Achieving Accuracy Guarantee for Answering Batch Queries with Differential Privacy

Dong Huang(✉), Shuguo Han, and Xiaoli Li

Institute for Infocomm Research, Singapore, Singapore
{huangd,shan,xlli}@i2r.a-star.edu.sg

Abstract. In this paper, we develop a novel strategy for the privacy budget allocation on answering a batch of queries for statistical databases under differential privacy framework. Under such a strategy, the noisy results are more meaningful and achieve better utility of the dataset. In particular, we first formulate the privacy allocation as an optimization problem. Then derive explicit approximation of the relationships among privacy budget, dataset size and confidence interval. Based on the derived formulas, one can automatically determine optimal privacy budget allocation for batch queries with the given accuracy requirements. Extensive experiments across a synthetic dataset and a real dataset are conducted to demonstrate the effectiveness of the proposed approach.

1 Introduction

Differential privacy (DP), is a promising strategy for providing privacy for data publishing and data queries [6,8]. A simple but feasible method to achieve differential privacy is to insert noises to the query outputs [4]. Currently, most of the related work focus on privacy protection but don't further analyze how useful of the noisy results. If these noise results show what level of accuracy can be achieved, they will help data analysts further investigate and improve the effectiveness of them. Moreover, the privacy allocation is very important to how useful of the noisy results. To answer multiple queries, a simple way is to allocate the privacy budget to these queries *equally* [7]. However, such a strategy may cause some noisy results to be unmeaningful due to large noise magnitude relative to the original results.

We use the following intuitive example to illustrate the problem. The *Adult* dataset, extracted from UCI machine learning repository [2], has $32,561$ individuals. Suppose we want to use two queries $q = [q_1, q_2]$ on the *Adult* dataset to infer the real values of the whole population, where q_1 is the proportion of individuals with Sex="Male", Race="Black" and Income=">50K" over the number of individuals with Sex="Male" and Income=">50K", which is equal to 0.0446; q_2 is the proportion of individuals with Sex="Male", Race="White" and Income=">50K" over the number of individuals with Sex="Male" and Income=">50K", which is equal to 0.9140.

We observe that the returned real value from q_1 (i.e. 0.0446) is much smaller than that from q_2 (i.e. 0.9140). Under the aforementioned "uniformly split" strategy, equal budgets will be assigned to q_1 and q_2, e.g. each gets a budget 0.5.

T. Cao et al. (Eds.): PAKDD 2015, Part II, LNAI 9078, pp. 305–316, 2015.
DOI: 10.1007/978-3-319-18032-8_24

Table 1. Statistics of the *Adult* dataset

Sex	Race	Income	
		$> 50K$	$<= 50K$
Male	Amer-Indian-Eskimo	24	168
Male	Black	297	1272
Female	Other	6	103
Male	White	6089	13085
...	...	...	...

Correspondingly, the noise magnitude for q_1 will be larger compared with its smaller value, making its noisy result less useful, although the q_2's noisy result could be relatively close to its true value. Ideally, we should assign a higher budget for q_1 and lower budget for q_2, such that the lower noise magnitude will be added to q_1, making both of them are useful. As such, how to reasonably allocate the limited privacy budget to multiple queries is a crucial problem to ensure the overall accuracy guarantee for all the queries.

For the impact of noise on the noisy results, while most existing work provide the analysis of the upper error bound [3,17] in implementing differential privacy, this is not sufficient for the utility evaluation of the noisy results. From data mining perspective, it will be helpful for data analysts to understand the utility of the noisy results if they can visualize the level of accuracy achieved after adding noise.

The objective of this research is thus to design a framework to allocate privacy budget among the queries with differential privacy and further provides analysis of how useful of the noisy results. We have further investigated the framework proposed by A. Smith [14]. We consider the problem of multiple queries with ϵ-differential privacy under this framework, where the queries studied in the paper are the ratios of multiple subsets to the given dataset. The contributions of this paper can be summarized as follows:

- We formulate the optimization problem with accuracy guarantee in terms of confidence interval (CI). This enables data analysts to better understand what are the accuracy guarantee of the noisy statistical results.
- We formulate the noisy results with normal-Laplace distribution. This property enables us to derive its cumulative distribution function (i.e. cdf).
- We further approximate the minimum privacy budget required for given level of accuracy with explicit formulas.

