Modeling Privacy Erosion: Differential Privacy Dynamics in Machine Learning

Reza Shokri

Data Privacy and Trustworthy ML Research Lab

National University of Singapore

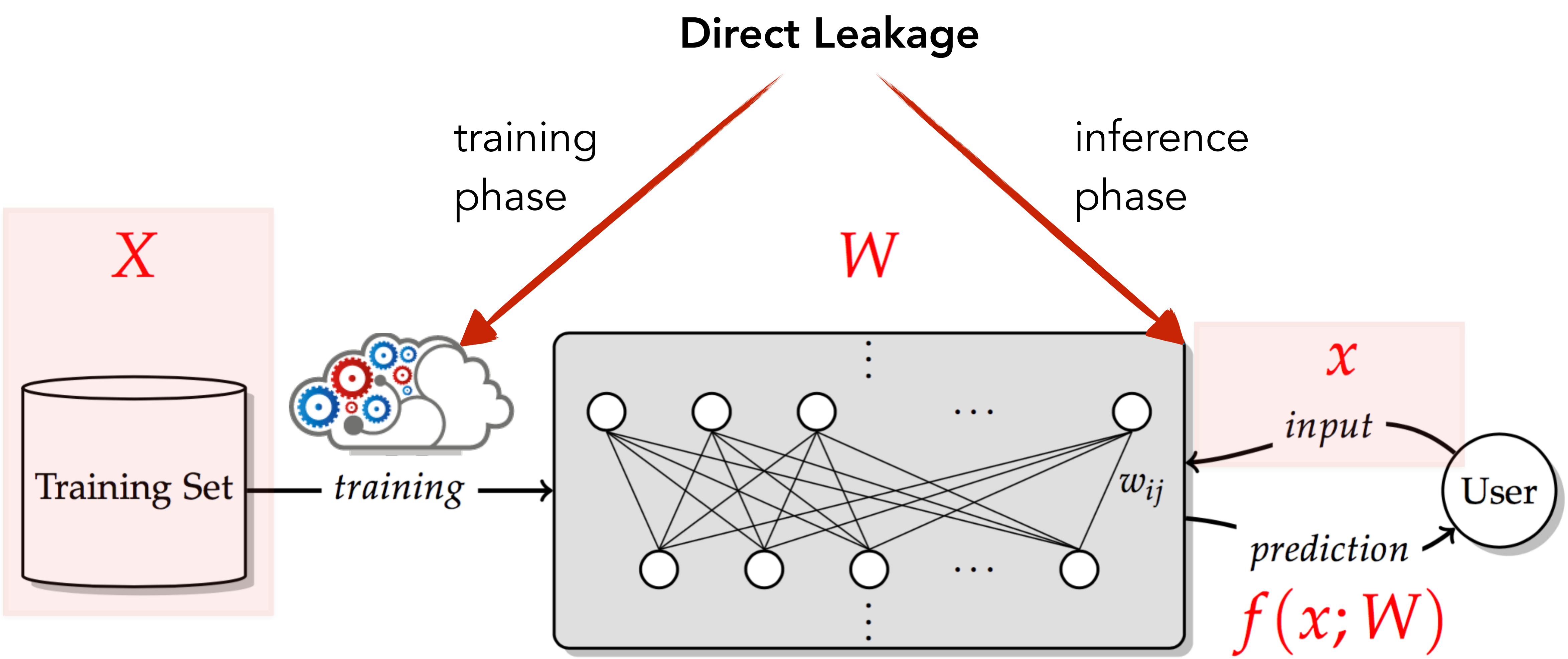
reza@comp.nus.edu.sg @rzshokri



Based on joint work with: Rishav Chourasia*, and Jiayuan Ye*

https://arxiv.org/abs/2102.05855

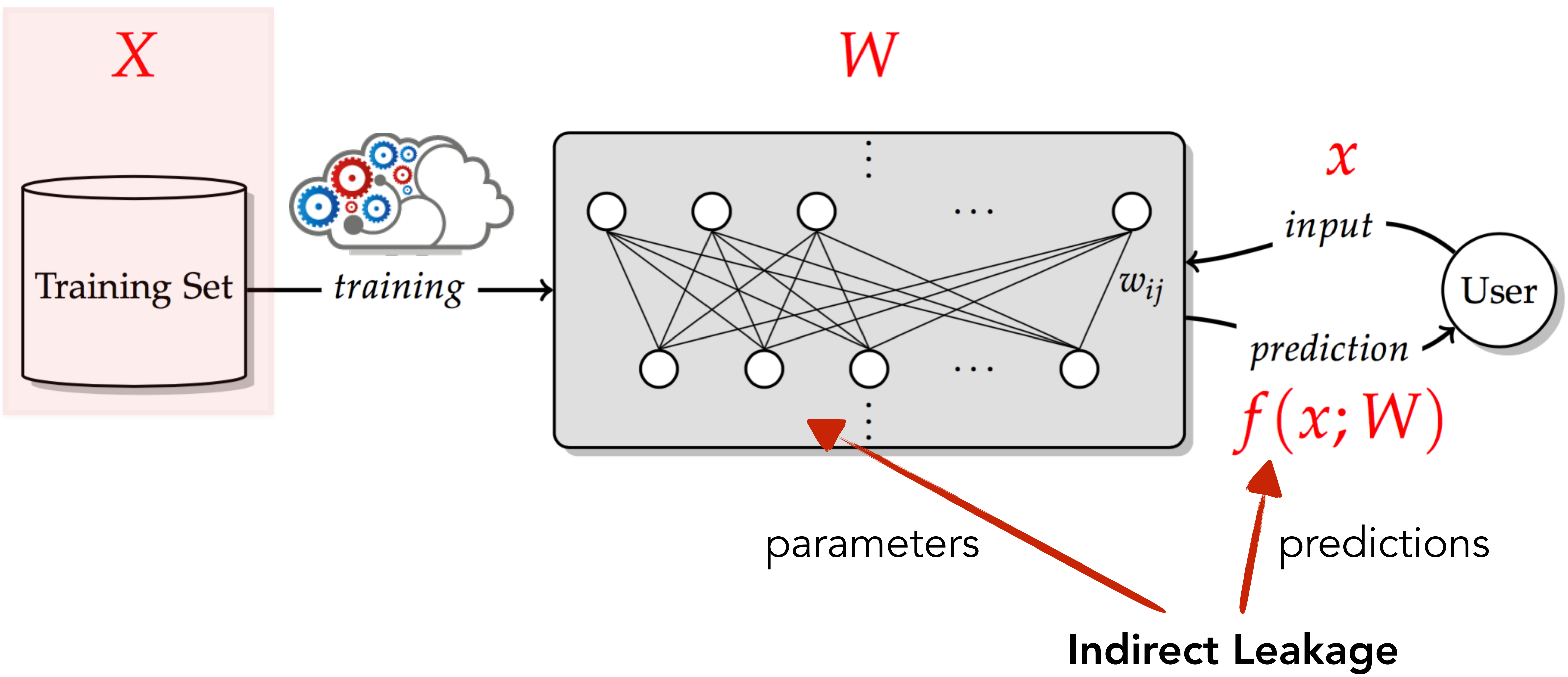
Privacy Risks in Machine Learning



<u>How to prevent this leakage?</u> Secure multi-party computation, homomorphic encryption, trusted hardware, …

Privacy Risks in Machine Learning

What is leakage? Inferring information about members of X, beyond what can be learned about its underlying distribution

