Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 7962489, 13 pages
https://doi.org/10.1155/2021/7962489

Research Article

The Discrete Gaussian Expectation Maximization (Gradient) Algorithm for Differential Privacy

Weisan Wu

School of Mathematics and Statistics, Baicheng Normal University, Baicheng, China

Correspondence should be addressed to Weisan Wu; wuws009@outlook.com

Received 24 November 2021; Accepted 13 December 2021; Published 30 December 2021

Academic Editor: Ahmed Mostafa Khalil

In this paper, we give a modified gradient EM algorithm; it can protect the privacy of sensitive data by adding discrete Gaussian mechanism noise. Specifically, it makes the high-dimensional data easier to process mainly by scaling, truncating, noise multiplication, and smoothing steps on the data. Since the variance of discrete Gaussian is smaller than that of the continuous Gaussian, the difference privacy of data can be guaranteed more effectively by adding the noise of the discrete Gaussian mechanism. Finally, the standard gradient EM algorithm, clipped algorithm, and our algorithm (DG-EM) are compared with the GMM model. The experiments show that our algorithm can effectively protect high-dimensional sensitive data.

1. Introduction

Now, big data have spread to every field and organization in our society, generating large amounts of personal data every day, which people use and analyse to enable the rapid development of society and technology. However, it is expected that some personal private data will be protected from being hacked or made public when it is collected. Therefore, how to effectively protect the privacy of data, not to be attacked, and can be effectively used, has gradually been paid attention to. Dwork et al. [1] introduced the concept and basic theoretical framework of differential privacy, which can effectively protect users' data privacy and has a strict and elegant mathematical theoretical framework and guarantees.

Gradient EM algorithm is one of the most important statistical models, and Wang et al. [2] recently applied sensitive data for privacy protection. Before this, people used the original EM algorithm and gradient EM algorithm, and there is no statistical guarantee. Until Balakrishnan et al. [4] gave the statistical guarantee of EM algorithm, Wang et al. [3] gave the guarantee of gradient EM algorithm based on it and extended it to the data privacy protection theory. However, just like most scholars, Gaussian noise with continuous distribution is added to the data, while in practice, the data output queries are often discrete, such as the number of records in the database that meets certain conditions. For this reason, Canonne et al. [5] proposed to use a discrete Gaussian mechanism to add discrete Gaussian noise to the data and to ensure that it has the same excellent accuracy as adding continuous Gaussian noise.

In this paper, we design a discretized Gaussian algorithm based on the gradient EM algorithm for differential privacy calculation based on [2]. Our algorithm has a good practical effect and can be extended to the general standard model. Meanwhile, the corresponding statistical guarantee of the algorithm is given in this paper. The structure of this paper is as follows: in the second part, we first introduce some theories of gradient EM algorithm, discrete Gaussian, and differential privacy, as well as some works related to this paper. In the third part, we introduce our model, namely, differential privacy discrete Gaussian EM (Gradient) algorithm (DG-EM), and the relevant statistical guarantee theorem. In the fourth part, we give the data simulation of the sensitivity, sample size, and dimension of the aggregated data, and the discussion of the model and future work are shown in the fifth part. Finally, we add the proof of some lemmas in the appendix.

2. Preliminaries

2.1. Gradient EM Algorithm. Assume that (X, Z) is complete data, where X is an observing sample and called Z as a latent variable. They are generally unobservable because they are missing or have underlying data structures. We denote $\mathcal{X}$ and $\mathcal{Z}$ as the sample space for variables X, Z, respectively. Suppose that (X, Z) has a joint density function $p_{\theta_0}(x, z)$; it belongs to some parameterized distribution family $\{p_{\theta_0} | \theta_0 \in \Omega\}$. For convenience, the variable X has a margin density function $\pi_\theta(x) = \int_{\mathcal{Z}} p_\theta(x, z)\mathrm{d}z$, and $\pi_\theta(z|x) = p_\theta(x, z)/\pi_\theta(x)$ is a Z' s conditional density function which is under $X = x$. Suppose that the given observer samples are $x_1, \ldots, x_n$ from population X. The EM algorithm needs to maximize the log-likelihood function $\ell_n(\theta) = \log\ p_\theta(x, z)$. Through Jensen's inequality, the lower bound of the log-likelihood function can be writen as follows:

$$\begin{aligned}&\frac{1}{n}\{\ell_n(\theta) - \ell_n(\theta')\}\\&\geq \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{Z}} \pi_{\theta'}(z \mid x_i)\log\ p_\theta(x_i, z)\mathrm{d}z - \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{Z}} \pi_{\theta'}(z|x_i)\log\ p_{\theta'}(x_i, z)\mathrm{d}z,\end{aligned} \tag{1}$$

where

$$q_i(\theta, \theta') = \sum_{i=1}^{n}\int_{\mathcal{Z}} \pi_{\theta'}(z|x_i)\log\ p_\theta(x_i, z)\mathrm{d}z, \tag{2}$$

$$Q_n(\theta, \theta') = \frac{1}{n}\sum_{i=1}^{n} q_i(\theta, \theta'). \tag{3}$$

The expectation of $Q_n(\theta, \theta')$ is denoted as

$$Q(\theta, \theta') = \mathbb{E}_{x\sim\pi_{\theta'}(x)}\int_{\mathcal{Z}} \pi_{\theta'}(z|x)\log\ p_\theta(x, z)\mathrm{d}z. \tag{4}$$

To maximize equation (3), the left term of the inequality can be sufficiently large by iteratively increasing the lower bound on the right term. The standard EM algorithm [6–9] estimates the function $Q_n(\theta, \theta^{(t)})$ by E-step at each iteration, then the parameters are estimated in M-step to make the parameter values of this iteration maximize the function $Q_n(\theta, \theta^{(t)})$ and denote the parameter as $\theta^{(t+1)} = \max_{\theta\in\Omega} Q_n(\theta, \theta^{(t)})$. The gradient EM algorithm is usually used to achieve higher accuracy and faster global maximum if the function is differentiable at each iteration step. The gradient EM algorithm is usually stated as follows: when the function $Q_n(\theta, \theta^{(t)})$ is differentiable at the t-th iteration, we can update the current parameter $\theta^{(t)}$ to $\theta^{(t+1)}$ by the following steps:

E-step: compute $Q_n(\theta, \theta^{(t)})$,
M-step: update $\theta^{(t+1)} = \theta^{(t)} + \eta\nabla Q_n(\theta^{(t)}, \theta^{(t)})$,

where η is a parameter which calls step size.

2.2. Discrete Gaussian. The study of discrete distributed forms of noise has received more attention this year. In the literature, people studied discrete Laplace distribution, discrete binomial distribution, and discrete Gaussian distribution and applied them to the field of cryptography.

In this paper, the differential privacy model is studied based on Gaussian mechanism. The noise with normal distribution makes the model have many elegant mathematical properties. Although the discrete Laplace noise mechanism and the discrete Gaussian noise mechanism cannot be compared in the same model, since they are used in different privacy mechanisms, we are still willing to use the discrete Gaussian noise in order to obtain aesthetic mathematical conclusions [10–13].

In this paper, we need to add noise to have discrete Gaussian distribution to specially treated sample. Firstly, we will give the definition of the discrete Gaussian distribution and some useful related theories.

Definition 1. Let $\mu, \sigma^2 \in \mathbb{R}, \sigma > 0$, if random variable X has probability mass function as follows:

$$\Pr(X = x) = \frac{\exp\{-(x-\mu)^2/(2\sigma^2)\}}{\sum_{y\in\mathbb{Z}}\exp\{-(y-\mu)^2/(2\sigma^2)\}}, \forall x \in \mathbb{Z}. \tag{5}$$

On the integers support set, then we call it is a discrete Gaussian distribution with location parameter μ and scale parameter σ^2 and denoted $N_{\mathbb{Z}}(\mu, \sigma^2)$.

2.3. Some Basic Theories on Differential Privacy. In this part, we will give some basic theories on differential privacy [14, 15].

Definition 2. A randomized algorithm $\mathcal{M}: \mathcal{X} \longrightarrow \mathcal{Y}$ satisfies (ϵ, δ)-differential privacy (DP) if for all neighboring datasets ,$D, D' \subset \mathcal{X}$, differing on a single entry. For all events S in the space $\mathcal{Y}$, we have $\Pr(\mathcal{M}(D) \in S) \leq e^\epsilon \Pr(\mathcal{M}(D') \in S) + \delta$. Moreover, we called its approximate differential privacy, if $\delta > 0$, and we called its pure or point-wise ϵ-differential privacy in the case of $(\epsilon, 0)$-differential privacy.