The remaining parts of the paper are organized as follows: First, section 2 provides a brief discussion about the related work. Then, we describe the background information in Section 3. Next, section 4 presents the differential privacy framework and discusses the normal-Laplace distribution. Section 5 introduces our novel approximation formulas for accuracy guarantee. Finally, we evaluate the proposed approach and conclude the paper in section 6 and section 7 respectively.

2 The Related Work

Currently, most of existing work studied the noise reduction in terms of sensitivity. For example, a data publishing technique, *Privelet*, based on wavelet transforms, was proposed in [16]. *Privelet* not only ensures ϵ- differential privacy, but also guarantees that the variance of the noisy results is polylogarithmic in terms of m where m denotes the number of queries. There are some database/ data mining applications, where the given dataset is a correlated time-series data or the dataset is distributively collected. For such types of the applications, a differential privacy framework, PASTE, was proposed in [12]. To provide differential privacy for time-series data, PASTE developed a Fourier perturbation algorithm. For the case of absence of a trusted central server, PASTE used a distributed Laplace perturbation algorithm to guarantee differential privacy. In order to publish cuboids for data cubes with small noise, an efficient method was proposed in [5]. The proposed method ensures that the maximal noise in all published cuboids will be within a factor $(\ln|\mathcal{L}|+1)^2$ of the optimal, where $|\mathcal{L}|$ is the number of cuboids to be published. To handle the problem of differential private data release for a class of counting queries, a new computationally efficient method based on learning thresholds was proposed in [9].

We notice that privacy budget has important impact on the noise magnitude. A few related work on this topic have been investigated. For example, K. Nissim *et al.* have proposed a framework, subsample and aggregate, to reduce the noise magnitude [11,14]. In such a framework, the dataset is first divided into k groups. Then it estimates the parameters based on the k results. Compared with traditional Laplace mechanism in [8], the framework reduces the error dramatically, where the errors decrease with the increasing number of data. The GUPT was proposed in [10] to allocate the privacy budget by ensuring the same noise magnitude for each query.

3 Background

Definition 1 ([15]). *Two databases x, $x' \subseteq D^n$ are neighbouring databases if they differ on exactly one record, i.e.,*

$$x = \{x_1, \ldots, x_i, \ldots, x_n\} \quad \text{and} \quad x' = \{x_1, \ldots, x'_i, \ldots, x_n\}$$

From the definition, note that two neighbouring databases differ only one record while they have the same cardinality. Before discussing the Laplace mechanism, we first give the definition of ϵ-differential privacy proposed by Dwork [8].

Definition 2 ([8]). *With $\epsilon > 0$, a randomized algorithm $\mathcal{K} : D^n \to \mathbb{R}^l$ is said to satisfy ϵ-differential privacy, if for any two neighbouring databases x, $x' \subseteq D^n$ and for any subset of outputs $S \subseteq Range(\mathcal{K})$, the following condition holds:*

$$\frac{Pr(\mathcal{K}(x) \in S)}{Pr(\mathcal{K}(x') \in S)} \leq \exp(\epsilon) \tag{1}$$

where the probability is taken over the randomness of $\mathcal{K}$.

A simple way to achieve ϵ-differential privacy is to insert noise to the true value. For instance, the Laplace mechanism (LM) generates the noise by using a random variable of Laplace distribution with mean equal to 0 and scale parameter equal to S_T/ϵ,where S_T is the sensitivity of $T(X)$.

Definition 3 ([11]). *The sensitivity of a function $T : D^n \rightarrow \mathbb{R}^d$ is*

$$S_T = \max_{x,x':d(x,x')=1} \|T(x) - T(x')\|. \tag{2}$$

Lemma 1 (Laplace Mechanism [8]). *Given $\epsilon > 0$, a statistic $T(X) \in \mathbb{R}^l$ and its sensitivity S_T, the noisy result*

$$T^*(X) = T(X) + Lap(S_T/\epsilon)) \tag{3}$$

satisfies ϵ-differential privacy.

For the convenience of discussion, we define the Laplace mechanism with accuracy guarantee as follows.

