

# Efficient Byzantine-Resilient Stochastic Gradient Descent

Kaiyun Li<sup>1,2</sup>, Xiaojun Chen<sup>1,✉</sup>, Ye Dong<sup>1,2</sup>, Peng Zhang<sup>3</sup>, Dakui Wang<sup>1</sup>, Shuai Zeng<sup>1</sup>

<sup>1</sup>Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China.

<sup>2</sup>School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China.

<sup>3</sup>Guangzhou University, Guangzhou, China.

{likaiyun, chenxiaojun, dongye, wangdakui, zengshuai}@ie.ac.cn  
p.zhang@gzhu.edu.cn

## Abstract

*Distributed Learning* often suffers from Byzantine failures, and there have been a number of works studying the problem of distributed stochastic optimization under Byzantine failures, where only a portion of workers, instead of all the workers in a distributed learning system, compute stochastic gradients at each iteration. These methods, albeit workable under Byzantine failures, have the shortcomings of either a sub-optimal convergence rate or high computation cost. To this end, we propose a new Byzantine-resilient stochastic gradient descent algorithm (BrSGD for short) which is provably robust against Byzantine failures. BrSGD obtains the optimal statistical performance and efficient computation simultaneously. In particular, BrSGD can achieve an order-optimal statistical error rate for strongly convex loss functions. The computation complexity of BrSGD is  $O(md)$ , where  $d$  is the model dimension and  $m$  is the number of machines. Experimental results show that BrSGD can obtain competitive results compared with non-Byzantine machines in terms of effectiveness and convergence.

## 1 Introduction

In the fields of recommendation systems, natural language processing, and computer vision, it is often the case that we need to build complex models from large-scale datasets. As the available data for training models continuously grow, it is urgent to use distributed learning to digest these large-scale data. In a distributed learning system, robustness and security issues have become a major concern. In particular, individual computing units, i.e., *worker machines*, may exhibit abnormal behavior due to data corruption, hardware/software malfunction, communication delay, etc. This abnormal (even worse adversarial) behavior is typically modeled as *Byzantine failure* [Lamport *et al.*, 1982]. It is well-known that a single Byzantine-faulty machine can arbitrarily skew standard distributed learning algorithms based on naive gradients aggregation, e.g., an average of the gradients collected from all the workers [Blanchard *et al.*, 2017]. Moreover, Byzantine failure is exacerbated in Federated Learning (FL), where training

data resides at autonomous worker machines, and a central server facilitates the learning process [Konečný *et al.*, 2015; McMahan *et al.*, 2017; Kairouz and McMahan, 2021]. The behavior of a worker machine under Byzantine failure is often unpredictable, and even becomes susceptible to malicious and coordinated attacks [Aono *et al.*, 2017; Zhao *et al.*, 2020]. Therefore, it is increasingly important to develop robust distributed algorithms in the adversarial setting.

Existing works on [Blanchard *et al.*, 2017; Feng *et al.*, 2014; Chen *et al.*, 2017; Xie *et al.*, 2018; Yin *et al.*, 2018] suffer from the shortcomings of either a sub-optimal statistical guarantees or high computation cost. They are inapplicable to distributed settings where. As a result, we wish to develop distributed learning algorithms that can achieve two objectives simultaneously. The first objective is **Byzantine-resilience** which refers to achieving the best performance even though a relatively large fraction of the workers are Byzantine. The second objective is **high computation efficiency** which refers to preserving as much as possible the run-time speedup by using distributing computation across multiple workers.

To deal with Byzantine failures, an intuitive solution is to use coordinate-wise median [Yin *et al.*, 2018] and its variants [Chen *et al.*, 2017; Xie *et al.*, 2018]. However, these methods independently consider each single exception dimension, which requires heavy computation cost when there are a large number of dimensions. For example, to compute the median value of a single exception dimension, the easiest way is to apply a sorting algorithm to the dimension, which yields a time complexity of  $O(dm \log m)$ . This time complexity is unaffordable when  $d$  is very large.

In this paper, we define a new metric for the gradient vector to evaluate whether it is a Byzantine worker. Our key insight is to directly use the  $\ell_1$ -norm constraint to eliminate abnormal gradients, which can greatly simplify the calculation. Specifically, let  $\mathbf{g}^i$  be a stochastic gradient computed by a worker  $i \in [m]$ , then  $\mathbf{g}^i$  shall concentrate around  $\mathbf{g}^{med}$ , where  $\mathbf{g}^{med}$  is the median value of  $\mathbf{g}^i, i \in [m]$ . In other words, if  $\|\mathbf{g}^i - \mathbf{g}^{med}\|_1 > 2\mathfrak{T}$ , where  $\mathfrak{T} > 0$  is a hyperparameter, then we can declare worker  $i$  as Byzantine.

