EDGE AI Academy

Summer School

7-8 July 2025, Pisa, Italy

Edge AI Academy Summer School

Privacy in Federated Learning

Chuan Xu, Université Côte d'Azur, France

UNIVERSITÉ CÔTE D'AZUR Inria COATI cnrs i3S sophia antipolis

7-8 July 2025 Pisa, Italy

Outline

- Introduction to Federated Learning
- Privacy vulnerabilities in federated learning
  - Privacy attacks (Data reconstruction attack)
  - Threat models
  - Attack methodologies
- Privacy preserving mechanisms
  - Differential privacy

Copyright © 2025

Introduction

❖Federated Learning

[Li et al, 2019]

- Enables collective learning **without** transferring data
- Joint global mode

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) = \sum_{c \in \mathcal{C}} p_c \mathcal{L}_c(\theta) = \sum_{c \in \mathcal{C}} p_c \left(\frac{1}{|\mathcal{D}_c|} \sum_{x \in \mathcal{D}_c} l(\theta, x) \right)$$

Copyright © 2025

Introduction

❖Federated Learning

[Li et al, 2019]

- Real world applications
  - ❑ Google: Gboard keyboard, Pixel Phones
  - ❑ Apple: Hey Siri
  - ❑ Healthcare: Owkin, Doc.ai

Copyright © 2025

Lifecycle of federated learning

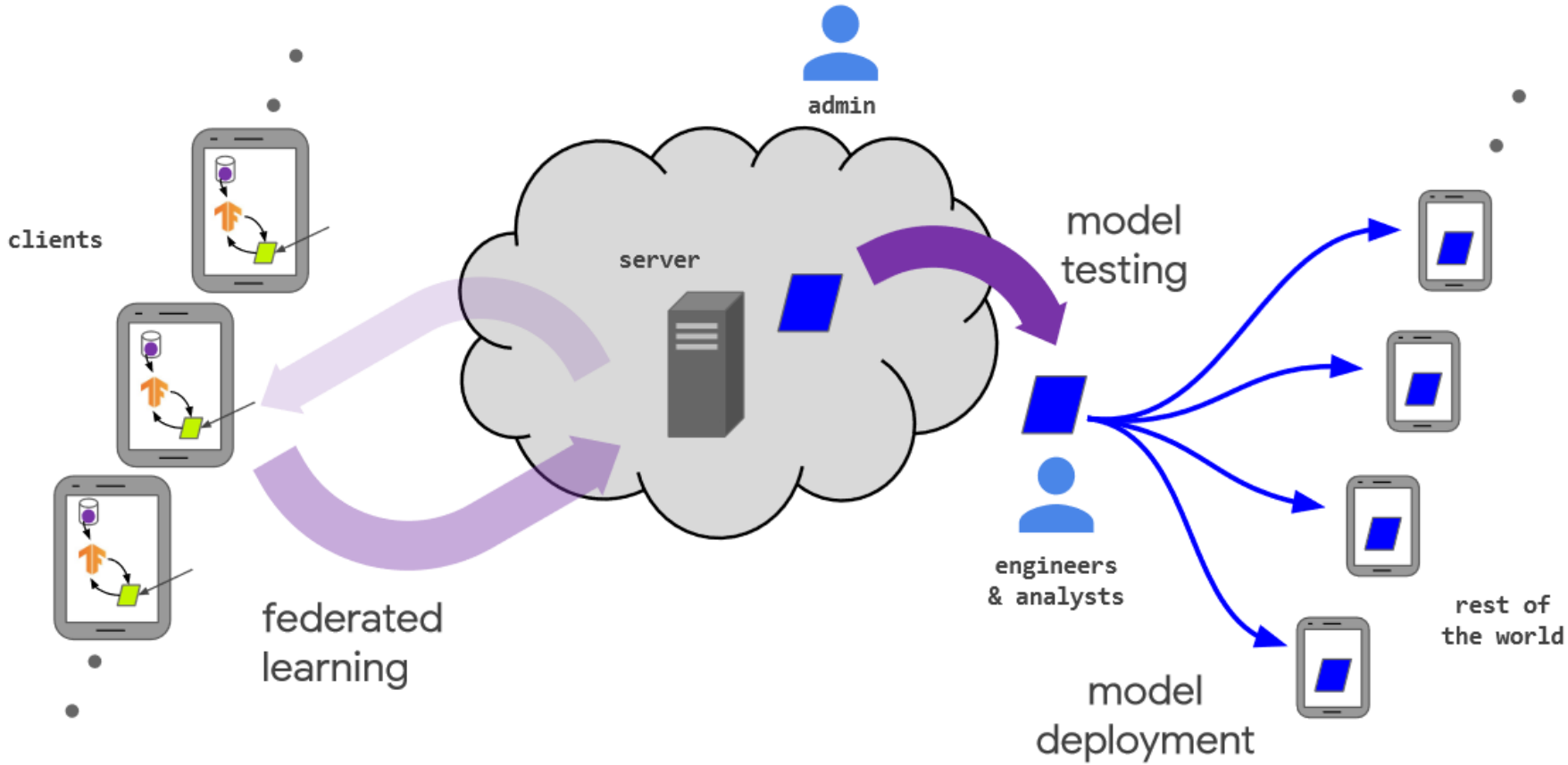
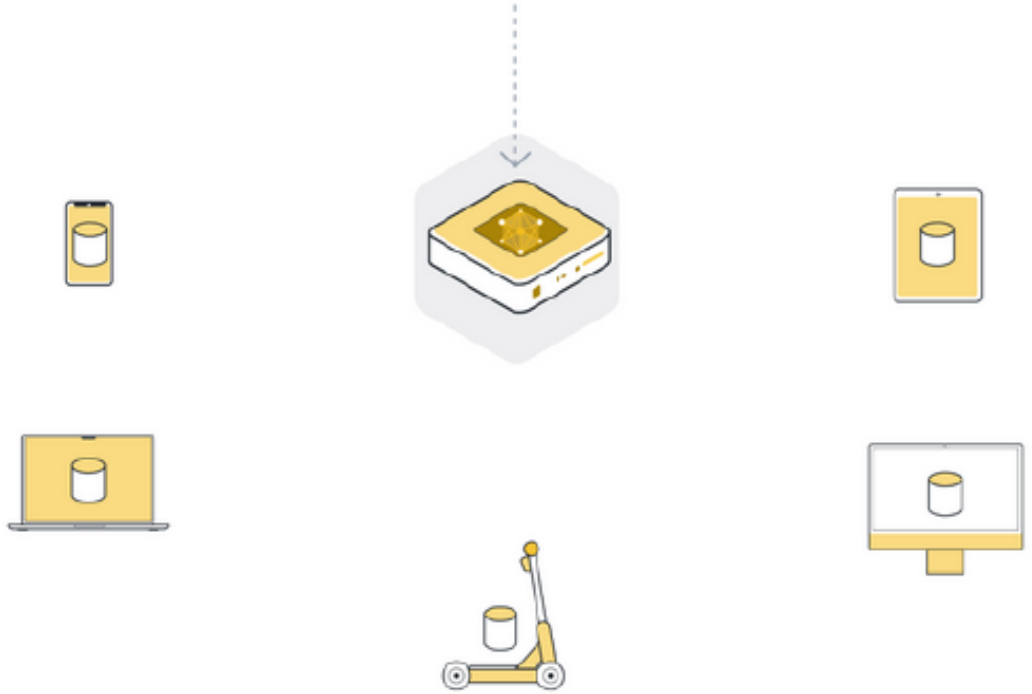


