Generalized Gaussian Mechanism for Differential Privacy

Fang Liu

Abstract—Assessment of disclosure risk is of paramount importance in data privacy research and applications. The concept of differential privacy (DP) formalizes privacy in probabilistic terms and provides a robust concept for privacy protection. Practical applications of DP involve development of DP mechanisms to release data at a pre-specified privacy budget. In this paper, we generalize the widely used Laplace mechanism to the family of generalized Gaussian (GG) mechanism based on the l_p global sensitivity of statistical queries. We explore the theoretical requirement for the GG mechanism to reach DP at prespecified privacy parameters, and investigate the connections and differences between the GG mechanism and the Exponential mechanism based on the GG distribution. We also present a lower bound on the scale parameter of the Gaussian mechanism of (ϵ, δ)-probabilistic DP as a special case of the GG mechanism, and compare the utility of sanitized results in the tail probability and dispersion between the Gaussian and Laplace mechanisms. Lastly, we apply the GG mechanism in three experiments and compare the accuracy of sanitized results in the l_1 distance and Kullback-Leibler divergence, and examine the prediction power of a SVM classifier constructed with the sanitized data relative to the original results.

Index Terms—probabilistic differential privacy, l_p global sensitivity, privacy budget, Laplace mechanism, Gaussian mechanism

1 Introduction

When releasing information publicly from a database or sharing data with collaborators, data collectors are always concerned about exposing sensitive personal information of individuals who contribute to the data. Even with key identifiers removed, data users may still identify a participant in a data set such as via linkage with public information. Differential privacy (DP) provides a strong privacy guarantee to data release without making assumptions about the background knowledge or behavior of data intruders (adversaries) [1], [2], [3]. For a given privacy budget, information released via a differentially private mechanism guarantees no additional personal information of an individual in the data can be inferred, regardless how much background information adversaries already possess about the individual. DP has spurred a great amount work in the development of differentially private mechanisms to release results and data, including the Laplace mechanism [1], the Exponential mechanism [4], [5], the medium mechanism [6], the multiplicative weights mechanism [7], the geometric mechanism [8], the staircase mechanism [9], the Gaussian mechanism [10], and applications of DP for private and secure inference in a Bayesian setting [11], among others.

In this paper, we unify the Laplace mechanism and the Gaussian mechanism in the framework of a general family, referred to as the generalized Gaussian (GG) mechanism. The GG mechanism is based on the l_p global sensitivity (GS) of queries, a generalization of the l_1 GS. We demonstrate the nonexistence of a scale parameter that would lead to a GG mechanism of pure ϵ-DP in the case of $p \neq 1$ if the results to be released are unbounded, but suggest the GG mechanism of (ϵ, δ)-probabilistic DP (pDP) as an alternative in such cases. For bounded data we introduce the truncated GG mechanism and the boundary inflated truncated GG mechanism that satisfy pure ϵ-DP. We investigate the connections between the GG mechanism and the Exponential mechanism when the utility function in the latter is based on the Minkowski distance, and establish the relationship between the sensitivity of the utility function in the Exponential mechanism and the l_p GS of queries. We then take a closer look at the Gaussian mechanism (the GG mechanism of order 2), and derive a lower bound on the scale parameter that delivers (ϵ, δ)-pDP. The bound is tighter than the bound for satisfying (ϵ, δ)-approximate DP (aDP) in the Gaussian mechanism [10], implying that less noise is injected in the sanitized results with the new bound. We compare the utility of sanitized results, in terms of the tail probability and dispersion or mean squared errors (MSE), from independent applications of the Gaussian mechanism and the Laplace mechanism. Finally, we run 3 experiments on the mildew, Czech, and adult data, respectively, and sanitize the count data via the Laplace mechanism, the Gaussian mechanisms of (ϵ, δ)-pDP and (ϵ, δ)-aDP. We compare the accuracy of sanitized results in terms of the l_1 distance and Kullback-Leibler divergence from the original results, and examine how sanitization affects the prediction accuracy of support vector machines constructed with the sanitized data in the adult experiment.

The rest of the paper is organized as follows. Section 2 defines the l_p GS and presents the GG mechanism of (ϵ, δ)-pDP, the truncated GG mechanism, and the boundary inflated truncated GG mechanism that satisfy pure ϵ-DP. It also connects and differentiates between the GG mechanisms and the Exponential mechanism when the utility function in the latter is based the Minkowski

- *F. Liu is an associate professor in the Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556*
E-mail: fang.liu.131@nd.edu

distance. Section 3 takes a close look at the Gaussian mechanism of (ϵ,δ)-pDP, and compares it with the Gaussian mechanism of (ϵ,δ)-aDP. It also compares the tail probability and the dispersion of the noises injected via the Gaussian mechanism and the Laplace mechanism. Section 4 presents the findings from the 3 experiments. Concluding remarks are given in Section 5.

2 Generalized Gaussian Mechanism

2.1 Differential Privacy (DP)

DP was proposed and formulated in Dwork et al. [1] and Dwork [12]. A perturbation algorithm $\mathcal{R}$ gives ϵ-differential privacy if for all data sets $(\mathbf{x},\mathbf{x}')$ that differ by only one individual ($d(\mathbf{x},\mathbf{x}')=1$), and all possible query results $Q\subseteq\mathcal{Y}$ to query $\mathbf{s}$ ($\mathcal{Y}$ denotes the output range of $\mathcal{R}$),

$$\left|\log\left(\frac{\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x}))\in Q)}{\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x}'))\in Q)}\right)\right|\le\epsilon, \tag{1}$$

where $\epsilon>0$ is the privacy "budget" parameter. $\mathbf{s}$ refers to queries about data, we also use it to denote the query results (unless stated otherwise, the domain of the query results is the set of all real numbers). $d(\mathbf{x},\mathbf{x}')=1$ is often defined in two ways in the DP community: $\mathbf{x}$ and $\mathbf{x}'$ are of the same size and differ in exactly one record (row) in at least one attributes (columns); and $\mathbf{x}$ has one less (more) record than $\mathbf{x}'$ and is the same as $\mathbf{x}'$ otherwise. Mathematically, Eqn (1) states that the probabilities of obtaining the same query result perturbed via $\mathcal{R}$ are roughly the same regardless of whether the query is sent to $\mathbf{x}$ or $\mathbf{x}'$. In layman's terms, DP implies the chance that an individual will be identified based on the perturbed query result is very low since the query result would be about the same with or without the individual in the data. The degree of "roughly the same" is determined by the privacy budget ϵ. The lower ϵ is, the more similar the probabilities of obtaining the same query results from $\mathbf{x}$ and $\mathbf{x}'$ are. DP provides a strong and robust privacy guarantee in the sense that it does not assume anything regarding the background knowledge or the behavior on data users.

In addition to the "pure" ϵ-DP in Eqn (1), there are softer versions of DP, including the (ϵ,δ)-approximate DP (aDP) [13], the (ϵ,δ)-probabilistic DP (pDP) [14], the (ϵ,δ)-random DP (rDP) [15], and the (ϵ,τ)-concentrated DP (cDP) [16]. In all the relaxed versions of DP, one additional parameter is employed to characterize the amount of relaxation on top of the privacy budget ϵ. Both the (ϵ,δ)-aDP and the (ϵ,δ)-pDP reduce to ϵ-DP when $\delta=0$, but are different with respect to the interpretation of δ. In (ϵ,δ)-aDP,

$$\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x}))\in Q)\le e^{\epsilon}\Pr(\mathcal{R}(s(\mathbf{x}'))\in Q)+\delta; \tag{2}$$

while a perturbation algorithm $\mathcal{R}$ satisfies (ϵ,δ)-pDP if

$$\Pr\left(\left|\log\left(\frac{\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x}))\in Q)}{\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x}'))\in Q)}\right)\right|>\epsilon\right)\le\delta; \tag{3}$$

that is, the probability of an output generated by $\mathcal{R}$ belonging to the disclosure set is bounded below δ, where the disclosure set contains all the possible outputs that leak information for a given privacy budget ϵ. The fact that probabilities are within $[0,1]$ puts constraints on the values of $\epsilon,\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x}')\in Q)$, and δ in the framework of (ϵ,δ)-aDP. By contrast, (ϵ,δ)-pDP seems to be less constrained and more intuitive with its probabilistic flavor. When δ is small, (ϵ,δ)-aDP and (ϵ,δ)-aDP are roughly the same. The (ϵ,δ)-rDP is also a probabilistic relaxation of DP; but it differs from (ϵ,δ)-pDP in that the probabilistic relaxation is with respect to data generation. In (ϵ,τ)-cDP, privacy cost is treated as a random variable with an expectation of ϵ, and Prob(the actual cost $>\epsilon$)$>a$ is bounded by $e^{-(a/\tau)^2/2}$. The (ϵ,τ)-cDP, similar to the (ϵ,δ)-pDP, relaxes the satisfaction of DP with respect to $\mathcal{R}$ and is broader in scope.

