Privacy Auditing

Tudor Cebere

July 6, 2023

Presentation

- Started my PhD in November 2022 under the supervision of Aurélien Bellet
- Topic of my thesis: Privacy Preserving Machine Learning
- Generally interested in Differential Privacy

The right to be left alone

Privacy in Machine Learning

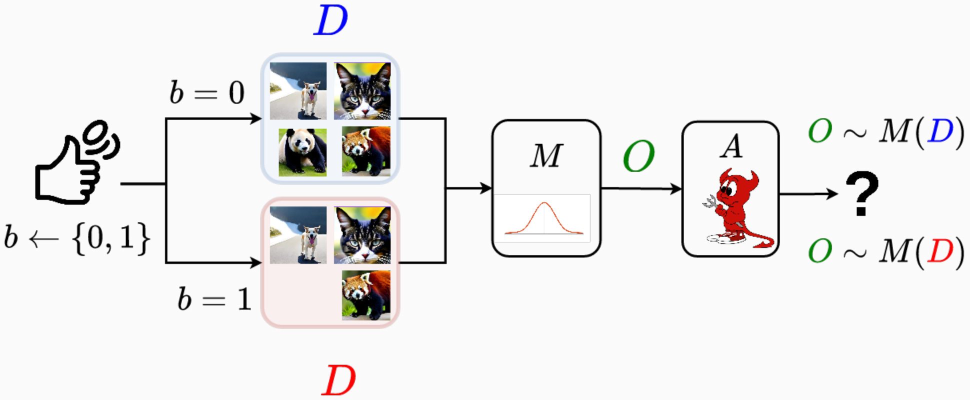
Evaluating and Testing Unintended Memorization in Neural Networks - N. Carlini

1. Introduction

2. Privacy Auditing

3. Auditing DP-SGD

Adversarial Game of Privacy

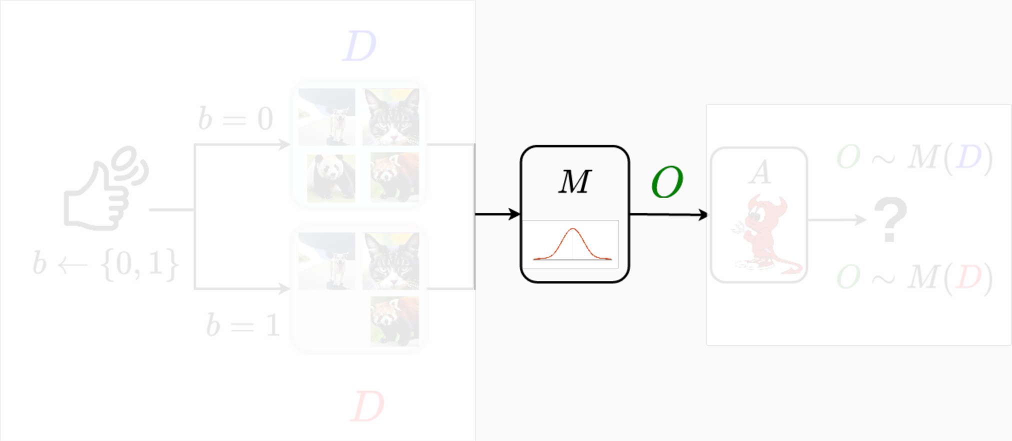


(ϵ, δ) Differential Privacy (DP)

A mechanism $M : \mathcal{X}^* \rightarrow \mathcal{Y}$ is (ϵ, δ)-DP if for all neighboring datasets D and D' the following inequality holds for all $S \in \mathcal{Y}$:

$$P[M(D) \in S] \leq e^{\epsilon} P[M(D') \in S] + \delta \tag{1}$$

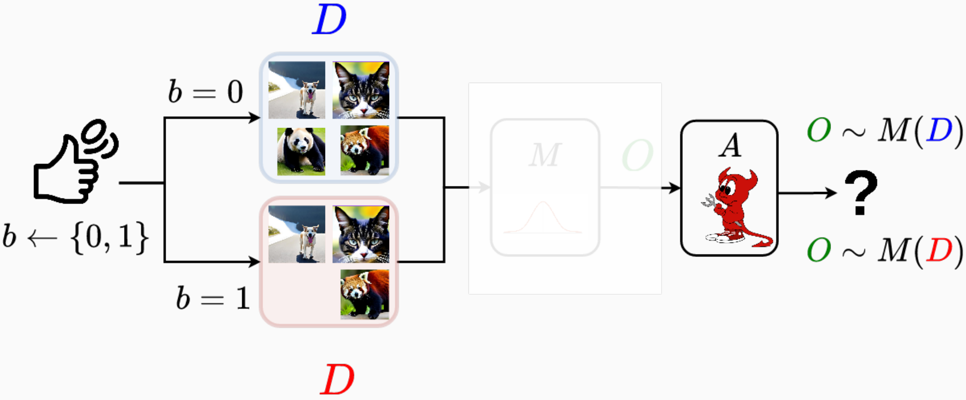
Differential Privacy



Auditing Differential Privacy

Given a blackbox mechanism $M : \mathcal{X}^* \to \mathcal{Y}$, what are the privacy guarantees $(\overline{\epsilon}, \delta)$ of M on a fixed dataset D and an adversary A?

