Privacy in AI

Toniann Pitassi Richard Zemel

CSC 2541

October 8, 2019

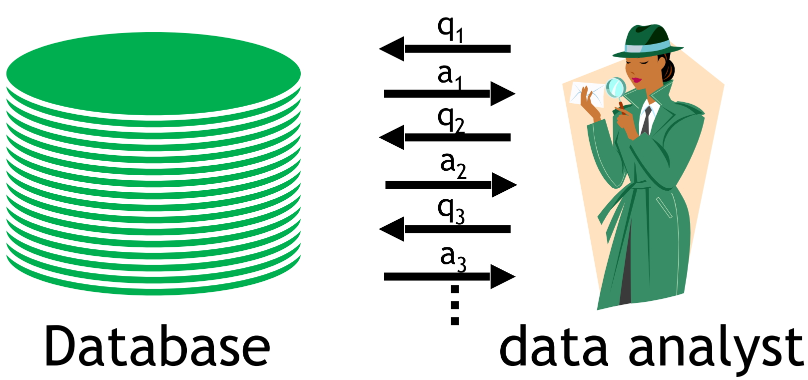
Why Privacy?

Microsoft : tool for diagnosing pancreatic cancer
by monitoring Bing queries

Netflix : film recommender algorithm "anonymized"

Model inversion attacks :
train ML model using sensitive information
hackers can invert model to recover very
sensitive individual info (credit card number)

Privacy-Preserving Data Analysis



- Census, epidemic detection based on OTC drug purchases; analysis of loan application data for evidence of discrimination,…
- 50+ year old problem

What analyses on a database might violate privacy? What analyses are privacy-preserving?

what to promise?

delete identifying information

maybe not

Latanya Sweeney's Attack (1997)

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
			09/27/64	female	02139	divorced	hypertension
			09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	....
....	....	....	....	....	....	....	
....	....	....	....	....	....	....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	....
....	....	....	....	....	....	....	

Figure 1: Re-identifying anonymous data by linking to external data

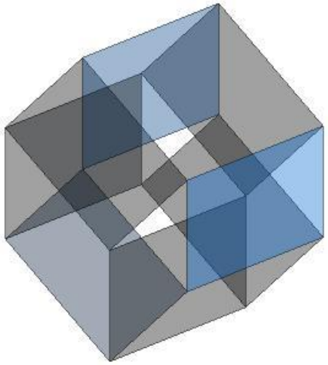
Public voter dataset

K-Anonymity: Intuition

- The information for each person contained in the released table cannot be distinguished from at least k-1 individuals whose information also appears in the release
  - Example: you try to identify a man in the released table, but the only information you have is his birth date and gender. There are k men in the table with the same birth date and gender.
- Any quasi-identifier present in the released table must appear in at least k records

Curse of Dimensionality

Aggarwal (VLDB 2005)



- Generalization fundamentally relies on spatial locality
  - Each record must have k close neighbors
- Real-world datasets are very sparse
  - Many attributes (dimensions)
    - Netflix Prize dataset: 17,000 dimensions
    - Amazon customer records: several million dimensions
  - "Nearest neighbor" is very far
- Projection to low dimensions loses all info $\Rightarrow$ k-anonymized datasets are useless

what to promise?

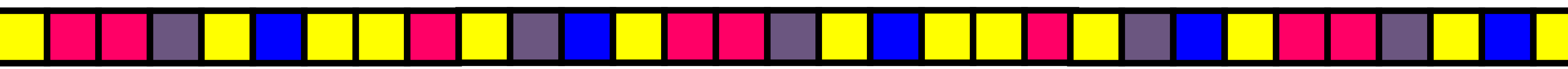
only ask questions that pertain to large populations

maybe not

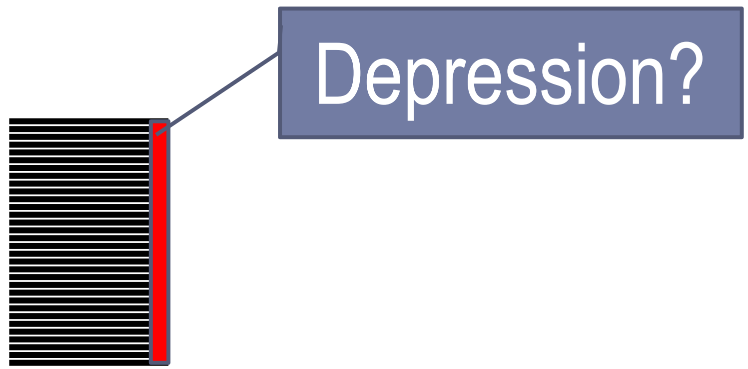
The Statistics Masquerade

- Differencing Attack
  - *How many members of House of Representatives have sickle cell trait?*
  - *How many members of House, other than the Speaker, have the trait?*

- Needle in a Haystack
  - Determine presence of an individual’s genomic data in GWAS case group



- The Big Bang attack
  - Reconstruct “depression” bit column



Dinur and Nissim’03 Homer+’08

Fundamental Law of Info Recovery

- “Overly accurate” estimates of “too many” statistics is blatantly non-private.

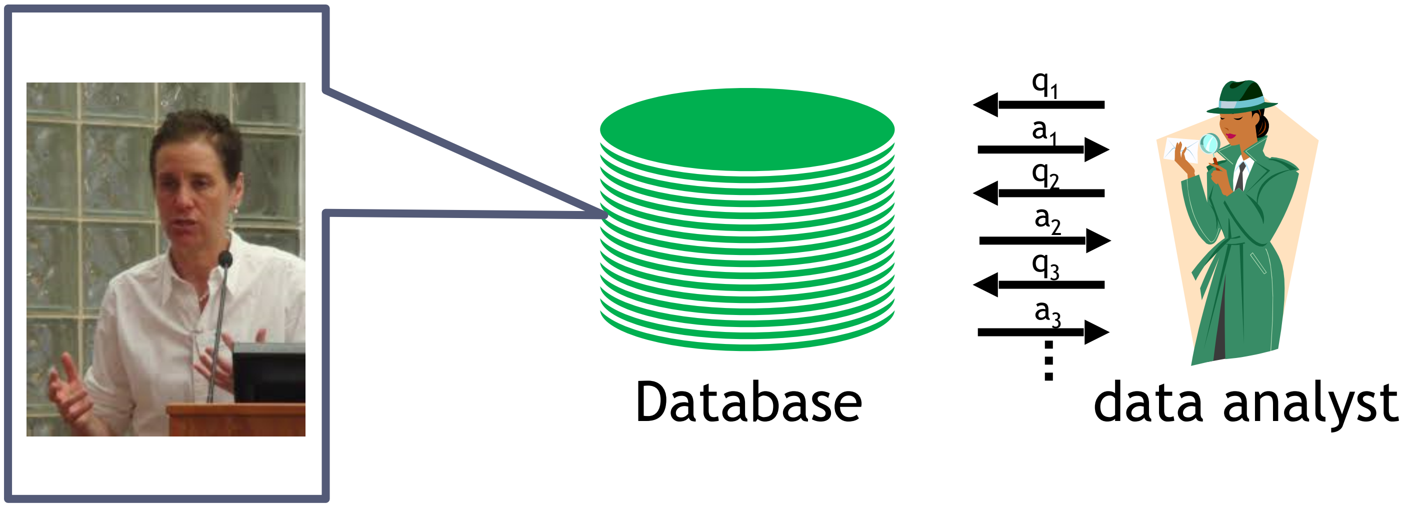
what to promise?

access to the output should not enable one to learn anything about an individual that could not be learned without access

cryptographic definition

is this desirable?

Privacy-Preserving Data Analysis?