2.2 l_p Global Sensitivity

Definition 1 (**l_p Global Sensitivity**)**.** For all $(\mathbf{x},\mathbf{x}')$ that is $d(\mathbf{x},\mathbf{x}')=1$, the l_p-global sensitivity (GS) of a query set $\mathbf{s}$ with r elements is

$$\Delta_p=\max_{\substack{\mathbf{x},\mathbf{x}'\\ d(\mathbf{x},\mathbf{x}')=1}}\|\mathbf{s}(\mathbf{x})-\mathbf{s}(\mathbf{x}')\|_p= \tag{4}$$
$$\max_{\substack{\mathbf{x},\mathbf{x}'\\ d(\mathbf{x},\mathbf{x}')=1}}\left(\textstyle\sum_{k=1}^r|s_k(\mathbf{x})-s_k(\mathbf{x}')|^p\right)^{1/p}\text{ for integer }p>0.$$

Δ_p is the maximum difference measured by the Minkowski distance in query results $\mathbf{s}$ between two neighboring data set $\mathbf{x},\mathbf{x}'$ with $d(\mathbf{x},\mathbf{x}')=1$. The sensitivity is "global" since it is defined for all possible data sets and all possible ways that $\mathbf{x}$ and $\mathbf{x}'$ differ by one. The higher Δ_p is, the more disclosure risk there is on the individuals from releasing the original query results $\mathbf{s}$. The l_p GS is a key concept in the construction of the generalized Gaussian mechanism in Section 2.

The l_p GS is a generalization of the l_1 GS [1], [12] and the l_2 GS [10]. The "difference" between $\mathbf{s}(\mathbf{x})$ and $\mathbf{s}(\mathbf{x}')$ measured by Δ_1 is the largest among all Δ_p for $p\ge1$ since that $\|\mathbf{s}\|_{p+a}\le\|\mathbf{s}\|_p$ for any real-valued vector $\mathbf{s}$ and $a\ge0$. In addition, Δ_1 is also the most "sensitive" measure given that the rate of change with respective to any s_k is the largest among all $p\ge1$. When s is a scalar, $\Delta_p=\Delta_1$ for all $p>0$. When $\mathbf{s}$ is multi-dimensional, an easy upper bound for l_1 GS Δ_1 is $\sum_{k=1}^r\Delta_{1,k}$, the sum of the l_1 GS of each element k in $\mathbf{s}$, by the triangle inequality. Lemma 2 gives an upper bound on Δ_p for a general p that includes $p=1$ as a special case (the proof is provided in Appendix A).

Lemma 2 (**Upper Bound for Δ_p**)**.** $\left(\sum_{k=1}^r\Delta_{1,k}^p\right)^{1/p}$ is an upper bound for Δ_p, where $\Delta_{1,k}$ is the l_1 GS of s_k.

The upper bound given in Lemma 2 can be conservative in cases where the change from $\mathbf{x}$ to $\mathbf{x}'$ does not necessarily alter every entry in the multidimensional $\mathbf{s}$. For example, the l_p GS of releasing a histogram with r bins is 1 (if $d(\mathbf{x},\mathbf{x}')=1$ is defined as $\mathbf{x}'$ is one record less/more than $\mathbf{x}$). In other words, the GS is not $r^{1/p}$ even though there are r counts in the released histogram, but is the same as in releasing a single cell because removing one record only alters the count in a single bin.

It is obvious that each element s_k in $\mathbf{s}$ for $k = 1, \ldots, r$ needs to be bounded to obtain a finite Δ_p. The most extreme case is the change from $\mathbf{x}$ to $\mathbf{x}'$ makes s_k jump from one extreme to the other, implying the range of s_k can be used as an upper bound for $\Delta_{k,1}$, which, combined with Lemma 2, leads to the following claim.

Claim 3 (Upper Bound for Δ_p for Bounded Statistics). Denote the finite bounds for statistic s_k by $[c_{k0}, c_{k1}]$. The GS for s_k is $\Delta_k \leq c_{k1} - c_{k0}$ and the GS for $\mathbf{s} = \{s_k\}_{k=1,\ldots,r}$ is $\Delta_p \leq (\sum_{k=1}^{r}(c_{k1} - c_{k0})^p)^{1/p}$.

2.3 Generalized Gaussian Distribution

The GG mechanism is defined based on the GG distribution $\mathrm{GG}(\mu, b, p)$ with location parameter μ, scale parameter $b > 0$, shape parameter $p > 0$. The probability density function (pdf) is

$$f(x|\mu, b, p) = \frac{p}{2b\Gamma(p^{-1})} \exp\left\{-\left(\frac{|x-\mu|}{b}\right)^p\right\}.$$

The mean and variance of x are μ and $b^2\Gamma(3/b)/\Gamma(1/b)$, respectively ($\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$ is the Gamma function). When $p = 1$, the GG distribution is the Laplace distribution with mean μ and variance $2b^2$; when $p = 2$, the GG distribution becomes the Gaussian distribution with mean 0 and variance $b^2/2$. Figure 1 presents some examples of the GG distributions at different p. All the distributions in the left plot have the same scale $b = \sqrt{2}$ and location 0, and those in the right plot have the same variance 1 and location 0. When the scale parameter is the same (the left plot), the distributions become less spread as p increases, and the Laplace distribution ($p = 1$) looks very different from the rest. When the variance is the same (the right plot), the Laplace distribution is most likely to generate values that are close to the mean, followed by the Gaussian distribution ($p = 2$).

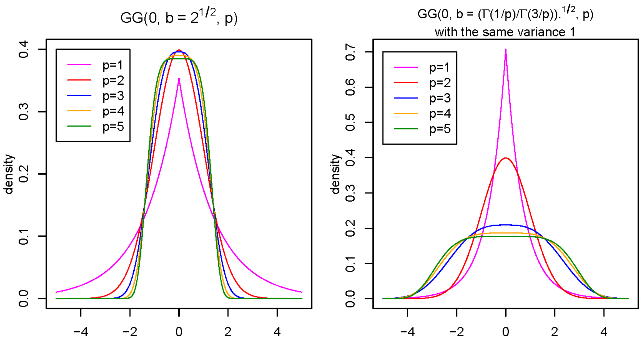

Fig. 1: Density of GG distributions

2.4 GG Mechanism of ϵ-DP

Query $\mathbf{s}$ needs to bounded to calculate the l_p GS, but the bounding requirement does not necessarily goes into formulating the GG distribution for the GG mechanism in the first place. A well-known example along these lines is the Laplace mechanism, which employs a Laplace distribution with support $(-\infty, \infty)$, though its scale parameter $b = \Delta_1/\epsilon$ requires $\mathbf{s}$ to be bounded for Δ_1 to be calculated. We thus first examine the GG mechanism of ϵ-DP with the domain for sanitized $\mathbf{s}^*_k$ defined on $(-\infty, \infty)^r$. If bounding for $\mathbf{s}^*$ is necessary, it can be incorporated in a post-hoc manner after being generated from the GG mechanism.

Eqn (5) presents the GG distribution from which sanitized $\mathbf{s}^*$ would be generated to satisfy ϵ-DP given the original $\mathbf{s}$, assuming b exists.

$$\begin{aligned} f(\mathbf{s}^*) &\propto \exp\{-\left(\|\mathbf{s}^* - \mathbf{s}\|_p/b\right)^p\} \\ &\propto \textstyle\prod_{k=1}^{r} \exp\{-(|s^*_k - s_k|/b)^p\} \\ &= \textstyle\prod_{k=1}^{r} \frac{p}{2b\Gamma(p^{-1})} \exp\{-(|s^*_k - s_k|/b)^p\} \\ &= \textstyle\prod_{k=1}^{r} \mathrm{GG}(s_k, b, p) \end{aligned} \tag{5}$$

Claim 4 (Nonexist of b to Achieve ϵ-DP for $p \neq 1$). There does not exist a lower bound on b when $p \neq 1$ for the GG distribution in Eqn (5) that generates $\mathbf{s}^*$ with ϵ-DP. When $p = 1$, the lower bound on b that leads to ϵ-DP is $\epsilon^{-1}\Delta_1$.

