CSCI 699: Privacy Preserving Machine Learning - Week 5

Gaussian DP and Privacy Auditing

Sai Praneeth Karimireddy, Sep 27 2024

Recap

- Composition: simple - $k\varepsilon$-DP

ε-DP

Theorem. Advanced Composition

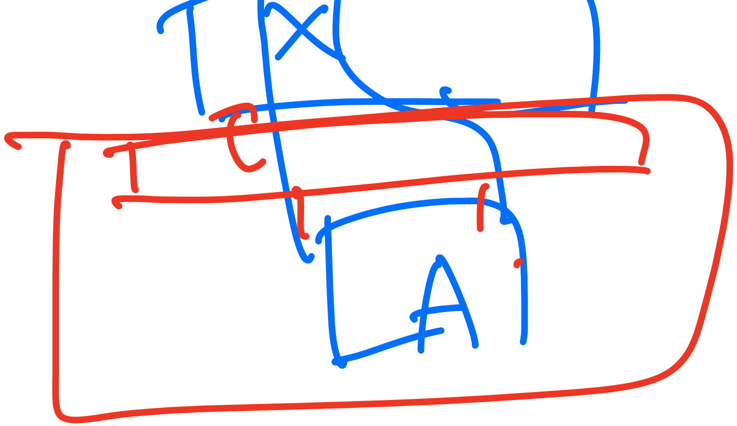
A combination of $A_1 \circ A_2 \circ A_k$, each of which is (ε, δ)-DP is $(\tilde{\varepsilon}, \tilde{\delta})$-DP where

$$\tilde{\varepsilon} = \varepsilon\sqrt{2k\ln(1/\delta')} + k\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1} \quad \text{and} \quad \tilde{\delta} = k\delta + \delta'$$

For any choice of δ'.

ε-DP

Recap



- Subsampling amplification

Theorem. Subsampling Amplification

Composing an (ε, δ)-DP A with a sampling rate of q results in an $(\tilde{\varepsilon}, \tilde{\delta})$-DP algorithm where

$$\tilde{\varepsilon} = \log(1 - q + qe^{\varepsilon}) = O(q\varepsilon) \quad \text{and} \quad \tilde{\delta} = q\delta$$

Recap

- Private SGD with clipping L1 norm:
  - $\theta_t = \theta_{t-1} - \gamma \text{Clip}_\tau \left(\nabla_\theta \ell(f(x_t; \theta), y_t) \right) + Lap(2\tau/\varepsilon)$ δ
- With $q = 1/n$, k rounds satisfies $(O(\varepsilon/n\sqrt{k \ln(1/\delta)}), \delta)$-DP for any $\delta > 0$.
- Can also clip L2 norm and use Gaussian mechanism.

  $N\left(0, \left(\frac{c\sqrt{\log\frac{1}{\delta}}}{\varepsilon}\right)^2\right)$
- Q: what did you observe empirically L1 vs. L2?

