CSCI 699: Privacy Preserving Machine Learning - Week 4

Gaussian DP and Privacy Auditing

Sai Praneeth Karimireddy, Sep 22 2025

Recap

- Approximate differential privacy

[Dwork and Roth 2014]

Let us draw a variable $t \sim A(D)$. Then the privacy loss random variable.

$$\mathscr{L}_{D,D'} = \ln\left(\frac{Pr[A(D) = t]}{Pr[A(D') = t]}\right)$$

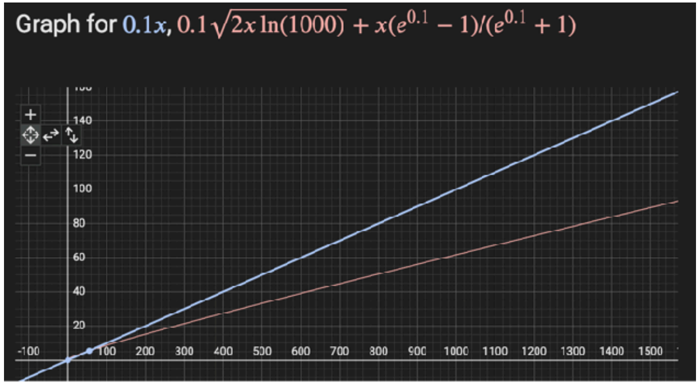
A satisfies (ε, δ)-DP iff for any similar/neighboring datasets $D, D' \in \chi^n$ we have $Pr\left[\mathscr{L}_{D,D'} \geq \varepsilon\right] \leq \delta$

Recap

$\varepsilon_1 + \varepsilon_2 + \varepsilon_3$

max ε

$\le \varepsilon_{max}$ k



- Composition: simple - $k\varepsilon$-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'.

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:
    - $\theta_t = \theta_{t-1} - \frac{1}{n}\sum_{i=1}^{n} \gamma \text{Clip}_\tau\left(\nabla_\theta \ell(f(x_i;\theta), y_i)\right) + Lap(2\tau/\varepsilon)$
- With $\cancel{q = 1/n}$, 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.
- Q: what did you observe empirically L1 vs. L2?

q-amplification

Recap

Poisson subsampling disadvantages

- $\theta_t = \theta_{t-1} - \gamma \left(\left[\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \text{Clip}_\tau \left(\nabla_\theta \ell(f(x_i; \theta), y_i) \right) \right] + \mathcal{N}(0, \tau^2 \rho^2) \right)$

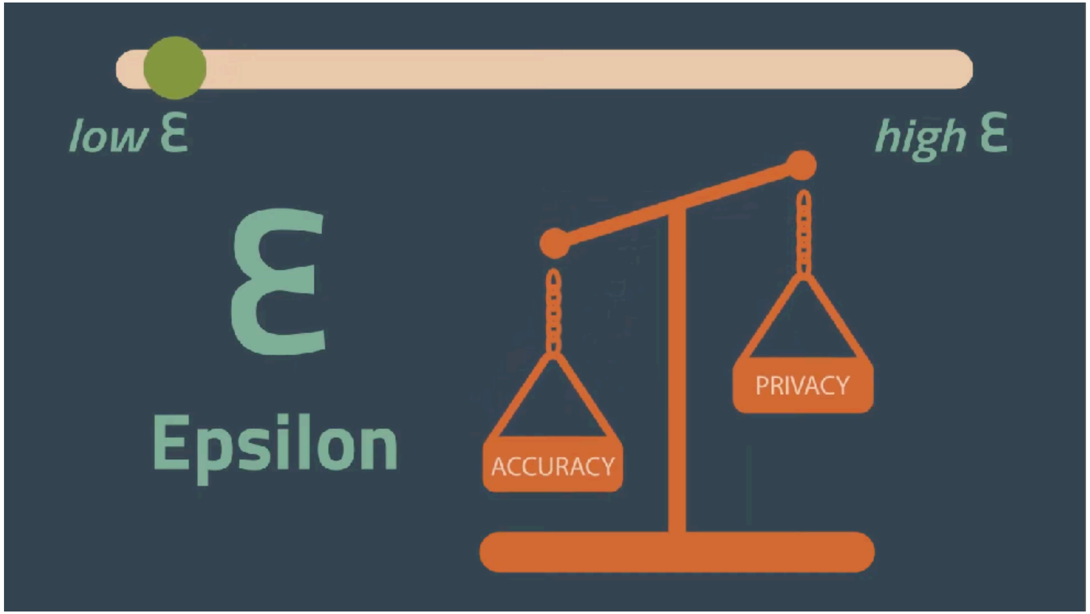
- I cannot set $\rho \propto |\mathcal{B}|^{-1}$ - mechanism **cannot be data-dependent.**
  It should work for the worst case i.e. when $|\mathcal{B}| = 1$.

Agenda for today

Analyzing privacy of ML training

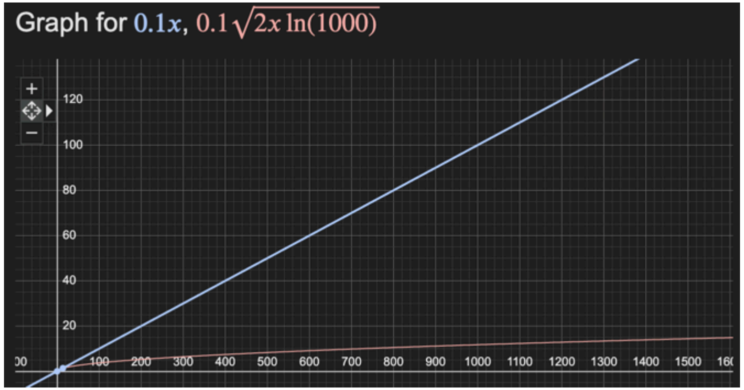
- Improving composition
- Gaussian DP
- Privacy Auditing
- HW1 solutions

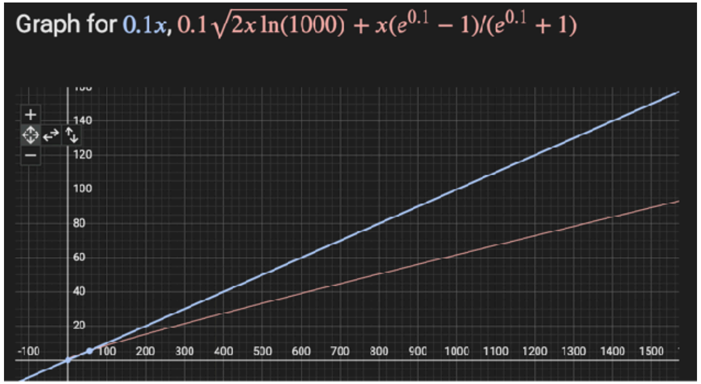
Better composition



Approximate DP analysis is loose

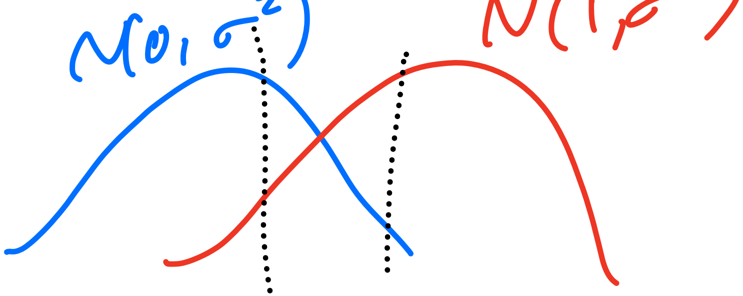
- After k steps of DP-SGD, we had $O(\varepsilon\sqrt{2k\ln(1/\delta)} + k\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}, \delta)$
- The extra k seems unnecessary advanced composition is too lose.