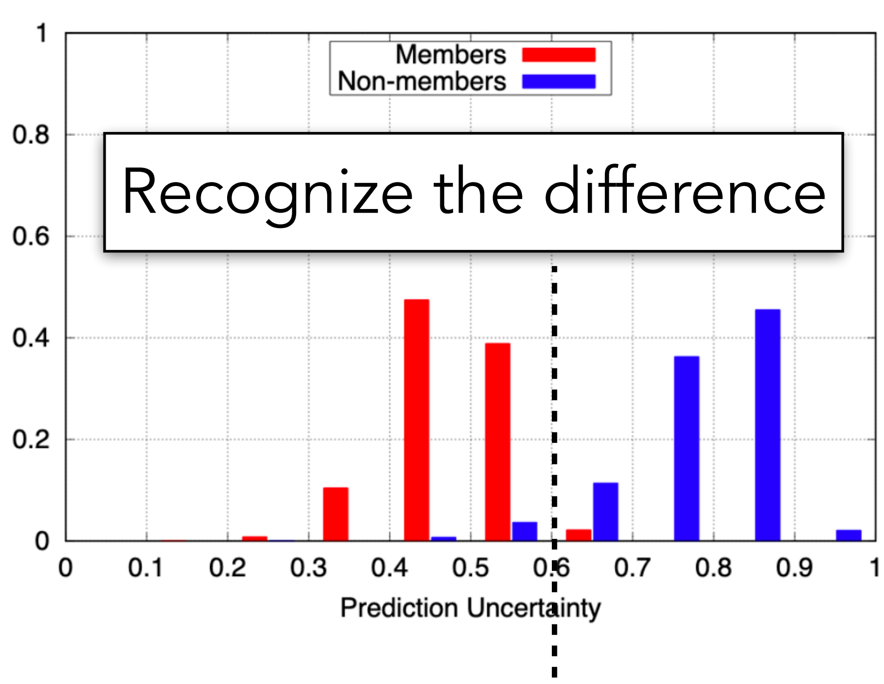


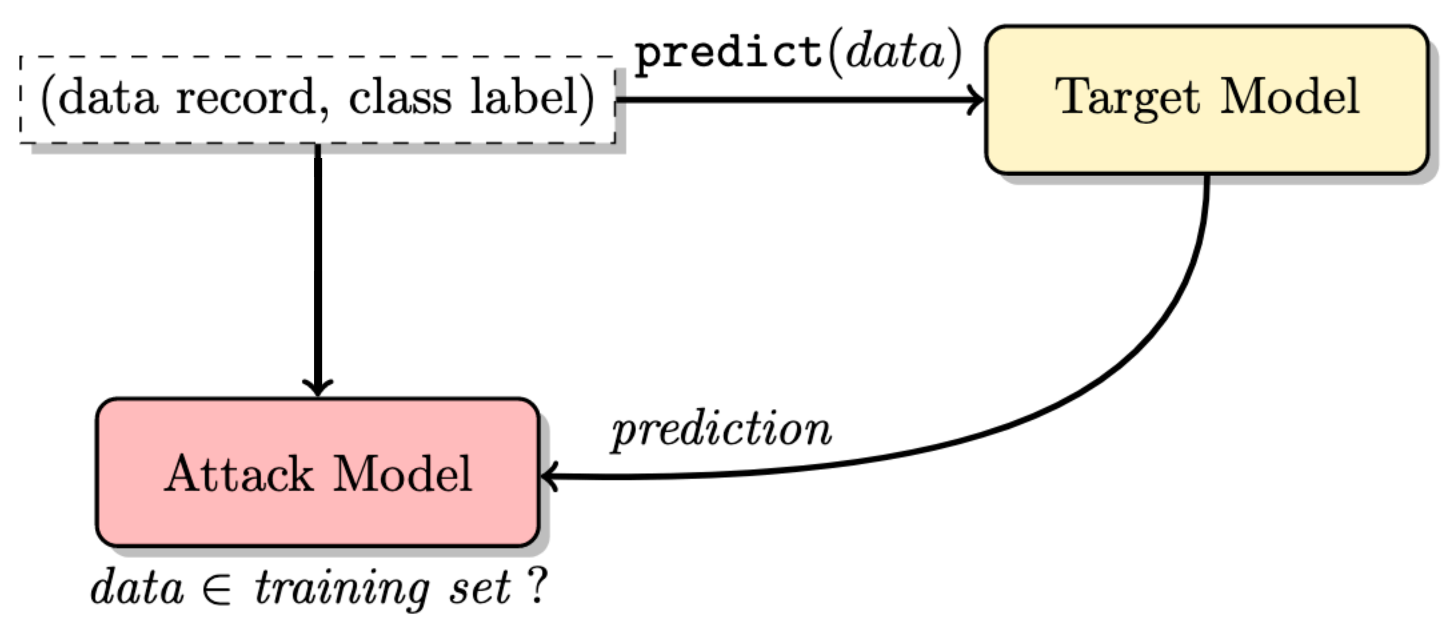
[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP'17

[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

How to Quantify the Leakage?

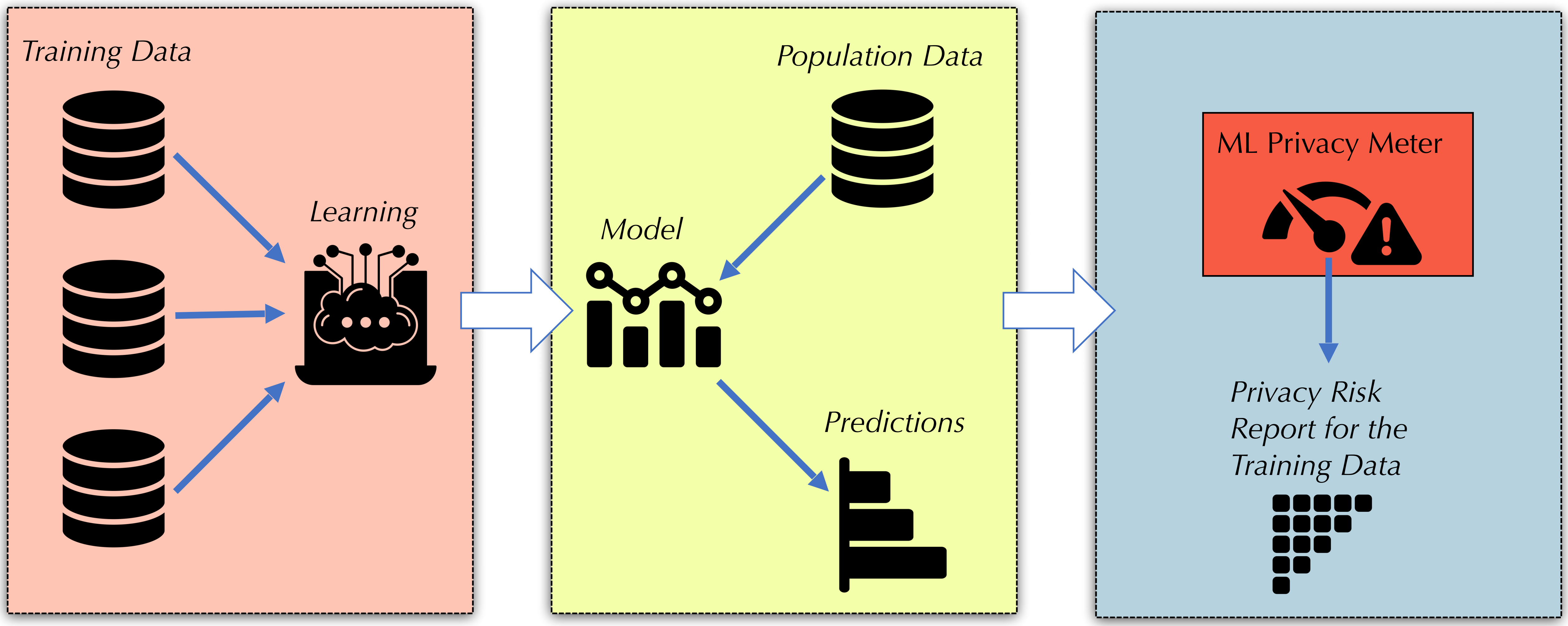
- Indistinguishability game: Can an adversary distinguish between two models that are trained on two neighboring datasets (only one includes data point x)?
    - <u>Membership inference</u>: Given a model, can an adversary infer whether data point x is part of its training set?





