

\section*{ABSTRACT} Differentially stochastic gradient descent (DP-SGD) that perturbs the clipped gradients is a popular approach for private machine learning. Gaussian mechanism GM, combining with the moments accountant (MA), has demonstrated a much better privacy-utility tradeoff than using the advanced composition theorem. However, it is unclear whether the tradeoff can be further improved by other mechanisms with different noise distribution. To this end, we extend GM (p = 2) to the generalized p-power exponential mechanism (pEM with \( p > 0 \)) family and show its privacy guarantee. Straightforwardly, we can enhance the privacy-utility tradeoff of GM by searching noise distribution in the wider mechanism space. To implement pEM in practice, we design an effective sampling method and extend MA to pEM for tightly estimating privacy loss. Besides, we formally prove the non-optimality of GM based on the variation method. Numerical experiments validate the properties of pEM and illustrate a comprehensive comparison between pEM and the other two state-of-the-art methods. Experimental results show that pEM is preferred when the noise variance is relatively small to the signal and the dimension is not too high.

\section*{INDEX TERMS} Privacy Protection, Privacy-utility Trade-off, Noise Variance, Gaussian Mechanism, Moments Accountant.

\section*{I. INTRODUCTION} Machine learning has been highly successful across a variety of applications, including computer vision\cite{1}, image understanding\cite{2}, natural language processing\cite{3}. The great success is thanks to the advances of computing power, the breakthroughs in algorithmic design, as well as the availability of massive data. However, data collection and mining have raised severe privacy issues\cite{4, 5, 6}. Ideally, it is wished that, after well training, the output model generalizes away from the specific data of any individual user. However, information related to the training data is memorized by deep network models\cite{7}. Even worse, some attacks exploiting this implicit memorization have demonstrated that private, sensitive training data can be recovered from model parameters or updated gradients\cite{8, 9}, which further exacerbate the privacy concerns in the development and applications of machine learning. Recently, differential privacy (DP) has been proposed to protect the privacy of training data by introducing noises to machine learning process. Machine learning with DP enjoys the high effectiveness and efficiency in privacy preservation, thus attracting increasingly attentions from both the academia and industry\cite{10, 11, 12, 13}. Due to the added noise, DP based machine learning algorithms often suffer from the model utility degradation and a fundamental problem is \textit{how to improve the trade-off between model utility and privacy loss}. Traditionally, DP for machine learning can be achieved by output perturbation or objective perturbation, i.e., adding calibrated noises to the final model parameters\cite{14} or the optimization objective\cite{15}. However, for the popular deep learning tasks that the training process is more like a black box, it is very difficult to characterize the dependence of the final parameters or the objective on the training data. Differently, gradient perturbation that adds noise to the gradients during training is more flexible than output or objective perturbation for general machine learning.
models [16]. A popular gradient perturbation method is DP-SGD [11], [12], [17], which adds noises to the clipped gradients.

Among the existing research, Gaussian mechanism (GM) which draws noises from Gaussian distribution, comimg with MA [11] which provides a tight estimates of privacy loss, has been widely applied in various SGD based machine learning tasks [5], [11], [12], [18], [17], [19]. Besides, considerable research has been conducted to improve the tradeoff from other perspectives, which can be categorized as sensitivity calibration [19], [20] that aims to improve the utility by adaptively calibrate the sensitivity, dimension reduction [21] and transformations [22] that aim to reduce the sensitivity in the lower-dimension or the transformed space instead of the original space, correlation exploration [23], [24] that aims to introduce fewer noises according to the correlation between gradients and target parameters, post-processing that applies denoising technique on the noisy gradient to improve the trained model utility [25]. However, these strategies are usually problem dependent. To use them, one has to analyze the specific features of data space, query function, and exploit the underlying relationship. This limits the generality of these methods.

A. MOTIVATION

Given the exact $\sigma$, does GM is the optimal $(\varepsilon, \delta)$-DP? If this is not true, what is the better mechanism and how to efficiently apply it in practice?

Some remarks are in order. First, the criterion of "optimal mechanism" or "better mechanism" is defined as a mechanism which achieves the lowest or lower privacy loss under the given mean square error. Second, the "efficiently apply" means one can generate noises from the proposed mechanism at least the same as generating noises from GM. Third, we use MA to track the privacy loss, unless otherwise indicated.

B. BASIC IDEA

To simply answer the above question, a straightforward approach is to find an instance which is better than GM. That is, we should explore other mechanisms rather than exploiting GM itself. As Fig. 1 shows, the basic idea is to generalize the density function $h(x) \sim \exp (-\|x\|^2 / \beta)$ (where $\| \cdot \|$ denotes the $l_2$ norm as default) to $h(x) \sim \exp (-\|x\|^p / \beta)$ with $p > 0$, meanwhile, using MA to tightly track the privacy loss. Intuitively, it is possible to select a better mechanism from the richer distribution space. The possibility is based on the fact that each $p$EM has a different function shape and the corresponding richer moment information about the privacy loss variable. Therefore, we can make a tighter estimate of $\varepsilon$ based on the richer moment information. Figure 2(a) illustrates the density function shape of $p$EM and the corresponding privacy loss. Besides, Figure 2(b) illustrates the application scope of $a$GM, MA, and the proposed $p$EM, based on the results of parameter analysis in Section VI.

C. CHALLENGES AND CONTRIBUTIONS

Till now, we have illustrated that GM is not the optimal and we can select a better one from $p$EM with $p \neq 2$. It seems that the problem has been done. However, two main challenges remain.

1. Note that one can replace the mean square error by other utility measures such as absolute error, entropy or others.

2. For a given mechanism, using different accountant methods, such as the advanced composition and MA, will cause different privacy loss. We adopt the later method, considering that MA ensures a tight estimate.
emerge when we make a deeper analysis about \( pEM \). One challenge is how to use MA for multi-dimensional \( pEM \)? The main difficulty lies in the inefficient or even prohibitive computation when solving the high-dimension integration in MA. The other challenge is how to formally prove the non-optimality of GM. The main difficulty is how to solve the restricted optimization problem with a maximin objective. We aim to solve the challenges and main contributions of the work are as follows.

1) We generalize GM to \( pEM \), which ensures \( \varepsilon \)-DP when \( p \in (0, 1] \) and \( (\varepsilon, \delta) \)-DP when \( p > 1 \). This significantly enriches existing DP mechanisms and allows us to find a mechanism from the broader mechanism family to ensure a better privacy-utility tradeoff than GM.

2) We apply MA to \( pEM \) while preserving a high computational efficiency. In particular, the estimate of \( \varepsilon \) reduces from a \( d \)-dimensional integration to a double integral whenever the dimension \( d \geq 2 \). It is the first time to apply MA to other DP mechanisms, beyond GM. In addition, we show how to efficiently sample noises from \( pEM \).

3) We derive the necessary conditions for optimal \( (\varepsilon, \delta) \)-DP based on the variation method and prove that GM is not the optimal \( (\varepsilon, \delta) \)-DP mechanism when \( \varepsilon \) is tracked by the moment accountant.

4) Experimental results validate the proposed formula for tracking privacy loss and show that \( pEM \) outperforms other two state-of-the-art DP methods, especially when the signal-to-noise ratio is relatively large and the dimension is not too high.

The remainder is organized as follows. Section II introduces related work about how to improve the privacy-utility tradeoff. Section III gives some preliminaries about DP, MA and SGD. In Section IV, we propose the \( pEM \) and extend MA to general \( pEM \). In Section V we prove the non-optimality of GM. We conducted extensive experiments to validate \( pEM \) in Section VI and summary our study in Section VII.

II. RELATED WORK

There have been a variety studies about how to improve the privacy-utility tradeoff in DP. Because DP-SGD is an instance of DP application, we review the related references including but not limited to the machine learning from three aspects.

A. PRIVACY LOSS ACCOUNTANT

To improve the tradeoff, one approach is using different methods to estimate the privacy loss. For a mechanism satisfying \( (\varepsilon_0, \delta_0) \)-DP at each fold, the simple composition ensures \( (T\varepsilon_0, T\delta_0) \)-DP after \( T \) folds. As shown in [27], this can be improved to \( (\varepsilon', T\delta_0 + \delta') \) under the advanced composition with \( \varepsilon' = \sqrt{2T \ln(1/\delta')/\varepsilon_0 + T\varepsilon_0(\varepsilon_0 - 1)} \). However, this only exploits the first order moment of privacy loss variable. Furthermore, by tracking the higher moments information, the authors [11] propose MA for achieving a tighter estimate. Meanwhile, MA has been implemented in the Tensorflow Privacy library. Other relaxation conceptions of DP, such as RDP [28] and CDP [29] are also proposed to account the privacy loss. All conceptions [11], [28], [29] can give a tighter estimates for the privacy loss than the advanced composition, but the privacy loss measurements in [28], [29] are inconsistent with the original conception of DP and a conversion is needed.

B. SENSITIVITY CALIBRATION

Under a given privacy loss tracking method, one can still calibrate the sensitivity to improve the tradeoff. When the sensitivity is large and the worst-case rarely occurs, using the global sensitivity will introduce overestimating noises and destroy the model utility. To alleviate it, the smooth sensitivity [30] is proposed to reduce the noise variance. In machine learning, the clipping strategy [11] is a common method to calibrate the influence of training data on gradients or parameters. However, because using the fixed clipping threshold can not accurately capture the varying characteristics of gradients or parameters, several adaptive methods are proposed to further improve the model utility [19], [20], [31]. Besides directly calibrating the sensitivity, other research with the similar idea has been conducted to improve the tradeoff based on dimension reduction [32], transformation [33], and correlation exploration [13].

C. NOISE DISTRIBUTION DESIGN AND OPTIMAL MECHANISM

Under the given privacy loss accountant and techniques described above, one orthogonal perspective to improve the tradeoff is designing noise distributions. When outputs are real values, Laplace mechanism [34] and GM [27] are two typical mechanisms, ensuring \( \varepsilon \)-DP and \( (\varepsilon, \delta) \)-DP respectively. When outputs are discrete values, Geometry mechanism [35] and Binomial mechanism [36] are the corresponding two mechanisms. At the first time, [37] proves that it is impossible to design the universally optimal mechanism. Nonetheless, analyzing the optimal mechanism in certain conditions is possible. For \( \varepsilon \)-DP, [39] proves that Geometry mechanism is optimal for the single count query and [38] presents the staircase-shaped mechanism is optimal for the single real-valued query. Besides, Laplace mechanism is proved to be the optimal under the entropy-minimizing criterion [39] or the constraint of \( \varepsilon \)-Lipschitz privacy [40].

For \( (\varepsilon, \delta) \)-DP, although many researchers have studied the theoretical bound of the optimal mechanism [41], [42], [43], [44], they mainly focus on the non-adaptive linear queries but seldom propose practical mechanism. In machine learning, the most common mechanism is the GM [12], [17], [45], [46]. In [26], the authors develop aGM in which the noise variance is derived from the cumulative density function instead of using the tail bound approximation. In [47], the authors propose a novel approach R\(^2\)DP in which a two-fold distribution may approximately cover the search space of all distributions, leading to a high efficient computation. Different from [11], [26], [47] which exploits GM itself, we...
propose new mechanisms and analyze the optimality of GM based on the proposed mechanisms.

III. PRELIMINARIES

Before giving the formal analysis of the optimal mechanism, we review some basic concepts about the mini-batch SGD, DP, and MA. We use the bold font to denote vector and vector function, such as $x$, $g(\cdot)$. By default, $\| \cdot \|$ means $\| \cdot \|_2$. For convenience, we list the used acronyms in Table 1 and variables in Table 2.

A. MINI-BATCH SGD

For a machine learning task, let $F(w, \xi)$ be the loss function with $w \in \mathbb{R}^d$ is the model parameter and $\xi$ is the training sample. In practice, we usually obtain a training set of $n$ i.i.d. samples $\xi_1, \cdots, \xi_n$ following the unknown $P$ and aim to minimize the empirical risk $R(w) = \frac{1}{n} \sum_{i=1}^{n} F(w, \xi_i)$.

The iteration update is

$$w_{t+1} \leftarrow w_t - \frac{\gamma_t}{b} \sum_{\xi \in B_t} \nabla F(w_t, \xi), \quad (1)$$

where $\gamma_t$ is a positive stepsize and $B_t$ is the sampled mini-batch of size $b$ which is usually small compared to $n$. Mini-batch SGD has two advantages [48], the high efficiency due to the computations of the first-order gradients and the reliability of convergence due to the reduced variance of the stochastic gradient estimates. Therefore, it has been extensively used in practice [5], [11], [12], [13], [17], [19] and is adopted in our analysis. Considering the DP-SGD, i.e., adding noise to gradients to protect the training data privacy, Eq. (1) is rewritten as

$$w_{t+1} \leftarrow w_t - \frac{\gamma_t}{b} \sum_{\xi \in B_t} (\nabla F(w_t, \xi) + \xi_t), \quad (2)$$

where $\xi_t, i = 1, \cdots, b$ are noises independently drawn from a given distribution (e.g., Gaussian distribution) with zero mean at the $t$-th iteration.