The concept of concentrated differential privacy given by Bun et al. [14] as follows:

Definition 3. A randomized algorithm $\mathcal{M}: \mathcal{X} \longrightarrow \mathcal{Y}$ satisfies ρ-concentrated differential privacy if for neighboring datasets $D, D' \subset \mathscr{X}$, and for any $\alpha \in (1, \infty)$, we have $D_\alpha(\mathscr{M}(D) \| \mathscr{M}(D')) \leq \rho$, where $D_\alpha(P\|Q) = (1/\alpha - 1)\log\sum_y (P(y)/Q(y))^\alpha Q(y)$ is the Renyi divergence of order α of the distribution form the distribution.

From these definitions, we have the conclusion that pure-DP can imply ρ-CDP, and ρ-CDP can imply $(\rho + 2\sqrt{\rho \log \delta^{-1}}, \delta)$-DP, where δ is a positive constant.

In order to ensure the consistency of the parameters of our model, we need some basic definitions and assumptions based on [4].

Definition 4 (self-consistent). We called the function $Q(\cdot; \theta^*)$ is self-consistent if $\theta^* = \operatorname{argmax}_{\theta \in \Omega} Q(\theta; \theta^*)$.

Definition 5 (Lipschitz-gradient-2 $(L, \mathscr{B})$). We called the function $Q(\cdot; \cdot)$ is Lipschitz-gradient-2 $(L, \mathscr{B})$, if we have the following inequality for parameter θ^* and $\theta \in \mathscr{B}$:

$$\left\|\nabla Q_n(\theta; \theta^*) - \nabla Q_n(\theta; \theta)\right\|_2 \leq L\left\|\theta - \theta^*\right\|_2. \tag{6}$$

Definition 6 (μ-smooth). We call the function $Q(\cdot; \cdot)$ is μ-smooth, if for any parameters $\theta, \theta' \in \mathscr{B}$, we have the inequality

$$Q(\theta; \theta') \geq Q(\theta'; \theta^*) + (\theta - \theta')^T \nabla Q(\theta'; \theta^*) - \frac{\mu}{2}\left\|\theta - \theta'\right\|_2^2. \tag{7}$$

Definition 7. (λ-strongly concave). We call the function $Q(\cdot; \theta^*)$ is λ-strongly concave, if for any parameters $\theta, \theta' \in \mathscr{B}$, we have the inequality

$$Q(\theta; \theta') \leq Q(\theta'; \theta^*) + (\theta - \theta')^T \nabla Q(\theta'; \theta^*) - \frac{\lambda}{2}\left\|\theta - \theta'\right\|_2^2. \tag{8}$$

Assumption 1. We assume that the function $Q(\cdot; \cdot)$ is self-consistent, Lipschitz-gradient-2 $(L, \mathscr{B})$, μ-smooth, and λ-strongly concave on some parameter sets $\mathscr{B}$.

3. Differential Privacy Discrete Gaussian EM (Gradient) Model

We will mention that the EM algorithm based on [2] and use the discrete Gaussian noise mechanism of high-dimensional truncation algorithm, which satisfies the centralized differential privacy (CDP). Like Wang et al. [2], we have first considered one coordinate case that is 1-dimensional random variable x. Let $x_1, \ldots, x_n$ be i.i.d. sampled from x. We get the clipped estimator as follows:

Step 1. For the sample x_i, we take a soft truncation function $h(x)$ which is defined by Catoni and Giulini [16],

$$h(x) = \begin{cases} -\dfrac{2\sqrt{2}}{3}, x < -\sqrt{2} \\ x - \dfrac{x^3}{6}, -\sqrt{2} \leq x \leq \sqrt{2} \\ \dfrac{2\sqrt{2}}{3}, x > \sqrt{2} \end{cases} \tag{9}$$

Then, we take some mild constant ω and rescaled sample x_i by dividing ω to get $h(x_i/\omega)$; through this approach, we can get the truncated mean as follows:

$$\frac{\omega}{n}\sum_{i=1}^{n} h\left(\frac{x_i}{\omega}\right) \approx \mathbb{E}(X). \tag{10}$$

From the expression of the function $h(x)$, we know $h(x)$ is bounded by $(2\sqrt{2}/3)$, so the sensitivity is $(4\sqrt{2}/3)$.

Step 2. Generate random noises $o_1, \ldots, o_n$ from a common distribution $o \sim \chi$ with $\mathbb{E}(o) = 0$. For data x_i, we get a new data $x_i(1 + o_i)$ though multiply the noise factor $1 + o_i$, and we get term $h(x_i(1 + o_i)/\omega)$ by scaling and truncation step. Finally, we get

$$\tilde{x}(o) = \frac{\omega}{n}\sum_{i=1}^{n} h\left(\frac{x_i(1 + o_i)}{\omega}\right). \tag{11}$$

Multiplicative noise is an effective method to ensure the estimation effect of typical points and increase the estimation effect of outliers as much as possible. It was first proposed by Srivastava et al. [17], and the motivation of using Gaussian multiplicative noise comes from [18].

Step 3. Finally, we take the expectation for the distributions with arrive multiplicative noise as follows:

$$\hat{x} = \mathbb{E}(\tilde{x}(o)) = \frac{\omega}{n}\sum_{i=1}^{n}\int h\left(\frac{x_i(1 + o_i)}{\omega}\right) d\chi(o_i). \tag{12}$$

Like Catoni and Giulini [16], taking $\chi \sim N(0, (1/\beta))$, we take the distribution χ following the discrete Gaussian distribution as $\chi \sim N_{\mathbb{Z}}(0, (1/\beta))$. Easily, for any given constant $a, b > 0$, we also have

$$\mathbb{E}_\chi(h(a + b\sqrt{\beta}o)) = a\left(1 - \frac{b^2}{2}\right) - \frac{a^3}{6} + R(a, b), \tag{13}$$

where $R(a,b)$ is a correction term $R(a,b) = T_1 + T_2 + T_3 + T_4 + T_5$. Signs $T_1 - T_5$ are respectively denoted as

$$\begin{aligned} T_1 &= \frac{2\sqrt{2}}{3}\left(F_- - F_+\right), \\ T_2 &= -\left(a - \frac{a^3}{6}\right)\left(F_- + F_+\right), \\ T_3 &= \frac{b}{\sqrt{2\pi}}\left(1 - \frac{a^2}{2}\right)\left(E_- - E_+\right), \\ T_4 &= \frac{ab^2}{2}\left(F_- + F_+ + \frac{1}{\sqrt{2\pi}}\left(V_+E_+ - V_-E_-\right)\right), \\ T_5 &= \frac{b^3}{6\sqrt{2\pi}}\left(\left(2 + V_-^2\right)E_- - \left(2 + V_+^2\right)E_+\right). \end{aligned} \tag{14}$$

Also, the notation is defined by

$$\begin{aligned} V_- &= \frac{\sqrt{2} - a}{b}, \\ V_+ &= \frac{a + \sqrt{2}}{b}, \\ E_- &= \exp\left(-\frac{V_-^2}{2}\right), \\ E_+ &= \exp\left(-\frac{V_+^2}{2}\right), \\ F_- &= \Phi\left(-V_-\right), \\ F_+ &= \Phi\left(-V_+\right). \end{aligned} \tag{15}$$

Unproved, we have the following estimation error Lemma 1 which is like Lemma 5 in Holland [19], and we gave the proof of it in Appendix A.

Lemma 1. *Let $x_1, \ldots, x_n$ be i.i.d. sampled form $x \sim \mu$. Assume $\mathbb{E}_\mu x^2 \le \tau$, and the upper bound has known. Given a number $0 < \gamma < 1$, for $\beta = 2\log(\gamma^{-1})$ and $\omega = \sqrt{(n\tau/2\log(\gamma^{-1}))}$, we have*

$$\left|\hat{x} - \mathbb{E}_\mu(x)\right| \le O\left(\sqrt{\frac{\tau\log\left(\gamma^{-1}\right)}{n}}\right), \tag{16}$$

with probability at least $1 - \gamma$.

From the soft truncation function and the multiplicative noise algorithm, we know that the sensitivity of the processed observation samples is $(4\sqrt{2}\,s/3n)$. Next, we need to add discrete Gaussian noise to the observations and obtain that the query

$$\mathscr{M}(D) = \hat{x} + Y, Y \sim N_{\mathbb{Z}}\left(0, \sigma^2\right), \sigma^2 = O\left(\frac{\omega^2\log\left(\delta^{-1}\right)}{\epsilon^2 n^2}\right), \tag{17}$$

will be (ϵ, δ)-DP, which leads the following Lemma 3; we give the proof in Appendix B.

