Lecture 7 Modern Tools for Privacy Accounting

Yu-Xiang Wang

COMPUTER SCIENCE
UC SANTA BARBARA
Computing. ReInvented.

Recap: last lecture

- Advanced Composition
- Gaussian mechanism
- Concentration inequalities
- Renyi Differential Privacy

Recap: Advanced Composition

- Martingale concentration of privacy loss random variables
- Bound mean: KL-divergence of two distributions with bounded ratio
- Bound deviation from the mean: O(sqrt{k})

Recap: Gaussian mechanism

$f(x) + N(0, I_d \sigma^2)$

- The privacy loss RV of Gaussian mechanism is Gaussian.

$N(y; \frac{\Delta^2}{2\sigma^2})$ where $y = \frac{\Delta^2}{\sigma^2}$ $\Delta = \|f(x) - f(x')\|_2$

- Gaussian mechanism is better than the Laplace when
  - its L2 sensitivity is substantially smaller than L1
  - we compose over many rounds

Recap: Concentration Inequalities

- From Markov to Chernoff

$$P(X>t) = P(e^{uX} > e^{ut}) \leq \frac{\mathbb{E}(e^{uX})}{e^{ut}}$$

- Concentration of subgaussian random variables

- Idea:
  - Gaussian mechanism is quite special, with its PLRV being Gaussian.
  - Are there other DP mechanisms with their privacy loss r.v.'s tails behaving like that of the Gaussian mechanism?

Recap: Concentrated DP and Renyi DP

- Renyi Divergence:
  - Basically a transformation of MGF

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log \underbrace{E_Q\left(\frac{P}{Q}\right)^\alpha}_{E_Q e^{\alpha \log \frac{P}{Q}}} = E_P\left(\frac{P}{Q}\right)^{\alpha-1}$$

$$\| \quad E_P e^{(\alpha-1)\log\frac{P}{Q}}$$

- Formalizing the idea of subgaussian PLRV:
  - zCDP: A linear upper bound of the Renyi divergence
  - Renyi DP: A pointwise upper bound of the Renyi divergence

$\forall P=N(x)\ Q=N(x')$

$D_\alpha(P\|Q) \le \rho\cdot\alpha$

x, x' are neighbors

$$D_\alpha(P\|Q) \le \varepsilon(\alpha)$$

$\alpha, \varepsilon(\alpha)$-RDP

- Advanced composition for Gaussian mechanism

This lecture

- Equivalent definitions of approximate DP (ε, δ 1-DP)
  - Hockey-Stick divergence / privacy profiles
  - Tradeoff functions / f-DP
  - Difference of the CDFs
- Application: Analytical Gaussian mechanism
- Mechanism specific analysis
  - More precise representations
  - Their composition

Readings

- “Improving the Gaussian Mechanism for Differential Privacy” by Balle and W. https://arxiv.org/abs/1805.06530

- “Gaussian Differential Privacy” by Dong, Roth and Su https://arxiv.org/abs/1905.02383

- “Optimal accounting of Differential Privacy Via Characteristic Function” by Zhu, Dong and W. https://arxiv.org/abs/2106.08567

Recall: definition of 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.

- What is this set S that attains the bound?
- Can we redefine approximate DP in the same way Renyi DP is defined?

$\sup_{x,y \text{ neighbors}} D_\varepsilon(\mathcal{M}(x) \| \mathcal{M}(y)) < \infty$

Detour: f-divergence

- For any convex function $f : \mathbb{R}_+ \to \mathbb{R}$, satisfying $f(1) = 0$ you may define a divergence

$$D_f(P\|Q) = \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q$$

- When densities exists:

$$D_f(P\|Q) = \int q\, f\left(\frac{p}{q}\right) \mathrm{d}\mu$$

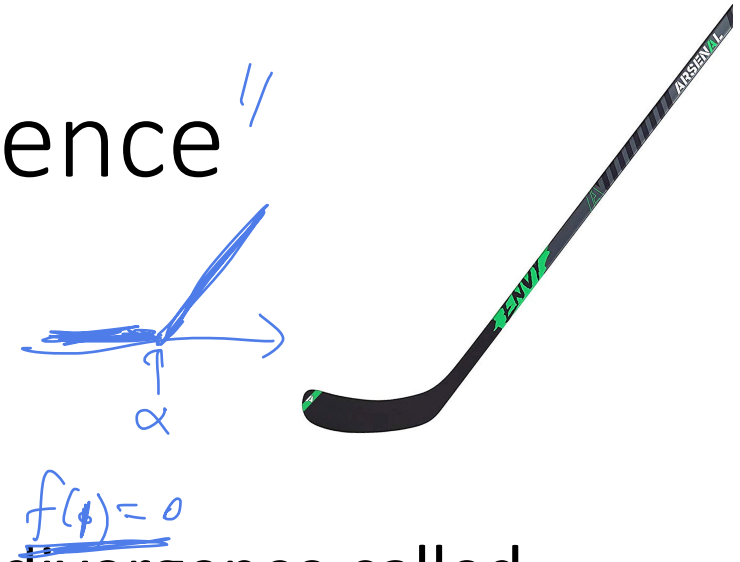
- Very general family with many nice properties:

See Sason and Verdu: https://arxiv.org/pdf/1508.00335.pdf

Detour: Example of f-divergence

- KL-divergence $f(t) = t \log t$, $\int q f(\frac{p}{q}) d\mu = \int q \frac{p}{q} \log \frac{p}{q} d\mu = E_P(\log \frac{dP}{dQ})$
- xi^2 divergence $f(t) = (t-1)^2$, $\int q (\frac{p}{q} - 1)^2 d\mu = \int q (\frac{p^2}{q^2} - 2\frac{p}{q} + 1) d\mu$
  $= \int \frac{p^2}{q} d\mu - 1 \in \chi^2$-divergence
- Total variation distance $f(t) = |t-1|$, $\int q |\frac{p}{q} - 1| d\mu = \int |p - q| d\mu = TV(P \| Q)$
- Hellinger divergence $f(t) = (\sqrt{t} - 1)^2$

Hockey-Stick divergence



- When $f(x) = (x - \alpha)_+$

- The f-divergence is a special divergence called "Hockey-Stick divergence"