Figure from Survey « Advances and open problems in federated learning », Kairouz et al

Copyright © 2025

First learning algorithm for FL

- FedAvg

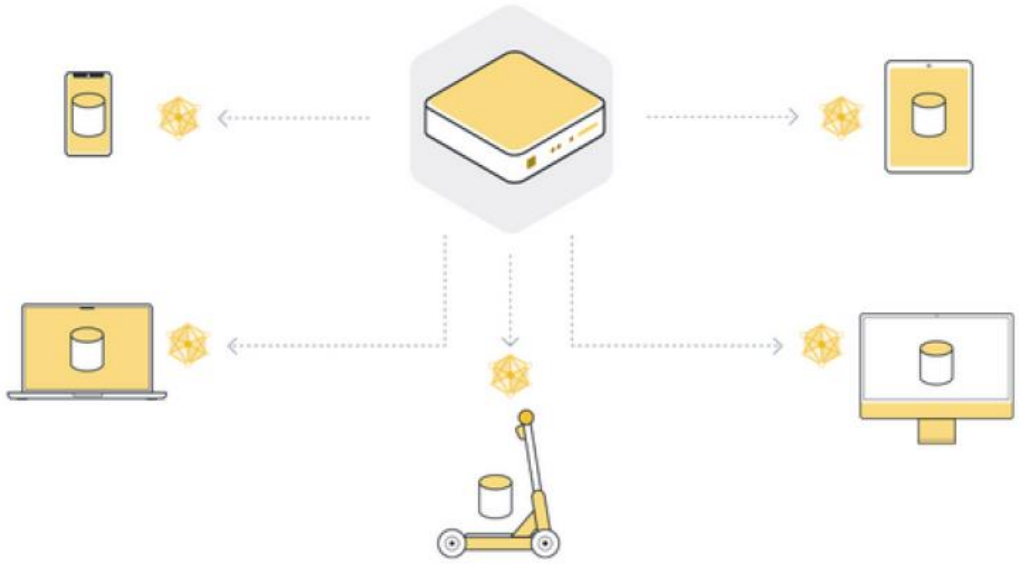


Communication-Efficient Learning of Deep Networks from Decentralized Data, McMahan et al, AISTATS 2017

Copyright © 2025

First learning algorithm for FL

- FedAvg



Communication-Efficient Learning of Deep Networks from Decentralized Data, McMahan et al, AISTATS 2017

Copyright © 2025

First learning algorithm for FL

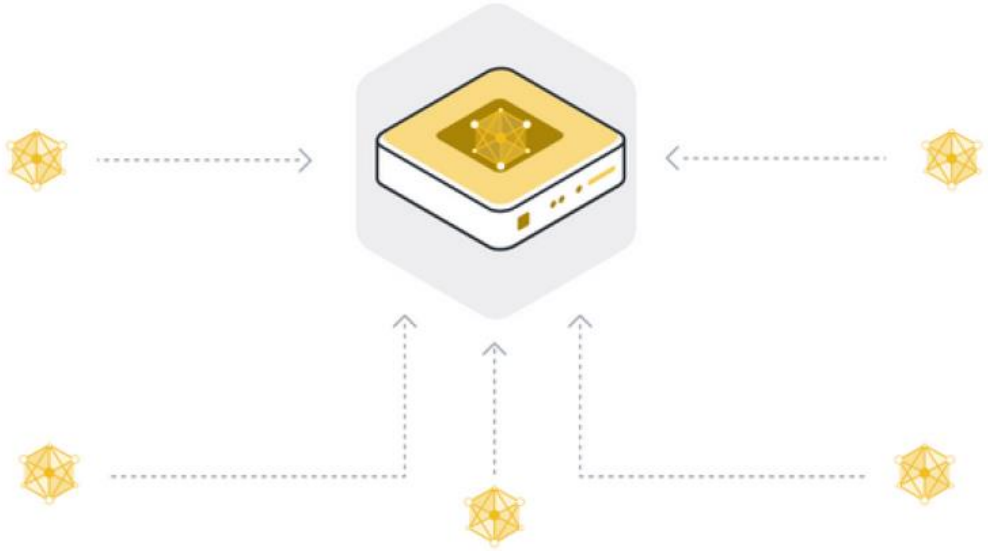
- FedAvg

Communication-Efficient Learning of Deep Networks from Decentralized Data, McMahan et al, AISTATS 2017