[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP’17

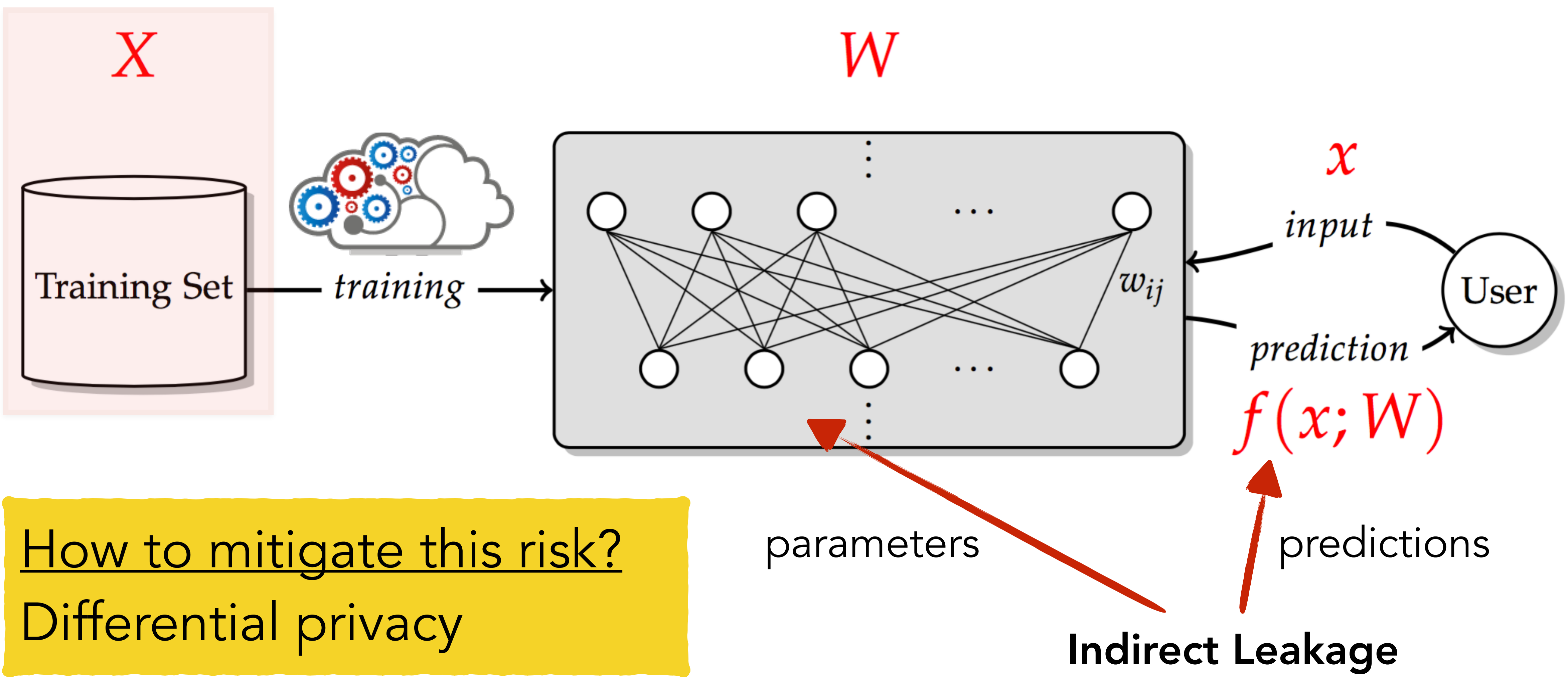
Tool: ML Privacy Meter



ML Privacy Meter is a Python library (`ml_privacy_meter`) that enables quantifying the privacy risks of machine learning models. https://github.com/privacytrustlab/ml_privacy_meter

Privacy Risks in Machine Learning

What is leakage? Inferring information about members of X, beyond what can be learned about its underlying distribution



How to mitigate this risk?
Differential privacy

[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP'17

[Nasr, Shokri, Houmansadr] Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, SP'19

Differential Privacy

- A randomized algorithm $\mathscr{A}$ satisfies (ϵ, δ)-DP, if for any two neighboring datasets D, D', and all sets S

$$\Pr[\mathscr{A}(D) \in S] \leq e^{\epsilon} \Pr[\mathscr{A}(D') \in S] + \delta$$

[Dwork, McSherry, Nissim, Smith] Calibrating Noise to Sensitivity in Private Data Analysis, TCC 2006

Renyi Differential Privacy

- A randomized algorithm $\mathscr{A}$ satisfies (α, ϵ)-Renyi DP, for $\alpha > 1$, if for any two neighboring datasets D, D',

$$R_\alpha(\mathscr{A}(D) \,|\, \mathscr{A}(D')) \leq \epsilon$$

- $R_\alpha(P \| Q)$ is the α-Renyi divergence of P with respect to Q

$$R_\alpha(P \| Q) = \frac{1}{\alpha - 1} \log \mathop{\mathbb{E}}_{z \sim Q} \left[\left(\frac{P(z)}{Q(z)} \right)^\alpha \right]$$

- (α, ϵ)-RDP is $(\epsilon + \dfrac{\log 1/\delta}{\alpha - 1}, \delta)$-DP for any δ

