Lecture 5 Private selection (Part II), Privacy Loss RV and Advanced Composition

Yu-Xiang Wang

COMPUTER SCIENCE
UC SANTA BARBARA
Computing. ReInvented.

A few logistic notes

- Time to think about your course project
  - I shared a list of ideas on Piazza
  - Form your own team (up to 3 people), or do an independent project.
  - Discuss your idea with me (piazza / email / in-person)!

- Scribing:
  - Scribes for lecture 1 and 2 now available.
  - Send me the latex file when you are done.

Recap: last lecture

- Sparse Vector
  - Everything in the proof works for general low-sensitivity queries (possibly non-linear queries)
- The problem of linear query release
- Private multiplicative weights
  - Use sparse vector (“NumericSparse”)
  - A cute application of a no-regret online learning method

Recap: Private Multiplicative Weight

- Learn a synthetic dataset while answering queries.
- Use SVT to check if the error of the synthetic data is large.
- Show that the number of rounds is 1/alpha^2 by the regret bound, thus after that many rounds of queries, the synthetic dataset will take over.

Summary of the problem of private query release

$|\mathcal{X}| = 2^d$

	Laplace (release query)	Laplace (release data / contingency table)	Private Multiplicative Weights
Error (normalized query)	$\frac{k \log(k/\delta)}{n\epsilon}$	$\frac{\sqrt{\lvert\mathcal{X}\rvert} \log(k/\delta)}{n\epsilon}$	$\left(\frac{\log(\lvert\mathcal{X}\rvert) \log(k/\delta)}{n\epsilon}\right)^{1/3}$ (d)
Computational complexity (per query)	$O(n)$	$O(\lvert\mathcal{X}\rvert)$	$O(\max\{\lvert\mathcal{X}\rvert, n\})$

Recap: Exponential mechanism

- Global sensitivity of the utility function

$$\Delta u \equiv \max_{r \in \mathcal{R}} \max_{x,y:\|x-y\|_1 \leq 1} |u(x,r) - u(y,r)|.$$

- The exponential mechanism samples an output from a "Gibbs distribution":

$$\mathcal{M}(x) \sim p(r|x) \propto \exp\left(\frac{\varepsilon u(x,r)}{2\Delta u}\right)$$

Proof: 1. Bound the ratio of the above exponentiated utility function (up to scale).
2. Bound the ratio of the normalization constant.

This lecture

- Privacy selection (Part II)
  - Utility of Exponential mechanism.
  - Application: SmallDB
  - ReportNoisyMax
- Advanced Composition
  - Privacy loss random variable
  - Advanced composition for pure-DP
  - Linear Query Release under Approximate DP

Readings:

- Report Noisy Max / Exponential mechanism
  - Dwork and Roth 3.3 – 3.4
- SmallDB
  - Dwork and Roth 4.1
- Advanced Composition for pure-DP
  - Vadhan 2.2. (Specifically, Lemma 2.4)

Randomized response and Laplace mechanism are instances of exponential mechanisms!

$u(x,y) = \mathbb{1}(y=x)$

$p(y|x) \propto e^{\varepsilon \mathbb{1}(y=x)}$

$\Delta u = 1$

- Randomized Response

RR outputs x w.p. $\frac{e^\varepsilon}{1+e^\varepsilon}$.

$1-x$ w.p. $\frac{1}{1+e^\varepsilon}$.