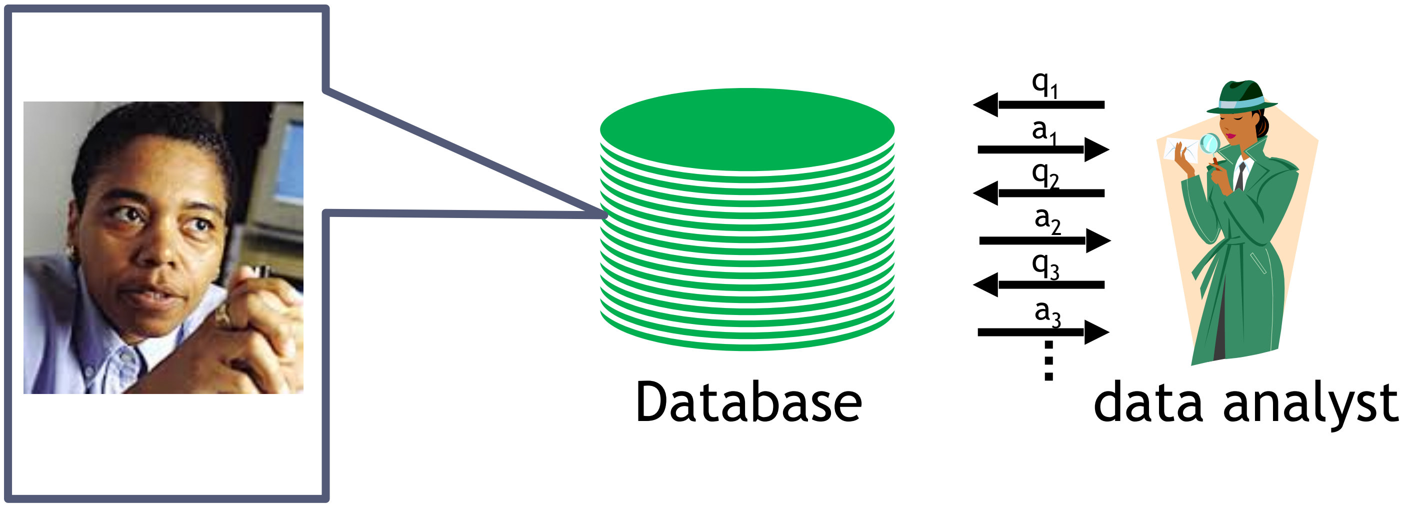


- "Can't learn anything new about Helen"?
- Then what is the point?

what to promise?

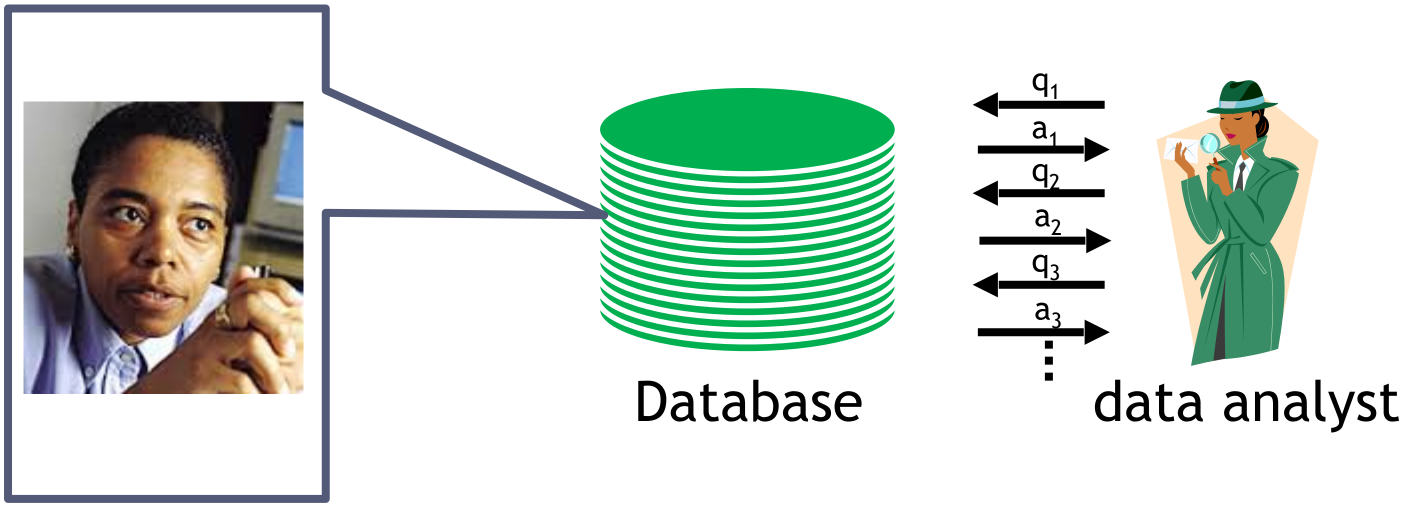
access to the output should not enable one to learn much more about an individual than could be learned via the same analysis omitting that individual from the database

Privacy-Preserving Data Analysis?



- Ideally: learn same things if Helen is replaced by another random member of the population ("stability")

Privacy-Preserving Data Analysis?



- Stability preserves Helen's privacy AND prevents over-fitting
- Privacy and Generalization are aligned!

statistical database model

X set of possible entries/rows

one row per person

database x a set of rows; $x \in \mathbb{N}^{|X|}$ (histogram)

name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
Ellen Jones	4/24/59	F	160	Y	Y
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N

neighboring databases

what's a small change?

require nearly identical behavior on neighboring databases differing by the addition or removal of a single row:

$$||x - y||_1 \leq 1$$

for $x, y \in \mathbb{N}^{|X|}$

differential privacy

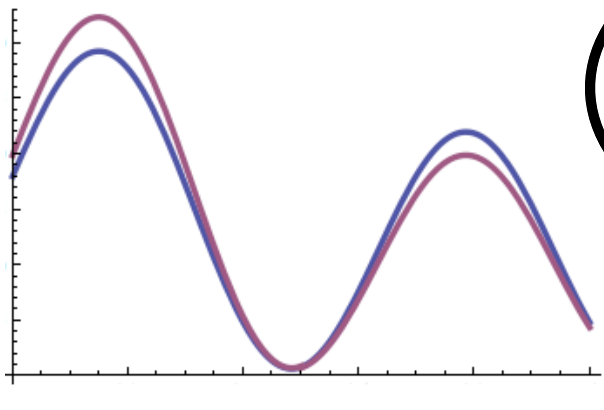
[DinurNissim03, DworkNissimMcSherrySmith06, Dwork06]

ε-Differential Privacy for algorithm M:

for any two neighboring data sets x_1, x_2, differing by the addition or removal of a single row

