Lecture 6

Recap

- Nice properties of DP
  - Composition
  - Post-Processing
  - Group Privacy
    - ↳ Interpreting DP.

- Solving Seletion Problem with DP.

Definition. (Differential Privacy)

A is ε-differentially private if

for all neighbors x and x'

for all subsets E of outputs

$$\mathbb{P}[A(x) \in E] \leq e^{\varepsilon} \, \mathbb{P}[A(x') \in E]$$

$\downarrow$

How small can ϵ be?

Adaptive Composition (of 2 mechanisms)

Suppose $A_1 : \mathcal{X}^n \longmapsto \mathcal{Y}_1$ is ε_1 - DP.

$A_2 : (\mathcal{Y}_1 \times \mathcal{X}^n) \longrightarrow \mathcal{Y}_2$ satisfies ε_2 - DP $(\forall y_1 \in \mathcal{Y}_1)$.

Then. $A(x)$:

$a_1 \leftarrow A_1(x)$

$a_2 \leftarrow A_2(a_1, x)$

return (a_1, a_2)

is $(\varepsilon_1 + \varepsilon_2)$ - DP.

Composition of k algorithms $A_1, \dots, A_k$

The choice of A_i depends on $A_1, \dots A_{i-1}$'s outputs

The "adaptive" composition of $A_1, \dots, A_k$

is $\left(\sum_{i=1}^{k} \varepsilon_i\right)$

where each A_i is ε_i- DP.

Proof by induction.

Post-Processing Lemma

post-processing

$$x \longrightarrow \underset{\text{DP}}{\boxed{A}} \longrightarrow a \longrightarrow B \longrightarrow b$$

Lemma. If $A : \mathcal{X}^n \mapsto \mathcal{Y}$ is ε-DP, then $B(A(\cdot))$ is $\underline{\varepsilon\text{-DP}}$ for any $\underline{B : \mathcal{Y} \mapsto \mathcal{Y}'}$.

Proof. Fix $\underset{\text{neighbors}}{x, x'}$, any event $E \subseteq \mathcal{Y}'$.

Case 1: First, focus on B that is deterministic.

Let $B^{-1}(E) = \{a \mid B(a) \in E\}$

$$\mathbb{P}[B(A(x)) \in E] = \mathbb{P}[A(x) \in B^{-1}(E)]$$

$$(\varepsilon\text{-DP of } A) \quad \leq e^{\varepsilon}\, \mathbb{P}[A(x') \in B^{-1}(E)]$$

$$= e^{\varepsilon}\, \mathbb{P}[B(A(x')) \in E]$$

For B to randomized, $B(a) = \underset{\text{deterministic}}{f}(a, \underset{\text{sources of randomness}}{R})$

$A'(x) = (A(x), \underset{\substack{\text{independent of } x \\ 0\text{-DP}}}{R})$ is ε-DP. by composition.

$$B(A(x)) = \underbrace{\underset{\text{deterministic}}{f}(\underbrace{A(x), R}_{\varepsilon\text{-DP}})} \text{ is } \varepsilon\text{-DP}$$

by post-processing Lemma of Case 1.

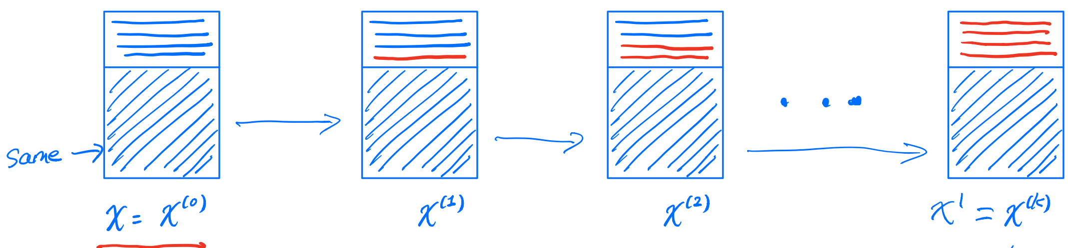
Group Privacy

"What is revealed about k people?"

Lemma. Let $A : \mathcal{X}^n \to \mathcal{Y}$ be ε-DP.
If x and x' differ by k records,
then for any $E \subseteq \mathcal{Y}$

$$\mathbb{P}[A(x) \in E] \le e^{k\varepsilon}\, \mathbb{P}[A(x') \in E]$$

Proof by picture



$$\mathbb{P}[A(x^{(i)}) \in E] \le e^{\varepsilon}\, \mathbb{P}[A(x^{(i+1)}) \in E]$$

(Standard DP)

$$\mathbb{P}[A(x) \in E] \le e^{k\varepsilon}\, \mathbb{P}[A(x') \in E]$$

Observation: Any pair of data sets $x, \tilde{x}$ differ by at most n records

$$\Rightarrow \quad \mathbb{P}[A(x) \in E] \le e^{n\varepsilon}\, \mathbb{P}[A(\tilde{x}) \in E]$$

If $\varepsilon \ll \frac{1}{n}$, two prob. are almost the same.
"No useful info is released."

Interpreting Differential Privacy.

— What should privacy mean?

Naive hope:

You cannot learn anything about me.

