PAPER REVIEW

ONE-SHOT EMPIRICAL PRIVACY ESTIMATION FOR FEDERATED LEARNING

GALEN ANDREW, PETER KAIROUZ, SEWOONG OH,
ALINA OPREA, H. BRENDAN MCMAHAN, VINITH SURIYAKUMAR

Speaker: Jonggyu Jang

2023.12.21 @ POSTECH

Discussion

1. 세미나 주기 변경 + 하루에 발표하는 인원 변경
   - 옵션 1: 주 2회, 하루에 1명 발표
   - 옵션 2: 주 1회, 하루에 2명 발표

3. 발표 목적 변경 (아이디어 → **논문 리뷰 위주**)

4. 발표 소요 시간 변경 (??? → **30분 + a**)

5. 발표 논문 선정 방식 리뷰할 논문의 **리스트**를 만들고 논문 별 **발표자 할당**

6. 모두 그 논문을 읽어오는걸 권장, 강요 X **(안 읽어오면 시간낭비 가능성 높음)**

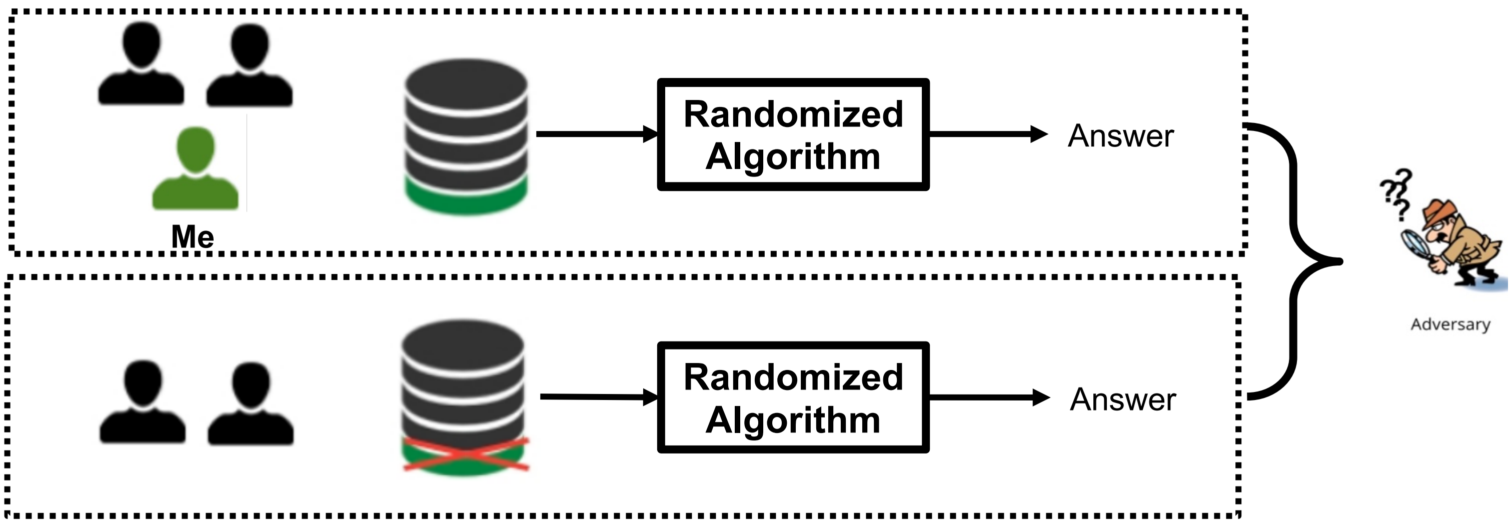
One-shot Empirical Privacy Estimation for Federated Learning

Galen Andrew*§ Peter Kairouz* Sewoong Oh*
Alina Oprea*† H. Brendan McMahan* Vinith M. Suriyakumar‡

Abstract

Privacy estimation techniques for differentially private (DP) algorithms are useful for comparing against analytical bounds, or to empirically measure privacy loss in settings where known analytical bounds are not tight. However, existing privacy auditing techniques usually make strong assumptions on the adversary (e.g., knowledge of intermediate model iterates or the training data distribution), are tailored to specific tasks, model architectures, or DP algorithm, and/or require retraining the model many times (typically on the order of thousands). These shortcomings make deploying such techniques at scale difficult in practice, especially in federated settings where model training can take days or weeks. In this work, we present a novel "one-shot" approach that can systematically address these challenges, allowing efficient auditing or estimation of the privacy loss of a model during the same, single training run used to fit model parameters, and without requiring any *a priori* knowledge about the model architecture, task, or DP training algorithm. We show that our method provides provably correct estimates for the privacy loss under the Gaussian mechanism, and we demonstrate its performance on well-established FL benchmark datasets under several adversarial threat models.