Privacy Auditing



Applications: Correctness

Backpropagation Clipping for Deep Learning with Differential Privacy

Timothy Stevens, Ivoline C. Ngong, David Darais, Calvin Hirsch, David Slater, Joseph P. Near

We present backpropagation clipping, a novel variant of differentially private stochastic gradient descent (DP-SGD) for privacy-preserving deep learning. Our approach clips each trainable layer's inputs (during the forward pass) and its upstream gradients (during the backward pass) to ensure bounded global sensitivity for the layer's gradient; this combination replaces the gradient clipping step in existing DP-SGD variants. Our approach is simple to implement in existing deep learning frameworks. The results of our empirical evaluation demonstrate that backpropagation clipping provides higher accuracy at lower values for the privacy parameter ϵ compared to previous work. We achieve 98.7% accuracy for MNIST with $\epsilon = 0.07$ and 74% accuracy for CIFAR-10 with $\epsilon = 3.64$.

Comments: **We found a bug in our implementation code that invalidates our experimental results**

Debugging Differential Privacy: A Case Study for Privacy Auditing - F. Tramer et al.

Applications: Correctness

Privacy Leakage at low sample size #571

Open tudorcebere opened this issue on Mar 3 · 6 comments

tudorcebere commented on Mar 3 • edited

Bug

When using opacus at low sample sizes (~2-3 samples), I managed to leak more privacy than the accounting described:

Link: https://colab.research.google.com/drive/1gZVrg9kPIWjibApBkEnKNQqaIn8kUySs?usp=sharing

The privacy estimation is made as in:
https://proceedings.neurips.cc/paper/2020/file/fc4ddc15f9f4b4b06ef7844d6bb53abf-Paper.pdf

Applications: Tightness

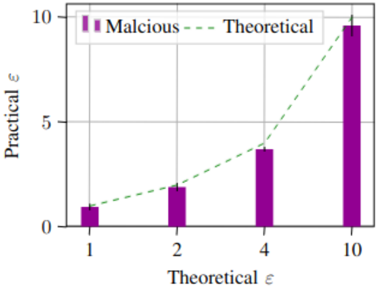


Fig. 8: **Malicious dataset attack**: the adversary creates a custom dataset to reduce the effect of other samples on the inserted watermark. This verifies the DP-SGD privacy is tight.

Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning - M. Nasr et al.

1. Introduction

2. Privacy Auditing

3. Auditing DP-SGD

(ϵ, δ) Differential Privacy in Hypothesis Testing

Differential Privacy as Hypothesis Testing

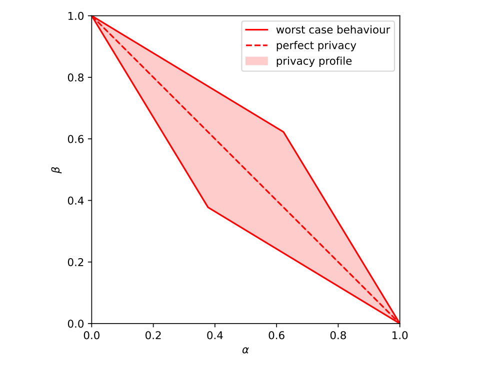
Given a random output O of a (ϵ, δ)-DP mechanism M, consider the following hypothesis testing experiment:

$$\begin{aligned} &\text{H0:} O \text{ was computed on } D \\ &\text{H1:} O \text{ was computed on } D' \end{aligned} \tag{2}$$

Any rejection rule and it's expectation of Type I (α) and II (β) errors, satisfies:

$$\begin{aligned} \alpha + e^{\epsilon}\beta \geq 1 - \delta \\ \beta + e^{\epsilon}\alpha \geq 1 - \delta \end{aligned} \tag{3}$$