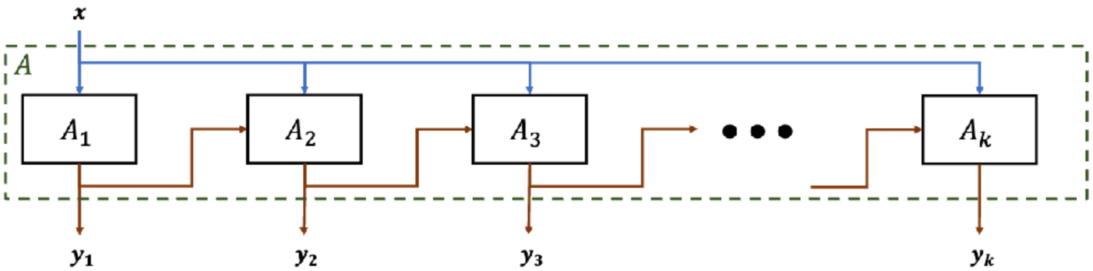




Advanced composition

Proof sketch





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

$O(n)$

-1

- Privacy random variable of composition:

$$R = \sum_{i=1}^{k} \log \left(\frac{Pr[A_i(D) = t_i]}{Pr[A_i(D') = t_i]} \right) = \sum_{i=1}^{k} R_i$$

- If $R_i \in [-\varepsilon, \varepsilon]$, 0-mean, conditionally independent, we get $O(\varepsilon\sqrt{k})$

- With bias, we get $O(\varepsilon\sqrt{k} + E[R] \cdot k)$

Advanced composition

Proof sketch

- Privacy random variable of composition:

$$R = \sum_{i=1}^{k} \log\left(\frac{Pr[A_i(D) = t_i]}{Pr[A_i(D') = t_i]}\right) = \sum_{i=1}^{k} R_i$$

- What is the bias i.e. $E[R_i] = ?$
- $E_t[\mathscr{L}] = E_{t\sim y}[\log(P[y = t]/P[y' = t])] = \text{KL}(y\|y')$
  - where $y = A(D)$ and $y' = A(D')$
- Let's compute it

$e^{-\varepsilon} \leq \frac{y}{y'} \leq e^{\varepsilon}$

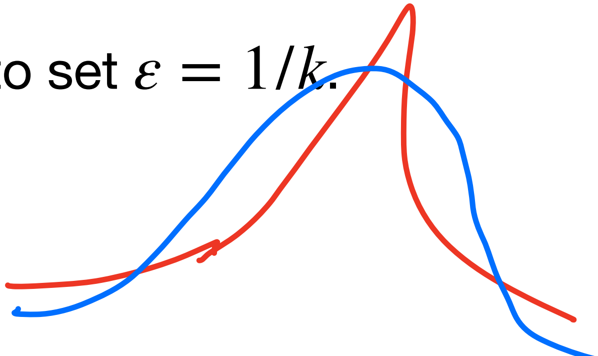
$\leq \varepsilon(e^{\varepsilon}-1)$

Advanced composition

Proof sketch

$\frac{\Delta}{\varepsilon}$

- Worst-case: $D_{\mathrm{KL}}\left(y\, y'\right) \leq \varepsilon(e^{\varepsilon} - 1)$
- KL-divergence between two Laplace distributions with different means
  - $\leq \Delta$
  - $D_{\mathrm{KL}}\left(\text{Laplace}(\mu_1, b) \,\|\, \text{Laplace}(\mu_2, b)\right) = \dfrac{|\mu_1 - \mu_2|}{b} + e^{-|\mu_1 - \mu_2|/b} - 1.$
  - $= \varepsilon + e^{-\varepsilon} - 1 \approx O(\varepsilon)$
- After k rounds, $O(\varepsilon\sqrt{k} + \varepsilon k) = O(\varepsilon k)$. Need to set $\varepsilon = 1/k$.



Advanced composition

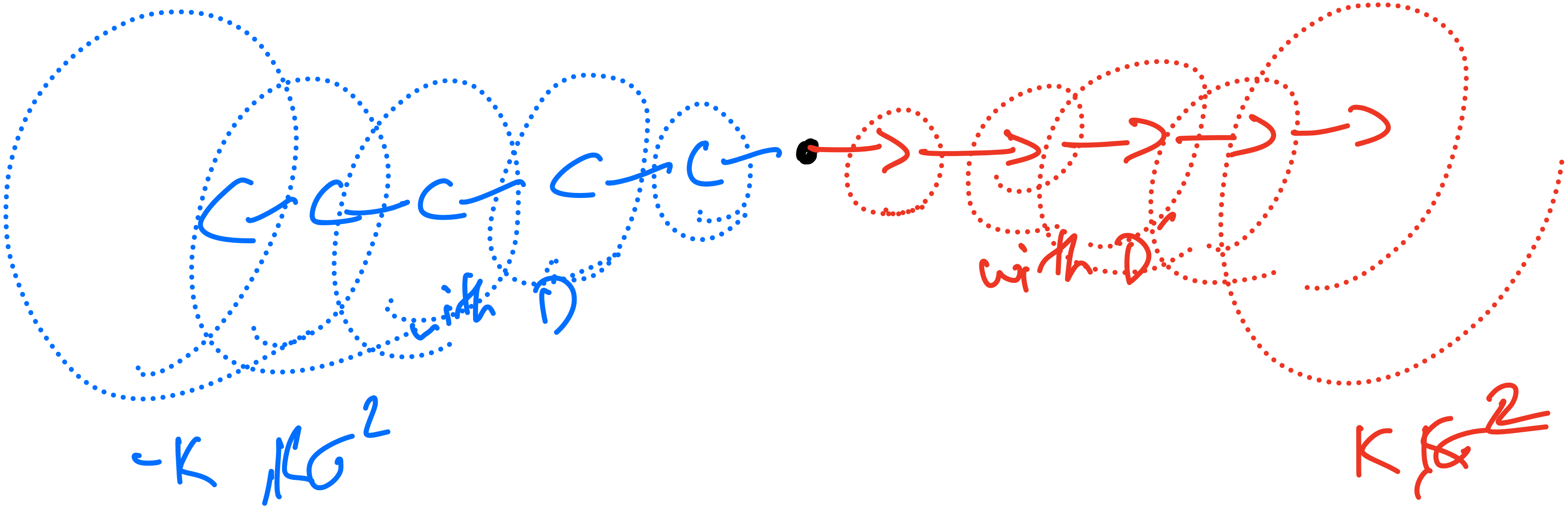
$\varepsilon^2 k + \varepsilon\sqrt{k}$

Proof sketch

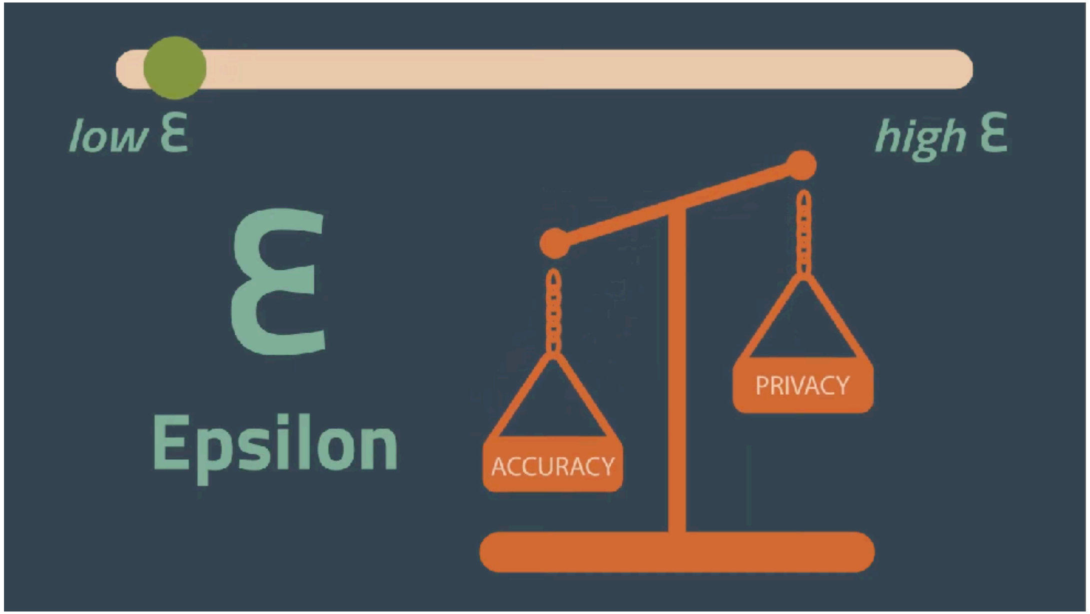
- KL-divergence between two Gaussian distributions with different means
  - $D_{\mathrm{KL}}\left(\mathcal{N}(\mu_1,\sigma^2) \,\|\, \mathcal{N}(\mu_2,\sigma^2)\right) = \dfrac{(\mu_1-\mu_2)^2}{2\sigma^2}$. $\approx o(\varepsilon^2)$
  - $= O(\varepsilon^2)$ since recall $\sigma = \dfrac{\Delta_2\sqrt{2\ln(1.25/\delta)}}{\varepsilon}$
  - After k rounds, $O(\varepsilon\sqrt{k} + \varepsilon^2 k)$. Sufficient to set $\varepsilon = 1/\sqrt{k}$!

