

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]

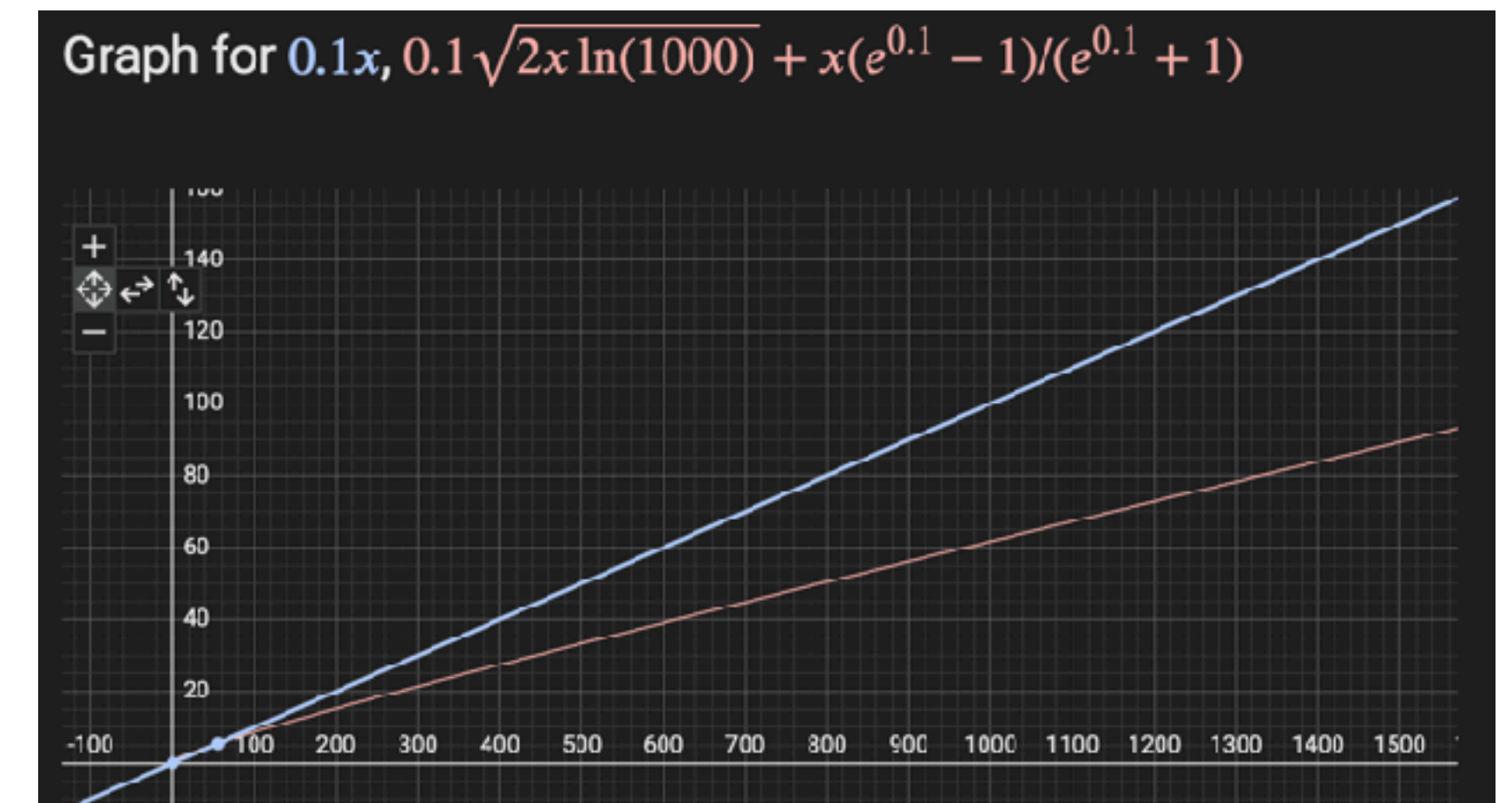
Let us draw a variable $t \sim A(D)$. Then the **privacy loss random variable**.

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

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

Recap

- 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 $(\tilde{\epsilon}, \tilde{\delta})$ -DP algorithm where

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

Recap

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

Recap

Poisson subsampling disadvantages

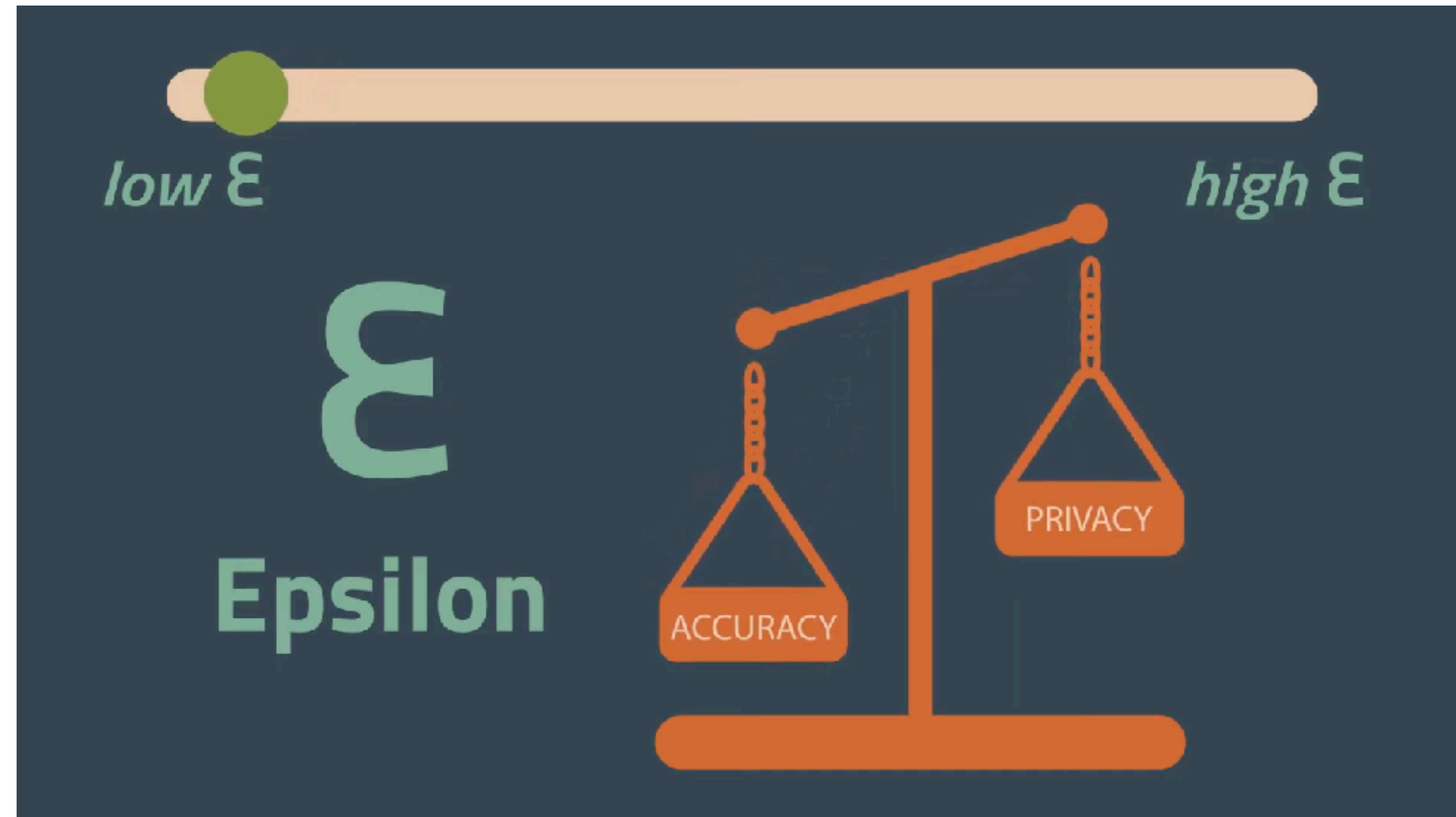
- $\theta_t = \theta_{t-1} - \gamma \left(\left[\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \text{Clip}_\tau \left(\nabla_\theta \ell(f(x_t; \theta), y_i) \right) \right] + \mathcal{N}(0, \tau^2 \rho^2) \right)$
- I **cannot** set $\rho \propto |\mathcal{B}|^{-1}$ - mechanism **cannot be data-dependent**.
It should work for the **worst case** i.e. when $|\mathcal{B}| = 1$.

Agenda for today

Analyzing privacy of ML training

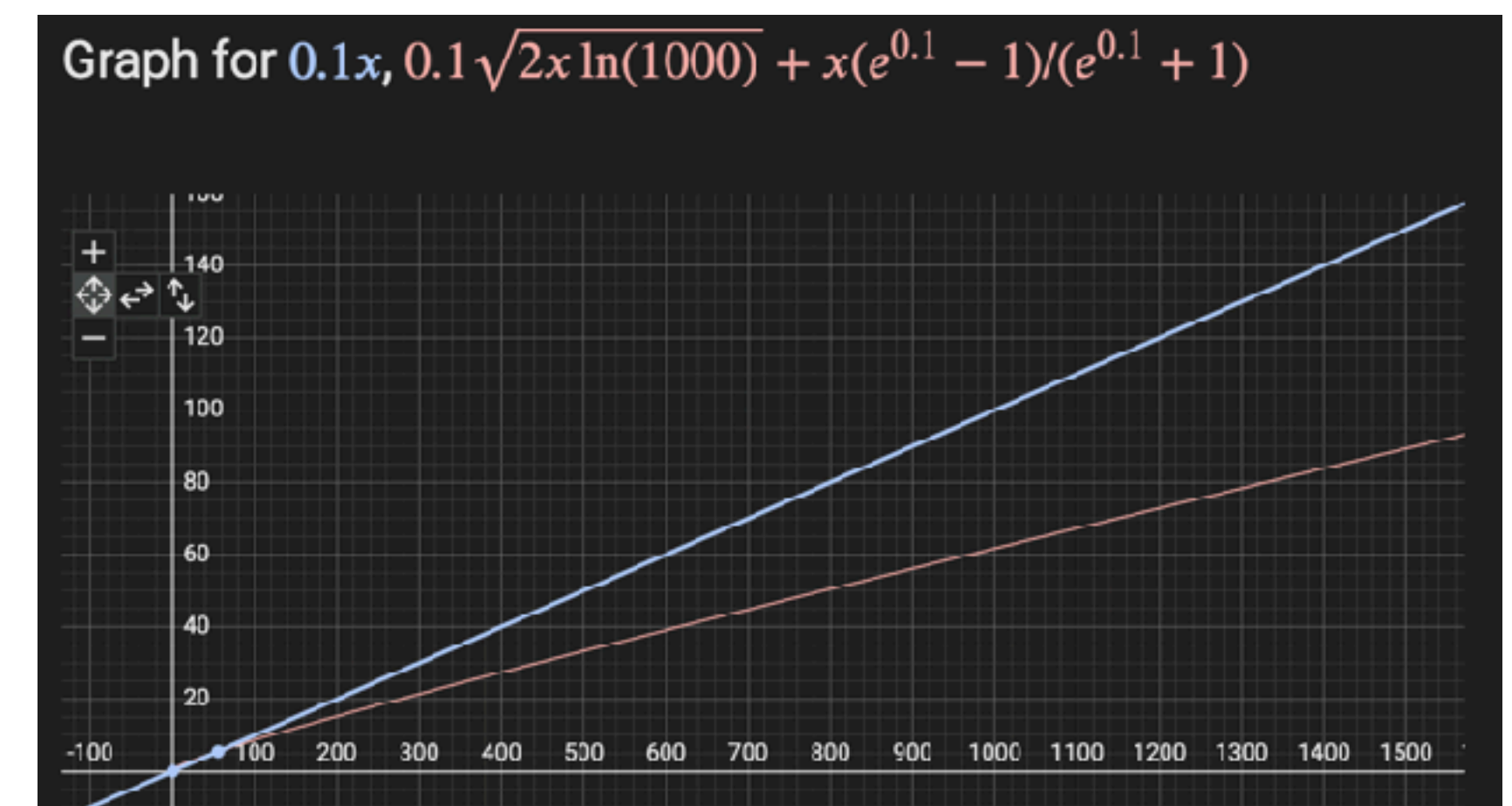
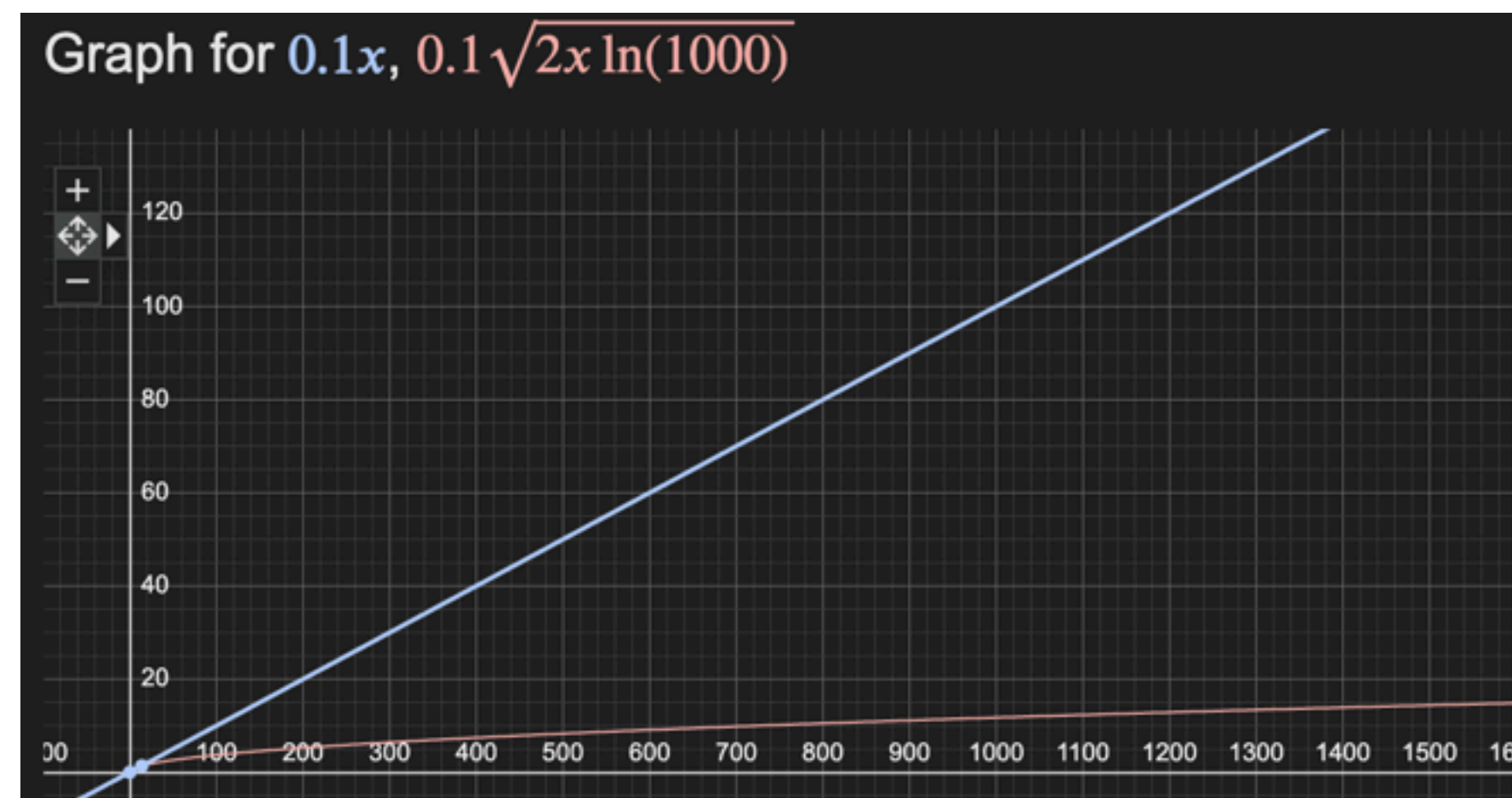
- Improving composition
- Gaussian DP
- Privacy Auditing
- HW1 solutions