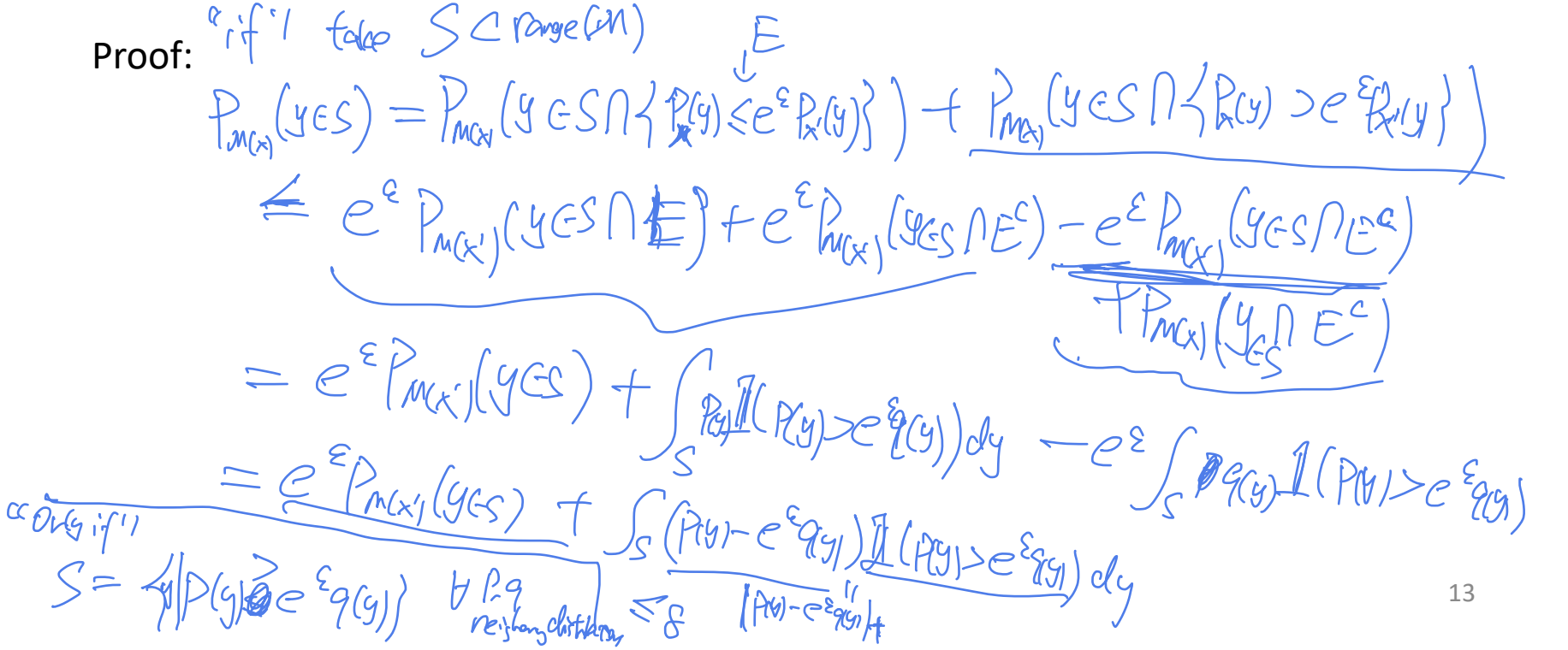
$$H_\alpha(P\|Q) := \mathbb{E}_{o\sim Q}[(\frac{\mathrm{d}P}{\mathrm{d}Q}(o) - \alpha)_+]$$

Approximate DP is equivalent to a bound of Hockey Stick divergence.

- M is (ε,δ)-DP if and only if

$$\sup_{D\simeq D'} H_{e^\epsilon}(\mathcal{M}(D)\|\mathcal{M}(D')) \le \delta.$$

Proof:



Rewriting it in terms of the privacy loss random variable

$\sup_{D \simeq D'} H_{e^\epsilon}(\mathcal{M}(D) \| \mathcal{M}(D')) \leq \delta.$

$P \| Q$

$$\int q \left(\frac{p}{q} - e^\varepsilon\right)_+ d\mu = \int (p - e^\varepsilon q)_+ d\mu = \int (p - e^\varepsilon q) \mathbb{1}(p > e^\varepsilon q)\, d\mu$$

$$= \int p\, \mathbb{1}(p > e^\varepsilon q)\, d\mu - e^\varepsilon \int q\, \mathbb{1}(p > e^\varepsilon q)\, d\mu$$

$$= P_P\left(\frac{p}{q} > e^\varepsilon\right) - e^\varepsilon P_Q\left(\frac{p}{q} > e^\varepsilon\right)$$

$$= P_P\left(\log \frac{p}{q} > \varepsilon\right) - e^\varepsilon P_Q\left(\log \frac{p}{q} > \varepsilon\right)$$

$\log \frac{q}{p} < -\varepsilon$

$$\Pr_{o \sim \mathcal{M}(D)}[L_{P,Q} > \epsilon] - e^\epsilon \Pr_{o \sim \mathcal{M}(D')}[L_{Q,P} < -\epsilon] \leq \delta$$

for all neighboring D, D'.

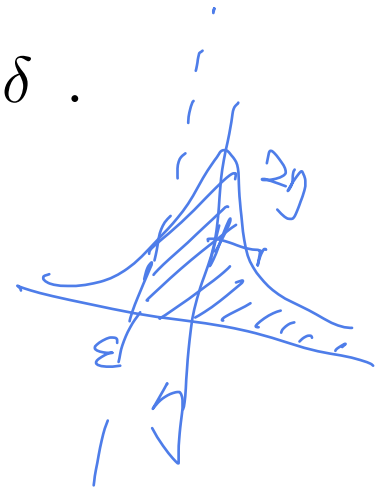
Application: Analytical Gaussian mechanism

- Recall PLRV of Gaussian mechanism is

$$\mathcal{N}(\eta, 2\eta) \ \textit{with}\ \eta = D^2/2\sigma^2, \ \textit{where}\ D = \|f(x) - f(x')\|.$$

- Applying the above

$$\mathbb{P}[L_{M,x,x'} \geq \varepsilon] - e^{\varepsilon}\mathbb{P}[L_{M,x',x} \leq -\varepsilon] \leq \delta \ .$$



- It's easy to show that

$$\mathbb{P}[L_{M,x,x'} \geq \varepsilon] = \Phi\left(\frac{D}{2\sigma} - \frac{\varepsilon\sigma}{D}\right) \ ,$$

$$\mathbb{P}[L_{M,x',x} \leq -\varepsilon] = \Phi\left(-\frac{D}{2\sigma} - \frac{\varepsilon\sigma}{D}\right) \ .$$