Advanced composition $\sigma\sqrt{K}$ $2K$

Geometric intuition for gaussians



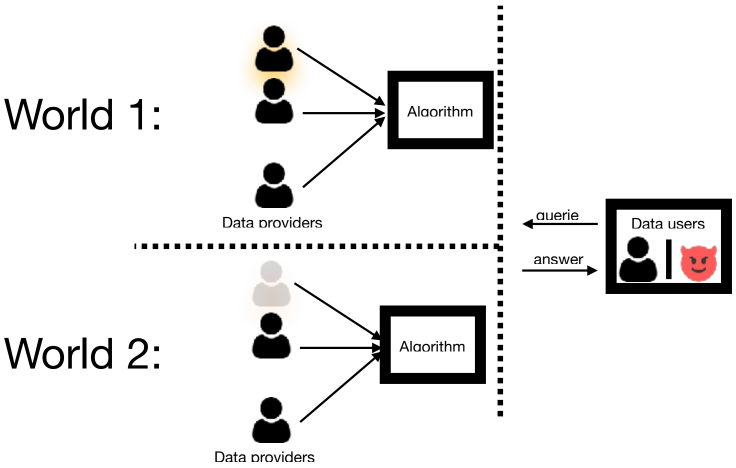
Gaussian Differential Privacy

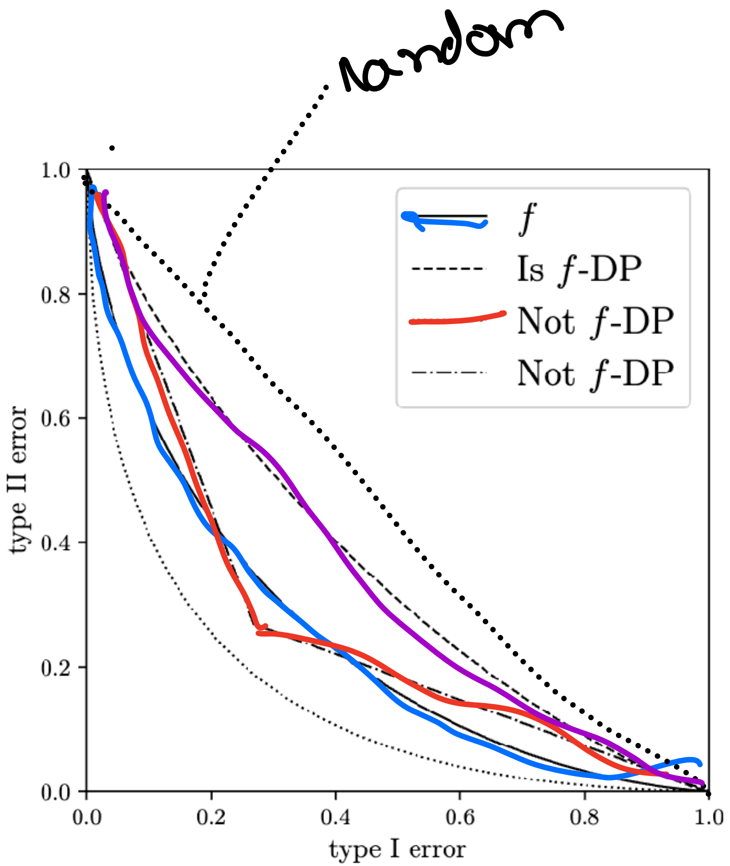


f-DP

Most general privacy definition

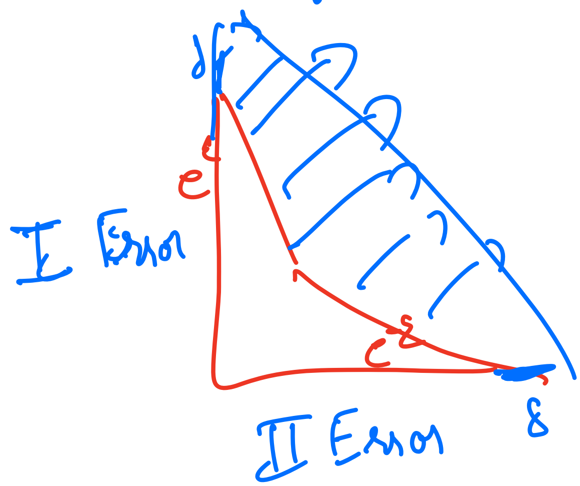
- **Definition.** Given a function f, we say an algorithm is f-DP if the tradeoff curve of an optimal distinguisher is strictly above f.

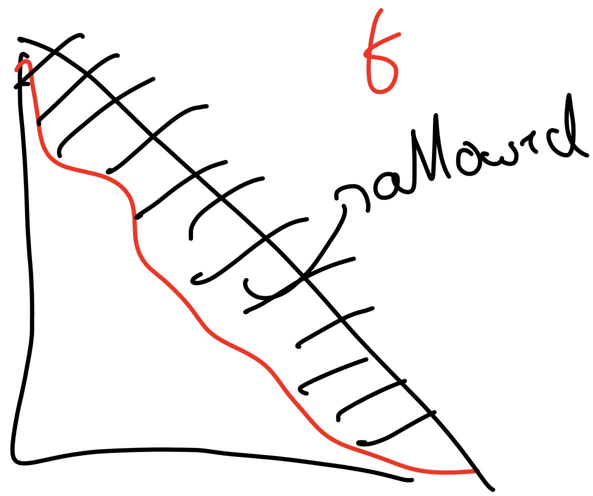




f for ε-DP?

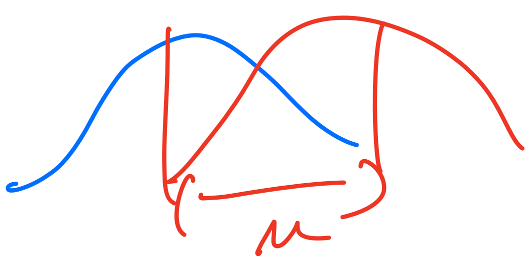
$$\int \max\left(e^{\varepsilon}x + y,\ x + e^{\varepsilon}y\right) \le 1$$

