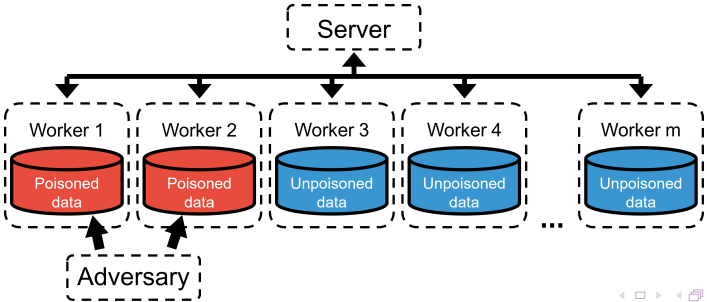
Near Optimal Robust Federated Learning Against Data Poisoning Attack

Jingfan Yu Zhixuan Fang

Tsinghua University
Shanghai Qi Zhi Institute

yujf20@mails.tsinghua.edu.cn
zfang@mail.tsinghua.edu.cn

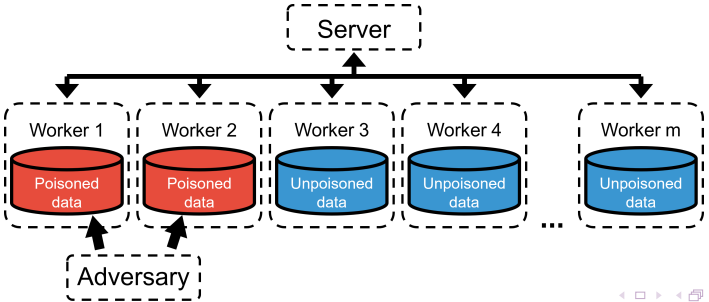
Data Poisoning Attacks in Federated Learning



Data Poisoning Attacks in Federated Learning

Data poisoning attacks:

- Attacks where an adversary modifies the datasets of compromised workers.
- The affected workers are honest, but their raw data is unknowingly corrupted by the adversary.

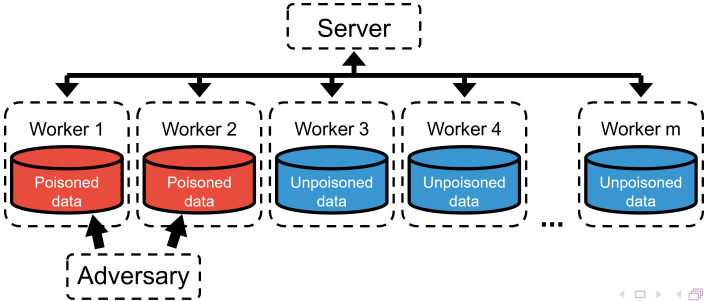


Data Poisoning Attacks in Federated Learning

Data poisoning attacks:

- Attacks where an adversary modifies the datasets of compromised workers.
- The affected workers are honest, but their raw data is unknowingly corrupted by the adversary.

Compared with model poisoning attacks, in which adversaries directly manipulate model parameters, data poisoning attacks are weaker but more practical to carry out.



Two-fold Design Goal: High Accuracy, Low Adversary Gain

Utility: Attack loss

- The attack loss measures the increase in target model error caused by the attack.

For a mechanism $\mathcal{M}$ that outputs h under attack,

$$\text{Attack loss} = \mathrm{Err}(h) - \mathrm{Err}(h^*),$$

where h^* is the optimal hypothesis.

Two-fold Design Goal: High Accuracy, Low Adversary Gain

Utility: Attack loss

- The attack loss measures the increase in target model error caused by the attack.

For a mechanism $\mathcal{M}$ that outputs h under attack,

$$\text{Attack loss} = \text{Err}(h) - \text{Err}(h^*),$$

where h^* is the optimal hypothesis.

Robustness: Effective poison rate (EPR)

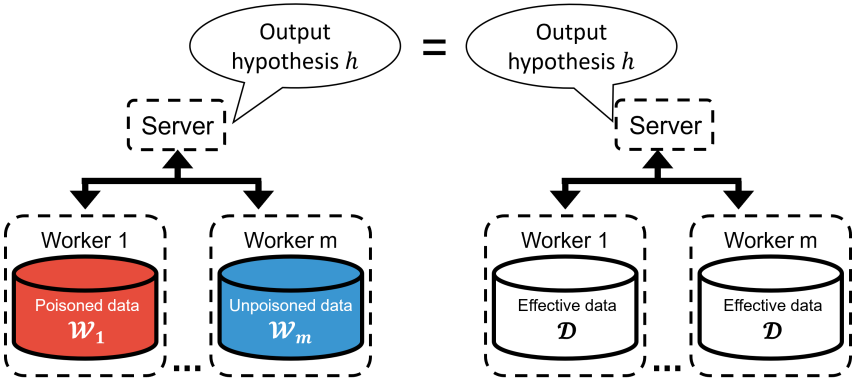
- EPR characterizes the adversary's possible gain from the attack.
- Addressing the problem that the adversary can achieve a high gain from the attack while maintaining a low attack loss.