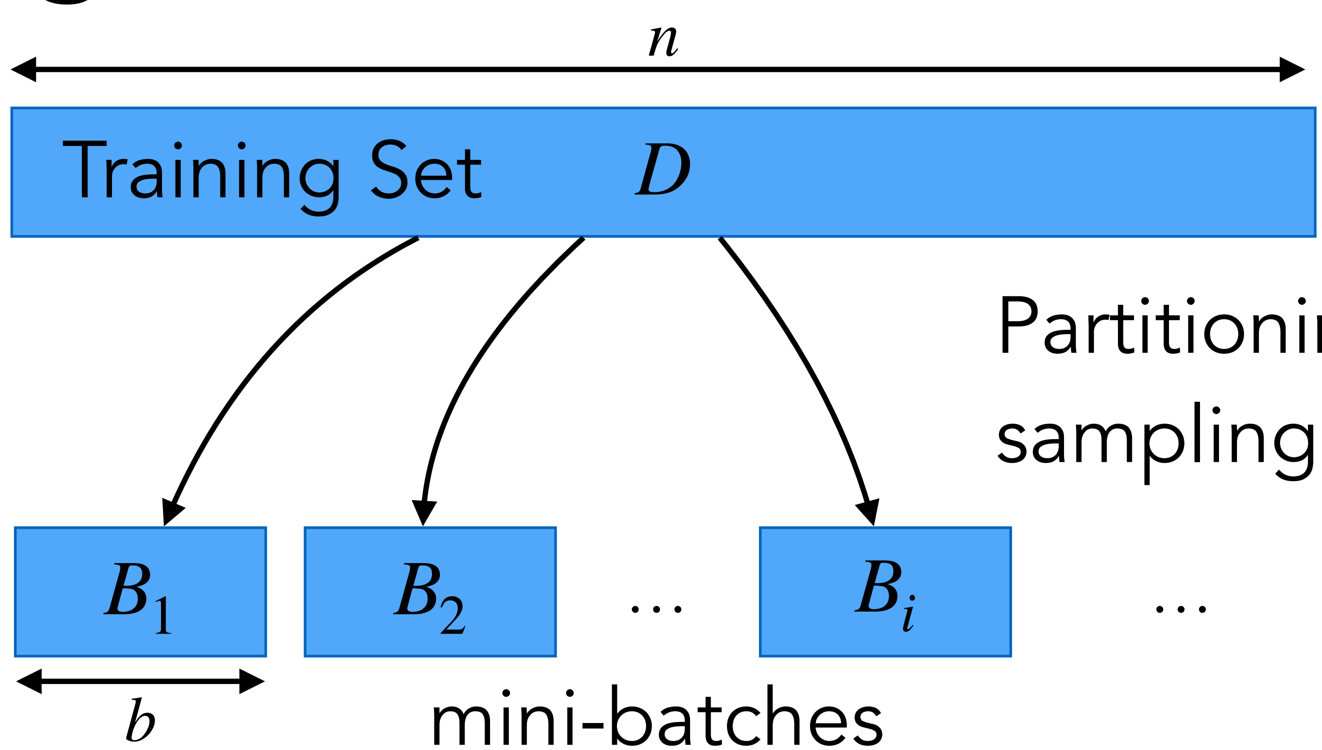
- Composition of (α, ϵ_1)-RDP and (α, ϵ_2)-RDP is $(\alpha, \epsilon_1 + \epsilon_2)$-RDP

[Mironov] Rényi differential privacy. CSF 2017

Learning with Differential Privacy

One iteration:

n/b mini-batches



Partitioning, or
sampling with replacement

Parameter Update

Average gradient

$$g_i(\theta_i; \mathcal{D}) = \frac{1}{b} \sum_{j \in B_i} \nabla \ell(\theta_i; x_j)$$

$$\theta_{i+1} = \theta_i - \eta \left(g_i(\theta_i; \mathcal{D}) + \mathcal{N}(0, \sigma^2 \mathbb{I}_d) \right)$$

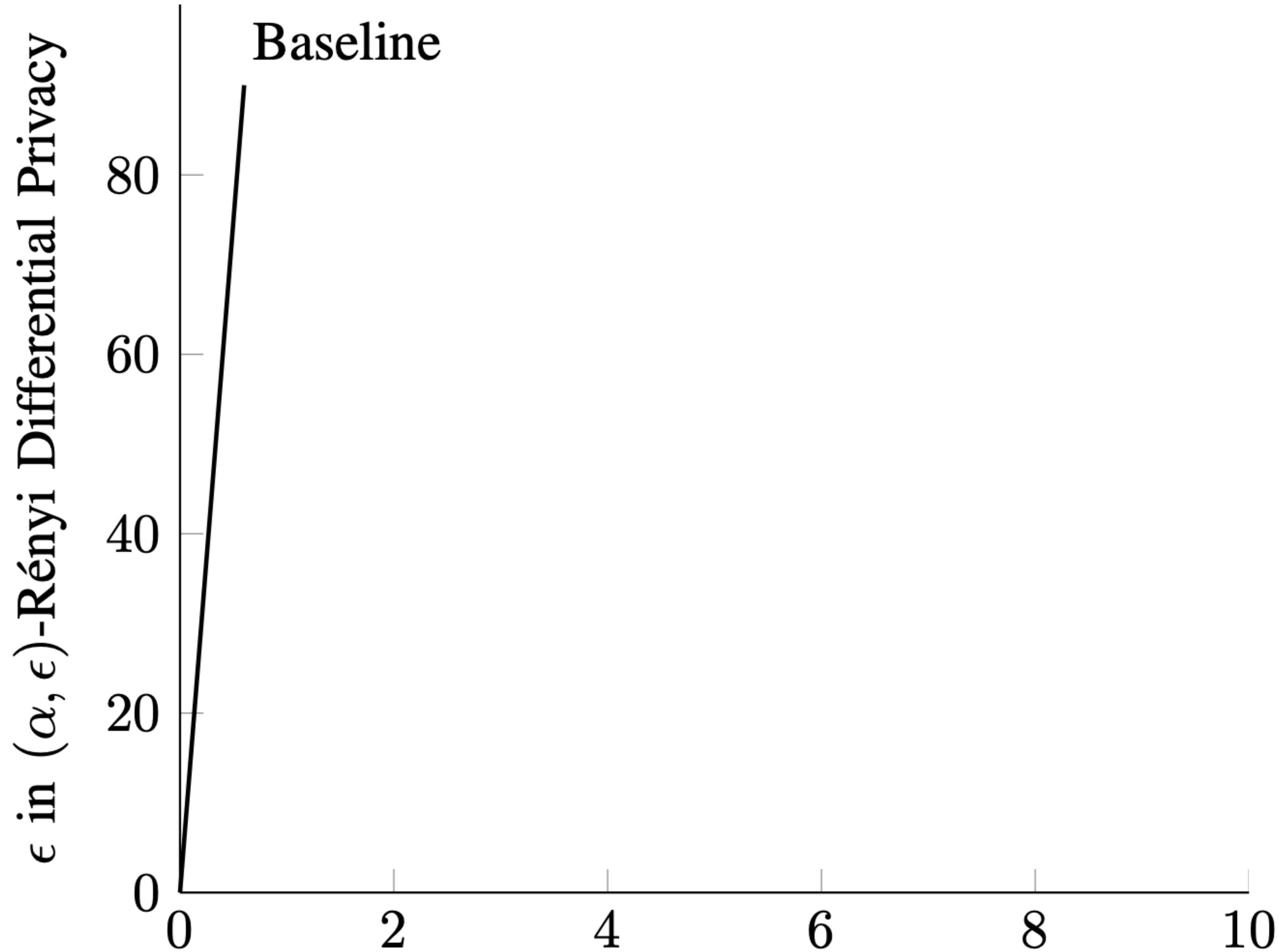
Model parameters

Step-size

Privacy Loss (Composition)