Appendix B lists the detailed steps that lead to Claim 4. In brief, to achieve ϵ-DP, we need $b^{-p}\left(\sum_{k=1}^{r}\sum_{j=1}^{p-1}\binom{p}{j}|s^*_k - s_k|^{p-j}\Delta^j_{1,k} + \Delta^p_p\right) \leq \epsilon$ (Eqn B.4). However, this inequality depends on the random GG noise $e_k = s^*_k - s_k$ for $k = 1, \ldots, r$, the support of which is $(-\infty, \infty)^r$. In other words, there does not exist a random noise-free solution on b, unless $p = 1$ in which case the inequality no longer involves the error terms and the GG mechanism reduces to the familiar Laplace mechanism of ϵ-DP. We propose two approaches to fix the problem and achieve DP through the GG mechanism. The first approach leverages the bounding requirement for $\mathbf{s}$ and builds in the requirement in the GG distribution in the first place to get a lower bound on b and generate $\mathbf{s}^*$ with ϵ-DP, assuming that $\mathbf{s}^*$ and $\mathbf{s}$ share the same bounded domain (Section 2.5). The second approach still employs the GG distribution in Eqn (5) to sanitize $\mathbf{s}$, but satisfying (ϵ, δ)-pDP instead of the pure ϵ-DP (Section 2.6). The sanitized $\mathbf{s}^*$ can be bounded in a post-hoc manner, as needed.

2.5 Truncated GG Mechanism and Boundary Inflated Truncated GG Mechanism of ϵ-DP

Definition 5 (Truncated Generalized Gaussian Mechanism). Denote the bounds on query result $\mathbf{s}$ by $[c_{k0}, c_{k1}]_{k=1,\ldots,r}$. For integer $p \geq 1$, the truncated GG mechanism $\mathcal{T}^p_\epsilon$ generates $\mathbf{s}^* \in [c_{k0}, c_{k1}]_{k=1,\ldots,r}$ by drawing from the truncated GG distribution

$$f(\mathbf{s}^*|c_{k0} \leq s^*_k \leq c_{k1}, \forall\, k = 1, \ldots, r) = \tag{6}$$
$$\prod_{k=1}^{r} \frac{b^{-1}p\exp\{-(|s^*_k - s_k|/b)^p\}}{\gamma[p^{-1}, (|c_{k1} - s_k|/b)^p] + \gamma[p^{-1}, (|s_k - c_{k0}|/b)^p]},$$

where γ is the lower incomplete gamma function and the scale parameter is

$$b \geq \left(2\epsilon^{-1}\left(\sum_{k=1}^{r}\sum_{j=1}^{p-1}\binom{p}{j}|c_{k1} - c_{k0}|^{p-j}\Delta^j_{1,k} + \Delta^p_p\right)\right)^{1/p}, \tag{7}$$

Proposition 6. The truncated GG mechanism $\mathcal{T}^p_\epsilon$ satisfies ϵ-differential privacy.

The proof of ϵ-DP of $\mathcal{T}^p_\epsilon$ is given in Appendix C. $\mathcal{T}^p_\epsilon$ perturbs each element in $\mathbf{s}$ independently; thus Eqn (6) involves the product of r independent density functions. Though the closed interval $[c_{k0}, c_{k1}]$ is used to denote the bounds on s_k, Definition 6 remains the same regardless of whether the interval is closed, open, or half-closed since the GG distribution is defined on a continuous domain. If s_k is discrete in nature such as counts, post-hoc rounding on perturbed $\mathbf{s}^*_k$ can be applied. The lower bound on b in Eqn (7) depends on Δ_p. We may apply Lemma 2 and set Δ^p_p at its upper bound $\sum_{k=1}^r \Delta^p_{1,k}$ to obtain a lower bound on b.

$$b \geq \left(2\epsilon^{-1}\left(\sum_{k=1}^r \sum_{j=1}^p \tbinom{p}{j} |c_{k1} - c_{k0}|^{p-j} \Delta^j_{1,k}\right)\right)^{1/p}. \quad (8)$$

Definition 7 (BIT Generalized Gaussian Mechanism). Denote the bounds on query result s_k by $[c_{k0}, c_{k1}]$ for $k = 1, \ldots, r$. For integer $p \geq 1$, the boundary inflated truncated (BIT) GG mechanism $\mathcal{B}^p_\epsilon$ sanitizes $\mathbf{s}$ by drawing perturbed $\mathbf{s}^*$ from the following piecewise distribution

$$\begin{aligned} & f(\mathbf{s}^* | c_{k0} \leq s^*_k \leq c_{k1}, \forall\ k = 1, \ldots, r) = \quad (9) \\ & \prod_{k=1}^r \left\{ p_k^{\mathrm{I}(s^*_k = c_{k0})} q_k^{\mathrm{I}(s^*_k = c_{k1})} \right. \\ & \qquad \left. \left(\frac{p \exp\{(|s^*_k - s_k|/b(\epsilon))^p\}}{2b(\epsilon)\Gamma(p^{-1})} \right)^{\mathrm{I}(c_{k0} < s^*_k < c_{k1})} \right\}, \end{aligned}$$

where $p_k = \Pr(s^*_k < c_{k0}; s_k, p, b) = \frac{1}{2} - \gamma[p^{-1}, (|s_k - c_{k0}|/b)^p](2\Gamma(p^{-1}))^{-1}$ and $q_k = \Pr(s^*_k > c_{k1}; s_k, p, b) = \frac{1}{2} - \gamma[p^{-1}, (|c_{k1} - s_k|/b)^p](2\Gamma(p^{-1}))^{-1}$, $b(\epsilon)$ indicates that the scale parameter b is a function of ϵ, γ is the lower incomplete gamma function, Γ is the gamma function, and $\mathrm{I}()$ is the indicator function that equals 1 if the argument in the parentheses is true, 0 otherwise.

In a nutshell, the BIT GG mechanism replaces out-of-bound values with the boundary values and keeps the within-bound values as is, leading to a piecewise distribution. This is in contrast to the truncated GG mechanism which throws away out-of-bound values. The challenge with perturbing $\mathbf{s}$ via Eqn (9) lies in solving for a lower bound b as a function of ϵ that satisfies ϵ-DP from

$$\log\left|\frac{f(\mathbf{s}^* | c_{k0} \leq s^*_k \leq c_{k1}, \forall\ k = 1, \ldots, r)}{f(\mathbf{s}'^* | c_{k0} \leq s^*_k \leq c_{k1}, \forall\ k = 1, \ldots, r)}\right| \leq \epsilon \quad (10)$$

where $\mathbf{s}^* = \{s^*_k\}$ and $\mathbf{s}'^* = \{s'^*_k\}$ are the sanitized results from data $\mathbf{x}$ and $\mathbf{x}'$ that are $d(\mathbf{x}, \mathbf{x}') = 1$, respectively. The lower bound given in Eqns (7) and (8) can be used when the output subset Q is a subset of $(c_{10}, c_{11}) \times \cdots \times (c_{r0}, c_{r1})$ (open intervals). However, when Q is $\{s_k = c_{k0}\ \forall\ k = 1, \ldots, r\}$ and $\{s_k = c_{k1}\ \forall\ k = 1, \ldots, r\}$, respectively, there are no analytical solutions on b in either Eqns (11) or (12)

$$\log\left|\prod_{i=1}^r \frac{1/2 - \gamma(p^{-1}, ((s_k - c_{k0})/b)^p)(2\Gamma(p^{-1}))^{-1}}{1/2 - \gamma(p^{-1}, ((s'_k - c_{k0})/b)^p)(2\Gamma(p^{-1}))^{-1}}\right| \leq \epsilon \quad (11)$$

$$\log\left|\prod_{i=1}^r \frac{1/2 - \gamma(p^{-1}, ((s_k - c_{k0})/b)^p)(2\Gamma(p^{-1}))^{-1}}{1/2 - \gamma(p^{-1}, ((s'_k - c_{k0})/b)^p)(2\Gamma(p^{-1}))^{-1}}\right| \leq \epsilon. \quad (12)$$

The most challenging situation is when Q is a mixture set of (c_{k0}, c_{k1}), c_{k0}, and c_{k1} for different $k = 1, \ldots, r$. In summary, the BIT GG mechanism is not very appealing from a practical standpoint given the difficulty in solving for b, though in theory such a b exists to achieve ϵ-DP.

2.6 GG Mechanism of (ϵ, δ)-pDP

The second approach for obtaining a lower bound on the scale parameter b for the GG distribution in Eqn (5) when $p \geq 2$ is to employ a soft version of DP. Proposition 8 presents a solution on b that satisfies (ϵ, δ)-pDP.

Proposition 8 (GG Mechanism of (ϵ, δ)-pDP). If the scale parameter b in the GG distribution in Eqn (5) satisfies

$$\Pr\left(\sum_{k=1}^r \sum_{j=1}^{p-1} \tbinom{p}{j} |s^*_k - s_k|^{p-j} \Delta^j_{1,k} > b^p\epsilon - \Delta^p_p\right) < \delta, \quad (13)$$

then the GG mechanism satisfies (ϵ, δ)-pDP when $p \geq 2$.