any $S \subseteq$ range(M),

$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S]$

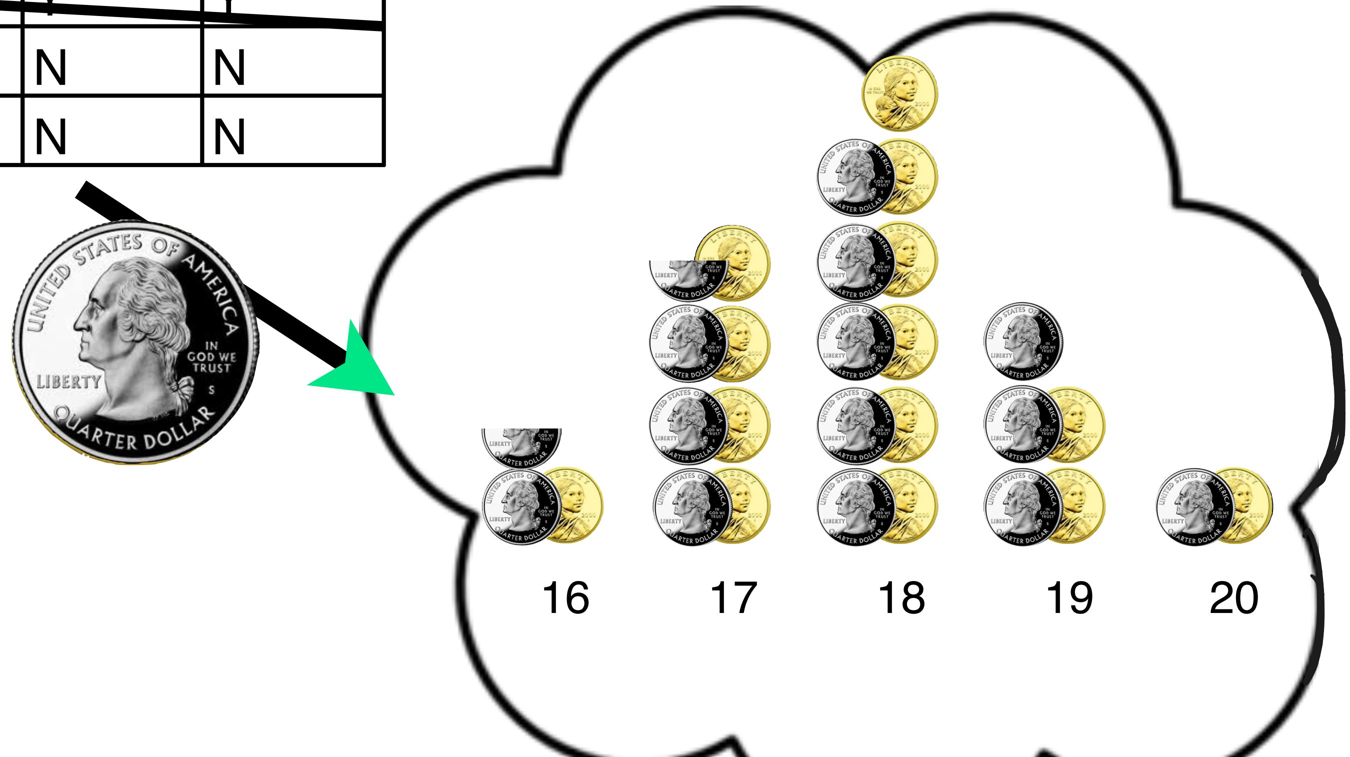


$e^{\varepsilon} \sim (1 + \varepsilon)$

differential privacy

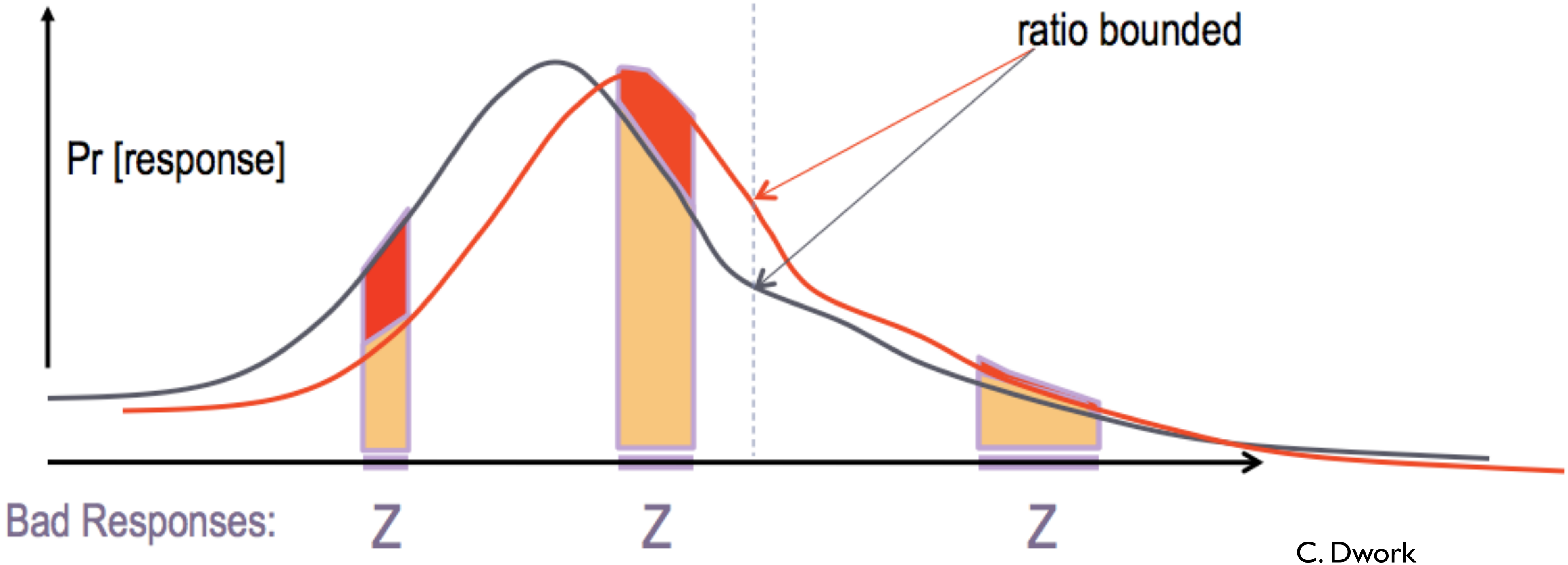
$$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S]$$

name	DOB	sex	weight	smoker	lung cancer
John Doe	12/1/51	M	185	Y	N
Jane Smith	3/3/46	F	140	N	N
~~Ellen Jones~~	~~4/24/59~~	~~F~~	~~160~~	~~Y~~	~~Y~~
Jennifer Kim	3/1/70	F	135	N	N
Rachel Waters	9/5/43	F	140	N	N



differential privacy

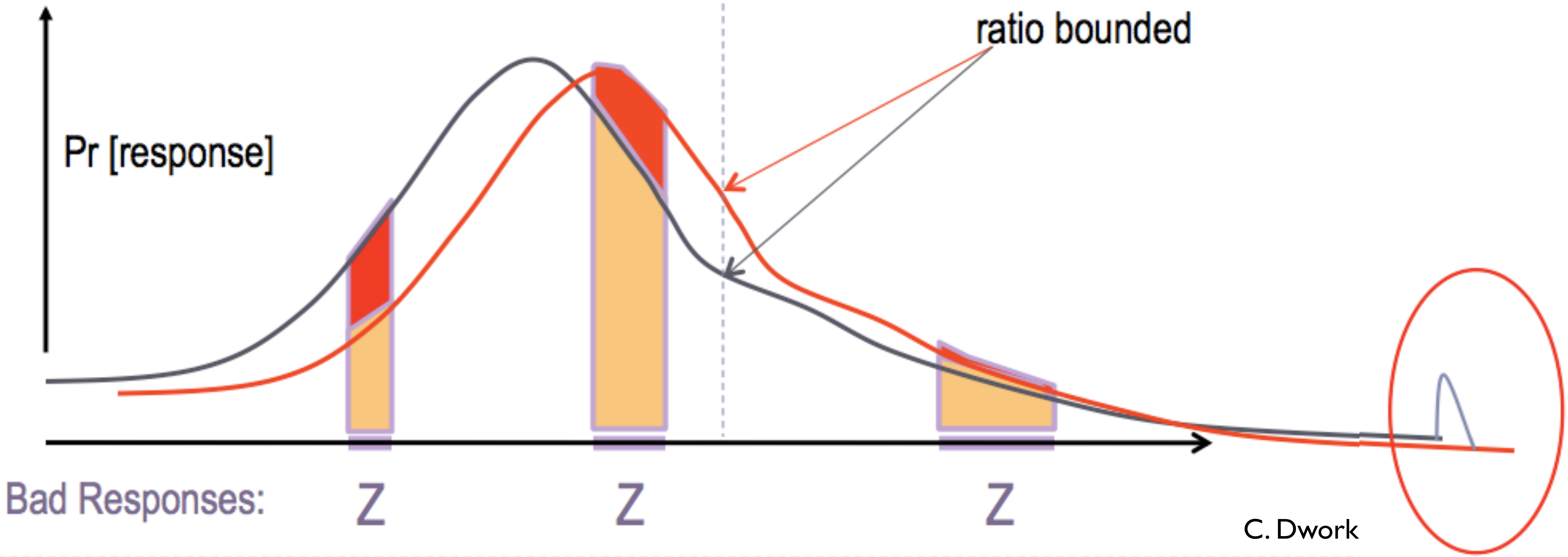
$$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S]$$



C. Dwork

(ε, δ)-differential privacy

$$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S] + \delta$$



differential privacy

$$\Pr[M(x_1) \in S] \leq e^{\varepsilon} \Pr[M(x_2) \in S]$$

promise: if you leave the database, no outcome will change probability by very much

is this achievable
with high accuracy?

yes!

