HW8b: Local DP

CS 208 Applied Privacy for Data Science, Spring 2022

Version 1.0: Due Fri, April 15, 5:00pm.

1. **DP Histogram in the Shuffle Model.** In this problem, you will construct an algorithm to release a DP histogram of categorical outcomes in the Shuffle Model, and will evaluate the accuracy and compare it against the accuracy under the Local Model and Central Model.

   The dataset `FultonPUMS5full.csv` provides the 5% PUMS Census file for Fulton. In class, we have seen an example[1] of releasing a Local-DP histogram of `educ` clamped to the interval $[1, 16]$.

   (a) Convert the Local-DP algorithm to a Shuffle-DP algorithm using Privacy Amplification by Shuffling. Specifically:

      i. Add a shuffling step that randomly permutes the locally randomized vectors obtained from each user's data.

      ii. Given desired shuffle-privacy parameters (ε, δ), determine how to set the parameters of the local randomizer using the following privacy amplification by shuffling theorem [Feldman, McMillan, Talwar 2021]:[2] If R is ε_0-DP, then for every $\delta \in (0, 1)$ such that $\varepsilon_0 \leq \log\left(\frac{n}{16\log(2/\delta)}\right)$, $M(x_1, \ldots, x_n) = \text{Shuffle}(R(x_1), \ldots, R(x_n))$ is (ε, δ)-DP for

$$\varepsilon \leq \log\left(1 + \frac{\exp(\varepsilon_0) - 1}{\exp(\varepsilon_0) + 1} \cdot \left(\frac{8\sqrt{\exp(\varepsilon_0)\log(4/\delta)}}{\sqrt{n}} + \frac{8\exp(\varepsilon_0)}{n}\right)\right).$$

      iii. For post-processing the shuffled vectors to estimate the histogram, you can use the same post-processing that was used for the local DP algorithm.

   (b) Compare the performance of the Shuffle-DP algorithm, the Local-DP algorithm, and the Central-DP algorithm[3] on subsamples of size n of the dataset, varying n from 200 to 20000. Throughout use privacy parameters $\varepsilon = 1$ and $\delta = 10^{-5}$. On the same graph, plot the sample size n versus error for all three algorithms, where we measure error by the maximum over all bins of the difference between the true count and the DP count.

   From your plot, at what value of n does the Shuffle-DP algorithm start to outperform the local-DP algorithm?

[1] Local-DP histogram: `https://github.com/opendp/cs208/blob/main/spring2022/examples/wk9_local_model.ipynb`

[2] `https://arxiv.org/pdf/2012.12803.pdf`

[3] Code for the Central-DP histogram: `https://github.com/opendp/cs208/blob/main/spring2022/examples/wk3_laplace_mechanism_and_opendp.ipynb`