Better composition



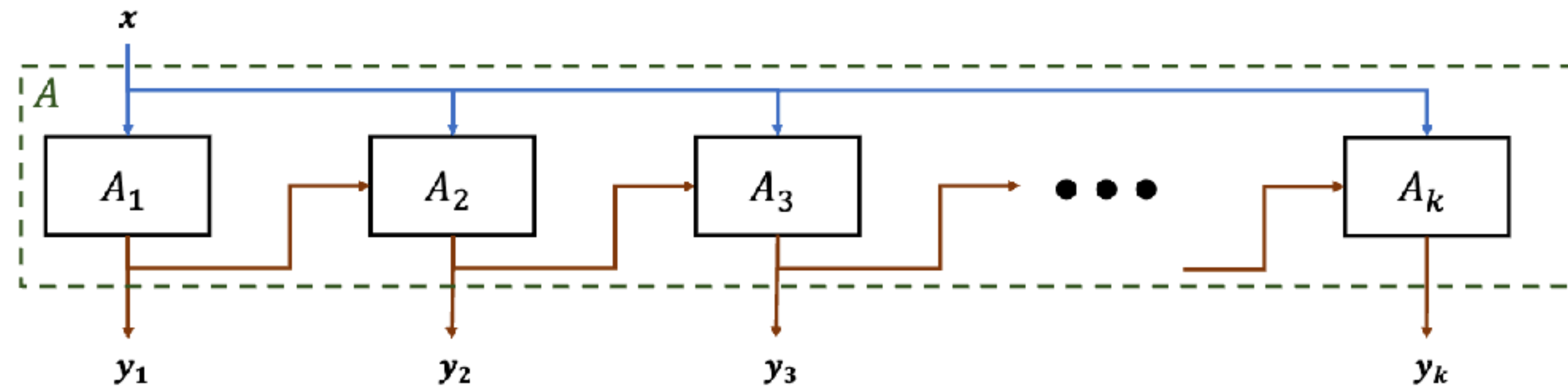
Approximate DP analysis is loose

- 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 loose.



Advanced composition

Proof sketch



- Privacy random variable of composition:

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

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

Advanced composition

Proof sketch

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

Advanced composition

Proof sketch

- KL-divergence between two Gaussian distributions with different means

- $D_{\text{KL}}(\mathcal{N}(\mu_1, \sigma^2) \parallel \mathcal{N}(\mu_2, \sigma^2)) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}.$

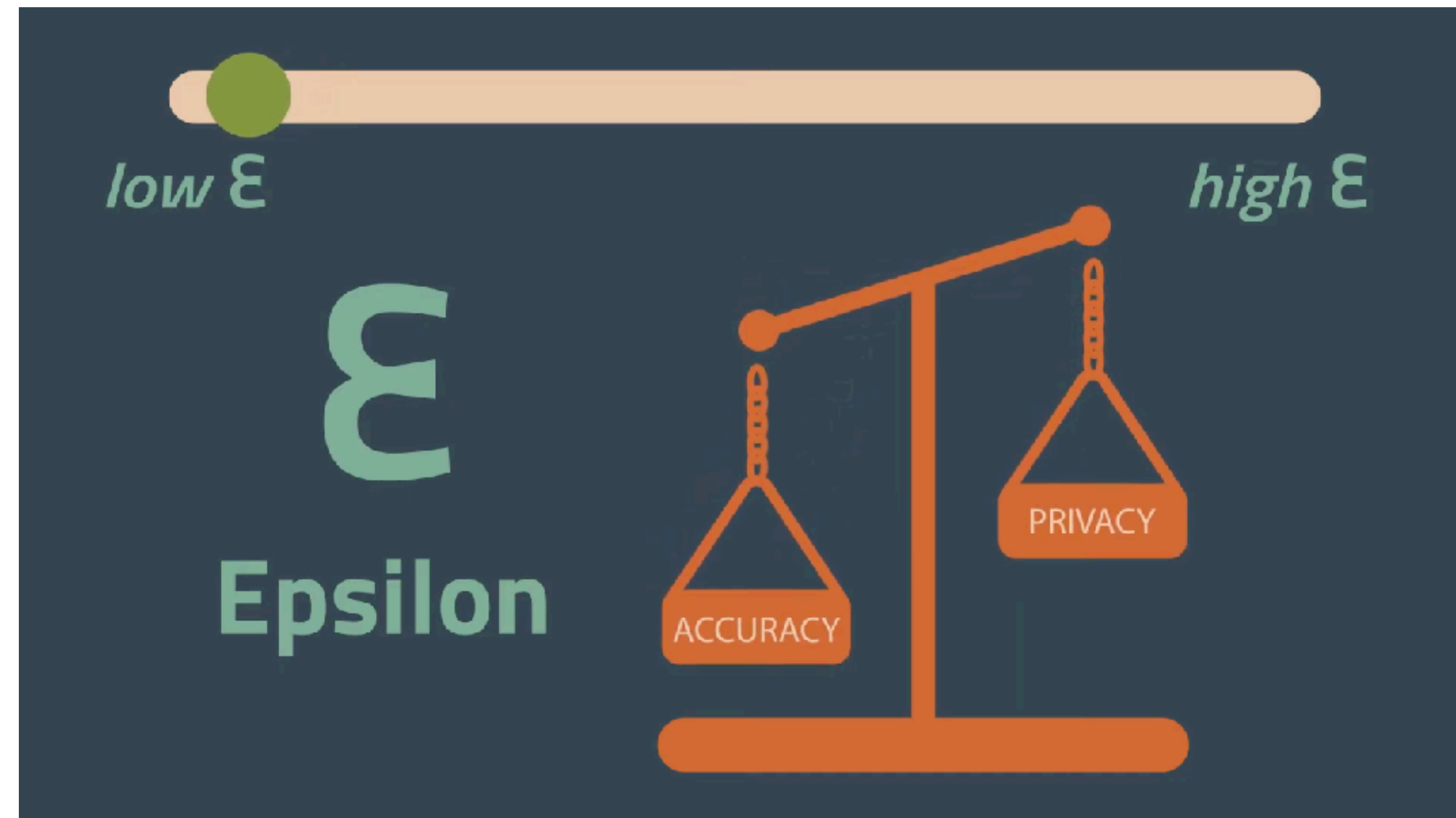
- $= O(\varepsilon^2)$ since recall $\sigma = \frac{\Delta_2 \sqrt{2 \ln(1.25/\delta)}}{\varepsilon}$

- After k rounds, $O(\varepsilon\sqrt{k} + \varepsilon^2 k)$. Sufficient to set $\varepsilon = 1/\sqrt{k}$!

Advanced composition

Geometric intuition for gaussians

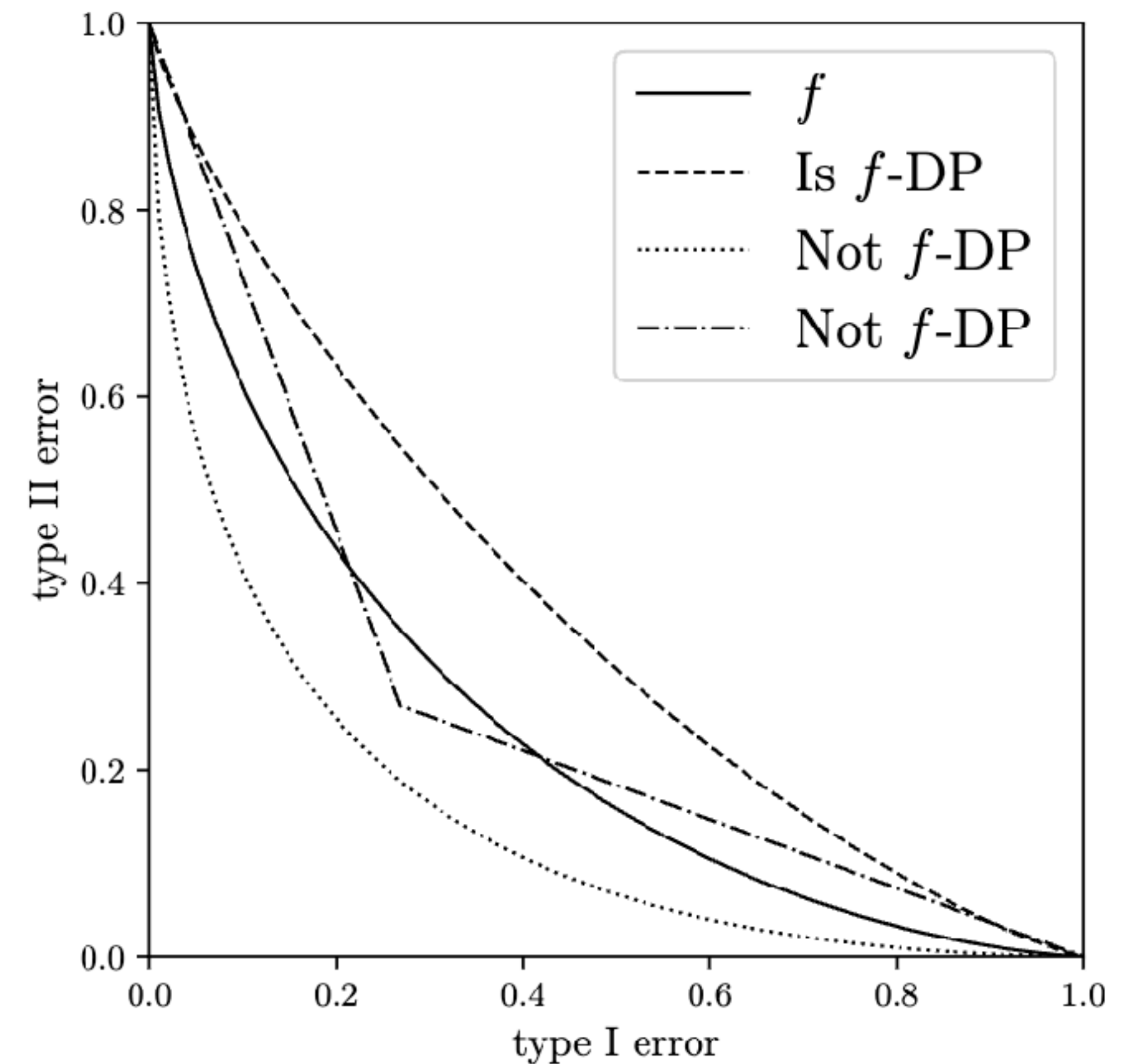
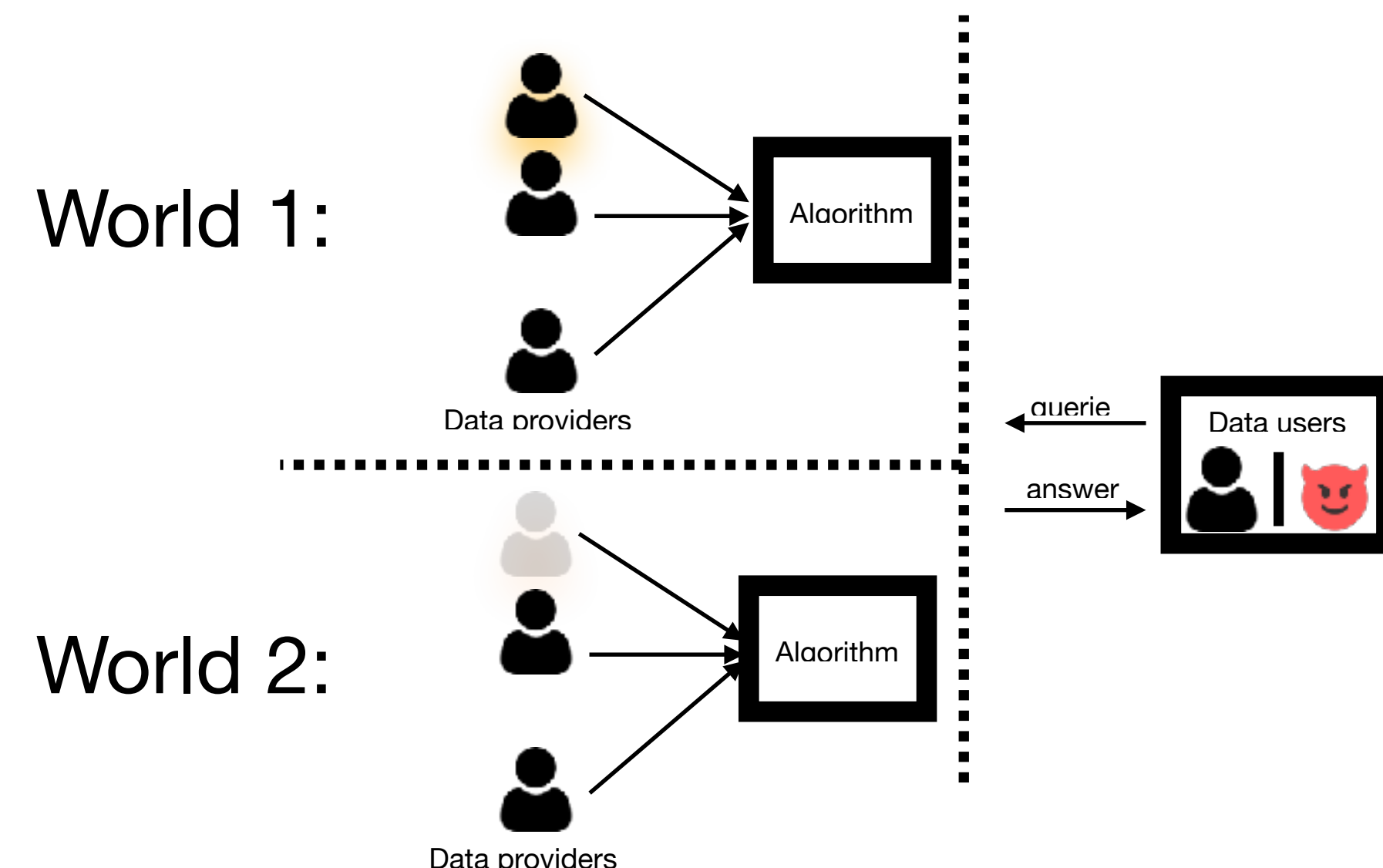
Gaussian Differential Privacy



