CSCI 699: Privacy Preserving Machine Learning - Week 4

Algorithms for Differentially Privacy and Machine Learning

Sai Praneeth Karimireddy, Sep 20 2024

Recap

- Approximate differential privacy

Lemma 3.17 [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 $D, D' \in \chi^n$ we have $Pr\left[\mathcal{L}_{D,D'} \geq \varepsilon\right] \leq \delta$

Recap

Private mean estimation

- Output $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \text{clip}_\tau(x_i) + \mathcal{N}(0, \rho^2)$ for $\rho = 2\tau \log(2/\delta)/n\varepsilon$.

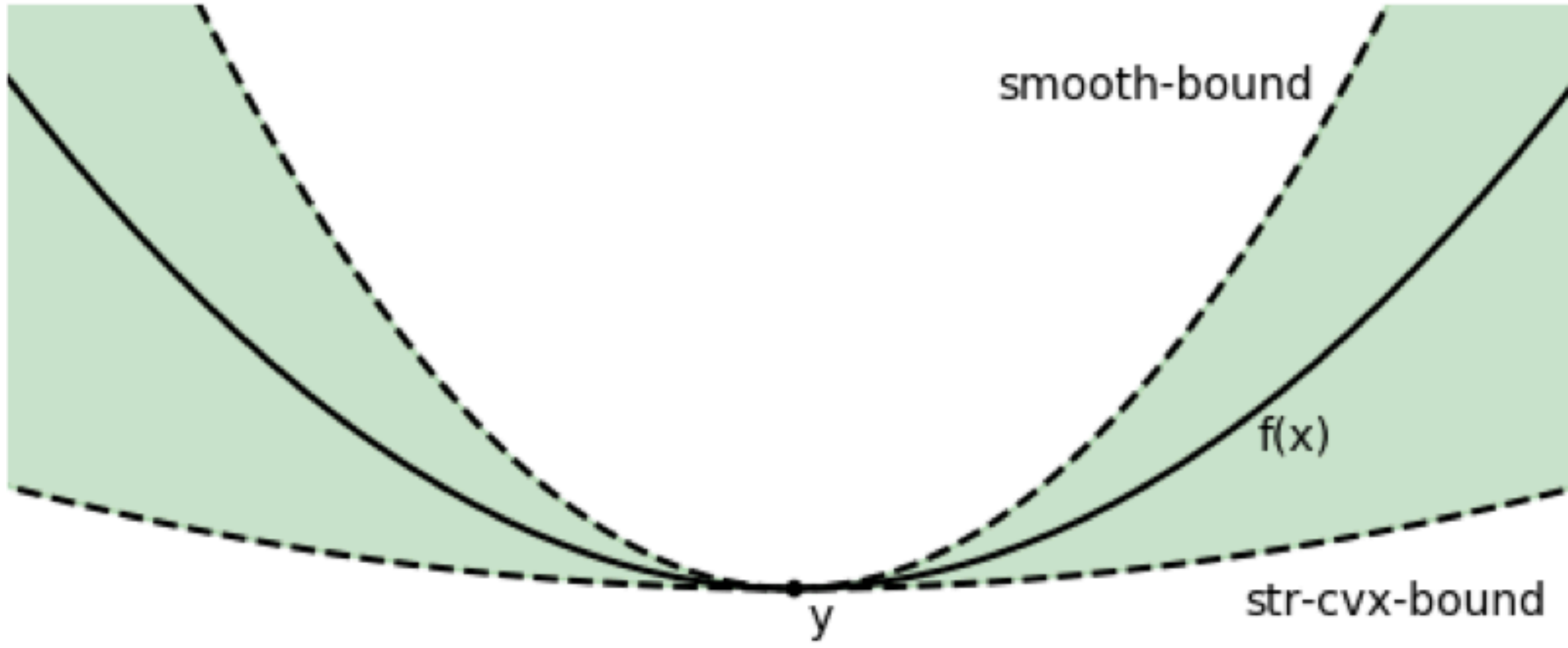
Theorem

$\hat{\mu}$ with $\tau = O(\sigma\sqrt{n\varepsilon}/d^{1/4})$ satisfies (ε, δ)-DP and has an error

$$E[(\hat{\mu} - \mu)^2] \leq O\left(\frac{\sigma^2}{n} + \frac{\sigma^2\sqrt{d}\log(1/\delta)}{n\varepsilon}\right)$$

Recap

Gradient descent



- $\theta_t = \theta_{t-1} - \gamma_t \nabla L(\theta_{t-1})$
- $\frac{\mu}{2}\|\Delta\theta\|_2^2 \geq L(\theta_t + \Delta\theta) - \left(L(\theta_t) + \nabla L(\theta_t)^\top \Delta\theta\right) \leq \frac{\beta}{2}\|\Delta\theta\|_2^2$

μ-strongly-convex

β-Smoothness

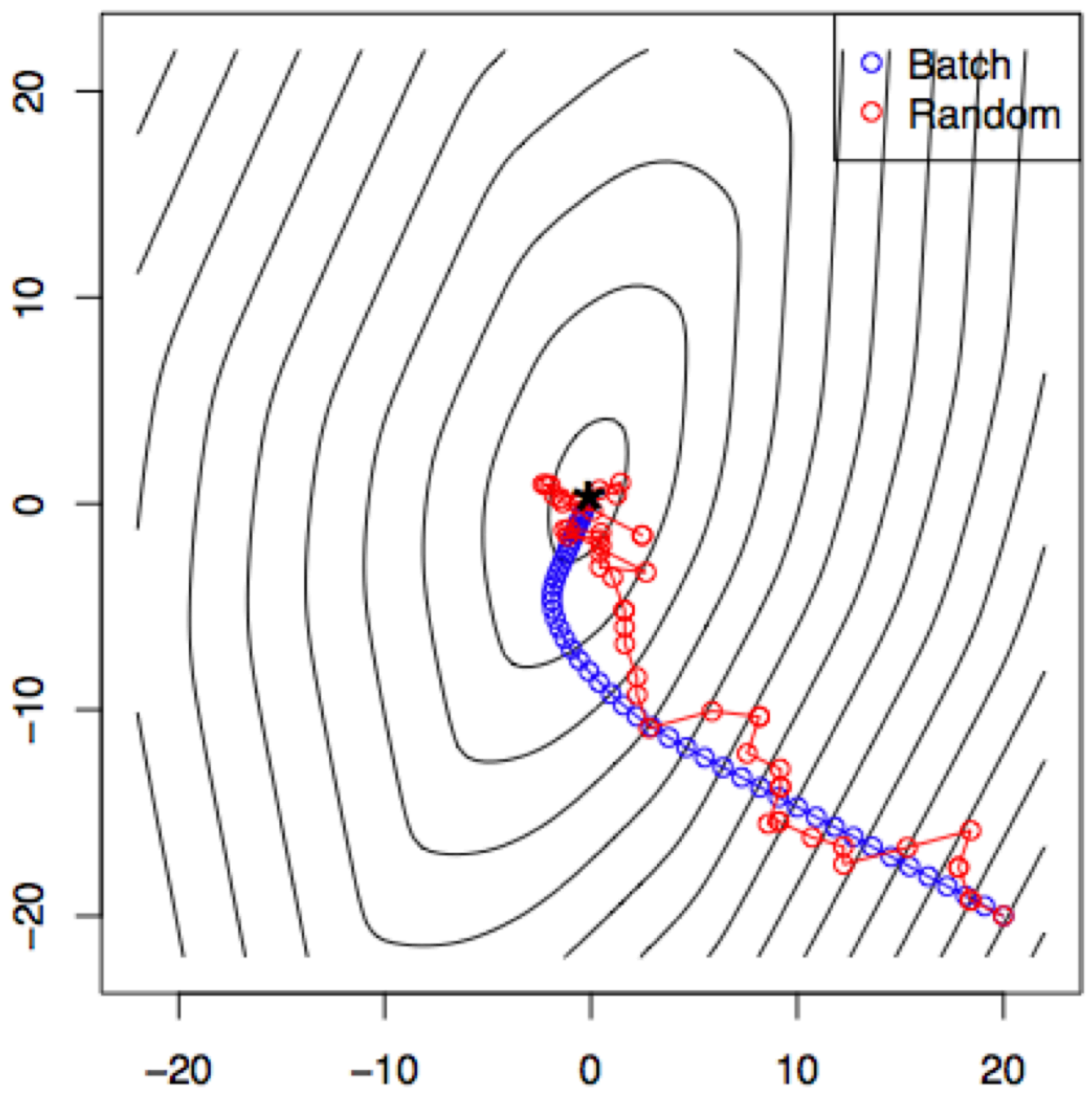
Theorem

If L is β-smooth and μ-strongly convex, gradient descent with $\gamma_t = 1/\beta$ converges as

$$L(\theta_t) - \min_\theta L(\theta) \leq \left(1 - \frac{\mu}{\beta}\right)^t \|\theta_0 - \theta^\star\|_2^2$$