A challenge to the  $\ell_1$ -norm method is that some Byzantine workers may hide without violating the constraint. Then, we further define a score for each gradient. As shown in Fig.1, the master machine needs to collect the gradients that are flat

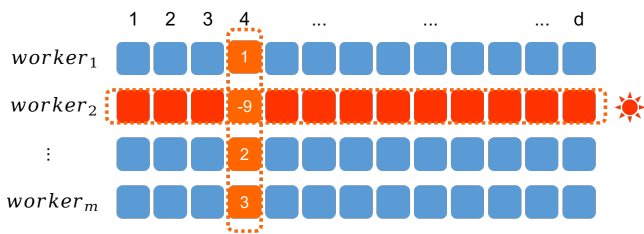


Figure 1: The gradients are collected and combined into a matrix. Taking a column of the matrix, the values uploaded by the normal workers in this column has little difference.

tened into a  $d$ -dimensional vector from all the workers, and then align these vectors into a new matrix by columns. As shown in Fig.2, when taking one column of the matrix, we can observe that a subset with a smaller number of elements must deviate heavier from the global mean value than a larger number of elements. Based on the observations that the *honest majority* and the values uploaded by the normal workers are with little difference, a subset with a smaller number of elements is more likely to contain Byzantine workers. Therefore, we can judge the outliers according to each dimension of the gradient and count the number of outliers as a score. We keep the  $\beta$ -fraction gradients with the highest scores.

Noted that Byzantine workers may be falsely labeled as good ones. Fortunately, the convergence rates of our method are not impacted significantly. Especially, the subroutine for calculating the score has only one round of comparison and one round of averaging. Thus, it is computationally efficient, almost the same as a naive method implemented through averaging [Polyak and Juditsky, 1992]. The contributions of the paper can be summarized as follows:

- We present a new robust aggregation rule for distributed synchronous Stochastic Gradient Descent algorithm BrSGD in an adversarial setting. The new algorithm can handle Byzantine resilience. Moreover, the computation complexity of the proposed algorithm is  $O(md)$ .
- We theoretically analyze the statistical error rates of our method on strongly convex loss functions. In particular, our algorithm can achieve an order-optimal statistical error rate for strongly convex losses.
- We also demonstrate the convergence of the proposed algorithm by conducting empirical studies. The experimental results on four types of Byzantine attacks match the results of none-Byzantine machines in terms of effectiveness and convergence.

The rest of this paper is organized as follows. Related work is summarized in Section 2. In Section 3, we formulate the problem of this work. Section 4 presents our method and analyzes the statistical error rates. In Section 5, we present the empirical evaluations, and we conclude this work in Section 6.

## 2 Related Work

Recently there have been a large number of works studying the problem of Byzantine stochastic optimization [Blan-

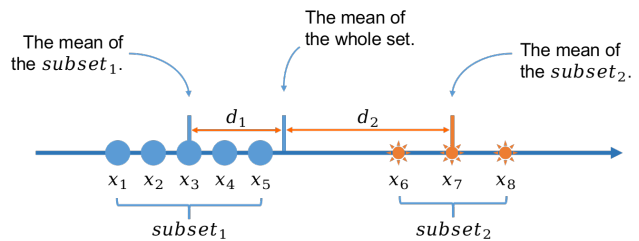


Figure 2: The goal of Byzantine workers (subset<sub>2</sub>) is to make the mean value of the set  $\{x_1, \dots, x_8\}$  deviating from the *real* mean of the subset<sub>1</sub> as much as possible. However, based on the *honest majority* assumption, the number of elements in subset<sub>1</sub> is greater than subset<sub>2</sub>, i.e.,  $d_1 < d_2$ . Thus, we can infer whether the Byzantine worker is included according to the number of elements in the two subsets divided by the mean value.

chard *et al.*, 2017; Feng *et al.*, 2014; Su and Vaidya, 2016; Chen *et al.*, 2017; Alistarh *et al.*, 2018; Su and Vaidya, 2019; Xie *et al.*, 2018; Yin *et al.*, 2018; Su and Xu, 2019; Chen *et al.*, 2018]. Here, we compare existing Byzantine-robust distributed learning algorithms that are most relevant to our work and summarize the comparison in this section.

According to different ideas, existing algorithms can be roughly divided into the following three categories, i.e., *Median-based* method, *Anomaly-detection-based* method, and *Redundant-gradients-based* method.