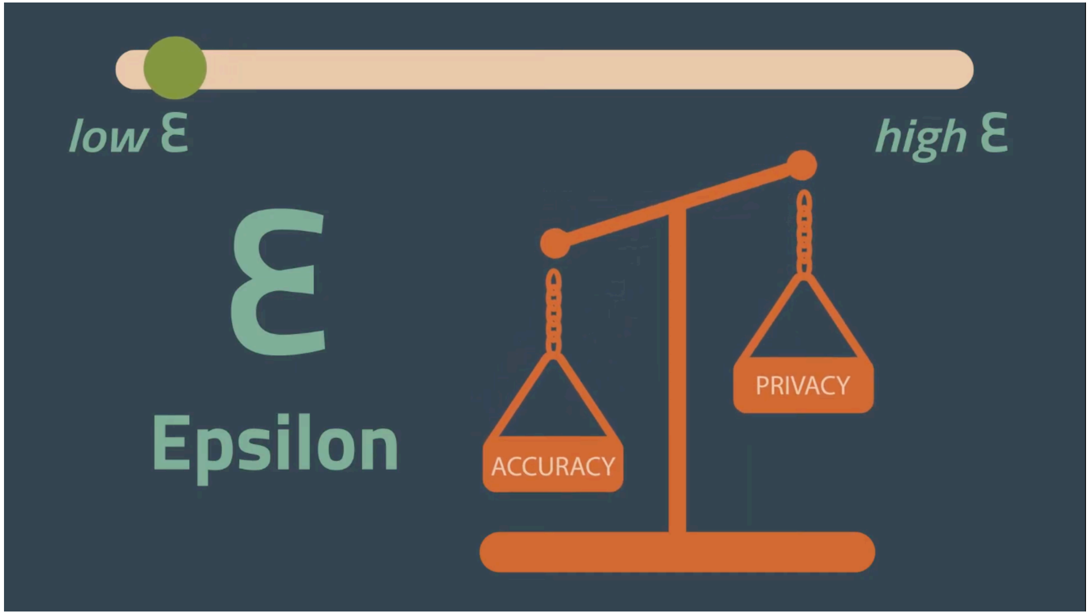
$\forall \delta$

Agenda for today

Analyzing privacy of ML training

- Gaussian DP
- Privacy Auditing
- Presentations + discussions
- Auditing Practical - next week (needs HW2 soln)

Gaussian Differential Privacy



Drawbacks of Approximate DP

- After k steps of Lap-SGD, we were able to show $(\boxed{\varepsilon\sqrt{2k\ln(1/\delta)}}, \delta)$-DP
- But advanced composition is too lose.

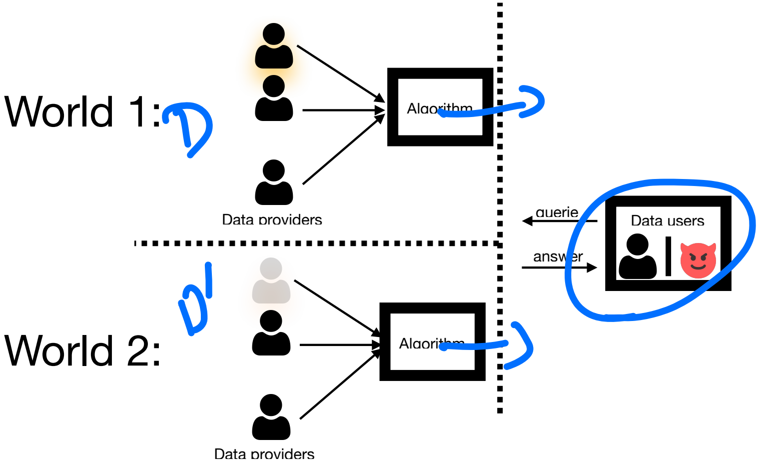
$+ \cancel{k\left(\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}\right)}$

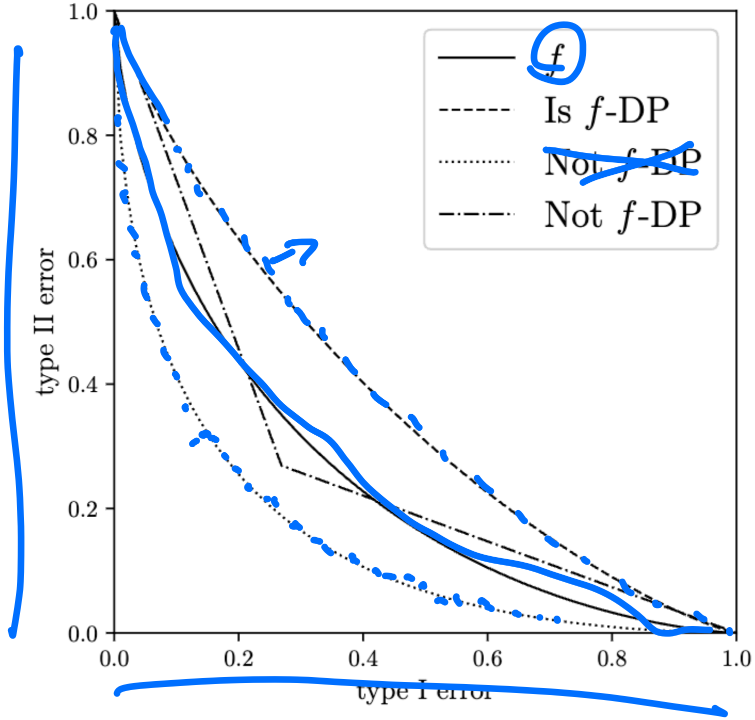
- Renyi DP
- Concentrated DP

f-DP

Most general privacy definition

- **Definition.** Given a function f, we say an algorithm is f-DP if the tradeoff curve of an optimal distinguisher is strictly above f.