f -DP

Most general privacy definition

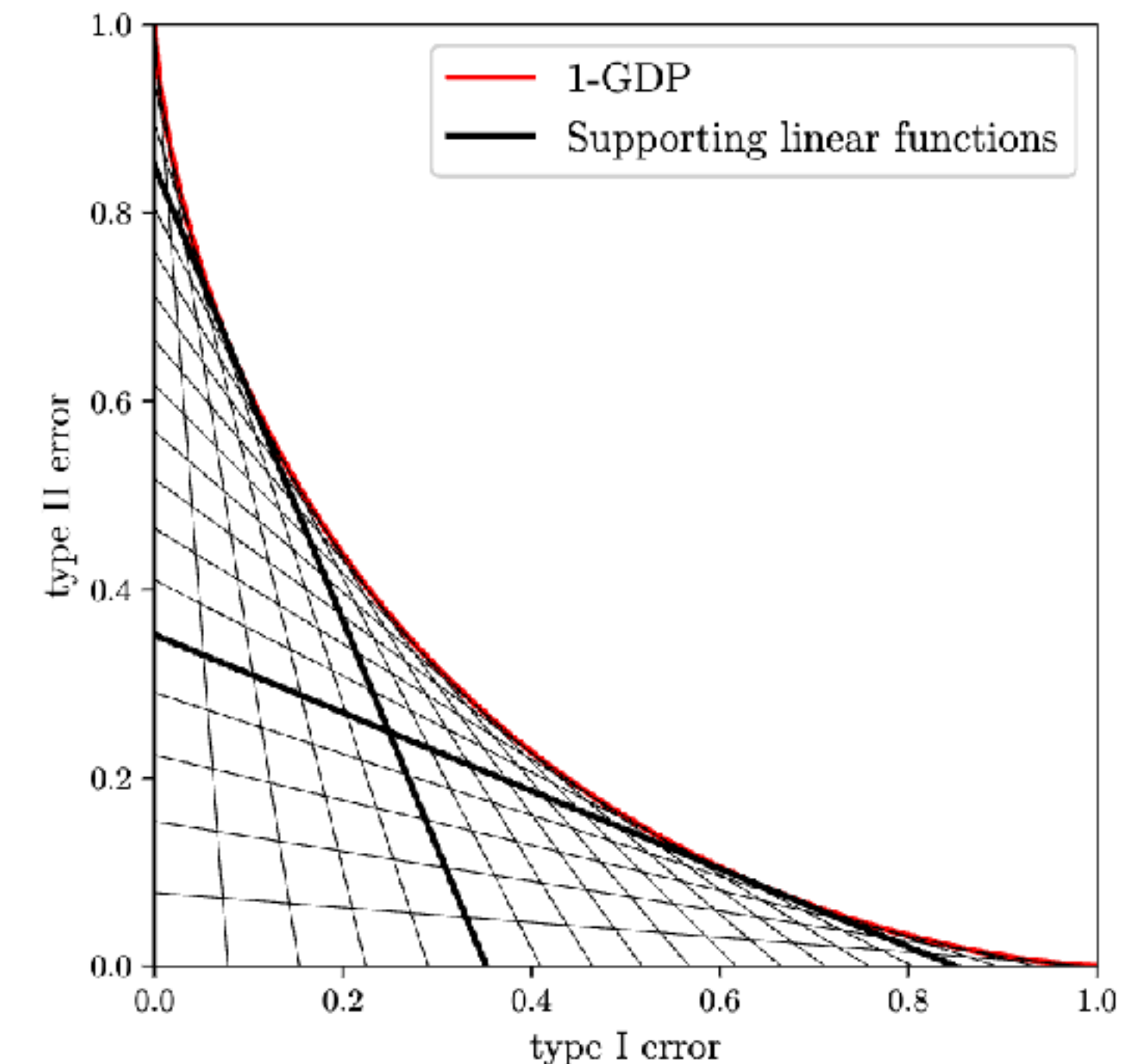
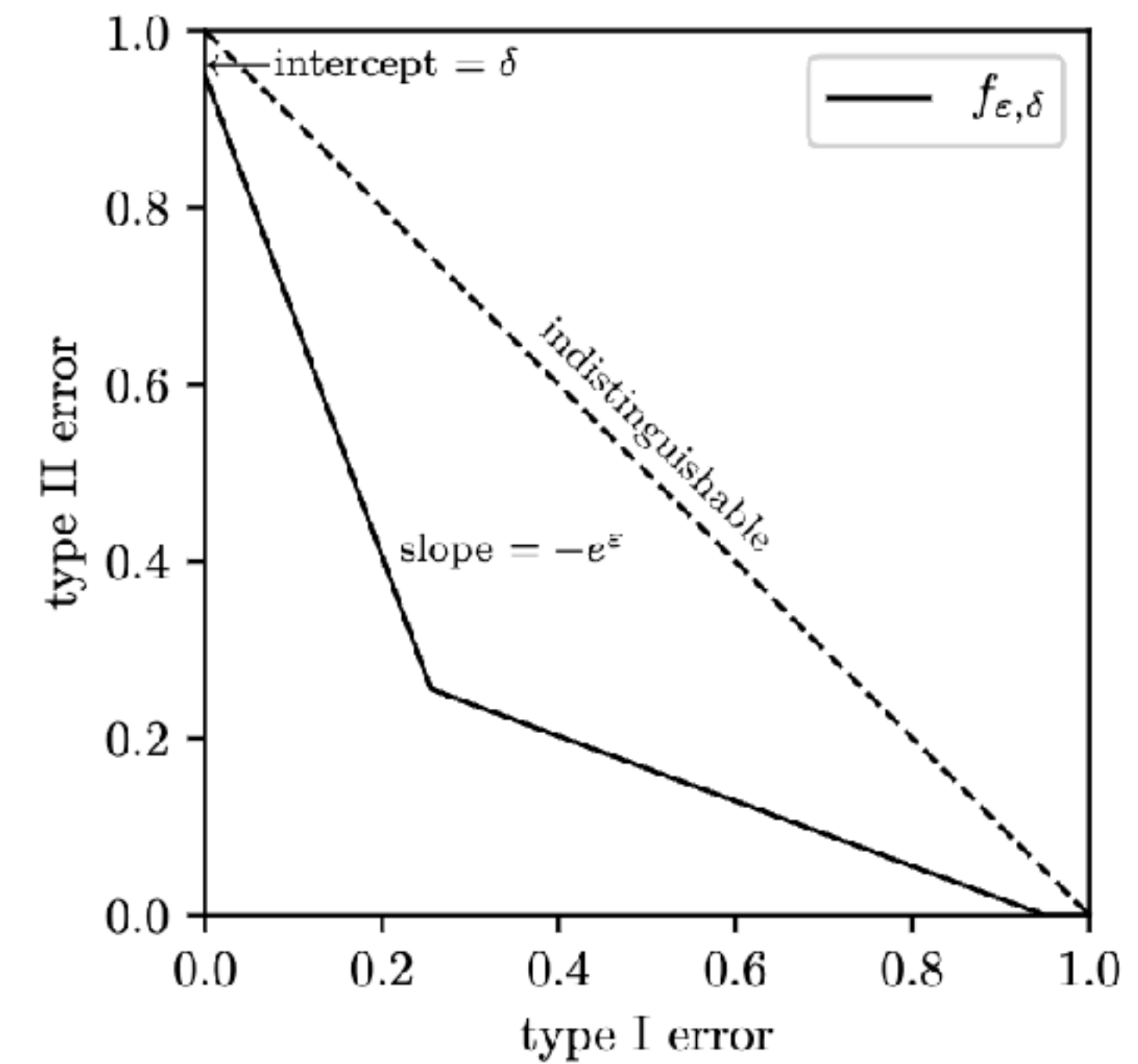
- **Definition.** Given a function f , we say an algorithm is f -DP if the tradeoff curve of an optimal distinguisher is strictly above f .



f-DP

Generalization (ϵ, δ) -DP

- **Prop 2.5 [WZ10].** A is (ϵ, δ) -DP iff it satisfies $f_{\epsilon, \delta}$ -DP for
$$f_{\epsilon, \delta} = \max(1 - \delta - e^\epsilon x, (1 - \delta - x)/e^\epsilon)$$
- **Prop 2.12 [DRS19]** A is f -DP iff it satisfies $(\epsilon, \delta_f(\epsilon))$ -DP for $\forall \epsilon \geq 0$ and
$$\delta_f(\epsilon) = 1 + f^*(-e^\epsilon).$$