Need to maximize over two neighboring datasets

- Lemma: The following function is monotonically increasing in when ε>0 and η>0

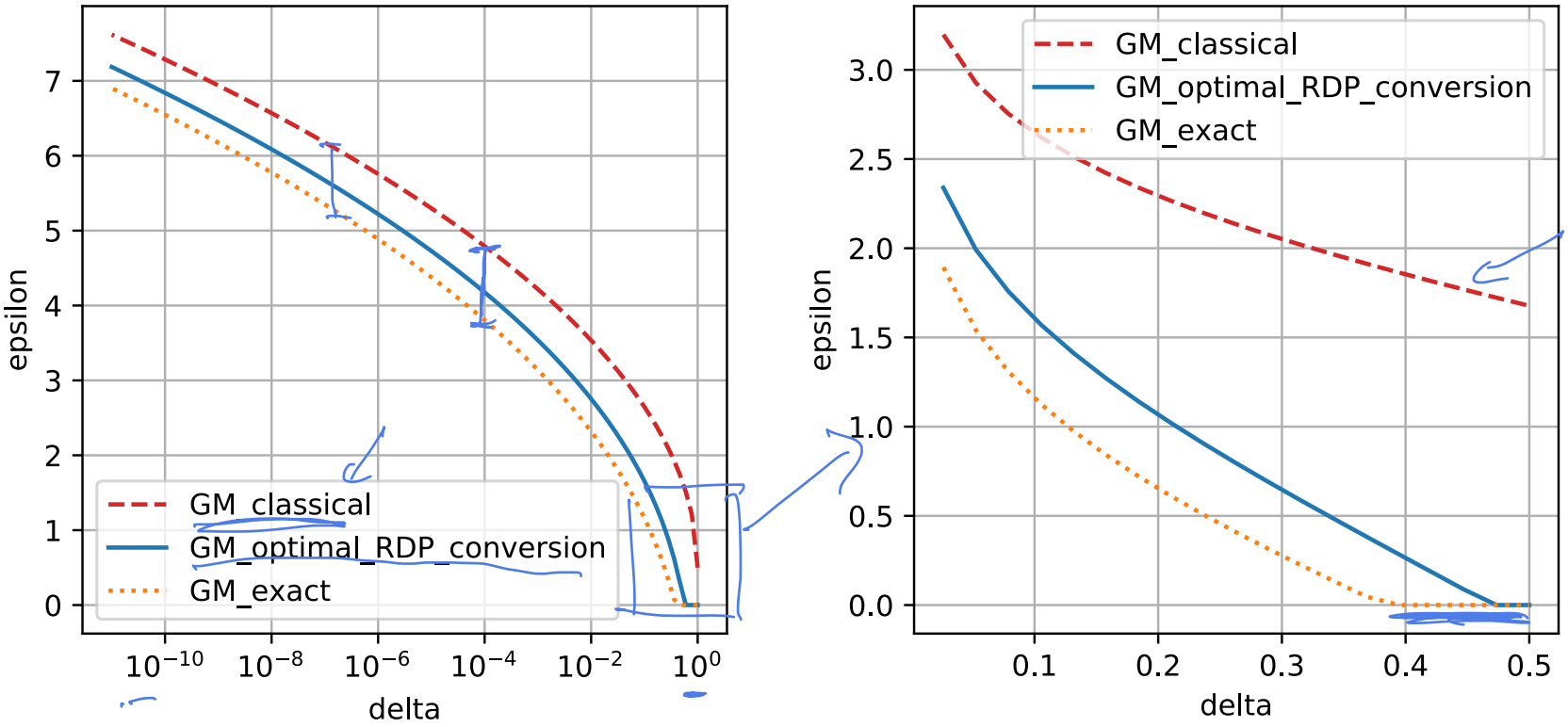
$$h(\eta) = \mathbb{P}[\mathcal{N}(\eta, 2\eta) \geq \varepsilon] - e^{\varepsilon}\mathbb{P}[\mathcal{N}(\eta, 2\eta) \leq -\varepsilon] \ .$$

- **Theorem**: Gaussian mechanism is (ε,δ)-DP if and only if

$$\Phi\left(\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) - e^{\varepsilon}\Phi\left(-\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) \leq \delta \ .$$

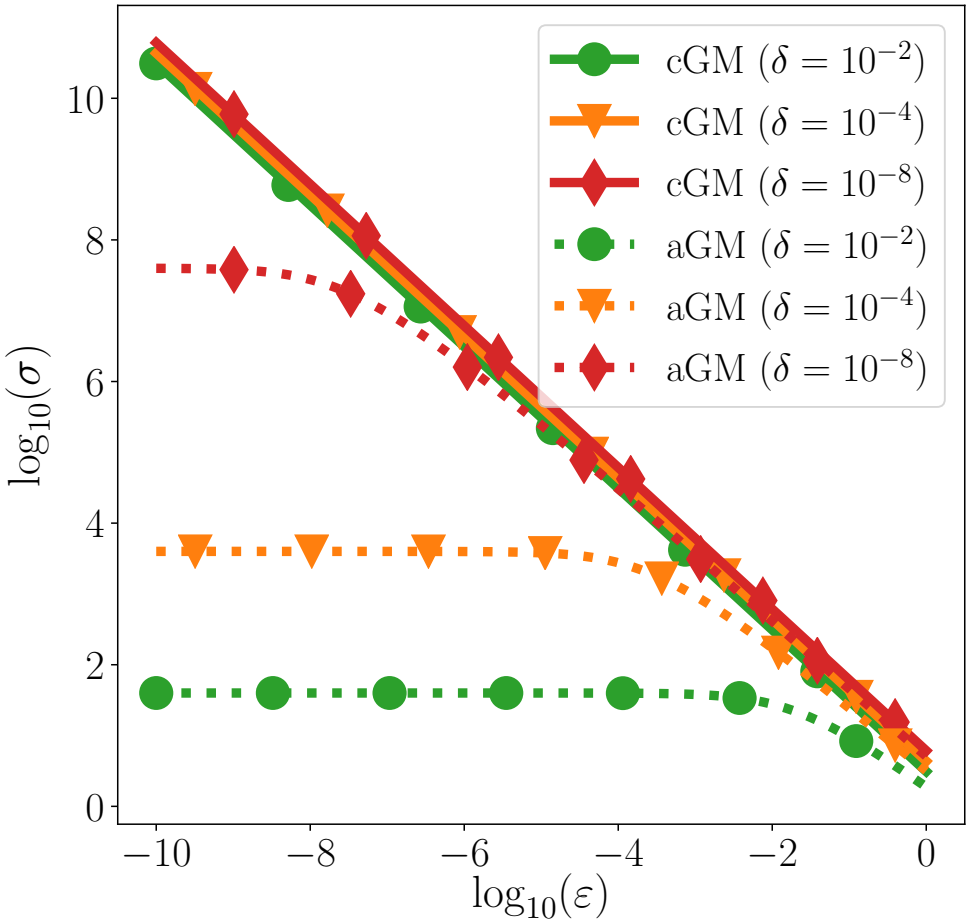
Balle and W. (2018): https://arxiv.org/pdf/1805.06530.pdf

Improvements from the Analytical Gaussian mechanism is substantial



Improvements from the Analytical Gaussian mechanism is substantial

- Calibrate noise to privacy requirements



Checkpoint 1: HS Divergence as an equivalent definition of DP

- M is (ε,δ)-DP if and only if

$$\sup_{D \simeq D'} H_{e^\epsilon}(\mathcal{M}(D) \| \mathcal{M}(D')) \leq \delta.$$

And if and only if

$$\Pr_{o \sim \mathcal{M}(D)}[L_{P,Q} > \epsilon] - e^\epsilon \Pr_{o \sim \mathcal{M}(D')}[L_{Q,P} < -\epsilon] \leq \delta$$

$$for\ all\ neighboring\ D, D'.$$

- Obtain exactly optimal Gaussian mechanism.