f-DP

Generalization (ε, δ)-DP

- **Prop 2.5** [WZ10]**.** A is (ε, δ)-DP iff it satisfies $f_{\varepsilon,\delta}$-DP for

$$f_{\varepsilon,\delta} = \max(1 - \delta - e^{\varepsilon}x \ , \ (1 - \delta - x)/e^{\varepsilon})$$

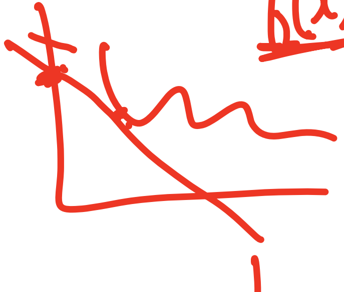
Primal - Dual / Convex conjugate

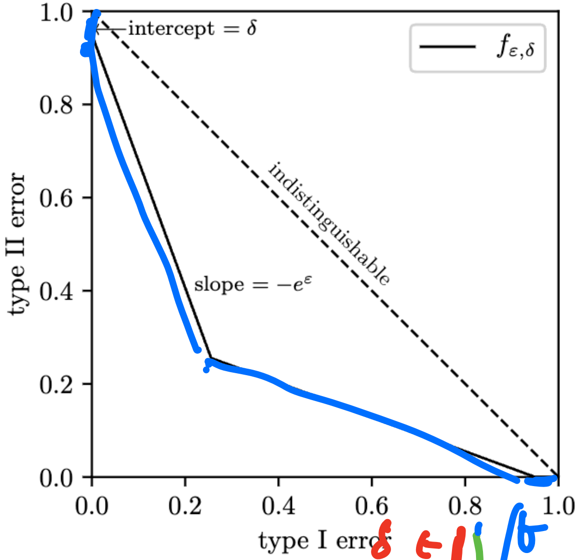
- **Prop 2.12** [DRS19] A is f-DP iff it satisfies $(\varepsilon, \delta_f(\varepsilon))$-DP for $\forall \varepsilon \geq 0$ and

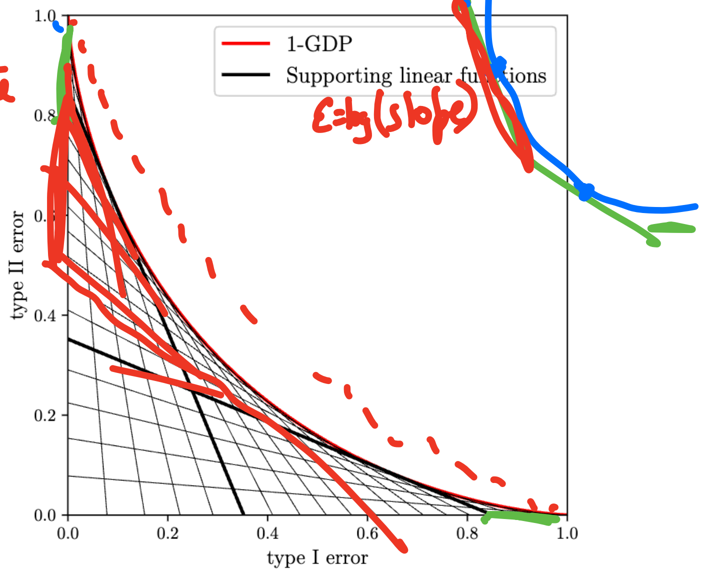
$$\delta_f(\varepsilon) = 1 + f^*(-e^{\varepsilon}).$$