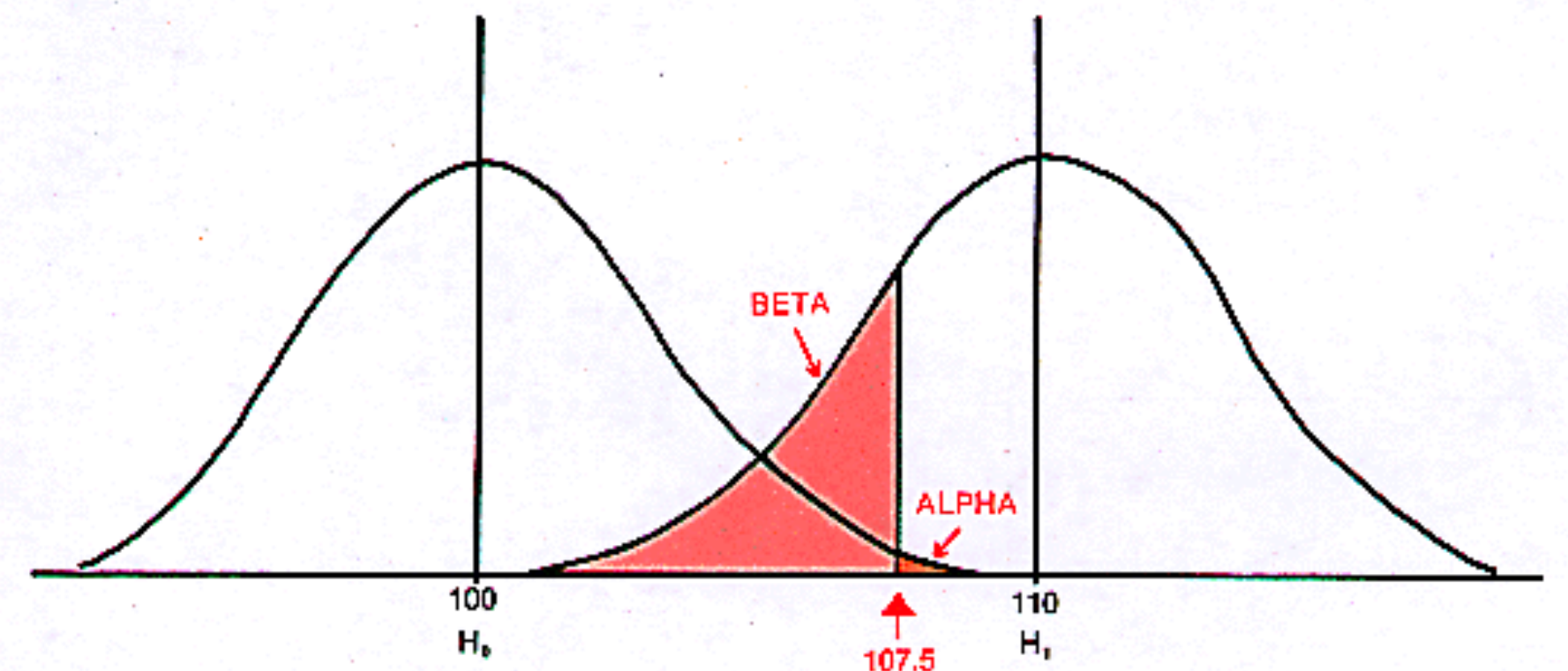
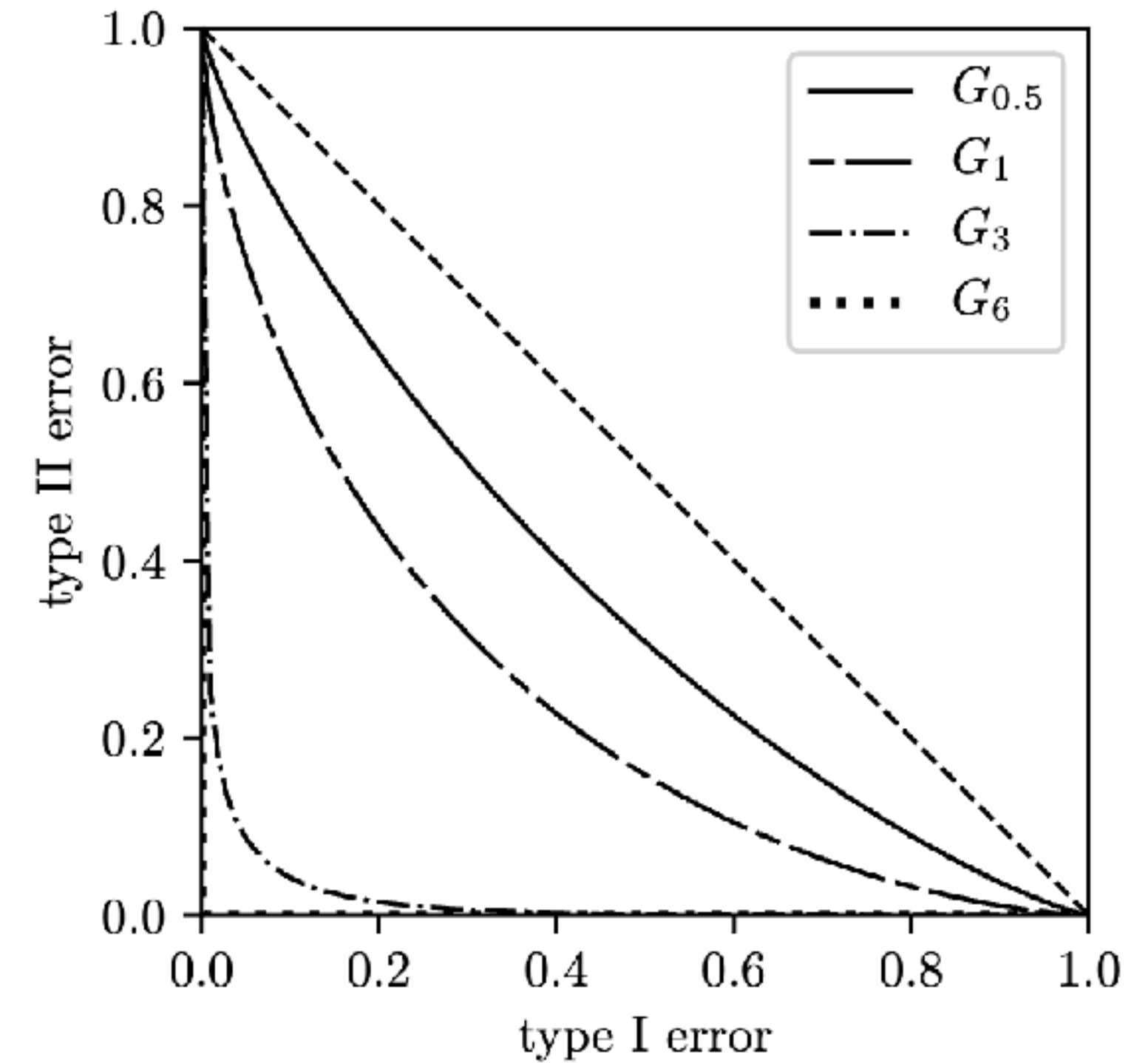


Gaussian-DP

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

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

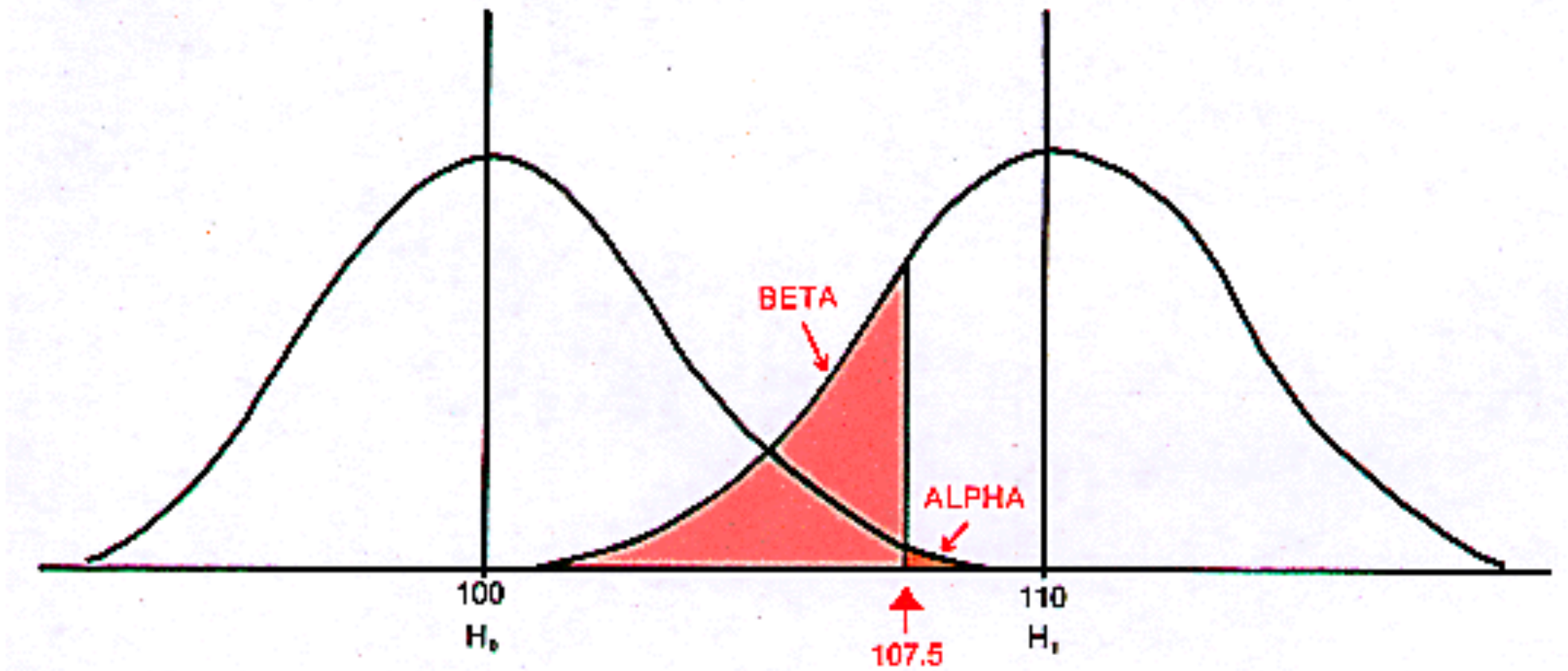
- $\alpha(\tau) = 1 - \Phi(\tau)$ and $\beta(\tau) = \Phi(\tau - \mu)$



Gaussian-DP

Gaussian mechanism

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



Theorem. Gaussian mechanism

Given $f : \mathcal{X}^n \rightarrow \mathbb{R}^d$ 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 $A_1 \circ A_2 \dots \circ A_k$, each of which is μ_i -GDP is $\sqrt{\sum_{i=1}^k \mu_i^2}$ -GDP.

Gaussian Differential Privacy

Canonical f

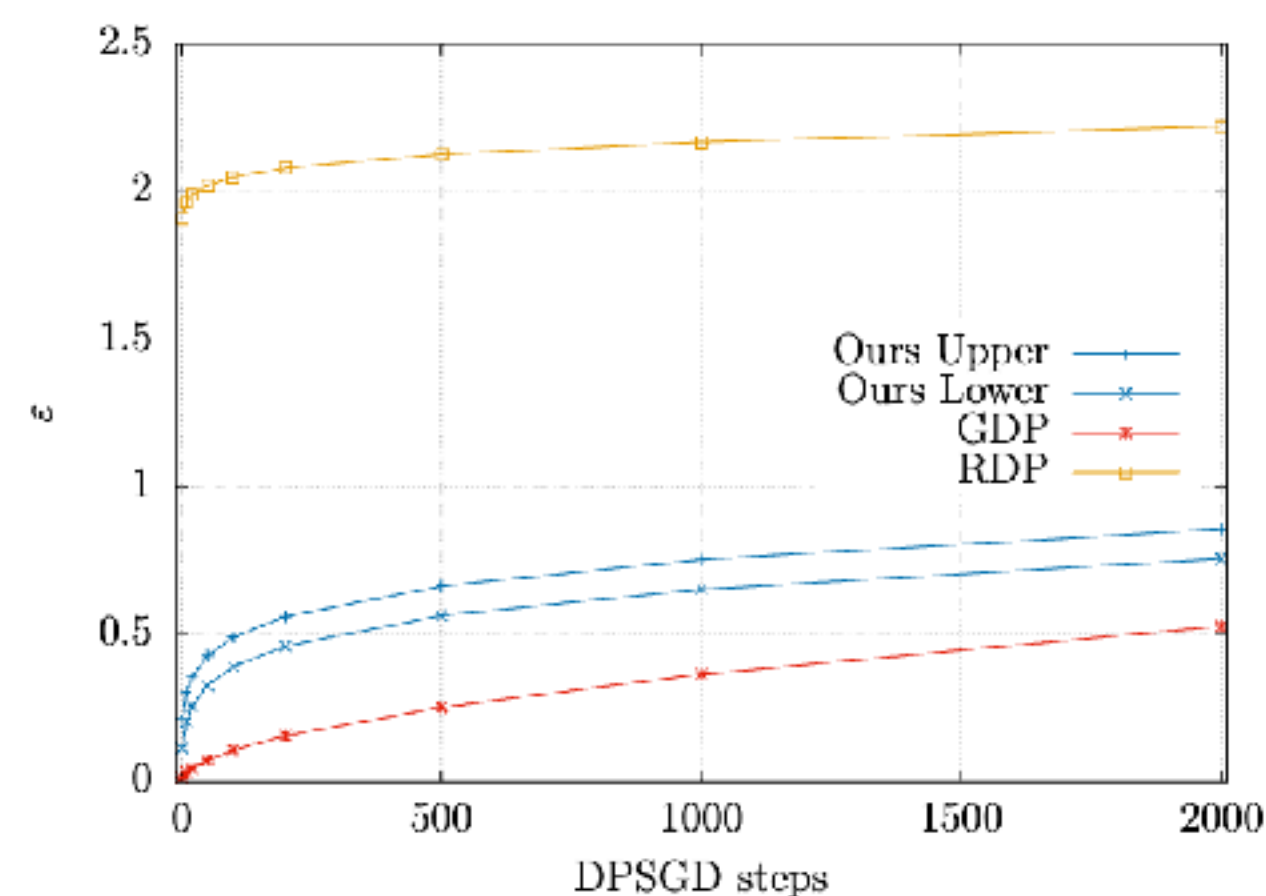
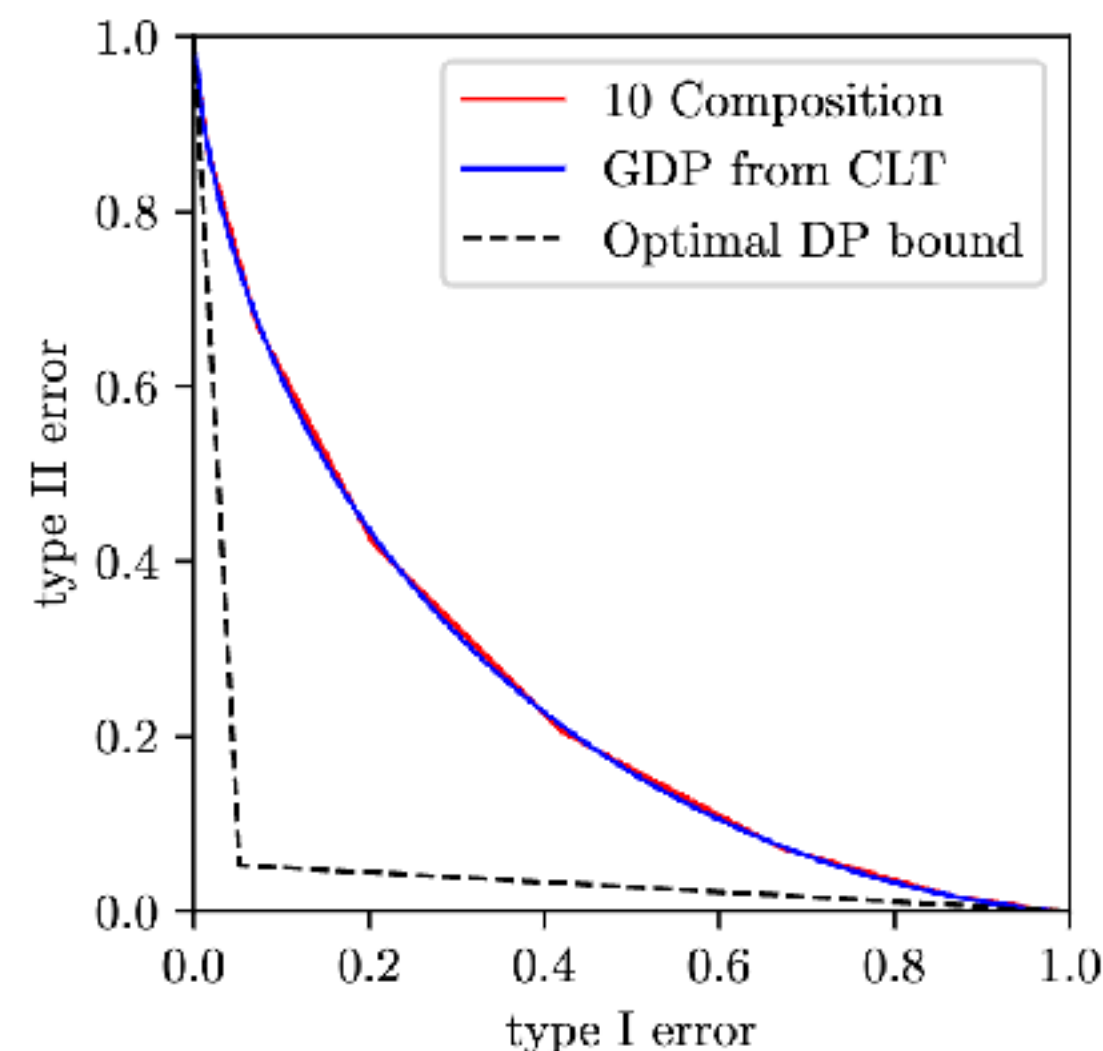
Theorem 3.4 [DRS19] Central limit theorem of composition

