Cutting Through Privacy: A Hyperplane-Based Data Reconstruction Attack in Federated Learning

Francesco Diana[1,2], André Nusser [1,2,3,4], Chuan Xu[1,2,3,4], Giovanni Neglia[1,2]
[1]Université Côte d'Azur , [2]Inria, [3]CNRS, [4]I3S

uai2025

Context

- In Federated Learning (FL), clients collaborate to learn a global model θ which minimizes the empirical risk:

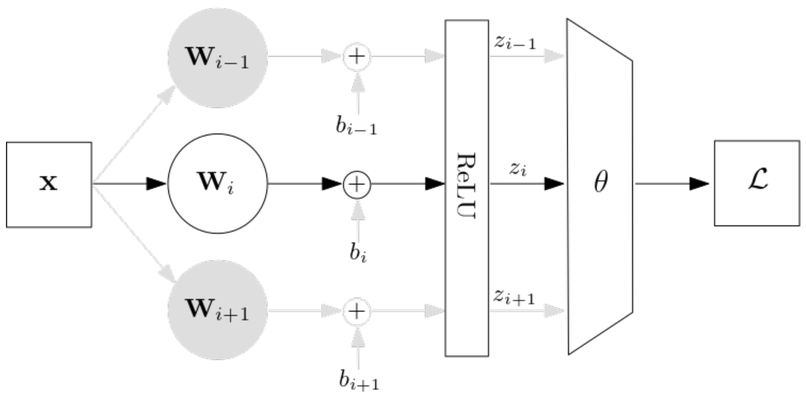
$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) = \sum_{u \in \mathcal{U}} p^u \mathcal{L}(\theta, \mathcal{D}^u)$$

- Honest-but-curious attackers can reconstruct training data from the updates.
- A **malicious** server can manipulate model parameters to increase data leakage.

Data Reconstruction Attacks

Two strategies:

- **Optimization-based [1]**: server iteratively optimizes initial dummy inputs to match client updates.
- **Analytical [2]**: server exploits properties of fully connected (FC) layers with ReLU activations in the model to invert the input.



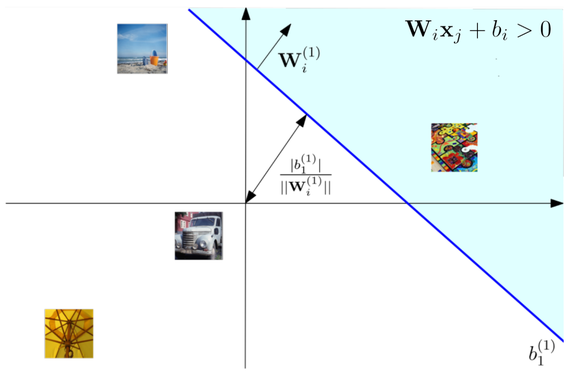
The server receives:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_i} = \frac{1}{n} \sum_{j=1}^{n} \frac{\partial \mathcal{L}_j}{\partial b_i} \mathbf{x}_j, \qquad \frac{\partial \mathcal{L}}{\partial b_i} = \frac{1}{n} \sum_{j=1}^{n} \frac{\partial \mathcal{L}_j}{\partial b_i}$$

and recovers:

$$\mathbf{g}_i = \frac{\partial \mathcal{L}}{\partial \mathbf{W}_i} \cdot \left(\frac{\partial \mathcal{L}}{\partial b_i} \right)^{-1} = \sum_{j=1}^{n} \alpha_j \mathbf{x}_j, \qquad \alpha_j = \frac{\frac{\partial \mathcal{L}_j}{\partial b_i}}{\sum_{k=1}^{n} \frac{\partial \mathcal{L}_k}{\partial b_i}}$$

In **sparsity-based** attacks [2],server's goal is to have, for a neuron i:
$\mathbf{W}_i \mathbf{x}_j + b_i > 0$ and
$\mathbf{W}_i \mathbf{x}_k + b_i < 0$, i.e.
$\alpha_j = 1$ and $\alpha_k = 0$, for
$k \in \{1, \ldots, n\} \setminus \{j\}$.



Motivations

- Optimization-based attacks do not achieve high-quality recovery of the inputs for large batches.
- Existing analytical attacks present intrinsic limitations or require assumptions on clients' data distribution.

