Lecture 2 Differential Privacy Basics

Yu-Xiang Wang

COMPUTER SCIENCE
UC SANTA BARBARA
Computing. ReInvented.

Recap: last lecture

- The challenge of privacy in the big data era
  - Remove PII?
  - Reveal only aggregate statistics?
  - Reveal ML models

- Dinur-Nissim attack
  - “Revealing too much information too accurately results in blatant-non-privacy”

This lecture

1. Differential privacy: Definition and interpretations
2. The curator model of private data analysis
3. Mechanism:
   1. Randomized Response, revisited
   2. Laplace Mechanism
4. Applying RR and Laplace mechanism for linear query release

Readings

- Dwork and Roth textbook. Chapter 2 and 3.1-3.3
- Supplementary reading:
  - Differential privacy: A primer for non-technical audience
  - On the `semantic` of differential privacy

How do we formally define privacy?

- We have seen:
  - (“Dinur-Nissm”) Data reconstruction attack
  - Data linkage attack (IMDB → Netflix)
  - Membership inference attack (a small sample of training data / non-training data)
  - …
- It is insufficient to defend against one specific attack.
- Idea: separate “privacy definition” from the actual algorithm that implements the defense.

k-anonymity and composition attack

- K-anonymity (informally): any person's non-sensitive attribute be binned into size >= K
- An example of K-anonymous outputs

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

(a)

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

(b)

Side information: Alice's boss knows she is 28, lives in 13012, and go to both hospitals.

Example from: Ganta, Kasiviswanathan, and Smith. "Composition attacks and auxiliary information in data privacy." In *KDD* 2008.

Any reasonable privacy definition should satisfy the following.

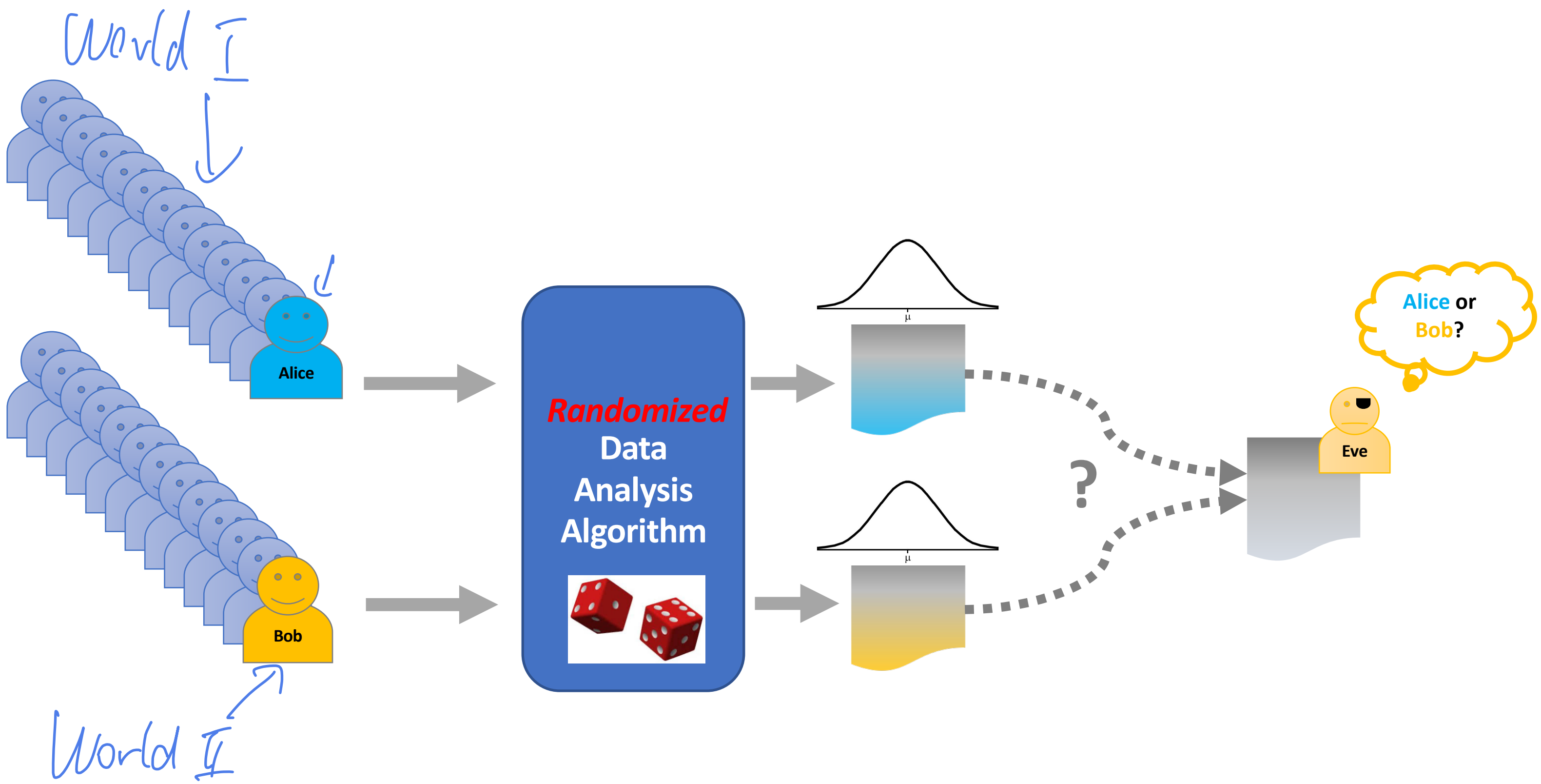
1. Protect against most (if not all) attacks known to date

