Differentially Private Algorithms in DPSQL

Explanation of terms

- private data
  - The goal of privacy protection is to prevent re-identification, and the general protection entities are users, devices, companies, etc.
- privacy mechanism
  - A program capable of processing private data and outputting privacy-preserving results
  - This mechanism is generally relatively standard, such as laplace, gaussian, etc
  - This mechanism can handle aggregated results such as count, sum, and some can handle a single data object, such as gradient
  - Each execution of the privacy mechanism will incur privacy budget cost
- epsilon
  - Epsilon is a measure of privacy leakage
  - Smaller epsilon means less privacy leakage and more noise is introduced
- delta
  - Indicates the probability that the result cannot be protected by epsilon, which is generally set to be very small (e.g., $\frac{1}{n^2}$, where n is the total number of records)
- privacy budget
  - When multiple queries are allowed, the noisy results of multiple queries can be combined for re-identification. Thus, in addition to measuring the privacy loss each time, it is also necessary to measure the total privacy loss after multiple queries
  - Privacy budget refers to the total amount of privacy loss that may occur for all queries before privacy data can no longer be accessed, which is an expected upper limit.
- composition
  - Describe how the privacy budget costs of multiple queries are composed

together to get the total cost

- sensitivity
  - The maximum change of the result of a particular query f when deleting or adding any piece of data in the dataset.
  - Sensitivity is a key parameter in determining the amount of added noise.

DP Mechanisms

mechanisms	Laplace	Gaussian
definition	$Lap(b)$	$\mathcal{N}(0,\sigma^2)$
parameter	$b=\frac{\Delta f}{\varepsilon}$	$\sigma=\frac{\Delta f\sqrt{2log(1.25/\delta)}}{\varepsilon}$
DP	ε-DP	(ε,δ)-DP
variance	$\frac{2\Delta f^2}{\varepsilon^2}=O\left(\frac{1}{\varepsilon^2}\right)$	$\frac{2\Delta f^2}{\varepsilon^2}\log\frac{1.25}{\delta}=O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$
(α,β)	$\frac{\Delta f}{\varepsilon}\log\frac{1}{\beta}$	$\frac{2\Delta f}{\varepsilon}\sqrt{log\frac{1.25}{\delta}}erf^{-1}(1-\beta)\le\frac{2\Delta f}{\varepsilon}\sqrt{\log\frac{1}{\beta}\log\frac{1.25}{\delta}}$

1. Laplace Mechanism Introduction

- Laplace distribution
  - Probability density function $Lap(x\mid b)=\frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)$
- Laplace mechanism
  - Noise distribution $Lap(b)$, scale parameter $b=\Delta f/\varepsilon$, satisfies ε-DP
- Usability
  - Variance, $2b^2=\frac{2\Delta f^2}{\varepsilon^2}=O\left(\frac{1}{\varepsilon^2}\right)$
  - (α,β)-accuracy, $\alpha=\frac{\Delta f}{\varepsilon}\log\frac{1}{\beta}$

Usability Analysis:

Laplace mechanism adding noise N_L *obeying distribution* $Lap\left(b = \frac{\Delta f}{\varepsilon}\right)$

Laplace distribution satisfies $P[|N_L| \geq t * b] = \exp(-t)$. *If* $\exp(-t)$ *is replaced with* β, *then there is* $P\left[|N_L| \geq \frac{\Delta f}{\varepsilon}\log\frac{1}{\beta}\right] = \beta$.