Given some regularity assumptions, composition of $A_1 \circ A_2 \dots \circ A_k$, each of which is f_i -DP is approximately μ -GDP for

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

Gaussian Differential Privacy

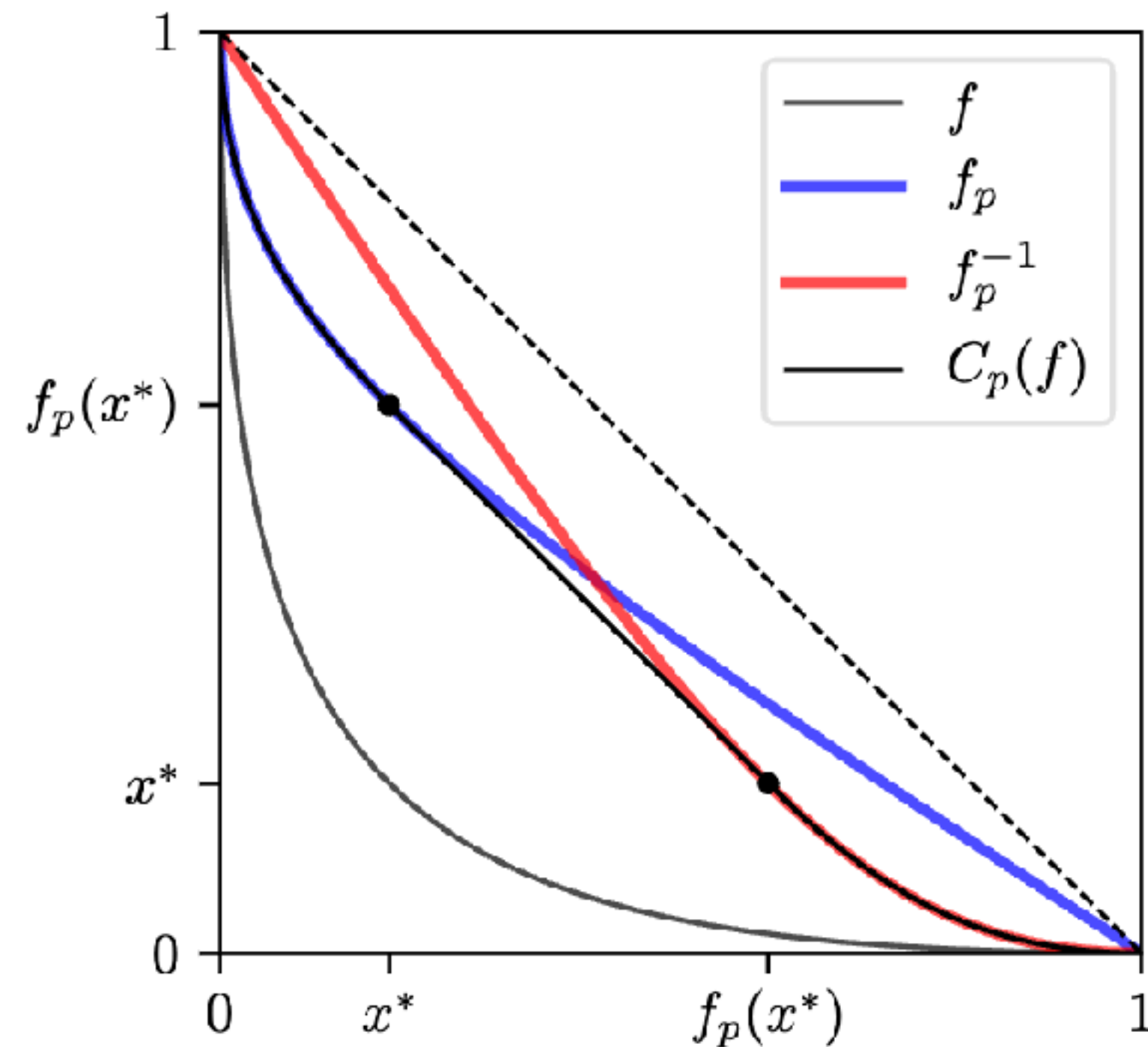
Canonical f



- In stats, combining many random variables \approx Gaussian by CLT. In DP, composing many DP steps \approx 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 $(\min(f_p, f_p^{-1}))^{**}$ -DP
- Unfortunately, no closed form for GDP, compute numerically.

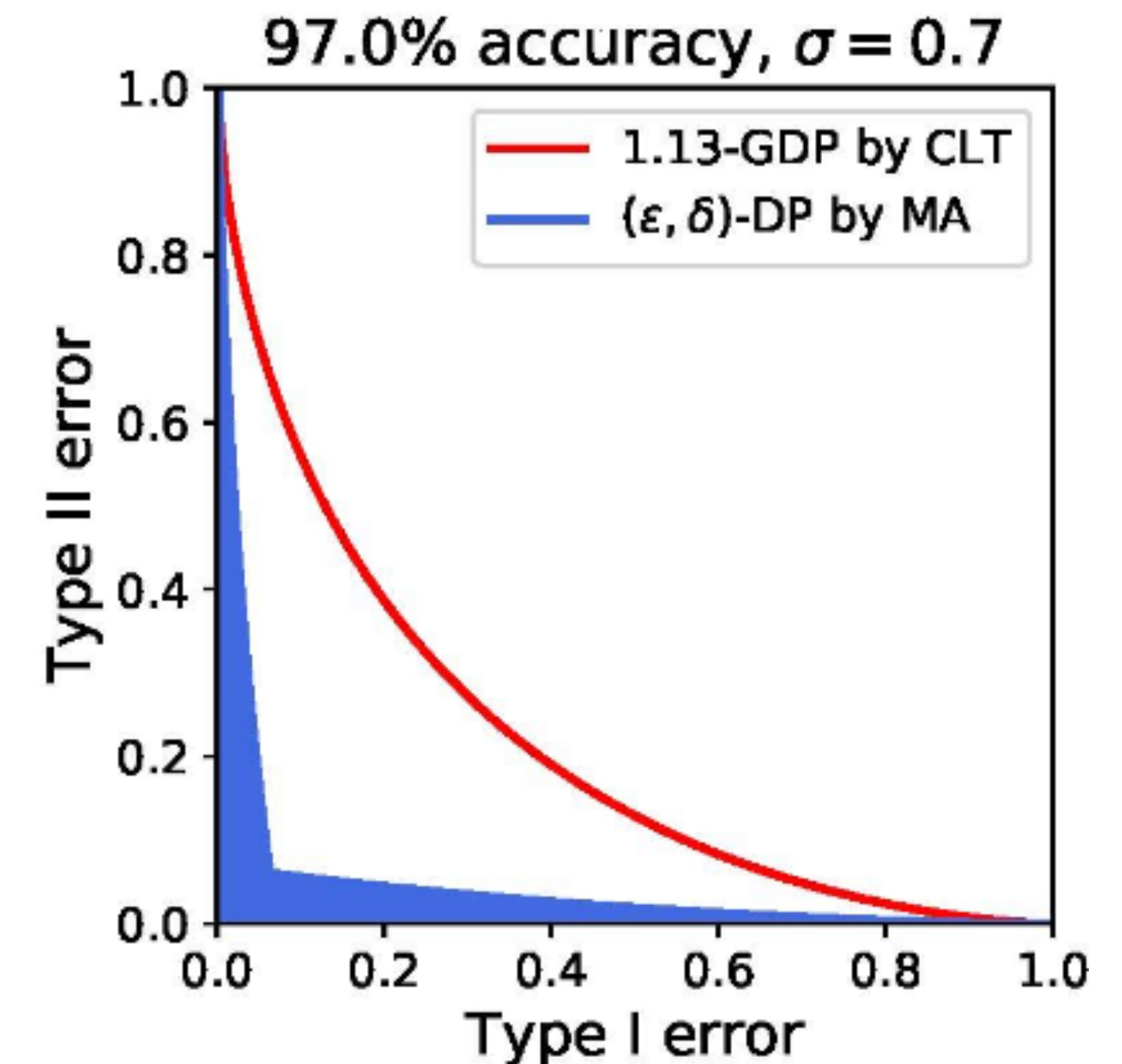
Private SGD

Using Gaussian-DP

Corollary 5.4 [DRS19] Subsampled Composition

Suppose each A_i is μ -GDP. Then, composing q -sampled A_i is asymptotically

$$\left(q\sqrt{k}\sqrt{e^{\mu^2}\Phi(3\mu/2) + 3\Phi(-\mu/2) - 2} \right)\text{-GDP.}$$



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

Aside: Communicating Privacy

Odds ratio

If you **do not participate**,
39 out of 100 potential reports will lead your manager to believe you responded NO.

If you **participate**,
61 out of 100 potential reports will lead your manager to believe you responded NO.

