Attribute Inference Attacks for Federated Regression Tasks

Francesco Diana[1,2], Othmane Marfoq[3], Chuan Xu[1,2,4,5], Giovanni Neglia[1,2], Frédéric Giroire[1,2,4,5], Eoin Thomas[6]

[1]Université Côte d'Azur , [2]Inria, [3]Meta, [4]CNRS, [5]I3S, [6]Amadeus

Introduction

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

$$\min_{\theta\in\mathbb{R}^d} \mathcal{L}(\theta) = \sum_{c\in\mathcal{C}} p_c \left(\frac{1}{S_c} \sum_{i=1}^{S_c} \ell(\theta, \mathbf{x}_c(i), y_c(i)) \right),$$

  where S_c is client c's dataset size.
- No formal privacy guarantees in FL.
- Clients' private information can be leaked.

Attribute Inference Attacks (AIA)

An adversary leverages public information $\{(\mathbf{x}_c^p(i), y_c^p(i)), i = 1, ..., S_c\}$ and exchanged updates $\mathcal{M}_c$ to recover the sensitive attributes $s_c(i)$.

Two existing approaches:

- **Gradient-based [1]** Select the sensitive attribute values that yield virtual gradients closely resembling the client's model updates, by solving

$$\underset{\{s_c(i)\}_{i=1}^{S_c}}{\operatorname{argmax}} \sum_{t\in\mathcal{T}} \textbf{CosSim}\left(\frac{\partial \ell(\theta^t, \{(\mathbf{x}_c^p(i), s_c(i), y_c^p(i))\})}{\partial \theta^t}, \theta^t - \theta_c^t \right)$$

- **Model-based [2]** In centralized training, the adversary solves

$$\underset{s_c(i)}{\operatorname{argmin}}\, \ell(\theta, (\mathbf{x}_c^p(i), s_c(i), y_c^p(i))), \quad \forall i \in \{1, ..., S_c\}$$

Motivations

- Reconstruction attacks in FL have not been explored for **regression** tasks.
- Accuracy of SOTA gradient-based attack for FL drops to random guess.

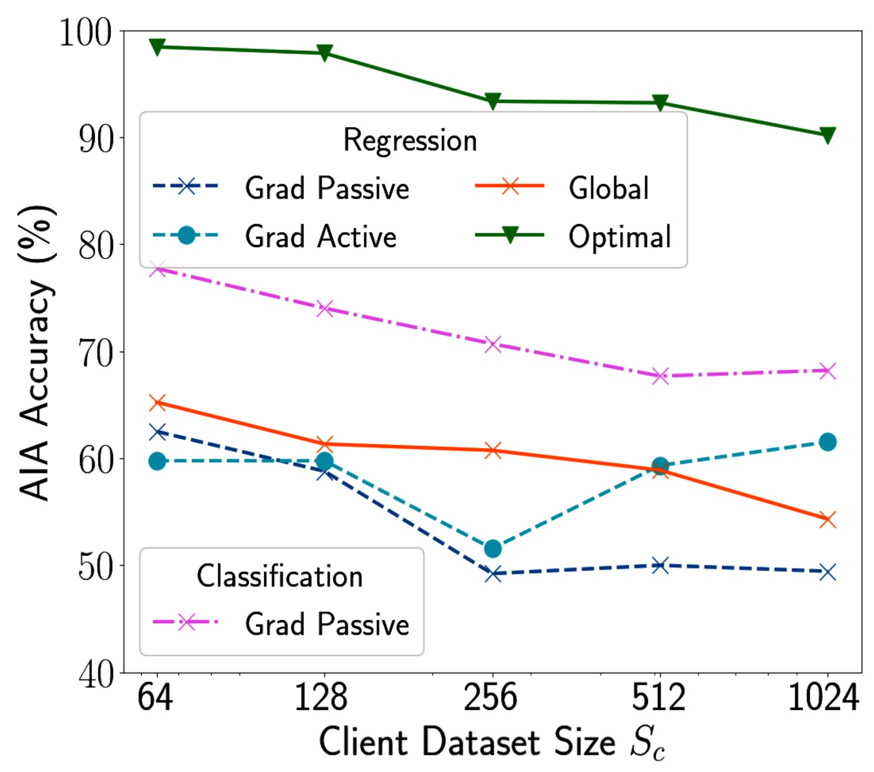

Figure 1: Effect of client dataset size on different AIAs.