2. Not making strong assumptions about the adversary

3. Not making strong assumptions about the input data

4. Graceful degradation over composition

The idea of differential privacy --- the indistinguishability of two worlds

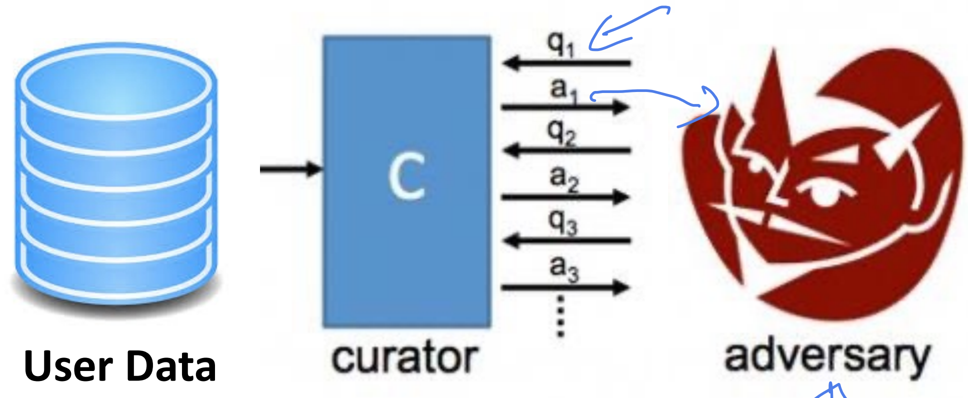


A subtle change of paradigm

- k-anonymity is a definition that covers a property that the (sanitized) output should satisfy, and it does not control how these outputs are obtained.

- In contrast, differential privacy is a property of the algorithm that publishes information from the dataset.

Basic terms: The curator model



interactive query

q_1

(a_1)

$q_2\ (q_1, a_1)$

$q_k \leftarrow (q_1 a_1, q_2 a_2, \ldots q_{k-1} a_{k-1})$

$\boxed{q \in Q}$

offline or non-interactive case.

Defining the jargon. (What do we mean when we talk about the following?)

- Query, trusted curator, query, privacy mechanism, release

Different modes of operations:

- Interactive vs non-interactive query release
- Synthetic data generation
- Training machine learning models

Who is the adversary?