2. **Gauss Mechanism Introduction**

- Gauss distribution
  - Probability density function $\mathcal{N}(x|\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)$
- Gauss mechanism
  - Noise distribution $\mathcal{N}(0, \sigma^2)$, Standard Deviation parameter $\sigma = \frac{\Delta f\sqrt{2log(1.25/\delta)}}{\varepsilon}$, satisfies (ε, δ)-DP
- Usability
  - Variance, $\sigma^2 = \frac{2\Delta f^2\log(1.25/\delta)}{\varepsilon^2} = O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$
  - (α, β)-accuracy, $\alpha = \frac{2\Delta f}{\varepsilon}\sqrt{log\frac{1.25}{\delta}}erf^{-1}(1-\beta) \leq \frac{2\Delta f}{\varepsilon}\sqrt{\log\frac{1}{\beta}\log\frac{1.25}{\delta}}$

Usability Analysis:

The Gauss mechanism adds noise N_G*obeying distribution* $\mathcal{N}(0, \sigma^2)$.

Gauss distribution of the inverse cumulative distribution function is $F(x) = \frac{1}{2} + \frac{1}{2}erf\left(\frac{x}{\sigma\sqrt{2}}\right)$,

where erf *is the Gaussian distribution error function error function:* $erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$.

By setting $F(\alpha) = 1 - \frac{\beta}{2}, F(\alpha) - F(-\alpha) = 1 - \beta$, *we have*

$P[|N_G| \leq \alpha] = 1 - \beta$ *Of which* $\alpha = \frac{2\Delta f}{\varepsilon}\sqrt{log\frac{1.25}{\delta}}erf^{-1}(1-\beta) \leq \frac{2\Delta f}{\varepsilon}\sqrt{\log\frac{1}{\beta}\log\frac{1.25}{\delta}}$

Composition theorem (k adaptive queries, where there is a correlation between queries)

- If we consider privacy loss as a random variable: for one output o, the privacy loss of the mechanism M on two neighboring datasets X, X' is $L^{o}_{M(X),M(X')} = \ln \frac{Pr[M(X)=o]}{Pr[M(X')=o]}$
- (ε, δ)-DP indicates that for any two neighboring datasets, the above random variable is greater than ε at most probability δ.

1. Basic composition

- A set of k mechanisms satisfying $(\varepsilon_i, \delta_i)$-DP, for the same dataset for k queries, k-fold adaptive composition satisfies $(\sum\varepsilon_i, \sum\delta_i)$-DP
- The basic composition is to accumulate the upper bound of the random variable privacy loss (based on KL divergence).

2. Advanced composition 1

- A set of k mechanisms satisfying (ε, δ)-DP, for k queries on the same dataset, k-fold adaptive composition satisfies $(\varepsilon', k\delta + \delta')$-DP, where $\varepsilon' = \varepsilon\sqrt{2k\ln(1/\delta')} + k\varepsilon(e^{\varepsilon} - 1)$
- When $\varepsilon = 10^{-5}$, the latter item $k\varepsilon(e^{\varepsilon} - 1)$ is approximately 0, then $\varepsilon' = \varepsilon\sqrt{2k\ln 1/\delta'} = O(\sqrt{k})$. Compared to the basic composition of $O(k)$, it saves more privacy budget. Thus, Advanced composition 1 is significantly better than the basic composition when k is large.
- This method accumulates the upper bound of privacy loss expectations (based on max divergence)

3. Advanced composition 2

- A set of k mechanisms satisfying (ε, δ)- DP, for k queries on the same dataset, k-fold adaptive composition satisfies $\left((k-2i)\varepsilon, 1-(1-\delta)^k(1-\delta_i)\right)$- DP, where for any $i = \{0,1,\dots,\lfloor k/2 \rfloor\}$, $\delta_i = \frac{\sum_{l=0}^{i-1}\binom{k}{l}\left(e^{(k-l)\varepsilon} - e^{(k-2i+l)\varepsilon}\right)}{(1+e^{\varepsilon})^k}$
- A tighter bound, dependent on the operational interpretation of the privacy as hypothesis testing, is given to prove the optimality of this result

Self-Adaptative Selection:

- A set of k mechanisms satisfying (ε, δ)-DP, for k queries on the same dataset, k-

fold adaptive composition satisfies $\left(\varepsilon', 1-(1-\delta)^k(1-\delta_i)\right)$- DP, where

$$\varepsilon' = min\left\{k\varepsilon, \frac{\varepsilon k(e^\varepsilon-1)}{e^\varepsilon+1} + \varepsilon\sqrt{2k\ln(1/\delta')}, \frac{\varepsilon k(e^\varepsilon-1)}{e^\varepsilon+1} + \varepsilon\sqrt{2k\ln(e+\frac{\sqrt{k\varepsilon^2}}{\delta'})}\right\}$$

- When k mechanisms respectively satisfy $(\varepsilon_i, \delta_i)$- DP , the set of mechanisms satisfy $\left(\varepsilon', 1-(1-\delta')\prod_{i=1}^k(1-\delta_i)\right)$-DP, where $\varepsilon' =$

$$min\left\{\sum_{i=1}^k \varepsilon_i, \sqrt{\sum_{i=1}^k 2\varepsilon_i^2\ln(1/\delta')} + \sum_{i=1}^k \frac{\varepsilon_i(e^{\varepsilon_i}-1)}{(e^{\varepsilon_i}+1)}, \sqrt{\sum_{i=1}^k 2\varepsilon_i^2\ln\left(e+\frac{\sqrt{\sum_{i=1}^k \varepsilon^2}}{\delta'}\right)} + \sum_{i=1}^k \frac{\varepsilon_i(e^{\varepsilon_i}-1)}{(e^{\varepsilon_i}+1)}\right\}$$

- $\varepsilon' = min\left\{k\varepsilon, \frac{\varepsilon k(e^\varepsilon-1)}{e^\varepsilon+1} + \varepsilon\sqrt{2k\ln(1/\delta')}, \frac{\varepsilon k(e^\varepsilon-1)}{e^\varepsilon+1} + \varepsilon\sqrt{2k\ln(e+\frac{\sqrt{k\varepsilon^2}}{\delta'})}\right\}$

Reference:

The Composition Theorem for Differential Privacy

Sensitivity

1. Global sensitivity

- Function $f: D^n \to \mathbb{R}^d$, the input is a dataset, the output is a d dimensional vector, for any two neighboring datasets $x, y \in D^n$, function f's global sensitivity is

$$GS_f = \max_{x,y:d(x,y)=1} \|f(x)-f(y)\|_1$$

- The global sensitivity of function f is determined by itself

2. Local sensitivity

- Function $f: D^n \to \mathbb{R}^d$, input a dataset $x \in D^n$, output a d dimensional vector, for a given dataset x and its neighboring datasets y, the local sensitivity of function f on x is

$$LS_f(x) = \max_{y:d(x,y)=1} \|f(x)-f(y)\|_1$$

- Local sensitivity is determined by the function $f$and given dataset x. Local sensitivity is usually much smaller than global sensitivity due to the data distribution characteristics of the dataset.

Relation to global sensitivity:

$$GS_f = \max_x(LS_f(x))$$

- If the local sensitivity is directly applied to calculate the amount of noise, the sensitive information of the dataset will be leaked, so the smooth upper bound of the local sensitivity is used to determine the noise size together with the local sensitivity.

3. **Smooth sensitivity**

reference: smooth sensitivity

- Smooth upper bound of LS
  - Given a dataset $x \in D^n$ and any neighboring dataset $y: d(x, y) = 1$, function f's local sensitivity is $LS_f(x)$. For $\beta > 0$, if function $S: D^n \to \mathbb{R}^d$ satisfies $S_f(x) \geq LS_f(x), S_f(x) \leq e^{\beta} S_f(y)$, S is called the f's local sensitivity of the β-smooth upper bound
- Smooth sensitivity
  - Given the neighboring datasets x, y,

$$S_{f,\beta}(x) = \max_{y \in D^n} \left(e^{-\beta d(x,y)} LS_f(y) \right)$$

- Calculation of Smooth Sensitivity
  - Defined in distance k, the sensitivity is

$$A_f^{(k)}(x) = \max_{y \in D^n : d(x,y) \leq k} LS_f(y)$$

  - $S_{f,\beta}(x) = \max_{k=0,1,\ldots,n} e^{-\beta k} \left(\max_{y: d(x,y)=k} LS_f(y) \right) = \max_{k=0,1,\ldots,n} e^{-\beta k} A_f^{(k)}(x)$