Copyright © 2025

First learning algorithm for FL

- FedAvg



Communication-Efficient Learning of Deep Networks from Decentralized Data, McMahan et al, AISTATS 2017

Copyright © 2025

Privacy Attacks

- Inversion attacks
  - Reconstruct the data points (data reconstruction attack)
  - Reconstruct sensitive attributes
- Membership inference attack
- Source inference attack
- Property inference attack

* Distinguished from poisioning/backdoor attacks

Copyright © 2025

Privacy Attacks

- Inversion attacks
  - **Reconstruct the data points (data reconstruction attack)**
  - Reconstruct sensitive attributes
- Membership inference attack
- Source inference attack
- Property inference attack

* Distinguished from poisioning/backdoor attacks

Copyright © 2025

Threat models

- Honest-but-curious
  - The rest of the world: access to the final deployed model
  - Client: root access to the client device
  - Engineers & Analysts: multiple outputs from the system
  - Administrator: root access to the server

- Malicious
  - Clients: send malicious updated models
  - Administrator: send malicious global models

Copyright © 2025

Threat models

- Honest-but-curious
  - **The rest of the world: access to the final deployed model**
  - Client: root access to the client device
  - Engineers & Analysts: multiple outputs from the system
  - **Administrator: root access to the server**

- Malicious
  - Clients: send malicious updated models
  - **Administrator: send malicious global models to clients**

Copyright © 2025

Data reconstruction attack – The rest of the world

- Access to the parameters of the final deployed model
- Methodology: Model-based

  - Invert the model to recover private information
  - Given a model θ:

    To reconstruct sample for each class k:

    [Fredrikson et al, 2015]

$$\hat{\mathbf{x}}(k) = \arg\max_{\mathbf{x}(k)} \mathbf{h}_\theta(\mathbf{x}(k))[k], \quad \forall k \in \{1, \ldots, K\}.$$

Copyright © 2025

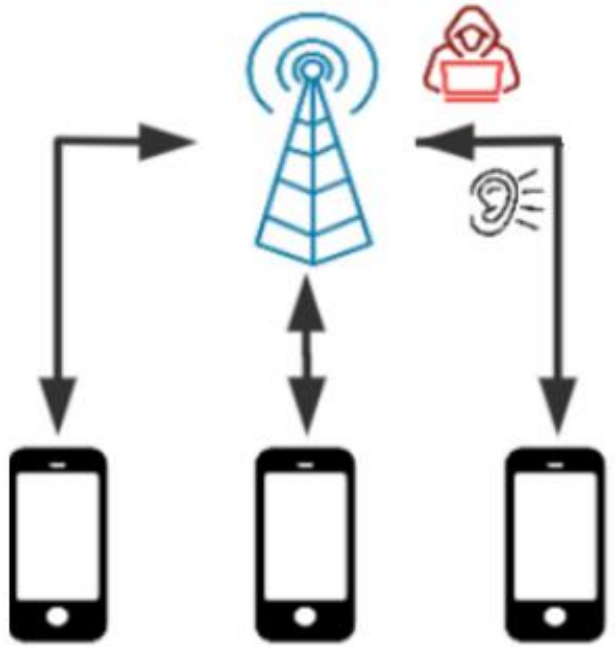
Data reconstruction attack – The rest of the world

- Access to the parameters of the final deployed model
- Limitations:
  - Works only when model overfits the data, classification problems
  - The number of recovered images are limited
  - Class representative datasets (such as MNIST, Facial datasets)

Copyright © 2025

Data reconstruction attack - Honest-but-curious administrator

- Knows the model structure, loss function and training algorithm
- Inspects raw exchanged messages (global model, local updated models)
- Steal the target client's dataset



Copyright © 2025

Data reconstruction attack - Honest-but-curious administrator

- Methodology

Gradient-based: Reconstruct the private information X whose yield virtual gradient is as close as to the observed "pseudogradient".

$$\hat{X} = \arg\max_{X} \textbf{Similarity}\left(\frac{\partial L(\theta(t), X)}{\partial \theta}, (\theta(t) - \theta_c(t))/\eta\right)$$

Copyright © 2025

Data reconstruction attack - Honest-but-curious administrator

- Performance

[Geiping et al, 2020][Yin et al, 2021]

Figure: 5 most recognizable images when attacking the target client with 100 images from 100 classes in CIFAR100, FedAvg with batch size 100, one local step, ResNet32 [Geiping et al, 2020]

Inverting gradients - how easy is it to break privacy in federated learning?, Geiping et al, NeurIPS 2020

See through gradients: Image batch recovery via gradinversion, Yin et al, CVPR 2021

Copyright © 2025

Data reconstruction attack - Honest-but-curious administrator

- Limitations - large batch size

Figure: Gradient-based inversion attack: Images recovered when training a logistic regression model on AT&T facial dataset among 3 clients under batch size 32 (middle) and 256 (right) when training a logistic regression model for AT&T dataset among 3 clients. The data is split non-iid by labels.

