Homework 1 of DSC 291 Introduction to Differential Privacy (Fall 2024)

University of California, San Diego

Assigned on Oct 4, 2024(Friday)

Due at 11:59 pm on Oct 18, 2021 (Friday)

Notes:

- Be sure to read "Policy on Academic Integrity" on the course syllabus.
- There are ***[100 points]*** in this homework.
- You need to submit your homework via Gradescope.
- Contact the instructor or TA if you spot typos. Any updates or correction will be posted on the course Announcements page and piazza, so check there occasionally.

---

0. Acknowledgment *[0 points]*

For each question in this HW, please list all your collaborators and reference materials (beyond those specified on the website) that were used for this homework.

1. **List of Collaborators** List the names of all people you have collaborated with and for which question(s).
2. **List of Acknowledgements.** If you find an assignment's answer or use a another source for help, acknowledge for which question and provide an appropriate citation (there is no penalty, provided you include the acknowledgement). If not, then write "none".

1 Differential Privacy basics [30 Pts]

(I have covered many of these in the lectures. The point of this question is to make sure you understand the lecture. If you don't follow the lecture, feel free to read the textbook, understand the proof and write them down.)

(a) (6 pts) Prove that differential privacy satisfies closure to post-processing.

(b) (6 pts) Prove that ϵ-DP mechanisms for add/remove one individual satisfies $(k\epsilon)$-DP for adding or removing a maximum of k individuals.

(c) (6 pts) Prove that Randomized Response with parameter $p > 0.5$ satisfies ϵ-DP. Parameterize ϵ as a function of p and come up with a formula to choose p according to a privacy budget ϵ.

(d) (6 pts) Prove that Laplace mechanism satisfies ϵ-DP.

(e) (6 pts) Prove that exponential mechanism satisfies ϵ-DP.

2 Utility of Randomized Response Data Release in the Query Release problem [28 pts]

(a) (7 pts) Consider *simple mathematical model* of a dataset as in the Dinur-Nissm attack, i.e., the dataset x is a bit vector of length n. Suppose we use the Randomized Response with parameter $p > 0.5$ to release every coordinate of x. Let the output vector be $x' \in \{0,1\}^n$. Construct an estimator $\hat{x} \in \mathbb{R}^n$ using x' and p such that $\mathbb{E}[\hat{x}] = x$.

(Hint: $\hat{x}$ is a linear transformation of x'.)

(b) (7 pts) Show that for any fixed linear query $q \in \{0,1\}^n$, your constructed synthetic data $\hat{x}$ from Part (a) satisfies that $\mathbb{E}[q^T\hat{x}] = q^T x$.

(c) (7 pts) Apply *Hoeffding's inequality* to obtain a high probability bound of $q^T\hat{x}$ around the ground truth $q^T x$ Write your bound explicitly as a function of p, n and a failure probability β (so the statement would start with "with probability at least $1-\beta$,...").

(d) (7 pts) Now let's say you are given an arbitrary set of k queries (denoted in the query matrix $Q \in \{0,1\}^{k\times n}$) which your mechanism answers with $Q\hat{x}$. Prove a high probability upper bound of the worst-case error $\|Q\hat{x} - Qx\|_\infty$ using your result in Part (c) and the *union bound*.

3 Implement differential private data release for COVID'19 data / an introduction to *autodp* [42 Pts]

Set up your favorite Python programming environment. Install "numpy, scipy, pandas". Also install "autodp" by following instructions here `https://github.com/yuxiangw/autodp`.

Read the notebook here (thanks EricThomson for making it available) for `https://github.com/EricThomson/covid-mapping/blob/master/covid_map.ipynb`, modify it so that you are plotting the state-level daily COVD cases (normalized by the popular of each state) for 09/30/2021.

In this question we will implement how to differential-privately release and visualize this dataset. Answer the following questions using Jupyter notebook and save the notebook as a pdf file (combine with the rest of your written answers), then submit through Gradescope. Please also submit your code separately.

(a) (7 pts) Formulate the problem as a numerical query where f takes the dataset x of individuals and output the daily COVID case count by states on 09/30/2020. Write down the mathematical formula for this query f. You may use your favorite definition of the dataset, but please specify the data-domain and how the output can be computed from x using formal mathematical notation. Explain any choices that are non-standard.

(b) (7 pts) Under the add/remove neighboring relationship, work out the L1 sensitivity of f. You may specify any additional assumption that you are willing to make.

(c) (7 pts) Figure out how to generate a Laplace random variable using `numpy.random`, and implement the Laplace mechanism to privately release $f(x)$ using the formula we learned from the class on how to calibrate the noise to a pre-defined privacy budget ϵ.

(d) (7 pts) Quantitatively measure the accuracy of the output. Plot the ground truth and the differentially private release of the output in terms of the mean normalized L1-distance

$$\frac{1}{\text{Number of states}} \sum_i \frac{|y_i - f(x)_i|}{|\text{Population of State } i|}.$$

Plot the above metric against $\epsilon = [1e-3, 1e-2, 1e-1, 1, 10, 100]$ with the horizontal and vertical axis on log-log-scale.

(e) (7 pts) Qualitatively visualizing the accuracy of the output. Plot the ground truth and the differentially private release of the output on the map using the bubble plot side-by-side for the case when $\epsilon = 0.5$.

(f) (7 pts) ((ϵ, δ)-DP of Laplace Mechanism) Review the first example in this autodp tutorial `https://github.com/yuxiangw/autodp/blob/master/tutorials/tutorial_new_api.ipynb`, use the b parameter you calculated from Part (c) for $\epsilon = 0.5$ to instantiate a `Mechanism` object using `autodp.mechanism_zoo.LaplaceMechanism`. Plot the ϵ parameter of the Laplace mechanism for $\delta = [0.5, 0.4, 0.3, 0.2, 0.1, 1e-8]$.