- Examples: Scientists, Readers of the released statistics, users of a recommender system, etc…

Mathematical notations

Probability Simplex

$\mathcal{B}$ $\quad \Delta(\mathcal{B})$

- Output space and a sigma-field:

Range(M)

Output space

- Randomized algorithm: M: Dataspace $\to \Delta(\mathcal{B})$

$M(x)$ is an R.V. $\quad M(x) = Y \sim M(x)$

"Draw" "Sample"

- Data space, individuals, dataset

Example $\mathcal{X} = \{$apple, orange, pear$\}$

$\mathcal{X} \qquad i \in \mathcal{X} \qquad x \in \mathbb{N}^{|\mathcal{X}|}$

Dataset: [apple, pear, apple, orange] $\leftarrow$ Data Matrix

$x = [2, 1, 1] \in \mathbb{N}^{|\mathcal{X}|}$

- Individual vs. data row / data point of an individual

$\tilde{\mathcal{X}} = \{1, 2, 3, \ldots, N\}$

("What's your favorite fruit")

56 $\quad$ 78 $\quad$ 101 $\quad$ 1001

Alternative Representation of the Individuals

564 $\quad$ 784 $\quad$ 101 $\quad$ 1001

$[0 \ldots 0 \, 1 \, 0 \ldots 0 \, 1 \, 0 \ldots 0 \, 1 \, 0 \ldots 0 \, 1 \, 0 \ldots] \in \{0, 1\}^{\tilde{\mathcal{X}}}$

More mathematical notations

- Distance between two datasets

$x, y \in \mathbb{N}^{\mathcal{X}}$ $\|x-y\|_1 = \sum_{i=1}^{|\mathcal{X}|} |x_i - y_i| \Leftarrow$ # of people you need to add/remove to go from x to y

- Neighboring relationship

"Replace One": swapping one individual with another

"Add/Remove": $x \overset{\text{neighbor}}{\simeq} y$ iff $\|x-y\|_1 \leq 1$

Formal definition of differential privacy

Definition 2.4 (Differential Privacy). A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ε, δ)-differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$:

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta,$$

where the probability space is over the coin flips of the mechanism $\mathcal{M}$. If $\delta = 0$, we say that $\mathcal{M}$ is ε-differentially private.

$\delta = 0$ (ε)

$$\log \frac{\Pr(\mathcal{M}(x) \in S)}{\Pr(\mathcal{M}(y) \in S)} \leq \varepsilon$$

"pure"

$$\phi \cup \bigcup_{n=1}^{\infty} \mathcal{X}^n$$

- A few remarks

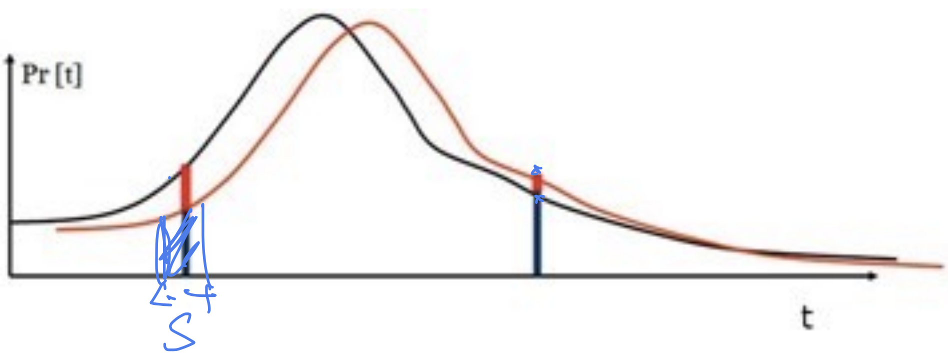
- The randomness is **only** coming from the randomized algorithm.
- We may define "neighboring relationship" differently to encode different granularity of the DP guarantee: e.g., "Add / remove", "Replace"
- This need to hold for **any pairs** of neighboring inputs and **any set** of outputs

Making intuitive sense of the guarantee

Definition 2.4 (Differential Privacy)**.** A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ε, δ)-differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$:

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta,$$

where the probability space is over the coin flips of the mechanism $\mathcal{M}$. If $\delta = 0$, we say that $\mathcal{M}$ is ε-differentially private.



Privacy parameters (ε, δ) measure the "loss of privacy".

- Reasonable ranges of privacy parameter
  - ε is a small constant. ≤ 1 $\quad \varepsilon = O(1)$
  - δ should be very small. o(1/poly(n)) in theory, o(1/n) in practice.