B. DIFFERENTIAL PRIVACY (DP)

DP is a rigorous privacy notion that has emerged from a line of work in theoretical computer science and cryptography, aiming to limit the information disclosure in computing results via adding proper noises.

**Definition 1 (DP [11]).** Let $\varepsilon > 0$ and $\delta \in [0, 1]$. A mechanism $M$ is said to be $(\varepsilon, \delta)$-DP if for all neighboring datasets $D$ and $D'$ that differ in a single entry and for any measurable subset $E$ in the output space, we have

$$\Pr[M(D) \in E] \leq e^{\varepsilon} \Pr[M(D') \in E] + \delta. \quad (3)$$

The probability is taken over the random coins of $M$. When $\delta = 0$, we say that $M$ preserves pure DP which is denoted as $\varepsilon$-DP. When $\delta > 0$, we say that $M$ preserves approximate DP which is denoted as $(\varepsilon, \delta)$-DP.

In this paper, we consider the multi-dimensional noise addition mechanism for privacy preservation. That is, we aim to find an optimal noise distribution (denoted by the density function $h(x)$), from which we draw noise $x \in \mathbb{R}^d$ and add it to the assumed real-vector $f(D)$ for privacy protection. Then Eq. (3) can be expressed as $\int h(x)dx \leq e^{\varepsilon} \int h(x-\bar{g})dx + \delta$, where $\bar{g} = f(D) - f(D')$. One common mechanism used in DP-SGD is GM, which has the following conclusion.

**Theorem 1 (GM [27]).** Let $\varepsilon \in (0, 1)$ be arbitrary. For $c^2 > 2 \ln(1.25/\delta)$, the mechanism $M$ that adds noise drawn from Gaussian distribution $N(0, \Sigma)$ preserves $(\varepsilon, \delta)$-DP, where $\Sigma$ is a diagonal matrix with entries $\sigma^2$ and $\sigma \geq c\Delta/\varepsilon$.

It has been proven that the estimate of $\varepsilon$ in Theorem 1 is loose. To improve the tradeoff, MA is used for tracking a tight estimate for $\varepsilon$, especially after $T$ folds.

C. MA WITH PRIVACY AMPLIFICATION

MA is a privacy account method using Markov’s inequality to track the detailed information of privacy loss distribution [11]. Formally, the privacy loss variable is defined as

$$c_M(o; D, D', aux, \lambda) \triangleq \ln \frac{\Pr[M(aux, D) = o]}{\Pr[M(aux, D') = o]}, \quad (4)$$

where $aux$ is the side information and $o \in \mathbb{R}^d$ is the noisy output. When noise is drawn from distribution with probability density $h(x)$, Eq. (4) is expressed as $c_M(o; x, D, D', \lambda) = \ln \frac{h(x)}{h(x-\bar{g})}$. Then, the $\lambda$-th moment of random variable $c_M(o; \cdot)$ is

$$\alpha_M(o; \lambda; h(x)) = \max_{D, D'} \ln \mathbb{E}_{x \sim h(x)} \exp[\lambda c_M(o; x, D, D')] \quad (4)$$

To compute privacy loss $\varepsilon$ for the given failure probability $\delta$ after $T$ iterations, one needs to solve the following optimization

$$\varepsilon = \min_{\lambda \in \mathbb{R}^+} \frac{T \alpha_M(o; \lambda; h(x)) + \ln(1/\delta)}{\lambda}. \quad (5)$$

Without considering the amplification, $\varepsilon$ is determined by $h(x)$ and its translation $h(x-\bar{g})$ with $\|\bar{g}\| \leq \Delta$. 

TABLE 1. List of Acronyms

| Acronym | Description |
|---------|-------------|
| DP      | Differential privacy [22] |
| SGD     | Stochastic gradient descent |
| MA      | Moments account method [11] |
| GM or cGM | Classical Gaussian mechanism [27] |
| aGM     | Analytical Gaussian mechanism [26] |
| pEM     | $p$-power exponential mechanism (ours) |

TABLE 2. Variables and Descriptions

| Variable | Description |
|----------|-------------|
| $F(w, \xi)$ | Loss function of model parameters $w$ and sample $\xi$ |
| $\nabla F(w, \xi), d$ | Stochastic gradient of $F(w, \xi)$ with dimension $d$ |
| $c, \sigma^2$ | Clipping bound for $\nabla F(w, \xi)$ and noise variance |
| $B_t, b$ | Sampled mini-batch at $t$-th iteration with size $b$ |
| $n, q, T$ | Dataset size, sampling ratio, and iteration number |
| $\varepsilon, \delta, \Delta$ | Privacy loss, failure probability, and global sensitivity |
| $p_0, p_1, p_s$ | Range of $p$-EM $[p_0, p_1]$ with step size $p_s$ |

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
where $\Delta$ is the global sensitivity. However, when considering amplification, $\varepsilon$ will be further reduced and is determined by $h(x)$ and $h_{\hat{q}, \hat{g}}(x) \Delta \geq (1 - q)h(x) + qh(x - \hat{g})$, where $q$ is the sampling ratio \cite{11, 49}. Then, the calculation of $\alpha_M(\lambda; h(x))$ becomes

$$\begin{align*}
\alpha_M(\lambda; h(x)) &= \max_{\beta, \beta^\prime} \ln \int_{\mathbb{R}^d} \left[ \frac{h(x)}{h_{\hat{q}, \hat{g}}(x)} \right]^\lambda h(x)dx \\
&= \max\{\ln I(h, h_{q, \hat{g}}), \ln I(h_{q, \hat{g}}, h)\},
\end{align*}$$

where

$$\begin{align*}
I(h, h_{q, \hat{g}}) &= \int_{\mathbb{R}^d} \left[ \frac{h(x)}{h_{q, \hat{g}}(x)} \right]^\lambda h(x)dx, \\
I(h_{q, \hat{g}}, h) &= \int_{\mathbb{R}^d} \left[ \frac{h_{q, \hat{g}}(x)}{h(x)} \right]^\lambda h_{q, \hat{g}}(x)dx
\end{align*}$$

are two important conceptions when analyzing MA and also in our analysis. Note that Eq. (5) is a unimodal/quasi-convex function with respect to $\lambda$ \cite{20}. Therefore, there exists a unique solution of Eq. (5). If we extend $\lambda \in \mathbb{Z}^+$ to $\lambda \geq 1, \lambda \in \mathbb{R}$, then MA is extended to Rényi DP \cite{28} and Eq. (5) can be solved to an arbitrary accuracy $\tau$ in time $\log(\lambda^*/\tau)$ with $\lambda^*$ the optimum.

**D. PIPELINE OF DP-SGD**

Based on Sections \[ III-A \] \[ III-B \] \[ III-C \] we show the pipeline of DP-SGD in machine learning, as shown in the top block (light orange) of Figure 3. In particular, it consists of five steps: (1) sample mini-batch $B_i$ with ratio $q$; (2) compute gradients $\nabla F(w_t, \xi)$ where $\xi \sim B_i$; (3) clip $\nabla F(w_t, \xi)$ with threshold $c$; (4) add noise to clipped gradient; (5) use average noisy gradient to update $w_t$. Note that DP is used to add noise to the clipped gradient in step (4), and mini-batch SGD is used to update the model parameters in step (5). Additionally, MA is used to track the privacy loss of the accumulated noises, based on Eq. (5). In traditional, the used DP mechanism in step (4) is Laplacian or Gaussian. In the paper, we propose to use a new family of DP mechanism, pEM, to protect the training process by adding noise to the clipped gradient. The aim is to further improve the privacy-utility tradeoff.

**IV. $p$-POWER EXPONENTIAL MECHANISM FAMILY**

In the section, we show how to design and apply pEM to DP-SGD. As the bottom block (light green) of Figure 3 shows, the process of pEM consists of four parts, A to D. They are corresponding to Sections \[ IV-A \] \[ IV-B \] \[ IV-C \] \[ IV-D \] respectively. Note that A and B are two components of C. In Section \[ IV-A \] we propose and analyze the general $p$-power exponential mechanism (pEM) family. Obviously, the optimal mechanism selected from pEM family will achieve a better tradeoff than GM under any given criterion. In Section \[ IV-B \] we apply MA to pEM for tightly tracking the privacy loss. In Section \[ IV-C \] we show how to select the optimal $p^*$ based on given parameters. In Section \[ IV-D \] we show how to efficiently generate noise from pEM.

**A. DEFINITION AND PRIVACY GUARANTEE**

pEM is an extension of GM. Therefore, based on Gaussian distribution, we propose to define pEM as follows. This section corresponds to box A in Figure 3.

**Definition 2 (pEM Family).** A random mechanism $M$ is called pEM if it draws noise from the distribution with density function $h(x) = 1/\alpha \exp(-\|x\|^p/\beta), p > 0, x \in \mathbb{R}^d$, where $\alpha > 0$ is the normalization and $\beta > 0$ is the scale parameter.

Obviously, pEM contains GM ($p = 2$) as a specific instance. In the following, we further demonstrate that pEM satisfies $\varepsilon$-DP when $p \leq 1$ and $(\varepsilon, \delta)$-DP when $p > 1$. That is, pEM can ensure both pure and approximate DP for different values $p > 0$. The proof when $p \leq 1$ is straightforwardly derived from \cite{13}, while the proof when $p > 1$ is based on Lemma \[ 1 \]. In particular, Lemma \[ 1 \] shows that $\|x\|$ follows a generalized gamma distribution $\gamma(r; k, \beta, p) = \frac{1}{\Gamma(\frac{r}{k})} e^{-\frac{k^p r}{\beta}} \frac{\beta^p}{\Gamma(\frac{r}{k})}, r > 0, k = d/p$ and $\frac{1}{\Gamma(r/k)} = \frac{r^p}{\Gamma(r/k)}$. Based on Lemma \[ 1 \] we deduce the following.
properties of pEM.

**Theorem 2.** For a random mechanism $M$ belonging to the pEM family, that is, $M: f(D) \rightarrow f'(D)+x$, where the density function of $x$ is $h(x) = 1/\alpha \exp(-\|x\|^p/\beta)$ and $x \in \mathbb{R}^d$. Let $\Delta$ be the global sensitivity of vector function $f(D)$, i.e., $\Delta = \max_{D,D'} |f(D)-f(D')|$, where $D$ and $D'$ are two adjacent data sets. If the scale parameter $\beta$ is set as:

$$\beta = \Delta^p/\varepsilon, \quad 0 < p \leq 1, \quad (8)$$

$$\beta = \frac{\Delta^p}{\varepsilon p}, \quad p > 1, \quad (9)$$

where $z_* = \Delta \left[ \left( 1 + \frac{\varepsilon p}{\|d/p\|\ln(\|d/p\|/\delta)} \right)^{1/p} - 1 \right]^{-1}$, then $M$ satisfies the impacts of scaling skills.

**Proof.** Refer to Appendix B1.

Although Theorem 2 shows how to set $\beta$ for the required privacy protection, Eq. (9) gives a loose upper bound because of the relaxation in proof. As a special case, when $p = 2$, Theorem 2 provides a larger noise variance than Theorem 1. In particular, based on $z_*^2 \approx \Delta^2/[d/2] \ln((d/2)/\delta)/\varepsilon^2$ when $p = 2$. Therefore, the coordinate noise variance $\sigma^2 = \beta^2/2$ is about $2\Delta^2/[d/2] \ln((d/2)/\delta)/\varepsilon^2$ which saves a factor of $O(d \ln(d))$. The reason is that the probability inequality for proving general pEM is looser than that of the specific GM (Theorem 1 [27]). To further improve the privacy-utility tradeoff from the privacy loss perspective, we extend MA to the pEM family. This enables us to tightly track the privacy loss of pEM, ignoring the impacts of scaling skills.