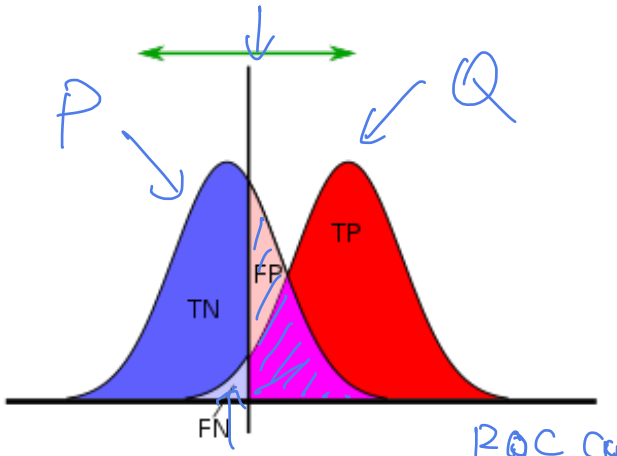
Adversary's point of view: Hypothesis testing interpretation of DP

- Suppose an attacker wants to make inference about whether Alice is in the dataset
  - Type I error (false positive rate): Probability of predicting "Yes" when "Alice is not in the data".
  - Type II error (false negative rate): Probability of predicting "No" when "Alice is in the data"

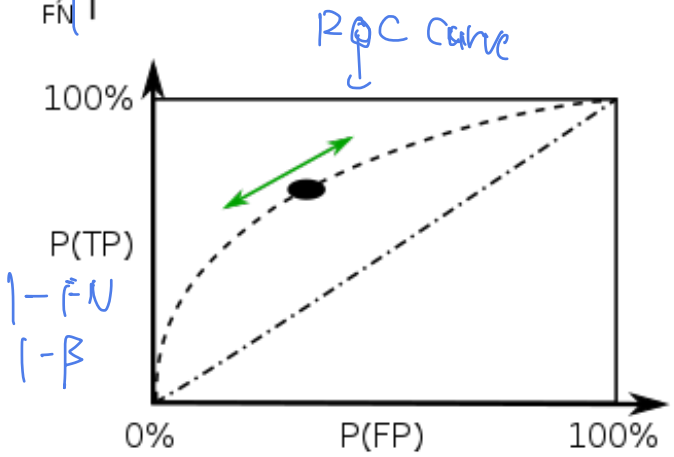
- Trade-off function (the performance of the optimal classifier)

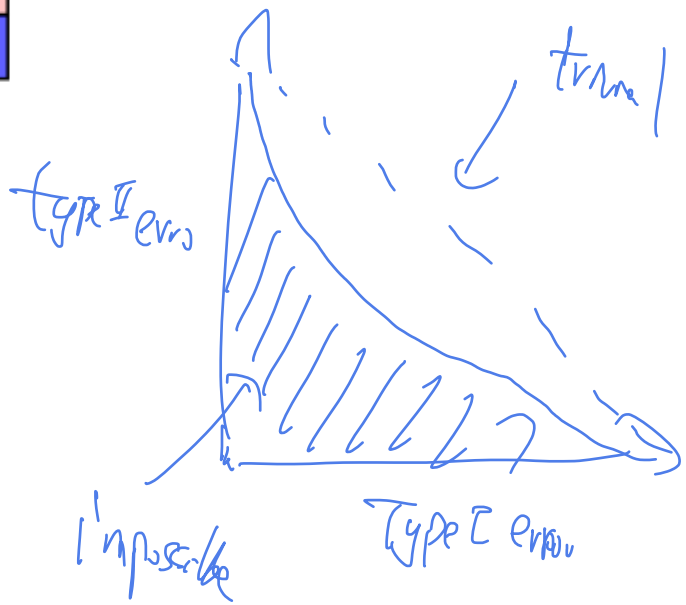
$$T_{P,Q}(\alpha) := \inf_{\phi}\{\beta_\phi \mid \alpha_\phi \leq \alpha\}$$

Illustration of the Type I vs Type II error and the tradeoff function



TP	FP
FN	TN





(Figures taken from Wikipedia https://en.wikipedia.org/wiki/Type_I_and_type_II_errors)

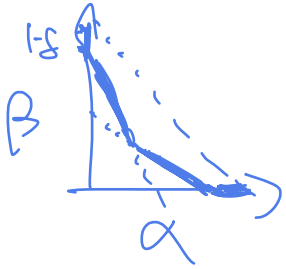
"=>" Differential Privacy implies a lower bound on the tradeoff

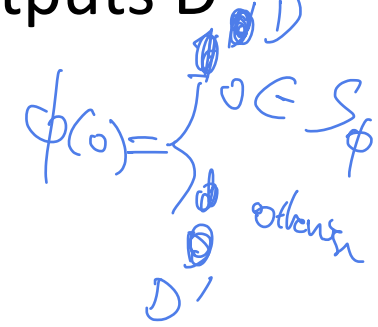
- From DP to tradeoff function

$$\mathbb{P}[\mathcal{M}(D) \in S] \leq e^{\epsilon}\mathbb{P}[\mathcal{M}(D') \in S] + \delta$$

- Take S to be the event when classifier ϕ outputs D

$$1 - \alpha_\phi \leq e^{\epsilon}\beta_\phi + \delta$$





Swap D, D'

$$\beta_\phi \geq e^{-\epsilon}\alpha_\phi(1-\alpha_\phi) - \frac{\delta}{e^{\epsilon}}$$

$$1-\beta_\phi \leq e^{\epsilon}\alpha_\phi + \delta \Rightarrow \beta_\phi \geq -e^{\epsilon}\alpha_\phi + 1 - \delta$$

"<=" Tradeoff function also implies DP

- Note that the above covers any classifier
- For each set S, define a classifier
  - $\phi(o)$ = {Predict D' if o is in S; and D otherwise}
- The tradeoff function definition implies that

$$1 - \alpha_\phi \leq e^\epsilon \beta_\phi + \delta$$

for each S, define $\phi_S = \begin{cases} D' \text{ if } o \in S \\ D \text{ otherwise} \end{cases}$