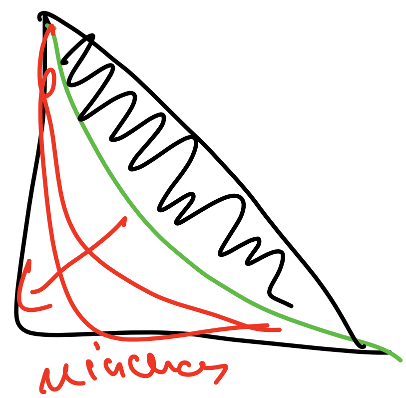


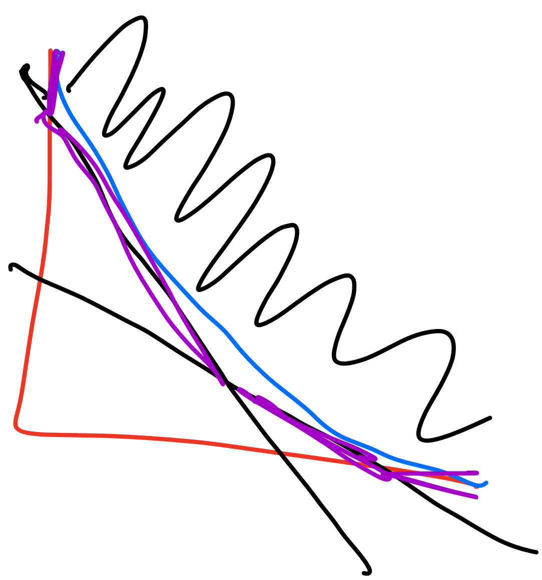


μ- Gaussian-DP

$$f = T\left(N(0,1),\ N(\mu,1)\right)$$



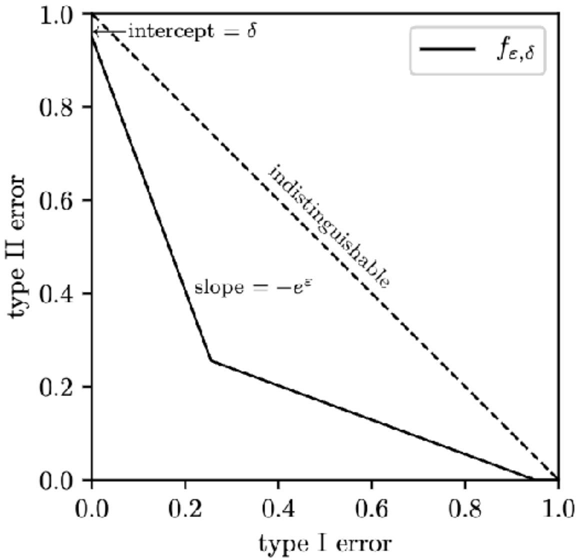




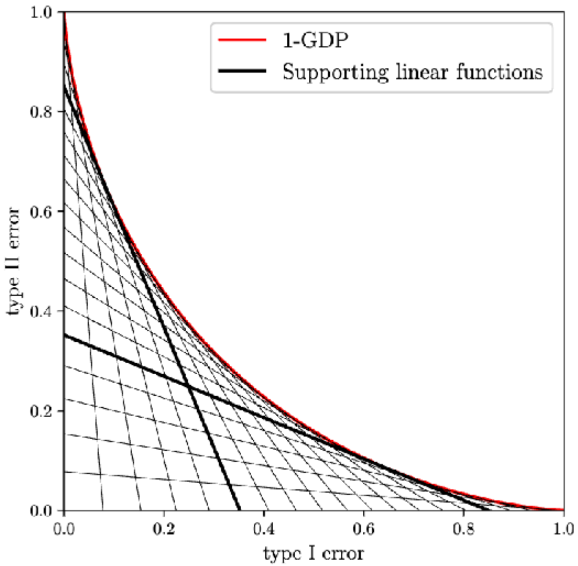
f-DP

Generalization (ε, δ)-DP

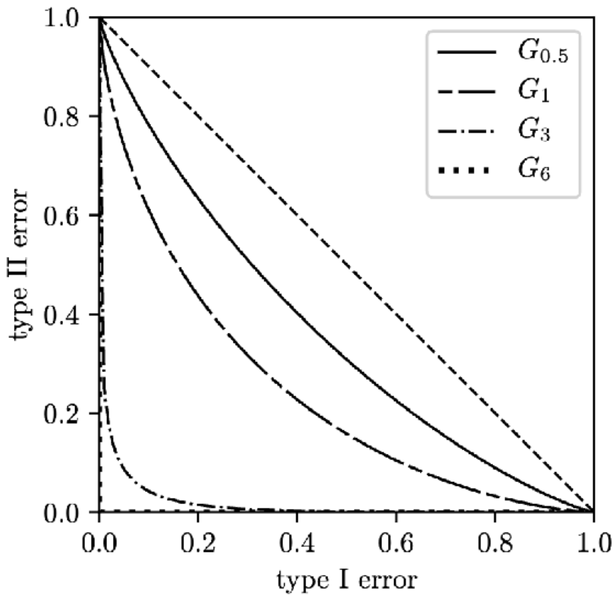
- **Prop 2.5** [WZ10]. A is (ε, δ)-DP iff it satisfies $f_{\varepsilon,\delta}$-DP for
$f_{\varepsilon,\delta} = \max(1 - \delta - e^{\varepsilon}x \ , \ (1 - \delta - x)/e^{\varepsilon})$



- **Prop 2.12** [DRS19] A is f-DP iff it satisfies $(\varepsilon, \delta_f(\varepsilon))$-DP for $\forall \varepsilon \geq 0$ and
$\delta_f(\varepsilon) = 1 + f^*(-e^{\varepsilon})$.

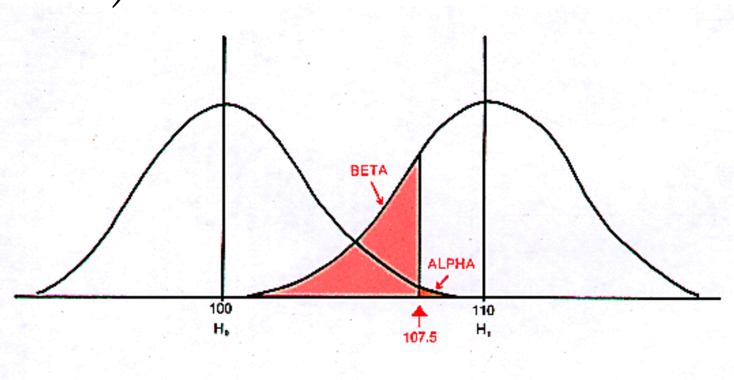


Gaussian-DP



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

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



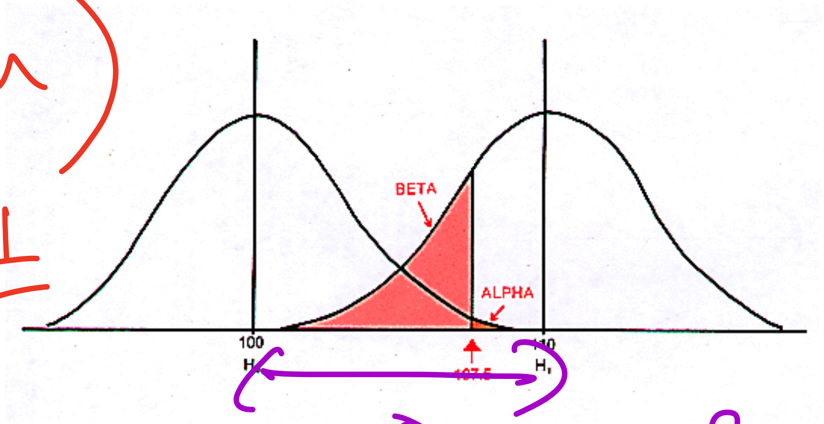
- $\alpha(\tau) = 1 - \Phi(\tau)$ and $\beta(\tau) = \Phi(\tau - \mu)$

Gaussian-DP

Gaussian mechanism

$\left(\frac{\Delta}{\sigma} = \mu\right)$

$\frac{1}{\sigma}$



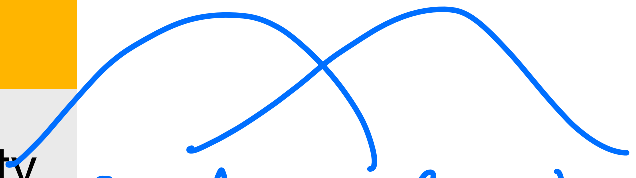
Δ

σ^2

- **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.