Privacy Profiles

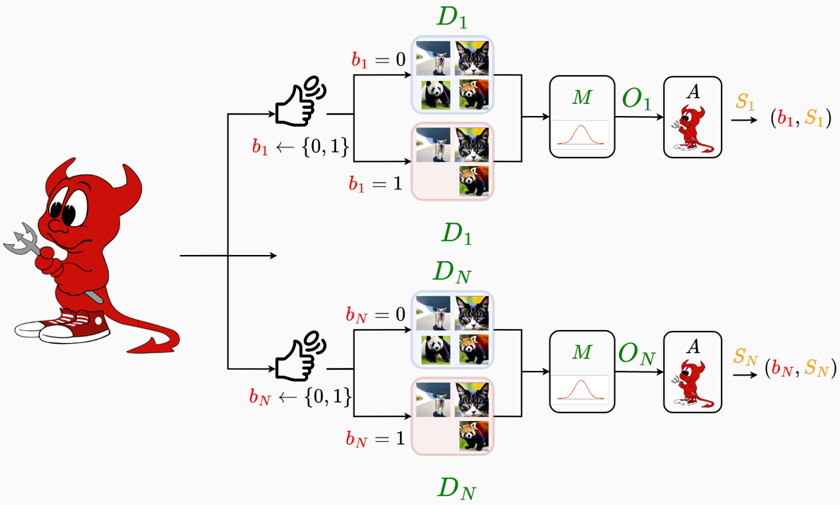


The Composition Theorem of Differential Privacy - P. Kairouz et al.

Auditing Pipeline overview

- Adversary & Sample Gathering
- Bounding
- Conversion to Differential Privacy

Auditing Pipeline: Gathering Samples



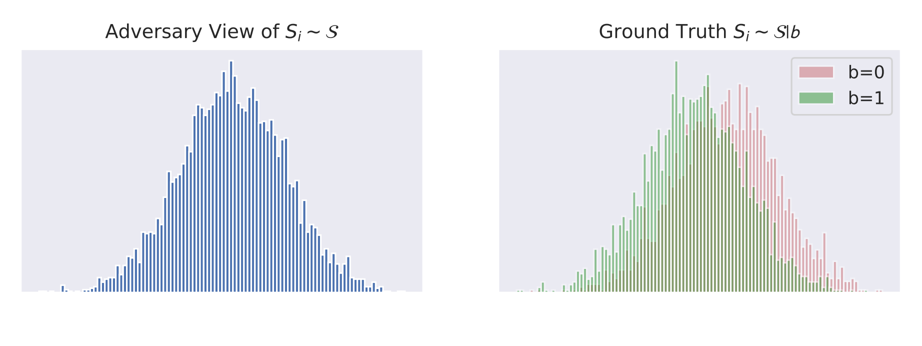
Auditing Pipeline: Gathering Samples Advances

- Multisample Testing: Zanella-Béguelin and K. Pillutla et al.
- Auditing via Generalization Bounds: Steinke et al.

Bayesian Estimation of Differential Privacy - S. Zanella-Béguelin et al.
Unleashing the Power of Randomization in Auditing Differentially Private ML - K. Pillutla et al.

Privacy Auditing with One (1) Training Run - T. Steinke et al.

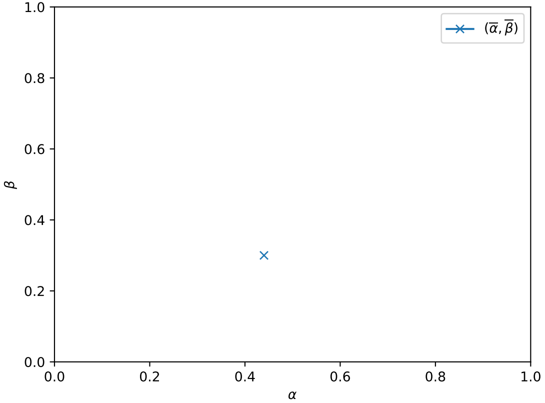
Auditing Pipeline: Selecting a Rejection Rule



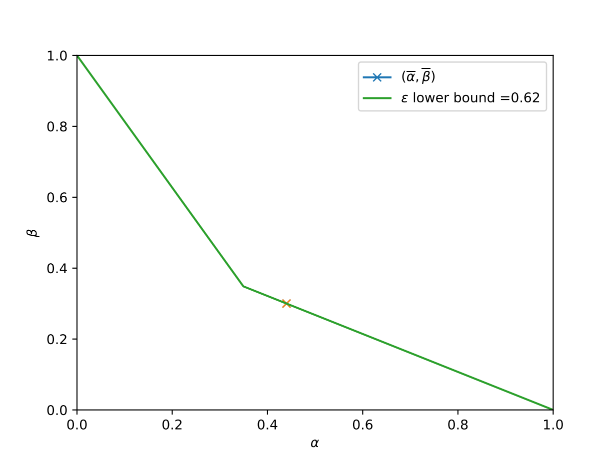
Auditing Pipeline: Bounding

$$\longrightarrow (TN, TP, FN, FP) \leftarrow \phi(S, b) \longrightarrow \begin{cases} \underline{\alpha}, \overline{\alpha} \leftarrow CI(FN, FN + TP) \\ \underline{\beta}, \overline{\beta} \leftarrow CI(FP, FP + TN) \end{cases}$$