Copyright © 2025

Data reconstruction attack - Honest-but-curious administrator

- Limitations – unknown private labels

CIFAR10. BATCH SIZE 16

(b) Reconstructions with and without private labels

Evaluating gradient inversion attacks and defenses in federated learning, Huang et al, NeurIPS 2020

Copyright © 2025

Data reconstruction attack - Honest-but-curious administrator

❖A series of works

- Hongxu Yin et al, See through gradients: Image batch recovery via gradinversion. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), **2021**
- Dimitar Iliev Dimitrov et al. Data leakage in federated averaging. Transactions on Machine Learning Research, **2022**
- Sanjay Kariyappa et al, Cocktail party attack: Breaking aggregation-based privacy in federated learning using independent component analysis, Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, **2023**.
- Dimitar Iliev Dimitrov et al. SPEAR: Exact gradient inversion of batches in federated learning. Thirtyeighth Annual Conference on Neural Information Processing Systems, **2024**

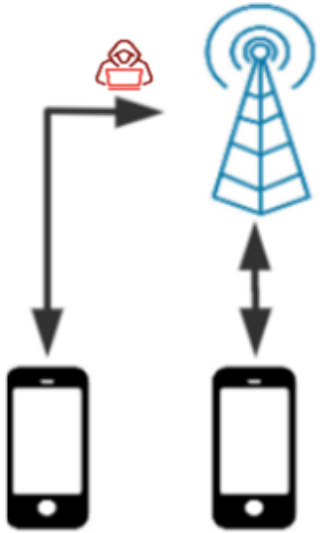
SOTA:

For a batch of >128 ImageNet images, the reconstructed images are very noisy!

Copyright © 2025

Data reconstruction attack – Malicious server

- The adversary can send arbitrary model parameters to the target clien
- Make the pseudo-gradient easily leaks data points



Two Methodologies:

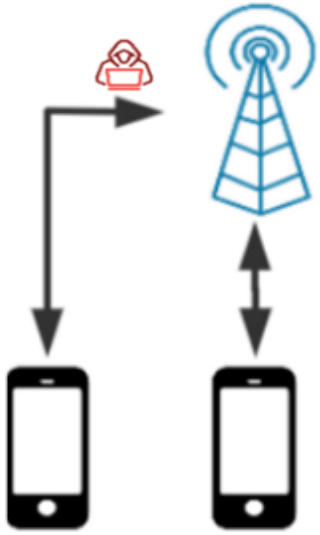
- *Gradient-based optimization*
  malicious parameters to amplify the gradient contribution of a single input

Wen, Y et al, Fishing for User Data in Large-Batch Federated Learning via Gradient Magnification. ICML 2022
Kostadin Garov et al, Hiding in plain sight: Disguising data stealing attacks in federated learning. ICLR 2024

Copyright © 2025

Data reconstruction attack – Malicious server

- The adversary can send arbitrary model parameters to the target clier
- Make the pseudo-gradient easily leaks data points



Two Methodologies:

- *Gradient-based optimization*

  SOTA: For a batch of 512 CIFAR10 images, 50% of the images are be reconstructed with high quality.

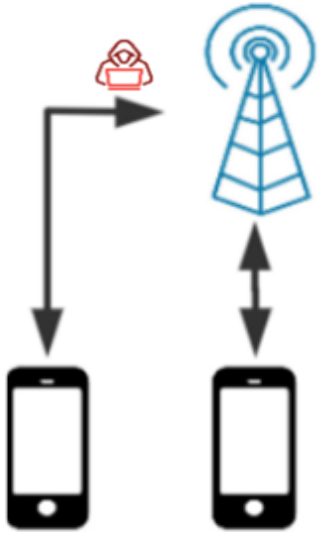
  For a batch of 256 ImageNet images, the reconstructed outputs exhibit reduced noise but generally lack sufficient quality

Wen, Y et al, Fishing for User Data in Large-Batch Federated Learning via Gradient Magnification. ICML 2022
Kostadin Garov et al, Hiding in plain sight: Disguising data stealing attacks in federated learning. ICLR 2024

Copyright © 2025

Data reconstruction attack – Malicious server

- The adversary can send arbitrary model parameters to the target clier
- Make the pseudo-gradient easily leaks data points



Two Methodologies:

- *Gradient-based optimization*

  Limitations:

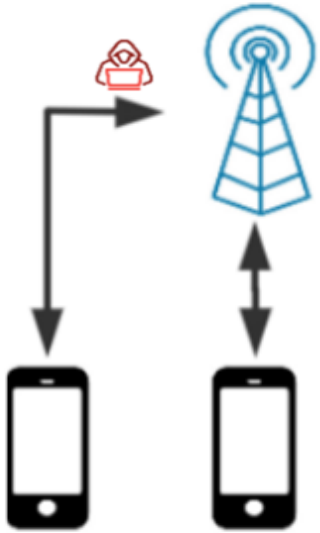
  - Prior knowledge on the dataset
  - Still struggle with large batch

Wen, Y et al, Fishing for User Data in Large-Batch Federated Learning via Gradient Magnification. ICML 2022
Kostadin Garov et al, Hiding in plain sight: Disguising data stealing attacks in federated learning. ICLR 2024

Copyright © 2025

Data reconstruction attack – Malicious server

- The adversary can send arbitrary model parameters to the target clier
- Make the pseudo-gradient easily leaks data points



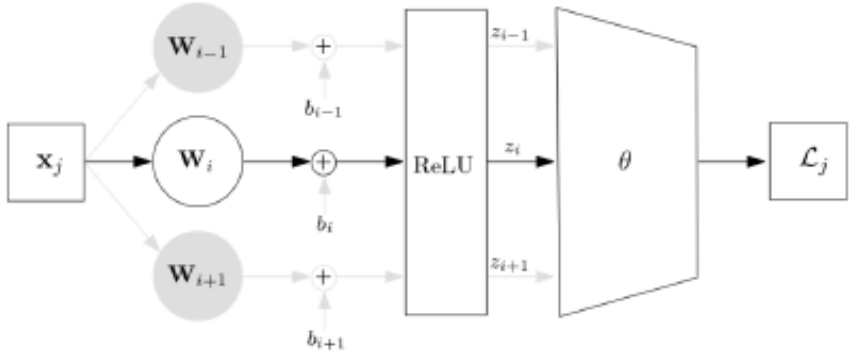
Two Methodologies:

- *Analytical attacks*
  - Observation: input to a biased linear layer (ReLU) can be **exactly reconstructed** from the gradient

Copyright © 2025

Data reconstruction attack – Malicious server

Input to a biased linear layer (ReLU) can be **exactly** reconstructed from the gradient [2]



$$\frac{\partial \mathcal{L}_j}{\partial \mathbf{W}_i} = \frac{\partial \mathcal{L}_j}{\partial z_i}\frac{\partial z_i}{\partial \mathbf{W}_i} = \frac{\partial \mathcal{L}_j}{\partial z_i}\mathbf{x}_j$$

$$\frac{\partial \mathcal{L}_j}{\partial b_i} = \frac{\partial \mathcal{L}_j}{\partial z_i}\frac{\partial z_i}{\partial b_i} = \frac{\partial \mathcal{L}_j}{\partial z_i}$$

If neuron i is activated by x_j

$$z_i = \mathbf{W_i}\mathbf{x}_j + b_i$$

$$\rightarrow \mathbf{x}_j = \frac{\partial \mathcal{L}_j}{\partial \mathbf{W}_i}\left(\frac{\partial \mathcal{L}_j}{\partial b_i}\right)^{-1}$$

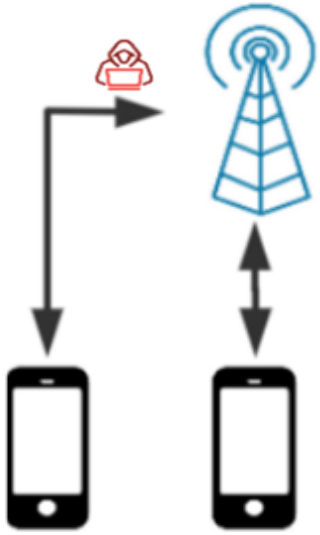
The server receives the batch gradient $\frac{\partial \mathcal{L}}{\partial \mathbf{W}_i}$ instead of $\frac{\partial \mathcal{L}_j}{\partial \mathbf{W}_i}$!

[2] Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). Inverting gradients - how easy is it to break privacy in federated learning? In Proceedings of the 34th International Conference on Neural Information Processing Systems

Copyright © 2025

Data reconstruction attack – Malicious server

- The adversary can send arbitrary model parameters to the target clien
- Make the pseudo-gradient easily leaks data points



Two Methodologies:

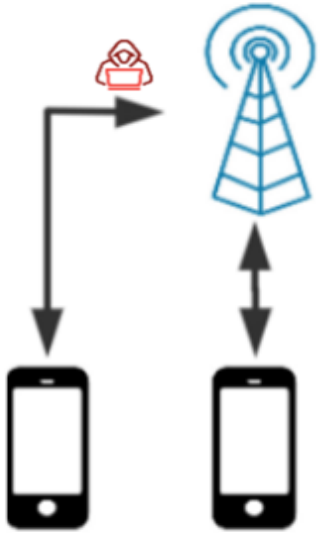
- *Analytical attacks*
  - Without any prior knowledge on the dataset
  - Perfect reconstruction for any batch size n given O(log n) roun
  - Succeed to recover exactly **all** in a batch of size **1024 ImageN** **images**, using **10** communication rounds

Cutting Through Privacy: A Hyperplane-Based Data Reconstruction Attack in Federated Learning
Francesco Diana, André Nusser, Chuan Xu, Giovanni Neglia, UAI, 2025

Copyright © 2025

Data reconstruction attack – Malicious server

- The adversary can send arbitrary model parameters to the target clien
- Make the pseudo-gradient easily leaks data points



Two Methodologies:

- *Analytical attacks*
  - Limitation:
    - Works only for classification tasks
    - Not generalized to any model structure

Cutting Through Privacy: A Hyperplane-Based Data Reconstruction Attack in Federated Learning
Francesco Diana, André Nusser, Chuan Xu, Giovanni Neglia, UAI, 2025