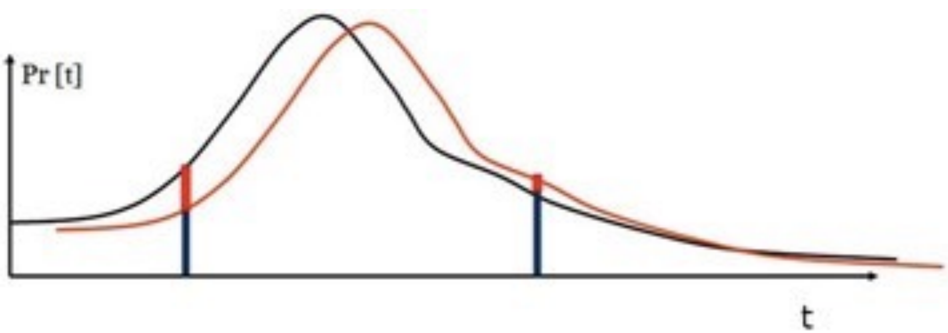
$\delta = \frac{1}{n^2}$

choose $\delta > \frac{1}{n}$

$(0, \frac{1}{n})$-DP

Randomly output (i, x_i)

at random $i \in \text{Uniform}\{1, 2, \ldots, n\}$



We will focus on (pure) ε-DP for the first few lectures.

Making sense of the side-information from a Bayesian interpretation of DP

- Adversary has a prior belief.

$\pi(S)$ X : dataset

- Adversary finds the posterior belief by conditioning on the output

Receives $y \sim M(x)$

$\pi[S \mid M(x)=y]$

- Whether or not "Alice" is in the dataset, the posterior beliefs are about the same.

$$\sup_{x, \text{Alice}} \sup_{\pi} TV[\pi(S \mid M(x)=y), \pi(S \mid M(x_{\text{remove Alice}})=y)] \le e^{\varepsilon}-1$$

- The prior belief can encode any side information.

Kasiviswanathan, S. P., & Smith, A. (2014). On the'semantics' of differential privacy: A bayesian formulation. *Journal of Privacy and Confidentiality*, *6*(1).

Robustness to side-information is a consequence of the worst-case nature of the DP definition.

- Let’s say that there is a distribution the data is sampled from.

$x \sim D$ $\quad X = (x_1, \dots, x_n) \sim D^n$ $\quad \left\{ \begin{array}{l} X \sim D \\ y \sim M(X) \end{array} \right.$

- Knowing any side information allows the adversary to condition on this information, which could change the distribution

Aux $\quad \left\{ \begin{array}{l} X \sim D(\cdot \mid Aux) \\ y \sim M(X) \end{array} \right.$

- But DP applies to all datasets…

Desirable properties of DP

$f \circ M = f(M(\cdot))$

1. Closure to post-processing

M is (ε,δ)-DP $\Rightarrow$ $f \circ M$ is (ε,δ)-DP $\forall f$

Proof: $\underline{\Pr(f \circ M(x) \in S)} = \underline{\Pr(M(x) \in T)} \leq e^{\varepsilon} \Pr(M(y) \in T) + \delta = \underline{e^{\varepsilon} \Pr(f \circ M(y) \in S) + \delta}$

$S \subseteq \mathrm{Range}(f)$, $T = f^{-1}(S) = \{t \in \mathrm{Range}(M) \mid \text{s.t. } f(t) \in S\}$

preimage

2. Composition

Adaptive

M_1 (ε_1,δ_1)-DP, M_2: (ε_2,δ_2)-DP

(M_1, M_2) is $(\varepsilon_1+\varepsilon_2, \delta_1+\delta_2)$-DP.