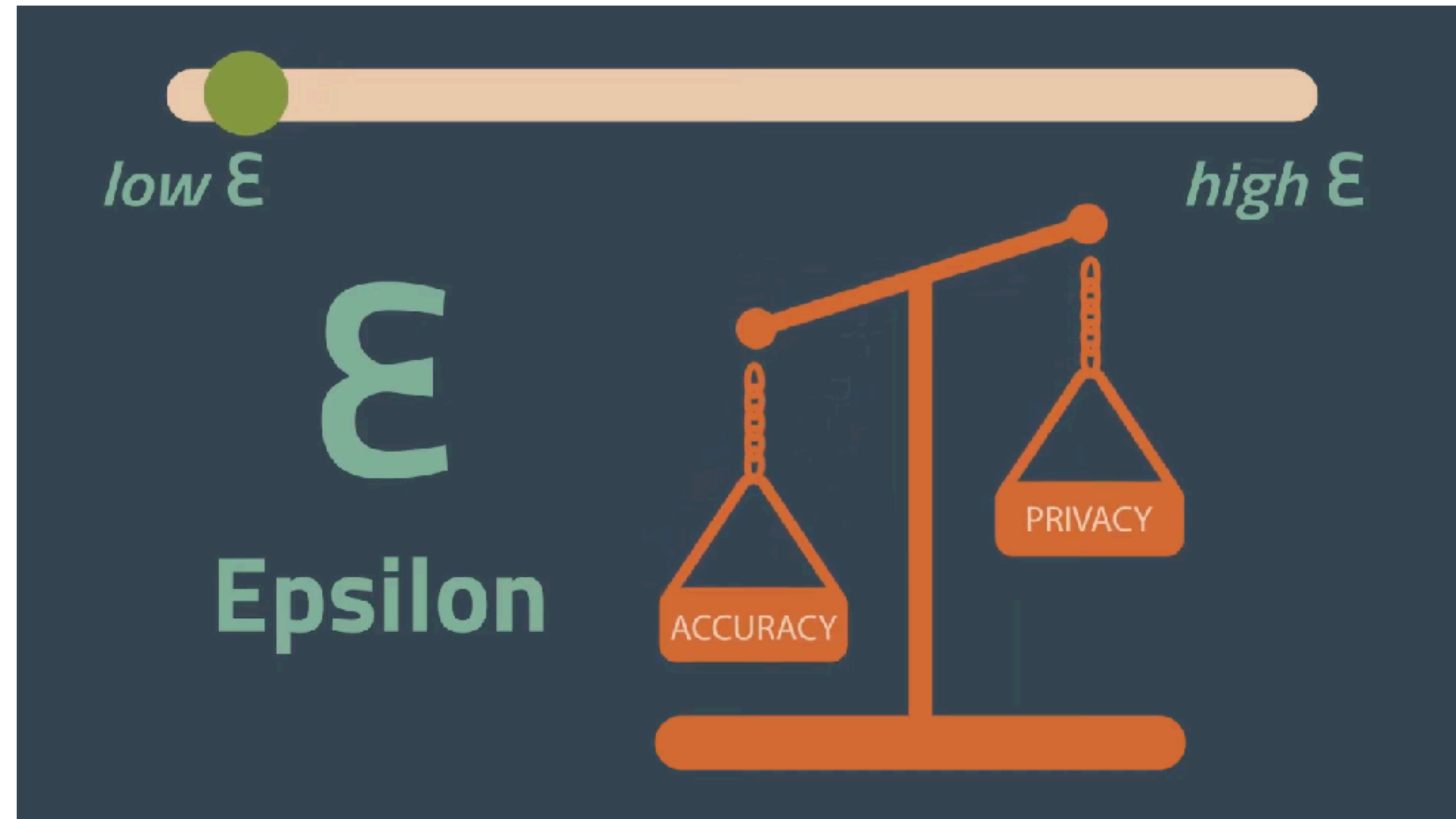
(a) ODDS-TEXT



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

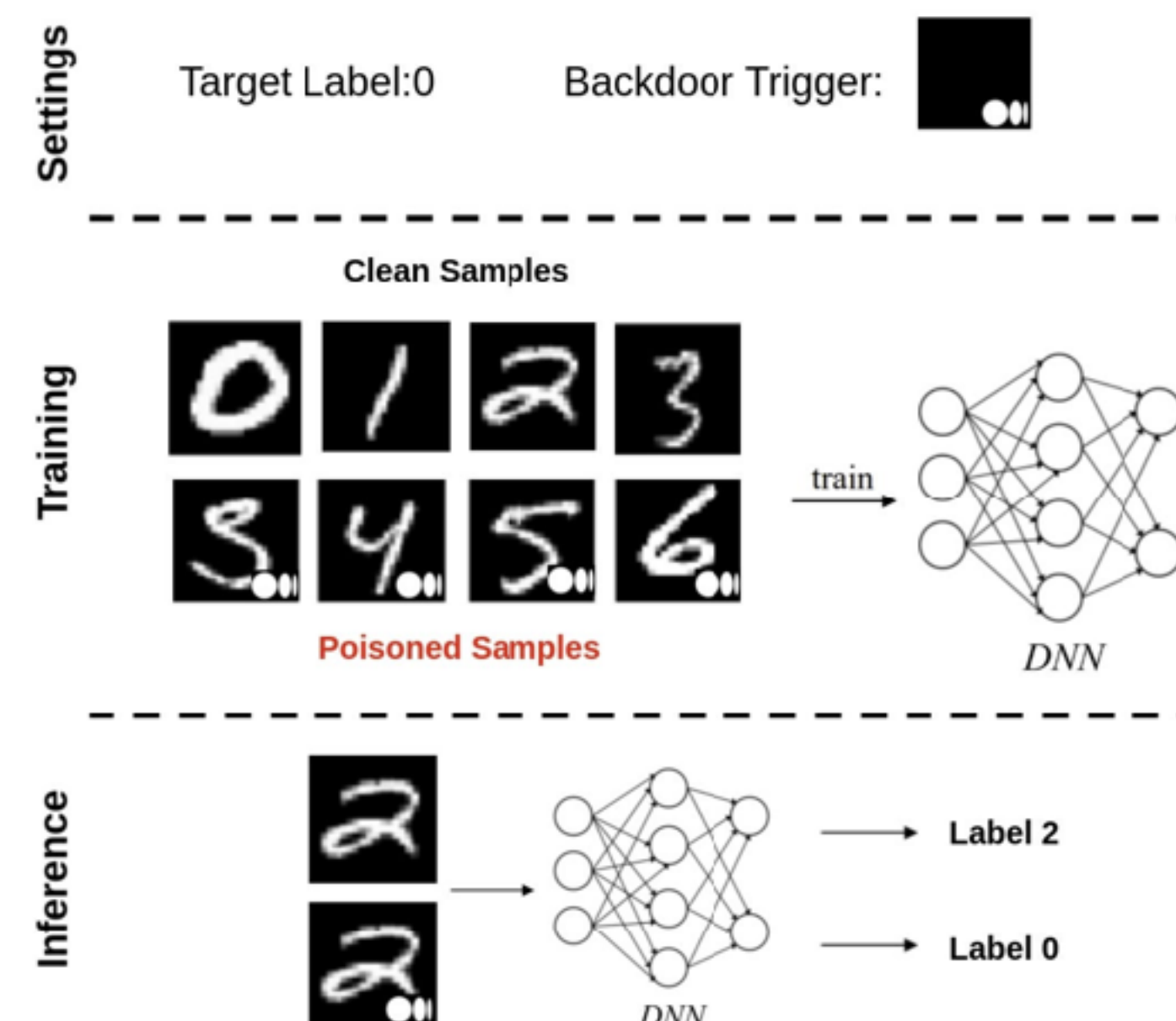
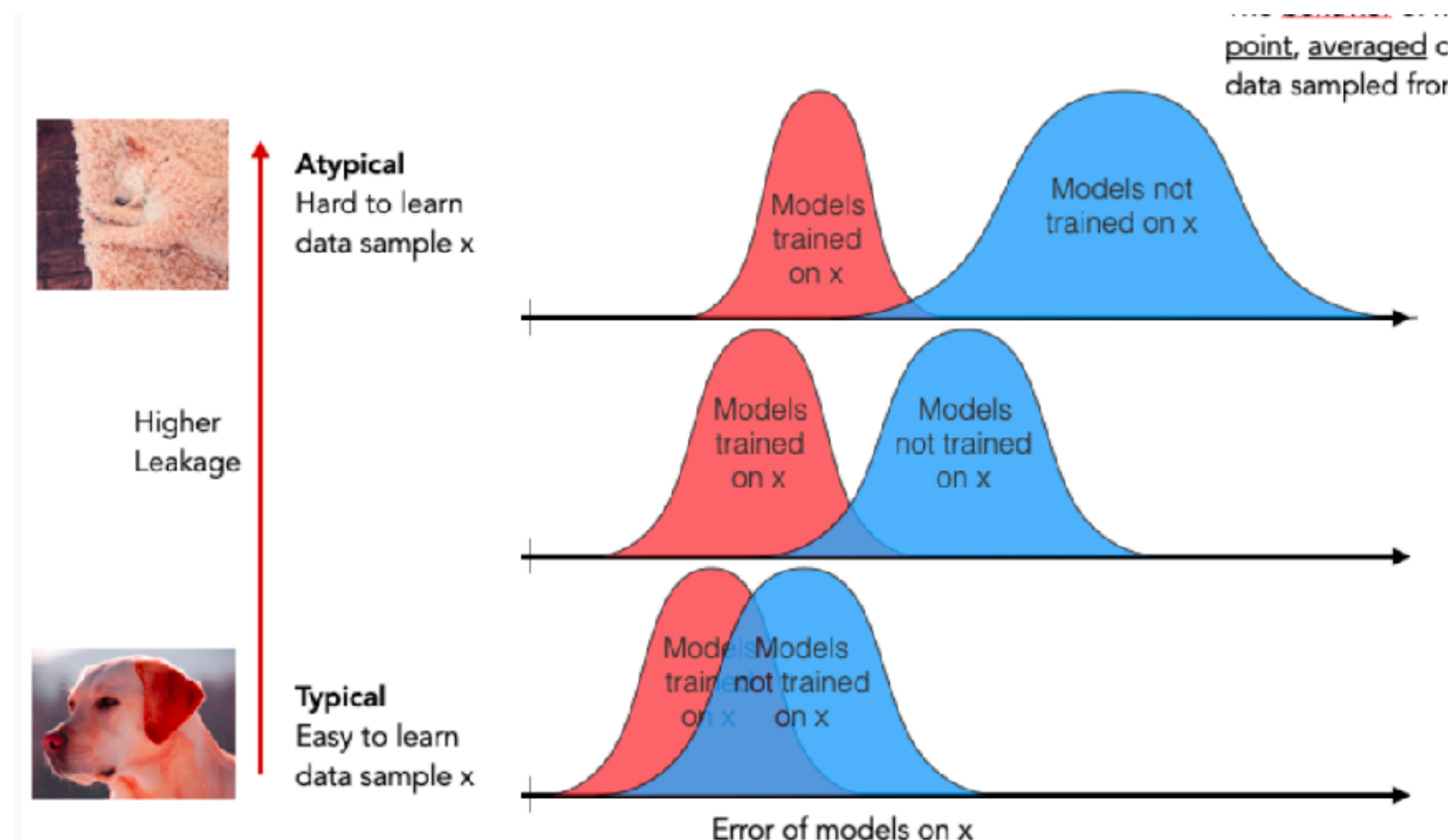
Privacy Auditing

- 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



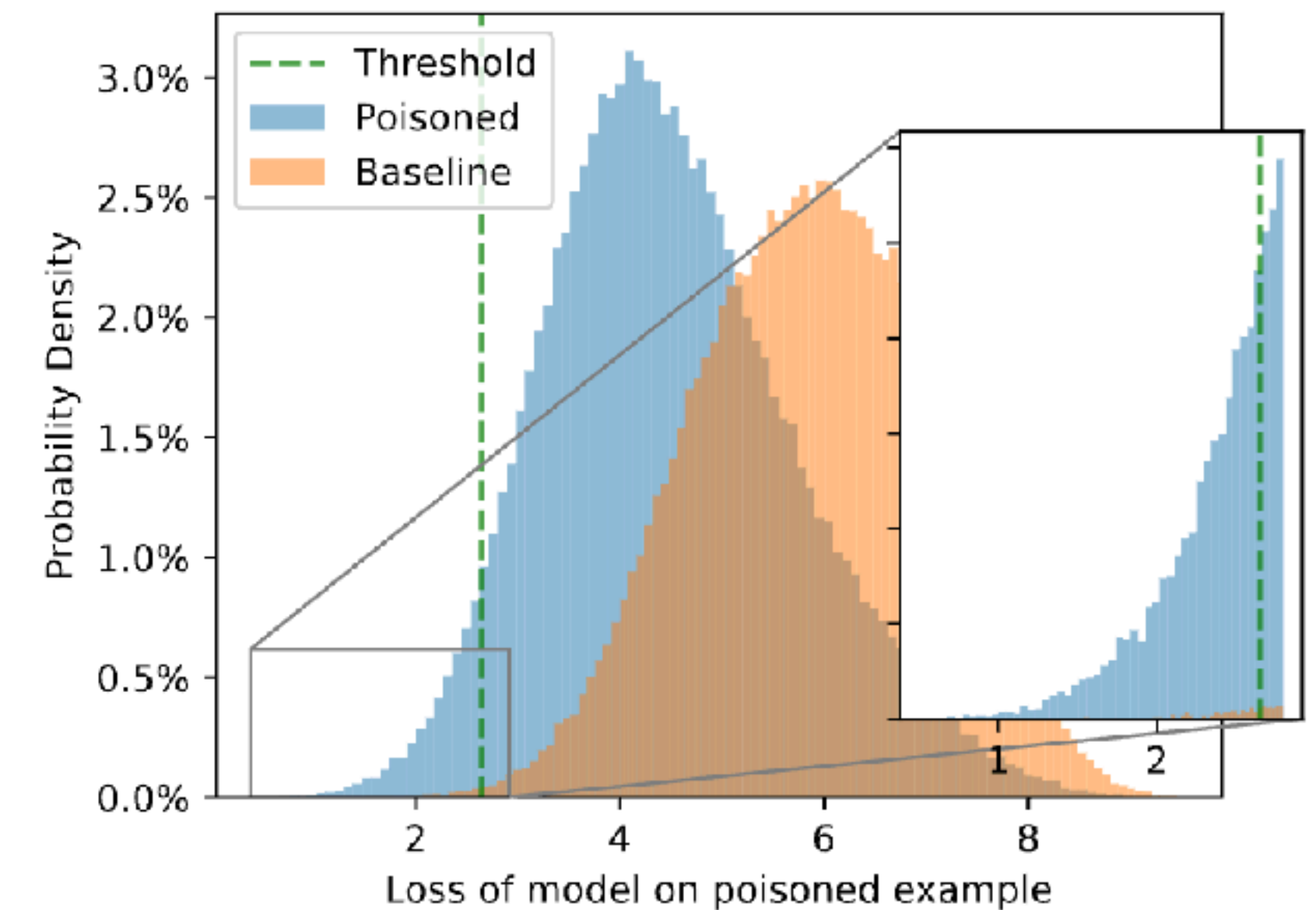
Privacy Auditing

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



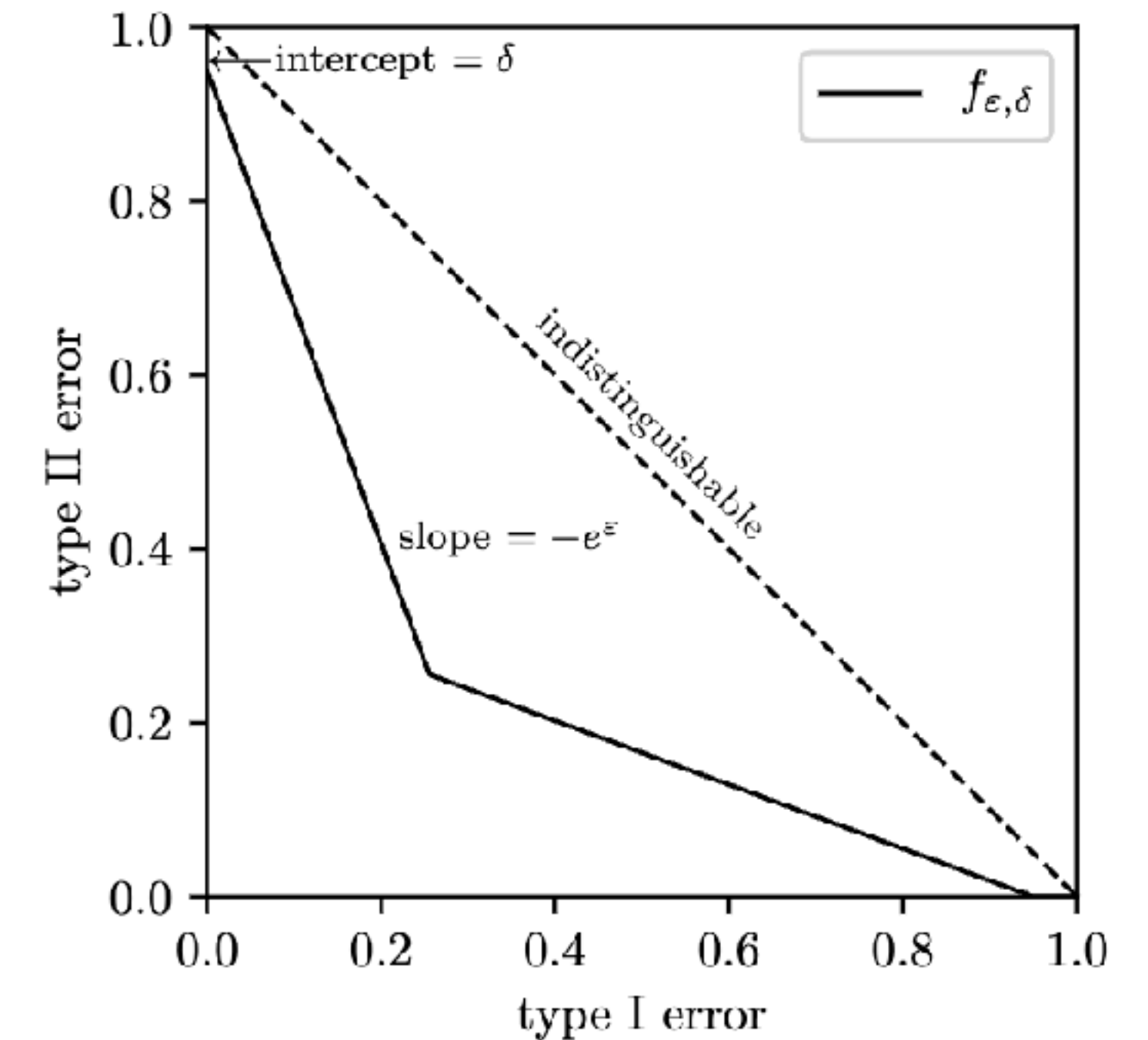
Privacy Auditing

- 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.