Properties of Differential Privacy

- group Privacy
- postprocessing
- Composition

group privacy

Thm. Any $(\varepsilon, 0)$-DP mechanism M is $(k\,\varepsilon, 0)$-DP for groups of size k. i.e., for all

$$||x - y||_1 \leq k$$

and any $S \subseteq$ range(M),

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

post-processing

Thm. Let $M : \mathbb{N}^{|X|} \rightarrow R$ be (ε, δ)-DP.

Let f: $R \rightarrow R'$ be an arbitrary randomized mapping.

Then $f \circ M : \mathbb{N}^{|X|} \rightarrow R'$ is (ε, δ)-DP.

composition

[DworkKenthapadiMcSherryMironovNaor06,DworkLei09]

Thm. For $i \in [k]$, let $M_i : \mathbb{N}^{|X|} \rightarrow R_i$ be $(\varepsilon_{\mathrm{i}}, \delta_{\mathrm{i}})$-DP. Then the mechanism $(M_1(x), \ldots, M_{\mathrm{k}}(x))$ is $(\sum_i \varepsilon_i, \sum_i \delta_i)$-DP.

DP Mechanisms

- Randomized Response
- Laplacian (+ gaussian) Mechanism
- Noisy Max
- Exponential Mechanism
- (Better) composition

DP Mechanisms

- Randomized Response
- Laplacian (+ gaussian) Mechanism
- Noisy Max
- Exponential Mechanism
- (Better) Composition

Randomized Response
[Warner65]

flip a coin

if tails, respond truthfully

if heads, flip a second coin and respond "yes" if heads; respond "no" if tails

Claim. Randomized Response is (ln 3, 0)-DP.

Proof. $$\frac{\Pr[\text{Response} = \text{Yes}|\text{Truth} = \text{Yes}]}{\Pr[\text{Response} = \text{Yes}|\text{Truth} = \text{No}]}$$

$$= \frac{3/4}{1/4} = \frac{\Pr[\text{Response} = \text{No}|\text{Truth} = \text{No}]}{\Pr[\text{Response} = \text{No}|\text{Truth} = \text{Yes}]} = 3.$$

Randomized Response

Given database $X = X_1, \ldots, X_n$ where $X_i \in \{0,1\}$

(say $X_i = 1$ if person commited crime
$= 0$ if person did not commit crime)

Query: $\sum_i X_i / n$ (= fraction of people that commited crime)

Mechanism:

Step 1. For $i = 1 \ldots n$

Let $Y_i = X_i$ with probability $3/4$
$Y_i = 1 - X_i$ with probability $1/4$

Step 2. Let $f(Y_1 \ldots Y_n) = \sum_i Y_i / n$
Output $f(Y_1 \ldots Y_n)$

<u>Lemma</u> Mechanism is $(\ln 3, 0)$-dp

<u>pf</u> First we show that the output of step 1 $y_1 \ldots y_n$ is $(\ln 3, 0)$-dp. Then by post processing, $f(y_1 \ldots y_n)$ is also $(\ln 3, 0)$-dp.

consider 2 neighboring databases

$$\left.\begin{aligned} x &= x_1 x_2 \ldots x_n \\ x' &= x_1 \ldots \bar{x}_i \ldots x_n \end{aligned}\right\} \text{differ only on coord. } i$$

show $\forall y_1 \ldots y_n$ $\dfrac{\Pr(y_1 \ldots y_n \mid x_1 \ldots x_n)}{\Pr(y_1 \ldots y_n \mid x_1 \ldots \bar{x}_i \ldots x_n)} \leq 3$

$$\frac{\Pr(y_1 \ldots y_n \mid x_1 \ldots x_n)}{\Pr(y_1 \ldots y_n \mid x_1 \ldots \bar{x}_i \ldots x_n)} = \frac{\Pr(y_1 \mid x_1) \cdot \Pr(y_i \mid x_i) \ldots \Pr(y_n \mid x_n)}{\Pr(y_1 \mid x_1) \ldots \Pr(y_i \mid \bar{x}_i) \ldots \Pr(y_n \mid x_n)} = \frac{\Pr(y_i \mid x_i)}{\Pr(y_i \mid \bar{x}_i)}$$

$$\frac{\Pr(Y_i \mid X_i)}{\Pr(Y_i \mid \bar{X}_i)} \leq \frac{3/4}{1/4} = 3 = e^{\varepsilon}$$

$$\Rightarrow \quad \varepsilon = \ln 3$$

---

accuracy of Randomized Response:

Let n' = # of respondents who say 1 $\left(= \sum_i Y_i\right)$

Let $p = \sum_i X_i / n$

$$E(n') = (pn)\tfrac{3}{4} + (1-p)n\tfrac{1}{4} = \frac{pn}{2} + \frac{n}{4}$$

So max likelihood estimator of p, $\hat{p}$ is $\left(n' - \frac{n}{4}\right)\frac{2}{n} = \frac{2n'}{n} - \frac{1}{2}$

and variance of $\hat{p}$ $= \dfrac{p(1-p)}{n} + \dfrac{\frac{3}{4}\left(\frac{1}{4}\right)}{n\left(2\cdot\frac{3}{4}-1\right)^2} = \dfrac{p(1-p)}{n} + \dfrac{3}{4n}$

DP Mechanisms

- Randomized Response also locally DP!
- Laplacian (+ gaussian) Mechanism
- Noisy Max
- Exponential Mechanism
- (Better) composition

DP Mechanisms

- Randomized Response
- Laplacian (+ gaussian) Mechanism
- Noisy Max
- Exponential mechanism
- (Better) Composition

ℓ_1- sensitivity of a function f

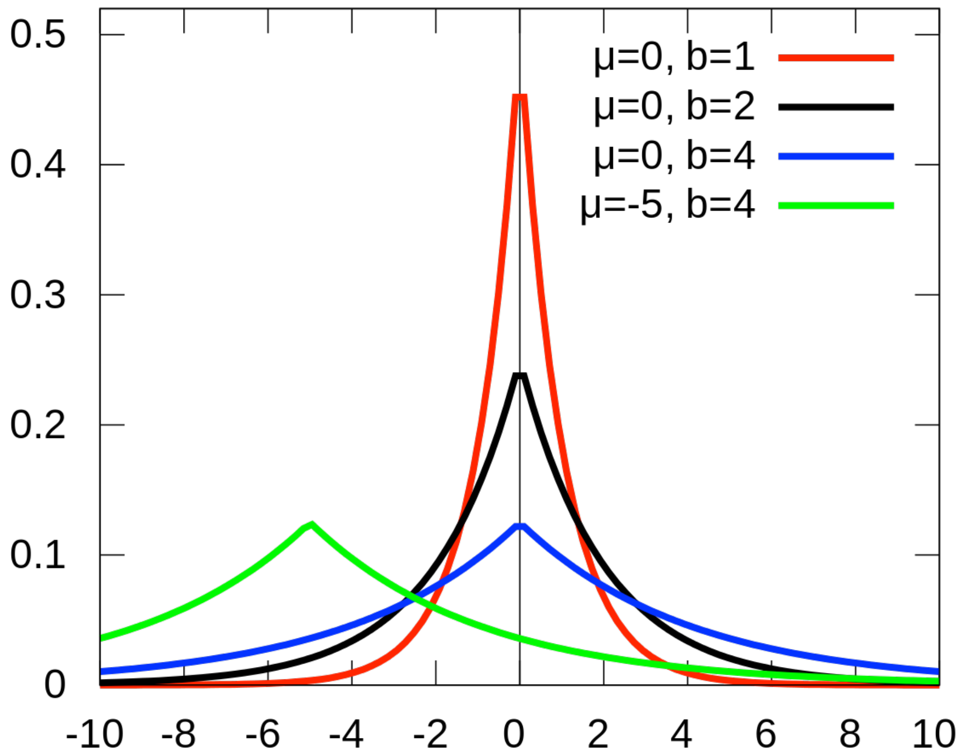
$$\Delta f = \max_{x1,\ x2} \ |f(x_1) - f(x_2)|_1$$

for neighboring data sets $x_1,\ x_2$

measures how much one person can affect output

sensitivity is $1/|x|$ for queries returning the average value of count queries mapping X to $\{0,1\}$