Lemma 2. *Let $\epsilon > 0$; let the function $q\colon \mathscr{X}^n \longrightarrow \mathbb{Z}$ be an operator algorithm which is defined by Steps 1–3, satisfying $|q(x) - q(x')| \le \Delta$ for any $x, x' \in \mathscr{X}^n$; the query can be writen as randomized algorithm $\mathscr{M}\colon \mathscr{X}^n \longrightarrow \mathbb{Z}$ by $\mathscr{M}(D) = q(x) + Y$, where $Y \sim N_{\mathbb{Z}}(0, \sigma^2)$, then $\mathscr{M}$ satisfies (ϵ, δ)-DP.*

Furthermore, these results imply the following lemma.

Lemma 3. *Under the assumptions in Assumption 1, with probability at least $1 - \gamma$, the following holds:*

$$|\mathscr{M}(D) - \mathbb{E}(x)| \le O\left(\frac{\Delta\log\left(\delta^{-1}\right)}{\epsilon^2}\right). \tag{18}$$

After the estimation of the univariate private data, in the t-th iteration of Algorithm 1, we use the univariate estimation method for each coordinate of the gradient $\nabla Q_n(\theta^{(t)}; \theta^{(t)})$ and then get the estimation of the gradient $\nabla Q_n(\theta^{(t)}; \theta^{(t)})$. Finally, step M is performed.

Lemma 4. *For any $0 < \epsilon < 1$, let $D_\alpha(\mathscr{M}(x)\|\mathscr{M}(x')) \le \varpi$; for any $\alpha \in (1, \infty), \epsilon \ge 0$, and $x, x' \in \mathscr{X}^n$, Algorithm 1 satisfies (ϵ, δ)-DP for*

$$\delta = \frac{\exp\left((\alpha - 1)(\varpi - \epsilon)\right)}{\alpha - 1}\left(1 - \frac{1}{\alpha}\right)^\alpha, \tag{19}$$

where $Y \sim N_{\mathbb{Z}}(0, \sigma^2)$.

For Algorithm 1, the next theorem shows that the parameter estimation is consistent if the initial parameter θ^{Init} is close to the true parameter θ^* enough. After some simple calculations, we conclude that in Lemma 2, the upper bound is $\Delta = (n\tau + \omega_{op}^2/n\omega_{op})$ $\left\{1 + [(1/4)\log(3n\tau/2\omega_{op}^2) + \log(\gamma^{-1})]^{-1}\right\}$, where ω_{op} is the optimal numerical solution to the equation

$$2\omega^2 + n\mathbb{E}_\mu\left(x^2\right) = \omega^2\log\left(\frac{3n\mathbb{E}_\mu\left(x^2\right)}{2\omega^2}\gamma^{-2}\right). \tag{20}$$

Lemma 5. *Let $\mathscr{B} = \{\theta\colon \|\theta - \theta^*\|_2 \le R\}$ denote a parameter set with $R = \kappa\|\theta^*\|_2^2, \kappa \in (0, 1)$ which is a positive constant. Assume parameters $L, \mathscr{B}, \mu, \lambda, \tau$ satisfying condition of $1 - 2(\lambda - L/\lambda + \mu) \in (0, 1)$. If $\|\theta^{Init} - \theta^*\|_2 \le R/2$ and n is a large number such that*

$$\tilde{\Omega}\left(\left(\frac{1}{\lambda - L}\right)^2\frac{d^2T\tau\log\left(\gamma^{-1}\right)}{\epsilon^2R^2}\right) \le n. \tag{21}$$

```
Input: $D = \{x_i\} \subset \mathbb{R}^d, i = 1, \ldots, n$, privacy parameter $\epsilon, \delta, Q(\cdot;\cdot)$ and $q_i(\cdot;\cdot)$, initial parameter $\theta^{Init} \in \mathscr{B}$ and $\tau$ satisfy Assumption 1, the number of iterations $T$, step size $\eta$, and failure probability $\gamma > 0$.
(1) Let $\overline{\epsilon} = \sqrt{2(\log(\delta^{-1}) + \epsilon)} - \sqrt{2\log(\delta^{-1})}, \omega = \sqrt{(n\tau/2\log(d/\gamma))}, \beta = \log(d/\gamma)$,
(2) for $t = 1, \ldots, T$ do
(3)     For each $j \in [d]$, calculate the robust gradient and add a discrete Gaussian noise, that is,
        $g_j^{(t-1)}(\theta^{(t-1)}) = (\omega/n)\sum_{i=1}^{n}[\nabla_j q_i(\theta^{(t-1)};\theta^{(t-1)})(1 - (\nabla_j^2 q_i(\theta^{(t-1)};\theta^{(t-1)})/2\omega^2\beta)) - (\nabla_j^3 q_i(\theta^{(t-1)};\theta^{(t-1)})/6\omega^2)]$
        $+(\omega/n)\sum_{i=1}^{n}\sum^{n} R((\nabla_j q_i(\theta^{(t-1)};\theta^{(t-1)})/\omega), (|\nabla_j q_i(\theta^{(t-1)};\theta^{(t-1)})|/\omega\sqrt{\beta})) + Y_j^{(t-1)}$
        where $Y_j^{(t-1)} \sim N_{\mathbb{Z}}(0, \sigma^2), \sigma^2 = (8\tau\, dT/9\beta n\overline{\epsilon}^2)$.
(4)     Let vector $\check{\nabla} Q_n(\theta^{(t-1)}) \in \mathbb{R}^d$ denote $\check{\nabla} Q_n(\theta^{(t-1)}) = (g_1^{(t-1)}(\theta^{(t-1)}), g_2^{(t-1)}(\theta^{(t-1)}), \ldots, g_d^{(t-1)}(\theta^{(t-1)}))$.
(5)     Update $\theta^{(t)} = \theta^{(t-1)} + \eta\check{\nabla} Q_n(\theta^{(t-1)})$.
(6) end for
```

Algorithm 1: Differentially private DG-EM (gradient) algorithm.

We have $\Pr(\theta^{(t)} \in \mathscr{B}) \geq 1 - 2T\gamma$ for all $t \in [T]$. Furthermore, if we take $T = O((\lambda + \mu/\lambda - L)\log(n))$ and $\eta = (2/\lambda + \mu)$, we have

$$\left\|\theta^{(T)} - \theta^*\right\|_2 \leq \widetilde{O}\left(R\sqrt{\frac{\lambda + \mu}{(\lambda - L)^3}}\;\frac{d\log\left(\delta^{-1}\right)\log\left(\gamma^{-1}\sqrt{\tau}\right)}{\sqrt{n\epsilon^2}}\right). \tag{22}$$

Lemma 6. *Let $(\|\theta^*\|/\sigma) \geq r$, then there exists a constant C such that the properties of self-consistent Lipschitz-gradient-$2(L, \mathscr{B})$, μ-smoothness, and λ-strongly concave hold for the function $Q(\cdot;\cdot)$ with $L = \exp(-Cr)$, $\mu = \lambda = 1$, $R = \kappa\|\theta^*\|_2$, $\kappa = 1/4$, $\mathscr{B} = \{\theta\colon \|\theta - \theta^*\| \leq R\}$, where r is a enough large constant means that the minimum signal-to-noise ratio (SNR).*

Furthermore, we can get Theorems 1 and 2. The proof of these theorems is very simple; we do not list the detailed proof procedure here. In fact, we only need to replace the upper bound on the variance of the discrete noise in [2] with a single coordinate with $3\exp(-1/2\sigma^2)$.

