CSCE 689: Special Topics in Modern Algorithms for Data Science

Lecture 33

Samson Zhou

Presentation Schedule

- November 27: Chunkai, Jung, Galaxy AI
- November 29: STMI, Anmol, Jason
- December 1: Bokun, Ayesha, Dawei, Lipai

Last Time: Differential Privacy

- [DMNS06] 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 $E \subseteq Y$,

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

Last Time: Differential Privacy

- [DMNS06] 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 $E \subseteq Y$,

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

$A($  $) \approx A($  $)$

Last Time: Differential Privacy

- [DMNS06] 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 $E \subseteq Y$,
$$\Pr[A(f) \in E] \leq e^{\varepsilon} \cdot \Pr[A(f') \in E] + \delta$$

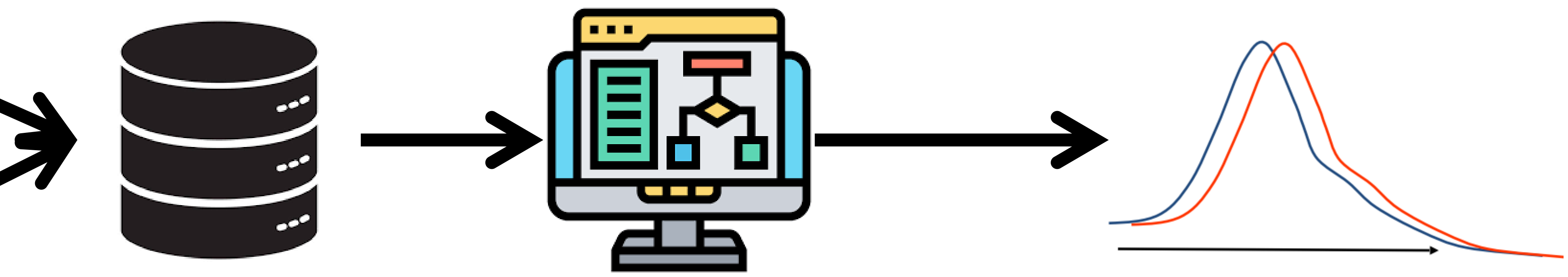
- For small ε, can think of e^{ε} as $1 + \varepsilon$

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

Last Time: Differential Privacy

- [DMNS06] 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 $E \subseteq Y$,

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



Last Time: Differential Privacy

- [DMNS06] Given $\varepsilon > 0$ and $\delta \in (0,1)$, a randomized algorithm $A: U^* \rightarrow Y$ is (ε, δ)-differentially private if, for every neighboring frequency vectors f and f' and for all $E \subseteq Y$,
$$\Pr[A(f) \in E] \leq e^{\varepsilon} \cdot \Pr[A(f') \in E] + \delta$$

- Implication: Deterministic algorithms cannot be differentially private unless they are a constant function

Differential Privacy Properties

- What properties would we like from a rigorous definition of privacy?

Differential Privacy Properties

- Privacy loss measure ε accumulates across multiple computations and datasets
  - 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 privacy loss $\varepsilon_1 + \varepsilon_2$
- Ability to handle post-processing
  - If mechanism M_1 has privacy loss ε_1 and we release $f(M_1)$, then we have privacy loss ε_1

Counting

- How many people in the population satisfy some property?

- How many people in this class have a pet?

Counting

- How many people in this class have a pet?

- What happens if each person answers with their truth?

Counting

- How many people in this class have a pet?

- What happens if each person flips a coin and answers with the coin flip?
- Think of your favorite (integer) number:
  - If it is even, answer YES
  - Otherwise if it is odd, answer NO

Counting

- How many people in this class have a pet?

- Think of your home address:
  - If it is even, answer truthfully
  - Otherwise, proceed below
- Think of your phone number:
  - If it is even, answer YES
  - Otherwise if it is odd, answer NO

Counting

- How to estimate the true number?