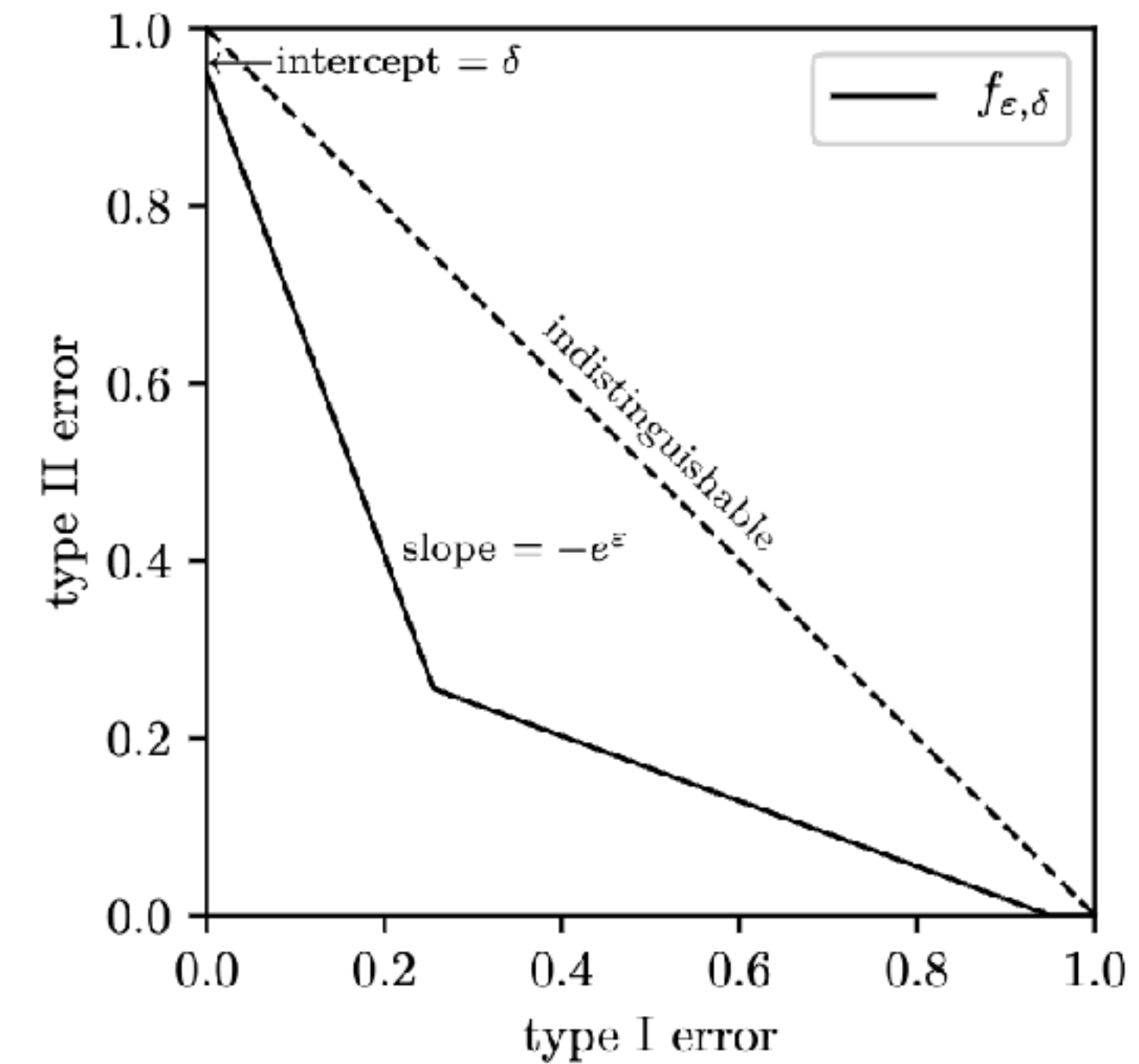
Privacy Auditing

- Claimed privacy: $(0.21, 10^{-5})$ -DP.
- With a threshold $\tau = 2.64$, attack had true positive rate of 4.922% and false positive rate of 0.174%.
- Is this possible?



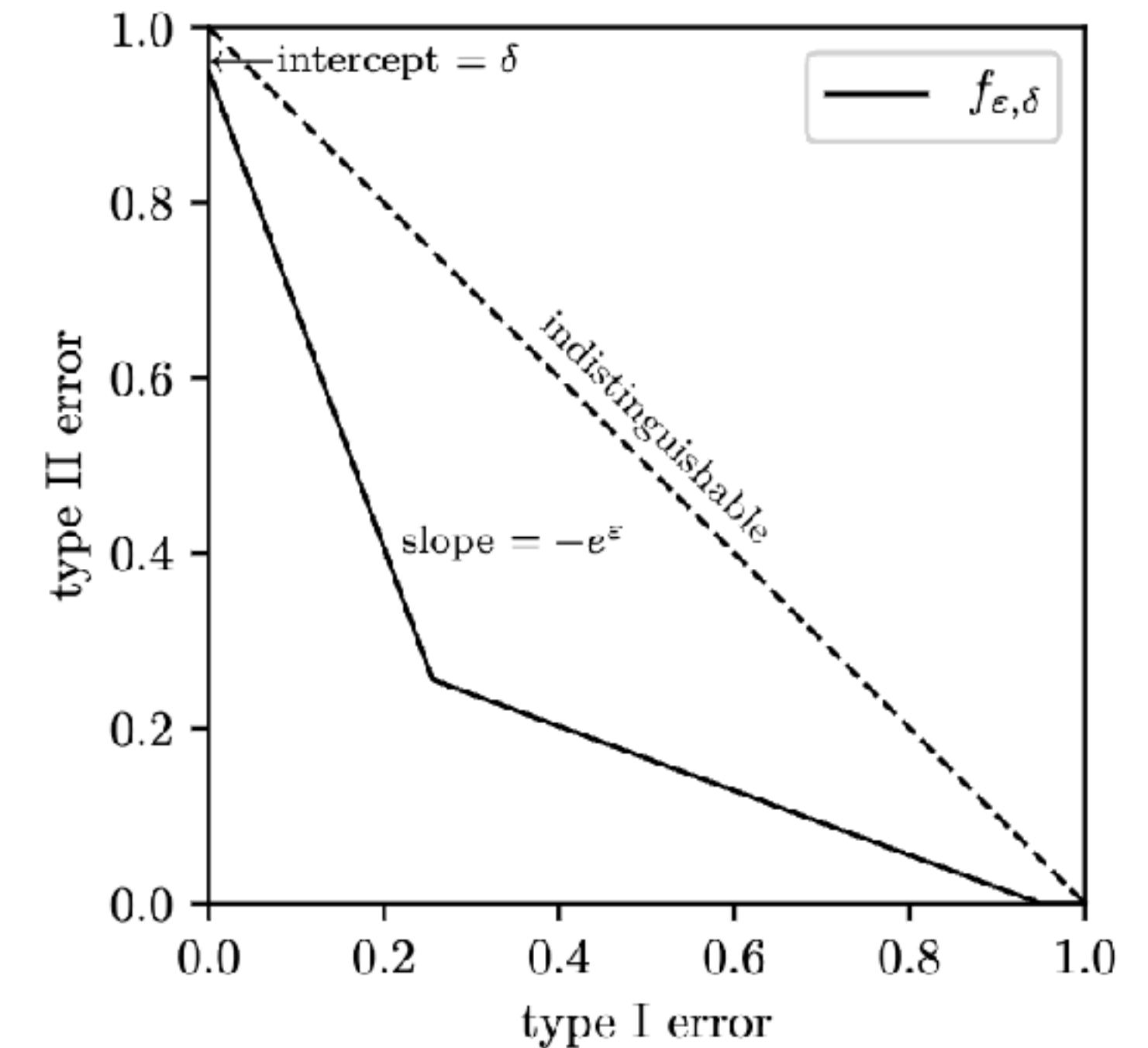
Privacy Auditing

- 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
- Can be due to sampling?



Privacy Auditing

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

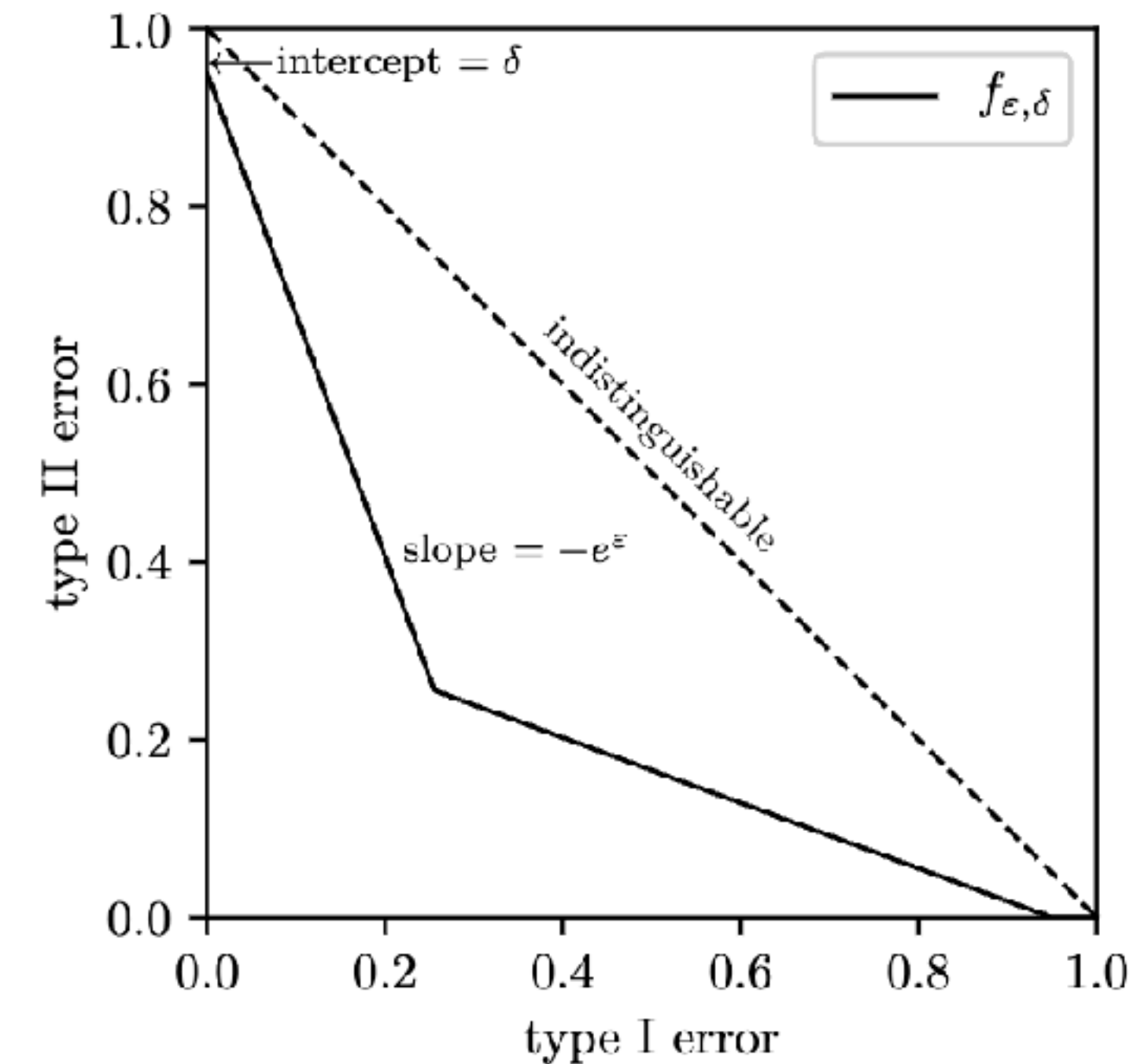


Aside: Clopper-Pearson “exact” method

- $Y = \frac{1}{n} \sum_{i=1}^n X_i$, where $X_i \sim \text{Bern}(\alpha)$. 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.