- Laplace mechanism

Lap Mech $f(x)$ Δf

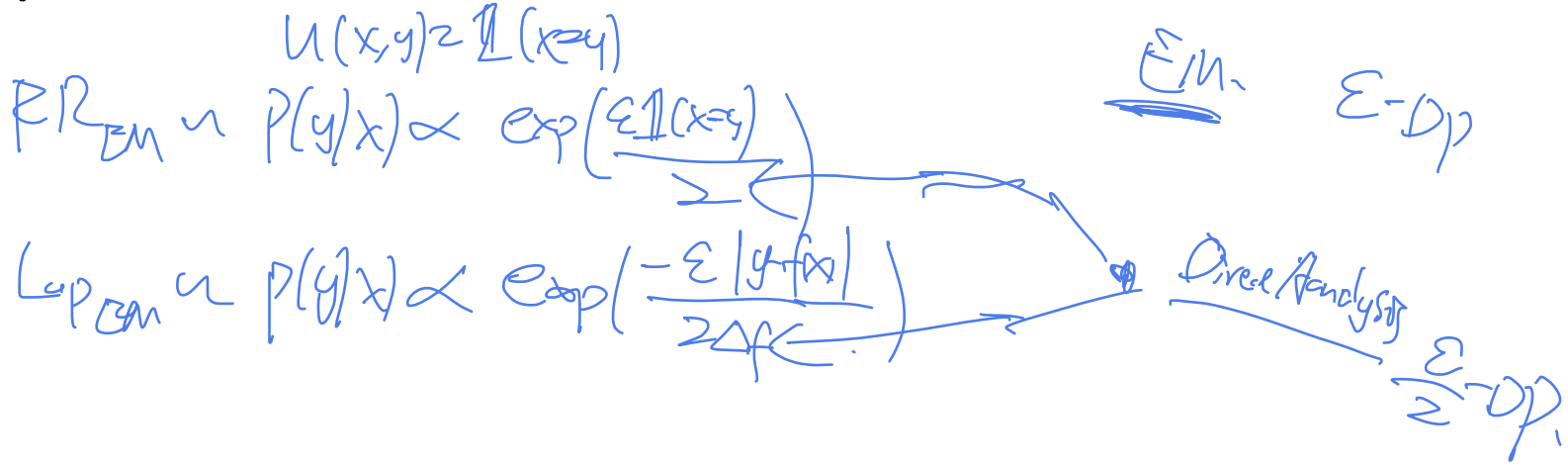
outputs $y := f(x) + \mathrm{Lap}\left(\frac{\Delta f}{\varepsilon}\right)$

$p(y|x) = \frac{\varepsilon}{2\Delta f} \exp\left(-\frac{\varepsilon |f(x)-y|}{\Delta f}\right)$

$u(x,y) = -|f(x)-y|$

$\Delta u \leq \Delta f$

Detour: Applying the results of exponential mechanism to these two.



Detour: What is really happening? There are two different types of "exponential mechanism"

- Type I ("exponential mechanism" with insensitive utility function)

$$\mathcal{M}(x) \sim p(r|x) \propto \exp(\frac{\varepsilon u(x,r)}{2\Delta u})$$

- Type II ("exponential mechanism" with insensitive log-probabilities)

$$\mathcal{M}(x) \sim p(r|x) = \exp(\frac{\varepsilon \tilde{u}(x,r)}{2\Delta \tilde{u}})$$

$\forall u$ s.t. Δu
$\exists \tilde{u}$ s.t. $\Delta\tilde{u} \leq 2\Delta u$

Recommended further reading: Dong's blog on exponential mechanism
https://dongjs.github.io/2020/02/10/ExpMech.html
Also, Durfee and Rogers (2019): https://arxiv.org/abs/1905.04273

Utility of exponential mechanism

Theorem 3.11. Fixing a database x, let $\mathcal{R}_{\mathrm{OPT}} = \{r \in \mathcal{R} : u(x,r) = \mathrm{OPT}_u(x)\}$ denote the set of elements in $\mathcal{R}$ which attain utility score $\mathrm{OPT}_u(x)$. Then:

$$\Pr\left[u(\mathcal{M}_E(x,u,\mathcal{R})) \leq \mathrm{OPT}_u(x) - \frac{2\Delta u}{\varepsilon}\left(\ln\left(\frac{|\mathcal{R}|}{|\mathcal{R}_{\mathrm{OPT}}|}\right) + t\right)\right] \leq e^{-t}$$

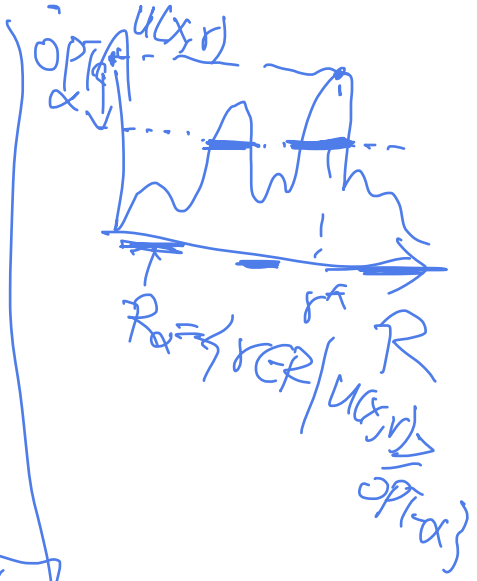
Proof:

$$\Pr[M(x) \notin R_\alpha] = \sum_{r \notin R_\alpha} P(M_{x,E} r) = \frac{\sum_{r \notin R_\alpha} \exp\left(\frac{\varepsilon u(x,r)}{2\Delta u}\right)}{\sum_{r \in R} \exp\left(\frac{\varepsilon u(x,r)}{2\Delta u}\right)}$$

$$\leq \frac{\sum_{r \notin R_\alpha} \exp\left(\frac{\varepsilon(\mathrm{OPT}-\alpha)}{2\Delta u}\right)}{\sum_{r \in R_0} \exp\left(\frac{\varepsilon \mathrm{OPT}}{2\Delta u}\right)} \leq \frac{|R|}{|R_0|} \exp\left(\frac{-\varepsilon\alpha}{2\Delta u}\right)$$

$|R_0| = |R_{\mathrm{OPT}}|$

$$\alpha = \frac{2\Delta u}{\varepsilon}\left(\ln\left(\frac{|R|}{|R_{\mathrm{OPT}}|\delta}\right)\right)$$



Applying Exponential Mechanism to (offline) Linear Query Release

- Given a fixed set of queries of size k
- Run exponential mechanism to select a dataset most consistent with the answers to these queries

Algorithm 4 The Small Database Mechanism

SmallDB$(x, \mathcal{Q}, \varepsilon, \alpha)$

Let $\mathcal{R} \leftarrow \{y \in \mathbb{N}^{|\mathcal{X}|} : \|y\|_1 = \frac{\log|\mathcal{Q}|}{\alpha^2}\}$

Let $u : \mathbb{N}^{|\mathcal{X}|} \times \mathcal{R} \to \mathbb{R}$ be defined to be:

$$u(x, y) = -\max_{f \in \mathcal{Q}} |f(x) - f(y)|$$

Sample And Output $y \in \mathcal{R}$ with the exponential mechanism $\mathcal{M}_E(x, u, \mathcal{R})$

Analyzing the smallDB mechanism

- Privacy guarantee follows from that of the exponential mechanism

- Analyzing the sensitivity $u(x, y) = -\max_{f \in Q} |f(x) - f(y)|$

Utility of the SmallDB in answering linear queries

- Notice that we are restricting the sample size
  - If $\log |Q| / \alpha^2 < n$, we need to work out the optimal solution (even if we output argmin in the clear)

- Claim: There always exists a smallDB that is α accurate.
  - Idea: randomly sample the dataset (with replacement).

Apply the utility theorem of exponential mechanism

$\Delta u = \frac{1}{n}$

w.p $1-\delta$ $u(x, M(x)) \geq u(x, r^x) - \frac{2\Delta u}{\varepsilon} \log\left(\frac{|R|}{\delta}\right)$

$= \frac{2}{n\varepsilon} \log\left(\frac{|R|}{\delta}\right)$

$u(x, M(x)) \geq OPT - OPT + u(x, r^f) - \frac{2}{n\varepsilon} \log\left(\frac{|R|}{\delta}\right)$

$- \alpha$

$|R| = |X|^m = |X|^{\frac{\log|Q|}{\alpha^2}}$

$- \frac{2}{n\varepsilon} \frac{\log|Q|}{\alpha^2} \log|X|$

$\max_{f \in Q} |f(M(x)) - f(x)| \leq \alpha + \frac{2}{n\varepsilon\alpha^2} \log|Q| \log|X|$

$\alpha = O\left(\sqrt{\frac{\log|Q| \log|X|}{n\varepsilon}}\right)$

$1 - \beta - \frac{4}{n\varepsilon\alpha^3} \log|Q| \log|X| = 0$

Checkpoint: SmallDB vs Private Multiplicative Weights

- Both achieve the same asymptotic error
- MW also works for an online sequence of adaptively chosen query
- Neither is computationally efficient

Alternative algorithm for private selection: ReportNoisyMax

- For each $r \in \mathcal{R}$

$$\hat{u}(r, x) = u(r, x) + \mathrm{Lap}(2\Delta u/\epsilon)$$

- Output $\arg\max_{r \in \mathcal{R}} \hat{u}(r, x)$

Privacy analysis is similar to that of SVT

- What is the output space?