Main Contributions

- Upper bound on the accuracy of data reconstruction attacks.
- We show how a malicious server can control each data point's contribution to the client's update.
- A perfect reconstruction algorithm for classification tasks, agnostic to input dimensionality, enabling full-batch recovery.
- Our attack fully recovers batches of up to 4096 data points (both images or tabular records).

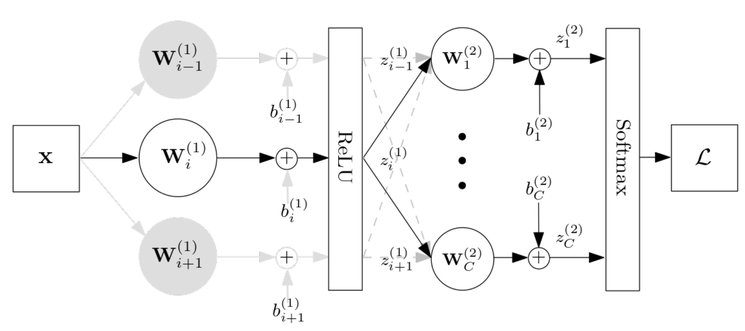
Theorem: Consider d-dimensional data and FedSGD with full-batch updates with size n. The expected number of inputs an attacker can recover by isolating individual samples is:

1. $O(n^{(d-1)/(d+1)})$ if samples are drawn uniformly from the unit ball;
2. $O(\log^{d-1} n)$ if samples are drawn uniformly from the unit hypercube;
3. $O(\log^{(d-1)/2} n)$ if samples follow a centered Gaussian with covariance $\mathbf{I}_d$.

Our Method

Clients perform one full-batch gradient update.

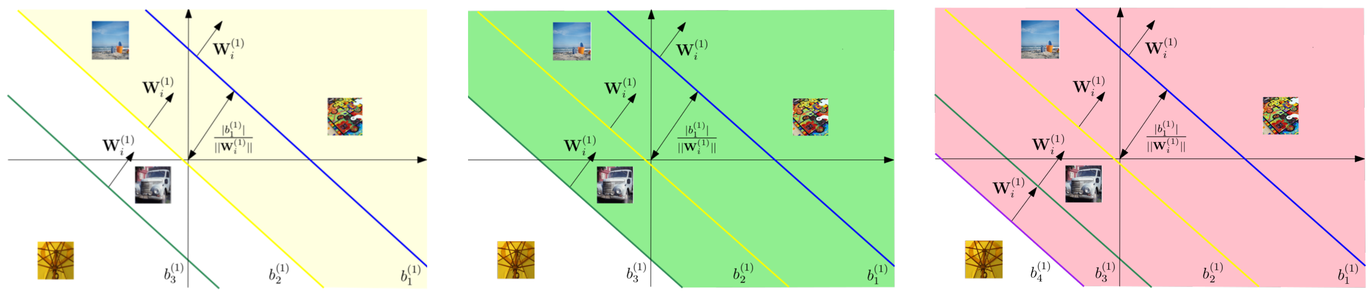
Assumption: if neuron i is activated by the input $\mathbf{x}_j$ for a given value of $b_i^{(1)}$, the derivative $\partial \mathcal{L}_j / \partial b_i^{(1)}$ remains independent of $b_i^{(1)}$.



- To satisfy the assumption, the server sets large values for $\mathbf{b}^{(2)}$, such that:

$$\frac{\partial \mathcal{L}_j}{\partial b_i^{(1)}} = -w_{y_j,i}^{(2)} + \sum_{k \in \mathcal{C}} w_{k,i}^{(2)} \frac{\exp(z_k^{(2)})}{\sum_{c \in \mathcal{C}} \exp(z_c^{(2)})} \approx -w_{y_j,i}^{(2)} + \sum_{k \in \mathcal{C}} w_{k,i}^{(2)} \frac{1}{|\mathcal{C}|}.$$

- Server fixes a random direction $\mathbf{W}_i^{(1)}$ and varies $b_i^{(1)}$ across rounds, sweeping a hyperplane through the input space.