The proof is straightforward. Specifically, rather than setting the left side of Eqn (B.4) $\leq \epsilon$, we attach a non-zero probability of achieving the inequality, that is, Pr(Eqn (B.4)$< \epsilon$) $> 1-\delta$, leading to Eqn (13). The (ϵ, δ)-pDP does not apply to the Laplace mechanism ($p = 1$) at least in the framework laid out in Proposition 8. When $p=1$, Eqn (B.1) becomes $b^{-1}\sum_{k=1}^r \big||e_k| - |e_k + d_k|\big| \leq b^{-1}\sum_{k=1}^r |d_k| \leq b^{-1}\Delta_1$, which does not involve the random variable $\mathbf{s}^*$; in other words, as long as $b^{-1}\Delta_{\mathbf{s},1} \leq \epsilon$, the pure ϵ-DP is guaranteed.

Proposition 8 does not list a closed-form solution on b as it is likely that only numerical solutions exist in most cases. Given that s^*_k is independent across $k = 1, \ldots, r$, $a_k = \sum_{j=1}^{p-1} \binom{p}{j} |s^*_k - s_k|^{p-j} \Delta^j_{1,k}$ a function of s^*_k, is also independent across k. Therefore, the problem becomes searching for a lower bound on b where the probability of a sum of r independent variables $(a_1, \ldots, a_r)$ exceeding $b^p - \Delta^p_p\epsilon$ is smaller than δ. If there exists a closed-form distribution function for $\sum_{k=1}^r a_k$, an exact solution on b can be obtained. When $p = 2$, an analytical lower bound b can be obtained (see Section 3); when $p > 2$ we only manage to obtain the distribution function for $\binom{p}{j} |s^*_k - s_k|^{p-j} \Delta^j_{1,k}$, but not for a_k or $\sum_{k=1}^r a_k$ at the current stage. A relatively simple case is when the elements of statistics $\mathbf{s}$ are calculated on disjoint subsets of the original data, thus removing one individual from the data only affects one element out of r, $\Delta_1 = \Delta_p = \Delta_{1,k'}$, leading to the Corollary 9.

Corollary 9 (Lower Bound for b with Disjoint Queries). When elements in $\mathbf{s}$ are based on non-overlapping disjoint subsets of the data, the lower bound on b satisfies $\Pr(\sum_{j=1}^p \binom{p}{j} |s^*_{k'} - s_{k'}|^{p-j} \Delta_{1,k'} > b^p\epsilon) < \delta$ in the GG mechanism of (ϵ, δ)-DP, where $k' = \mathrm{argmax}_k \Delta_{1,k}$.

The proof of Corollary 9 is trivial. With disjoint queries, only one element in $\mathbf{s}$ is affected by changing from $\mathbf{x}$ to $\mathbf{x}'$ (if the definition $\mathbf{x}$ has one more/less record than $\mathbf{x}'$ is used) while the remaining elements in Eqn (B.2) in Appendix B are 0 as $s_k(\mathbf{x}) = s_k(\mathbf{x}')$, and Eqn (B.2) $= b^{-p}\sum_{j=1}^p \binom{p}{j} |e_{k'}|^{p-j} |d_{k'}|^j \leq |b^{-p}\sum_{j=1}^p \binom{p}{j} |e_{k'}|^{p-j} \Delta_{1,k'}$.

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/TKDE.2018.2845388, IEEE Transactions on Knowledge and Data Engineering

A special case of Corollary 9 is when the query is a histogram, $\Delta_1=\Delta_p=\Delta_{1,k'}=1$, and the lower bound b for (ϵ,δ)-pDP can be derived from $\Pr(\sum_{j=1}^{p}\binom{p}{j}|e_{k'}|^{p-j}>b^p\epsilon)<\delta$.

Numerical approaches can be applied to obtain a lower bound on b when the closed-form solutions are difficult to attain. Figure 2 depicts the lower bounds on b at different p and (ϵ,δ) obtained via the Monte Carlo approach. We set $\Delta_{1,k}$ at $1, 0.1, 0.05$ for $k=1,2,3$, respectively and applied Lemma 2 to obtain an upper bound on Δ_p for a given p value. As expected, the lower bound on b increases with decreased ϵ (lower privacy budget) and decreased δ (reduced chance of failing the pure ϵ-DP). The results also suggest b increases with p to maintain (ϵ,δ)-pDP in the examined scenarios.

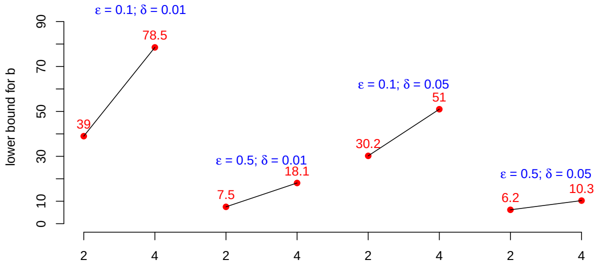

Fig. 2: Numerical Lower bound on b from Corollary 8

$\mathbf{s}^*$ sampled from the GG mechanism of (ϵ,δ)-pPD in Eqn (5) ranges $(-\infty,\infty)$. To bound $\mathbf{s}^*$, it is straightforward to apply a post processing procedure such as the truncation and the BIT procedure [17]. The truncation procedure throws away the out-of-bounds values and only keeps those in bounds while the BIT procedure sets the out-of-bounds values at the bounds. If the bounds are noninformative in the sense that the bounds are global and do not contain any data-specific information, then neither one of the two post-hoc bounding procedures will leak the original information or compromise the established (ϵ,δ)-pDP.

2.7 Connection Between GG Mechanism and Exponential Mechanism

The exponential mechanism was introduced by McSherry and Talwar [4]. Let $\mathcal{S}$ denote the set containing all possible output $\mathbf{s}^*$. The exponential mechanism releases $\mathbf{s}^*$ with probability

$$f(\mathbf{s}^*)=\exp\left(u(\mathbf{s}^*|\mathbf{x})\frac{\epsilon}{2\Delta_u}\right)(A(\mathbf{x}))^{-1} \tag{14}$$

to ensure ϵ-DP. $A(\mathbf{x})$ is a normalizing constant so that $f(\mathbf{s}^*)$ sums or integrates to 1, and equals to $\sum_{\mathbf{s}^*\in\mathcal{S}}\exp\left(u(\mathbf{s}^*|\mathbf{x})\frac{\epsilon}{2\Delta_u}\right)$ or $\int_{\mathbf{s}^*\in\mathcal{S}}\exp\left(u(\mathbf{s}^*|\mathbf{x})\frac{\epsilon}{2\Delta_u}\right)d\mathbf{s}^*$, depending on whether $\mathcal{S}$ is a countable/discrete sample space, or a continuous set, respectively. u is the utility function and assigns a "utility" score to each possible outcome $\mathbf{s}^*$ conditional on the original data $\mathbf{x}$, and $\Delta_u=\max_{\mathbf{x},\mathbf{x}',d(\mathbf{x},\mathbf{x}')=1,\mathbf{s}^*\in\mathcal{S}}|u(\mathbf{s}^*|\mathbf{x})-u(\mathbf{s}^*|\mathbf{x}')|$ is the maximum change in the utility score across all possible output $\mathbf{s}^*$ and all possible data sets $\mathbf{x}$ and $\mathbf{x}'$ that is $d(\mathbf{x},\mathbf{x}')=1$. From a practical perspective, the scores should properly reflect the "usefulness" of $\mathbf{s}^*$. For example, "usefulness" can be measured by the similarity between perturbed $\mathbf{s}^*$ and original $\mathbf{s}$ for numerical $\mathbf{s}$. The closer $\mathbf{s}^*$ is to the original $\mathbf{s}$, the larger $u(\mathbf{s}^*|\mathbf{x})$ is, and the higher the probability of releasing $\mathbf{s}^*$ is.

The Exponential mechanism can be conservative (See the online Supplementary Materials), in the sense that the actual privacy cost is lower than the nominal privacy budget ϵ, or more than necessary amount of perturbation is injected to preserve ϵ-DP. Despite the conservativeness, the Exponential mechanism is a widely used mechanism in DP with its generality and flexibility as long as the utility function u is properly designed.