**B. MA FOR pEM**

The section corresponds to box B in Figure 3. According to Eq. (6) and Eq. (7) in Section III-C, we have to compute two $d$-dimensional integrals $I(h, q, \tilde{g})$ and $I(h, q, \tilde{g}, h)$ to obtain the optimal $\varepsilon$. Obviously, it is inefficient or even prohibitive to compute these integrals due to the possibly high dimension $d$. To address it, we use high-dimensional polar transformation and properties of general gamma distribution to reduce it to a double integral. Recall that $\varepsilon$ is calculated as follows:

$$\varepsilon = \min_{\lambda \in \mathbb{Z}^+} \frac{T \alpha_M(\lambda; h(x)) + \ln(1/\delta)}{\lambda},$$

where $\alpha_M(\lambda; h(x)) = \max\{\ln I(h, h, q, \tilde{g}), \ln I(h, q, \tilde{g}, h)\}$. Therefore, The above equation is rewritten as

$$\varepsilon = \max\{\varepsilon_1, \varepsilon_2\},$$

where $\varepsilon_1 = \min_{\lambda \in \mathbb{Z}^+} \frac{T \ln I(h, h, q, \tilde{g}) + \ln(1/\delta)}{\lambda}$ and $\varepsilon_2 = \min_{\lambda \in \mathbb{Z}^+} \frac{T \ln I(h, q, \tilde{g}, h) + \ln(1/\delta)}{\lambda}$, and $h, q, \tilde{g}(x) = (1 - q)h(x) + q\tilde{g}(x - \tilde{g})$. Obviously, the above two optimization problems need to be solved to obtain $\varepsilon$. Each of them contains a $d$-dimensional integral. Fortunately, Theorem 3 shows that $I(h, q, \tilde{g}) > I(h, h, q, \tilde{g})$ holds for general pEM. This enables us only to compute $\varepsilon_2$ to obtain $\varepsilon$.

**Theorem 3.** For $I(h, h, q, \tilde{g})$ and $I(h, q, \tilde{g}, h)$ defined as Eq. (6) and Eq. (7) respectively, where $\lambda \in \mathbb{Z}^+$, we have

1. $I(h, q, \tilde{g}) > I(h, h, q, \tilde{g}) \geq 1$;
2. $I(h, q, \tilde{g}, h)$ is convex.

**Proof.** Refer to Appendix B2.

Based on the first conclusion of Theorem 3, the computation of $\varepsilon$ reduces to $\varepsilon_2$. From the second conclusion of Theorem 3, $\varepsilon_2$ is an unimodal/quasi convex expression. Therefore, there exists the single value $\lambda^*$ subject to $\varepsilon$ achieves the minimum, which can be solved through bisec tion method. Theorem 4 shows that the $d$-dimensional integral $I(h, q, \tilde{g}, h)$ will reduce to a double integral whenever $d \geq 2$.

**Theorem 4 (Double Integral for Computing Privacy Loss of pEM).** According to MA, privacy loss $\varepsilon$ of pEM after $T$-fold adaptive composition can be tracked by

$$\varepsilon = \min_{\lambda} T \ln I(h, q, \tilde{g}, h) + \ln(1/\delta), \quad (10)$$

where $\lambda$ is selected to minimize $\varepsilon$. Furthermore, $\varepsilon_2$ is an unimodal/quasi convex expression. Therefore, there exists the single value $\lambda^*$ subject to $\varepsilon$ achieves the minimum, which can be solved through bisection method. Theorem 4 shows that the $d$-dimensional integral $I(h, q, \tilde{g}, h)$ will reduce to a double integral whenever $d \geq 2$.

**Proof.** Refer to Appendix B3.

The double integral in Theorem 4 is used for tracking the privacy loss when $d \geq 2$. When $d = 1$, one can directly use the definite integral to track it. That is, $\varepsilon = \int I(h, q, \tilde{g}, h) + \ln(1/\delta)$ with

$$I(h, q, \tilde{g}, h) = \int_{\mathbb{R}} \left[ \frac{h_{\Delta}(x)}{h(x)} \right]^{\lambda} \lambda dx$$

$$= \int_{\mathbb{R}} \left[ 1 - q + \frac{h(x - \Delta)}{h(x)} \right]^{\lambda} \left[ (1 - q)h(x) + qh(x - \Delta) \right] dx$$

and $h(x) = 1/\alpha \exp(-\|x\|^p/\beta)$. Fig. 2(a) shows that, comparing with $p = 2$, the privacy loss $\varepsilon$ tracked by the above formula can be improved from about 1.32 to 0.36.

**C. MECHANISM SELECTION**

The section corresponds to box C in Figure 3. Based on Eq. (10), the computation of $\varepsilon$ depends on several parameters: dimension $d$, iteration number $T$, global sensitivity $\Delta$, noise variance $\sigma^2$, sampling ratio $q$, and failure probability $\delta$. We show how to select a better mechanism from pEM based on these parameters. Recall that “better” in this paper means achieving smaller value of $\varepsilon(T)$ under $\mathcal{E}(\|x\|^2)$. Furthermore, based on Lemma 2, $\mathcal{E}(\|x\|^2) \leq \sigma^2/(\alpha^2 p)$. Due to that

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
\[ \mathbb{E}[\|x\|^2] = \sum_{i=1}^{d} \mathbb{E}x_i^2 = d \mathbb{E}x_i^2 \quad \text{and} \quad \mathbb{E}x_i = 0 \]

The coordinate noise variance \( \sigma^2 \) is

\[ \sigma^2 = \frac{\beta^2/p \Gamma((d + 2)/p)}{d \Gamma(d/p)}. \] (11)

For convenience, we will use \( \sigma^2 \) to control \( \mathbb{E}[\|x\|^2] \) based on the relation \( \mathbb{E}[\|x\|^2] = d \sigma^2 \). Therefore, a better mechanism can be selected via the following process, as the bottom block (light green) of Figure 3 shows.

- Each \( \rho \text{EM} \) with \( p \in [p_0, p_1] \) and step size \( p_s \),
  1. Use Eq. (11) to compute \( \beta(p) = \sigma^2 \left[ \frac{d \Gamma(d/p)}{\Gamma(d+2)/p} \right]^{p/2} \), based on \( \sigma^2, \delta, \) and \( d \);
  2. Define density function \( h(x) \) based on \( \beta(p) \) and \( p \);
  3. Compute the total privacy loss based on Eq. (10), where \( \Delta = c \) and \( T = t \);
- Return the optimal \( p^* \text{EM} \) with \( p^* = \arg \min_p \{ \varepsilon(p) \} \).

Among parameters appeared in the process, \( q, \sigma^2, T \) can control the privacy-utility tradeoff. (1) A large sampling ratio \( q \) causes more privacy loss but a better utility. (2) A large noise variance \( \sigma^2 \) cause less privacy loss but a worse utility. (3) A large iteration number \( T \) causes more privacy loss but a better utility (Note that is not always true in practice due to overfitting on training dataset).

**Remark 1.** The above procedures are used for searching a better \((\varepsilon, \delta)\)-DP mechanism from \( \rho \text{EM} \), where \( p > 0 \) in general. Note that this is not contradictory to that \( \rho \text{EM} \) is \( \varepsilon \)-DP when \( 0 < p \leq 1 \). Because we can use MA to track privacy loss \( \varepsilon \) for any \( \rho \text{DP} \) mechanism with a given \( \delta \).

For example, if a task \( T \) has parameters as shown in Fig. 4, we can use 1.3EM to achieve a better tradeoff than GM, where dimension \( d = 20 \).

**D. COMPUTATIONAL METHOD**

The section corresponds to box D in Figure 3. Till now, we have shown how to track the privacy loss \( \varepsilon(p) \) for \( \rho \text{EM} \) and select a better \( p^* = \arg \min_p \{ \varepsilon(p) \} \). The only remaining problem in practice is how to efficiently generate noise from \( \rho \text{EM} \). We prove that the computational efficiency of \( \rho \text{EM} \) is the same as Gaussian distribution. That is, we can efficiently generate random numbers from \( \rho \text{EM} \) without extra complex computation. The formal statement is as follows.

**Theorem 5 (Computational Method).** Given two random variables, \( r \in \mathbb{R}^+ \) and \( x \in \mathbb{R}^d \), where \( r \) is drawn from the generalized gamma distribution \( \gamma(r) \propto r^{d-1} e^{-r^p/\beta} \) and \( x \) is drawn from Gaussian distribution \( \mathcal{N}(0, I_d) \), then, the random variable \( x \sim x/\|x\| \cdot r \) has a \( \rho \text{EM} \) distribution with parameters \( p \) and \( \beta \).

**Proof.** Refer to Appendix B4.

Therefore, one can generate noises as follows.

**Step 1.** Generate \( x \) from Gaussian distribution \( \mathcal{N}(0, I_d) \).

**Step 2.** Generate \( r \) from the generalized gamma distribution with density \( \gamma(r) = \frac{1}{\beta} r^{d-1} e^{-r^p/\beta} \) (Lemma 1).

**Step 3.** Rescale \( x \leftarrow x/\|x\| \cdot r \).

**V. ANALYSIS OF OPTIMAL MECHANISM**

Based on Fig. 2(a) and Fig. 4, we have known that GM is not the optimal. However, this lacks the theoretical support. This motivates us to analyze the optimality of GM. We first model it as a restricted optimization problem in Section V-A. Based on variation method, we then derive the necessary conditions for the optimal solution in Section V-B and finally prove the non optimality of GM in Section V-C.

**A. PROBLEM FORMULATION**

In this section, we analyze the optimal mechanism in the whole space of continuous density function, beyond \( \rho \text{EM} \). The reason is that even if we could prove the optimality of GM in \( \rho \text{EM} \), this is not sufficient for ensuring the optimality in the whole space. However, the opposite is true. Therefore, we model it as the following optimization (Problem 1) in the general expression. For convenience, let \( h_{q,\beta}(x) \equiv (1 - q) h(x) + q h(x - \tilde{\beta}) \).

**Problem 1 (Optimal Density Function).**

\[
\min_{h(x)} \varepsilon = \min_{\lambda \in \mathbb{Z}^+} \frac{T \alpha_M(\lambda; h(x)) + \ln(1/\delta)}{\lambda}
\]

\[ \text{s.t.} \quad \int_{\mathbb{R}^d} h(x) \, dx = 1, \]

\[ \int_{\mathbb{R}^d} \|x\|^2 h(x) \, dx = A, \]

\[ \alpha_M(\lambda; h(x)) = \max_{p, \delta} \ln \int_{\mathbb{R}^d} \left[ \frac{h(x)}{h_{q,\beta}(x)} \right]^\lambda \, dh(x). \]

In Problem 1, the objective function \( \varepsilon \) is presented as an optimization with respect to \( \lambda \in \mathbb{Z}^+ \), with three constraint conditions. The first is the constraint of the probability density and the second is about the model utility where \( A \) is a given constant value. The last is about the high-order moments information. Recall in Section III-C

\[ \alpha_M(\lambda; h(x)) = \max \{ \ln I(h, h_{q,\beta}), \ln I(h_{q,\beta}, h) \}, \]
where \( I(h, h_{\theta, \tilde{g}}) = \int_{\mathbb{R}^d} \left[ \frac{h(x)}{h_{\theta, \tilde{g}}(x)} \right]^\lambda h(x) \, dx \), and

\[
I(h_{\theta, \tilde{g}}, h) = \int_{\mathbb{R}^d} \left[ \frac{h_{\theta, \tilde{g}}(x)}{h(x)} \right]^\lambda h_{\theta, \tilde{g}}(x) \, dx.
\]

Therefore, Problem \( \text{I} \) can be further decomposed to two subsequent optimization problems. Assume \( \varepsilon^* \) is its optimal value, that is,

\[
\varepsilon^* = \min_{\lambda \in \mathbb{Z}^+} \min_{h(x)} T_{\alpha_M}(\lambda; h(x)) + \ln(1/\delta).
\]

Furthermore, Problem \( \text{I} \) is equivalent to \( \varepsilon^* = \min_{\lambda \in \mathbb{Z}^+} \varepsilon^*(\lambda) \), where \( \varepsilon^*(\lambda) = \max\{\varepsilon_{\lambda, 1}^*, \varepsilon_{\lambda, 2}^*\} \) and \( \varepsilon_{\lambda, i}^*, i = 1, 2 \) are two minima corresponding to the following two optimizations.

**Problem 2 (Sub Optimization 1 for Solving \( \varepsilon_{\lambda, 1}^* \)).**

\[
\min_{h(x)} \varepsilon_{\lambda}^1(h(x)) = \frac{1}{\lambda} \left( \ln \frac{1}{\delta} + T \ln \int_{\mathbb{R}^d} \left[ \frac{h(x)}{h_{\theta, \tilde{g}}(x)} \right]^\lambda h(x) \, dx \right)
\]

s.t. \( \int_{\mathbb{R}^d} h(x) \, dx = 1 \),

\( \int_{\mathbb{R}^d} \|x\|^2 h(x) \, dx = A \).

**Problem 3 (Sub Optimization 2 for Solving \( \varepsilon_{\lambda, 2}^* \)).**

\[
\min_{h(x)} \varepsilon_{\lambda}^2(h(x)) = \frac{1}{\lambda} \left( \ln \frac{1}{\delta} + T \ln \int_{\mathbb{R}^d} \left[ \frac{h_{\theta, \tilde{g}}(x)}{h(x)} \right]^\lambda h_{\theta, \tilde{g}}(x) \, dx \right)
\]

s.t. \( \int_{\mathbb{R}^d} h(x) \, dx = 1 \),

\( \int_{\mathbb{R}^d} \|x\|^2 h(x) \, dx = A \).

