Differential Privacy Dynamics of Langevin Diffusion and Noisy Gradient Descent

Rishav Chourasia*, Jiayuan Ye*, Reza Shokri
Department of Computer Science, National University of Singapore
{rishav1, jiayuan, reza}@comp.nus.edu.sg
*Equal contribution. Alphabetical Order.

Privacy Risks of ML Algorithms

Privacy Risk: output model leaks information about the individual members of its training dataset

- Membership inference attacks Shokri, Stronati, Song, Shmatikov (2017) [5]
- Reconstruction attacks Carlini, Tramèr, et al. (2021) [2]

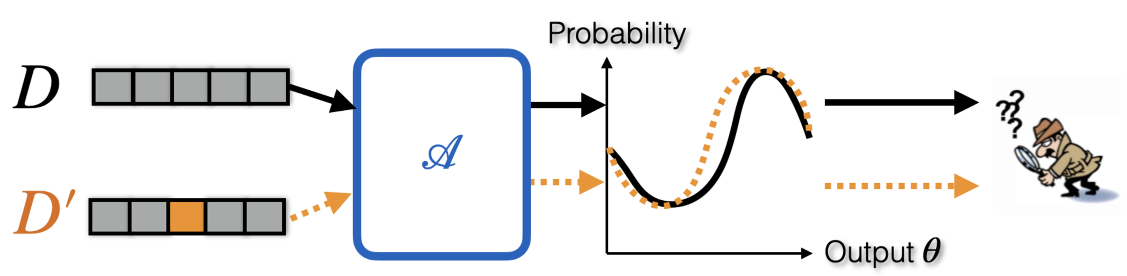
Differential Privacy

- Differential Privacy the distribution of the output for algorithm $\mathcal{A}$, on any neighboring input datasets, are indistinguishable.
- Rényi Differential Privacy[3] We say algorithm $\mathcal{A}$ satisfies (α, ϵ)-Rényi DP, if for any neighboring datasets D and D',

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

- Rényi Divergence

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



How to Train Privacy-preserving Model

- $\theta_0 \leftarrow$ initialization
- Dataset $D = (x_1, \cdots, x_n)$
- For $k = 1, \cdots, K$ do
  - θ_{k+1} = Update (θ_k, D) + Noise — (α, ϵ) - Rényi DP
- Output θ_K and $\theta_{K-1}, \cdots, \theta_1$ — $(\alpha, \epsilon \cdot K)$ - Rényi DP

Has a Complicated Distribution

DP Composition Analysis

≥

Problem: how to bound the Rényi privacy loss $R_\alpha(\theta_K\|\theta'_K)$

How to Compute a Better Bound

In this paper, we offer a new privacy analysis for the Noisy Gradient Descent on a certain class of loss functions, that

- analyzes the privacy loss for revealing the final model θ_K
- assumes hidden intermediate models $\theta_1, \cdots, \theta_{K-1}$

Input: Dataset $\mathcal{D} = (x_1, x_2, \cdots, x_n)$, loss function ℓ, learning rate η, noise variance σ^2, initial parameter vector θ_0.