$(0,1)$ $(\mu,1)$

Gaussian Differential Privacy

Tight composition

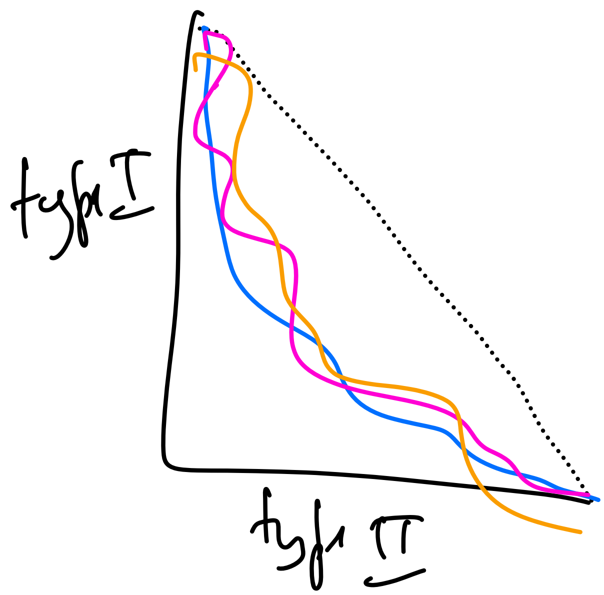
μ- GDP

Theorem. GDP Composition

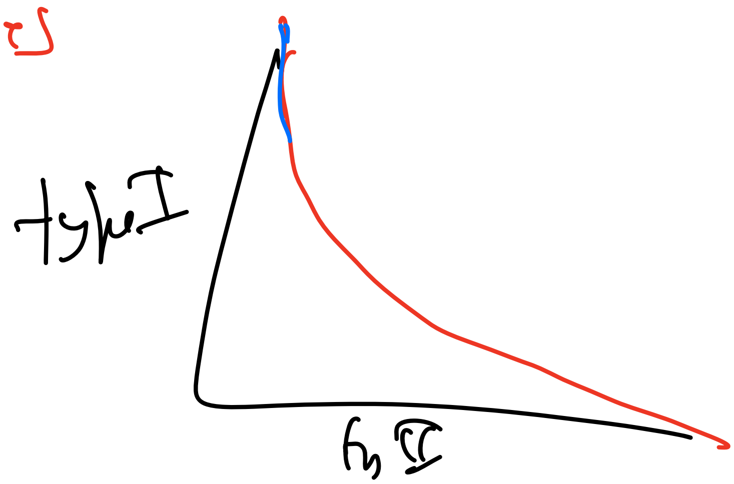
Composition of $A_1 \circ A_2 \ldots \circ A_k$, each of which is μ_i-GDP is $\sqrt{\sum_{i=1}^{k} \mu_i^2}$-GDP.

$\mu\sqrt{k}$

$$\theta_t = \theta_1 - \mathrm{Adam}\left(g_t + \mathrm{noise}\right)$$



100K $\xRightarrow{\text{becomes}}$



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

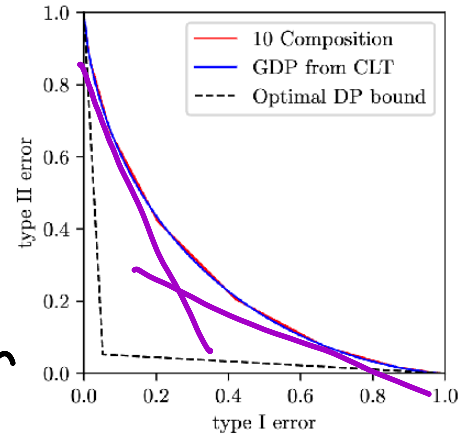
$\mu = \dfrac{2\sqrt{k}\kappa_1}{\kappa_1 - \kappa_2}$ for $\kappa_1 = -\int_0^1 \log|f'(x)|\,dx$ and $\kappa_2 = -\int_0^1 \log^2|f'(x)|\,dx$.

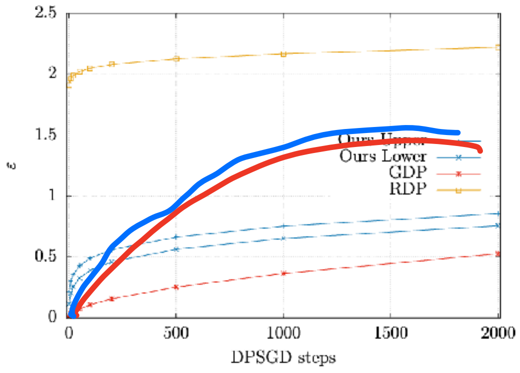
Gaussian Differential Privacy

Canonical f

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

- Canonical
- Composition
- tighter



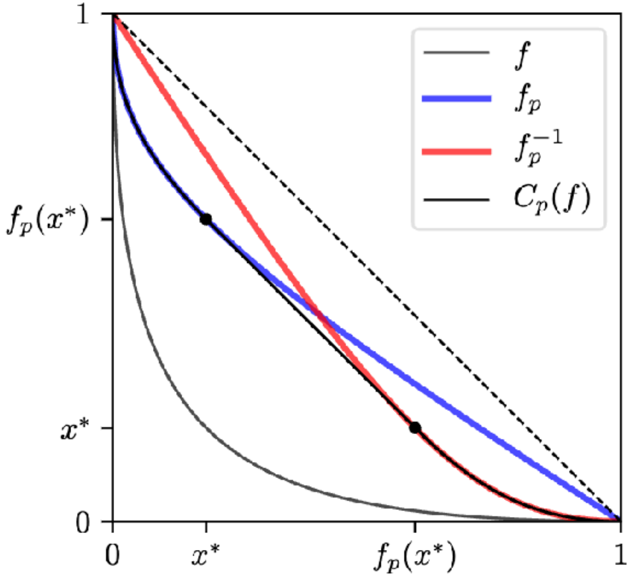


- In stats, combining may random variables ≈ Gaussian by CLT. In DP, composing many DP steps ≈gDP.

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

Gaussian Differential Privacy

Amplification by subsampling



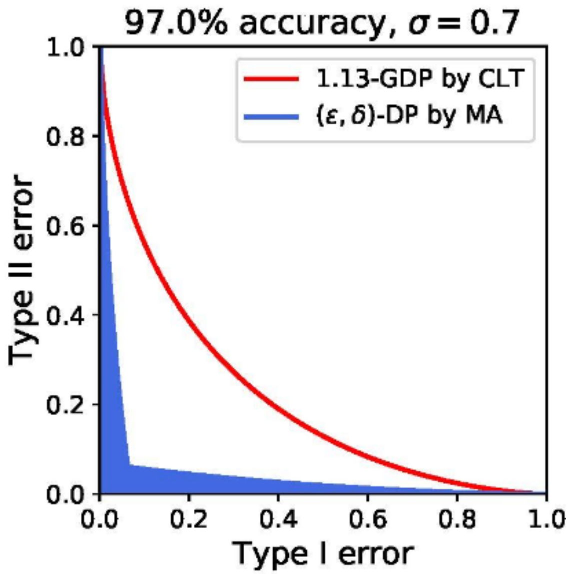
- Define $f_q(x) = qf(x) + (1-q)(1-x)$ and f_q^{-1}
- **Theorem 4.2** [DRS19]
  Composing q-sampling with f-DP, is $\left(\min(f_p, f_p^{-1})\right)^{**}$-DP
- Unfortunately, no closed form for GDP, compute numerically.

Private SGD

Using Gaussian-DP

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.