4. **Elastic sensitivity**

reference: elastic sensitivity

- Elastic Sensitivity is an Upper Bound on Local Sensitivity
- Definition
  - Elastic sensitivity: Query q's elasticity sensitivity at distance k from the real database x is $\hat{S}^{(k)}(q, x)$
  - Elastic stability: relational transformation r's elastic stability is $\hat{S}_R^{(k)}(r, x)$
  - Maximum frequency: for database x's relationship r, the maximum frequency of values of attribute a is $mf(a, r, x)$
  - Maximum frequency at distance k: at distance k from real database x, in

relationship r, the maximum frequency of values of attribute a is $mf_k(a,r,x)$, where

$$mf_k(a,r,y) \geq \max_{y:d(x,y)\leq k} mf(a,r,y)$$

- Calculate elastic sensitivity

Maximum frequency	1. $mf_k(a,t,x) = mf(a,t,x) + k$ 2. $mf_k\left(a_1, r_1 \underset{a_2=a_3}{\bowtie} r_2, x\right) = mf_k(a_1,r_1,x)mf_k(a_3,r_2,x)\ \ a_1 \in r_1$ $mf_k\left(a_1, r_1 \underset{a_2=a_3}{\bowtie} r_2, x\right) = mf_k(a_1,r_2,x)mf_k(a_2,r_1,x)\ \ a_1 \in r_2$ 3. $mf_k(a, \Pi_{a_1,\dots,a_n} r, x) = mf_k(a,r,x)$ 4. $mf_k(a, \sigma_\varphi r, x) = mf_k(a,r,x)$ 5. $mf_k(a, Count(r), x) = \perp$
Elastic stability	1. $\hat{S}_R^{(k)}(t,x) = 1$ 2. $\hat{S}_R^{(k)}\left(r_1 \underset{a=b}{\bowtie} r_2, x\right) = \max\left(mf_k(a,r_1,x)\hat{S}_R^{(k)}(r_2,x), mf_k(b,r_2,x)\hat{S}_R^{(k)}(r_1,x)\right),$ $\lvert\mathcal{A}(r_1) \cap \mathcal{A}(r_2)\rvert = 0$ $\hat{S}_R^{(k)}\left(r_1 \underset{a=b}{\bowtie} r_2, x\right) = mf_k(a,r_1,x)\hat{S}_R^{(k)}(r_2,x) + mf_k(b,r_2,x)\hat{S}_R^{(k)}(r_1,x) +$ $\hat{S}_R^{(k)}(r_1,x)\hat{S}_R^{(k)}(r_2,x)$, $\lvert\mathcal{A}(r_1) \cap \mathcal{A}(r_2)\rvert > 0$ 3. $\hat{S}_R^{(k)}(\Pi_{a_1,\dots,a_n} r, x) = \hat{S}_R^{(k)}(r,x)$ 4. $\hat{S}_R^{(k)}(\sigma_\varphi r, x) = \hat{S}_R^{(k)}(r,x)$ 5. $\hat{S}_R^{(k)}(Count(r)) = 1$
Elastic sensitivity	1. $\hat{S}^{(k)}(Count(r), x) = \hat{S}_R^{(k)}(r,x)$ 2. $\hat{S}^{(k)}\left(Count(r), x\right) = 2\hat{S}_R^{(k)}(r,x)$

- Smooth the elastic sensitivity
  - $S = \max_{k=0,1,\dots,n} e^{-\beta k}\hat{S}^{(K)}, \beta = \frac{\varepsilon}{2\ln(2/\delta)}, \quad \delta = 10^{-8}$
  - The scale parameter of the Laplace mechanism $b = \frac{2S}{\varepsilon}$