$\delta_1 = \delta_2 = 0$

$\Pr[(M_1(x), M_2(x)) \in S_1 \times S_2]$

$= \sum_{s \in S_1} \underbrace{\Pr[M_1(x) = s]}_{\leq e^{\varepsilon_1} \Pr[M_1(y) = s]} \cdot \underbrace{\Pr[M_2(x) \in S_2 \mid M_1(x)]}_{\leq e^{\varepsilon_2} \Pr(M_2(y) \in S_2 \mid M_1(s))}$

$\leq e^{\varepsilon_1+\varepsilon_2} \sum_{s \in S_1} \Pr[M_1(y) = s] \Pr[M_2(y) \in S_2 \mid M_1(s)]$

$= e^{\varepsilon_1+\varepsilon_2} \Pr[(M_1(y), M_2(y)) \in (S_1, S_2)]$

3. Small group privacy

$$\boxed{\begin{array}{l} M \text{ is } \varepsilon\text{-DP on Add/Remove One Person} \\ M \text{ is } k\varepsilon\text{-DP on Add/Rem any group of } k \text{ ppl.} \end{array}}$$

An important disclaimer: DP does not prevent all harms of a data analysis

- Example: medical study.
  - A study conducted differential privately may conclude that “Smoking causes lung cancer”
  - Alice is a smoker.
  - Due to this study, Alice’s insurance company increases the premium for all smokers.

- Does this break DP?

The promise of differential privacy

- Decouples the risk of the study itself and the risk of participation.

$\Pr[M(x) \in S] \leq e^{\varepsilon} \Pr[M(y) \in S]$

- Privacy loss ε as a risk multiplier.
  - Any bad things that could happen without your participation can happen at most exp(ε) times higher probability.

- Hides the information specific to individuals, but permits information about the population to be learned accurately.

Checkpoint: qualitative properties of DP

1. Protection against arbitrary risk, not just against re-identification.
2. Automatic neuralization of linkage-attacks from any datasets / other side information
3. Quantifiable privacy loss
4. Composition with graceful degradation
5. Group privacy
6. Closure under post-processing

Remainder of the lecture

- Randomized Response
- Laplace mechanism
- Apply to answering linear queries

Randomized Response, revisited

- Do you like Justin Bieber?
  - Space of the answer: {0,1}

1. Each individual tosses an independent coin with probability p > 0.5
2. If "head", keep your answer.
3. Otherwise, flip your answer.

$RR: \{0,1\} \rightarrow \Delta(\{0,1\})$ Input X Output Y

$E[Y \mid X=1] = p \cdot 1 + (1-p) \cdot 0 = p$

$E[Y \mid X=0] = p \cdot 0 + (1-p) \cdot 1 = 1-p$

estimator: $\hat{X} = 0.5 + \frac{Y-0.5}{2(p-0.5)}$

$E[\hat{X} \mid X=0] = 0.5 + \frac{1-p-0.5}{2(p-0.5)} = 0$

$E[\hat{X} \mid X=1] = 0.5 + \frac{p-0.5}{2(p-0.5)} = 1$

Randomized response satisfies differential privacy!

$RR(p): X \to Y$

- Some questions to address:
  - What is the dataset here? $x \in \{0,1\}$
  - What is the mechanism? $RR(p)$: output $\begin{cases} x & \text{w.p. } p \\ 1-x & \text{w.p. } 1-p \end{cases}$
  - What is the neighboring relationship to define DP? "Replace One"
- What is the privacy parameter of RR(p)?

$\underline{x=1,\ y=0}$ or $\underline{x=0,\ y=1}$

$$P[Y=1 \mid X=1] = p = \frac{p}{1-p}(1-p) = e^{\log\frac{p}{1-p}}(1-p) = \boxed{e^{\log\frac{p}{1-p}}}\ \boxed{P[Y=1 \mid X=0]}$$

$$P[Y=0 \mid X=1] = 1-p = \frac{1-p}{p}\cdot p = e^{\log\frac{1-p}{p}}\cdot p = \boxed{e^{\log\frac{1-p}{p}}}\ P[Y=0 \mid X=0]$$

Case when $x=0,\ y=1$

$$\varepsilon = \log\frac{p}{1-p} \iff e^{\varepsilon} = \frac{p}{1-p} \iff p = \frac{e^{\varepsilon}}{e^{\varepsilon}+1}$$

$\boxed{RR(p) \text{ is } \varepsilon\text{-DP with } \varepsilon = \log\frac{p}{1-p}.}$