### B. NECESSARY CONDITIONS

The optimal mechanism \( h(x) \) must exist in the sets of Problems \( \text{I} \) and \( \text{II} \), which can be derived by solving two functional extreme value problems based on the variation method. With respect to Problems \( \text{I} \) and \( \text{II} \), we construct two augment Lagrange functions formulated as

\[
J_i(h(x)) = \varepsilon_{\lambda, i}^*(h(x)) + \mu_1 \left( \int_{\mathbb{R}^d} h(x) \, dx - 1 \right) + \mu_2 \left( \int_{\mathbb{R}^d} \|x\|^2 h(x) \, dx - A \right), i = 1, 2.
\]

Here, \( \mu_1 \) and \( \mu_2 \) are two coefficients to be determined. By letting the derivative of Eq. (13) be zero, we will obtain the necessary conditions for optimal solution to Problem \( \text{I} \).

**Theorem 6.** If a function \( h(x) \) is the optimal solution to Problem \( \text{I} \) then at least one of the following conditions holds.

\[
\mathbb{E}_{X \sim h(x)} \left[ (\lambda + 1) f^{\lambda}(x) - \lambda \left( (1 - q) f^{\lambda+1}(x) + q f^{\lambda+1}(x + \tilde{g}) \right) \right] + \mathbb{E}_{X \sim h_\theta(x)} f^\lambda(x) (\mu_1 + \mu_2 \mathbb{E}_{X \sim h(x)} \|x\|^2) = 0,
\]

\[
\mathbb{E}_{X \sim h_{\theta, \tilde{g}}(x)} \left[ (\lambda + 1) (1 - q) u^{\lambda}(x) + qu^{\lambda}(x + \tilde{g}) \right] - \mu u^{\lambda+1}(x) \right] + \mathbb{E}_{X \sim h_{\theta, \tilde{g}}(x)} u^{\lambda}(x) (\mu_1 + \mu_2 \mathbb{E}_{X \sim h_{\theta, \tilde{g}}(x)} \|x\|^2) = 0,
\]

where \( u(x) = f^{-1}(x) = h_{\theta, \tilde{g}}(x)/h(x) \) with \( h_{\theta, \tilde{g}}(x) = (1 - q)h(x) + q h(x - \tilde{g}) \), and \( \lambda, \mu_1, \mu_2 \) are coefficients in Lagrange functions Eq. (13) to be determined.

**Proof.** Refer to Appendix B5 □

Ideally, we want to obtain the closed form of \( h(x) \) and determine the corresponding coefficients \( \mu_1 \) and \( \mu_2 \). However, due to the concurrence of terms \( g, x - \tilde{g} \) and \( x + \tilde{g} \) in function \( f() \), as well as the concurrence of \( \mathbb{E}_x f^{\lambda}(\cdot) \) and \( \mathbb{E}_x f^{\lambda+1}(\cdot) \), it is hardly to obtain the closed expression of \( h(x) \) straightforwardly. Nevertheless, Theorem 6 still can give us two theoretical implications. On one hand, it enables us to search the optimal mechanism through numerical calculation. For example, one can construct a series of functions which converge to the solution of equations in Theorem 6 and generate noise from the convergence function to improve the tradeoff. On the other hand, we can use these necessary conditions to determine the optimality of some specfic mechanisms, such as GM.

### C. OPTIMALITY OF GM

In this section, we prove that GM is not the optimal mechanism, based on Theorems 3[6] and Lemma 3[6].

To address it, we need to verify that \( h(x) = \frac{1}{2} e^{-\|x\|^2/\beta} \) does not satisfy the both conditions in Theorem 6. Fortunately, Theorem 3[6] indicates that \( I(h_{\theta, \tilde{g}}, h) > I(h, h_{\theta, \tilde{g}}) \geq 1 \) for EM family. That is, it suffices to verify the second condition. Furthermore, Lemma 3[6] indicates that the sampling ratio \( q \) does not change the optimal mechanism. This is because for all mechanisms achieving \( (\varepsilon, \delta) \) with different \( \varepsilon \) and the fixed \( \delta \), the mechanisms after subsampling with probability \( q \) satisfy \((\log(1 + q(\varepsilon - 1)), q\delta)\)-DP. Therefore, sampling does not change the optimality. Then, substituting \( h(x) = 1/\alpha \exp(-\|x\|^2/\beta) \) into the second condition with \( q = 1 \) and via a straightforward calculation, the verification of GM finally reduces to verify whether equality \( e^{2\Delta^2/\beta} = 1 + 2\Delta^2/\beta \) holds. Because \( e^{2\Delta^2/\beta} \neq 1 + 2\Delta^2/\beta \) whatever \( \Delta^2/\beta \) is, we obtain the following conclusion.

**Theorem 7.** GM is not the optimal \( (\varepsilon, \delta) \)-DP mechanism which achieves the minimal privacy loss under the given model utility.

**Proof.** Refer to Appendix B6 □

Although GM is not optimal, it should note that \( e^{2\Delta^2/\beta} \approx 1 + 2\Delta^2/\beta \) when \( 2\Delta^2/\beta \approx 0 \). Therefore, when signal-to-noise ratio \( \Delta/\sigma \) is small, conditions in Theorem 6 approximately holds for Gaussian distribution. That is, GM is the nearly optimal when \( \Delta/\sigma \) is small.

### VI. PERFORMANCE EVALUATION

In the section, we conducted extensive experiments to validate properties of EM and give a comprehensive comparison with other state-of-the-art methods. We implemented the python codes of EM, and embedded them to the TensorFlow Privacy library by modifying the noise generating and privacy ledger modules. Our codes are publicly available[5] and can support DP-SGD based algorithms (Algorithm 1[1]), where the input parameters are defined in Table 2[4]. In all experiments, a better value \( p^* \) is selected by searching the interval \([1, 2]\) with step size 0.1 (i.e., \( p_0 = 1, p_1 = 2, p_s = 0.1 \) in Algorithm 1[1]), and the optimal value \( \lambda \in \mathbb{Z}^+ \)

https://github.com/private-mechanism/p_exponential_mechanism
As mentioned above, several parameters have impacts on privacy loss. As shown in Fig. 5, the parameters $\Delta$, $T$, $\sigma^2$, $q$, and $\epsilon$ when dimension $d$ is fixed as 20. We set $q = 10^{-3}$, $\delta = 10^{-5}$, $T = 10^4$, $\sigma^2 = 4^2$ as the baseline, and varied each of them in the range shown in Fig. 5. Overall, three conclusions are observed from Fig. 5. Firstly, $\epsilon$ increases with $q$, $T$ (Fig. 5(a) and Fig. 5(c)) but decreases with $\delta$, $\sigma^2$ (Fig. 5(b) and Fig. 5(d)). Fig. 5(a) is consistent with Lemma 3 that the larger $q$, the larger $\epsilon$. Figs. 5(b) and 5(c) are consistent with formula $\epsilon = \min_{\lambda} \frac{T \ln(4.3 \sqrt{\epsilon} + \lambda + 10)}{\min(\Delta, \epsilon)}$, where the numerator increases with $T$ and decreases with $\delta$. Fig. 5(d) is consistent with the definition of DP that a larger noise variance corresponds the stricter privacy guarantee (lower $\epsilon$). Secondly, there exists a better mechanism such as $p = 1, 2, 1.4$ which achieves significantly lower $\epsilon$ than GM (i.e., $p = 2$). For example, in Fig. 5(b) where $\Delta = 1$, the privacy loss is saved about 50%. Thirdly, the advantage in Figs. 5(b) and 5(d) is decreasing as the parameters (i.e., $\delta$, $T$, $\sigma^2$) increasing, especially in Figs. 5(c) and 5(b). The varying tending of Fig. 5(d) is in accordance with definition of DP. That is, when $\delta < 1$ is large, each DP mechanism can use a smaller $\epsilon$ to satisfy $\Pr[M(D) \in E] \leq e^\epsilon \Pr[M(D') \in E] + \delta$ and therefore the differences among mechanisms decrease. Fig. 5(c) can be explained by the idea of MA which exploits the moment information of privacy loss variable. As pointed out in [51], the variable will converge to the normal distribution as $T \to \infty$. Therefore, the differences will decrease. Fig. 5(d) is indicated by analysis of Theorem 7 that when $\Delta / \sigma$ is small, GM is the nearly optimal. Then all $p$EMs trend to the same level of the privacy loss.

### 3. COMPARISON RESULTS BETWEEN CGM, AGM, AND pEM

In this section, we compare $p$EM with the cGM, aGM, and MA (exchangeable denoted as 2EM). We illustrate a comprehensive comparison in Section VI-C1 where various several were considered. Then, we compare cGM/aGM/MA/pEM by conducting several machine learning tasks in Section VI-C2.

1) **Comprehensive Numerical Comparison**

For a given $(\epsilon_0, \delta_0)$ at each fold, due to the different tracking methods, cGM, aGM, and pEM obtain the different $(\epsilon, \delta)$ after $T$ folds. In particular, by applying advanced composition to cGM and aGM, both $\epsilon$ and $\delta$ vary with $T$. However, by using MA in pEM, only $\epsilon$ varies with $T$ but $\delta$ is fixed. We design the following pipeline to make a fair comparison. The main purpose is to fix $\delta$ for all cGM, aGM and pEM after $T$ folds, and we only need to compare $\epsilon$. In the subsection, $\delta$ was fixed as $10^{-4}$.

**Step 1.** Set $\delta_0$. For the given $\delta$ after $T$ folds and amplification ratio $q$, set $\delta_0 = \sqrt{2qT}/(2qT)$ for both cGM and aGM.

**Step 2.** Set $\sigma^2$. For the derived $\delta_0$, and given $\epsilon_0, \Delta$, set $\sigma^2_c$ and $\sigma^2_a$ for cGM and aGM, based on $\sigma_c = \sqrt{2 \ln(1.25/\delta_0)} \Delta / \epsilon_0$ and $\sigma_a = \Phi \left( \frac{\Delta - \epsilon_0^2}{\sqrt{\sigma_c^2}} \right) - e^{\delta_0} \Phi \left( \frac{-\Delta - \epsilon_0^2}{\sqrt{\sigma_a^2}} \right) \leq \delta_0$.

**Step 3.** Track $\epsilon$. Based on $\sigma_c, \sigma_a,$ and $q$, using advanced composition and Lemma 3 to track $\epsilon$ for cGM and for aGM. Meanwhile, based on $\sigma_a$ and $\delta$, using MA to track $\epsilon$ for pEM.

Note that $\sigma_c$ and $\sigma_a$ vary with $T$. Table 4 shows the range of $\sigma_a$ when $T$ increases up to 10,000 with step 200. From the vertical, $\sigma_a$ increases with $\epsilon_0$. From the horizontal, $\sigma_a$ decreases with $q$. Focusing on each cell, $\sigma_a$ increase with $T$. This indicates that cGM, aGM,
and \( pEM \) for all settings in Table 4. Some intuitively correct results are observed from Fig. 6. (1) From left to right, \( \varepsilon \) decreases with the amplification ratio \( q \). (2) In each of plots, \( \varepsilon \) increases with \( \varepsilon_0 \) and \( T \). The underlying reasons have been interpreted in Section VI-B.

Besides, we present a comprehensive comparison of cGM, aGM, and \( pEM \). Firstly, within cGM and aGM, the latter outperforms the former all along. Therefore, we can drop cGM from the comparison. Secondly, within \( pEM \) family, there exists a better mechanism than 2EM when \( \varepsilon_0 = 3 \), 1. In such case where \( \sigma/\Delta \) ranges about 1 to 5 (rows 1-2 of Table 4), we can use the \( pEM \) with \( p \neq 2 \) to replace 2EM. Thirdly, between aGM and \( pEM \), it is observed that aGM outperforms \( pEM \) when \( T \) and \( \varepsilon_0 \) are small. For example, in Fig. 6(c) when \( \varepsilon_0 = 1, 0.1, 0.01 \), the curves of aGM are lower than \( pEM \) when \( T \) is near coordinates \((0, 0)\). Elsewhere, \( pEM \) outperforms aGM always. Note that the three red boxes in Figs. 6(c) and 6(d) indicate that \( pEM \) is significantly worse than aGM. The reason is that we fix \( \lambda \) in a fixed range \([1, \ldots, 32]\) in all cases. In fact, when \( q \) is diminishing, \( \lambda \) corresponding to the minimal \( \varepsilon \) should increase up to infinity. However, we note that these cases have little practical significance. As red cells in Table 4 show, in such case, the noise-to-signal ratio \( \Delta/\sigma \) is too large to destroy the outputting utility in practice. Finally, we illustrate a sketch map about the application scope of aGM and \( pEM \) in Fig. 2(b), which shows that \( pEM \) is a preferred choice when \( \varepsilon_0 \) and \( T \) are moderate.

