CS591S1 Homework 2: More on Differential Privacy

Jiawen Liu
Collaborators: none.

1 Problem 1

1. (Sensitivity)

   Given data set $\mathbf{x}$, for arbitrary $y \in \{1, \cdots, R\}$ and adjacent dataset $\mathbf{x}'$ by insertion or deletion of one data point x_k, we have following cases for insertion (deletion will be symmetric):

   - $y = x_k$:

   $$q(y;\mathbf{x}') = -|\sum_{i=0}^{n} sign(y-x_i)+sign(y-x_k)| = -|\sum_{i=0}^{n} sign(y-x_i)+0| = -|\sum_{i=0}^{n} sign(y-x_i)| = q(y;\mathbf{x})$$

   - $y < x_k$

   $$\begin{array}{rcl} q(y;\mathbf{x}') & = & -|\sum_{i=0}^{n} sign(y-x_i) + sign(y-x_k)| \\ & = & -|\sum_{i=0}^{n} sign(y-x_i) - 1| \ (\star) \end{array}$$

   By triangle inequality, we have:

   $$\begin{array}{rcl} (\star) & \leq & -|\sum_{i=0}^{n} sign(y-x_i)| + 1 = q(y;\mathbf{x}) + 1 \\ (\star) & \geq & -|\sum_{i=0}^{n} sign(y-x_i)| - 1 = q(y;\mathbf{x}) - 1 \end{array}$$

   Then we can get:

   $$-1 \leq q(y;\mathbf{x}') - q(y;\mathbf{x}) \leq 1$$

   - $y > x_k$

   $$\begin{array}{rcl} q(y;\mathbf{x}') & = & -|\sum_{i=0}^{n} sign(y-x_i) + sign(y-x_k)| \\ & = & -|\sum_{i=0}^{n} sign(y-x_i) + 1| \ (\star) \end{array}$$

   By triangle inequality, we have:

   $$\begin{array}{rcl} (\star) & \leq & -|\sum_{i=0}^{n} sign(y-x_i)| + 1 = q(y;\mathbf{x}) + 1 \\ (\star) & \geq & -|\sum_{i=0}^{n} sign(y-x_i)| - 1 = q(y;\mathbf{x}) - 1 \end{array}$$

   Then we can get:

   $$-1 \leq q(y;\mathbf{x}') - q(y;\mathbf{x}) \leq 1$$

   The Deletion is symmetric where we can get: $-1 \leq q(y;\mathbf{x}) - q(y;\mathbf{x}') \leq 1$ in the same way. Then, we can conclude from all cases, the $|q(y;\mathbf{x}) - q(y;\mathbf{x}')| \leq 1$, i.e., the sensitivity be at most 1.

2. *Proof.* By the definition of $rank_{\mathbf{x}}(y)$, we have:

$$|rank_{\mathbf{x}}(y) - \frac{n}{2}| = -q(y; \mathbf{x}).$$

Then, we know:

$$\begin{aligned} Pr_{y\sim A_\epsilon(\mathbf{x})}[|rank_{\mathbf{x}}(y) - \tfrac{n}{2}| > c \cdot \tfrac{\ln(R)+\ln(1/\beta)}{\epsilon}] &\equiv Pr_{y\sim A_\epsilon(\mathbf{x})}[-q(y;\mathbf{x}) > c \cdot \tfrac{\ln(R)+\ln(1/\beta)}{\epsilon}] \\ &= Pr_{y\sim A_\epsilon(\mathbf{x})}[q(y;\mathbf{x}) \le -c \cdot \tfrac{\ln(R)+\ln(1/\beta)}{\epsilon}] \end{aligned}$$

By definition of exponential mechanism, we have:

$$\begin{aligned} Pr_{y\sim A_\epsilon(\mathbf{x})}[q(y;\mathbf{x}) \le -c \cdot \tfrac{\ln(R)+\ln(1/\beta)}{\epsilon}] &= \sum_{y|q(y;\mathbf{x})<-c\cdot\frac{\ln(R)+\ln(1/\beta)}{\epsilon}} \frac{\exp\left(q(y;\mathbf{x})\epsilon/2S\right)}{\sum_y \exp\left(q(y;\mathbf{x})\epsilon/2S\right)} \\ &\le R\frac{\exp\left(-c\cdot\frac{\ln(R)+\ln(1/\beta)}{\epsilon}\frac{\epsilon}{2S}\right)}{\sum_y \exp\left(q(y;\mathbf{x})\epsilon/2S\right)} \\ &= R\frac{\exp\left(\frac{c}{2}(\ln(\frac{1}{R})+\ln(\beta))\right)}{\sum_y \exp\left(q(y;\mathbf{x})\epsilon/2S\right)} \;(\star) \end{aligned}$$