Laplace distribution $\mathrm{Lap}(\mu, b)$



$$\mathrm{pdf}(z) = \frac{1}{2b} \exp\left(\frac{-|z-\mu|}{b}\right)$$

$$\text{variance} = 2b^2$$

For $Y \sim \mathrm{Lap}(b)$, $\Pr[|Y| \geq bt] = \exp(-t)$

Laplace mechanism

Def. Given $f : \mathbb{N}^{|X|} \rightarrow R^k$ the Laplace Mechanism is defined as

$$M_{\mathrm{L}}(x,\ f(.), \varepsilon) = f(x) + (\mathrm{Y}_1,\ \ldots,\ \mathrm{Y}_\mathrm{k})$$

where the Y_i are iid random draws from $\mathrm{Lap}(b)$ with $b = \Delta f/\varepsilon$.

(If we want discrete output space, subsequently round accordingly.)

Laplace mechanism: Privacy

Thm. The Laplace Mechanism preserves

$(\varepsilon, 0)$-differential privacy.

Laplace Mechanism: Privacy

<u>Thm</u> The Laplace Mechanism preserves $(\varepsilon, 0)$-dp

<u>Pf</u> Let x, x' be neighboring databases, so $\|x - x'\|_1 \leq 1$

Let $f: \mathbb{N}^{|X|} \to \mathbb{R}$ $(k=1)$

Let p_x be prob. density function of $M_L(x, f, \varepsilon)$

$p_{x'}$ " " " " " $M_L(x', f, \varepsilon)$

Let $s \in \mathbb{R}$

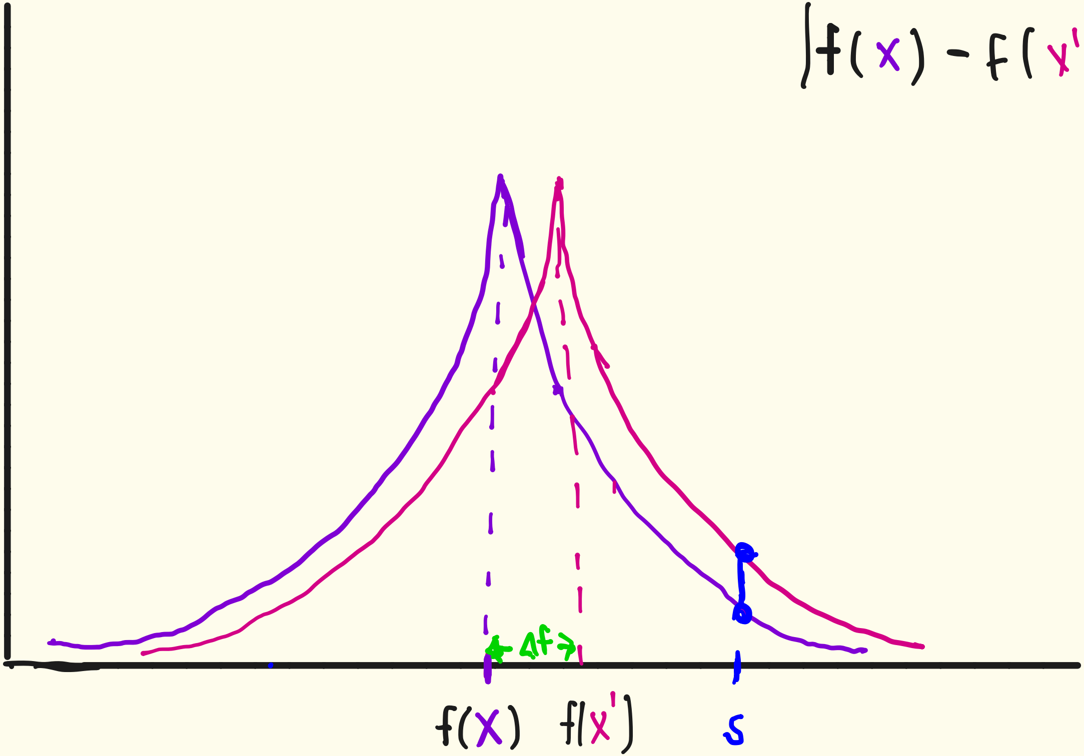
$$\frac{\Pr(s|x)}{\Pr(s|x')} = \left[\frac{\exp\left(-\frac{\varepsilon|f(x) - s|}{\Delta f}\right)}{\exp\left(-\frac{\varepsilon|f(x') - s|}{\Delta f}\right)}\right] = \exp\left[\frac{\varepsilon\left(|f(x') - s| - |f(x) - s|\right)}{\Delta f}\right]$$

$$\leq \exp\left(\frac{\varepsilon|f(x) - f(x')|}{\Delta f}\right) = \exp\left(\frac{\varepsilon\|f(x) - f(x')\|_1}{\Delta f}\right)$$

$$\leq \exp(\varepsilon)$$

$x \qquad x'$

$$|f(x) - f(x')| \leq \Delta f$$



$$\ln\left(\frac{\Pr(s|x)}{\Pr(s|x')}\right) = \frac{\varepsilon\left(|f(x')-s| - |f(x)-s|\right)}{\Delta f} \leq \frac{\varepsilon \Delta f}{\Delta f} \leq \varepsilon$$

Laplace mechanism: Accuracy

Thm. The Laplace Mechanism preserves accuracy

Laplace Mechanism - Accuracy

Thm Let $f : \mathbb{N}^{|X|} \to \mathbb{R}^k$, $y = M_L(x, f, \varepsilon)$. Then $\forall \delta \in [0,1]$:

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

Pf

$$\Pr\left[\|f(x) - y\|_\infty \geq \ln\left(\frac{k}{\delta}\right)\left(\frac{\Delta f}{\varepsilon}\right)\right] = \Pr\left(\max_{i \in [k]} |Y_i| \geq \ln\left(\frac{k}{\delta}\right)\left(\frac{\Delta f}{\varepsilon}\right)\right)$$

$$\leq k \cdot \Pr\left[|Y_i| \geq \ln\left(\frac{k}{\delta}\right)\left(\frac{\Delta f}{\varepsilon}\right)\right]$$

$$= k\left(\frac{\delta}{k}\right)$$

$$= \delta$$

$\Pr\left[|Y| \geq t b\right] = \exp(-t)$

$b = \Delta f / \varepsilon$

Notes

1. Could replace Laplacian by gaussian Noise

   add noise scaled to $N(0, \sigma^2)$, $\sigma \sim \Delta f \ln(1/\delta)/\varepsilon$

   gives (ε, δ)-dp

2. The simpler randomized response algorithm

   is Local

   However overall its accuracy is worse.

DP Mechanisms

- Randomized Response
- Laplacian (+ gaussian) Mechanism
- Noisy Max
- Exponential Mechanism
- (Better) composition

example

Suppose we wanted to determine the most commonly-"liked" Facebook page, subject to DP

> could give a DP count of the number of likes for each page, but sensitivity would grow with the max number of "likes" a person could give (bad)
>
> but we only want to know the max, not *every* count—could that be easier?

reportNoisyMax

For m count queries add noise $\mathrm{Lap}(1/\varepsilon)$ to each, and report the index of the largest noised query.