### TABLE 4. The range of noise-to-signal ratio under different settings of sampling ratio \( q \) and initial privacy budget \( \varepsilon_0 \).

| \( \varepsilon_0 \) | \( q = 1 \) | \( q = 0.1 \) | \( q = 0.01 \) | \( q = 0.001 \) |
|------------------|-----------|-------------|--------------|--------------|
| 3                | 1.27-1.86 | 1.09-1.72   | 0.89-1.58    | 0.65-1.45    |
| 1                | 3.36-5.22 | 2.76-4.80   | 2.09-4.36    | 1.33-3.88    |
| 0.1              | 26.47-47.25 | 19.64-37.86 | 11.93-37.86 | 4.48-32.48 |
| 0.01             | 194.98-246.65 | 117.84-377.67 | 43.76-323.75 | 7.30-263.45 |

### TABLE 5. Settings of noise variance and iteration number in Figure 7.

| Dataset | Model | \( cEM \), aGM | \( 2EM \), pEM |
|---------|-------|---------------|---------------|
| SAHeart | LR    | 500, 6.2, 2.4 | 1.6           |
|         | SVM   | 3409, 6.8, 2.6 | 1.6           |
| Adult   | LR    | 500, 6.7, 2.8 | 1.4           |
|         | SVM   | 3409, 6.9, 2.6 | 1.4           |
|         | 2EM   | 1950, 1990    |               |
|         | 3EM   | 700, 1100     |               |
|         | SVM   | 3409, 6.9, 2.6 | 1.4           |

2) Machine Learning Accuracy Comparison

In this section, we applied cGM, aGM, and \( pEM \) on two datasets, Adult and SAHeart. Each of datasets was trained by logistic regression (LR) and support vector machine (SVM) 50 times, with regularized coefficient \( \rho \) = 0.001 and learning rate 0.01. The purpose is to protect the training process, where noises were added into the gradients. Experimental settings are as follows.

- [http://archive.ics.uci.edu/ml](http://archive.ics.uci.edu/ml)
- [http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart](http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart)

- [http://www.stat.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart](http://www.stat.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart)
• Privacy loss \((\varepsilon, \delta)\). When using SVM method, we aimed to achieve \((1, 10^{-4})\)-DP and \((1, 10^{-2})\)-DP for Adult and SAHeart, respectively. When using LR method, the privacy loss was set as \((0.5, 10^{-4})\) for Adult and \((0.5, 10^{-2})\)-DP for SAHeart.
• Global sensitivity \(\Delta\). For SVM method, \(\Delta\) was set as 1 via clipping gradient with threshold 1. For LR method, \(\Delta\) was set as 1/4 + 0.001 based on that the smoothness constant is 1/4 + \(\rho\).
• Sampling ratio \(q\). For Adult dataset, \(q\) was set as 0.0044 (batch size 128 to dataset size 29304). For SAHeart, it was set as 0.01 (batch size 5 to data size 462).
• Noise variance \(\sigma^2\) and iteration number \(T\). In each of four settings (i.e., LR on SAHeart, SVM on SAHeart, LR on Adult, and SVM on Adult), we compared the proposed \(p\)EM with cGM and aGM. To ensure the same privacy loss in each setting, \(\sigma^2/\Delta\) and \(T\) were set in Table 5. The settings are based on method used in numerical comparison section (i.e., Section VI-C1).

Fig. 7 shows the average test accuracy with respect to privacy loss \(\varepsilon\), which denotes the impact of accumulated noise added to gradient. Experimental results are almost consistent with the numerical comparison in Section VI-C1 and three conclusions are observed.

Firstly, aGM outperforms cGM in most cases. This is because aGM uses a smaller variance than cGM under the given privacy loss \(\varepsilon\). Especially in Figs. 7(a) and 7(b) due to the large noise variance, cGM has a fast increasing phase which followed by a fluctuated plateau. Secondly, there exist a better mechanism 1.4EM which achieves the higher test accuracy than 2EM. Especially in Fig. 7(a) the smaller noise variance ensures 1.4EM has the faster and better performance. In other three plots, 1.4EM also has the faster training speed than 2EM but with same final utility. Thirdly, \(p\)EM outperforms aGM in overall. This is because we use MA to track the privacy loss in \(p\)EM. The tighter estimate of \(\varepsilon\) ensures more iterations for \(p\)EM than aGM. However, in Fig. 7(d) aGM has a faster training speed than \(p\)EM. This is mainly because the task can be efficiently trained, even with small total privacy loss.

VII. CONCLUSION

In the paper, we clearly claim that GM is not the optimal \((\varepsilon, \delta)\)-DP and propose a general \(p\)EM family to improve the tradeoff of GM. However, two challenges emerge when we make a full analysis of \(p\)EM. One is how to extend MA for tracking privacy of \(p\)EM, and the other is how to prove the non-optimality of GM. We use the high-dimensional poplar transformation and properties of gamma distribution to solve the first challenge, and use the variation method to solve a derived restricted optimization problem of the second challenge. Besides the above theoretical analysis, we present how to apply \(p\)EM in practice and show that one can efficiently generate noises from \(p\)EM the same as generating normal noises. We conducted extensive numerical experiments to validate the properties of \(p\)EM and show that \(p\)EM is preferred when ratio of noise variance to signal is relatively small and the dimension is not high. In contrast, the improvement of \(p\)EM is slight. The reason is that when the ratio is large, GM is proved to be the nearly optimal mechanism, therefore, \(p\)EM cannot obtain a significant improvement. On the other hand, when dimension is too high, \(p\)EM with different \(p\) has similar moments information, therefore, \(p\)EM performs similarly to GM. In future, we will develop \(p\)EM from two aspects. In theory, we will analyze the impacts of parameters on the privacy loss and aim to find a closed solution of the optimal \(p\)EM. In practice, we will apply \(p\)EM to federated learning to improve the existing privacy-utility tradeoff of GM.

REFERENCES

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in Proc. of CVPR. IEEE, 2014, pp. 1701–1708.
[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. of CVPR. IEEE, 2016, pp. 770–778.
[3] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” nature, vol. 521, no. 7553, p. 436, 2015.
[4] P. Mohassel and Y. Zhang, “Secureml: A system for scalable privacy-preserving machine learning,” in IEEE Security Privacy. IEEE, 2017, pp. 19–38.
[5] L. Zhao, Q. Wang, Q. Zou, Y. Zhang, and Y. Chen, “Privacy-preserving collaborative deep learning with unreliable participants,” IEEE Trans. Inf. Theory, pp. 1–1, 2019.
[6] M. Nasr, R. Shokri, and A. Houmansadr, “Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against
centralized and federated learning,” in *IEEE Security Privacy*. IEEE, 2019, pp. 739–753.

[7] N. Carlini, C. Liu, J. Kos, Ů. Erlingsson, and D. Song, “The secret sharer: Measuring unintended neural network memorization & extracting secrets,” *arXiv preprint arXiv:1802.08222*, 2018.

[8] M. Fredrikson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” in *Proc. of CCS*. New York, NY, USA: ACM, 2015, pp. 1322–1333.

[9] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Mori, “Privacy-preserving deep learning via additively homomorphic encryption,” *IEEE Trans. Inf. Theory*, vol. 13, no. 5, pp. 3333–3348, 2018.

[10] J. Sharma, D. Kim, A. Lee, and D. Seo, “On differential privacy-based framework for enhancing user data privacy in mobile edge computing environment,” *IEEE Access*, vol. 9, pp. 38 107–38 118, 2021.

[11] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. “Deep learning with differential privacy,” in *Proc. of CCS*. New York, NY, USA: ACM, 2016, pp. 308–318.

[12] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. “Learning differentially private recurrent language models,” in *Proc. of ICLR*. OpenReview.net, 2018.

[13] G. Qiu and Y. Shen, “Mobility-aware differentially private trajectory for privacy-preserving continuous crowdsourcing,” *IEEE Access*, vol. 9, pp. 26 362–26 376, 2021.

[14] J. Zhang, K. Zheng, W. Mou, and L. Wang, “Efficient private erm for smooth objectives,” in *Proc. of ICALI*. ijcai.org, 2017, pp. 3922–3928.

[15] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, “Differentially private empirical risk minimization,” *JMLR*, vol. 12, no. Mar, pp. 1069–1109, 2011.

[16] D. Yu, H. Zhang, W. Chen, T.-Y. Liu, and J. Yin, “Gradient perturbation is underrated for differentially private convex optimization,” *arXiv preprint arXiv:1911.11363*, 2019.

[17] H. B. McMahan and G. Andrew, “A general approach to adding differential privacy to iterative training procedures,” *arXiv preprint arXiv:1812.06210*, 2018.

[18] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, “Semi-supervised knowledge transfer for deep learning from private training data,” *arXiv:1610.05755*, 2016.

[19] O. Thakkar, G. Andrew, and H. B. McMahan, “Differentially private learning with adaptive clipping,” *arXiv preprint arXiv:1905.03871*, 2019, 2019.

[20] V. Pichapati, A. T. Suresh, F. X. Yu, S. J. Reddi, and S. Kumar, “Adaclip: Adaptive clipping for private sgd,” *arXiv preprint arXiv:1908.07643*, 2019.

[21] C. Dwork, K. Talwar, A. Thakurta, and L. Zhang, “Analyze gauss: Optimal bounds for privacy-preserving principal component analysis,” *Proc. of STOC*, pp. 11–20, 05 2014.

[22] Y. Yan, L. Zhang, Q. Z. Sheng, B. Wang, X. Gao, and Y. Cong, “Dynamic release of big location data based on adaptive sampling and differential privacy,” *IEEE Access*, vol. 7, pp. 164 962–164 974, 2019.

[23] X. Zhang, M. M. Khalili, and M. Liu, “Improving the privacy and accuracy of admm-based distributed algorithms,” in *Proc. of ACM ICML*, 2018.

[24] D. Lv and S. Zhu, “Achieving correlated differential privacy of big data publication,” *Computers & Security*, vol. 82, pp. 184–195, 2019.

[25] B. Wang, Q. Gu, M. Boedihardjo, L. Wang, F. Barekat, and S. J. Osher, “Dp-lssgd: A stochastic optimization method to lift the utility in privacy-preserving erm,” in *Mathematical and Scientific Machine Learning*. PMLR, 2020, pp. 328–351.

[26] B. Balie, G. Barthe, and M. Gaboardi, “Privacy amplification by subsampling: Tight analyses via couplings and divergences,” in *Proc. of NeurIPS*, 2018, pp. 6277–6287.

[27] C. Dwork, A. Roth et al., “The algorithmic foundations of differential privacy,” *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[28] I. Mironov, “Rényi differential privacy,” in *Proc. of IEEE CSF*. IEEE, 2017, pp. 263–275.

[29] C. Dwork and G. N. Rothblum, “Concentrated differential privacy,” *arXiv eprint 1603.01887*, 2016.

[30] K. Nissim, S. Rashidnikova, and A. Smith, “Smooth sensitivity and sampling in private data analysis,” in *Proc. of STOC*. ACM, 2007, pp. 75–84.

[31] Y. Li, S. Yang, X. Ren, and C. Zhao, “Asynchronous federated learning with differential privacy for edge intelligence,” *arXiv:1912.07902*, 2019.
YANAN LI received his in Bachelor and Master degree from Henan Normal University of China in 2004 and 2007, respectively. He is currently working towards the PhD degree in the School of Mathematics and Statistics at Xi’an Jiaotong University. Before that, he worked as a lecturer in Henan Polytechnic University from 2007 to 2017. His research interests include differential privacy, edge-cloud intelligence.

XUEBIN REN received his PhD degree in the Department of Computer Science and Technology from Xi’an Jiaotong University, China in 2017. Currently, he is an associate professor in Xi’an Jiaotong University and a member of National Engineering Laboratory for Big Data Analytics (NEL-BDA). He has been a visiting PhD student in the Department of Computing at Imperial College London from 2016 to 2017. His research interests focus on data privacy protection, federated learning and privacy-preserving machine learning. He is a member of the IEEE and the ACM.

FANGYUAN ZHAO received his Bachelor’s degree in the School of Mathematics and Statistics from Xi’an Jiaotong University, China in 2018. Currently, he is working towards the PhD degree in the School of Computer Science and Technology at Xi’an Jiaotong University. His research interests include text analysis, privacy-preserving machine learning, federated learning and fair machine learning.

SHUSEN YANG received his Ph.D. degree in Computing from Imperial College London in 2014. He is a professor and director of the National Engineering Laboratory for Big Data Analytics, and deputy director of Ministry of Education (MoE) Key Lab for Intelligent Networks and Network Security, both at Xi’an Jiaotong University (XJTU), Xi’an, China. He is also the deputy director of the Industrial Artificial Intelligence Center at Pazhou Laboratory, Guangzhou, China.