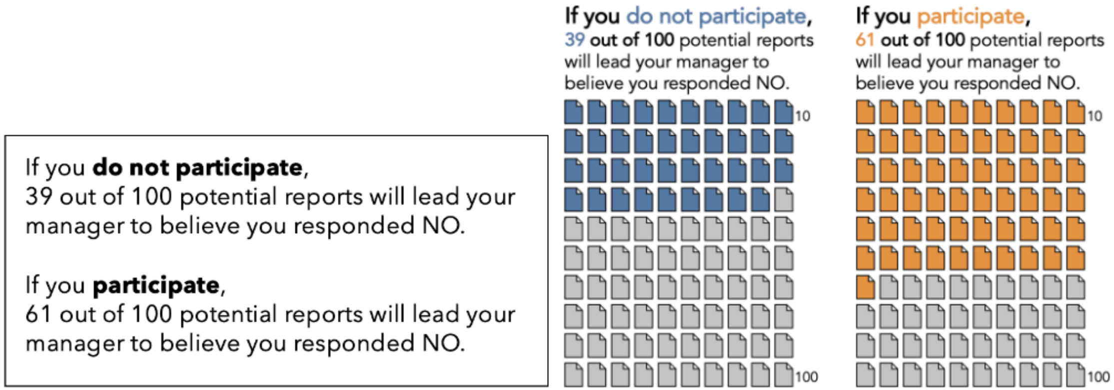
Tightest privacy bound [B+'20].
But, only asymptotically valid.

pe^{ε}

Aside: Communicating Privacy

$p \cdot e^{10}$

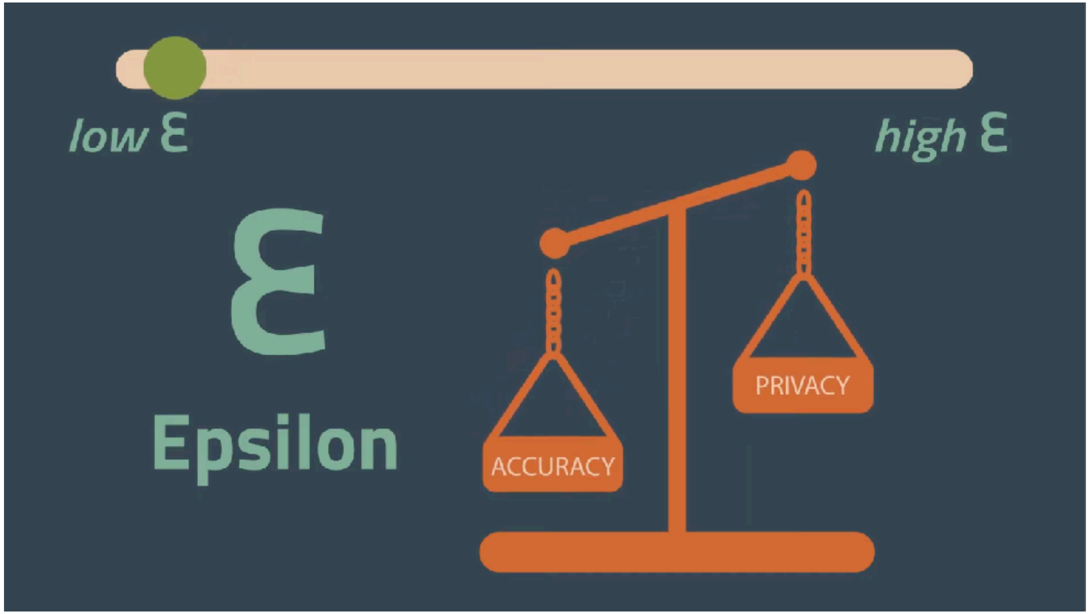
Odds ratio



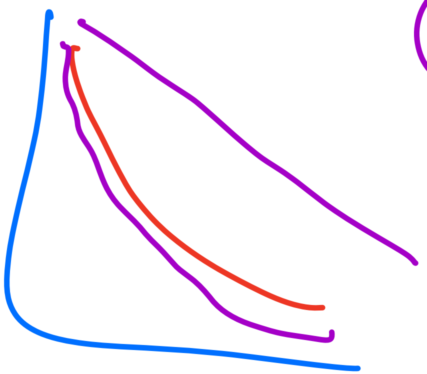
(a) ODDS-TEXT (b) ODDS-VIS

- 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



$\varepsilon = 10$

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

$x \leftrightarrow$ ML mod

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

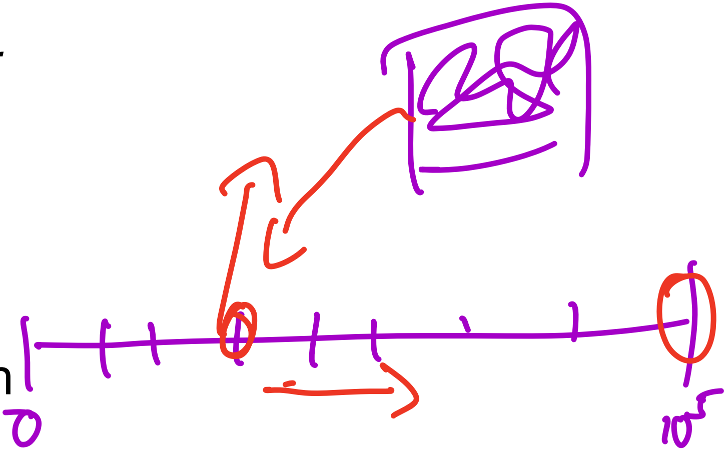
- Consider the following test:
  - D = MNIST dataset: 60k images
  - D’ = Add (x', y').
  - Train a CNN θ using [S+22] to get 0.98 acc and (0.21, 10–5)-DP.
  - Check $\ell_\theta(x', y') \leq \tau$. If D’ will be smaller.
  -

Privacy Auditing

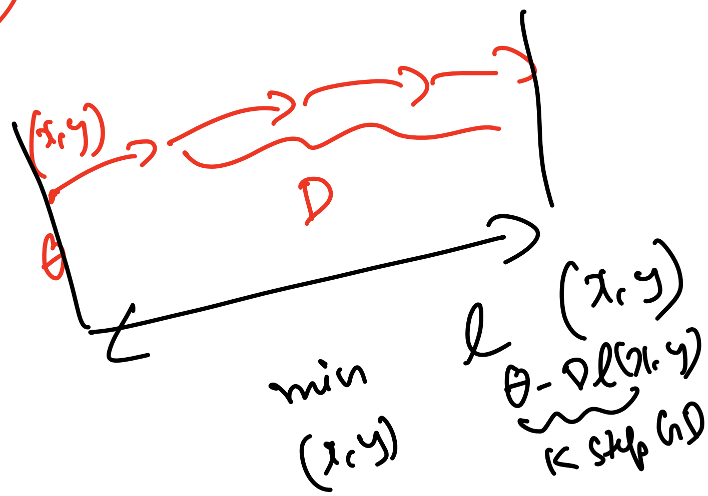
D – MNIST 60K

D' – 1 sample modified

- Some decisions to make
    - Which (x', y')? Called canary
    - insert an *unique* image which model is likely to memorize. i.e. insert a *backdoor attack*
    - Try a few images (~25) on an initial 2k training runs.
    - chose to insert a "checkerboard" pattern in x and incorrect label as y