dual

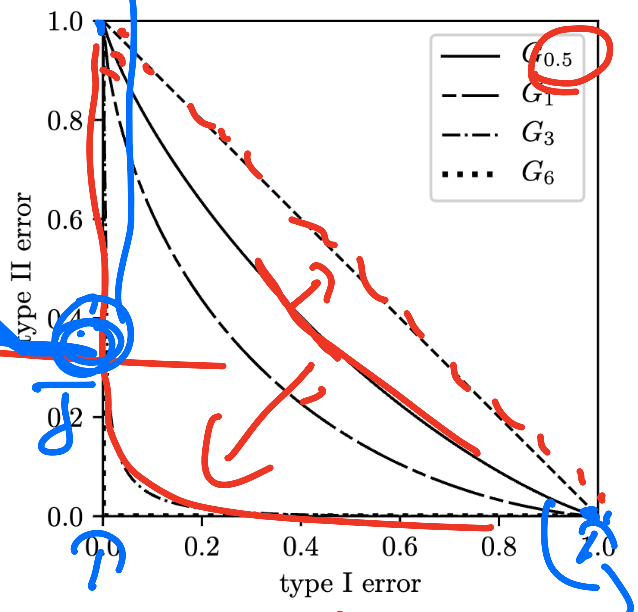
$\delta(x) =$





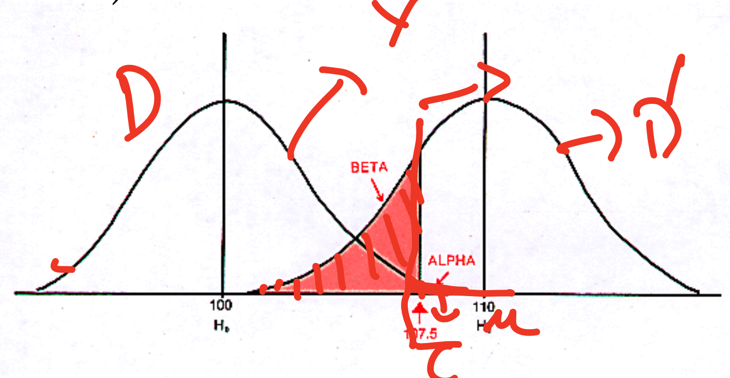


Gaussian-DP



- **Definition.** A is μ-GDP if it satisfies f_μ DP for $f_\mu = T\left(\mathcal{N}(0,1)\ ,\ \mathcal{N}(\mu,1)\right)$

- $$\frac{Pr[A(D)=t]}{Pr[A(D')=t]} \le \frac{Pr[\mathcal{N}(0,1)=t]}{Pr[\mathcal{N}(\mu,1)=t]} = \exp\left(\frac{1}{2}(\mu^2 - 2\mu t)\right)$$ $\forall t$

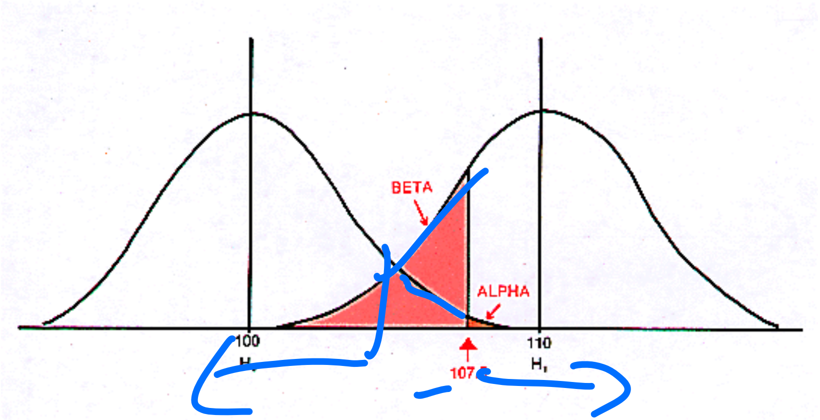


- $\alpha(\tau) = 1 - \Phi(\tau)$ and $\beta(\tau) = \Phi(\tau - \mu)$

$\beta(\alpha) = \phi(\phi^{-1}(1-\alpha) - \mu)$

Gaussian-DP
Gaussian mechanism

- **Definition.** A is μ-GDP if it satisfies f_μ-DP for $f_\mu = T\left(\mathcal{N}(0,1), \mathcal{N}(\mu,1)\right)$