Recap

Stochastic gradient descent



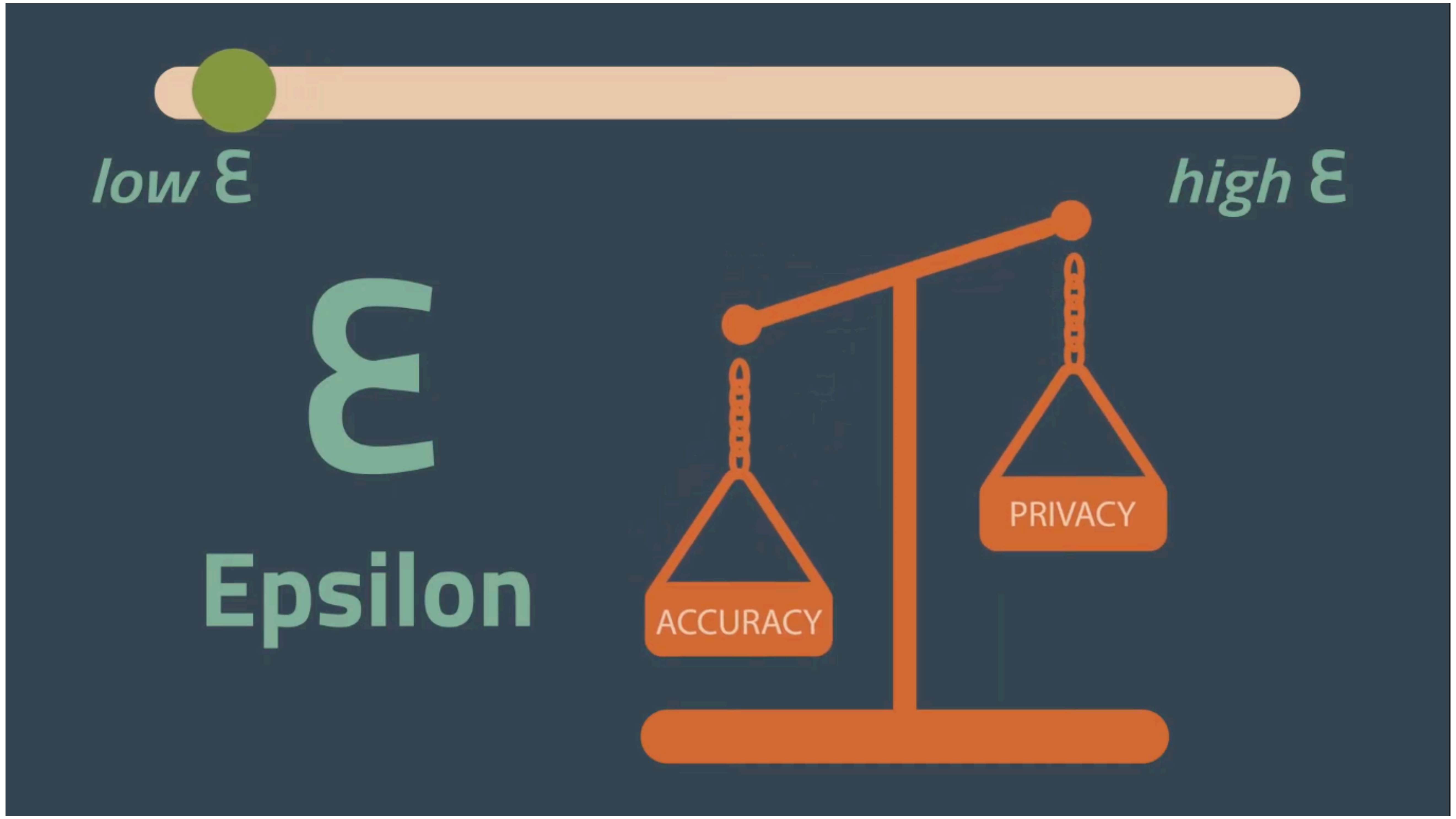
- We are do not know $L(\theta) = E_{(x,y)}[\ell(f(x;\theta), y)]$, only samples.
- For t = 1,…, n
  - Sample a data point (x_t, y_t)
  - $\theta_t = \theta_{t-1} - \gamma_t \nabla_\theta \ell(f(x_t;\theta_{t-1}), y_t) = \theta_{t-1} - \gamma_t \nabla \ell_t(\theta_{t-1})$
- Question: how do we make this private?

Agenda for today

Analyzing privacy of ML training

- Analysis of private GD: Composition
- Analysis of private SGD: Subsampling amplification
- Privacy-utility tradeoff for mean
- DP-deep learning with Opacus

Making Gradient Descent Private: Composition



Gradient Descent Variants

- we are given n samples $(x_1, y_1), \ldots, (x_n, y_n)$
- We have a few options:
  - Exact gradient: $\nabla_\theta E_{x,y}[\ell(f(x;\theta), y)]$
  - Stochastic gradient: for a random sample (x_i, y_i), $\nabla_\theta \ell(f(x_i;\theta), y_i)$
  - Full-batch gradient: $\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(f(x_i;\theta), y_i)$
  - Mini-batch gradient: for a sample $\mathscr{B}$, $\frac{1}{|\mathscr{B}|}\sum_{i\in\mathscr{B}} \nabla_\theta \ell(f(x_i;\theta), y_i)$

Private full-batch gradient descent
Algorithm

- Starting from θ_0, at each time step we update
  - $\theta_t = \theta_{t-1} - \gamma \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \ell(f(x_i;\theta), y_i)$
- To make it private
  - $\theta_t = \theta_{t-1} - \gamma \frac{1}{n} \sum_{i=1}^{n} \text{Clip}_\tau \left(\nabla_\theta \ell(f(x_i;\theta), y_i)\right)$ + noise
  - Assume scalar for now. So noise = $Lap(??)$

Private full-batch gradient descent

One-step privacy

- Suppose we just run step of
$\theta_t = \theta_{t-1} - \gamma \frac{1}{n} \sum_{i=1}^{n} \text{Clip}_\tau \left(\nabla_\theta \ell(f(x_i; \theta), y_i) \right) + Lap(??)$
- Sensitivity? How much noise?
- How to reason about what happens across time steps?

