HW 4: Differential Privacy Foundations 2

CS 208 Applied Privacy for Data Science, Spring 2022

Version 1.0: Due Fri, Feb. 25, 5:00pm.

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

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

   (This is an instantiation of a continuous version of the exponential mechanism.)

   In HW3, we have identified some of the above mechanisms that do not meet the definition of $(\epsilon, 0)$-differential privacy. For those mechanisms, calculate the smallest value of δ (again possibly as a function of n) for which they satisfy (ϵ, δ) differential privacy for a finite value of ϵ.

2. **Regression:** Consider a dataset where each of its n rows is a pair of real numbers (x_i, y_i), each from an interval $[-b, b]$. Suppose we wish to find a best-fit linear relationship $y_i \approx \beta x_i$ between the y's and the x's. Non-privately, a standard way to estimate β is via the OLS regression formula
   $$\hat{\beta} = \hat{\beta}(x, y) = \frac{S_{xy}}{S_{yy}} = \frac{\sum_i x_i y_i}{\sum_i x_i^2}.$$

   This is called *ordinary least-squares (OLS)* regression, since $\hat{\beta}$ is the minimizer of the mean-squared residuals
   $$\frac{1}{n} \sum_i (y_i - \hat{\beta} x_i)^2. \tag{1}$$

(a) Show that the function $\hat{\beta}(x, y)$ has infinite global sensitivity, and hence we cannot get a useful DP estimate of it via a direct application of the Laplace or Gaussian mechanisms.

(b) Show that S_{xy} and S_{xx} have global sensitivity that is bounded solely as a function of b, and hence each of these can be approximated in a DP manner using the Laplace mechanism.

(c) Using Part 2b together with basic composition and post-processing, devise and implement an ϵ-DP algorithm for approximating $\hat{\beta}$ on an arbitrary dataset with $x_i, y_i \in \mathbb{R}$. In addition to the dataset $((x_1, y_1), \ldots, (x_n, y_n))$, your implementation should take as input parameters a clipping bound b and the privacy-loss parameter ϵ.

(d) Evaluate the performance of your algorithm using a Monte Carlo simulation with synthetic data. Set $b = \epsilon = 1$, generate the x_i's uniformly at random from $[-1/2, 1/2]$, and generate the y_i's according to a linear model with slope 1 and Gaussian noise, but clipped to $[-1, 1]$ to satisfy the range requirements:

$$y_i = [x_i + \mathcal{N}(0, .02)]_{-1}^{1}.$$

For each $n = 100, 200, 300, \ldots, 5000$, run many Monte Carlo trials to estimate and plot the bias and standard deviation of both the OLS estimate $\hat{\beta}$ and the DP estimate $\tilde{\beta}$.

(If $\hat{\theta} = \hat{\theta}(z)$ is an estimator of a population parameter θ based on a dataset z, then the *bias* of $\hat{\theta}$ is $\mathrm{E}[\hat{\theta} - \theta]$, where the expectation is taken over both the dataset z and any randomization used by estimator $\hat{\theta}$. The "bias-variance tradeoff" says that the MSE of an estimator is the sum of its squared bias and its variance; in previous homeworks, we evaluated the (R)MSE of DP estimators, now we are doing a finer analysis by separating the MSE into the bias and variance.)

(e) Try to give an intuitive explanation of the source of the bias you see in Part 2d and on what kinds of dataset distributions this might be largest. How might bias in particular (not just MSE) have an impact on downstream applications?

3. **DP vs. Reconstruction Attacks:** Suppose $M : \{0,1\}^n \to \mathcal{Y}$ is an (ϵ, δ)-DP mechanism and $A : \mathcal{Y} \to \{0,1\}^n$ is an adversary that is trying to reconstruct the sensitive bits in the dataset $x \in \{0,1\}^n$ from the output $M(x)$. Suppose the dataset is a random variable $X = (X_1, \ldots, X_n)$ consisting of n iid draws from a Bernoulli(p) distribution, for a known value of p. Prove that the expected fraction of bits that the adversary successfully reconstructs is not much larger than the trivial bound of $\max\{p, 1-p\}$ (which can be achieved by guessing the all-zeroes or all-ones dataset). Specifically:

$$\mathrm{E}\left[\#\{i \in [n] : A(M(X))_i = X_i\}/n\right] \le e^{\epsilon} \cdot \max\{p, 1-p\} + \delta.$$

(Hint: write the quantity inside the expectation as an average of indicator random variables, and for each i, consider running M on the dataset $X^{(i)}$ where we replace the i'th row of X with the fixed value 0.)

4. **Final Project Ideas** The final projects are an important focus of this course, and we want you to start thinking about yours as soon as possible. Please read the "Final Project Guidelines" (`https://github.com/opendp/cs208/blob/main/spring2022/final%20project/Final%20Project%20Guidelines.pdf`) document on the course website and submit about a paragraph as described in the "Topic Ideas" bullet.