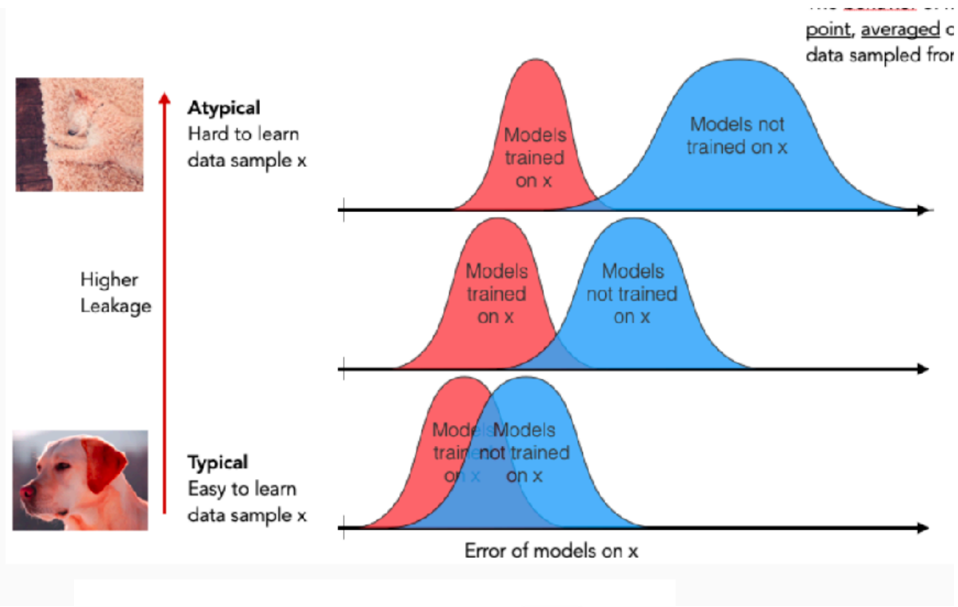


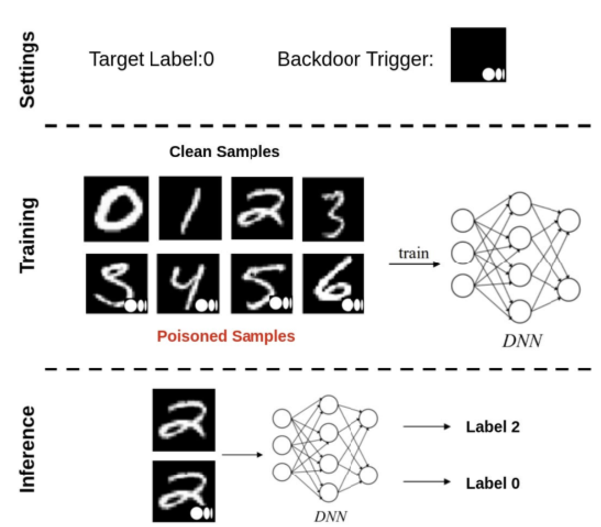
$$-\max_{(x,y)} \|\nabla \ell_\theta(x,y)\|^2$$



Privacy Auditing

- What makes a good canary?
    - Memorable to the model
    - “data poisoning” or “backdoor insertion” attacks make for great canaries

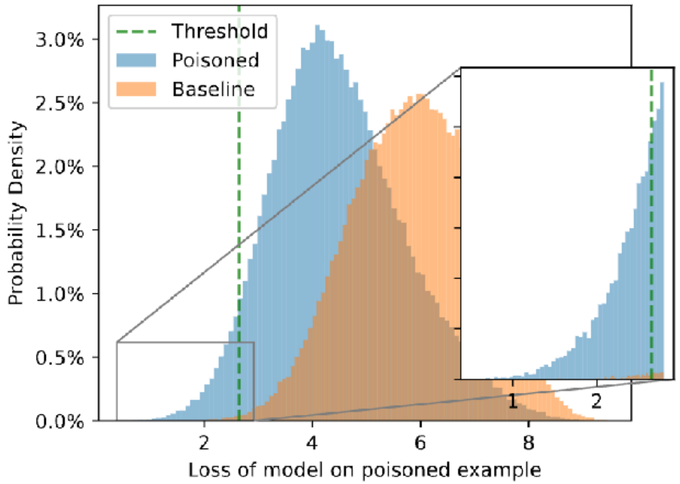




Privacy Auditing

D D'

- Some decisions to make
  - Measure loss on canary $\ell_\theta(x', y')$
  - Repeat 100k on D and 100k on D'.
  - Classify as D' if $\ell_\theta(x', y') \leq \tau$
  - Which τ? Pick best using validation training runs.



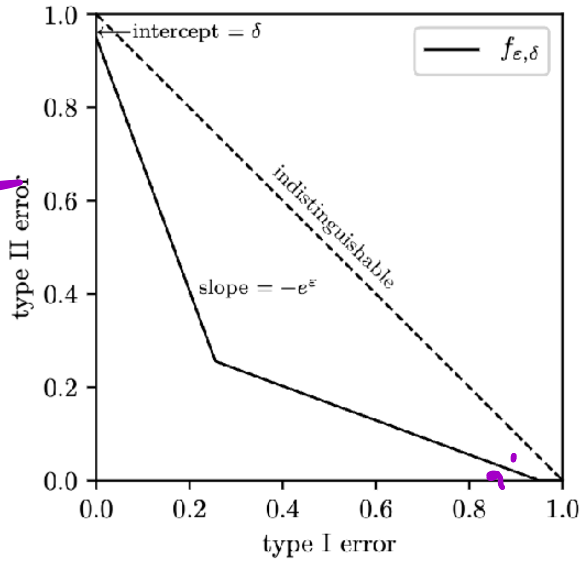
Privacy Auditing

$\alpha = 1 - 0.04922$

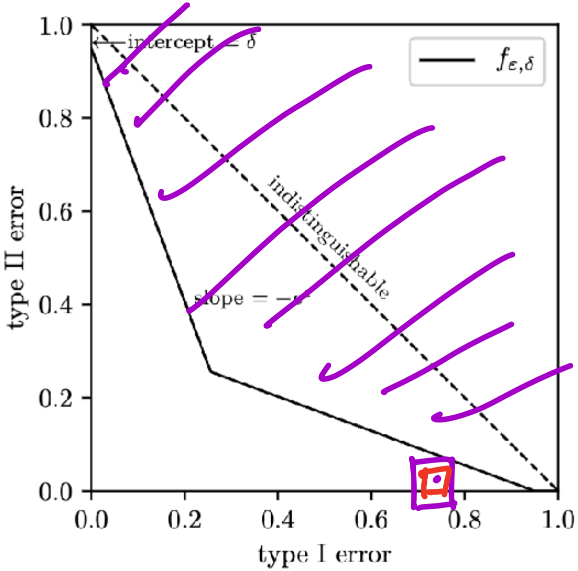
- Claimed privacy: (0.21, 10–5)-DP.
- With a threshold τ = 2.64 , attack had true positive rate of 4.922% and false positive rate of 0.174%.

$\beta = 0.0017$

- Is this possible?



Privacy Auditing



- We have claimed $\beta = 0.00174$ and α = 1- 4.922/100 = 0.95078.
- We have claimed privacy of (0.21, 10–5)-DP.
- $\beta \geq \max(1 - 10^{-5} - e^{0.21}0.95078\ ,\ (1 - 10^{-5} - 0.95078)/(e^{0.21})$ = 0.03988885074
- Can be due to sampling?

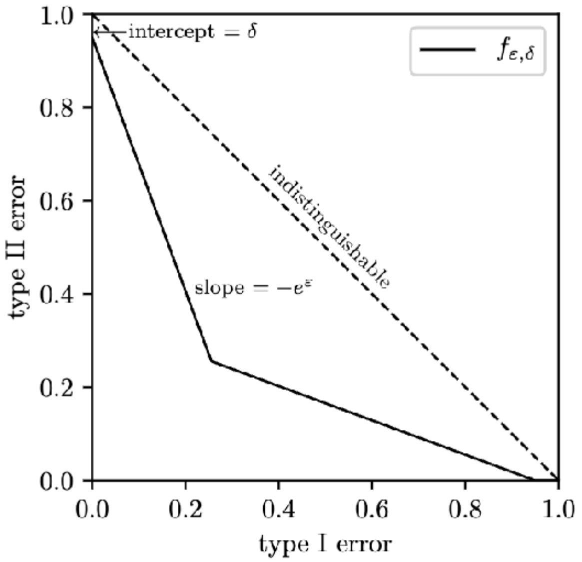
$\alpha \in (\alpha^-, \alpha^+)$

$1-\beta$,

$\beta \in [\beta^-, \beta^+]$

Privacy Auditing

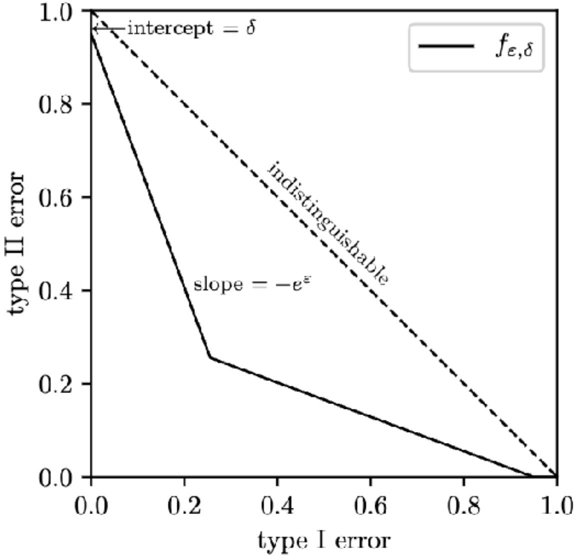
- Define $X = 1\{\text{predicted } D \,|\, \text{was } D'\}$ on a training run.
- False positive rate $\alpha = E[X]$ i.e. $X \sim \text{Ber}(\alpha)$
- We have 100k iid samples $X_1, \ldots, X_{100k} \sim \text{Ber}(\alpha)$
- How far can empirical $\hat{\alpha}$ and true α be?