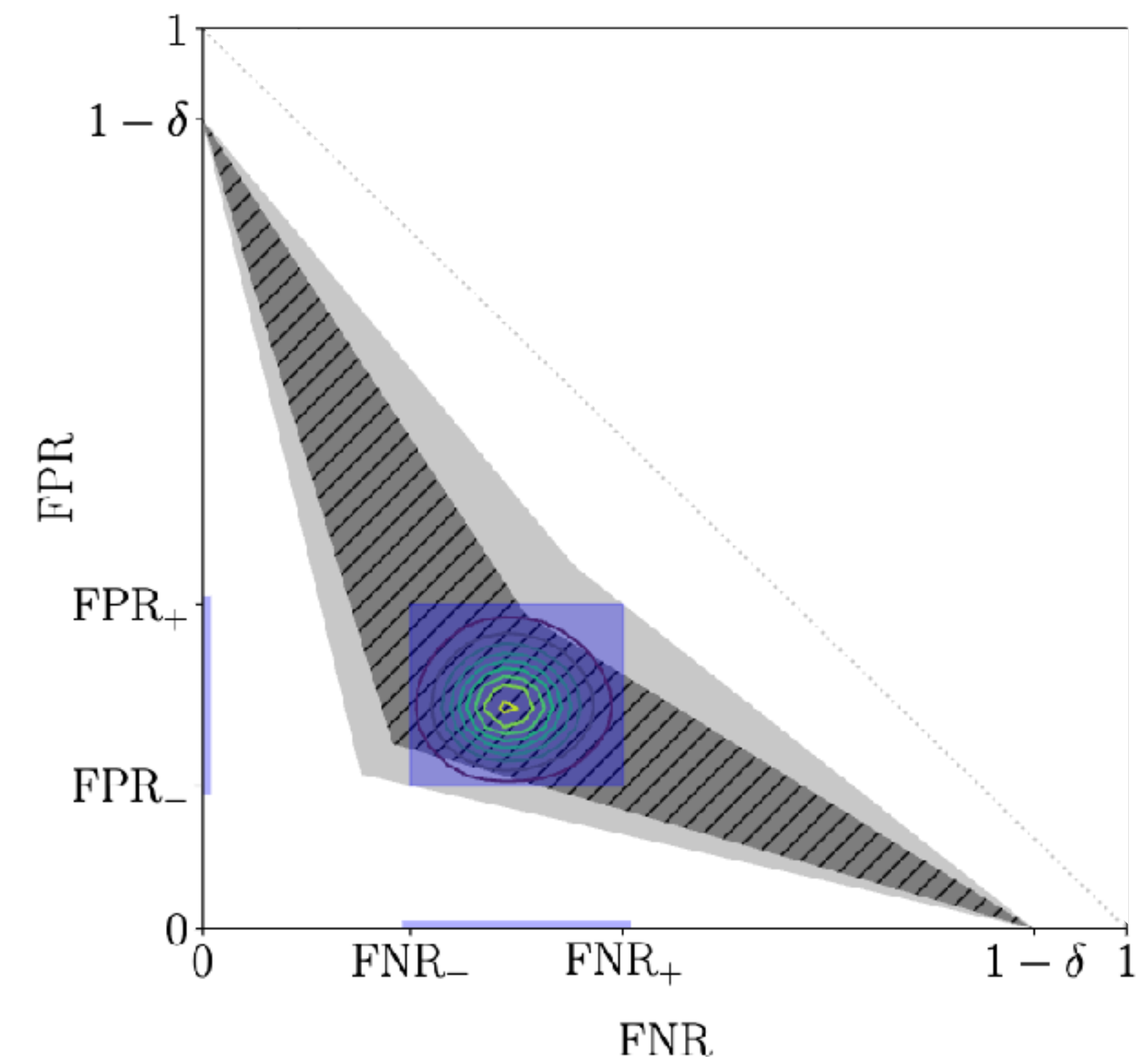
Privacy Auditing

- 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!!



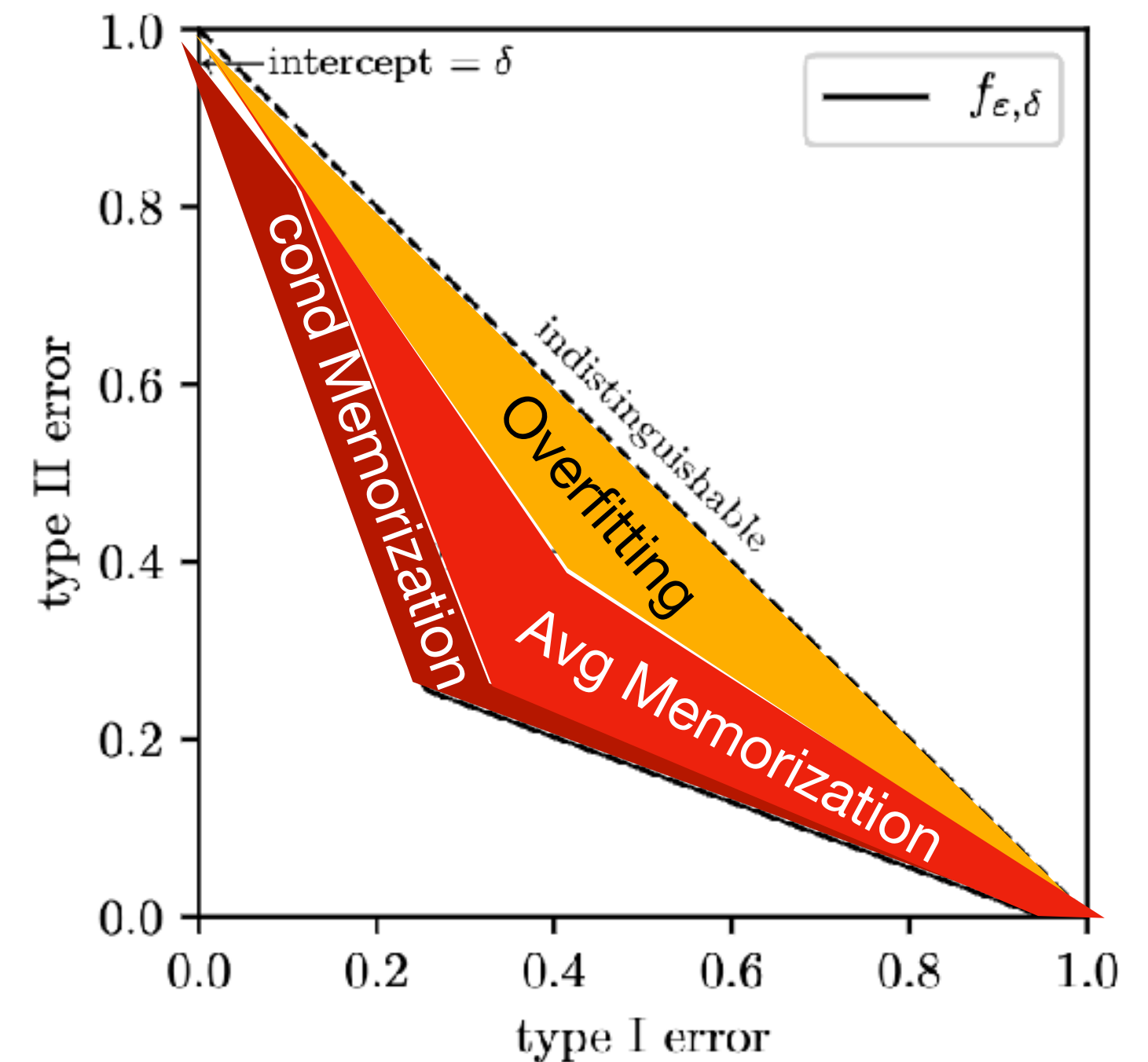
Improvements: better stats

- Do we really need α^+, β^- ?
 - 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

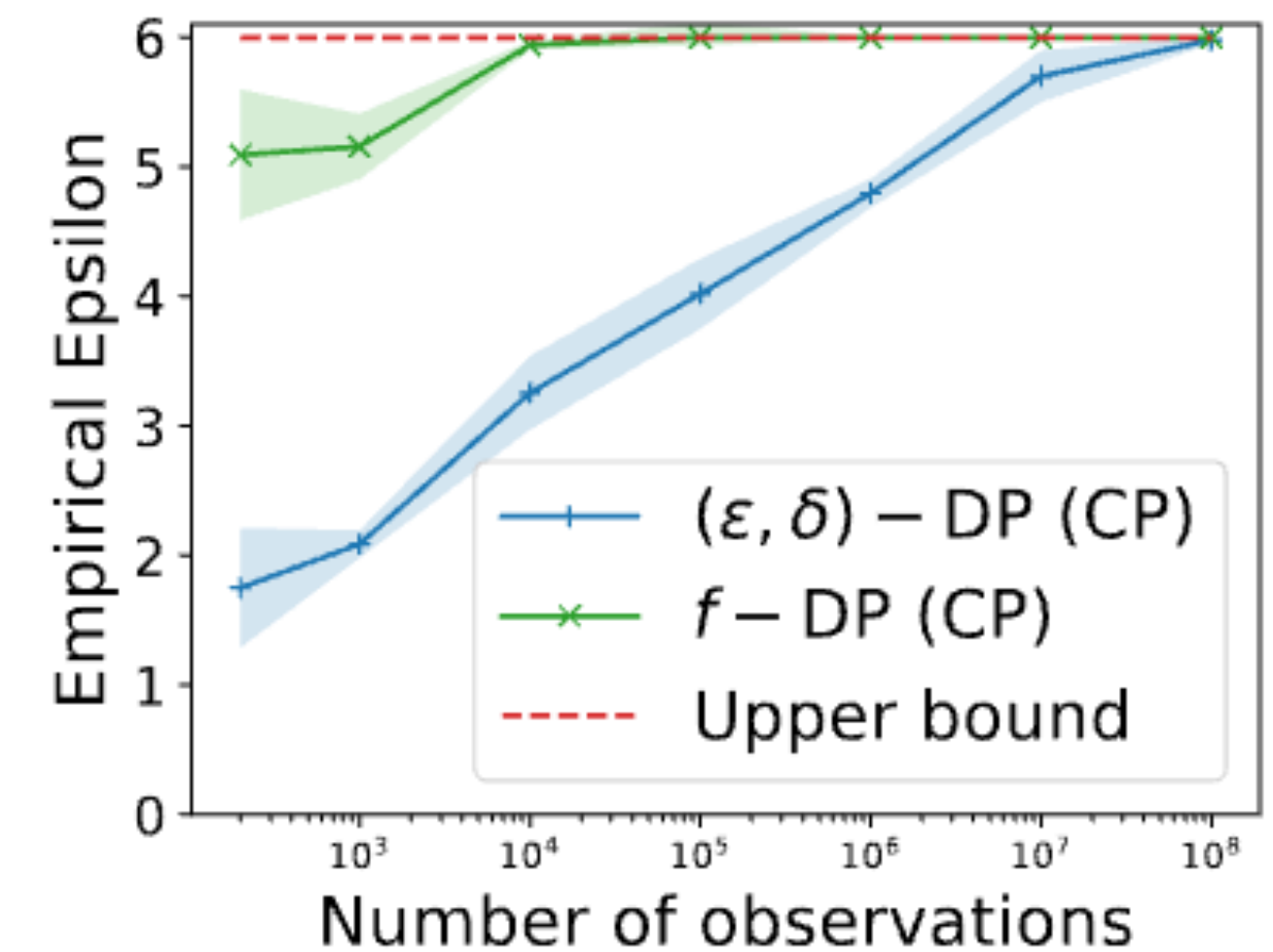
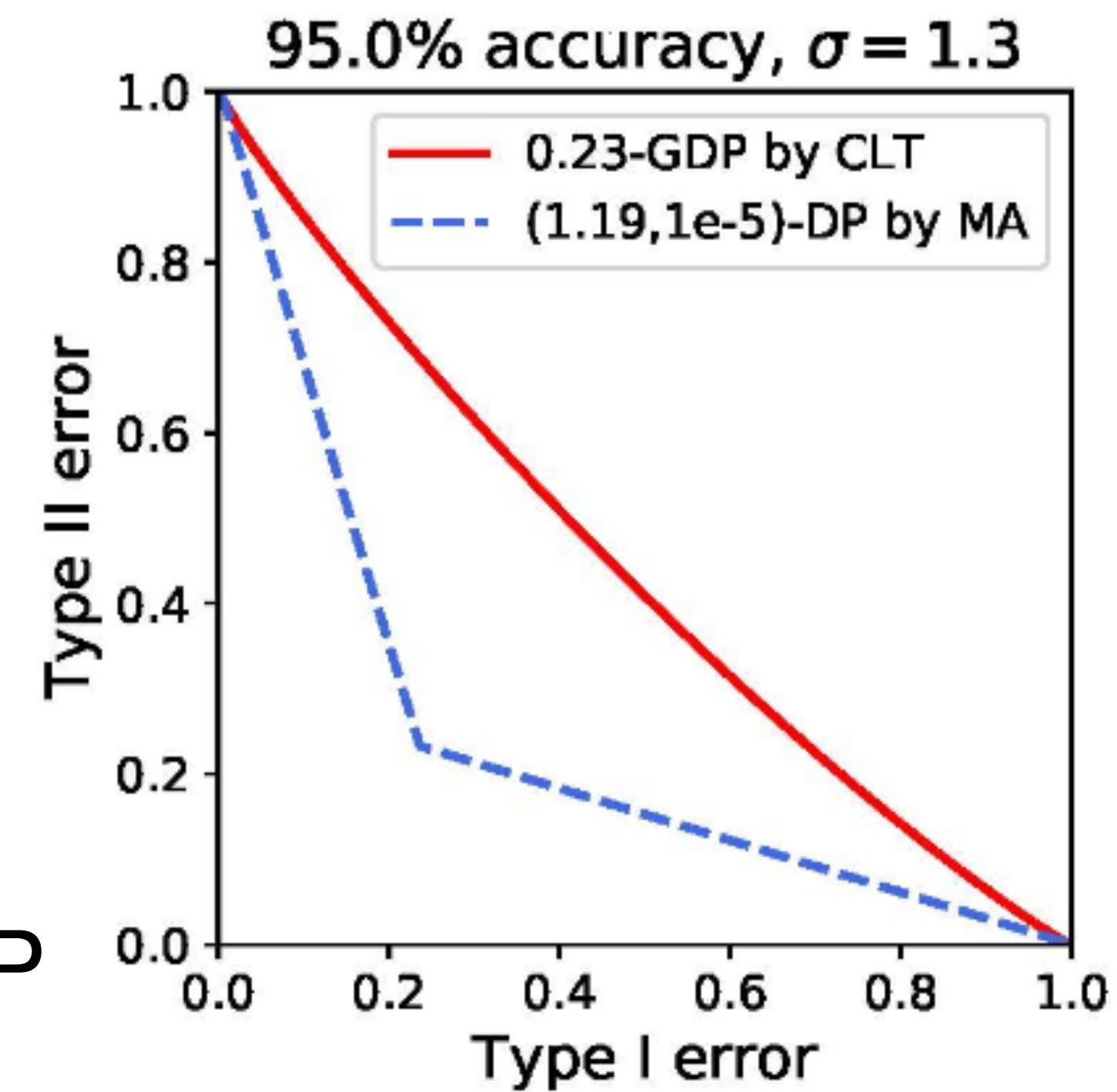
- Picking the right (x', y') is an art
 - Very similar to backdoor attacks
- Goal is to test for *conditional memorization*
- Means searching for a “planted signal”
 - when detected, we are sure. i.e. low type I
 - but can miss a lot i.e. high type II
 - what if $\delta \geq \alpha$?



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) $\epsilon = 6$