HW 3: Differential Privacy Foundations

CS 208 Applied Privacy for Data Science, Spring 2022

Version 1.0: Due Fri, Feb. 18, 5:00pm.

Instructions: Submit a single PDF file to Gradescope containing your solutions, plots, and analyses. Submit any code files and notebooks separately on Gradescope. Make sure to list all collaborators and references.

1. **Mechanisms:** Consider the following mechanisms M that takes a dataset $x \in [0,1]^n$ and returns an estimate of the mean $\bar{x} = (\sum_{i=1}^{n} x_i)/n$.

   i $M(x) = [\bar{x} + Z]_0^1$, for $Z \sim \text{Lap}(2/n)$, where for real numbers y and $a \leq b$, $[y]_a^b$ denotes the "clamping" function:

   $$[y]_a^b = \begin{cases} a & \text{if } y < a \\ y & \text{if } a \leq y \leq b \\ b & \text{if } y > b \end{cases} .$$

   ii $M(x) = \bar{x} + [Z]_{-1}^1$, for $Z \sim \text{Lap}(2/n)$.

   iii

   $$M(x) = \begin{cases} 1 & \text{w.p. } \overline{x} \\ 0 & \text{w.p. } 1 - \overline{x}. \end{cases} .$$

   iv $M(x) = Y$ where Y has probability density function f_Y given as follows:

   $$f_Y(y) = \begin{cases} \frac{e^{-n|y-\bar{x}|/10}}{\int_0^1 e^{-n|y-\bar{x}|/10} dy} & \text{if } y \in [0,1]. \\ 0 & \text{if } y \notin [0,1]. \end{cases}$$

   (a) Which of the above mechanisms meet the definition of ϵ-differential privacy for a finite value of ϵ, and what is the smallest value of ϵ (possibly as a function of n) for which they do? As in class, here we are treating n as public knowledge (so it is not a privacy violation to reveal n), and working with the "change-one" definition of DP.

   (b) Consider the algorithms that satisfy ϵ-DP from Problem 1a. Describe how you would modify these algorithms to have a tunable privacy parameter ϵ and a tunable data domain $[a, b]$ (rather than $[0, 1]$).

   (c) Of the algorithms from Problem 1b, which do you consider to be "best" for releasing a DP mean and why? (There is not a single "right" answer for this problem.)

2. **Evaluating DP Algorithms with Synthetic Data:** Consider a dataset $x \in \mathbb{N}^n$ drawn from a Poisson process, which has probability distribution $\Pr[x_i = k] = 10^k e^{-10}/k!$ for natural numbers k (where we consider $k = 0$ to be a natural number and define $0! = 1$).

   (a) **Generate data.** Write a *data generating process* (DGP) function that generates a dataset $x \in \mathbb{N}^n$ according to the above Poisson process.

   (b) **Write a DP function.** Pick one of your differentially private mechanisms from question Problem 1b (generalized to allow for arbitrary choices of ϵ and data range $[a, b]$ as parameters) that releases an estimate of $\bar{x}$. Write a function that takes in a vector of values $x \in [a, b]^n$ and an ϵ, and outputs a differentially private mean. To apply your mechanism to unbounded data $x \in \mathbb{R}^n$, you will have to clamp x to a chosen range $[a, b]$. For simplicity, we will fix $a = 0$ and only consider varying b.

   (c) **Plot the results of your function.** Recall the discussion on clamping from class: if the range is large, the global sensitivity increases, so noise increases and utility drops. However, if you clip the values too aggressively the answer will be biased, and again utility will drop. We study the effect of varying the upper bound b. Vary b in your function and plot the root mean squared error of the results against b for the setting $n = 200$ and $\epsilon = .5$. Identify the approximately optimal value b^* of b for this data distribution (in the sense of minimizing the RMSE).

   (d) **Explain a potential issue.** Suppose we have an actual (not synthetic) dataset $x \in \mathbb{N}^n$ for which we want to release a differentially private mean, and we don't know the exact underlying distribution of x. Again, we need to select the parameter b and want to do so in a way that minimizes the error. A natural idea is to use a *(nonparametric) bootstrap*[1] to generate many datasets that are "similar" to x in place of the data-generating process above, and optimize the choice of b as above. Once we find an optimal value b^*, we then do our differentially private release on the dataset x itself. Explain why this approach is not safe in general and may violate differential privacy.

   (e) **Propose a method.** Propose some alternative methods for determining a good upper bound b for a given sensitive dataset $x \in \mathbb{N}^n$, while continuing to provide the protections of differential privacy. (There is no single "right answer" for this problem.)

3. **Translating DP.** Consider how you would translate the mathematical definition and properties of differential privacy into societal terms. For example, what does it mean to define privacy semantically (as a property of the algorithm or information flow) rather than syntactically (as a property of a dataset, statistical release, or information output)? In one paragraph, reflect on how differential privacy comports with your personal views of privacy as both a technical and societal concept.

[1] In a nonparametric bootstrap, we generate new datasets by sampling with replacement from x itself. You may find the code to implement the bootstrap in the notebook on 2/10.