Data 102, Fall 2023
Homework 6

Due: **5:00 PM** Friday, December 1, 2023

Submission Instructions

Homework assignments throughout the course will have a written portion and a code portion. Please follow the directions below to properly submit both portions.

Written Portion:

- Every answer should contain a calculation or reasoning.
- You may write the written portions on paper or in LaTeX.
- If you type your written responses, please make sure to put it in a markdown cell instead of writing it as a comment in a code cell.
- Please start each question on a new page.
- It is your responsibility to check that work on all the scanned pages is legible.

Code Portion:

- You should append any code you wrote in the PDF you submit. You can either do so by copy and paste the code into a text file or convert your Jupyter Notebook to PDF.
- Run your notebook and make sure you print out your outputs from running the code.
- It is your responsibility to check that your code and answers show up in the PDF file.

Submitting:

You will submit a PDF file to Gradescope containing all the work you want graded (including your math and code).

- When downloading your Jupyter Notebook, make sure you go to File → Save and Export Notebook As → PDF; do not just print page from your web browser because your code and written responses will be cut off.
- Combine the PDFs from the written and code portions into one PDF. Here is a useful tool for doing so. As a Berkeley student, you get free access to Adobe Acrobat, which you can use to merge as many PDFs as you want.
- Please see this guide for how to submit your PDF on Gradescope. In particular, for each question on the assignment, please make sure you understand how to select the corresponding page(s) that contain your solution (see item 2 on the last page).

Late assignments will count towards your slip days; it is your responsibility to ensure you have enough time to submit your work.

Data science is a collaborative activity. While you may talk with others about the homework, please write up your solutions individually. If you discuss the homework with your peers, please include their names on your submission. Please make sure any handwritten answers are legible, as we may deduct points otherwise.

Private Mean Estimation

One of the most important techniques in data analysis and machine learning is mean estimation. It is used a subroutine in essentially every task. In this question, we will explore how to incorporate differential privacy into mean estimation. En route, we will explore the Laplace mechanism, which is one of the fundamental tools in building differentially private algorithms.

Let $S = \{X_1, \dots X_n\}$ be i.i.d. samples from a Bernoulli distribution with unknown mean p. Recall, from HW4, that the sample mean

$$p_n(S) = \frac{1}{n} \sum_{x \in S} x \tag{1}$$

satisfies $|p_n - p| \leq cn^{-1/2}$ with probability 0.99 for some constant c.

In order to incorporate privacy, the main idea is to add noise to the estimator Equation (1). For the noise distribution, we will use the Laplace distribution, which has density given by

$$f_{\mu,b}(x) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right).$$

We will denote this distribution as $\text{Lap}(\mu, b)$. The mean of the distribution is μ and the variance is $2b^2$. The differentially private estimator is given by

$$\hat{p}_{\epsilon,n}(S) = p_n(S) + Y$$

where Y is sampled from $\text{Lap}\left(0, \frac{1}{\epsilon n}\right)$. Here ϵ is a parameter that will control the privacy.

(a) (1 point) Let S_1 and S_2 be two data sets with n binary samples ($\{0,1\}$-valued) each. Additionally, also assume that S_1 and S_2 differ at most by one item. More precisely, we can construct S_2 by removing one element from S_1 and adding another binary value (0 or 1).

Show that the sample means for the two sets are close. Specifically, show:

$$\left|p_n(S_1) - p_n(S_2)\right| \leq \frac{1}{n}. \tag{2}$$

This is referred to as p_n having sensitivity n^{-1}.

(b) (1 point) For any fixed S, explain why $\hat{p}_{\epsilon,n}(S)$ is distributed according to a Laplace distribution. What are the corresponding parameters?

(c) (2 points) First, we will show that the above estimator is still fairly accurate. Show that with probability 0.995 (over the sampling of the noise), for every S, we have

$$\left|p_n(S) - \hat{p}_{\epsilon,n}(S)\right| \leq \frac{20}{\epsilon n}.$$

You may find it especially useful to apply a concentration inequality we learned about in class.

(d) In this part, we will see that the mechanism is ϵ-differentially private. Let us recall the definition of differential privacy in this context. An estimator g is ϵ-differentially private if for all sets $A \subset \mathbb{R}$, we have

$$\Pr[g(S_1) \in A] \leq \exp(\epsilon) \cdot \Pr\left[g(S_2) \in A\right]$$

where S_1, S_2 are two data sets that differ at most by one item.

(i) (3 points) Let $Y_1 \sim \text{Lap}(\mu_1, b)$ and $Y_2 \sim \text{Lap}(\mu_2, b)$. Show that

$$\Pr[Y_1 \in A] \leq \exp\left(\frac{|\mu_1 - \mu_2|}{b}\right) \cdot \Pr[Y_2 \in A].$$

This hints at why the Laplace distribution is particularly well suited for differential privacy.

Hint: Find a bound on the likelihood ratio, and relate that to the inequality above

(ii) (2 points) Using Equation (2) and earlier parts of the question, show that the estimator $\hat{p}_{\epsilon,n}$ is ϵ-differentially private.

(iii) (1 points) Put these steps together show that $\hat{p}_{\epsilon,n}$ is a ϵ-DP estimator for p with error

$$\left|p - \hat{p}_{\epsilon,n}\right| \leq O\left(\frac{1}{\sqrt{n}} + \frac{1}{n\epsilon}\right)$$

with a probability of at least 0.98 over the randomness of the sample and the mechanism.

(e) (1 point) Now, suppose that instead of Bernoulli, the individual samples X_i were real-valued random variables taking values in $[0, 5]$. Which part(s) of the analysis above (if any) would change? You don't need to redo the analysis.

Hint: Start by revisiting part (a), and consider the downstream effects of each revision and/or correction you make.