Laplace mechanism

- Consider the query aims at releasing real value(s)

$$f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$$

- L1 Sensitivity of the query:

$$\Delta f = \max_{\substack{x,y \in \mathbb{N}^{|\mathcal{X}|} \\ \underbrace{\|x-y\|_1 = 1}_{\text{neighboring relationships}}}} \|f(x) - f(y)\|_1$$

- Laplace mechanism returns

$$f(x) + Z \text{ where } \underbrace{Z_i \sim \mathrm{Lap}(\Delta f/\epsilon)}_{\epsilon\text{-DP}} \text{ i.i.d. for } i \in [k]$$

$$\boxed{b \doteq \frac{\Delta f}{\varepsilon}}$$

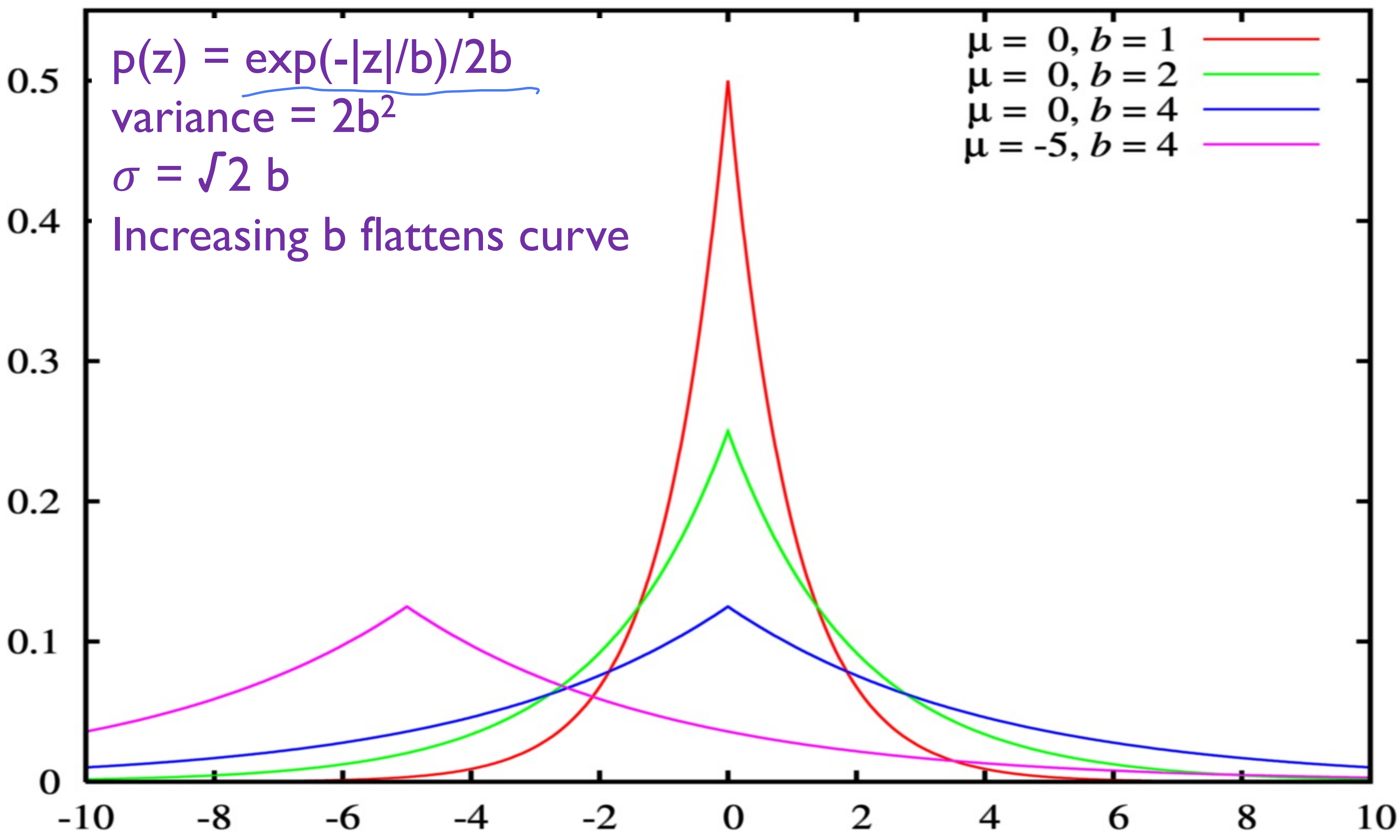
The Laplace distribution

$p(z) = \exp(-|z|/b)/2b$

variance = $2b^2$

$\sigma = \sqrt{2}\, b$

Increasing b flattens curve



(Figure from Wikipedia)