Before joining XJTU, He worked as a lecturer (assistant professor) at the University of Liverpool from 2015 to 2016, and a research associate at Intel Collaborative Research Institute (ICRI) on sustainable connected cities from 2013 to 2014. He is a DAMO Academy Young Fellow, and an honorary research fellow at Imperial College London. He is a senior member of IEEE and a member of ACM. His research focuses on distributed systems and data sciences, and their applications in industrial scenarios, including data-driven network algorithms, distributed machine learning, Edge-Cloud intelligence, industrial internet and industrial intelligence.

APPENDIX. SUPPLEMENTARY MATERIAL FOR THEORETICAL CONCLUSIONS

A. THREE LEMMAS

In this section, we propose two lemmas (Lemma 1, 2) about the proposed pEM and adopt a lemma (Lemma 3) about the privacy loss amplification after sampling. In particular, for pEM with density function \( h(x) = 1/\alpha e^{-||x||^p/\beta} \), Lemma 1 shows the density function about \( ||x|| \) and Lemma 2 shows the result about \( \mathbb{E}||x||^2 \). In proof of both Lemma 1 and Lemma 2, the following high-dimensional polar transformation is necessary. When \( x = (x_1, \cdots, x_d)^T \in \mathbb{R}^d \), let

\[
\begin{align*}
x_1 &= r \cos(\theta_1) \\
x_2 &= r \sin(\theta_1) \cos(\theta_2) \\
\vdots \\
x_{d-1} &= r \sin(\theta_1) \cdots \sin(\theta_{d-1}) \cos(\theta_{d-1}) \\
x_d &= r \sin(\theta_1) \cdots \sin(\theta_{d-1}) \sin(\theta_{d-1})
\end{align*}
\]

(14)

where \( \theta_i \in [0, \pi], i = 1, \cdots, d-2, \theta_{d-1} \in [0, 2\pi] \), and the Jacobin determinant is

\[
J(d) = r^{d-1} \sin(\theta_1)^{d-2} \sin(\theta_2)^{d-3} \cdots \sin(\theta_{d-1}).
\]

Furthermore, we denote the superficial area of unit sphere in \( \mathbb{R}^d \), i.e., \( \int_0^\pi d\theta_1 \cdots \int_0^\pi d\theta_{d-2} \int_0^{2\pi} \frac{r^d}{r^d} d\theta_{d-1} \), as \( A(d) \) for simplicity.

**Lemma 1.** If \( x \in \mathbb{R}^d \) has the density function \( h(x) = 1/\alpha e^{-||x||^p/\beta} \), then variable \( r = ||x|| \) follows the generalized gamma distribution with the density function \( \gamma(r; k, \beta, p) = 1/Nr^{kp-1}e^{-r^p/\beta}, r > 0 \), where \( N \) is the normalization with \( 1/N = \frac{p/\beta^k}{\Gamma(k)} \) and \( k = d/p \).

**Proof.** The cumulative probability distribution of \( ||x|| \) is (when \( z > 0 \))

\[
F_{||x||}(z) = \Pr(||x|| \leq z) = \int_{||x|| \leq z} \frac{1}{\alpha} e^{-||x||^p/\beta} dx.
\]

Based on Eq. [14], we have

\[
\begin{align*}
\int_{||x|| \leq z} \frac{1}{\alpha} e^{-||x||^p/\beta} dx &= \int_0^z \frac{1}{\alpha} e^{-r^p} J(d) dr \int_0^\pi d\theta_1 \cdots \int_0^\pi d\theta_{d-2} \int_0^{2\pi} \frac{r^d}{r^d} d\theta_{d-1} \\
&= \int_0^z \frac{1}{\alpha} e^{-r^p} r^{d-1} dr \cdot A(d).
\end{align*}
\]

That is, \( F_{||x||}(z) = \frac{A(d)}{\alpha} \int_0^z e^{-r^p} r^{d-1} dr. \) Taking derivative with respect to \( z \), we deduce the probability density function \( f_{||x||}(z) = \frac{dF_{||x||}(z)}{dz} = \frac{A(d)}{\alpha} e^{-z^p/\beta} z^{d-1}. \) Replacing \( z \) with \( r \), we further obtain

\[
f_{||x||}(r) = \frac{A(d)}{\alpha} r^{d-1} e^{-r^p/\beta}.
\]

Comparing \( f_{||x||}(r) \) with the density of generalized gamma distribution \( \gamma(r; k, \beta, p) = 1/Nr^{kp-1}e^{-r^p/\beta}, r > 0 \), where \( N \) is the normalization with \( 1/N = \frac{p/\beta^k}{\Gamma(k)} \), we deduce that...
Theorem 4: If the random variable $x$ has the probability density function $h(x) = \frac{1}{\alpha} e^{-|x|^p/\alpha}$, $x \in \mathbb{R}^d$, where $\alpha$ is the normalization, then the $m$-th order moment is

$$E\|x\|^m = \frac{\beta^m}{\alpha} \Gamma(m + d/\alpha)/\Gamma(d/\alpha).$$

Proof. Based on Eq. (14), we have

$$E\|x\|^m = \int_{\mathbb{R}^d} \|x\|^m \frac{1}{\alpha} e^{-|x|^p/\alpha} dx = \frac{\Gamma(d/\alpha)}{\Gamma(d/\alpha)} \int_{\mathbb{R}^d} r^m e^{-r^p} dr.$$

Note that $\int_{\mathbb{R}^d} r^m e^{-r^p} dr = \frac{\beta^m}{\alpha} \Gamma(d+m/\alpha)$, we obtain

$$E\|x\|^m = \frac{\Gamma(d/\alpha)}{\Gamma(d/\alpha)} \frac{\beta^m}{\alpha} \Gamma(d+m/\alpha).$$

At the end of Lemma 1, we have proved $A(d) = \frac{\Gamma(d/\alpha)}{\Gamma(d/\alpha)} \frac{\beta^m}{\alpha} \Gamma(d+m/\alpha)$, where $k = d/p$. Then, substituting $A(d) = \frac{\Gamma(d/\alpha)}{\Gamma(d/\alpha)} \frac{\beta^m}{\alpha} \Gamma(d+m/\alpha)$ into Eq. (16), we obtain Eq. (15).

Lemma 3: If $M$ is $(\varepsilon, \delta)$-DP, then $M'$ that applies $M$ after subsampling with probability $q$ obeys $(\varepsilon', \delta')$-DP with $\varepsilon' = \log(1 + q(e^{\varepsilon} - 1))$ and $\delta' = q\delta$.

B. PROOFS OF THEOREMS

1) Proof of Theorem 2

Proof. According to the definition of DP, we need to verify that privacy loss variable $c_M(o; D, D') = \left| \ln \frac{Pr(f(D) = x_i | o) - Pr(f(D') = x_i | o)}{Pr(f(D) = x_i | o) + Pr(f(D') = x_i | o)} \right| \leq \varepsilon$ holds for any possible output $o \in$ Range($M$) $\subseteq \mathbb{R}^2$, where $M : f(\cdot) \rightarrow f(\cdot) + x$ with $x$ drawn from $pEM$. That is, we need to verify

$$c_M(o; D, D') = \frac{\|o - f(D)\|^p - \|o - f(D')\|^p}{\beta} \leq \varepsilon$$

holds, where $v = f(D') - f(D)$.

Case 1). When $0 < p \leq 1$, based on the triangle inequality, $c_M(o; D, D')$ in Eq. (17) satisfies

$$c_M(o; D, D') \leq \frac{\|x + v\|^p - \|x\|^p}{\beta} \leq \frac{\|v\|^p}{\beta} \leq \frac{\Delta^p}{\beta}.$$

Therefore, if we set $\beta = \Delta^p/\varepsilon$, then $M$ is $\varepsilon$-DP. To deduce the second inequality, we use the fact that $f(x) = x^p, x \geq 0$ is a concave function when $0 < p \leq 1$.

Case 2) When $p > 1$, based on triangle inequality, the privacy loss $c_M(o; D, D')$ also satisfies $c_M(o; D, D') \leq \frac{\|x + v\|^p - \|x\|^p}{\beta}$. Therefore, if Eq. (18) and Eq. (19) hold when $z_*>0$, then $M$ is $(\varepsilon, \delta)$-DP.

$$z_*>0, then M is (\varepsilon, \delta)$-DP.

Based on Lemma 1, the probability density function of $\|x\|$ is $f_{|x|}(z) = \frac{p^{p/\beta} e^{-z^p/\beta}}{\Gamma(d/\alpha)}$, $z > 0$. We first estimate $z_*$ by Eq. (19) and then calculate $\beta$ by Eq. (18).

Based on Eq. (19), we have

$$Pr(\|x\| \leq z_*) = \int_0^{z_*^p} f_{|x|}(z) dz = \int_0^{z_*^p} \frac{p^{p/\beta} e^{-z^p/\beta}}{\Gamma(d/\alpha)} du.$$

Note that the probability density function of $Gamma(d/\alpha, 1)$ just is $\frac{p^{p/\beta}}{\Gamma(d/\alpha)}$. Based on the tail estimation, $Pr(\|X\| \leq k\theta \ln(k/\delta)) \geq 1 - \delta$ if $X \sim Gamma(d/\alpha, 1)$, we have

$$Pr(\|X\| \leq z_*) = \int_0^{z_*^p} f_{|X|}(z) dz = \int_0^{z_*^p} \frac{p^{p/\beta} e^{-z^p/\beta}}{\Gamma(d/\alpha)} du.$$

2) Proof of Theorem 3

Proof. Case 1) We first prove $I(h, h_{q, g}) \geq 1$ and then $I(h, h_{q, g}) \geq I(h, h_{q, g})$. For the former, based on Eq. (6), i.e., $I(h, h_{q, g}) = E_{x \sim h(x)} \left[ \frac{h(x)}{h_{q, g}(x)} \right] \geq 1$.

$$E_{x \sim h(x)} \left[ \frac{1}{1 - q + qh_{q, g}(x)/h(x)} \right] \geq 1.$$
To differentiate scale and vector, we rewrite \( x = x \) and \( \hat{g} = \Delta \). For an intuitive cognition, we combine Fig. 8 to complete the prove. Uniformly splitting the region \((-\infty, \Delta/2]\) and \([\Delta/2, +\infty)\) with length 1, we obtain
\[
I(h(x), h_{\Delta}(x)) = \lim_{i \to 1} \sum_{i=0}^{\infty} \left( \frac{1}{1 - q + qa_i} \right)^\lambda A_i + \left( \frac{1}{1 - q + qa_i} \right)^\lambda A_i,
\]
where \( a_i = h_{\Delta}(x_i)/h(x_i) \leq 1 \) and \( 1/a_i = h_{\Delta}(x_i)/h(x_i) \geq 1 \), \( A_i \) and \( A_i' \) are probabilities of \( h(x) \) on interval that contain \( x \) and \( x' \) respectively. Result about \( I(h_{\Delta}(x), h(x)) \) can be similarly obtained. Next, we prove that \( I(h_{\Delta}(x), h(x)) \geq I(h(x), h_{\Delta}(x)) \). For simplicity, we ignore variable \( x \) and denote \( f(a_i) \overset{\Delta}{=} (1 - q + qa_i)^\lambda \). Then, we have
\[
I_{gm}(h, h_{\Delta}) = f(a_i) \left[ (1 - q)A_i + qA_i \right] + f(a_i') \left[ (1 - q)A_i + qA_i' \right],
\]
\[
I_{gm}(h, h_{\Delta}) = f^{-1}(a_i)A_i + f^{-1}(1/a_i)A_i.
\]
To prove \( I_{gm}(h, h_{\Delta}) > I_{gm}(h, h_{\Delta}) \), it reduces to
\[
\begin{align*}
\left\{ (1 - q)f(a_i) + qf(1/a_i) > 1/f(a_i), \\
qf(a_i) + (1 - q)f(1/a_i) > 1/f(1/a_i).
\end{align*}
\]
Replacing \( a_i \) with \( 1/a_i \) in Eq. (23), we have
\[
\begin{align*}
\left\{ (1 - q)f(1/a_i) + qf(a_i) > 1/f(a_i), \\
qf(1/a_i) + (1 - q)f(a_i) > 1/f(a_i).
\end{align*}
\]
Taking summation of Eq. (23) and Eq. (24), it reduces to
\[
f(a_i) + f(1/a_i) > 1/f(a_i) + f(1/a_i) = f(a_i) + f(1/a_i).
\]
Therefore, it is equivalent to prove \( f(a_i) + f(1/a_i) \geq 1 \), based on definition of \( f(a_i) \) and \( \lambda \geq 1 \), this further reduces to prove
\[
g(a_i) \overset{\Delta}{=} (1 - q + qa_i)(1 - q + qa_i) > 1, 0 < a_i \leq 1.
\]
Taking derivative of \( g(a_i) \) with respect to \( a_i \), we have
\[
g'(a_i) = q(1 - q + qa_i) - (1 - q + qa_i)1/a_i^2 = (q - 1)1/a_i^2.
\]
Because the sampling ratio \( q \in [0, 1] \) and \( a_i \in (0, 1] \), then \( g'(a_i) \geq 0 \) and hence \( g(a_i) \) is increasing with \( a_i \). Based on definition of \( g(1) \), we have \( g(1) = 1 \). Then, \( g(a_i) \geq 1 \) and the equality holds only if \( a_i = 1 \). Therefore, \( I_{gm}(h, h_{\Delta}) \geq I_{gm}(h, h_{\Delta}) \) and then \( I(h_{\Delta}, h) \geq I(h, h_{\Delta}) \).