The median-based method is used in the context of Byzantine-tolerant distributed learning in papers [Blanchard *et al.*, 2017; Feng *et al.*, 2014]. The median-based method can be further sub-divided into *Geometric Median* [Chen *et al.*, 2017] and *Marginal Median* [Xie *et al.*, 2018; Yin *et al.*, 2018]. For geometric median, the  $(1 + \epsilon)$ -approximate geometric median can be computed in  $O(dm \log^3 \frac{1}{\epsilon})$  [Cohen *et al.*, 2016]. An algorithm called the *selection algorithm* [Blum *et al.*, 1973] with average time complexity  $O(m)$  ( $O(m^2)$  is the worst case) to obtain median values. Anomaly detection is performed by running the  $T$  iterations [Alistarh *et al.*, 2018; Allen-Zhu *et al.*, 2021]. Although computation complexity is considerable for anomaly detection methods, each gradient needs to be checked, and inner product operations are also involved. These operations are very time-consuming in practical. In these algorithms, the most time-consuming is via redundant gradients methods (ideas from coding theory) [Chen *et al.*, 2018].

The closest literature is the work of Yin *et al.* [Yin *et al.*, 2018]. They consider a similar Byzantine model, but for *gradient descent* (GD). Reference [Blanchard *et al.*, 2017] proposes a general Byzantine-resilient gradient aggregation rule called *Krum*. This rule has expensive complexity  $O(m^2(d + \log m))$  and does not provide a characterization of the statistical errors. Recent work [Chen *et al.*, 2017] only applies in the strongly convex setting and is sub-optimal convergence rate. In contrast, our algorithm is efficient ( $O(md)$ ) in computational complexity and achieves order-optimal statistical error rates for strongly convex losses.

---

**Algorithm 1: Byzantine-resilient SGD (BrSGD)**

---

**Input:** initialize parameters  $\mathbf{w}^0 \in \mathcal{W}$ , step-size  $\eta$ , hyperparameters  $0 < \beta < \frac{1}{2}$ , thresholds  $\mathfrak{T} > 0$ , and iterations  $T$ ;

*Master machine:* send  $\mathbf{w}^0$  to all the worker machines;

**for**  $t = 0, 1, \dots, T - 1$  **do**

**for all**  $i \in [m]$  **do**

*Worker machine*  $i$ : compute local gradient;

$$\mathbf{g}^i(\mathbf{w}^t) = \begin{cases} \nabla \ell_i(\mathbf{w}^t), & i \in [m] \setminus \mathcal{B}, \\ *, & i \in \mathcal{B}. \end{cases}$$

    upload  $\mathbf{g}^i(\mathbf{w}^t)$  to master machine;

**end**

*Master machine:* compute aggregate gradient

$$\mathbf{g}(\mathbf{w}^t) \leftarrow \mathcal{A}_{\beta, \mathfrak{T}}(\{\mathbf{g}^i(\mathbf{w}^t) : i \in [m]\}),$$

  and then send  $\mathbf{g}(\mathbf{w}^t)$  to all worker machines;

*Worker machine:* update model parameter

$$\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta \mathbf{g}(\mathbf{w}^t)$$

**end**

---

### 3 Preliminaries

First of all, we introduce the used notations. A vector is denoted as  $\mathbf{w}$ , and  $w_i$  denotes the  $i$ -th element of  $\mathbf{w}$ . Matrices are denoted as  $\mathbf{M}$ . We let  $M_{i,j}$  denote its entry at location  $(i, j)$ ,  $\mathbf{M}_{i,\cdot}$  denote its  $i$ -th row, and  $\mathbf{M}_{\cdot,j}$  denote its  $j$ -th column.  $[N]$  denotes  $\{1, 2, \dots, N\}$ . For any differentiable function  $f$ , we denote its partial derivative for the  $i$ -th argument by  $\partial_i f$ . We index the set of worker machines by  $[m]$ , and denote the set of Byzantine machines by  $\mathcal{B} \in [m]$  (thus  $|\mathcal{B}| = \alpha m$ ).

We consider a distributed computation model with one master machine and  $m$  worker machines and that an  $\alpha$ -fraction of the workers may be Byzantine (where  $\alpha < 1/2$ ). Each worker machine stores  $n$  samples, each of which is sampled independently from an unknown distribution  $\mathcal{D}$ .  $\ell_i(\mathbf{w}; \mathbf{x})$  be a loss function of  $i$ -th worker machine associated with parameter vector  $\mathbf{w} \in \mathcal{W} \in \mathbb{R}^d$  and the sample  $\mathbf{x}$ , where  $\mathcal{W}$  is the parameter space. Denoted the  $j$ -th data on the  $i$ -th worker machine by  $\mathbf{x}^{i,j}$ , and  $F_i(\mathbf{w}) := \frac{1}{n} \sum_{j=1}^n \ell_i(\mathbf{w}, \mathbf{x}^{i,j})$  denote the empirical risk function for the  $i$ -th worker. Our goal is to learn a model defined by the parameter that minimizes the population loss:

$$\mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathbb{R}^d} F(\mathbf{w}), \quad (1)$$

where

$$F(\mathbf{w}) := \mathbb{E}[F_i(\mathbf{w})]. \quad (2)$$

The parameter space  $\mathcal{W}$  is compact with diameter  $D$ , i.e.,  $\|\mathbf{w} - \mathbf{w}'\|_2 \leq D, \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}$ .