Post-processing and composition

Post-processing

- You can never undo the output of a DP-algorithm

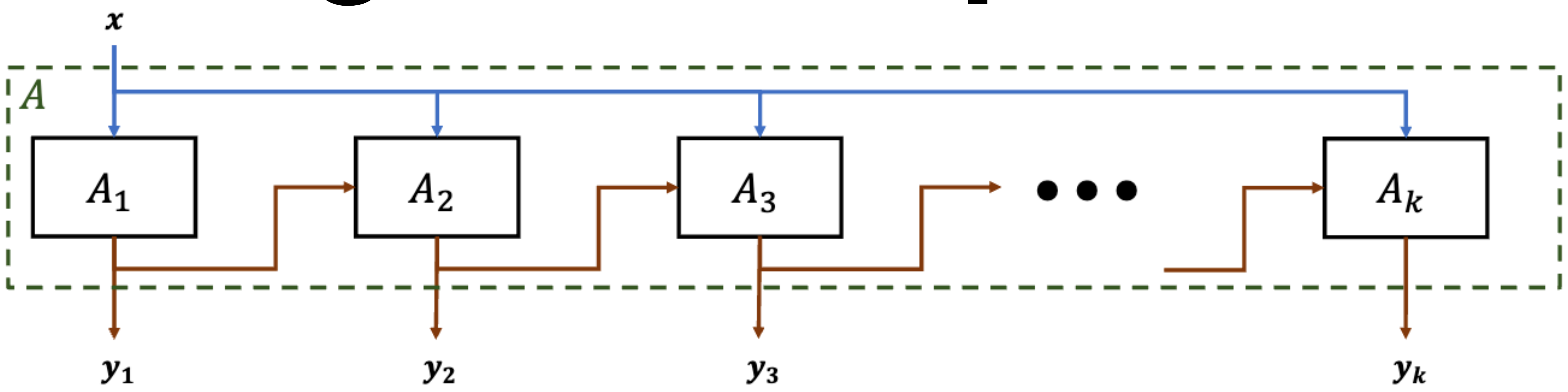
Theorem

$A : \mathcal{X}^n \to \mathbb{R}^d$ is a (ε, δ)-DP algorithm and f is a mapping independent of $\mathcal{X}$, then $f \circ A$ is (ε, δ)-DP

- Upshot: we can plug in our private gradients into any optimizer (e.g. AdamW).

Post-processing and composition

Composition



- What if the new function also depends on our data?

Theorem

$A : \mathcal{X}^n \to \mathbb{R}^d$ is a $(\varepsilon_1, 0)$-DP algorithm and
$B : \mathcal{X}^n \to \mathbb{R}^d$ is a $(\varepsilon_2, 0)$-DP algorithm, then
$(A, B) : \mathcal{X}^n \to \mathbb{R}^d \times \mathbb{R}^d$ is $(\varepsilon_1 + \varepsilon_2, 0)$-DP

Private full-batch gradient descent

Multi-step privacy

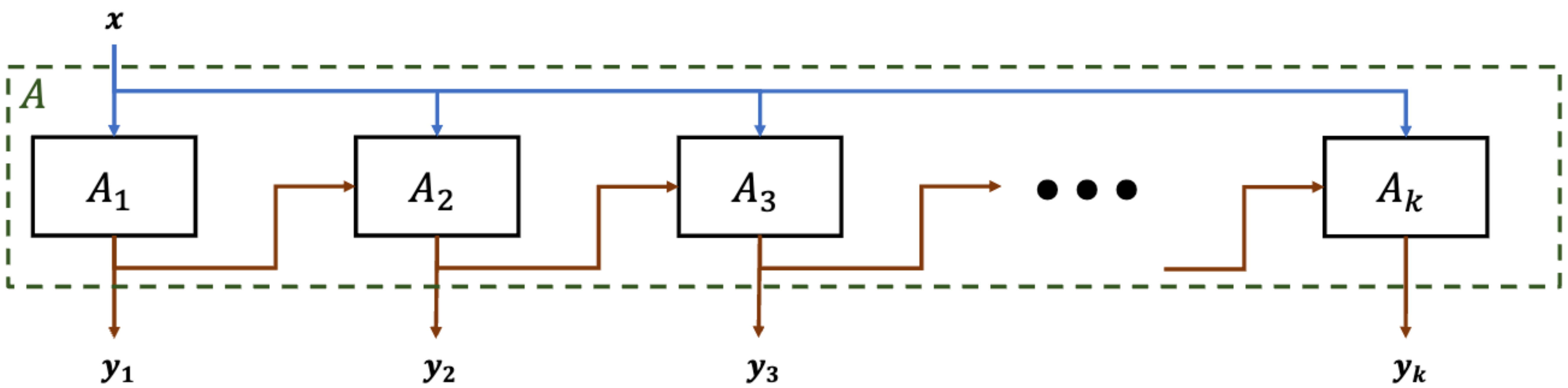
- One step is $(\varepsilon, 0)$-DP
  $\theta_t = \theta_{t-1} - \gamma \frac{1}{n} \sum_{i=1}^{n} \text{Clip}_\tau \left(\nabla_\theta \ell(f(x_i; \theta), y_i) \right) + Lap(2\tau / n\varepsilon)$

- k-steps of full-batch gradient descent is $(k\varepsilon, 0)$-DP.

- We can do better!

Private full-batch gradient descent

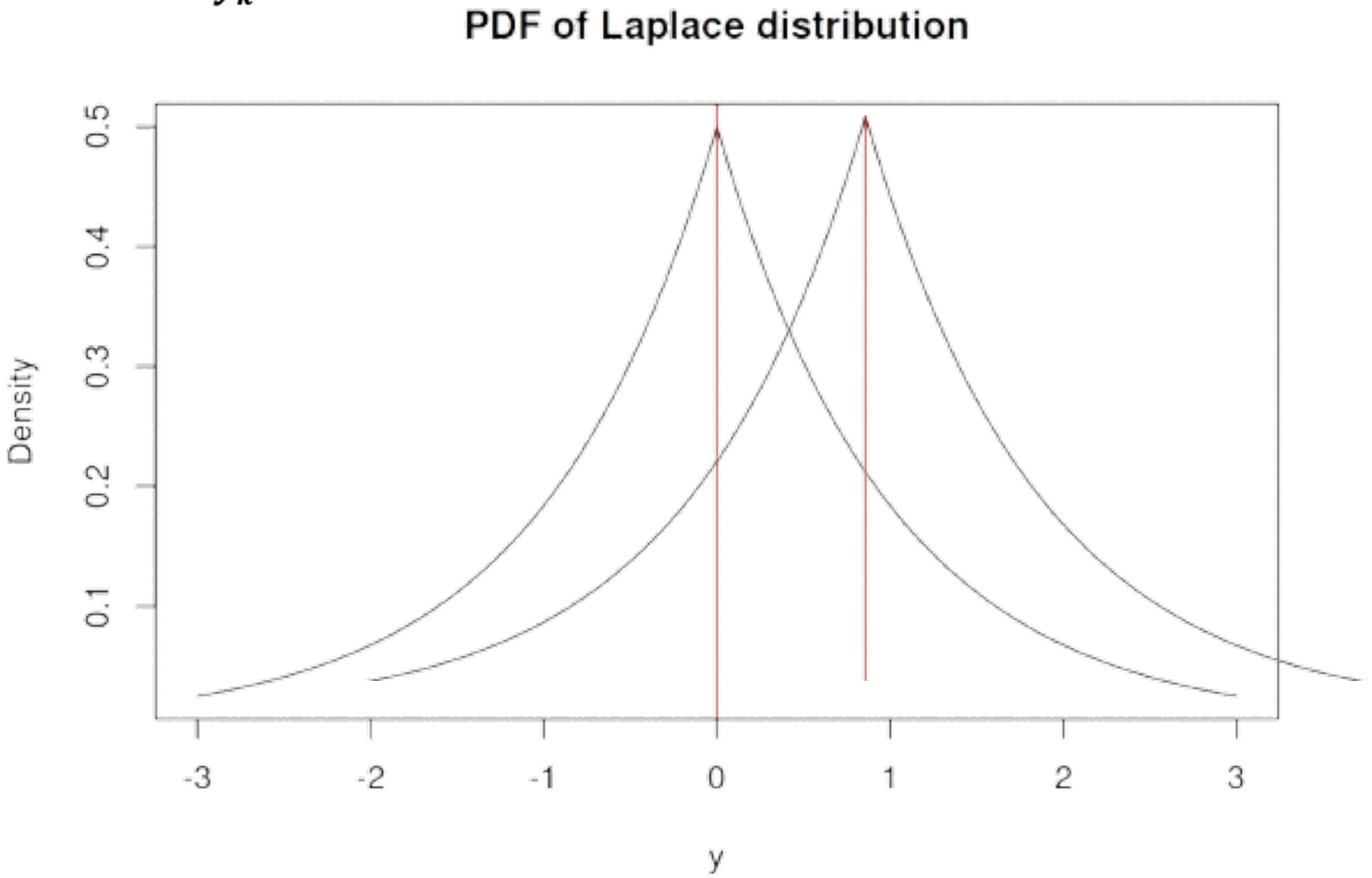
Advanced composition



- Let us compute the privacy random variable:

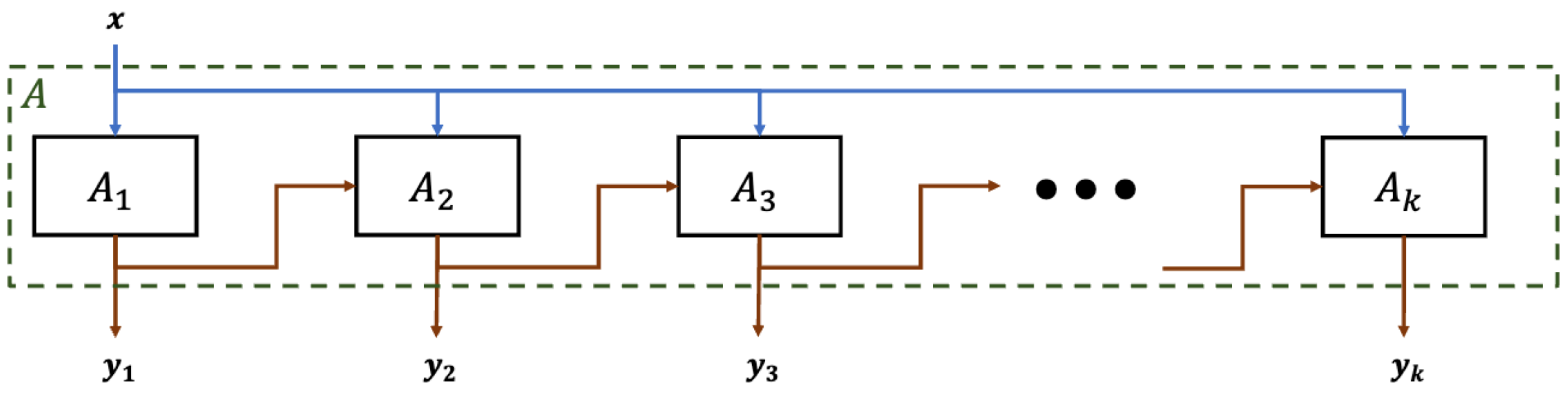
$$R = \log \left(\frac{Pr[A(D) = t]}{Pr[A(D') = t]} \right) \qquad \text{for } t \sim A(D)$$

- $R \in [-\varepsilon, \varepsilon]$ and has mean 0.



Private full-batch gradient descent

Advanced composition



- Privacy random variable of composition:

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

- $R_i \in [-\varepsilon, \varepsilon]$, 0-mean, conditionally independent.

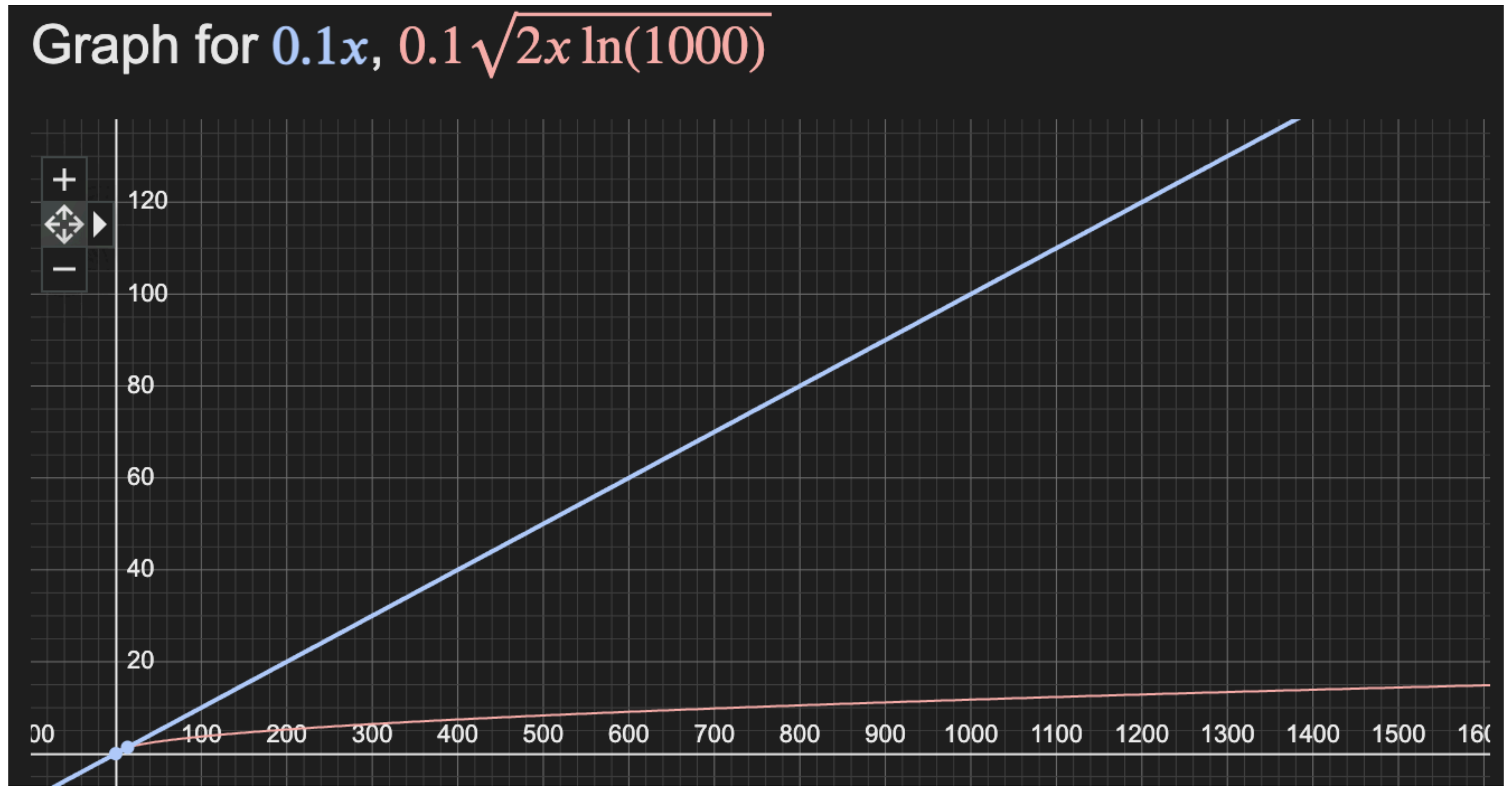
Private full-batch gradient descent

Aside: Azuma's inequality

Azuma's inequality

Given $X_1, \ldots, X_n$ where $E[X_i \,|\, \text{past}] = 0$, $|X_i| \leq \varepsilon_i$.
Then,