Main Contributions

- An analytical lower bound for model-based AIA accuracy in the least squares regression problem, motivating the adversary's strategy to approximate the client's optimal local model in federated regression tasks.
- Methods for approximating optimal local models where adversaries can either eavesdrop on exchanged messages or directly interfere with the training process.
- Experiments show that our model-based AIAs are better candidates for empirically quantifying privacy leakage for federated regression tasks.

Reconstructing the local model

Passive approach for linear least squares

- The adversary knows the trained model structure, the loss function, and the training algorithm.
- He has access to $\mathcal{M}_c$ but does not interfere with the training process.

Algorithm 1: Reconstruction of client-c local model by a passive adversary for federated least squares regression

Input: the server models $\theta^{t_i}(0) = \theta_c^{t_i}(0)$ and the local updated models $\theta_c^{t_i}(K)$ at all the inspected rounds $t_i \in \mathcal{T}_c = \{t_1, t_2, \dots, t_{n_c}\}$.

1: Let $\Theta_{\text{in}} = [\theta_c^{t_1}(0)\ \theta_c^{t_2}(0)\ ...\ \theta_c^{t_{n_c}}(0)]^T \in \mathbb{R}^{n_c\times d}$

2: Let $\Theta_{\text{out}} = \begin{bmatrix} (\theta_c^{t_1}(0) - \theta_c^{t_1}(K))^T & 1 \\ \vdots & \\ (\theta_c^{t_{n_c}}(0) - \theta_c^{t_{n_c}}(K))^T & 1 \end{bmatrix} \in \mathbb{R}^{n_c\times(d+1)}$

3: $(\hat{\theta}_c^*)^T \leftarrow$ last row of $((\Theta_{\text{out}}^T\Theta_{\text{out}})^\dagger \Theta_{\text{out}}^T \Theta_{\text{in}})$

4: Return $\hat{\theta}_c^*$ as the estimator for client c's local model

By eavesdropping on $n_c > d$ message exchanges between client c and the server, the error of the reconstructed model $\hat{\theta}_c^*$ of Alg. 1 is upper bounded w.p. $\geq 1-\delta$ when $\eta \leq \frac{S_c}{2\lambda_{\max}(\mathbf{x}_c^T\mathbf{x}_c)}$ and

$$\left\|\hat{\theta}_c^* - \theta_c^*\right\|_2 = \mathcal{O}\left(\eta\sigma d \sqrt{dE\left\lceil\frac{S_c}{B}\right\rceil \frac{d+1+\ln\frac{2d}{\delta}}{n_c\cdot\underline{\lambda}}} \right),$$

where d is the rank of $\mathbf{x}_c$, σ is the noise scale of the stochastic gradient, n_c is the number of messages, $\underline{\lambda}$ is a lower bound on the eigenvalues of matrix $\frac{\Theta_{\text{out}}^T\Theta_{\text{out}}}{n_c}$, B is the training batch size, and E is the number of local epochs.

Active approach

Algorithm 2: Reconstruction of client-c local model by an active adversary a

Input: Let $\mathcal{T}_c^a$ be set of rounds during which the adversary attacks client c and θ_c^a be the corresponding malicious model.

1: $\theta_c^a \leftarrow$ latest model received from client c

2: **for** $t \in \mathcal{T}_c^a$ **do**

3: a sends the model θ_c^a to client c,

4: a waits the updated model from θ_c from client c,

5: a computes the pseudo-gradient $\theta_c^a - \theta_c$ and updates θ_c^a and the corresponding moment vectors following Adam,