When u is defined as the negative p^{th} power of the p^{th}-order Minkowski distance between $\mathbf{s}^*$ and $\mathbf{s}$, that is, $u(\mathbf{s}^*|\mathbf{s})=-\|\mathbf{s}^*-\mathbf{s}\|_p^p$, the Exponential mechanism generates perturbed $\mathbf{s}^*$ from the GG distribution

$$\begin{aligned}f(\mathbf{s}^*|\mathbf{s})&=(A(\mathbf{s}))^{-1}\exp\left(-\|\mathbf{s}^*-\mathbf{s}\|_p^p\tfrac{\epsilon}{2\Delta_u}\right) \\ &=(A(\mathbf{s}))^{-1}\textstyle\prod_{k=1}^{r}\exp\left(-\frac{|s_k^*-s_k|^p}{2\Delta_u\epsilon^{-1}}\right)=\prod_{k=1}^{r}\text{GG}(s_k,b,p)\end{aligned} \tag{15}$$

with $A(\mathbf{s})=\left(p^{-1}2b\Gamma(p^{-1})\right)^r$ and $b^p=2\Delta_u\epsilon^{-1}$. The scale parameter b in Eqn (15) is a function of the GS of the utility function Δ_u and the privacy budget ϵ. For bounded data $s_k^*\in[c_{k0},c_k]$ for $k=1,\ldots,r$, the Exponential mechanism based on the GG distribution is

$$\begin{aligned}&f(\mathbf{s}^*|\mathbf{s}^*\in[\mathbf{c}_0,\mathbf{c}_1])= \\ &(A(\mathbf{s}))^{-1}\textstyle\prod_{k=1}^{r}(B(s_k))^{-1}\exp\left(-\frac{|s_k^*-s_k|^p}{2\Delta_u\epsilon^{-1}}\right),\end{aligned} \tag{16}$$

where $B(s_k)=\Pr(s_k^*\in[c_{k0},c_k])$ is calculated from the pdf $\text{GG}(s_k,b,p)$. Compared to the truncated GG mechanism in Definition 6, the only difference in the Exponential mechanism in Eqn (16) is how the scale parameter b is defined. In Definition 6, b depends on the GS of $\mathbf{s}$ (Δ_p) while it is a function of the GS of the utility function u (Δ_u) in the Exponential mechanism ($b^p=2\epsilon^{-1}\Delta_u$ in the Exponential mechanism, and Eqn (7) gives the lower bound on b in $\mathcal{T}_\epsilon^p$). While both mechanisms lead to the satisfaction of ϵ-DP, the one with a smaller b at the same ϵ is preferable. The magnitude of b in each case depends on the bounds of $\mathbf{s}$, and the order p, in addition to Δ_u or Δ_p. Though not a direct comparison on b, Lemma 10 explores the relationship between Δ_u and Δ_p, with the hope to shed light on the comparison of b (the proof is in Appendix D).

Lemma 10. Relationships between Δ_u and Δ_p Let $[c_{k0},c_{k1}]$ denote the bounds on s_k for $k=1,\ldots,r$, and $u=-\|\mathbf{s}^*-\mathbf{s}\|_p^p$.

a) When $p=1$, $\Delta_u\le\Delta_1$, and the GG mechanism and the GG-distribution based Exponential mechanism both reduce to the truncated Laplace mechanism with the same b.

b) When $p=2$, $\Delta_u\le2\sum_{k=1}^{r}\Delta_{1,k}|c_{k1}-c_{k0}|$.

c) When $p\geq3$, $\Delta_u\le\sum_{k=1}^{r}\sum_{j=1}^{p}\binom{p}{j}\max\{|c_{k0}|,|c_{k1}|\}^{p-j}\Delta_{1,k}^{(j)}$, where $\Delta_{1,k}^{(j)}=\max_{\mathbf{x},\mathbf{x}',d(\mathbf{x},\mathbf{x}')=1}|(s_k(\mathbf{x}))^j-(s_k(\mathbf{x}'))^j|$ is the l_1 GS of $(s_k)^j$.

As a final note on the GG-distribution based Exponential mechanism, we did not use the negative Minkowski distance directly as the utility function due to a couple of potential practical difficulties with this approach. First, Δ_u can be difficulty to obtain. Second, $f(\mathbf{s}^*) \propto \exp\{-\left(\sum_{k=1}^r |s_k^* - s_k|^p\right)^{1/p} \epsilon(2\Delta_u)^{-1}\}$, does not appear to be associated with any known distributions (except when $p = 1$), and additional efforts are required to study the properties of $f(\mathbf{s}^*)$ and to develop an efficient algorithm to draw samples from it.

3 Gaussian Mechanism

A special case of the GG mechanism is the Gaussian mechanism when $p = 2$ that draws s_k^* independently from a Gaussian distribution with mean s_k and variance $\sigma^2 = b^2/2$ for $k = 1, \ldots, r$.

Applying Eqn (6) with b defined in Eqns (7) and (8), we can obtain the truncated Gaussian mechanism of ϵ-DP for bounded $\mathbf{s} \in [c_{10}, c_{11}] \times \cdots \times [c_{r0}, c_{r1}]$

$$
\begin{aligned}
f(\mathbf{s}^*|\mathbf{s}) = \textstyle\prod_{k=1}^r \Big\{ &\left(\Phi(c_{k1};\mu,\sigma^2) - \Phi(c_{k0};\mu,\sigma^2)\right)^{-1} \\
&\phi(s_k^*;\mu = s_k, \sigma^2 = b^2/2)\Big\}\text{ , where} \\
b^2 \geq 2\epsilon^{-1} &\left(2\textstyle\sum_{k=1}^r |c_{k1} - c_{k0}|\Delta_{1,k} + \Delta_2^2\right) \\
\geq 2\epsilon^{-1} &\textstyle\sum_{k=1}^r \left(2|c_{k1} - c_{k0}|\Delta_{1,k} + \Delta_{1,k}^2\right),
\end{aligned}
$$

where ϕ and Φ are the pdf and the CDF of the Gaussian distribution, respectively. An analytical solution on the lower bound of b for the Gaussian mechanism of (ϵ, δ)-pDP is provided in Proposition 11 (the proof is available in Appendix E).

Proposition 11 (Lower Bound on b for Gaussian Mechanism of (ϵ, δ)-pDP). The lower bound on the scale parameter b for the Gaussian mechanism of (ϵ, δ)-pDP is $b \geq 2^{-1/2}\epsilon^{-1}\Delta_2\left(\sqrt{(\Phi^{-1}(\delta/2))^2 + 2\epsilon} - \Phi^{-1}(\delta/2)\right)$.

Given the relationship between b and the standard deviation of the Gaussian distribution $\sigma = b/\sqrt{2}$, the lower bound can also be expressed in σ,

$$
\sigma \geq (2\epsilon)^{-1}\Delta_2\left(\sqrt{(\Phi^{-1}(\delta/2))^2 + 2\epsilon} - \Phi^{-1}(\delta/2)\right). \quad (17)
$$

The pDP lower bound given in Eqn (17) is different from the lower bound

$$
\sigma > \epsilon^{-1}\Delta_2 c,\ \text{ with } \epsilon \in (0,1) \text{ and } c^2 > 2\ln(1.25/\delta). \quad (18)
$$

in Dwork and Roth [10] for (ϵ, δ)-aDP (Eqn (2)). The pDP bound in Eqn (17) is tighter than the aDP bound in Eqn (18) for the same set of (ϵ, δ) (note the interpretation of δ in pDP and aDP is different, but the DP guarantee is roughly the same when δ is small). In addition, the pDP bound does not constrain ϵ to be < 1 as required in the aDP bound. Figure 3 compares the two two lower bounds at several $\epsilon \in (0, 1)$ and $\delta \in (0, 0.5)$. As observed, the ratio between the aPD vs. pDP lower bounds is always < 1 for the same (ϵ, δ). The smaller ϵ is, or the larger δ is, the smaller the ratio is and the larger the difference is between the two bounds.

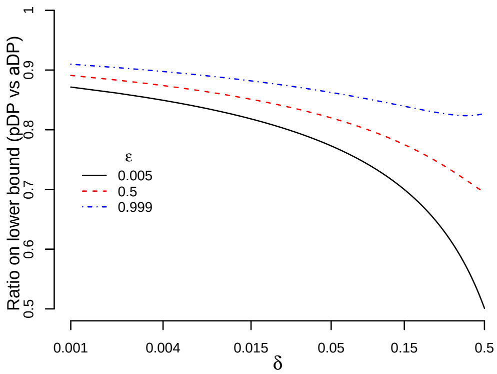

Fig. 3: Comparison of pDP lower bound (Eqn 17) vs. aDP bound (Eqn 18) on σ in the Gaussian mechanism for $\epsilon < 1$ (the aDP bound requires $\epsilon < 1$)

Dwork and Roth [10] list several advantages of the Gaussian noises. For example, the Gaussian noise is a "familiar" type of noise as many noise sources in real life can be well approximated by Gaussian distributions; the sum of Gaussian variable is still a Gaussian; and finally, in the case of multiple queries or when δ is small, the pure-DP guarantee in the Laplace mechanism and the pDP guarantee in the Gaussian mechanism are very similar. A theoretical disadvantage to the Gaussian mechanism is that it does not guarantee DP in some cases (e.g., Report Noisy Max) [10].

