Attacks on Federated Learning

- Attack 1: Data poisoning attack [1].
- Attack 2: Model poisoning attack [2].

- Defense 1: Server check validation accuracy.
- Defense 2: Server check gradient statistics.
- Defense 3: Byzantine-tolerant aggregation [3, 4, 5].

References

1. Shafah and others: Poison frogs! targeted clean-label poisoning attacks on neural networks. In *NIPS*, 2018.
2. Bhagoji and others: Analyzing federated learning through an adversarial lens. In *ICML*, 2019.
3. Blanchard, Guerraoui, & Stainer: Machine learning with adversaries: Byzantine tolerant gradient descent. In *NIPS*, 2017.
4. Chen, Su, & Xu: Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *In Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2017.
5. Yin and others: Byzantine-robust distributed learning: Towards optimal statistical rates. In *ICML*, 2018.