Resilient to Byzantine Attacks Finite-Sum Optimization over Networks

Zhaoxian Wu

School of Data and Computer Science, Sun Yat-Sen University

Joint work with
Qing Ling (SYSU) Tianyi Chen (RPI) Georgios B. Giannakis (UMN)

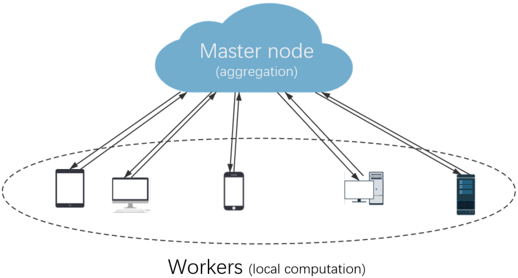
45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2020)

Outline

Background

- Federated learning: A promising distributed learning framework [1]
  - (Distributed) Major computations are carried at workers locally
  - (Privacy-preserving) Data in workers are kept private
- There may be Byzantine attackers biasing the learning process

How to alleviate the negative effect caused by Byzantine attacks?



Problem Statement

- Consider a network with one master node and W workers, among which B workers are Byzantine attackers
- The goal is to find the solution of the following optimization problem:

$$x^* = \arg\min_x \; f(x) := \frac{1}{W-B} \sum_{w \notin \mathcal{B}} f_w(x) \tag{1}$$

- Notations
  - $f_w(x) := \frac{1}{J} \sum_{j=1}^{J} f_{w,j}(x)$: local finite-sum objective
  - $\mathcal{B}$: set of Byzantine workers, with $|\mathcal{B}| = B$
  - $x \in \mathbb{R}^p$: optimization variable

Revisiting SGD

- Stochastic Gradient Descent (SGD): a popular solver of problem (1)
- SGD updates at time slot k
  - Master node broadcasts x^k to workers
  - Worker w uniformly at random chooses a local data sample (or a mini batch) with index i_w^k and computes stochastic gradient
  - Worker w communicates $f'_{w,i_w^k}(x^k)$ back to the master node
  - Master node updates the model as

$$x^{k+1} = x^k - \gamma^k \frac{1}{W} \sum_{w=1}^{W} f'_{w,i_w^k}(x^k) \tag{2}$$

  where γ^k is the non-negative step size
- SGD is vulnerable to Byzantine attacks

Illustration of Byzantine Attacks

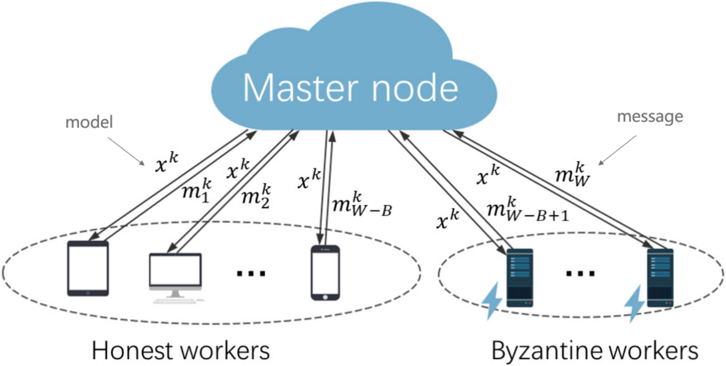


Figure 1: Illustration of SGD in federated learning framework and Byzantine attacks model. In practice, the identities of Byzantine attackers are unknown to the master node

Byzantine Attacks Model

- Byzantine workers can send arbitrary messages to the master node
- Let m_w^k denote the message that worker w sends to the master node at slot k, given by

$$m_w^k = \begin{cases} f'_{w,i_w^k}(x^k), & w \notin \mathcal{B} \\ *, & w \in \mathcal{B} \end{cases} \tag{3}$$

- The update rule of SGD can be written as

$$x^{k+1} = x^k - \gamma^k \frac{1}{W} \sum_{w=1}^{W} m_w^k \tag{4}$$

- Even only one Byzantine attacker can lead SGD to fail
  - $m_{w_b}^k = -\sum_{w \neq w_b} m_w^k$ yields $x^{k+1} = x^k$
  - $m_{w_b}^k = +\infty$ yields $x^{k+1} = +\infty$