Claim: reportNoisyMax is $(\varepsilon,\, 0)$-differentially private; and accurate

DP Mechanisms

- Randomized Response
- Laplacian (+ gaussian) Mechanism
- Noisy Max
- Exponential Mechanism
- (Better) composition

Ok, but I wanted to use my data for a scenario where direct noise addition doesn't make sense

selecting from among discrete set of alternatives

small perturbation in outcome space could be disastrous for outcome quality

The Exponential Mechanism

- A mechanism $M: \mathbb{N}^{|X|} \to R$ for some abstract range R.
  - i.e. $R = \{\text{Red, Blue, Green, Brown, Purple}\}$
  - $R = \{\$1.00, \$1.01, \$1.02, \$1.03, \ldots\}$
- Paired with a *quality score:*

$$q: \mathbb{N}^{|X|} \times R \to \mathbb{R}$$

$q(D, r)$ represents how good output r is for database D.

The Exponential Mechanism

- Relative parameters for privacy, solution quality:
  - Sensitivity of q:
    $$GS(q) = \max_{r \in R, D, D' : ||D - D'||_1 \leq 1} |q(D, r) - q(D', r)|$$
  - Size and structure of R.
    - How many elements of R are high quality? How many are low quality?

The Exponential Mechanism

Exponential$(D, R, q: \mathbb{N}^{|X|} \rightarrow R, \epsilon)$:

1. Let $\Delta = GS(q)$.
2. Output $r \sim R$ with probability proportional to:

$$\Pr[r] \sim \exp\left(\frac{\epsilon q(D,r)}{2\Delta}\right)$$

$$\Pr[r] = \frac{\exp(\frac{\epsilon q(D,r)}{2\Delta})}{\sum_{r' \in R} \exp(\frac{\epsilon q(D,r')}{2\Delta})}$$

The Exponential Mechanism

$\text{Exponential}(D, R, q: \mathbb{N}^{|X|} \to R, \epsilon)$:

1. Let $\Delta = GS(q)$.
2. Output $r \sim R$ with probability proportional to:

$$\Pr[r] \sim \exp\left(\frac{\epsilon q(D, r)}{2\Delta}\right)$$

Idea: Make high quality outputs exponentially more likely at a rate that depends on the sensitivity of the quality score (and the privacy parameter)

Thm. The exponential mechanism preserves $(\varepsilon, 0)$-differential privacy.

The Exponential Mechanism

$\text{Exponential}(D, R, q: \mathbb{N}^{|X|} \rightarrow R, \epsilon)$:

1. Let $\Delta = GS(q)$.
2. Output $r \sim R$ with probability proportional to:

$$\Pr[r] \sim \exp\left(\frac{\epsilon q(D, r)}{2\Delta}\right)$$

But is the answer any good?

The Exponential Mechanism

$\text{Exponential}(D, R, q: \mathbb{N}^{|X|} \rightarrow R, \epsilon)$:

1. Let $\Delta = GS(q)$.
2. Output $r \sim R$ with probability proportional to:

$$\Pr[r] \sim \exp\left(\frac{\epsilon q(D,r)}{2\Delta}\right)$$

But is the answer any good?

It depends…

The Exponential Mechanism

Define:

$$OPT_q(D) = \max_{r \in R} q(D, r)$$

$$R_{OPT} = \{r \in R : q(D, r) = OPT_q(D)\}$$

$$r^* = \text{Exponential}(D, R, q, \epsilon)$$

output of exponential mech

Theorem:

$$\Pr\left[q(r^*) \leq OPT_q(D) - \frac{2\Delta}{\epsilon}\left(\log\left(\frac{|R|}{|R_{OPT}|}\right) + t\right)\right] \leq e^{-t}$$

The Exponential Mechanism

Theorem:

$$\Pr\left[q(r^*) \leq OPT_q(D) - \frac{2\Delta}{\epsilon}\left(\log\left(\frac{|R|}{|R_{OPT}|}\right) + t\right)\right] \leq e^{-t}$$

Corollary:

$$\Pr\left[q(r^*) \leq OPT_q(D) - \frac{2\Delta}{\epsilon}(\log(|R|) + t)\right] \leq e^{-t}$$

Proof:

$|R_{OPT}| \geq 1$ by definition.

Private PAC Learning (using Exponential Mech)

Labelled example : $(x, y) \in \mathcal{X} \times \{0,1\}$

Let $\mathcal{D}$ be a distribution over labelled examples.

Algorithm A PAC Learns a class of functions C (over d dimensions, so $x \in \{0,1\}^d$) if $\forall \alpha, \beta > 0$ $\exists m = \text{poly}(d, 1/\alpha, \log(1/\beta))$ s.t. for every distribution $\mathcal{D}$, A takes m labelled examples D from $\mathcal{D}$, and outputs $f \in C$ such that with prob $\geq 1 - \beta$

$$err(f, D) \leq \min_{f^* \in C} err(f^*, \mathcal{D}) + \alpha$$

Private PAC Learning (using Exponential Mech)

Labelled example : $(x, y) \in \mathcal{X} \times \{0,1\}$

Let $\mathcal{D}$ be a distribution over labelled examples.

Algorithm A PAC Learns a class of functions C (over d dimensions, so $x \in \{0,1\}^d$) if $\forall \alpha, \beta > 0$ $\exists m = \mathrm{poly}(d, 1/\alpha, \log(1/\beta))$ s.t. for every distribution $\mathcal{D}$, A takes m labelled examples D from $\mathcal{D}$, and outputs $f \in C$ such that with prob $\geq 1 - \beta$

$$\underbrace{\mathrm{err}(f, D)}_{\frac{1}{|D|}\left|\{(x,y) \in D \mid f(x) \neq y\}\right|} \leq \min_{f^* \in C} \underbrace{\mathrm{err}(f^*, \mathcal{D})}_{\Pr_{(x,y) \sim \mathcal{D}}[f(x) \neq y]} + \alpha$$

Private PAC Learning

Now A is randomized. Takes m samples, D, from $\mathcal{D}$. Should output $f \in C$

differential privacy: $\forall$ Neighboring D, D'

$$\Pr[A(D) = f] \approx \Pr[A(D') = f]$$

Q: How many additional samples are required to privately learn?

Private PAC Learning

(1) Use exponential mechanum: $R = C$

$$q(D, f) = -\frac{1}{|D|} \left|\{(x,y) \in D \mid f(x) \neq y\}\right|. \quad \text{sensitivity} : \frac{1}{m}$$

with high prob. exponential mech returns some $f \in C$ s.t.

$$err(f, D) \leq \min_{f^* \in C} err(f^*, D) + O\left(\frac{\log|C|}{\varepsilon m}\right)$$

Private PAC Learning

① Use exponential mechanism: $R = C$

$$q(D, f) = -\frac{1}{|D|} \left|\{(x,y) \in D \mid f(x) \neq y\}\right|. \quad \text{sensitivity}: \frac{1}{m}$$

with high prob. exponential mech returns some $f \in C$ s.t.

$$\mathrm{err}(f, D) \le \min_{f^* \in C} \mathrm{err}(f^*, D) + O\left(\frac{\log|C|}{\varepsilon m}\right) \leftarrow m \ge \frac{\log|C|}{\varepsilon\alpha}$$

② generalization $\forall f \in C$:

$$\left|\mathrm{err}(f, D) - \mathrm{err}(f, \mathcal{D})\right| \le O\left(\sqrt{\frac{\log C}{m}}\right) \longleftarrow m \ge \frac{\log|C|}{\alpha^2}$$

$\therefore\ m \ge O\left(\max\left(\frac{\log|C|}{\varepsilon\alpha}, \frac{\log|C|}{\alpha^2}\right)\right)$ to get error within α of OPT

Private PAC Learning

So exponential mechanism gives private PAC learning algorithms with little increase in sample complexity!