Definition 4 (Accuracy Guarantee). *Let $\theta \in \mathbb{R}^l$ be the true value of the statistic $T(X)$. Suppose $\Phi(W)$ is the cumulative distribution function of a random variable of Laplace distribution with mean equal to 0 and scale parameter equal to S_T/ϵ. If there exist $d = (d_1, \ldots, d_l) > 0$ and $\alpha = (\alpha_1, \ldots, \alpha_l) \in (0, 1)$ such that*

$$\Phi(d_i) \geq 1 - \alpha_i/2, \forall i = 1, 2, \ldots, l,$$

then the noisy results obtained from Eq. (3) is at least $100(1-\alpha)\%$ to be $\pm d$ of the true value θ.

Note that the accuracy achieved by LM may be smaller than the required accuracy since there exists estimation error between the statistic $T(X)$ and the true value θ but the error is not considered for the accuracy estimation (i.e., only the variance of the Laplace noise is considered here). We need to find a more accurate formula to describe the relationships among the accuracy achieved and other related parameters in order to ensure the accuracy achieved is close to the required accuracy. Specifically, we focus on the maximum likelihood estimator (MLE), which is obtained by maximizing the likelihood function.

Theorem 1 (Asymptotic Distribution). *Let $x_1, x_2, \ldots, x_n$ be independently identically distributed with density $f(x|\theta)$, $\theta \in \Theta$ and let θ_0 denote the true value of θ. Suppose the MLE estimator of θ_0 is $T(x)$. Then the probability distribution of*

$$\sqrt{nI(\theta_0)}(T(x) - \theta_0)$$

tends to be a standard normal distribution, i.e.,

$$\sqrt{nI(\theta_0)}(T(x)) - \theta_0) \xrightarrow{D} N(0, 1)^l, \tag{4}$$

where $I(\theta_0)$ is Fisher information.

4 Differential Privacy Framework

In this section, we first formulate the optimization problem and then show the relationship between the problem and normal-Laplace distribution.

4.1 Problem Definition

The scenarios we consider in the paper is illustrated in Figure 1. Users first send queries with accuracy requirements to the database server, then the database server passes the requirements to the computing server to optimize the parameters such as privacy budget and confidence interval in order to minimize the expectation of the errors in Eq. (5). Finally, the database server executes queries based on the optimized parameters and returns users the noisy results with accuracy description.

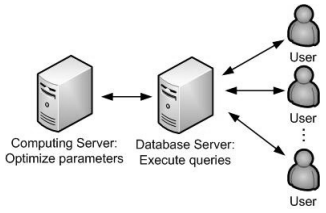


Fig. 1. The query execution model

Given a dataset $x = (x_1, \ldots, x_n) \in D^n$, the value of x_i in D is a real number, where D is a value space of x_i. Here, we use the capital letter $X = (X_1, \ldots, X_n)$ to denote a random vector variable and lower case $x = (x_1, \ldots, x_n) \in D^n$ to denote a specific value in D^n. Suppose $X = (X_1, \ldots, X_n)$ is drawn according to the distribution $f(x|\theta)$, where $\theta \in \mathbb{R}^l$ is unknown parameter vector. Let

$$T(X) = [T_1(X), \ldots, T_l(X)]$$

be the estimator of θ. In this paper, we study the problem of how to ensure that the parameter estimations under ϵ-differential privacy satisfy the given level of precision. In other words, we wish to estimate θ using an estimator based on the given dataset $x = (x_1, \ldots, x_n) \in D^n$ with $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_l]$ confidence interval to be $\pm\mathbf{d}$, where $\mathbf{d} = [d_1, \ldots, d_l]$, of the true value θ_0. Here, we want to minimize the expected squared deviation from the real parameter θ. Specifically, we wish to minimize the following objective function:

$$\textbf{min} \quad J_\theta(T^*(X)) = \mathrm{E}\{\|T^*(X) - \theta\|^2\} \tag{5}$$

$$\text{s.t} \quad \Pr(|T_i^*(X) - \theta_i| \leq d_i) \geq 1 - \alpha_i, \forall i \in \{1, 2, \ldots, l\} \tag{6}$$

$$\sum \epsilon_i = \epsilon_{total}, \forall i \in \{1, 2, \ldots, l\} \tag{7}$$

$$0 < k \leq n \tag{8}$$

where ϵ_{total} is the total privacy budget and k is the number of blocks. In order to solve this problem, we need to derive explicit formula to characterize the relationships among the privacy budget, the number of blocks, the statistics of the data and accuracy guarantee. Without loss of generality, we assume the solution of the problem in Eqs. (5)-(8) always exists. Specifically, the multiple queries we consider in this paper are the ratios of multiple subsets to the given dataset.