Proof that the Laplace mechanism is differentially private

$$\boxed{b = \frac{\Delta f}{\varepsilon}}$$

- Recall the mechanism returns:

$$f(x) + Z \text{ where } Z_i \sim \mathrm{Lap}(\Delta f/\epsilon) \text{ i.i.d. for } i \in [k]$$

$$M_f(x) = f(x) + Z \qquad P(Z_i) = \frac{e^{-\frac{|Z_i|}{b}}}{2b}$$

$$\boxed{P(M_f(x)=y) \le e^{\varepsilon} P(M_f(x')=y)}$$

$$= \prod_{i=1}^{k} \frac{e^{-\frac{|y-f(x)|_i}{b}}}{2b} = \prod_{i=1}^{k} \frac{e^{-\frac{|y-f(x)|_i - |y-f(x')|_i + |y-f(x')|_i}{b}}}{2b}$$

$$= \left(\frac{1}{2b}\right)^k e^{\sum_{i=1}^{k} \frac{|y-f(x')|_i - |y-f(x)|_i}{b} - \frac{|y-f(x')|_i}{b}}$$

$$\le \left(\frac{1}{2b}\right)^k e^{\sum_{i=1}^{k} \frac{|f(x')-f(x)|_i}{b} - \frac{|y-f(x')|_i}{b}}$$

$$\le \left(\frac{1}{2b}\right)^k e^{\frac{\Delta f}{b}} \cdot e^{-\frac{\sum |y-f(x')|_i}{b}} = e^{\varepsilon} P(M_f(x')=y) \qquad \square$$

Utility of the Laplace Mechanism

- CDF of the Laplace distribution:

$$\begin{cases} \frac{1}{2}\exp\left(\frac{x-\mu}{b}\right) & \text{if } x \leq \mu \\ 1 - \frac{1}{2}\exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \geq \mu \end{cases}$$

Theorem 3.8. Let $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$, and let $y = \mathcal{M}_L(x, f(\cdot), \varepsilon)$. Then $\forall \delta \in (0, 1]$:

$$\Pr\left[\|f(x) - y\|_\infty \geq \ln\left(\frac{k}{\delta}\right) \cdot \left(\frac{\Delta f}{\varepsilon}\right)\right] \leq \delta$$

Example applications of Laplace mechanism. What is the L1 sensitivity?

- Linear query (from the last lecture)

- Histograms: distribution of grades in a class

- Demographics statistics over map:
  - Number of people living in different zip code by race and gender

- COVID'19 Hospitalization Data:
  - Number of active patients in the ICU of each hospital

Apply Laplace mechanism to answer many linear queries

1. Set privacy budget, and number of queries

2. Decide how much noise to add

3. Work out the error bound

4. Error bound => sample complexity

Apply randomized response to answer linear queries

- Answering a single linear query

Hoeffding's inequality: Suppose that $X_1, \ldots, X_n$ are independent and that, $a_i \le X_i \le b_i$, and $\mathbb{E}[X_i] = \mu$. Then for any $t > 0$,

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| \ge t\right) \le 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \quad \text{where } \overline{X}_n = n^{-1}\sum_i X_i.$$

Apply randomized response to answer linear queries

- Answering many linear query

- Question: does it cost any additional privacy?

Comparing randomized response and Laplace mechanism in answering linear queries.

What can we still do?

Target accuracy	k = O(2^n) linear queries	k = O(n) linear queries	k << n linear queries
$\alpha = O(1)$ (any non-trivial error)	Blatantly non-private	?	?
$\alpha = O(1/\sqrt{n})$ (statistical error)	Blatantly non-private	Blatantly non-private	DP / Laplace mech
$\alpha = o(1/\sqrt{n})$ (<< statistical error)	Blatantly non-private	Blatantly non-private	DP / Laplace mech