Theorem 1. *With the same condition as in Lemma 4, for any $\theta \in \mathscr{B}$, the j-th coordinate of $\nabla q(\theta;\theta)$ satisfies the following results:*

$$\mathbb{E}_y\left(\nabla_j q(\theta;\theta)\right)^2 \leq O\left(\left\|\theta^*\right\|_\infty^2 + 3\exp\left(-\frac{1}{2\sigma^2}\right)\right). \tag{23}$$

Theorem 2. *With the same conditions in Lemma 3, we assume that $\|\theta^{Init} - \theta^*\|_2 \leq (\|\theta^*\|_2^2/8)$ in Algorithm 1, and n is a large enough number such that*

$$\widetilde{\Omega}\left(\frac{n\left(\left\|\theta^*\right\|_\infty^2 + 3\exp\left(-\left(1/2\sigma^2\right)\right)\right) + \omega_{op}^2}{\omega_{op}\epsilon^2\left\|\theta^*\right\|_2^2}d^2\left\{1 + \left[\frac{1}{4}\log\left(\frac{3n\tau}{2\omega_{op}^2}\right) + \log\left(\gamma^{-1}\right)\right]^{-1}\right\}\right) \leq n. \tag{24}$$

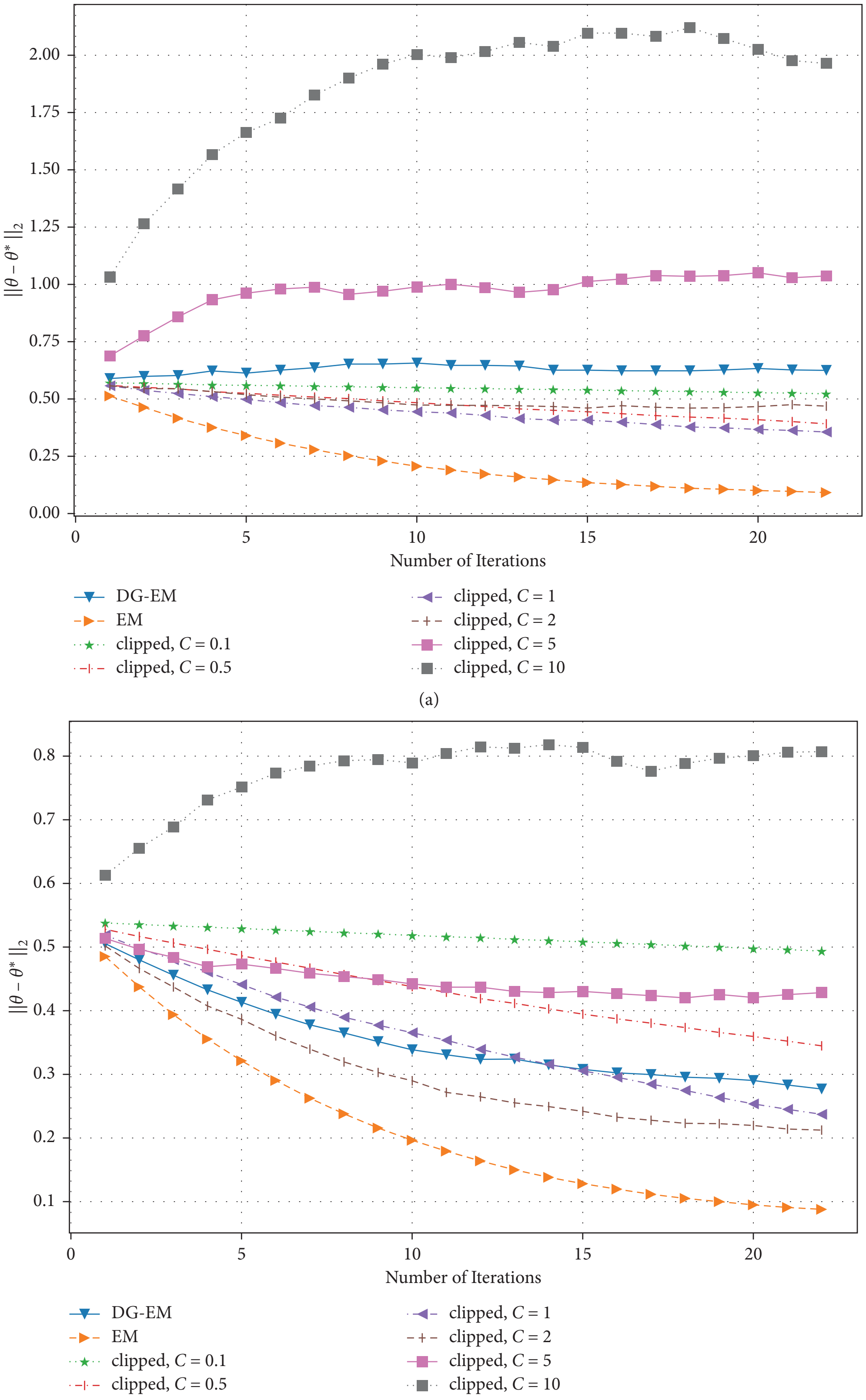
If we take $T = O(\log(n))$ and the ratio as $\eta = O(1)$, then for a failure probability γ, we have with probability at least $1 - 2T\gamma$

$$\left\|\theta^{(T)} - \theta^*\right\|_2 \leq \widetilde{O}\left(\left\|\theta^*\right\|_2\frac{n\sqrt{\left\|\theta^*\right\|_\infty^2 + 3\exp\left(-\left(1/2\sigma^2\right)\right)} + \omega_{op}^2}{\sqrt{n\epsilon^2}\,\omega_{op}}d\{1+\left[\frac{1}{4}\log\left(\frac{3n\tau}{2\omega_{op}^2}\right) + \log\left(\gamma^{-1}\right)\right]^{-1}\}\right). \tag{25}$$

We note that Lemmas 3–6 and Theorems 1 and 2 are easy to get through Lemmas 1 and 2. Due to limited space, we delete these proofs here, and readers can prove them by themselves. It is only necessary to pay attention to the upper bound of the ℓ_2-norm between the iterative values of parameters and the truth values in the process of proof.

4. Experiments and Results

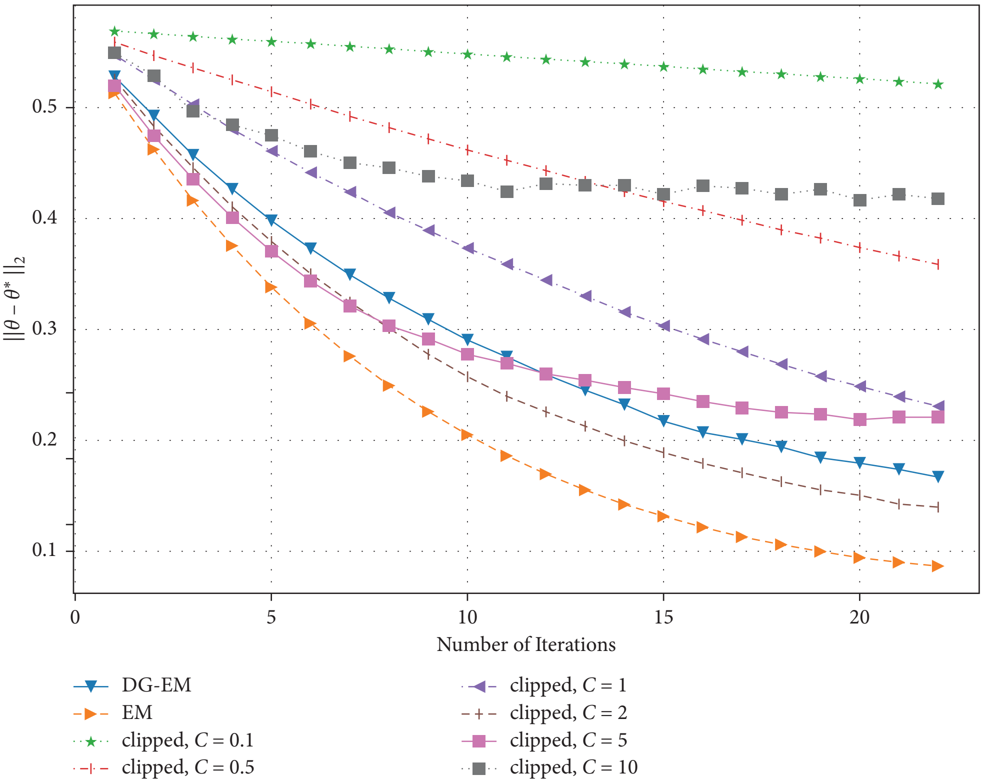
In this section, we will evaluate the performance of Algorithm 1 on the GMM model based on these methods. We will study the statistical setting and theoretical behavior of this algorithm on synthetic data.