Specifically, suppose users send queries $Q_1(\mathbf{q}, \boldsymbol{\chi})$, where $\mathbf{q} = [q_1, \ldots, q_l]$ denotes the query vector while $\boldsymbol{\chi}$ represents the corresponding accuracy constraints. The database server then passes the queries with privacy budget, $Q_2(\mathbf{q}, \boldsymbol{\chi}, \boldsymbol{\vartheta}, \epsilon_{total})$, where $\boldsymbol{\vartheta}$ denotes the required statistics of related dataset for parameter optimization, to the computing server. It optimizes the privacy budget among the queries by solving the problem in Eqs. (5)-(8) and returns the execution queries with optimized parameters, $Q_3(\mathbf{q}, \boldsymbol{\chi}, \boldsymbol{\epsilon}, \varsigma)$, where $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_l]$ denotes the privacy budget allocation for the query vector $\mathbf{q}$ and ς is the number of blocks, to the execution server. Finally, the execution server executes the queries according to given optimized parameters and returns the noisy results, $R(\mathbf{q}, \boldsymbol{\chi}')$, where $\boldsymbol{\chi}'$ is the accuracy obtained, to users. In order to illustrate the model clearly, we use an example to show how is the process of our proposed model.

Example 1. Consider the Adult dataset. Suppose users are interested in two queries $\mathbf{q} = [q_1, q_2]$ where q_1 is the ratio of individuals with race="black", sex= "female" and income="> 50K" to those with sex="female" and income="> 50K" and q_2 is the ratio of individuals with race="white", sex="female" and income="> 50K to those with sex="female" and income="> 50K. The corresponding accuracy requirement for the two queries is $\boldsymbol{\chi} = [\chi_1, \chi_2]$, where χ_1 is that the noisy result should be $\pm d_1$ with $d_1 = 0.05$ of the true value with $\alpha_1 = 95\%$ and χ_2 denotes that the noisy result should be $\pm d_2$ with $d_2 = 0.1$ of the true value with $\alpha_2 = 90\%$. Suppose $\epsilon_{total} = 1$. The database server will pass $Q_2(\mathbf{q}, \boldsymbol{\chi}, \boldsymbol{\vartheta}, 1)$ to the computing server after it receives the queries $Q_1(\mathbf{q}, \boldsymbol{\chi})$. Here $\boldsymbol{\vartheta}$ may include the mean and variance of an estimator, sample size and the sensitivity of a query. The computing server then optimizes the privacy budget between the two queries by solving the problem in Eqs. (5)-(8). Finally, the database server returns users the query results, $R(\mathbf{q}, \boldsymbol{\chi}')$.

4.2 Differential Privacy Framework

In this paper, we apply the differential privacy framework proposed by [14], called "sample and aggregate" [11]. It is an effective method to decrease the noise magnitude, where it randomly divides the data set into k blocks with size roughly equal to n/k. Then the estimation is applied in each block and finally the estimates are aggregated by using a differentially private function. Especially, the MLE estimator developed by Algorithm 1 [15] can asymptotically approach the true value θ_0.

Lemma 2 ([15]). *Algorithm 1 satisfies ϵ-differential privacy.*

Algorithm 1. An ϵ-Differential Privacy Algorithm

Input: $x = (x_1, \ldots, x_n) \in D^n$, $\epsilon > 0$
Output: $T_i^*(x)$, $i = 1, \ldots, m$
1: Let Γ be the range of $T_i(x)$ or diameter of the parameter space
2: Suppose $T_1(x), \ldots, T_m$ are the sufficient statistics for a set of parameters $\theta_1, \ldots, \theta_m$.
3: Calculate T_i, $i = 1, \ldots, m$ based on the input data x
4: **for** $i = 1$ to m **do**
5: Draw a random observation R_i from a laplace distribution with mean 0 and standard deviation $\sqrt{2}\Gamma/(n\epsilon)$
6: **end for**
7: Output $T_i^*(x) = T_i(x) + R_i$