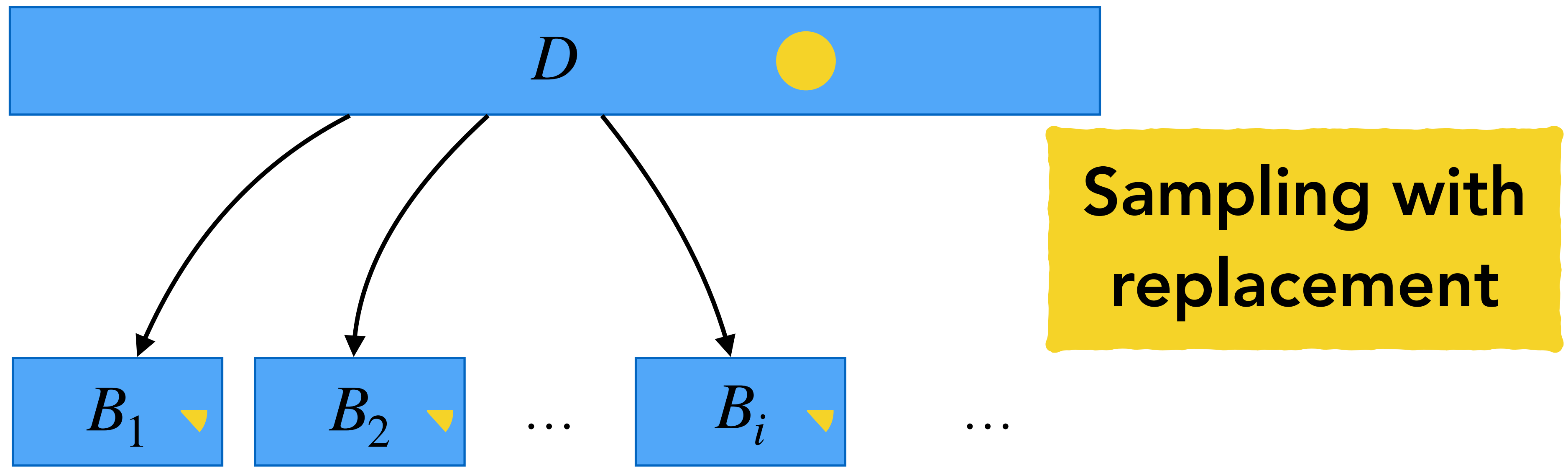
Leakage of one update

$$\boxed{\varepsilon = \frac{2\alpha L^2}{\sigma^2} \cdot \frac{n}{b} \cdot K}$$



Iterations: K (passes over the whole data)

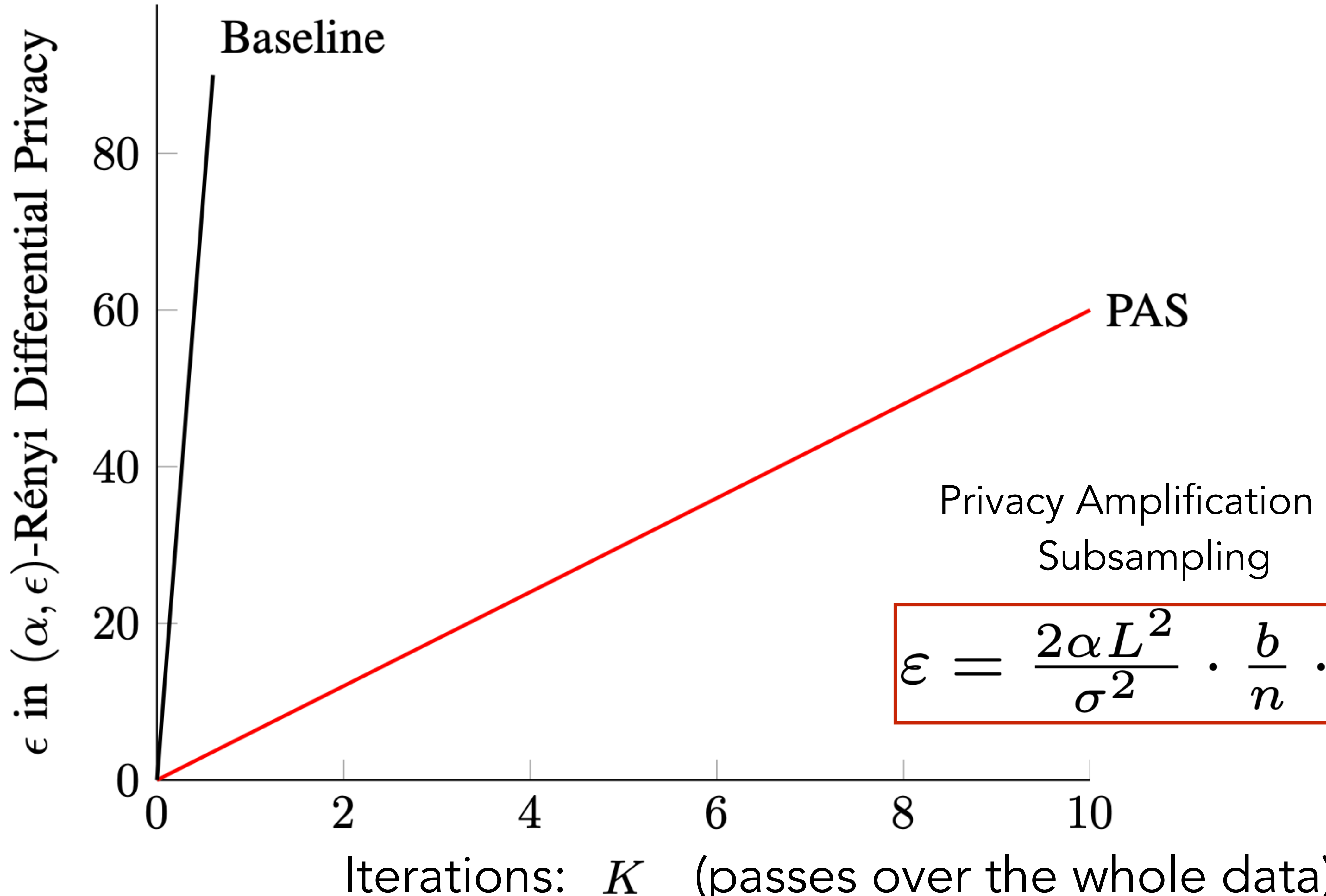
Learning with Differential Privacy



Privacy Loss (Composition)

Leakage of one update

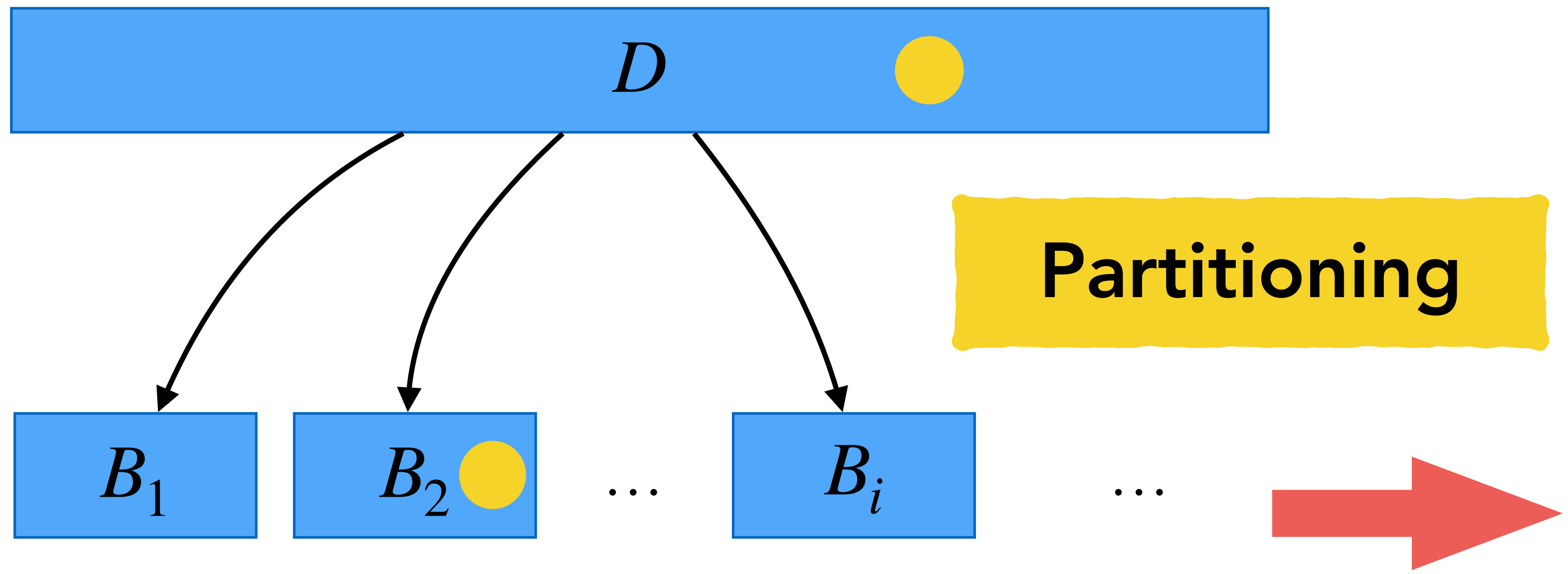
$$\varepsilon = \frac{2\alpha L^2}{\sigma^2} \cdot \frac{n}{b} \cdot K$$