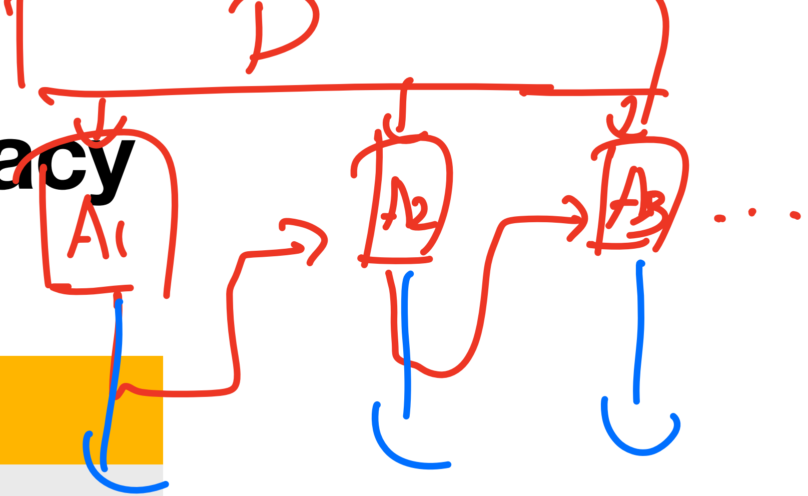


Theorem. Gaussian mechanism

Given $f : \mathcal{X}^n \to \mathbb{R}^d$ with Δ bounded ℓ_2-sensitivity, $f(D) + \mathcal{N}\left(0, \frac{\Delta^2}{\mu^2} I_d\right)$ is μ-GDP.

Gaussian Differential Privacy

Tight composition



Theorem. GDP Composition

Composition of $A_1 \circ A_2 \ldots \circ A_k$, each of which is μ_i-GDP is $\boxed{\sqrt{\sum_{i=1}^{k} \mu_i^2}\text{-GDP.}}$

$\mu\sqrt{k}$ - GDP

$A_1 \to \mu_1$-GDP

$A_2 \to \mu_2$-GDP

$(A_1(D), A_2(D))$ vs. $(A_1(D'), A_2(D'))$

$\Pr\{A_1(D) = t_1, A_2(D) = t_2\}$

Gaussian Differential Privacy

Tight composition

Theorem. GDP Composition

Composition of $A_1 \circ A_2 \ldots \circ A_k$, each of which is μ_i-GDP is $\sqrt{\sum_{i=1}^k \mu_i^2}$-GDP.

$$\Pr[A_1(D')=t_1, A_2(D')=t_2]$$

$$= \frac{\Pr[A_1(D)=t_1]}{\Pr[A_1(D')=t_1]} \cdot \frac{\Pr[A_2(D)=t_2]}{\Pr[A_2(D')=t_2]}$$

$$\le \exp(\mu_1^2 - 2\mu_1 t_1) \cdot \exp(\mu_2^2 - 2\mu_2 t_2)$$

$$= \exp\left(\left(\sqrt{\mu_1^2+\mu_2^2}\right)^2 - 2(\mu_1 t_1 + \mu_2 t_2)\right)$$

$$\le \exp\left(\sqrt{\mu_1^2+\mu_2^2}^2 - 2\sqrt{\mu_1^2+\mu_2^2}\sqrt{t_1^2+t_2^2}\right)$$

$$a_1 b_1 + a_2 b_2 \le \sqrt{a_1^2+a_2^2} \cdot \sqrt{b_1^2+b_2^2}$$

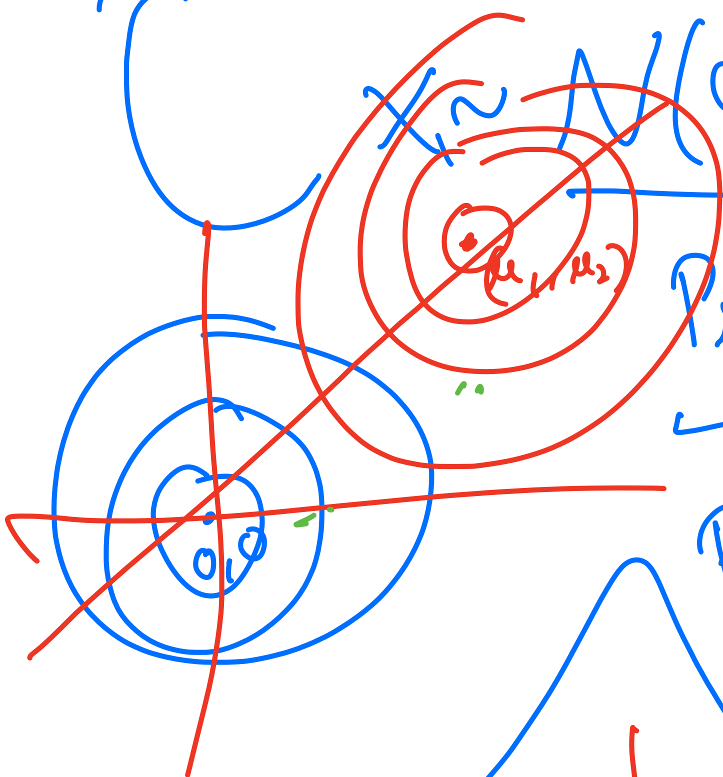
$$\exp\left(\sqrt{\mu_1^2+\mu_2^2} - 2\sqrt{\mu_1^2+\mu_2^2}\,\boxed{\sqrt{t_1^2+t_2^2}}\right)$$