We investigate the accuracy of $\mathbf{s}^*$ by examining the tail probability and the dispersion of the noises injected via the ϵ-DP Laplace mechanism and the (ϵ, δ)-pDP Gaussian mechanism. Denote the noise drawn from the Laplace distribution by e_1 and that from the Gaussian distribution by e_2; the location parameters in both are 0. The tail probability $p_1 = \Pr(e_1 > |t|) = \exp(-|t|\epsilon/\Delta_1)$ in the Laplace distribution and $p_2 = \Pr(e_2 > |t|) = 2\Phi(-|t|/\sigma)$ in the Gaussian distribution, where σ is given in Eqn (17). Since the CDF $\Phi()$ does not have a close-formed expression, we examine several numerical examples to compare p_1 and p_2 (Figure 4). We set ϵ to be the same (0.1, 1, 2, respectively) between the two mechanisms and examine $\delta = (1\%, 5\%, 10\%, 20\%)$ for the (ϵ, δ)-pDP Gaussian mechanism. If the ratio $p_1 : p_2$ is < 1, it implies that the Laplace mechanism is less likely to generate more extreme $\mathbf{s}^*$ compared to the Gaussian mechanism at the same privacy specification of ϵ. We should focus on the meaningful case where noise $|t|$ has a non-ignorable chance to occur in either mechanism. We used cutoff 10^{-4}; that is, either $p_1 > 10^{-4}$ or $p_2 > 10^{-4}$ (other cutoffs can be used, depending on how "non-ignorable" is defined). It is interesting to observe that after the initial take-off at 1 at $|t| = 0$, the ratio continues to decrease until hitting the bottom and then bounds back with some cases eventually exceeding 1 at some value of $|t|$. The smaller ϵ or δ is, the longer it takes for the bounce-back to occurs. The observation suggests that the Laplace mechanism in some cases is more likely to generate sanitized results $\mathbf{s}^*$ that are far away from $\mathbf{s}$.

We also compare the privacy cost ϵ between the two mechanisms when they yield the same tail probability.

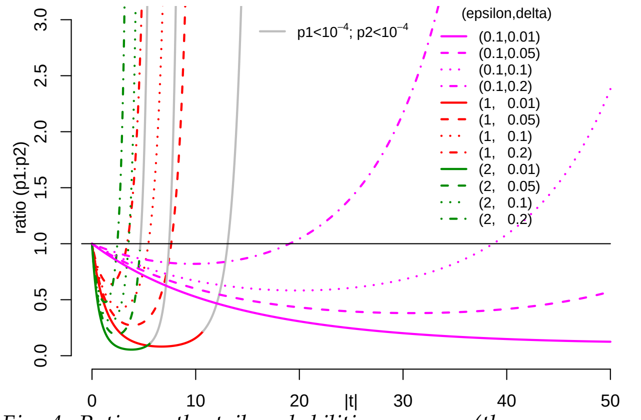

Fig. 4: Ratio on the tail probabilities $p_1 : p_2$ (the gray curves represent the unlikely cases where both p_1 and p_2 are $< 10^{-4}$)

Figure 5 shows the calculated ϵ_2 value associated with the Gaussian mechanism of (ϵ_2, δ)-DP for a given δ that yields $\Pr(e_2 < |t|) = \Pr(e_1 < |t|)$ for the Laplace mechanism of ϵ_1-DP. If the ratio of $\epsilon_2 : \epsilon_1 < 1$ at some $|t|$ and a somewhat small ignorable δ, it implies the same tail probability can be achieved with less privacy cost with the Gaussian mechanism compared to the Laplace mechanism. Figure 5 suggests that at the same $|t|$, the more relaxation of the pure ϵ-DP is allowed (i.e., the larger δ is), the smaller ϵ_2 is (relative to baseline ϵ_1), which is expected as the ϵ and δ together determine the noise released in the Gaussian mechanism.

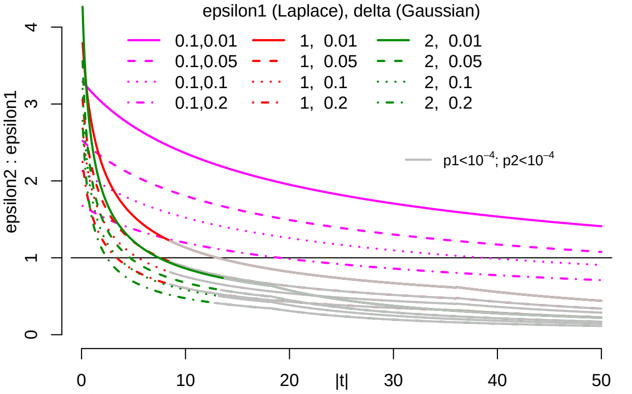

Fig. 5: Relative privacy cost $\epsilon_2 : \epsilon_1$ (the gray curves represent the unlikely cases where both p_1 and p_2 are $< 10^{-4}$)

Proposition 12 (Precision of Sanitized s^* in Gaussian Mechanism of (ϵ, δ)-pDP and Laplace Mechanism of ϵ-DP). Between the Gaussian mechanism of (ϵ, δ)-pDP and the Laplace mechanism of ϵ-DP for sanitizing a statistic s, when $\delta < 2\Phi(\sqrt{2}) \approx 0.157$, the variance of the injected Gaussian noise is always greater than the variance of the Laplace noise.

The proof is provided in Appendix F. If the associated Laplace distribution and the Gaussian distribution have the same location parameter, a larger variance also implies a smaller mean squared error (MSE) of the injected noise. Proposition 12 suggests that there is more dispersion in the perturbed s^* released by the Gaussian mechanism of $(\epsilon, \delta < 0.157)$-pDP than the Laplace mechanism of ϵ-DP. In other words, if there are multiple sets of s^* released via the Gaussian and the Laplace mechanisms respectively, then the former sets would have a wider spread than the latter. Since (ϵ, δ)-pDP provides less privacy protection than ϵ-pDP, together with the larger MSE, it can be argued that the Laplace mechanism is superior to the Gaussian mechanism (which is also reflected in the 3 experiments in Section 4). It should be noted that $\delta < 0.157$ in Proposition 12 is a sufficient but not necessary condition. In other words, the Gaussian mechanism may not be less dispersed than the Laplace mechanism when $\delta \geq 0.157$. Furthermore, since δ needs to be small to provide sufficient privacy protection in the setting of (ϵ, δ)-pDP, it is very unlikely that δ as large as > 0.157 will be used in practical applications. Also noted is that the setting explored in Proposition 12, where the focus is on examining the precision (dispersion) of a single perturbed statistic given specific privacy parameters when the sample size of a data set is public, is different from the work on bounding sample complexity (required sample size) to reach a certain level of a statistical *accuracy* in perturbed results with ϵ-DP or (ϵ, δ)-aDP [18] (refer to Section 5 for more discussions).

4 Experiments

We run three experiments on the mildew data set, the Czech data set, and the Census Income data set (a.k.a. the adult data). The mildew data contains information of parental alleles at 6 loci on the chromosome for 70 strands of barley powder mildew [19]. Each loci has two levels, yielding a very sparse 6-way cross-tabulation (22 cells out of the 64 are non-empty with low frequencies in many other cells). The Czech data contains data collected on 6 potential risk factors for coronary thrombosis for 1841 workers in a Czechoslovakian car factory [19]. Each risk factor has 2 levels (Y or N). The cross-tabulation is also 6-way with 64 cells, the same as the mildew data, but table is not as sparse with the large n (only one empty cell). The adult data was extracted from the 1994 US Census database to yield a set of reasonably clean records that satisfy a set of conditions [20]. The data set is often used to test classifiers by predicting whether a person makes over 50K a year. We used only the completers (without missing values on the attributes) in the adult data and then split them to 2/3 training (20009 subjects) and 1/3 testing (10005 subjects).

