Gradients Look Alike: Sensitivity is Often Overestimated in DP-SGD

Anvith Thudi, Hengrui Jia, Casey Meehan, Ilia Shumailov, Nicolas Papernot

UNIVERSITY OF TORONTO

VECTOR INSTITUTE

UNIVERSITY OF CALIFORNIA SAN DIEGO

UNIVERSITY OF OXFORD

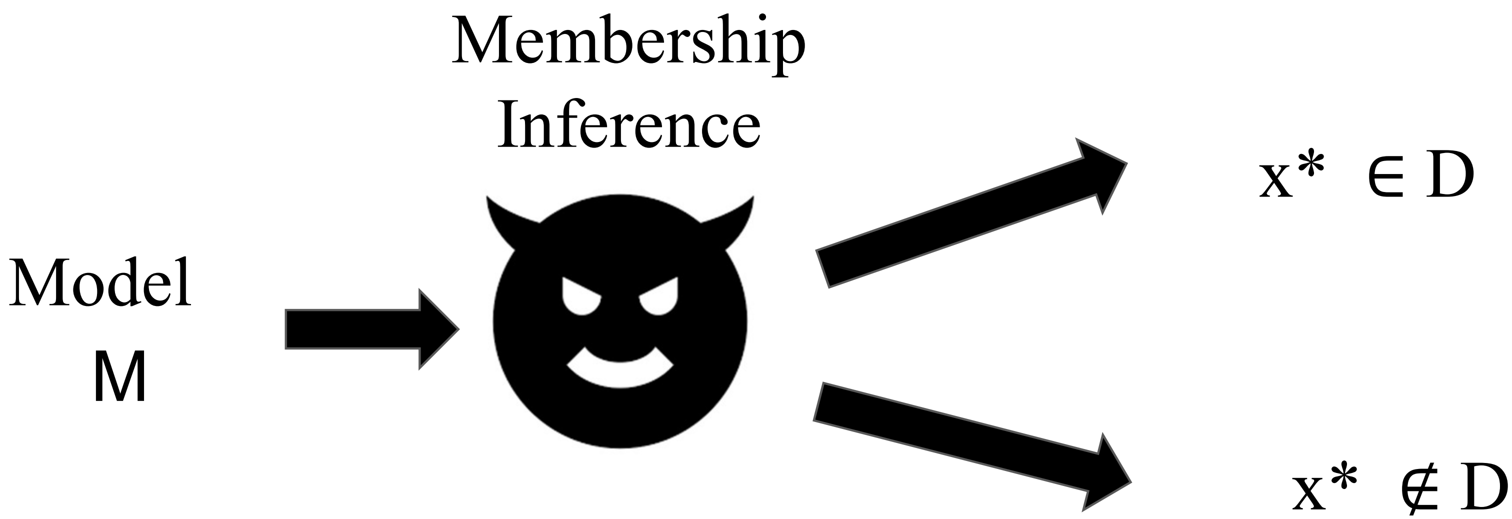
Outline

1. Primer on Private ML and an Open Problem

1. Explaining Gaps with Data-Dependent Analysis

Primer on Privacy

The Adversary



- Implies other privacy attacks

Main Q: How to protect against this adversary?

Differential Privacy

Renyi DP: For ***ALL*** adjacent training datasets D,D'

$$\frac{1}{\alpha - 1} \ln \mathbb{E}_{f(D')} \left(\frac{f(D)}{f(D')} \right)^{\alpha} \leq \epsilon$$

Model Training Algorithm

Bounds the adversary for all datapoints

How to Obtain DP: DP- SGD

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

Clip Gradients Per Example

Add Noise

“Deep Learning with Differential Privacy” [ACGMMTZ] CCS 2016

Private ML in the Wild

1) Can match the worst case guarantee of DP-SGD:
- *“Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning”* Nasr et al. **IEEE S&P**

1) But in most settings attacks are far away from the bound
- *For most Models, D,D’ pairs, we empirically don’t reach the bound on privacy leakage*