$$\mathbf{g}_2 = \alpha_{2,1}\mathbf{x}_1 + \alpha_{2,2}\mathbf{x}_2 \quad \neq \quad \mathbf{g}_3 = \alpha_{3,1}\mathbf{x}_1 + \alpha_{3,2}\mathbf{x}_2 + \alpha_{3,3}\mathbf{x}_3 = \mathbf{g}_4 = \alpha_{3,1}\mathbf{x}_1 + \alpha_{3,2}\mathbf{x}_2 + \alpha_{3,3}\mathbf{x}_3$$

- Final hyperplane positions $b_1, \ldots, b_n$ are found through binary search.
- Iterative reconstruction of $\mathbf{x}_1, ..., \mathbf{x}_k, ..., \mathbf{x}_n$.
- By $\partial \mathcal{L}_{k+1} / \partial b_{k+1} = n(h_{k+1} - h_k)$, where $h_k = \partial \mathcal{L} / \partial b_k$:

$$\alpha_{j,k} = \frac{\frac{\partial \mathcal{L}_j}{\partial b_k}}{\sum_{l=1}^{n} \frac{\partial \mathcal{L}_l}{\partial b_k}} \qquad \mathbf{x}_{k+1} = \frac{\mathbf{g}_{k+1} - \sum_{j=1}^{k} \alpha_{j,k+1} \mathbf{x}_j}{\alpha_{j,k+1}}.$$

Experiments

We tested our attack on ImageNet and HARUS datasets. Our baseline is the *Curious Abandon Honesty* (CAH) attack [2].

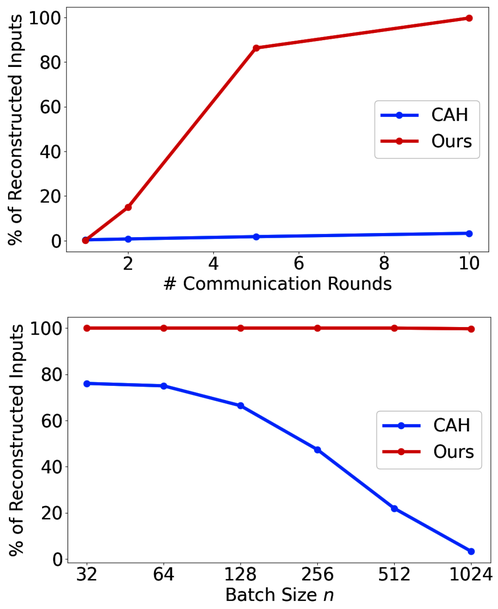


Figure 1: Effect of communication rounds and batch size on ImageNet.

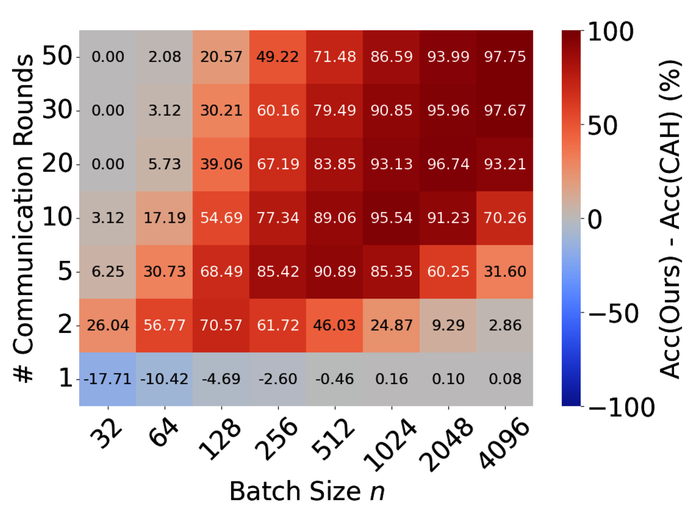


Figure 2: Reconstruction accuracy difference on HARUS dataset.

References

[1] Yuxin Wen, Jonas A Geiping, Liam Fowl, Micah Goldblum, and Tom Goldstein. Fishing for user data in large-batch federated learning via gradient magnification. In International Conference on Machine Learning, pages 23668–23684. PMLR, 2022.
[2] Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, and Nicolas Papernot. When the curious abandon honesty: Federated learning is not private. In 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P), pages 175–199, 2023.