta^*\|_2 \leq \tilde{O}\big(\|\beta^*\|_2\frac{d\sqrt[4]{\log\frac{1}{\delta}}\log\frac{1}{\zeta}\sqrt{\|\beta^*\|_\infty^2 + \sigma^2}}{\sqrt{n}\epsilon}\big). \quad (20)$$



(a) GMM, $n = 1000, d = 20, \epsilon = 0.2$ (b) MRM, $n = 1000, d = 20, \epsilon = 0.2$ (c) RMC, $n = 1000, d = 20, \epsilon = 0.2$

Figure 1. Estimation error of Algorithm 1 (clipped) v.s. iteration t under different clipping threshold C

Remark 2. Note that if we assume that $\sigma, \|\beta^*\|_2 = O(1)$, then the error in (20) is upper bounded by $\tilde{O}(\frac{d}{\sqrt{n\epsilon}})$. This means that to achieve the error of $\alpha \in (0,1)$, the sample complexity is $\tilde{O}(\frac{d^2}{\alpha^2\epsilon})$. It is notable that for GMM, the near optimal rate is $\tilde{O}(d^2(\frac{1}{\alpha^2} + \frac{1}{\alpha\epsilon})$ [13]. Thus when ϵ 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.

5.2 *Mixture of Regressions Model*

Lemma 4 ([3, 31])**.** If $\frac{\|\beta^*\|_2}{\sigma} \geq r$, where r is a sufficiently large constant denoting the required minimal signal-to-noise ratio (SNR), then function $Q(\cdot;\cdot)$ of the Mixture of Regressions Model has the properties of self-consistent, Lipschitz-Gradient-2$(\gamma, \mathcal{B})$, μ-smoothness, and υ-strongly with $\gamma \in (0, \frac{1}{4}), \mu = \upsilon = 1, \mathcal{B} = \{\beta : \|\beta - \beta^*\|_2 \leq R\}, R = k\|\beta^*\|_2$, and $k = \frac{1}{32}$.

Theorem 6. With the same notations as in Lemma 4, in Algorithm 2 assume that $\|\beta^0 - \beta^*\|_2 \leq \frac{1}{64}\|\beta^*\|_2$ and n is large enough so that

$$\tilde{\Omega}(\frac{d^2 \max\{(\|\beta^*\|_2^2 + \sigma^2)^2, d\|\beta^*\|_2^2\}\sqrt{\log\frac{1}{\delta}}\log\frac{1}{\zeta}}{\epsilon\|\beta^*\|_2^2}) \leq n.$$

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

$$\|\beta^T - \beta^*\|_2 \leq \tilde{O}(\frac{d\|\beta^*\|_2 \sqrt[4]{\log\frac{1}{\delta}}\sqrt{\max\{\|\beta^*\|_2^2 + \sigma^2, d\|\beta^*\|_2^2\}}}{\sqrt{n\epsilon}}).$$

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[3]: 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 ϵ. 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.

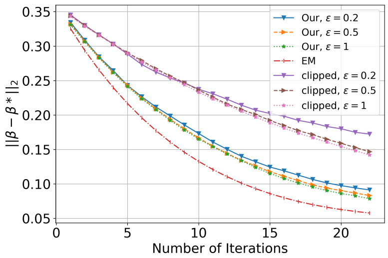
In Figure 2, 3 and 4, we test how privacy budget ϵ, data dimension d and data size n affect the estimation error $\|\beta - \beta^*\|_2$ of all algorithms on three canonical models over iteration t. We can see that the estimation error of our proposed algorithm in each of the three models decreases when ϵ 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.

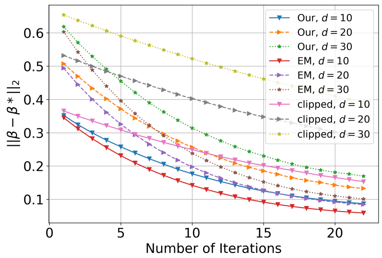
7 Conclusion

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.

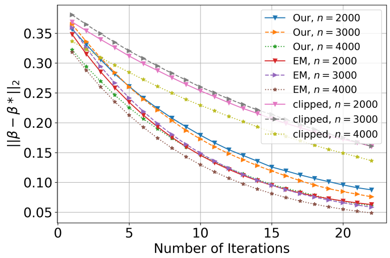
[3] http://archive.ics.uci.edu/ml/datasets/Adult, http://international.ipums.org



