Data Privacy Homework 2

1. (15') Laplace mechanism

(a) (5') Given the function $f(x) = \frac{1}{6}\sum_{i=1}^{6} x_i$, where $x_i \in \{1, 2, \ldots, 10\}$ for $i \in \{1, 2, \ldots, 6\}$, compute the global sensitivity and local sensitivity when $x = \{3, 5, 4, 5, 6, 7\}$.

(b) (10') Given a database x where each element is in $\{1, 2, 3, 4, 5, 6\}$, design ϵ-differentially private Laplace mechanisms corresponding to the following queries, where $\epsilon = 0.1$:

1. $q_1(x) = \sum_{i=1}^{6} x_i$
2. $q_2(x) = \max_{i \in \{1,2,\ldots,6\}} x_i$

2.(15') Exponential mechanism

ID	sex	Chinese	Mathematics	English	Physics	Chemistry	Biology
1	Male	96	58	80	53	56	100
2	Male	60	63	77	50	59	75
3	Female	83	86	98	69	80	100
...							
4000	Female	86	83	98	87	82	92

Table 1: Scores of students in School A

Table 1 records the scores of students in School A in the final exam. We need to help the teacher query the database while protecting the privacy of students' scores. The domain of this database is $\{Male, Female\} \times \{0, 1, 2, \ldots, 100\}^6$. Answer the following questions.

(a) (5') What is the sensitivity of the following queries:

1. $q_1(x) = \frac{1}{4000}\sum_{ID=1}^{4000} Physics_{ID}$
2. $q_2(x) = \max_{ID \in \{1,2,\ldots,4000\}} Biology_{ID}$

(b) (10') Design ϵ-differential privacy mechanisms corresponding to the two queries in (a), where $\epsilon = 0.1$. (Using Laplace mechanism for q_1 and Exponential mechanism for q_2.)

3.(20') Composition

Theorem 3.16. Let $\mathcal{M}_i : \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}_i$ be an $(\varepsilon_i, \delta_i)$-differentially private algorithm for $i \in [k]$. Then if $\mathcal{M}_{[k]} : \mathbb{N}^{|\mathcal{X}|} \to \prod_{i=1}^{k} \mathcal{R}_i$ is defined to be $\mathcal{M}_{[k]}(x) = (\mathcal{M}_1(x), \ldots, \mathcal{M}_k(x))$, then $\mathcal{M}_{[k]}$ is $(\sum_{i=1}^{k} \varepsilon_i, \sum_{i=1}^{k} \delta_i)$-differentially private.

Theorem 3.20 (Advanced Composition)**.** For all $\varepsilon, \delta, \delta' \geq 0$, the class of (ε, δ)-differentially private mechanisms satisfies $(\varepsilon', k\delta + \delta')$-differential privacy under k-fold adaptive composition for:

$$\varepsilon' = \sqrt{2k \ln(1/\delta')}\varepsilon + k\varepsilon(e^{\varepsilon} - 1).$$

(a) (10') Given a database $x = \{x_1, x_2, \ldots, x_{2000}\}$ where $x_i \in \{0, 1, 2, \ldots, 100\}$ for each i and privacy parameters $(\epsilon, \delta) = (1.25, 10^{-5})$, apply the Gaussian mechanism to protect 100 calls to the query $q_1(x) = \frac{1}{2000}\sum_{i=1}^{2000} x_i$. Determine the noise variances σ^2 of the Gaussian mechanism to ensure (ϵ, δ)-DP based on the composition and advanced composition theorems, respectively.

(b) (10') Determine the noise variances σ^2 of the Gaussian mechanism to protect 100 calls to the query $q_2(x) = \max_{i\in\{1,2,\ldots,2000\}} x_i$ to ensure $(1.25, 10^{-5})$-DP based on the composition and advanced composition theorems, respectively, where x is the database in (a).

4. (25') Randomized Response for Local DP

Consider a population of n users, where the true proportion of males is denoted as π. Our objective is to gather statistics on the proportion of males, prompting a sensitive question: "Are you male?" Each user responds with either a yes or no, but due to privacy concerns, they refrain from directly disclosing their true gender. Instead, they employ a biased coin with a probability of landing heads denoted as p, and tails as $1 - p$. When the coin is tossed, a truthful response is given if heads appear, while the opposite response is provided if tails come up.

(a) (10') Demonstrate that the aforementioned randomized response adheres to local differential privacy and determine the corresponding privacy parameter, ϵ.

(b) (15') Employing the perturbation method outlined above to aggregate responses from the n users yields a statistical estimate for the number of males. Assuming the count of "yes" responses is n_1, construct an unbiased estimate for π based on n, n_1, p. Calculate the variance associated with this estimate.

5. (10') Accuracy Guarantee of DP

Consider the application of an (ϵ, δ)-differentially private Gaussian mechanism denoted by $\mathcal{M}$ to protect the mean estimator $\bar{x} = \frac{1}{n}\Sigma_{i=1}^n x_i$ of a d-dimensional input database x, where $x_i \in \{0, 1, \ldots, 100\}^d$ for each i. Let $\mathcal{M}(x)$ represent the output of this Gaussian mechanism. Utilize both the tail bound and the union bound to derive the L_∞-norm error bound of $\mathcal{M}$, denoted by $\|\mathcal{M}(x) - \bar{x}\|_\infty$, ensuring a probability of at least $1 - \beta$. Specifically, solve for the bound $\mathcal{B}$ such that

$$\Pr[\|\mathcal{M}(x) - \bar{x}\|_\infty \leq \mathcal{B}] \geq 1 - \beta.$$

Hint: Refer to https://zhuanlan.zhihu.com/p/425562737 for descriptions of statistical inequalities.

6. (15') Personalized Differential Privacy

Consider an n-element dataset D where the i-th element is owned by a user $i \in [n]$, where $[n] = \{1, 2, \ldots, n\}$ and the privacy requirement of user i is ϵ_i-DP. A randomized mechanism $\mathcal{M}$ satisfies $\{\epsilon_i\}_{i\in[n]}$-personalized differential privacy (or $\{\epsilon_i\}_{i\in[n]}$-PDP) if, for every pair of neighboring datasets D, D' differing at the j-th element for an arbitrary $j \in [n]$, and for all sets S of possible outputs,

$$P[\mathcal{M}(D) \in S] \leq e^{\epsilon_j} Pr[\mathcal{M}(D') \in S].$$

(a) (5') Prove the composition theorem of PDP: If a mechanism is $\{\epsilon_i^{(1)}\}_{i\in[n]}$-PDP and another is $\{\epsilon_i^{(2)}\}_{i\in[n]}$-PDP, then publishing the result of both is $\{\epsilon_i^{(1)} + \epsilon_i^{(2)}\}_{i\in[n]}$-PDP.

(b) (10') Given a dataset D and a privacy requirement set $\{\epsilon_i\}_{i\in[n]}$, the *Sample mechanism* works as follows: 1) We pick an arbitrary threshold value $t > 0$; 2) We sample a subset $D_S \subset D$ where the probability that the i-th element of D is contained in D_S equals $\frac{e^{\epsilon_i}-1}{e^t-1}$ if $\epsilon_i < t$ and 1 otherwise; 3) We output $\mathcal{M}(D_S)$, where $\mathcal{M}$ is a t-differentially private mechanism. Prove that the Sample mechanism with any $t > 0$ is $\{\epsilon_i\}_{i\in[n]}$-PDP. **Hint:** Use the Bayes formula.