(a)

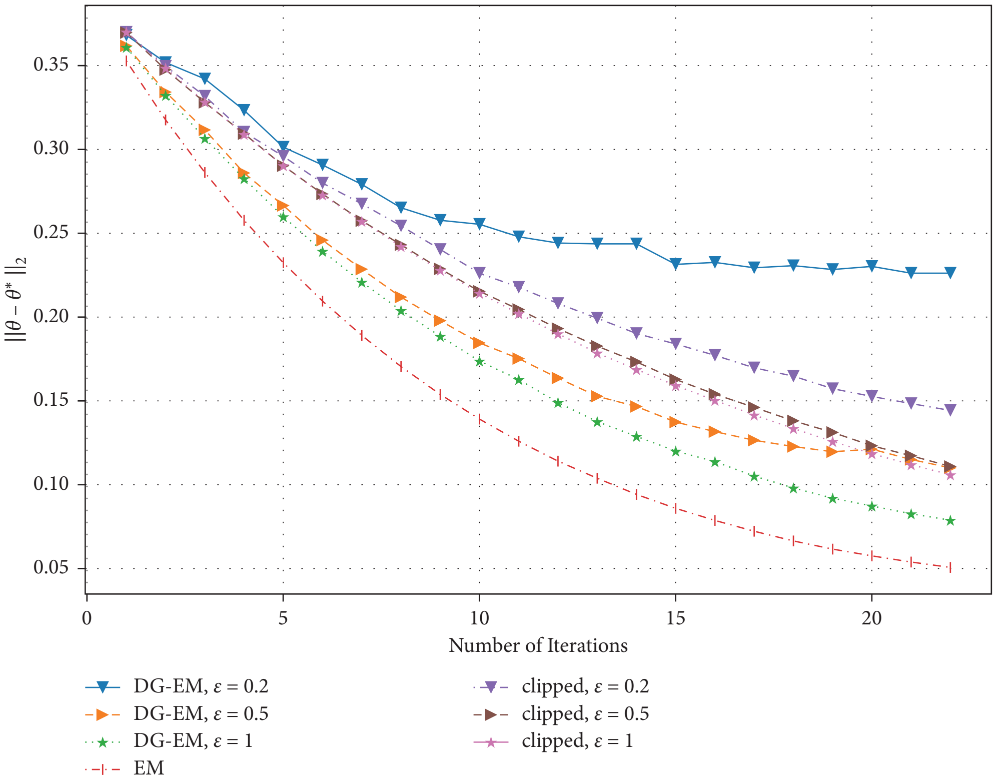
(b)

Figure 1: Continued.



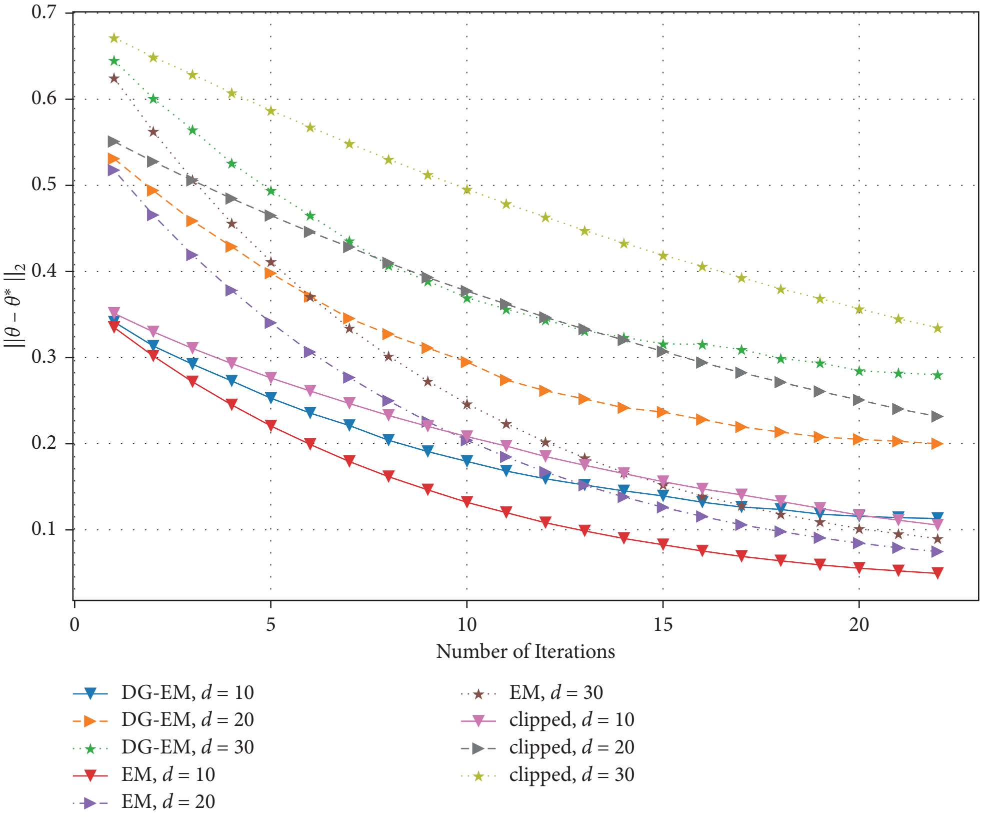
(c)

Figure 1: Estimation error of GMM clipped vs. iteration t under different clipping threshold C and budgets ϵ. (a) $n = 1000$; $d = 20$; $\epsilon = 0.2$, (b) $n = 1000$; $d = 20$; $\epsilon = 0.5$, and (c) $n = 1000$; $d = 20$; $\epsilon = 1$.

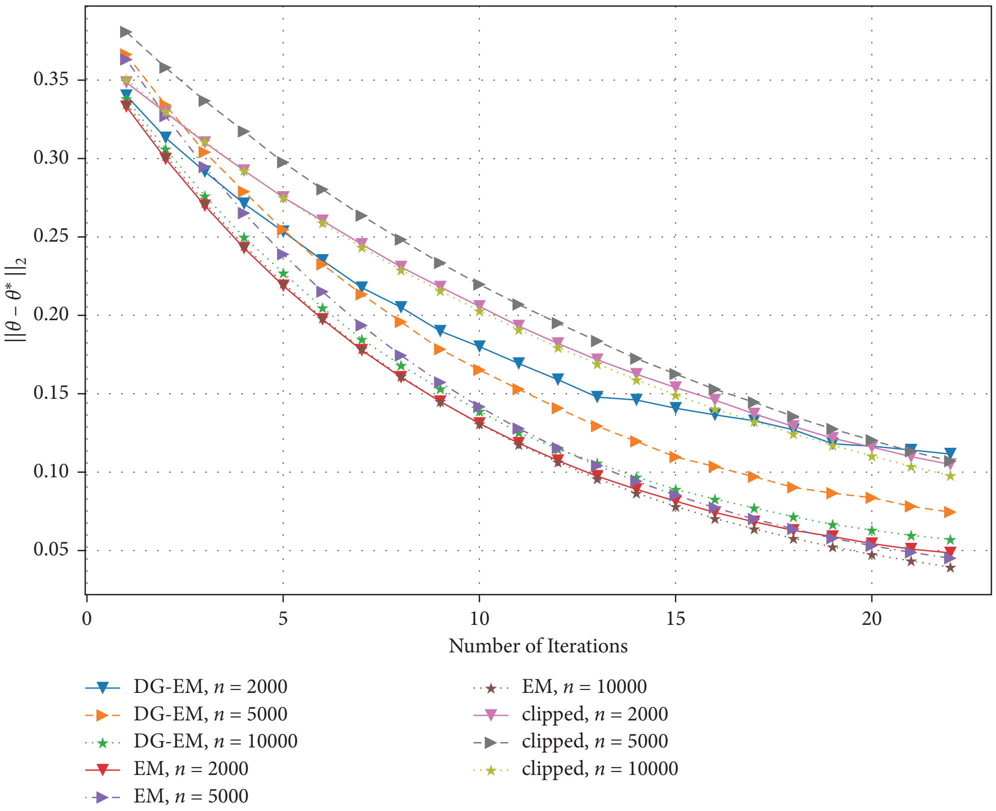


(a)

Figure 2: Continued.

