CSCE 689: Special Topics in Modern Algorithms for Data Science Fall 2023

Lecture 33 — November 17, 2023

Prof. Samson Zhou *Scribe: Ryan King*

1 Differential Privacy

From the last lecture we have that given $\varepsilon > 0$ and $\delta \in (0, 1)$, a randomized algorithm $A : U^* \to Y$ is (ε, δ)-differentially private if, for every neighboring frequency vectors f and f' and for all $E \subseteq Y$,

$$\mathbf{Pr}[A(f) \in E] \leq e^{\varepsilon} \cdot \mathbf{Pr}[A(f') \in E] + \delta$$

Intuitively, this means that the algorithm A should produce a similar output if a neighboring element of f is selected instead. For small ε, we can think of e^{ε} as $1 + \varepsilon$ or

$$\mathbf{Pr}[A(f) \in E] \leq (1 + \varepsilon) \cdot \mathbf{Pr}[A(f') \in E] + \delta$$

An implication of this is that a deterministic algorithm cannot be differentially private unless they are a constant function.

2 Differential Privacy Properties

What properties would we we like from a rigorous definition of privacy? One property we would like is for privacy loss measures ε to accumulate across multiple computations and dataset. If mechanism M_1 has privacy loss ε_1 and mechanism M_2 has privacy loss ε_2, then releasing the results of both M_1 and M_2 has a privacy loss of $\varepsilon_1 + \varepsilon_2$.

We also like to have the ability to handle post-processing. If a mechanism M_1 has privacy loss of ε_1 and we release $f(M_1)$, then we have privacy loss ε_1.

3 Example Problem

Let's create an example problem. Let's start by asking how many people in a group have a pet? What happens if each person answer with their truthful answer? We'd expect each person to give us the right answer but now we have some privacy loss.

Instead, what would happen if each person flips a coin and answers based on the coin flip? Then the answers aren't useful since everyone answered randomly but the results are private.

Now, what happens if we combine the two approaches. Now each person will think of their home address. If their address is even then they will answer truthfully, otherwise they will think of their

phone number and answer yes if it is even and no otherwise. Now we have some truth and some randomness but how do we estimate the true number if we use this method?

For any person i, let $X_i \in \{0, 1\}$ be the true answer and let $Y_i \in \{0, 1\}$ be the reported answer. In our example above we have that:

$$\mathbf{Pr}[Y_i = X_i] = \frac{3}{4}$$
$$\mathbf{Pr}[Y_i = 1 - X_i] = \frac{1}{4}$$
$$\mathbb{E}[Y_i] = \frac{3}{4} \cdot X_i + \frac{1}{4} \cdot (1 - X_i) = \frac{X_i}{2} + \frac{1}{4}$$

Let $Y = \frac{Y_1 + \cdots + Y_n}{n}$ and $X = \frac{X_1 + \cdots + X_n}{n}$

$$\mathbb{E}[Y_i] = \frac{X}{2} + \frac{1}{4}$$

Report $2(Y - \frac{1}{4})$ for the true fraction.

4 Randomized Response

The example above is called randomized response. In the example, we have the $\mathbf{Pr}[Y_i = 1 | X_i = 1] = \frac{3}{4}$ and $\mathbf{Pr}[Y_i = 1 | X_i = 0] = \frac{1}{4}$. Using the definition of (ε, δ)-differentially private, we can see that

$$\mathbf{Pr}[Y_i = 1 | X_i = 0] \leq 3 \cdot \mathbf{Pr}[Y_i = 1 | X_i = 1]$$
$$\mathbf{Pr}[Y_i = 1 | X_i = 1] \leq 3 \cdot \mathbf{Pr}[Y_i = 1 | X_i = 0]$$

so the privacy loss is $ln(3)$

5 Local Differential Privacy (LDP)

Given $\varepsilon > 0$ and $\delta \in (0, 1)$, a randomized algorithm $A : U^* \to Y$ is (ε, δ)-differentially private if, for every pairs of users' possible data x and x' and for all $E \subseteq Y$,

$$\mathbf{Pr}[A(x) \in E] \leq e^{\varepsilon} \cdot \mathbf{Pr}[A(x') \in E] + \delta$$

Here this the algorithm takes a single user's data compared to the previous definition of DP, where the algorithm takes all users' data.

A example of LDP is in Mobile Data Analytics where LDP can be applied to data collected from mobile devices to allow analysis of aggregate movement patterns and trends without compromising the privacy of individual users. Example of Mobile Data Analytics include location-based services and user behavior analysis.

6 Privacy and Noise

Our goal is to release a private approximation to $f(x)$. The intuition is that $f(x)$ can be released accurately if the function f is not sensitive to changes by any of the individuals $x = x_1, \cdots, x_n$. We can measure the sensitivity of a pair of users using the following:

$$\sigma_f = \max_{neighborsx,x'} ||f(x) = f(x')||_1$$

Suppose that a study is conducted that measures the height of individuals, ranging from 1 to 300 centimeters.

1. What is the sensitivity of the maximum height query? 300
2. What is the sensitivity of the average height query? 25

7 Laplace Mechanism

The goal of our algorithm is to compute $f(x)$ and release $f(x) + Z$, where Z $Lap(\frac{\sigma_f}{\varepsilon})$ where Lap is a Laplacian distribution with a probability density function of:

$$p(x) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$$

The Laplace mechanism is used here because it is ε-differentially private (pure DP).

8 Exponential Mechanism

What if the output is not a scalar, e.g. a vector? Suppose the outputs lie in some space Y. Or suppose a study is conducted that finds the current location of individuals, in the two-dimensional plane. Who is the closest individual to a query location? In these cases the Laplace mechanism doesn't work so we have to use an exponential mechanism. To use this mechanism we have to choose a score function $S : (Y, X^n) \to \mathbb{R}$ and a sensitivity σ. Sample $y \in Y$ with probability proportional to

$$\exp\left(\frac{\varepsilon}{2\sigma} S(y, x)\right)$$

The exponential mechanism is ε-differentially private (pure DP). In fact, when Y is the set of the real numbers, there is a setting of the score function S for which the exponential mechanism reduces to the Laplace mechanism. However, the sampling process may be inefficient.