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

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

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

4: Output θ_K

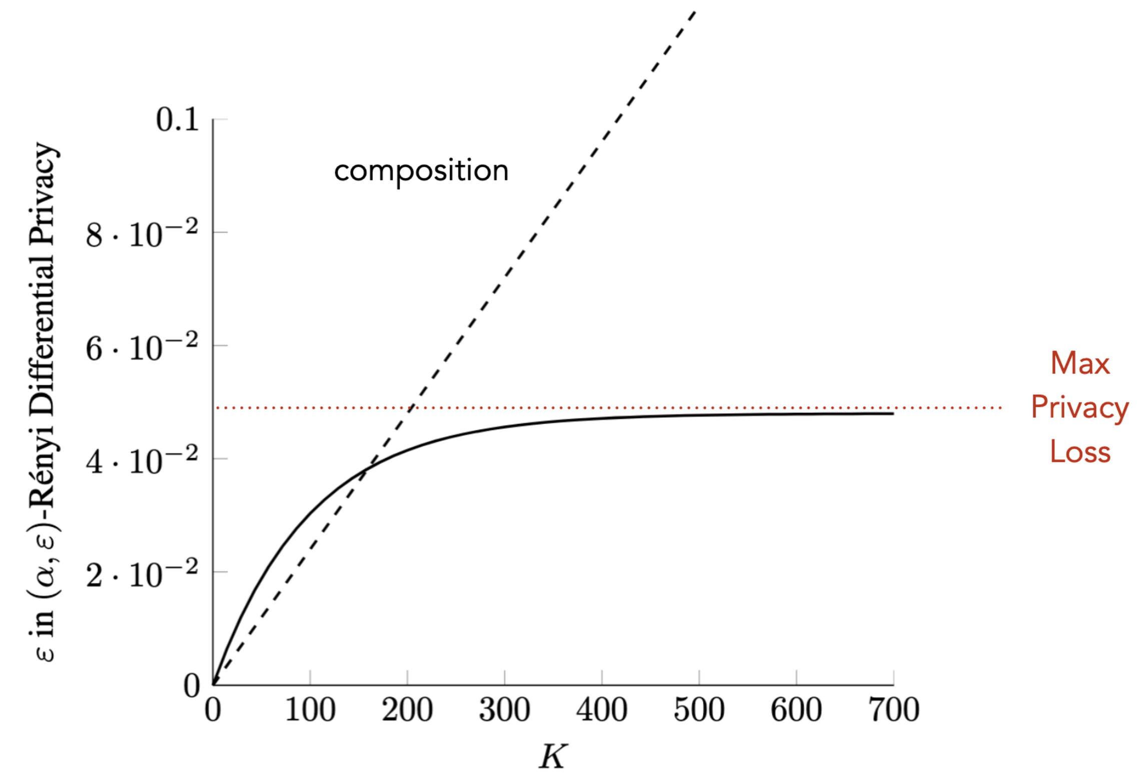
Privacy Dynamics Bound

Main Theorem: Noisy GD on λ-strongly convex β-smooth loss functions with gradient sensitivity $S_g = \max_{D,D'}\|g(\theta; D) - g(\theta; D')\|_2$, step-size $\eta \leq 1/\beta$ and K iterations satisfies (α, ϵ)-Rényi DP

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

Max Privacy Loss — Privacy Loss Convergence Rate

Parameters:
$\alpha = 30$,
$\sigma = 0.02$,
$S_g = 4$,
$\eta = 0.02$,
$\lambda = 1$
Size of dataset:
$n = 5000$



Our Privacy Analysis is Tight

- Exact Privacy Loss Lower Bound: compute exact privacy loss for noisy GD on the squared norm loss function $\ell(\theta; x) = \|\theta - x\|^2/2$, where the output distribution is Gaussian

$$\epsilon \geq \frac{\alpha S_g^2}{4\sigma^2 n^2} \cdot (1 - e^{-\eta K})$$

- Privacy Dynamics Bound:

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

- Tightness: the upper bound matches the lower bound up to a small constant of 4

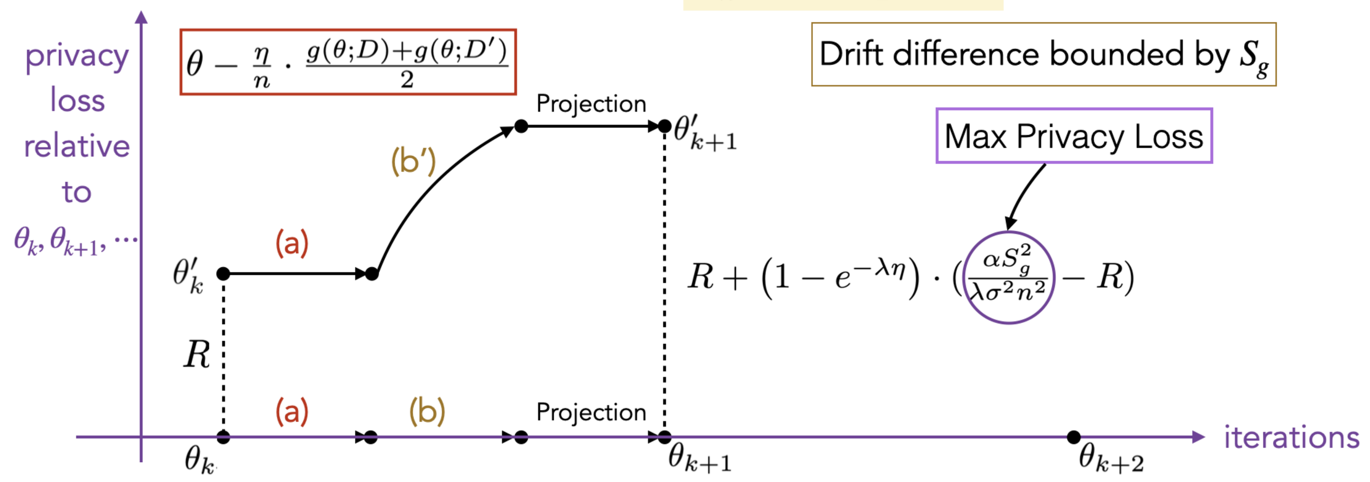
How to Prove Privacy Dynamics

Sketch: recursively bound the change of privacy loss in one update

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

Technique: decompose the update in the k-th iteration on neighboring datasets into three steps

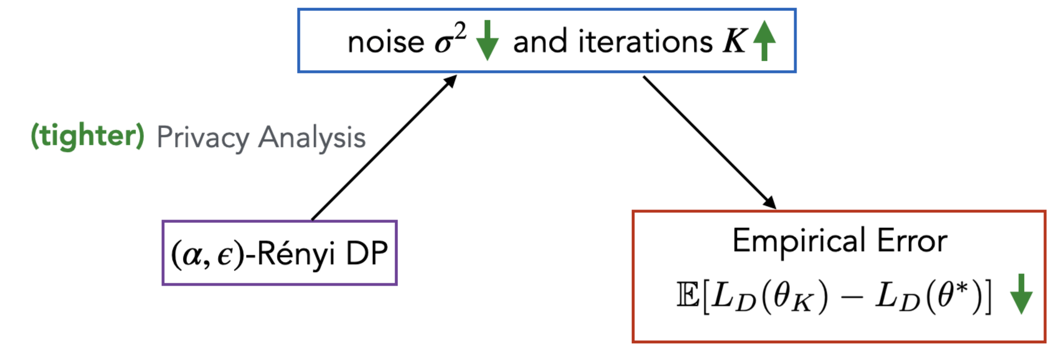
- (b) Langevin diffusion with drift $-\frac{1}{n} \cdot \frac{g(\theta_k; D) - g(\theta_k; D')}{2}$
- (b') Langevin diffusion with drift $-\frac{1}{n} \cdot \frac{g(\theta'_k; D') - g(\theta'_k; D)}{2}$



Utility Analysis

Goal

Analyze how does the added randomness required for achieving privacy by a privacy analysis affect the error of the algorithm's output?



Utility Gain From Our Tight Privacy Analysis

Privacy dynamics analysis facilitates a better privacy-utility tradeoff, under (α, ϵ)-Rényi DP than the composition analysis for strongly convex smooth loss functions.

$$\mathbb{E}[L_D(\theta_{K^*}) - L_D(\theta^*)] \leq \frac{\alpha}{\epsilon} \cdot \frac{\beta d L^2}{\lambda^2 n^2}$$

$poly(n)$ smaller runtime — $poly \log n$ smaller error

Matching Lower Bound in Previous Works

This error matches the lower bound [1] for (ϵ, δ)-differentially private empirical risk minimization for Lipschitz, strongly convex, and smooth loss function, up to a constant of $\log(1/\delta)$.

Summary

- We need more precise estimates of the privacy loss for differentially-private machine learning algorithms
  - How much does a trained model leak about its training data?
  - Assuming that intermediate steps of the training algorithm are private and not visible to adversary.
- We present a new tight converging privacy dynamics theorem for noisy gradient descent algorithms on strongly convex smooth loss functions
- Open problem: Privacy dynamics under relaxed conditions

References

[1] Raef Bassily, Adam Smith, and Abhradeep Thakurta.
Private empirical risk minimization: Efficient algorithms and tight error bounds.
In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464--473. IEEE, 2014.

[2] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al.
Extracting training data from large language models.
arXiv preprint arXiv:2012.07805, 2020.

[3] Ilya Mironov.
Rényi differential privacy.
In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263--275. IEEE, 2017.

[4] Claude E. Shannon.
A mathematical theory of communication.
Bell System Technical Journal, 27(3):379--423, 1948.

[5] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov.
Membership inference attacks against machine learning models.
In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3--18. IEEE, 2017.