Alice is a smoker.

Smoking ⟶ Lung Cancer

BRITISH MEDICAL JOURNAL

LONDON SATURDAY SEPTEMBER 30 1950

SMOKING AND CARCINOMA OF THE LUNG

PRELIMINARY REPORT

BY

RICHARD DOLL, M.D., M.R.C.P.

Member of the Statistical Research Unit of the Medical Research Council

AND

A. BRADFORD HILL, Ph.D., D.Sc.

Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

In England and Wales the phenomenal increase in the number of deaths attributed to cancer of the lung provides one of the most striking changes in the pattern of mortality recorded by the Registrar-General. For example

whole explanation, although no one would deny that it may well have been contributory. As a corollary, it is right and proper to seek for other causes.

But we learn about this whether or not Alice's data is in the study

Differential Privacy Implication

$\boxed{\text{We learn (almost) the same thing about Alice whether or not her data was used.}}$

Formalize w/ Bayesian Stats.

prior $\mathbb{P}[X]$, show $\underbrace{\mathbb{P}[X \mid A(x)]} \underset{e^{2\varepsilon}}{\approx} \underbrace{\mathbb{P}[X \mid A(x')]}$

posterior.

Frank McSherry blog post.

Variations on DP? .

$\longrightarrow$ Additive variation?

Stability property.

$$\mathbb{P}[A(x) \in E] \leq \mathbb{P}[A(x') \in E] \underbrace{+\delta}_{\text{Additive approx}}$$

Still has: Composition, post-processing, group privacy

If $\delta \ll \frac{1}{n}$, $\mathbb{P}[A(x) \in E] \sim \mathbb{P}[A(\tilde{x}) \in E]$ $\forall x, \tilde{x}$

then A is not useful.

"Name & Shame" Algorithm

$NS_\delta(x_1, x_2, \ldots, x_n)$

For each $i = 1, \ldots, n$

Release $Y_i = \begin{cases} X_i & \text{w.p. } \delta \\ \perp & \text{w.p. } (1-\delta) \end{cases}$

For δ in order of $\frac{1}{n}$ (e.g. $\frac{20}{n}$)

NS_δ exposes some individuals' data in the clear.

Approximate DP.

$\forall$ neighbors x & x', $E \subseteq Y$.

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

(ε, δ) - differential privacy.

$\delta \ll \frac{1}{n}$. for A to be meaningfully private.

Selection Problem

Heavy Hitter

Example. A set of websites $\{1,\dots,d\}$

Each user submits $x_i \subseteq \{1,\dots,d\}$ ← list pages they visit.

Winner: website with the highest score: $\forall j \in \{1,\dots,d\}$

$$q(j;x) = |\{i \mid j \in x_i\}|. \quad \text{"\# users visiting"}$$

minimize → $\text{Error} = \max_j q(j;x) - q(A(x);x).$

→ Randomized Response on each x_i

→ Laplace mechanism on q's.

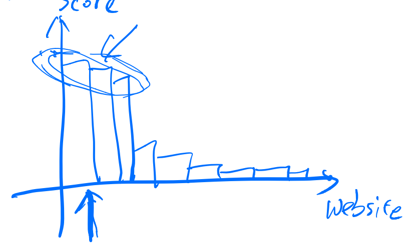
↓

How much noise do you add?

$$\boxed{GS_{(q(1),\dots,q(d))} = d.}$$

→ err scales w/ d.

exp mech has err scaling $\sim \ln(d)$



Example 2: Pricing a digital good.

- Selling an app; what price?
- n people's valuations: "How much are they willing to pay?" $\to x_i$

Revenue:

$$f(p; x) = p \cdot \#\{i : x_i \geq p\}$$

Error

$$\max_p f(p, x) - f(A(x), x)$$

3 people

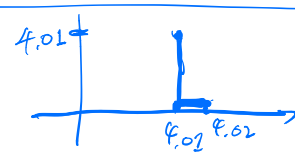
$x_1 = 1$

$x_2 = 1$

$x_3 = 1$

$x_4 = 4.01$

optimal price?



Formulation: Selection Problem

$\mathcal{Y}$: possible outcomes

$q : \underline{\mathcal{Y}} \times \underline{\mathcal{X}^n} \to \mathbb{R}$ "score" function

measures how good y is on dataset x.

q is Δ-sensitive if $\forall y \in \mathcal{Y}$

$q(y;\cdot)$ has $GS_q \le \Delta$.

Exponential Mechanism. $A_{EM}(x, q, \varepsilon, \Delta)$

Output an outcome y with prob. $\propto \exp\left(\frac{\varepsilon}{2\Delta} q(y;x)\right)$

proportional to

When is the prob. distribution well-defined?

→ 1) Finite $\mathcal{Y}$: $P[A(x)=y] = \frac{1}{C_x} \exp\left(\frac{\varepsilon}{2\Delta} q(y;x)\right)$

with $C_x = \sum_{y'} \underline{\exp\left(\frac{\varepsilon}{2\Delta} q(y';x)\right)}$

2) Infinite $\mathcal{Y}$ or Continuous $\mathcal{Y}$.

depends on C_x is well-defined.