6: Return θ_c^a as the estimator for client c's local model

Experiments

AIA (%) / Datasets		Income-L	Income-A	Medical
Passive	Grad	60.36±0.67	54.98±0.29	87.26 ±0.92
	Grad-w-O	71.44±0.33	**56.10±1.12**	91.06 ±0.55
	Ours	**75.27 ±0.32**	55.75 ±0.17	**95.90 ±0.04**
Active (10 Rnds)	Grad	60.24 ±0.60	54.98 ±0.29	87.26 ±0.92
	Grad-w-O	80.69 ±0.55	56.10 ±1.12	91.06 ±0.55
	Ours	**82.02 ±0.85**	**63.53 ±0.73**	**95.93 ±0.07**
Active (50 Rnds)	Grad	60.24 ±0.60	53.36 ±0.40	87.26 ±0.92
	Grad-w-O	80.69 ±0.55	56.12 ±0.12	91.06 ±0.55
	Ours	**94.31 ±0.11**	**78.09 ±0.25**	**96.79 ±0.79**
Model-w-O		94.31 ±0.11	78.31 ±0.07	96.79 ±0.79

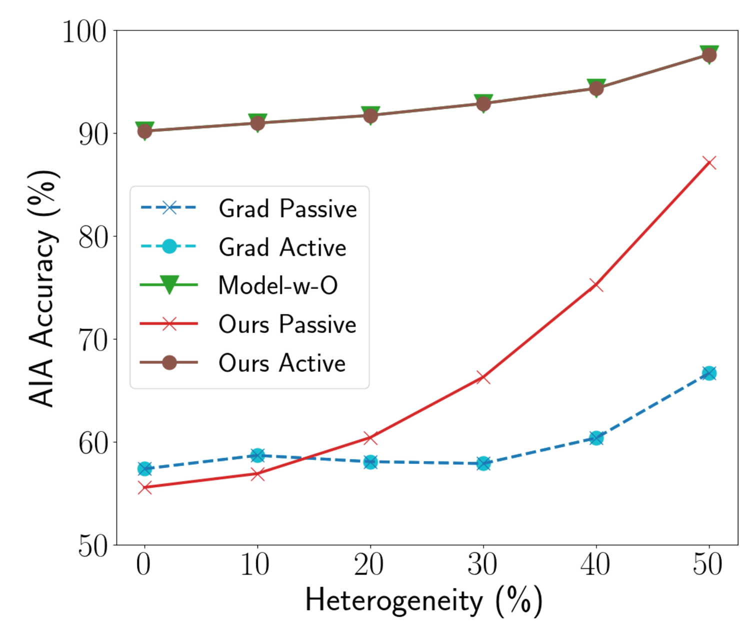

Figure 2: Effect of local dataset heterogeneity on Income-L.

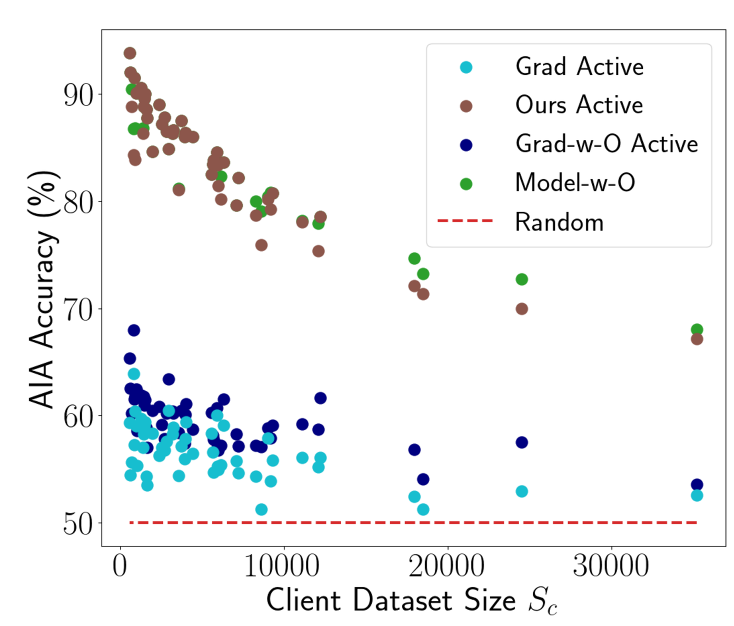

Figure 3: Effect of local dataset size on Income-A.

References

[1] Chen Chen, Lingjuan Lyu, Han Yu, and Gang Chen. "Practical Attribute Reconstruction Attack Against Federated Learning." In: IEEE Transactions on Big Data. 2024.

[2] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. "Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing." In: 23rd USENIX Security Symposium. 2014.