When \( x \in \mathbb{R}^d, d > 1 \), the proof is similar except replacing the scale \( \Delta \) with the vector \( \hat{g} \) satisfying \( ||\hat{g}|| = \Delta \), and replacing symmetric points (with respect to \( x = \Delta/2 \)) \( x, \hat{x} \in \mathbb{R}^d \) with symmetric points \( x' \in \mathbb{R}^d \) with respect to \( ||x'|| = \Delta/2 \).

**Case 2)** Based on Eq. (7), i.e., \( I(h_{\hat{g}}, h) = \mathbb{E}_{x \sim h_{\hat{g}}(x)} \left[ 1 - q + q \frac{h(x - \hat{g})}{h(x)} \right]^{\lambda} \), we have
\[
I(h_{\hat{g}}, h) = \mathbb{E}_{x \sim h_{\hat{g}}(x)} \sum_{i=0}^{\lambda} \left( \frac{\lambda}{i} \right) (1 - q)^{\lambda - i} q^{i} \left[ \frac{h(x - \hat{g})}{h(x)} \right]^{i}.
\]
Furthermore, \( \mathbb{E}_{x \sim h_{\hat{g}}(x)} \left[ \frac{h(x - \hat{g})}{h(x)} \right]^{i} \) can be written as
\[
(1 - q)^{i} \mathbb{E}_{x \sim h_{\hat{g}}(x)} \left[ \frac{h(x - \hat{g})}{h(x)} \right]^{i} + q \mathbb{E}_{x \sim h_{\hat{g}}(x)} \left[ \frac{h(x - \hat{g})}{h(x)} \right]^{i - 1},
\]
where \( \mathbb{E}_{x \sim h_{\hat{g}}(x)} \left[ \frac{h(x - \hat{g})}{h(x)} \right]^{i} = \mathbb{E}_{x \sim h_{\hat{g}}(x)} \left[ \frac{h(x - \hat{g})}{h(x)} \right]^{i - 1} \). Then, \( I(h_{\hat{g}}, h) \) can be expressed as a combination of \( \mathbb{E}_{x \sim h_{\hat{g}}(x)} \left[ \frac{h(x - \hat{g})}{h(x)} \right]^{i} \). Due to \( \ln \mathbb{E}_{x \sim h_{\hat{g}}(x)} \left[ \frac{h(x - \hat{g})}{h(x)} \right]^{i} \) is convex function and \( \lambda \in \mathbb{R} \) (refer to Lemma 36 in [49]), so function \( \exp(\cdot) \) is also convex, then \( I(h_{\hat{g}}, h) \) is convex.

3) Proof of Theorem 4

Proof. For greater clarity, we recall several expressions at the beginning.

- \( \varepsilon_1 = \min_{\lambda \in \mathbb{R}} \frac{\lambda}{T} \ln I(h, h_{\hat{g}}) \) denotes privacy loss tracked by MA when considering \( I(h, h_{\hat{g}}) \);
- \( \varepsilon_2 = \min_{\lambda \in \mathbb{R}} \frac{\lambda}{T} \ln I(h, h_{\hat{g}}) \) denotes privacy loss tracked by MA when considering the contrary case \( I(h_{\hat{g}}, h) \);
- \( \varepsilon = \max\{\varepsilon_1, \varepsilon_2\} \) denotes the final privacy loss tracked by MA;
- \( I(h, h_{\hat{g}}) = \int_{\mathbb{R}^d} \left[ \frac{h(x)}{h_{\hat{g}}(x)} \right]^{\lambda} h'(x)dx \) is the integral used to calculate \( \varepsilon_1 \);
- \( I(h_{\hat{g}}, h) = \int_{\mathbb{R}^d} \left[ \frac{h_{\hat{g}}(x)}{h(x)} \right]^{\lambda} h_{\hat{g}}(x)dx \) is the integral used to calculate \( \varepsilon_2 \).

Now, we begin to prove the required conclusions.

First, we prove that Eq. (10) of Theorem 4 holds. That is, \( \varepsilon_2 > \varepsilon_1 \). Since Theorem 3 shows \( I(h_{\hat{g}}, h) > I(h, h_{\hat{g}}) \), we can obtain that \( \varepsilon_2 > \varepsilon_1 \). Then, \( \varepsilon = \max\{\varepsilon_1, \varepsilon_2\} = \varepsilon_2 \).

Second, we prove the accurate expression of \( I(h_{\hat{g}}, h) \) shown in Theorem 4. Recall that \( h_{\hat{g}}(x) \) is defined as \( h_{\hat{g}}(x) = (1 - q)h(x) + qh(x - \hat{g}) \). Based on that
\[
I(h, h_{\hat{g}}) = \int_{\mathbb{R}^d} \left[ \frac{h(x)}{h_{\hat{g}}(x)} \right]^{\lambda} h(x)dx, \quad I(h_{\hat{g}}, h) = \int_{\mathbb{R}^d} \left[ \frac{h_{\hat{g}}(x)}{h(x)} \right]^{\lambda} h_{\hat{g}}(x)dx
\]
we have
\[
I(h, h_{\hat{g}}) = (1 - q)\mathbb{E}_{x \sim h}(x) \left[ 1 - q + q \frac{h(x - \hat{g})}{h(x)} \right]^{\lambda} + q \mathbb{E}_{x \sim h}(x) \left[ 1 - q + q \frac{h(x - \hat{g})}{h(x)} \right]^{\lambda - 1},
\]
\[
\mathbb{E}_{x \sim h}(x) \left[ 1 - q + q \frac{h(x - \hat{g})}{h(x)} \right]^{\lambda} = \int_{\mathbb{R}^d} \left[ 1 - q + q \exp \left( \frac{-||x||^p}{\beta} \right) \right]^{\lambda} \cdot \frac{1}{\alpha} \exp \left( -\frac{||x||^p}{\beta} \right) dx.
\]
Due to the rotation invariant of sphere surface in $L_2$ space, we set vector $\mathbf{g} \triangleq \Delta, 0, \cdots, 0 \top$ for simplicity. By using the high-dimensional polar transformation Eq. (14), we have

$$
\|x\|^p - \|x - \mathbf{g}\|^p = \left(\sum_{i=1}^{d} x_i^2\right)^{p/2} - \left(\sum_{i=1}^{d} (x_i - \Delta)^2 + \sum_{i=2}^{d} x_i^2\right)^{p/2}
= r^p - (r^2 + \Delta^2 - 2r\Delta \cos(\theta_1))^{p/2},
$$

and $\frac{1}{\alpha} \exp \left(-\frac{\|x\|^p}{\beta}\right) dx = \frac{1}{\alpha} \exp(-\frac{\|\mathbf{g}\|^p}{\beta}) J(d) dr d\theta_1 \cdots d\theta_{d-1},$ where $J(d) = \frac{1}{\alpha} \exp(-\frac{\|\mathbf{g}\|^p}{\beta}) J(d) dr d\theta_1 \cdots d\theta_{d-1}$ is the Jacobian determinant. Let $f_1 \triangleq \left(\frac{r^p - (r^2 + \Delta^2 - 2r\Delta \cos(\theta_1))^{p/2}}{\beta}\right)$ for simplicity. Note that $f_1$ only contains $r$ and $\theta_1$. Therefore, we can integrate $\frac{1}{\alpha} \exp \left(-\frac{\|x\|^p}{\beta}\right)$ with respect to $\theta_2, \cdots, \theta_{d-1}$ at first and we have

$$
\int_0^\pi d\theta_2 \cdots \int_0^\pi d\theta_{d-2} \int_0^{2\pi} \frac{1}{\alpha} \exp \left(-\frac{\|x\|^p}{\beta}\right) d\theta_1 = \frac{1}{\alpha} \exp \left(-\frac{\|\mathbf{g}\|^p}{\beta}\right) \int_0^\pi \sin(\theta_1) d\theta_1 \int_0^{\pi} 1 \frac{1}{\sqrt{\pi}} (d-1)2 d\theta_1 = \frac{1}{\alpha} \exp \left(-\frac{\|\mathbf{g}\|^p}{\beta}\right) \frac{1}{\sqrt{\pi}} (d-1)2 d\theta_1 = \frac{1}{\alpha} \exp \left(-\frac{\|\mathbf{g}\|^p}{\beta}\right) \frac{1}{\sqrt{\pi}} (d-1)2 d\theta_1,
$$

where $1/N = \frac{b/\beta}{1/(k(x))}$ with $k = d/p$ (refer to Lemma 1).

Lettings $g(\theta_1, d) \triangleq \sin(\theta_1)^{d-2} 2\pi/(d-1)/2), we obtain

$$
E_{x \sim h(x)} \left[1 - q + \frac{h(x - \mathbf{g})}{h(x)}\right]^\lambda = \int_0^{\pi} g(\theta_1, d) d\theta_1 \int_0^{\pi} \left[1 - q + \frac{1}{\alpha} \exp(f_1)\right] d\gamma(r; d/p, \beta, p).
$$

Similarly to analysis of $E_{x \sim h(x)} \left[1 - q + \frac{h(x - \mathbf{g})}{h(x)}\right]^\lambda$, we obtain the following result that

$$
E_{x \sim h(x)} \left[1 - q + \frac{h(x + \mathbf{g})}{h(x + \mathbf{g})}\right] = \int_0^{\pi} g(\theta_1, d) d\theta_1 \int_0^{\pi} \left[1 - q + \frac{1}{\alpha} \exp(f_2)\right] d\gamma(r; d/p, \beta, p),
$$

where $f_2$ is defined as

$$
\frac{1}{\alpha} \exp \left(-\frac{\|\mathbf{g}\|^p}{\beta}\right) \frac{1}{\sqrt{\pi}} (d-1)2 d\theta_1 = \frac{1}{\alpha} \exp \left(-\frac{\|\mathbf{g}\|^p}{\beta}\right) \frac{1}{\sqrt{\pi}} (d-1)2 d\theta_1.
$$

Replacing results of $E_{x \sim h(x)} \left[1 - q + \frac{h(x - \mathbf{g})}{h(x)}\right]^\lambda$ and $E_{x \sim h(x)} \left[1 - q + \frac{h(x + \mathbf{g})}{h(x)}\right]^\lambda$ back into $I(h, q, \mathbf{g})$, and replacing $\theta_1$ with $x$ in $g(\theta_1, d)$, we complete the proof. \hfill $\square$

4) Proof of Theorem 3

Proof. Denote $h(x), \gamma(r), f(x)$ as the probability density function of pEM, generalized matrix distribution, and uniform distribution on $d$-dimensional sphere surface $S^d$, with radius $r$ and area $dC_d r^{d-1}$. Note that $C_d$ is a constant only related to dimension $d$. The proof includes two steps.

Step 1. Randomly sample two points from $S^d$ with the same probability, where $S^d$ is obtained by rescaling point $\mathbf{x} \sim \mathcal{N}(0, I_d)$ with length $r$ (i.e., $r\mathbf{x}/\|\mathbf{x}\| \in S^d$). Then, the remaining is to prove $f(x_1) = f(x_2), \forall x_1, x_2 \in S^d$ and $x_1 \neq x_2$. Step 2. The probability of sampling $\mathbf{x}$ from $pEM$ is the production of the probability product of sampling $\|\mathbf{x}\|$, (Lemma 1) and a point from $S^d$. That is, it requires to prove $h(x) = \gamma(\|\mathbf{x}\|) f(x)$.

For step 1, since the density of a point $x_1 \in S^d$ equals the integral of $\mathcal{N}(0, I_d)$ on a ray $l(x_1) \{ \rho x_1, \rho > 0 \}$, i.e., $f(x_1)$ is proportional to

$$
\int_{l(x_1)} e^{-\|\mathbf{x}\|^2/2} d\mathbf{s} = \int_{l(x_1)} e^{-\rho^2 \|\mathbf{x}\|^2/2} d\mathbf{x} = \sqrt{\pi/2},
$$

where the result is independent with $x_1$, then we have $f(x_1) = f(x_2), \forall x_1 \neq x_2$. This step is due to Lemma 1 and $f(x) = (dC_d)^{-1} \|\mathbf{x}\|^{-1}$, based on independence between $\gamma(r)$ and $f(x)$, we have $\gamma(\|\mathbf{x}\|) f(x) \propto e^{-\|\mathbf{x}\|^2/2}$. Meanwhile, due to $h(x) \propto e^{-\|\mathbf{x}\|^2/2}$ (pEM definition), we deduce that $h(x) = \gamma(\|\mathbf{x}\|) f(x)$.