Private PAC Learning

So exponential mechanism gives private PAC learning algorithms with little increase in sample complexity!

BAD NEWS : very inefficient

But can often do much better

DP Mechanisms

- Randomized Response
- Laplacian (+ gaussian) Mechanism
- Noisy Max
- Exponential Mechanism
- (Better) composition

Basic composition

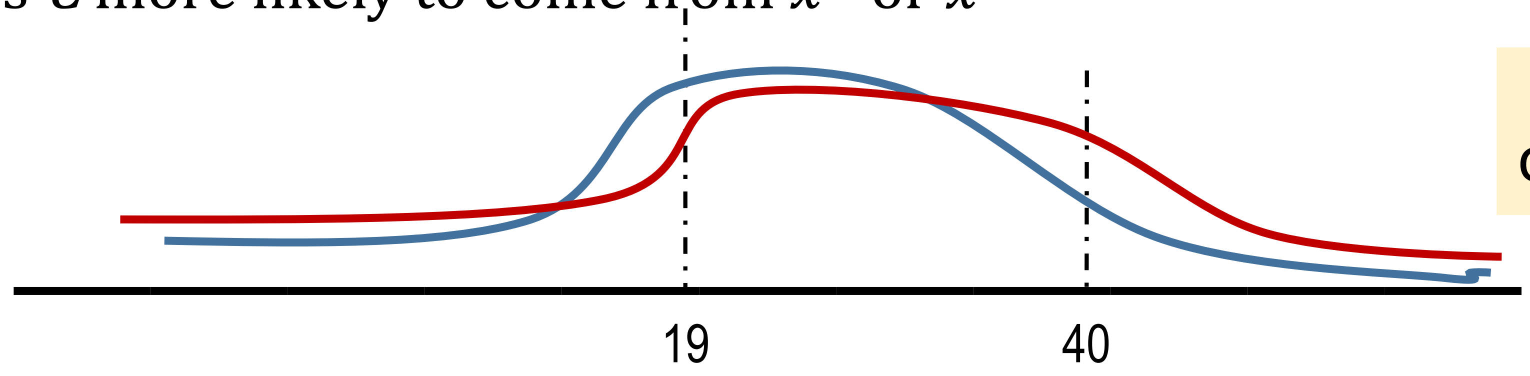
- Setting:
  - M_i be (ϵ_i, δ_i)-differentially private
  - M applies $M_1, \ldots, M_t$ on its input (the inner $M_1, \ldots, M_t$ use independent randomness).
- Basic composition theorem [DMNS06, DL09]:
  - M is $(\sum_i \epsilon_i, \sum_i \delta_i)$-differentially private

What is privacy loss?

- Measured by the 'privacy loss' parameter ϵ
- Fix adjacent x^0, x^1, draw $C \leftarrow M(x_0)$
  - Is C more likely to come from x^0 or x^1

"19" more likely as output on x^0 than on x^1

"40" more likely as output on x^1 than on x^0



- Define $Loss(C) = \ln\left[\frac{\Pr[M(x^0)=C]}{\Pr[M(x^1)=C]}\right]$
  - $(\varepsilon, 0) - DP: w.p. 1\ over\ C,\ |Loss(C)| \leq \varepsilon$
  - $(\varepsilon, \delta) - DP^{*}: w.p. 1 - \delta\ over\ C,\ |Loss(C)| \leq \varepsilon$

Log of likelihood ratio

What is privacy loss?

- Fix adjacent x^0, x^1, draw $C \leftarrow M(x_0)$

$$Loss(C) = \ln\left[\frac{\Pr\left[M\left(x^0\right) = C\right]}{\Pr[M(x^1) = C]}\right]$$

- In multiple independent executions *loss* accumulates
  - Worst case: *Loss*= ε for every execution (as in analysis of basic composition)
  - This is pessimistic: *Loss* can be positive, negative → cancellations
  - Random variable, has a mean ([DDN03, DRV10]…)

Privacy Loss in Randomized Response

(general case follows similar argument)

$$RR_\varepsilon(x_i) = \begin{cases} x_i & \text{w.p. } \frac{e^\varepsilon}{e^\varepsilon+1} \\ \neg x_i & \text{wp } \frac{1}{e^\varepsilon+1} \end{cases}$$

Privacy loss

$$\ln\left[\frac{\Pr[y_i=0 \mid x_i=0]}{\Pr[y_i=0 \mid x_i=1]}\right] = \ln\left[e^\varepsilon\right] = \varepsilon$$

$$\ln\left[\frac{\Pr[y_i=0 \mid x_i=1]}{\Pr[y_i=0 \mid x_i=0]}\right] = \ln\left[e^{-\varepsilon}\right] = -\varepsilon$$

$c_i \approx$ privacy loss of step i

so $\quad -\varepsilon \le c_i \le \varepsilon$

Privacy Loss in Randomized Response

So $-\varepsilon \leq c_i \leq \varepsilon$

$$E[c_i] = \varepsilon \cdot \frac{e^{\varepsilon}}{e^{\varepsilon}+1} - \varepsilon\left[\frac{1}{e^{\varepsilon}+1}\right] \approx \frac{\varepsilon(1+\varepsilon-1)}{e^{\varepsilon}+1} \sim \varepsilon^2$$

so $E\left[\sum_{i=1}^{k} c_i\right] = \sum_{i=1}^{k} E[c_i] \sim k \cdot \varepsilon^2$

$\therefore$ Expected cumulative loss $E\left[\sum c_i\right] \sim k\varepsilon^2$

and $\left|\sum_{i=1}^{j+1} c_i - \sum_{i=1}^{j} c_i\right| \leq \varepsilon$

So this is a Martingale

Azuma's Inequality

Let $c_1, c_2, .. c_k$ be real valued r.v.'s satisfying this ε-Lipshitz property: $\forall j$

$$\left| \sum_{i=1}^{j+1} c_i - \sum_{i=1}^{j} c_i \right| \leq \varepsilon$$

Then $\forall t \geq 0$

$$\Pr\left[\sum_{i=1}^{k} c_i \geq E\left[\sum_{i=1}^{k} c_i\right] + t \right] \leq 2^{-t^2/2k\varepsilon^2}$$

Azuma's Inequality

Let $c_1, c_2, .. c_k$ be real valued r.v.'s satisfying this ε-Lipshitz property: $\forall j$

$$\left| \sum_{i=1}^{j+1} c_i - \sum_{i=1}^{j} c_i \right| \leq \varepsilon$$

Then $\forall t \geq 0$

$$\Pr\left[\sum_{i=1}^{k} c_i \geq E\left[\sum_{i=1}^{k} c_i\right] + t \right] \leq 2^{-t^2/2k\varepsilon^2}$$

We have $E\left[\sum_{i=1}^{k} c_i\right] \sim k\varepsilon^2$

choose $t \approx \sqrt{k \log \frac{1}{\delta}}\ \varepsilon$ gives

$$\Pr\left[\sum_{i=1}^{k} c_i \geq \underbrace{k\varepsilon^2 + \sqrt{k \log \tfrac{1}{\delta}} \cdot \varepsilon}_{\varepsilon'}\right] \leq \delta$$

So we have (ε', δ)-dp

Advanced Composition [DRV10]

Composing k pure-DP algorithms (each ε_0-DP):

$$\varepsilon_g = O\left(\sqrt{k \cdot \ln \frac{1}{\delta_g}} \cdot \varepsilon_0 + k \cdot \varepsilon_0^2\right)$$ with all but δ_g probability.

