Solution to Hw2

jdliu

May 2020

1 Q1

1) $f(x) = q(x), \Delta f(x) = 5 - 1 = 4, \Delta f/\epsilon = 40$
Laplace Mechanism: $M_L(x, f(\cdot), \epsilon) = f(x) + Y$, where Y are i.i.d. random variables drawn from $Lap(40)$

2)Laplace Mechanism: $f(x) = q(x), \Delta f(x) = 1, \Delta f/\epsilon = 10$
$M_L(x, f(\cdot), \epsilon) = f(x) + Y$, where Y are i.i.d. random variables drawn from $Lap(10)$

Exponential Mechanism: Define the score function $u(x, r) = I(r = \max(x))$, where $I(\cdot)$ is the indicator function. Then $\Delta u = 1$, by the definition of Exponential Mechanism, we can output $r \in \{1, \ldots, 5\}$ with probability proportional to $\exp(\frac{\epsilon u(x,r)}{2\Delta u})$.

output	1	2	3	4	5
probability	0.1980	0.1980	0.1980	0.1980	0.2080

2 Q2

1) For $k = 1$, $\Delta_T = \Delta_m \leq 2L\eta_1$. Thus the proposition holds. By induction, the proposition holds for $T = km$, $k = 1, 2, \ldots$.

2) For $k = 1$, $\Delta_T = \Delta_m \leq 2L\eta(1 - n\gamma)^{m-1}$. Thus the proposition holds. By induction, the proposition holds for $T = km$, $k = 1, 2, \ldots$.

3 Q3

1) From Algorithm 1, we can find that the sensitivity of each update is bounded by C. By Theorem A.1 in the textbook, each update is (ϵ, δ)-DP if $C^2\sigma^2 \geq 2\ln(1.25/\delta)C^2/\epsilon^2$, or equivalently, $\sigma^2 \geq 2\ln(1.25/\delta)/\epsilon^2$.

2) By the composition theorem, each update should be $(\epsilon/T, \delta/T)$-DP so that the algorithm is (ϵ, δ)-DP. By results in 1), we have $\sigma^2 \geq 2.7 \times 10^9$.

3) By the advanced composition theorem, if each update (ϵ, δ)-DP, the algorithm is $(\epsilon^{'}, (T+1)\delta)$-DP, where

$$\epsilon^{'} = \sqrt{2T\ln(1/\delta)}\epsilon + T\epsilon(e^{\epsilon} - 1).$$

By setting $(\epsilon^{'}, (T+1)\delta) = (1.25, 10^{-5})$, we can get $(\epsilon, \delta) = (1.89 \times 10^{-3}, 1.00 \times 10^{-9})$ by solving the equation above. Thus it holds that $\sigma^2 \geq 1.2 \times 10^7$.

4 Q4

Bernoulli distribution: $Pr[x = 1] = p, Pr[x = 0] = 1 - p$.
for $\tilde{x} \in \tilde{X}$,

$$\begin{aligned} Pr[\tilde{x} = 1] &= Pr[x = 1] \times \frac{e^{\epsilon}}{1 + e^{\epsilon}} + Pr[x = 0] \times \frac{1}{1 + e^{\epsilon}} \\ &= \frac{1 + p(e^{\epsilon} - 1)}{1 + e^{\epsilon}} \end{aligned}$$

$$\begin{aligned} Pr[\tilde{x} = 0] &= Pr[x = 0] \times \frac{e^{\epsilon}}{1 + e^{\epsilon}} + Pr[x = 1] \times \frac{1}{1 + e^{\epsilon}} \\ &= \frac{e^{\epsilon} + p(1 - e^{\epsilon})}{1 + e^{\epsilon}} \end{aligned}$$

if $p = 0.5$, $Pr[\tilde{x} = 1] = 0.5, Pr[\tilde{x} = 0] = 0.5$, $E(\sum_{x \in \tilde{X}}) = 0.5n$, $E(\sum_{x \in X}) = 0.5n$
if $p = 0.1$, $Pr[\tilde{x} = 1] = 0.46, Pr[\tilde{x} = 0] = 0.54$, $E(\sum_{x \in \tilde{X}}) = 0.46n$, $E(\sum_{x \in X}) = 0.1n$.
if $p = 0.9$, $Pr[\tilde{x} = 1] = 0.54, Pr[\tilde{x} = 0] = 0.46$, $E(\sum_{x \in \tilde{X}}) = 0.54n$, $E(\sum_{x \in X}) = 0.9n$.