Copyright © 2025

Outline

- Introduction to Federated Learning

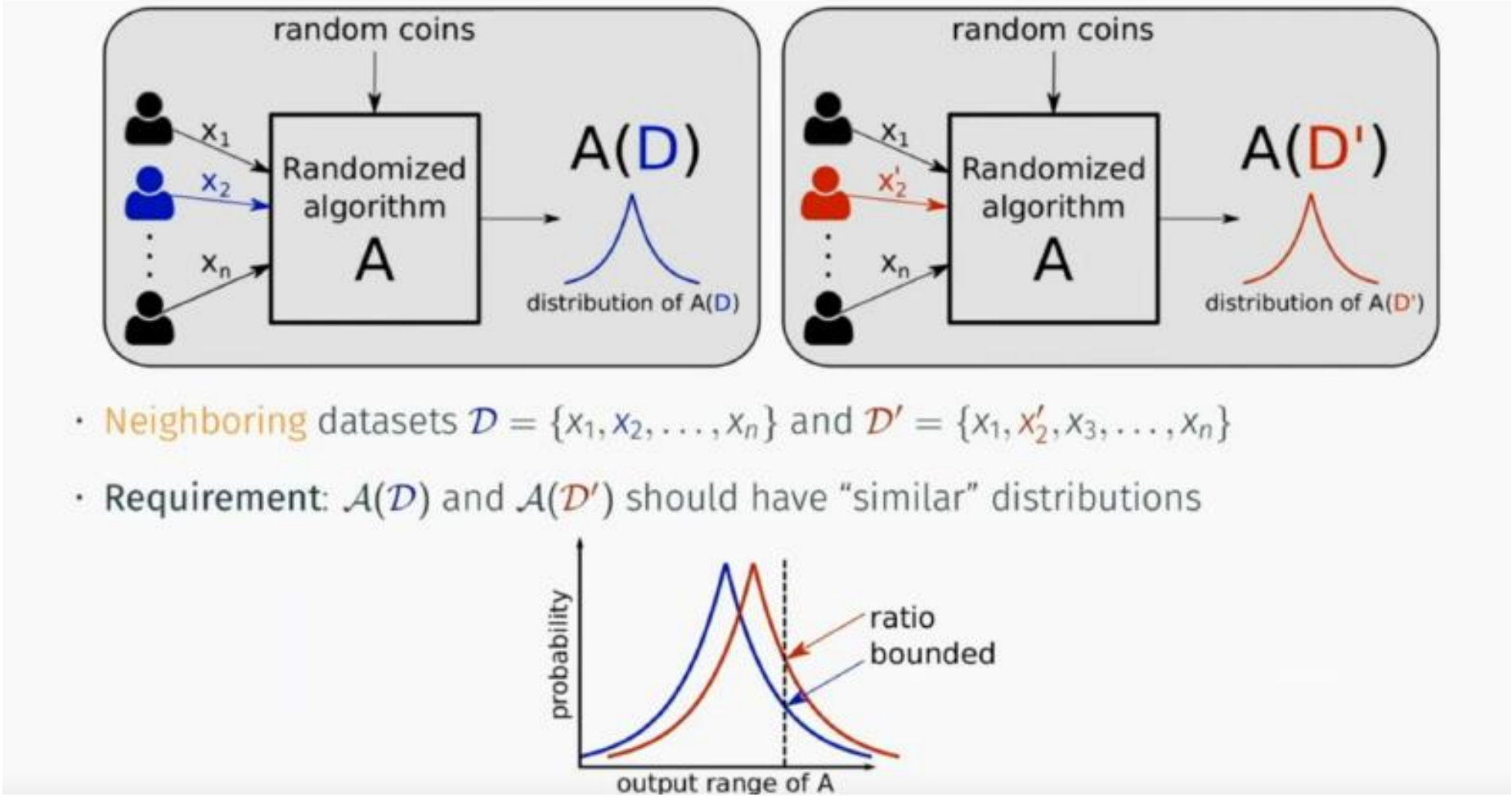
- Privacy vulnerabilities in federated learning
  - Privacy attacks (Data reconstruction attack)
  - Threat models
  - Attack methodologies

- **Privacy preserving mechanisms**
  - Differential privacy

Copyright © 2025

Privacy preserving mechanisms

- Differential privacy



Copyright © 2025

Differential privacy (DP)

- a mathematical framework for formally defining and achieving privacy [Dwork et al, 2006]

Definition ((ϵ, δ)-DP)

A randomized algorithm A is (ϵ, δ)-DP if for all pairs of neighboring datasets D, D' and every subset S, we have

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

[Dwork et al 2016] The Algorithmic Foundations of Differential Privacy, Cynthia Dwork and Aaron Roth

Copyright © 2025

Differential privacy

Definition ((ϵ, δ)-DP)

A randomized algorithm A is (ϵ, δ)-DP if for all pairs of neighboring datasets D, D' and every subset S, we have

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

- must hold for all pairs of neigboring datasets
- must hold for all possible outputs
- Algorithm must be random

Copyright © 2025

Property of DP

- Immune to post-processing: if A is (ϵ, δ)-DP, then any $f(A)$ is also (ϵ, δ)-DP
- Composition property:

Theorem 3.16. Let $\mathcal{M}_i : \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}_i$ be an $(\varepsilon_i, \delta_i)$-differentially private algorithm for $i \in [k]$. Then if $\mathcal{M}_{[k]} : \mathbb{N}^{|\mathcal{X}|} \to \prod_{i=1}^{k} \mathcal{R}_i$ is defined to be $\mathcal{M}_{[k]}(x) = (\mathcal{M}_1(x), \ldots, \mathcal{M}_k(x))$, then $\mathcal{M}_{[k]}$ is $(\sum_{i=1}^{k} \varepsilon_i, \sum_{i=1}^{k} \delta_i)$-differentially private.

- Why composition is important? Attacks with multiple observations
- More observations, more risk

Copyright © 2025

Gaussian Mechanism

- Considering A takes a dataset as input and returning a p-dimensional real vector
- Denote the function sensitivity Δ as

$$\Delta = \max_{D,D'} ||A(D) - A(D')||_2$$

Theorem (Gaussian Mechanism)

Let $\epsilon < 1$. *For* $c^2 > 2\ln(1.25/\delta)$, $A(D) + \mathcal{N}(0, \sigma^2)$ *is* $(\epsilon, \delta) - DP$ *where* $\sigma = \frac{c\Delta}{\epsilon}$.

Theorem (Subsampling theorem)

If A is executed on a random fraction q of D, then it satisfies $O(q\epsilon, q\delta)$*-DP.*

Copyright © 2025

How to make training DP?

- What to protect in the training process? ***Gradient***
- How to bound the sensitivity? ***Bound the norm of the gradient***