For all δ_g simultaneously

Dominant if $k \ll \frac{1}{\epsilon_0^2}$

Dominant if $k \gg \frac{1}{\epsilon_0^2}$

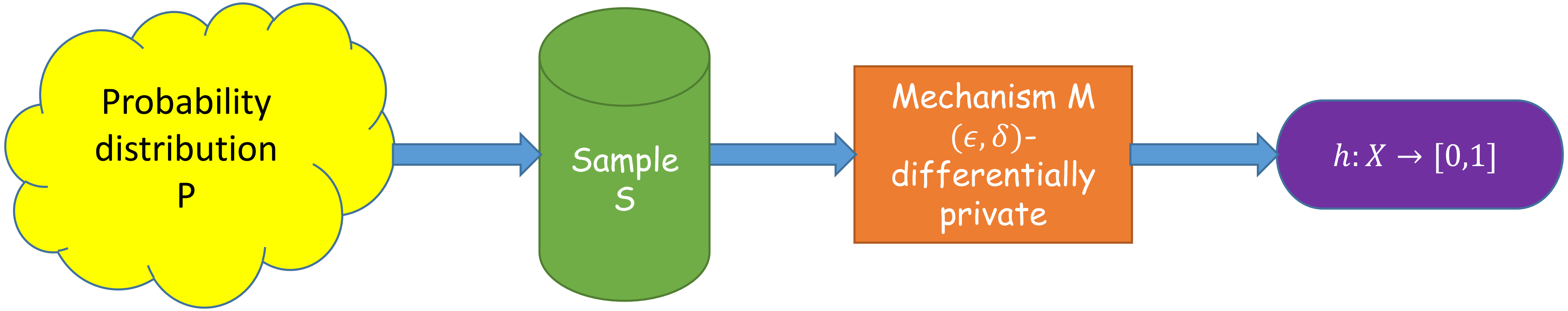
DP $\Rightarrow$ generalization

What is generalization?

Say we train a model on training set X,
where $X = n$ labelled examples
ie. X_i : choose $u \sim P$, $X_i = (u, f(u))$

If model is accurate on X, then we want to conclude model is accurate on whole distribution

Differential privacy → generalization "on average"



- Intuition: "Overfitting is a common enemy"
- Theorem [McSherry, folklore]: $\left|\mathbb{E}[h(S)] - \mathbb{E}[h(P)]\right| \leq \epsilon + \delta$

Differential privacy → generalization "on average"

- Theorem: $\left|\mathbb{E}[h(S)] - \mathbb{E}[h(P)]\right| \leq 2\epsilon + \delta$
- Proof:

$$\mathbb{E}[h(S)] = \mathop{\mathbb{E}}_{S\sim P}\ \mathop{\mathbb{E}}_{h\leftarrow M(S)}[h(S)]$$

$$= \mathop{\mathbb{E}}_{S\sim P}\ \mathop{\mathbb{E}}_{h\leftarrow M(S)}\ \mathop{\mathbb{E}}_{i\in_R[n]}[h(x_i)]$$

(reorder expectations)

$$= \mathop{\mathbb{E}}_{S\sim P}\ \mathop{\mathbb{E}}_{i\in_R[n]}\ \mathop{\mathbb{E}}_{h\leftarrow M(S)}[h(x_i)]$$

(consider M' that takes output of M and applies it on x_i, then apply proposition)

$$\leq \mathop{\mathbb{E}}_{S\sim P}\ \mathop{\mathbb{E}}_{i\in_R[n]}\left[e^{\epsilon} \mathop{\mathbb{E}}_{z\sim P;\, h\leftarrow M(S\setminus\{x_i\}\cup\{z\})}[h(x_i)] + \delta\right]$$

(rename z and x_i as $(S,z) \equiv (S \setminus \{x_i\} \cup \{z\}, x_i)$)

$$= \mathop{\mathbb{E}}_{S\sim P}\ \mathop{\mathbb{E}}_{i\in_R[n]}\left[e^{\epsilon} \mathop{\mathbb{E}}_{z\sim P;\, h\leftarrow M(S)}[h(z)] + \delta\right]$$

($\mathop{\mathbb{E}}_{z\sim P}[h(z)] = h(P)$)

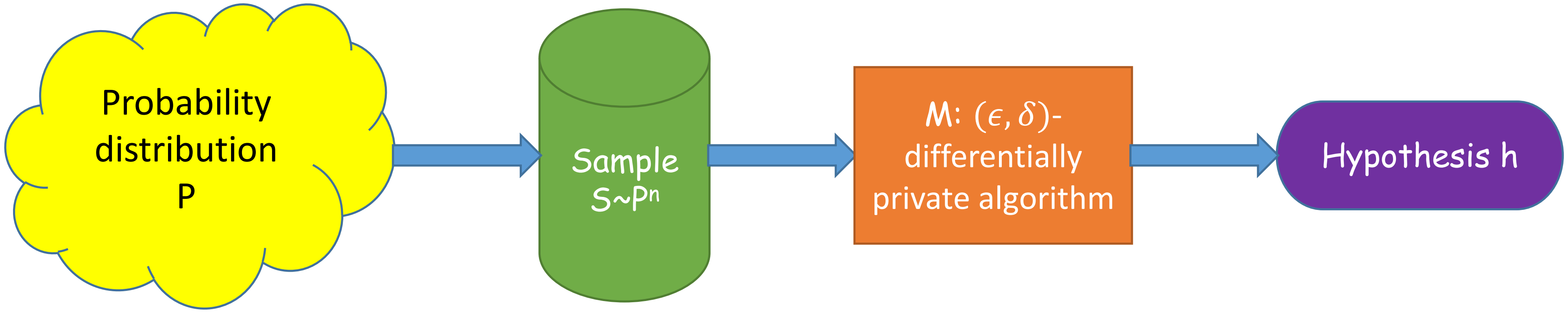
$$= e^{\epsilon} \mathop{\mathbb{E}}_{S\sim P}\ \mathop{\mathbb{E}}_{h\leftarrow M(S)} h(P) + \delta$$

($e^{\epsilon} \leq 1 + 2\epsilon$ for $\epsilon < 1$)

$$= \mathop{\mathbb{E}}_{S\sim P}\ \mathop{\mathbb{E}}_{h\leftarrow M(S)} h(P) + 2\epsilon + \delta$$

(for other direction: let $h'(x) = 1 - h(x)$)

Differential privacy → generalization (summary)



- Define: $h(S) = \frac{1}{n} \sum h(s_i)$ and $h(P) = \Pr_{s \sim P}[h(s)]$

Theorem [McSherry, folklore]:	$\mathbb{E}_{\substack{S \sim P \\ h \leftarrow M(S)}}[h(S)] \approx \mathbb{E}_{\substack{S \sim P \\ h \leftarrow M(S)}}[h(P)]$
Theorem [DFHPRR’15]:	$\Pr_{\substack{S \sim P \\ h \leftarrow M(S)}}[\lvert h(S) - h(P) \rvert > \epsilon] \leq \delta^{\epsilon}$
Tight theorem [BNSSSU’16] $(n \geq O(\frac{\ln \frac{1}{\delta}}{\epsilon^2}))$:	$\Pr_{\substack{S \sim P \\ h \leftarrow M(S)}}[\lvert h(S) - h(P) \rvert > \epsilon] \leq \delta / \epsilon$

$\underbrace{\text{first row}}$: Expectation

$\underbrace{\text{second and third rows}}$: High probability

Application to adaptive querying

- Differential privacy closed under post processing
  - Robust generalization: further post-processing unlikely to generate a non-generalizing hypothesis!
  - In standard learning, a model (that generalizes) may inadvertently reveal the sample, and hence lead to a non-generalizing hypothesis!
- Differential privacy closed under adaptive composition
  - [DFHPRR'15]: Even adaptive querying with differential privacy would not lead to a non-generalizing hypothesis

SUMMARY

Many DP mechanisms that can be mixed & matched:

- Laplace, gaussian
- Sparse Vector
- Subsampling
- Advanced composition
- Exponential Mechanism

SUMMARY

Many DP mechanisms that can be mixed & matched:

- Laplace, gaussian
- Sparse Vector
- Subsampling
- Advanced composition
- Exponential Mechanism

DP connected to generalization in ML
and to hypothesis testing

Next Class

Private Machine Learning

- Motivation,
- Theory, &
- Practice