Differential Privacy



Definition: Differential Privacy (DP)

Let us assume X and X' are neighboring datasets. We say randomized mechanism M is ϵ-DP if, for all X' and $\mathrm{R} \subset \mathcal{R}$, we have

$$\Pr\left[M(X) \in R\right] \leq e^{\epsilon} \Pr\left[M(X') \in R\right]$$
$$\Pr\left[M(X') \in R\right] \leq e^{\epsilon} \Pr\left[M(X) \in R\right]$$

Approximated DP and Gaussian Mechanism

Definition: Approximated DP

Let us assume X and X' are neighboring datasets. We say randomized mechanism M is (ϵ, δ)-DP if, for all X' and $\mathrm{R} \subset \mathcal{R}$, we have

$$\Pr\left[M(X) \in R\right] \leq e^{\epsilon} \Pr\left[M(X') \in R\right] \boxed{+\, \delta}$$
$$\Pr\left[M(X') \in R\right] \leq e^{\epsilon} \Pr\left[M(X) \in R\right] \boxed{+\, \delta}$$

$$\Pr_{r \sim M(X)}\left[\left|\log \frac{\Pr\left[M(X) \in r\right]}{\Pr\left[M(X') \in r\right]}\right| > \epsilon\right] < 1 - \delta$$

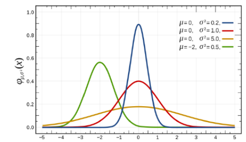
Definition: Gaussian Mechanism

Let $f\ : \mathcal{X}^n \to R^k$. The Gaussian mechanism is defined as

$$M(X) = f(X) + (Y_1, \dots, Y_k),$$

where Y_i are independent $\boldsymbol{N}\left(\mathbf{0}, \mathbf{2}\,\mathbf{ln}\left(\frac{\mathbf{1.25}}{\boldsymbol{\delta}}\right)\frac{\boldsymbol{\Delta}_2^2}{\boldsymbol{\epsilon}^2}\right)$ and Δ_2^2 denotes ℓ_2-sensitivity.

The Gaussian mechanism is (ϵ, δ)-DP.



What is Empirical Privacy Estimation ?

Analytic Way

- Privacy Accountant
- Composition Method
- **Not a tight bound**

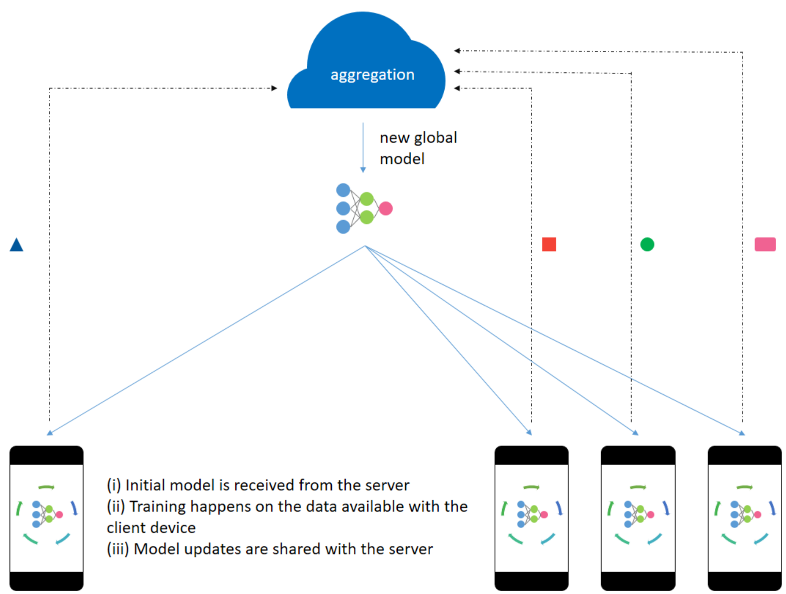
Empirical Way

- Privacy Auditing
- Canaries
- **A tight bound**

		auditor controls	auditor receives
Central	Jagielski et al. [2020]	train data	final model
	Zanella-Beguelin et al. [2023]	train data	final model
	Pillutla et al. [2023]	train data	final model
	Steinke et al. [2023]	train data	intermediate/final model
	Jagielski et al. [2023]	train data	intermediate models
	Nasr et al. [2023]	train data, privacy noise, minibatch	intermediate/final models
FL	Algorithm 2 (Ours)	client model update	final model
	Algorithm 3 (Ours)	client model update	intermediate models
	CANIFE [Maddock et al., 2022]	client sample, privacy noise, minibatch	intermediate models

Federated Learning

What is Federated Learning?



Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:
  initialize w_0
  for each round $t = 1, 2, \dots$ **do**
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of m clients)
    for each client $k \in S_t$ **in parallel do**
      $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
    $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$

ClientUpdate(k, w): // *Run on client k*
  $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
  for each local epoch i from 1 to E **do**
    for batch $b \in \mathcal{B}$ **do**
      $w \leftarrow w - \eta \nabla \ell(w; b)$
  return w to server

Challenges in Federated Learning:

1. Analytic way is possible, but loose bound
2. Existing work: Modify training data is impossible
3. Existing work: Intermediate gradient is required (further privacy leakage)

Method

Canary

You’ve probably heard the phrase “the canary in the coal mine” and know it refers to advanced warning of a danger. In the centuries before air quality instruments, miners carried canaries in cages into the mines to detect carbon monoxide and methane before they reached dangerous levels for humans.

Method

1. What we want to do.

For $\delta = 1e-5$, corresponding DP is $\epsilon = 0.3$!

Additive Noise

$z_i \sim N(0, \sigma^2)$

Data (Gradient)

$x_1, x_2, \ldots, x_n$

Perturbed Data (Gradient)

$x'_1, x'_2, \ldots, x'_n$

Method

2. Canary Approach

Canaries = Virtual Client

Data (Gradient)
$x_1, x_2, \ldots, x_n$

+

c_1

c_2

…

c_k

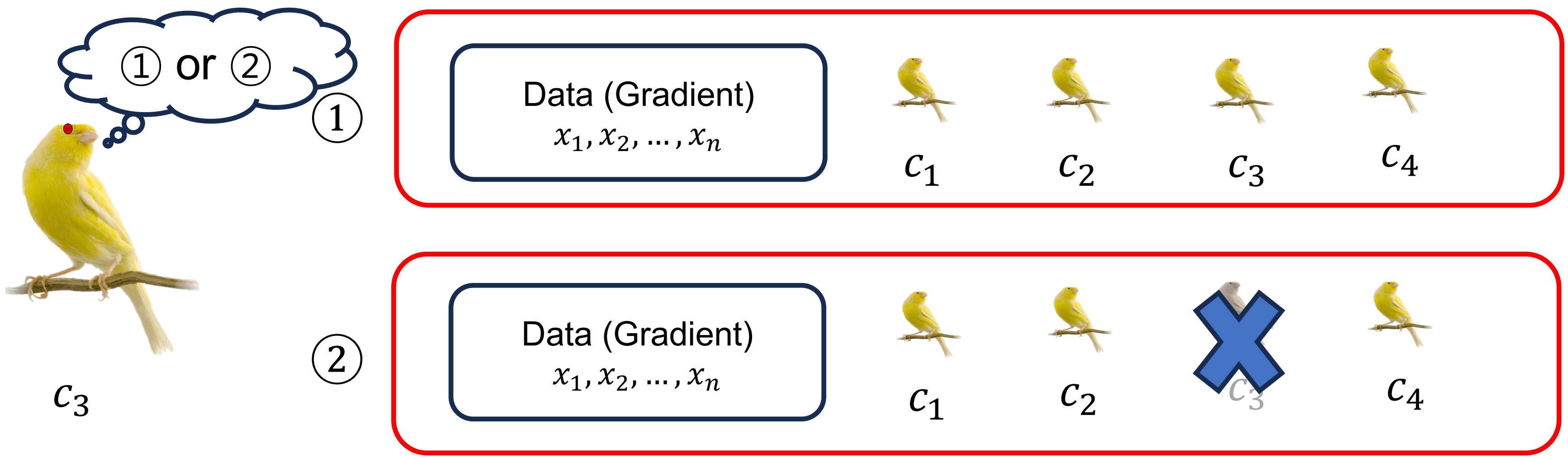
Sum-query ground-truth: $\rho \leftarrow \sum_i x_i$

Canary addition: $\rho \leftarrow \rho + \sum_j c_j$

Additive noise: $\rho \leftarrow \rho + N(0, \sigma^2)$

Method

3. Canary Detection



Confusing : Good, privacy is protected
Trivial : Bad, privacy is leaked

Method

4. How to detect?

Cosine similarity…

c_i

Obtained gradient @ server

$$\rho = \sum_i x_i + \sum_j c_j + N(0, \sigma^2)$$

$$\rho' = \sum_i x_i + \sum_{j \neq k} c_j + N(0, \sigma^2)$$

This data is used!
I confirm!