4.3 The Normal-Laplace Distribution

Suppose an MLE estimator is used to estimate the ratios of multiple subsets to a given dataset in Algorithm 1. Then the output T^* is the summation of two independent random variables Z and Y, where Z is drawn from the normal distribution with $N(\mathrm{E}_\theta(T(X)), \mathrm{Var}(T(X)))$ and Y is drawn from the Laplace distribution with $\mathrm{Lap}(\lambda)$. The distribution of T^* is called normal-Laplace distribution [13]. In general, let $W = Z + Y$, where Z and Y are independent random variables with $Z \sim N(\mu, \sigma^2)$ and Y with following an asymmetric Laplace distribution with pdf

$$f_Y(y) = \begin{cases} \frac{\eta}{2} e^{\eta y}, & \text{for } y \leq 0 \\ \frac{\eta}{2} e^{-\eta y}, & \text{for } y > 0 \end{cases}$$

The distribution of W is called normal-Laplace distribution. We use

$$W \sim \mathrm{NL}(\mu, \sigma^2, \eta, \eta)$$

to denote such a distribution.

From the properties of characteristic function [1], we can derive the mean and variance of W as

$$\mathrm{E}\{W\} = \mu, \quad \text{and} \quad \mathrm{Var}(W) = \sigma^2 + 2/\eta^2.$$

A closed-form expression for the cumulative distribution function of the normal-Laplace distribution can be expressed as [13]

$$F(W) = \Phi(\frac{W-\mu}{\sigma}) - \phi(\frac{W-\mu}{\sigma}) \frac{R(\varphi_1) - R(\varphi_2)}{2} \tag{9}$$

with $\varphi_1 = \eta\sigma - (W-\mu)/\sigma$ and $\varphi_2 = \eta\sigma + (W-\mu)/\sigma$, where Φ and ϕ are the cdf and the pdf of a standard normal random variable, respectively. R is *Mill's ratio*.

5 Accuracy Guarantee

Suppose μ and σ are the mean and standard deviation of variable X. Let $T_1(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the estimator of μ. The noisy result is derived from Algorithm 1. We approximate the minimum privacy budget required for given level of precision requirement according to Eq. (9). In general, constructing an exact confidence interval requires complete information about the distribution of the variable. However, this information is not available in practice. Note that it is not easy to derive $W_{\alpha/2}$ such that $F(W_{\alpha/2}) = 1 - \alpha/2$ in Eq. (9). A feasible way is to construct confidence interval based on the large sample theory. Suppose $\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0)$ is approximately the standard normal distribution, then we get

$$\Pr(-y_{\alpha/2} \leq \sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0) \leq y_{\alpha/2}) \approx 1 - \alpha.$$

That is, we can get an approximate $100(1-\alpha)\%$ confidence interval such that

$$\hat{\theta} - y_{\alpha/2}\frac{1}{\sqrt{nI(\hat{\theta})}} \leq \theta_0 \leq \hat{\theta} + y_{\alpha/2}\frac{1}{\sqrt{nI(\hat{\theta})}}.$$

Infinite Case. Consider the population is infinite. Let X be a variable. Assume X has a normal, bell-shaped frequency distribution. We wish to estimate the mean of the population subject to the following constraint

$$\Pr(\hat{\theta} - y_{\alpha/2}s_{\hat{\theta}} < \theta < \hat{\theta} + y_{\alpha/2}s_{\hat{\theta}}) = 1 - \alpha,$$

where $s_{\hat{\theta}} = \frac{1}{\sqrt{nI(\hat{\theta})}}$ is the estimated standard deviation. We can determine the sample size by

$$n = (\frac{y_{\alpha/2}s}{d})^2,$$

where d is the desired absolute error and s is the standard deviation. Suppose $X \sim \mathcal{NL}(\mu, \frac{\sigma^2}{n}, \frac{k\epsilon}{\Gamma}, \frac{k\epsilon}{\Gamma})$. Then we get $Y \sim \mathcal{NL}(0, 1, \frac{k\epsilon\sigma}{\Gamma\sqrt{n}}, \frac{k\epsilon\sigma}{\Gamma\sqrt{n}})$. Here, we can characterize the accuracy guarantee as

$$y_{\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n} + \frac{2\Gamma^2}{k^2\epsilon^2}} \leq d.$$