### 4 Our Method

This section presents our Byzantine-Resilient stochastic gradient descent algorithm (named BrSGD) and briefly summarizes our convergence results on its performance.

#### 4.1 BrSGD Algorithm

The master machine broadcasts the initialized model, and all the workers receive and accept it as the local initialized model. Then, the master and the workers update the model iteratively. At each iteration, the normal workers compute the gradients of their local loss functions and then upload to the master. The Byzantine workers may send any messages. Next, the master performs a robust aggregator to compute gradients for model updating and send the gradient to all workers. After receiving the gradient from the master, the workers update its parameters in the way of gradient descent and moves into the next iteration until the whole algorithm is completed. The formulation is illustrated in Algorithm 1.

In the following context, we focus on the aggregation method construction. Intuitively, the aggregation rule should return a vector  $\mathbf{g}$  that is not too far from the *real* gradient  $\nabla F(\mathbf{w})$ . More precisely,  $\mathbf{g}$  should approximate the steepest direction of the loss function being optimized. This is expressed as upper bound  $\|\mathbf{g}^i - \mathbf{g}^{med}\|_1 \leq 2\mathfrak{T}$  on worker's gradient vector  $\mathbf{g}^i$  (*Constraint 1*). If the gradient satisfies *Constraint 1*, we add it to the candidates  $\mathcal{C}_1$ . To further eliminate the impacts of potential anomaly values in candidates, we set a score to each  $\mathbf{g}^i$  and keep the  $\beta$ -fraction gradients with larger scores (*Constraint 2*). Concretely, the master collect the gradients that are flattened into a  $d$ -dimensional vector from all workers and then align them into a new matrix  $\mathbf{G}$  by column. Then, we utilize the observation (cf. Section 1) to mark the anomaly values of  $\mathbf{G}$  column-wise. We divide one column of data into two subsets using their mean and set the subset with a large number of elements to 1, otherwise 0. Finally, a scoring matrix  $\mathbf{M}$  composed of 0 and 1 will be generated, and the result of adding the rows will be used as a score for the corresponding gradient. For each worker  $i$ , we define the score  $s_i = \sum_{j \in [d]} M_{i,j}, i \in [m]$ . For the details, please see Algorithm 2.

#### 4.2 Convergence Analysis

In this part, we provide statistical guarantees on the error rates (defined as the distance between  $\mathbf{w}^T$  and the optimal solution  $\mathbf{w}^*$ ) of the algorithms. Throughout we assume that the loss function  $F(\mathbf{w})$  is smooth.

**Assumption 1.** *In each iteration  $t$ , each normal worker machine  $i \in [m]$  gives back a vector  $\nabla \ell_i(\mathbf{w}^t) \in \mathbb{R}^d$  satisfying  $\|\nabla \ell_i(\mathbf{w}^t) - \nabla F(\mathbf{w}^t)\|_1 \leq \mathcal{V}$ .*

One can instead assume  $Pr[\|\nabla \ell_i(\mathbf{w}^t) - \nabla F(\mathbf{w}^t)\|_1 \geq t] \leq 2 \exp(-t^2/2\mathcal{V}^2)$  and the results of this paper continue to hold up to logarithmic factors. But we assumes  $F(\mathbf{w}_t)$  is Lipschitz continuous, which implies  $\|\nabla \ell_i(\mathbf{w}^t) - \nabla F(\mathbf{w}^t)\|_1$  is bounded. To present the simplest theory, we assumed it is bound with probability 1.

Firstly, we introduces Lemma 1, which is used to bound the output of algorithm 2.