- 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

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

Counting

- $\Pr[Y_i = X_i] = \frac{3}{4}$ and $\Pr[Y_i = 1 - X_i] = \frac{1}{4}$
- $\mathrm{E}[Y_i] = \frac{3}{4} \cdot X_i + \frac{1}{4} \cdot (1 - X_i) = \frac{X_i}{2} + \frac{1}{4}$
- Let $Y = \frac{Y_1 + \ldots + Y_n}{n}$ and $X = \frac{X_1 + \ldots + X_n}{n}$

- $\mathrm{E}[Y] = \frac{X}{2} + \frac{1}{4}$
- Report $2\left(Y - \frac{1}{4}\right)$ for true fraction

Randomized Response

- $\Pr[Y_i = 1 \mid X_i = 1] = \frac{3}{4}$
- $\Pr[Y_i = 1 \mid X_i = 0] = \frac{1}{4}$

- $\Pr[Y_i = 1 \mid X_i = 0] \leq 3 \cdot \Pr[Y_i = 1 \mid X_i = 1]$
- $\Pr[Y_i = 1 \mid X_i = 1] \leq 3 \cdot \Pr[Y_i = 1 \mid X_i = 0]$
- Privacy loss $\ln 3$

Differential Privacy

- [DMNS06] Given $\varepsilon > 0$ and $\delta \in (0,1)$, a randomized algorithm $A: U^* \rightarrow Y$ is (ε, δ)-differentially private if, for every neighboring frequency vectors f and f' and for all $E \subseteq Y$,

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

Local Differential Privacy (LDP)

- [KLNRS08] Given $\varepsilon > 0$ and $\delta \in (0,1)$, a randomized algorithm $A: U^* \rightarrow Y$ is (ε, δ)-differentially private if, for every pairs of users' possible data x and x' and for all $E \subseteq Y$,

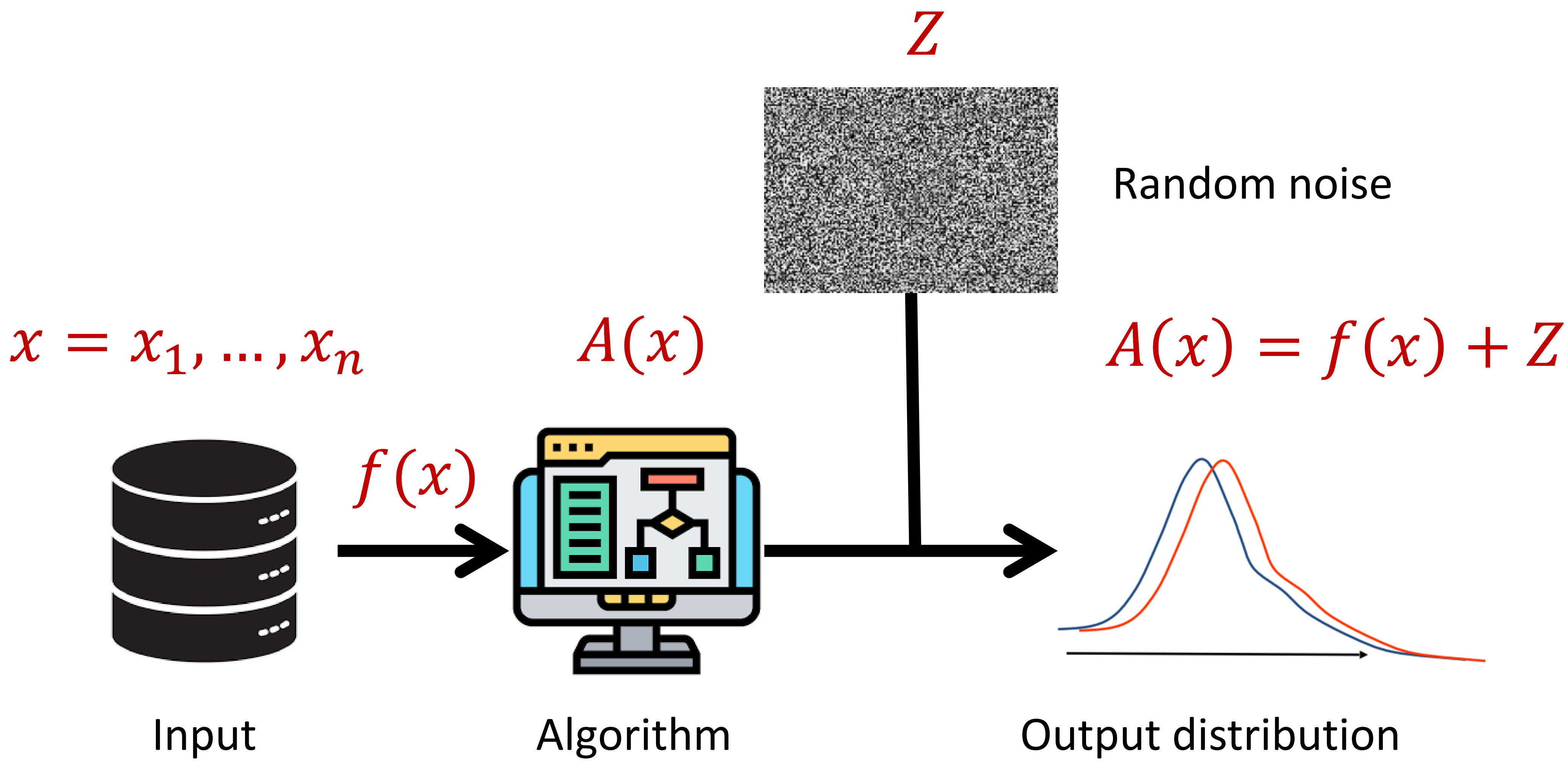
$$\Pr[A(x) \in E] \leq e^{\varepsilon} \cdot \Pr[A(x') \in E] + \delta$$