Given accuracy requirement and dataset size, minimum the privacy budget ϵ required is expressed as

$$\epsilon = \phi_1(n, \sigma^2, \Gamma, d, y_{\alpha/2}) = \frac{\Gamma}{k} \cdot \sqrt{\frac{2}{(d/y_{\alpha/2})^2 - \sigma^2/n}}. \tag{10}$$

Finite Case. When the population is finite, the accuracy guarantee is different. Suppose the population is N. We need to derive explicit formula to express the relationships among those parameters discussed above for given level of precision

$1-\alpha$. Let $x_1, \ldots, x_N$ be the population and $X_1, \ldots, X_n$ be the variables selected for estimation. Let $p_i \in \{0, 1\}$ be the indicator variable. $p_i = 1$ if x_i belongs to a given sample. Then we can see that $\sum_{i=1}^{n} X_i = \sum_{i=1}^{N} p_i x_i$. Therefore,

$$\mathrm{E}\{1/n \sum_{i=1}^{n} X_i\} = 1/n \cdot n/N \sum_{i=1}^{N} x_i = m.$$

Let $X \sim \mathcal{NL}(\mu, \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}, \frac{k\epsilon}{\Gamma}, \frac{k\epsilon}{\Gamma})$. If $Y = (X - \mu)/(\frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}})$, then we have $Y \sim \mathcal{NL}(0, 1, \frac{k\epsilon}{\Gamma} \cdot \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}, \frac{k\epsilon}{\Gamma} \cdot \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}})$. The accuracy guarantee can be expressed as

$$y_{\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} + \frac{2\Gamma^2}{k^2\epsilon^2}} \leq d.$$

Here, we also derive similar function such that

$$\epsilon = \phi_2(n, \sigma^2, \Gamma, d, y_{\alpha/2}) = \frac{\Gamma}{k} \cdot \sqrt{\frac{2}{(d/y_{\alpha/2})^2 - \sigma^2 \cdot (N-n)/(n \cdot (N-1))}} \quad (11)$$

for the minimum privacy budget required.

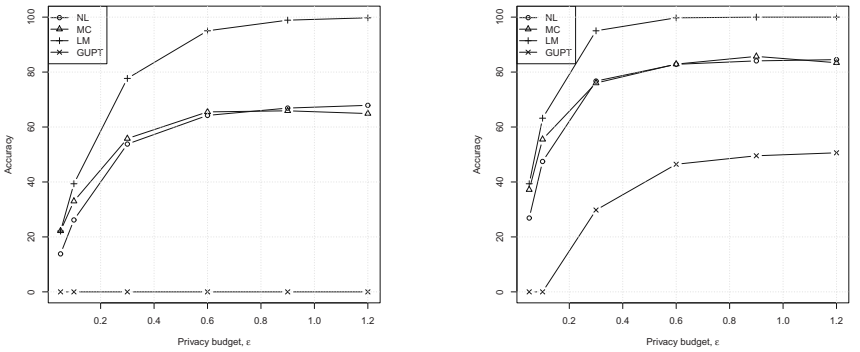
We have derived explicit formulas to describe given accuracy guarantee in terms of the privacy budget, number of blocks and dataset size. These formulas characterize how the parameters affect mutually. Thus, we can solve the optimization problem in Eqs. (5)-(8)) based on Lagrangian method. In the following section, we conduct simulations to demonstrate the effectiveness and feasibility of them.

6 Empirical Evaluations

In this section, we evaluate the performance of the proposed algorithm (denoted as NL) by comparing it with two state-of-the-art mechanisms, including LM and GUPT, which are proposed in [8] and [10] respectively. Particularly, we first evaluate the effectiveness and feasibility of the proposed algorithm based on a synthetic data and a real dataset. Then we further study the privacy budget allocation for the optimization problem in Eqs. (5)-(8).