Existing Algorithm: SGD with Geometric Median

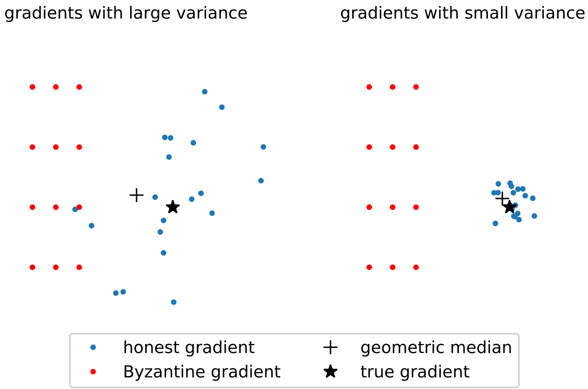
- To replace the mean with geometric median, we can robustify distributed SGD [2]
- The geometric median of $\{z, z \in \mathcal{Z}\}$ is

$$\underset{z \in \mathcal{Z}}{\mathrm{geomed}}\{z\} := \arg\min_{y} \sum_{z \in \mathcal{Z}} \|y - z\| \tag{5}$$

- With geometric median, the distributed SGD in (2) can be modified to its Byzantine attack resilient form as

$$x^{k+1} = x^k - \gamma^k \cdot \underset{w \in \mathcal{W}}{\mathrm{geomed}}\{m_w^k\} \tag{6}$$

Impact of Gradient Noise of SGD



- While geometric median can resist Byzantine attacks to some extent, its performance strongly depends on the variance of gradient
- Smaller noise may make the same Byzantine attacks less effective

Revisiting Variance Reduction Techniques

- Our key idea is to reduce the variance of stochastic gradients in order to enhance robustness to Byzantine attacks
- An effective approach to alleviating stochastic gradient noise in SGD is through variance reduction
- Existing variance reduction techniques in stochastic optimization include mini-batch, SAG, SVRG, SAGA, SDCA, SARAH, Katyusha, to list a few

Revisiting Distributed SAGA with Mean

- In SAGA [3], worker w stores a gradient table $f'_{w,j}(\phi^k_{w,j})$ locally

$$\phi^0_{w,j} = x^0, \quad \phi^{k+1}_{w,j} = \begin{cases} \phi^k_{w,j}, & j \neq i^k_w \\ x^k, & j = i^k_w \end{cases} \tag{7}$$

- $f'_{w,j}(\phi^k_{w,j})$ refers to the previously stored stochastic gradient of the j-th data sample prior to slot k on worker w
- Worker w will send the corrected stochastic gradient g^k_w to center

$$g^k_w := f'_{w,i^k_w}(x^k) - f'_{w,i^k_w}(\phi^k_{w,i^k_w}) + \frac{1}{J}\sum_{j=1}^{J} f'_{w,j}(\phi^k_{w,j})$$

- The model update of SAGA is hence

$$x^{k+1} = x^k - \gamma \cdot \frac{1}{W}\sum_{w=1}^{W} g^k_w \tag{8}$$

where $\gamma > 0$ is the constant step size

Byrd-SAGA: Distributed SAGA with Geometric Median

- Let m_w^k denote the message from worker w at time slot k

$$m_w^k = \begin{cases} g_w^k, & w \notin \mathcal{B} \\ *, & w \in \mathcal{B} \end{cases} \tag{9}$$

- Similar to distributed SGD, we can equip distributed SAGA with geometric median to defend against Byzantine attacks

$$x^{k+1} = x^k - \gamma \cdot \underset{w \in \mathcal{W}}{\mathrm{geomed}}\{m_w^k\} \tag{10}$$

- This leads to the proposed **<u>By</u>zantine-attack <u>r</u>esilient <u>d</u>istributed form of SAGA** (**Byrd-SAGA**)

Convergence Analysis: Assumptions

Assumption 1. *(Strong convexity and Lipschitz continuous gradients)*

The function f is μ-strongly convex and has L-Lipschitz continuous gradients.

Assumption 2. *(Bounded outer variation)*

For any $x \in \mathbb{R}^p$, variation of the aggregated gradients at the honest workers with respect to the overall gradient is upper-bounded by $E_{w \notin \mathcal{B}} \|f'_w(x) - f'(x)\|^2 \le \delta^2$.

Assumption 3. *(Bounded inner variation)*

For every honest worker w and any $x \in \mathbb{R}^p$, the variation of its stochastic gradients with respect to its aggregated gradient is upper-bounded by $E_{i_w^k} \|f'_{w,i_w^k}(x) - f'_w(x)\|^2 \le \sigma^2, \forall w \notin B$.