5) Proof of Theorem 6

Proof. Based on Eq. (12), the optimal solution for $\varepsilon$ can be decomposed as the following sequential problems:

- For a given $\lambda$, solve Eq. (25) and Eq. (26).

$$
\alpha_3^1(\lambda; h(x)) = \min_{h(x)} \int_{\mathbb{R}^d} \left[ \frac{h(x)}{h_q, \mathbf{g}(x)} \right]^\lambda h(x) dx,
(25)
$$

$$
\alpha_3^2(\lambda; h(x)) = \min_{h(x)} \int_{\mathbb{R}^d} \left[ \frac{h_q, \mathbf{g}(x)}{h(x)} \right]^\lambda h_q, \mathbf{g}(x) dx.
(26)
$$

where $h_q, \mathbf{g}(x) = (1 - q) h(x) + q h(x - \mathbf{g})$.

- Searching the following optimizations with respect to $\lambda \in \mathbb{Z}^+$,

$$
\varepsilon_1 = \min_{\lambda \in \mathbb{Z}^+} \frac{\lambda}{T \alpha_3^1(\lambda; h(x)) + \ln(1/\delta)},
(27)
$$

$$
\varepsilon_2 = \min_{\lambda \in \mathbb{Z}^+} \frac{\lambda}{T \alpha_3^2(\lambda; h(x)) + \ln(1/\delta)}.
(28)
$$

Finally, we obtain the minimal $\varepsilon^* = \max(\varepsilon_1, \varepsilon_2)$. Therefore, the optimal solution $h(x)$ must satisfy Eq. (25) and Eq. (26), with constraint conditions $\int_{\mathbb{R}^d} [|x|] h(x) dx = A$. That is, any optimal solution has to satisfy the following optimization problems at first.

$$
\lambda = 1, \min_{h(x)} \int_{\mathbb{R}^d} \left[ \frac{h(x)}{h_q, \mathbf{g}(x)} \right]^\lambda h(x) dx \quad \text{or} \quad \lambda = 2, \min_{h(x)} \int_{\mathbb{R}^d} \left[ \frac{h_q, \mathbf{g}(x)}{h(x)} \right]^\lambda h_q, \mathbf{g}(x) dx
$$

s.t. $\int_{\mathbb{R}^d} |x|^2 h(x) dx = A$.

Obviously, this is a functional extreme value problem and we construct the following Lagrange functional,

$$
\mathcal{J}_1(h(x)) = \ln \int_{\mathbb{R}^d} \left[ \frac{h(x)}{h_q, \mathbf{g}(x)} \right]^\lambda h(x) dx + \mu_1 \left( \int_{\mathbb{R}^d} |x|^2 h(x) dx - A \right) + \mu_2 \left( \int_{\mathbb{R}^d} |x|^2 h(x) dx - A \right),
$$
Taking derivative of $h(x) = \int_{R^d} \left[ \frac{h_{q,g}(x)}{h(x)} \right]^\lambda \frac{h_{q,g}(x)}{h(x)} \cdot h_{q,g}(x) \cdot dx + \mu_1 \left[ \int_{R^d} h(x) \cdot dx - 1 \right] + \mu_2 \left[ \int_{R^d} \|x\|^2 \cdot h(x) \cdot dx - A \right].$

To obtain the optimal solution, we first calculate $\frac{\partial J_1(h(x))}{\partial h(x)}$ and $\frac{\partial J_2(h(x))}{\partial h(x)}$, then let them equal 0. By functional derivation, we can obtain the conditions by the following equation $(i = 1, 2)$,

$$
\frac{d J_i(h(x) + \tau \delta h(x))}{d \tau} \bigg|_{\tau = 0} = \int_{R^d} \frac{\partial J_i(h(x))}{\partial h(x)} \delta h(x) \cdot dx, \tag{29}
$$

where $\tau$ is a real number and $\delta h(x)$ is the variation of $h(x)$. For simplicity, let $\tilde{h}(\tau) = h(x) + \tau \delta h(x)$, $h_{q,g}(\tau) = (1 - q) [h(x) + \tau \delta h(x)] + q [h(x - \tilde{g}) + \tau \delta h(x - \tilde{g})]$, and $I_1(\tau) = \int_{R^d} \left| h_{q,g}(\tau) \right|^\lambda \tilde{h}(\tau) \cdot dx$. Taking derivative of $J_1(\tilde{h}(\tau))$ with respect to $\tau$, we have

$$
\frac{d J_1(\tilde{h}(\tau))}{d \tau} = \frac{1}{I_1(\tau)} \int_{R^d} \left[ \frac{\tilde{h}(\tau)}{h_{q,g}(\tau)} \right]^\lambda \left[ \frac{\delta h(x) \cdot \tilde{h}(\tau)}{h_{q,g}(\tau)} \right] \cdot \delta h(x) \cdot dx + \mu_1 \left[ \int_{R^d} \delta h(x) \cdot dx + \mu_2 \int_{R^d} \|x\|^2 \cdot \delta h(x) \cdot dx \right].
$$

Let $\tau = 0$ and rearrange $\delta h(x)$ and $\delta h(x - \tilde{g})$. Note that $\tilde{h}(0) = h(x)$ and $h_{q,g}(0) = h_{q,g}(x)$, then we have

$$
\frac{d J_1(\tilde{h}(\tau))}{d \tau} = \frac{1}{I_1(0)} \int_{R^d} \left[ \frac{h(x)}{h_{q,g}(x)} \right]^\lambda q \delta h(x - \tilde{g}) \cdot \delta h(x) \cdot dx - \frac{1}{I_1(0)} \int_{R^d} \left[ \frac{h(x)}{h_{q,g}(x)} \right]^\lambda q \delta h(x) \cdot \delta h(x - \tilde{g}) \cdot dx + \mu_1 \int_{R^d} \delta h(x) \cdot dx + \mu_1 \int_{R^d} \|x\|^2 \cdot \delta h(x) \cdot dx.
$$

Therefore, based on Eq. (29), we have

$$
\frac{\partial J_1(h(x))}{\partial h(x)} = \frac{\lambda + 1}{I_1(0)} \left[ \frac{h(x)}{h_{q,g}(x)} \right]^\lambda \cdot (1 - q) \left[ \frac{h(x)}{h_{q,g}(x)} \right] - q \left[ \frac{h(x + \tilde{g})}{h_{q,g}(x + \tilde{g})} \right]^\lambda + \mu_1 + \mu_2 \cdot \|x\|^2.
$$

Let $\frac{\partial J_1(h(x))}{\partial h(x)} = 0$, and denote $f(x) = h(x)/h_{q,g}(x)$ for simplicity. We deduce that the necessary condition for achieving the minimum $\delta x$ is $\left( \lambda + 1 \right) f^\lambda(x) - \lambda \left[ (1 - q) f^\lambda(x) + q \left[ f^\lambda(x + \tilde{g}) \right] + (\mu_1 + \mu_2 \cdot \|x\|^2)I_1(0) \right] = 0$. Note that $I_1(0) = \|x\|^2 \cdot f^\lambda(x)$. Multiplying both sides of the above equation with $h(x)$ and taking expectation with respect to $x \sim h(x)$, we have

$$
(\lambda + 1) \mathbb{E}_{x \sim h(x)} f^\lambda(x) + \mathbb{E}_{x \sim h(x)} f^\lambda(x) \cdot (\mu_1 + \mu_2 \mathbb{E}_{x \sim h(x)} \cdot \|x\|^2) - \lambda \left[ (1 - q) \mathbb{E}_{x \sim h(x)} f^\lambda(x) + q \mathbb{E}_{x \sim h(x)} f^\lambda(x + \tilde{g}) \right] = 0.
$$

That is,

$$
\mathbb{E}_{x \sim h(x)} \left[ (\lambda + 1) f^\lambda(x) - \lambda \left( (1 - q) f^\lambda(x) + q f^\lambda(x + \tilde{g}) \right) \right] + \mathbb{E}_{x \sim h(x)} f^\lambda(x) \cdot (\mu_1 + \mu_2 \mathbb{E}_{x \sim h(x)} \cdot \|x\|^2) = 0.
$$

Similar to computation of $J_1(h(x))$, we obtain the following necessary condition for $J_2(h(x))$.

$$
\mathbb{E}_{x \sim h(x)} \left[ (\lambda + 1) (1 - q) u^\lambda(x) + q u^\lambda(x + \tilde{g}) + \mu_1 + \mu_2 \mathbb{E}_{x \sim h(x)} \cdot \|x\|^2 = 0.
$$

where $u(x) = f^{-1}(x) = h_{q,g}(x)$ and $h_{q,g}(x) = (1 - q) h(x) + q h(x - \tilde{g})$. These two equations are necessary conditions for achieving the minimum $\varepsilon^*$, and we complete this theorem. }

6) Proof of Theorem

Proof. Without loss of generality, assume GM has density function $h(x) = \frac{1}{\alpha} \exp(-\|x\|^2/\beta)$, where $\alpha$ is the normalization and $\int_{R^d} \|x\|^2 h(x) \cdot dx = A$. Next we verify that $h(x)$ does not satisfy the second condition of Theorem 6. That is, GM is not the optimal mechanism. Recall that the second condition is

$$
(\lambda + 1) \left[ (1 - q) \mathbb{E}_{x \sim h(x)} u^\lambda(x) + q \mathbb{E}_{x \sim h(x)} u^\lambda(x + \tilde{g}) \right] + \mathbb{E}_{x \sim h(x)} u^\lambda(x) \cdot (\mu_1 + \mu_2 \mathbb{E}_{x \sim h(x)} \cdot \|x\|^2) = 0.
$$

Let $x = [x_1, \cdots, x_d]^\top$, $\tilde{g} = [\Delta, 0, \cdots, 0]^\top$ with $\Delta = \|\tilde{g}\|^2$. Based on straightforward calculations about Gaussian distribution, we have

$$
\mathbb{E}_{x \sim h(x)} \left[ \frac{h(x)}{h(x + \tilde{g})} \right]^\lambda = \mathbb{E}_{x \sim h(x)} \left[ \frac{h(x)}{h(x + \tilde{g})} \right]^\lambda \mathbb{E}_{x \sim N(\Delta, \beta/2)} \mathbb{E}^{2 \lambda \Delta_1 + 1 \Delta^2 / \beta} = e \cdot \lambda \Delta^2 \cdot \mathbb{E}_{x \sim N(\Delta, \beta/2)} \mathbb{E}^{2 \lambda \Delta_1}.
$$

Similarly,

$$
\mathbb{E}_{x \sim h(x)} \left[ \frac{h(x)}{h(x - \tilde{g})} \right]^\lambda = \frac{e \cdot \lambda \Delta^2}{\beta} \cdot \mathbb{E}_{x \sim N(\Delta, \beta/2)} \mathbb{E}^{2 \lambda \Delta_1}.
$$

With respect to the remaining $\mathbb{E}_{x \sim h(x)} \cdot \|x\|^2$, we have

$$
\mathbb{E}_{x \sim h(x)} \cdot \|x\|^2 = \mathbb{E}_{x \sim h(x)} \cdot \|x + \tilde{g}\|^2 = \mathbb{E}_{x \sim h(x)} \cdot \|x\|^2 + 2 \Delta x_1 + \Delta^2 = d\beta / 2 + \Delta^2,
$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3129130, IEEE Access
where we use $E_{x \sim h(x)} \|x\|^2 = \frac{\beta \Gamma(d/2+1)}{\Gamma(d/2)} = d\beta/2$ based on Lemma 2 ($p = 2, m = 2$), and $E_{x \sim h(x)} x_1 = 0$. Substituting these results of expectations into Eq. (30), we have

$$
(\lambda + 1)e^{2\lambda \Delta^2/\beta} - \lambda e^{(2\lambda + 2)\Delta^2/\beta} + [\mu_1 + \mu_2 (d\beta/2 + \Delta^2)] = 0.
$$

With respect to Eq. (31), replacing $e^{2\lambda \Delta^2/\beta}$ with its Taylor expansion on $\lambda$, we have the following equivalent equations.

$$
\begin{align*}
1 + \mu_1 + \mu_2 (d\beta/2 + \Delta^2) &= 0, \\
1 + 2\Delta^2/\beta &= e^{2\Delta^2/\beta}.
\end{align*}
$$

(32)

It is obvious observed that the second equation of Eq. (32) can not hold. Therefore, GM is not the optimal mechanism.

\[ \square \]

***