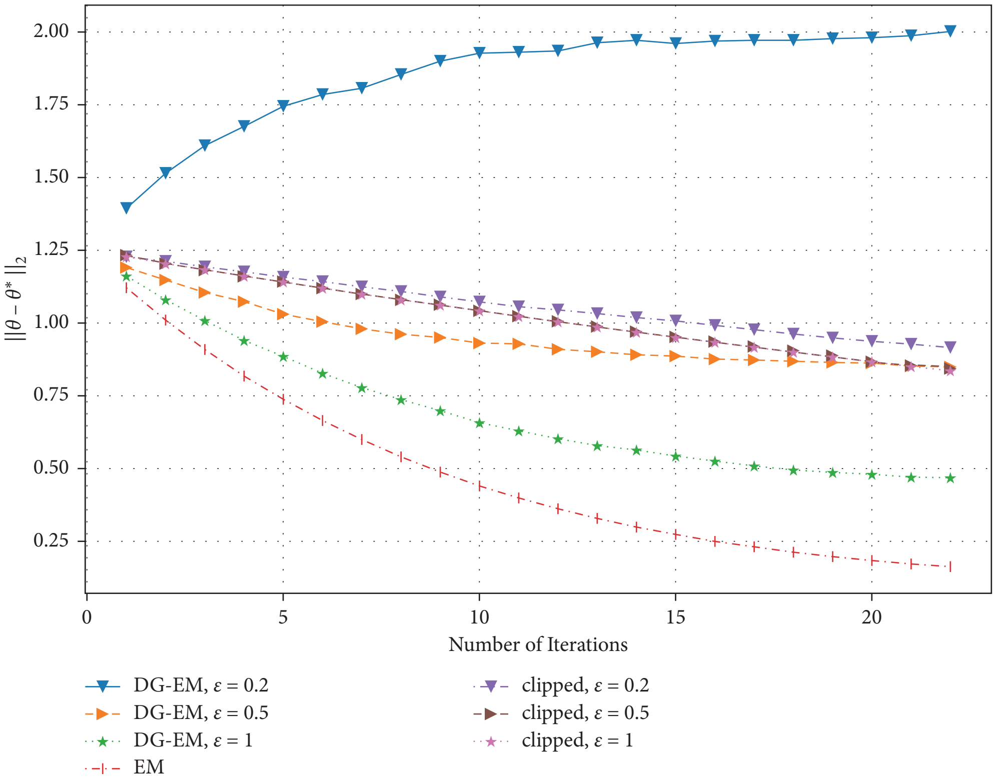


(b)

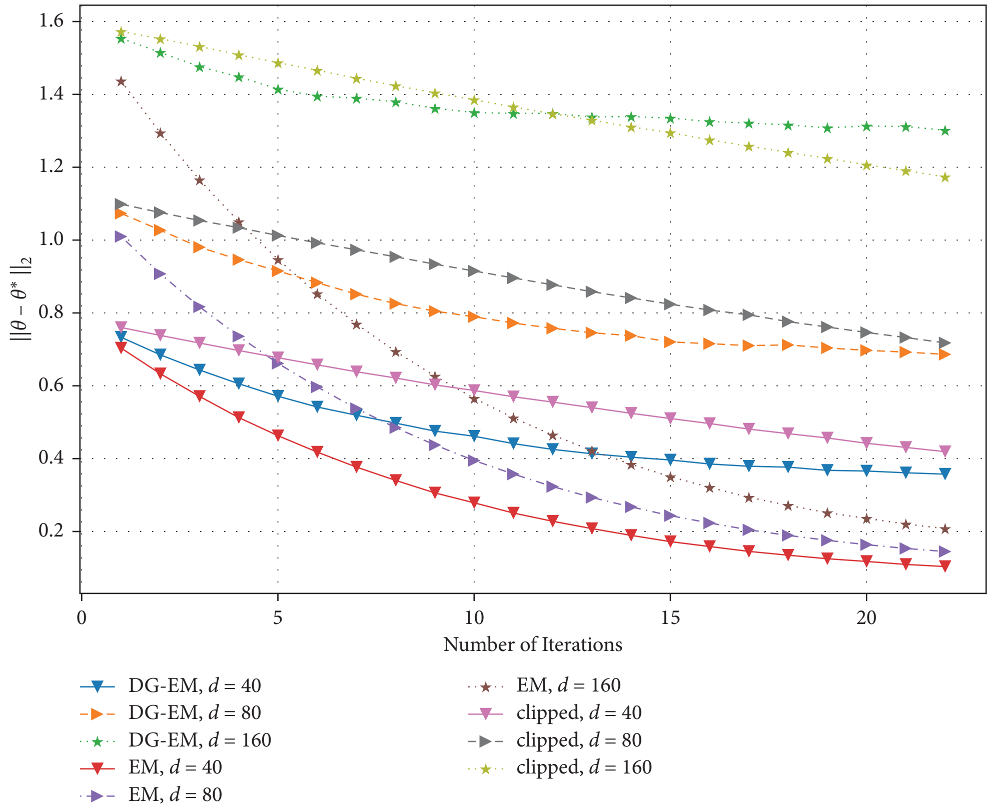


(c)

FIGURE 2: Estimation error of GMM w.r.t. privacy budget ϵ, data dimension (lower) d, data size n, and iteration t. (a) $n = 2000$; $d = 10$, (b) $n = 2000$; $\epsilon\ = 0.5$, and (c) $d = 10$, $\epsilon = 0.5$.

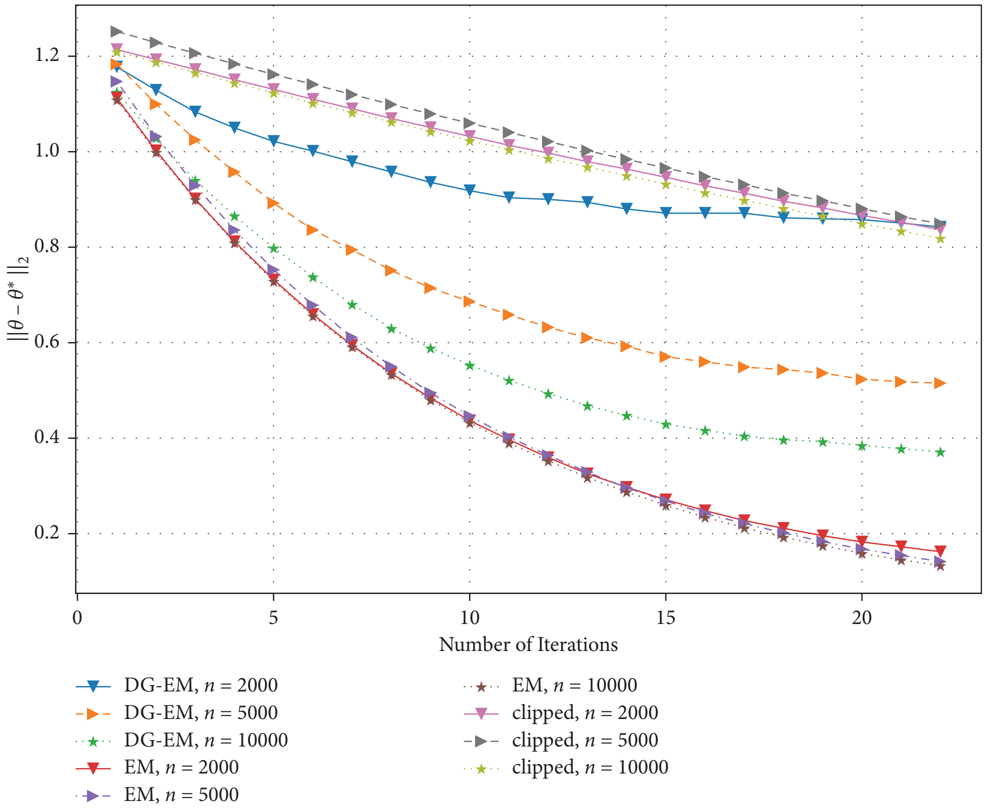


(a)



(b)

Figure 3: Continued.



(c)

Figure 3: Estimation error of GMM w.r.t. privacy budget ϵ, data dimension (higher) d, data size n, and iteration t. (a) $n = 2000$; $d = 100$, (b) $n = 2000$; $\epsilon = 0.5$, and (c) $d = 100$, $\epsilon = 0.5$.

4.1. Baseline Methods. In this part, we will compare the two methods primarily. For convenience, we will refer to the gradient EM algorithm as EM, which will serve as a nonprivate baseline method. The other is the clipped differential private EM algorithm, which we still refer to as clipped [20], which will serve as our privacy baseline approach.

4.2. Experimental Settings. In this experiment, we generate the synthetic data of the mixed distribution of two components. To generate each of the algorithm, we consider the random initialization method for the selection of initial parameter values. In the results, we used to measure the resulting estimation error. We set signal-to-noise ratio $(\|\theta^*\|/\sigma) = 3$. For the privacy parameter ϵ, we set $\epsilon = \{0.5, 0.8, 1\}$, and then the parameter $\delta = \Pr(Y > (\epsilon\sigma^2/\Delta) + (\Delta/2))$ needs to calculate because it is the function of ϵ.