The answer of "what do you find" is not fixed. A reasonable answer will be accepted. for example:
For count queries, LDP performs best when the data is balanced, i.e., $p = 0.5$.

5 Q5

1)

Let $\mathcal{X} = \{0,1\}$ and consider the two datasets $x = 0^n$ and $x' = 10^{n-1}$. Now define $S = \{z \in \{0,1\}^m \mid z \neq 0^m\}$. Then for every ϵ and every $\delta < m/n$

$$e^{\varepsilon} \Pr[A(x) \in S] + \delta = \delta < \frac{m}{n} = \Pr[A(x') \in S],$$

contradicting (ε, δ)-dp of M.

2)

We'll use $T \subseteq \{1, \ldots, n\}$ to denote the identities of the m-subsampled rows (i.e. their row number, not their actual contents). Note that T is a random variable, and that the randomness of A' includes both the randomness of the sample T and the random coins of A. Let $x \sim x'$ be adjacent databases and assume that x and x' differ only on some row t. Let x_T (or x'_T) be a subsample from x (or x') containing the rows in T. Let S be an arbitrary subset of the range of A'. For convenience, define $p = m/n$

To show $(p(e^{\varepsilon} - 1), p\delta)$-dp, we have to bound the ratio

$$\frac{\Pr[A'(x) \in S] - p\delta}{\Pr[A'(x') \in S]} = \frac{p \Pr[A(x_T) \in S \mid i \in T] + (1-p)\Pr[A(x_T) \in S \mid i \notin T] - p\delta}{p \Pr[A(x'_T) \in S \mid i \in T] + (1-p)\Pr[A(x'_T) \in S \mid i \notin T]}$$

by $e^{p(e^{\varepsilon}-1)}$. For convenience, define the quantities

$$\begin{aligned}
C &= \Pr[A(x_T) \in S \mid i \in T] \\
C' &= \Pr[A(x'_T) \in S \mid i \in T] \\
D &= \Pr[A(x_T) \in S \mid i \notin T] = \Pr[A(x'_T) \in S \mid i \notin T]
\end{aligned}$$

We can rewrite the ratio as

$$\frac{\Pr[A'(x) \in S]}{\Pr[A'(x') \in S]} = \frac{pC + (1-p)D - p\delta}{pC' + (1-p)D}$$

Now we use the fact that, by (ε, δ)-dp, $A \leq e^{\varepsilon} \min\{C', D\} + \delta$. The rest is a calculation:

$$\begin{aligned}
& pC + (1-p)D - p\delta \\
&\leq p(e^{\varepsilon} \min\{C', D\} + \delta) + (1-p)D - p\delta \\
&\leq p(\min\{C', D\} + (e^{\varepsilon} - 1)\min\{C', D\}) + \delta) + (1-p)D - p\delta \\
&\leq p(\min\{C', D\} + (e^{\varepsilon} - 1)(pC' + (1-p)D) + \delta) + (1-p)D - p\delta \\
& \qquad\qquad\qquad\qquad \text{(Because } \min\{x,y\} \leq \alpha x + (1-\alpha)y \text{ for every } 0 \leq \alpha \leq 1) \\
&\leq p(C' + (e^{\varepsilon} - 1)(pC' + (1-p)D) + \delta) + (1-p)D - p\delta \qquad \text{(Because } \min\{x,y\} \leq x) \\
&\leq p(C' + (e^{\varepsilon} - 1)(pC' + (1-p)D)) + (1-p)D \\
&\leq (pC' + (1-p)D) + (p(e^{\varepsilon} - 1))(pC' + (1-p)D) \\
&\leq (1 + p(e^{\varepsilon} - 1))(pC' + (1-p)D) \\
&\leq e^{p(e^{\varepsilon}-1)}(pC' + (1-p)D)
\end{aligned}$$

So we've succeeded in bounding the necessary ratio of probabilities. Note, if you are willing to settle for $(O(\varepsilon m/n), O(\delta m/n))$-dp the calculation is much simpler. All this algebra is mostly just to get the tight bound.