6.1 Approximation Formulas Evaluation

We evaluate the relationships among accuracy, dataset size and privacy budget for the infinite case through synthetic data. We first generate the synthetic data with binomial distribution and dataset size $n_1 = 200$ and $n_2 = 100$ by Monte Carlo method. Two cases are considered, where $p = 0.4$. We wish to estimate the mean here. Then we test the cumulative accuracy of noisy results derived from Algorithm 1 falling into the given interval with $d = 0.05$, where the number



(a) $n = 100$, $\mu = 0.4$ and $d = 0.05$ (b) $n = 200$, $\mu = 0.4$ and $d = 0.05$

Fig. 2. Comparison of privacy budget, ϵ, versus accuracy for the infinite case

of generations is set to 1000. The results are shown in Figure 2, where NL and MC denote the theoretical results obtained by the proposed algorithm and the true results obtained by Monte Carlo method, LM and GUPT are the results obtained by Laplace mechanism and the GUPT algorithm, respectively.

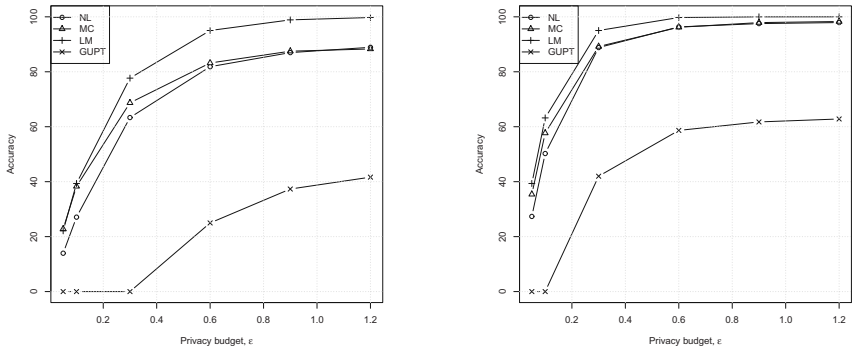
From the above figure, the accuracy increases with the increase of ϵ. The results obtained by NL are the most close to those of MC. Especially in the case of high accuracy. In contrast, the theoretical accuracy obtained by LM is much higher than the true accuracy while the theoretical accuracy obtained by GUPT is much lower. This demonstrates that NL is able to achieve higher accurate estimation than the two state-of-the-art techniques.

Next, we employ a real dataset, (i.e., *Adult* dataset from UCI dataset), to further prove the correctness of the approximation for finite case. Consider the estimation of the proportion of individuals with race="black" and sex="females" with income="> 50K" in terms of race="black" and income="> 50K". The total number of individuals with race="black" and income=">50K" is $N = 387$. We first randomly select $n_1 = 100$ and $n_2 = 200$ samples from the 387 individuals. Then we calculate the theoretical accuracy by using NL, LM and GUPT for different privacy budget.

The results are shown in Figure 3. The results are very similar to the infinite cases, as shown in Figure 2. This means that the proposed NL algorithm accurately characterizes the relationships among the three parameters.

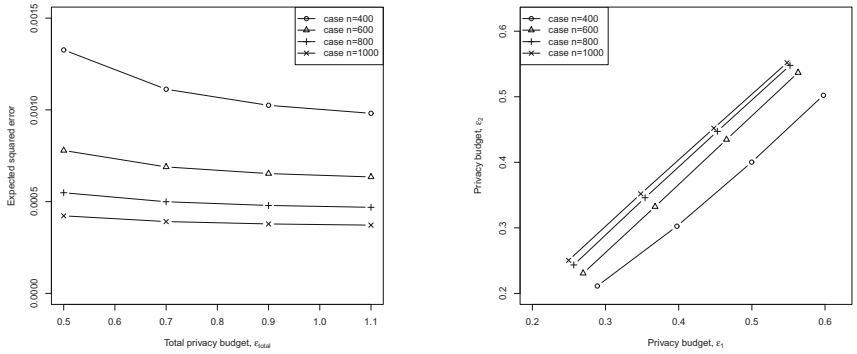
6.2 Privacy Budget Allocation for Multiple Queries