4.3. Experimental Results. As can be seen from Figure 1, we fixed $n = 1000$, $d = 20$. When the budget of our method is set at different values, the estimation error decreases significantly with the increase of iteration time. When the budget is 0.2, 0.5, and 1, the optimal value is 1, 2, and 2, respectively. It is difficult for us to determine the optimal value C.

In Figure 2, under the lower dimension case, we test how the data dimension d, privacy budget ϵ, and data size n affect the estimation error $\|\theta - \theta^*\|_2$ of algorithms on the Gaussian mixture model over iteration t. We can see that the estimation error of Algorithm 1 in GMM decreases when ϵ increases, n increases, or d decreases. However, we can see that when the budget ϵ is small, the effect of our algorithm is performed badly, and the estimation error declines unstably with the increase of the number of iterations.

In Figure 3, we can see that, in the face of high-dimensional data, the effect of estimation error $\|\theta - \theta^*\|_2$ needs a relatively large sample to be guaranteed. We conducted experiments with higher dimensions $d = 40, 80, 160$ and different sample sizes of 2000, 5000, and 10 000, respectively. It can be seen that when the sample size n is large enough, the estimation error can be guaranteed to decrease significantly with the number of iterations t. As shown in Figure 3, with the increase of sample size, our algorithm is equally effective in high-dimensional space, which is not comparable with Wang et al.'s [2] algorithm.

5. Conclusions

In this paper, we study the differential privacy model with discrete Gaussian mechanism noise. Through the process of data scaling and truncation, the model effectively solves the influence of high-dimensional data on the model. Through the experimental part and theoretical proof, we can see that the estimation error of the model adding discrete Gaussian noise is faster than that of the model adding continuous Gaussian noise in the low dimension than that of the clipped model. The effect is much better than that of [2] in the case of high dimension. At the same time, in the previous lemma section, we can see that our model has more compact bounds, because of the smaller variance of discrete Gaussian noise.

Appendix

A

Proof of Lemma 1

Proof. In order to make the conclusion universal, we make some necessary assumptions. Firstly, let $\mathscr{P}(\mathbb{R})$ denote all probability measures on $\mathbb{R}$, and we assumed it has an appropriate σ-field. Consider any two measures $\nu, \nu' \in \mathscr{P}(\mathbb{R})$, and $f_0: \mathbb{R} \longrightarrow \mathbb{R}$ is a ν'-measurable function. We take the form of a cumulative generating function as

$$\sup_{\nu}\left(\int f_0(u)\mathrm{d}\nu(u) - D_\alpha\left(\nu\|\nu'\right)\right) = \log\left(\int \exp\left(f_0(u)\right)\mathrm{d}\nu'(u)\right), \tag{A.1}$$

through a Legendre transform of the mapping $\nu \longrightarrow D_\alpha(\nu\|\nu')$ like [16], where $D_\alpha(\nu\|\nu')$ denotes the Renyi divergence between ν and ν'.

Because $h(x_i(1+o_i)/\omega), i \in [n]$ is depend on two random quantities x_i and the noise o_i, we write $f(o,x) \triangleq h(x(1+o)/\omega), o, x \in \mathbb{R}$.

By the definition of function $h(\cdot)$ before, the function $f: \mathbb{R} \longrightarrow \mathbb{R}$ is measurable and bounded with $(2\sqrt{2}/3)$. Next, we let

$$f_0(o) = \sum_{i=1}^{n} f(o, x_i) - c(o), \tag{A.2}$$

where $c(o)$ is a term needs to be determined later. Inserting $f_0(o)$ to (A.1), we have

$$\begin{aligned} B &\triangleq \sup_{\nu}\left(\int f_0(o)\mathrm{d}\nu(o) - D_\alpha\left(\nu\|\nu'\right)\right) \\ &= \log\left(\int \exp\left(\sum_{i=1}^{n} f(o,x_i) - c(o)\right)\mathrm{d}\nu(o)\right). \end{aligned} \tag{A.3}$$

Furthermore, we have

$$\begin{aligned} \mathbb{E}_\mu(\exp(B)) &= \mathbb{E}_\mu\left(\int \frac{\exp\left(\sum_{i=1}^{n} f(o,x_i)\right)}{\exp(c(o))\mathrm{d}\nu(o)}\right) \\ &= \int \prod_{i=1}^{n} \mathbb{E}_\mu \frac{\exp(f(o,x_i)))}{\exp(c(o))\mathrm{d}\nu(o)}. \end{aligned} \tag{A.4}$$

If

$$c(o) = n\log\left(\mathbb{E}_\mu \exp(f(o,x))\right), \tag{A.5}$$

we have

$$\mathbb{E}_\mu(\exp(B)) = \int\left(\prod_{i=1}^{n} \frac{\mathbb{E}_\mu \exp(f(o,x_i))}{\left[\mathbb{E}_\mu \exp(f(o,x))\right]^n}\right)\mathrm{d}\nu(o) = 1. \tag{A.6}$$

So,

$$\begin{aligned} \Pr\left(B \geq \log\left(\gamma^{-1}\right)\right) &= \Pr\left(\exp(B) \geq \gamma^{-1}\right) \\ &= \mathbb{E}_\mu I\{\gamma \exp(B) \geq 1\} \\ &\leq \mathbb{E}_\mu(\gamma \exp(B)) \\ &= \gamma. \end{aligned} \tag{A.7}$$

Because $c(o)$ is ν-measurable, the $f_0(o)$ is ν-measurable. We have

$$\sup_{\nu}\left(\int f_0(o)\mathrm{d}\nu(o) - D_\alpha\left(\nu\|\nu'\right)\right) = \log\left(\gamma^{-1}\right), \tag{A.8}$$

with probability at least $1-\gamma$. We have the following inequality:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\int f(o,x_i)\mathrm{d}\nu(o) \leq& \int \log\left(\mathbb{E}_\mu \exp(f(o,x))\right)\mathrm{d}\nu(o) \\ &+ \frac{D_\alpha\left(\nu\|\nu'\right) + \log\left(\gamma^{-1}\right)}{n}. \end{aligned} \tag{A.9}$$

Since the noise terms $o_1, \ldots, o_n$ are independent and follow distribution $o \sim \nu$, we can get

$$\begin{aligned} \hat{x} &= \frac{s}{n}\sum_{i=1}^{n}\int h\left(\frac{x_i(1+o_i)}{\omega}\right)\mathrm{d}\nu(o_i) \\ &= \frac{s}{n}\sum_{i=1}^{n}\int f(o,x_i)\mathrm{d}\nu(o_i). \end{aligned} \tag{A.10}$$

Thus, we have the bound from equation (A.9) as follows:

$$\begin{aligned} \hat{x} \leq& s\int \log\left(\mathbb{E}_\mu \exp\left(h\left(\frac{x(1+o)}{\omega}\right)\right)\right)\mathrm{d}\nu(o) \\ &+ \frac{s}{n}\left[D_\alpha\left(\nu\|\nu'\right) + \log\left(\gamma^{-1}\right)\right], \end{aligned} \tag{A.11}$$

and then we need to analyse the first term and the second term on the right-hand side of the top inequality (A.11).

For the first term, from the definition of the truncation function $h(\cdot)$, by (A.11), we have

$$\begin{aligned}
&\int \log\left(\mathbb{E}_\mu \exp\left(h\left(\frac{x(1+o)}{\omega}\right)\right)\right)\mathrm{d}\nu(o)\\
&\le \int \log\left[\frac{(1+o)\mathbb{E}_\mu(x)}{s} + \frac{(1+o)^2\mathbb{E}_\mu\left(x^2\right)}{s^2}\right]\mathrm{d}\nu(o)\\
&= \frac{\mathbb{E}_\nu(1+o)\mathbb{E}_\mu(x)}{s} + \frac{\mathbb{E}_\nu(1+o)^2\mathbb{E}_\mu\left(x^2\right)}{s^2}\\
&= \frac{\mathbb{E}_\mu(x)}{s} + \frac{\mathbb{E}_\mu\left(x^2\right)}{s^2}\left(1+\frac{1}{\theta}\right).
\end{aligned} \tag{A.12}$$

Since $o \sim \nu = N(0, (1/\theta))$, the expectation and variance of the $1 + o$ are as follows:

$$\mathbb{E}_\nu(1+o) = 1, \mathbb{E}_\nu(1+o)^2 = \frac{1}{\theta} + \left[\mathbb{E}_\nu(1+o)\right]^2 = \frac{1}{\theta} + 1. \tag{A.13}$$