Towards Explaining This

1) Bounding Membership Inference Accuracy:
- *“Optimal Membership Inference Bounds for Adaptive Composition of Sampled Gaussian Mechanisms”* Mahloujifar et al. **Preprint**
- *“From Differential Privacy to Bounds on Membership Inference: Less can be More”* Thudi et al. **TMLR**

1) Bounding Reconstruction Attacks:
- *“Bounding Training Data Reconstruction in Private (Deep) Learning”* Guo et al. **ICML**

1) DP-like Guarantee with Additional Assumptions:
- *“Individual Privacy Accounting for Differentially Private Stochastic Gradient Descent”* Yu et al. **TMLR**

Either not Individual, Attack specific, or Weaker than the DP inequality

The Problem: Per-Instance DP

Show that, for many ***specific*** adjacent pair D,D'= D ∪ x*

$$D_\alpha(f(D)||f(D')) \ll \epsilon$$

Smaller than the worst case for DP-SGD

Implications

Memorization:

- Performance change between training with or without a specific point
- Leaks privacy hence bounded by Per-Instance DP

Unlearning:

- Change in models between training with or without a specific point
- Leaks privacy hence bounded by Per-Instance DP

How does a dataset give more privacy to a point?

DP-SGD Analysis

Bounding the Renyi Divergence for DP-SGD follows in two steps:

1) Bounds on the per-step divergence
2) Bounds on the composition of per-step divergences

So how can a dataset D make a point x* more private?

Datasets can mask per-step updates

Classical Analysis: Clipping uniformly bounds the sensitivity to any point

Observation: What happens if many other datapoints in the dataset give a similar update?

Sensitivity Distributions: Can derive per-step analysis with the distribution of updates coming from the dataset

A Sensitivity Distribution

Difference in || || minus || || of difference

$$\Delta_{U,\alpha}(X_B{}^{\tilde{\alpha}}, X'_B) \;:= \sum_i ||U(X_B{}^i)||_2^2 - (\alpha-1)||U(X'_B)||_2^2 - ||\Delta_\alpha(X_B{}^{\tilde{\alpha}}, X'_B)||_2^2$$

α mini batches from D, 1 from D'

$$\text{where } \Delta_\alpha(X_B{}^{\tilde{\alpha}}, X'_B) = (\textstyle\sum_i U(X_B{}^i)) - (\alpha-1)U(X'_B).$$

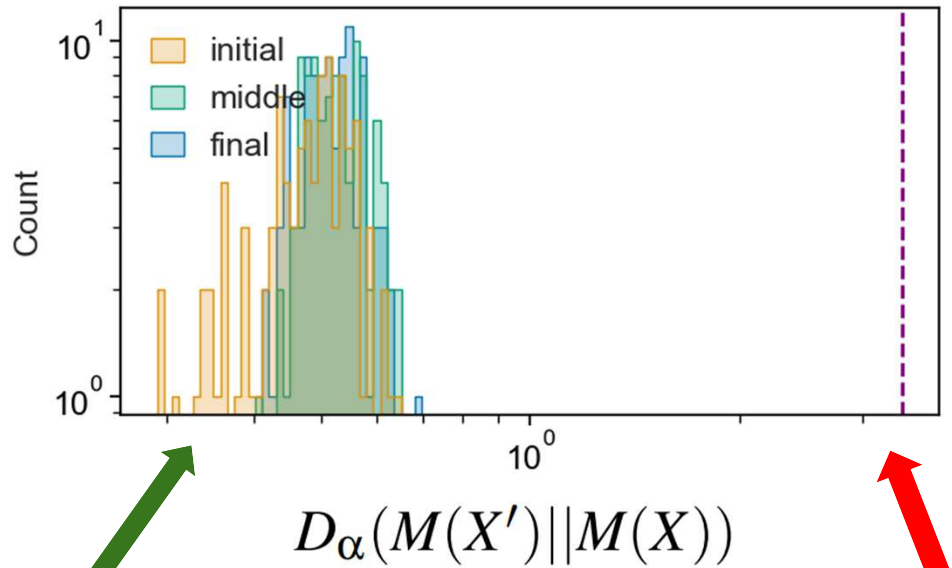
The Guarantee

Theorem 3.6. *Let $\alpha > 1$ be an integer. Given two arbitrary datasets X, X', the sampled Gaussian mechanism M with noise σ satisfies:*