(a) $n = 2000, d = 10$

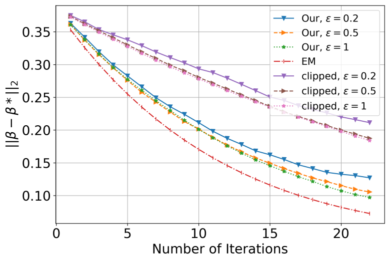


(b) $n = 2000, \epsilon = 0.5$

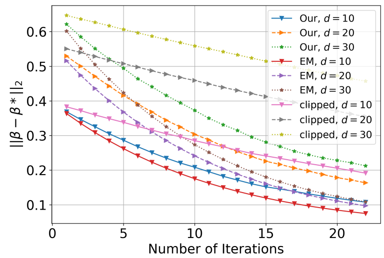


(c) $d = 10, \epsilon = 0.5$

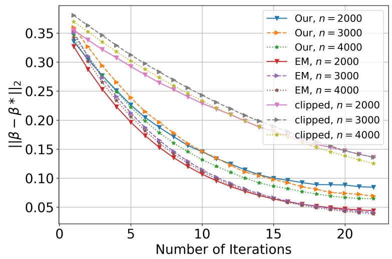
Figure 2. Estimation error of GMM w.r.t. privacy budget ϵ, data dimension d, data size n and iteration t



(a) $n = 2000, d = 10$

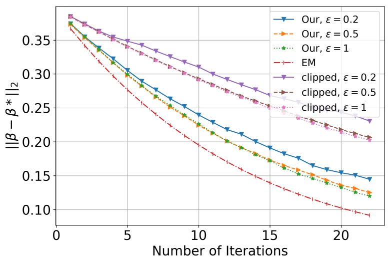


(b) $n = 2000, \epsilon = 0.5$

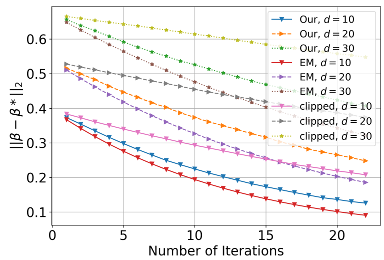


(c) $d = 10, \epsilon = 0.5$

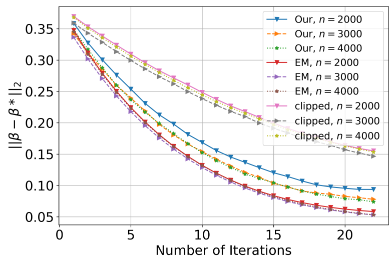
Figure 3. Estimation error of MRM w.r.t. privacy budget ϵ, data dimension d, data size n and iteration t



(a) $n = 2000, d = 10$

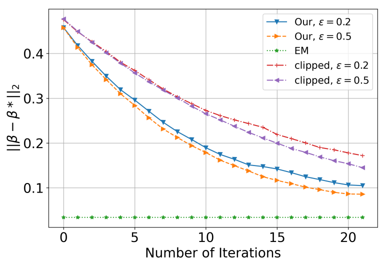


(b) $n = 2000, \epsilon = 0.5$

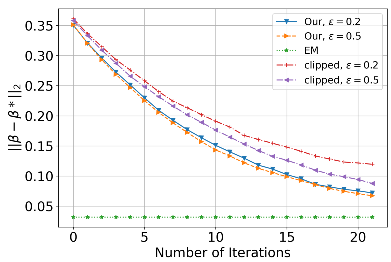


(c) $d = 10, \epsilon = 0.5$

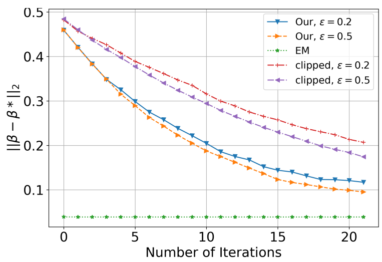
Figure 4. Estimation error of RMC w.r.t. privacy budget ϵ, data dimension d, data size n and iteration t



(a) ADULT



(b) IPUMS-US



(c) IPUMS-BR

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

Acknowledgements

Di Wang and Lijie Hu were supported in part by the baseline funding BAS/1/1689-01-01, funding from the CRG grand URF/1/4663-01-01, FCC/1/1976-49-01 from CBRC, and funding from the AI Initiative REI/1/4811-10-01 of King Abdullah University of Science and Technology (KAUST). Di Wang was also supported by the funding of the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI). The research of the last author was supported in part by NSF through grants IIS-1910492 and CCF-2200173 and by KAUST through grant CRG10-4663.2.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, 'Deep learning with differential privacy', in *CCS*, pp. 308–318, (2016).

[2] Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh, 'Computationally efficient robust sparse estimation in high dimensions', in *Conference on Learning Theory*, pp. 169–212, (2017).

[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] Raef Bassily, Adam Smith, and Abhradeep Thakurta, 'Private empirical risk minimization: Efficient algorithms and tight error bounds', in *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 464–473. IEEE, (2014).

[5] Olivier Catoni, *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*, volume 1851, Springer Science & Business Media, 2004.

[6] Olivier Catoni and Ilaria Giulini, 'Dimension-free pac-bayesian bounds for matrices, vectors, and linear least squares regression', *arXiv preprint arXiv:1712.02747*, (2017).

[7] Alexander Philip Dawid and Allan M Skene, 'Maximum likelihood estimation of observer error-rates using the em algorithm', *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **28**(1), 20–28, (1979).