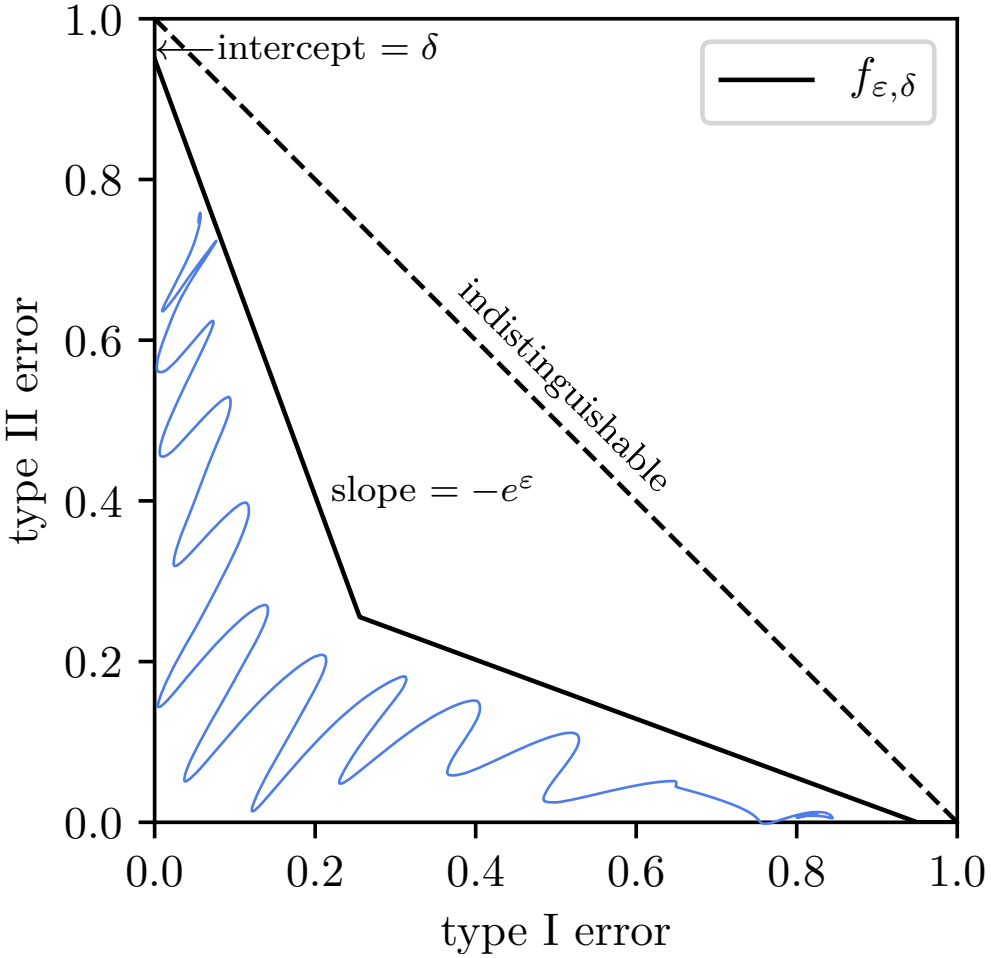
$P(M(x) \in S_\phi) = 1 - \alpha_\phi \leq e^\epsilon \beta_\phi + \delta = e^\epsilon P(M(x') \in S_\phi) + \delta$

What is the optimal classifier?

- (Neyman-Pearson) The uniform most-powerful test of two point hypotheses is the likelihood ratio test.
  - It means that the optimal test is constructed by thresholding the log-likelihood ratio.

return P if $\log \frac{dP(o)}{dQ(o)} > \text{Threshold}_\alpha$

On the figure of Type II error vs Type I error



Checkpoint 2: DP is characterized by how it prevents attackers from making accurate inference.

Theorem: M is (ε,δ)-DP if and only if

1. $\sup_{D \simeq D'} H_{e^\epsilon}(\mathcal{M}(D) \| \mathcal{M}(D')) \leq \delta.$
2. $\sup_{D \simeq D'} T_{\mathcal{M}(D),\mathcal{M}(D')}(\alpha) \geq \max\{0, 1-\delta-e^\epsilon\alpha, e^{-\epsilon}(1-\delta-\alpha)\}.$
3. $\Pr_{o\sim\mathcal{M}(D)}[L_{P,Q} > \epsilon] - e^\epsilon \Pr_{o\sim\mathcal{M}(D')}[L_{Q,P} < -\epsilon] \leq \delta$ *for all neighboring* D, D'.

How to compose over multiple rounds?

Remainder of the lecture

- Mechanism-specific analysis

- Mechanism-specific privacy accounting
  - RDP-accountant
  - f-DP accountant
  - Fourier-accountant

What makes advanced composition suboptimal?

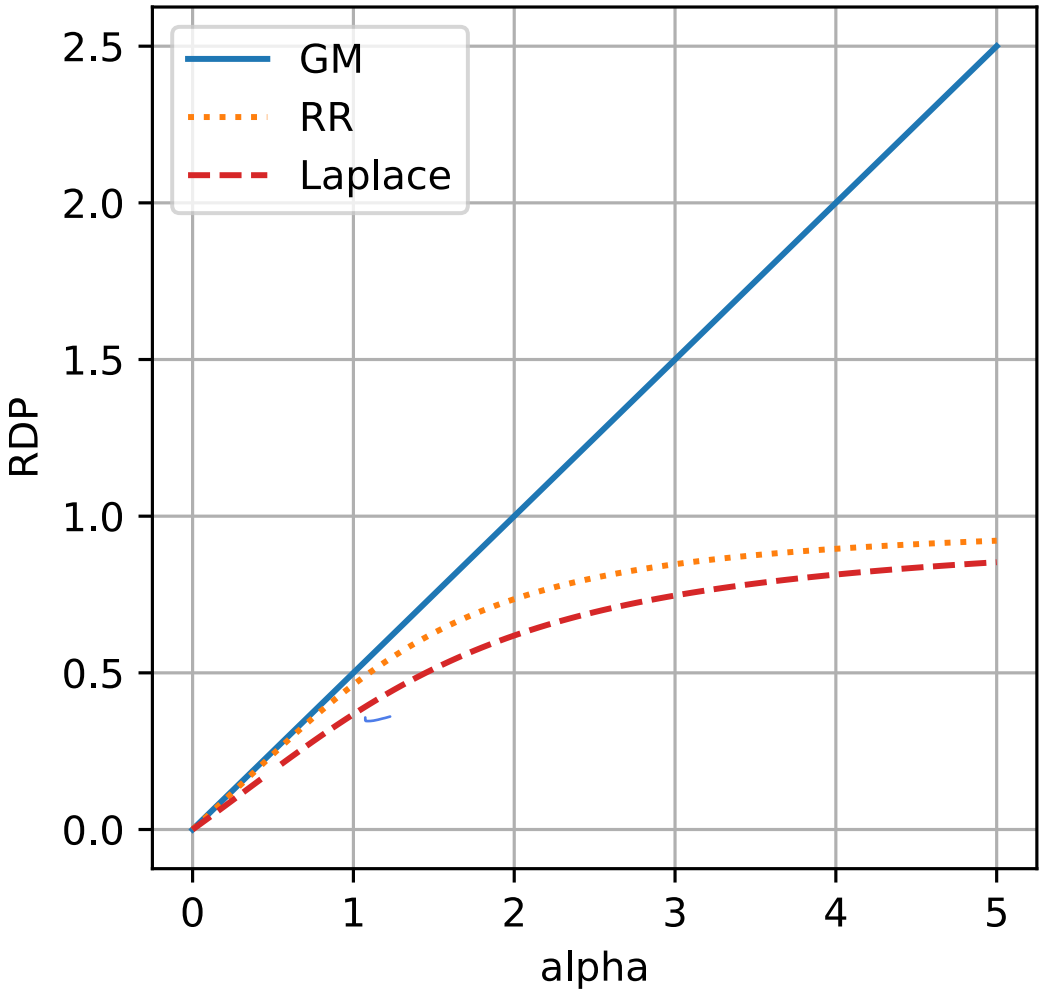
- Murtagh and Vadhan: Computing optimal composition of approximate-DP mechanisms is #P-hard.