“Expectation” of sensitivity

$$D_\alpha(M(X')||M(X)) \leq \frac{1}{(\alpha-1)} \mathbb{E}_{X_B}(\ln(\mathbb{E}_{{X'_B}^{\tilde{\alpha}}}(e^{\frac{-1}{2\sigma^2}\Delta_{U,\alpha}({X'_B}^{\tilde{\alpha}}, X_B)})))$$

Per-Step: Most Points are Better Than Worst Case



Improvement Across Training Steps

Data-Independent Bound

Results for CIFAR10

Datasets can lead to more private models

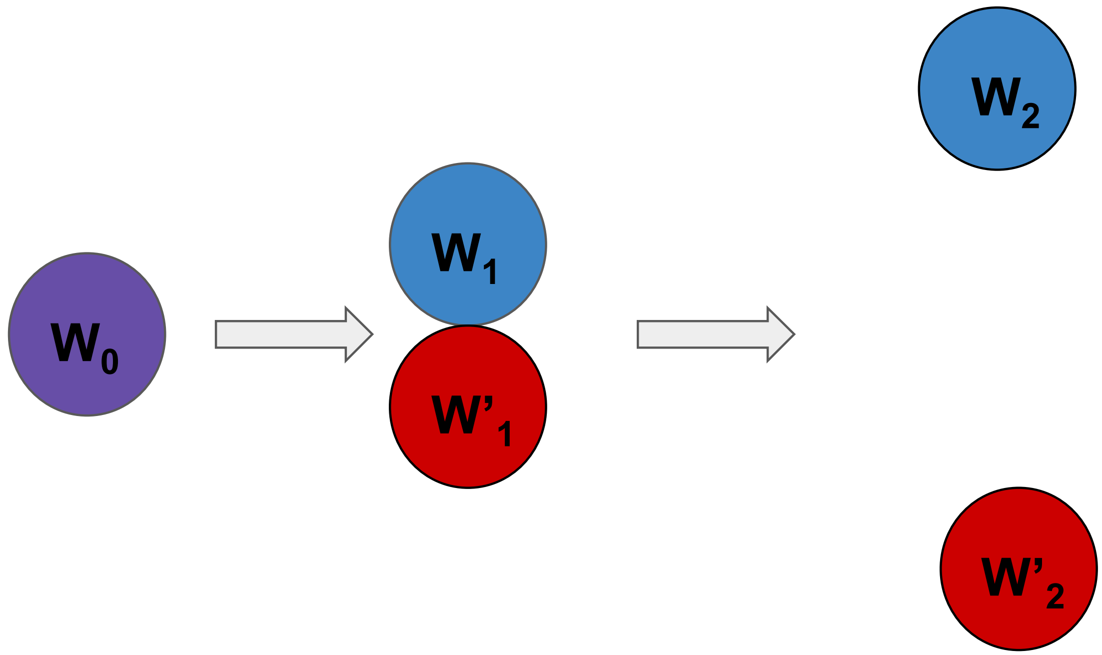
The per-step guarantee depends on a given model

Classical Analysis: models reached during training are always worst-case for the datapoint

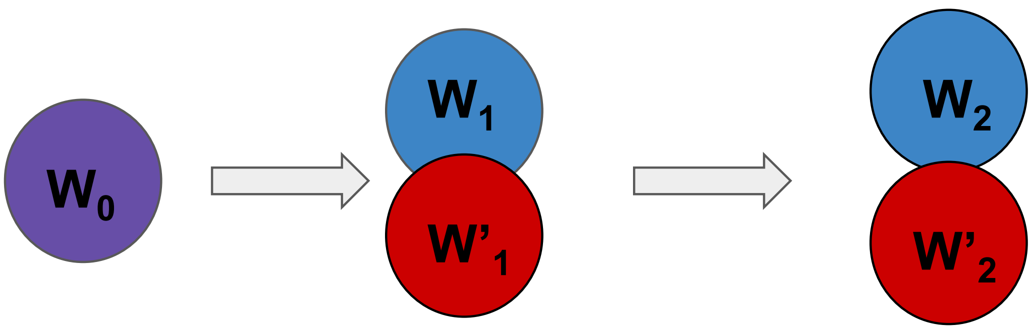
Observation: But what if most models reached during training have better guarantees?