Informally, EPR is the $\mathcal{H}$-divergence between the ground truth distribution and the effective dataset $\mathcal{D}$:

$$\text{EPR}(\mathcal{M}) = d_{\mathcal{H}}\left(\mathcal{D}_{\text{ground truth}}, \mathcal{D}\right).$$

Robustness Goal: Effective Poison Rate (EPR)

EPR is the $\mathcal{H}$-divergence between the ground truth distribution and the effective dataset $\mathcal{D}$, $d_{\mathcal{H}}\left(\mathcal{D}_{\text{ground truth}}, \mathcal{D}\right)$.



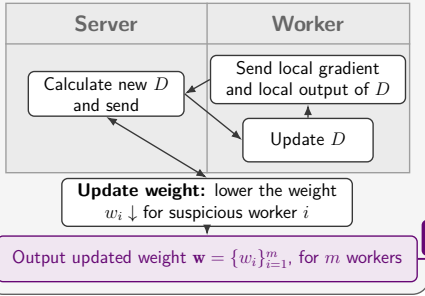
Algorithm Overview

Phase 1: Trustworthiness Weight Update

Train discriminator D by maximizing variance:

$$D^* := \arg\max_{D \in \mathcal{H}} Var_D(\{\mathcal{W}_i\}_{i=1}^m),$$

where $Var_D(\{\mathcal{W}_i\}_{i=1}^m)$ is the variance of the discriminator D's outputs across the workers' datasets.



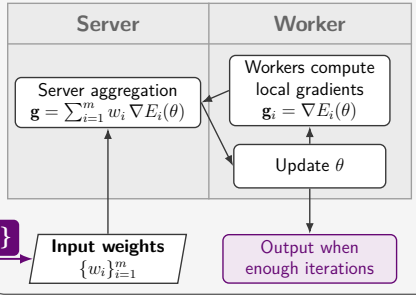
$\{w_i\}$

Phase 2: Target Model Training

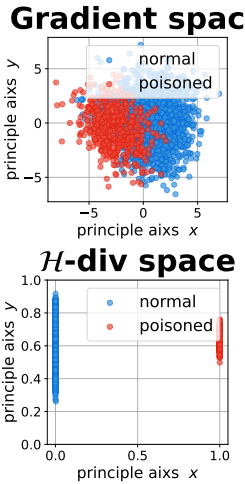
Train target model θ by solving weighted objective:

$$\theta^* = \arg\min_{\theta \in \Theta} \sum_{i=1}^m w_i E_i(\theta).$$

where $E_i(\theta)$ is the local error of worker i and w_i is the weight assigned to worker i in Phase 1.

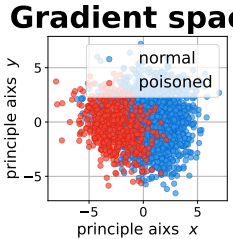


Intuition: $\mathcal{H}$-divergence can be a more effective measure for outlier detection than the Euclidean distance of gradients.



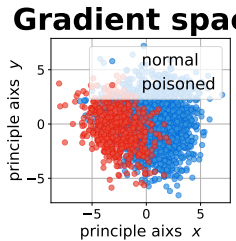
- Each point refers to a worker and its dataset. The poisoned worker (red) is more distinguishable from normal workers (blue) in $\mathcal{H}$-divergence space than in gradient space.

Intuition: $\mathcal{H}$-divergence can be a more effective measure for outlier detection than the Euclidean distance of gradients.



- Each point refers to a worker and its dataset. The poisoned worker (red) is more distinguishable from normal workers (blue) in $\mathcal{H}$-divergence space than in gradient space.
- In our algorithm, the discriminator model is trained to map each worker's dataset to a representation in the $\mathcal{H}$-divergence space and downweight workers whose representations are far from the majority.

Intuition: $\mathcal{H}$-divergence can be a more effective measure for outlier detection than the Euclidean distance of gradients.



- Each point refers to a worker and its dataset. The poisoned worker (red) is more distinguishable from normal workers (blue) in $\mathcal{H}$-divergence space than in gradient space.
- In our algorithm, the discriminator model is trained to map each worker's dataset to a representation in the $\mathcal{H}$-divergence space and downweight workers whose representations are far from the majority.
- Thus, our approach filters out poisoned workers much more effectively than conventional gradient-based defenses.

Theoretical Results: Asymptotically Optimal

- We establish the lower bound of the minimax attack loss. It depends on the correlation among the normal workers' datasets, which can be characterized by the number of data samples n and the non-IID level γ of the datasets of workers.