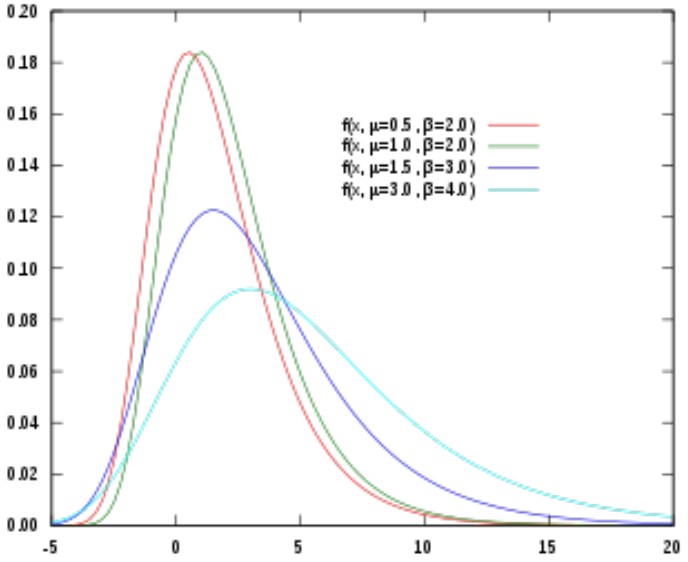
- When does the algorithm output r?

- The same trick of change of variable applies.

Remarks about RNM

1. If u(x,r) is a count for each r, then you can improve RNM by a factor of 2
   - More generally, it applies when u(x,r) is monotonic.
   - Similar proof.

2. You can also add other noise
   - Add (one sided) exponential noise
   - Add Gumbel noise



Application: Private voting (One vote per person)

Trump, Biden, Kanye

- One vote one person

- Two ways of releasing the results
  - Privately publish the histogram with Laplace mechanism
  - Privately publish the winner with ReportNoisyMax

Both ε-DP

Application: Private voting (Vote for as many people as you like!)

- Two ways of releasing the results
  - Privately publish the histogram with Laplace mechanism

  - Privately publish the winner with ReportNoisyMax

Remainder of today's lecture

- Advanced composition
  - Privacy loss random variable
  - Prove advanced composition for pure-DP
  - Linear Query Release under Approximate DP

(ε, δ)-DP

$\delta > 0$

Recall: Summary of the problem of private query release

	Laplace (release query)	Laplace (release data / contingency table)	Private Multiplicative Weights
Error (normalized query)	$\frac{k\log(k/\delta)}{n\epsilon}$	$\frac{\sqrt{\lvert\mathcal{X}\rvert}\log(k/\delta)}{n\epsilon}$	$\left(\frac{\log(\lvert\mathcal{X}\rvert)\log(k/\delta)}{n\epsilon}\right)^{1/3}$
Computational complexity (per query)	$O(n)$	$O(\lvert\mathcal{X}\rvert)$	$O(\max\{\lvert\mathcal{X}\rvert, n\})$

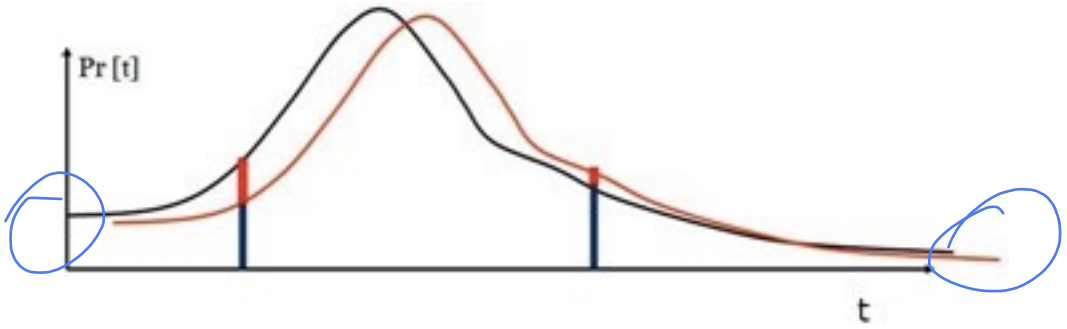
Can we do better under (ε,δ)-DP?

Recap: Approximate DP

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.