Composition with "Expectations": We can bound composition by only considering "expected" guarantees at each step.

Worst Case View:



Expected View:



Free Variable p > 1

Expected guarantee at (n-i) step

$$D_\alpha(X||Y)$$

$$\leq \frac{1}{\alpha-1}\left(\sum_{i=0}^{n-2}\frac{(p-1)^i}{p^{i+1}}\ln(\mathbb{E}_{X_1,\cdots X_{n-(i+1)}}((e^{(g_p^i(\alpha)-1)D_{g_p^i(\alpha)}(X_{n-i}||Y_{n-i})})^p))\right)$$

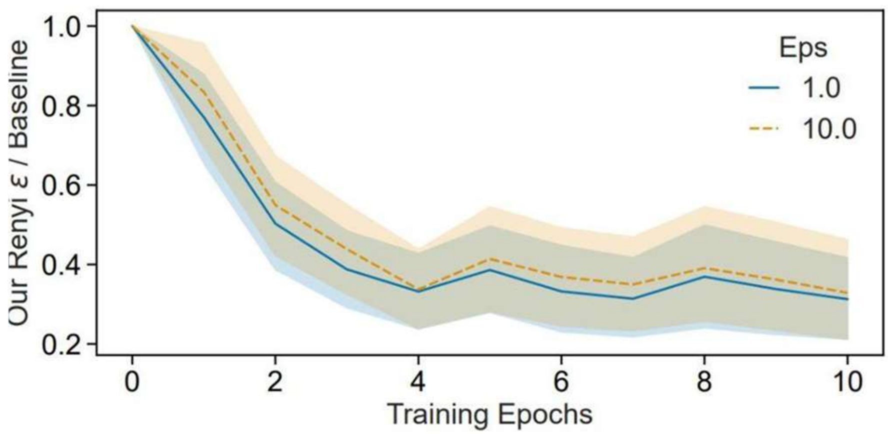
$$+\frac{1}{\alpha-1}\left(\frac{(p-1)^{n-1}}{p^n}\right)\ln((e^{(g_p^{n-1}(\alpha)-1)D_{g_p^{n-1}(\alpha)}(X_1||Y_1)})^p)$$

N steps with D and D'

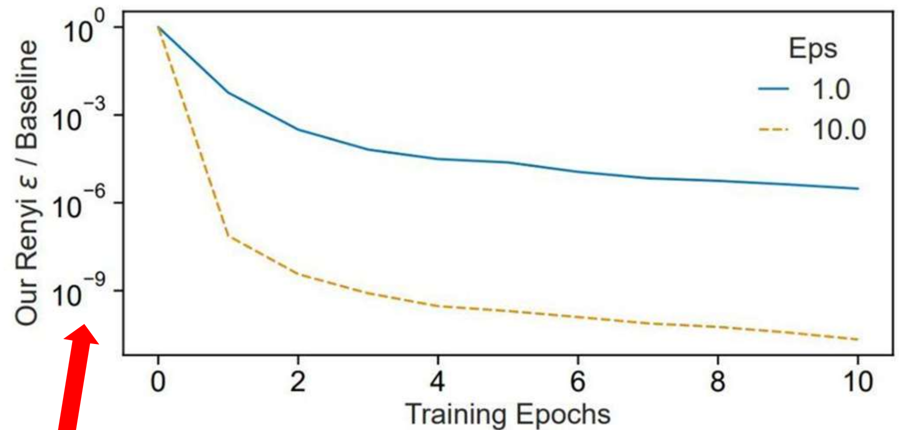
where $g_p(\alpha)=\frac{p}{p-1}\alpha-\frac{1}{p}$ *and* g_p^i *is* g_p *composed* i *times, where we defined* $g_p^0(\alpha)=\alpha$

