Homework 4

Lijie Hu

1 Problem 1

We first prove it is ϵ-LDP, by the definition of Y_i we known that $\Pr(Y_i = 1), \Pr(Y_i = 0) \in [\frac{1}{e^\epsilon+1}, \frac{e^\epsilon}{e^\epsilon+1}$ no matter what x_i is. Thus we always have $e^{-\epsilon} \leq \frac{\Pr(Y_i=1|x_i)}{\Pr(Y_i=1|x'_i)} \leq e^\epsilon$ for any pair of x_i, x'_i.

We then consider the unbiaseness, we have

$$\begin{aligned}\mathbb{E}(\hat{\mu}) &= \frac{m}{n}(\sum_{i=1}^{n} \mathbb{E}[Y_i]\frac{e^\epsilon+1}{e^\epsilon-1} - \frac{1}{e^\epsilon-1}) \\ &= \frac{m}{n}(\sum_{i=1}^{n}(\frac{1}{e^\epsilon+1} + \frac{x_i(e^\epsilon-1)}{m(e^\epsilon+1}))\frac{e^\epsilon+1}{e^\epsilon-1} - \frac{1}{e^\epsilon-1}) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i.\end{aligned}$$

Finally we will show the variance, first we have

$$\hat{\mu} - \mu = \frac{m}{n}\sum_{i=1}^{n}(Y_i - \mathbb{E}(Y_i))\frac{e^\epsilon+1}{e^\epsilon-1}. \tag{1}$$

Thus,

$$\mathrm{Var}(\hat{\mu}) = \frac{m^2}{n^2}(\frac{e^\epsilon+1}{e^\epsilon-1})^2\sum_{i=1}^{n}\mathrm{Var}(Y_i) = O(\frac{m^2}{n\epsilon^2}\mathrm{Var}(Y_i)). \tag{2}$$

For each i since $Y_i \in [0, 1]$ thus we have

$$\mathrm{Var}(Y_i) \leq 1$$

. Thus we have the proof.

2 Problem 2

We first proof it is ϵ-LDP. By the definition of z_i we known that $\Pr(z_i = z) \in [\frac{1}{e^\epsilon+1}, \frac{e^\epsilon}{e^\epsilon+1}]$ for any possible output, no matter what x_i is. Thus we always have $e^{-\epsilon} \leq \frac{\Pr(Y_i=1|x_i)}{\Pr(Y_i=1|x'_i)} \leq e^\epsilon$ for any pair of x_i, x'_i.

For unbiaseness, if $x = 0$, then we can see that $\mathbb{E}[z] = 0$. Otherwise by definition

$$\mathbb{E}[z_i] = \frac{1}{m}\sum_{j=1}^{m}(0, 0, \cdots, \mathbb{E}[z_{i,j}], 0, \cdots, 0) = \frac{1}{m}\sum_{j=1}^{m}(0, 0, \cdots, mx_{i,j}, 0, \cdots, 0) = x_i.$$

3 Problem 3

I mainly follow the idea in [2]. Consider a fixed distance k, the worst case is that we can change up to k entries in $x_1, ..., x_n$ to 0 or Λ and then change the median value from x_{m-k} to x_{m+k}. Therefore, when the median is an end point of a large empty interval the local sensitivity at distance k is maximized. In order to achieve that, we can modify entries $x_{m-k+1}, \cdots, x_{m-1+t}$ for some $t = 0, \cdots, k+1$ to 0 or Λ. By the definition of the smooth sensitivity we have

$$A^{(k)}(x) = \max_{y:d(x,y)\leq k} LS(y) = \max_{0\leq t\leq k+1}(x_{m+t} - x_{m+t-k-1})$$

And, then we have:

$$S_{f,\epsilon)(D)} = \max_{k=0,\cdots,n}(e^{-k\epsilon} \max_{0\leq t\leq k+1}(x_{m+t} - x_{m+t-k-1}). \tag{3}$$

4 Problem 4

To prove the algorithm is (ϵ, δ)-DP, we know that using the AboveThreshold for finding one query that is larger than c is ϵ'-DP with $\epsilon' = \frac{\epsilon}{2\sqrt{2c\log 2/\delta}}$. Thus, by the composition theorem (Corollary 7.4) we can see that the algorithm is (ϵ, δ)-DP. The same for the ϵ-DP.

Next we focus on the accuracy, first by Theorem 9.2 we know that Abovethreshold is (α, β)-accurate for $\alpha = \frac{8(\log k + \log 2/\beta)}{\epsilon}$. Since Algorithm 2 is c compositions of Abovethreshold. Thus here $\epsilon = \epsilon' = \frac{\epsilon}{2\sqrt{2c\log 2/\delta}}$ and $\beta = \frac{\beta}{c}$ since we need the accuracy holds for each query. Thus, for (ϵ, δ)-DP, Algorithm 2 will be (α, β)-accurate for $\alpha = \frac{16\sqrt{2c\log 2/\delta}(\log k + \log 2c/\beta)}{\epsilon}$. Thus same for ϵ-DP.