# ByGARS: Byzantine SGD with Arbitrary Number of Attackers

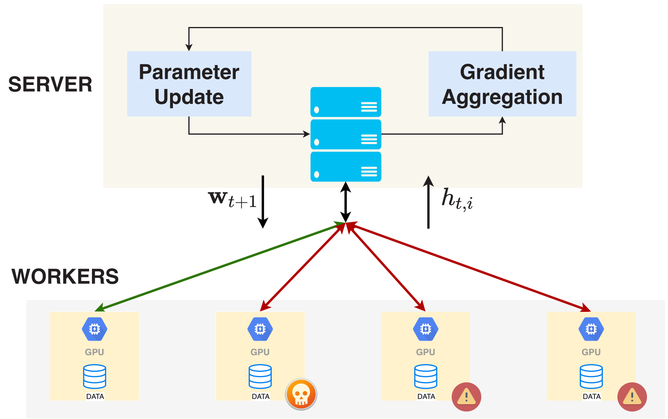
Jayanth Regatti, Hao Chen, Abhishek Gupta

Electrical and Computer Engineering, The Ohio State University

The Ohio State University
College of Engineering

SCAN ME

## PROBLEM




$$\mathbf{w}_{t+1} = \mathbf{w}_t - \gamma_t \ \texttt{Aggr}(h_{t,1}, ..., h_{t,m})$$

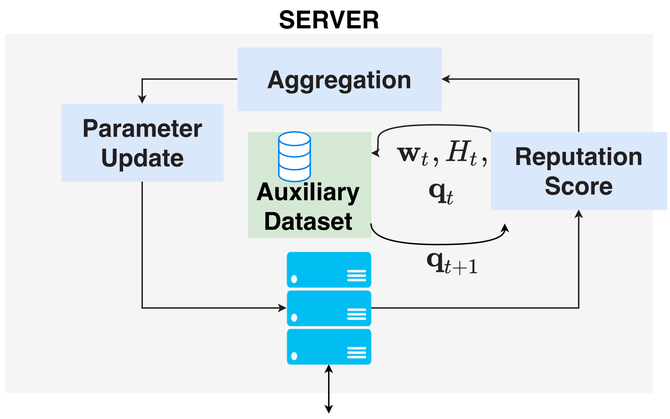
**Standard setting (fraction of adversaries $f < 0.5$)**

- Median based [1], Majority Voting based [2], Trimmed methods [3], Krum [4], etc.

**A more practical setting ( $f > 0.5$)**

- Not explored in detail
- Existing works such as [5] require additional knowledge on number of adversaries

## PROPOSED SOLUTION: REPUTATION SCORE BASED AGGREGATION




$$\mathbf{w}_{t+1} = \mathbf{w}_t - \gamma_t \sum_{i=1}^{m} h_{t,i} q_i = \mathbf{w}_t - \gamma_t H_t^T \mathbf{q}_t$$

A1 Auxiliary dataset available at server
A2 Worker behavior is stationary
$\Longrightarrow$ **Reputation scores can be learned**

## REPUTATION SCORES

$$\mathbf{q}_{t+1} = \arg\min_{\mathbf{q}\in\mathbb{R}^m} L_t\left(\mathbf{w}_t - \gamma_t \sum_{i=1}^{m} h_{t,i} q_i\right) = \arg\min_{\mathbf{q}\in\mathbb{R}^m} L_t(\mathbf{w}_t - \gamma_t H_t^T \mathbf{q})$$

where $H_t^T = [h_{t,1}, \cdots, h_{t,m}] \in \mathbb{R}^{d\times m}$ are stochastic gradients