Advanced Composition

$(M_1, M_2, \cdots, M_k)$

Theorem: The adaptive composition of k (ε,δ)-DP mechanisms satisfies $(\tilde{\varepsilon}, \tilde{\delta}) - DP$ where

$$\tilde{\varepsilon} = \varepsilon\sqrt{2k\log(1/\delta')} + 2k\varepsilon^2, \qquad \tilde{\delta} = k\delta + \delta'$$

for any $\varepsilon, \delta \geq 0, \delta' > 0$

Simple Composition: $\tilde{\varepsilon} = k\varepsilon$ $\quad \tilde{\delta} = k\delta$

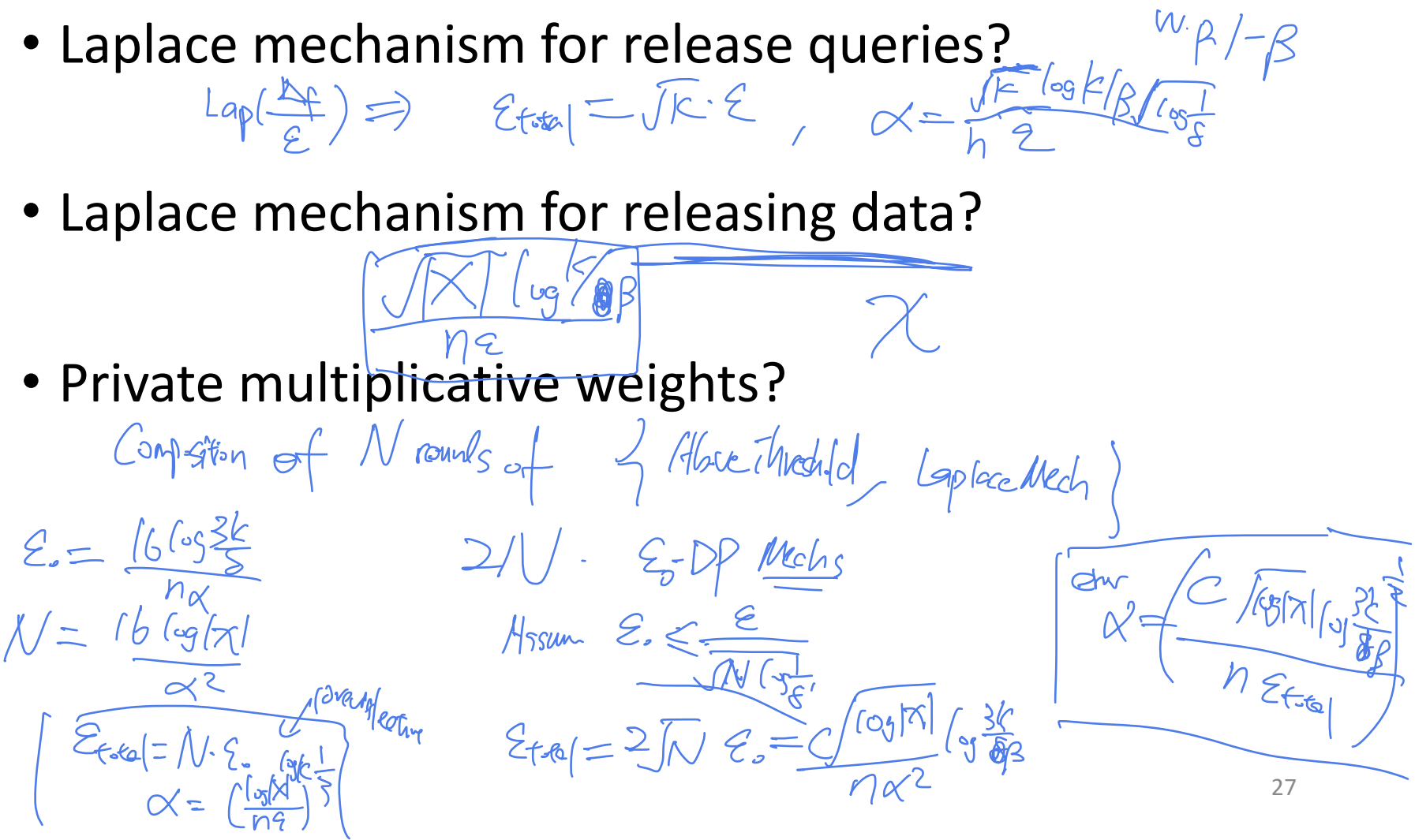
$\varepsilon = \frac{\varepsilon'}{\sqrt{2k\log(1/\delta')}}$ $\quad \delta = 0$

$\tilde{\varepsilon} = \varepsilon' + \frac{2k(\varepsilon')^2}{\left(\sqrt{2k\log(1/\delta')}\right)^2} = \varepsilon' + \frac{(\varepsilon')^2}{\log\frac{1}{\delta'}} \leq 2\varepsilon'$

where $\varepsilon' \leq \sqrt{\log\frac{2}{\delta'}}$

Application to linear query release

- Laplace mechanism for release queries?
- Laplace mechanism for releasing data?
- Private multiplicative weights?