- It is more worst-case than needed.
  - Example: composition of a sequence of Gaussian Mechanisms
  - We know the sequence of mechanisms we are using!

Mechanism specific analysis describes mechanisms by **a function**

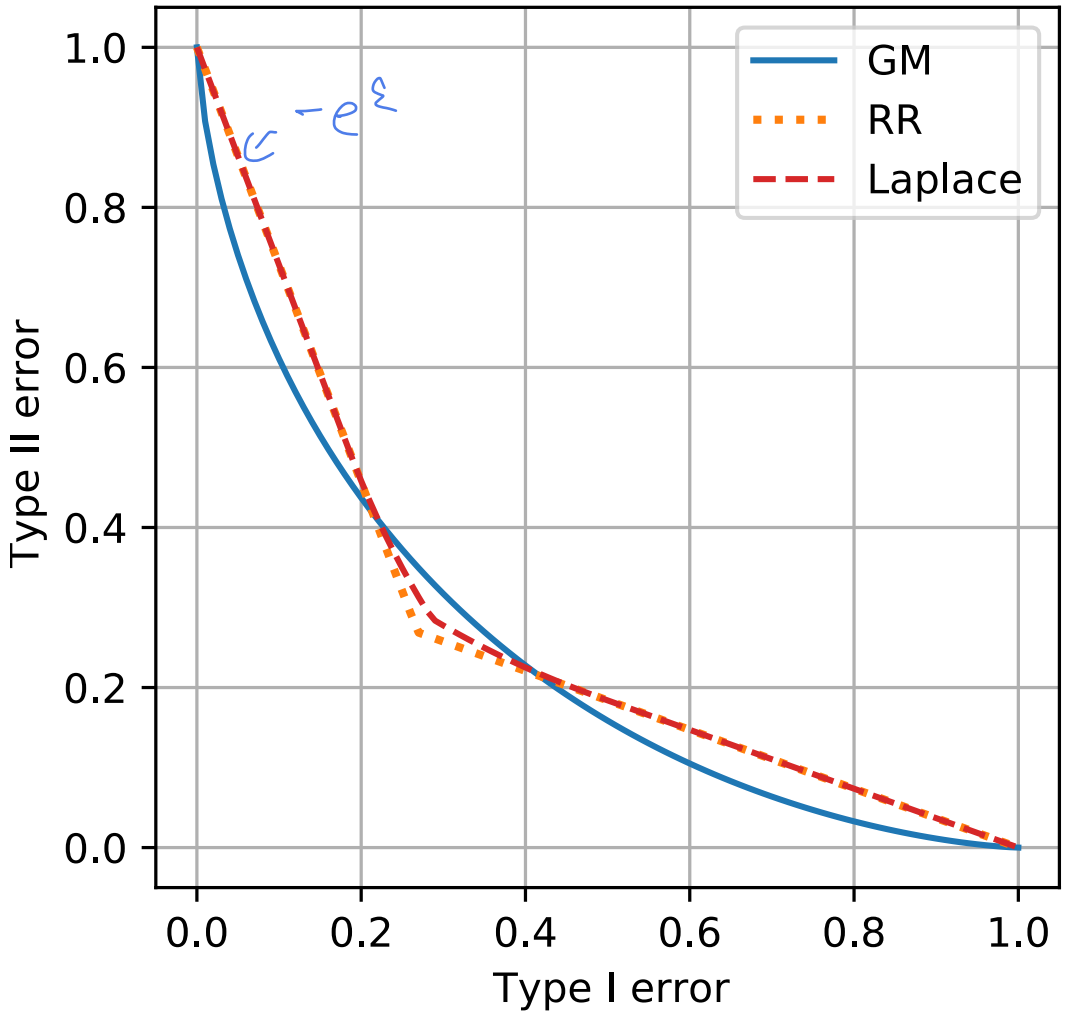
- Renyi DP as a function of order α
- Tradeoff function: Type II error as a function of Type I error
- Privacy profile: HS-divergence bound as a function of α