**Lemma 1.** *In our algorithm, suppose that the one dimensional samples on all the normal worker machines are independent and identically distributed (i.i.d). Satisfy  $v$ -sub-exponential with mean  $\mu$ . For any  $t \geq 0$  and  $s \geq 0$ , let  $|\frac{1}{|\mathcal{C}_1 \cap \mathcal{C}_2}| \sum_{i \in [m] \setminus \mathcal{B}} \mathbf{g}^i - \mu| \leq t$  and  $\max_{i \in [m] \setminus \mathcal{B}} |\mathbf{g}^i - \mu| \leq s$ .*

---

**Algorithm 2:** Aggregator  $\mathcal{A}_{\beta, \mathfrak{T}}(\{g^i(\mathbf{w}^t) : i \in [m]\})$ 

---

**Inputs:** gradient set  $\{g^i(\mathbf{w}^t), i \in [m]\}$ , hyperparameters  $0 < \alpha < \beta \leq \frac{1}{2}$ , and thresholds  $\mathfrak{T} > 0$ ;

**Return:** aggregation gradient  $g(\mathbf{w}^t)$ ;

$\mathbf{a} = 0, \mathbf{s} = 0, \mathbf{n} = 0, \mathbf{M} = \mathbf{0}$ ; # Initialize the mean vector  $\mathbf{a}$ , score vector  $\mathbf{s}$ ,  $\ell_1$ -norm vector  $\mathbf{n}$ , and score matrix  $\mathbf{M}$ ;

**for**  $c = 0, 1, \dots, d$  **do**

**for**  $r = 0, 1, \dots, m$  **do**

$\mathbf{a}_c = \mathbf{a}_c + \mathbf{a}_r$ ;

**end**

$\mathbf{a}_c = \mathbf{a}_c / m$ ; # calculate the average value of the arranged gradient matrix by column;

**end**

**for**  $c = 0, 1, \dots, d$  **do**

$\text{counter} = 0$ ;

**for**  $r = 0, 1, \dots, m$  **do**

**if**  $g_c^r(\mathbf{w}^t) \geq \mathbf{a}_c$  **then**

$\mathbf{M}_{r,c} = 1$ ;

$\text{counter} = \text{counter} + 1$ ;

**end**

**else**

$\mathbf{M}_{r,c} = 0$ ;

**end**

**end**

**if**  $\text{counter} < m/2$  **then**

$\mathbf{M}_{\cdot,d} = -\mathbf{M}_{\cdot,d}$ ; # obtain the score matrix based on observation 1;

**end**

$\mathbf{s} = \mathbf{s} + \mathbf{M}_{\cdot,d}^T$ ; # calculate the score for each gradient;

**end**

$\mathbf{g}_{med} = \text{median}\{g^i(\mathbf{w}^t) : i \in [m]\}$ ;

$\mathcal{C}_1 = \{i : \|g^i(\mathbf{w}^t) - \mathbf{g}_{med}\|_1 \leq 2\mathfrak{T}\}$ ; # Constraint 1;

$\mathcal{C}_2 = \{i : \text{Top}_{max}^\beta(\mathbf{s}_i), i \in [m]\}$ ; # Constraint 2;

$g(\mathbf{w}^t) = \text{mean}\{g^i(\mathbf{w}^t) : i \in \mathcal{C}_1 \cap \mathcal{C}_2\}$ ;

---

If we set  $\mathfrak{T} = s \leq \mathcal{V}$ , then we have

$$|\mathcal{A}_{\beta, \mathfrak{T}}(\{g^i : i \in [m]\}) - \mu| \leq t + 3\beta s \quad (3)$$

*Proof.* To simplify notation, we define  $\mathcal{M} = [m] \setminus \mathcal{B}$  as the set of all normal worker machines,  $\mathcal{C} = \{\mathcal{C}_1 \cap \mathcal{C}_2\} \subseteq [m]$  as the set of candidates, and  $\mathcal{R} \subseteq [m]$  as the set of all removed machines. The estimator simply computes

$$\mathcal{A}_{\beta, \mathfrak{T}}(\{g^i : i \in [m]\}) = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} g^i \quad (4)$$

Condition 1 and Assumption 1 ensures that the boundary of each dimension of the gradient in the candidate set does not

exceed  $2s$ . Then, We have

$$\begin{aligned} |\mathcal{A}_{\beta, \mathfrak{T}}(\{g^i : i \in [m]\}) - \mu| &= \left| \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} g^i - \mu \right| \\ &= \frac{1}{|\mathcal{C}|} \left| \sum_{i \in \mathcal{M}} (g^i - \mu) - \sum_{i \in \mathcal{M} \cap \mathcal{R}} (g^i - \mu) + \sum_{i \in \mathcal{B} \cup \mathcal{C}} (g^i - \mu) \right| \\ &\leq \frac{1}{|\mathcal{C}|} ((1 - \alpha)mt + \beta ms + 2\alpha ms) \\ &\leq t + 3\beta s \end{aligned} \quad (5)$$

where the last inequality is by  $\beta \geq \alpha$ . We directly obtain the desired result.  $\square$

**Assumption 2** (Smoothness of  $\ell$  and  $F$ ). For any  $\mathbf{x} \in \mathcal{X}$ , the partial derivative of  $\ell(\cdot; \mathbf{x})$  with respect to the  $k$ -th coordinate of its first argument, denoted by  $\partial_k \ell(\cdot; \mathbf{x})$ , is  $L_k$ -Lipschitz continuity for each  $k \in [d]$ , and the function  $\ell(\cdot; \mathbf{x})$  is  $G$ -Lipschitz smoothness. Let  $L := \sqrt{\sum_{k=1}^d L_k^2}$ . Also assume that the population loss function  $F(\cdot)$  is  $G_F$ -Lipschitz smoothness. It is easy to see that  $G_F \leq G \leq L$ .