	Attack loss ($m \to \infty$)	
	iid	non-iid
Our algorithm	$\tilde{O}((\frac{1}{n})^{\frac{1}{2}})$	$\tilde{O}((\frac{1}{\gamma})^{\frac{1}{2}} + (\frac{1}{n})^{\frac{1}{2}})$
Lower bound	$\Omega((\frac{1}{n})^{\frac{1}{2}})$	$\Omega((\frac{1}{\gamma})^{\frac{1}{2}} + (\frac{1}{n})^{\frac{1}{2}})$

Theoretical Results: Asymptotically Optimal

- We establish the lower bound of the minimax attack loss. It depends on the correlation among the normal workers' datasets, which can be characterized by the number of data samples n and the non-IID level γ of the datasets of workers.
- Our algorithm asymptotically matches the lower bound in both IID and non-IID settings, when the number of workers m approaches infinity.

	Attack loss ($m \to \infty$)	
	iid	non-iid
Our algorithm	$\tilde{O}((\frac{1}{n})^{\frac{1}{2}})$	$\tilde{O}((\frac{1}{\gamma})^{\frac{1}{2}} + (\frac{1}{n})^{\frac{1}{2}})$
Lower bound	$\Omega((\frac{1}{n})^{\frac{1}{2}})$	$\Omega((\frac{1}{\gamma})^{\frac{1}{2}} + (\frac{1}{n})^{\frac{1}{2}})$

Theoretical Results: Asymptotically Optimal

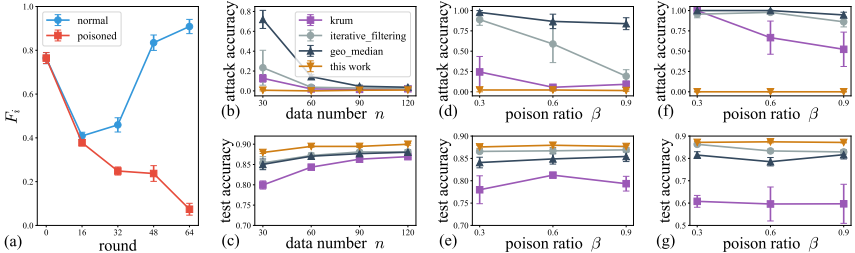
- We establish the lower bound of the minimax attack loss. It depends on the correlation among the normal workers' datasets, which can be characterized by the number of data samples n and the non-IID level γ of the datasets of workers.
- Our algorithm asymptotically matches the lower bound in both IID and non-IID settings, when the number of workers m approaches infinity.

	Attack loss ($m \to \infty$)	
	iid	non-iid
Our algorithm	$\tilde{O}((\frac{1}{n})^{\frac{1}{2}})$	$\tilde{O}((\frac{1}{\gamma})^{\frac{1}{2}} + (\frac{1}{n})^{\frac{1}{2}})$
Lower bound	$\Omega((\frac{1}{n})^{\frac{1}{2}})$	$\Omega((\frac{1}{\gamma})^{\frac{1}{2}} + (\frac{1}{n})^{\frac{1}{2}})$

Compared to the baselines, our algorithm achieves a lower attack loss in both settings and successfully mitigates the adversary's gain. Moreover, it is theoretically insensitive to model dimension, making it scalable to massive models.

Evaluation

We evaluate robust federated learning mechanisms on the MNIST and CIFAR-10 datasets with the flip-label attack and backdoor attack. The following figure shows the result of defending against the flip-label attack on the MNIST dataset.



Our algorithm maintains higher test accuracy and lower attack accuracy than the baselines in both IID and non-IID settings.

Conclusion

- We establish the minimax lower bound of the target model attack loss under data poisoning attacks. We show that the worst-case defense performance of any algorithm depends on the correlation among the normal workers' datasets.

Conclusion

- We establish the minimax lower bound of the target model attack loss under data poisoning attacks. We show that the worst-case defense performance of any algorithm depends on the correlation among the normal workers' datasets.
- We propose an algorithm to asymptotically match the lower bound. Our algorithm thoroughly exploits the correlation among the normal workers' datasets. Our approach is specifically tailored to defend against data poisoning attacks and is more effective than existing defenses.

Conclusion

- We establish the minimax lower bound of the target model attack loss under data poisoning attacks. We show that the worst-case defense performance of any algorithm depends on the correlation among the normal workers' datasets.
- We propose an algorithm to asymptotically match the lower bound. Our algorithm thoroughly exploits the correlation among the normal workers' datasets. Our approach is specifically tailored to defend against data poisoning attacks and is more effective than existing defenses.
- We propose the notion of Effective Poison Rate (EPR), that characterizes the adversary's gain from the attack. We show that our algorithm mitigates the adversary's gain both theoretically and empirically.

Thanks