- DP-SGD

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

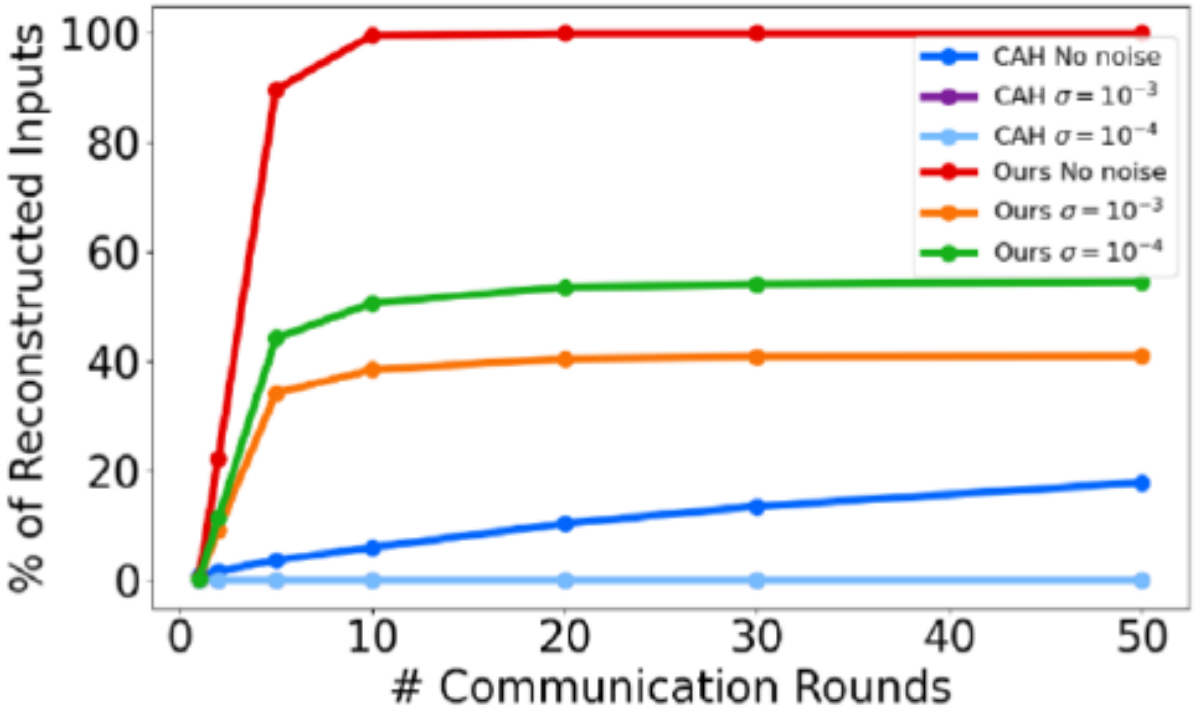
$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Deep Learning with Differential Privacy, Abadi et al, CCS 2016

Copyright © 2025

Trade off between privacy and model utility

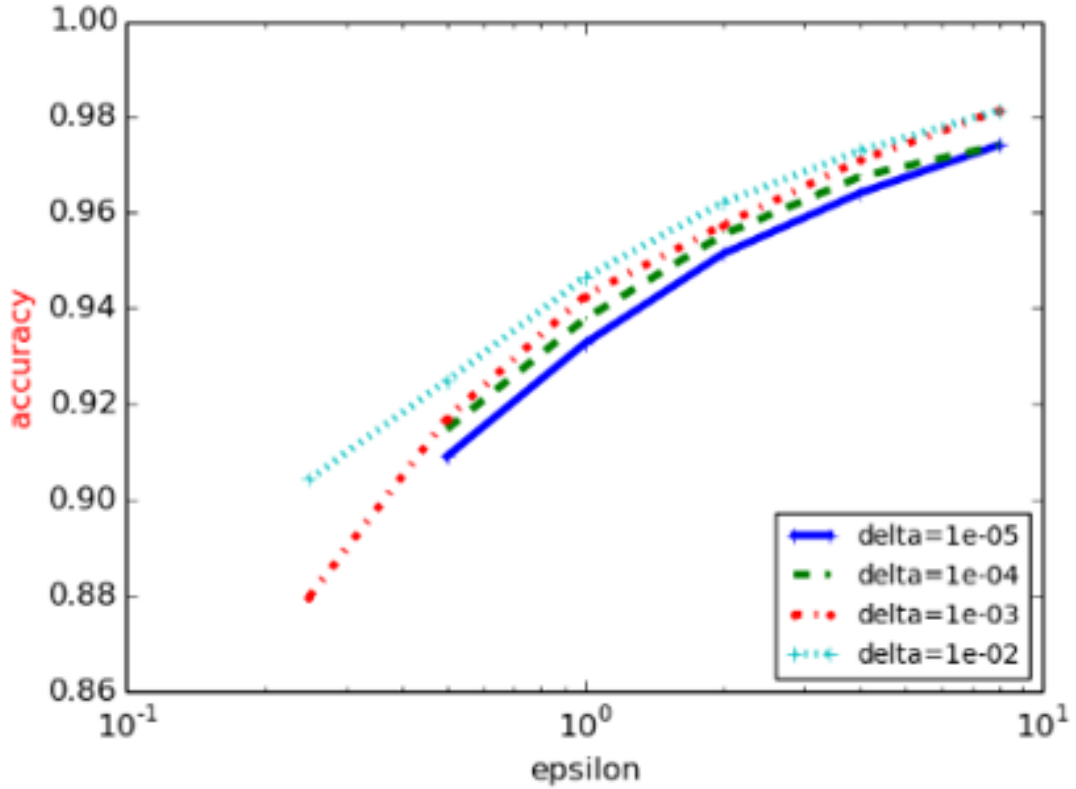
❖DP is effective against the FL attack



Copyright © 2025

Trade off between privacy and model utility

- But the model utility is affected



Model accuracy under various privacy guarantees on MNIST dataset

Copyright © 2025

Other privacy enhancing technologies

- Secure aggregation (additive mask)
- Homomorphic encryption
- Secure multi-party computation
- Trusted execution environement (TEE)

Copyright © 2025

Thank You

@

EAI

https://edge-ai-tech.eu/

The two-day innovative educational initiative is designed for students, young researchers, and professionals eager to expand their understanding and skills in these cutting-edge technologies.

Participants will engage in interactive discussions, hands-on sessions, and lectures covering foundational concepts and innovative applications of edge AI across various domains with leading experts, gain insights into real-world applications, and work on team-based projects to foster collaboration and innovation.

