Byzantine-Robust Learning on Heterogeneous Datasets via Resampling

Lie He*, Sai Praneeth Karimireddy*, Martin Jaggi
MLO, EPFL
{lie.he, sai.karimireddy, martin.jaggi}@epfl.ch

NEURAL INFORMATION PROCESSING SYSTEMS

Objectives

- Demonstrate the chanlleges of non-iid data for Byzantine-robust learning.
- Propose a novel *mimic* attack on non-iid data.
- Propose a resampling-based general defense framework.

Background

- **Byzantine-robust aggregation rules.**
  - Identifying the outliers by pairwise distance [1];
  - using robust statistics such as median [2] and geometric median [3].
  - ...

  However, most assume the iid inputs.
- **Challenge of the non-iid data.**
  - Good gradients not clustered around the global mean.
  - One good gradient does not represent all the good workers.

Mimic Attack

- All Byzantine workers mimic gradients from a specific worker.
- Throughout training, median-based and middle-seeking aggregation rules only pick the gradients from this worker.

Methodology

Solution. We propose to use *resample* the gradients before using the aggregation rules (Algorithm 1).

1. workers compute updates locally;
2. workers send gradients to a central server;
3. the server resamples the gradients (Algorithm 1);
4. the server robustly aggregates the resampled gradients.
5. workers pull the latest updates from the server.

Algorithm 1 Resampling with s-replacement

Input: $\{\boldsymbol{g}_i : i \in [n]\}$, $T = n$, s, $\{c[i] = 0 : i \in [n]\}$
for $t := 1, \dots, T$ **do**
  for $i := 1, \dots, s$ **do**
    while Select $j_i \sim \mathbf{Uniform}([n])$ **do**
      if $c[j_i] < s$ **then**
        $c[j_i] \mathrel{+}= 1$
        If $c[j_i] == s$ Break;
Return $\{\bar{\boldsymbol{g}}_t : t \in [T]\}$, $\{j_i^t : t \in [T], i \in [s]\}$

Takeaway

- **Chanlleges.** On non-iid data, existing Byzantine-robust aggregation rules
  - fail to satisfy the assumption of iid input.
  - May fail even without attackers.
  - Are vulnerable to *mimic attack*.
- **Defense.** We propose to *resample* the worker gradients before aggregation.
  - Resampled gradients are identically distributed.
  - Resampling reduce the variance among gradients.
  - Resampling theoretically and empirically improves the existing aggregation rules on non-iid data and against the mimic attack.
  - Resampling works in a plug-and-play fashion.

References

[1] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *NeurIPS - Advances in Neural Information Processing Systems 30*, pages 119–129, 2017.

[2] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. *arXiv preprint arXiv:1803.01498*, 2018.

[3] Krishna Pillutla, Sham M. Kakade, and Zaid Harchaoui. Robust Aggregation for Federated Learning. *arXiv preprint arXiv:1912.13445*, 2019.

Empirical Results

We train models on MNIST with 8 good nodes and f Byzantine nodes. We compare KRUM [1], CM [2], and RFA [3] on iid and non-iid data, with or without resampling (RS).