Convergence Analysis: SAGA vs SGD

Theorem 1: SAGA with geometric median (Byrd-SAGA)

Under Assumptions 1 and 2, if the number of Byzantine attackers satisfies $B < \frac{W}{2}$ and the step size satisfies $\gamma \leq \frac{\mu}{8J^2 C_\alpha L^2}$, then for Byrd-SAGA with geometric median aggregation, it holds that

$$E\|x^k - x^*\|^2 \leq O((1 - \frac{\gamma\mu}{2})^k) + O(\delta^2). \tag{11}$$

Theorem 2: SGD with geometric median

Under Assumptions 1, 2 and 3, if the number of Byzantine attackers is $B < \frac{W}{2}$ and the step size satisfies $\gamma < \frac{L}{2\mu^2}$, then for Byzantine attack resilient SGD with geometric median aggregation, it holds that

$$E\|x^k - x^*\|^2 \leq O((1 - \gamma\mu)^k) + O(\sigma^2 + \delta^2). \tag{12}$$

Asymptotic error: from $O(\sigma^2 + \delta^2)$ to $O(\delta^2)$

- ℓ_2-regularized logistic regression on IJCNN1 dataset
- $W - B = 50$ honest workers and $B = 20$ Byzantine workers (if any)
- Benchmark algorithms: SGD, mini-Batch SGD with batch size 50 (BSGD), SAGA equipped with mean or geometric median
- Attacks
    - Gaussian attacks: m_w^k drawn from a Gaussian distribution with mean $\frac{1}{W-B}\sum_{w'\notin\mathcal{B}} m_{w'}^k$ and variance 30
    - Sign-flipping attacks: $m_w^k = u \cdot \frac{1}{W-B}\sum_{w'\notin\mathcal{B}} m_{w'}^k$, where the magnitude $u = -3$
    - Zero gradient attacks: $m_w^k = -\frac{1}{B}\sum_{w'\notin\mathcal{B}} m_{w'}^k$

Numerical Experiments: ℓ_2-regularized Logistic Regression

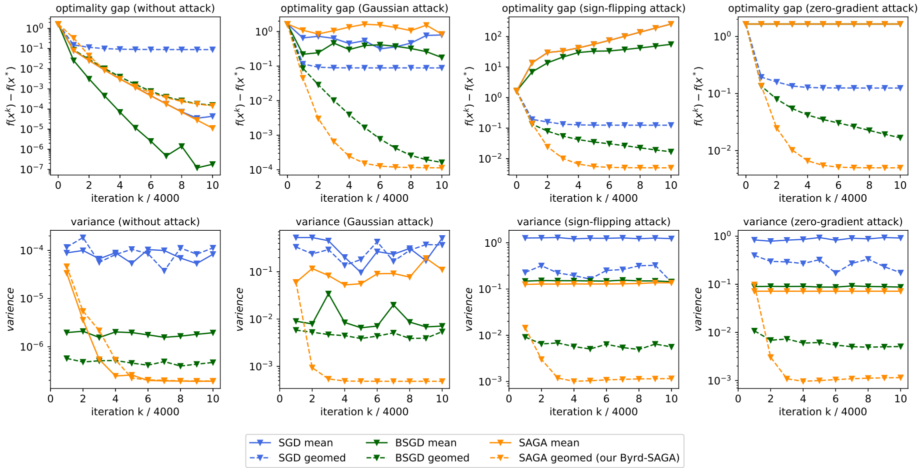


Figure 2: The first row shows the optimal gap $f(x^k) - f(x^*)$ and the second row shows the variance of honest workers

Conclusions

- For the first time, we reveal the relation between gradient noise and resilience to Byzantine attacks
- We design a novel algorithm, Byrd-SAGA, to defend against Byzantine attacks in federated learning
- Byrd-SAGA reduces the gradient noise in SGD and achieves better resilience to Byzantine attacks

Thanks

References

J. Konecny, H. B. McMahan, D. Ramage, and P. Richtarik, "Federated optimization: Distributed machine learning for on-device intelligence," arXiv Preprint arXiv:1610.02527, 2016.

Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proceedings of SIGMETRICS*, Phonenix, Arizona, USA, Jun. 2019.

A. Defazio, F. R. Bach, and S. Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," *Proceedings of NIPS*, Montreal, Canada, Dec. 2014.