Example: Renyi DP of our familiar mechanisms

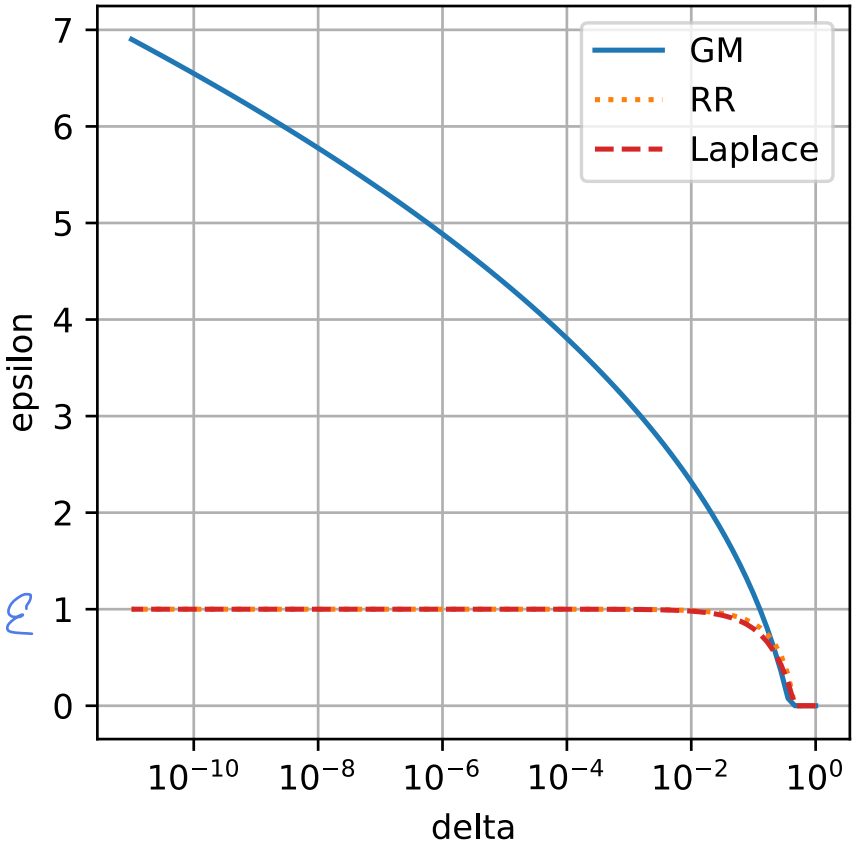


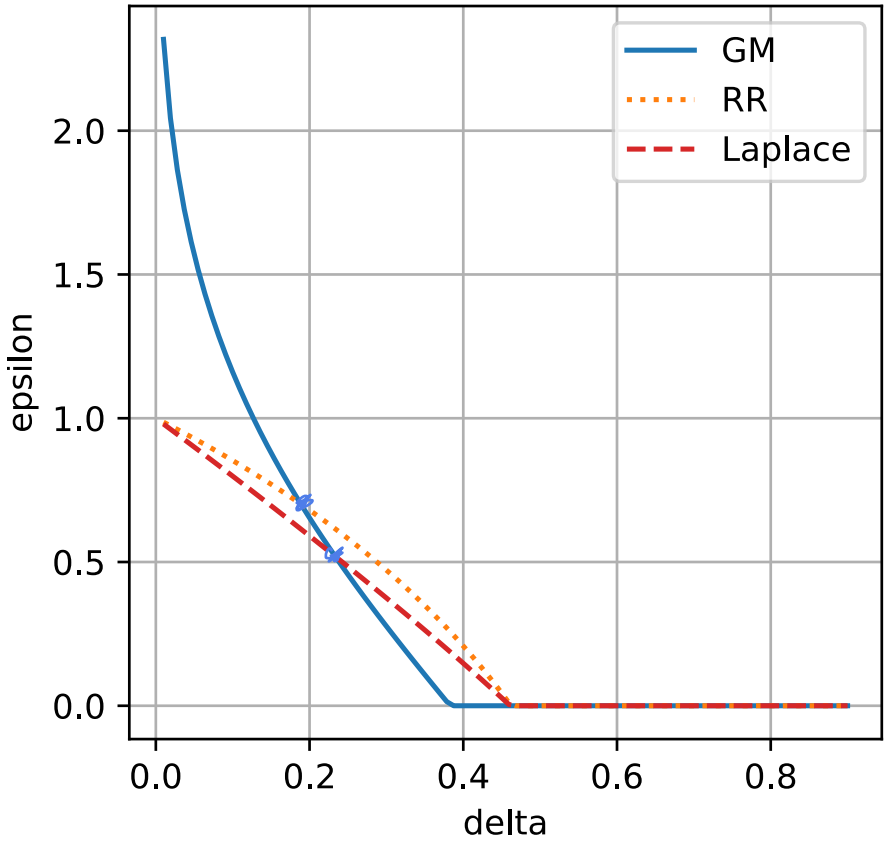
- RR and Laplace mechanism are calibrated to satisfy pure DP with ε=1
- GM and RR are calibrated to satisfy zCDP with ρ = 1/2

Example: The tradeoff function for our familiar mechanisms

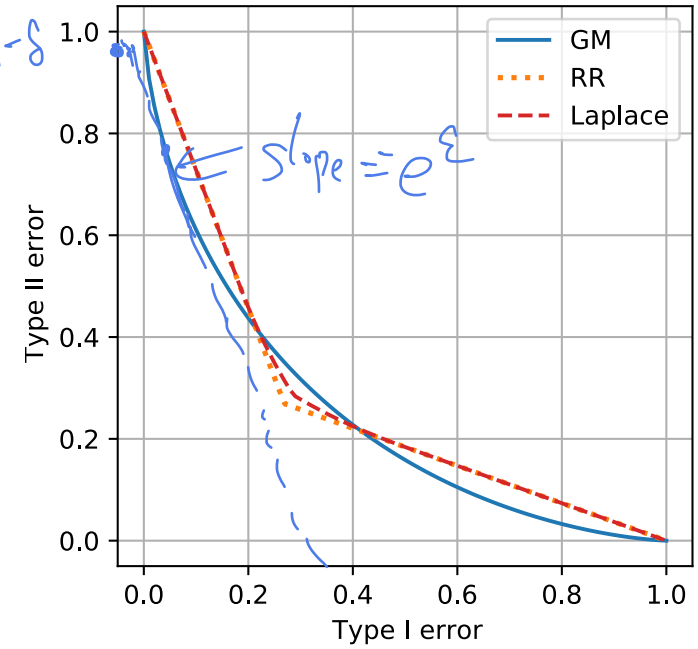


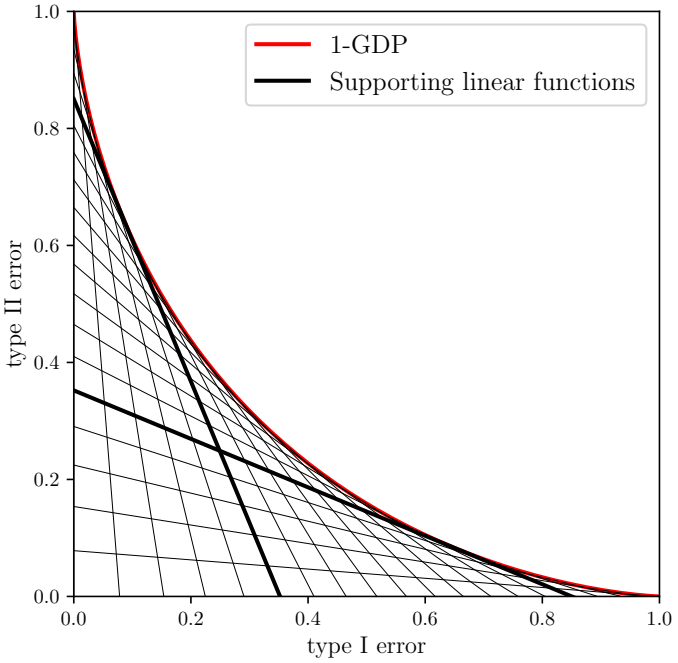
Example: Privacy profile of our familiar mechanisms





A mechanism-specific tradeoff function characterizes the the family of all (ε,δ)-DP a mechanism satisfies!