For the second term in (A.11), we need to evaluate $D_\alpha(\nu\|\nu')$. We take $\nu' = N(0, (1/\theta))$; through simple computations, we can get

$$\begin{aligned}
D_\alpha\left(\nu\|\nu'\right) &== \frac{1}{\alpha-1}\log\left(\int (\mathrm{d}\nu)^\alpha \left(\mathrm{d}\nu'\right)^{1-\alpha}\mathrm{d}u\right)\\
&= \frac{(1-0)^2}{2\theta^{-1}}\alpha\\
&= \frac{\theta\alpha}{2}.
\end{aligned} \tag{A.14}$$

Thus, we can take the upper bound form as

$$\widehat{x} \le \mathbb{E}_\mu(x) + \frac{\mathbb{E}_\mu\left(x^2\right)}{2s}\left(1+\frac{1}{\theta}\right) + \frac{s}{n}\left(\frac{\theta\alpha}{2} + \log\left(\gamma^{-1}\right)\right). \tag{A.15}$$

We take the differential for the variable s; we have

$$s^2 = \left(1+\frac{1}{\theta}\right)\mathbb{E}_\mu\left(x^2\right)\left(\frac{\theta\alpha}{2} + \log\left(\gamma^{-1}\right)\right)^{-1}, \tag{A.16}$$

and with respect to θ, we have

$$\theta^2 = \frac{n\mathbb{E}_\mu\left(x^2\right)}{\alpha s^2}. \tag{A.17}$$

Plugging equation (A.17) into the setting of s, we can get

$$s^2 = \frac{n\mathbb{E}_\mu\left(x^2\right)}{2\log\left(\gamma^{-1}\right)}. \tag{A.18}$$

We can get the setting of s from equation (A.18); equation (A.15) has upper bound with form as follows:

$$\widehat{x} \le \mathbb{E}_\mu(x) + \sqrt{\frac{2\mathbb{E}_\mu\left(x^2\right)\log\left(\gamma^{-1}\right)}{n}} + \sqrt{\frac{\mathbb{E}_\mu\left(x^2\right)}{n}}. \tag{A.19}$$

To get lower bounds on $\widehat{x} - \mathbb{E}_\mu(x)$, we need to get the upper bounds on $-\widehat{x} - \mathbb{E}_\mu(x)$. Similar to the analysis above, we get the upper bound of $-\widehat{x}$ through

$$-\widehat{x} \le s\int \log\left(\mathbb{E}_\mu \exp\left(-h\left(\frac{x(1+o)}{\omega}\right)\right)\right)\mathrm{d}\nu(o) + \frac{s}{n}\left[D_\alpha\left(\nu\|\nu'\right) + \log\left(\gamma^{-1}\right)\right] \tag{A.20}$$

By the fact

$$-\log\left(\frac{1-x+x^2}{2}\right) \le \psi(x) \le \log\left(\frac{1-x+x^2}{2}\right), \tag{A.21}$$

we have

$$\begin{aligned}
&\int \log\left(\mathbb{E}_\mu \exp\left(-h\left(\frac{x(1+o)}{\omega}\right)\right)\right)\mathrm{d}\nu(o)\\
&\le \int \log\left[1 + \frac{-(1+o)\mathbb{E}_\mu(x)}{s} + \frac{(1+o)^2\mathbb{E}_\mu\left(x^2\right)}{s^2}\right)\right]\mathrm{d}\nu(o),
\end{aligned} \tag{A.22}$$

$$\widehat{x} \le -\mathbb{E}_\mu(x) + \sqrt{\frac{2\mathbb{E}_\mu\left(x^2\right)\log\left(\gamma^{-1}\right)}{n}} + \sqrt{\frac{\mathbb{E}_\mu\left(x^2\right)}{n}}. \tag{A.23}$$

Putting the above analysis together, we have

$$\hat{x} - \mathbb{E}_{\mu}(x) \geq \sqrt{\frac{2\mathbb{E}_{\mu}(x^2)\log(\gamma^{-1})}{n}} + \sqrt{\frac{\mathbb{E}_{\mu}(x^2)}{n}}. \tag{A.24}$$

Then, with probability at least $1 - \gamma$, the following holds

$$\left|\hat{x} - \mathbb{E}_{\mu}(x)\right| \leq \sqrt{\frac{2\mathbb{E}_{\mu}(x^2)\log(\gamma^{-1})}{n}} + \sqrt{\frac{\mathbb{E}_{\mu}(x^2)}{n}}. \tag{A.25}$$

□

B

Proof of Lemma 2

The proof process of Lemma 2 needs the next proposition [5]:

Proposition 1. *Let $\sigma, \alpha \in \mathbb{R}$ with $\sigma > 0$ and $\alpha \geq 1$. Let $\mu_1, \mu_2 \in \mathbb{Z}$. Then,*

$$D_{\alpha}\left(N_{\mathbb{Z}}(\mu_1, \sigma^2) \middle\| N_{\mathbb{Z}}(\mu_2, \sigma^2)\right) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}\alpha. \tag{B.1}$$

Furthermore, this inequality is an equality whenever $\alpha(\mu_1 - \mu_2)$ is an integer.

Proof of Lemma 2: we can get Lemma 2 easily though Proposition 1 and Definition 3.

Data Availability

The data in this paper are random numbers generated by statistical software R.

Conflicts of Interest

The author declares no conflicts of interest.

Acknowledgments

This research was supported by the Higher Education Project of Jilin Province (Grant no. JGJX2019D253).

References

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," *Theory of Cryptography*, vol. 3876, pp. 265–284, 2006.

[2] D. Wang, J. Ding, Z. Xie, M. Pan, and J. Xu, "Differentially private (gradient) expectation maximization algorithm with statistical guarantees," 2020, https://arxiv.org/abs/2010.13520.

[3] D. Wang, M. Ye, and J. Xu, "Differentially private empirical risk minimization revisited: Faster and more general," 2018, https://arxiv.org/abs/1802.05251.

[4] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the EM algorithm: from population to sample-based analysis," *The Annals of Statistics*, vol. 45, no. 1, 2017.

[5] C. L. Canonne, G. Kamath, and T. Steinke, "The discrete gaussian for differential privacy," 2020, https://arxiv.org/abs/2004.00010.

[6] G. Mclachlan and T. Krishnan, *The EM Algorithm and Extensions*, Wiley-Interscience, Hoboken, NJ, USA, 2nd edition, 2007.

[7] S. Dasgupta and L. Schulman, "A two-round variant of em for gaussian mixtures," 2013, https://arxiv.org/abs/1301.3850.

[8] S. Vempala and G. Wang, "A spectral algorithm for learning mixture models," *Journal of Computer and System Sciences*, vol. 68, no. 4, pp. 841–860, 2004.

[9] I. Naim and D. Gildea, "Convergence of the EM algorithm for Gaussian mixtures with unbalanced mixing coefficients," in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, vol. 2, Edinburgh, Scotland, June 2012.

[10] Y. Du, B. Fan, and B. Wei, "An improved exact sampling algorithm for the standard normal distribution," 2020, https://arxiv.org/abs/2008.03855.

[11] M. Abadi, A. Chu, I. Goodfellow et al., "Deep learning with differential privacy," in *Proceedings of the ACM Conference on Computer and Communications Security*, pp. 308–318, Vienna, Austria, October 2016.

[12] A. Koskela, J. Jälkö, L. Prediger, and A. Honkela, "Tight approximate differential privacy for discrete-valued mechanisms using FFT," 2020, https://arxiv.org/abs/2006.07134.

[13] R. K. Zhao, R. Steinfeld, and A. Sakzad, "COSAC: Compact and scalable arbitrary-centered discrete Gaussian sampling over integers," *Post-Quantum Cryptography*, vol. 12100, pp. 284–303, 2020.

[14] M. Bun and T. Steinke, "Average-case averages: private algorithms for smooth sensitivity and mean estimation," 2019, https://arxiv.org/abs/1906.02830.

[15] C. Dwork and R. Aaron, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, pp. 211–407, 2014.

[16] O. Catoni and I. Giulini, "Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression," 2017, https://arxiv.org/abs/1712.02747.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[18] E. Nalisnick, A. Anandkumar, and P. Smyth, "A scale mixture perspective of multiplicative noise in neural networks," 2015, https://arxiv.org/abs/1506.03208.

[19] M. J. Holland, "Robust descent using smoothed multiplicative noise," 2020, https://arxiv.org/abs/1810.06207.

[20] S. Song, O. Thakkar, and A. Thakurta, "Characterizing private clipped gradient descent on convex generalized linear problems," 2020, https://arxiv.org/pdf/2006.06783.