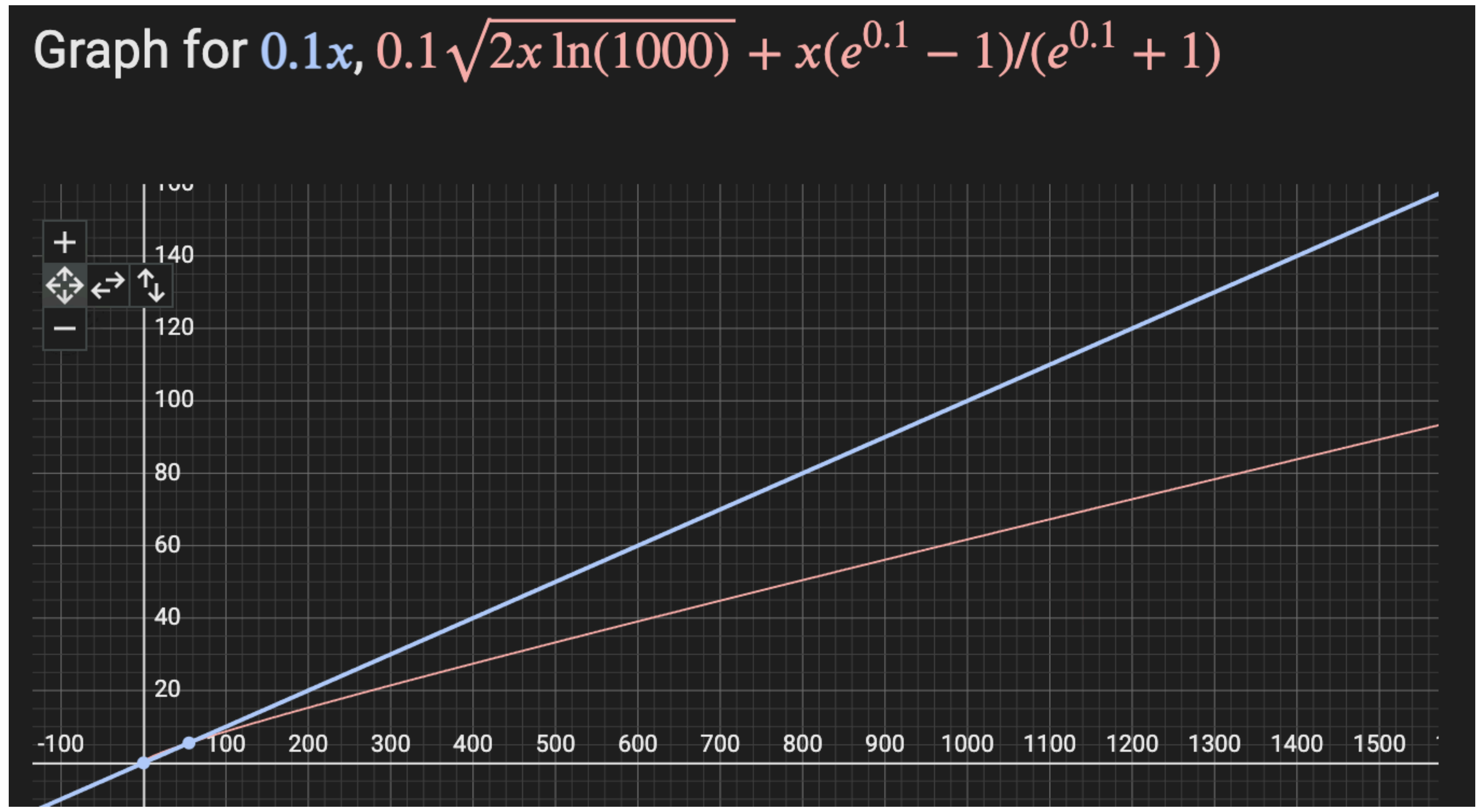
$$Pr[\sum_{i=1}^{k} X_i \geq \Delta] \leq \exp(-\Delta^2/2\sum_{i=1}^{k} \varepsilon_i^2)$$



- $R_i \in [-\varepsilon, \varepsilon]$, 0-mean, conditionally independent.
- $Pr[\sum_{i=1}^{k} R_i \geq \varepsilon\sqrt{2k\log(1/\delta)}] \leq \delta$ i.e. we have $(\varepsilon\sqrt{2k\ln(1/\delta)}, \delta)$-DP!

Private full-batch gradient descent

Advanced composition



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

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

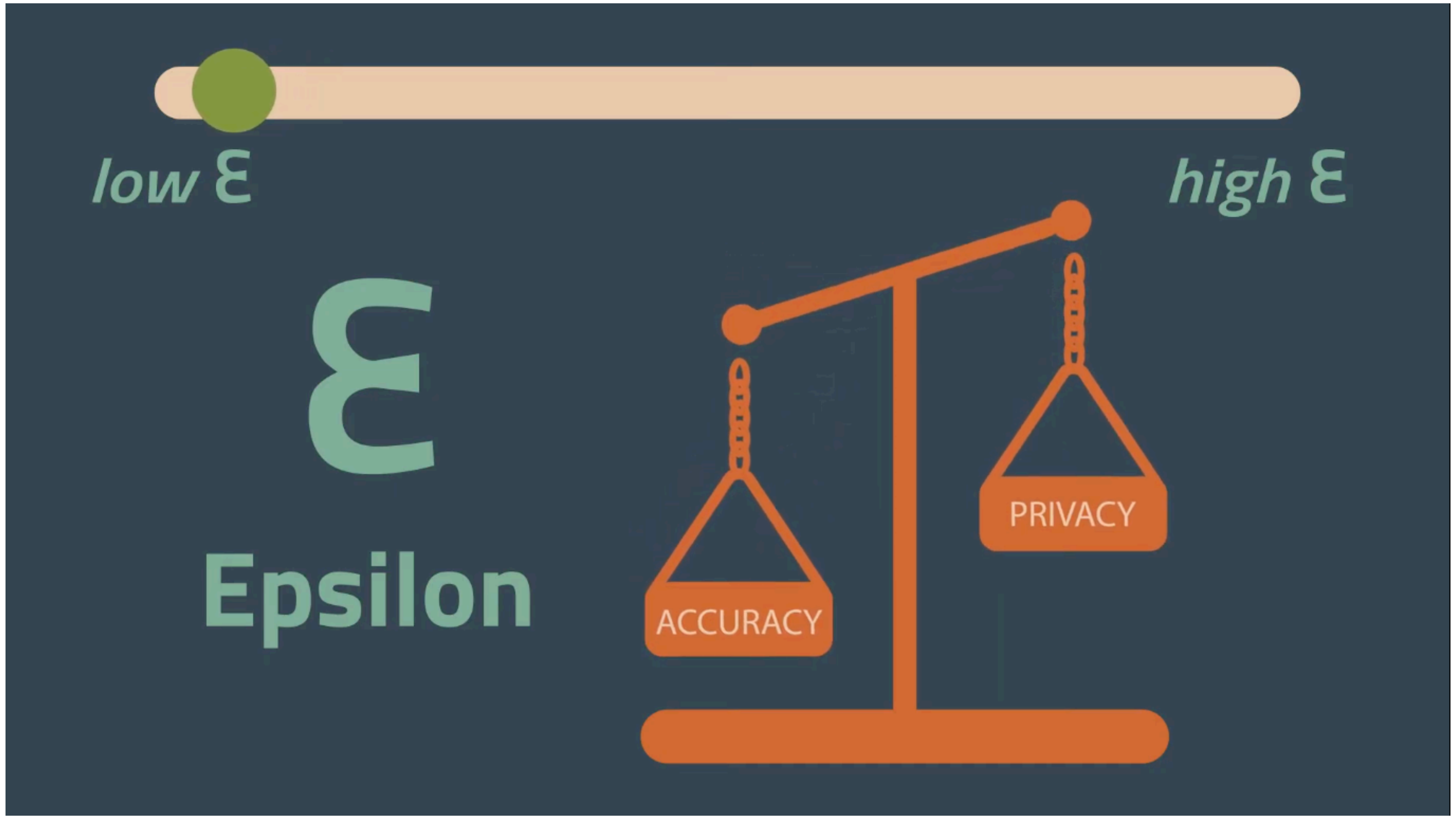
For any choice of δ'.

Private full-batch gradient descent

Multi-step privacy

- One step is $(\varepsilon, 0)$-DP
  $\theta_t = \theta_{t-1} - \gamma \frac{1}{n} \sum_{i=1}^{n} \text{Clip}_\tau \left(\nabla_\theta \ell(f(x_i; \theta), y_i) \right) + Lap(2\tau/n\varepsilon)$
- k-steps of full-batch gradient descent is $(\varepsilon\sqrt{2k \ln(1/\delta)}, \delta)$-DP.
- How about with Gaussian-noise and vectors?