In each experiment, we run the Laplace mechanism of ϵ-DP, the Gaussian mechanism of (ϵ, δ)-pDP presented in Section 3, and the Gaussian mechanism of of (ϵ, δ)-aDP [10] to sanitize count data. We examined $\epsilon = 0.5, 1, 2$ and $\delta = 0.01, 0.05, 0.1, 0.25$. To examine the variation of noises, we run 500 repeats and computed the means and standard deviations of the l_1 distances between the sanitized and the original counts and the Kullback-Leibler (KL) divergence between the empirical distributions of the synthetic data and the original data over the 500 repeats. In addition, we tested the GG mechanism of order 3 ($p = 3$) in the mildew data, and compared

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/TKDE.2018.2845388, IEEE Transactions on Knowledge and Data Engineering

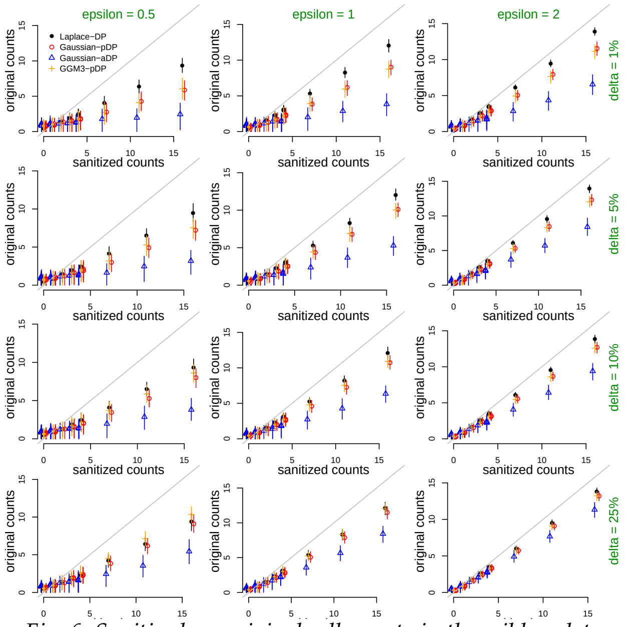

Fig. 6: Sanitized vs. original cell counts in the mildew data

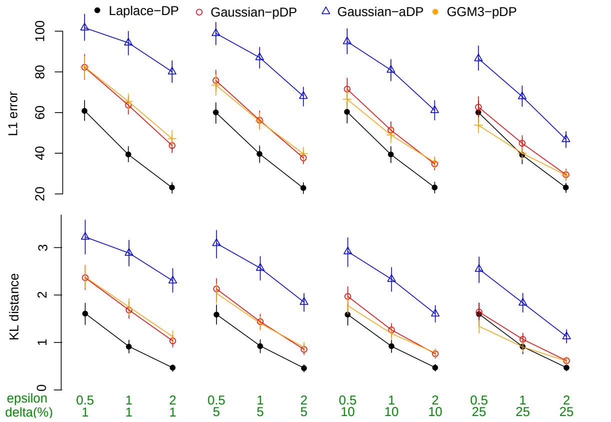

Fig. 7: l_1 Distance and KL divergence between sanitized and original counts in the mildew data

the classification accuracy of the income outcome in the testing data set in the adult experiment based on the support vector machines (SVMs) trained with the original data and the sanitized data, respectively. The KL distance was calculated using the `KL.Dirichlet` command in R package `entropy` The SVMs were trained using the `svm` command in R package `e1071`. In all experiments, $\Delta_p = 1$ for all p since the released query is a histogram and the bin counts are based on disjoint subsets of data. The scale parameters of the Laplace mechanism and the Gaussian mechanisms were obtained analytically ($\Delta_1\epsilon^{-1}$, Eqns (17) and (18), respectively). A grid search and the MC approach were applied to obtain the lower bound b for GGM-3 via Corollary 9. In the mildew and Czech experiments, we sanitized all bins in the histograms, including the empty bins, assuming all combinations of the 6 attributes in each case are practically meaningful (in other words, the empty cells are sample zeros rather than population zeros). In the adult data, there are 14 attributes and $\sim 1.944 \times 10^{13}$ bins in the 14-attribute histogram, a non-ignorable portion of which do not make any practical sense (e.g., a 90-age works $>$ 80 hours per week). For simplicity, we only sanitized the 17,985 nonempty cells in the training data. After the sanitization, we set the out-of-bound synthetic counts < 0 at 0 and those $> n$ at n, respectively, and normalized the sanitized counts to sum up to the original sample size n in all 3 experiments, assuming n itself is public or does not carry privacy information.

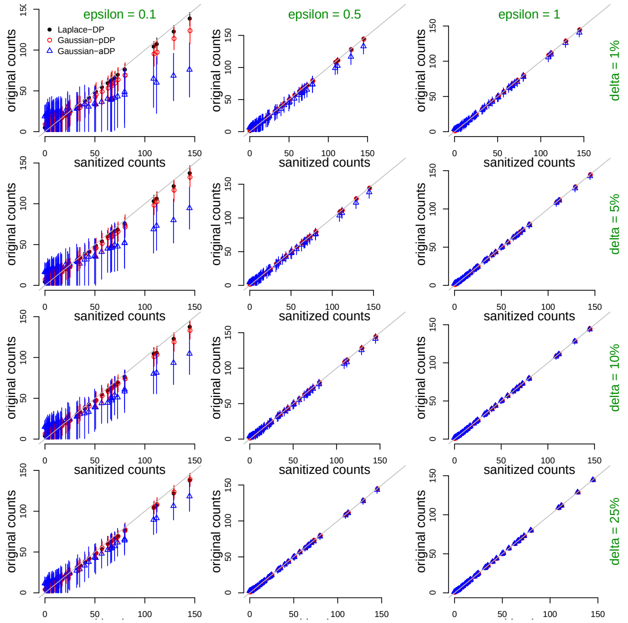

Fig. 8: Sanitized vs. original cell counts in the Czech data

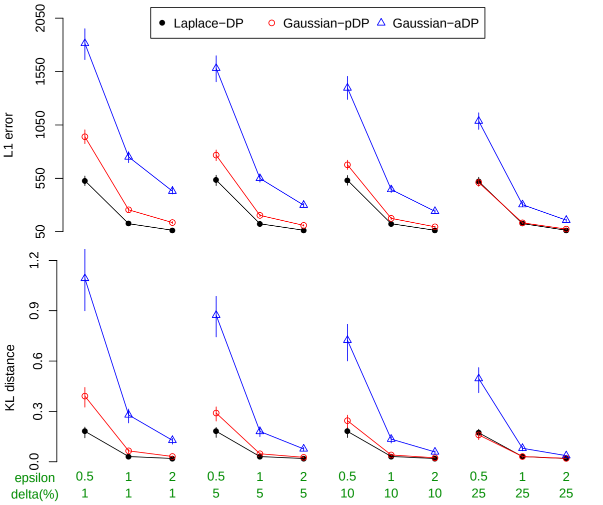

Fig. 9: l_1 Distance and KL divergence between sanitized and original counts in the Czech data

The results are given in Figures 6 to 12. In Figures 6, 8 and 10, the closer the points are to the identity line, the more similar are the original and sanitized counts. The Laplace sanitizer is the obvious winner in all 3 cases, producing the sanitized counts closest to the original with the smallest l_l error and the KL divergence, followed by a similar performance from the Gaussian mechanism of (ϵ, δ)-pDP and GGM3 of (ϵ, δ)-pDP in the mildew data; the Gaussian mechanism of (ϵ, δ)-aDP is the worst performer. Specifically, in the mildew experiment, the

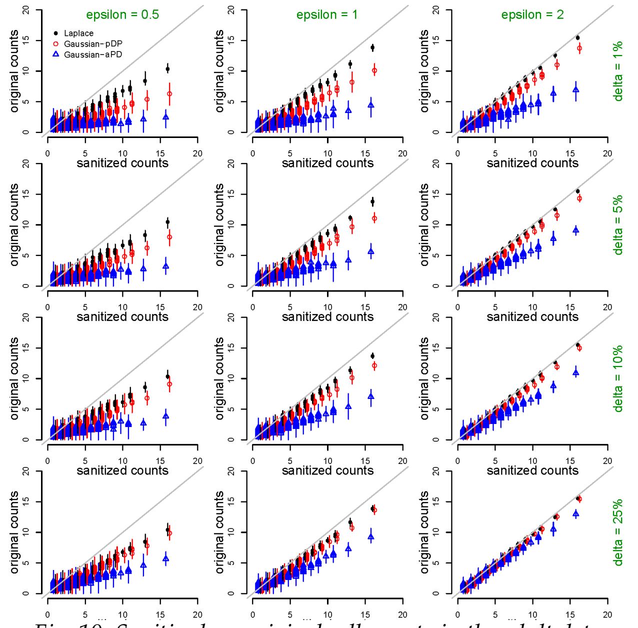

Fig. 10: Sanitized vs. original cell counts in the adult data

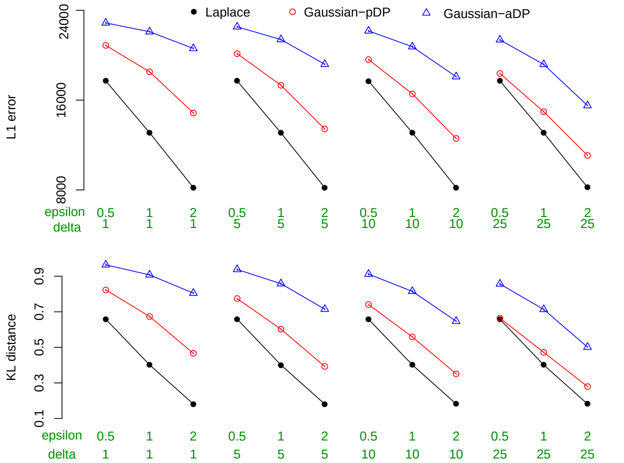

Fig. 11: l_1 Distance and KL divergence between sanitized and original counts in the adult data

performance of the Laplace sanitizer and the Gaussian mechanism of (ϵ, δ)-pDP is similar when $\epsilon = 2$ or $\delta \geq 0.1$. The l_1 error and the KL divergence seems to decrease more or less in a linear manner as ϵ increases from 0.5 to 1 to 2, while the impact of δ seemed to have less a profound impact on the l_1 error and the KL divergence. In the Czech experiment, the sanitized counts approach the original counts more quickly than the mildew case with increased ϵ and δ, but there is significantly more variability for small ϵ (0.1); and the l_1 error and the KL divergence no longer decreases in a linear fashion, but drastically from $\epsilon = 0.5$ to 1 and much more slowly from $\epsilon = 1$ to 2. The differences in the results between the mildew and the Czech experiments can be explained by the larger n in the latter. In the adult experiment, Figure 12 suggests the prediction accuracy via the SVMs built on sanitized data is barely affected compared to the original accuracy regardless of the mechanism. There are some decreases in the accuracy rates from the original, but they are largely ignorable (on the scale of 0.25% to 1%), even with the variation taken into account. In addition, the Gaussian mechanism of (ϵ, δ)-aDP, though being the worst in preserving the original counts measured the l_1 distance and KL divergence, is no worse than the two Gaussian mechanisms in the SVM prediction.