- Algorithm takes a single user's data
- Compared to previous definition of DP, where algorithm takes all users' data

Local Differential Privacy (LDP)

- Mobile Data Analytics: 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
  - Location-based services
  - User behavior analysis

Privacy and Noise



Privacy and Noise

- Goal: release private approximation to $f(x)$
- Intuition: $f(x)$ can be released accurately if the function f is not sensitive to changes by any of the individuals $x = x_1, \dots, x_n$
- Sensitivity: $\sigma_f = \max_{\text{neighbors } x,x'} \|f(x) - f(x')\|_1$

Sensitivity

- Sensitivity: $\sigma_f = \max_{\text{neighbors } x, x'} \| f(x) - f(x') \|_1$

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

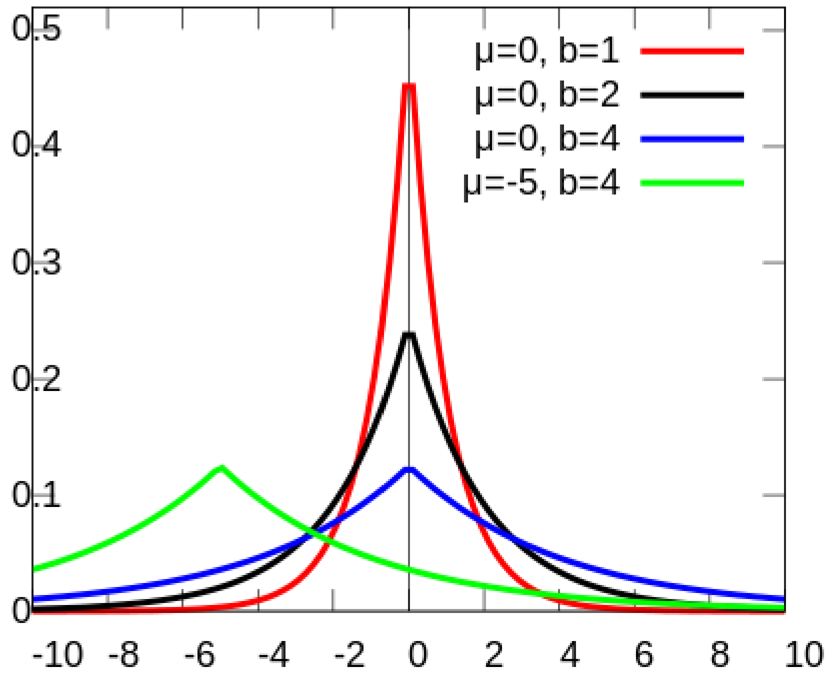
- What is the sensitivity of the maximum height query?
- What is the sensitivity of the average height query?

Laplace Mechanism

- Goal: Algorithm computes $f(x)$ and releases $f(x) + Z$, where $Z \sim \mathrm{Lap}\left(\frac{\sigma_f}{\varepsilon}\right)$

- Laplacian distribution: Probability density function for $\mathrm{Lap}(b)$ is

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



Laplace Mechanism

- What does the Laplace mechanism do in the following cases?

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

- What is the sensitivity of the maximum height query?
- What is the sensitivity of the average height query?

Laplace Mechanism

- Theorem: Laplace mechanism is ε-differentially private (pure DP)

Beyond Laplace Mechanism

- What if the output is not a scalar, e.g., a vector?

- Suppose the outputs lie in some space Y

Beyond Laplace Mechanism

- 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?

Exponential Mechanism

- Choose a score function $S: (Y, X^n) \to \mathbb{R}$ and global sensitivity σ

- Sample $y \in Y$ with probability proportional to $\exp\left(\frac{\varepsilon}{2\sigma} S(y, x)\right)$

Exponential Mechanism

- Theorem: 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 down to the Laplace mechanism

- Downside: sampling process may be inefficient