$T\ t$

$\exp\left(\sqrt{\mu_1^2+\mu_2^2} - 2\sqrt{\mu_1^2+\mu_2^2}\sqrt{t_1^2+t_2^2}\right)$

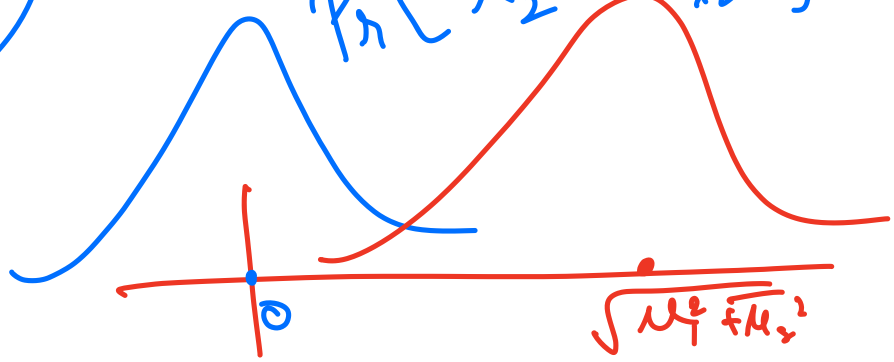
X_2

$X_1 \sim \mathcal{N}(0, I_2)$ vs. $\mathcal{N}\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, I_2\right)$



$\Pr\left[X_1 = \begin{bmatrix}t_1\\ t_2\end{bmatrix}\right]$

$\Pr\left[X_2 = \begin{bmatrix}t_1\\ t_2\end{bmatrix}\right]$



$\Rightarrow \sqrt{\mu_1^2+\mu_2^2}$ - GDP

| p

Gaussian Differential Privacy

Canonical f

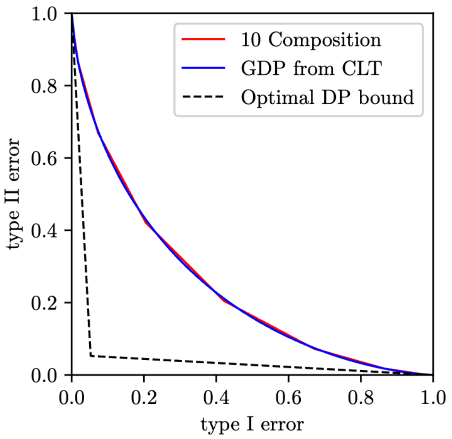
Theorem 3.4 [DRS19] Central limit theorem of composition

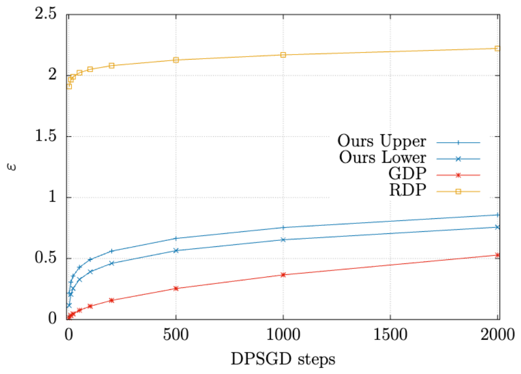
Given some regularity assumptions, composition of $A_1 \circ A_2 \ldots \circ A_k$, each of which is f_i-DP is approximately μ-GDP for $\mu = \frac{2\sqrt{k}\kappa_1}{\kappa_1 - \kappa_2}$ for $\kappa_1 = -\int_0^1 \log|f'(x)|\,dx$ and $\kappa_2 = -\int_0^1 \log^2|f'(x)|\,dx$.