Aside: Clopper-Pearson “exact” method

- $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i \sim$ Bern(α). α is unknown.
- Given Y for n observations, what can we say about α?
- Clopper-Pearson gives intervals $\alpha \in [\alpha^-, \alpha^+]$ with probability $\geq 1 - p$
- No closed form - need to compute numerically.

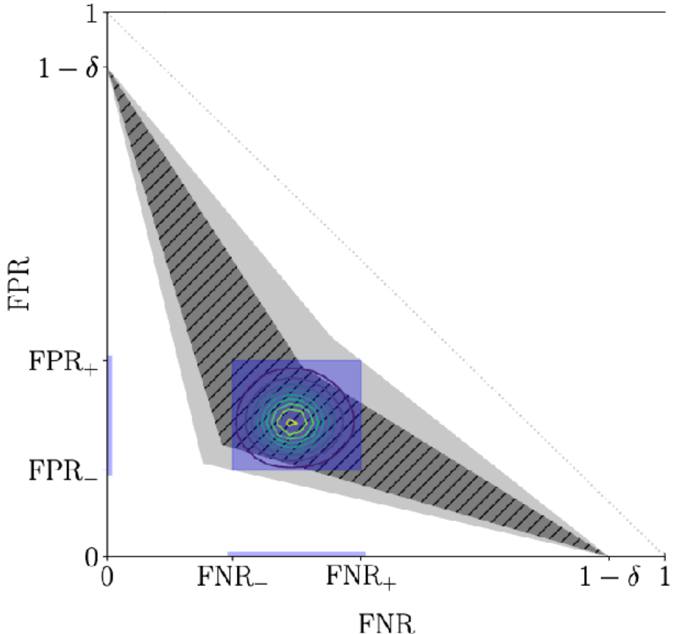
Privacy Auditing



- We have claimed $\beta = 0.00174$ and $\alpha = 1 - 4.922/100 = 0.95078$.
- We have claimed privacy of (0.21, 10–5)-DP.
- $\beta \geq \max(1 - 10^{-5} - e^{0.21} 0.95078\ ,\ (1 - 10^{-5} - 0.95078)/(e^{0.21})$ = 0.03988885074
- By Clopper-Pearson, $\alpha^+ \leq 0.95509$, $\beta^- \geq 0.00274$ with $p = 10^{-10}$
- Later, they found a bug and retracted the paper. Very common in DP!!

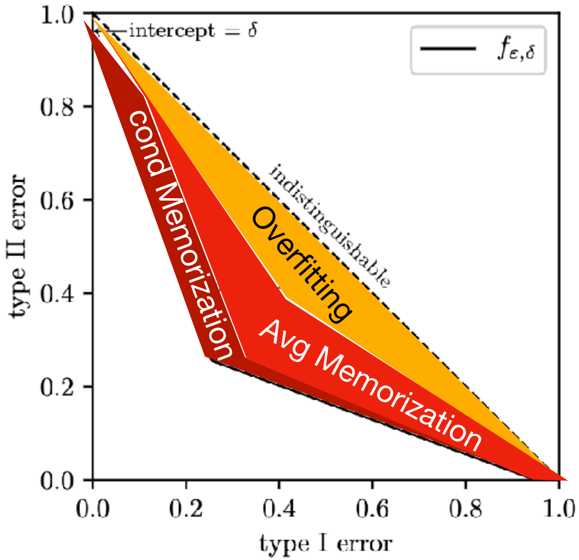
Improvements: better stats

- Do we really need α^+, β^-? vs ε
  - Directly bound $\log(\frac{1-\delta-\beta}{\alpha})$ using *Log-Katz confidence intervals*.
- Incorporate priors [ZB+23]:
  - Use Bayesian approach
  - Compute joint posterior of $\alpha, \beta, \varepsilon$
- Your favorite stats trick



Improvements: picking canaries

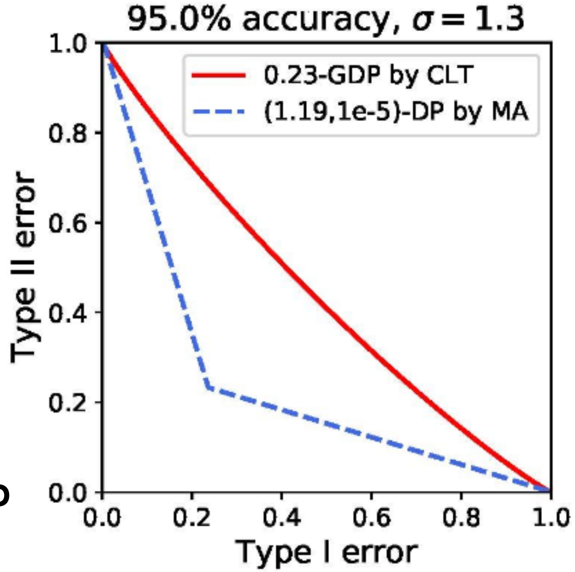
- Picking the right (x', y') is an art
  - Very similar to backdoor attacks
- Goal is to test for *conditional memorization*
- Means searching for a "planted signal"
  - when detected, we are sure. i.e. low type I
  - but can miss a lot i.e. high type II
  - what if $\delta \geq \alpha$?

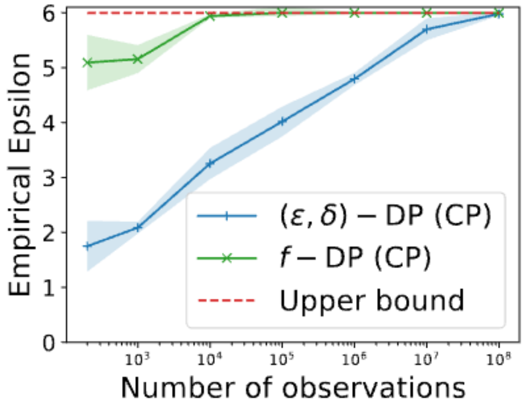


Gaussian Membership Inference

More improvements

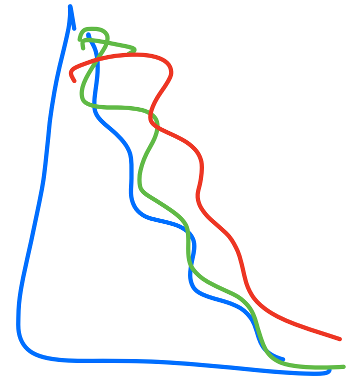
- Test for GDP instead:
  - Suppose some Gaussian mechanism claims (ε, δ)-DP
  - Calculate corresponding μ-GDP
  - Check if empirical α, β allows such μ
    $\mu^{-} = \Phi^{-1}(1 - \alpha^{+}) - \Phi^{-1}(\beta^{-})$
  - Reduces number of runs by 10,000x [N+23]





(d) $\varepsilon = 6$

$$\underbrace{\theta_{t+1} = \theta_t - \nabla \ell(\theta_t)}_{\text{repeat } k\text{-times}}$$



(ε, δ)