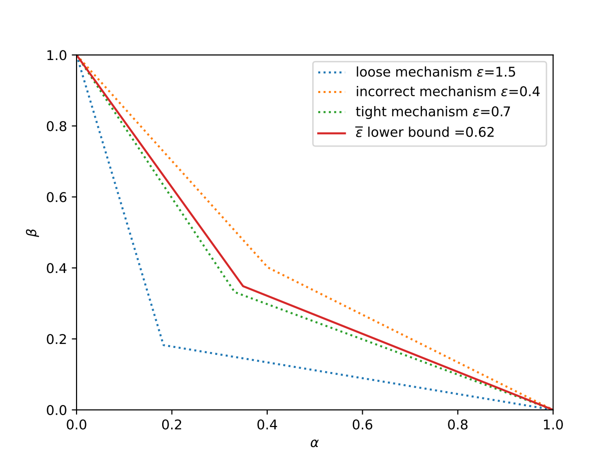
Auditing Pipeline: Conversion



Auditing Pipeline: Conversion



Auditing Pipeline: Conversion



Takeaways

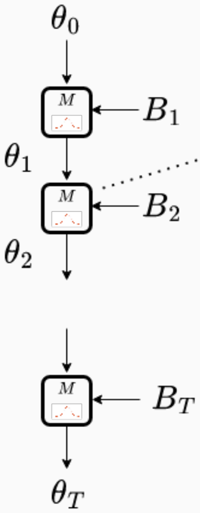
- Tight auditing is sample expensive.
- Generating samples, depending on the underlying mechanism and the threat model, can be very expensive.
- A. Gilbert shows that there is No Free Lunch Theorem in Auditing

1. Introduction

2. Privacy Auditing

3. Auditing DP-SGD

DP-SGD



$$\theta_2 = \theta_1 - \xi(\nabla(loss(\theta_1, B_1)) + \mathcal{N}(0, \sigma))$$

Tightness of DP-SGD

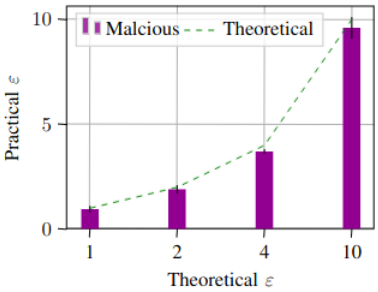
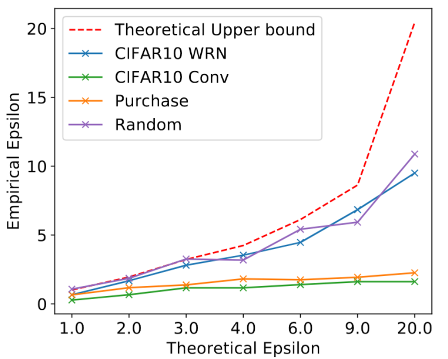


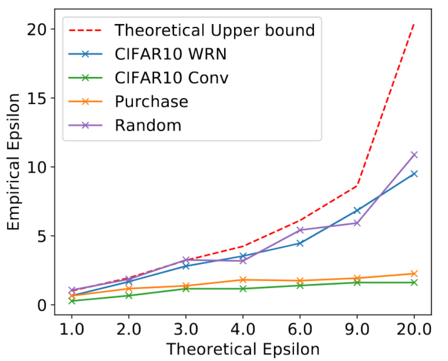
Fig. 8: **Malicious dataset attack**: the adversary creates a custom dataset to reduce the effect of other samples on the inserted watermark. This verifies the DP-SGD privacy is tight.

Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning - M. Nasr et al.

Tightness of DP-SGD



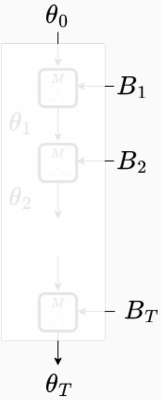
Tight Auditing of Differentially Private Machine Learning - M. Nasr et al.



Tight Auditing of Differentially Private Machine Learning - M. Nasr et al.

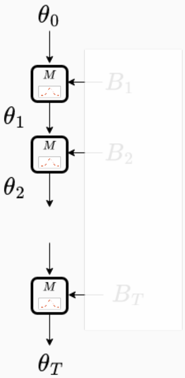
End of story for differentially private machine learning?

DP-SGD + Amplification by Iteration



Privacy Amplification by Iteration - V. Feldman et al.

DP-SGD + Amplification by Subsampling



What Can We Learn Privately? - S. Kasiviswanathan et al.

Privacy Amplification by Subsampling: Tight Analyses via Couplings and Divergences - B. Balle et al.

Takeaways

- How to do private learning is still unclear (what other privacy amplifications are there?).
- The threat model guarantees and assumptions are still unclear.
- Plenty of research yet to be done

Questions