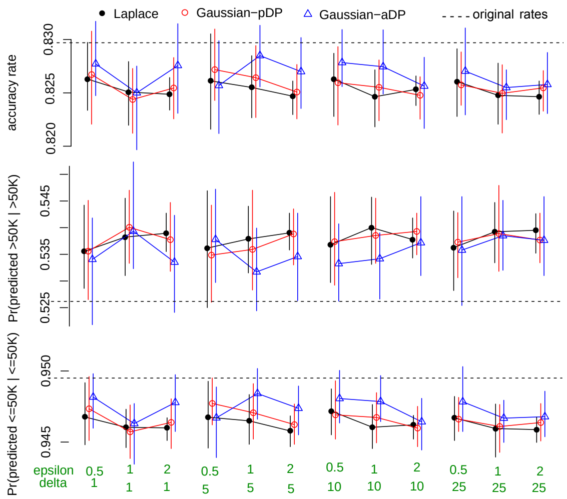

Fig. 12: Prediction accuracy in testing data via SVMs trained on sanitized and original data in the adult data

5 Discussion

We introduced a new concept called the l_p GS, and unified the Laplace mechanism and the Gaussian mechanism in the family of the GG mechanism. For bounded data, we discussed the truncated and the BIT GG mechanisms to achieve ϵ-DP. We also proposed (ϵ, δ)-pDP as an alternative paradigm to the pure ϵ-DP for the GG mechanism for order $p \geq 2$. We showed the connections and distinctions between the GG mechanism and the Exponential mechanism when the utility function is defined as the negative p^{th}-power of the Minkowski distance between the original and sanitized results. We also presented the Gaussian mechanism as an example of the GG mechanism and derived a lower bound for the scale parameter of the associated Gaussian distribution to achieve (ϵ, δ)-pDP. The bound is tighter than the lower bound for the Gaussian mechanism of (ϵ, δ)-aDP. We compared the tail probability and the dispersion of the the noise generated via the Gaussian mechanism of (ϵ, δ)-pDP and the Laplace mechanism. We finally applied the Gaussian mechanisms of (ϵ, δ)-pDP and (ϵ, δ)-aDP and the Laplace mechanism of ϵ-DP in three real-life data sets.

The GG mechanism is based on the l_p "global" sensitivity of query results in the sense that the sensitivity is independent of any specific data. Though the employment of the GS is robust in terms of privacy protection, it could result in a large amount of noises being injected to query results. There is work that allows the sensitivity of a query to vary with data ("local" sensitivity) [21], [22] with the purpose to increase the accuracy of sanitized results. How to develop the GG mechanism in the context of local sensitivity is a topic for future investigation.

The examination on the tail probability and dispersion of the sanitized results in the Gaussian mechanism in Section 3 has a different focus from, though related to, the work on bounding the sample complexity that examines the required sample size n to reach a certain level of accuracy α with (ϵ, δ)-privacy guarantee for count queries [18], [23], [24]. α often refers to the accuracy of the perturbed results in the DP literature, such as the worst case accuracy l_∞, the average accuracy l_1, or the tail probability and the MSE of the released data, among others. The existing work on sample complexity focuses on bounding n given ϵ (and δ) and α, while the results in Section 3 focus on the the accuracy and precision of sanitized results given ϵ (and δ) and n. If the bias from perturbed results (relative to the original results) are the same between the two mechanisms, a larger precision is equivalent to a smaller MSE.

Acknowledgments

This work was supported by the NSF Grants #1546373, #1717417, and the University of Notre Dame Faculty Research Support Initiation Grant Program. We also thank the editor, associate editor, and two reviewers for their valuable comments and suggestions that greatly improved the quality of the manuscript.

References

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

[2] C. Dwork, "Differential privacy: A survey of results," *Theory and Applications of Models of Computation*, vol. 4978, pp. 1–19, 2008.

[3] ——, "Differential privacy," in *Encyclopedia of Cryptography and Security*. Springer, 2011, pp. 338–340.

[4] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *48th Annual IEEE Symposium on Foundations of Computer Science*, 2007, pp. 94–103.

[5] F. McSherry, "Privacy integrated queries: an extensible platform for privacy-preserving data analysis," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, 2009, pp. 19–30.

[6] A. Roth and T. Roughgarden, "Interactive privacy via the median mechanism," in *Proceedings of the 42nd ACM Symposium on Theory of Computing*, June 5-8, 2010.

[7] M. Hardt, K. Ligett, and F. McSherry, "A simple and practical algorithm for differentially private data release," *arXiv:1012.4763v2*, 2012.

[8] A. Ghosh, T. Roughgarden, and M. Sundararajan, "Universally utility-maximizing privacy mechanisms," *SIAM Journal on Computing*, vol. 41, no. 6, pp. 1673–1693, 2012.

[9] Q. Geng and P. Viswanath, "The optimal noise-adding mechanism in differential privacy," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 925–951, 2016.

[10] C. Dwork and A. Roth, *The Algorithmic Foundation of Differential Privacy*. Now Publishes, Inc., 2014.

[11] C. Dimitrakakis, B. Nelson, A. Mitrokotsa, and B. Rubinstein, "Robust and private bayesian inference," in *Algorithmic Learning Theory ALT 2014*, P. Auer, A. Clark, T. Zeugmann, and S. Zilles, Eds. Spring, Cham, 2014.

[12] C. Dwork, "Differential privacy," in *Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP)*. Springer-Verlag ARCoSS, 2006, pp. 1–12.

[13] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: privacy via distributed noise generation," in *Advances in Cryptology: Proceedings of EUROCRYPT*. Springer Berlin Heidelberg, 2006, pp. 485–503.

[14] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets practice on the map," *IEEE ICDE 24th International Conference*, pp. 277 – 286, 2008.

[15] R. Hall, A. Rinaldoy, and L. Wasserman, "Random differential privacy," *Journal of Privacy and Confidentiality*, vol. 4, no. 2, pp. 43–59, 2012.

[16] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," *arXiv:1603.01887v2*, 2016.

[17] F. Liu, "Noninformative bounding in differential privacy and its impact on statistical properties of sanitized results in truncated and boundary-inflated-truncated laplace mechanisms," *arXiv:1607.08554*, 2016.

[18] T. Steinke and J. Ullman, "Between pure and approximate differential privacy," *arXiv:1501.06095v1*, 2015.

[19] A.-S. Charest, "Empirical evaluation of statistical inference from differentially-private contingency tables," in *Proceeding of International Conferency on Privacy in Statistical Databases*, 2012, pp. 257–272.

[20] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[21] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth sensitivity and sampling in private data analysis," *Proceedings of the 39th ACM Symposium on Theory of Computing*, p. 7584, 2007.

[22] C. Dwork and J. Lei, "Differential privacy and robust statistics," *Proceedings of the 41rd ACM symposium on Theory of computing*, pp. 371–380, 2009.

[23] M. Hardt and K. Talwar, "On the geometry of differential privacy," *Proceedings of the Forty-second ACM Symposium on Theory of Computing, STOC '10*, pp. 705–714, 2010.

[24] B. Mark, J. Ullman, and S. Vadhan, "Fingerprinting codes and the price of approximate differential privacy," *arXiv:1311.3158v2*, 2015.

PLACE
PHOTO
HERE

Fang Liu Fang Liu is an Associate Professor in the Department of Applied and Computational Mathematics and Statistics at the University of Notre Dame. She obtained her Ph.D. from University of Michigan, Ann Arbor. Her research interests include data privacy and statistical disclosure limitation, statistical learning of big data and model regularization, Bayesian methodology, and modelling and analysis of missing data.