### ByGARS

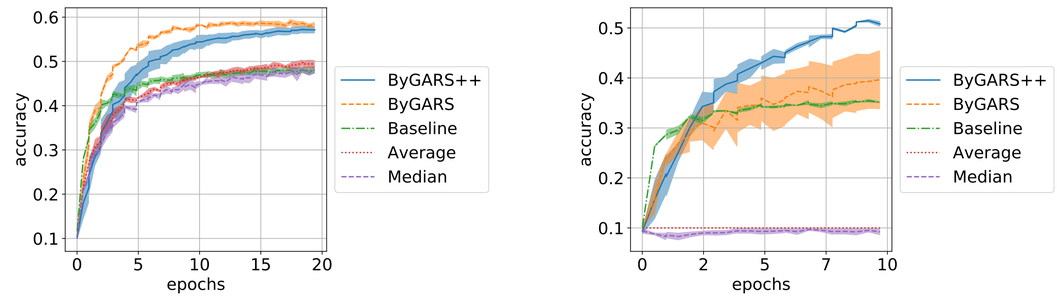
$$\widehat{\mathbf{w}}_{t+1} \leftarrow \mathbf{w}_t - \gamma_t H_t^T \mathbf{q}_{t+1}^{i-1}$$
$$\mathbf{q}_{t+1}^{i} \leftarrow \mathbf{q}_{t+1}^{i-1} + \alpha_t \gamma_t H_t \nabla L_t(\widehat{\mathbf{w}}_{t+1})$$

### ByGARS++

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \gamma_t H_t^T \mathbf{q}_t$$
$$\mathbf{q}_{t+1} \leftarrow (1-\alpha_t)\mathbf{q}_t + \alpha_t H_t \nabla L_t(\mathbf{w}_t)$$

**Theorem:** $L$ strongly convex $\Longrightarrow$ $\mathbf{w}_t \to \mathbf{w}^*$ almost surely.

## EXPERIMENTAL RESULTS




(a) Label Flip Attack with $< 0.5$ adversaries (b) Mixed Attack with $> 0.5$ adversaries

Figure: Defense against different types of attacks using ByGARS/ByGARS++ on CIFAR-10 dataset

Table: Summary of various attacks that ByGARS or ByGARS++ is robust to

| Type | Attack | Fraction of Adversaries $f$ $f < 0.5$ | $f \in [0.5, 1)$ | $f = 1$ |
|---|---|---|---|---|
| Omniscient / Collusion | Inner Product Manipulation | $\surd$ | $\surd$ | - |
| Omniscient / Collusion | LIE | $\surd$ | $\times$ | - |
| Omniscient / Collusion | OFOM | $\surd$ | $\surd$ | - |
| Omniscient / Collusion | PAF | $\surd$ | $\surd$ | - |
| Local / Failure | Sign Flip/Reverse Attack | $\surd$ | $\surd$ | $\surd$ |
| Local / Failure | Random Sign Flip Attack | $\surd$ | $\surd$ | $\surd$ |
| Local / Failure | Gaussian Attack | $\surd$ | $\surd$ | - |
| Local / Failure | Constant Attack | $\surd$ | $\surd$ | - |
| Data Poisoning | Label Flipping | $\surd$ | $\surd$ | - |
| *Mixed Attacks* | Multiple types of attacks | $\surd$ | $\surd$ | $\surd$ |

## STRENGTHS

- ByGARS++ has same complexity as normal averaging SGD
- Performance of ByGARS/ByGARS++ under *No attack* matches averaging SGD
- Need only a small amount of auxiliary data at the server to compute reputation scores efficiently
- ByGARS/ByGARS++ defend almost all state-of-the-art attacks even with an arbitrary number of adversaries

## REFERENCES


- Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 2, pp. 1–25, 2017.
- J. Bernstein, J. Zhao, K. Azizzadenesheli, and A. Anandkumar, "signSGD with majority vote is communication efficient and fault tolerant," *arXiv preprint arXiv:1810.05291*, 2018.
- D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," *arXiv preprint arXiv:1803.01498*, 2018.
- P. Blanchard, R. Guerraoui, J. Stainer, *et al.*, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems*, pp. 119–129, 2017.
- C. Xie, O. Koyejo, and I. Gupta, "Zeno: Byzantine-suspicious stochastic gradient descent," *arXiv preprint arXiv:1805.10032*, 2018.


The authors thank ARPA-E NEXTCAR program and ARO Grant W911NF1920256 for supporting the research. Results presented in this poster were obtained using the Chameleon testbed supported by the National Science Foundation.