Privacy Amplification by Subsampling

$$\varepsilon = \frac{2\alpha L^2}{\sigma^2} \cdot \frac{b}{n} \cdot K$$

Learning with Differential Privacy



Reveal parameters only
at the end of an iteration

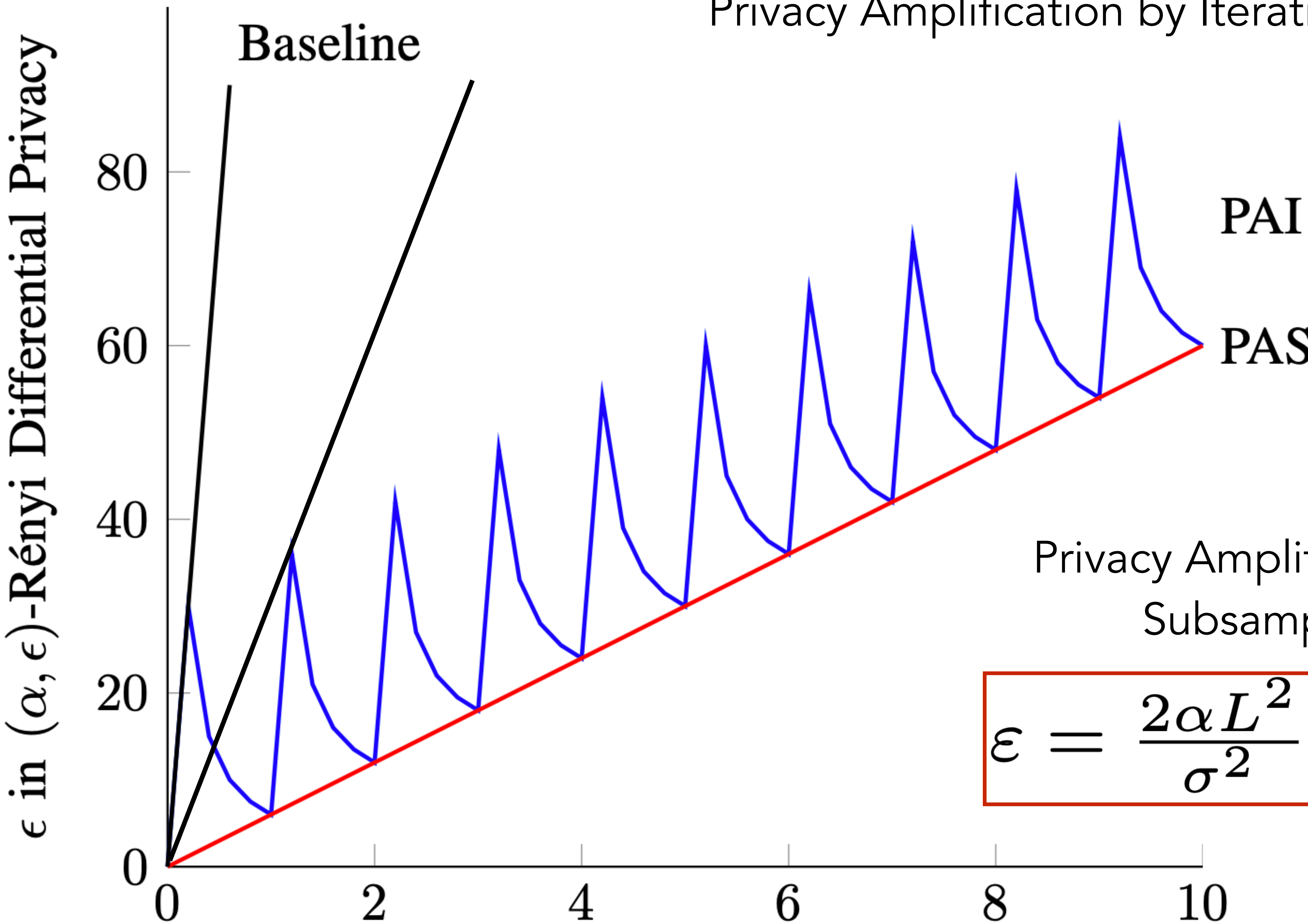
Privacy Loss (Composition)

Leakage of one update

$$\varepsilon = \frac{2\alpha L^2}{\sigma^2} \cdot K$$

$$\varepsilon = \frac{2\alpha L^2}{\sigma^2} \cdot \frac{b}{n} \cdot (K-1) + \frac{2\alpha L^2}{i \cdot \sigma^2}$$

Privacy Amplification by Iteration



Privacy Amplification by Subsampling

$$\varepsilon = \frac{2\alpha L^2}{\sigma^2} \cdot \frac{b}{n} \cdot K$$

Iterations: K (passes over the whole data)