Making SGD Private: Subsampling



Private stochastic gradient descent

Algorithm

- Starting from θ_0, at each time step
  - sample (x_i, y_i) randomly from $(x_1, y_1), \ldots, (x_n, y_n)$
  - $\theta_t = \theta_{t-1} - \gamma \nabla_\theta \ell(f(x_i; \theta), y_i)$
- To make it private
  - $\theta_t = \theta_{t-1} - \gamma \text{Clip}_\tau \left(\nabla_\theta \ell(f(x_i; \theta), y_i) \right)$ + noise
  - Assume scalar for now. So noise = $Lap(??)$

Private stochastic gradient descent

One-step privacy

- Suppose we just run step:
  - $\theta_t = \theta_{t-1} - \gamma \text{Clip}_\tau\left(\nabla_\theta \ell(f(x_t;\theta), y_t)\right) + Lap(2\tau/\varepsilon)$
- No improvement due to n
- Important note: use poisson sampling! Not uniform.
- This makes analyzing what happens to each data-point independent.

Privacy amplification via subsampling

- Given a dataset $D \in \mathcal{X}^n$, and $m \in [n]$
- We define S to be a random m-subsample of D
- Is releasing S private?
- Now suppose A is ε-DP on D. What is the privacy of A composed with subsampling?

Privacy amplification via subsampling

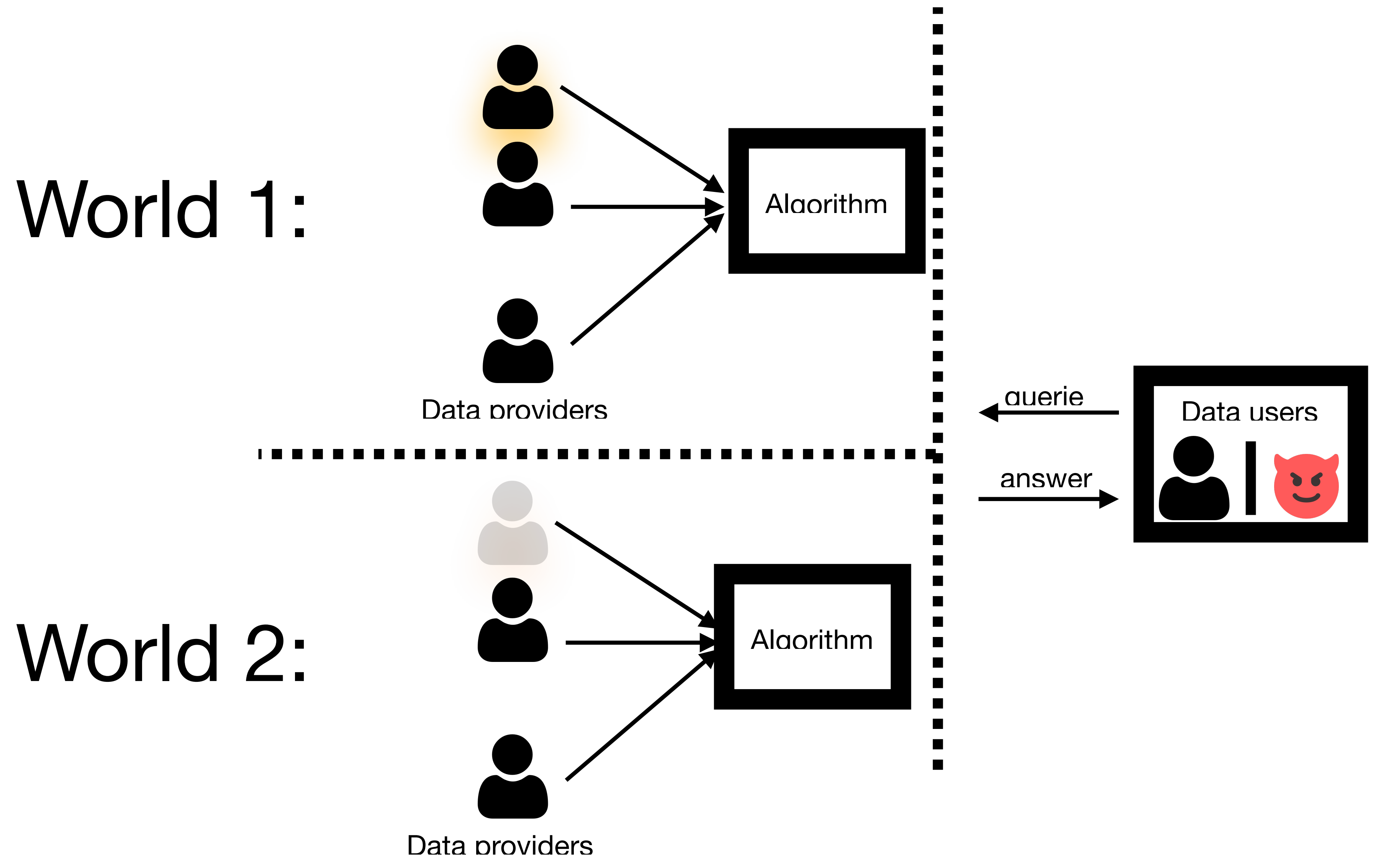
Theorem. Subsampling Amplification

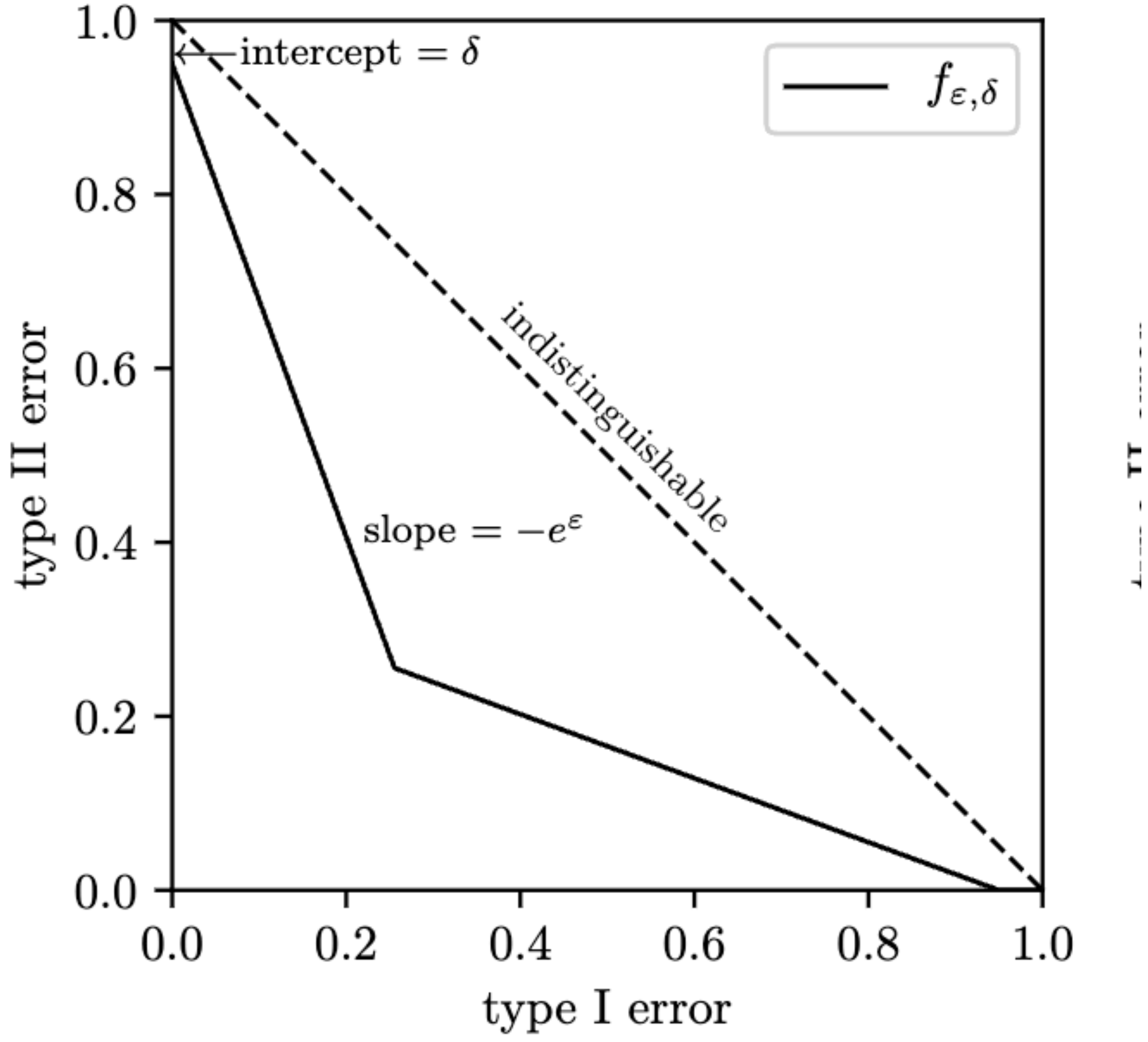
Composing an (ε, δ)-DP A with a sampling rate of q results in an $(\tilde{\varepsilon}, \tilde{\delta})$-DP algorithm where