Privacy loss random variable

- PLRV is the log probability ratio as a random variable

$$\boldsymbol{\varepsilon}_{\mathcal{M}}^{x,x'} = \log\left(\frac{p(\mathbf{y})}{p'(\mathbf{y})}\right) \text{ where random variable } \mathbf{y} \sim \mathcal{M}(x).$$

- Tail bound of privacy loss r.v. implies DP

Lemma 1 (Tail bound to (ϵ, δ)-DP conversion)**.** *Let* $\boldsymbol{\varepsilon}_{\mathcal{M}}^{x,x'}$ *be the privacy loss RV defined above. If*

$$\mathbb{P}(\boldsymbol{\varepsilon}_{\mathcal{M}}^{x,x'} > \epsilon) \leq \delta$$

for all *pair of neighboring* x, x' *then* $\mathcal{M}$ *satisfies* (ϵ, δ)*-DP.*

(You are to prove this in HW1.)

Two more properties of privacy loss r.v.

- Expected value of a privacy loss is KL-divergence

$$\mathbb{E}_{y\sim M(x)}\left[\log \frac{P_x(y)}{P_{x'}(y)}\right] = D_{KL}(P_x \| P_{x'}) = \int P_x(y) \log \frac{P_x(y)}{P_{x'}(y)}$$

- PLRV under composition

$(M_1, M_2, \ldots, M_k)$

$$\varepsilon_{(M_1,\ldots,M_k)}^{x,x'}(y) = \log \frac{P_{M_1(x)}(y_1)\cdot P_{M_2(x)}(y_2)\cdots P_{M_k(x)}(y_k)}{P_{M_1(x)}(y_1)\, P_{M_2(x)}(y_2)\cdots P_{M_k(x)}(y_k)} = \sum_{i=1}^{k} \varepsilon_{M_i}^{x,x'}(y_i)$$

M_i adaptively chosen. $y_1 \sim y_{i-1}$

$$\varepsilon_{M_i|_h}^{x,x'}(y) = \log \frac{P_{M_1(x)}(y_1)\, P_{M_2(x)}(y_2 \mid M_1(x)=y_1)\cdots P_{M_k(x)}(y_k \mid y_1 \sim y_{k-1})}{\text{martingale}}$$

Proof Idea of Advanced Composition

- Observation 1: sometimes PLRV is positive, other times negative. They cancel with each other
- Observation 2: as k gets larger, the sum of PLRV concentrates around its mean.
- Observation 3: the adaptivity means that the PLRV will depend on the past

Martingale

- We say that a sequence of r.v. $X_1,\ldots,X_n,\ldots$ is a Martingale if for any n

$$\mathbf{E}(|X_n|) < \infty$$

$$\mathbf{E}(X_{n+1} \mid X_1, \ldots, X_n) = X_n.$$

- Example:
  - Random-walk: Total number of heads minus tails in n coin tosses

Azuma-Hoeffding's inequality

- **Azuma-Hoeffding's inequality**: Assume X_1, …, X_n are Martingale differences

$$S_n = X_1 + ... + X_n$$

$$\mathbb{P}\left[S_n \geq \epsilon\right] \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

- Apply Azuma-Hoeffding's inequality to our problem

Proof for the advanced composition for pure DP mechanisms

- Fix x, x’, apply Azuma-Hoeffding’s inequality

Bounding the KL-divergence

- **Lemma** (Pinsker's inequality)

$$\|P - Q\|_1 \leq \sqrt{2D_{KL}(P\|Q)}$$

- **Corollary**: KL-divergence is nonnegative.
- Now's let's bound the expected value of PLRV:

Next lecture

- Gaussian mechanism
- CDP and Renyi DP
- Composition of Gaussian mechanism