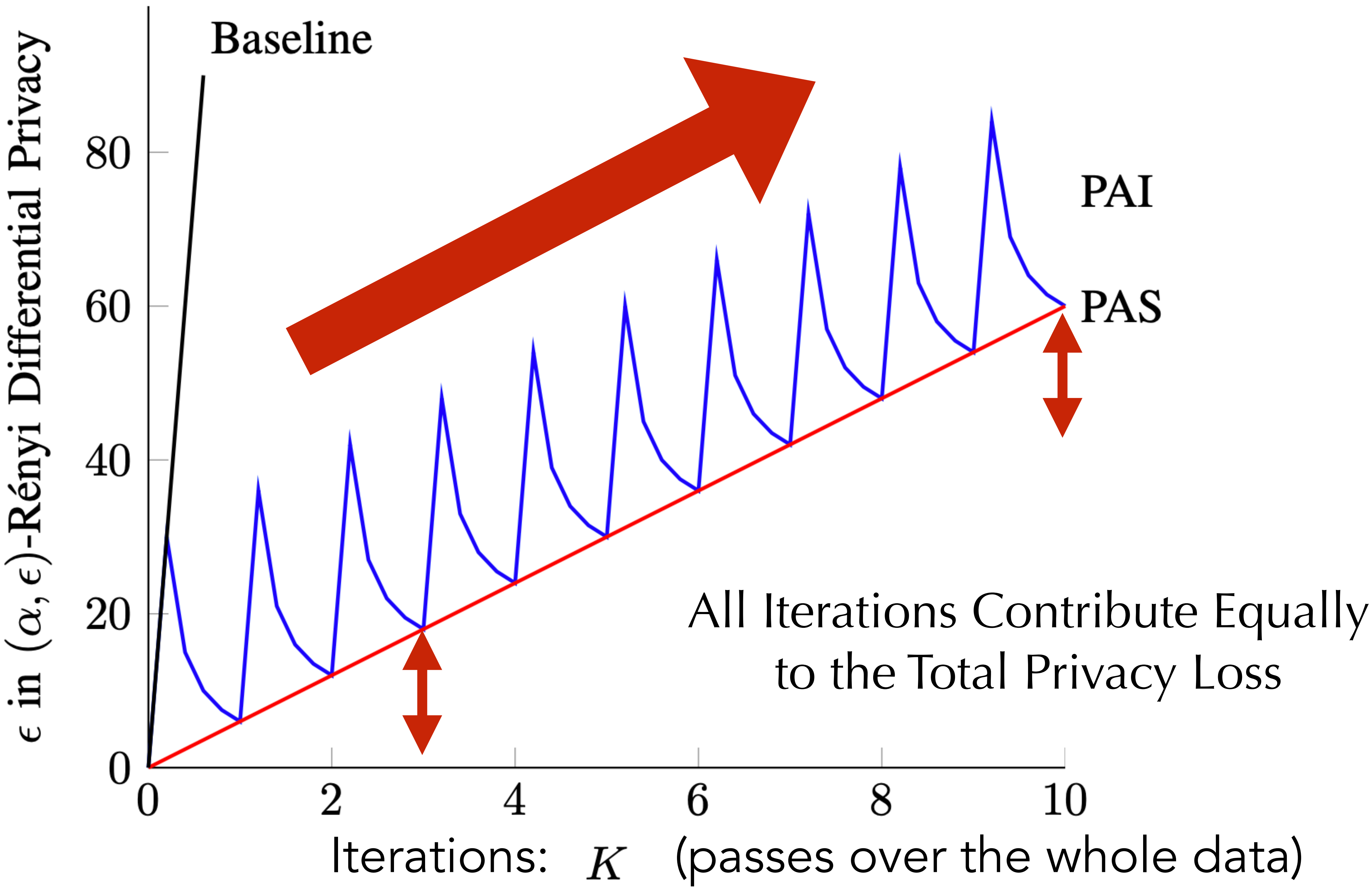
Observation 1:

This analysis accounts for privacy loss of all iterations,

even if the only observables are the model parameters at the final iteration K

Observation 2:

Privacy loss increases with a linear rate



Differential Privacy Dynamics

- Assume that adversary observes the model parameters at iteration K, and the state of the algorithm is **private** throughout the training
- How does privacy loss change over time?
  - What is the difference between $R_\alpha(\mathscr{A}_{K-1}(D) \,|\, \mathscr{A}_{K-1}(D'))$ and $R_\alpha(\mathscr{A}_K(D) \,|\, \mathscr{A}_K(D'))$ for various K?

Noisy Gradient Descent

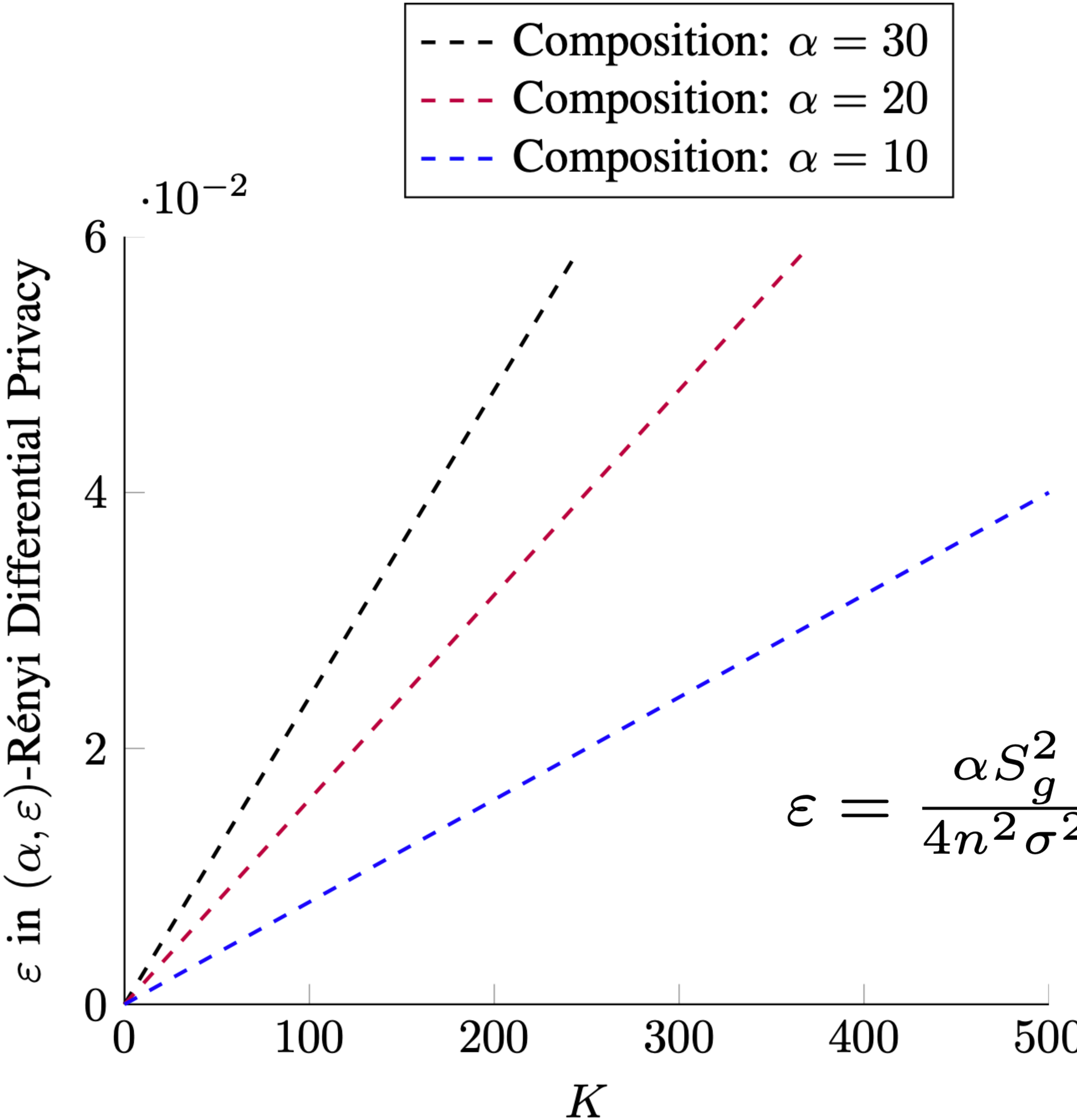
Input: Dataset $\mathcal{D} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$, loss function ℓ, learning rate η, noise variance σ^2, initial parameter vector θ_0.