$\sum_{i=1}^{k} R_i$

Gaussian Differential Privacy

Canonical f



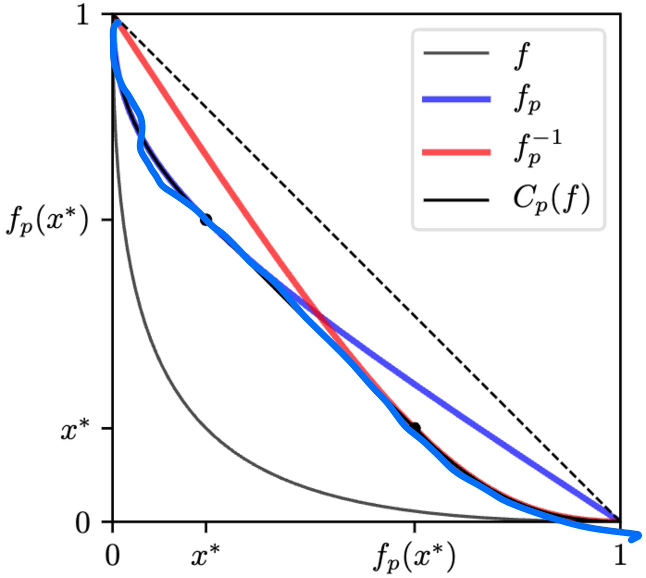


- In stats, combining may random variables ≈ Gaussian by CLT. In DP, composing many DP steps ≈gDP.

- Caution: just like CLT sometimes fails, Thm 3.4 is sometimes fails and underestimates privacy [GLW21].

Gaussian Differential Privacy

Amplification by subsampling



- Define $f_q(x) = qf(x) + (1-q)(1-x)$ and f_q^{-1}
- **Theorem 4.2** [DRS19] Composing q-sampling with f-DP, is $\left(\min(f_p, f_p^{-1})\right)^{**}$-DP
- Unfortunately, no closed form for GDP, compute numerically.

Private SGD

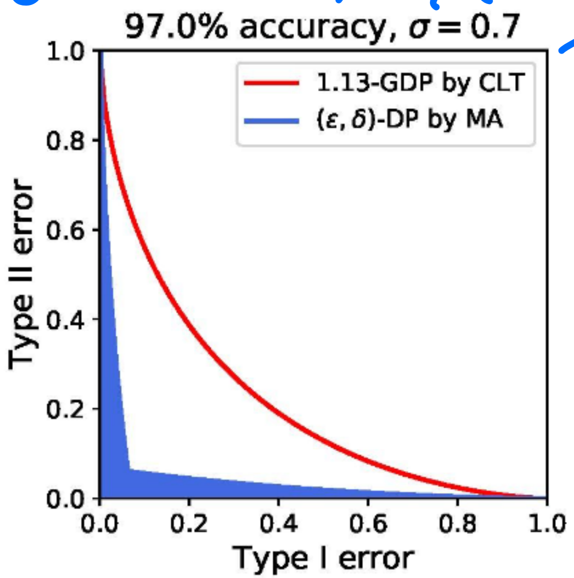
Using Gaussian-DP

Clip $\| \|_2 - C$

add $N(0, \frac{C^2}{\mu^2} I)$

Corollary 5.4 [DRS19] Subsampled Composition

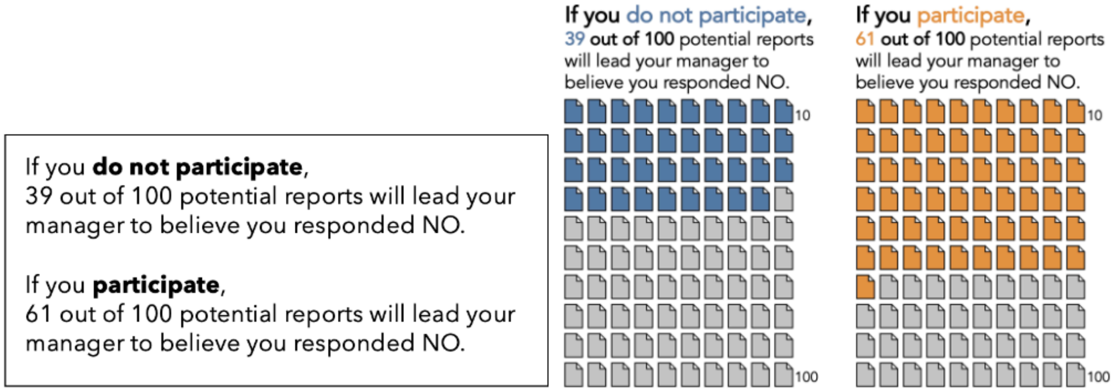
Suppose each A_i is μ-GDP. Then, composing q-sampled A_i is asymptotically $\left(q\sqrt{k}\sqrt{e^{\mu^2}\Phi(3\mu/2) + 3\Phi(-\mu/2) - 2}\right)$-GDP.