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

Recall

Membership Inference definition of privacy





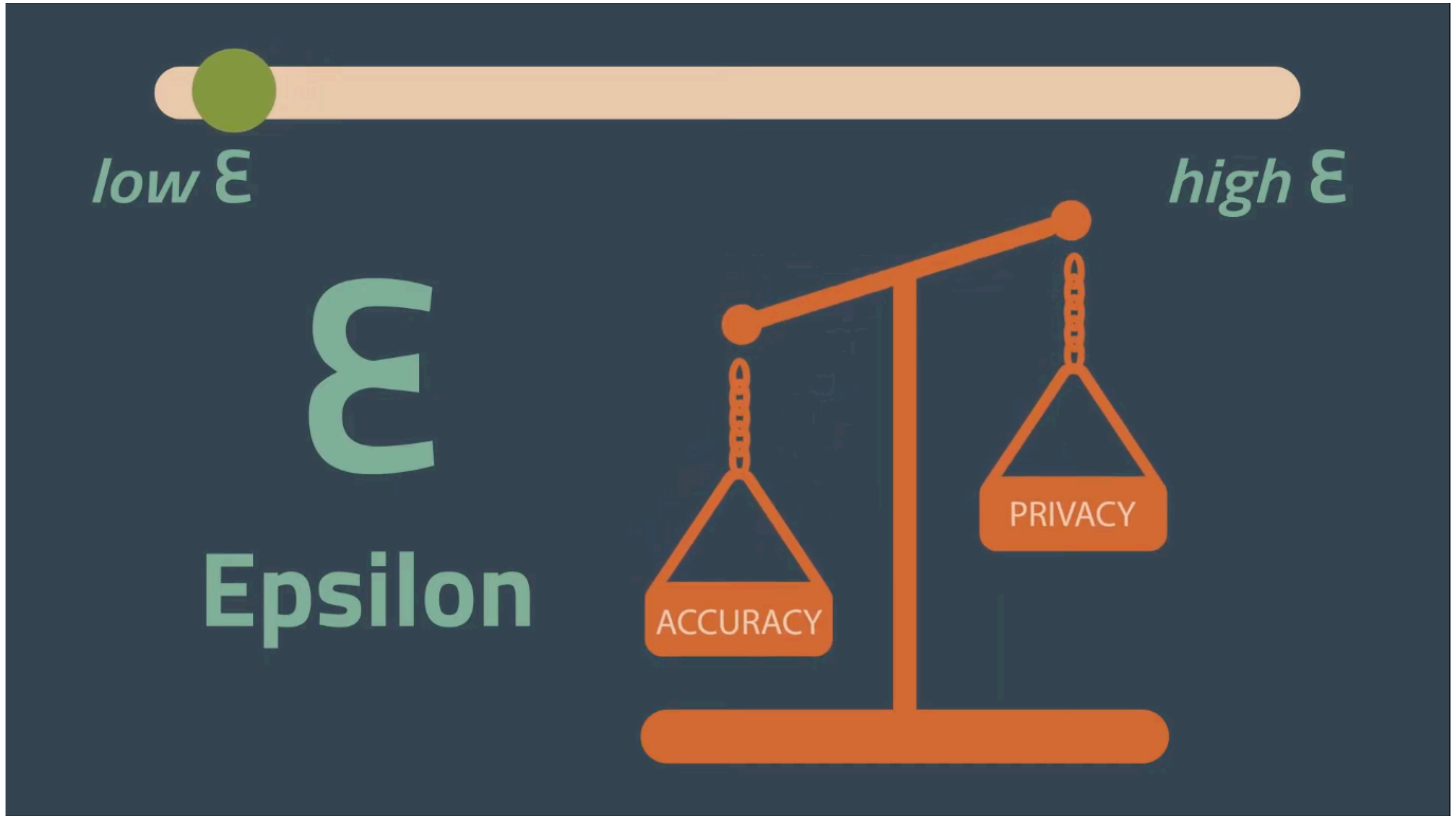
- Claim: $\beta + (1 - q + qe^{\varepsilon})\alpha \geq 1 - \delta$
- and, $(1 - q + qe^{\varepsilon})\beta + \alpha \geq 1 - \delta$, where $\alpha =$ type I error, $\beta =$ type II error

Private stochastic gradient descent

One-step privacy

- Suppose we just run step:
  - $\theta_t = \theta_{t-1} - \gamma \text{Clip}_\tau \left(\nabla_\theta \ell(f(x_t; \theta), y_t) \right) + Lap(2\tau/\varepsilon)$
- We have $q = 1/n$. So, we have $\tilde{\varepsilon} = \log(1 - 1/n + e^\varepsilon/n) = O(\varepsilon/n)$
- Adding in advanced composition, k rounds of SGD satisfies $(O(\varepsilon/n\sqrt{k \ln(1/\delta)}), \delta)$-DP
- Compare n steps of SGD with 1 step of full-batch. In practice, much better utility.

Analyzing Private learning



Private learning analysis

Private mean estimation

- Output $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \text{clip}_\tau(x_i) + \mathcal{N}(0, \rho^2)$ for $\rho = 2\tau \log(2/\delta)/n\varepsilon$.

Theorem

$\hat{\mu}$ with $\tau = O(\sigma\sqrt{n\varepsilon}/d^{1/4})$ satisfies (ε, δ)-DP and has an error