1: **for** $k = 0, 1, \cdots, K-1$ **do**

2: $\quad g(\theta_k; \mathcal{D}) = \sum_{i=1}^{n} \nabla \ell(\theta_k; \mathbf{x_i})$

3: $\quad \theta_{k+1} = \theta_k - \frac{\eta}{n} g(\theta_k; \mathcal{D}) + \sqrt{2\eta\sigma^2}\mathcal{N}(0, \mathbb{I}_d)$

4: Output θ_K



$$\varepsilon = \frac{\alpha S_g^2}{4n^2\sigma^2} \cdot \eta K$$

Dynamics of RDP in Noisy GD

- Noisy GD is a discrete-time stochastic process
- **Coupled stochastic processes**: Let D, D' be two neighboring datasets. Let $\{\Theta_k\}_{k\geq 0}$ and $\{\Theta'_k\}_{k\geq 0}$ be the sequence of probability distributions over training iterations on D, D', respectively. We assume $\Theta_0 = \Theta'_0$ are initial parameter distributions.
- Renyi divergence $R_\alpha(\Theta_K | \Theta'_K)$ reflects the privacy loss at iteration K

Dynamics of RDP in Noisy GD

- We trace the changes in privacy loss of this discrete-time stochastic process, with a continuous-time stochastic process, which matches the probability distributions at each iteration

$$\frac{\partial R(\alpha, t)}{\partial t} \leq \frac{\alpha S_g^2}{2\sigma^2 n^2 (1-\eta\beta)^2} - \sigma^2 c\left[\frac{R(\alpha, t)}{\alpha} + (\alpha - 1)\frac{\partial R(\alpha, t)}{\partial \alpha}\right]$$

Rate of Privacy Loss

Privacy Loss

RDP for Noisy GD

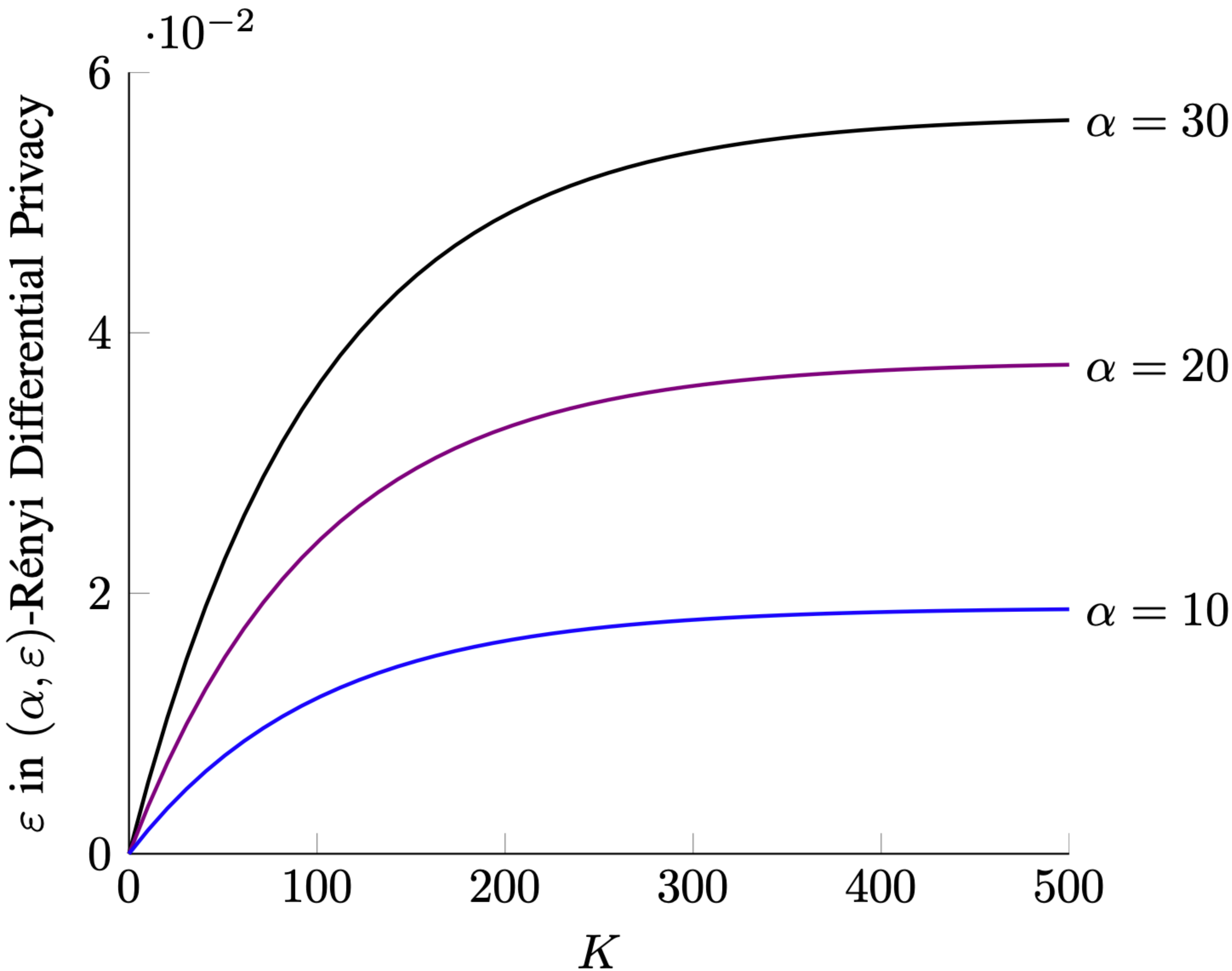
Theorem *(RDP for Noisy GD) The noisy GD algorithm with loss function $\ell(\theta; \mathbf{x})$, learning rate η, and noise variance σ^2, satisfies (α, ε) Rényi differential privacy with*

$$\boxed{\varepsilon \geq \frac{\alpha S_g^2}{\lambda \sigma^2 n^2 (1 - \eta\beta)^2} \left(1 - e^{-\lambda \eta K/2}\right),}$$

If

1) $\theta_0 \sim \mathcal{N}\left(0, \frac{2\sigma^2}{\lambda} \mathbb{I}_d\right)$,
2) *loss function $\ell(\theta; \mathbf{x})$ is λ-strongly convex and β-smooth, and total gradient $g(\theta; \mathcal{D}) = \sum_{\mathbf{x}_i \in \mathcal{D}} \nabla \ell(\theta; \mathbf{x}_i)$ has a finite sensitivity S_g,*
3) *update step-size $\eta < \frac{1}{\beta}$.*

$$\varepsilon = \frac{\alpha S_g^2}{\lambda \sigma^2 n^2 (1-\eta\beta)^2} \cdot \left(1 - e^{-\lambda\eta K/2}\right)$$



Lower Bound on RDP

Theorem 5 (Lower Bound on Rényi DP for Noisy GD). *There exist two neighboring datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^n$, a start distribution p_0, and a β-smooth loss function $\ell(\theta; \mathbf{x})$ which has a total gradient $g(\theta; \mathcal{D})$ with finite sensitivity S_g, such that for some constants $a_1, a_2 > 0$, and for for any $K \in \mathbb{N}$, the Rényi privacy loss of $\mathcal{A}_{Noisy\text{-}GD}$ on $\mathcal{D}, \mathcal{D}'$ with step-size η and noise variance σ^2 is lower-bounded by*

$$\boxed{R_\alpha\left(\Theta_{\eta K} \middle\| \Theta'_{\eta K}\right) \geq \frac{a_1 \alpha}{\sigma^2 n^2}\left(1 - e^{-a_2 \eta K}\right).}$$

Tightness Analysis