[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith, 'Calibrating noise to sensitivity in private data analysis', in *Theory of cryptography conference*, pp. 265–284. Springer, (2006).

[9] Cynthia Dwork, Aaron Roth, et al., 'The algorithmic foundations of differential privacy.', *Foundations and Trends in Theoretical Computer Science*, **9**(3-4), 211–407, (2014).

[10] Susana Faria and F Gonçalves, 'Financial data modeling by poisson mixture regression', *Journal of Applied Statistics*, **40**(10), 2150–2162, (2013).

[11] Matthew J Holland, 'Robust descent using smoothed multiplicative noise', in *22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pp. 703–711, (2019).

[12] Lijie Hu, Shuo Ni, Hanshen Xiao, and Di Wang, 'High dimensional differentially private stochastic optimization with heavy-tailed data', in *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 227–236, (2022).

[13] Gautam Kamath, Or Sheffet, Vikrant Singhal, and Jonathan Ullman, 'Differentially private algorithms for learning mixtures of separated gaussians', in *Advances in Neural Information Processing Systems*, pp. 168–180, (2019).

[14] Nan M Laird, 'The em algorithm in genetics, genomics and public health', *Statistical Science*, 450–457, (2010).

[15] Jaewoo Lee and Daniel Kifer, 'Concentrated differentially private gradient descent with adaptive per-iteration privacy budget', in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1656–1665, (2018).

[16] Geoffrey McLachlan and Thriyambakam Krishnan, *The EM algorithm and extensions*, volume 382, John Wiley & Sons, 2007.

[17] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith, 'Smooth sensitivity and sampling in private data analysis', in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84, (2007).

[18] Mijung Park, James Foulds, Kamalika Choudhary, and Max Welling, 'Dp-em: Differentially private expectation maximization', in *Artificial Intelligence and Statistics*, pp. 896–904, (2017).

[19] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate, 'Stochastic gradient descent with differentially private updates', in *2013 IEEE Global Conference on Signal and Information Processing*, pp. 245–248. IEEE, (2013).

[20] Shuang Song, Om Thakkar, and Abhradeep Thakurta, 'Characterizing private clipped gradient descent on convex generalized linear problems', *arXiv preprint arXiv:2006.06783*, (2020).

[21] Jinyan Su, Lijie Hu, and Di Wang, 'Faster rates of private stochastic convex optimization', in *International Conference on Algorithmic Learning Theory*, pp. 995–1002. PMLR, (2022).

[22] Youming Tao, Yulian Wu, Xiuzhen Cheng, and Di Wang, 'Private stochastic convex optimization and sparse learning with heavy-tailed data revisited', in *International Joint Conferences on Artificial Intelligence Organization*, (2022).

[23] Di Wang, Jiahao Ding, Lijie Hu, Zejun Xie, Miao Pan, and Jinhui Xu, 'Differentially private (gradient) expectation maximization algorithm with statistical guarantees', *arXiv preprint arXiv:2010.13520*, (2020).

[24] Di Wang, Xiangyu Guo, Shi Li, and Jinhui Xu, 'Robust high dimensional expectation maximization algorithm via trimmed hard thresholding', *Machine Learning*, **109**(12), 2283–2311, (2020).

[25] Di Wang, Hanshen Xiao, Srini Devadas, and Jinhui Xu, 'On differentially private stochastic optimization with heavy-tailed data', in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 12-18 July 2020, Virtual Conference*, (2020).

[26] Di Wang, Minwei Ye, and Jinhui Xu, 'Differentially private empirical risk minimization revisited: Faster and more general', in *Advances in Neural Information Processing Systems*, pp. 2722–2731, (2017).

[27] Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu, 'High dimensional em algorithm: Statistical optimization and asymptotic normality', in *Advances in neural information processing systems*, pp. 2521–2529, (2015).

[28] CF Jeff Wu et al., 'On the convergence properties of the em algorithm', *The Annals of statistics*, **11**(1), 95–103, (1983).

[29] Zihang Xiang, Tianhao Wang, Wanyu Lin, and Di Wang, 'Practical differentially private and byzantine-resilient federated learning', *Proceedings of the ACM on Management of Data*, **1**(2), 1–26, (2023).

[30] Hanshen Xiao, Zihang Xiang, Di Wang, and Srinivas Devadas, 'A theory to instruct differentially-private learning via clipping bias reduction', in *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 2170–2189. IEEE Computer Society, (2023).

[31] Xinyang Yi and Constantine Caramanis, 'Regularized em algorithms: A unified framework and statistical guarantees', in *Advances in Neural Information Processing Systems*, pp. 1567–1575, (2015).

[32] Zhe Zhang and Linjun Zhang, 'High-dimensional differentially-private em algorithm: Methods and near-optimal statistical guarantees', *arXiv preprint arXiv:2104.00245*, (2021).

[33] Rongda Zhu, Lingxiao Wang, Chengxiang Zhai, and Quanquan Gu, 'High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm', in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 4180–4188. JMLR. org, (2017).