Tightest privacy bound [B+'20].
But, only asymptotically valid.

Aside: Communicating Privacy

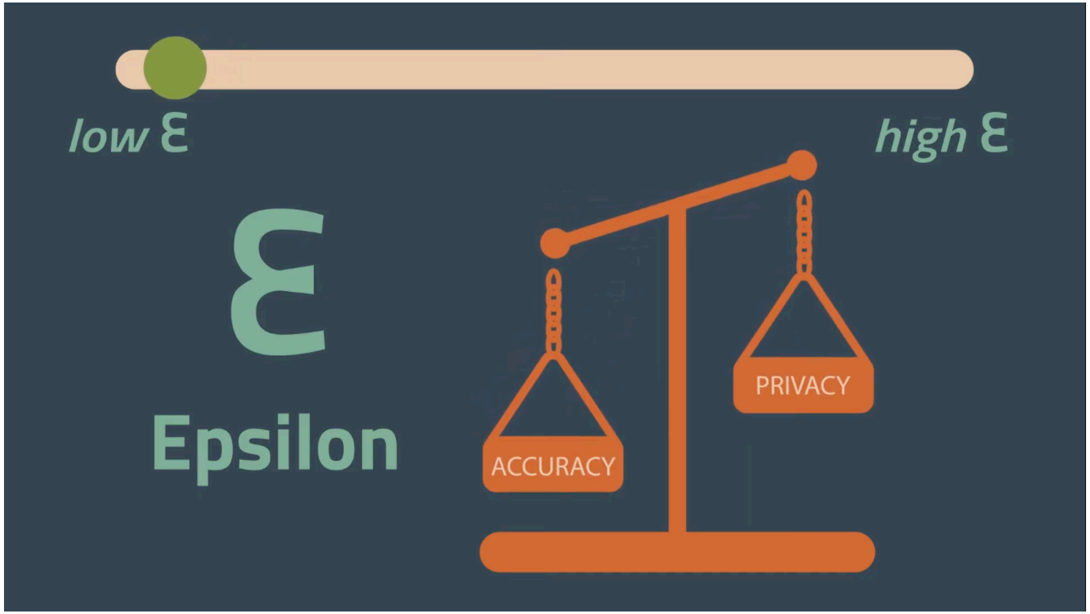
Odds ratio



(a) ODDS-TEXT (b) ODDS-VIS

- How do you communicate privacy risk to your friends?
- Excellent study: [N+UseNIX'23]
- Using odds ratio leads to increased understanding of risks and willingness to share data.
- How to explain ε-DP and μ-GDP? Need to incorporate prior knowledge of attacker.

Privacy Auditing



Drawbacks of pure theory

- Bounds always loose
  - people assume this and train models with high theoretical ε
- Maybe my implementation is incorrect
- Why should I trust your claim?

Backpropagation Clipping for Deep Learning with Differential Privacy

Timothy Stevens* — *University of Vermont*
Ivoline C. Ngong* — *University of Vermont*
David Darais — *Galois, Inc.*
Calvin Hirsch — *Two Six Technologies*
David Slater — *Two Six Technologies*
Joseph P. Near — *University of Vermont*

- In 2022, proposed to integrate clipping into forward/backward pass directly
- SOTA accuracy with 30x smaller ε

Privacy Auditing

Debugging Differential Privacy: A Case Study for Privacy Auditing

Florian Tramèr, Andreas Terzis, Thomas Steinke, Shuang Song, Matthew Jagielski, Nicholas Carlini*
Google Research

- Consider the following test:
  - D = MNIST dataset: 60k images
  - D’ = Add (x', y').
  - Train a CNN θ using [S+22] to get 0.98 acc and (0.21, 10–5)-DP.
  - Check $\ell_\theta(x', y') \leq \tau$. If D’ will be smaller.
  - Repeat 100k on D and 100k on D’.