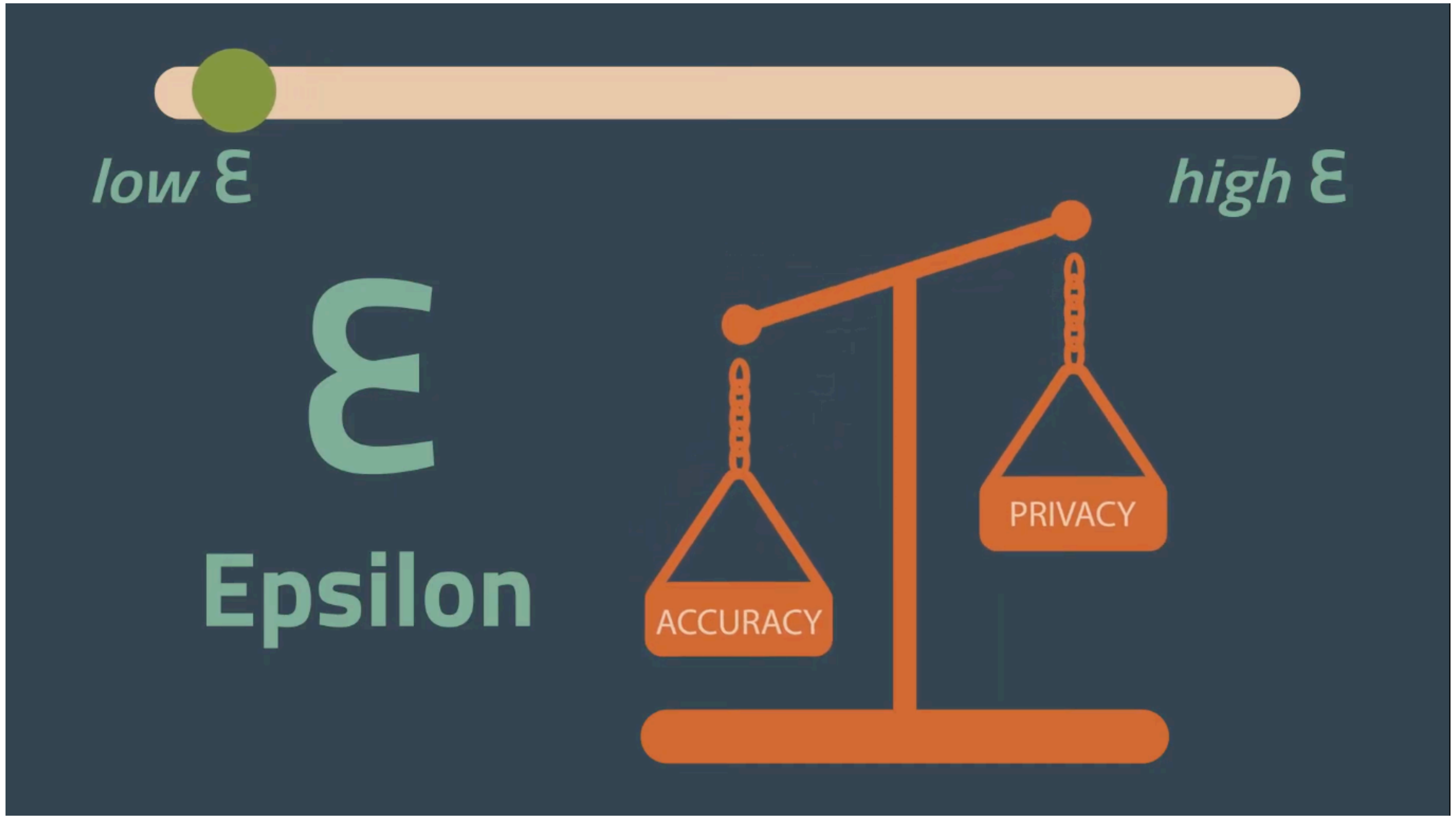
$$E[(\hat{\mu} - \mu)^2] \leq O\left(\frac{\sigma^2}{n} + \frac{\sigma^2\sqrt{d}\log(1/\delta)}{n\varepsilon}\right)$$

Private learning analysis

Private mean estimation

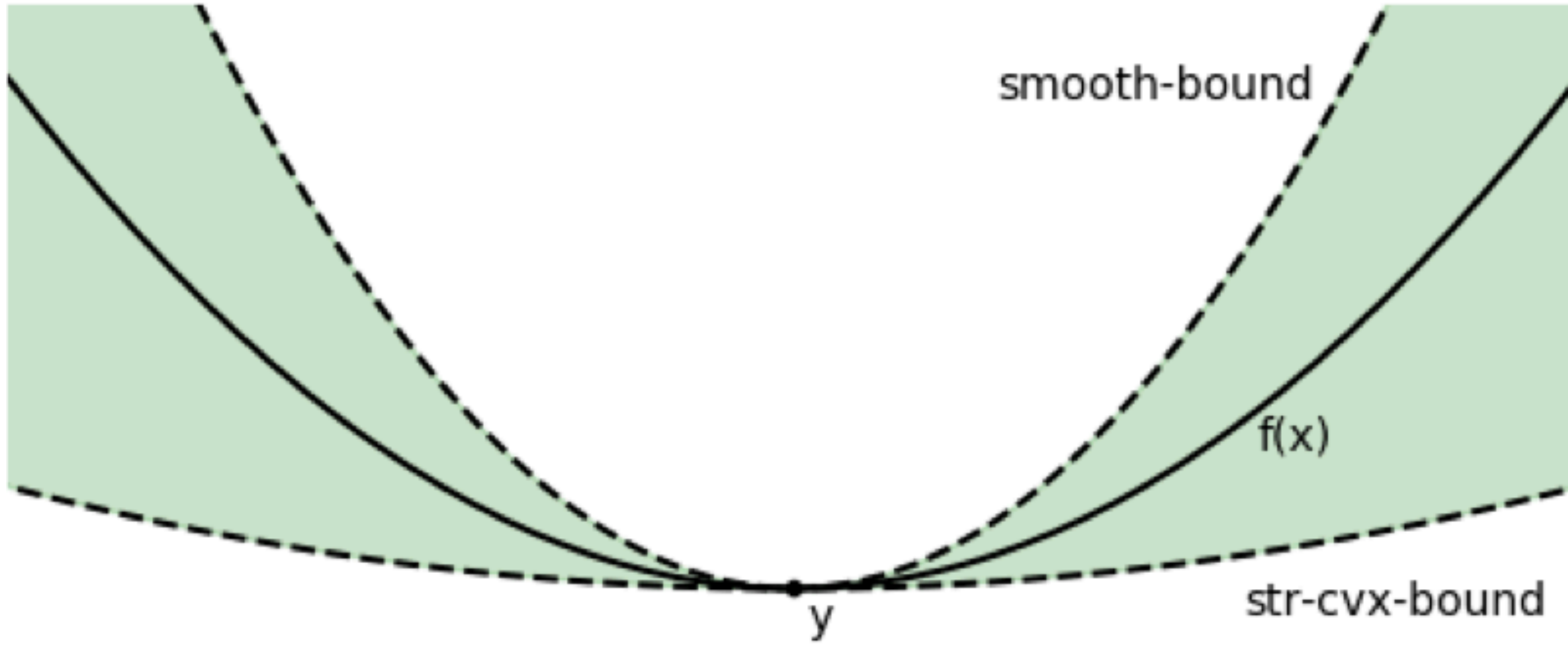
- Output $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \text{clip}_\tau(x_i) + \mathcal{N}(0, \rho^2)$.
- What if we think of this as an iterative algorithm of n steps with $\gamma = \frac{1}{n}$:
  - $\hat{\mu}_t = \hat{\mu}_{t-1} - \gamma\left(\text{clip}_\tau(x_i) + \mathcal{N}(0, \rho^2)\right)$
  - Privacy analysis?
  - Error analysis?

Bonus



Convergence analysis

Gradient descent



- $\theta_t = \theta_{t-1} - \gamma_t \nabla L(\theta_{t-1})$
- $\frac{\mu}{2}\|\Delta\theta\|_2^2 \geq L(\theta_t + \Delta\theta) - \left(L(\theta_t) + \nabla L(\theta_t)^\top \Delta\theta\right) \leq \frac{\beta}{2}\|\Delta\theta\|_2^2$

μ-strongly-convex

β-Smoothness

Theorem

If L is β-smooth and μ-strongly convex, gradient descent with $\gamma_t = 1/\beta$ converges as

$$L(\theta_t) - \min_\theta L(\theta) \leq \left(1 - \frac{\mu}{\beta}\right)^t \|\theta_0 - \theta^\star\|_2^2$$

Understanding Gradient Descent

Convergence analysis

- One final assumption: how bad is this approximation?
- $$\max_{\theta} E\|\nabla \ell_t(\theta) - \nabla L(\theta)\|_2^2 \leq \sigma^2$$
- Proofs cheat sheet: https://gowerrobert.github.io/pdf/M2_statistique_optimisation/grad_conv.pdf

Theorem

If L is β-smooth and μ-strongly convex, SGD with step-size γ converges as

$$E\|\theta^t - \theta^\star\|_2^2 \leq (1 - \gamma\mu)E\|\theta^{t-1} - \theta^\star\|_2^2 + \gamma^2\sigma^2$$