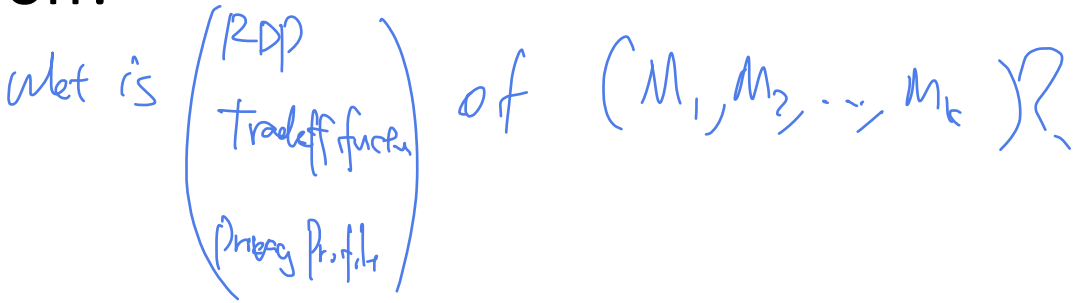


Proposition 2.12 (Primal to Dual). *For a symmetric trade-off function f, a mechanism is f-DP if and only if it is $(\varepsilon, \delta(\varepsilon))$-DP for all $\varepsilon \geqslant 0$ with* $\delta(\varepsilon) = 1 + f^*(-e^{\varepsilon})$.

Composition of Mechanism Specific Representations

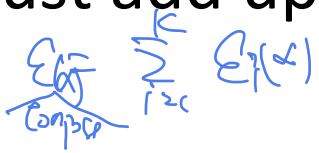
- If we have functional representations $f_1, f_2, \ldots, f_k$ of a mechanism M_1, M_2 ,…, M_k

- What is a functional representation of their composition?



Composition of Mechanism Specific Representations

1. RDP function: Just add up!
   - But… not tight



2. Tradeoff function:
   - Somewhat complex.
   - A central limit theorem exists

3. Privacy profile:
   - Convolution of the distribution of PLRVs for each pair of neighboring datasets

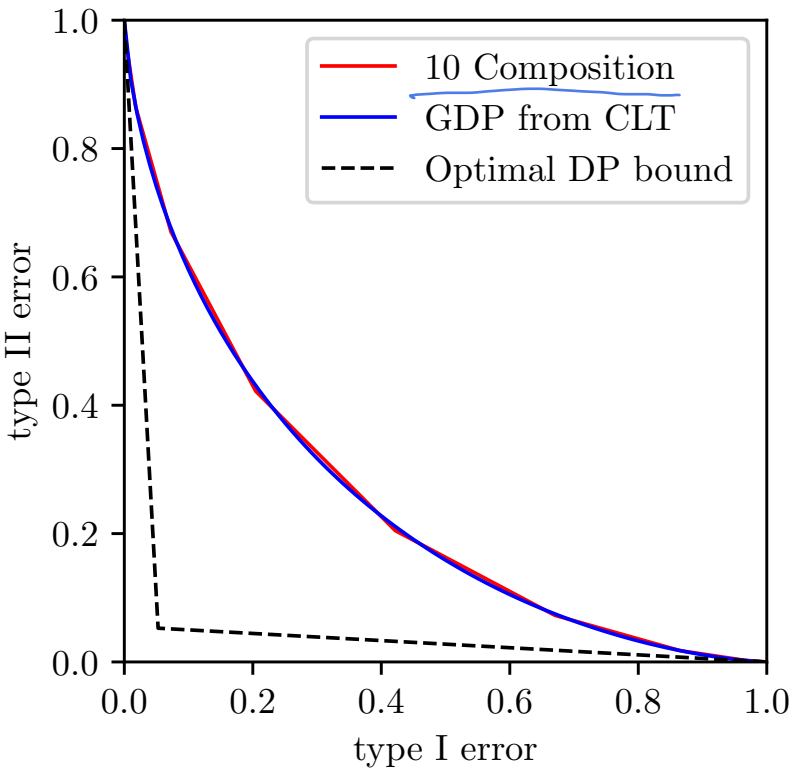
Composition of Tradeoff functions

- The composed mechanism

$$M : X \to Y_1 \times \cdots \times Y_n \ \textit{is} \ f_1 \otimes \cdots \otimes f_n\text{-}DP.$$

where $f \otimes g := T(P \times P', Q \times Q')$.

- A central limit theorem
  - Theorem 3.4 in Dong et al. 2019.



Composition of Privacy Profiles

- Somewhat tricky.
  - In principle, one can take all neighboring pairs of datasets, compose M1, …, Mk by adding up the PLRVs.
  - Requires us to know the worst-pair of datasets to make it efficient.
  - Even then, how do we know the composition of worst-case pair for each is a worst-cases pair for the composition?

Dominating pair distributions

- We say that P,Q is a dominating pair of Mechanism M if for all α>0 $$\sup_{D \simeq D'} H_\alpha(\mathcal{M}(D) \| \mathcal{M}(D')) \leq H_\alpha(P \| Q).$$
- When equal sign holds for all α>0, then it is a tight dominating pair.

Theorem 1: tight dominating pair always exists.

Theorem 2: Dominating pairs compose adaptively.

Theorem 3: (P,Q) is dominating if and only if M satisfies f-DP with f = T[P,Q].

(Zhu, Dong and W. https://arxiv.org/abs/2106.08567)

Two satisfying consequences

- Advanced composition for (ε,δ)-DP
  - One particular mechanism that attains the privacy-profile / tradeoff function pointwise.

Outcome	P_U	P_V
0	$\frac{e^{\varepsilon}(1-\delta)}{e^{\varepsilon}+1}$	$\frac{1-\delta}{e^{\varepsilon}+1}$
1	$\frac{1-\delta}{e^{\varepsilon}+1}$	$\frac{e^{\varepsilon}(1-\delta)}{e^{\varepsilon}+1}$
"I am U"	δ	0
"I am V"	0	δ

Leaky Randomized Response is a dominating pair for all (ε,δ)-DP mechanisms.

- Composition of Gaussian mechanism
  - Simply adding the PLRV or individual GMs

Next lecture

- Tools for modern privacy accounting
- Start differentially private machine learning