### Strongly Convex Losses

We consider the case where the population loss function  $F(\cdot)$  is strongly convex.

**Theorem 1.** suppose that Assumption 2 hold,  $F(\cdot)$  is  $\lambda_F$ -strongly convex, and  $\alpha \leq \frac{1}{2} - \epsilon$  for some  $\epsilon > 0$ . In addition, we assume that  $\partial_k \ell(\mathbf{w}; \mathbf{x})$  is  $v$ -sub-exponential for any  $k \in [d]$  and  $\mathbf{w} \in \mathcal{W}$ . Choose step-size  $\eta = 1/G_F$ . Then, with probability at least  $1 - \frac{4d}{(1+nmLD)^d}$ , after  $T$  parallel iterations, we have

$$\|\mathbf{w}^T - \mathbf{w}^*\|_2 \leq \left(1 - \frac{\lambda_F}{G_F + \lambda_F}\right)^T \|\mathbf{w}^0 - \mathbf{w}^*\|_2 + \frac{2}{\lambda_F} \Delta$$

where

$$\Delta = \tilde{O}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{nm}}\right)$$

Note that this paper [Yin *et al.*, 2018] provide a lower bound showing that the error rate of any algorithm is  $\tilde{\Omega}\left(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}}\right)$ . Therefore, our algorithm achieved order-optimal statistical error rate.

*Proof.* The proof is essentially the same as [Yin *et al.*, 2018]. Therefore, we omit the details of the analysis here. Choosing

$$t = v \max \left\{ \frac{8d}{nm} \log(1 + nmLD), \sqrt{\frac{8d}{nm} \log(1 + nmLD)} \right\},$$

and

$$s = v \max \left\{ \frac{4}{n} (d \log(1 + nmLD) + \log m), \sqrt{\frac{4}{n} (d \log(1 + nmLD) \log m)} \right\}.$$

By using the simple algebra, we have

$$\Delta = \tilde{O}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{nm}}\right)$$

$\square$

Algorithms	Ours				Median			Mean			Krum		
	$\alpha$	0	10%	25%	50%	10%	25%	50%	10%	25%	50%	10%	25%
Gaussian	84.16	83.42	84.56	83.29	83.73	85.40	86.78	9.93	9.93	9.93	82.04	79.56	79.62
Model Negation	85.87	83.89	84.07	82.13	83.29	84.47	82.84	10.84	10.84	10.84	80.44	81.18	79.36
Gradient Scale	85.29	84.33	83.67	82.87	81.00	61.13	9.80	85.87	82.47	60.67	77.20	81.33	70.76
Label Inverse	85.11	82.93	82.49	83.93	84.91	86.00	82.09	72.71	74.93	80.11	81.82	78.56	68.76

Table 1: Test results of LeNet on FashionMNIST in terms of accuracy using stochastic gradient descent.

## 5 Experiments

We implemented BrSGD in Python3. Our experiments are executed on Intel® Xeon® CPU E5-2650 v3@ 2.30GHz servers with 64GB RAM. We simulate  $m = 20$  clients. The model is learned atop Pytorch v1.6.0 equipped with CUDA v10.2 and one 12G memory TITAN Xp GPU. the C-S connection is over WAN with 50Mbps bandwidth and 50ms RTT.

### 5.1 Experimental settings

We conduct experiments to show the Byzantine resilience in the following four classic adversarial settings:

- We consider the attackers that replace some of the gradient vectors with Gaussian random vectors with zero mean and isotropic covariance matrix with standard deviation 200. We refer to this kind of attack as *Gaussian Attack*.
- For each Byzantine gradient vector, the gradient is replaced by the negative sum of all the correct gradients, scaled by a large constant (1e10 in the experiments). We call this attack *Model Negation Attack*.
- We generate the Byzantine machines in the following way: we replace every training label  $y$  on these machines with  $9 - y$ , e.g., 1 is replaced with 8, 7 is replaced with 2, etc., and the Byzantine machines compute gradients on these data. Denote this kind of attack by *Label Shift*.
- Named after *Gradient Scale Attack*, an attacker replaces some of the gradient vectors with scaled by a constant (1e10 in the experiments).