Since the only one optimal output candidate is the median value where $q(y, \mathbf{x}) = 0$, so we have:

$$(\star) \quad \le \quad \frac{\exp\left(\frac{c}{2}(\ln(\frac{1}{R})+\ln(\beta))\right)}{\exp(0)}$$

In order to have this probability be at most β, we take the equality and get:

$$\begin{aligned} \frac{\exp\left(\frac{c}{2}(\ln(\frac{1}{R})+\ln(\beta))\right)}{\exp(0)} &= \beta \\ R^{1-\frac{c}{2}} &= \beta^{1-\frac{c}{2}} \end{aligned}$$

Since we have $R \ge 1$ and $\beta \in (0, 1)$, there exists $c = 2$ which can make the equation holds. □

2 Problem 2

Proof. The proof are developed by two cases where the insertion and deletion whether or not make changes on empty bins:

case changes on nonempty bins

Taking arbitrary data set $\mathbf{x}$ and one possible output S, there are two cases of insertion to create the adjacent data set $\mathbf{x}'$:

subcase insertion

Inserting one data that makes an empty bin ($k \in \mathcal{X}$) be nonempty, we have the

$$\begin{aligned} \frac{Pr[A(\mathbf{x}')=S]}{Pr[A(\mathbf{x})=S]} &= \frac{Pr[A(\mathbf{x})=S \wedge \tilde{c_k'} < \tau]}{Pr[A(\mathbf{x})=S]} \\ &= \frac{Pr[A(\mathbf{x})=S]\cdot Pr[\tilde{c_k'}<\tau]}{Pr[A(\mathbf{x})=S]} \\ &= Pr[1 + Lap(\tfrac{1}{\epsilon}) < \tau] \\ &\le (1 - \tfrac{\delta}{2}) \end{aligned}$$

subcase deletion.

Deleting one data point that makes the existing nonempty bins ($k \in \mathcal{X}$) be empty bins, we have the probability as:

$$\begin{aligned}
\frac{Pr[A(\mathbf{x}')=S]}{Pr[A(\mathbf{x})=S]} &= \frac{Pr[A(\mathbf{x}')=S]}{Pr[A(\mathbf{x}')=S \ \wedge \ \tilde{c_k}<\tau]} \\
&= \frac{Pr[A(\mathbf{x}')=S]]}{Pr[A(\mathbf{x}')=S]\cdot Pr[\tilde{c_k}<\tau]} \\
&= \frac{1}{Pr[1+Lap(\frac{1}{\epsilon})<\tau]} \\
&\geq \frac{1}{1-\frac{\delta}{2}}
\end{aligned}$$

case no change on nonempty bins

Taking arbitrary data set $\mathbf{x}$ and a possible output S, There are two cases of creating the adjacent data set $\mathbf{x}'$:

subcase insertion

insert one data which makes an empty bin ($k \in \mathcal{X}$) be nonempty, we have the

$$\begin{aligned}
\frac{Pr[A(\mathbf{x}')=S]}{Pr[A(\mathbf{x})=S]} &= \frac{Pr[A(\mathbf{x}\backslash k)=S\backslash k \wedge \tilde{c_k'}=\tilde{c_k}]}{Pr[A(\mathbf{x})=S]} \\
&= \frac{Pr[A(\mathbf{x}\backslash k)=S\backslash k]\cdot Pr[\tilde{c_k'}=\tilde{c_k}]}{Pr[A(\mathbf{x})=S]} \\
&= \frac{Pr[c_k'+Lap(\frac{1}{\epsilon})=\tilde{c_k}]}{Pr[c_k+Lap(\frac{1}{\epsilon})=\tilde{c_k}]} \\
&\leq e^{\epsilon}
\end{aligned}$$

subcase deletion

insert one data point in the existing nonempty bins ($k \in \mathcal{X}$). We have the probability as:

$$\begin{aligned}
\frac{Pr[A(\mathbf{x}')=S]}{Pr[A(\mathbf{x})=S]} &= \frac{Pr[A(\mathbf{x}\backslash k)=S\backslash k \wedge \tilde{c_k'}=\tilde{c_k}]}{Pr[A(\mathbf{x})=S]} \\
&= \frac{Pr[A(\mathbf{x}\backslash k)=S\backslash k]\cdot Pr[\tilde{c_k'}=\tilde{c_k}]}{Pr[A(\mathbf{x})=S]} \\
&= \frac{Pr[c_k'+Lap(\frac{1}{\epsilon})=\tilde{c_k}]}{Pr[c_k+Lap(\frac{1}{\epsilon})=\tilde{c_k}]} \\
&\geq e^{-\epsilon}
\end{aligned}$$

By summarization, the probability of failure would be δ above cases. So we have the algorithm be (ϵ, δ)-DP. □