Assume c_i is randomly chosen from unit sphere space $\mathcal{S}^{d-1}$

If c_i is included : $g_k = \frac{\rho^T c_k}{||\rho||} \sim N(\hat{\mu}, \hat{\sigma})$ Empirical Estimation

If c_i is not included : $g_k = \frac{\rho'^T c_k}{||\rho||} \approx N\left(0, \frac{1}{d}\right)$ Analytic Estimation

Guess Why!

Method

Algorithm 1 One-shot privacy estimation for Gaussian mechanism.

1: **Input:** Vectors $x_1, \cdots, x_n$ with $\|x_i\| \leq 1$, DP noise variance σ^2, and target δ
2: $\rho \leftarrow \sum_{i \in [n]} x_i$
3: **for** $j \in [k]$ **do**
4: Draw random $c_j \in \mathbb{S}^{d-1}$ uniformly from unit sphere
5: $\rho \leftarrow \rho + c_j$
6: Release $\rho \leftarrow \rho + \mathcal{N}(0, \sigma^2 I)$
7: **for** $j \in [k]$ **do**
8: $g_j \leftarrow \langle c_j, \rho \rangle / ||\rho||$
9: $\hat{\mu}, \hat{\sigma} \leftarrow \textbf{mean}(\{g_j\}), \textbf{std}(\{g_j\})$
10: $\boxed{\hat{\varepsilon} \leftarrow \varepsilon(\mathcal{N}(0, 1/d) \,||\, \mathcal{N}(\hat{\mu}, \hat{\sigma}^2); \delta)}$

What..?

Find smallest ϵ satisfying the DP condition

Definition: Approximated DP

Let us assume X and X' are neighboring datasets. We say randomized mechanism M is (ϵ, δ)-DP if, for all X' and $R \subset \mathcal{R}$, we have

$$\Pr[M(X) \in R] \leq e^{\epsilon} \Pr[M(X') \in R] + \delta$$
$$\Pr[M(X') \in R] \leq e^{\epsilon} \Pr[M(X) \in R] + \delta$$

$$\Pr_{r \sim M(X)}\left[\left|\log \frac{\Pr[M(X) \in r]}{\Pr[M(X') \in r]}\right| > \epsilon\right] < 1 - \delta$$

$$\Pr[Z_1 > \varepsilon] - e^{\varepsilon} \Pr[-Z_2 > \varepsilon] \leq \delta \text{ and } \Pr[Z_2 > \varepsilon] - e^{\varepsilon} \Pr[-Z_1 > \varepsilon] \leq \delta.$$

$$\begin{aligned}\log \delta &\geq \log\left(\Pr[Z_1 > \varepsilon] - e^{\varepsilon} \Pr[-Z_2 > \varepsilon]\right) \\ &= \log \Pr[Z_1 > \varepsilon] + \log\left(1 - \exp(\varepsilon + \log \Pr[-Z_2 > \varepsilon] - \log \Pr[Z_1 > \varepsilon])\right).\end{aligned}$$

Results

Dataset: Stackoverflow Word prediction dataset

https://github.com/google-research/federated/blob/master/utils/datasets/stackoverflow_word_prediction.py

- 2048 rounds with 167 clients per round.
- Each of 341k clients participates in exactly one round. (single epoch)
- 1k canaries

		Baseline (global)	Proposed	
Noise	analytical ε	ε_{lo}-all	ε_{est}-all	[redacted]
0	∞	6.240	45800	[redacted]
0.0496	300	6.238	382	[redacted]
0.0986	100	5.05	89.4	[redacted]
0.2317	30	0.407	2.693	[redacted]

We note that across the range of noise multipliers, the participation of 1k canaries had no significant impact on model accuracy – at most causing a 0.1% relative decrease.