Copyright © 2025

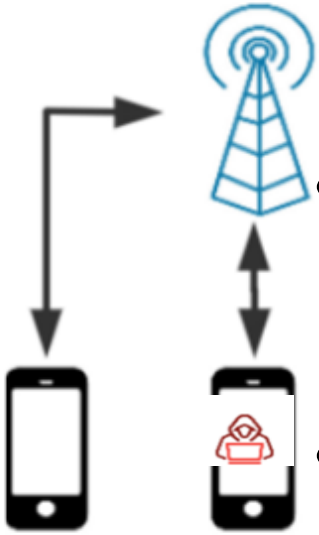
EDGE AI Academy 2025

Summer School Objectives

- The EDGE AI Academy Summer School aims to empower participants with advanced knowledge, hands-on experience and an in-depth understanding of edge AI, which combines Artificial Intelligence (AI) with edge computing and the Internet of Things (IoT) to process data closer to its source.
- Edge AI technology is gaining prominence due to its ability to enhance real-time decision-making, reduce latency, improve data privacy, and diminish reliance on continuous internet connectivity.
- The topics will extend to RISC-V architectures as ultra-low-power high-performance RISC-V-based processors and ultra-low-power vector-based accelerators are advancing smart IoT and edge AI computing capabilities.

Copyright © 2025

Data reconstruction attack – Malicious clients

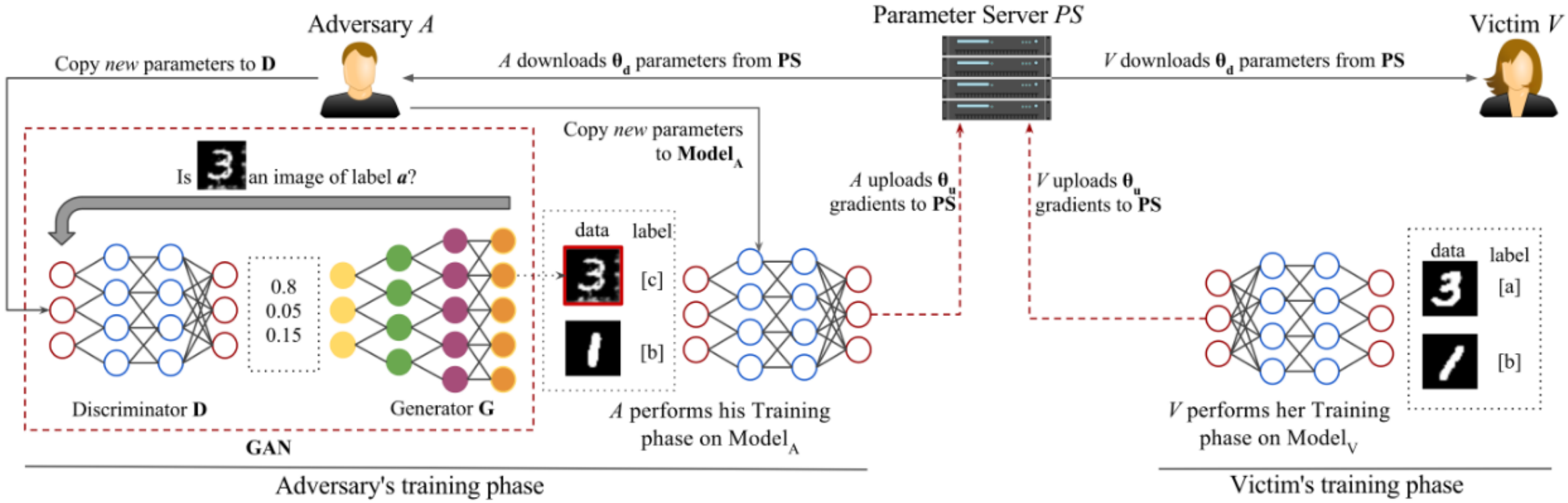


- The adversary can send arbitrary model parameters to the server
- Reconstruct another client's (victim) data points of some classes, D_C
- **Methodology:**
  - fool the victim to release more information on D_C
  - provide a wrong updated local model misclassified D_C

Deep models under the GAN: information leakage from collaborative deep learning, Hitaj et al, CCS 2017

Copyright © 2025

Data reconstruction attack – Malicious clients



- Limitations:
  - Works only for class representative datasets (such as MNIST, Facial datasets)
  - Hard for large scale of clients

Copyright © 2025

Backup

Local differential privacy in FL

Definition (FL algorithm $\mathcal{A}$ is (ϵ, δ)-DP:)

During the execution of $\mathcal{A}$, let Z_i be the function that outputs the transcript of communications between the server and client i. $\mathcal{A}$ is (ϵ, δ)-DP if for all the clients i, Z_i satisfies (ϵ, δ)-DP w.r.t. the dataset held by client i.

How to design an FL algorithm $\mathcal{A}$ which satisfies (ϵ, δ)-DP?

Copyright © 2025

Backup

(ϵ, δ)-DP FL algorithm

- The output of the client to the server should be random: **add the noise**
- The output should not be too sensitive to one data point: **bound the output**

DP-FedAvg:

for $k = 1, ..., K$ **do**

Data subsampling by user: $S_i^k \subset D_i$ of size $\lfloor sR \rfloor$

for *sample* $j \in S_i^k$ **do**

Compute gradient: $g_{ij} \leftarrow \nabla f_i(y_i^{k-1}, d_j^i)$

Clip gradient: $\tilde{g}_{ij} \leftarrow g_{ij} / \max\left(1, ||g_{ij}||_2 / \mathcal{C}\right)$

Add DP noise to local gradients: $\tilde{H}_i^k \leftarrow \frac{1}{sR} \sum_{j \in S_i^k} \tilde{g}_{ij} + \frac{2\mathcal{C}}{sR} \mathcal{N}(0, \sigma_g^2)$

$y_i^k \leftarrow y_i^{k-1} - \eta_l \tilde{H}_i^k$

Differentially Private Federated Learning on Heterogeneous Data, Noble et al, 2023

Copyright © 2025