We instantiate  $m = 20$  workers and one master node in experiments, and train a convolutional neural network model (LeNet [LeCun *et al.*, 1998]) on the FashionMNIST dataset [Xiao *et al.*, 2017] using mini-batch stochastic gradient descent to compare the test accuracies in those above four adversarial settings. We compare accuracy and convergence with *Mean* (find the average of all gradients), *Krum*[Blanchard *et al.*, 2017], and coordinate-wise median[Yin *et al.*, 2018].

### 5.2 Experimental results

As is shown in Figure 3, we observe that even with a single attacker performing a *Gaussian Attack* or *Model Negation Attack*, using the *Mean* method directly is devastating (a0,a1). Although the median-based method has high accuracy, the convergence is very slow (b1,b3). *Krum* and our method have good accuracy, but the effect of *Krum* will drop precipitously when the number of Byzantine worker machines is

large ( $\alpha = 50\%$ ,c3). In contrast, we can see that our algorithm has the same convergence as the average method without an attack ( $\alpha = 0$ ). When the  $\alpha = 50\%$ , the model still achieves high accuracy. We also include the full test accuracy comparison table in Table 1.

In conclusion, the distributed gradient descent algorithm suffers from severe performance loss in adversarial settings. Moreover, show our algorithm can indeed defend against Byzantine failures. Compared with the median-based and *Krum* methods, our algorithm has excellent accuracy and faster convergence for the four attacks.

## 6 Conclusion

In this paper, we study a new distributed stochastic optimization algorithm in an adversarial setting with the purpose of obtaining the optimal statistical results and computation efficiency simultaneously. Based on the honest-majority assumption, we propose a new stochastic gradient descent algorithm BrSGD. We show that the method can achieve an order-optimal  $\tilde{O}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{nm}}\right)$  for strongly convex losses, and the computation complexity of our algorithm is  $O(md)$ . Moreover, we conduct extensive experiments to show that our method outperforms the state-of-the-art methods in terms of effectiveness and convergence.

## References

- [Alistarh *et al.*, 2018] Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [Allen-Zhu *et al.*, 2021] Zeyuan Allen-Zhu, Faeze Ebrahimiaghazani, Jerry Li, and Dan Alistarh. Byzantine-resilient non-convex stochastic gradient descent. In *International Conference on Learning Representations*, 2021.
- [Aono *et al.*, 2017] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 13(5):1333–1345, 2017.
- [Blanchard *et al.*, 2017] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant

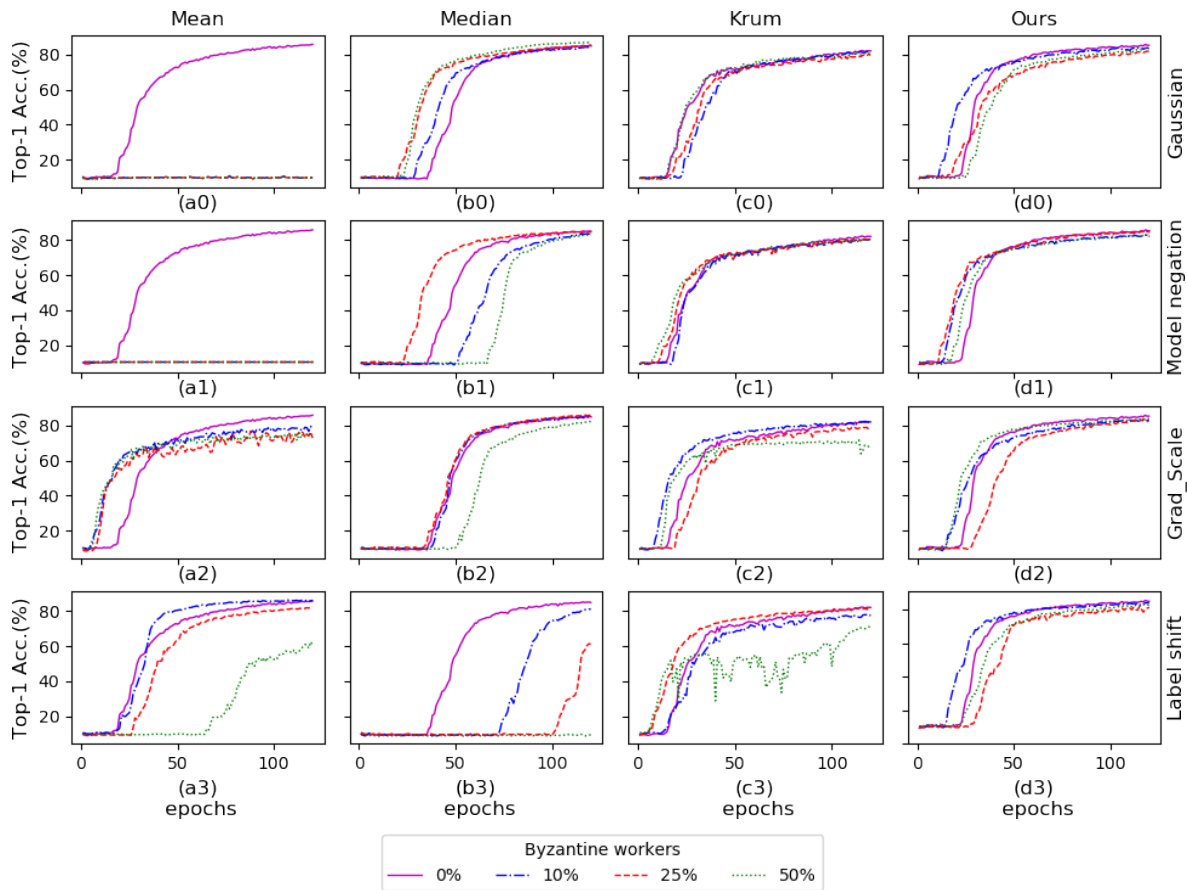


Figure 3: The results *w.r.t.* accuracy vary with the epochs under Byzantine attacks, where we set step-size  $\eta = 0.03$  for the LeNet on FashionMNIST. Curves correspond to losses, and columns correspond to different aggregation methods. Byzantine workers perform four attacks with 10%, 25%, and 50% attackers, respectively. We set  $\beta = 1/2$  for our algorithm.

gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 118–128, 2017.

[Blum *et al.*, 1973] Manuel Blum, Robert W. Floyd, Vaughan R. Pratt, Ronald L. Rivest, and Robert Endre Tarjan. Time bounds for selection. *J. Comput. Syst. Sci.*, 7(4):448–461, 1973.

[Chen *et al.*, 2017] Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017.

[Chen *et al.*, 2018] Lingjiao Chen, Hongyi Wang, Zachary Charles, and Dimitris Papailiopoulos. Draco: Byzantine-resilient distributed training via redundant gradients. In *In-*

*ternational Conference on Machine Learning*, pages 903–912. PMLR, 2018.

[Cohen *et al.*, 2016] Michael B Cohen, Yin Tat Lee, Gary Miller, Jakub Pachocki, and Aaron Sidford. Geometric median in nearly linear time. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 9–21, 2016.

[Feng *et al.*, 2014] Jiashi Feng, Huan Xu, and Shie Mannor. Distributed Robust Learning. *arXiv e-prints*, page arXiv:1409.5937, September 2014.

[Kairouz and McMahan, 2021] Peter Kairouz and H. Brendan McMahan. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1):–, 2021.

- [Konečný *et al.*, 2015] Jakub Konečný, Brendan McMahan, and Daniel Ramage. Federated optimization: Distributed optimization beyond the datacenter. *arXiv preprint arXiv:1511.03575*, 2015.
- [Lamport *et al.*, 1982] Leslie Lamport, Robert Shostak, and Marshall Pease. The byzantine generals problem. *ACM Trans. Program. Lang. Syst.*, 4(3):382–401, July 1982.
- [LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.
- [Polyak and Juditsky, 1992] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.
- [Su and Vaidya, 2016] Lili Su and Nitin H Vaidya. Fault-tolerant multi-agent optimization: optimal iterative distributed algorithms. In *Proceedings of the 2016 ACM symposium on principles of distributed computing*, pages 425–434, 2016.
- [Su and Vaidya, 2019] Lili Su and Nitin H Vaidya. Defending non-bayesian learning against adversarial attacks. *Distributed Computing*, 32(4):277–289, 2019.
- [Su and Xu, 2019] Lili Su and Jiaming Xu. Securing distributed gradient descent in high dimensional statistical learning. *Proc. ACM Meas. Anal. Comput. Syst.*, 3(1), March 2019.
- [Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv e-prints*, page arXiv:1708.07747, August 2017.
- [Xie *et al.*, 2018] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Generalized Byzantine-tolerant SGD. *arXiv e-prints*, page arXiv:1802.10116, February 2018.
- [Yin *et al.*, 2018] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659. PMLR, 2018.
- [Zhao *et al.*, 2020] B. Zhao, Konda Reddy Mopuri, and Hakan Bilen. Idlg: Improved deep leakage from gradients. *ArXiv*, abs/2001.02610, 2020.