We further investigate the expected squared estimation errors of multiple queries from the optimization problem in Eqs. (5)-(8)). We consider two queries $q = [q_1, q_2]$ with $\alpha = (0.05, 0.1)$ and $d = (0.05, 0.1)$. Figure 4(a) shows the comparison of the cases with different dataset size. It can be observed that the expected squared errors obtained from different cases decrease with the increasing ϵ_{total}. Particularly, given a ϵ_{total}, the expected squared errors decrease with the increasing dataset size. Figure 4(b) shows the comparison of the corresponding privacy budget allocation under different datasets for the two queries. It can be seen that the



(a) $n = 100$ and $d = 0.05$ (b) $n = 200$ and $d = 0.05$

Fig. 3. Comparison of privacy budget, ϵ, versus accuracy for the finite case



(a) Expected squared. (b) Privacy budget allocation.

Fig. 4. Performance comparison of privacy budget allocation

privacy budget allocated to q_2 linearly increases with the privacy budget allocated to q_1. Moreover, given a total privacy budget ϵ_{total}, when the dataset size increases, the privacy allocated to it decreases while the privacy allocated to q_2 increases.

In summary, the above simulation results demonstrate that the proposed NL algorithm accurately describes the relationships among the parameters, namely the privacy budget, dataset size, accuracy and confidence interval, as well as how the privacy budget varies with the accuracy requirement.

7 Conclusion

In this paper, we have investigated the problem of how to allocate privacy budget among a batch of queries under the differential privacy framework. Particularly,

we formulated the level of accuracy in terms of privacy budget and dataset size, and we proposed a novel NL algorithm to determine the optimal privacy budget for the given accuracy guarantee. We further derived explicit formulas to accurately characterize the relationships among three parameters.

Acknowledgments. We thank Xiaokui Xiao for his valuable comments and suggestions.

References

1. Billingsley, P.: Probability and measure. John Wiley & Sons (2008)
2. Blake, C.L., Merz, C.J.: UCI repository of machine learning databases (1998); Robustness of maximum boxes
3. Chaudhuri, K., Monteleoni, C., Sarwate, A.D.: Differentially private empirical risk minimization. Journal of Machine Learning Research: JMLR **12**, 1069 (2011)
4. Clifton, C., Tassa, T.: On syntactic anonymity and differential privacy. In: First Workshop on Privacy-Preserving Data Publication and Analysis at ICDE, pp. 8–12 (2013)
5. Ding, B., Winslett, M., Han, J., Li, Z.: Differentially private data cubes: optimizing noise sources and consistency. In: SIGMOD Conference, pp. 217–228 (2011)
6. Dwork, C.: Differential Privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006)
7. Dwork, C.: A firm foundation for private data analysis. Communications of the ACM **54**(1), 86–95 (2011)
8. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006)
9. Hardt, M., Rothblum, G.N., Servedio, R.A.: Private data release via learning thresholds. In: Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 168–187 (2012)
10. Mohan, P., Thakurta, A., Shi, E., Song, D., Culler, D.: Gupt: privacy preserving data analysis made easy. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 349–360 (2012)
11. Nissim, K., Raskhodnikova, S., Smith, A.: Smooth sensitivity and sampling in private data analysis. In: Proceedings of the 39th Annual ACM Symposium on Theory of Computing, pp. 75–84 (2007)
12. Rastogi, V., Nath, S.: Differentially private aggregation of distributed time-series with transformation and encryption. In: Proceedings of the International Conference on Management of Data, pp. 735–746. ACM (2010)
13. Reed, W.J.: The normal-laplace distribution and its relatives. In: Advances in Distribution Theory, Order Statistics, and Inference, pp. 61–74. Springer (2006)
14. Smith, A.: Efficient, differentially private point estimators. arXiv preprint arXiv:0809.4794 (2008)
15. Vu, D., Slavkovic, A.: Differential privacy for clinical trial data: Preliminary evaluations. In: IEEE International Conference on Data Mining Workshops, pp. 138–143 (2009)
16. Xiao, X., Wang, G., Gehrke, J.: Differential privacy via wavelet transforms. IEEE Transactions on Knowledge and Data Engineering **23**(8), 1200–1214 (2011)
17. Zhang, J., Zhang, Z., Xiao, X., Yang, Y., Winslett, M.: Functional mechanism: regression analysis under differential privacy. Proceedings of the VLDB Endowment **5**(11), 1364–1375 (2012)