Initial steps are weighted higher

Better privacy for many datapoints than worst-case



Average Trend for 500 points

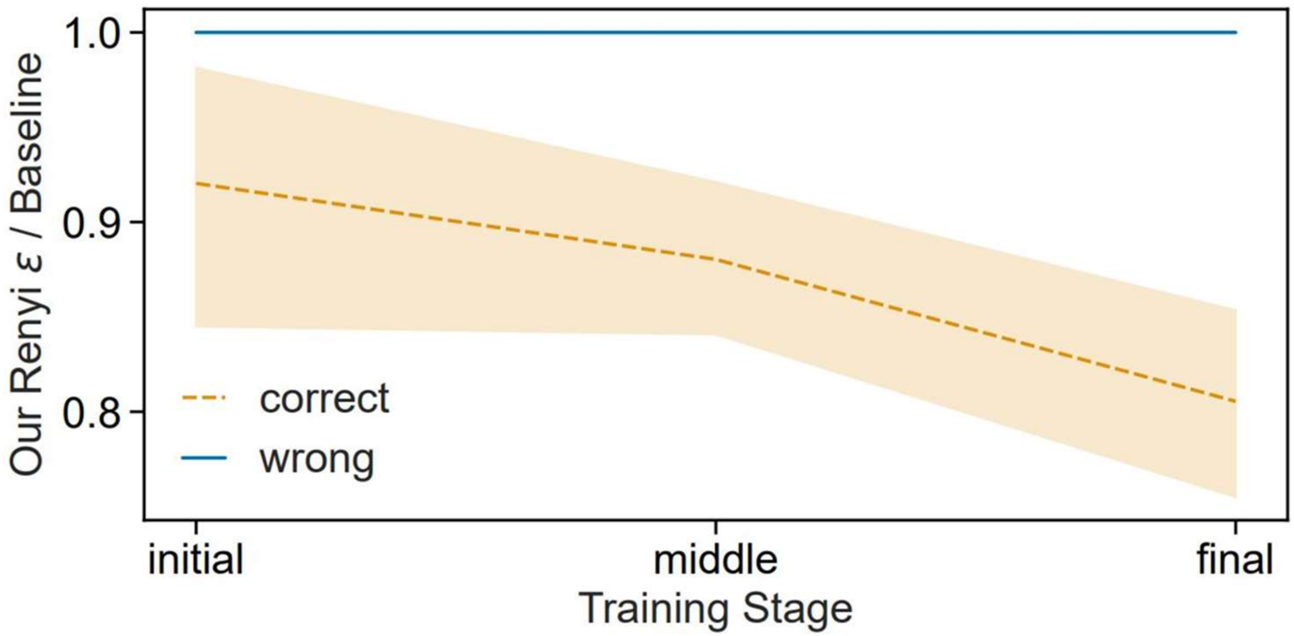


Log Scale

10th Percentile

Results for MNIST

Correct Points Benefit More



Results for CIFAR10

Takeaways

1. Many datapoints are harder to attack than the worst case
   - Datasets can mask updates from datapoint
   - Datasets can lead to favourable models for the datapoint

1. Analogously: many datapoints are easier to unlearn

1. Open Problem: How tight is this per-instance analysis?

1. Open Problem: How to check data-dependent privacy efficiently?
   - Current approach is expensive, useful for existence

Warning: On Data Dependency

Data dependent guarantees have known security issues

- E.g., releasing data-dependent guarantee leaks privacy

But useful quantity in the study of Trustworthy ML

Future Work: to better understand the utility of per-instance DP in Trustworthy ML

Thank You!

Contact: anvith.thudi@mail.utoronto.ca , nickhengrui.jia@mail.utoronto.ca

Paper

Code